Using the systematic nature of errors in NGS data to efficiently detect mutations : computational methods and application to early cancer detection

Tiffany Delhomme

N° d'ordre NNT : 2019LYSE1098

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED340
BIOLOGIE MOLÉCULAIRE INTÉGRATIVE ET CELLULAIRE (BMIC)

Spécialité de doctorat : Bioinformatique

Soutenue publiquement le 01/07/2019, par :
Tiffany Delhomme

Using the systematic nature of errors in NGS data to efficiently detect mutations: computational methods and application to early cancer detection.

Devant le jury composé de :

Maucort-Boulch Delphine, Professeur des Universités, Université Lyon 1	Président(e)
Nikolski Macha, Directrice de Recherche, Université de Bordeaux	Rapporteure
Thierry-Mieg Nicolas, Chargé de Recherche, Université Grenoble Alpes	Rapporteur
Blum Michael, Directeur de Recherche, Université Grenoble Alpes	Examinateur
Maucort-Boulch Delphine, Professeur des Universités, Université Lyon 1	Examinatrice
McKay James, Chercheur/Chef de groupe, IARC	Directeur de thèse
Foll Matthieu, Chercheur, IARC	Co-directeur de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directrice Générale des Services	Mme Dominique MARCHAND
COMPOSANTES SANTE	
Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur G.RODE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. X. PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT
COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE	
Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Abstract

Comprehensive characterization of DNA variations can help to progress in multiple cancer genomics fields. Next Generation Sequencing (NGS) is currently the most efficient technique to determine a DNA sequence, due to low experiment cost and time compared to the traditional Sanger sequencing. Nevertheless, detection of mutations from NGS data is still a difficult problem, in particular for somatic mutations present in very low abundance like when trying to identify tumor subclonal mutations, tumor-derived mutations in cell free DNA, or somatic mutations from histological normal tissue. The main difficulty is to precisely distinguish between true mutations from sequencing artifacts as they reach similar levels. In this thesis we have studied the systematic nature of errors in NGS data to propose efficient methodologies in order to accurately identify mutations potentially in low proportion. In a first chapter, we describe *needlestack*, a new variant caller based on the modelling of systematic errors across multiple samples to extract candidate mutations. In a second chapter, we propose two post-calling variant filtering methods based on new summary statistics and on machine learning, with the aim of boosting the precision of mutation detection through the identification of non-systematic errors. Finally, in a last chapter we apply these approaches to develop cancer early detection biomarkers using circulating tumor DNA.

Résumé

La caractérisation exaustive des variations de l'ADN peut aider à progresser dans de nombreux champs liés à la génomique du cancer. Le séquençage nouvelle génération (NGS en anglais pour *Next Generation Sequencing*) est actuellement la technique la plus efficace pour déterminer une séquence ADN, du aux faibles coûts et durées des expériences comparé à la méthode de séquençage traditionnelle de Sanger. Cependant, la détection de mutations à partir de données NGS reste encore un problème difficile, en particulier pour les mutations somatiques présentes en très faible abondance comme lorsque l'on essaye d'identifier des mutations sous-clonales d'une tumeur, des mutations dérivées de la tumeur dans l'ADN circulant libre, ou des mutations somatiques dans des tissus normaux. La difficulté principale est de précisement distinguer les vraies mutations des artefacts de séquençage du au fait qu'ils atteignent des niveaux similaires. Dans cette thèse nous avons étudié la nature systématique des erreurs dans les données NGS afin de proposer des méthodologies efficaces capables d'identifier des mutations potentiellement en faible abondance. Dans un premier chapitre, nous decrivons *needlestack*, un nouvel outil d'appel de variants basé sur la modélisation des erreurs systématiques sur plusieurs échantillons pour extraire des mutations candidates. Dans un deuxième chapitre, nous proposons deux méthodes de filtrage des variants basées sur des résumés statistiques et sur de l'apprentissage automatique, dans le but de d'améliorer la précision de la détection des mutations par l'identification des erreurs non-systématiques. Finalement, dans un dernier chapitre nous appliquons ces approches pour développer des biomarqueurs de détection précoce du cancer en utilisant l'ADN circulant tumoral.

À mon oncle, Franck...

Contents

Contents v

List of figures ix

List of Tables xi

1 Introduction 1

1.1 Scientific context . . . 3

1.1.1 Cancer . . . 3

1.1.2 Genomic variations causing cancer . . . 3

1.1.3 Next-generation sequencing . . . 8

1.2 Errors in NGS data . . . 13

1.2.1 Types of NGS errors . . . 13

1.2.2 Prevalence of NGS errors . . . 16

1.2.3 Consequences of NGS errors . . . 17

1.3 Detection of mutations from NGS data . . . 18

1.3.1 Variant calling . . . 18

1.3.2 Variant filtering . . . 23

1.4 Early cancer detection . . . 27

1.4.1 Definitions . . . 27

1.4.2 Objective . . . 27

1.4.3 Methods . . . 27

1.4.4 Circulating tumor DNA . . . 28

1.4.5 CtDNA as a biomarker for early cancer detection . . . 29

1.5 Global aim of the study . . . 31

2 Dealing with systematic errors: needlestack, a multi-sample sensitive variant caller 33

2.1 Scientific context . . . 34

2.2 Scientific contribution . . . 35

2.2.1 Statistical algorithm . . . 35

2.2.2 A robust implementation . . . 37

2.2.3 Needlestack performance . . . 38

2.3 Article A (Submitted in *Nucleic Acid Research*) . . . 40

2.4 Discussion . . . 67

3 Dealing with pseudo and non-systematic errors: variant filtering methodologies 73

3.1 Scientific context . . . 74

3.2 Scientific contribution . . . 75

3.2.1 Variant filtering for deep targeted sequencing data . . . 75

3.2.2 Application to ctDNA data . . . 84

3.2.3 Article B (in preparation) . . . 89

3.2.4 Variant filtering for germline data . . . 103

3.3 Discussion . . . 110

4 Applications to circulating-tumor DNA data 115

4.1 Scientific context . . . 116

4.2 Scientific contribution A . . . 117

4.2.1 Article C . . . 117

4.3 Scientific contribution B . . . 132

4.3.1 Article D . . . 132

4.4 Scientific contribution C . . . 157

4.4.1 Article E . . . 157

4.5 Scientific contribution D . . . 165

4.5.1 Article B . . . 167

4.6 Discussion . . . 167

Global discussion **169**

Conclusion **177**

A Annexes **I**

A.1 Supplementary work I

A.1.1 TCGA germline variant calling for rare variant susceptibility project . . . I

A.1.2 IARC bioinformatics pipeline homogenization IV

A.2 Publications from other collaborations VI

A.2.1 Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors (George *et al.*, Nature Communications, 2018) VI

A.2.2 Integrative and comparative genomic analyses identify clinically relevant groups of pulmonary carcinoids and unveil the existence of supra-carcinoids (Alcala *et al.*, Nature Communications, 2019) XX

B List of acronyms **LI**

C Bibliography **LV**

List of figures

1.1 DNA double helix 4
1.2 Genomic variations 6
1.3 NGS experiment 9
1.4 Germline VAF 12
1.5 Somatic VAF 13
1.6 Strand Bias 15
1.7 Number of variant reads for error and true mutation 17
1.8 GATK-HC-Reference Confidence Model overview 20
1.9 Effect on VAF hard threshold on variant calling accuracy 24
1.10 Example of a machine learning application 26
1.11 Release and extraction of cfDNA from the blood 29
1.12 Correlation between tumor volume and VAF from plasma samples 30

2.1 The needlestack workflow 38
2.2 SER coefficient of variation 69
2.3 Consequences of pre-computed error rates 70
2.4 Validation of Whole-Exome sequencing (WES) somatic mutations 71

3.1 Example of the distribution of number of called mutations per library 77
3.2 RVSB depending on sequenced reads strand repartition 79
3.3 example of LCAP values 81
3.4 example of LCAP values 82
3.5 Venn diagram of removed ctDNA mutations for each filter 86

3.6 Repartition of removed ctDNA mutations in cases and controls 86

3.7 Permutation test to estimate the expected number of duplicated errors 88

3.8 General Architecture of random forest 108

3.9 Paired representation of false and true called mutations for three variant statistics 109

3.10 Random forest on germline variant filtering Recall-Precision curve 109

3.11 Consistency of low confidence alterations across multiple runs 111

4.1 Development of the Circulating Tumor DNA (ctDNA) biomarker using genetic scores 166

A.1 IARC bioinformatics pipeline implementation pattern V

List of Tables

1.1 Reference genomes 5

1.2 Contingency table for strand bias Fisher test 15

3.1 Random forest features for germline variant filtering 107

A.1 Platypus parameters for WES germline variant calling III

A.2 IARC Bioinformatics pipelines XLIX

Chapter 1

Introduction

" Almost in the beginning was curiosity "

Isaac Asimov

Contents

1.1 Scientific context 3

1.1.1 Cancer 3

1.1.2 Genomic variations causing cancer 3

1.1.3 Next-generation sequencing 8

1.2 Errors in NGS data 13

1.2.1 Types of NGS errors 13

1.2.2 Prevalence of NGS errors 16

1.2.3 Consequences of NGS errors 17

1.3 Detection of mutations from NGS data 18

1.3.1 Variant calling 18

1.3.2 Variant filtering 23

1.4 Early cancer detection 27

1.4.1 Definitions 27

1.4.2 Objective 27

1.4.3 Methods 27

1.4.4 Circulating tumor DNA 28

1.4.5 CtDNA as a biomarker for early cancer detection 29

1.5 Global aim of the study 31

1.1 Scientific context

1.1.1 Cancer

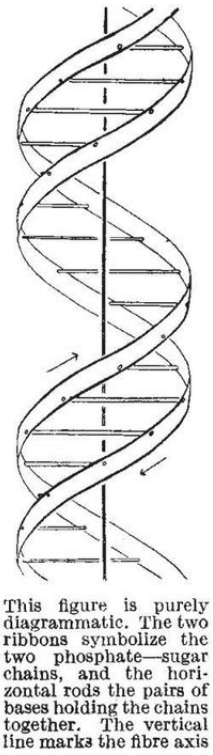
Cancer is a leading cause of death worldwide, responsible for around 10 million of deaths in 2018, which corresponds to 1 death over 6 dues to cancer [23]. Cancer can be defined as a set of related genetic diseases that can affect multiple parts of a living organism. In this work, we will focus on human cancers. An important measure that can expose the large diversity of cancer is the total number of classified types of cancer: there are more than one hundred distinct classes of cancers according to the National Cancer Institute (NCI) [1]. The common factor shared by every class of cancer is the type of initiation and the mode of proliferation [32]. The first step to develop a cancer, which corresponds to the type of disease initiation, consists in the transformation of a normal cell into a tumor cell [36]. A set of tumor cells defines a tumor, and there is two distinct types of tumors: benign and malignant [32]. Benign tumors remains located at the original place whereas malignant tumors are able to invading the surrounding normal tissue and to proliferate throughout other body areas. This process is called metastasis [27]. The major difference between a normal and a tumor cell is the capacity of tumor cells to grow out of control and to become invasive: they can ignore the biological signals conducting into apoptosis, also known as programmed cell death [46]. Cancer is mainly a genetic disease [144]. This means that cancers are mostly caused by genetic (or genomic) changes.

1.1.2 Genomic variations causing cancer

Genome

The term *genome* was firstly defined by a german botanist from the university of Hamburg, Hans Winkler, in 1920 [143]. The genome corresponds to the genetic material of a living organism. It is constituted of deoxyribonucleic acid (DNA), and is found in the cell nucleus. It is represented by a consecutive serie of nucleotides, or often simply called bases. These bases are biochemical entities and are usually represented by an alphabetical letter: the base *Adenine* is encoded as *A*, *Thymine* as *T*, *Cytosine* as *C* and *Guanine* as *G*. Bases are grouped into

two biochemical classes: purine bases (Adenine and Guanine) and pyrimidine bases (Cytosine and Thymine) that present different molecular properties. The DNA molecule that constitute the genome is composed of two strands, the forward strand and the reverse strand, coiling around each other to form the well-known DNA double helix. This structure of DNA was firstly described in 1953 by James Watson and Francis Crick [140] (figure 1.1).



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Figure 1.1 – Figure from Watson and Crick [140] representing the DNA double strands

Genome variations

Contrary to studies in the field of population genomics that are focused on large-scale comparisons of genomes between multiple populations, here we will focus on individual genomic variations. Such genomic variations correspond to changes of nucleotides when comparing a particular part of an individual genome to a *reference* genome. This comparison between a reference genome and an individual genome is inherent to the Next-Generation Sequencing (NGS) technology (see paragraph 1.1.3) as this requires a step of alignment. Nevertheless, such comparison is possible only when a reference genome is available, such as for human or mouse, and for species with an unknown reference, a step of *de-novo* assembly is needed to define the genomic sequence of the sample [16]. The reference reliability is a

current source of discussion and recently a group of researchers have proposed to re-define the reference genome using a consensus approach [17]. The human reference genome is currently defined by the Genome Reference Consortium (GRC) [30] [119], and represents a reference sequence of the human DNA. Coding regions are the most studied part of the genome, and the genomic coding region set is defined as the *exome*, which accounts for approximatively 1-2% of the genome. Nevertheless, recent studies show that variations found in non-coding regions of the genome can drive a cancer [38] [54]. The human genome is versioned (Table 1.1), and a particular version of the human genome corresponds to a fixed consensus DNA sequence.

Table 1.1 – Details of available reference genomes, identified by GRC version

Release Name	**Date of release**	**UCSC version**	**Total number of bases**
GRCh38	Dec. 2013	hg38	3,209,286,105
GRCh37	Feb. 2009	hg19	3,137,144,693
NCBI Build 36.1	Mar. 2006	hg18	3,104,054,490
NCBI Build 35	May 2004	hg17	3,091,649,889
NCBI Build 34	Jul. 2003	hg16	3,091,959,510

There exists three major types of genomic variation: Single Nucleotide Variation (SNV), Insertion or deletion (indel) and structural variations (figure 1.2). A SNV corresponds to any one-base-pair change, and there are two subgroups of SNVs: transitions, that corresponds to a change between the same biochemical class (purine into purine or pyrimidine into pyrimidine), and transversions, that corresponds to a change between the two different biochemical classes (purine into pyrimidine or pyrimidine into purine). An indel is defined by a loss or a gain of nucleotides that ranges from two to hundreds of bases in length. Finally, a structural variation describes a genomic variation of a larger size. It includes both chromosomal rearrangements and DNA Copy Number Variation (CNV). In this work we will focus on *i.e.* SNVs and indels.

Germline and somatic mutations

A genomic variation when comparing an individual genome to a reference genome can be called a mutation. Mutations are grouped into two classes depending on the type of cell carrying the mutation: if the mutation was acquired from the parents of the individual, *i.e.*

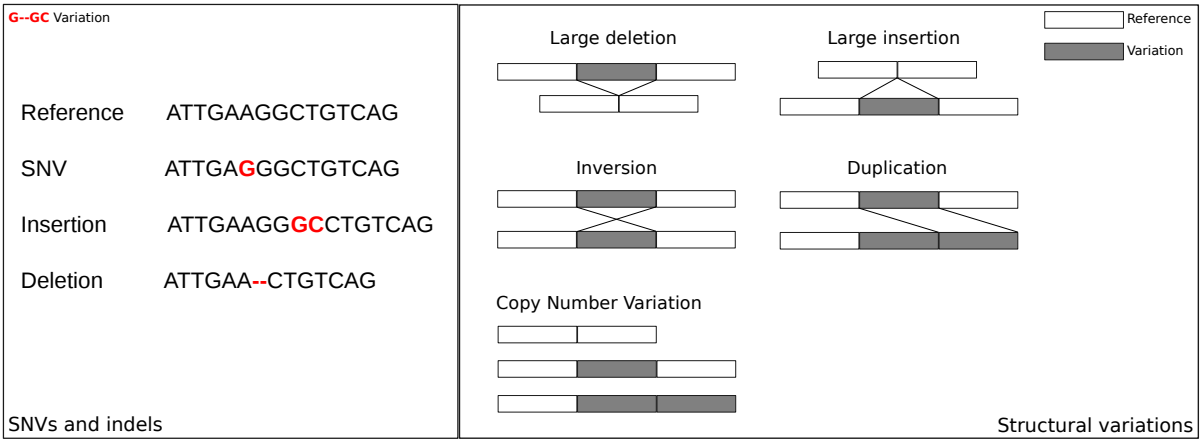


Figure 1.2 – Schematic representation of the three different classes of genomic variations

presents in their germ cells it is called a *germline* mutation. Germline mutations are expected to be present in each cell of the individual, *i.e.* in both normal and tumor cells. A germ cell can also acquire a mutation during the lifetime of an individual, and this type of germline mutation is called a *de novo* mutation. If a germline mutation is identified in an individual but is not present when analyzing DNA of the parents, this corresponds to such *de novo* mutation that one of the parent has acquired [53]. More rarely, a *de novo* mutation can arise not in the germ cell of one parent but directly inside individual cells at early development stages, and in this case the mutation is called a *post-zygotic de novo* mutation and are responsible for somatic mosaicism [4]. If the mutation in acquired during the lifetime of the individual in a non-germ cell, it is called a *somatic* mutation. This means that given a particular sample of cells from an individual, all the cells will carry the germline mutations, but it is expected that only a fragment of them will carry a specific somatic mutation. Germline and somatic mutation proportions in a human sample are not similar. This leads to a difference in term of observed number of mutations. When analyzing the genome of a particular cell (normal or tumor cell), it is expected to detect around one germline mutation every 1,000 base pairs, which is equivalent to a rate of 1/Kb [66],[21]. When analyzing a tumor cell, it is expected to detect in addition around one somatic mutation every 1,000,000 base pairs, which is equivalent to a rate of 1/Mb [9]. This leads to an expectation of around 30,000 germline mutations in a human exome and 30 somatic mutations. This is just an order of magnitude, and therefore these numbers are expected to vary largely across individuals and cancer types [9].

Variations causing cancer

The evolutionary perspective on cancer stipulates that cancer initiation and progression is due to an accumulation of selectively advantageous mutations [18]. Indeed, a normal cell can acquire one or multiple genomic alterations, and these variations can give to the cell the capacity to divide abnormally faster than a normal cell and to proliferate, and subsequently causes cancer. Such mutations that decrease the fitness of a cell are called deleterious mutations [91].

Genes that promote a tumor initiation when altered by a deleterious genomic variation are called *driver* genes. Two types of genes playing a role in cancer has been reported: oncogenes and tumor suppressor genes. Oncogenes are genes that can help abnormal cells to grow when activated and Tumor Suppressor Gene (TSG) are genes that control cell divisions, DNA repair and apoptosis when activated. This leads to a difference in term of expected type of mutations observed in these genes in a tumor sample: mutations in oncogenes should activate the protein function to help cancer to progress contrary to mutations in TSG that should inactivate the protein function. It has also been reported that some genes can show both characteristics [120]. In 2013, there were around 140 genes recorded as cancer drivers [137]. In 2018, a new study updated this and described a total of 299 driver genes [15]. The top-five genes identified as driver in most cancer types among the 33 studied types are *TP53* (most extreme case, driver in 27 cancer types), *PIK3CA*, *KRAS*, *PTEN*, and *ARID1A*. Mutations affecting the production of a protein are called driver mutations, and other mutations that do not confer any growth advantage, are called *passenger* mutations.

Due to genetic code redundancy [78], it is accepted that missense variations *i.e.* mutations that change the particular amino acid (entity that form a protein) encoded, indels and non-sense mutations *i.e.* a mutation that introduces a stop codon and truncate the protein can alter the protein. To predict the deleterious power of a particular mutation, multiple methods have tried to build databases of pathogenicity of human variations, such as SIFT [102], PolyPhen [6],[5], or more recently REVEL [67], a tool predicting the pathogenicity of missense mutations based on a random forest learning on multiple databases of rare neutral and disease missense variants.

In 1971, the geneticist Alfred Knudson proposed the hypothesis that most genes require two mutations to be inactivated, that is also known as the *two-hits* hypothesis [73]. This hypothesis is based on the observation that most of deleterious mutations in TSG are recessive [97] and therefore the two alleles of a TSG should be mutated to activate tumor proliferation. Based on his hereditary observations on retinoblastoma, Knudson proposed that a TSG is inactivated from both a deleterious germline mutation in one allele and a deleterious somatic mutation in the second allele.

In summary, cancer is a two-component disease. First component is defined by individual susceptibility, *i.e.* the inherited mutations that can participate to the alteration of a gene leading to the development of a cancer. Second component is defined by individual capacity to acquire mutations. This second component depends both on randomness and on environment.

1.1.3 Next-generation sequencing

Definition

The identification of genomic variation was made possible by the development of sequencing techniques. The aim of such techniques is to determine the genomic sequence of an individual DNA, that can then be compared with a defined reference to identify variations.

The first robust DNA sequencing method was developed during the 1970s at the Medical Research Council Center in Cambridge, UK, by Frederick Sanger [115], who received for this innovative discovery the Nobel Price in 1980. The method is based on the incorporation of complementary nucleotides during and *in vitro* DNA replication, that can be determined at the end of the experiment. The Sanger sequencing method was the one used to produce the first human genome in 2001 [81]. Nevertheless, the main disadvantage of this sequencing technique is its cost. Although the per genome cost of such methods was divided by 10,000 from 2001 to 2011 [141], the total cost of a genome sequencing with the Sanger technology is still highly expensive (around $10,000).

This need has driven the development of new low-cost sequencing methods, known as High-Throughput Sequencing (HTS) or NGS[123]. The price of a whole genome sequencing

using these methods was decreased until reaching only $1,000 in 2015. Multiple distinct NGS technologies have been developed since early 2000's, and in this work we will focus on two particular sequencers, the Illumina [121] and Ion Torrent next-generation sequencers. Whatever the sequencing technology, NGS method of DNA sequencing follows three steps: library preparation, amplification of DNA fragments and sequencing of these fragments [123]. The resulting entity is called a *sequencing read*, which identify the sequence of a particular segment of DNA, and a NGS experiment creates a massive amount of reads, up to several hundreds of millions of reads when sequencing a complete genome. The capacity of NGS technologies to sequence millions of DNA fragments defines a major improvement when comparing NGS and traditional Sanger sequencing: NGS is offering a unique ability to detect minor variants in a DNA sample [43]. Figure 1.3(A) summarizes the different steps forming a NGS experiment.

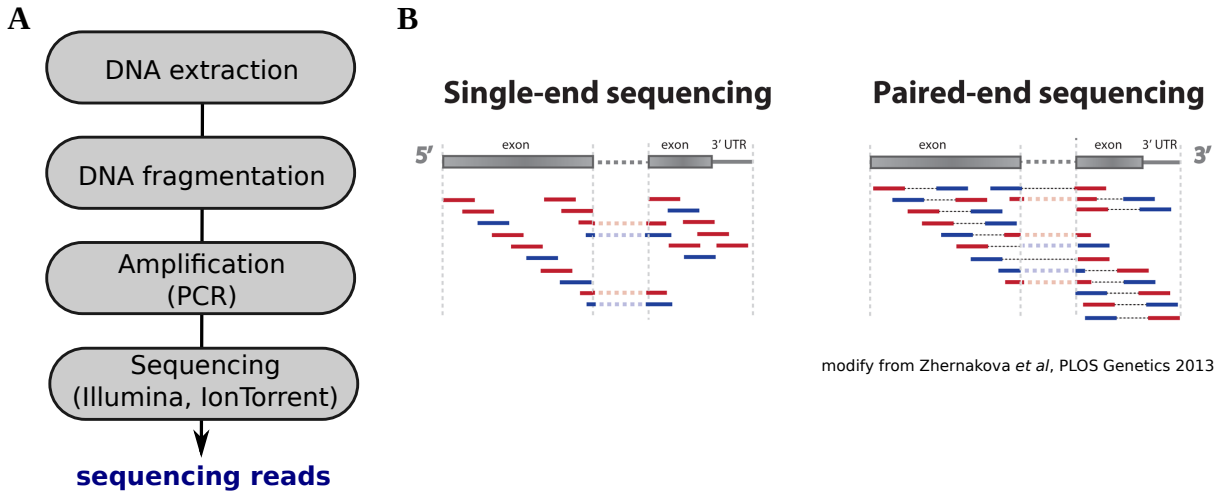


Figure 1.3 – Representation of a NGS experiment

A subsequent required step after the raw sequencing to obtain data that could be analyzed is called the read *alignment*. The raw sequencing only outputs small segments corresponding to initial DNA, but these small fragments need to be mapped to the reference DNA to then reconstruct entirely the DNA sequence 1.3(B) [148]. There exists multiple methods of NGS alignment based on different algorithms, the main challenges justifying these developments being the computing time and the accuracy [51]. The most commonly used alignment algorithm is Burrows-Wheeler Aligner (BWA), which was developed by Heng Li at Harvard University. This method proposes two different algorithms to maximize the ac-

curacy of the alignment depending on the length of the sequencing reads: for reads up to 100bp, the BWA-SW algorithm, based on the Burrows-Wheeler Transform is recommended [88]. The BWA-MEM algorithm was designed for longer reads [85].

Ion Torrent sequencers implement their own technology-specific aligner which is based on the BWA algorithm and is called *TMAP* [2]. This step is an automatic part of the Ion Torrent sequencing process. Illumina and IonTorrent Proton sequencer have many differences. First, they do not use the same technology to read the DNA sequence. Illumina uses a fluorescence-based method whereas the IonTorrent Proton uses pH measurement [79]. There are also some differences in the type of data generated by these sequencing technologies. Illumina sequenced reads have the same length whereas IonTorrent Proton reads have lengths that differ. Both sequencing technologies can generate "paired-end" reads (corresponding to two extremities of a particular DNA fragment) or "single-end" reads (Figure 1.3(B), [148]).

There are three main types of study design for sequencing projects based on the length of DNA that the experiment aims to determine: Whole-Genome Sequencing (WGS), that refers to the sequencing of the entire DNA molecule present in a cell, WES, that refers to the sequencing of exomes, the coding part of the genome that contributes to the production of the protein, and finally *target sequencing*, that refers to the sequencing of only a small previously determined part of the genome, frequently corresponding to specific genes. This does not depend on the sequencing technologies, but due to cost divergences, in this work we used data from Illumina sequencers for WES and WGS and Ion Torrent for target sequencing. Deep sequencing usually generates coverages corresponding to a couple thousand (or tens of thousands) sequenced aligned reads per position, compared to only a few dozens or hundreds reads for WGS and WES.

Sequencing of variations

As mentioned in the paragraph 1.1.3, a genomic variation can be defined by a difference when comparing a given DNA sample to a reference DNA. In NGS, each sequenced genomic variation is associated to a metric, the Variant Allelic Fraction (VAF). The VAF of a mutation

m at a genomic position p can be defined as the following:

$$\text{VAF}_m = \frac{\text{AO}_m}{\text{DP}_p}$$

with AO_m being the number of alternative reads aligned at the position of the mutation, and DP_p being the total number of aligned reads at the position. DP_p is also named *coverage* or *depth*. VAF_m corresponds to the proportion of sequenced alleles that are carrying the mutation m. The VAF is therefore a proxy of the proportion of sampled cells that are carrying the mutation, integrating in addition the status of the mutation: heterozygous or homozygous. The VAF of a germline mutation only depends on the status of the mutation, due to the fact that all of the cells in a biological sample are expected, under normal conditions, to contain this mutation inherited from parental germ cells. This VAF is expected to be equals either to 100% if the individual is homozygous for the mutation *i.e.* the two alleles of the chromosome are mutated, or to 50% if the individual is heterozygous for the mutation *i.e.* only one of the two alleles is mutated. The sequencing of germline DNA mutations can be modelled by a binomial sampling (B) and therefore the expected number of alternative reads is defined as the following:

$$\text{AO}_m \sim \text{B}(\text{DP}_p, \text{P}) \text{ with } \text{P} \in [0.5, 1]$$

The variance of such a distribution is equal to $\text{DP}_p * \text{P} * (1 - \text{P})$ and therefore the variance of germline VAF depends on the coverage and is expected to be null for heterozygous mutations. Figure 1.4 represents the distribution of germline VAF under the expectation of a binomial sampling, for homozygous mutations (simulation of 250 mutations with a coverage uniformly varying between 100 and 200, in orange) and heterozygous mutations (simulation of 1,000 mutations with a coverage uniformly varying between 100 and 200, blue).

In the case of a somatic mutation, the expected VAF is unknown [26], [126]. Indeed, firstly, somatic mutations are typically analyzed on tumor sample from a tumor bulk. Nevertheless, such samples are basically composed of both normal and tumor cells. The proportion of tumor cells among the total number of extracted cells from a tumor bulk corresponds

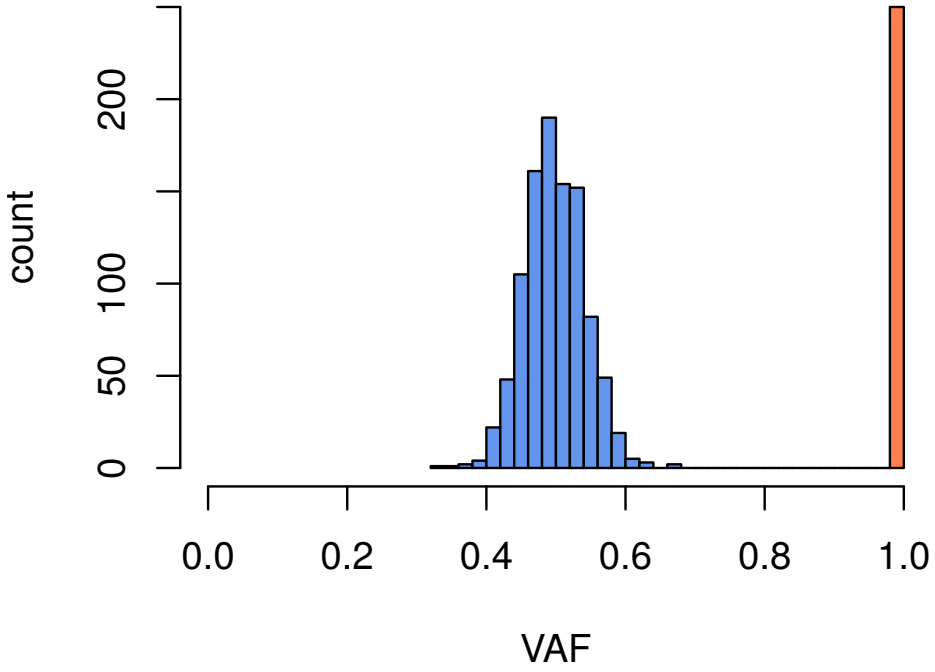


Figure 1.4 – Expected distribution of germline VAF for homozygous mutation (orange) and heterozygous mutations (blue) under assumption of a binomial sampling of sequencing reads. Distributions correspond to a simulation of 1,000 heterozygous mutations and 250 homozygous mutations with coverages varying uniformly between 100 and 200.

to the purity P of the sample [127]. The value of the exact purity is unknown and only an estimation can be obtain, either through a pathology visualization or through computational analysis [127].

Secondly, a somatic mutation in contrary to an inherited mutation is not expected to be found in all of the cells of the tumor except if this mutation is the tumor initiating event. This also contributes to the deviation of the somatic VAF from the expected VAF of 50 or 100%. The proportion of cancer cells that are carrying a particular somatic mutation corresponds to the *subclonality* S of the mutation [95], [142]. Lower the subclonality, more recent the mutation in term of the tumor event timeline, higher the deviance of VAF from the expectation.

Finally the last component of the deviation of the somatic VAF from the expected values is the possibility of variation in the number of copy of the allele carrying the mutation [122], [58]. Finally, the VAF of a somatic mutation *sm* can be defined as the following:

$$\text{VAF}_{sm} = 0.5 * \text{CNV} * \text{P} * \text{S}$$

As a descriptive example, Figure 1.5 [131] is showing the VAF distribution of somatic mutations found in a lung adenocarcinoma tumor sample from The Cancer Genome Atlas (TCGA) [35].

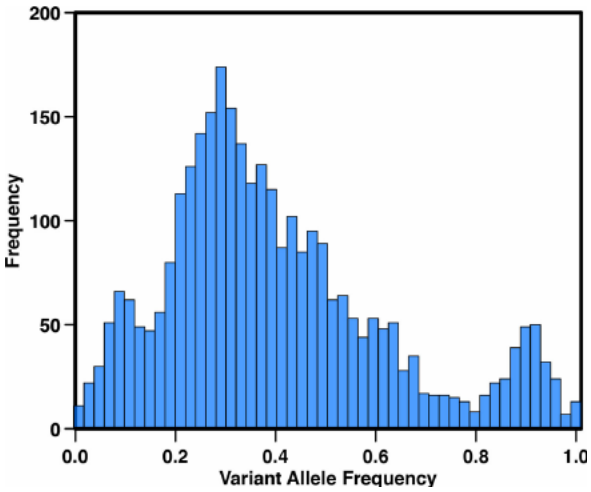


Figure 1.5 – Distribution of VAF of somatic mutation from one TCGA lung adenocarcinoma patient [35].

1.2 Errors in NGS data

1.2.1 Types of NGS errors

Next-generation sequencing technologies has revolutionized the genome analyses due to its ability to produce high number and low cost DNA sequences in a short time comparing to the traditional Sanger method. Unfortunately, NGS technologies are producing a higher number of artifacts in output data.

These errors can be *systematic*, *non-systematic* or *pseudo-systematic*. This error feature is a *read-level* feature: the **systematic nature** of an error is defined by its tendency to occur in multiple samples.

Systematic errors can be defined as artifacts that occur in all the samples with a particular rate that can be statistically modelled. This is the case of the Sequencing Error Rate (SER) (see paragraph 1.2.2).

Non-systematic errors can be defined as artifacts specific to one particular sample. Typically, the most common source of non-systematic error from the sequencing is induced by

the Polymerase Chain Reaction (PCR) processes. PCR steps are required upstream of the sequencing process to the amplify the input DNA molecules in order to be sure that each DNA molecule present in the input sample would be sequenced even if present in a very low proportion. Nevertheless, these steps of PCR during the library preparation introduce multiple punctual errors due to mistakes from the DNA polymerase enzyme at a varying rate basically comprised between 1/Mb and 0.1/Mb [33]. These errors are amplified in each subsequent cycle of PCR, conducting to a number of mutated reads in the same order than the one expected in the case of a true DNA mutation. It has been shown that sequencing the DNA in replicates *i.e.* from multiple independent library preparations should reduce artifacts from PCR process [111].

Some errors tend to occur in multiple samples without presenting any constant rate across the samples. They can be defined as *pseudo-systematic* errors. These errors can depend on the sequence of the DNA and are called in this case Context-specific Error (CSE) [11]. As an example, if a DNA sequence carries a mutation that increases the complexity of the nearby region of the sequence by generating a repetition of a particular nucleotide, this can create alignment artifacts on this region presenting low VAF [124] [145]. These artifacts would be observed only in samples carrying a mutation in this region and so these artifacts are pseudo-systematic. Several methods have been developed recently to increase the accuracy of the alignment step specially for indel detection by proposing to integrate a re-alignment process, such as methods based on consensus sequence correlation maximization [65] or assembly-based methods [99]. Unfortunately, these methods only reduce these artifacts but does not remove them if falling in a highly complex region. A particular metric can identify these CSE: the strand bias. A variant is found in strand bias when the alternative reads are not dispersed on the two strand in the same way than reference reads. Indeed, a CSE is caused by the DNA sequence preceding it and not following it (reading direction of the sequencing) and, then, if a specific context prone to errors, such as repeated regions, the error should be present in reads of one direction only (forward or reverse). This defines a sequencing strand bias (1.6).

The common measure of strand bias for low coverage data is the Fisher exact test statis-

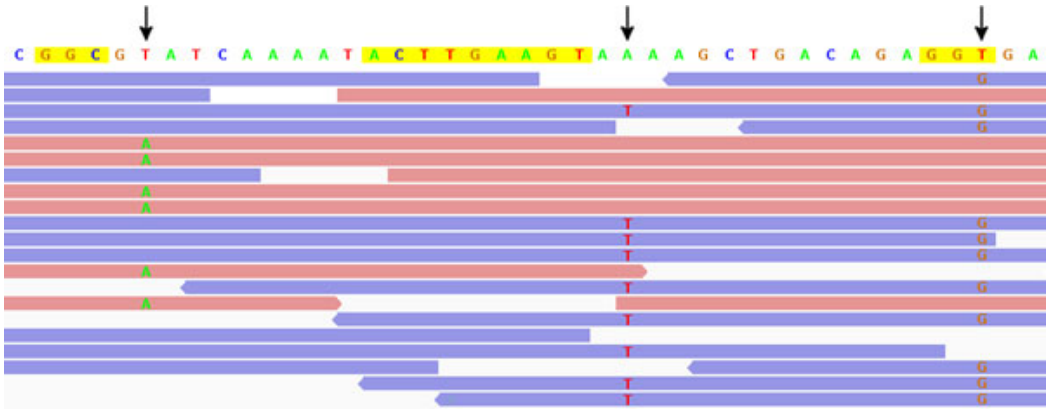


Figure 1.6 – Hypothetical reads of two directions (red: forward; blue: backward) are aligned to a reference genome shown on top. Nucleotides within reads indicate mismatches to the forward reference. Three genome positions with extreme strand bias are marked by arrows [11]. Created with Integrative Genomics Viewer (IGV) [113], [130].

tic [50]. In this case such a Fisher exact test *p*-value is computed on the contingency table 1.2.

Table 1.2 – Contingency table on which the Fisher exact test *p*-value is computed to estimate strand bias of a genomic variation on low coverage data. AO defines the number of reads carrying the variation and RO defines the number of reads carrying the reference base at the position.

	Variant	Reference
Forward	AO_f	RO_f
Reverse	AO_r	RO_r

Nevertheless, the Fisher exact test is extremely sensitive to very low differences in strand repartition such as found in high coverage data. An alternative measure of strand bias, the Relative Variant Strand Bias (RVSB) statistic, is better adapted to high coverage data [49]. For a given mutation, the RVSB statistic is defined by:

$$\text{RVSB} = \frac{max(\text{AO}_f * \text{DP}_r, \text{AO}_r * \text{DP}_f)}{\text{AO}_f * \text{DP}_r + \text{AO}_r * \text{DP}_f}$$

where *AO* corresponds to the number of alternative reads aligned at the position of the mutation, *DP* to the total number of aligned reads, *f* to reads aligned on the forward strand and *r* aligned to the reverse strand. RVSB is comprised between 0.5 and 1, and higher the RVSB higher the strand bias. In such tests, the null hypothesis is defined by an absence of strand bias, and therefore the test *p*-value can be compared to a pre-defined threshold to decide if the variant presents a strand bias or not. Interestingly, it has been reported that the strand

bias is not related to the alignment method [60].

1.2.2 Prevalence of NGS errors

Systematic sequencing errors in NGS data appear with a particular rate, called the SER. The SER at a position p can be defined as the following:

$$\mathrm{SER}_p = \frac{\mathrm{EO}_p}{\mathrm{DP}_p}$$

with EO_p begin the number of reads aligned at the position presenting a error base for this position, and DP_p being equals to $\mathrm{EO}_p + \mathrm{RO}_p$, with RO_p being the number of reads aligned at the position presenting the reference base for the position.

The SER reflects the prevalence of a given NGS error. This prevalence is varying among sequencing technologies [104],[22] but also among DNA positions. Current reported per-base SER are ranging from 0.18 to 1.17% depending on the sequencer, with higher rates observed in Ion Torrent technologies [14], [104], [107]. However, these studies only reported *per-base* error rates, considering that a sequencing error only depends on the genomic position and not on the alteration, *i.e.* at a specific position, there is the same probability to observed an error independently of a particular base change.

When considering the base change in addition to the position, a different formulation of the SER can be defined:

$$\mathrm{SER}_{p,m} = \frac{\mathrm{EO}_{p,m}}{\mathrm{DP}_p}$$

where m is defining the base change. With this definition, the SER is modelled independently for each possible base change, *i.e.* at a given DNA position, $\mathrm{SER}_{p,\mathrm{G}\rightarrow\mathrm{T}}$ should potentially be different than $\mathrm{SER}_{p,\mathrm{G}\rightarrow\mathrm{A}}$.

1.2.3 Consequences of NGS errors

As mentioned earlier, the way to detect a DNA mutation from a NGS experiment consists in analyzing the sequenced reads and observed the variations present in these reads compared to a DNA reference. Nevertheless, taking into account the fact that NGS techniques intrinsically produce errors, one available metric that can be used to classify a variation as being a true mutation or a sequencing artifact is the prevalence of the variation. Indeed, sequencing artifact are expected to be rare compared to the expected proportion of the sequenced reads carrying a true mutation, even though this highly depend on the proportion of sequenced cells that are carrying the mutation (see paragraph 1.1.3). This suggests that in the case of an high SER potentially in the same range than the VAF of a true mutation, the detection of the mutation would not be possible due to the failure to distinguish between the observed number of variable reads due to a true mutation and the observed number of variable reads due to sequencing errors. This potential similarity in prevalence between true mutations and sequencing errors is the major consequence of NGS error production. The figure 1.7 is showing an schematic example of observed number of aligned reads and their variations in a position p and for a base change m, for two distinct cases of $\text{SER}_{p,m}$ and $\text{VAF}_{p,m}$ combinations: $\text{VAF}_{p,m}$ higher than $\text{SER}_{p,m}$ and $\text{VAF}_{p,m}$ in the same range than $\text{SER}_{p,m}$.

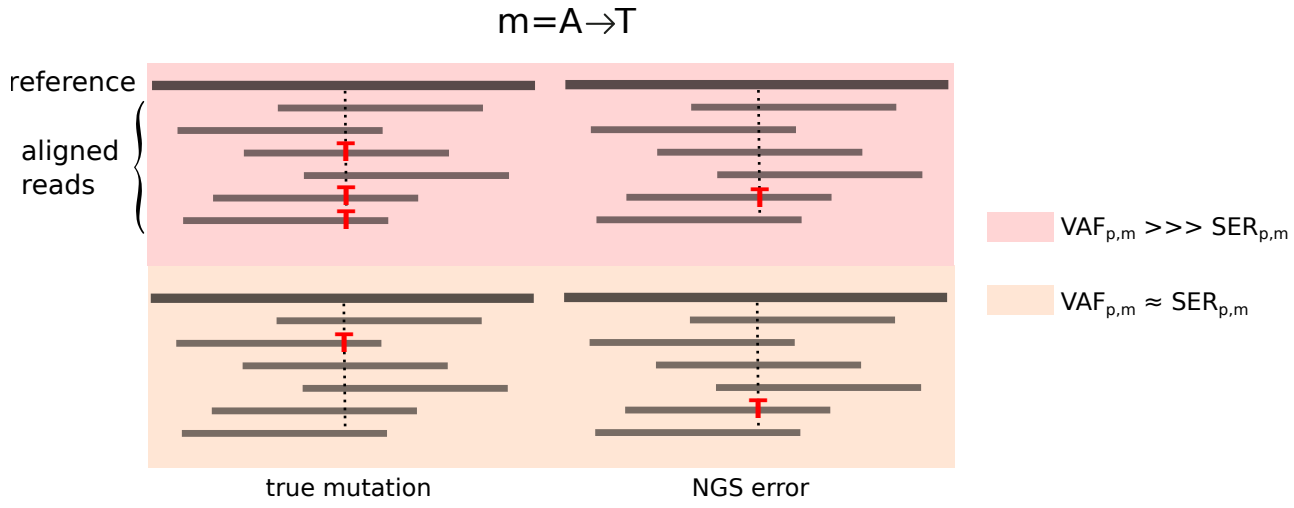


Figure 1.7 – Schematization of the number of variable reads for the two cases of combination of SER and VAF

1.3 Detection of mutations from NGS data

The sequencing of DNA is necessary to identify genomic variations, but the sequencing process itself only produces raw data about the DNA sequence of an individual: if should be associated with downstream analyses to obtain the set of mutations identified from a DNA sample. The major subsequent step is the variant calling, a analytical process that aim to detect DNA variation compared to a reference. This process does not correct for all types of NGS errors, and therefore is often coupled with a step of variant filtering to remove the false discoveries of a variant calling procedure.

1.3.1 Variant calling

Variant calling is a generic term grouping the computational techniques that identify DNA mutations from NGS experiments [103], [101], [146]. The key issue of the variant calling is first the identification of DNA variations and then the distinction between a variation corresponding to an error from the NGS experiment and a variation corresponding to a true biological DNA mutation. For this, variant calling methodologies are mostly based on the analyze of the prevalence of the variation in the NGS experiment (the prevalence of a true DNA mutation corresponds to the VAF). Therefore, variant calling algorithms are designed according to three distinct types of DNA mutations defined by their expected prevalence in the NGS data, that can be defined as *germline*, *somatic*, and *low VAF somatic* mutations.

As exposed in the paragraph 1.1.2, a germline mutation corresponds to a mutation inherited from the parents of the individual and is expected to be present in all of the sequenced cells and to harbor an *a-priori* VAF either equals to 0.5 (heterozygous mutation) or equals to 1 (homozygous mutation). Combinatorial methods for detecting germline mutations from NGS data are typically based on a *bayesian inference model* using the expected VAF in the likelihood function of the bayesian model. This method has been adopted for example by *Strelka*, *Freebayes* and GATK UnifiedGenotyper [117], [55], [106]. At each position and for each possible variation, the probability of observing a genotype G can be computed

with the Bayes'rule:

$$P(G|D) = \frac{P(G) * P(D|G)}{P(D)}$$

$$= \frac{P(G) * P(D|G)}{\sum_{i=1}^{n} P(D|G_i)P(G_i)}$$

With G_i referring to the i^{th} genotype over n possible genotypes, D corresponding to the observed data *i.e.* the aligned reads (quality of the sequencing can be taking into the counting of reads to remove potentially artifact reads) and G corresponding to the estimated genotype (none, heterozygous mutation or homozygous mutation).

Existing germline variant calling methods mainly differ on the way of computing both the prior on genotypes P(G) and P(D|G) the likelihood of the observations, which should incorporate a specific error model.

As an example, Genome Analysis ToolKit (GATK) HaplotypeCaller (GATK-HC) [96], [106], a method for germline variant calling, can be defined as 4 separated steps (figure 1.8):

- Identification of variable genomic regions ("ActiveRegions")
- Determination of all possible haplotypes through a re-assembly of variable regions
- Per-read likelihood estimation depending on the possible haplotypes, using the Pair Hidden Markov Model (PairHMM) algorithm. This part is computing P(D|G).
- Attribution of the genotypes using the Bayes'rule.

The underlying assumptions for somatic variant calling is totally different from germline variant calling. Indeed, there are some expectations on the germline VAF but somatic VAF are more variable and predictions are more difficult [26], [126]. This is a key issue in variant calling, because contrary to germline variant which are expected to be found in high proportion (50% or 100%), a somatic variant could be present in a different abundance potentially reaching the SER (1.2.3, 1.1.3). This requires a more sensitive statistical model.

Most of somatic variant callers are based on a paired variant calling which is realized on a tumor sample and on its matched normal sample to identify tumor somatic variations. They benefit from the availability of both normal and tumor samples to increase the sensitiv-

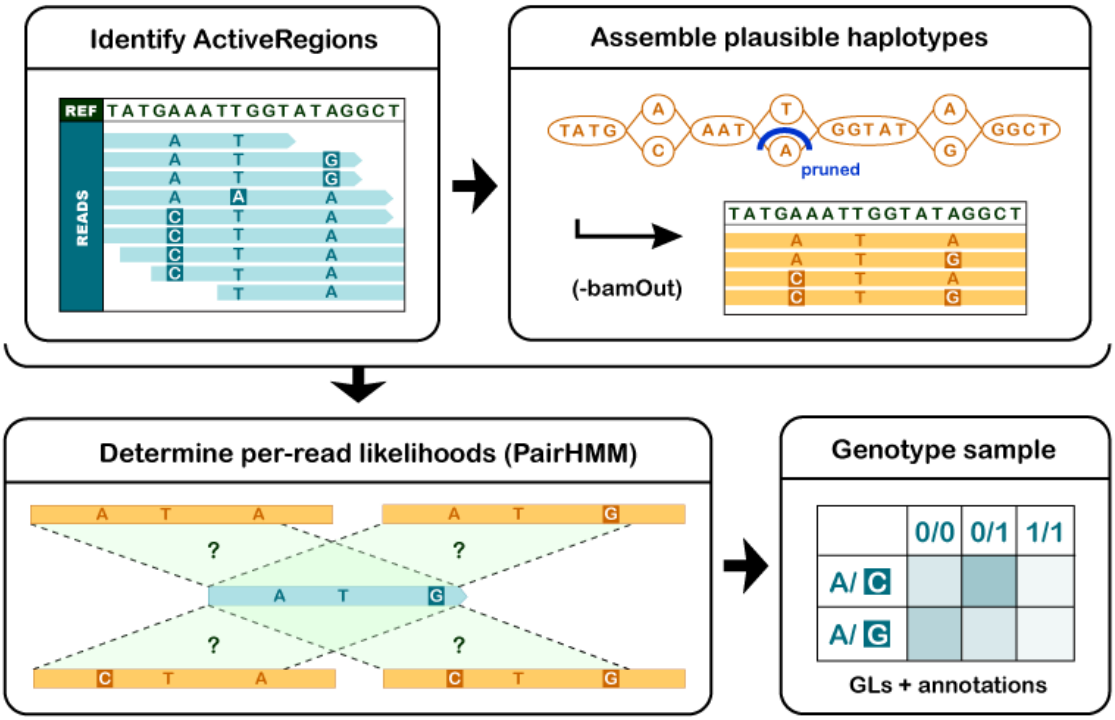


Figure 1.8 – Four steps of the GATK-HC algorithm to detect germline variant from NGS data from [106]

ity of the germline variant calling by directly comparing the two samples (indeed, a germline mutation is expected to be found in both samples), instead of calling mutations separately and removing the mutations identified in the normal samples from the mutations identified in the tumor samples to compute the set of somatic mutations. This approach is essential given the fact that, as exposed in the paragraph 1.1.2, germline mutations are expected to be found in a proportion 1,000 times higher than somatic ones, and, then, a decreased germline sensitivity of x compared to the maximum of 1 would leads to a somatic false discovery rate of $\frac{x*10^3}{1+(x*10^3)}$. This mean that if the two callings are not paired, even a germline sensitivity of 99.9% would lead to a somatic false discovery rate of 50% (corresponding to 0.1% of the germline mutations classified as somatic).

VarScan2 [74] and VarDict [80] have chosen a frequentist approach to filter sequencing reads based on fixed thresholds on read statistics and then identify potential variants and excluded the errors. These potential variants are classified as somatic based on a statistical test comparing the tumor and normal sample for this variant. Some somatic variant callers are based on a probabilistic framework, such as SomaticSniper [83], JointSNVMix [114] or CaVEMan [69]. The core concept of these methods is the assumption of diploidy in the tu-

mor.

Nevertheless, the diploidy assumption underlying these algorithm is violated in the case of subclonality or low sample purity leading to the dilution of somatic mutations and to the reduction of the VAF (1.1.3). To be able to efficiently detect all types of somatic mutations, some other methods prefer to base their probabilist model on VAF instead of genotypes. MuTect [31], Strelka [117], and MuSE [48] are based on such a method. The common idea of these variant callers is basically to identify potential true variants first and then to compare the observed VAF in the somatic sample to the VAF in the normal sample to extract somatic mutations.

The most commonly used variant caller for somatic mutations, MuTect, computes for each possible mutation m and for each sample identified by its VAF f a Logarithm of Odds (LOD) score that is comparing the likelihood of the model of errors $\mathrm{L}(\mathrm{M}_0)$ to the likelihood of a true mutation model $\mathrm{L}(\mathrm{M}_f^m)$:

$$\mathrm{LOD}_\mathrm{T}(m,f) = log_{10}\left(\frac{\mathrm{L}(\mathrm{M}_f^m)}{\mathrm{L}(\mathrm{M}_0)}\right)$$

with

$$\mathrm{L}(\mathrm{M}_f^m) = \prod_{i=1}^{d} \mathrm{P}(b_i|e_i,r,m,f)$$

and

$$\mathrm{L}(\mathrm{M}_0) = [\frac{e_i}{3}]^d$$

$r \in \mathrm{A,C,G,T}$ denotes the reference allele, d the coverage at the position of the mutation m, b_i and e_i respectively the called base and its quality for the read $i \in [1,d]$.

The statistic $\mathrm{LOD}_\mathrm{T}(m,f)$ is then compared to the threshold θ_T and if $\mathrm{LOD}_\mathrm{T}(m,f) \geqslant \theta_\mathrm{T}$, the algorithm declares m as a candidate variant. In the method paper, the authors propose a threshold of $\theta_\mathrm{T} = 6.3$ which corresponds to an expected mutation frequency of 3/Mb.

MuTect uses in its LOD score the base qualities provided by the sequencing machine.

This statistic corresponds to the probability for the base to be false, and is given by the sequencing machine. They recommend in the *GATK best practices* for somatic variant calling to execute a Base Quality Score Recalibration (BQSR) step before the computations of the LOD scores, because the base qualities are not correctly calibrated by the sequencing machine []. The BQSR uses a machine learning approach to estimate the influence of particular covariates, such as the sequence context or the position in the read, on the confidence on bases, and then recalibrate the initial base qualities. Nevertheless, to be efficient, this step requires a large panel of sequenced bases to be performed, and, as mentionned by the authors, is not adapted to small gene panels.

A second limitation is that the method assumes that all the substitution errors occur with the same probability $\frac{e_i}{3}$ [31]. This is a large assumption which is inherent to the method and which has not be proved.

Another limitation of this method is when trying to identify mutations present in very low proportion. Indeed, when f tends to zero, $\mathrm{P}(b_i|e_i,r,m,f)$ tends to $\frac{e_i}{3}$ and $\mathrm{L}(\mathrm{M}_f^m)$ tends to $\mathrm{L}(\mathrm{M}_0$ (see paper methods for details [31]), and therefore this leads to uncertainty of the low abundance mutation calling. To deal with this issue, typically it is not recommended to consider mutations called with a VAF lower than 5%.

The most accurate variant caller available to detect low VAF somatic mutations is ShearwaterML [92]. This variant caller proposes a method based on multiple samples to estimate the SER at each position and to call variant as being different from this error model. ShearwaterML is the maximum-likelihood adaptation of the original Shearwater algorithm published previously [57]. ShearwaterML is based on a beta-binomial regression to estimate the error rate for the observed mutation and attributes at each sample a p-value corresponding to the probability for the sample to carry the mutation. The algorithm models the errors independently on the two DNA strands (X_{ijk} models errors on the forward strand and X'_{ijk} on the reverse strand) as the following:

$$\mathrm{X}_{ijk} \backsim \mathrm{BetaBin}(n_{ij}, v_{ijk}, p_{jk})$$

$$\mathrm{X}'_{ijk} \backsim \mathrm{BetaBin}(n'_{ij}, v'_{ijk}, p_{jk})$$

with i the sample, j the position, k the alternative base, n the coverage, v the VAF and p the overdispersion parameter of the beta-binomial regression. Given this, then shearwaterML computes the likelihood of the observation considering that the observed number of alternative reads is drawn from this model of error and the likelihood considering that the number of alternative reads is drawn from a beta-binomial regression with mean equals to the VAF. Using a likelihood ratio test, shearwaterML finally generates a p-value for each sample and for each possible mutation.

1.3.2 Variant filtering

Variant calling methods are suitable to detect potential mutations by the study of quantitative features such as the read count to distinguish between mutations and sequencing errors. Nevertheless, there still remain false positive observations after the variant calling, typically non-systematic and pseudo-systematic errors appeared at a former step, such as the PCR-induced errors. Secondary genomic analyses should be therefore oriented, in addition to sequencing alignment and variant calling, toward *variant filtering*. Various methodologies have been developed recently in the aim at increasing the *specificity* of the variant calling by removing remaining artifacts.

The naive method to reduce false positives by filtering variants consists in the manual inspection of the aligned reads. This should help to decide if a variant can be considered with confidence or not. It has been recently proposed to use the IGV software [113], [130] to visualize potential variants and classify them into false or true observations [112]. The main problem of this strategy is both the lack of objectivity and the lack of references to helping to make a decision. A second method is proposing a Standard Operating Procedure which integrates an IGV plug-in named IGVNavigator to refine the manual inspection of variant. More advanced variant filtering method can be classified into two groups:

- hard-filtering methods on key statistics
- machine-learning-based filtering methods

Hard-filtering methods consist in removing variants based on variant summary statistics. The main idea is first to select these summary statistics grouped as the set S, and then

to remove variant that harbor a value of the statistic s_v higher than a *pre-defined* threshold s_t. A variant is not removed if it validates the following condition:

$$\forall s \in S, s_v < s_t$$

MuTect variant caller proposes to remove variant based on six different statistics with *pre-defined* thresholds: *proximal gap, strand bias, poor mapping, triallelic site, clustered positions* and *observation in controls*. Generally, variant callers propose a set of thresholds for each of tunable variant statistic, but these thresholds can be also defined by the user using independent studies such as simulations [128].

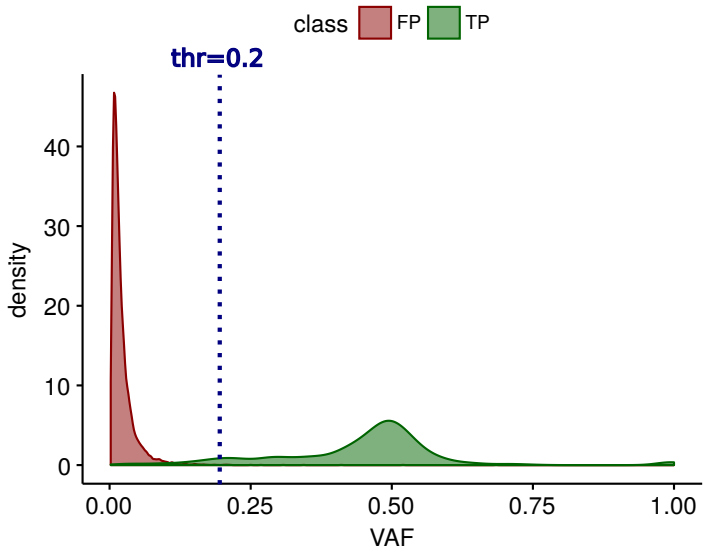


Figure 1.9 – Hard threshold filtering on the VAF of a set of false positive (VAF distribution is shown in red) and true positive (VAF distribution is shown in green) variants. VAF lower than the fixed threshold are removed, and therefore removed false positives correspond to the red distribution at the left of the threshold (gain of specificity) and removed true positives to the green one (loss of sensitivity). The accuracy of this variant filtering is given by the combination of these two removed sets.

The figure 1.9 is presenting the distribution of VAF of both true positive variants (in green) and false positive variants (in red) from real samples of whole-exome NGS data. A true positive is defined as a true mutation detected from the variant calling algorithm and a false positive is defined as a false mutation detected from the variant calling. A threshold of 0.2 has been chosen to remove potentially false variants, and the effect on sensitivity and specificity can be induced by respectively the true and false variants that would be removed by this statistic threshold. In addition to hard pre-defined thresholds on variant statistics,

GATK has implemented the Variant Quality Score Recalibration (VQSR) step [40]. This step consists in associating to each putative variant a new statistic that can be used then to filter the variants. The variant recalibration step uses a fit of a Gaussian mixture model on a set of *a-priori* true variants from external studies to compute a probability of being a true positive variant to each mutation identify in the current variant calling. Taking the example in figure 1.9, the VQSR would fit one Gaussian on the false positive variants (in red) and one gaussian one the true positive variants (in green), and then would determine the best threshold that separate the two distributions. This can be generalized in more than one dimension (here the VAF), leading to a fit of a Gaussians mixture. The main drawback of the GATK VQSR step is that it requires large input data, at least one WGS or more than 30 WES to be able to accurately annotate the variants [128].

Advanced filtering methods based on machine learning algorithms have been recently proposed to deal with remaining false positive subsequent to the variant calling step [135],[106], [133], [108], [8]. The aim of a machine learning algorithm is to use information on statistical variables (called "features") of known entities to predict the status of unknown entities. Machine learning methods work as the following:

- Definition of a set of known entities **E**
- Definition of a set of statistical features **F**
- For each known entities $e \in \mathbf{E}$ and each feature $f \in \mathbf{F}$, computation of f_e
- **Training** of a machine-learning model (*e.g* a random forest)
- For each unknown entities $e' \in$ **E'** and each feature $f \in \mathbf{F}$, computation of $f_{e'}$
- **Application** of the trained model on each e'

Figure 1.10 from [90] is presenting an example of such a machine learning application in genomic analyses. In its second version, Strelka has implemented a machine-learning-based variant scoring step, directly inside the variant calling [71]. This variant scoring uses a pre-trained random forest algorithm on multiple sequencing conditions of known germline variants from Platinum Genomes [44] sample NA12878 for the germline variant filtering and pre-trained on curated tumor-cell lines for somatic variant filtering. The most important features of the model are (1) the genotype probability computed by the core variant proba-

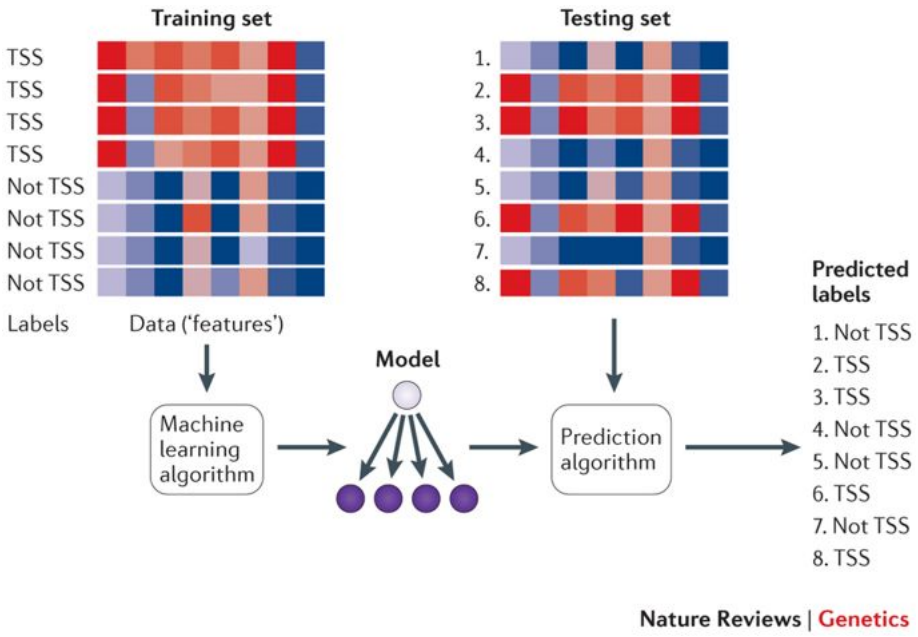


Figure 1.10 – A training set of DNA sequences is provided as input to a learning procedure, along with binary labels indicating whether each sequence is centred on a transcription start site (TSS) or not. The learning algorithm produces a model that can then be subsequently used, in conjunction with a prediction algorithm, to assign predicted labels (such as 'TSS' or 'not TSS') to unlabeled test sequences. In the figure, the red–blue gradient might represent, for example, the scores of various motif models (one per column) against the DNA sequence [90].

bility model, (2) root-mean-square mapping quality, (3) strand bias, (4) the fraction of reads consistent with locus haplotype model, and (5) the complexity of the reference context as measured by metrics such as homopolymer length. This step produces a single aggregate score for each putative variant, that can be used to remove potentially false observations.

Unfortunately, variant filtering methods based on machine learning algorithms present two major limitations. Firstly, these models are feature-fixed in the sense that the features incorporated in the model should be available for the dataset on which the model would be applied to predict the classes. Secondly, it is theoretically feasible for every type of data to construct a machine learning model: the main part consists in the decision of the features. But practically, it is much more difficult to construct the model: a sufficient amount of data should be available for every class and every feature to correctly train the model. These drawbacks make the variant filtering a difficult task due to its lack of genericity.

1.4 Early cancer detection

1.4.1 Definitions

An individual cancer can be scored depending on the anatomic disease extent, *i.e.* its size and its spreading. This is called the cancer *stage*. There exist many cancer staging systems depending on the cancer type, and one specific staging system can be used for every cancer type: the *TNM* classification developed by the Union for International Cancer Control [24]. This classification system attributes a stage to cancer depending on three features: the primary tumor site (the first reached tissue) T, the regional lymph node involvement N, and the metastatic spread M. A cancer stage is ranging from I to IV. According the NCI [1], an early stage cancer is a cancer that did not have spread to other part of the body than the primary site, this means that early stages typically group I and II. Detecting an early stage cancer is defining the *early cancer detection*.

1.4.2 Objective

A cancer diagnosed at early stages, *i.e.* with a small size and not spread in other tissues is more likely to be successfully treated. The aim of early cancer detection is to improve cancer patient survival, and it has been shown for at least some cancer types that the reduction of a cancer stage when cancer is detected is correlated with an increase of patient survival [29].

1.4.3 Methods

According to the World Health Organization (WHO), early cancer detection can be divided into two major components: early cancer diagnosis and population screening for cancer. Early cancer diagnosis aims at increasing the awareness of early sign of cancer, and this requires acts from a patient and the health care providers such as physicians. In this work we will focus on population screening for early cancer detection. General definition of screening is the usage of tests inside an healthy population aiming at detecting a particular disease on individuals that does not present any symptom corresponding to the disease. Population screening generally targets high-risk individuals, such as individuals presenting a germline

(heritable) genomic variant [59] or individuals expose to a particular carcinogen (*e.g.* tobacco smoke [105]), in order to increase reduce the impact of the false positive rate. Recently, the National Lung Screening Trial (NLST) reported a lung cancer mortality reduction of 20% associated with a low-dose Computed Tomography (CT) cancer screening [129].

A particular set of cancer screening tests that emerged in the last few years and that seems to be very promising are the *body liquid-based* tests, also known as *liquid biopsies* [45], [63]. The aim of such tests is to detect the presence of a tumor from the analysis of an individual body-liquid sample. This offers a non-invasive and more specific method than traditional CT scan [64]. Currently, two types of body liquid are studied to build early cancer detection screening tests: urine [98], blood [49], [25], [34] or even cerebrospinal fluid [94]. At the moment early detection based on urine sample is mainly studied in the case of urological cancer [98] whereas blood-based tests are not limited to a particular cancer type [45].

The challenge of these tests is to detect molecular biomarkers of cancer in liquid samples. These biomarkers include proteins, DNA and RNA (transcripted version of DNA that is converted into a protein). In the work presented here we will focus on DNA biomarkers from blood samples.

1.4.4 Circulating tumor DNA

The DNA biomarker that can be extracted from a blood sample is called ctDNA. It corresponds to the fraction of the Circulating cell-Free DNA (cfDNA), the set of DNA molecules coming from the degradation of cells and freely circulating in the bloodstream, that is attribuable to cancer cells (Figure 1.3, [37]).

If an individual has a tumor in a particular tissue, the tumor-derived genomic variations identified in the ctDNA can be considered as a proxy of the genomic variations present in the tumor. This gives the potential to ctDNA to be a cancer molecular biomarker. The major limitation of the ctDNA is its detectability. Indeed, in cancer patients, only a small fraction of the total amount of cfDNA corresponds to ctDNA, and this fraction can be as low as 0.01% [41]. This fraction varies depending on the cancer stage [19]. A promising clinical application of ctDNA as a cancer biomarker is its usage for the detection of small tumor in early stage can-

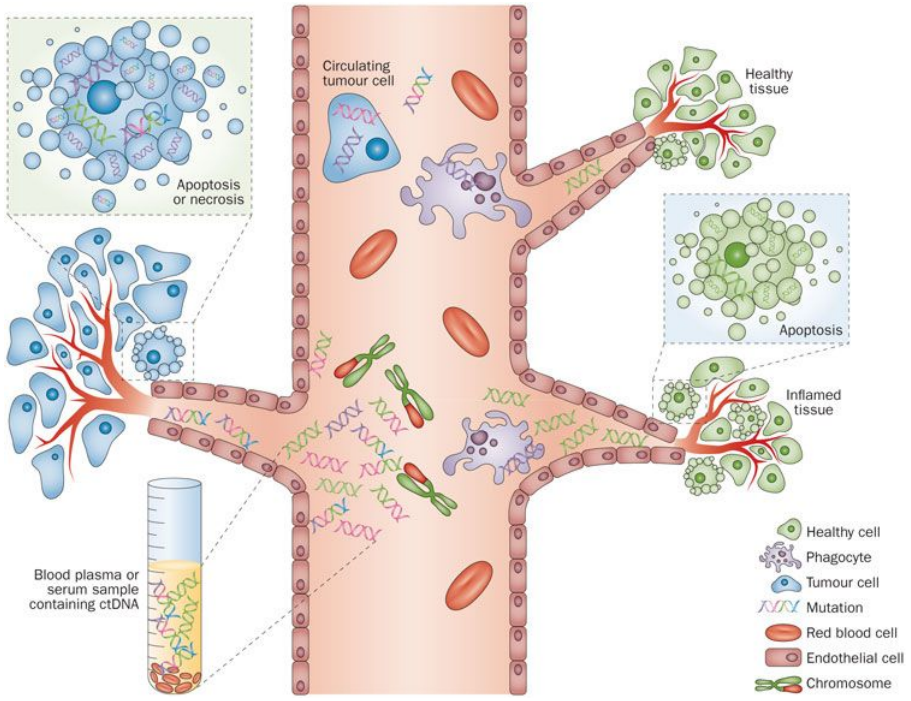


Figure 1.11 – Figure from Crowley et al. 2013 [37]. Extraction of cfDNA and identification of tumor-derived genomic variations from blood sample. Tumour-derived genetic alterations that can be detected in the blood include point mutations (consecutive purple, red, green and blue DNA strands), copy number fluctuations (red portion of chromosomes) and structural rearrangements (green and red DNA strands).

cers, before the appearance of any observable symptoms [19]. The development of ctDNA as a biomarker for early detection is currently an exciting scientific challenge [13].

1.4.5 CtDNA as a biomarker for early cancer detection

The first major component of the pertinence of the usage of ctDNA as a early cancer biomarker is the availability of the tumor-derived genetic material in the studied biological samples, *i.e.* is there a sufficient amount of ctDNA inside the blood sample from a cancer patient to be able to be detected? The notion of VAF can be used to answer this question. As exposed in the paragraph 1.1.3, the VAF is a proxy of the proportion of sequenced cells that are carrying the mutation corrected by the heterozygous or homozygous status of the mutation. As an example, define a blood sample that contains 1,000 cfDNA molecules from which 10 are ctDNA molecules. If the individual is homozygous for the mutation, the VAF will corresponds to $10/1,000 * 50\% = 0.5\%$, and if the individual is heterozygous for the mutation, the VAF will be $10/1,000 * 100\% = 1\%$. It has been reported in a recent study [3] that in patients

with detectable ctDNA from blood samples, the pathologic tumor size is correlated with the average VAF of ctDNA mutations (figure 1.12): higher the tumor size, higher the proportion of ctDNA molecules in the blood. Finally, by translation, because the tumor size is correlated with the cancer stage (by definition), the amount of ctDNA is correlated with the tumor stage: higher the stage, higher the amount of ctDNA, and higher the potential to detect it in the blood.

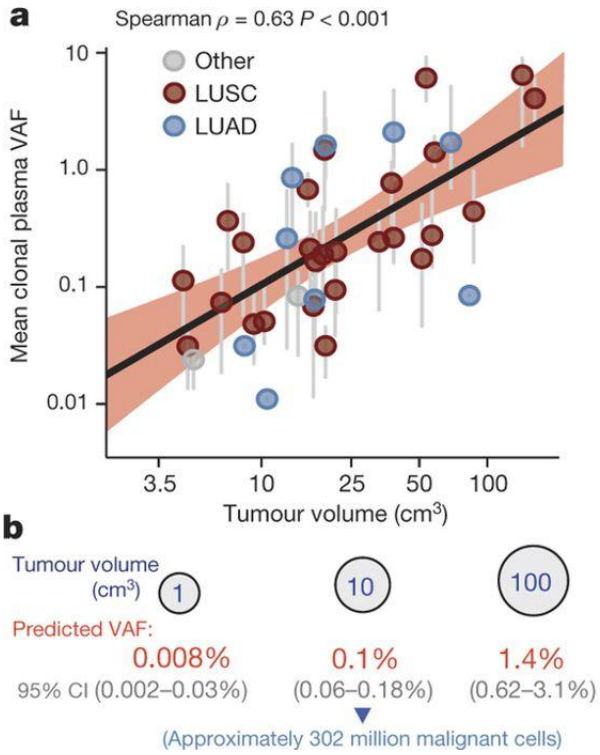


Figure 1.12 – Figure from Abbosh et al. 2017 [3]. Tumour volume cm3 measured by computed tomography (CT) volumetric analysis correlates with mean clonal plasma VAF. n=37, grey vertical lines represent range of clonal VAF, red shading indicates 95% confidence intervals.

Using ctDNA as an early cancer biomarker is also limited by the availability of algorithmic methods that can detect tumor-derived mutations from extracted cfDNA. Indeed, as mention previously, tumor-derived mutations present in cfDNA samples are found in low proportion an subsequently would harbor very low VAF. This means that it is necessary to use computational methods that present a good sensitivity when dealing with ctDNA data in the context of early cancer detection.

1.5 Global aim of the study

The aim of the thesis was to understand and identify the errors on genomic sequences found in NGS data, to enable the accurate detection of mutations. The thesis will not expose the detailed *causes* of errors, we only tried to build statistical models explaining their *appearance* in order to identify them. The main developments can be separated into two paired approaches: *variant calling* to detect candidate mutations in NGS data and *variant filtering* to boost the precision of the variant calling step. These developments were finally applied in a third chapter to four studies on circulating tumor DNA data in the context of biomarker development for early cancer detection.

Chapter 2

Dealing with systematic errors: needlestack, a multi-sample sensitive variant caller

Contents

2.1 Scientific context **34**

2.2 Scientific contribution **35**

2.2.1 Statistical algorithm 35

2.2.2 A robust implementation 37

2.2.3 Needlestack performance 38

2.3 Article A (Submitted in *Nucleic Acid Research*) **40**

2.4 Discussion **67**

2.1 Scientific context

NGS has revolutionized the way to infer genomic variation from human genomes by its ability to identify DNA sequences from multiple samples in a short time scale and for a reduced cost compared to the traditional Sanger sequencing (see Figure 1.3). Nevertheless, the NGS technology is more prone to sequencing errors. The comprehensive characterization of DNA variations by screening cancer genomes can help to understand cancer appearance and progression but also to identify predictive biomarkers such as ctDNA[19] and to study mutations from histological normal tissue [92], [93]. In 2015, the International Cancer Genome Consortium launched a large benchmarking operation with the objective of identifying and resolving issues of detecting variants from NGS data [10]. The conclusion of this study was that detecting somatic mutations in cancer genomes remains a considerable challenge due to the high complexity of cancers. This challenge is exacerbated when trying to identify DNA mutations in very low abundance that presents therefore very low VAF that are deviating from the expectation 1.1.3. This deviation makes the detection of mutations a complex task due to a lack of theoretical models that could infer the candidate DNA variations. In addition of such difficulties, there exists still a lack in availability of integrative variant calling methods that can identify all possible types of genomic variations, *i.e.* simultaneously germline mutations, somatic and potentially low abundance mutations [146].

In this chapter, we propose to study the *systematic* errors found in NGS data. For this, we designed and implemented a robust model of systematic errors. True mutations are predicted as diverging from this model. To build the statistical model of systematic errors, we use the information of read counts across multiple sample to obtain a powerfull approach that is able to precisely estimate the SER. Because the SER is varying both across DNA positions and depending on the base change of the candidate mutation, we propose to estimate it for each pair of position and nucleotide variation. As levels of error can reach the proportion of reads truly mutated, the variant calling process is similar to finding a needle in a needlestack, and therefore we named our pipeline *needlestack*.

2.2 Scientific contribution

2.2.1 Statistical algorithm

Errors found in NGS data can be represented as count data. A classic way to model individual count data is to use a Binomial distribution, and, to model multiple binomial distributions, the regression model is applied. The binomial regression is used in statistics to model N independent predictor variables X_i and their response variables Y_i if Y_i is a results of n Bernoulli trials. A Bernoulli trial is a random experiment with only two possible outcomes : success or failure. In the context of modelling sequencing errors, a success would be an error and a failure a correct sequenced base. In a serie of n Bernoulli trials that would be in our case the sequenced reads, each trial has the same probability of success p. In the binomial regression models, each Y_i corresponds to the number of successes of a serie of X_i Bernoulli trials with a probability of success equals to p:

$$Y_i \sim B(X_i, p)$$

with p corresponding to the SER in our case.

In our case, because the sequencing error rate is expected to be very low, X_i tend to be high and p to be low. In such a way, the binomial distribution approaches a Poisson distribution with parameter λ, with $\lambda = X_i * p$, and the sequencing error rate can be modelled using a Poisson regression.

Because NGS error count data are over-dispersed data [62], the Poisson regression is not adapted to model the sequencing errors because of the assumption of equal mean and variance. By contrast to the Poisson regression, the negative-binomial regression takes into account the over-dispersion:

$$AO_{ijk} \sim NB(\mu_{ijk}, \sigma_{jk})$$

with $i = 1...N$ the index of the sample from a sequenced panel of size N, j the genomic position and k the potential nucleotide change: $k \in (A, T, C, G, ins, del)$. ins and del are de-

noted respectively the set of observed insertions and deletions in the sequenced data. The over-dispersion parameter is denoted as σ_{jk} and $\mu_{ijk} = e_{jk} * \mathrm{DP}_{ijk}$ is corresponding to the expected number of reads supporting alteration k across samples with a coverage equals to DP_{ijk}.

Due to the fact that, in addition to errors, there are potentially true mutations in the sequencing data set that can influence the fitting of the regression model as being outliers, we used a previously published *robust* negative binomial regression from Aeberhard *et al* [7]. We adapted the original implementation as an R package as the following in order to fit correctly our data and to improve the speed of execution:

- we modified the model logarithm link function into a linear link between AO_{ijk} and DP_{ijk}
- we adapted the code to constrain the intercept coefficient of the regression to be null, *i.e.* we expect zero error reads if the coverage is null:

$$\mathrm{AO}_i = e_{jk} * \mathrm{DP}_{ijk} + 0$$

  with e_{jk} corresponding to the SER at the position j for the alteration k.
- we proposed an estimation of initial SER that is used in their maximum likelihood approach equals to the median of individual error rates e_i after the Tukey's outlier filter [134]:

$$e_{init} = mean(\mathbf{e})$$
$$\text{with } \forall i, e_i \in \mathbf{e} \text{ and } e_i \leqslant \mathrm{Q}_3 + 1.5\ \mathrm{IQR}$$

  with Q_3 the third quantile and IQR the interquartile range equals to the difference between the first and the third quartiles Q_3.
- instead of using sums to compute integrals in the maximum likelihood estimations, that can be time consuming in case of high values of DP_{ijk}, we used interpolation of the points with the *spline* function in R

2.2.2 A robust implementation

To implement needlestack, we emphasized on four main concepts:

- efficiency
- robustness
- reproducibility
- user-friendliness

To follow these guidelines, we used *nextflow*. Nextflow is a Domain Specific Language that enables the writing of scalable and reproducible scientific workflows in an easy and efficient manner [132]. To maximize the efficiency of the computations, we allow the user to run needlestack in parallel. Indeed, our needlestack algorithm is launched independently on each position, and therefore it could be run in parallel on sets of positions, so a user can input both a set of positions and a number of sets to split them. To maximize the robustness of the pipeline, we benefited from the nextflow language that enables the deployment of the pipeline on multiple types of environments such as local computers, HPC (high performance computing) environments or cloud instances such as Amazon Web Services instances. The reproducibility is maintained by the interaction between nextflow and GitHub, a web plateform that store the code of the pipeline with versioning. Reproducibility is also maintained by the availability of Docker [20] and Singularity [76] containers, which allow to package up the pipeline with all its dependencies into a fixed version.

Finally, we also maintained user-friendliness by providing a pipeline which is runnable in only one command line, without the need to install each specified dependency if Docker or Singularity is available in the running environment.

The complete pipeline is defined as a chain of piped commands (see figure 2.1): first, *samtools mpileup* [84] command computes for each of the input BAM files, the list of read nucleotides overlapping the input positions. Then, samtools output is translated into a big table with samples in columns and positions in lines. Finally, needlestack uses its own R script to run the variant calling and produce a Variant Call Format (VCF) [39] that will be merged with all the others parallel VCFs. Needlestack versioned source code is maintained and freely available on GitHub (https://github.com/IARCbioinfo/needlestack).

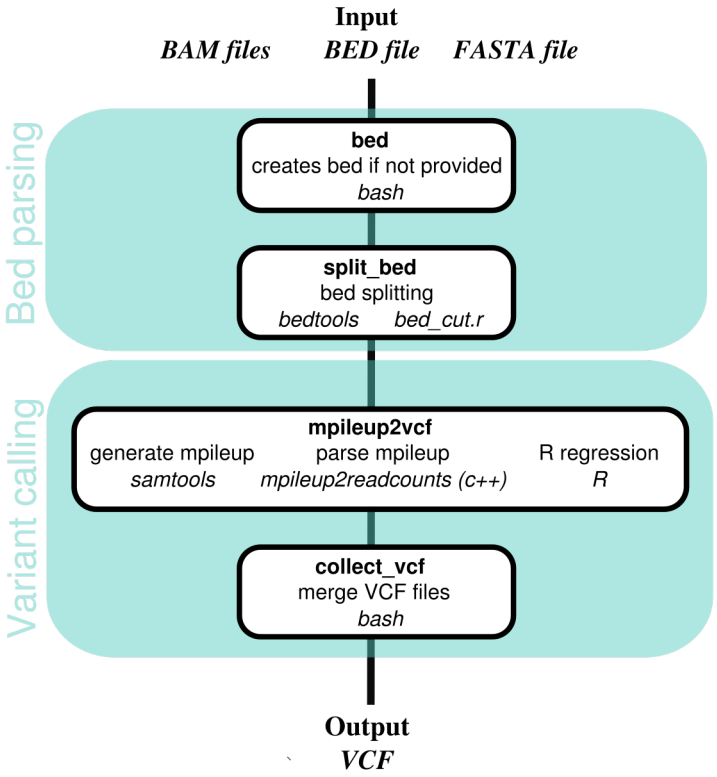


Figure 2.1 – The needlestack workflow.

2.2.3 Needlestack performance

We assessed the performance of needlestack independently for rare germline variant calling and (potentially) low VAF somatic variant calling using both simulated and real data.

To evaluate the accuracy of needlestack to identify rare (in the sense "low population frequency") germline mutations, we used a total of 62 WES of normal samples. We compared the needlestack calls to GATK Haplotype Caller (GATK-HC) [96], [106] using both raw mutations and mutations validated by a gold standard dataset defined as an Illumina bead-array data (available for 33/62 samples). Needlestack and GATK-HC sensitivities were found to be similar when using our gold standard dataset (around 95% for both methods). Without taking into account the bead-array that can be bias toward evident variations, we reported a concordance rate of 97.3% for SNVs and 70.3% for the indels.

In term of somatic mutation performance, first, we used a total of 35 lung cancer patient samples of both cfDNA and tumor in order to validate in the tumor the called cfDNA mutations that are deleterious (we expect that a deleterious cfDNA mutation should come

from a tumor). Each of deleterious cfDNA mutations was validated.

We also used simulated data in order to estimate the performance of needlestack on multiple VAF, down to 0.01%. For this, we used 125 plasma samples from healthy patient sequenced at the *TP53* gene. We introduced 1,000 *in-silico* mutations using BAMsurgeon [47] and repeated this process 10 times to increase the power of our performance computations. Source code to simulate these tumor data is available on github: `https://github.com/IARCbioinfo/bamsurgeon-nf`. We showed that the sensitivity of needlestack depends of the combination of the target VAF and the SER, and for example needlestack detects more than 99% of mutations with a VAF greater than 1%. We compared our results with the shearwaterML algorithm [92], [57], a competing variant caller suitable to detect very low VAF. shearwaterML is also based on multiple sampling to estimate the SER at each pair of position and nucleotide change, but, contrary to needlestack, shearwaterML uses a beta-binomial regression instead of a negative binomial regression. There is no robust version of the beta-binomial regression described in the literature, and therefore to keep robustness in case of outliers (true mutations), shearwaterML proposes to use an *a-priori* threshold on the SER to first remove potentially true variants and then fit efficiently the model without outliers. We then showed that, contrary to shearwaterML, needlestack false discovery rate does not depend of the SER.

2.3 Article A (Submitted in *Nucleic Acid Research*)

Needlestack: an ultra-sensitive variant caller for multi-sample next generation sequencing data

Tiffany M. Delhomme[1], Patrice H. Avogbe[1], Aurélie Gabriel[1], Nicolas Alcala[1], Noemie Leblay[1], Catherine Voegele[1], Maxime Vallée[2], Priscilia Chopard[2], Amélie Chabrier[1], Behnoush Abedi-Ardekani[1], Valerie Gaborieau[2], Ivana Holcatova[3], Vladimir Janout[4], Lenka Foretova[5], Sasa Milosavljevic[6], David Zaridze[7], Anush Mukeriya[7], Elisabeth Bambrilla[8], Paul Brennan[2], Ghislaine Scelo[2], Lynnette Fernandez-Cuesta[1], Graham Byrnes[9], Florence Le Calvez-Kelm[1], James D. McKay[1*], Matthieu Foll[1*✉]

[1] Genetic Cancer Susceptibility Group, Section of Genetics, International Agency for Research on Cancer (IARC-WHO), 150 cours Albert Thomas, 69008, Lyon, France
[2] Genetic Epidemiology Group, Section of Genetics, International Agency for Research on Cancer (IARC-WHO), 150 cours Albert Thomas, 69008, Lyon, France
[3] Institute of Public Health and Preventive Medicine, Charles University, 2nd Faculty of Medicine, Prague, Czech Republic
[4] Faculty of Health Sciences, Palacky University, Olomouc, Czech Republic
[5] Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic
[6] International Organization for Cancer Prevention and Research (IOCPR), Belgrade, Serbia
[7] Russian N.N. Blokhin Cancer Research Centre, Moscow, The Russian Federation
[8] Centre Hospitalier Universitaire de Grenoble Département d'Anatomie et Cytologie Pathologiques, CS 10217 Grenoble, France
[9] Section of Environment and Radiation, International Agency for Research on Cancer (IARC-WHO), 150 cours Albert Thomas, 69008, Lyon, France
[*] Jointly supervised this work
[✉] To whom correspondence should be addressed. Email: follm@iarc.fr

ABSTRACT

The emergence of Next-Generation Sequencing (NGS) has revolutionized the way of reaching a genome sequence, with the promise of a potentially comprehensive characterization of DNA variations. Nevertheless, detecting somatic mutations is still a difficult problem, in particular when trying to identify low abundance mutations such as subclonal mutations, tumour-derived alterations in cell-free DNA or somatic mutations from histological normal tissue. The main challenge is to precisely distinguish between sequencing artefacts and true mutations, particularly when the latter are so rare they reach similar abundance levels as artefacts. Here, we present needlestack, a highly sensitive variant caller, which directly learns from the data the level of systematic sequencing errors to accurately call mutations. Needlestack is based on the idea that the sequencing error rate can be dynamically estimated from analyzing multiple samples together. We show that the sequencing error rate varies across alterations, illustrating the need to precisely estimate it. We evaluate the performance of needlestack for various types of variations, and we show that needlestack is robust among positions and outperforms existing state-of-the-art method for low abundance mutations. Needlestack, along with its source code is freely available on the GitHub plateform: https://github.com/IARCbioinfo/needlestack.

INTRODUCTION

Massive parallel sequencing, or next generation sequencing (NGS), has revolutionized the manner in which genetic variation can be explored, due to a large increase in throughput compared to the

traditional Sanger sequencing, and at a greatly reduced cost per sequenced base. However, because these new technologies are prone to errors, identifying genetic variants from NGS data remains a considerable challenge (1). This is particularly true in heterogeneous samples, where the variant allelic fractions (VAF, the ratio of the number of sequencing reads carrying the mutant allele to the total read count) deviate away from the expectations of a diploid genome (0%, 50% or 100% for the three possible diploid genotypes), until the point where the mutant alleles make up only a small fraction of the sequenced reads, approaching the background error rate. Nevertheless, robustly identifying low VAF sequence variants in such heterogeneous settings can be highly informative, for example allowing insights into the clonal evolution of tumours (2), analyzing the cell-free DNA in order to identify tumour-derived footprints (3), or evaluating somatic mutations in histologically normal material (4).

The error rate of next generation sequencing is known to vary across DNA base pairs and even across multiple base changes at the same DNA position (5,6). NGS errors originate from many of the steps in the sequencing process, stemming from the quality of the template DNA, its subsequent fragmentation, the library preparation, the base calling, or the alignment step subsequent to the sequencing of raw reads. Some of these errors have a tendency to reoccur consistently across samples whereas others have a more unpredictable appearance. The net effect of NGS being made of errors from multiple sources is that they become highly difficult to distinguish or correct for (7). Variant identification methods that consider this highly variable error pattern may improve our ability to robustly detect true sequence variants even when their abundance is low. Most current algorithms use a probabilistic model on VAF independently across samples to distinguish between sequencing artifacts and true variations (8), while methods to detect low abundance mutation, like shearwaterML (9,10), propose to benefit from the shared knowledge on errors across samples, but are limited by the requirement of a prior threshold on the error rate.

Here, we have explored the approach of using multiple samples analyzed concurrently to develop an error model for each potential base change. Sequence variants are identified as outliers relative to this robust error model. This method, which we call needlestack, allows the definition of sequencing variants in a dynamic manner relative to the variable error pattern found in NGS data, and particularly variants that are rare in the sequenced material. By combining this method with additional laboratory processing for further error correction (11) and very deep next generation sequencing, we are able to robustly identify VAFs well below 1% while maintaining acceptable false discovery rates. We conducted multiple rigorous performance estimations and comparisons with methods for both somatic and germline variant detection. We deployed our pipeline focusing on efficiency and robustness using the Domain-Specific Language (DSL) nextflow (12), and on reproducibility by providing Docker and Singularity images. Source code is versioned and freely available on GitHub (https://github.com/IARCbioinfo/needlestack).

MATERIAL AND METHODS

Needlestack overview

Needlestack estimates for each candidate alteration, *i.e.* each pair of position and base change (the three non-reference nucleotides and each observed insertions and deletions) the systematic sequencing error rate across a series of samples, typically more than twenty to ensure a reasonable estimation of this metric. Then, for each sample, it computes the *p*-value for the observed reads under the null hypothesis of this estimated model of errors, and transforms this *p*-value into a Phred-scale *Q*-value reported as a variant quality score (QVAL) for the candidate mutation. As such it measures the evidence that the observed mutation is not explained by the error model, and should therefore be considered a mutation.

Needlestack takes as input a series of BAM files, and is based on three main piped processes, the generation of the mpileup file containing read counts at the target positions using samtools (13), the reformatting of this file into readable tabulated file and finally the estimation of the error model using our R regression script (see below) coupled with the computations of Q-values (supp figure 1). Needlestack is highly parallelizable as input positions are analyzed independently. As an output, needlestack provides a multi-sample VCF file containing all candidate variants that obtain a QVAL higher than the input threshold in at least one sample, general information about the variant in the INFO field (*e.g.* error rate estimation, maximum observed QVAL) and individual information in the GENOTYPE field (*e.g.* QVAL of the sample, coverage of the sample at the position).

The Needlestack algorithm

Let *i=1...N* be the index of the sample taken from an aligned sequenced panel of size *N*, *j* the genomic position considered and *k* the potential alteration, with $k \in (A, T, C, G, ins, del)$, *ins* and *del* covering respectively every insertion and deletion observed in the data at position *j*. Let DP_{ij} denote the total number of sequenced reads at position *j* for the sample *i*, AO_{ijk} the reads count supporting alteration *k* and e_{jk} the corresponding error rate. We model the sequencing error distribution using a negative binomial (NB) regression:

$$AO_{ijk} \sim NB(\mu_{ijk}, \sigma_{jk})$$

with σ_{kj} the over-dispersion parameter and $\mu_{ijk} = e_{jk} * DP_{ijk}$ corresponding to the expected number of reads supporting alteration *k* across samples with a coverage D_{ijk} . A robust negative binomial regression method (14) is employed to ensure that the outliers from this error model, such as true mutations, are not biasing the regression parameters estimates. This model is based on a robust weighted maximum likelihood estimator (MLE) for the over-dispersion parameter σ_{jk} . We modified the original implementation of this regression to fit the need of our model here with: (*i*) a linear link function, (*ii*) a zero intercept, as a null coverage will exhibit a null read count, and (*iii*) an approximation of the bounding functions to allow the MLE to run efficiently for high coverage data (see supplementary methods).

For each position *j* and alternative *k*, we perform this robust negative binomial regression to estimate parameters e_{jk} and σ_{kj} . We then consider a sample *i* as carrying a true mutation *k* at the position *j* when being an outlier from the corresponding error model. We calculate for each sample a *p*-value for being an outlier using the estimated parameters that we further transform into *q*-values using the Benjamini and Hochberg procedure (15) to account for multiple testing and control the false discovery rate.

Importantly, because true mutations are identified as the outliers from the error model fitted using a robust regression, this approach is more suited to detect rare mutations. Common mutations (for example germline SNPs with common allele frequencies) will be observed in the error model and therefore not detected as outliers by needlestack. In practice we found that mutations with a minor allele frequency below 10% can be accurately detected (see below). Additionally, while allowing over-dispersion, our model assumes that the error rate e_{jk} is constant across samples for a given alteration. This means that it should be applied to a homogeneous series of samples (that is prepared using comparable laboratory techniques and sequencing machines etc.). Importantly other types of errors that have less tendency to reoccur uniformly across samples are identified by needlestack as outliers.

Sequencing data for performance evaluation

125 cell-free DNA (cfDNA) samples from healthy donors were used to study the distribution of error rates estimated by needlestack and to estimate its accuracy to detect low VAF using *in-silico* mutations. We also obtained 46 cfDNA samples from 18 small-cell lung cancer (SCLC) patients and 28 squamous-cell carcinoma (SCC) patients, two cancer types that harbour a high prevalence of *TP53* mutations (respectively 99% (16) and 81% (17)). In order to validate in the tumour the low VAF mutations identified by needlestack in the cfDNA, we also sequenced tumour samples for these patients. Each of the cfDNA samples was sequenced in the *TP53* exonic regions (exons 2-11, which corresponds to 1,704 base pairs with a median coverage of around 10,000X) using the IonTorrent Proton technology, in two technical duplicates in order to account for potential errors during library preparation. Details about cfDNA sequencing steps and tumour sequencing method are provided in the Supplementary Material.

Additionally, we performed whole-exome sequencing (WES) from the blood of 62 samples from an independent cohort in order to estimate the performance of needlestack on germline mutations. As a gold standard, we used genotypes derived from Illumina SNP array (Illumina 5M beadarray) that were available for 33 of these 62 samples.

Comparison with other variant callers

We used BAMsurgeon software (18) to introduce SNVs at varied VAF in the 125 cfDNA samples in *TP53* in order to benchmark and compare the method through *in-silico* simulations. BAMsurgeon presents the advantage of synthetic benchmarking methods that allow the simulation of mutations for

which gold standards don't exist to evaluate the performance (here low VAF, that are in addition challenging to validate), while maintaining the real data background such as the true error profiles. We introduced 1000 SNVs at random positions in the gene in random samples, and we replicated this process in ten batches. As each sample has been sequenced twice, we introduced each *in-silico* mutation in the two technical duplicates of a sample. We took benefit from the variable coverage among samples and genomic positions to study the sensitivity of our method down to VAF=10^{-4}. For each mutation *m*, the VAF was simulated using a log-uniform distribution: $VAF_m = 10^{-u}$ with $u \sim uniform(0,4)$. Mutations were only introduced at positions where at least five mutated reads would be observed. This means that a mutation with a VAF=10^{-4} would be introduced only in positions with a coverage of at least 50,000X. To compare needlestack with a similar variant caller, we ran ShearwaterML (4,10) on the same ten batches (see supplementary methods). ShearwaterML is based on a beta-binomial regression and requires an *a-priori* threshold *t* for the error rate. ShearwaterML excludes each sample having a number of alternative bases higher than *t**coverage, aiming at removing potential true mutations that act as outliers in the regression to robustly estimate the error rate. To compute the global performance of both methods, the ten simulation batches were merged, and only mutations detected in both technical duplicates were considered. In-silico simulations were repeated for 1-base pair insertions and deletions (indels) for needlestack. In this case, the total number of *in-silico* mutations was reduced to minimize the potential alignment artifacts created by the introduction of two indels close together. For that, using the same initial data, 100 insertions and 100 deletions were added again in ten simulations batches (total of twenty batches).

To estimate the ability of needlestack to detect rare germline variations, we used the 62 WES from blood samples. Needlestack variant calling was performed using our germline recommendations (see supplementary methods). GATK variant calling was performed using HaplotypeCaller best practice workflow (see supplementary methods). From the 3,446,898 bead array non-reference genotypes (0/1 or 1/1) distributed over 113,232 positions in the 33 individuals, we selected 20,439 genotypes with a sufficient coverage (see supplementary methods). In a second part, to account for possible bias in the array, variant calls from both needlestack and GATK were compared independently of the array data, on a total of 44,314,972 exonic positions. To compare only rare germline variants, we removed common variants from each calling set (bead array, GATK calling and needlestack calling, see supplementary methods).

Error rate estimation

To estimate the error rate variability across positions, we computed with needlestack the sequencing error rates from two data sets of the *TP53* gene sequenced with two different technologies (on the 62 blood samples and on the 125 cfDNA samples). Error rates were estimated at each position of the gene and for each substitution, totaling *1704*3=5,112* values. We were then interested in estimating the contribution of each possible nucleotide change on the error rate. We therefore computed, for

each error-rate range *e* in [{10^{-5},10^{-4}};{10^{-4},10^{-3}};{10^{-3},10^{-2}};{10^{-2},10^{-1}}] and for each possible base change *b* in [G>T, C>A, ..., A>C, T>G]:

$$prop_{e,b} = \#ER_{e,b} / \#ER_e$$

with $\#ER_{e,b}$ being the number of estimated error rates in the class *e* observed for a base change *b*, and $\#ER_e$ being the total number of estimated error rates in the class *e.*

In the case of the Ion Torrent sequencing, we observed a sufficiently high number of single nucleotide variations (SNVs) (n=5,112) to also compute the distribution of error rate depending on the 96 possible SNVs taking into account the preceding and following bases to evaluate the effect of the sequence context. Similarly, the high number of insertions (n=7,662) and deletions (n=1,724) detected allowed us to also compute the distribution of estimated error rates (i) as a function of the length of the inserted/deleted sequences; and (ii) as a function of the length of homopolymer regions for the insertion/deletion of one base pair.

cfDNA and matched tumour analysis for validation

Observed deleterious mutations in the *TP53* gene of a cfDNA lung cancer patient are generally expected to be derived from their tumour (but see Fernandez-Cuesta et al. 2016 Ebiomedicine) (19). Therefore we used the tumour samples as a proxy for validation of the identified cfDNA mutations. To limit our false discovery rate, we considered only cfDNA mutations that passed post-calling filters, *i.e.* a RVSB (Relative Variant Strand Bias) (19) lower than 0.85, no high-VAF variant (*i.e.* a VAF ten times higher than the candidate mutation) within 5 base-pairs upstream or downstream, and a VAF higher than 10% if the mutation is found in a low confidence base change (*i.e.* where technical duplicates don't cluster together; see Supplementary Methods). We independently performed the needlestack variant calling on the cfDNA samples and the matched tumour samples.

RESULTS

Sequencing error rates depend on the alteration type

Globally, 95% of the error rates across alterations were estimated as lower than $10^{-2.5}$ in both sequencing technologies (Figure 1A). Nevertheless, the error rates varied importantly across the target sequences and alterations. For the amplicon-based Ion Torrent sequencing, transitions had 5-fold higher error rates than transversions (Figure 1A), on average, although not clearly influenced by the sequence context when considering the flanking 3′ and 5′ bases (Figure S2). For exome-capture sequencing, a bulk in the distribution of transversion-like errors is observed at an error rate in the order of $10^{-2.5}$ (Figure 1A). When looking at the proportion of different nucleotide substitutions across multiple ranges of sequencing error rates (Figure 1B), we observed that in this range (10^{-2} – 10^{-3}) the majority of substitutions correspond to G>T transversions, previously reported and suggested to be related to DNA sonication(20).

As previously reported, we observed a large number of indels (9,389) in the Ion Torrent sequencing data (21). We found that the error rate is dependent of their length: long indel (with a size greater than

3bp) error rates are around 100-fold lower than 1bp indel error rate (Figure S3A). As previously reported (21), the error rate also increases with the length of homopolymer region, reaching 1% for repetitions of four nucleotides (Figure S3B).

Variant detection limit depends on the error rate

Importantly, errors identified in the previous section are classified as such by needlestack, and not as potential variants, even when the error rate is high, as opposed to traditional variant callers considering samples individually and that rely mostly on the VAF (20). Figure 2A illustrates a position at which needlestack identifies a high error rate ($e_{jk}=3.8$) without reporting any variant, even though alternate reads are observed in individuals VAF's up to ~ 9%. Figure 2B illustrates a position with a very different estimated error rate ($e_{jk}=10^{-4}$) where a putative very rare variant is identified. It is also noteworthy that the variant identified in Figure 2B has a VAF ten times lower (10^{-3}) than the error rate estimated in Figure 2A, and thus the sensitivity to detect a variant is considerably improved at the site with the lower error rate, highlighting the need to quantify the error rate distributions for each candidate mutation independently.

Technical replicates reduce low VAF false calls

We noted that the majority of variants detected by needlestack in the cfDNA of healthy patients harbour a particularly low VAF, typically under 0.5% (Figure 3A, black solid line). Importantly, the majority of these variants are not present in a second library preparation (a technical duplicate) of the same sample (Figure 3 blue line). Such variants illustrate an additional type of errors found in NGS data that do not consistently re-occur in the samples and that are not validated when sequencing a technical replicate of the sample, for example those introduced by polymerase chain reaction (PCR) amplification errors. These non-systematic artefacts are not expected to be captured by our error model and should be detected by needlestack as outliers (see Figure 2C for such an example). Importantly, we showed that this high number of calls not validated in a technical replicate of the sample is not dependent on our method (Figure 3A, blue lines). Subsequently, here, for the evaluation of needlestack's ability to detect efficiently low VAF mutations, we added the condition that variants are also detected in the technical duplicates to account for this type of error (Figure 3A, blue line).

Performance evaluation using *in-silico* simulation of somatic mutations

From the 10,000 mutations introduced by BAMsurgeon, needlestack detected 5% of mutations with a VAF lower than 0.1%, 51.4% of mutations with a VAF between 0.1% and 1%, and 99.4% of mutation with a VAF higher than 1%. As expected, the sensitivity of needlestack is highly dependent on the sequencing error rate. Indeed, needlestack does not call a mutation if the sequencing error rate for that alteration is greater than or in the same range as the VAF of the candidate mutation (Figure 4). As

an example, needlestack detected 0%, 6.5%, and 47.8% of SNVs with a VAF of 0.1% at positions where the sequencing error rate was higher than 0.1%, between 0.1% and 0.01%, and lower than 0.01%, respectively. When comparing needlestack and shearwaterML, we found that globally needlestack sensitivity was higher than that of ShearwaterML, and for example, ShearwaterML detected 7.7% of all inserted mutations with a VAF at 10^{-3} whereas needlestack detected 16.8% of these mutations. Given *t* the shearwaterML *a-priori* threshold on the sequencing error rate (Figure 3B red line) and *e* the observed sequencing error rate, we showed that the false positive rate of shearwater is markedly increased when *t>e*, whereas needlestack's false positive rate is stable across the whole range of error rates (Figure 3B).

Detection of tumour-derived mutations from cell-free DNA

Next, we tested needlestack's detection of very low VAF mutations in a biologically relevant setting. For this we screened cfDNA extracted from plasma samples from 35 lung cancer patients where the matched tumour sample was analyzed concurrently, and considered the concordance between the identified variants. A total of 22 *TP53* mutations from 18 samples (9 SCLC and 9 SCC) where identified in the cfDNA. 16/22 (70%) mutations were called in the tumour of the same patient. All the 12/22 cfDNA mutations considered as deleterious (*i.e* indels, non synonymous SNVs with a REVEL score higher than 0.5, stopgain or stoploss variants) (22) were present in the tumour. cfDNA and tumour VAF were found to be moderately correlated, which is concordant with previously reported results (23) (Pearson correlation coefficient equals to 0.59, Figure S4A). Details of the 22 cfDNA mutations and corresponding observations in the tumour matched samples are provided in supplementary table S1. The needlestack plots of a low VAF cfDNA mutation validated in the tumour are shown in supplementary Figure S4B.

Application to germline variant calling

For rare germline variants from 33 whole exomes, needlestack has a sensitivity of 95.64% to detect non-reference genotypes when using bead array data as a gold standard, which is quite similar to the GATK-HC Haplotype Caller results (95.48%). GATK-HC and needlestack variants concordant with the bead array (19,515 of the 20,439 variants) had VAF distributed around 50% and 100%, as expected for germline variants (Figure 5A). Most of the few calls that were not validated in the array were also centred around 50% and found by both variant callers, implying that they certainly contain additional heterozygotes that the SNP array failed to detect. Finally, the majority of variants not identified with NGS had no sequencing reads supporting the alternative allele detected by the array (841/892 variants), suggesting that these variants are potentially false positive results from the SNP array (Figure 5A).

Because SNP arrays are biased toward sites amenable to the design of Illumina BeadArrays (24), we also undertook needlestack and GATK germline genotyping of SNVs and indels calls across 62 exomes. Respectively 97.3% and 70.3% of the SNV and indel calls were concordant (Figure 5B-C)

with VAFs around 50%, whereas the genotypes identified uniquely by one of the two methods tended to have low VAF. For indel calling, 46% of calls unique to needlestack and 34% of calls unique to GATK are more than 10bp long, compared to only 12% of common calls. This suggests that discrepancies among the methods can be partially explained by longer indels difficult to align and call. For 66% of uniquely called indels by GATK-HC, no alternate reads were present in the BAM file used by needlestack, suggesting divergences in the assembly steps (haplotyper Caller versus ABRA). Interestingly, for 52% of the SNVs detected by GATK-HC and not by needlestack, needlestack estimated an error rate higher than 1%, pointing to possible false positives in the GATK calls (Figure S6).

DISCUSSION

The needlestack method is based on the notion that, as error rates strongly vary along the genome, their dynamic estimation from multiple samples, for each potential base change at a given DNA position, may assist in identifying sequence variants. Here, we have demonstrated that, even within a single gene (*TP53*), and even if the sequencing error rate is generally low, it varies importantly across positions and base changes (Figure 1). Needlestack implements a robust negative binomial regression for this purpose, and the ability of the method to identify variants will be dependent upon the error rate at that particular site and for that base change. By identifying sequence variants as outliers relative to the error model, needlestack maximizes the sensitivity to detect variants in a dynamic manner relative to the error rate in that particular setting. As such, low allelic fraction variants are identified from sites with low errors rates, whereas in settings where error rates are high, needlestack maintains reasonable false discovery rates (Figures 3 and 4).

We have benchmarked our method using both simulated and real data from different sequencing platforms. First, we have tested our method on low VAF mutations using BAMsurgeon to generate *in-silico* mutations and have compared our findings to variants identified by a similar rare variant orientated algorithm shearwaterML. We have shown that our method outperforms shearwaterML for VAF lower than 10^{-2}. We also have shown that the performance of shearwaterML highly depends on the difference between the error rate *e* and the error rate *a-priori* threshold *t* (see methods for details). Contrary to shearwaterML, needlestack's false discovery rate is not dependent on the sequencing error rate. In addition, needlestack also considers indel mutations. For this type of variant, the sensitivity of needlestack is slightly reduced compared to SNVs (Figure S5A-B). This is potentially due to the increased complexity of the assembly step around indels compared to SNVs. Moreover, needlestack detects a high number of indels replicated in the two technical duplicates that were not *in-silico* introduced (around 8 by samples in average), whereas *TP53* is not expected to harbour many indels in healthy patients. These mutations can be moderated using a filter on the strand bias, as previously reported (Figure S5C-D) (25).

The true specificity of needlestack cannot be achieved with BAMsurgeon simulations, due to a probably very low presence of true mutations in the cfDNA of healthy patients that is difficult to determine *a priori* (19). We therefore have estimated the validation rate in the tumour of deleterious

cfDNA mutations identified by needlestack in 35 lung patient cfDNA samples. All of these 12 mutations were validated in the tumour.

Finally, we have benchmarked needlestack on germline mutations using SNP array data to validate the mutations detected in WES of 33 individuals, and showed an excellent concordance when results are compared with both a SNP array as a gold standard set and calls from GATK HaplotypeCaller. This illustrates that needlestack, even if based on a totally different approach to detect variants, can reach similar performance to state-of-the-art germline variant callers.

The needlestack method nevertheless has several limitations. Even though needlestack is extremely sensitive, it is suited to detect rare mutations rather than common germline variants or highly re-occurring hotspot mutations. Adding an *a-priori* threshold for the error rate (extra_robust mode – see supplementary methods) can partially offset this limitation, but is only applicable to particular situations for the reasons explained above. More importantly the inherent logic of the needlestack approach corrects for errors that have a tendency to reoccur, as such errors that are rarer are identified as outliers in the regression. Following this, needlestack does not correct for sample-specific artifacts such as (i) (sample specific) stochastic alignment errors and we recommend to use it in conjunction with an assembly based re-alignment method (26); (ii) polymerase errors introduced in PCR amplification step; (iii) complex errors leading to features like strand bias. Such errors remain a feature in NGS data (Figures 2C and 4A), thus additional error correction (27,28) and/or validation techniques are needed. This can be achieved with hard filtering on the output statistics such as the VAF or the strand bias, but also with machine-learning-based approaches applied to multiple variant summary statistics when validated data are available to inform the model. Here we have controlled for these errors by undertaking technical duplicate of each sample and conditioning on the requirement that the variant must be present in each preparation.

Our pipeline is implemented using nextflow (12), to facilitate its scientific reproducibility but also efficient parallel computations. Needlestack is also provided with Docker (29) and singularity (30) containers to avoid installation of dependencies and produce perfectly reproducible results. Needlestack is a user-friendly pipeline that can be run in one command line. In addition, needlestack implements a power calculation to estimate if the coverage is sufficient to call a mutation (see supplementary methods for details). Using this power analysis, it can predict the germline or somatic status of a mutation when applied to tumour-matched normal mode. This also allows needlestack to flag mutations with an “unpredictable” status (when the coverage is to low) to accurately control the false discovery rate. Source code is available on GitHub and is versioned using a stable git branching model. Importantly, this approach is relatively computationally efficient and parallelisable. This allows error models to be built even across large target stretches of DNA, enabling applications at the exome level, genome levels or to most forms of sequencing data. As an example, needlestack takes around 20 hours to analyze 100 WES when launched on 100 CPUs.

In summary, needlestack uses a robust model of sequencing errors to accurately identify DNA mutations potentially in very low abundance. The model takes the advantage of batch sequencing of multiple samples to precisely estimate the error rate for each candidate alteration. Needlestack can be applicable to various types of studies such as cfDNA, histological normal tissue investigation or high

precision tumour subclonality estimation by providing a high sensitivity for low allelic fraction mutations.

AVAILABILITY

needlestack is an open source collaborative initiative and is available in the GitHub repository (https://github.com/IARCbioinfo/needlestack).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We would like to acknowledge Alain Viari, Dariush Nasrollahzadeh Nesheli and David Muller for their helpful inputs and feedbacks.

FUNDING

La Ligue Nationale (Française) Contre le Cancer [to T.M.D.]; US National Cancer Institute [5R21CA175979-02].

DISCLAIMER

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

REFERENCES

1. Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun*, **6**, 10001.
2. Greaves, M. and Maley, C.C. (2012) Clonal evolution in cancer. *Nature*, **481**, 306-313.
3. Schwarzenbach, H., Hoon, D.S. and Pantel, K. (2011) Cell-free nucleic acids as biomarkers in cancer patients. *Nature reviews. Cancer*, **11**, 426-437.

4. Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R. *et al.* (2018) Somatic mutant clones colonize the human esophagus with age. *Science (New York, N.Y.)*, **362**, 911-917.
5. Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P. and Tyson, G.W. (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology*, **9**, e1003031.
6. Pfeiffer, F., Grober, C., Blank, M., Handler, K., Beyer, M., Schultze, J.L. and Mayer, G. (2018) Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific reports*, **8**, 10950.
7. Fox, E.J., Reid-Bayliss, K.S., Emond, M.J. and Loeb, L.A. (2014) Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications*, **1**.
8. Xu, C. (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, **16**, 15-24.
9. Gerstung, M., Papaemmanuil, E. and Campbell, P.J. (2014) Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics (Oxford, England)*, **30**, 1198-1204.
10. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M. *et al.* (2015) Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (New York, N.Y.)*, **348**, 880-886.
11. Shi, W., Ng, C.K.Y., Lim, R.S., Jiang, T., Kumar, S., Li, X., Wali, V.B., Piscuoglio, S., Gerstein, M.B., Chagpar, A.B. *et al.* (2018) Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. *Cell reports*, **25**, 1446-1457.
12. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. and Notredame, C. (2017) Nextflow enables reproducible computational workflows. *Nature biotechnology*, **35**, 316-319.
13. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078-2079.
14. Aeberhard, W.H., Cantoni, E. and Heritier, S. (2014) Robust inference in the negative binomial regression model with an application to falls data. *Biometrics*, **70**, 920-931.
15. Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289-300.
16. George, J., Lim, J.S., Jang, S.J., Cun, Y., Ozretic, L., Kong, G., Leenders, F., Lu, X., Fernandez-Cuesta, L., Bosco, G. *et al.* (2015) Comprehensive genomic profiles of small cell lung cancer. *Nature*, **524**, 47-53.
17. Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519-525.
18. Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature methods*, **12**, 623-630.
19. Fernandez-Cuesta, L., Perdomo, S., Avogbe, P.H., Leblay, N., Delhomme, T.M., Gaborieau, V., Abedi-Ardekani, B., Chanudet, E., Olivier, M., Zaridze, D. *et al.* (2016) Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. *EBioMedicine*, **10**, 117-123.
20. Chen, L., Liu, P., Evans, T.C., Jr. and Ettwiller, L.M. (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (New York, N.Y.)*, **355**, 752-756.
21. Laehnemann, D., Borkhardt, A. and McHardy, A.C. (2016) Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Briefings in bioinformatics*, **17**, 154-179.
22. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D. *et al.* (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American journal of human genetics*, **99**, 877-885.
23. Nong, J., Gong, Y., Guan, Y., Yi, X., Yi, Y., Chang, L., Yang, L., Lv, J., Guo, Z., Jia, H. *et al.* (2018) Circulating tumor DNA analysis depicts subclonal architecture and genomic evolution of small cell lung cancer. *Nat Commun*, **9**, 3114.

24. LaFramboise, T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, **37**, 4181-4193.
25. Allhoff, M., Schonhuth, A., Martin, M., Costa, I.G., Rahmann, S. and Marschall, T. (2013) Discovering motifs that induce sequencing errors. *BMC bioinformatics*, **14 Suppl 5**, S1.
26. Mose, L.E., Wilkerson, M.D., Hayes, D.N., Perou, C.M. and Parker, J.S. (2014) ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics (Oxford, England)*, **30**, 2813-2815.
27. Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, **9**, 72-74.
28. Ravasio, V., Ritelli, M., Legati, A. and Giacopuzzi, E. (2018) GARFIELD-NGS: Genomic vARiants FIltering by dEep Learning moDels in NGS. *Bioinformatics (Oxford, England)*, **34**, 3038-3040.
29. Boettiger, C. (2015) An introduction to Docker for reproducible research. *SIGOPS Oper. Syst. Rev.*, **49**, 71-79.
30. Kurtzer, G.M., Sochat, V. and Bauer, M.W. (2017) Singularity: Scientific containers for mobility of compute. *PloS one*, **12**, e0177459.

TABLE AND FIGURES LEGENDS

Figure 1: Sequencing error rates estimated by needlestack across the *TP53* gene. A: distribution of sequencing error rates in log-10 scale across the 1,704 positions accounting for a total of 5,112 values. Results are stratified by type of base change: transition or transversion (x-axis) and by type of sequencing technology (IonTorrent Proton amplicon-based data in violet and Illumina exome capture data in yellow). Horizontal black lines correspond to the 5% quantiles of each of the sequencing error rate distribution. B: contribution of each of the 12 possible base changes on the estimated error rate. Error rates are stratified by ranges ($[10^{-5},10^{-4}];[10^{-4},10^{-3}];[10^{-3},10^{-2}];[10^{-2},10^{-1}]$, x-axis). Base change contributions are colored according to DNA strand equivalences (*e.g.* G to T and C to A are both colored in blue). As an example, around 80% of alterations with an estimated error rate between 10^{-3} and 10^{-2} in exome capture data correspond to a G to T transversion.

Figure 2: needlestack regression plot for three independent genomic alterations. A: example of a G to T transversion from exome-hydrid capture Illumina sequencing where the sequencing error rate is estimated as 4.10^{-2} and no variant is detected. B: Example of a duplicated mutation (*i.e.* found in the two technical replicates of the same sample) with a VAF at around 10^{-3} with a corresponding sequencing error rate estimated at 10^{-4}. C: Example of a non-replicated mutation with a VAF at 10^{-4} in the positive library. The second library was covered at more than 18,000X suggesting that the mutation would have been detected if truly present in the DNA sample. Each dot corresponds to the library of a sample and the dots are colored according to the Q-values attributed by needlestack. Red dots are libraries identified as carrying the mutation by needlestack (their Q-values are higher than 50).

Figure 3: needlestack and shearwaterML variant calling false discovery overview from *in-silico* simultations with BAMsurgeon on 125 duplicated samples of circulating cell-free DNA from control individuals. A: cumulative number of detected mutations that were not introduced by BAMsurgeon as

a function of the VAF (in log10 scale) of the mutations, for both methods (needlestack in plain lines and shearwaterML in dashed lines). This number is computed as the average per library when considering all mutations (black lines) and as the average per sample when considering duplicated mutations (blue lines). B: False positive rate (per alteration) for both needlestack and shearwaterML, depending on the estimated error rate at the position. A false positive is defined as a variant not introduced with BAMsurgeon. The red line corresponds to the error rate threshold *t* used for shearwaterML (0.005). ShearwaterML uses this threshold to remove *a-priori* true variants, *i.e.* samples with a VAF>*t*, to then estimate the error rate.

Figure 4: Performance of needlestack for somatic mutation calling using simulated data. The sensitivity of needlestack is presented for multiple values of VAF (in log10 scale, x-axis) of *in-silico* simulated mutations. A total of 10x1,000 SNVs were introduced using the BAMsurgeon software, on a set of 125 samples sequenced at the *TP53* gene locus with the IonTorrent Proton technology. Needlestack sensitivity was computed independently for different error rate ranges (*e*, red, blue and green lines). Black line corresponds to the global sensitivity for all the mutations independently of the sequencing error rate. Global sensitivities of shearwaterML for the same data are shown in grey.

Figure 5: germline variant calling comparison between needlestack and GATK-HC across 62 samples. Both distributions of the VAF and Venn diagrams presenting the concordance of called mutations are shown. VAF distributions are colored according to the Venn diagram. A: comparison between both methods and an Illumina bead array containing gold standard genotypes available for a total of 33 samples. B and C: comparison between needlestack and GATK-HC called mutations without any reference gold standard for both SNVs (B) and indels (C).

Figure 1:

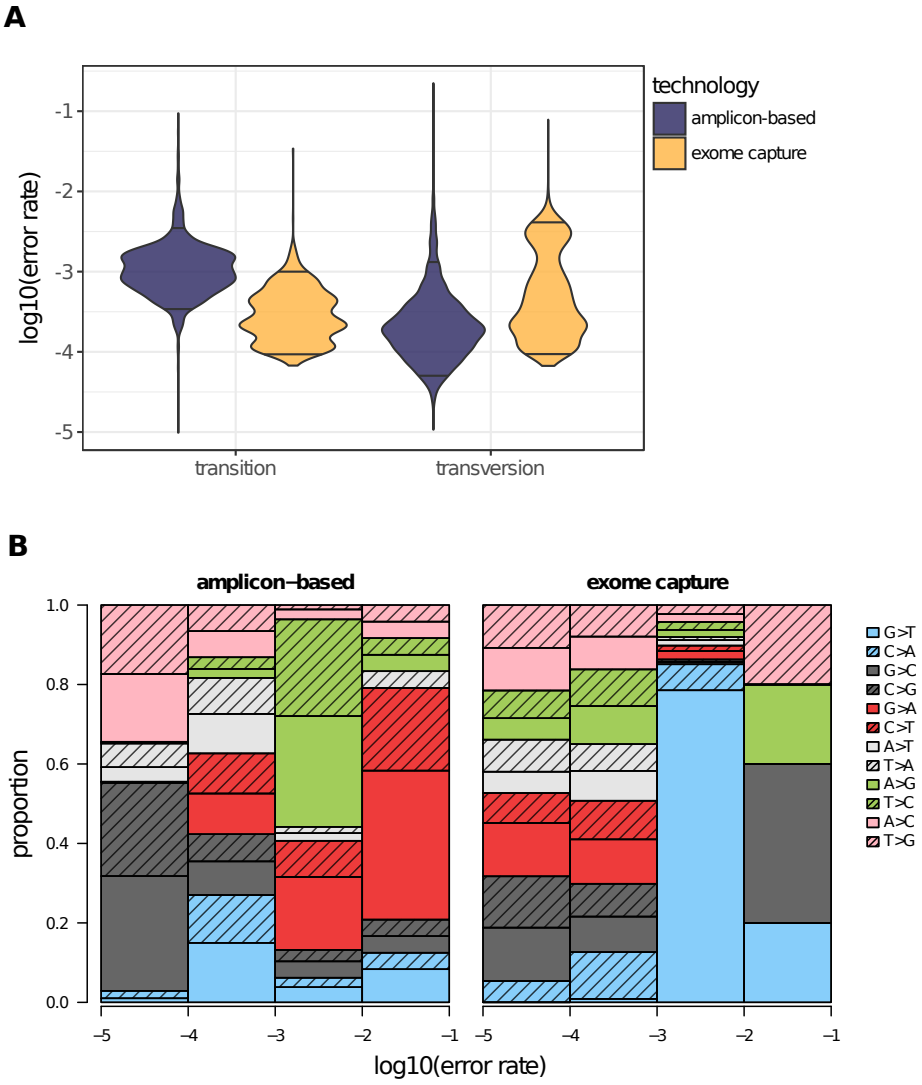

A
technology
amplicon-based
exome capture
log10(error rate)
-1
-2
-3
-4
-5
transition
transversion
B
amplicon−based
exome capture
proportion
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
log10(error rate)
G>T
C>A
G>C
C>G
G>A
C>T
A>T
T>A
A>G
T>C
A>C
T>G

Figure 2:

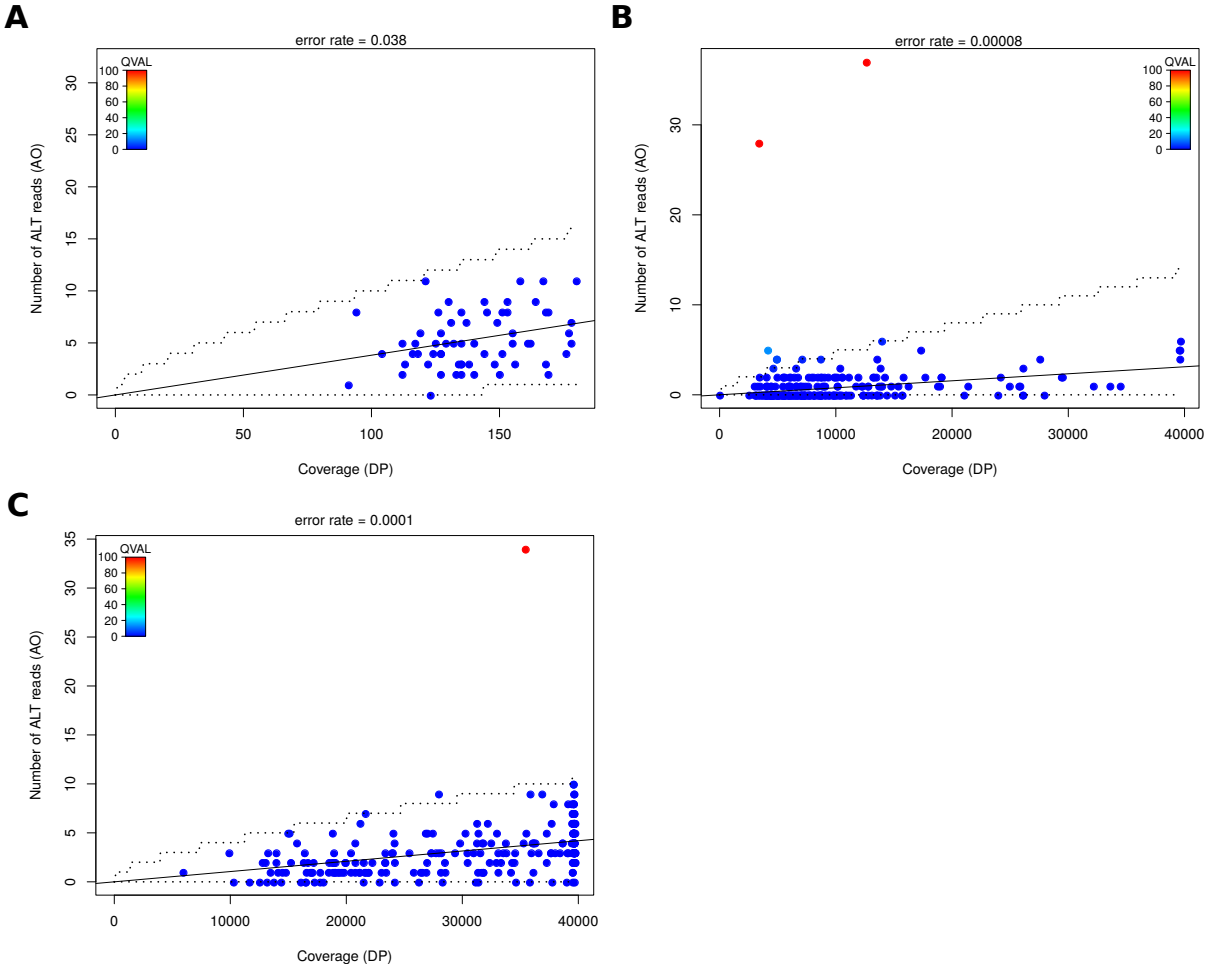


Figure 3:

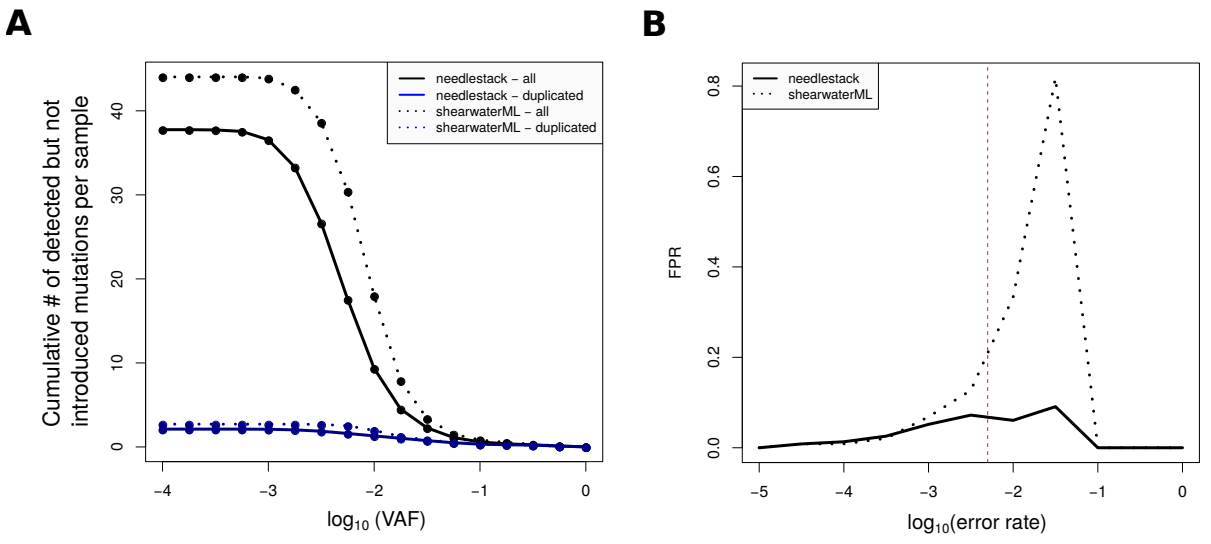


Figure 4:

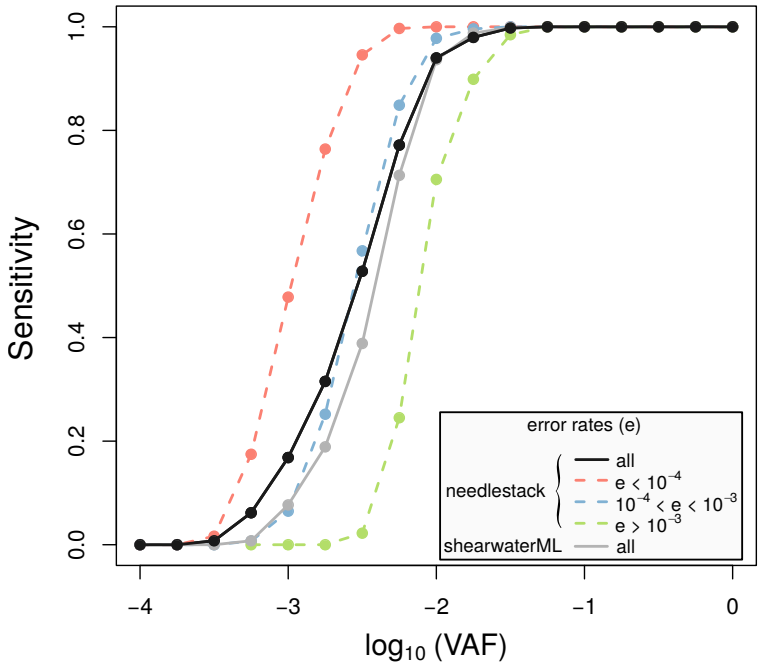
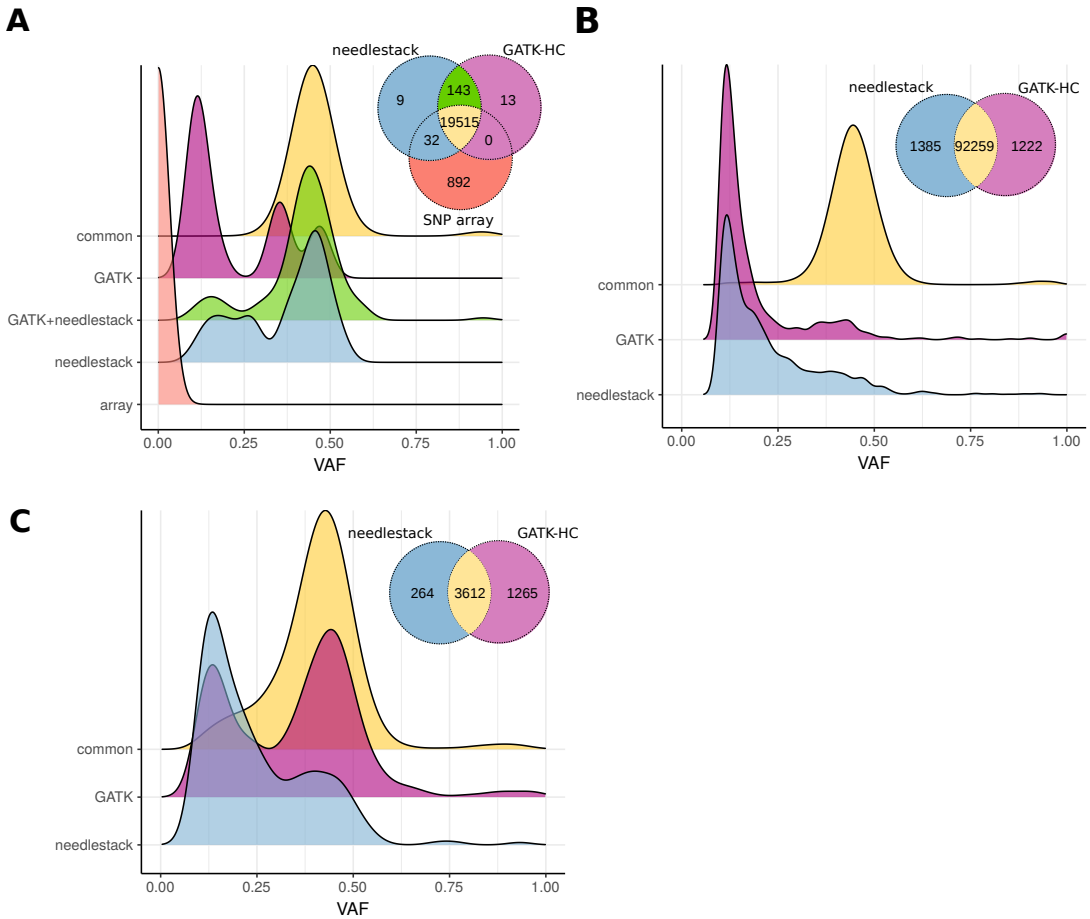


Figure 5:



Supplementary Material

Supplementary Methods

Robust Negative Binomial regression adaptation

The original method was established on falls data where the predictor variable took values from 0 to a couple of hundreds. Here we need to take into account cases where we sequenced deeply and therefore the predictor variable DP can be up to hundreds of thousands. The first model uses integrals of bounding functions for the maximum likelihood estimation (MLE) of e_{jk} to keep robustness, which can take a very long time for high coverage data. To save computing time, we approximate the calculation of integrals required for the MLE of e_{jk} . Instead of computing the sum of all values corresponding to the integral, we interpolate the points using the *spline* function in R, and compute the sum of a set of sampling points with a reasonable size (default is 100). The idea of this robust model is to correctly initiate parameter values and then to perform a MLE to update these values. The initial estimation of e_{jk} is based on a Poisson model, and because of a lack of robustness the following MLE of σ_{jk} can take a lot of time. We thus define our initiation of e_{jk} as the mean of observed e_{ijk} after passing the Tukey's outlier filter, *i.e* an observed e_{ijk} is taken into account for the mean computation if and only if it verifies $e_{ijk} \leq Q_3 + 1.5 * IQR$, with $Qi = i^{th} quartile$ and $IQR = Q_3 - Q_1$.

Implementation

Needlestack is implemented as one major process, which can be executed in parallel for multiple input chunks, each corresponding to a set of called positions. This process is defined as a chain of piped commands: firstly, it runs samtools mpileup utility to compute, for each of the input BAM files, the list of read nucleotides overlapping the input positions. Then, it translates the samtools output into an easier to process format through the mpileup2readcount tool [REF]. Finally, needlestack uses its own R script to run the variant calling independently at each position and for each observed alternative base change, and produces a resulting VCF file that will be merged with other created VCF if run in parallel mode. See supp figure 5 for details on the pipeline. Needlestack is written in the nextflow domain-specific language, allowing high scalability and reproducibility, but also efficient parallel execution. Needlestack source code is freely available on Github, and a Docker container image is hosted on DockerHub. This docker image is based on Bioconda, a sustainable and comprehensive collection of bioinformatics software that helps to easily install workflow dependencies.

Tumour-Normal pairs method

We have implemented, in addition to our basic model, a method to classify any observed variant as somatic or germline when needlestack is launched in tumour-normal pairs mode (supplementary table S2), in order to control our false discovery rate in case of low coverage positions. For this, needlestack introduces the

concept of “power to detect a variant” at a particular position. Indeed, in that case, not observing a variant in the normal sample should not induce a somatic status, due to the uncertainty to have a sufficiently covered position in the normal sample to have the power to detect the variant if present. Our power metric is based on the expected Q-value of the variant if truly present, which depends on the coverage of the sample at the site: if this Q-value is too low, we consider that the site was not enough covered, which induces a lack of power. For a particular individual sequenced for a tumour-normal pair, needlestack classifies its variants as follows: if a variant is observed in the tumour sample, it is classified as “somatic” if not observed in the normal with a sufficient power, and as “unknown” in a case of a lack of power in the normal sample. If a variant is observed in the normal sample, it will be labeled as “germline”, and sublabeled as “confirmed” if also found in the tumour, “unconfirmed” if not in the tumour whereas power was satisfactory, and “unconfirmable” if there was not enough power in the tumour to detect it if present.

To establish if, at a particular site and for a particular base change, a sample is sufficiently covered to validate a variant observed in a matched biological sample, *i.e.*, if the power of detection is sufficient, we compute an expected *Q*-value based on the expected variant allelic fraction in the observed sample. To conclude on the power of detection of a particular variant, we compute the expected *Q*-value and compare it to a threshold.

In the case of a germline this statistic is computed as follows:

$Q-value_{normal}=NB(\mu,\sigma)$ with *NB*=negative binomial distribution, μ=0.5xcoverage at the position, and σ=dispersion parameter (by default=0.1).

In the case of a tumour variant, we computed the expected *Q*-value as the following:

$Q-value_{tumor}=B(n,p)$ with *B*=binomial distribution, *n*=coverage at the position, and *p*=minimum variant allelic fraction expected (by default=0.01).

cfDNA and tumour sequencing for *in-silico* simulations and tumour validation

CfDNA was extracted from 0.8-1.3 mL of plasma using the QIAamp DNA Circulating Nucleic Acid kit (Qiagen) following manufacturer’s instructions. CfDNA was eluted into 100 µL of elution buffer and quantified with the Qubit DNA high-sensitivity assay kit (Invitrogen Corporation). Twenty-one amplicons of 150 bp in size were designed (Eurofins Genomics Ebersberg, Germany) to cover exons 2 to 11 of *TP53*. The GeneRead DNAseq Panel PCR Kit V2 (Qiagen) was used for target enrichment. A validated in-house protocol was used to set up multiplex PCRs in 10 µL reaction volume, containing 5 ng cfDNA, 60 nM of primer pool and 0.73 µL of HotStarTaq enzyme. The experiments were carried out in two physically isolated laboratory spaces: one for sample preparation and another one for post-amplification steps. Amplification was carried out in a 96-well format plates DNA engine Tetrad 2 Peltier Thermal Cycler (BIORAD) as follows: 15 min at 95^{o}C and 30 cycles of 15 seconds at 95^{o}C and 2 min at 60^{o}C and 10 min at 72^{o}C. Two technical duplicates were undertaken for each cfDNA sample including amplification, library preparation, and sequencing. Each technical duplicate pair was assessed on two separate plates to limit the possibility of a contamination.

For the tumour sequencing, eighty nanograms of each DNA sample was used as template to set up four separate PCR reactions (20ng/pool) using the Qiagen GeneRead DNAseq Panel PCR Kit V1 and primer mix (Qiagen), following manufacturer's instructions. The amplified PCR products were then pooled, purified with the Serapure magnetic beads and subjected to library preparation including adapter ligation, purification, and amplification using the NEBNext Fast DNA Library Preparation Kit (New England Biolabs). About 200 ng of individual libraries were pooled into a single tube and size selection (230~250 bp) of pooled libraries was

performed using 100µL aliquot of pooled libraries onto a 2% agarose gel and MinElute Gel Extraction Kit (Qiagen).

Template preparation was done on the Ion OneTouch2 instrument using the Ion PGM Template OT2 200 Kit, followed by sequencing on an Ion Torrent PGM sequencer using the Ion PGM Sequencing 200 Kit v2 (Life Technologies), aiming for mean depth of 500X.

Bioinformatics processing

Short reads from NGS sequencing were aligned to the hg19 human reference genome using the Torrent Suite software (v4.4.2) with default parameters. Somatic mutations were detected with needlestack using the version 1.0 and a QVAL threshold at 50. As recommended by Martincorena *et al.* we used a threshold of 20 for the shearwaterML statistic.

CfDNA samples from lung cancer patients that harbored a high number of raw mutations (>100) in at least one of the two technical replicates were excluded. This removed 4 SCC and 7 SCLC from the 46 matched samples. CfDNA mutations associated with a low confidence base change were removed. A low confidence base change satisfied $P<0.05$ with P given by:

$$P = \sum_{p=p_{obs}}^{p_{max}} \left[\prod_{i=0}^{p-1} \binom{k-2i}{2} \left[\prod_{j=0}^{k-p-1} (2N-2j) \right] \frac{1}{p!\,k!\binom{2N}{k}} \right]$$

with N the total number of sequenced libraries, k the number of libraries being positive for the mutation, p_{obs} the number of paired called mutations (paired in the sense found in the two technical replicates) and p_{max} the total number of possible pairs from k entities (independently of the data). The detailed source code for cfDNA mutation analysis including all quality filtering step description is available on GitHub at: https://github.com/IARCbioinfo/target-seq.

For the germline analysis, GATK-HC variant calling was performed using version 3.4 and the HaplotypeCaller algorithm with default parameters, followed by the joint genotyping step. Finally, Variant Quality Score Recalibration from the GATK best practices was applied, using dbSNP 138, HapMap 3.3, 1000 Genomes phase 1 and OMNI 2.5 databases. Options provided were « -tranche 100 -tranche 99.9 -tranche 99.0 -tranche 90.0 » for both INDEL and SNP modes. GATK-HC variant calls were filtered on PASS and on Phred-scaled likelihood (PL) more than 20. BAM files were locally reassembled with ABRA version 1.0 before launching the variant calling by needlestack. Needlestack germline calling was launched using our recommendations for germline detection, *i.e.* QVAL>20 and the option *–-extra-robust*. This option helps needlestack to correctly estimate the error rate in the case of common germline variants (defined when more than 10% of the samples present a VAF higher than 20%) that tend to bias this estimation towards high values. For each of these base changes independently, this process first eliminates these germline samples and then estimates the error rate on remaining samples. In this germline analysis, both positions and variants with respectively a median coverage and an individual coverage less than 50 were removed from the whole analysis. Coverages were computed with samtools mpileup, counting only reads with a mapping quality higher than 20 and a base quality higher than 13. We considered as variant frequency the maximum proportion of samples carrying the variant estimated by both methods and then filtered out germline variants

with a frequency higher than 10% to consider only rare variations.

Computation of ShearwaterML statistic used in the BAMsurgeon *in-silico* simulations

We launched the shearwaterML algorithm on the simulation data sets to compare its global performance with needlestack. We used default thresholds except that we increased *maxvaf* to 1 so that not to filter on VAF, and set *truncate* to 0.005 to avoid true mutations present initially at low VAF to enter the background error model and potentially reduce the sensitivity, as recommended in Martinorena *et al.*. ShearwaterML produced *p*-values instead of the shearwater Bayes factor, that we corrected for multiple testing using the Benjamini-Hochberg method which produces then *Q*-values that we finally transformed into Phred scale *Q*-values (QVAL).

Using needlestack to compute the error rate distribution

Needlestack can be forced to compute the error rate for every query position. For this, the user needs to ask needlestack not to only consider variable positions to output the informations in the VCF file. This can be achieved by launching needlestack with parameters --all_SNVs, --min_ao 1 and --min_dp 1. This way, in our analysis, the 5112 error rates across the TP53 gene were computed, and those that are precisely equal to zero and were set as NA in the VCF.

Supplementary legends

Supplementary figure 1: needlestack workflow description. First step corresponds to the creation of a BED file containing the DNA positions on which the calling should be launched using the fasta index of the input reference genome; this step is optional, only performed if target positions are not provided by the user. Second step splits the BED file into multiple sets of positions to run the algorithm independently on each set in parallel; number of position sets is given in input by the user. Third step runs the variant calling from three piped substeps: (i) the mpileup file building using samtools, (ii) the parsing of the mpileup to produce count data per sample in a tabulated readable file, (iii) and the running of the regression in R on each tested mutation to estimate the error rate. Fourth and last step merges VCF files previously produced in parallel and outputs the global result. Workflow orchestration is done thanks to the nextflow domain specific language. Not that nextflow manages all the execution of the pipeline from one unique user command line.

Supplementary figure 2: estimated error rate distributions from amplicon-based sequencing of *TP53* gene (median coverage around 10,000X). Distributions of error rates are shown for each of the 96 possible base variation (with flanking 3' and 5' bases), and are colored by DNA base changes.

Supplementary figure 3: estimated error rate distributions for both SNVs and indels from the same data as used in supplementary figure 1. A: error rate distributions are shown as a function of the type of SNV (yellow for transversions and green for transitions), the length of the insertion (pink) and the length of the deletion (blue), with *n* indicating the total number of error rates used to compute the distribution. B: error rate distributions restricted to insertions and deletions, as a function of the size of the homopolymer region (i.e. depending on the number of repeated nucleotides) at the position.

Supplementary figure 4: validation of cfDNA mutations in the matched tumour sample from a total of 11 SCLC cases and 24 SCC cases. A: correlation of cfDNA and tumour VAF for deleterious validated mutations (total=12). Pearson correlation coefficient was estimated as 0.59. Grey dashed line corresponds to the fitted linear regression characterized by the values *a* (slope) and *b* (intercept). B: needlestack regression plots of a validated deleterious SNV (left panel corresponds to cfDNA data and right panel to tumour data).

Supplementary figure 5: sensitivity of needlestack as a function of the VAF of the *in-silico* simulated insertions (A) and deletions (B), depending on the error rate for the mutation. Data used are described in the BAMsurgeon *in-silico* simulations material and method paragraph. Cumulative number of false discoveries *i.e.*, detected but not introduced is shown for insertions (C) and deletions (D) per sample, depending on the VAF of the detected mutation. This number was computed firstly for all detected indels (the result is per library) and secondly for indels validated in the second library (the result is per sample). Dashed lines correspond to the number of indels not introduced by BAMsurgeon that are however not in strand bias (RVSB<0.85).

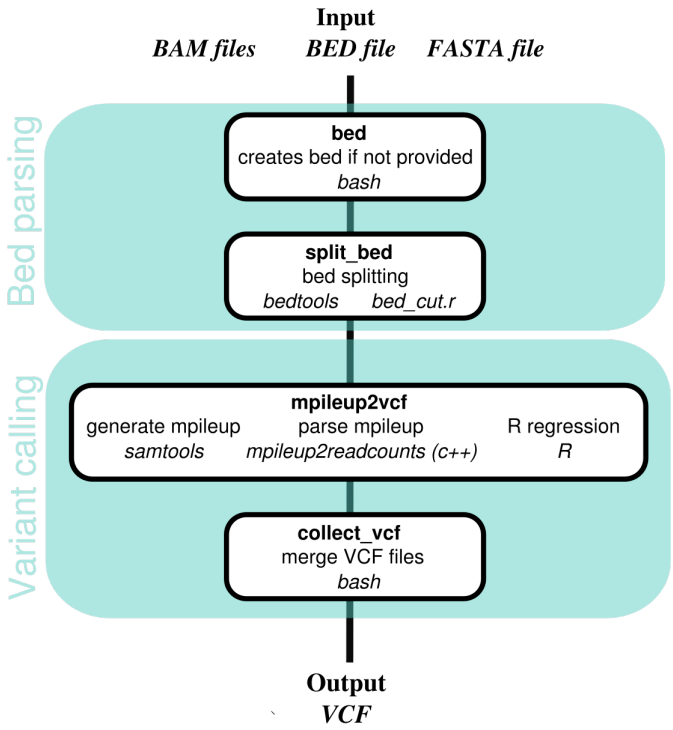
Supplementary figure 6: A: distribution of the sequencing error rates for SNVs detected by GATK-HC but not detected by needlestack from 62 WES samples (total of 1385 mutations), estimated using kernel density estimation. B: needlestack regression plots for one particular position where GATK called 3 variants (in purple). Needlestack estimated a high sequencing error rate (around 1%) for this mutation and therefore did not call it, highlighting the fact that estimating the systematic error rate across multiple sample should reduce the false discovery rate of the method.

Table S1: description of the 22 mutations identified by needlestack in the cfDNA of 35 lung cancer patients.

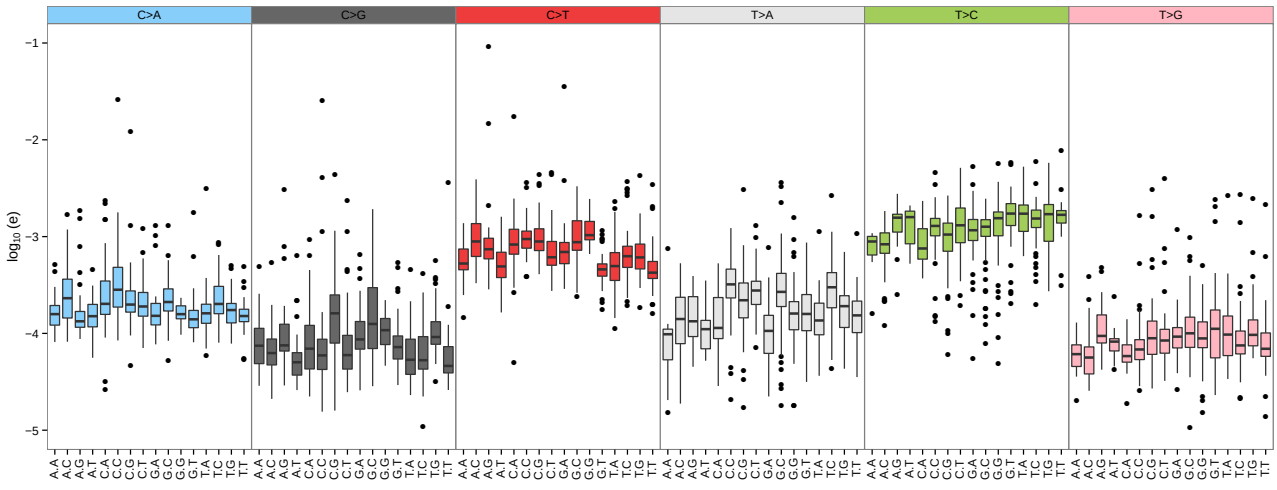
Table S2: variant status and genotype attributed by needlestack as a function of variant detection and the power to detect variants in tumour and matched normal samples.

Supplementary Figures

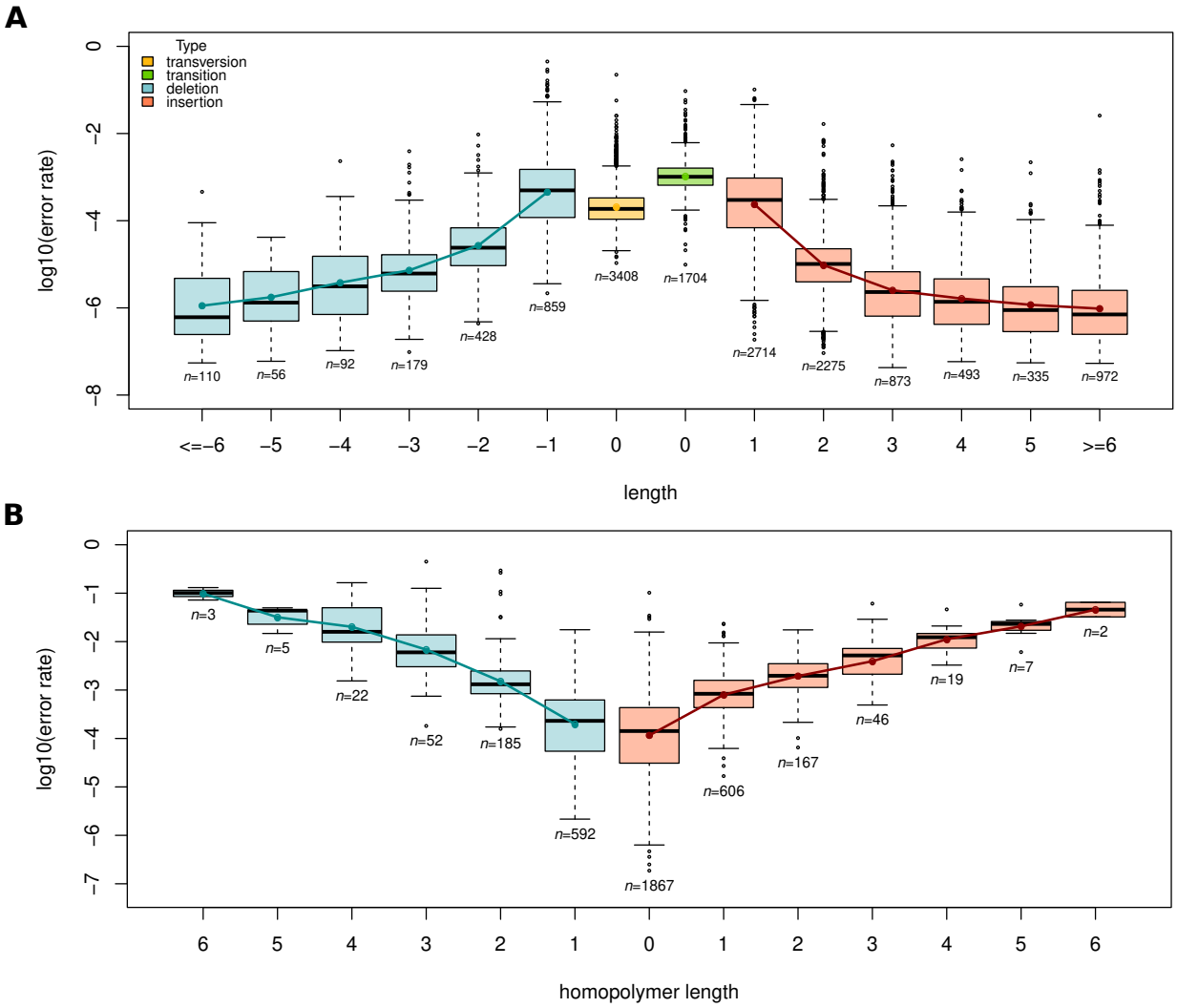
Supplementary Figure 1:



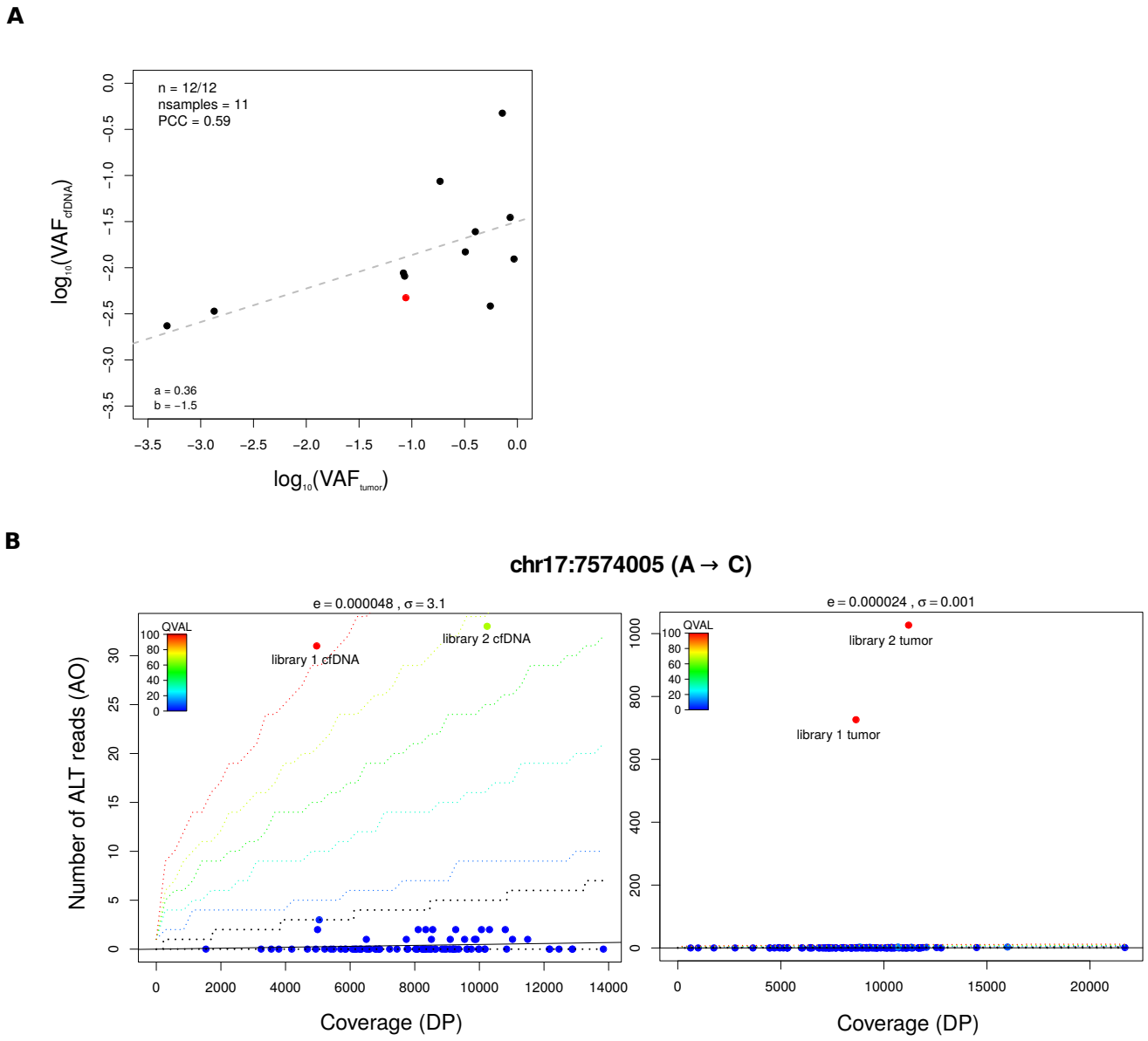
Supplementary Figure 2:



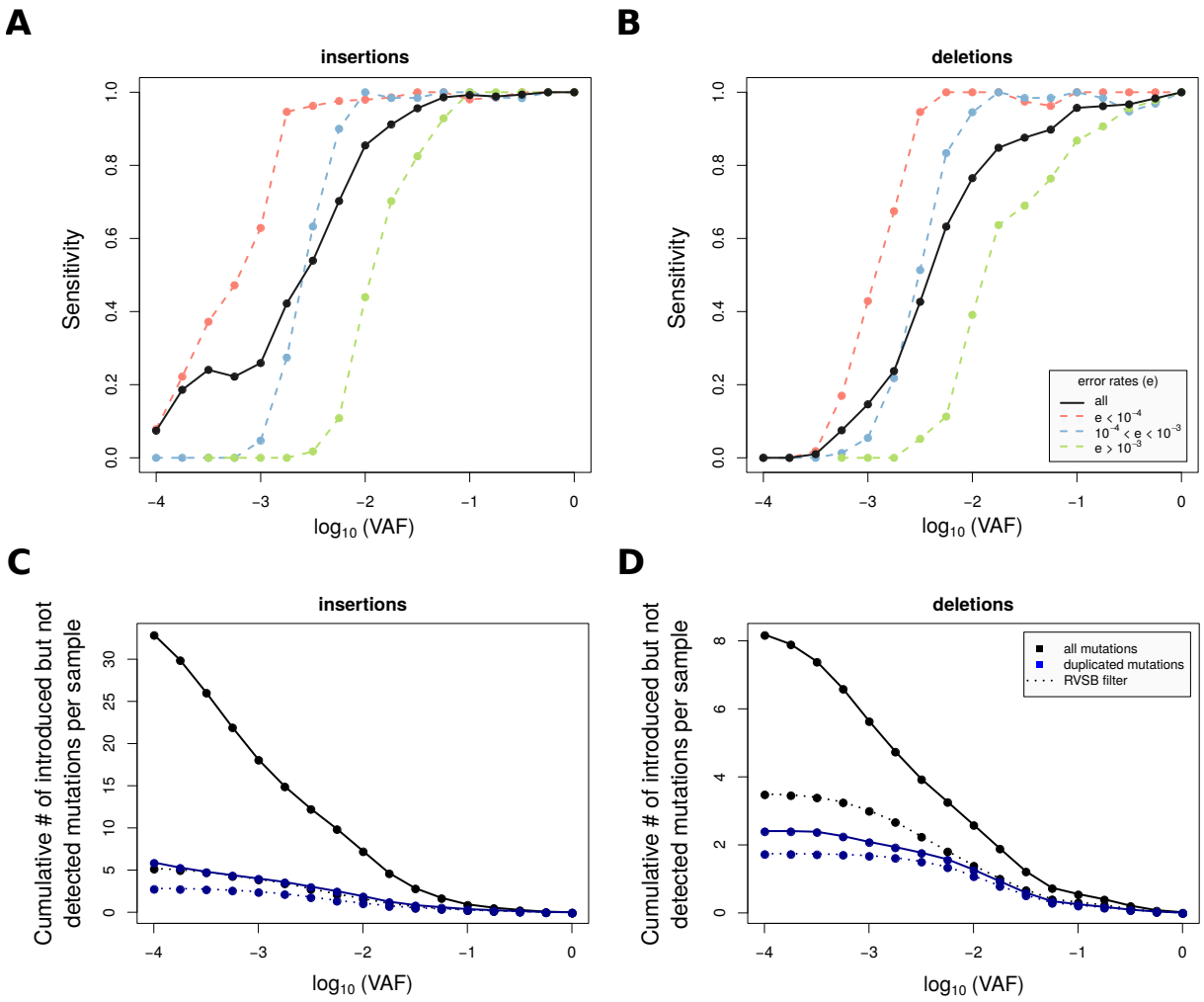
Supplementary Figure 3:



Supplementary Figure 4:



Supplementary Figure 5:



2.4 Discussion

In this first chapter we presented needlestack, a sensitive integrative variant caller. The algorithm is based on the application of a robust negative binomial regression to estimate the rate of systematic errors across a panel of samples. Contrary to the existing variant callers, the regression used in our algorithm is *robust* in the sense that the model is not biased by the presence of true mutations that tend to influence the estimation of the SER to higher values. The main drawback of non-robust methods is that the number of missed true mutations would increase with the total number of true mutations at a particular position. Methods that requires a prior on the SER to keep robustness by removing potential *a priori* true mutations are exposed to an increase of the number of false positive for position harboring a SER higher than this prior. To deal with such unknown variabilities, needlestack directly learn the SER from the data. For this, it requires a panel of samples sequenced in a similar way to precisely estimate this variable. Because we model the error rate for each observed alteration independently, the number of times this process needs to be done is huge, and therefore we developed needlestack with the aim of proposing a method that can be applied in a practical manner.

The main advantage of needlestack compared to other existing variant caller is its ability to detect very low VAF mutations. Nevertheless, needlestack can detect a very low VAF mutation only if the SER is sufficiently low compared to this VAF. This means that needlestack accuracy depends of the SER at the position, but it also depends of the depth at the position. As an example, a mutation with VAF at 0.1% requires a depth at the position of 1000X to observed 1 sequenced read that contains the mutation (in average). An interesting perspective could be to study the influence of the coverage on the sensitivity of our method.

By design, needlestack only detect systematic errors across multiple samples. This also means that it can not detect recurrent mutations across these samples, that would be incorporated in the error model, and therefore needlestack is limited to *rare* mutations, rare in the sense "low population frequency". To deal with common mutations such as hotspots or common germline genetic variations [136], we have implemented the option *extra_robust_gl*. When this option is activated, needlestack will remove potentially true mutations in

high proportion (same idea than shearwaterML algorithm) to then estimate the SER on the remaining samples. These potentially true mutations are defined as the following:

- harboring a VAF higher than υ.
- being present in a proportion between ρ_{min} and ρ_{max}.

The aim of using these parameters is to efficiently distinguish between errors and true mutations: (i) ρ_{min} parameter controls the fact that, this "tuned" robustness should be run only if a sufficiently proportion of mutated samples can influence the SER estimation; (ii) ρ_{max} parameter controls the fact that, if a variation is observed in the majority of the samples, needlestack would not be able to distinguish it from errors; (iii) υ parameter defines the maximum expected error rate. By default, $\rho_{min} = 10\%$, $\rho_{max} = 50\%$ and $\upsilon = 20\%$. We suggest to use this *extra_robust_gl* option in the case of hotspot variations or in the case of common germline genetic mutations where the VAF is expected to be high and different from the SER.

Our method shows the advantage to be integrative, in a sense that needlestack can call both germline and somatic mutation at the same time. Needlestack will assign a status *germline* or *somatic* to each called mutation if the user provides paired data, *i.e.* tumor and normal sample for each individual.

Nevertheless, because the error rate is, in a sense, stochastic across sequencing runs and genomic positions, needlestack requires a particular type of data, that is composed of similarly sequenced samples at the same positions. Another possibility would be to use precomputed estimations of the SER. This would avoid the need of such sequencing design that could be potentially difficult to validate, but for this the SER should be quite constant across sequencing runs. We then computed the Coefficient of Variation (CV) of the SER for SNVs across 5711 genomic positions sequenced independently ten times (figure 2.2). Let E defines the set of estimated SER across the ten sequencing runs for a given alteration. The corresponding CV is defined as the following

$$CV = \frac{\sigma}{\mu} \text{ with } \sigma = \sqrt{\frac{\sum_{i=1}^{|E|}(e_i - \overline{e})^2}{|E| - 1}} \text{ and } \mu = \overline{e} = \frac{\sum_{i=1}^{|E|} e_i}{|E|}$$

The median CV was estimated at -0.63 in 10-logarithm scale, suggesting that, for a given

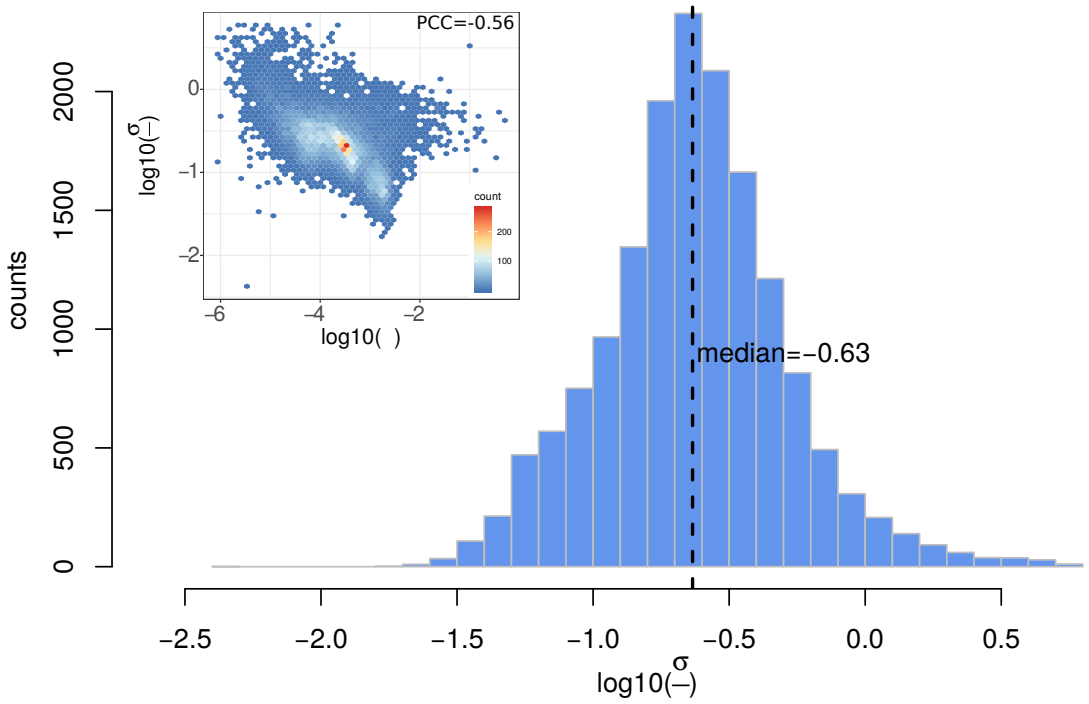


Figure 2.2 – Distribution of the Coefficient of variation (CV) of the SER based on 5711 genomic positions sequenced independently 10 times, in 10-logarithm scale. Median of CV was estimated at −0.63. Small left panel corresponds to the distribution of CV according to the SER mean, and shows no evidence of correlation (Pearson Correlation Coefficient (PCC)=-0.56).

alteration, the SER would differ of around 20% between two sequencing runs. In addition, there is no evidence of correlation between the CV and the mean of the SER (figure 2.2 small panel, (PCC=-0.56)), which means that the variation of SER is relatively constant whatever the SER.

To conclude, using using a "catalogue" of error rates as a reference is possible, but this would have a major consequence on the accuracy of mutation detection: if the pre-computed SER is greater than the true SER, this will potentially create false negatives (figure 2.3 A), and, if the pre-computed SER is lower than the true SER, this will potentially create false positives (figure 2.3 B). Nevertheless, these findings are only applicable on IonTorrent sequencings, further investigations on the stability of the SER are required for Illumina sequencing data.

Needlestack can be applied on traditional tumor and normal datasets to estimate the somatic mutations attributed to cancer cells. For example, we have used needlestack (tumor-normal pairs mode, see supplementary material in the paper for more details) to efficiently

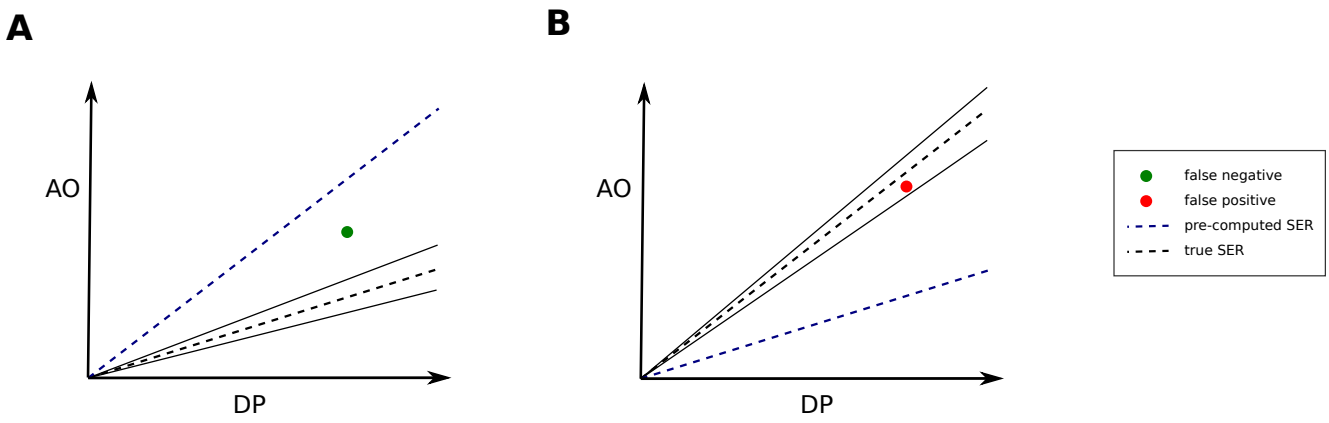


Figure 2.3 – Consequences on the mutation detection accuracy of using pre-computed SER.

identify somatic mutations in WES data from 21 atypical carcinoid samples and 10 typical carcinoid samples, in order to perform integrative and comparative genomic analyses of pulmonary carcinoids (see annexes A.2.2, Alcala *et al.* Nature Communications 2019.). We have also validated the somatic mutations called in the WES with needlestack using a target sequencing approach (IonTorrent Proton technology). For these 220 SNVs and 30 indels, we have reported a validation rate higher than 95% for VAF higher than 10%, and we have shown that this validation depends of the VAF of the mutations (see figure 2.4), as we already reported in the needlestack paper.

Needlestack can also be used in other type of studies such as those on somatic mutations in histologically normal tissues that are expected to be found in low proportion, studies on subclonality of tumors due to its ability to detect mutations with high sensitivity, or even studies on circulating tumor DNA that carries very low abundance tumor-derived mutations.

Needlestack is based on the estimation of systematic errors to efficiently call mutations as being different from these errors. It does not correct for non-systematic or pseudo-systematic errors 1.2.1. To correct for these types of errors and reduce the falsely called mutations, we proposed to use *a posteriori* steps of *variant filtering*. This second type of methodology to refine the detection of mutation from NGS data is presented in the second chapter.

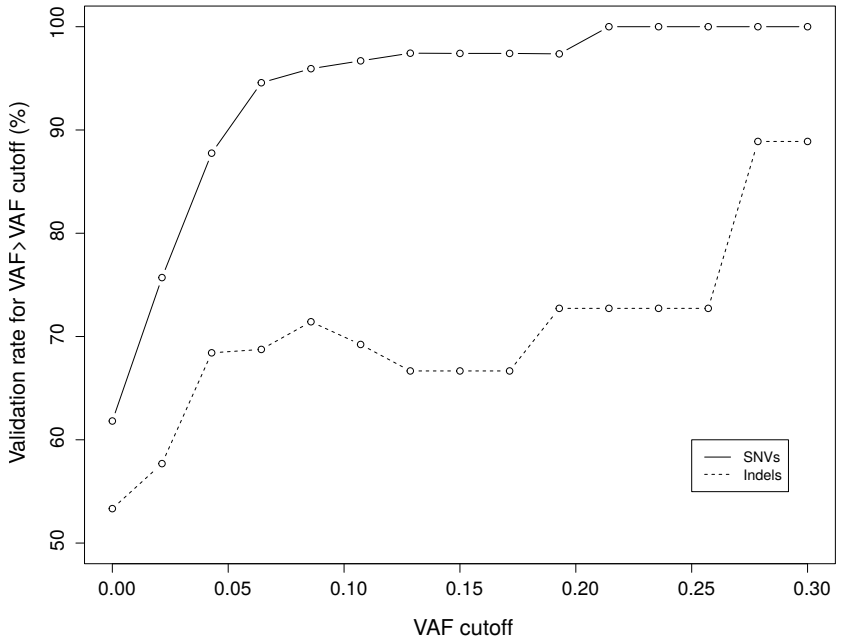


Figure 2.4 – Validation of somatic mutations from WES of pulmonary tumor and matched normal samples using a targeted sequencing approach.

Chapter 3

Dealing with pseudo and non-systematic errors: variant filtering methodologies

Contents

3.1 Scientific context 74

3.2 Scientific contribution 75

3.2.1 Variant filtering for deep targeted sequencing data 75

3.2.2 Application to ctDNA data 84

3.2.3 Article B (in preparation) 89

3.2.4 Variant filtering for germline data 103

3.3 Discussion 110

3.1 Scientific context

As exposed in the introduction chapter, errors found in NGS data can be classed into two distinct groups: (i) systematic errors; (ii) and pseudo and non-systematic errors. In the chapter 2, we described the development of a sensitive variant calling algorithm, needlestack, which estimates the systematic errors across samples and then call mutations as being different. Needlestack, by design, classify errors as such only if they are recurrent across multiple samples. Therefore, such rare variant calling method is not appropriate to accurately distinguish between pseudo or non-systematic errors and true mutations.

To reduce the potentially false calls from the variant calling process, a subsequent step of *variant filtering* is required. As presented in the introduction of the thesis, the aim of the variant filtering process is basically to boost the precision of the variant calling step with a minimal decrease of sensitivity. Contrary to our needlestack method, variant filtering algorithms are not based on the systematic nature of errors that they try to remove. However, it is possible to use statistics on error proportion across the samples.

Roughly, variant filtering methods try to use the knowledge on known mutations, and conversely, the knowledge on expected errors to remove falsely called variants, based on predefined statistics. Due to the fact that these statistics can be related to one particular variant calling method (as an exemple the QVAL statistic given by needlestack), the variant filtering algorithm is adapted to the variant calling method. This means that the variant filtering step is highly dependent on the variant caller and therefore that new methodologies should be developed when no algorithm is suitable.

In this chapter, we propose to present two different developed methodologies of post-calling variant filtering. Due to data availability, we propose two independent filtering methodologies, (i) for somatic deep targeted sequencing data; (ii) and for germline sequencing data both analyzed with needlestack. Methodologies have been tested on IonTorrent Proton sequencing data, but used statistical variables do not depend on the sequencing technology and therefore our methods can be easily transposed on other type of data such as data from Illumina sequencers.

3.2 Scientific contribution

3.2.1 Variant filtering for deep targeted sequencing data

In this first part, we will present our filtering methodology for deep targeted sequencing data. The method is restricted to data analysed with our needlestack variant caller, but the scripts can easily be adapted to other variant caller if needed. In term of data, the method is restricted to sequencing data presenting a deep coverage of typically several thousands reads.

We have previously shown in the needlestack paper that an efficient approach to remove PCR errors is to sequence each DNA sample twice. One DNA sample containing $n * p$ DNA molecules can be divided into n sets containing p molecules on average, after the extraction and before the PCR steps. Each DNA set is called a library. Therefore, we developed our approach based on duplicated libraries, *i.e.* $n = 2$ DNA sets per sample.

Our approach is composed of five filters which are based independently on:

- the number of mutations per library
- the concordance between the two libraries
- the strand bias
- noisy positions
- the genomic distance from a true variant

As a first *a priori* quality control step, we have adapted the QC3 software from Guo *et al.* [61]. We were interested in using this tool to estimate the median coverage of each sample across the sequenced positions, and then remove samples not well covered, that are considered as non-analyzable. The main improvement was to propose two new options:

- *-d_cumul* to output the cumulative coverages, *i.e* for a list of coverages threshold output the percentage of positions covered by at least these coverages
- *-nod* to do not compute the coverage in non-target regions, in the aim at reducing the computing time

We maintain and propose a free access to the adapted source code on GitHub: https://github.com/IARCbioinfo/QC3.

The first filter subsequent to the variant calling that we have developed is based on the quality of the libraries. Indeed, degradation of DNA should create artifact DNA variations and lead to false calls from the variant calling. This has mainly been observed in formalin-fixed paraffin-embedded (FFPE) tissues [42] because these samples are fragmented and contain DNA lesions. Nevertheless, there is no actual evidence that the increased number of artifact mutations due to DNA degradation could not happen in every type of samples.

This first filter is a sample-scale filter. It does not remove falsely called mutations but will remove low confidence samples. Some of the mutations called in these low confidence samples can be attributed to the DNA degradation, and there is no particular variable able to distinguish true from false variations in such cases. Consequently, we decided to remove these samples.

To determine the set of low confidence sample to be removed from the analyses, we computed a threshold on the maximum number of expected mutations per library from the variant calling. Each library harboring an unexpected high number of mutation is considered as non analyzable. Due to the fact that we require two library per sample, our filter will remove each sample presenting at least one such low confidence sequencing library. This maximum value of expected number of mutations per library is expected to depend on the sequencing batch and probably on the type of sample, so an ideal scenario would be to compute it for each sequencing batch, supposing that one batch contains the same type of analyzed sample, *i.e.* normal tissues, tumor biopsies, *etc.* Naturally, if the number of libraries per batch is low, multiple batches can be merged to efficiently compute this value if realized in identical conditions.

To compute the threshold on the number of expected mutations, we propose to fit a robust negative binomial distribution on the number of observed mutations per library. Indeed, we have observed that this number of mutations, that can be modeled using a Poisson distribution, harbor a larger variance than expected by this model, and then the negative binomial model is more adapted. In addition, we require robustness to control for possible unexpected high number of observed mutation, in the case of degraded DNA for example. We then compute the 95^{th} percentile of the fitted negative binomial distribution as the fil-

tering threshold and identify low confidence libraries as the one presenting a number of observed mutations higher than this computed threshold.

As a descriptive example, figure 3.1 is showing the distribution of the number of mutations per library for four merged sequencing batches and the corresponding exclusion area. Data are generated from the sequencing of the 1704 positions of the *TP53* gene of 209 samples (total=2∗209=418) on a IonTorrent Proton sequencing platform.

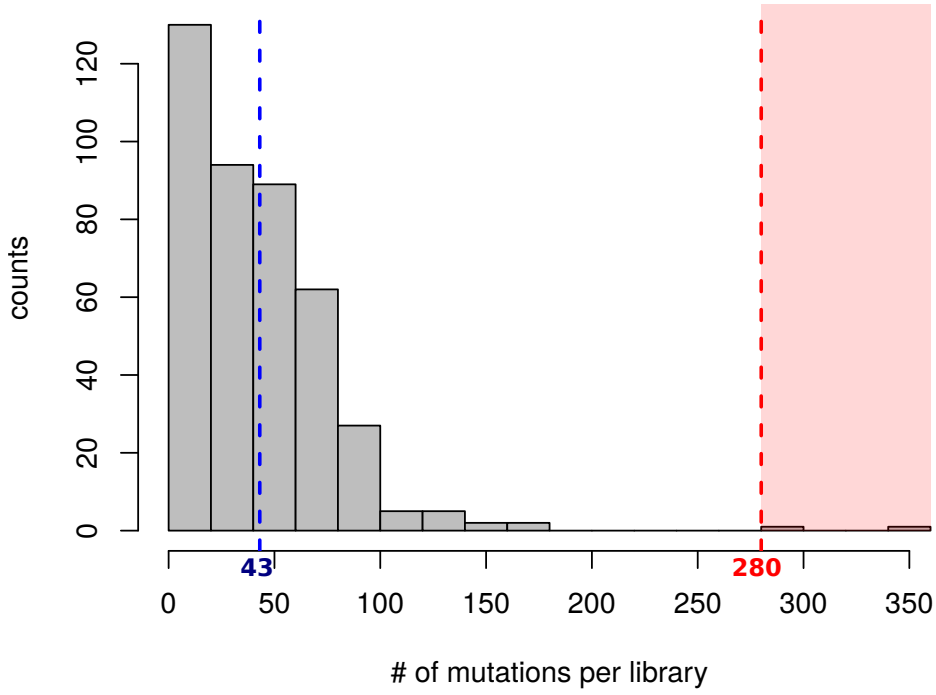


Figure 3.1 – Plot example of the distribution of number of called mutations per library. Data contains a total number of 418 analyzed libraries. Dashed blue and red lines correspond respectively to the mean and the 95^{th} percentile of the fitted negative binomial distribution. Excluded libraries by filters are present in the light red area.

As a second filter, we propose to use the presence of the mutations in the technical replicates (sequencing libraries) of the same samples. As described in the introduction, sequencing the same DNA in replicates should reduce artifacts on observed mutations, notably the one coming from the PCR processes [111]. As a matter of fact, a DNA variation coming from an error of the DNA replication during one of the PCR series is not attributable to the biology of the sample and is not expected to be shared across multiple samples, it then can be

classified as a totally non-systematic error. In addition, such non-biological DNA variations are expected to be random and then not replicated across independent libraries.

To benefit from the availability of technical replicates of the same sample in order to reduce the set of false called mutations from the variant calling process, we propose to consider that a mutation found in one of the n libraries should be validated in each of the other libraries of the sample. This leads to the fact that if one mutation is observed in only one specific replicated libraries, this can be a PCR artifact and consequently this mutation should be differently considered from the other called mutations. If $n = 2$, we propose naturally to exclude called mutations not validated twice, *i.e* not called in the two libraries.

This approach could be easily extended in the presence of more than two replicated libraries: for each called mutation m, let n be the total number of technical replicates, n_{obs} the number of technical replicates from which the mutation m was called and p_m the minimum percentage of the total number of technical replicates that is expected to share the called mutation m. Then, the mutation m will be filtered if and only if:

$$n_{min} = n * p_m < n_{obs}$$

The third filter that we propose is based on the strand repartition of each called mutation m. The sequencing of DNA is expected to be equally spread on the forward and reverse strands (see figure 1.1 for an explanation of DNA strands). This means that whatever the allelic fraction of a called mutation, the repartition of the sequencing reads on both forward and reverse strands is a random process, and then the proportion of forward (and reverse) reads follow a Binomial distribution with a mean equals to 0.5 and with a particular unknown variance.

As discussed in the introduction and as previously reported [60], the strand bias is not expected to be consistent across the samples and therefore it would not be incorporated in the background systematic error rate, such as the one estimated by needlestack to filter out systematic errors in NGS data. Thus, the strand bias should be analyzed separately and should be considered as a per-mutation variable.

Several measures of strand bias have been proposed, but for deep targeted sequencing

the common measure is the RVSB [49]. RVSB measure is computed as the following (see paragraph 1.2.1 for detailed explanations of the formula):

$$\text{RVSB} = \frac{max(\text{AO}_f * \text{DP}_r, \text{AO}_r * \text{DP}_f)}{\text{AO}_f * \text{DP}_r + \text{AO}_r * \text{DP}_f}$$

RVSB measures the difference of strand repartition between the mutated sequenced reads and the total sequenced reads. Figure 3.2 shows multiple measures of RVSB when these two values are varying.

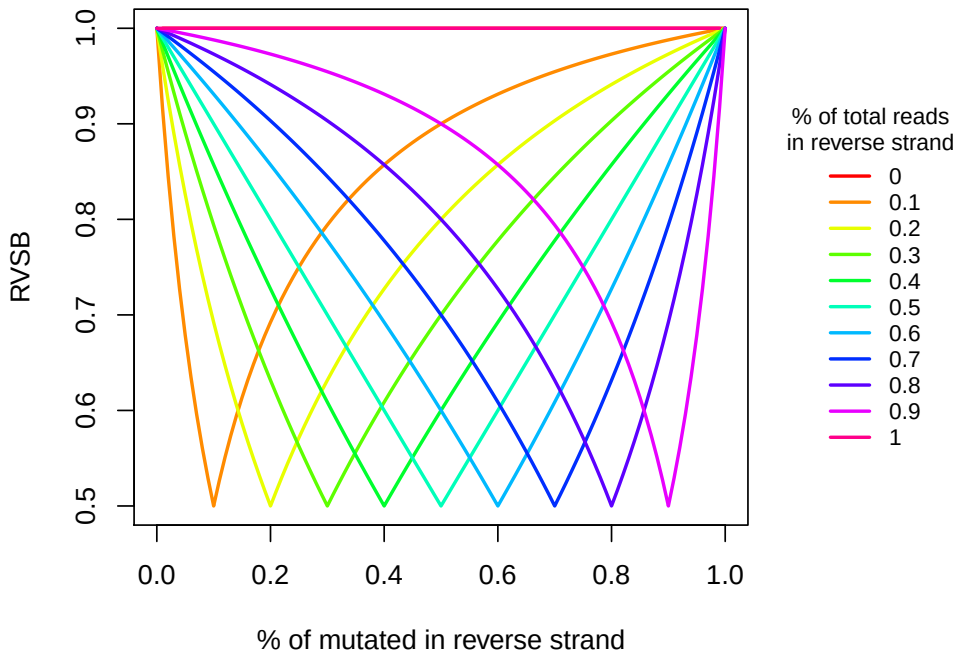


Figure 3.2 – RVSB measure depending on the strand repartition of the mutated reads. The RVSB is computed for multiple values of the strand repartition of the totality of sequenced reads. Strand bias is equal to 0.5 when the strand repartition is the same for mutated reads compared to all the reads. Higher the difference of repartition, higher the deviation of RVSB from 0.5.

The filtering step based on RVSB consists in removing each mutation m presenting RVSB_m which validates:

$$\text{RVSB}_m > \text{RVSB}_{max}$$

with RVSB_{max} being the *a priori* threshold on the strand bias.

The fourth filter that we present here is an alteration-level filter. Indeed, we have observed that some alterations tend to re-occur in multiple sample but are not validated when a second replicate of the same sample is also sequenced, whereas a true mutation is expected to be detected in both technical replicates. The aim of this filter is to remove a called mutation when it corresponds to a low-confidence alteration in term of sequencing. This does not mean that these mutations are false but it means that the confidence associated with them is not sufficiently high to keep them in the analysis. In this case, the best choice should be to do not consider them.

To deal with such pseudo-systematic errors, not replicated across technical replicates, we have developed a metric, the Low-Confidence Alteration Probability (LCAP). This part was the major scientific contribution of our variant filtering methods for deep targeted sequencing data in term of development. The LCAP measures the probability that an observed alteration corresponds to "noisy sequencing". Indeed, if each sample is duplicated, a true mutation is expected to be detected in both technical replicates. As a true mutation is expected to be detected in both technical replicates, by negative logical equivalence, a false mutation is expected to be randomly detected in replicates, an therefore not clustered in pairs of libraries (when two technical replicates are sequenced). To estimate this random spreading across samples, we compute the probability that, for a given called alteration, the observed repartition in pairs can be explained by a random process (null hypothesis).

First, we compute C_p the probability of observing p clusters of pairs when randomly picking k elements from a total of 2N elements:

$$C_p = \prod_{i=0}^{p-1} \binom{k-2i}{2} \left[\prod_{j=0}^{k-p-1} (2N-2j) \right] \frac{1}{p!k!\binom{2N}{k}}$$

with in our case k the number of libraries which are positive for the mutation, p the number of observed pairs of mutations (found in the two technical replicates) and N the number of samples.

This is analogous to the problem of "socks": C_p corresponds to the probability to obtain p pairs when picking up k socks in a drawer containing N pairs of socks broken.

LCAP is a p-value, *i.e.* corresponds the probability that the number of pairs observed

under the null hypothesis (random picking) would be greater than or equal to the observed number of pairs, and is computed as follows:

$$\text{LCAP} = \sum_{p=p_{obs}}^{p_{max}} C_p$$

with p_{obs} the number of paired called mutations (paired in the sense found in the two technical replicates) and p_{max} the total number of possible pairs from k entities (independently of the data).

Finally, when LCAP $< p_t$ with p_t the p-value threshold, we reject the null hypothesis and therefore don't consider it as a low confidence alteration.

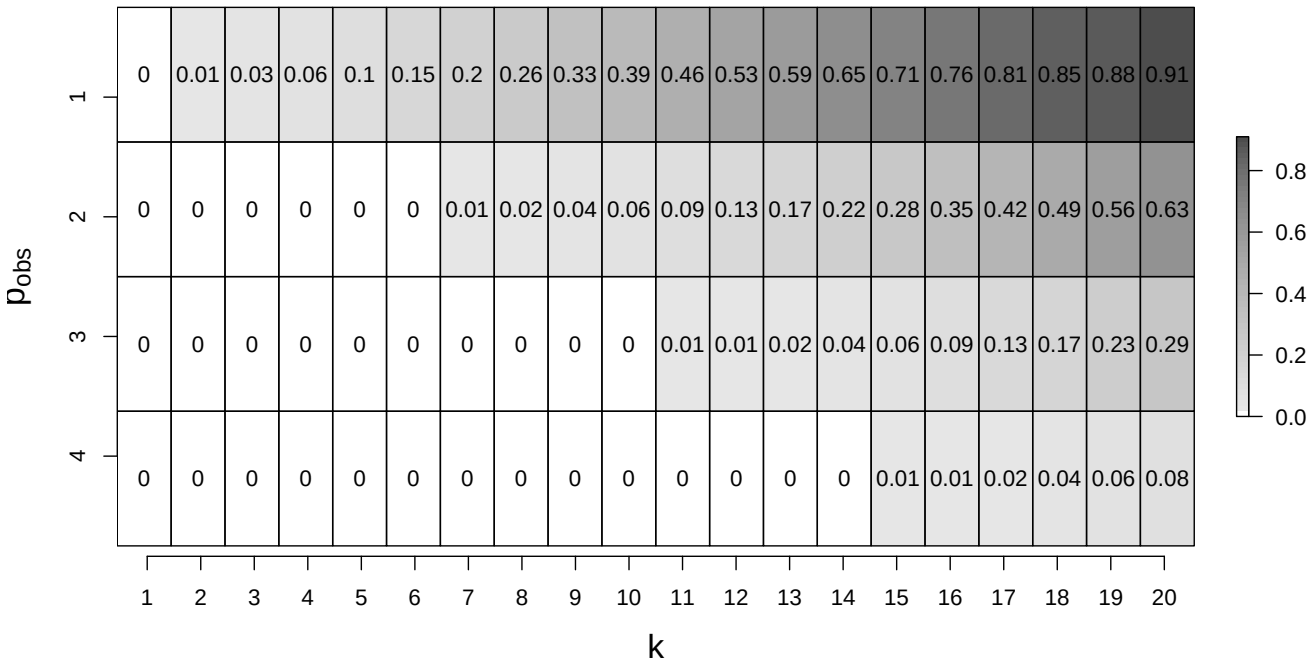


Figure 3.3 – LCAP values computed for $k \in [1;20]$ and $p_{obs} \in [1;4]$, for a set of 50 pairs of technical replicates. As an example, the probability to observe at least 2 duplicated entities when drawing 8 entities from a total of 50 pairs (initially duplicated entities) is equals to 0.02. Boxes are colored according to the corresponding LCAP value.

The figure 3.3 is showing examples of LCAP values for a set of 50 pairs of technical replicates. The LCAP values are computed for combinations of $k \in [1;20]$ (in x-axis, corresponding to the called mutations) and $p_{obs} \in [1;4]$.

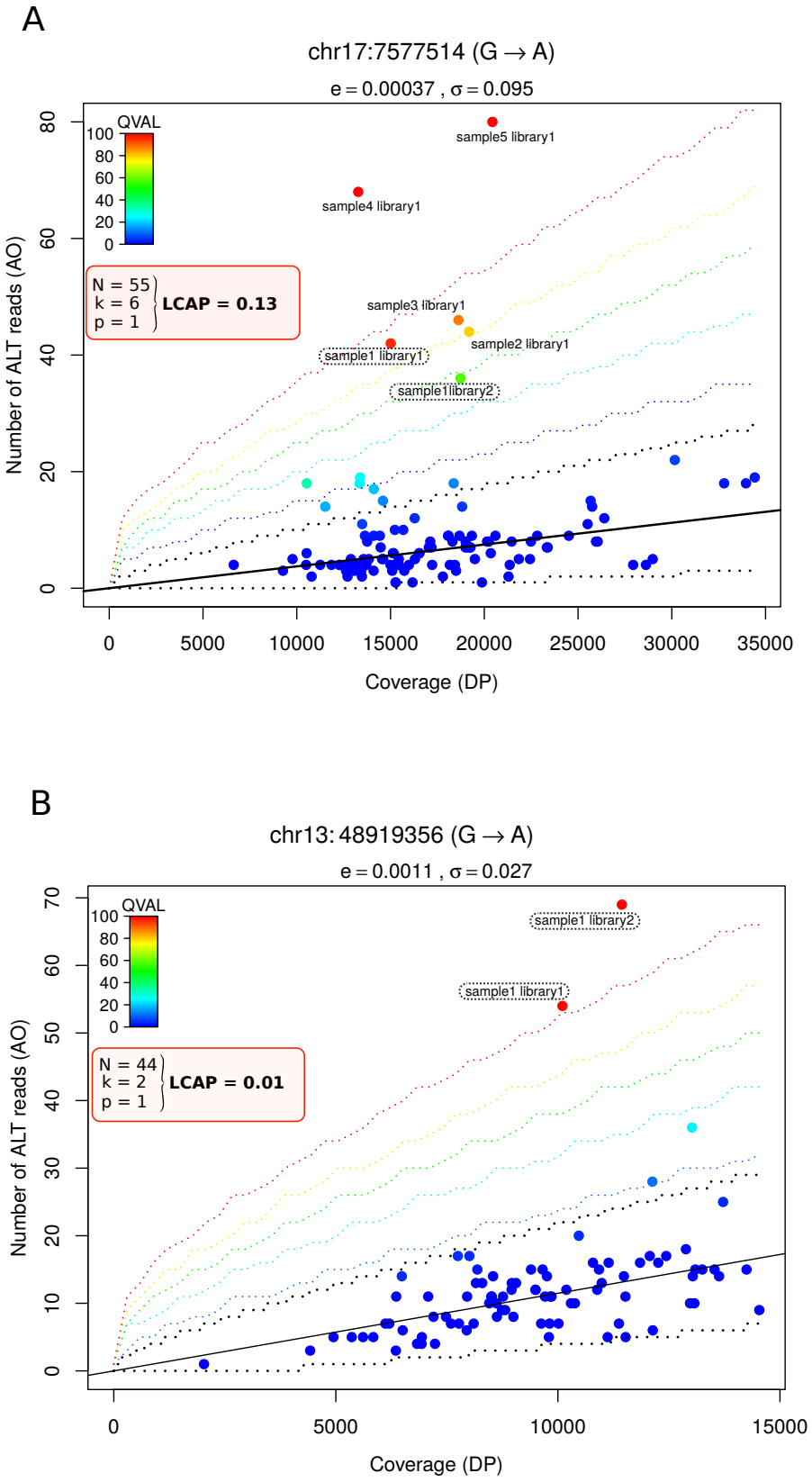


Figure 3.4 – Example of our LCAP statistic on two sequenced positions. Dashed lines surrounds technical duplicates. A: example of a position excluded based on low confidence. Six variants are detected, only two of them are found in technical replicates of the same sample, which corresponds to a p-value of 0.13 (>0.05, position is removed). B: example of a position maintained in the analysis. Three variants are identified and they form one pair, which corresponds to a p-value of 0.03 (<0.05, position is kept).

The figure 3.4 shows two examples of our LCAP statistic (with plots from needlestack), for two independent base changes. The panel A corresponds to an example of a low confidence alteration, and the panel B an example of a confident alteration kept in the analysis.

The fifth and last filter proposed in this part is based on the genomic distance of a mutation from a true DNA variation. A true mutation observed in one sample at a position p_t can create a false mutation in the same sample at a position p_f. This false mutation at p_f can be caused by an alignment artifact created by true variations on the same reads that corresponds to the true mutation at p_t [124] [145]. If r denotes the average length of the sequencing reads at the position p_t, we expect $|p_t - p_f| < r$. Errors caused by such alignment artifacts can be defined as both a pseudo and a non-systematic error.

Such errors created by an alignment artefact can potentially be found in very low proportion of sequenced reads [145], and this proportion is expected to be related to the proportion of the true mutation that is responsible for the alignment artifact.

Let ptm denotes one particular probably true mutation, pfm a particular possibly false mutation in the same individual and VAF_{ptm} and VAF_{pfm} their VAFs, respectively. A called mutation can be considered as "probably true" based on specific variables, and we propose to use the VAF variable for this task. Indeed, we propose that a "probably true" mutation should validate:

$$\text{VAF}_{ptm} >= e_{max}$$

With e_{max} corresponding to a *a priori* maximum proportion of reads that can correspond to errors from the NGS experiment.

The first "naive" way to define a possibly false mutation pfm is to consider each called mutation in the same sample a probably true mutation as defined above.

We can also define more specifically a possibly false mutation pfm: if a corresponds to the maximum expected alignment error rate, the "possibly false" nature of a called mutation can be defined validating the following rule:

$$\text{VAF}_{pfm} < \text{VAF}_{ptm} * a$$

A last step should be realized once both the set of (i) probably true mutations and (ii) the set of possibly false mutations are defined (these mutations should be found in the same sample and should respect the previously described rules). This last step corresponds to the classification of the possibly false mutations into finally true and false mutations. For this, we propose to compute for each possibly false mutation the smallest distance with a probably true mutation, and then filter on this statistic using a pre-defined threshold.

The scripts that implement this variant filtering framework are freely available and maintained on the GitHub plateform: https://github.com/IARCbioinfo/target-seq.

3.2.2 Application to ctDNA data

With collaborators, we have conducted a first study published in 2016 [49] (see details on chapter 4). In this study we were interested in the assessment of *TP53* gene variations in the blood of Small Cell Lung Cancer (SCLC) cases and control samples, as a proxy of the presence of a tumor. We reported *TP53* mutations in the blood of 49% SCLC patients and 11.4% of non-cancer controls (results were replicated in an independent validation cohort). Following these first results, we were suspicious concerning the fact that non-cancer controls seem contain *TP53* deleterious mutations in the blood, which was not expected given the actual state-of-the-art. We then decided to conduct a new study on an independent cohort of SCLC cases and controls (followed by a replication on a validation cohort) with (i) the addition of the *RB1* gene; (ii) and the application of our variant filtering methodology, in order to increase the potential specificity of the ctDNA as a tumor biomarker.

In this second study, we analyzed two independent cohorts, a discovery cohort and a replication cohort to validate the results. 253 samples were available for the discovery cohort and 172 samples for the validation cohort. Each samples was sequenced in two technical replicates. The quality control step based on the median coverage computed with the adapted QC3 tools on the two sequenced genes *TP53* and *RB1* removed 12 samples from the discovery cohort and 3 samples from the replication cohort (we required a median coverage of at least 1,000X in both the two technical replicates). Finally, a total of 241 discovery samples and 169 replication samples were analyzed.

To detect the potential mutations from these data, our needlestack algorithm was launched on BAM files containing sequencing reads with the IonTorrent Proton sequencer that were aligned against the human reference genome hg19 with the Torrent Suite Software [2], the default aligner provided with the sequencing on the IonTorrent Proton sequencer. Read bases with low sequencing confidence (base quality lower than 13 in Phred-scale) were not considered. Needlestack was applied independently on each sequencing run to avoid potential sequencing batch effect. Indeed, we assumed that the SER is not replicable across multiple independent runs, and therefore merging multiple runs can reduce the sensitivity of the algorithm to detect low VAF mutations that can potentially reach the error rate attributed to a different run. Needlestack Q-value threshold was set at 50 (default value). Then, the variant filtering methods described in the part 3.2.1 was applied on called data with needlestack.

Firstly, the samples presenting an unexpected high number of mutations in at least one of the two technical replicates were removed from the analysis. According to this, a total of 8 samples from the discovery cohort and 2 samples from the replication cohort were defined as non analyzable. This led to a discovery cohort totalling 50 cases and 183 controls and a validation cohort totalling 51 cases and 116 controls.

Then, we applied the other filters as the following:

- mutations found in one of the two replicates were not considered
- mutations with a strand bias (RVSB higher than 0.85) in at least one of the two replicates were removed
- alterations with low-confidence were not considered (LCAP$\geqslant$0.05)
- potential variant close to less than 5 base-pairs from a called mutation with a VAF higher than 10% were removed (MIN_DIST$\leqslant$5)

Figure 3.5 is a Venn diagram showing the number of removed mutations according to the filter applied and the corresponding overlapping between filters. Interestingly, the filter which removes the higher number of mutation independently from the others is the LCAP statistic. The major overlap between these filters is the intersection of LCAP statistic and the replicate requirement, suggesting that majority of non replicated mutations fall in low-confidence regions.

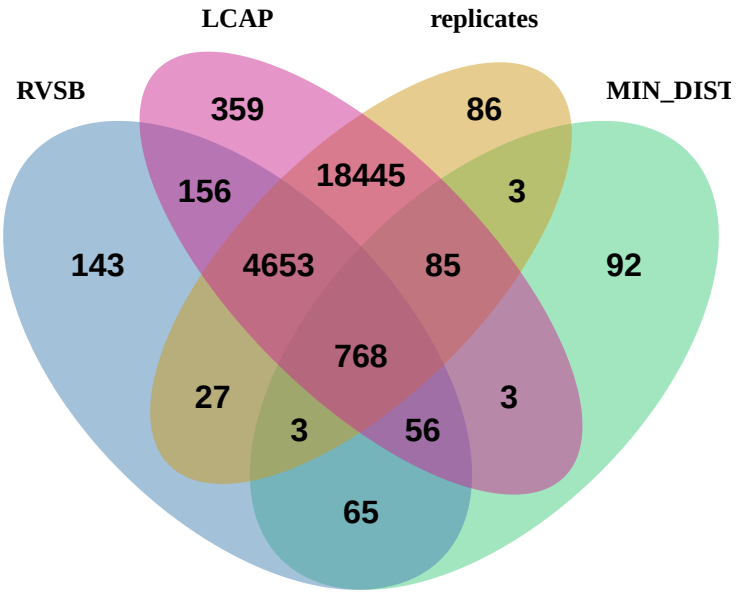


Figure 3.5 – Venn diagram presenting the logical relations in the collection of per-filter removed mutations from our ctDNA data (discovery cohort). A given ellipse correspond to a given filter and overlap between multiple ellipses correspond to commonly removed mutations.

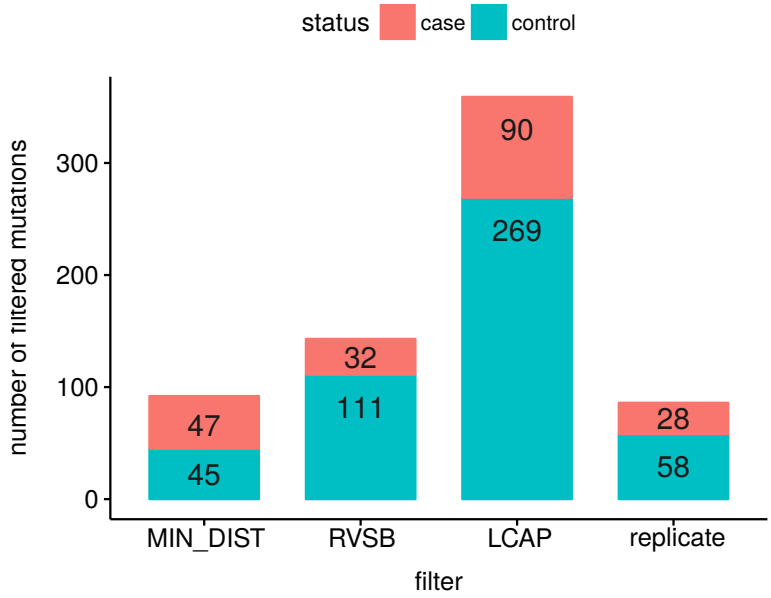


Figure 3.6 – Repartition of per-filter removed mutations in cases and controls from our ctDNA data (discovery cohort). The comparison between this repartition and the repartition of sequenced samples in cases and controls gives an information on the accuracy of the filter to removed errors (expected to be equally present in cases and controls) and not true mutations (expected to be more present in cases).

To test the accuracy of our filters in this study, we did not have validation of variants but nevertheless we have benefited from the availability of the case/control status of each mutations that passed the filtering step. Indeed, the errors should be randomly spread into cases and controls contrary to the true mutations that are expected to be found more in cases. This suggests that if a filter removes the errors and not the true mutations, the removed mutations should be found in the same proportion in cases versus controls than the initial proportion of cases and control samples. To validate the filters, we then computed the repartition in cases and controls of the removed mutations for each given filter. To estimate the accuracy of this filter independently of the others, we considered the mutations removed only by the filter (Figure 3.6). Fisher exacts test p-value on these repartition was significant (< 0.05) only for MIN_DIST filter, which is expected: indeed, this filter should remove artifacts created by true variant which are expected to be more present in cases.

Statistical validation

The global aim of this collaborative study was to developed a non-invasive biomarker based on the detection of tumor-derived mutations present in blood samples. For this, we used targeted NGS sequencing of both *TP53* and *RB1* genes from circulating cell-free DNA samples extracted from plasma. For this, we have developed a coupled laboratory and computational framework based on a case-control study. Nevertheless, according to our first published work [49], controls still contains mutations in these genes, which is not a result commonly demonstrated in the literature.

As we showed previously, errors are not expected to appear randomly across a genomic sequence (variability of the error rate) but tend to cluster at certain positions. We then wanted to test if the observed number of mutations in the technical replicates n_{obs} were in the same order than the number of duplicated mutations n_{exp} if we randomize the library labels. This would mean that our duplicated mutations can be the consequence of clustering of errors, that increase the probability to validate them in the two technical duplicates. For this, we performed a permutation test (randomization of library labels) to get the expected distribution of the number of mutations (the mean corresponds to n_{exp}).

Let S be the set of samples with s the total number of samples. $\forall i \in S$, $n_{i,lib1}$ and $n_{i,lib2}$ denotes respectively the number of mutations in the first and in the second library of the sample i. We picked randomly $n_{i,lib1}$ and $n_{i,lib2}$ mutations in a urn of size $\sum_{i=1}^{s}(n_{i,lib1}+n_{i,lib2})$ containing all the observed mutations. Then we computed pm_i denotes the number of duplicated mutations (see figure 3.7 for schematic representation of our simulations).

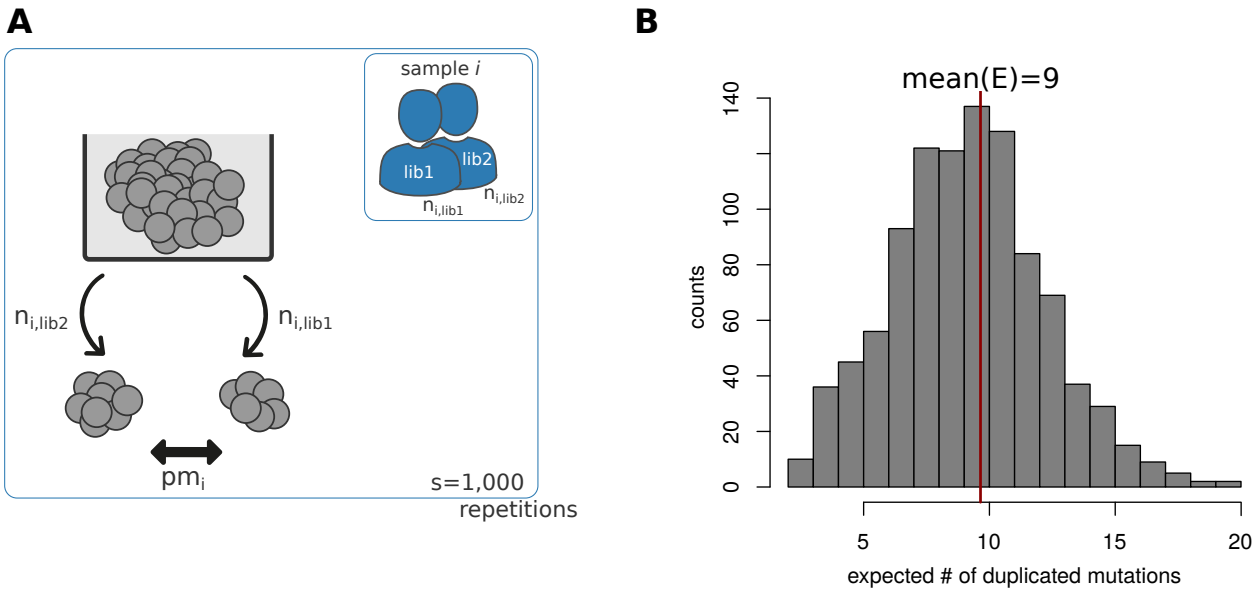


Figure 3.7 – A: Schematic representation of our permutation test design to estimate the expected number of duplicated mutations under a random process. This number corresponds to the expected number of duplicated errors, and can be used to infer the trustability of our final mutations. B: distribution of expected duplicated errors by random using our permutation test.

Finally the number of expected duplicated errors in the data would be defined as the following:

$$E = \sum_{i=1}^{s} pm_i$$

We repeated these simulations 1,000 times computed the distribution of E, that had an average of 9 (figure **??**). In our study we found a total of 162 duplicated mutations, which is significantly different from 9 (p-value < 0.001).

3.2.3 Article B (in preparation)

Diagnostic accuracy of joined *TP53* and *RB1* mutations in circulating tumor DNA for the non-invasive detection of small cell lung cancer

Patrice H. Avogbe[1*], Tiffany M. Delhomme[1*], Aurélie Gabriel[1], Dariush Nasrollahzadeh Nesheli[1], Florence Guida[1], Amélie Chabrier[1], Valerie Gaborieau[1], Behnoush Abedi-Ardekani[1], David Zaridze[2], Anush Mukeria[2], Ghislaine Scelo[1], Mattias Johansson[1], Florence Le Calvez-Kelm[1], Lynnette Fernandez-Cuesta[1], Paul Brennan[1], Matthieu Foll[1], James D McKay[1#]

[1]International Agency for Research on Cancer (IARC-WHO), 150 Cours Albert Thomas, 69008, Lyon, France
[2]Russian N.N. Blokhin Cancer Research Centre, Moscow, Russian Federation

* The two first authors contributed equally

[#] Corresponding author: James D McKay, email: McKayJ@iarc.fr

INTRODUCTION

Circulating tumor DNA (ctDNA) was currently emerging as a new non-invasive biomarker for cancer detection and treatment follow up, and its accuracy has been tested across several studies [1], [2]. As a potential cancer detection biomarker, ctDNA should both discriminate cancer patients from healthy individuals but also should make the distinction between a targeted cancer type and others. Therefore, this discrimination power is important to measure to find how accurate its diagnostic ability is. Nevertheless, a few studies reported mutations in the cell-free DNA (cfDNA) of controls that would be attributed to the presence of a tumor, because of the mutated gene function [3-5]. Interestingly, a recent study also reported cancer driver genes mutations in the cfDNA of healthy controls who remained cancer free after a 6 years follow-up [6].

A part of cancer driver gene mutations would possibly accounted for germline mutations, therefore it is important to also consider these type of variations, to remove the potentially false discoveries in term of expected tumor derived mutations. This process can be done by using public databases of common germline genetic variations [7] by can be improved by sequencing the matched white blood cell (WBC) samples to also detect non-common germline mutations.

In addition. cancer driver gene mutations can also be the consequence of clonal hematopoiesis and in the case should not be attributed to the presence of a tumor. Thus, extracting DNA from WBC can help resolving it by detection the mutations seen in WBC but without a germline origin.

One particular difficulty in term of methodology when trying to detect mutations in cfDNA is that the expected low proportion of such true mutations can reach the level of sequencing artefacts, and then it is crucial to precisely estimate the sequencing error rate in order to accurately detect mutations in low abundance, *i.e. harbouring* a low variant allelic fraction (VAF). In addition to the need of a precise variant calling algorithm that can detect low VAF mutations, to boost the precision of the mutation detection process in reduce the potentially false calls that can fall into control samples and change the overall interpretations of the results, a variant filtering step should be added subsequently to the variant calling process. This can be done by laboratory processes, such as sequence each sample in independent technical replicates to reduce

the errors from the library preparation, but this can also be done through statistical analyses. Even if the technical replicates of a same sample should reduce false calls, it can be possible that randomly a same error is found in the two technical replicates, and then it is important to consider and if possible to evaluate the proportion of random duplicated errors expected in the analysed dataset.

An important aspect to consider is the functionality of the detected mutations. Indeed, it is possible that mutations found in control samples are not functionally important compared to the one found in cancer patient samples. This can be done using a ctDNA score attributed to each sample that take into account the deleterious power of its mutations.

Small cell lung cancer (SCLC) accounts for about 15% of all lung tumors and harbour a very low 5-year survival, estimated below 5%. This cancer type is therefore a good candidate for developing a non-invasive cancer biomarker that can potentially detect cancer in early stages. The genomic architecture of SCLC tumors is characterized by recurrent somatic mutations in RB1 and TP53 [8]. We have recently conducted a study on the identification of ctDNA *TP53* mutations for the early detection of SCLC, and we have reported a propotion of positive controls at around 10% (with a validation in an independent cohort). We were interested in adding the second most mutated gene in SCLC, *RB1*, which was not reported to by highly mutated in other cancer types, in order to increase the specificity or our biomarker.

In this study we combined both laboratory and statistical frameworks in order to evaluate the diagnostic accuracy of both *TP53* and *RB1* mutations in the cfDNA of both SCLC cases and controls, aiming at developing a non-invasive biomarker for the detection of SCLC potentially in early stages. For this we developed a amplicon-based deep sequencing methodology in order to sequence the *TP53* and *RB1* genes from plasma samples, and a variant filtering methodology subsequently to an ultra-sensitive variant calling step with our needlestack algorithm [9] in order to precisely, both in term of sensitivity and specificity, identify ctDNA mutations. We tested the accuracy of our biomarker on both TCGA data, in order to estimate the capacity to distinguish between SCLC and other cancer types (using tumor data), and on two case-control cohorts, in order to estimate the capacity to attribute correctly a case-control status (using cfDNA data).

In spite of difference in prevalence of TP53 (56.9) and RB1 (45.1) mutations among cases (p=0.02), a similar proportion of controls with cfDNA mutations was observed for RB1 (17.2) and TP53 (16.4). Adding functional score to the mutations did not change this picture. We conclude that presence of mutations among controls when using genes instead of variant as a targeted sequencing results in limitations. This finding is in line with TCGA data in which for cancers with low TP53 mutations no privilege was obtained by adding RB1 data. Also we observed that TP53 alone can discriminate cases from control using scores equal to combination of TP53 and RB1.

MATERIAL AND METHODS

Study population and sample collection in the discovery and validation cohorts

SCLC cases and controls were recruited through two large case-control studies (Bardin-Mikolajczak et al., 2007; Fernandez-Cuesta et al., 2016; Wozniak et al., 2015) coordinated by IARC: the Russia multicenter study conducted between 2006 and 2012 (discovery cohort) and the CEE study conducted between 1998 and 2001 (validation cohort). Briefly, each study centre followed an identical protocol and was responsible for recruiting a consecutive group of newly diagnosed cases of lung cancer and a comparable group of controls with no known history of cancer. Controls were from the same hospitals or neighboring

general hospitals where the cases originated. Cases were recruited before they receive surgery or any adjuvant treatment. Clinical staging of the SCLC cases was done following recommendations of the International Association for the Study of Lung Cancer (IASLC) (Nicholson et al., 2016). The recruitment involved collection of smoking history and other epidemiological data, blood samples (10-15 ml in EDTA tubes) as well as, wherever possible, collection of a surgical resection of the tumors. Blood samples were centrifuged at 2,000xg for 10 min at room temperature to separate plasma from peripheral blood cells and stored at –80ºC until use.

A total of 253 individuals were included in the discovery cohort but only 233 passed sequencing QC criteria: median age was 61.8 (range 38.0–78.4) and 65.6 (range 43.1–77.4) in SCLC and controls, respectively (Table 1). Age did not differ significantly between SCLC cases and controls. Of the 172 individuals included in the replication cohort, 167 passed all quality control steps. Baseline characteristics of patients are summarized in Table 1: median age was 58.0 (range 41.0–74.0) and 60.0 (range 38.0–74) in SCLC and controls, respectively. The most common stage at disease presentation in the study cohorts was stage III (51%, 50/98); 22.4% (22/98) had stage I-II tumors (clinical stage was unknown in 3 patients). All participants provided written informed consent and the study complied with the ethical guidelines of the declaration of Helsinki and was approved by relevant local ethical review committees and the IARC Ethics Committee.

TCGA data

We retrieved somatic mutations in *RB1* and *TP53* from the whole genome sequencing of 110 SCLC tumors [8], and from 10,202 cancer cases other than SCLC available in the TCGA database (33 cohorts), using the TCGAbiolinks Bioconductor package. Due to the difference in sequencing techniques between the SCLC dataset and the TCGA dataset (WGS vs WES respectively), we selected mutations located in coding and splicing regions. Also, as previously shown [11], WES based on different exome capture kits exhibit non-uniform coverage in several genes, including RB1. Thus, exons recurrently having a mean sequencing depth <15X across TCGA samples were removed from the analysis. Selected mutations were re-annotated with ANNOVAR, which allowed us to grade variants on the basis of their putative impact on the gene product.

Extraction of cfDNA and genomic DNA from tumor tissue and WBC

Cell-free DNA was purified from plasma (volumes range 0.4 –1 ml) using the QIAamp Circulating Nucleic Acid Kit (Qiagen). DNA from microdissected fresh-frozen tumor tissue (10 sections of 20-µm thick with >80% of tumor cell content) was extracted using the Gentra Puregene Tissue Kit (Qiagen).

For each cfDNA-positive individual (*i.e.*, patient with at least one mutation identified in their cfDNA), we also undertaken sequencing analyses of paired WBC DNA to exclude the possibility of somatic WBC mosaicism and germ line variants as a possible source of a positive ctDNA finding. Genomic DNA was extracted from white blood cells (WBC) using the QIAamp 96 DNA Blood Kits (Qiagen), according to the manufacturers' instructions. Extracted DNA samples were quantified using the Qubit dsDNA HS Assay kit (Invitrogen).

Primer design and amplification of targets

A total of 70 primers pairs (49 in *RB1* and 21 in *TP53*) were synthetized commercially by Eurofins Genomics (Ebersberg, Germany) for a total panel size of 5676 bp to cover the coding regions on *RB1* (83%, exon 2–

27) and *TP53* (94%, exon 2-11). Prior to undertaking ctDNA analysis, we optimized our multiplex PCR-based Gene-Read assay (Qiagen) and verified appropriate base coverage of all the target bases for sensitive and efficient variant detection (**Supplementary Figure 1**). We next dispensed 5 ng cfDNA in 96-well format plates and performed target enrichment using the GeneRead DNAseq Panel PCR Kit V2 (Qiagen). Briefly, single pool, multiplexed PCR reactions were performed in 10 µL with final concentrations of 30 nM of each primer, 4.4 U GeneRead HotStarTaq ® DNA Polymerase, 1X of the GeneRead DNAseq Panel 5X PCR buffer, and DNA template (5 ng). All amplifications were carried as follows: 15 min at 95 °C and 30 cycles of 15 s at 95 °C and 2 min at 60 °C and 10 min at 72 °C. The amplified DNA was then purified using the Serapure beads, and quantified by Qubit dsDNA HS Assay kit (Invitrogen).

Library preparation and sequencing

Libraries were constructed using the NEBNext Library Prep Set for Ion Torrent (BioLabs, New England) and 150–200 ng of purified PCR products. The amplicons were ligated to the specific adapters and individual IonXpress™ barcodes with a subsequent purification step using the Serapure beads. Adapter-carrying fragments were further amplified using the Q5 Hot Start High-Fidelity 2X Master Mix. Next, ~ 200 ng of individual libraries were pooled into batches of 45 samples. An aliquot of each batch was loaded onto a 2% agarose gel for electrophoresis (150 V, 1.3 h). Fragments of 180–220 bp were recovered from the gel using the QIAquick gel extraction kit (Qiagen). The quality and quantity of the library were then assessed on the Bioanalyzer 2100 platform (Agilent Technologies, USA). Purified libraries were enriched by clonal amplification using emulsion PCR on Ion Sphere particles, with subsequent elimination of non-templated Ion Sphere Particles beads by magnetic bead purification (Ion PI ™ Hi-Q™ OT2 200 Kit, Life Technologies Corp., USA). Finally, the target-enriched libraries were deep sequenced (sequencing depth > 10,000X) on the Ion Torrent™ Proton Sequencer using the Ion PI ™ Hi-Q™ Sequencing 200 Kit with the Ion PI v3 (Thermo Fisher Scientific, USA) following the manufacturer's protocol.

Each DNA sample was tested as technical duplicate, including amplification, library preparation and sequencing. PCRs and library duplicates were undertaken on physically two distinct 96-well format plates for each sample to minimize contamination. We only considered variants found in both libraries, to guard against rare errors specific to a particular library. All operators were blinded to case control status.

Variant detection

Needlestack [9], a recently developed low abundance mutation caller was used to perform the variant calling. Needlestack is based on the idea that analyzing multiple samples together can help estimating the distribution of sequencing errors to accurately identify variants present in very low proportion. At each position and for each candidate mutation, needlestack models sequencing errors using a robust negative binomial regression [3] with a linear link and a zero intercept. Variants are detected as being outliers from this error model for the corresponding mutation [Figure 1]. Needlestack calculates for each sample a *p*-value for being a variant (*i.e.* outlier from the regression) and transforms it into a *q*-value using the Benjamini and Hochberg method to account for multiple testing and control the false discovery rate. *Q*-values are then transformed into a Phred-scale: $QVAL = -10 * log_{10}(q\text{-value})$, and a sample is considered positive for the mutation if QVAL>50.

Detected mutations were annotated using Annovar [4], and variants with a minor allelic frequency (MAF) higher than 0.5% in genetic variant databases [REF] were rejected in order to remove potential

germline variants. We then applied a stepwise filtering strategy to the variants called in order to boost the precision of the mutation detection, and finally retained only validated variants found in the two technical duplicates of a sample. Filtering strategy is defined as follows: (1) removing variants in strand bias, *i.e.*, with a relative variant strand bias (RVSB) higher than 0.85; (2) filtering out any variants present in the 5 base pairs neighborhood of a strong SNV defined by a VAF higher than 10% to correct for misalignments (MIN_DIST filter); (3) removing low confidence base change in term of sequencing. A low confidence base change satisfied a low confidence alteration probability (*LCAP*) lower than 0.05 with *LCAP* defined as follows:

$$LCAP = \sum_{p=p_{obs}}^{p_{max}} \left[\prod_{i=0}^{p-1} \binom{k-2i}{2} \right] \left[\prod_{j=0}^{k-p-1} (2N-2j) \right] \frac{1}{p!\,k! \binom{2N}{k}}$$

This *LCAP* statistic corresponds to the probability of observing at least p_{obs} pairs of elements by a random sampling of k elements from a set of $2N$ elements (N pairs). In other words, the *LCAP* statistic can be defined as the probability that the variants identified in the two technical duplicates are due to a random sequencing noise, *i.e.*, correspond to a low confidence base change leading to misinterpretations. For these base changes, evident variants that harbour a VAF higher than 10% were kept.

Statistical validation of technical replicates

To evaluate the filter on technical duplicates, we calculated the expected probability of finding the same variants in two technical duplicates of a particular sample as a result of random errors (Figure 2). We considered only mutations that passed the three other filters in order to into account only the errors that remain after the filters. For each sample i, we computed the observed total number of mutations in the first and second libraries, respectively $n_{i,lib1}$ and $n_{i,lib2}$. Then, we built a subset of independently and randomly picked $n_{i,lib1}$ and $n_{i,lib2}$ mutations from the complete set of observed mutations across all individuals (Figure 2-A). Finally, we computed pm_i, the number of paired mutations in this subset (corresponding to random replicates, or duplicated errors) appearing in two different samples in order to do not consider true second replicates (Figure 2-B). This number pm_i corresponds to the expected number of duplicated errors for the sample i. We replicated this process 1,000 to take into account potential variability. To compute the total expected number of duplicated error in our series, we computed the mean over the 1,000 replications of the sum of all the pm_i across the samples (Figure 2-C).

Score development

First, we attributed to each detected mutation an impact value ranging from 0 to 2 reflecting the deleterious nature of the mutation: 0.5 for intronic or synonymous variants, 1 + REVEL score for missense variants and 2 for stopgain, splicing or frameshift variants. A sample-score, for each gene, was then computed as the maximum of the impact values when multiple mutations are called in the same gene of the same sample. A sample-score of 0 is hence attributed to samples without mutations. These sample-scores computed for *TP53* and *RB1* independently were used to build three logistic regression models: (i) a model including the sample-score of *TP53* mutations, (ii) a model including the sample-score of *RB1* mutations and (iii) a model combining the two sample-scores.

For each model we estimated the parameters of the regression on a subset of the data (training). These learned parameters allow us to compute a per-sample ctDNA genetic score when applied on a validation set (test). We then estimated the ability of our biomarker to distinguish between cases and controls, using ROC curves.

For the TCGA data, we performed a five-fold cross-validation to train the three regression models, predict cases statuses (SCLC vs the other cancer type) and to evaluate the performance of each model. The mean AUC value across the five folds was computed to assess models performances. To assess if the combination of both genes scores (third regression model) is relevant to discriminate SCLC cases, for each fold we compared the performance of the regression models considering each gene individually to that of the regression model considering the two genes. The difference was considered as significant if the median *p*-value was lower than 0.05.

For the case-control cohorts, we tested the two combinations, *i.e.* first, we used the cohort 1 to estimate the regression coefficients and applied them on the cohort 2, then secondly we used the cohort 2 as a training cohort and the cohort 1 as a validation cohort.

RESULTS

Sample data

Majority of patients in our discovery series were diagnosed at late stage III (56%), and stage I/II consisted 28% of discovery phase patients. (Table 1). The median age was 61.8 years (range 38.0–78.4) in SCLC and 65.6 years (range 43.1–77.4) in controls (Table 1). Twenty-three SCLC in this series had previously assessed, carrying a total of 28 *TP53* mutations (Fernandez-Cuesta et al., 2016), which allowed us to evaluate reproducibility of our method.

ctDNA variant detection and exclusion of variants of WBC origin

We identified 162 variants in 51.5% (120/233) of patients in the discovery cohort, and 125 variants in 44.3% (74/167) of patients and the validations cohorts. Subsequently, we undertook deep sequencing analyses of matched WBC DNA in cfDNA-positive individuals to exclude the possibility of somatic WBC mosaicism and germ line variants as a possible source of a positive ctDNA finding. These analyses allowed us to removed variants co-occurring in the cfDNA and the WBC: 11.7% (19/162) and 12.8% (16/125) in the discovery and validation cohorts, respectively. These mutations were found in 28.6% (18/63) and 18.4% (9/49) of controls, and 2.5% (2/81) and 11.7% (7/60) in SCLC in the discovery and validation cohorts, respectively. Overall, 24% (27/112) of variants found in non-cancer controls were explained by WBC analysis. These variants observed in cfDNA were also found in the WBC, either as germ line variants, but also frequently in very low abundance. We observed that the variant allelic fractions (VAFs) of the variants co-occurring in the cfDNA and the WBC are highly correlated (Supplementary Figure 1). The challenge faced here is to determine when a particular mutation is not observed in the WBC if this is due to a true biological absence, or to a lack of statistical power to detect it. This situation is further complicated given that the allelic fraction expected in the WBC is unknown, whereas for germ line variants we expect it to be 50% for heterozygotes. We then extended our approach and checked for each detected cfDNA variant, if we have the statistical power to detect this allele in the matched sequenced WBC, given the observed depth coverage and the error rate for this particular base-change as inferred by Needlestack. This analysis revealed that we were powered

enough to detect 93.8% (152/162) and 97.6% (122/125) of variants detected in the first and second cohorts, respectively. Finally, we detected 252 ctDNA variants in in the study cohorts that were included in down-stream analyses.

Characteristics of ctDNA mutations detected in the study cohorts

In the discovery cohort, we found 80 variants in SCLC (35 variants in *RB1* in 28 patients, including 7 SCLC with two variants, and 45 variants in *TP53* in 35 patients of which 8 had 2 variants and one had 3 variants); and 63 variants in controls (29 in *RB1* from 25 individuals, of which 4 had two variants; and 34 variants in *TP53* from 30 individuals, of which 4 had two variants). SCLC patients had significantly higher VAFs compared with controls (median 1.06%, range 0.055–80.82% versus median 0.27%, range 0.023–12.63%, respectively, $P = 3.30 \times 10^{-11}$, Table 2). We also observed high mutation burden in SCLC compared with controls (1.58 versus 0.34, respectively, $P = 3.05 \times 10^{-17}$, table 2).

In the validation cohort, 60 variants were found in SCLC: 25 SNVs in *RB1* in 23 patients, including 2 with 2 variants versus 35 variants in *TP53* in 29 SCLC, of which 4 had 2 variants and one had 3 variants. In controls, 49 variants were found: 26 in *RB1* from 20 individuals (4 had 2 variants and one had 3 variants) and 23 in *TP53* from 19 individuals, including 4 with 2 variants. VAFs were significantly higher in SCLC compared with controls (median 1.81, range 0.061–73.91 versus median 0.32, range 0.087–5.18, respectively, $P = 7.36 \times 10^{-16}$, table 2). The average number of mutations per sample was higher in SCLC than in controls (1.18 versus 0.42, respectively, $P = 9.31 \times 10^{-6}$, table 2). Analysing both cohorts together, we found that positive mutant *RB1* or *TP53* status was significantly associated with early stage SCLC, when comparing patients with stage I-II disease to healthy individuals (Fisher's exact test, all *P-values* were < 0.0006). VAFs were significantly higher in stage III-IV tumors compared with stage I-II, when analyzing results, either cohort individually or both cohorts together (all *P-values* were < 0.0001). Similarly, VAFs in *RB1* and *TP53* were significantly higher in SCLC compared with controls (figure 2C-D).

The proportions of missense, nonsense, indels, or splicing mutations as well as silent mutations were significantly elevated in SCLC patients compared with controls when analyzing results, either cohort individually (table 2). In SCLC patients, mutational profiles were dominated by nonsense, indel, or splicing mutations in *RB1* (53.3%) and by missense variants in *TP53* (55%), whereas the pattern was quite different in both genes in controls (figure 2A). Mutational pattern observed in *RB1* and *TP53* in the cfDNA of SCLC was similar to that obtained in SCLC tumors in a recent study (George et al., 2015), when considering the same genomic coordinates covered by our assay (80% versus 63%, respectively, figure 2A). Given this particular mutational profile, we compared AFs of missense and truncating variants and found that *TP53*-mutated SCLC with missense variants had significantly higher AFs compared with *RB1*-mutated SCLC or controls carrying missense variants ($P < 0.05$). Similarly, *RB1*-mutated SCLC with nonsense, indels, or splicing mutations had significantly higher AFs compared to *TP53*-mutated SCLC or controls ($P < 0.05$; figure 2E & 2F).

ctDNA detection rates in *RB1* and *TP53* in the study cohorts

In the discovery cohort, the ctDNA detection rate was 84% (42/50) in SCLC versus 27.3% (50/183) in controls. In stage I-II tumors, the ctDNA detection rate was 85.7% (12/14) versus 85.7% (30/35) in stage III-IV tumors. At gene-level analyses, we identified circulating *RB1* and *TP53* mutations in 56% (28/50) and 70% (35/50) of SCLC cases versus 13.7% (25/183) and 16.4% (30/183) in healthy controls, respectively. *RB1* co-

altered with *TP53* in 40% (20/50) of SCLC versus 3.3% (6/183) in controls.

In the validation cohort ctDNA positivity was 66.7% (34/51) in SCLC and 29.3% (34/116) in controls. The detection rate was 62.5% (5/8) in stage I-II tumors versus 70.7% (29/41) for stage III-IV tumors. At gene-level, *RB1* and *TP53* variants were detected in 45.1% (23/51) and 56.9% (29/51) of SCLC versus 17.2% (20/116) and 16.4% (19/116) healthy controls, respectively. *RB1* co-altered with *TP53* in 33.3% (17/51) of SCLC versus 4.3% (5/116) in controls.

In ctDNA positive patients, sensitivities ranged from 45 to 56% in *RB1* and 57 to 70% in *TP53*. These proportions matched those reported in the current study by Almodovar and colleagues (52% and 70%, respectively) (Almodovar et al., 2018). The proportion of SCLC with *TP53* mutations in this study is higher than in our previous study, whereas the fractions of controls with *TP53* variants were comparable across study cohorts (Fernandez-Cuesta et al., 2016).

Observed replicated variant are not random errors

The total expected number of duplicated errors across the first cohort was estimated as 9.6 mutations in average across our 1.000 replications, *i.e.* we expect around 10 errors being duplicated by randomness that would be considered as true mutations. In this first cohort, we have detected a total of 162 duplicated variants, suggesting that this finding should not be a result of randomly replicated errors (p-value < 10^{-3}, figure 3-C).

Variant filtering strategy

We applied a total of four filters in the initial set of candidate mutations, (i) strand bias filter, (ii) MIN_DIST filter, (iii) LCAP filter and (iv) technical duplicate requirement. We show with a Venn diagram (figure 4-A) the concordance of removed mutations given the previously described filters, and reported the maximum concordance for the pair of replicates and *LCAP* filters, which is expected due to the fact that *LCAP* statistic is based on the validation of the mutations in the technical replicate. Technical replicates can be used as a sort of validation of the other filters, because errors are not expected to be validated in the replicate, as we shown in the last paragraph. We also show that the repartition of the filtered mutations across case and control samples corresponds to to the case-control initial repartition of the samples (figure 4-B), suggesting that filtered mutations are randomly distributed across samples, which is expected if our filters remove errors. Nevertheless, our MIN_DIST filter is related to the presence of true high VAF variant, and so true errors removed by this filter should be more present in cases, as we show (figure 4-B).

Score on TCGA

We first compared the occurrence of *TP53* and *RB1* mutations in 110 SCLC tumors (George et al., 2015) to 33 other cancer types available in the TCGA database (figure 5). SCLC had the highest proportion (66%) of patients carrying concurrent mutations in *TP53* and *RB1*, which was significantly different than in the other TCGA cancer types (Fisher's test: all p-values < 10^{-13}). Of the 33 cohorts, only the bladder cancer (BLCA) cohort presented concurrent mutations in both genes in more than 10% of the cases (figure 4). Overall, our data confirm that occurrence of concurrent mutations in *TP53* and *RB1* genes is less frequent in cancer types other than SCLC.

We applied the three logistic regression models based on *TP53* and *RB1* mutations scores (see Score development method) in order to asses the utility of *TP53* and *RB1* to distinguish SCLC cases from

other cancer types. AUCs ranged from 0.84 to 0.98, 0.48 to 0.96, and 0.79 to 0.88 for the regression models based on both *TP53* and *RB1*, *TP53* only and *RB1* only respectively (figure 5). To distinguish SCLC from cancer types showing high proportions of patients having *TP53* mutations (e.g., UCS, OV, ESCA, LUSC), the performance of the two genes does not significantly differ from that of *RB1* alone, signifying that *RB1* alone might be sufficient rather than the combination (figure 5). Conversely, to distinguish SCLC from cancer types showing low proportions of patients having *TP53* mutations (e.g., UVM, TGCT, THCA), including only *TP53* in the model seems to be sufficient (figure 5). However, discriminating SCLC from cancer types such as BLCA, SKCM, LIHC, BRCA, UCEC and GBM requires the use of the two genes.

Score on the case-control cohorts

We evaluate our ctDNA genetic score our two independent case-control cohorts (dataset 1 and dataset 2), by computing the AUC of the ROC curves for our three models (*TP53*, *RB1*, and the combination of the two genes) and for a model without our ctDNA score (taking into account only presence or absence of mutations). We were interested in testing the effect in the AUC of the addition of *RB1* gene compared to the *TP53* gene. For this, we took as a reference the *TP53* model, and compared it with the other models by computing *p*-values comparing the AUCs of the models. We showed that adding the mutations of the *RB1* gene in our biomarker is significantly better than sequenced only *TP53* when the second cohort is used as a training cohort and the first as the validation cohort (table 3). If we reverse the cohorts, this finding is not significant, suggesting high instability of the results.

CONCLUSION

In spite of difference in prevalence of TP53 (56.9) and RB1 (45.1) mutations among cases (p=0.02), a similar proportion of controls with cfDNA mutations was observed for RB1 (17.2) and TP53 (16.4). Adding functional score to the mutations did not change this picture. We conclude that presence of mutations among controls when using genes instead of variant as a targeted sequencing results in limitations. This finding is in line with TCGA data in which for cancers with low TP53 mutations no privilege was obtained by adding RB1 data. Also we observed that TP53 alone can discriminate cases from control using scores equal to combination of TP53 and RB1.

REFERENCES

[1] Le Calvez-Kelm, F., Foll, M., Wozniak, M. B., Delhomme, T. M., Durand, G., Chopard, P., … Scelo, G. (2016). KRAS mutations in blood circulating cell-free DNA: a pancreatic cancer case-control. Oncotarget, 7(48).

[2] Cheng, F., Su, L., & Qian, C. (2016). Circulating tumor DNA: a promising biomarker in the liquid biopsy of cancer. Oncotarget, 7(30).

[3] Fernandez-Cuesta, L., Perdomo, S., Avogbe, P. H., Leblay, N., Delhomme, T. M., Gaborieau, V., … Brennan, P. (2016). Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. EBioMedicine, 10, 117–123.

[4] Newman, A. M., Lovejoy, A. F., Klass, D. M., Kurtz, D. M., Chabon, J. J., Scherer, F., … Alizadeh, A. A. (2016). Integrated digital error suppression for improved detection of circulating tumor DNA. Nature

Biotechnology, 34(5), 547–555. https://doi.org/10.1038/nbt.3520

[5] Schwaederle, M., Husain, H., Fanta, P. T., Piccioni, D. E., Kesari, S., Schwab, R. B., … Kurzrock, R. (2016). Detection rate of actionable mutations in diverse cancers using a biopsy-free (blood) circulating tumor cell DNA assay. Oncotarget, 7(9).

[6] Gormally, E., Vineis, P., Matullo, G., Veglia, F., Caboux, E., Le Roux, E., … Hainaut, P. (2006). TP53andKRAS2Mutations in Plasma DNA of Healthy Subjects and Subsequent Cancer Occurrence: A Prospective Study. Cancer Research, 66(13), 6871–6876.

[7] Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16), e164–e164.

[8] George, J., Lim, J. S., Jang, S. J., Cun, Y., Ozretić, L., Kong, G., … Thomas, R. K. (2015). Comprehensive genomic profiles of small cell lung cancer. Nature, 524(7563), 47–53

[9] Delhomme *et* al., submitted in Nucleic Acid Research in april 2018. Available at: https://github.com/IARCbioinfo/needlestack

[10] Wozniak, M. B., Scelo, G., Muller, D. C., Mukeria, A., Zaridze, D., & Brennan, P. (2015). Circulating MicroRNAs as Non-Invasive Biomarkers for Early Detection of Non-Small-Cell Lung Cancer. PLOS ONE, 10(5), e0125026.

[11] Bonfiglio, S., Vanni, I., Rossella, V., Truini, A., Lazarevic, D., Dal Bello, M. G., … Coco, S. (2016). Performance comparison of two commercial human whole-exome capture systems on formalin-fixed paraffin-embedded lung adenocarcinoma samples.

Figure 1:

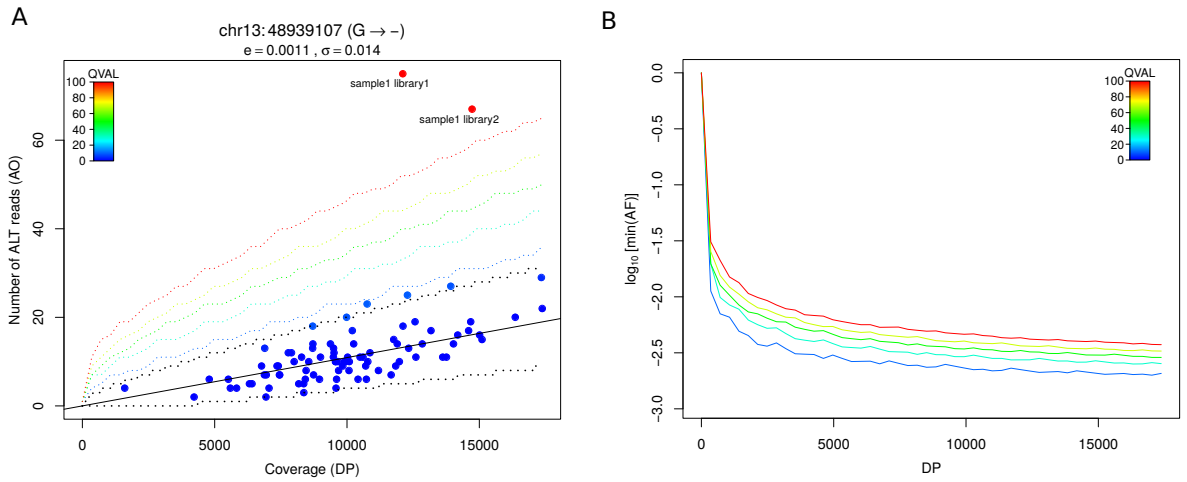


Figure 3:

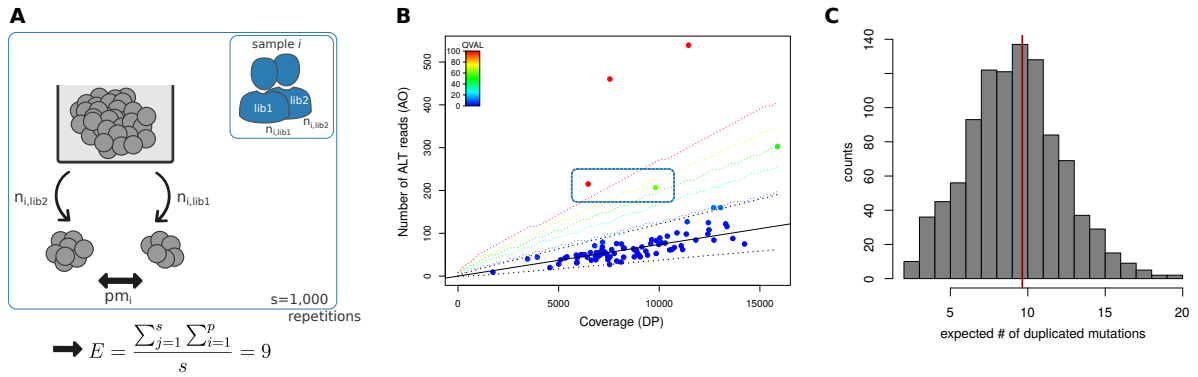


Figure 4:

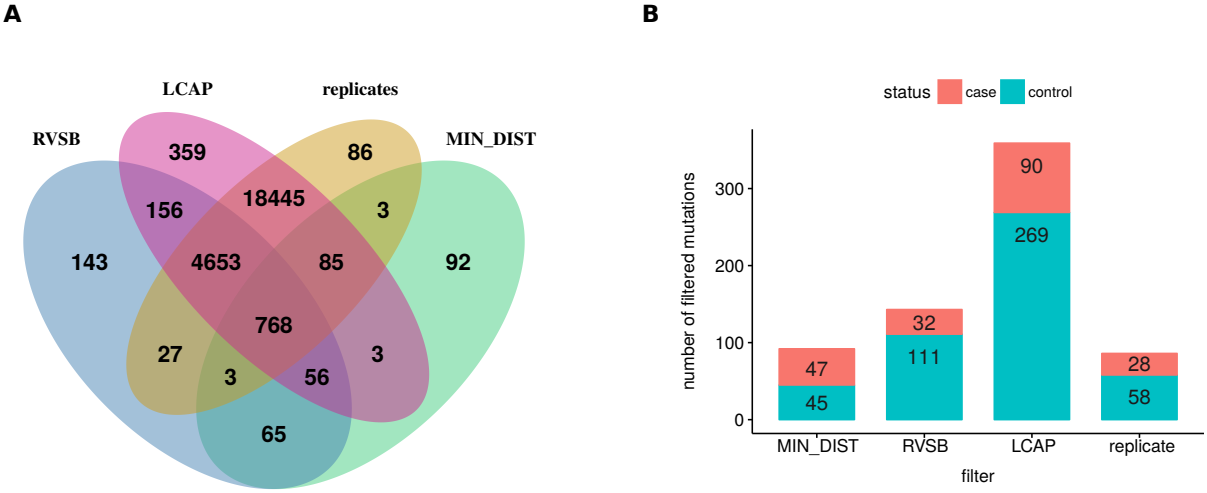


Figure 5:

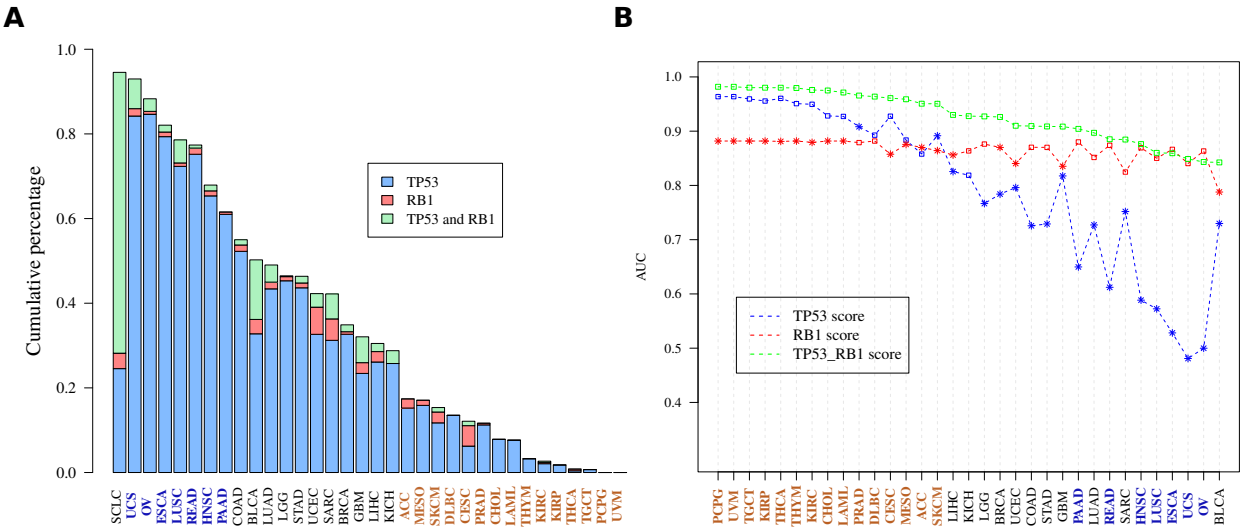


Table 1:

	First set		Second set	
	Cases (N=50)	**Controls (N=183)**	**Cases (N=51)**	**Controls (N=116)**
Age at diagnosis, years				
Median (range)	61.8 (38.0–78.4)	65.6 (43.1–77.4)	58.0 (41.0–74.0)	60.0 (38.0–74.0)
Sex, N (%)				
Male	40 (80.0%)	150 (82.0%)	34 (66.7%)	78 (67.2%)
Female	10 (20.0%)	33 (18.0%)	17 (33.3%)	38 (32.8%)
Smoking status, N (%)				
Never Smoker	5 (10.0)	55 (30.0%)	1 (2.0)	44 (37.9%)
Ex-smoker	5 (10.0)	54 (29.5%)	7 (13.7%)	42 (36.2%)
Current smoker	40 (80.0%)	71 (38.8%)	43 (84.3%)	30 (25.9%)
Unknown	0 (0.0%)	3 (1.6%)	–	–
Pack years				
Median (range)	36.75 (0–100)	16.98 (0–51.6)	32.5 (0–83.0)	10.0 (0–58.3)
Alcohol status, N (%)				
Never drinker	26 (52.0%)	63 (34.4%)	–	–
Ex-drinker	5 (10.0%)	27 (14.8%)	46 (90.2%)	104 (89.7%)
Current drinker	19 (38.0%)	93 (50.8%)	5 (9.8%)	12 (10.3%)
Tumor stage, N (%)				
I	7 (14.0%)	–	3 (5.9%)	–
II	7 (14.0%)	–	5 (9.8%)	–
III	28 (56.0%)	–	22 (43.1%)	–
IV	7 (14.0%)	–	19 (37.3%)	–
Unknown	1 (2.0%)	–	2 (3.9%)	–
TP53 and RB1 status				
TP53+, RB1-	15 (30)	23 (12.6)	12 (23.5)	14 (12.1)
TP53-, RB1+	7 (14)	19 (10.4)	6 (11.8)	15 (12.9)
TP53+ OR RB1+	42 (84)	48 (26.2)	35 (68.6)	34 (29.3)
TP53+ & RB1+	20 (40)	6 (3.3)	17 (33.3)	5 (4.3)
TP53 and RB1 score mean (SD)				
TP53 score	1.2 (0.9)	0.18 (0.46)	1.0 (0.9)	0.21 (0.5)
RB1 score	0.9 (0.9)	0.18 (0.49)	0.7 (0.9)	0.22 (0.5)

Table 2:

	Discovery cohort		Replication cohort	
	Cases (N=50)	**Controls (N=183)**	**Cases (N=51)**	**Controls (N=116)**
Age at diagnosis, years				
Median (range)	61.8 (38.0–78.4)	65.6 (43.1–77.4)	58.0 (41.0–74.0)	60.0 (38.0–74.0)
Sex, N (%)				
Male	40 (80.0%)	150 (82.0%)	34 (66.7%)	78 (67.2%)
Female	10 (20.0%)	33 (18.0%)	17 (33.3%)	38 (32.8%)
Smoking status, N (%)				
Never Smoker	5 (10.0)	55 (30.0%)	1 (2.0)	44 (37.9%)
Ex-smoker	5 (10.0)	54 (29.5%)	7 (13.7%)	42 (36.2%)
Current smoker	40 (80.0%)	71 (38.8%)	43 (84.3%)	30 (25.9%)
Unknown	0 (0.0%)	3 (1.6%)	–	–
Pack years				
Median (range)	36.75 (0–100)	16.98 (0–51.6)	32.5 (0–83.0)	10.0 (0–58.3)
Alcohol status, N (%)				
Never drinker	26 (52.0%)	63 (34.4%)	–	–
Ex-drinker	5 (10.0%)	27 (14.8%)	46 (90.2%)	104 (89.7%)
Current drinker	19 (38.0%)	93 (50.8%)	5 (9.8%)	12 (10.3%)
Tumor stage, N (%)				
I	7 (14.0%)	–	3 (5.9%)	–
II	7 (14.0%)	–	5 (9.8%)	–
III	28 (56.0%)	–	22 (43.1%)	–
IV	7 (14.0%)	–	19 (37.3%)	–
Unknown	1 (2.0%)	–	2 (3.9%)	–

Table 3:

	database1 as training set				database2 as training set			
	AUC(95%CI) no score	P-value	AUC(95%CI)-with score	P-value	AUC(95%CI) no score	P-value	AUC(95%CI) with score	P-value
N	167		167		233		233	
TP53	0.70 (0.62 - 0.78)	Referent	0.73 (0.65 - 0.81)	Referent	0.77 (0.70-0.84)	Referent	0.79 (0.72 - 0.88)	Referent
RB1	0.64 (0.56 - 0.72)		0.65 (0.57 - 0.73)	0.08	0.70 (0.63-0.77)	0.16	0.72 (0.64 - 0.80)	0.12
β1(TP53) +β2(RB1)+α	0.69 (0.62 - 0.77)	0.83	0.76 (0.67 - 0.84)	0.31	0.79 (0.61 - 0.75)	0.51	0.85 (0.78 - 0.91)	**0.01**

*set area for 1-specif city at 0.3, adjusted for co-variables (number of mutations: 0, 1, >1; smoking status: never, Ex, current)

3.2.4 Variant filtering for germline data

As a second variant filtering methodology, we were interested in the filtering of false calls from germline variant calling with needlestack on normal samples. In this part, we benefit from the availability of validated data as a gold standard set to use a machine learning model as a variant filtering method. Indeed, as explained in the introduction, variant filtering methods based on machine learning model are extremely robust and powerful but they need labelled data to be trained, and such data were available only for our project on germline variant calling (see annexes for details on this project).

The main idea of the methodology developed in this part is to use machine learning to predict false positive or true positive status of called mutations, using known status of mutations called from the same conditions (*i.e.* same variant caller needlestack, same sequencing technology, same coverage). As described in the introduction, the methodology is divided into the following steps:

- Definition of a set of known entities **E**: these would be the known mutations, with a known status st defined as false positive or true positive.
- Definition of a set of statistical features **F**: these would be features from both variant caller and sequencing machine. The difficult task is to find features with importance in the mutation status, *i.e.* features that can separate true and false called mutations.
- For each known entities $e \in \mathbf{E}$ and each feature $f \in \mathbf{F}$, computation of f_e, the value of the feature for the entity
- **Training** of a machine-learning model (*e.g* a random forest)
- For each unknown entities $e' \in$ **E'** and each feature $f \in \mathbf{F}$, computation of $f_{e'}$
- **Application** of the trained model on each e': this steps applies the trained random forest algorithm on the unknown mutation data frame. Once the model is trained and applied, it is possible to evalutate its accuracy based on known data using a k-fold cross validation. This consists in training the model on $(1-k)$% of the data, applying it on the k% remaining data, and finally repeating this process k times to compute the predicted status of each entity one time. Because the true status is known, this method enables the estimation of model accuracy.

Application

The development of this machine learning approach for germline variant filtering was integrated into a genetic susceptibility project based on rare variants. The global aim of this project is to identify new susceptibility genes from a list of candidate genes. This project was motivated by the potential of using genetic susceptibility information in the context of early cancer detection. A set of 86 potential susceptibility genes was available, and we sequenced these genes in two series of cases and controls in order to perform an association analysis using a burden test.

Available data

Two types of data were used: known data E to train the random forest model and target data E′ on which the trained random forest model would be applied in the aim at predicting their false positive or true positive status.

In this second part of this project, a first serie of 432 cases and 432 controls were sequenced on IonTorrent Proton sequencer on the 86 candidate genes. Among these samples, 55 (set of samples S) were independently sequenced in WES on an independent sequencing machine (Illumina HiSeq). Let I_s refers to the set of Illumina WES mutations of the sample $s \in S$. From these 55 samples, a total of N = 11,234 mutations were called using relaxed filters from the sequencing of the 83 genes on IonTorrent Proton, which corresponds to 204 mutations per sample in average (given expectations, this set should contains a lot of false positives). These 11,234 mutations forms the set E of known mutations. To decide the status of these mutations (true positive TP, false positive FP or non available status NA), we used the following rules:

- $\forall e_i \in E$, if $e_i \in I_s$ then the status of the mutation $st_{e_i} = TP$
- else, if $AO(e_{i,I_s}) \geqslant AO_{thr}$ with $AO_{thr} = 5$ then $st_{e_i} = FP$
- else, $st_{e_i} = NA$

The false positive status is attributed to a mutation found in the IonTorrent Proton sequencing at a position sufficiently covered in the Illumina sequencing to consider that if the

mutation is truly present in the sample, it would have been detected. For this we computed the expected minimum number of alternative reads corresponding to the mutation if present in the Illumina sequencing as the following:

$$\mathrm{AO}(e_{i,\mathrm{I}_s}) = q_{0.01}[\mathrm{NB}(\mu = 0.5 * \mathrm{DP}(e_{i,\mathrm{I}_s}), \sigma = 0.1)]$$

which corresponds to the 99^{th} percentile of a negative binomial distribution assuming an expected VAF at 50%. If the expected minimum number of alternative reads in the Illumina sequencing is higher than 5, we decided that the mutations found in the IonTorrent Proton sequencing is a false positive. Finally, the set of known mutation E contains a total of respectively 594 and 49 truly called SNV and indels, and respectively 8078 and 2513 falsely called SNV and indels.

In this first serie, each mutation found in a sample not sequenced in the independent Illumina WES is considered as an unknown mutation $e'_i \in \mathrm{E}'$.

In addition, to this first serie, a second serie of 576 cases and 579 controls has been sequenced in the IonTorrent Proton machine on the same genes. Nevertheless, average coverage in this second serie was different than average coverage in the first serie containing the gold standard known mutations. In order to maintain the random forest accuracy (coverages are variable between training data set and target data set), we downsampled the 55 replicated samples from the first serie to obtain a range of coverages similar to the second serie. We computed the ratio of median coverages between these 55 samples and the samples sequenced in the second serie. This ratio was estimated as 0.6, we then downsampled the first serie 55 samples at a rate of 60%.

Random forest algorithm

A random forest algorithm was used to build a machine learning model on the true and false positive status of called mutations. The random forest is a supervised learning algorithm, in the sense that it knows *a priori* the categories of the input entities, in our case false or true called mutations. By analogy to real life, a random forest is a mathematical object composed of a number T of trees. Each tree $t \in \mathrm{T}$ in the random forest is a decision tree, *i.e.* a math-

ematical object proposing a decision when it observes a set of data. The number of trees T is chosen *a priori*, and the random forest randomly subsets T times the data to create T decision trees. We propose to require N = 500 trees in the forest. We have used the R package *randomForest* [89] which subsets by default $\frac{|F|}{3}$ features to compute each tree. Due to the fact that the random forest algorithm is sensitive to unbalanced data and has a tendency to be biased for the class predominantly represented in the data used to train the model, we propose to re-equilibrate the classes when training the model, *i.e.* subset the major class to obtain equal size of false and true positives when computing the trees in the forest.

Once the model is built by training the algorithm and input dataset with known class for each entities (in our case class is the false or true positive status and entities are called mutations for which we have computed the status), the method is applied on unknown status data. In this step, each unknown entity would pass into each tree in the forest that give one particular decision *i.e.* which class the entities is more likely to belong. At the end, for each entity e' and class $c \in [\mathrm{TP}, \mathrm{FP}]$, the probability of belonging to the class is computed as the following:

$$\mathrm{P}_{\mathrm{TP}}(e') = \frac{|t_{\mathrm{TP}}(e')|}{\mathrm{T}}$$
$$\mathrm{P}_{\mathrm{FP}}(e') = \frac{|t_{\mathrm{FP}}(e')|}{\mathrm{T}}$$

with $|t_x(e')|$ the number of trees that decided to attribute the class x to the entity e'.

Figure 3.8 gives an overview of the architecture of the random forest algorithm, more precisely it shows how does the random forest choose the predicting class. However, by default the random forest returns directly the predicted class, but it is also possible to require the class probability for each class and then *a posteriori* choose the probability threshold to choose the class, which is by default 0.5.

Training and application of the model on our germline data

The first step to train such a model consists in the definition of the model features $f \in \mathbf{F}$. The following table describes the features used:

Table 3.1 – Details of features used to train the random forest model

STATISTIC	**DESCRIPTION**	**RATIONAL**
AF	Variant allelic fraction (AO/DP)	errors are not abundant
AO	Alternative Observation count	errors should not be replicated in a lot of reads
DP	sequencing Depth	low coverages can corresponds to individual sequencing issues
QVAL	Q-value, main statistic from needlestack	errors should harbor a low QVAL
FS	Fisher exact test statistic	errors can be in strand bias
RVSB	Relative Variant Strand Bias	errors can be in strand bias
MIN_DIST	minimum distance with a variant with a VAF 10 times higher	to detect alignment artefacts
QUAL	variant quality score (max QVAL)	low maximum QVAL can corresponds to errors
ERR	error rate estimated by needlestack	high error rate can generate significant QVAL for errors
medianDP	median coverage at the position	unexpected low or high median coverages can corresponds to global sequencing issues
maxRatioWin	maximum ratio of VAF in a window of 100bp	multiple low VAF can corresponds to sequencing issues
nbVarWin	number of variant in a window of 100bp	to detect alignment artefacts
IoD	index of model deviance (normalized variance)	to detect low goodness of fit of our model of errors
HpLength	size of the homopolymer region	high-length homopolymer can create errors
N_QVAL_INV_20_50	number of QVAL between 20 and 50	to detect low confidence alterations in term of sequencing (comparable to LCAP)

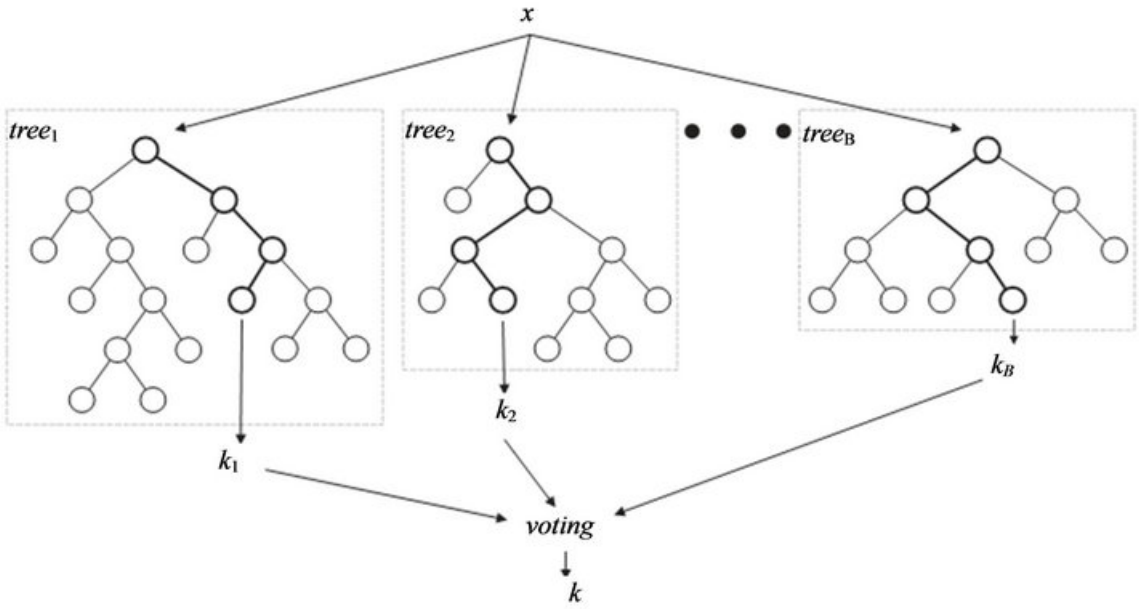


Figure 3.8 – General representation of the random forest algorithm from [**?**]. x is representing the unknown entity (denoted by e' in our methods) for which the class (or status) should be decided. k is representing the winner class, *i.e.* TP if $P_{TP}(e') > P_{FP}(e')$ and FP if $P_{FP}(e') > P_{TP}(e')$.

The first seven features are variant-level features whereas other ones are alteration-level features. To evaluate the distribution of these variables across known true and false positive called mutations (denoted respectively TP and FP), figure 3.9 is representing the values of the VAF, the RVSB and the Q-value for known mutations depending on the status. These plots illustrates the difficulty to determine hard-thresholds which would corresponds to lines on each axis of the plots to efficiently separate true and false calls. Most advanced methods such as machine learning approaches estimate possibly non-linear boundaries to correctly classify unknown entities.

As mentioned previously, we used the R package *randomForest* from [89] to train our model based on these features, requiring T = 500 trees in the forest and balancing the data by subsetting the major class (FP in our case). To test the ability of our model to class unknown status mutations, we used a k-fold cross validation strategy and then computed a Recall-Precision curve to estimate the global performance of the method. Indeed, using a k-fold method, each entity of the known data would have a predicted status, and this prediction can be then confronted with the truth to compute a Recall-Precision curve.

Figure 3.10 corresponds to the computed Recall-Precision curve based on a k-fold cross

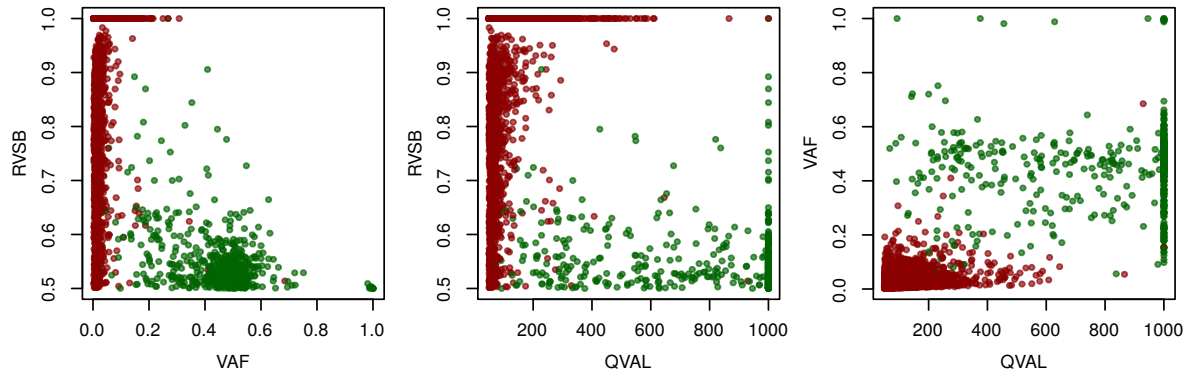


Figure 3.9 – Paired representation of false and true called mutations for three variant statistics: the VAF, the RVSB and the Q-value. Each dot corresponds to a mutation with a known status. Mutations are colored according to their status, false positives are shown in red and true positives are shown in green. Regular boundaries on this three variables can not separate the set of true mutations from the set of false ones.

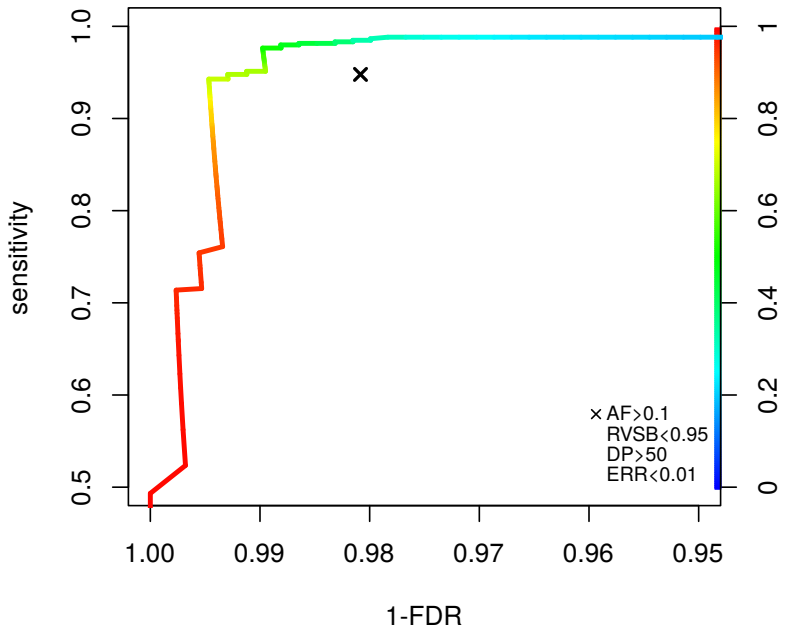


Figure 3.10 – Recall-Precision curve representing overall performance of the random forest algorithm developed for germline variant filtering. Recall-Precision curve is computed by varying the threshold on the probability to belonging to the class of TP (a called mutation is filter if its probability to be a TP is lower than this threshold). Closer to 1 the probability threshold, higher the number of filtered mutations, lower the sensitivity and lower the false discovery rate (FDR).

validation with $k = 10$ launched on the previously described known data. Performance results are shown for SNV due to reduced size of data in case of indels. In this study, SNV and indels were treated independently to train and apply the models. The cross in the plot corresponds to the measure of sensitivity and false discovery rate if hard-thresholds on VAF, coverage, RVSB and error rate were used (thresholds are shown in the figure) instead of machine learning. We were particularly interested in estimating if the random forest performs better than the hard-threshold that we used to apply. With a fixed false discovery rate at 2%, the gain of sensitivity of the random forest filtering is around 4% (adding around 20 true variants), and with a fixed sensitivity at 95%, the reduction of false discovery rate is around 1.3% (removing of around 100 false variants).

We had also the opportunity to validate our results on 28 samples that were sequenced twice. Using a threshold of probability at 0.5, *i.e.* for each tested mutations e', if $P_{FP}(e') > 0.5$ then e' would be considered as a false mutation otherwise it would be considered as a true mutation. Following this classification, we have computed the concordance of rare germline mutations (with a population frequency lower than 10%, which lead to a total of 367 mutations) which was estimated as **98%**.

3.3 Discussion

In this section we proposed two methodologies to efficiently filter variants as a subsequent step following variant calling with needlestack, our variant caller presented in the chapter 2. The first method is based on hard filtering on variant statistics for deep targeted sequencing data. The second method is based on a random forest algorithm for germline targeted sequencing data. Both methods have been developed based on data sequenced on a IonTorrent Proton sequencing machine. Nevertheless, the statistical variables used in the computation are not dependent on the sequencing technology, and therefore our methods can be easily used for other types of data.

In a first part we introduced a methodology to efficiently filter potentially false variants using pre-defined variant statistics. When computing the overlap of per-filter removed mutations (figure 3.5), we observed that our filtering based on our LCAP statistic and the tech-

nical replicates filter were extremely correlated. We were therefore interested in estimating the necessity of the duplicated libraries requirement, which is the only filter that increases a lot the cost of the experiment (it multiplies this cost by two). Because the LCAP statistic is based on the observed number of replicated mutations, the technical replication of samples is needed to compute this statistic. Nevertheless, if the LCAP statistic is consistent across multiple sequencing runs to identify low confidence alterations, using an LCAP "catalogue" built with estimations of the statistic would be a possible solution to do not require a duplicate sequencing. To estimate if such a catalogue can be built, we computed the percentage of LCAP values that are higher than 0.05 (*i.e.* proportion of low confidence alterations) in different proportions of runs p (Figure 3.11).

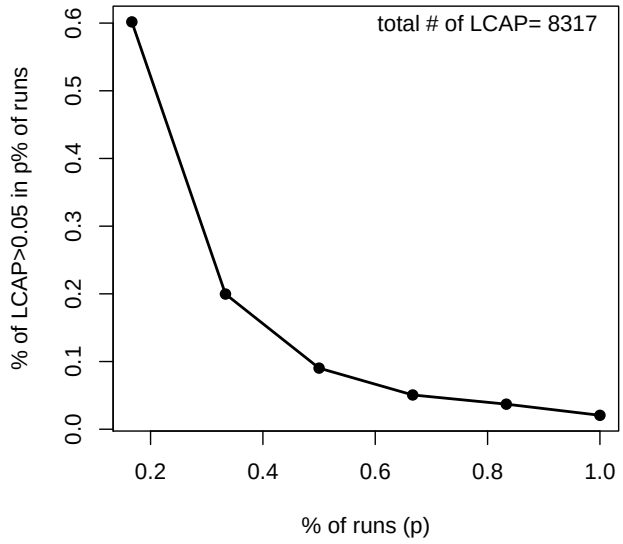


Figure 3.11 – Proportion of alterations (total is equal to 8317) that are identified as low confident (y-axis) in p% of sequencing runs. Computations were realized using six independent sequencing runs of *TP53* and *RB1* genes. As an example, the left-most dot indicates that 60% of alterations with an LCAP value lower than 0.05% only in one of the six runs (corresponding to 17% of the runs). This plot shows no evidence of LCAP consistency across multiple runs.

According to the Figure 3.11, only 10% of the alterations are consistently predicted as non-confident (*i.e* with a LCAP higher than 0.05) in at least half of sequencing runs and 2% are non-confident for all the runs. These results means the the LCAP statistic is not consistent across multiple runs. Therefore, a catalogue of LCAP value would not be informative. The variations of the LCAP can be due notably to the different numbers of sequenced sam-

ples between two runs, that can change the precision of the estimation of the statistic. In addition, the presence of a true mutation, detected in the two technical replicates, can also modify the LCAP statistic.

In this study we proposed the MIN_DIST variant statistic to identify low abundance alignment artefact created by a true variant. Recently, multiple re-alignment algorithms based on local assembly of small regions have been proposed to deal with alignment artefacts (see introduction). A possible perspective to estimate both the need and accuracy of our MIN_DIST statistic would be to test the accordance of removed mutations between filtering using this statistic and using re-assembly methods that commonly present the drawback of a long computation time.

A variant filtering approach that can reduce significantly the amount of false calls can have multiple applications, in particular it can be applied on ctDNA data in order to developed cancer biomarkers. Indeed, in such a case, the detection of very low VAF mutations is crucial, because tumor-derived mutation in body fluids such as plasma or urine are expected to be found in very low abundance. We have then applied our method on ctDNA data (the development of the biomarker in this study is presented in the next chapter). To test the accuracy of our variant filtering method on this data, we have performed permutation tests. For this, we have estimated the expected number of false mutations due to random attribution of errors in two technical replicates as 9. These false mutations are expected to be randomly distributed across cases and controls, and this would lead to 2 false mutations in cases and 7 false mutations in controls, corresponding to a false discovery rate of 3.8%. Nevertheless, a possibility to totally erase these potential false discoveries would be to increase the number of replicates for each sequenced sample and then adapt our simulations to compute the corresponding expected false discovery rate (see [34] where they sequenced 7 replicates).

We have also presented a smart variant filtering method based on machine learning algorithms. Because machine learning models require feature similarities between training and target datasets to perform correctly, trained models are not necessarily accurate for all types of data and it should be necessary to re-train a model for data presenting features dif-

ferently distributed than the ones used to train the model. Current classification algorithms are based on supervised learning, *i.e.* the models are trained on data with known labels. This requires the availability of validated data as a gold standard. As an example, in our study, we benefit from the concordance of datasets from two independent sequencing technologies to build a gold standard set of germline mutations used to trained our random forest algorithm. But gold standard data are not available for all types of mutations. Currently, the most widely used "truth" datasets are the Genome In A Bottle [149] and Platinium Genome [44] datasets which record genome variations of the human sample HG001 (NA12878). These consortia provide both a set of high-quality variants (used to assess sensitivity) and a set of confident regions to supply in addition position that does not present any variation (used to assess specificity). More recently, the Broad Institute of Harvard have provided a new gold standard dataset to deal with non "easy" variations contrary to the previously exposed datasets, called *SynDip* for "synthetic diploid" dataset [87]. Somatic truth data sets are quite less common. At the moment, majority of available truth somatic data sets come from somatic mutation detection competitions such as the PrecisionFDA or the ICGC-DREAM challenge. Nevertheless, these gold standard data sets are generated *in-silico* and are less accurate than the current germline sets. The germline and somatic gold standard datasets presented previously can therefore be used to train our machine learning model in case of other types of data *e.g.* data from other sequencing technology. Nevertheless, our method presents two major limitations. Firstly, it is important to remember that machine learning models require a sufficiently high size of data to be efficiently trained, and therefore using only a few samples to train the model should be suitable only for high number of sequenced positions. Secondly, because needlestack is a multi-sample variant caller, it can not analyze a sample alone. The only case where training of our model on one of these gold standard samples would be efficient is when data that need to be filtered are sequenced and analyzed in the same way than the gold standard sample used to train the model.

Finally, we did not tested the effect of different alignment tool on the accuracy of our variant filtering methodologies. Indeed, it is expected that using a different alignment algorithm on the same data can impact the mutation detection both through different vari-

ant calling and variant filtering results, notably when using a realigner based on local reassembly, due to the divergent nature of the algorithm compared to traditional aligners. As a perspective, it could be interesting to test the impact of the alignment on the variant filtering results to test the accuracy of the filters in the presence of other aligners.

Chapter 4

Applications to circulating-tumor DNA data

Contents

4.1 Scientific context 116

4.2 Scientific contribution A 117

4.2.1 Article C 117

4.3 Scientific contribution B 132

4.3.1 Article D 132

4.4 Scientific contribution C 157

4.4.1 Article E 157

4.5 Scientific contribution D 165

4.5.1 Article B 167

4.6 Discussion 167

4.1 Scientific context

DNA from tumor cells accounts for a small fraction of cfDNA, the DNA found in body liquids, such as in blood samples or urine samples, as a result of cell death and cell secretion [37]. ctDNA is currently emerging as a potential non-invasive biomarker of the tumor. It can be used in multiple areas such as cancer surveillance and also response to therapies. Recently, it has also been reported that ctDNA can be used as an cancer detection biomarker in order to reduce the mortality associated with cancer (see introduction chapter 1). More importantly, ctDNA can be used for early cancer detection, *i.e.*, for detecting cancer in early stages, before the apparition of symptoms that can be clinically identified by CT-scan, and that the proportion of tumor-derived DNA over the total amount of cfDNA is correlated with the stage of the tumor [19], [3]. This also means that using ctDNA as a cancer detection biomarker and potentially as an early cancer detection biomarker would requires a high sensitivity to detect mutations in low abundance. In addition, as it is required when developing a biomarker, and also because cancer does not have a high prevalence in the general population, it is important to obtain a good sensitivity but also a good specificity when identifying such mutations. In the case of early detection biomarker development, while sensitivity can be increased with a variant caller that can detect very low VAF such as needlestack, a high specificity can be achieved by selecting DNA positions or genes that are expected to be highly mutated in cancer cases and not in non-cancer individuals. In this chapter, we present four distinct applications of our needlestack algorithm to detect mutations in plasma cfDNA samples of cancer cases at multiple stages in order to estimate the possibility to use the ctDNA as a cancer biomarker, and even potentially as an early cancer biomarker. The first study estimates the accuracy of using *KRAS* mutations from blood samples to identify pancreatic cancer cases (particularly codons 12, 13 and 61 that are highly mutated in this cancer type). The second study described the UroMuTERT assay based on the detection of *TERT* promoter mutations in order to detect urothelial cancer patients from blood and urine samples. The two last studies describe the usage of both (i) *TP53* mutations, (ii) and the combination of *TP53* and *RB1* mutations in order to detect SCLC patients.

4.2 Scientific contribution A

In this study we were interested in the detection of *KRAS* mutations in plasma circulating cfDNA of pancreatic cancer patient, using a case-control cohort. Pancreatic cancer harbour one the poorest 5-year survival of all types of cancer (around 6%) in Europe, and, in addition, around 80% of patients die within a year following diagnostic [116]. It is then critical to detect earlier these types of cancer, and a promising non-invasive biomarker of the tumor is the ctDNA. A good candidate as a tumor-footprint potentially reachable in the cfDNA of pancreatic cancer samples would be the *KRAS* mutations, as the *KRAS* gene is mutated in the majority of pancreatic ductal adenocarcinomas (accounting for 90% of the pancreatic cancers) [139]. In addition, *KRAS* is known to present the earliest genetic alterations that drive pancreatic cancer [70].

We applied a *KRAS* amplicon-based deep sequencing approach (IonTorrent Proton sequencing technology) followed by our needlestack variant calling in order to detect cfDNA mutations in plasma samples of 437 pancreatic cancer cases, 141 chronic pancreatitis subjects, and 394 healthy controls. We found mutations in around 4% of non-pancreatic cancer individuals (healthy individuals or subjects with chronic pancreatitis), and in 21.1% of cases. 89.1% of these positive cases carried at least one mutation at codons 12, 13 or 61. Indeed, as previsously reported, these codons are expected to be highly mutated in the tumors of pancreatic cancer cases [138]. Reported VAF of case ctDNA mutations ranging from 0.08% to 79%, highlighting the fact that needlestack can detect very low VAF. Finally, we detected ctDNA mutations in 34% of advanced stages and in 10% of early stages, suggesting that (i) the limitation in sensitivity can be partially attributable to the biology of the tumor; (ii) and it is possible to find tumor footprints in the cfDNA of early stage pancreatic cancer patients, that can support the usage of ctDNA as an early cancer biomarker. Nevertheless, this would require an increased specificity.

4.2.1 Article C

Research Paper

KRAS mutations in blood circulating cell-free DNA: a pancreatic cancer case-control study

Florence Le Calvez-Kelm[1], Matthieu Foll[1], Magdalena B. Wozniak[1], Tiffany M. Delhomme[1], Geoffroy Durand[1], Priscilia Chopard[1], Maroulio Pertesi[1], Eleonora Fabianova[2], Zora Adamcakova[2], Ivana Holcatova[3], Lenka Foretova[4], Vladimir Janout[5,6], Maxime P. Vallee[1], Sabina Rinaldi[1], Paul Brennan[1], James D. McKay[1], Graham B. Byrnes[1], Ghislaine Scelo[1]

[1]International Agency for Research on Cancer (IARC), Lyon, France

[2]Regional Authority of Public Health, Banska Bystrica, Slovakia

[3]Charles University of Prague, First Faculty of Medicine, Institute of Hygiene and Epidemiology, Prague, Czech Republic

[4]Masaryk Memorial Cancer Institute and Medical Faculty of Masaryk University, Brno, Czech Republic

[5]Department of Preventive Medicine, Faculty of Medicine, Palacky University, Olomouc, Czech Republic

[6]Faculty of Medicine, University of Ostrava, Czech Republic

Correspondence to: *Florence Le Calvez-Kelm,* ***email:*** *lecalvez@iarc.fr*

Keywords: *cell-free DNA, KRAS mutations, plasma, pancreatic cancer detection*

Received: *June 29, 2016* ***Accepted:*** *September 19, 2016* ***Published:*** *October 01, 2016*

ABSTRACT

The utility of *KRAS* mutations in plasma circulating cell-free DNA (cfDNA) samples as non-invasive biomarkers for the detection of pancreatic cancer has never been evaluated in a large case-control series. We applied a *KRAS* amplicon-based deep sequencing strategy combined with analytical pipeline specifically designed for the detection of low-abundance mutations to screen plasma samples of 437 pancreatic cancer cases, 141 chronic pancreatitis subjects, and 394 healthy controls. We detected mutations in 21.1% (N=92) of cases, of whom 82 (89.1%) carried at least one mutation at hotspot codons 12, 13 or 61, with mutant allelic fractions from 0.08% to 79%. Advanced stages were associated with an increased proportion of detection, with *KRAS* cfDNA mutations detected in 10.3%, 17,5% and 33.3% of cases with local, regional and systemic stages, respectively. We also detected *KRAS* cfDNA mutations in 3.7% (N=14) of healthy controls and in 4.3% (N=6) of subjects with chronic pancreatitis, but at significantly lower allelic fractions than in cases. Combining cfDNA *KRAS* mutations and CA19-9 plasma levels on a limited set of case-control samples did not improve the overall performance of the biomarkers as compared to CA19-9 alone. Whether the limited sensitivity and specificity observed in our series of *KRAS* mutations in plasma cfDNA as biomarkers for pancreatic cancer detection are attributable to methodological limitations or to the biology of cfDNA should be further assessed in large case-control series.

INTRODUCTION

The latest estimates show that more than 330,000 cases of pancreatic cancer are diagnosed yearly worldwide, and approximately the same number of deaths are attributed to the disease (GLOBOCAN 2012 website: http://globocan.iarc.fr/, accessed on 9 Feb 2015). Disease survival is among the poorest of all cancers with 5-year survival at only 6 % in Europe and ~79 % of patients dying within a year following diagnosis [1, 2]. Improved survival is observed in patients that undergo surgical resection, but this therapeutic option is limited to cases with localized tumors [3]. Early detection has therefore the potential to reduce the mortality associated with pancreatic cancer. Endoscopic ultrasound has shown good sensibility and specificity to detect precancerous and cancerous lesions but this invasive technique has limited use for early detection in asymptomatic individuals [4]. Blood level of

the antigen CA 19-9 is the only validated tumor marker for pancreatic cancer with overall sensitivity of 79% (70-90%) and specificity of 82% (68%-91%) [5, 6]. However, non-specific expression in other benign or malignant diseases and absence of expression in Lewis (a-b-) blood phenotypes (~10-15% of the population) limit the use of this biomarker as a diagnostic test [7].

Cell-free DNA fragments (cfDNA) are released into the bloodstream and other body fluids as part of natural cell apoptosis, necrosis and active secretion. Gene mutations in cfDNA fragments have been found to be tumor-specific leading to the concept of circulating tumor DNA (ctDNA) and their potential utility as highly specific non-invasive biomarkers has raised in the recent years [8]. Pancreatic ductal adenocarcinoma (PDAC) accounts for more than 90% of all pancreatic cancer cases [9] and activating hotspot mutations in the *KRAS* gene are present in the majority of them, representing the most frequent [10] but also the earliest genetic alteration that drives pancreatic neoplasia [11–13]. Of the 596 PDAC cases sequenced within the International Cancer Genome Consortium (ICGC) project (https://icgc.org/, as of 23 Feb 2016), 534 (90%) harbored at least one *KRAS* mutation: 83%, 5.5% and 1.5% at codons 12, 61 and 13, respectively. *KRAS* mutations (often restricted to codon 12) have previously been detected in blood (plasma or serum) samples from patients with pancreatic cancer [14–26], showing large variations in the proportion of detected cases (27% to 93%) probably because of inter-laboratory variability, limited sample sizes, and variable sensitivities of the assays. Ultra-deep sequencing technologies allows the identification of low-abundance somatic variants and were shown to be applicable to ctDNA [26–31], but has so far been applied to sample series of limited size and lacking control groups. Here, we investigated whether deep sequencing of *KRAS* codons 12, 13 and 61 in cfDNA from plasma samples from a large series of more than 400 pancreatic cancer cases and 500 controls could represent a comprehensive assay for sensitive and specific detection of pancreatic cancer.

RESULTS

Subject characteristics, sequencing performance and inclusion criteria for analysis

Samples were included when cfDNA total yield was at least 4ng and when sequencing reads were above 1000 on average for all codons. In total, 96 samples (100%) from a pilot set and 903 samples (93.4%; 397 pancreatic cancer cases (94.2%); 132 chronic pancreatitis (91.0%) and 374 controls (93.3%)) from a validation set met the inclusion criteria. Table 1 provides the characteristics of cases and controls, as well as the average of cfDNA yields by status. Analysis of variance was used to compare (log-transformed) cfDNA concentrations by subject characteristics listed in table 1 and showed significant difference by status (with higher yields in pancreatic cancer cases versus controls; t-test $p<0.0001$), stage (higher yields in missing stages versus reported stages: $p<0.0001$), and center (higher yields in Prague and Olomouc when compared to other centers: $p<0.0001$). Other variables had no significant influence on cfDNA yield (Fisher test $p>0.05$).

The average mean depth of reads after filtering on mapping quality were, for the pilot and validation sets, respectively: 3992 (SD= 1123) and 2888 (SD=1259) at *KRAS* codon 12 c.34, and 2492 (SD=710) and 3765 (SD=1762) at codon 61 c.181.

Determination of the allelic fraction threshold for the detection of the *KRAS* p.G12V variant

The number of reads obtained from sequencing of 2ng of two independent serial dilutions (in duplicates) of *KRAS* c.35G>T; p.G12V mutated DNA was between 991 and 4205 with an average read depth of 2693 (Supplementary Table S1, Supplementary Data). There was a good correlation between expected and observed mutant allelic fractions ($r2=0.948$; Supplementary Figure S1). Needlestack analysis was performed independently on the 2 sets of data (Figure 1). Phred scale q-values (QVAL) determined by the Negative binomial regression show that the *KRAS* p.G12V mutation could be reliably detected down to a minor allele frequency of 0.2% when read depth was approximately of 2500 reads QVAL>30 for 3 of the replicates at 0.2%) (Supplementary Table S1, Supplementary Data).

Performance of *KRAS* mutations in cfDNA samples in complement to CA19-9 plasma levels as non-invasive pancreatic cancer biomarker

Applying a threshold of QVAL >30 to the sample set of the pilot series, *KRAS* mutations at hotspot codons reported in PDAC were identified in cfDNA plasma samples in 7 of 40 cases (sensitivity 17.5%) with PDAC and in 1 cfDNA of 27 patients with pancreatic benign neoplasms. None were detected in healthy controls, or in patients with chronic pancreatitis (overall specificity of 98.2%; of 100% against healthy controls) (Tables 2 and 3). All *KRAS* mutations were located at codon 12 (See Supplementary Data Supplementary Table S2 for the complete list of samples harboring cf DNA *KRAS* mutations). Investigating the presence of *KRAS* mutations at other screened codons (from *KRAS* codons 4 to 16 and from codons 51 to 69) and reported mutated for any cancer sites in the COSMIC database identified (i) 2 additional PDAC cases with cfDNA *KRAS* mutations (1 case with p.K5R and 1 case with p.K5R and p.G10R; leading to an overall sensitivity of 22.5%) and (ii) 1 additional mutation in a patient with benign neoplasm of the pancreas (p.A11P). All mutations except one had allelic fraction below 3% (Supplementary Table S2, Supplementary Data).

Table 1: Description of the study population

Characteristics	Pilot series (N=96)								Validation series (N=903)					
	Pancreatic cancer cases		Healthy controls		Chronic pancreatitis		Pancreatic benign neoplasms		Pancreatic cancer cases		Healthy controls		Chronic pancreatitis	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Total	**40**		**20**		**9**		**27**		**397**		**374**		**132**	
Sex														
Male	22	*55.0*	11	*55.0*	4	*44.4*	0	*0.0*	220	*55.4*	217	*58.0*	92	*69.7*
Female	18	*45.0*	9	*45.0*	5	*55.6*	0	*0.0*	177	*44.6*	157	*42.0*	40	*30.3*
Missing	0	*0.0*	0	*0.0*	0	*0.0*	27	*100.0*	0	*0.0*	0	*0.0*	0	*0.0*
Age at blood draw (mean, sd)	64.8 (10.6)		66.2 (8.7)		62.8 (8.2)		Missing		62.2 (10.2)		60.6 (11.9)		55.6 (12.9)	
BMI at blood draw (mean, sd)	24.7 (3.7)		27.4 (4.0)		23.2 (3.8)		Missing		25.1 (4.5)		28.2 (4.3)		24.4 (4.2)	
Recruiting country														
Czech Republic	40	*100.0*	20	*100.0*	9	*100.0*	27	*100.0*	298	*75.1*	248	*66.3*	47	*35.6*
Slovakia	0	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*	99	*24.9*	126	*33.7*	85	*64.4*
Tobacco smoking														
Never	20	*50.0*	9	*45.0*	3	*33.3*	0	*0.0*	167	*42.1*	175	*46.8*	45	*34.1*
Ex-smoker	10	*25.0*	6	*30.0*	4	*44.4*	0	*0.0*	123	*31.0*	113	*30.2*	24	*18.2*
Current smoker	10	*25.0*	5	*25.0*	2	*22.2*	0	*0.0*	107	*27.0*	86	*23.0*	63	*47.7*
Missing	0	*0.0*	0	*0.0*	0	*0.0*	27	*100.0*	0	*0.0*	0	*0.0*	0	*0.0*
Alcohol drinking														
Never	25	*62.5*	12	*60.0*	4	*44.4*	0	*0.0*	212	*53.4*	176	*47.1*	49	*37.1*
Ex-drinker	6	*15.0*	3	*15.0*	4	*44.4*	0	*0.0*	95	*23.9*	36	*9.6*	48	*36.4*
Current drinker	9	*22.5*	5	*25.0*	1	*11.1*	0	*0.0*	87	*21.9*	162	*43.3*	35	*26.5*
Missing	0	*0.0*	0	*0.0*	0	*0.0*	27	*100.0*	3	*0.8*	0	*0.0*	0	*0.0*
Tumor stage at diagnosis														
Local	6	*15.0*	-		-		-		33	*8.3*	-		-	
Regional	17	*42.5*	-		-		-		126	*31.7*	-		-	
Systemic	16	*40.0*	-		-		-		119	*30.0*	-		-	
Unknown	1	*2.5*	-		-		-		119	*30.0*	-		-	
Tumor histological type														
Ductal adenocarcinoma	40	*100.0*	-		-		-		243	*61.2*	-		-	
Other ductal carcinoma	0	*0.0*	-		-		-		19	*4.8*	-		-	
Endocrine	0	*0.0*	-		-		-		14	*3.5*	-		-	
Missing/Unknown	0	*0.0*	-		-		-		121	*30.5*	-		-	
Log10 cfDNA concentration, ng/mL plasma (mean, sd)	1.7 (0.5)		1.7 (0.5)		1.8 (0.3)		1.9 (0.7)		2.0 (0.7)		1.7 (0.6)		1.8 (0.7)	

The sensitivity and the overall specificity of plasma CA19-9 levels for detecting PDAC was 90.0% and 64.8% respectively (Table 3). Combining these so that the test was declared positive if a *KRAS* mutation was found at any COSMIC reported position or if the CA19-9 plasma level was positive enabled the detection of 2 additional PDAC cases (38/40) that were negative for CA19-9 plasma level but positive for cfDNA *KRAS* mutation, increasing the sensitivity to 95% (Tables 2 and 3). Comparisons of AUCs of the combined assays versus CA19-9 levels alone showed small increases, approximately 0.02 for each of the three comparisons (cancer cases vs. healthy controls; cancer cases vs. all other conditions; cancer cases vs. benign pancreatic conditions) and were non significant ($p>0.17$ for all comparisons).

Validation of the proportions of detectable cfDNA *KRAS* mutations in pancreatic cancer cases and controls

We extended the cfDNA *KRAS* mutation screening to the validation case-control series (N=903) (Table 1). Of the 397 patients with pancreatic cancer, 75 (18.9%) carried at least one cfDNA *KRAS* mutation at PDAC hotspot codons, a sensitivity close to that reported for the pilot series (17.5%). We also detected at least one *KRAS* mutations at PDAC hotspot codons in the plasma of 4/132 (3.0%) patients with chronic pancreatitis and of 9/374 (2.4%) healthy controls whereas none were detected in those subjects of the pilot series. Enlarging the search for *KRAS* mutations to other screened codons increased the sensitivity to 20.9% (83 patients with pancreatic cancer carrying at least one mutations in their cfDNA), but decreased the specificity with the detection of cfDNA *KRAS* mutations in 6/132 (4.5%) patients with chronic pancreatitis and in 14/374 (3.7%) healthy controls (Table 4).

Of note, we identified 3 subjects (2 cases and 1 control) with the silent base substitution c.24A>G p.V8V (at 46.38%, 11.46% and 46.98% allelic fractions respectively) which we considered as a rare SNP (rs147406419) as it was reported with an allelic frequency between 0.02% (Exome Variant Server ESP6500siv2) and 0.04% (Exome Aggregation Concortium ExAC) and classified as probably non-pathogenic impact by CLINSIG. This variant was ignored for the rest of the

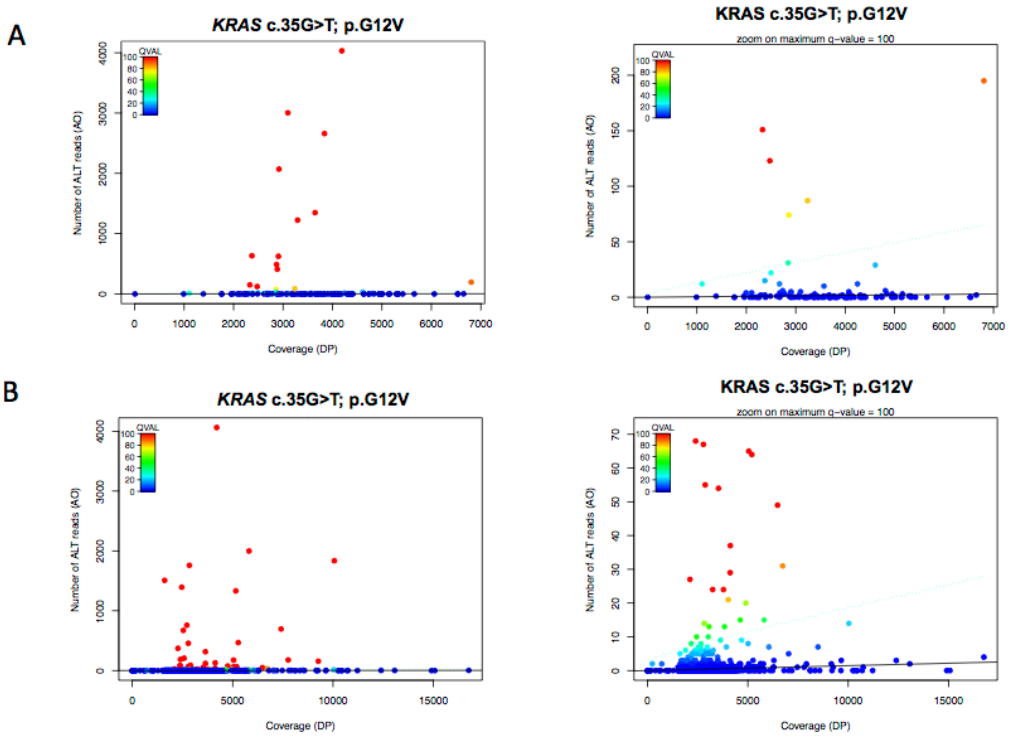


Figure 1: Mutation detection of *KRAS* c.35G>T; p.G12V in serial dilution and cfDNA samples using the Needlestack approach. Negative-binomial regression plot at *KRAS* c.35G>T; p.G12V displaying the total number of reads (coverage, DP) and the number of reads matching the candidate variant (AO). Black solid line: Estimated error rate (e) at the c.35 position for this G>T base change. Blue dashed line: Detection limit at q-values $<10^{-3}$; >30 in Phred scale (QVAL). Dots above the blue dashed line: Outliers of the regression (QVAL≥30), declared as mutant *KRAS* samples (c.35G>T; p.G12V). Dots below the blue dashed line: Inliers (QVAL<30) declared unmutated at this position for specified base change. **A.** Serial dilution of SW480 cell-lines in duplicates (N=28) and cfDNA from the pilot series (N=96) sequenced on a Ion Torrent PGM 316 Chip ($e= 4.2\times10^{-4}$); **B.** Serial dilution of SW480 cell-lines in duplicates (N=28) and cfDNA from the validation series (N=903) sequenced on Ion Torrent PGM 318 chips ($e=1.4\times10^{-4}$).

Table 2: *KRAS* mutations and CA19-9 plasma levels in the pilot series (N=94)

	cfDNA *KRAS* mutation at hotspot codons (12, 13, 61) reported in PDAC		cfDNA *KRAS* mutation at any screened codons reported in any cancer sites	
	N	%	N	%
Plasma CA19-9 positive level (≥37Ku/l)				
PDAC case, N=36	5	*13.9*	7	*19.4*
Healthy controls, N=3	0	*0.0*	0	*0.0*
Benign pancreatic neoplasm, N=11	1	*9.1*	1	*9.1*
Chronic pancreatitis, N=5	0	*0.0*	0	*0.0*
Plasma CA19-9 negative level (<37Ku/l)				
PDAC case, N=4	2	*50.0*	2	*50.0*
Healthy controls, N=17	0	*0.0*	0	*0.0*
Benign pancreatic neoplasm, N=14	0	*0.0*	1	*7.1*
Chronic pancreatitis, N=4	0	*0.0*	0	*0.0*
Total				
PDAC case, N=40	7	*17.5*	9	*22.5*
Healthy controls, N=20	0	*0.0*	0	*0.0*
Benign pancreatic neoplasm, N=25*	1	*4.0*	2	*8.0*
Chronic pancreatitis, N=9	0	*0.0*	0	*0.0*

*Two benign neoplasms were excluded from this analysis because CA19-9 plasma level measurements could not be performed.

Table 3: Performance of NGS-based assay for the detection of cfDNA *KRAS* mutations, CA19-9 plasma level and combined assays (40 PDAC, 20 healthy controls, 9 chronic pancreatitis subjects, and 25 benign neoplasm subjects)

	Sensitivity	Overall Specificity*	Specificity against healthy controls
cfDNA *KRAS* mutation			
at PDAC hotspot codons (12, 13, 61)	17.5%	98.2%	100.0%
at any screened codons reported in any cancer sites	22.5%	96.4%	100.0%
CA19-9 plasma level (≥37Ku/l)	90.0%	64.8%	85.0%
Combined cfDNA *KRAS* mutation and CA19-9 plasma level			
at PDAC hotspot codons (12, 13, 61)	95.0%	64.8%	85.0%
at any screened codons reported in any cancer sites	95.0%	63.0%[a]	85.0%

*against non-PDAC and controls

[a]Decreased specificity due to the detection of c.31G>C; p.A11P *KRAS* mutation in a patient with benign neoplasm negative for the plasma CA19-9 assay

analysis. Further restricting the analysis to missense *KRAS* mutations decreased false positive rates to 3.8% (5/132) and 3.2% (12/374) respectively (Table 4). The complete list of *KRAS* mutations identified in cfDNA of the validation series and corresponding allelic fractions is available in supplementary data (Supplementary Data,

Table 4: Proportion of subjects with *KRAS* mutations in their plasma cfDNA

	Pancreatic cancer cases						Chronic pancreatitis						Healthy controls					
	All N=437		Pilot N=40		Validation N=397		All N=141		Pilot N=9		Validation N=132		All N=394		Pilot N=20		Validation N=374	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Subjects with *KRAS* mutations in cell-free DNA	**92**	***21.1***	**9**	**22.5**	**83**	***20.9***	**6**	***4.3***	**0**	***0.0***	**6**	***4.5***	**14**	***3.6***	**0**	***0.0***	**14**	***3.7***
Numbers of mutation																		
Single	89	*20.4*	8	20.0	81	*20.4*	4	*2.8*	0	*0.0*	4	*3.0*	11	*2.8*	0	*0.0*	11	*2.9*
Multiple	3	*0.7*	1	2.5	2	*0.5*	2	*1.4*	0	*0.0*	2	*1.5*	3	*0.8*	0	*0.0*	3	*0.8*
Location																		
Mutation(s) at PDAC hotpot codon(s) 12, 13 or 61	81	*18.5*	7	17.5	74	*18.6*	4	*2.8*	0	*0.0*	4	*3.0*	8	*2.0*	0	*0.0*	8	*2.1*
Mutation(s) at other codon(s)*	10	*2.3*	2	5.0	8	*2.0*	2	*1.4*	0	*0.0*	2	*1.5*	5	*1.3*	0	*0.0*	5	*1.3*
Mutations at hotpot codons 12, 13 or 61 and others*	1	*0.2*	0	0.0	1	*0.3*	0	*0.0*	0	*0.0*	0	*0.0*	1	*0.3*	0	*0.0*	1	*0.3*
Type																		
Missense	92	*21.1*	9	22.5	83	*20.9*	5	*3.5*	0	*0.0*	5	*3.8*	12	*3.0*	0	*0.0*	12	*3.2*
Silent	0	*0.0*	0	0.0	0	*0.0*	1	*0.7*	0	*0.0*	1	*0.8*	2	*0.5*	0	*0.0*	2	*0.5*

* Codons reported mutated in COSMIC (all cancer sites)
We identified 3 silent base substitutions c.24A>G p.V8V (2 in cases and one in controls), which we considered as a rare SNP (rs147406419) as it was reported with an allelic frequency between 0.0002 (Exome Variant Server ESP6500siv2) and 0.0004 (Exome Aggregation Concortium ExAC) and classified as probably non-pathogenic impact by CLINSIG (Table S3, Supplementary Data). The 3 base substitutions are consequently not included in this table.

Supplementary Table S3). The lowest allelic fraction detected in the cfDNA samples was 0.08% in a plasma case (sample CA93) at *KRAS* p.G13R (Supplementary data, Supplementary Figure S2).

As for somatic *KRAS* mutations reported in PDAC (COSMIC and ICGC data) and chronic pancreatitis (COSMIC data), the majority of cfDNA *KRAS* mutations identified in the combined pilot and validation series were located at codon 12 (76.3 % in pancreatic cancer cases; 77.8% in chronic pancreatitis and 47.4% in healthy controls; Figure 2A). Similar proportions of *KRAS* mutations at codons 61 and 13 were observed in cfDNA of pancreatic cancer cases (7.2% and 3.1% respectively) as compared to PDAC ICGC tumors (6.1% and 1.7% respectively). However, while less than 1% of *KRAS* mutations reported in ICGC/ COSMIC data are located at other codons, 13% (13/97), 22% (2/9), and 31% (6/19) of such mutations were detected in the plasma samples of cancer cases, chronic pancreatitis, and controls, respectively (Figure 2A). The frequencies of the most predominant mutation types reported for PDAC in ICGC, i.e p.G12D, p.G12V, p.G12R, p.G12C followed by p.Q61H, p.Q61R and p.Q61L paralleled the frequencies of the cfDNA *KRAS* mutations in cases (Figure 2B) reflecting the probable tumor origin of the cfDNA *KRAS* mutations. In addition, one cancer case and one control harbored p.Q61P and p.Q61E in their cfDNA, respectively, two non-PDAC COSMIC missense substitutions previously reported in various cancer tissues (Figure 2 and Supplementary Data Supplementary Table S4).

We did not observe striking differences by histological groups. Amongst the 283 PDAC cases, 59 (20.8%) were detected with a cfDNA *KRAS* mutation,

all but three (p.M67L, p.M72V, and p.Q61P) reported as predominant PDAC mutations. Four "other ductal carcinoma" cases out of 19 (21.1%) were also detected with a single cfDNA mutation, all four reported as hotspot PDAC mutations. Amongst the 16 endocrine cases, 3 mutations were detected in 3 cases (18.7%), two of them not reported as hotspot PDAC mutations (p.G60D and p.A59G). The two cases with multiple mutations were found in the pancreatic cancer cases of unknown histological type, where 29 mutations were detected in 27/121 (22.3%) cases. Of these 29 mutations, five (p.A59E, p.E62D, p.Q61R, p.Q70P, and p.Y64D) were not reported as predominant PDAC mutations.

Advanced stages were significantly associated with an increased proportion of detection (*KRAS* cfDNA mutations were detected in 10.3% of cases diagnosed with local stage, 17,5% with regional stage, and 33.3% with systemic stage; chi-squared p=0.0009) (Table 5). Among detected cases, there was a non-significant trend of increased allelic fractions with stage (log10 of fractions were 0.1270, 0.1349, and 0.3047) on average, for local, regional and systemic disease, respectively; linear regression t-test p=0.3278). Allelic fractions correlated significantly with status (Table 6), pancreatic cancer cases carrying cfDNA *KRAS* mutations at higher allelic fractions than patients with chronic pancreatitis (t-test on log10(allelic fractions) p=0.0259) and healthy controls (p=0.0008). Healthy controls and chronic pancreatitis subjects had similar allelic fractions (p=0.8218). Of note, 3 PDAC cases were found to carry *KRAS* mutations in their plasma samples at allelic fractions higher than 50% reflecting gain of mutant *KRAS* copies. Other factors associated with allelic fractions were: histological type (with "other ductal carcinoma" cases having higher allelic fractions than PDAC (p=0.0016), endocrine (p=0.0078), and unknown/missing types (p=0.0004); sex (males having higher allelic fractions than females in healthy controls, p=0.0069); and age (borderline trend showing higher allelic fractions in older controls, p=0.0548).

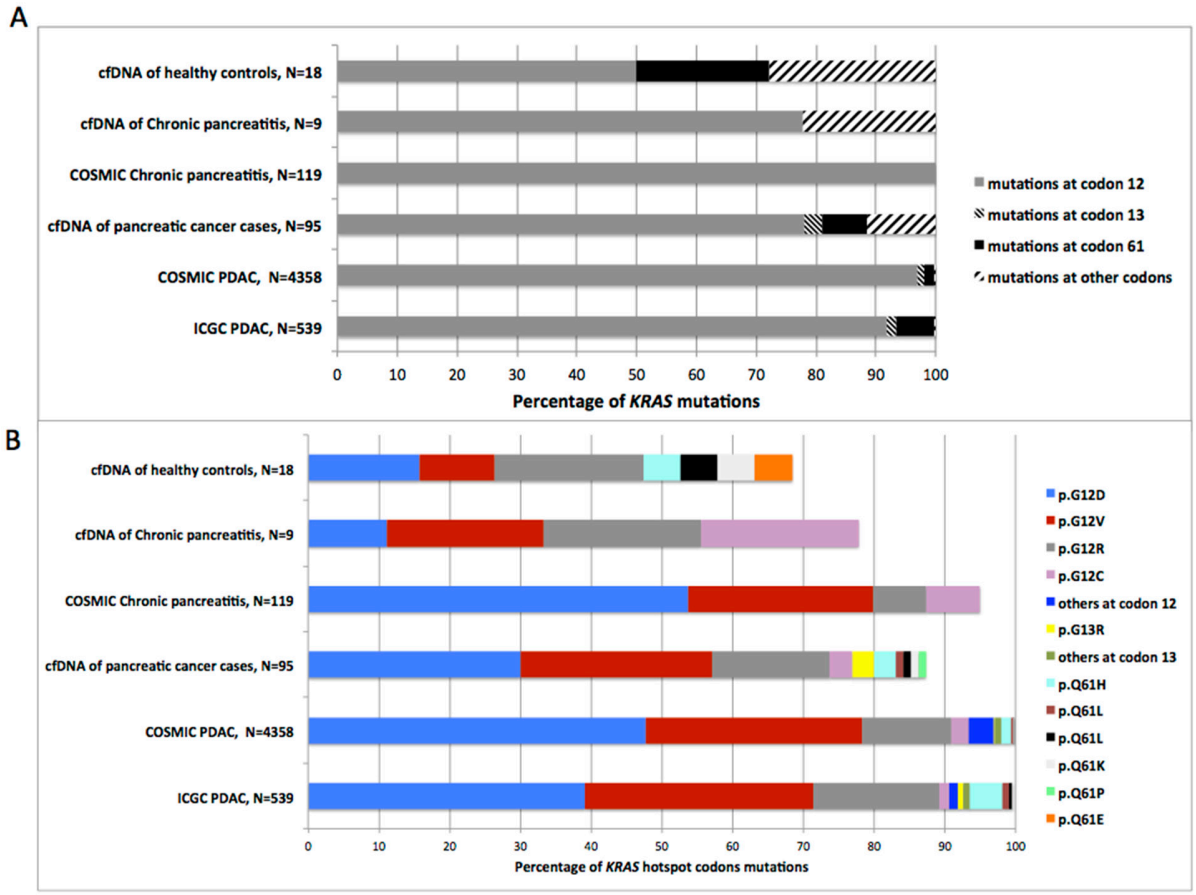


Figure 2: Distribution of *KRAS* mutations detected in plasma samples from pancreatic cases, chronic pancreatitis and healthy controls compared to somatic *KRAS* mutations reported in ICGC and COSMIC database. A. Comparison of cfDNA *KRAS* mutation location; **B.** Comparison of *KRAS* mutation spectrum at hotspot codons (12, 13 and 61). N= Number of *KRAS* mutations.

Table 5: Proportion of pancreatic cancer cases with *KRAS* mutations in their plasma cfDNA, by stage

Stage	Pilot series			Validation series			All		
	Total	cfDNA *KRAS* mutation		Total	cfDNA *KRAS* mutation		Total	cfDNA *KRAS* mutation	
	N	N	%	N	N	%	N	N	%
Local	6	1	*16.7*	33	3	9.1	39	4	*10.3*
Regional	17	1	*5.9*	126	24	19.0	143	25	*17.5*
Systemic	16	7	*43.8*	119	38	31.9	135	45	*33.3*
Unknown	1	0	*0.0*	119	18	15.1	120	18	*15.0*
All	**40**	**9**	***22.5***	**397**	**83**	**20.9**	**437**	**92**	***21.1***

KRAS Mutations identified at hotspot codons 12, 13 and 61 and at other codons reported mutated in COSMIC; the silent base substitution c.24A>G p.V8V was excluded from analysis.

Table 6: Proportion of subjects with cfDNA *KRAS* mutations at various allelic fractions

[a]AF (%)	Pancreatic cancer cases						Chronic pancreatitis						Healthy controls					
	All, N=93		Pilot, N=9		Validation, N=84		All, N=6		Pilot, N=0		Validation, N=6		All, N=14		Pilot, N=0		Validation, N=14	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
<0.2	4	*4.3*	0	*0.0*	4	*4.8*	1	*16.7*	0	*0.0*	1	*16.7*	4	*28.6*	0	*0.0*	4	*28.6*
[0.2-1]	35	*37.6*	3	*33.3*	32	*38.1*	4	*66.7*	0	*0.0*	4	*66.7*	7	*50.0*	0	*0.0*	7	*50.0*
[1.01-10]	40	*43.0*	5	*55.6*	35	*41.7*	0	*0.0*	0	*0.0*	0	*0.0*	2	*14.3*	0	*0.0*	2	*14.3*
[10.01-50]	11	*11.8*	1	*11.1*	10	*11.9*	1	*16.7*	0	*0.0*	1	*16.7*	1	*7.1*	0	*0.0*	1	*7.1*
[50.01-79]	3	*3.2*	0	*0.0*	3	*3.6*	0	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*

[a]AF : Allelic Fraction
KRAS p.V8V excluded
For samples with multiple variants, the mutation with the highest allelic frequency was taken into account

DISCUSSION

To our knowledge, our study is the largest screening of *KRAS* mutations in plasma samples of pancreatic cancer cases, other pathological pancreatic conditions and healthy controls allowing for the comprehensive assessment of the sensitivity and specificity of *KRAS* mutations as non-invasive biomarkers for the detection of pancreatic cancer. Using only 2ng/amplicon (4ng total) of cfDNA and amplicon sizes below the size of the most prominent peak (166 bp) of the recently reported narrow range distribution of cfDNA fragments size [32], our NGS-based *KRAS* mutation screening assay combined with our developed Needlestack variant caller algorithm proved to be a sensitive approach to detect low-allelic fraction *KRAS* mutations down to 0.08%; a detection limit comparable to other amplicon-based NGS sequencing methods [27, 30, 31, 33].

We demonstrated that cfDNA *KRAS* mutations were detectable at the time of diagnosis in the plasma of 20% of pancreatic cancer cases at PDAC hotspot codons (12, 13 and 61); a sensitivity which is more consistent with some studies (between 27 to 36%) [14, 16, 20, 25] than others (between 47 to 81%) [15, 17–19]. As previously reported, the majority of these alterations were located at the hotspot codon 12, the spectrum was concordant with the distribution of *KRAS* tumor mutation types from ICGC data [34–36], suggesting that *KRAS* mutations in the circulating DNA mainly originate from tumor cells. Interestingly, although it has been shown that 90% of patients with PDAC carry primary *KRAS* mutations at codons 12, 13 or 61, we identified 9 cfDNA variants outside of the predominantly mutated codons, not reported in the ICGC PDAC database but reported in the COSMIC database for other types of cancer, allowing for an increased sensitivity of 22.5%. Those non-hotspot cfDNA *KRAS* mutations identified in pancreatic cancer cases may reflect the heterogeneity of the tumors or the alterations of genetically different metastatic lesions. In agreement with previous reports, we also demonstrated

that the proportion of cases with detectable cfDNA *KRAS* mutations tended to increase with more advanced stages and that *KRAS* allelic fractions were higher in cases than controls or in patients with chronic pancreatitis [23, 26]. Using a ddPCR assay focusing on the four most common PDAC mutations (G12D, G12V, G12R, G13D) Takai and colleagues identified cfDNA *KRAS* mutations in PDAC patients with distant organ metastasis in higher proportion than us (58.9% and 33.3% respectively). However, both studies report similar proportion of detected cases in non-metastatic and localized disease; 8.3% of patients with resectable PDAC (stages I and II) in Takai study and 10.3% of patients with localized pancreatic cancer in our study [23]. While a recent study using ddPCR demonstrated a higher sensitivity (43%; 22 patients) for the detection of *KRAS* mutation in plasma samples of patients with localized PDAC, 10 patients harbored a mutation at an allelic fraction ≤ 0.08% [22]. As 0.08% represents the lowest allele fraction that we could detect with our NGS-based approach and Needlestack algorithm, it is likely that some true low-allelic fraction mutants were too close to the sequencing noise signals to be detected at QVAL> 30. A combined strategy of pre-screening by NGS-amplicon followed by ddPCR of suggestive but inconclusive samples for specific mutations (for example samples with 10<QVAL<30) could circumvent some limitations by discriminating true positive low-level allele fractions mutants from inconclusive or false negative NGS samples, providing that the amount of cfDNA obtained is not a limiting factor. Preanalytical parameters regarding blood processing are also known to affect cfDNA concentrations [37]. A limitation of our study is that we did not test whether removing cellular debris with a high speed centrifugation of plasma samples prior cfDNA isolation could improve the sensitivity. However, the low quantities of cfDNA we could extract from the plasma samples on average indicate that contamination by cellular DNA was minimal. It is possible that a proportion of KRAS mutant pancreatic cancer do not release KRAS mutant cfDNA in the bloodstream, in which case the main limiting factor would be the biology of the tumor rather than the technology. Whether those differences in the release process of ctDNA between patients are due to differences in tumor micro-environment, vascularization, molecular characteristics and/or clonality remains to be discovered [38, 39].

Our study highlights that at our level of detection, a non-negligible proportion of controls are detected. Sausen and colleagues report 99.9% specificity of their assay against matched tumor DNA but they have not evaluated the specificity of their method against plasma of healthy controls. This becomes of capital importance when ultra-low detection limit is required as the proportion of positive calls in non-cancer individuals is likely to increase significantly. The assessment of the biological specificity of mutations in cfDNA as a non-invasive biomarker is either inexistent or limited in size. This may be partly explained by the fact that somatic mutations are believed to occur at negligible frequencies in normal cell populations [40], and thus expected to derive exclusively from the tumor burden. Yet, using a technique of limited sensitivity, Gormally et al. reported the presence of *KRAS* (1%) and *TP53* (3.2%) mutations in plasma of individuals who had remained clinically cancer-free for more than five years [41]. Very recently two studies revealed low-abundant *TP53* somatic mutations in body fluids of non-cancer individuals [42,43]. In addition, while limited in sample size, two studies described circulating *KRAS* mutations in 5% (2/37) [14] and 13% (4/31) [17] of patients with chronic pancreatitis. In our series, we detected 3.7% (N=14) *KRAS* positive individuals in the healthy controls (N=9 at hotspot codons) and 4.3% (N=6) in subjects with chronic pancreatitis, three of them at PDAC hotspot codon with an allelic fraction >1%. Given the prevalence of *KRAS* mutated cancers (predominantly pancreas, colon and lung) in the population, we cannot exclude that a small proportion of these individuals were non-diagnosed *KRAS* mutated cancer cases. Cell-free DNA fragments released into the blood circulation represent a molecular footprint of the entire genome, potentially including somatic mutations that occur at a mosaic state e.g affecting a limited number of tissues and cells. Syndromes caused by mosaic mutations in the Ras/MAPK signaling pathway (Mosaic RASopathies) have been described as a rather frequent congenital disorder that results in special skin phenotypes, whose epidermal and sebaceous disorders have been recently attributed, among other mutations, to oncogenic mosaic *KRAS* mutations [44]. The relatively high incidence of the most frequent mosaic RASopathy; sebaceous nevi (1 in 1,000 births) suggest that *KRAS* mutations present at a mosaic state in humans may not be a rare phenomenon [45]. Moreover, mosaic RASopathies are predominantly reported as skin disorders because of the accessibility of the lesions but the frequency of those syndromes could be underestimated as mosaic RASopathies of internal organs have been poorly investigated. While there are no accurate estimates of the prevalence and pathogenicity of mosaic *KRAS* mutations in human, it is possible that a proportion of cancer-free individuals with detectable low allelic fractions mutations in circulating DNA could reflect somatic mosaicism.

In conclusion, at a detection limit of 0.08% allelic fraction, our amplicon-based *KRAS* mutations sequencing assay applied to a large case-control series of plasma samples showed a limited sensitivity of 21.1% for the detection of pancreatic cancer and was not as specific as anticipated.

We detected 34% of advanced stages and 10% of early stages, suggesting that the limitation in sensitivity is at least partially attributable to the biology of the pancreatic malignancies. Whether reaching a lower threshold of detection for cfDNA mutations could increase

the discriminatory performance of the test remains to be assessed. We evaluated whether the combination of the detection of circulating *KRAS* mutations and the plasma CA19-9 levels could improve the detection of pancreatic cancer. We confirm a good sensitivity (90%) but a poor specificity for the CA19-9 plasma levels (64.8%). Combining cfDNA *KRAS* mutations and CA19-9 levels improved the sensitivity to 95% but the overall performance of the combined biomarkers did not significantly improve as compared to CA19-9 alone. However, combining cfDNA *KRAS* mutations could potentially contribute to expanded panels of non-invasive biomarkers involving different tumorigenesis processes and/or different mechanisms of release in the bloodstream, such as protein-based [46], exosome-based [47], methylation-based [48] or RNA-based markers [49], for the risk assessment of the disease.

MATERIALS AND METHODS

Study population, sample selection and ethics statement

Samples were selected from a multi-center case-control study conducted in Czech Republic and Slovakia and described in detail elsewhere [50, 51] (Supplementary data).

We conducted this study in two phases, a pilot series where we screened for *KRAS* mutations and measured CA19-9 plasma levels in plasma samples of 96 subjects and a validation series where we extended our initial *KRAS* mutation screening to plasma samples of 967 subjects. For the pilot series, we selected subjects with available plasma and pancreatic tissue (tumor or juice) samples, hence limiting our series to subjects recruited in Czech Republic. We selected all such cases with a histologically-confirmed PDAC diagnosis (N=40) and the 9 subjects diagnosed with chronic pancreatitis (N=9). In addition, we randomly selected 20 healthy controls among 916 with available plasma samples, frequency matched for the 40 PDAC cases on sex, age, tobacco and alcohol consumption. Finally, we selected 27 subjects recruited into the study as pancreatic cancer in first instance, but who subsequently were re-classified as benign neoplasms of the pancreas. For the validation study, we selected all remaining cases with histologically/cytologically confirmed pancreatic cancer (N=421); chronic pancreatitis subjects (N=145); as well as 401 healthy controls among 896, frequency matched for the cancer and chronic pancreatitis subjects on center, sex and age. For pancreatic cancer cases, stage grouping was defined as local, regional, and systemic cancers, based on TNM staging (AJCC 6th edition) when available, and estimation by the clinician when formal TNM staging was not available or not complete.

The study protocol was approved by the institutional review boards of the International Agency for Research on Cancer and all collaborating centers/institutions, and written informed consent was obtained for all participating subjects.

Isolation of plasma cell-free DNA (cfDNA) and quantification

Peripheral blood from patients was collected in EDTA Vacutainer tubes (Becton Dickinson). Blood samples were processed within 12 h of collection by centrifugation at 2,000g for 10 min and stored frozen in 2mL cryotubes. Circulating DNA (cfDNA) was isolated from 0.6-2.0mL (pilot series; average: 1.4mL) and from 0.3-1.0mL (validation series; average: 0.9mL) plasma with the QIAamp Circulating Nucleic Acid Kit (Qiagen), following manufacturer's instructions [52]. The concentration of purified cfDNA was determined using the Quant-iT™ PicoGreenR dsDNA Assay (Molecular Probes, Invitrogen) PicoGreen® a dilution series of a standard lambda DNA and a Fluoroskan Ascent FL instrument (Thermo Fisher Scientific).

KRAS amplification, library construction and deep sequencing with Ion Torrent PGM

As the size of the cfDNA fragments in cancer patients was recently reported to follow a narrowed-range, unimodal distribution reaching a peak at 166bp [32], primers were designed to amplify exons 2 and 3 so that the amplicon size is < 130bp (79bp and 129bp respectively), covering from codons 4 to 16 (hg19: ch12: 25,398,271 - ch12: 25,398,309) and from codons 51 to 69 (hg19: ch12: 25,380,228 - ch12: 25,380,307), totalling 119 bp excluding primer regions. Forward and reverse primer sequences were 5'-GCCTGCTGAAAATGACTGAA-3' and 5'-AGCTGTATCGTCAAGGCACT-3' for the amplification of partial *KRAS* exon 2 and 5'-GCAAGT AGTAATTGATGGAGAAACC-3' and 5'-TTTATGGCA AATACACAAAGAAAG-3' for the partial amplification of *KRAS* exon 3. Independent PCR amplifications of the 2 exons were performed using 2ng of cfDNA, 5X AccuStart Buffer, 200 nM forward and reverse primers and 0.04 U/mL of AccuStart HiFi Taq Polymerase (Quanta BioSciences) with the following conditions: 2 min at 94ºC, 50 cycles of 30s at 94ºC, 30s at 58ºC and 40s at 72ºC and a final elongation of 5 min at 72ºC. Approximately 20% of the PCR products were quantified by Qubit™ dsDNA HS Assay Kit and (Invitrogen) and Qubit® 2.0 fluorometer and 20 ng of exon 2 and 3 were pooled together, purified with Serapure magnetic beads at a final concentration of 2.5X and 28% of isopropanol. Library preparation was done using the NEBNext NEB Next® Fast DNA Library Prep Set for Ion Torrent™ kit (New England Biolabs) with some modifications, where each volume of reagents was reduced by a factor 4. Briefly, 12.5µl of the 20µl purified products were end-repaired in 15µl, and added

to 8.6 µl of ligation reaction mix, 0.7µl of the Ion P1 Adapter and 0.7 µl of each Ion Barcode for the ligation step. The barcoded products were purified using Serapure magnetic beads at final concentration of 1.8X, amplified in 25µl and quantified using Qubit quantification system. 40 ng of amplified barcoded products were pooled into a single tube and the cleanup and size selection of pooled libraries (230~250 bp) was performed in a 2% agarose gel and MinElute Gel Extraction Kit (Qiagen). The pool of purified barcoded libraries was quantified using the Qubit quantification system and the assessment of the library quality (molarity and size analysis) was done using the Agilent® High Sensitivity DNA Kit and the Agilent Technologies 2100 Bioanalyzer™ (Agilent Technologies). The pool of purified barcoded libraries was diluted to 280 millions of molecules in 25µl and sequenced with the IonTorrent™ PGM sequencer (Thermo Fisher Scientific) at deep coverage using the Ion OneTouch 200 Template Kit v2 DL and Ion PGM Sequencing 200 Kit v2 with the 316 or 318 chips (Thermo Fisher Scientific), following manufacturer's instructions. Library preparation and sequencing conditions were adapted from previous protocols [43].

Detection Threshold

Genomic DNA from the cell-line SW480 harboring a hemizygous *KRAS* p.G12V (c.35G>T) mutation was serially diluted into genomic DNA of a human wild-type lymphoblastoid cell-line in order to assess the accuracy and the detection threshold of the Ion Torrent Sequencing for the measurement of the mutant allelic fraction. Mutant abundances were as follows: 100%, 50%, 20%, 10%, 5%, 2%, 1%, 0.5%, 0.2%, 0.1%, 0.05%, 0.02%, 0.01%. Four independent PCR amplifications were done for each serial diluted point and for six wild-type DNA samples to determine the read error rate for that specific genomic position. PCR amplifications from 2ng, library construction and deep sequencing were done following the same protocol as for the cfDNA.

Measurement of the CA19-9 plasma level

Measurements of CA19-9 were performed on plasma EDTA samples from the pilot study. Analyses were done using an immunoradiometric assay by Beckmann Coulter (Marseille, France). Samples have been randomized through the batches of analyses. We used the clinically accepted cut-off of 37 kU/l for CA19-9 positivity [53].

Bioinformatics and statistical analyses

We used Needlestack, a variant caller algorithm suitable for the detection of low-abundance mutations [43] (https://github.com/IARCbioinfo/needlestack). The approach is based on the inclusion of sequencing data of a sufficient number of samples to robustly estimate the sequencing error rates at each position considered and for each possible base change. Reads were mapped to the human whole genome and BAM files were generated by the Ion Torrent PGM server using default parameters. Reads with a mapping quality below 20 were excluded from subsequent analysis. At each position and for each candidate variant, sequencing errors are modeled using a robust negative binomial regression [54] to avoid bias of the over-dispersion parameter due to the potential presence of genetic variants. We use a linear link and a zero intercept, and detected variants as being outliers from this error model. We calculated for each sample a p-value for being a variant (outlier from the regression) that we further transformed into q-values to account for multiple testing. *q*-values are reported in Phred scale QVAL=-10 $\log_{10}$(*q*-value), and we used a threshold of QVAL>30 to call variants. For each variant, we also calculated the relative variant strand bias defined by:

$$RVSB = \frac{\max(AO_p DP_m, AO_m DP_p)}{AO_p DP_m + AO_m DP_p}$$

where *DP* and *AO* denote respectively the total number of reads and the number of reads matching the candidate variant, with the subscripts *p* and *m* referring to the forward and reverse strands respectively.

ACKNOWLEDGMENTS

We thank C. Voegele for precious help with operational tasks, H. Renard for the case-control study data management, and AS. Navionis for technical assistance.

CONFLICTS OF INTEREST

None.

GRANT SUPPORT

IARC; the recruitment of subjects was supported by the MH CZ - DRO (MMCI, 00209805) and TMD was supported by la Ligue Nationale (Française) Contre le Cancer.

REFERENCES

1. Sant M, Allemani C, Santaquilani M, Knijn A, Marchesi F, Capocaccia R. EUROCARE-4. Survival of cancer patients diagnosed in 1995-1999. Results and commentary. Eur J Cancer. 2009; 45: 931-91.
2. Thomson CS, Forman D. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EUROCARE results? Br J Cancer. 2009; 101 Suppl 2: S102-9.

3. Bliss LA, Witkowski ER, Yang CJ, Tseng JF. Outcomes in operative management of pancreatic cancer. J Surg Oncol. 2014; 110: 592-98.
4. Klapman J, Malafa MP. Early detection of pancreatic cancer: why, who, and how to screen. Cancer Control. 2008; 15: 280-7.
5. Goonetilleke KS, Siriwardena AK. Systematic review of carbohydrate antigen (CA 19-9) as a biochemical marker in the diagnosis of pancreatic cancer. Eur. J. Surg. Oncol. 2007; 33; 266-270.
6. Huang Z, Liu F. Diagnostic value of serum carbohydrate antigen 19-9 in pancreatic cancer: a meta-analysis. Tumour Biol. 2014; 35: 7459-65.
7. Ballehaninna UK, Chamberlain RS. The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal. J Gastrointest Oncol. 2012; 3: 105-19.
8. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer. 2011; 11: 426-37.
9. Partensky C. Toward a better understanding of pancreatic ductal adenocarcinoma: glimmers of hope? Pancreas. 2013; 42: 729-39.
10. Caldas C, Kern SE. K-ras mutation and pancreatic adenocarcinoma. Int J Pancreatol. 1995; 18: 1-6.
11. Kanda M, Matthaei H, Wu J, Hong SM, Yu J, Borges M, Hruban RH, Maitra A, Kinzler K, Vogelstein B, Goggins M. Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia. Gastroenterology. 2012; 142: 730-33.e9.
12. Maitra A, Fukushima N, Takaori K, Hruban RH. Precursors to invasive pancreatic cancer. Adv Anat Pathol. 2005; 12: 81-91.
13. Morris JPt, Wang SC, Hebrok M. KRAS, Hedgehog, Wnt and the twisted developmental biology of pancreatic ductal adenocarcinoma. Nat Rev Cancer. 2010; 10: 683-95.
14. Castells A, Puig P, Mora J, Boadas J, Boix L, Urgell E, Sole M, Capella G, Lluis F, Fernandez-Cruz L, Navarro S, Farre A. K-ras mutations in DNA extracted from the plasma of patients with pancreatic carcinoma: diagnostic utility and prognostic significance. J Clin Oncol. 1999; 17: 578-84.
15. Dianxu F, Shengdao Z, Tianquan H, Yu J, Ruoqing L, Zurong Y, Xuezhi W. A prospective study of detection of pancreatic carcinoma by combined plasma K-ras mutations and serum CA19-9 analysis. Pancreas. 2002; 25: 336-41.
16. Jiao L, Zhu J, Hassan MM, Evans DB, Abbruzzese JL, Li D. K-ras mutation and p16 and preproenkephalin promoter hypermethylation in plasma DNA of pancreatic cancer patients: in relation to cigarette smoking. Pancreas. 2007; 34: 55-62.
17. Maire F, Micard S, Hammel P, Voitot H, Levy P, Cugnenc PH, Ruszniewski P, Puig PL. Differential diagnosis between chronic pancreatitis and pancreatic cancer: value of the detection of KRAS2 mutations in circulating DNA. Br J Cancer. 2002; 87: 551-4.
18. Mulcahy HE, Lyautey J, Lederrey C, qi Chen X, Anker P, Alstead EM, Ballinger A, Farthing MJ, Stroun M. A prospective study of K-ras mutations in the plasma of pancreatic cancer patients. Clin Cancer Res. 1998; 4: 271-5.
19. Yamada T, Nakamori S, Ohzato H, Oshima S, Aoki T, Higaki N, Sugimoto K, Akagi K, Fujiwara Y, Nishisho I, Sakon M, Gotoh M, Monden M. Detection of K-ras gene mutations in plasma DNA of patients with pancreatic adenocarcinoma: correlation with clinicopathological features. Clin Cancer Res. 1998; 4: 1527-32.
20. Dabritz J, Preston R, Hanfler J, Oettle H. Follow-up study of K-ras mutations in the plasma of patients with pancreatic cancer: correlation with clinical features and carbohydrate antigen 19-9. Pancreas. 2009; 38: 534-41.
21. Mora J, Urgell E, Farre A, Comas L, Montserrat E, Gonzalez-Sastre F. Agreement between K-ras sequence variations detected in plasma and tissue DNA in pancreatic and colorectal cancer. Clin Chem. 2006; 52: 1448-9.
22. Sausen M, Phallen J, Adleff V, Jones S, Leary RJ, Barrett MT, Anagnostou V, Parpart-Li S, Murphy D, Kay Li Q, Hruban CA, Scharpf R, White JR, et al. Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. Nat Commun. 2015; 6: 7686.
23. Takai E, Totoki Y, Nakamura H, Morizane C, Nara S, Hama N, Suzuki M, Furukawa E, Kato M, Hayashi H, Kohno T, Ueno H, Shimada K, et al. Clinical utility of circulating tumor DNA for molecular assessment in pancreatic cancer. Sci Rep. 2015; 5: 18425.
24. Tjensvoll K, Lapin M, Buhl T, Oltedal S, Steen-Ottosen Berry K, Gilje B, Soreide JA, Javle M, Nordgard O, Smaaland R. Clinical relevance of circulating KRAS mutated DNA in plasma from patients with advanced pancreatic cancer. Mol Oncol. 2016; 10: 635-43.
25. Uemura T, Hibi K, Kaneko T, Takeda S, Inoue S, Okochi O, Nagasaka T, Nakao A. Detection of K-ras mutations in the plasma DNA of pancreatic cancer patients. J Gastroenterol. 2004; 39: 56-60.
26. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, Antonarakis ES, Azad NS, Bardelli A, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014; 6 :224ra24.
27. Couraud S, Vaca-Paniagua F, Villar S, Oliver J, Schuster T, Blanche H, Girard N, Tredaniel J, Guilleminault L, Gervais R, Prim N, Vincent M, Margery J, et al. Noninvasive diagnosis of actionable mutations by deep sequencing of circulating free DNA in lung cancer from never-smokers: a proof-of-concept study from BioCAST/IFCT-1002. Clin Cancer Res. 2014; 20 : 4613-24.
28. Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, Dunning MJ, Gale D, Forshew T, Mahler-Araujo B, Rajan S, Humphray S, Becq J, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med. 2013; 368 :1199-209.

29. Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, Duncan C, Antipova A, Lee C, McKernan K, De La Vega FM, Kinzler KW, Vogelstein B, Diaz LA, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. 2010; 2: 20ra14.

30. Narayan A, Carriero NJ, Gettinger SN, Kluytenaar J, Kozak KR, Yock TI, Muscato NE, Ugarelli P, Decker RH, Patel AA. Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer Res. 2012; 72: 3492-8.

31. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE, Shrager JB, Loo BW, Alizadeh AA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014; 20: 548-54.

32. Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, Wong GL, Chan SL, Mok TS, Chan HL, Lai PB, Chiu RW, Lo YM. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci U S A. 2015; 112.

33. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, Dawson SJ, Piskorz AM, Jimenez-Linan M, Bentley D, Hadfield J, May AP, Caldas C, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci Transl Med. 2012; 4: 136ra68.

34. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, Miller K, Wilson PJ, Patch AM, Wu J, Chang DK, Cowley MJ, Gardiner BB, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature. 2012; 491: 399-405.

35. Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K, Quinn MC, Robertson AJ, Fadlullah MZ, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature. 2015; 518: 495-501.

36. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, et al. International network of cancer genome projects. Nature. 2010; 464: 993-8.

37. El Messaoudi S, Rolet F, Mouliere F, Thierry AR. Circulating cell free DNA: Preanalytical considerations. Clin Chim Acta. 2013; 424:222-30.

38. Thierry AR, Mouliere F, El Messaoudi S, Mollevi C, Lopez-Crapez E, Rolet F, Gillet B, Gongora C, Dechelotte P, Robert B, Del Rio M, Lamy PJ, Bibeau F, et al. Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. Nat Med. 2014; 20: 430-5.

39. Yong E. Cancer biomarkers: Written in blood. Nature. 2014; 511: 524-6.

40. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013; 339: 1546-58.

41. Gormally E, Vineis P, Matullo G, Veglia F, Caboux E, Le Roux E, Peluso M, Garte S, Guarrera S, Munnia A, Airoldi L, Autrup H, Malaveille C, et al. TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study. Cancer Res. 2006; 66: 6871-6.

42. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, Loeb LA, Swisher EM, Risques RA. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proc Natl Acad Sci U S A. 2016.

43. Fernandez-Cuesta L, Perdomo S, Avogbe PH, Leblay N, Delhomme TM, Gaborieau V, Abedi-Ardekani B, Chanudet E, Olivier M, Zaridze D, Mukeria A, Vilensky M, Holcatova I, et al. Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. EBioMedicine. 2016; 117-23.

44. Hafner C, Toll A, Gantner S, Mauerer A, Lurkin I, Acquadro F, Fernandez-Casado A, Zwarthoff EC, Dietmaier W, Baselga E, Parera E, Vicente A, Casanova A, et al. Keratinocytic epidermal nevi are associated with mosaic RAS mutations. J Med Genet. 2012; 49: 249-53.

45. Hafner C, Groesser L. Mosaic RASopathies. Cell Cycle. 2013; 12: 43-50.

46. He XY, Liu BY, Yao WY, Zhao XJ, Zheng Z, Li JF, Yu BQ, Yuan YZ. Serum DJ-1 as a diagnostic marker and prognostic factor for pancreatic cancer. J Dig Dis. 2011; 12: 131-7.

47. Melo SA, Luecke LB, Kahlert C, Fernandez AF, Gammon ST, Kaye J, LeBleu VS, Mittendorf EA, Weitz J, Rahbari N, Reissfelder C, Pilarsky C, Fraga MF, et al. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. Nature. 2015; 523: 177-82.

48. Yi JM, Guzzetta AA, Bailey VJ, Downing SR, Van Neste L, Chiappinelli KB, Keeley BP, Stark A, Herrera A, Wolfgang C, Pappou EP, Iacobuzio-Donahue CA, Goggins MG, et al. Novel methylation biomarker panel for the early detection of pancreatic cancer. Clin Cancer Res. 2013; 19: 6544-55.

49. Liu R, Chen X, Du Y, Yao W, Shen L, Wang C, Hu Z, Zhuang R, Ning G, Zhang C, Yuan Y, Li Z, Zen K, et al. Serum microRNA expression profile as a biomarker in the diagnosis and prognosis of pancreatic cancer. Clin Chem. 2012; 58: 610-8.

50. Brenner DR, Wozniak MB, Feyt C, Holcatova I, Janout V, Foretova L, Fabianova E, Shonova O, Martinek A, Ryska M, Adamcakova Z, Flaska E, Moskal A, et al. Physical activity and risk of pancreatic cancer in a central European multicenter case-control study. Cancer Causes Control. 2014; 25: 669-81.

51. Urayama KY, Holcatova I, Janout V, Foretova L, Fabianova E, Adamcakova Z, Ryska M, Martinek A, Shonova O, Brennan P, Scelo G. Body mass index and body size in early adulthood and risk of pancreatic cancer in a central European multicenter case-control study. Int J Cancer. 2011; 129: 2875-84.

52. Devonshire AS, Whale AS, Gutteridge A, Jones G, Cowen S, Foy CA, Huggett JF. Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification. Anal Bioanal Chem. 2014; 406:6499-512.

53. Steinberg W. The clinical utility of the CA 19-9 tumor-associated antigen. Am J Gastroenterol. 1990; 85: 350-5.

54. Aeberhard WH, Cantoni E, Heritier S. Robust inference in the negative binomial regression model with an application to falls data. Biometrics. 2014; 70: 920-31.

4.3 Scientific contribution B

In this study, we have developed the UroMuTERT assay in order to detect *TERT* promoter mutations in the cfDNA from blood and urine samples and in the DNA from urinary exfoliated cells (cellDNA) of urothelial cancer patients in the context of non-invasive cancer detection. Indeed, it has been reported that between 60% and 85% of urothelial cancer patients have a tumor *TERT* promoter mutation, even in early stage [72].

We have analyzed a total of 93 cases and 94 controls in a first cohort (the DIAGURO French cohort) and 50 cases and 50 controls in a independent second cohort (the IPO-PORTO Portuguese cohort). For the DIAGURO cohort, tumor-urine-blood trios for cases and urine-blood duos for controls were available (cfDNA). For the PORTO cohort, urinary exfoliated cells from urine were available. Deep sequencing (IonTorrent Proton technology) was performed on two recurrently mutated genomic positions: C228T and C250T.

We used our needlestack variant caller to detect mutations potentially in low abundance. Our UroMuTERT assay could detect C228T and C250T mutations at VAF down to respectively 0.8% and 0.5%. The sensitivity of the assay in the DIAGURO cohort for urine samples was estimated at around 80%, and the specificity at around 97%, but for blood samples, the sensitivity was only 7%, suggesting low amounts of tumor-derived mutations as recently described [100]. The sensitivity of the assay in the IPO-PORTO cohort was estimated at around 68% (not significantly lower than for the DIAGURO cohort), and the specificity at around 98%. This study has shown an unprecedented performance of the UroMuTERT assay that quantifies tumor-derived TERT promoter mutation in urine samples for the detection urothelial carcinomas and that can be used for large-scale validation and biomarker development.

4.3.1 Article D

Urinary *TERT* promoter mutations as non-invasive biomarkers for the comprehensive detection of urothelial cancer

Patrice H. Avogbe Ph.D.[a], Arnaud Manel M.D.[b], Emmanuel Vian M.D.[b], Geoffroy Durand B.Sc.[a], Nathalie Forey B.Sc.[a], Catherine Voegele Ph.D.[a], Maria Zvereva Ph.D., D.Sc.[a,c], Md Ismail Hosen Ph.D.[a], Sonia Meziani B.Sc.[a], Berengere De tilly D.N.P.[b], Gilles Polo M.D.[b], Olesia Lole M.Sc.[a], Pauline Francois M.Sc.[a], Tiffany M. Delhomme M.Sc.[a], Christine Carreira B.Sc.[a], Sara Monteiro-Reis M.Sc.[d], Rui Henrique Ph.D.[d,e,f], Behnoush Abedi-Ardekani M.D.[a], Graham Byrnes Ph.D.[a], Matthieu Foll Ph.D.[a], Elisabete Weiderpass M.D., Ph.D.[a], James McKay Ph.D.[a], Carmen Jeronimo Ph.D.[d,e,f], Ghislaine Scelo Ph.D.[a], Florence Le Calvez-Kelm Ph.D.[a*]

[a] International Agency for Research on Cancer (IARC), Lyon, France

[b] Protestant Clinic of Lyon, Urology department, Lyon, France

[c] Faculty of Chemistry, Lomonosov Moscow State University, Moscow, Russia

[d] Portuguese Oncology Institute of Porto, Research Center (CI-IPOP), Porto, Portugal

[e] Portuguese Oncology Institute of Porto (IPOP), Department of Pathology, Porto, Portugal

[f] Institute of Biomedical Sciences Abel Salazar, University of Porto (ICBAS-UP), Porto, Portugal

Keywords: blood, cell DNA, cell-free DNA, *TERT* mutations, urine, urothelial cancer detection

Corresponding Author*:

Dr. Florence Le Calvez-Kelm
International Agency for Research on Cancer, Genetic Cancer Susceptibility group
150 Cours Albert Thomas
69372 Lyon Cedex 08, FRANCE
Phone: +33 4 72 73 83 58
Fax: +33 4 72 73 83 88
E-mail: lecalvezf@iarc.fr

Word count of the text (including Abstract): 2799
Word Count of Abstract: 300 without headings (315 with headings)
Number of figures and tables: 5

ABSTRACT

Background

Recurrent mutations in the promoter of the telomerase reverse transcriptase (*TERT)* gene (C228T and C250T) detected in tumor cells shed into urine of urothelial cancer (UC) patients are putative non-invasive biomarkers for UC detection and monitoring.

Objectives

To evaluate the clinical performance of a single-gene assay quantifying *TERT* promoter mutations in cell-free circulating DNA (cfDNA) in blood and urine, or DNA from urinary exfoliated cells (cellDNA) for the detection of primary and recurrent UC. To compare its performance with urine cytology.

Design, setting and participants

We developed a single-plex assay (UroMuTERT) for the detection of low-abundance *TERT* promoter mutations. We tested 93 primary and recurrent UC cases and 94 controls in France (prospectively collected blood, urine samples and tumors for the cases), and 50 primary UC cases and 50 controls in Portugal (retrospective urinary exfoliated cell samples).

Outcome measurements and statistical analysis

Sensitivity, specificity and accuracy of the liquid-based biomarkers. Association of mutation status with disease characteristics.

Results and limitations

In the French series, C228T or C250T were detected in urinary cfDNA or cellDNA in 81 cases (87.1%), and five controls (Specificity 94.7%), with 98.6% concordance in matched tumors. Detection rate in plasma cfDNA among cases was 7.1%. The UroMuTERT sensitivity was (i) highest for urinary cfDNA and cellDNA combined, (ii) consistent across primary and recurrent cases, tumor stages and grades, (iii) higher for low-risk non-muscle invasive UC (86.1%) than urine cytology (23.0%) (P<0·0001) and (iv) 93.9% when combined

with cytology. In the Portuguese series, the sensitivity and specificity for detection of UC with urinary cellDNA was 68.0% and 98.0%. Limitations include study size and inability to assess urinary cfDNA in the Portuguese series.

Conclusions

TERT promoter mutations detected by the UroMuTERT assay in urinary DNA (cfDNA or cellDNA) show excellent sensitivity and specificity for the detection of UC, significantly outperforming that of urine cytology notably for detection of low-grade early stages UC.

INTRODUCTION

Bladder cancer (BC), accounting for 90% of urothelial cancer (UC), has become a common cancer globally [1]. While 70-80% of BCs are non-muscle-invasive carcinoma [2], high rates of recurrence (50-70%) and progression to the muscle (10-20%) require close monitoring after first-line treatment. Upper tract urothelial cell carcinoma (UTUCC, 10% of UCs), while different in many aspects, shares many histological features and genetic alterations with BC [3]. UC detection relies on invasive cystoscopy, imaging approaches and noninvasive urine cytology, however the latter lacking sensitivity in detecting low-grade BC [4] and UTUCC [5]. Performance inconsistencies of FDA-approved urine-based biomarkers prevent their routine clinical use [6].

Mutations in the promoter of the Telomerase Reverse Transcriptase gene (*TERT*) are frequent in various human cancers. In both BCs and UTUCCs, they are observed in 60-85% of cases including in early stages [7, 8]. These mutations were detected in DNA from urinary exfoliated cells collected at diagnosis and during follow-ups [8-13]. Recently, urinary cell-free DNA (cfDNA) showed higher analytical sensitivity than DNA from exfoliated cells (cellDNA) for the detection of UC tumor-derived alterations [14]. Assessment of these mutations in different sources of DNA in urine and blood pairs has never been made in a case-control setting with a sensitive single-gene assay.

Because a sensitive and specific biomarker of UC might profoundly influence clinical practice, we developed a single-plex assay, UroMuTERT, based on *TERT* promoter ultra-deep sequencing and an algorithm for detection of low-abundance mutations [15]. We assessed *TERT* promoter mutations in DNA from various body fluids (cfDNA in blood and urine, and cellDNA) in two case-control series and compared UroMuTERT diagnostic performance to that of urine cytology.

MATERIALS AND METHODS

Study population and clinical specimens

Participants were recruited from two case-control studies. Written informed consent was obtained for all participants and details about ethical approvals of the study protocols are given in the supplement.

DIAGURO case-control study: Recruitment was conducted in the Protestant Clinic (Lyon, France) during 2016-2017. Clinical cases included patients with post-surgery histological confirmation of primary or recurrent UC (BC or/and UTUCC at any stage and grade). Controls were patients with urological pathological conditions other than UC or undergoing colonoscopy (Supplemental Fig.1). Clinical and epidemiological data were collected. Prospective sample collection included tumor-urine-blood trios for cases (before surgery) and urine-blood duos for controls (Supplemental Fig. 1). Blood and urine samples were processed within two hours of collection and DNA from plasmas, white blood cells (WBC), urine supernatants (US), urine pellets (UP) and tumor tissues were processed using standard protocols (Supplemental Fig. 2). A qualified pathologist performed histological review of the tumor tissues.

IPO-PORTO case-control replicative series: CellDNA from UP of 50 primary bladder cancer cases and 50 controls (healthy donors, with no history of cancer) were retrospectively selected from the Biobank of the Portuguese Oncology Institute of Porto. Clinical data were collected for all participants. Sample collection and processing are detailed in the supplement.

UroMuTERT assay and mutation analysis

A single-plex of 147bp was designed to be smaller than the 167bp average fragment size of cfDNA and to cover the C228T and C250T genomic positions. Experimental conditions for ultra-deep sequencing and assessment of detection thresholds are given in the supplement.

Variant calling was performed using our Needlestack algorithm specifically designed for the detection of low-allelic fraction mutations (MAFs) [15, 16]. It includes C228T and C250T and other rare BC mutations previously reported (C181T, C176G, C228A, CC242-243TT, G245T) [9, 11]. Reads with base quality below 13 at the evaluated positions were excluded. A p-value for being a variant (outlier from the regression) was calculated for each sample and transformed into q-values to account for multiple testing. A threshold of Phred scale *q*-values QVAL>20 was used to call variants (QVAL= -10 $\log_{10}$ (*q*-value)).

Statistical analyses

Mann-Whitney or Kruskal-Wallis tests were used for comparisons of quantitative variables between patient groups, Pearson χ^2 tests and two-tailed Fisher exact tests used for categorical variables. Sensitivities, specificities and accuracy of the putative biomarkers were calculated for the different sources of DNA with confidence intervals computed with the Clopper-Pearson method. Positive and negative predictive values (PPV and NPV) were calculated for patients at high-risk of BC, estimated at 30% for patients with hematuria or with lower urinary tract symptoms (LUTS) according to Springer and colleagues [12]. Confidence intervals for the predictive values are the standard logit confidence intervals given by Mercaldo et al. 2007 [17]. Analyses were conducted using IBM SPSS Statistics 20.

RESULTS

Performance of urinary UroMuTERT in detecting UC (DIAGURO cohort)

The DIAGURO cohort included 94 controls and 93 UC cases, of whom 93.5% had BC; 4.3% had mixture of BC and other urogenital tumors and 2.2% had UTUCC. 90.3% of cases were non-muscle-invasive UC (NMIUC) and 48.4% diagnosed with primary UC. Overall, cases and controls were balanced with respect to baseline characteristics (Table 1). CellDNA and

cfDNA yields were compared and evaluated for associations with clinical parameters [18] (Supplemental Figs. 3/4).

Technical validation showed that UroMuTERT could detect C228T and C250T mutations down to 0.8% and 0.5% MAFs respectively at sequencing read depth >10 000X (Supplemental Table 1, Fig. 5).

Sequencing results were available for 594 samples corresponding to US cfDNA (n=176), UP cellDNA (n=187), plasma cfDNA (n=148), and tumor DNA (n=83) at a mean depth of 9092X. The sensitivity of C228T and/or C250T was 81.8% for US and 83.5% for UP (Table 2). While the false positive rate was lower for the US analysis (2.3%) compared to the UP (5.4%), the US assay performance was hampered by the number of samples without sequencing data (N=11; 5.8%) which was more frequent than for UP (N=3; 1.6%). US and UP median MAF was 19.2%, (range 0.6%–68.8%) and 23.7% (range 1.0%–75.2%) with 34.7% and 22.1% of cases with MAF<10% respectively (Supplemental Table 2). In both US and UP, MAFs correlated with the tumor risk-score, with significantly higher mutational load in high-risk (pTa/pT1 high grade and any stage associated with CIS) compared with low-risk NMIUC (Low-grade pTa or pT1 tumors) (Supplemental Fig. 6). Combined urinary cfDNA and urine pellet DNA analysis outperforms either DNA types considered individually; overall sensitivity of 87.1% and specificity of 94.7%, with no missing data reported (Table 2). However, the differences were not statistically significant. Mutational status in US and UP was concordant in 79 of the 86 cases with sequencing data in both sample types (91.9%), of which 79.1% had *TERT* positive results (Figure 1, Supplemental Tables 3/4 for the lists of subjects with *TERT* variants). Five cases with mutations in UP were negative in US and two cases inversely. We noted comparable performance in detecting UC when rare but concomitant mutations to the predominant C228T/C250T detected in ten urinary DNAs of cases (C228A, CC242-243TT and a newly discovered G238A) were considered (Table 2,

Supplemental Table 3). There was no indication that UroMuTERT detection ability is modified by the primary or recurrent status of UC (Supplemental Table 5), neither by the tumor grade, risk score or muscle invasiveness (Supplemental Table 6) and the mutational pattern was equally distributed among those categories (Supplemental Fig. 7). We assessed the analytical sensitivity on the 83/93 available matched tumors and identified urinary *TERT* promoter mutation(s) in US or UP in 71 of the 72 TERT mutated tumors (analytical sensitivity of 98.6%; 95% CI, 92.5−99.96). Mutational status details between tumor and urine samples are provided in the supplement.

One of six controls positive in US or UP had a history of prostate cancer. None of mutated controls had however incidental detection of prostate cancer after prostate resection at inclusion (N=7) (Figure 1). As there was no difference in sensitivity for the detection of primary and recurrent cases (Supplemental Table 5), extrapolated PPV and NPV for patients at a hypothetical 30% UC risk, e.g with hematuria, LUTS or others [12] were calculated on the overall set of data. They were best for US (PPV: 93·9% and NPV: 92·6%) (Table 2) but did not consider missing data (N=10). The combined UP/US analysis overcame these limitations with PPV and NPV of 87·6% and 94·4% respectively.

Blood-based detection of *TERT* promoter mutations

In contrast to urine, a much lower performance was observed for plasma cfDNA (sensitivity of 7.1%; $P<0.001$). Importantly, the five cases with mutations in plasma cfDNA scored positive also for US or UP. The detection of concomitant C228T/C228A in plasma cfDNA, US, and UP at consistent levels (mean of 17.3%/0.4%) in a control prompted us to screen WBC to determine the origin of the multiple TERT positivity. WBC tested positive for C228T/C228A at similar levels, which is suggestive of mosaicism or clonal hematopoiesis associated with hematuria. Three cases tested positive in WBC DNA (Fig. 1) and plasma cfDNA at similar AFs, and in US, UP and tumor DNA, one of which with C228T levels

consistent with a germline or non-clonal mosaicism (MAF range 32.6%–45.8%). In the two other cases MAFs were higher in urine (4- and 6-fold) and tumors (2 and 14-fold) than in WBC and plasma, suggesting a dual contribution of mosaicism and tumorigenesis to the urinary mutational load (Supplemental Table 3).

Urinary *TERT* promoter mutations detection for primary UC (IPO-PORTO cohort)

The reproducibility of UroMuTERT was assessed in 50 primary UC cases and 50 healthy controls, where only urine cellDNA was available (Table 1). 76% of the tumors were classified as high-grade and 64% categorized as NMIUC. The overall sensitivity was 68.0% with a specificity of 98.0% (Tables 2, Supplemental Table 6). While no difference in sensitivity of detecting primary or recurrent UC was observed in the DIAGURO cohort (86.7% and 87.5%), the 68% estimate observed in the PORTO cohort was compared to the sensitivity obtained in the same conditions, e.g for DIAGURO primary UC detected with cellDNA only (84.1%). A borderline non-significant 16.1% difference in detecting primary UC with cellDNA between the two cohorts was observed (P= 0.07).

Comparison of UroMuTERT performance with urine cytology

Sensitivity of UroMuTERT in detecting low-risk NMIUC was significantly higher (86.1%) than that of urine cytology (23.0%, P<0.0001, Fig. 2), whereas no difference was observed in detecting high-risk NMIUC or MIUC. In the DIAGURO cohort, combined UroMuTERT and urine cytology enabled the detection of 62/66 cases compared to the UroMuTERT only where 59 patients had urine positive test(s) (sensitivity: 93.9%; 95% CI, 85.2−98.3 versus 89.4%; 95% CI, 79.4−95.6 respectively).

DISCUSSION

We developed UroMuTERT, a simple, non-invasive and sensitive assay with detection thresholds of 0.8% and 0.5% MAFs for C228T and C250T mutations. We evaluated its

clinical validity for the detection of UC against urine cytology. Our study shows excellent clinical sensitivity (87.1%), specificity (94.7%) and analytical sensitivity (98.6%) of a single-gene urinary biomarker based on tumor-derived *TERT* promoter mutations for the detection of all forms of UC. The diagnostic performance was best for urinary cfDNA and cellDNA combined. The ability of UroMuTERT to quantify low-level mutations enabled the detection of a significant proportion of cases with MAF<5% (26.4% in US and 13.0% in UP) and is therefore a critical parameter for enhanced sensitivity. Analyzing additional rare *TERT* promoter mutations did not improve UroMuTERT performance, as they were concomitant to C228T and/or C250T.

In previous studies, sensitivities and specificities of the same markers tested on alternative assays and only in exfoliated urothelial cells (cellDNA) varied from 55% to 62% and from 90% to 99% respectively in patients with incident or early BC and from 42% to 57% and 73% to 90% respectively in patients with recurrent BC [8, 9, 12, 13]. Two studies reported sensitivity of 80% using pre-surgery urine cellDNA but no precision was given on the primary or recurrence status [10, 11]. Our UroMuTERT assay demonstrated comparable performance to that of recently developed UroSEEK multiple markers assay (including C228T and C250T) for the detection of primary or early UC (sensitivity of 86.7% versus 83%; Specificity of 94.7% versus 93%) and higher sensitivity for the detection of UC recurrence (87·5% versus 68%) [12].

Importantly, our *TERT* mutation biomarkers achieved high specificity against patients with urological pathologies other than UC (including incidental prostate cancer cases) who may benefit from UroMuTERT screening as the symptoms may be similar to the ones observed in UC cases.

Consistent with previous findings [10], the added diagnostic value of the urinary *TERT* promoter mutations as biomarkers was particularly evident for the detection low-risk NMIUC

as compared to urine cytology (sensitivity of 86.1% versus 23.5%), where cytology demonstrated poor performance [4]. Combined UroMuTERT and cytology assays improved sensitivity up to 93.9%. UroMuTERT (cfDNA or cellDNA) PPV of 87.6% and NPV of 94.4% extrapolated to at high-risk subjects of UC (30% estimated risk [12]) reached 88.4% and 97.4% respectively when combined with cytology and assuming 100% specificity for cytology. Lower hypothetical risks of 20% and 5% for UC in patients with hematuria [19] and micro-hematuria [20], led to PPVs of 81.4% and 48.0% and NPVs of 98.4% and 99.7% respectively, which still demonstrates the superior diagnostic value of combined urinary UroMuTERT and cytology, which should be accurately assessed in large well-defined high-risk group populations [21]. We expect UroMuTERT to change UC detection by complementing cytology or replacing urine-based markers which lack performance for clinical utility [22]. Its high accuracy for early-stage UC should improve timely transurethral tumor resections, which in turn will contribute to reduced risk of progression and improved patients' survival. The high NPV in high-risk UC individuals may provide evidence for a reliable substitute to unnecessary cystoscopies to patients with negative tests, avoiding discomfort and risk of complications associated with invasive procedures while reducing high cost of clinical management of suspected UCs and patient non-adherence to screening or surveillance [23, 24]. We lay out a conceptual strategy integrating UroMuTERT as a primary diagnostic tool to individuals at high-risk or under surveillance for UC recurrence (Fig. 3). The sensitivity of our biomarkers in plasma is poor, reflecting low amounts of tumor-derived mutations in UC patients, such as recently described [25]. In addition, *TERT* positivity in plasma cfDNA can be confounded with rare leucocyte-derived mutations, whose influence on UC development should be further investigated. Rare mosaicism in patients with UC has been observed [26] and may add a layer of complexity in the interpretation of a urinary *TERT* positive test with negative subsequent cystoscopy or urography (Figure 3).

One limitation of our study is the inability to assess paired urinary cfDNA and cellDNA in the replication Portuguese series. Focusing on urinary cells we found a non-significant lower sensitivity in detecting Portuguese primary UC (66-68%) than French primary UC (84%), raising the question of whether differences in *TERT* promoter mutations frequency between the two cohorts exist or whether a subset of Portuguese urine samples carries mutations at MAF below detection thresholds.

CONCLUSIONS

Our study demonstrates unprecedented performance of a single-gene assay quantifying tumor-derived *TERT* promoter mutation load in urine for the detection of all forms of UC and lays the foundations for large-scale validation and clinical utility studies. The role of rare *TERT* promoter mutations in leucocytes on UC development and its impact on the clinical use of the biomarkers should be further examined.

Author Contributions: AM and FLCK supervised the project. AM, GB, EW, JM, CJ, GS, FLCK contributed to the study design. AM, EV, SM, NF, BDT, GP, SMR, RH, CJ contributed to recruitment of participants and collection of samples and medical data. GD, NF, OL, PF, CC designed and conducted the experiments. CV, TMD, MF did the bioinformatics data analysis. PHA and GB did the statistical analysis. BAA conducted pathological examinations. PHA, MZ, MIH, JM, CJ, GS AND FLCK interpreted the validation set. PHA and FLCK wrote the manuscript. All authors reviewed the manuscript and approved the final version.

Financial disclosures: we declare no competing interests

Funding/Support and role of the sponsor: The study was supported by the French Cancer League and the French Foster Research in Molecular Biology. SM was supported by the French Association for Research on Molecular Biology. OL, PF and TMD were supported by

the French Cancer League. The work reported in this paper was undertaken during the tenure of PHA's and MIH's postdoctoral fellowships from the International Agency for Research on Cancer, partially supported by the European Commission FP7 Marie Curie Actions – People – Co-funding of regional, national and international programmes (COFUND).

Acknowledgments: The authors would like to thank all patients participating in the case-control studies, Thierry Degoul, director of the Protestant clinic for his support to the DIAGURO project and the clinical research team members who supported the sample collection (From the Urology department of the Protestant Clinic of Lyon, France; Aurélie Couyotopoulo, Emilie Laurent, Emilie Morcillo, Noemie Tarride, Estelle Maillet, Véronique Richard, Audrey Franchi, Odette Jaume, Rachel Maynard). We also thank René Lattes for precious support in fund raising, Jean-Damien Combes for legal advice, Helene Renard for technical assistance in sample collection, Nolwenn Saunier and Isabelle Rondy for administrative support and Katarzyna Szymanska for scientific editing.

297 **REFERENCES**
298
299 [1] Antoni S, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F. Bladder Cancer
300 Incidence and Mortality: A Global Overview and Recent Trends. European urology.
301 2017;71:96-108.
302 [2] Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM. The 2016 WHO
303 Classification of Tumours of the Urinary System and Male Genital Organs-Part A:
304 Renal, Penile, and Testicular Tumours. European urology. 2016;70:93-105.
305 [3] Lee JY, Kim K, Sung HH, Jeon HG, Jeong BC, Seo SI, et al. Molecular
306 Characterization of Urothelial Carcinoma of the Bladder and Upper Urinary Tract.
307 Translational oncology. 2018;11:37-42.
308 [4] Lotan Y, Roehrborn CG. Sensitivity and specificity of commonly available bladder
309 tumor markers versus cytology: results of a comprehensive literature review and meta-
310 analyses. Urology. 2003;61:109-18; discussion 18.
311 [5] Baard J, de Bruin DM, Zondervan PJ, Kamphuis G, de la Rosette J, Laguna MP.
312 Diagnostic dilemmas in patients with upper tract urothelial carcinoma. Nature reviews
313 Urology. 2017;14:181-91.
314 [6] Zuiverloon TCM, de Jong FC, Theodorescu D. Clinical Decision Making in
315 Surveillance of Non-Muscle-Invasive Bladder Cancer: The Evolving Roles of Urinary
316 Cytology and Molecular Markers. Oncology (Williston Park, NY). 2017;31:855-62.
317 [7] Killela PJ, Reitman ZJ, Jiao Y, Bettegowda C, Agrawal N, Diaz LA, Jr., et al. TERT
318 promoter mutations occur frequently in gliomas and a subset of tumors derived from
319 cells with low rates of self-renewal. Proceedings of the National Academy of Sciences
320 of the United States of America. 2013;110:6021-6.
321 [8] Kinde I, Munari E, Faraj SF, Hruban RH, Schoenberg M, Bivalacqua T, et al. TERT
322 promoter mutations occur early in urothelial neoplasia and are biomarkers of early
323 disease and disease recurrence in urine. Cancer research. 2013;73:7162-7.
324 [9] Allory Y, Beukers W, Sagrera A, Flandez M, Marques M, Marquez M, et al.
325 Telomerase reverse transcriptase promoter mutations in bladder cancer: high
326 frequency across stages, detection in urine, and lack of association with outcome.
327 European urology. 2014;65:360-6.
328 [10] Descotes F, Kara N, Decaussin-Petrucci M, Piaton E, Geiguer F, Rodriguez-Lafrasse
329 C, et al. Non-invasive prediction of recurrence in bladder cancer by detecting somatic
330 TERT promoter mutations in urine. British journal of cancer. 2017;117:583-7.
331 [11] Hurst CD, Platt FM, Knowles MA. Comprehensive mutation analysis of the TERT
332 promoter in bladder cancer and detection of mutations in voided urine. European
333 urology. 2014;65:367-9.
334 [12] Springer SU, Chen CH, Rodriguez Pena MDC, Li L, Douville C, Wang Y, et al. Non-
335 invasive detection of urothelial cancer through the analysis of driver gene mutations
336 and aneuploidy. eLife. 2018;7.
337 [13] Ward DG, Baxter L, Gordon NS, Ott S, Savage RS, Beggs AD, et al. Multiplex PCR
338 and Next Generation Sequencing for the Non-Invasive Detection of Bladder Cancer.
339 PloS one. 2016;11:e0149756.
340 [14] Togneri FS, Ward DG, Foster JM, Devall AJ, Wojtowicz P, Alyas S, et al. Genomic
341 complexity of urothelial bladder cancer revealed in urinary cfDNA. European journal
342 of human genetics : EJHG. 2016;24:1167-74.
343 [15] Le Calvez-Kelm F, Foll M, Wozniak MB, Delhomme TM, Durand G, Chopard P, et
344 al. KRAS mutations in blood circulating cell-free DNA: a pancreatic cancer case-
345 control. Oncotarget. 2016;7:78827-40.
346 [16] Foll M. Needlestack: A multi-sample somatic variant caller. 2018.

[17] Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. Statistics in medicine. 2007;26:2170-83.
[18] Millan-Rodriguez F, Chechile-Toniolo G, Salvador-Bayarri J, Palou J, Algaba F, Vicente-Rodriguez J. Primary superficial bladder cancer risk groups according to progression, mortality and recurrence. The Journal of urology. 2000;164:680-4.
[19] Ngo B, Papa N, Perera M, Bolton D, Sengupta S. Bladder cancer diagnosis during haematuria investigation - implications for practice guidelines. BJU international. 2017;119 Suppl 5:53-4.
[20] Grossfeld GD, Litwin MS, Wolf JS, Hricak H, Shuler CL, Agerter DC, et al. Evaluation of asymptomatic microscopic hematuria in adults: the American Urological Association best practice policy--part I: definition, detection, prevalence, and etiology. Urology. 2001;57:599-603.
[21] Larre S, Catto JW, Cookson MS, Messing EM, Shariat SF, Soloway MS, et al. Screening for bladder cancer: rationale, limitations, whom to target, and perspectives. European urology. 2013;63:1049-58.
[22] Clinton T, Lotan Y. Review of the Clinical Approaches to the Use of Urine-based Tumor Markers in Bladder Cancer. Rambam Maimonides medical journal. 2017;8.
[23] Eble JN, Sauter G, Epstein J, Sesterhenn I. Pathology and genetics of tumours of the urinary system and male genital organs. 3rd edition ed. France: Lyon, Oxford, IARC Press 2004.
[24] Schrag D, Hsieh LJ, Rabbani F, Bach PB, Herr H, Begg CB. Adherence to surveillance among patients with superficial bladder cancer. Journal of the National Cancer Institute. 2003;95:588-97.
[25] Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Science translational medicine. 2018;10.
[26] Hafner C, Toll A, Real FX. HRAS mutation mosaicism causing urothelial cancer and epidermal nevus. The New England journal of medicine. 2011;365:1940-2.

Table 1. Patient's baseline characteristics

Characteristics	DIAGURO Cohort (N=187)		PORTO Cohort (N=100)	
	UC patients (N=93)	Controls (N=94)	UC patients (N=50)	Controls (N=50)
Median age (range)- yr	72 (42–95)	70 (34–93)	68 (37–91.4)	46 (38–62)
Sex - no. (%)				
Female	17 (18.3)	23 (24.5)	5 (10.0)	26 (52.0)
Male	76 (81.7)	71 (75.5)	45 (90.0)	24 (48.0)
Smoking status - n. (%)				
Never	23 (24.7)	39 (41.5)	–	–
Former	44 (47.3)	44 (46.8)	–	–
Current	21 (22.6)	11 (11.7)	–	–
Missing	5 (5.4)	–	–	–
Alcohol status - n. (%)				
Never	23 (24.7)	22 (23.4)	–	–
Ex-drinker	13 (14.0)	7 (7.4)	–	–
Current drinker	52 (55.9)	64 (68.1)	–	–
Missing	5 (5.4)	1 (1.1)		
Cancer history - n. (%)				
No	68 (73.1)	82 (87.2)	–	–
Yes	18 (19.4)	12 (12.8)	–	–
Missing	7 (7.5)	–	–	–
Diabetes - n. (%)				
No	69 (74.2)	82 (87.2)	–	–
Yes	18 (19.4)	12 (12.8)	–	–
Missing	6 (6.4)	–	–	–
Disease status - n. (%)				
Primary	45 (48.4)	–	50 (100.0)	–
Recurrence	48 (51.6)	–	0 (0.0)	–
Tumor stage - n. (%)				
CIS[a]	12 (12.9)	–	–	–
pTa	51 (54.8)	–	18 (36.0)	–
pTa–CIS	5 (5.4)	–	–	–
pT1	6 (6.4)	–	14 (28.0)	–
pT1–CIS	10 (10.8)	–	–	–
> pT1	6 (6.5)	–	18 (36.0)	
> pT1–CIS	3 (3.2)	–	–	–
Tumor grade - n. (%)				
Low	38 (40.9)	–	12 (24.0)	–
High	55 (59.1)	–	38 (76.0)	–
Tumor risk score - n. (%)				
Low-risk NMIUC[b]	36 (38.7)	–	12 (24.0)	
HIgh-risk NMIUC[c]	48 (51.6)	–	20 (40.0)	
MIUC[d]	9 (9.7)	–	18 (36.0)	
Urine cytology - n. (%)				
Negative	29 (31.2)	–	8 (16.0)	–

Atypical	6 (6.5)	–	–	–
Low grade	3 (3.2)	–	–	–
High grade	28 (30.1)	–	8 (16.0)	–
Missing	27 (29.0)	–	34 (68.0)	–
Median DNA yield (range) - ng/ml				
US cfDNA[e]	5.0 (0.3–808.9)	6.2 (0.1–1073.9)	–	–
UP cellDNA[f]	55.8 (1.1–460.5)	30.93 (1.9–389.8)	–	–
Plasma cfDNA	20.4 (9.3–8833.3)	20.7 (9.3–4466.7)	–	–

[a] Carcinoma In Situ
[b] Low-risk Non Muscle Invasive Urothelial Carcinoma (pTa/pT1, low grade)
[c] High-risk Non Muscle Invasive Urothelial Carcinoma (pTa/pT1, high grade with any stage associated with CIS)
[d] Muscle Invasive Urothelial Carcinoma
[e] Urine Supernatant cell-free DNA
[f] Urine Pellet cellular DNA

Table 2. Performance of body fluid-based *TERT* promoter mutations in detecting UC

	DIAGURO Cohort				PORTO cohort
	US cfDNA or UP cellDNA (N=187)	US cfDNA (N= 176)	UP cellDNA (N= 184)	Plasma cfDNA (N=148)	UP cellDNA (N=100)
C228T or C250T					
True Positive - no	81	72	76	5	33
True Negative - no	89	86	88	77	50
False positive - no	5	2	5	1	0
False negative - no	12	16	15	65	17
No data - no	0	11	3	39	0
Sensitivity (95% CI) - %	87.1 (78.6 - 93.2)	81.8 (72.2 - 89.2)	83.5 (74.27 - 90.47)	7.1 (2.4 -16.0)	66.0 (51.2 - 78.8)
Specificity (95% CI) - %	94.7 (88.0 - 98.3)	97.7 (92.0 - 99.7)	94.6 (87.9 - 98.2)	98.7 (93.1 - 100.0)	100.0 (92.9 - 100.0)
Positive likelihood ratio (95% CI) - %	16.4 (7.0 - 38.6)	36.0 (9.1 - 142.2)	15.5 (6.6 - 36.6)	5.6 (0.67 - 46.54)	-
Negative likelihood ratio (95% CI) - %	0.1 (0.1 - 0.2)	0.2 (0.1 - 0.3)	0.2 (0.1 - 0.3)	0.9 (0.9 - 1.0)	0.34 (0.2 - 0.5)
Positive predictive value* (95% CI) - %	87.6 (83.7 - 90.6)	93.9 (90.4 - 96.1)	86.6 (82.8 - 90.0)	70.0 (56.0 - 83.4)	100
Negative predictive value* (95% CI) - %	94.4 (92.7 - 95.8)	92.6 (90.7 - 94.1)	93.1 (91.3 - 94.6)	71.2 (70.6 - 72.0)	87.3 (85.4 - 89.0)
Accuracy* (95% CI) - %	92.4 (90.6 - 94.0)	92.9 (91.1 - 94.4)	91.3 (89.4 - 93.0)	71.2 (68.3 - 74.0)	89.8 (87.8 - 91.6)
All TERT mutations					
True Positive	81	72	77	5	34
True Negative	88	85	88	77	49
False positive	6	3	5	1	1
False negative	12	16	14	65	16
No data - no	0	11	3	39	0
Sensitivity (95% CI) - %	87.1 (78.6 - 93.2)	80.7 (70.9 - 88.3)	84.6 (75.5 - 91.3)	7.1 (2.4 -16.0)	68.0 (53.3 - 80.5)
Specificity (95% CI) - %	93.6 (86.6 - 97.6)	96.6 (90.4 - 99.3)	94.6 (87.9 - 98.2)	98.7 (93.1 - 100.0)	98.0 (89.3 - 100.0)
Positive likelihood ratio (95% CI) - %	13.7 (6.3 - 29.7)	24.0 (7.9 - 73.3)	15.7 (6.7 - 37.1)	5.6 (0.67 - 46.54)	34.0 (4.8 - 238.9)
Negative likelihood ratio (95% CI) - %	0.1 (0.1 - 0.2)	0.2 (0.1 - 0.3)	0.2 (0.1 - 0.3)	0.9 (0.9 - 1.0)	0.3 (0.2 - 0.5)

Positive predictive value* (95% CI) - %	85.3 (81.3 - 88.6)	91.1 (87.3 - 93.8)	87.0 (83.0 - 90.1)	70.0 (56.0 - 83.4)	97.1 (82.9 - 99.6)
Negative predictive value* (95% CI) - %	94.4 (92.6 - 95.8)	92.5 (90.6 - 94.0)	93.5 (91.7 - 95.0)	71.2 (70.6 - 72.0)	75.4 (67.1 - 82.1)
Accuracy* (95% CI) - %	91.6 (89.7 - 93.2)	92.1 (90.2 - 93.7)	91.6 (89.8 - 93.2)	71.2 (68.3 - 74.0)	83.0 (74.2 - 89.8)

US cfDNA: Urine Supernatant cell-free DNA
UP cellDNA : Urine Pellet cellular DNA
* Positive and negative predictive values were calculated for patients at high risk of developing bladder cancer, estimated at 30% for patients with hematuria or, patients with lower urinary tract symptoms or others according to Springer and colleagues.[12]
No data denotes samples that were run with the UroMuTERT assay at least twice with two independent amplification reactions and for which no sequencing reads were obtained.

Fig.1
Click here to download high resolution image

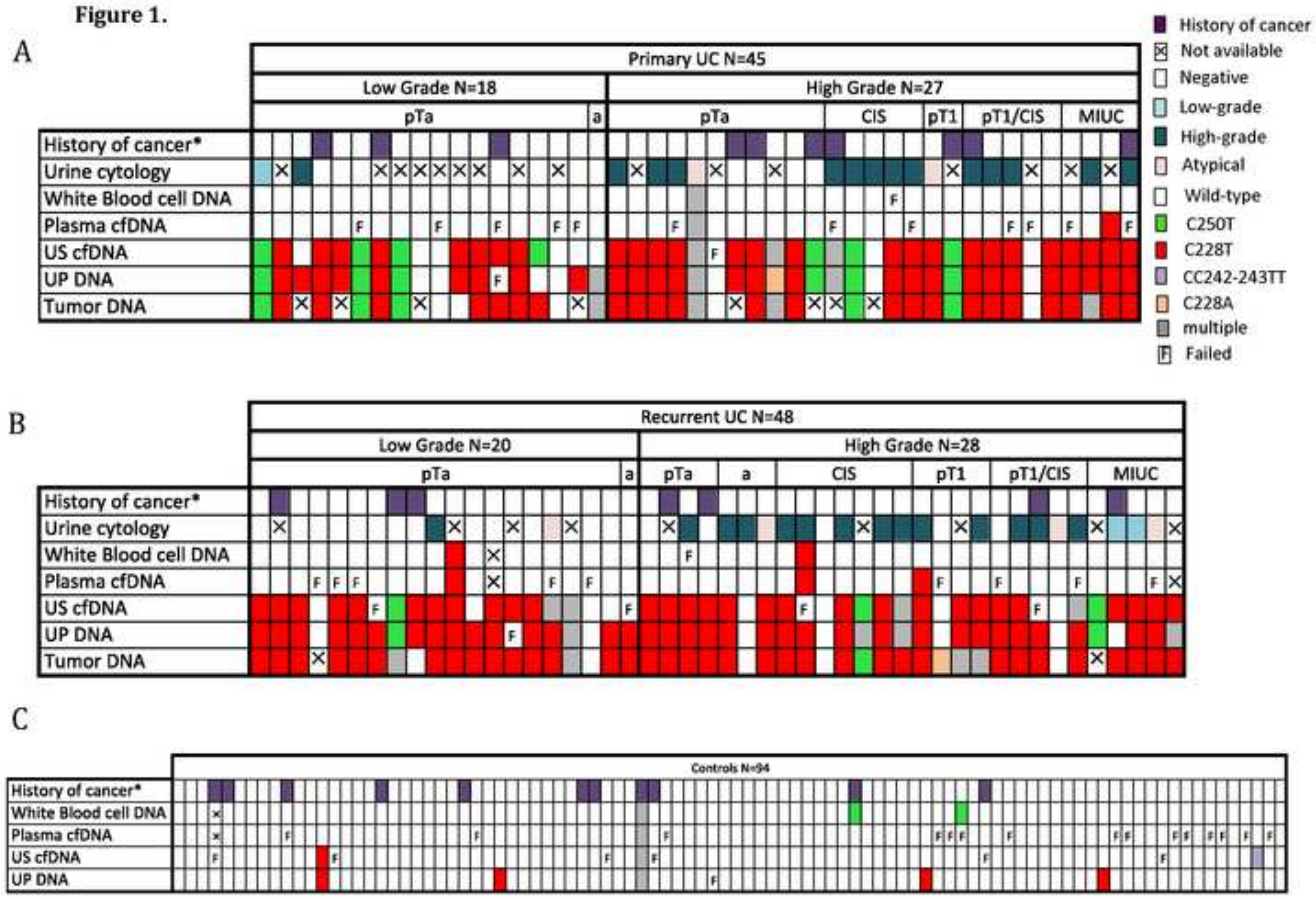


Fig.2
Click here to download high resolution image

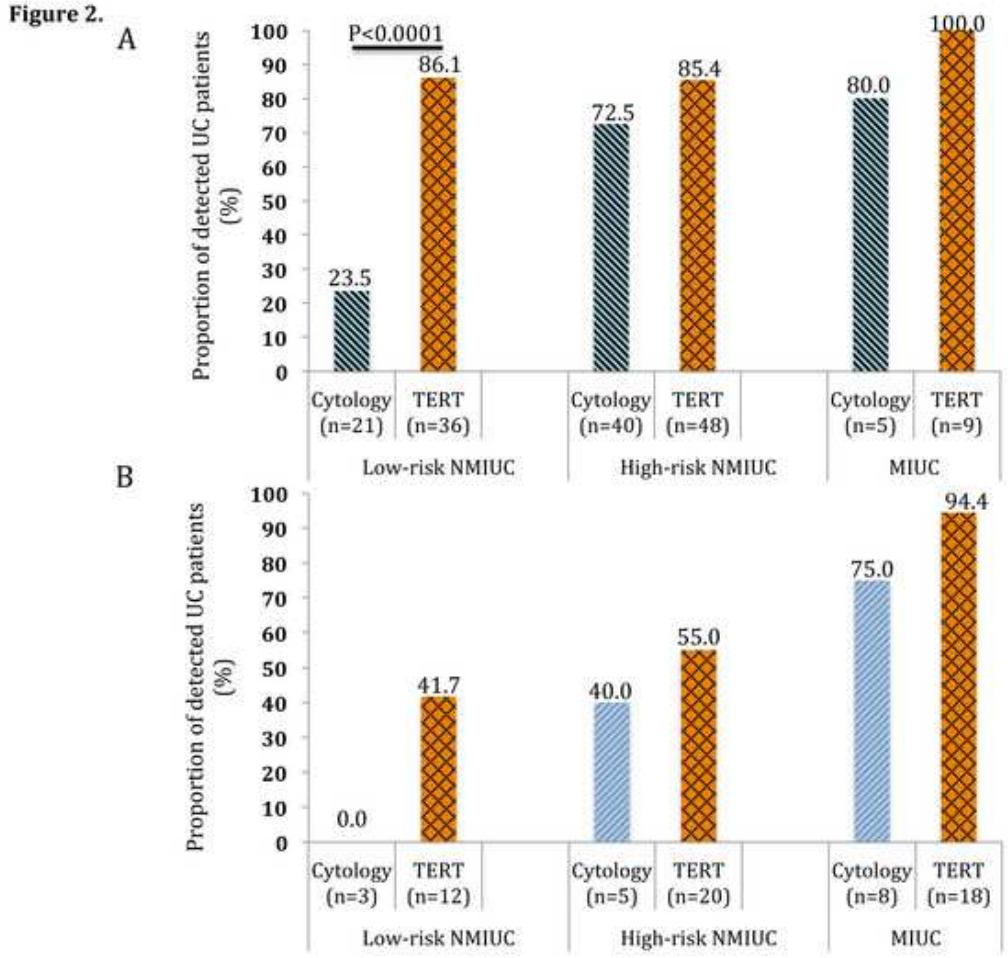
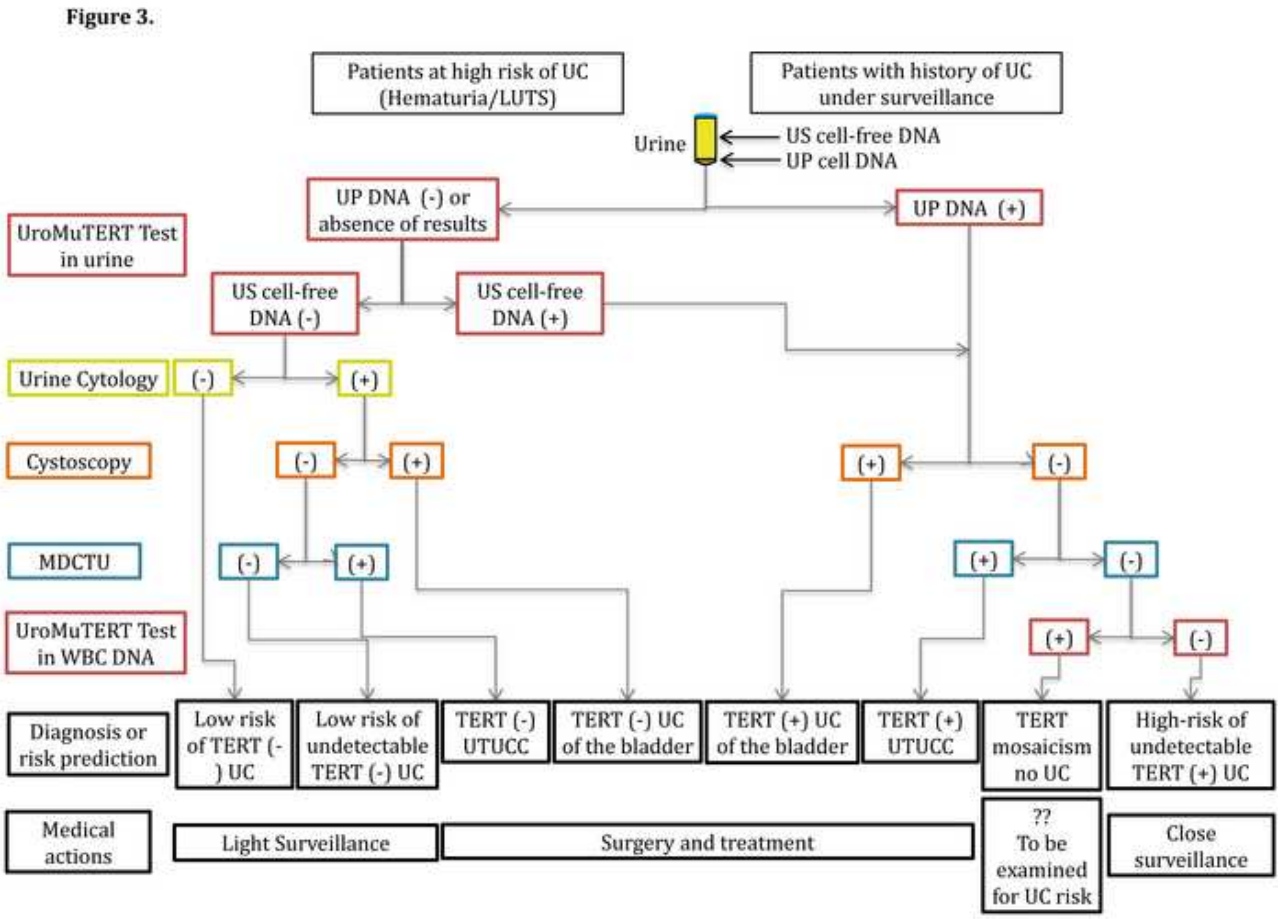


Fig.3
Click here to download high resolution image



Figures Legend

Figure 1. Overview of the detection of *TERT* promoter mutations by the UroMuTERT assay applied to body fluids and tumors of DIAGURO primary and recurrent UC cases and body fluids of controls.
UC denotes Urothelial Carcinoma; US cfDNA denotes Urine Supernatant cell-free DNA; UP DNA denotes Urine Pellet DNA; CIS denotes Carcinoma in situ; MIUC denotes Muscle-Invasive urothelial carcinoma and a stands for pTa/CIS
* other than UC

Figure 2. Performance of cytology and urinary *TERT* promoter mutations in detecting various risk categories of UC in the DIAGURO (A) and the PORTO (B) cohorts. Tumors are categorized in three groups: low-risk non–muscle- invasive urothelial cancer (NMIUC) (pTa/pT1, low grade), high-risk NMIUC (pTa/pT1 high grade and any stage associated with CIS), and muscle- invasive urothelial MIUC (pT2, pT3 or pT4). Risk classification of NMIUC was adapted from Millan-Rodriguez and colleagues. [18]

Figure 3. Proposed strategy integrating urinary *TERT* mutations analysis to current medical standards for the management of UC of the bladder and of the upper urinary tract
UC denotes Urothelial Cancer; UTUCC denotes Upper tract urothelial cell carcinoma; LUTS denotes Lower urinary tract symptoms; US denotes Urine Supernatant; UP denotes Urine Pellet; WBC denotes White Blood Cells; MDCTU denotes Multidetector Computed Tomographic Urography

4.4 Scientific contribution C

In this study we were interested in the identification of *TP53* mutations in the cfDNA of SCLC cancer cases for early detection. Compared to the previous studies that were focused on particular DNA positions (recurrently mutated positions in these types of cancer), here we targeted a whole gene, the *TP53* gene, because it has been reported that the majority of SCLC case tumors harbor at least one deleterious mutation in these gene [56]. In addition, because mutant *p53* proteins (due to deleterious mutations of the *TP53* gene) both lose their tumor suppressive role and can gain oncogenic functions that provide survival advantage to cells [110], non-cancer patients are not expected to present *TP53* deleterious mutations.

Here we estimated the presence of *TP53* mutations in the cfDNA of plasma samples from 51 SCLC cases and 123 non-cancer controls using a deep sequencing amplicon-based approach (IonTorrent Proton sequencing technology). For this, we used our needlestack algorithm to detect candidate mutations and we used basic thresholds on variant statistic in order to increase our specificity, because, contrary to previously presented studies, here we did not screened only a few positions but an entire gene. We filtered variants found to be in strand bias (corresponding to a RVSB>0.85) and kept only meaningful deleterious mutations, *i.e.*, found in the COSMIC database [52] and is a nonsense, indel, splicing or missense variant that is classified as deleterious by SIFT [102] or Polyphen [5]. Finally, we detected mutations in 49% of SCLC cases (in 5.7% of early-stage cases and in 54.1% of late-stage cases), with the lowest VAF detected around 0.1%. Interestingly, 11.4% of non-cancer controls harbored deleterious *TP53* mutations in their plasma cfDNA, and this result was replicated in an independent cohort (10.8% of a total of 102 non-cancer controls). This suggests that screening the *TP53* cfDNA mutations in order to develop a cancer biomarker presents challenges in term of biomarker specificity.

4.4.1 Article E

EBioMedicine 10 (2016) 117–123

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer

Lynnette Fernandez-Cuesta [a,1], Sandra Perdomo [a,b,1], Patrice H. Avogbe [a,1], Noemie Leblay [a], Tiffany M. Delhomme [a], Valerie Gaborieau [a], Behnoush Abedi-Ardekani [a], Estelle Chanudet [a], Magali Olivier [a], David Zaridze [c], Anush Mukeria [c], Marta Vilensky [d], Ivana Holcatova [e], Jerry Polesel [f], Lorenzo Simonato [g], Cristina Canova [g], Pagona Lagiou [h], Christian Brambilla [i], Elisabeth Brambilla [i], Graham Byrnes [a], Ghislaine Scelo [a], Florence Le Calvez-Kelm [a], Matthieu Foll [a], James D. McKay [a,*], Paul Brennan [a,*]

[a] International Agency for Research on Cancer (IARC-WHO), 150 cours Albert Thomas, 69008 Lyons, France
[b] Institute of Nutrition, Genetics and Metabolism Research, Universidad El Bosque, Bogotá, Colombia
[c] Russian N.N. Blokhin Cancer Research Centre, Moscow, Russian Federation
[d] Instituto Angel Roffo, Buenos Aires, Argentina,
[e] 1st Faculty of Medicine, Charles University, Prague, Czech Republic
[f] CRO Aviano National Cancer Institute, Aviano, Italy
[g] Department of Molecular Medicine, University of Padova, Padova, Italy,
[h] University of Athens Medical School, Greece
[i] CHU Grenoble, University Grenoble- Alpes, INSERM U823, Grenoble, France

ARTICLE INFO

Article history:
Received 25 May 2016
Received in revised form 20 June 2016
Accepted 23 June 2016
Available online 25 June 2016

Keywords:
ctDNA
cfDNA
Small-cell lung cancer
TP53 mutations
Early detection
Screening

ABSTRACT

Circulating tumor DNA (ctDNA) is emerging as a key potential biomarker for post-diagnosis surveillance but it may also play a crucial role in the detection of pre-clinical cancer. Small-cell lung cancer (SCLC) is an excellent candidate for early detection given there are no successful therapeutic options for late-stage disease, and it displays almost universal inactivation of *TP53*. We assessed the presence of *TP53* mutations in the cell-free DNA (cfDNA) extracted from the plasma of 51 SCLC cases and 123 non-cancer controls. We identified mutations using a pipeline specifically designed to accurately detect variants at very low fractions. We detected *TP53* mutations in the cfDNA of 49% SCLC patients and 11.4% of non-cancer controls. When stratifying the 51 initial SCLC cases by stage, *TP53* mutations were detected in the cfDNA of 35.7% early-stage and 54.1% late-stage SCLC patients. The results in the controls were further replicated in 10.8% of an independent series of 102 non-cancer controls. The detection of *TP53* mutations in 11% of the 225 non-cancer controls suggests that somatic mutations in cfDNA among individuals without any cancer diagnosis is a common occurrence, and poses serious challenges for the development of ctDNA screening tests.

1. Introduction

Cell-free DNA (cfDNA) refers to nucleic acids detected in body fluids and are thought to arise from two sources: passive release through cell death (Jahr et al., 2001), and active release by cell secretion (Stroun et al., 2000). DNA from cancer cells also contributes to the total load of cfDNA (Schwarzenbach et al., 2011), and the fraction of cfDNA that comes from cancer cells is called circulating-tumor DNA (ctDNA). ctDNA has been estimated to make up about 0.01%–1% of cfDNA for early-stage disease, reaching over 40% for late-stage disease (Beaver et al., 2014; Bettegowda et al., 2014; Couraud et al., 2014; Diehl et al., 2007; Forshew et al., 2012; Newman et al., 2014; Sausen et al., 2015). Despite its intrinsic limitations, including technical issues in the sample collection, detection, and identification of tumor origin, ctDNA is emerging as a key potential biomarker for monitoring response to treatment and relapse (Dawson et al., 2013; Esposito et al., 2014; Forshew et al., 2012; Garcia-Murillas et al., 2015; Murtaza et al., 2013; Roschewski et al., 2015; Siravegna et al., 2015). The potential of ctDNA is not limited to post-diagnosis surveillance but it may also play a crucial role in the detection of pre-clinical cancer. If successful, this could be translated

* Corresponding authors.
E-mail addresses: MckayJ@iarc.fr (J.D. McKay), BrennanP@iarc.fr (P. Brennan).
[1] Equally contributing authors.

http://dx.doi.org/10.1016/j.ebiom.2016.06.032

into much improved cancer survival, in particular for those cancer sites that are typically diagnosed at a late stage, and for which survival is poor, such as lung, pancreatic, or esophageal cancer (Brennan and Wild, 2015). However, implementation of ctDNA tests that detect pre-clinical disease in a non-symptomatic population will have to show an extremely high specificity if they are to provide meaningful results, or be part of a multi-modal screening program.

Very few studies have focused on the evaluation of ctDNA detection in early-stage cancers (*i.e.* stage I-II tumors) with even less data available on the detection of ctDNA in blood samples from pre-symptomatic cancer patients (Amant et al., 2015; Beaver et al., 2014; Bettegowda et al., 2014; Garcia-Murillas et al., 2015; Gormally et al., 2006; Jamal-Hanjani et al., 2016; Sausen et al., 2015); Table S1). In addition, these studies have aimed to detect specific mutations in cfDNA (most of them using digital droplet PCR) following previous assessment of the tumor mutational profile. This approach is only viable for cancers with common hot-spot mutations and is not amenable for most screening purposes. This is because early detection of pre-clinical cancer requires variant detection to be done without prior knowledge from tumor tissue of the expected mutations. Another limitation of these studies is the major assumption that circulating-mutated fragments would be absent (or very rare) in individuals without cancer. Demonstrating that any ctDNA detection marker has a specificity close to 100% would be of fundamental importance for large-scale utility in an asymptomatic population (Wentzensen and Wacholder, 2013).

Small-cell lung cancer (SCLC) accounts for about 15% of all lung tumors and has a 5-year survival below 5%. While SCLC tumors are initially sensitive to chemotherapy, they invariably relapse with a resistant and deadly disease. We and others have found that, contrary to lung adenocarcinomas and squamous-cell lung carcinomas, mutations in therapeutic targets are rare in SCLC (George et al., 2015; Peifer et al., 2012; Rudin et al., 2012). *TP53* is inactivated in virtually all SCLC cases, and *TP53* mutations are known to be an early event in the development of this disease. Given the almost uniform presence of *TP53* mutations in SCLC, we have investigated to what extent mutations in this gene can be identified in the cfDNA of patients with SCLC tumors. In addition, we have also assessed two independent series of non-cancer controls to evaluate the specificity of the approach.

2. Material and Methods

2.1. Study Population

Small-cell lung cancer (SCLC) patients and controls were recruited through an IARC case-control study coordinated in Moscow from 2006 to 2012 (Wozniak et al., 2015). Cases were incident cancer patients collected from the Russian N.N. Blokhin Cancer Research Centre and the Moscow City Clinical Oncology Dispensary. The staging of the SCLC cases is based on the recent recommendations of the International Association for the Study of Lung Cancer (IASLC) (Nicholson et al., 2016). Controls were recruited from individuals visiting two Moscow general hospitals for disorders unrelated to lung cancer and its associated risk factors (Table 1). The controls were matched for age, sex, and smoking status. All study participants provided written-informed consent and participated in an interview. Peripheral blood was collected in EDTA tubes at the time of interview and processed as rapidly as possible (generally within 2 h). For cases, blood draw was performed before surgery and any adjuvant treatment. Plasma samples were isolated by centrifugation of whole blood at 2000 ×*g* for 10 min at room temperature. Samples were stored at −80 °C. All specimens were obtained in accordance with the declaration of Helsinki guidelines and were approved by the local Institutional Review Board and the IARC Ethics Committee. A total of 52 SCLC cases and 165 controls were initially included but only 51 SCLC and 123 controls passed the sequencing QC criteria (see Sequencing Data Analyses, Annotation, and Filters), and were therefore included in down-stream analyses.

Table 1
Characteristics of small-cell lung cancer cases and controls from Russia, and additional replication controls from Greece, Czech Republic, Italy, and Argentina.

	Cases	Controls	Replication controls
Origin (country)			
Russia	51	123	
Greece			9 (8.8%)
Czech Republic			14 (13.7%)
Italy			40 (39.2%)
Argentina			39 (38.2%)
Total	51	123	102
Age at diagnosis			
<40	2 (3.9%)	2 (1.6%)	3 (2.9%)
40–49	4 (7.8%)	11 (8.9%)	15 (14.7%)
50–59	15 (29.4%)	42 (34.2%)	33 (32.4%)
60–69	22 (43.1%)	55 (44.7%)	33 (32.4%)
70+	8 (15.7%)	13 (10.6%)	18 (17.7%)
Sex			
Male	43 (84.3%)	107 (87.0%)	76 (74.5%)
Female	8 (15.7%)	16 (13.0%)	26 (25.5%)
Smoking status			
Never	5 (9.8%)	35 (28.4%)	34 (33.4%)
Former	6 (11.8%)	28 (22.8%)	25 (24.5%)
Current	40 (78.4%)	60 (48.8%)	43 (42.1%)
Alcohol status			
Never	30 (58.8%)	32 (26.0%)	16 (15.7%)
Former	4 (7.8%)	18 (14.6%)	14 (13.7%)
Current	17 (33.4%)	73 (59.4%)	72 (70.6%)
Tumor stage of cases			
I	7 (13.7%)		
II	7 (13.7%)		
III	28 (54.9%)		
IV	9 (17.6%)		
Disease type of hospital controls			
Infectious & parasitic diseases		0 (0.0%)	1 (1.0%)
Neoplasms		3 (2.4%)	0 (0.0%)
Endocrine, nutritional and metabolic diseases and immunity disorders		6 (4.9%)	1 (1.0%)
Diseases of blood and blood-forming organs		2 (1.6%)	1 (1.0%)
Diseases of the nervous system and sense organs		28 (22.8%)	6 (5.9%)
Diseases of the sense organs		0 (0.0%)	5 (4.9%)
Diseases of the circulatory system		25 (20.3%)	2 (2.0%)
Diseases of the respiratory system		3 (2.4%)	3 (2.9%)
Diseases of the digestive system		19 (15.4%)	19 (18.6%)
Diseases of the genitourinary system		19 (15.4%)	17 (16.7%)
Diseases of the skin and subcutaneous tissue		3 (2.4%)	2 (2.0%)
Diseases of the musculoskeletal system and connective tissue		9 (7.3%)	18 (17.6%)
Symptoms, signs and ill-defined conditions		0 (0.0%)	8 (7.8%)
Injury and poisoning		6 (4.9%)	18 (17.6%)
External causes		0 (0.0%)	1 (1.0%)

In order to further evaluate the prevalence of circulating-mutated fragments in non-cancer controls, 114 additional controls were retrieved from two large multicenter case-control studies coordinated by IARC. One was a study on alcohol-related cancers and genetic susceptibility in Europe (the ARCAGE study) that was conducted from 2002 to 2005 and from which we included hospital-based controls recruited in Prague (Czech Republic), Athens (Greece), Aviano, Padova and Turin (Italy). The second was the Latin American study of head and neck cancer conducted from 1998 to 2002 and from which we selected hospital-based controls from two institutions in Buenos Aires (Argentina). Additional details of the 2 large multicenter case-control studies are included elsewhere (Lagiou et al., 2009; Ribeiro et al., 2011). Out of the 114 controls initially included, 102 passed the sequencing QC criteria (see Sequencing Data Analyses, Annotation, and Filters), and were therefore included in down-stream analyses.

2.2. cfDNA Extraction

cfDNA was extracted from 0.8–1.3 mL of plasma using the QIAamp DNA Circulating Nucleic Acid kit (Qiagen) following manufacturer's instructions. cfDNA was eluted into 100 μL of elution buffer and quantified with the Qubit DNA high-sensitivity assay kit (Invitrogen Corporation). Details regarding amount of cfDNA are included in the Table S2.

2.3. Primer Design and Amplification of Targets

Twenty-one amplicons of 150 bp in size were designed (Eurofins Genomics Ebersberg, Germany) to cover exons 2 to 10 of *TP53*. The GeneRead DNAseq Panel PCR Kit V2 (Qiagen) was used for target enrichment. A validated in-house protocol was used to set up multiplex PCRs in 10 μL reaction volume, containing 5 ng cfDNA, 60 nM of primer pool and 0.73 μL of HotStarTaq enzyme. The experiments were carried out in two physically isolated laboratory spaces: one for sample preparation and another one for post-amplification steps. Amplification was carried out in a 96-well format plates DNA engine Tetrad 2 Peltier Thermal Cycler (BIORAD) as follows: 15 min at 95 °C and 30 cycles of 15 s at 95 °C and 2 min at 60 °C and 10 min at 72 °C.

2.4. Library Preparation and Sequencing

Following target enrichment, PCR products were purified using Serapure beads (prepared in-house and produced by ThermoFisher Scientific Inc.). A ratio of 2:1 of Serapure beads to PCR products was used. Purified amplicons were quantified with the Qubit DNA high-sensitivity assay kit (Invitrogen Corporation). Library preparation was done using 150 ng of purified PCR products and the NEBNext Fast DNA Library Prep Set for Ion Torrent (New England Biolabs, Ipswich, MA, USA) following manufacturer's instructions. Amplicons were end-repaired and ligated to the specific adapters and individual barcodes (designed in-house and produced by Eurofins MWG Operon, Ebersberg, Germany). Libraries were cleaned up, amplified with a final step of 6 PCR cycles, pooled in an equimolar way and loaded onto a 2% agarose gel and subjected to an electrophoresis at 150 V for 1.5 h. Fragments of 180–220 bp in size were selected and the pooled DNA library was recovered from the gel using the QIAquik Gel Extraction kit (Qiagen). The quality and quantity of the library was then assessed on the Agilent 2100 Bioanalyzer on-chip electrophoresis (Agilent Technologies, USA) for the absence of possible adapter dimers and heterodimers. The pooled libraries were sequenced on the Ion Torrent™ Proton Sequencer (Life Technologies Corp., USA) aiming for deep coverage (5,000×), using the Ion PI™ Hi-Q™ OT2 200 Kit and the Ion PI™ Hi-Q™ Sequencing 200 Kit with the Ion PI chip V3 (Life Technologies Corp., USA), following the manufacture's protocols.

2.5. Technical Duplication

Technical duplicates were undertaken for each sample including amplification, library preparation, and sequencing. Each technical duplicate pair was assessed on separate plates to limit the possibility of a contamination.

2.6. Sequencing Data Analyses, Annotation, and Filters

Short reads were aligned to the hg19 human reference genome and BAM files were generated using the Torrent Suite software (v4.4.2) with default parameters. Reads with a mapping quality below 20 were excluded from subsequent analysis. We also excluded those libraries for which the on-target median coverage was significantly lower in comparison to the other libraries sequenced in the same batch. On-target median coverage for both libraries is shown in Table S2.

For the calling of variants we used *Needlestack*, a recently developed ultra-sensitive variant caller, which estimates the distribution of sequencing errors across multiple samples to reliably identify variants present in very low proportion (https://github.com/IARCbioinfo/needlestack) (unpublished data; Delhomme et al.). Contrary to most existing algorithms, *Needlestack* can deal with both single nucleotide substitutions (SNVs) and short indels. At each position and for each candidate variant, sequencing errors are modeled using a robust negative binomial regression (Aeberhard et al., 2014), with a linear link and a zero intercept. True variants are outliers from this error model (Fig. 1a). The robust estimator of the over-dispersion parameter avoids bias due to these outliers (Aeberhard et al., 2014). For each sample a *p*-value against the null hypothesis of being a sequencing error is calculated, and further transformed into a *q*-value using the Benjamini and Hochberg false-discovery rate control method (Benjamini and Yosef, 1995). *Q*-values are reported as a Phred-scale quality score: $Q = -10\log_{10}(q\text{-value})$, and we used a threshold of $Q > 50$ to call variants. For each variant, we also calculated the relative variant strand bias defined by:

$$RVSB = \frac{max(AO_pDP_m, AO_mDP_p)}{AO_pDP_m + AO_mDP_{p.}}$$

where *DP* and *AO* denote respectively the total number of reads and the number of reads matching the candidate variant, with the subscripts *p* and *m* referring to the forward and reverse strands respectively. In the complete absence of strand bias, $RVSB = 0.5$ and $AO_p/AO_m = DP_p/DP_m$, whereas for a completely biased variant, $RVSB = 1$. We filtered out variants with $RVSB > 0.85$.

Variant calls were annotated using ANNOVAR (Yang and Wang, 2015). We only considered meaningful *TP53* mutations those that matched the following criteria: the mutation has already been reported in COSMIC and the mutation is a nonsense, indel, splicing or a missense variant that is classified as deleterious by SIFT or Polyphen.

2.7. Analyses of Technical Duplicates

While Needlestack models recurring errors, rare errors such as those generated by the DNA polymerase will be identified as variants. Such errors will be generally specific to a particular preparation. To filter out these rare errors, we required each of the individual's technical duplicate to be a Needlestack outlier (Fig. 1a). Additionally, we identified and excluded a few genomic positions that show a particularly high proportion (>10%) of these errors (*i.e.* higher than the estimated sequencing error rate, but not always replicable in two independent preparations).

2.8. Statistical Analyses

Effect of age, smoking, and alcohol status on the presence of *TP53* mutations was assessed using logistic regression adjusting one for the

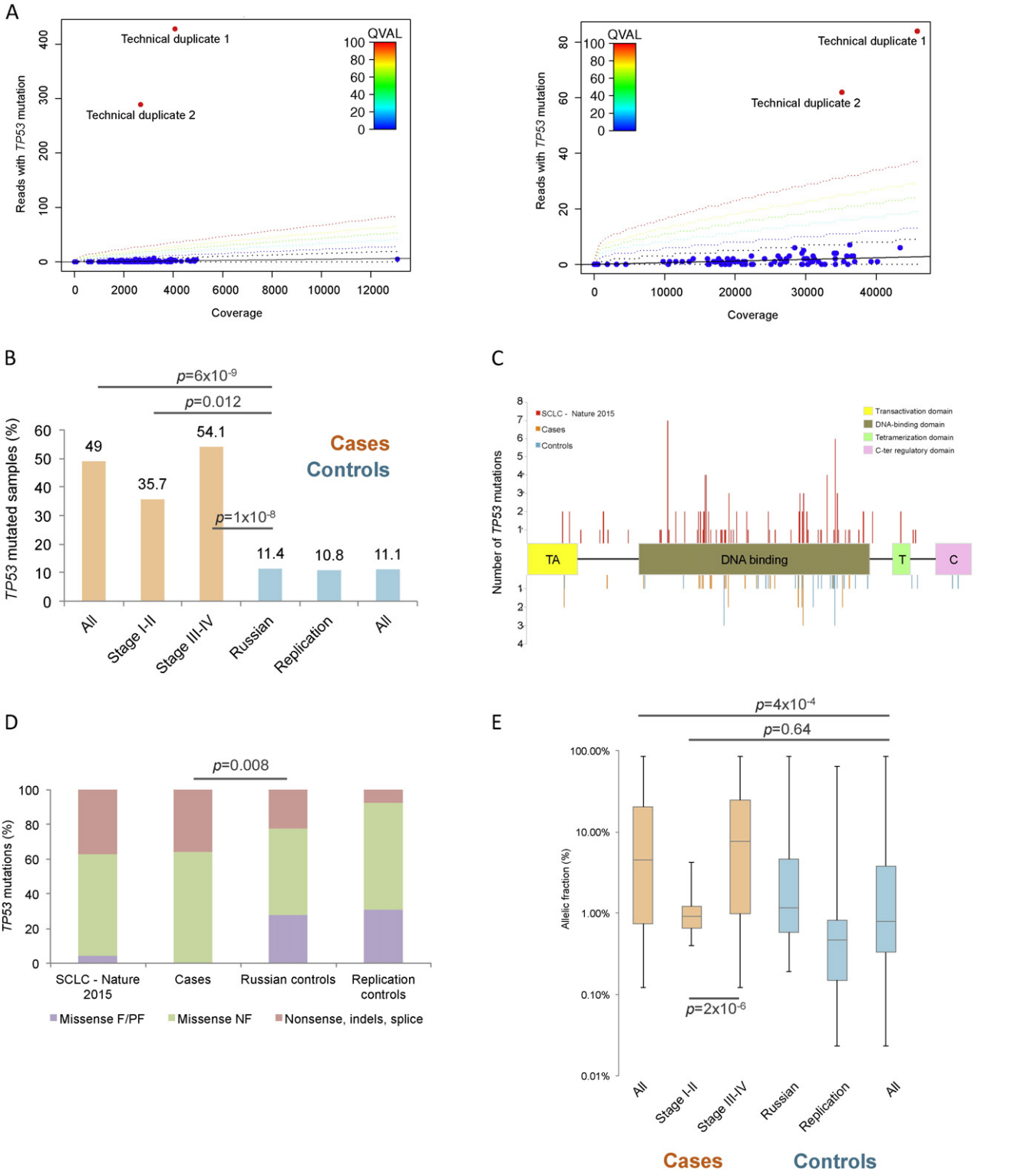


Fig. 1. Characteristics of *TP53* mutations in cases and controls (a) Two examples of variants called using Needlestack's regression model of sequencing error. Each dot represents a sequenced library (two dots per sample) colored according to its phred-scaled *q*-value. The black regression line shows the estimated sequencing-error rate along with the 99% confidence interval (black dotted lines) containing samples. Colored-dotted lines correspond to the limits of regions defined for different significance *q*-value thresholds. Both technical duplicates appear as outliers from the regression (in red), and are therefore classified as carrying the given mutation; (b) Percentage of *TP53* mutated samples in the cfDNA of Russian cases and controls, and replication controls; (c) Distribution of *TP53* mutations found in SCLC tumors (George et al., 2015) and in our series of cases and controls across the different p53 protein domains; (d) Type of mutations and functional impact of missense ones based on the IARC *TP53* database: F (functional), PF (partially functional), NF (non-functional); (e) Percentage of allelic fractions of the *TP53* mutations detected in the cfDNA of Russian cases and controls, and replication controls. The whiskers represent the minimum and maximum values.

others. Effect of case-status on the presence of *TP53* mutations was assessed using unadjusted logistic regression. *p*-Values to test the differences between the pattern of mutations in cases and controls are derived from Pearson's chi-square tests. All the analyses were conducted using SAS 9.4. For comparison of allelic fractions, we log-transformed the data and performed a *t*-test (2 tailed, unequal variance).

3. Results

The characteristics of the cases and controls are shown in Table 1. We detected 31 *TP53* mutations in 25 SCLC patients (49%, 25/51). When the 51 initial SCLC cases were stratified by stage, we found that 35.7% (5/14) of the stage I–II and 54.1% (20/37) of the stage III–IV, carried detectable *TP53* mutations in their cfDNA (Fig. 1b). While statistically significant in cases *versus* controls (p-value $= 6 \times 10^{-9}$), 18 *TP53* mutations were detected in 14 of the Russian non-cancer controls (11.4%, 14/123). The significance was also maintained when stratifying by stage (stage I–II *versus* controls, p-value $= 0.012$; stage III–IV *versus* controls, p-value $= 1 \times 10^{-8}$). We replicated these observations in an independent series of 102 controls, and found a comparable proportion of *TP53* mutated samples (10.8%, 13 *TP53* mutations in 11 controls).

Similarly to what is expected for *TP53* mutations present in cancer, most of the mutations in cases and controls altered amino acids coding for the *TP53* DNA-binding domain, which is critical for the transactivation activity of this gene (Fig. 1c). We next evaluated the characteristics of the mutations found in cases *versus* controls. Chi-square test analysis showed that there was a statistically significant difference between the mutational pattern found in cases *versus* controls (p-value $= 0.008$). The fraction of nonsense, indel, or splicing mutations found in the cases was similar to that previously reported for SCLC tumors (George et al., 2015) (35.5% *versus* 37% respectively), whereas this proportion was slightly lower in controls (22.2% in the Russian, and 7.7% in the replication controls; Fig. 1d). We used the IARC *TP53* database to classify the missense mutations in functional, partially functional, or non-functional based on the *in vitro* transcriptional activity of the resulting protein. Most missense mutations found in SCLC tumors (George et al., 2015) (92.6%) and cfDNA from cases (100%) were classified as resulting in a non-functional protein. However, controls had a higher proportion of missense mutations that retained some transcriptional activity (~30%; Fig. 1d).

We also compared the allelic fractions (AFs) of the *TP53* mutations found in the cfDNA of cases and controls. The AFs for a given mutation were similar in the two independent libraries, demonstrating the reproducibility of the assay (Table S3, S4, and S5). The AFs for the cases ranged from 0.12% to 84.81% (median 4.6%). In the Russian controls the AFs ranged from 0.19% to 84.94% (median 1.2%), and in the replication controls they ranged from 0.02% to 63.74% (median 0.5%) (Fig. 1e). The statistically significant difference in the AFs between cases and controls (p-value $= 4 \times 10^{-4}$) is explained by the presence of late-stage SCLC tumors, since the median AF of the *TP53* mutations detected in the five stage I–II SCLC (0.9%) is not statistically different from that found in controls (p-value $= 0.64$), while it differed from the median AF of stage III-IV SCLC tumors (8.2%; p-value $= 2 \times 10^{-6}$; Fig. 1e).

Finally, we sequenced the DNA extracted from the white-blood cells (WBC) of 39 cfDNA *TP53*-positive patients, from which material was available (19 cases and 20 controls). Five cfDNA *TP53* mutations (from one case and four controls) were detected in the WBC DNA, with similar AFs to those found in the cfDNA (Table 2). For one control (MLT-14), the AFs in both cfDNA and WBC DNA were around 50%, being consistent with a heterozygous germ-line variant. The other four mutations were detected at AFs consistent with a somatic origin (AFs below 11%) in both cfDNA and WBC DNA (Table S2).

Table 2
Overview of the cfDNA mutations also detected in the white-blood cells (WBC) DNA, and their corresponding allelic fractions in each technical duplicate (AFs in %).

Sample	*TP53* mutation	AFs detected in cfDNA		AFs detected in WBC	
SCLC-21	p.Y220C	0.90	1.27	0.50	0.70
MLT-6	p.R175G	4.09	4.41	4.40	4.50
MLT-14	p.G154S	47.17	50.58	52.10	54.90
ARG-1	p.R273C	5.22	5.58	7.30	10.40
ITA-8	p.V272M	0.78	0.80	0.90	1.40

Taken altogether, cancer-like *TP53* mutations were identified in 25 of the 225 non-cancer controls analyzed in this study (11.1%). We checked if the presence of *TP53* mutations in the controls was correlated with age, smoking status, or alcohol (adjusting one for the other), but none of these factors was found to be associated.

4. Discussion

Inactivation of *TP53* by mutation has been reported to occur in over 90% of SCLC cases (George et al., 2015). In this study we were able to detect *TP53* mutations in the cfDNA of 49% SCLC patients and, when stratifying by stage, in the cfDNA of 35.7% early-stage cases. These proportions matched those reported for other cancer types (Bettegowda et al., 2014). Unfortunately, we did not have the correspondent tumors to confirm that the *TP53* mutations detected in the cfDNA originate from the SCLC tumors. However, our method detected *TP53* mutations in 60% of the cfDNA samples of an independent French series of 10 SCLC patients (all of them carrying *TP53* somatic mutations in their tumors). Importantly, each of the *TP53* mutations found in the cfDNA matched the one found in the SCLC tumor (data not shown).

We also observed cfDNA *TP53*-mutated fragments in 11.4% of 123 matched non-cancer controls. Acknowledging the potential for bias in our selection of controls (such as differential performance in QC criteria or cfDNA amount, between cases and controls), we screened a second series of 102 non-cancer controls, and found a comparable proportion of *TP53* mutated samples in this independent group (13 *TP53* mutations in 11 controls, 10.8%). Altogether, the detection of *TP53* mutations in 11.1% of the 225 non-cancer controls, from two independent groups of samples, suggests that the presence of circulating-mutated fragments among individuals without any diagnosed cancer is a common occurrence, and poses serious challenges for the development of ctDNA screening tests for the early detection of cancer.

Only two other studies have explored the potential presence of circulating-mutated fragments in non-cancer subjects. A study within the EPIC prospective cohort (GENAIR) that used blood samples from controls, found that *KRAS* and *TP53* mutations were detectable in the cfDNA of 1% and 3.2% healthy subjects, respectively, without a cancer diagnoses five years subsequent to blood draw (Gormally et al., 2006). The higher percentage of *TP53*-positive controls in our analyses is likely to be explained by the fact that these analyses within EPIC were undertaken using DHPLC (denaturing high-pressure liquid chromatography) and Sanger sequencing, and these techniques are less sensitive and only allow for detection of mutations with allelic fractions of 3% or more. Further, only *TP53* exons five to nine were analyzed. If we limited our analysis to mutations from exons five to nine and AFs greater that 3%, we would have found a comparable number of *TP53*-positive controls (2.7%, 6/225). More recently, Krimmel and colleagues have reported extremely low-frequency cancer-like *TP53* mutations in the peritoneal fluid from both women with ovarian cancer and those with benign lesions, using duplex sequencing (Krimmel et al., 2016). They also showed that low frequency *TP53* mutagenesis increases with age and cancer. Overall, these results support the need for further ctDNA studies to incorporate series of non-cancer controls in order to improve validation of detection and analysis techniques.

A potential limitation of our study was the use of hospital controls as proxies of healthy people. Controls were admitted for a wide variety of routine conditions unrelated to tobacco and it is implausible that a high proportion of the controls with a detectable damaging *TP53* mutation developed a cancer in the short term. However, we cannot exclude this occurring in a small number of controls nor enriching for non-cancer diseases with unknown impact on the presence of circulating-tumor fragments. Nevertheless, as noted above, the prevalence of *TP53* mutations in our study is approximately equal to that of GENAIR (when applying the same detection thresholds). Prospective cohorts may help to overcome the limitations of using hospital controls and also help to

determine at what point in the development of the disease is ctDNA detectable in blood, and its correlation with a plausible diagnosis.

The source of circulating-mutated fragments in the cfDNA of apparently healthy people is still unknown. There is, however, accumulating evidence that clonal expansions are more frequent than originally thought. Martincorena and colleagues estimated that there are 9.5 clones per cm^2 of normal human skin carrying a driver mutation in *TP53* in a selected population for high-sun exposure (Martincorena and Campbell, 2015). Such clonal expansions might act as a reservoir of circulating-mutated fragments in cfDNA. In addition, several studies have shown that a subset of normal individuals could undergo clonal hematopoiesis with mutations in driver genes (Genovese et al., 2014; Jacobs et al., 2012; Jaiswal et al., 2014; Laurie et al., 2012; Wong et al., 2015; Xie et al., 2014). Consistent with this, we observed 4 cfDNA *TP53* mutations that appeared to be from clonal expansions in WBC. We also noted 2 *TP53* mutations in one SCLC case, apparently from different organs; one originating from WBC, the second we assume from the SCLC tumor. Such ambiguity around the tissue of origin of the circulating-mutated fragments adds another layer of complexity when using ctDNA for early detection.

The potential of ctDNA for early diagnosis of cancer is an area of much interest (The Lancet Oncology, 2016). While implementation in a screening setting will undoubtedly require more sensitive and specific tests as well as validation in pre-diagnostic blood samples, the unexpected presence of known cancer mutations in cfDNA among non-cancer controls represents an important challenge.

Funding Sources

This work has been funded by IARC. TMD was supported by la Ligue Nationale (Française) Contre le Cancer. The work reported in this paper was undertaken during the tenure of PHA's Postdoctoral Fellowship from the International Agency for Research on Cancer, partially supported by the European Commission FP7 Marie Curie Actions – People – Co-funding of regional, national and international programmes (COFUND).

Conflict of Interest

The authors declare no conflict of interests

Author Contributions

JDM and PB conceived and designed the study. GS coordinated the recruitment of samples and associated data. DZ, AM, MV, IH, JP, LS, CC, PL, CB, and EB provided samples and associated data. SP, PA, FLCK and JDM developed and undertook the laboratory procedures. SP, PA, and NL performed the experiments. TD, MF, and JDM developed Needlestack and performed sequencing analyses. LFC and SP analyzed and interpreted the data. VG and GB performed and provided statistical advice, respectively. EC, MO, and BAA provided intellectual input. LFC, SP, JDM, and PB wrote the manuscript. All authors reviewed and approved the final version for publication.

Acknowledgements

We acknowledge all the people who donated their biological specimens. We also acknowledge Priscilia Chopard, Helene Renard, Amelie Chabrier, Nathalie Forey, Geoffroy Durand, and Nicolas Lemaitre for their technical assistance.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ebiom.2016.06.032.

References

Aeberhard, W.H., Cantoni, E., Heritier, S., 2014. Robust inference in the negative binomial regression model with an application to falls data. Biometrics 70 (4), 920–931.

Amant, F., Verheecke, M., Wlodarska, I., Dehaspe, L., Brady, P., Brison, N., Van Den Bogaert, K., Dierickx, D., Vandecaveye, V., Tousseyn, T., Moerman, P., Vanderstichele, A., Vergote, I., Neven, P., Berteloot, P., Putseys, K., Danneels, L., Vandenberghe, P., Legius, E., Vermeesch, J.R., 2015. Presymptomatic identification of cancers in pregnant women during noninvasive prenatal testing. JAMA Oncol. 1 (6), 814–819.

Beaver, J.A., Jelovac, D., Balukrishna, S., Cochran, R.L., Croessmann, S., Zabransky, D.J., Wong, H.Y., Valda Toro, P., Cidado, J., Blair, B.G., Chu, D., Burns, T., Higgins, M.J., Stearns, V., Jacobs, L., Habibi, M., Lange, J., Hurley, P.J., Lauring, J., VanDenBerg, D.A., Kessler, J., Jeter, S., Samuels, M.L., Maar, D., Cope, L., Cimino-Mathews, A., Argani, P., Wolff, A.C., Park, B.H., 2014. Detection of cancer DNA in plasma of patients with early-stage breast cancer. Clin. Cancer Res. 20 (10), 2643–2650.

Benjamini, Y., Yosef, H., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57 (1).

Bettegowda, C., Sausen, M., Leary, R.J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B.R., Wang, H., Luber, B., Alani, R.M., Antonarakis, E.S., Azad, N.S., Bardelli, A., Brem, H., Cameron, J.L., Lee, C.C., Fecher, L.A., Gallia, G.L., Gibbs, P., Le, D., Giuntoli, R.L., Goggins, M., Hogarty, M.D., Holdhoff, M., Hong, S.M., Jiao, Y., Juhl, H.H., Kim, J.J., Siravegna, G., Laheru, D.A., Lauricella, C., Lim, M., Lipson, E.J., Marie, S.K.N., Netto, G.J., Oliner, K.S., Olivi, A., Olsson, L., Riggins, G.J., Sartore-Bianchi, A., Schmidt, K., Shih, I.M., Oba-Shinjo, S.M., Siena, S., Theodorescu, D., Tie, J., Harkins, T.T., Veronese, S., Wang, T.L., Weingart, J.D., Wolfgang, C.L., Wood, L.D., Xing, D., Hruban, R.H., Wu, J., Allen, P.J., Schmidt, C.M., Choti, M.A., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Papadopoulos, N., Diaz, L.A., 2014. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci. Transl. Med. 6 (224) (224ra24-224ra24).

Brennan, P., Wild, C.P., 2015. Genomics of cancer and a new era for cancer prevention. PLoS Genet. 11 (11), e1005522.

Couraud, S., Vaca-Paniagua, F., Villar, S., Oliver, J., Schuster, T., Blanche, H., Girard, N., Tredaniel, J., Guilleminault, L., Gervais, R., Prim, N., Vincent, M., Margery, J., Larive, S., Foucher, P., Duvert, B., Vallee, M., Le Calvez-Kelm, F., McKay, J., Missy, P., Morin, F., Zalcman, G., Olivier, M., Souquet, P.J., Bio, C.I.-i., 2014. Noninvasive diagnosis of actionable mutations by deep sequencing of circulating free DNA in lung cancer from never-smokers: a proof-of-concept study from BioCAST/IFCT-1002. Clin. Cancer Res. 20 (17), 4613–4624.

Dawson, S.-J., Tsui, D.W.Y., Murtaza, M., Biggs, H., Rueda, O.M., Chin, S.-F., Dunning, M.J., Gale, D., Forshew, T., Mahler-Araujo, B., Rajan, S., Humphray, S., Becq, J., Halsall, D., Wallis, M., Bentley, D., Caldas, C., Rosenfeld, N., 2013. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N. Engl. J. Med. 368 (13), 1199–1209.

Diehl, F., Schmidt, K., Choti, M.A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., Szabo, S.A., Kinzler, K.W., Vogelstein, B., Diaz Jr., L.A., 2007. Circulating mutant DNA to assess tumor dynamics. Nat. Med. 14 (9), 985–990.

Esposito, A., Bardelli, A., Criscitiello, C., Colombo, N., Gelao, L., Fumagalli, L., Minchella, I., Locatelli, M., Goldhirsch, A., Curigliano, G., 2014. Monitoring tumor-derived cell-free DNA in patients with solid tumors: clinical perspectives and research opportunities. Cancer Treat. Rev. 40 (5), 648–655.

Forshew, T., Murtaza, M., Parkinson, C., Gale, D., Tsui, D.W.Y., Kaper, F., Dawson, S.J., Piskorz, A.M., Jimenez-Linan, M., Bentley, D., Hadfield, J., May, A.P., Caldas, C., Brenton, J.D., Rosenfeld, N., 2012. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci. Transl. Med. 4 (136) (136ra68-136ra68).

Garcia-Murillas, I., Schiavon, G., Weigelt, B., Ng, C., Hrebien, S., Cutts, R.J., Cheang, M., Osin, P., Nerurkar, A., Kozarewa, I., Garrido, J.A., Dowsett, M., Reis-Filho, J.S., Smith, I.E., Turner, N.C., 2015. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci. Transl. Med. 7 (302) (302ra133).

Genovese, G., Kahler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., Purcell, S.M., Svantesson, O., Landen, M., Hoglund, M., Lehmann, S., Gabriel, S.B., Moran, J.L., Lander, E.S., Sullivan, P.F., Sklar, P., Gronberg, H., Hultman, C.M., McCarroll, S.A., 2014. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. 371 (26), 2477–2487.

George, J., Lim, J.S., Jang, S.J., Cun, Y., Ozretić, L., Kong, G., Leenders, F., Lu, X., Fernández-Cuesta, L., Bosco, G., Müller, C., Dahmen, I., Jahchan, N.S., Park, K.S., Yang, D., Karnezis, A.N., Vaka, D., Torres, A., Wang, M.S., Korbel, J.O., Menon, R., Chun, S.M., Kim, D., Wilkerson, M., Hayes, N., Engelmann, D., Pützer, B., Bos, M., Michels, S., Vlasic, I., Seidel, D., Pinther, B., Schaub, P., Becker, C., Altmüller, J., Yokota, J., Kohno, T., Iwakawa, R., Tsuta, K., Noguchi, M., Muley, T., Hoffmann, H., Schnabel, P.A., Petersen, I., Chen, Y., Soltermann, A., Tischler, V., Choi, C.M., Kim, Y.H., Massion, P.P., Zou, Y., Jovanovic, D., Kontic, M., Wright, G.M., Russell, P.A., Solomon, B., Koch, I., Lindner, M., Muscarella, L.A., la Torre, A., Field, J.K., Jakopovic, M., Knezevic, J., Castaños-Vélez, E., Roz, L., Pastorino, U., Brustugun, O.T., Lund-Iversen, M., Thunnissen, E., Köhler, J., Schuler, M., Botling, J., Sandelin, M., Sanchez-Cespedes, M., Salvesen, H.B., Achter, V., Lang, U., Bogus, M., Schneider, P.M., Zander, T., Ansén, S., Hallek, M., Wolf, J., Vingron, M., Yatabe, Y., Travis, W.D., Nürnberg, P., Reinhardt, C., Perner, S., Heukamp, L., Büttner, R., Haas, S.A., Brambilla, E., Peifer, M., Sage, J., Thomas, R.K., 2015. Comprehensive genomic profiles of small cell lung cancer. Nature 524 (7563), 47–53.

Gormally, E., Vineis, P., Matullo, G., Veglia, F., Caboux, E., Le Roux, E., Peluso, M., Garte, S., Guarrera, S., Munnia, A., Airoldi, L., Autrup, H., Malaveille, C., Dunning, A., Overvad, K., Tjonneland, A., Lund, E., Clavel-Chapelon, F., Boeing, H., Trichopoulou, A., Palli, D., Krogh, V., Tumino, R., Panico, S., Bueno-de-Mesquita, H.B., Peeters, P.H., Pera, G., Martinez, C., Dorronsoro, M., Barricarte, A., Navarro, C., Quiros, J.R., Hallmans, G., Day, N.E., Key, T.J., Saracci, R., Kaaks, R., Riboli, E., Hainaut, P., 2006. *TP53* and *KRAS2*

mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study. Cancer Res. 66 (13), 6871–6876.

Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.J., Cullen, M., Epstein, C.G., Burdett, L., Dean, M.C., Chatterjee, N., Sampson, J., Chung, C.C., Kovaks, J., Gapstur, S.M., Stevens, V.L., Teras, L.T., Gaudet, M.M., Albanes, D., Weinstein, S.J., Virtamo, J., Taylor, P.R., Freedman, N.D., Abnet, C.C., Goldstein, A.M., Hu, N., Yu, K., Yuan, J.M., Liao, L., Ding, T., Qiao, Y.L., Gao, Y.T., Koh, W.P., Xiang, Y.B., Tang, Z.Z., Fan, J.H., Aldrich, M.C., Amos, C., Blot, W.J., Bock, C.H., Gillanders, E.M., Harris, C.C., Haiman, C.A., Henderson, B.E., Kolonel, L.N., Le Marchand, L., McNeill, L.H., Rybicki, B.A., Schwartz, A.G., Signorello, L.B., Spitz, M.R., Wiencke, J.K., Wrensch, M., Wu, X., Zanetti, K.A., Ziegler, R.G., Figueroa, J.D., Garcia-Closas, M., Malats, N., Marenne, G., Prokunina-Olsson, L., Baris, D., Schwenn, M., Johnson, A., Landi, M.T., Goldin, L., Consonni, D., Bertazzi, P.A., Rotunno, M., Rajaraman, P., Andersson, U., Beane Freeman, L.E., Berg, C.D., Buring, J.E., Butler, M.A., Carreon, T., Feychting, M., Ahlbom, A., Gaziano, J.M., Giles, G.G., Hallmans, G., Hankinson, S.E., Hartge, P., Henriksson, R., Inskip, P.D., Johansen, C., Landgren, A., McKean-Cowdin, R., Michaud, D.S., Melin, B.S., Peters, U., Ruder, A.M., Sesso, H.D., Severi, G., Shu, X.O., Visvanathan, K., et al., 2012. Detectable clonal mosaicism and its relationship to aging and cancer. Nat. Genet. 44 (6), 651–658.

Jahr, S., Hentze, H., Englisch, S., Hardt, D., Fackelmayer, F.O., Hesch, R.D., Knippers, R., 2001. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res. 61 (4), 1659–1665.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., Higgins, J.M., Moltchanov, V., Kuo, F.C., Kluk, M.J., Henderson, B., Kinnunen, L., Koistinen, H.A., Ladenvall, C., Getz, G., Correa, A., Banahan, B.F., Gabriel, S., Kathiresan, S., Stringham, H.M., McCarthy, M.I., Boehnke, M., Tuomilehto, J., Haiman, C., Groop, L., Atzmon, G., Wilson, J.G., Neuberg, D., Altshuler, D., Ebert, B.L., 2014. Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. 371 (26), 2488–2498.

Jamal-Hanjani, M., Wilson, G.A., Horswell, S., Mitter, R., Sakarya, O., Constantin, T., Salari, R., Kirkizlar, E., Sigurjonsson, S., Pelham, R., Kareht, S., Zimmermann, B., Swanton, C., 2016. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. Ann. Oncol. 27 (5), 862–867.

Krimmel, J.D., Schmitt, M.W., Harrell, M.I., Agnew, K.J., Kennedy, S.R., Emond, M.J., Loeb, L.A., Swisher, E.M., Risques, R.A., 2016. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. Proc. Natl. Acad. Sci. U. S. A. 113 (21), 6005–6010.

Lagiou, P., Georgila, C., Minaki, P., Ahrens, W., Pohlabeln, H., Benhamou, S., Bouchardy, C., Slamova, A., Schejbalova, M., Merletti, F., Richiardi, L., Kjaerheim, K., Agudo, A., Castellsague, X., Macfarlane, T.V., Macfarlane, G.J., Talamini, R., Barzan, L., Canova, C., Simonato, L., Lowry, R., Conway, D.I., McKinney, P.A., Znaor, A., McCartan, B.E., Healy, C., Nelis, M., Metspalu, A., Marron, M., Hashibe, M., Brennan, P.J., 2009. Alcohol-related cancers and genetic susceptibility in Europe: the ARCAGE project: study samples and data collection. Eur. J. Cancer Prev. 18 (1), 76–84.

Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., Wei, Q., Wang, L.E., Lee, J.E., Barnes, K.C., Hansel, N.N., Mathias, R., Daley, D., Beaty, T.H., Scott, A.F., Ruczinski, I., Scharpf, R.B., Bierut, L.J., Hartz, S.M., Landi, M.T., Freedman, N.D., Goldin, L.R., Ginsburg, D., Li, J., Desch, K.C., Strom, S.S., Blot, W.J., Signorello, L.B., Ingles, S.A., Chanock, S.J., Berndt, S.I., Le Marchand, L., Henderson, B.E., Monroe, K.R., Heit, J.A., de Andrade, M., Armasu, S.M., Regnier, C., Lowe, W.L., Hayes, M.G., Marazita, M.L., Feingold, E., Murray, J.C., Melbye, M., Feenstra, B., Kang, J.H., Wiggs, J.L., Jarvik, G.P., McDavid, A.N., Seshan, V.E., Mirel, D.B., Crenshaw, A., Sharopova, N., Wise, A., Shen, J., Crosslin, D.R., Levine, D.M., Zheng, X., Udren, J.I., Bennett, S., Nelson, S.C., Gogarten, S.M., Conomos, M.P., Heagerty, P., Manolio, T., Pasquale, L.R., Haiman, C.A., Caporaso, N., Weir, B.S., 2012. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. 44 (6), 642–650.

Martincorena, I., Campbell, P.J., 2015. Somatic mutation in cancer and normal cells. Science 349 (6255), 1483–1489.

Murtaza, M., Dawson, S.-J., Tsui, D.W.Y., Gale, D., Forshew, T., Piskorz, A.M., Parkinson, C., Chin, S.-F., Kingsbury, Z., Wong, A.S.C., Marass, F., Humphray, S., Hadfield, J., Bentley, D., Chin, T.M., Brenton, J.D., Caldas, C., Rosenfeld, N., 2013. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature 497 (7447), 108–112.

Newman, A.M., Bratman, S.V., To, J., Wynne, J.F., Eclov, N.C.W., Modlin, L.A., Liu, C.L., Neal, J.W., Wakelee, H.A., Merritt, R.E., Shrager, J.B., Loo, B.W., Alizadeh, A.A., Diehn, M., 2014. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat. Med. 20 (5), 548–554.

Nicholson, A.G., Chansky, K., Crowley, J., Beyruti, R., Kubota, K., Turrisi, A., Eberhardt, W.E., van Meerbeeck, J., Rami-Porta, R., Staging and Prognostic Factors Committee, A. v. B., and Participating Institutions & Institutions, S. a. P. F. C. A. B. a. P., 2016. The international association for the study of lung cancer lung cancer staging project: proposals for the revision of the clinical and pathologic staging of small cell lung cancer in the forthcoming eighth edition of the TNM classification for lung cancer. J. Thorac. Oncol. 11 (3), 300–311.

Peifer, M., Fernández-Cuesta, L., Sos, M.L., George, J., Seidel, D., Kasper, L.H., Plenker, D., Leenders, F., Sun, R., Zander, T., Menon, R., Koker, M., Dahmen, I., Müller, C., Di Cerbo, V., Schildhaus, H.U., Altmüller, J., Baessmann, I., Becker, C., de Wilde, B., Vandesompele, J., Böhm, D., Ansén, S., Gabler, F., Wilkening, I., Heynck, S., Heuckmann, J.M., Lu, X., Carter, S.L., Cibulskis, K., Banerji, S., Getz, G., Park, K.S., Rauh, D., Grütter, C., Fischer, M., Pasqualucci, L., Wright, G., Wainer, Z., Russell, P., Petersen, I., Chen, Y., Stoelben, E., Ludwig, C., Schnabel, P., Hoffmann, H., Muley, T., Brockmann, M., Engel-Riedel, W., Muscarella, L.A., Fazio, V.M., Groen, H., Timens, W., Sietsma, H., Thunnissen, E., Smit, E., Heideman, D.A., Snijders, P.J., Cappuzzo, F., Ligorio, C., Damiani, S., Field, J., Solberg, S., Brustugun, O.T., Lund-Iversen, M., Sänger, J., Clement, J.H., Soltermann, A., Moch, H., Weder, W., Solomon, B., Soria, J.C., Validire, P., Besse, B., Brambilla, E., Brambilla, C., Lantuejoul, S., Lorimier, P., Schneider, P.M., Hallek, M., Pao, W., Meyerson, M., Sage, J., Shendure, J., Schneider, R., Büttner, R., Wolf, J., Nürnberg, P., Perner, S., Heukamp, L.C., Brindle, P.K., Haas, S., Thomas, R.K., 2012. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat. Genet. 44 (10), 1104–1110.

Ribeiro, K.B., Levi, J.E., Pawlita, M., Koifman, S., Matos, E., Eluf-Neto, J., Wunsch-Filho, V., Curado, M.P., Shangina, O., Zaridze, D., Szeszenia-Dabrowska, N., Lissowska, J., Daudt, A., Menezes, A., Bencko, V., Mates, D., Fernandez, L., Fabianova, E., Gheit, T., Tommasino, M., Boffetta, P., Brennan, P., Waterboer, T., 2011. Low human papillomavirus prevalence in head and neck cancer: results from two large case-control studies in high-incidence regions. Int. J. Epidemiol. 40 (2), 489–502.

Roschewski, M., Dunleavy, K., Pittaluga, S., Moorhead, M., Pepin, F., Kong, K., Shovlin, M., Jaffe, E.S., Staudt, L.M., Lai, C., Steinberg, S.M., Chen, C.C., Zheng, J., Willis, T.D., Faham, M., Wilson, W.H., 2015. Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B-cell lymphoma: a correlative biomarker study. Lancet Oncol. 16 (5), 541–549.

Rudin, C.M., Durinck, S., Stawiski, E.W., Poirier, J.T., Modrusan, Z., Shames, D.S., Bergbower, E.A., Guan, Y., Shin, J., Guillory, J., Rivers, C.S., Foo, C.K., Bhatt, D., Stinson, J., Gnad, F., Haverty, P.M., Gentleman, R., Chaudhuri, S., Janakiraman, V., Jaiswal, B.S., Parikh, C., Yuan, W., Zhang, Z., Koeppen, H., Wu, T.D., Stern, H.M., Yauch, R.L., Huffman, K.E., Paskulin, D.D., Illei, P.B., Varella-Garcia, M., Gazdar, A.F., de Sauvage, F.J., Bourgon, R., Minna, J.D., Brock, M.V., Seshagiri, S., 2012. Comprehensive genomic analysis identifies *SOX2* as a frequently amplified gene in small-cell lung cancer. Nat. Genet. 44 (10), 1111–1116.

Sausen, M., Phallen, J., Adleff, V., Jones, S., Leary, R.J., Barrett, M.T., Anagnostou, V., Parpart-Li, S., Murphy, D., Kay Li, Q., Hruban, C.A., Scharpf, R., White, J.R., O'Dwyer, P.J., Allen, P.J., Eshleman, J.R., Thompson, C.B., Klimstra, D.S., Linehan, D.C., Maitra, A., Hruban, R.H., Diaz Jr., L.A., Von Hoff, D.D., Johansen, J.S., Drebin, J.A., Velculescu, V.E., 2015. Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. Nat. Commun. 6, 7686.

Schwarzenbach, H., Hoon, D.S.B., Pantel, K., 2011. Cell-free nucleic acids as biomarkers in cancer patients. Nat. Rev. Cancer 11 (6), 426–437.

Siravegna, G., Mussolin, B., Buscarino, M., Corti, G., Cassingena, A., Crisafulli, G., Ponzetti, A., Cremolini, C., Amatu, A., Lauricella, C., Lamba, S., Hobor, S., Avallone, A., Valtorta, E., Rospo, G., Medico, E., Motta, V., Antoniotti, C., Tatangelo, F., Bellosillo, B., Veronese, S., Budillon, A., Montagut, C., Racca, P., Marsoni, S., Falcone, A., Corcoran, R.B., Di Nicolantonio, F., Loupakis, F., Siena, S., Sartore-Bianchi, A., Bardelli, A., 2015. Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. Nat. Med. 21 (7), 795–801.

Stroun, M., Maurice, P., Vasioukhin, V., Lyautey, J., Lederrey, C., Lefort, F., Rossier, A., Chen, X.Q., Anker, P., 2000. The origin and mechanism of circulating DNA. Ann. N. Y. Acad. Sci. 906, 161–168.

The Lancet Oncology, 2016. Liquid cancer biopsy: the future of cancer detection? Lancet Oncol. 17 (2), 123.

Wentzensen, N., Wacholder, S., 2013. From differences in means between cases and controls to risk stratification: a business plan for biomarker development. Cancer Discov. 3 (2), 148–157.

Wong, T.N., Ramsingh, G., Young, A.L., Miller, C.A., Touma, W., Welch, J.S., Lamprecht, T.L., Shen, D., Hundal, J., Fulton, R.S., Heath, S., Baty, J.D., Klco, J.M., Ding, L., Mardis, E.R., Westervelt, P., DiPersio, J.F., Walter, M.J., Graubert, T.A., Ley, T.J., Druley, T.E., Link, D.C., Wilson, R.K., 2015. Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. Nature 518 (7540), 552–555.

Wozniak, M.B., Scelo, G., Muller, D.C., Mukeria, A., Zaridze, D., Brennan, P., 2015. Circulating microRNAs as non-invasive biomarkers for early detection of non-small-cell lung cancer. PLoS One 10 (5), e0125026.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., Ozenberger, B.A., Welch, J.S., Link, D.C., Walter, M.J., Mardis, E.R., Dipersio, J.F., Chen, F., Wilson, R.K., Ley, T.J., Ding, L., 2014. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. 20 (12), 1472–1478.

Yang, H., Wang, K., 2015. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat. Protoc. 10 (10), 1556–1566.

4.5 Scientific contribution D

Finally, we were interested in the improvement of the approach presented in the section 4.4 (paper in preparation), in order to increase the specificity of the biomarker based on the *TP53* ctDNA mutations from plasma samples of SCLC cancer patients. We have then conducted a new study based on both the *TP53* gene and on the *RB1* gene (mutated in around 70% of SCLC tumors) to increase the specificity. We sequenced these genes from plasma samples of two independent cohorts: a first cohort composed of 50 cases and 183 controls, and a second cohort composed of 51 cases and 116 controls.

This study can be divided into two components:

- the accurate detection of mutations
- the development of a biomarker based on a genetic score

The accurate detection of mutations was presented in the chapter 3, and roughly was based on, firstly, the raw variant calling of mutations with needlestack, and, then, the boosting of the precision of mutation detection with the development and application of variant filtering methods, in order to reduce the potential presence of false calls (possibly in control samples). This accurate detection of mutations was finally validated using a simulation approach.

Here, we present the development of a ctDNA biomarker based on a genetic score (see picture 4.1). To build this genetic score, that is a per-sample statistic, and, that is then used to classify the samples into cases and controls, we used the mutations identified in the first part. We attributed a functional score to each mutation based on its deleterious power, as the following:

- 0.5 for synonymous or intronic variants
- 1 + REVEL (a score of the pathogenicity of a mutation) for missense variants
- 2 for stopgain, splicing or frameshift variants

Once each mutation obtained a functional score, we computed a sample-score by taking to the maximum of the functional score of mutations identified in the sample, independently

for the two genes. Then, we computed three logistic regression models based on the first cohort sample scores, using:

- only *TP53* sample-scores for the model *TP53*
- only *RB1* sample-scores for the model *RB1*
- the two TP53 and *RB1* sample-scores for the model *TP53* and *RB1*

We estimated, for the three models, the regression coefficients that maximize the correct attribution of the case/control status. We used these estimated regression coefficients in the second cohort to compute the ctDNA genetic scores, and then computed the accuracy of the three models (*i.e.* capacity to correctly attribute the case/control status to the samples), using *ROC* curves and their *AUC*s. We finally reported that, the addition of the *RB1* gene (third model), does not improve the performance of the biomarker using the ctDNA genetic scores compared to the usage of only the *TP53* gene (first model).

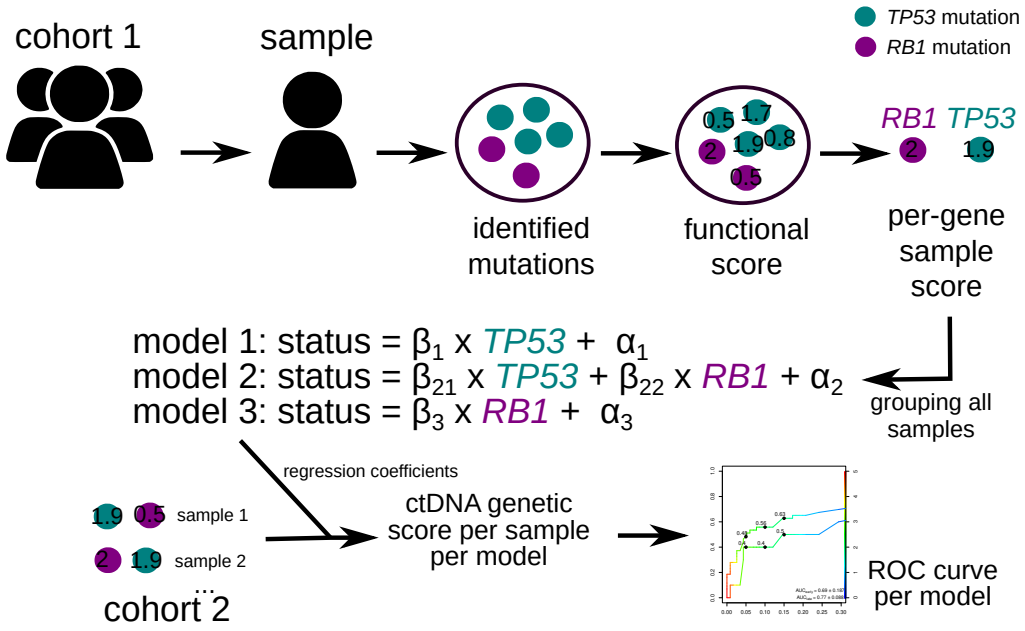


Figure 4.1 – Graphical visualization of the methodology used to developed our ctDNA biomarker, using genetic scores based on deleterious power of *TP53* and *RB1* mutations. Three models were explored, one based on the *TP53* gene alone, one based on the *RB1* gene alone, and one based on the combination of the two genes. Regression coefficient were estimated on the first cohort and applied on the second cohort to compute models accuracy (*i.e.* capacity to correctly attribute the case/control status to the samples).

4.5.1 Article B

See Chapter 3.

4.6 Discussion

The first study presented in this chapter was, at the time of publication, the largest screening of *KRAS* mutations in plasma samples of pancreatic cancer cases. The study has shown promising results in term of sensitivity and specificity of *KRAS* mutations as non-invasive biomarkers for the detection of pancreatic cancer, using a small amount of DNA, only 2 nanograms. Moreover, the lowest VAF detected was as low as 0.08%, suggesting a very good performance of our detection method. Nevertheless, a recent study has reported that half of the *KRAS* mutations found in cases harbored a a VAF lower than 0.08% [118], suggesting that our sensitivity could be improved if we could detect lower VAF, by increasing the coverage for example, or by using higher amount of input DNA to potentially increase the observed VAF.

In a second study, we developed UroMuTERT, a simple, non-invasive and sensitive assay for the detection of urothelial carcinomas using urine samples. Our study, based on the detection of *TERT* promoter mutations, has shown excellent sensitivity (87.1%) and specificity (94.7%). We reported a lowest VAF of 0.8% for the C228T mutation and a lowest VAF of 0.5% for the C250T mutation. Our performances were comparable to the recently developed UroSEEK assay (based on multiple genomic markers including C228T and C250T) [125].

Finally, we have conducted two studies aiming at developing a non-invasive SCLC detection biomarker. The first study was based on the identification of cfDNA *TP53* mutations from plasma samples. Nevertheless, a significant proportion of non-cancer patients were positive for cfDNA mutations, which was not expected (not reported in the litterature at this time, but later, other studies reported similar results [75]). We then decided to conduct a second study, with the objective of increasing the specificity of our biomarker. For this, firstly, we sequenced a second gene, *RB1*, that is recurrently mutated in SCLC tumors. Then, we developed a methodological framework based on efficient variant filtering in order to boost

the precision of the variant detection step. We also validated our results using simulations, to test if our variants can be explained by a random selection of errors. We developed a ctDNA genetic score in order to better discriminate between cases and controls than when using the raw mutations. Nevertheless, we reported a non-significant increase in sensitivity and specificity (ROC curve comparisons) when adding the *RB1* gene. This can be explained by the difference of mutations prevalences between the two cohorts, the first being used to estimate our regression coefficient that are then applied in the second. Indeed, we have estimated the accuracy of our biomarker when switching the two cohorts, *i.e.* when using the second cohort to estimate the parameters and applying them in the first cohort. In such a case, our biomarker harbor a better sensitivity and specificity when adding the *RB1* gene. This suggest a high instability of the results, that seem to be highly dependent of the data. This can be explained by the size of our cohort, the non-individually matched controls (we do not have one particular control matched to one particular case), and possible environmental differences between the two cohorts. Replicated studies can help to understand the lack of stability of our results.

Global discussion

“ We can only see a short distance ahead, but we can see plenty there that needs to be done ”

Alan Turing

In this thesis we have presented two computational approaches in order to efficiently identify mutations in NGS data, using the sytematic nature of errors to propose accurate ways of modelling them. In a first chapter, we have presented needlestack, a sensitive multi-sample variant caller. In a second chapter, we have presented two different methodologies of variant filtering. Finally, in a third and last chapter, we have presented the application of our variant calling and variant filtering methods to ctDNA data in order to develop an early cancer detection biomarker. Particular points of discussion have been handled at the end of each chapter. Here in this section we propose to provide a more global discussion around the thesis about inherent limitations and motivating perspectives.

Alignment and reference genome

Each of the developed methodologies as well as the biological application presented in the thesis are relying on a crucial step inherent to the NGS technology: the alignment of raw sequencing reads to a reference genome [147]. Because our methods are based on NGS data, they also are dependent on this alignment step: each identify variation is actually defined by "a variant observation compared to the reference genome". We can discuss the limitations of two components of the alignment step: the reference genome and the alignment algorithm.

In our analyses we have used one particular version of the human genome, *hg19* (also

named *GRCh37*), that have been released for the first time on 2009. The first human genome sequence was proposed in 2001 from a huge international effort, the Human Genome Project [81]. Several versions of the human genome have been proposed after, and the current available version is the *GRCh38 patch 12*, that have been released at the end of 2017. Indeed, we can not use universal constants for defining the genome as in the case of kilogram definition, as discussed in Ballouz *et al.* [17], and therefore it does not exist one particular fixed version of the human genome that can represent the whole diversity of the human genomes. The fact that the reference genome is idiosyncratic means that biological results of a particular application based on NGS data, such as the identification of genetic variants, can potentially differ when using different versions of the genome. In such a matter, both the alignment of reads can vary (modifying potentially the coverage and the VAF) and the variant itself. Obviously the impact of modifying the human genome version on biological results can be estimated when looking at the differences between versions. Another consequence of the idiosyncrasy nature of the reference genome is that this genome is not a baseline: it has been built at 70% from a single individual, and it has been reported that this individual has a high risk for diabetes [28]. This means that, firstly, the reference genome can contain "errors", *i.e.* individual variations, and secondly that the population-based referent DNA of an analyzed sample can vary from this human genome reference. In needlestack, to address this issue, we propose to inverse the reference DNA base if the majority of the analyzed samples (more than 50% by default) harbour an alternative base at a high proportion (more than 80% by default). This idea of a population-based reference is also developed in Ballouz *et. al.* [17], where the authors propose to build a consensus genome as a reference, based both on the existing reference and in addition on the population allele frequencies, leading to multiple reference genomes associated to multiple populations. Measuring the expected differences of results when using a population-based genome reference could be of interest.

The alignment step also requires to choose a particular alignment algorithm. As described in 2010 in a survey from Heng Li, the principal author of the extensively used BWA software, multiple alignment methods have been proposed since the emergence of NGS technology, and these methods vary mostly on the core algorithm and on the computation

time [86]. Majority of these methods are based either on hash tables such as the well-know Basic Local Alignment Search Tool (BLAST) aligner [12], or on prefix/suffix tries such as BWA [85], [88]. In addition to these methods, recently several re-alignment software have been developed, in order to take into account the correlation of reads mapped at the same locus, particularly to improve the detection of indels [99], [86]. The methods developed in this thesis are not directly linked to the alignment step, *i.e.* these methods does not require any particular alignment method, but, nevertheless, modifying the aligner can potentially modify the results of our methods. Needlestack can have different results both in term of sequencing error rate estimates and on identified variants when used in conjunction to different alignment methods. Indeed, observed read counts used in the needlestack model are totally dependent of the read alignment. In the applications presented in the thesis, we have used the BWA aligner, and the TMAP aligner that incorporates the BWA algorithms, and for germline analysis we in addition have used the assembly-based re-aligner ABRA [99], which was mostly developed for low coverage data. An interesting supplementary work could be to test the impact of aligner variation on the results of needlestack, particularly how changing the sequence alignment may result in differences between individuals, and, might influence the potential of a given individual to be an outlier in the needlestack regression. On another hand, concerning our variant-filtering methods, the machine-learning based approach result (the trained model) should vary when modifying the alignment method. Indeed, the model parameters (the trees in the forest) depend on the training data variables, and, some of them directly depend on the alignment, such as variables based on the coverage. It could be interesting to estimate if a machine learning model trained on data aligned with a particular aligner can be used efficiently on data aligned with a different alignment method.

Validation of the mutations

Even though we have provided in this thesis multiple methods to efficiently identify the DNA mutations, one particular difficulty when trying to estimate the accuracy of the methods was the technical validation of the mutations. For needlestack, we have proposed several ways to estimate our performance, such as:

- using BAMsurgeon to control the mutations and estimate the sensitivity of needlestack
- using the validation in the tumor of cfDNA mutations
- using a BeadArray to validate the germline mutations

Nevertheless, these approaches are limited: BAMsurgeon approach can not be used to estimate the specificity, only a few number of cfDNA mutations were used (also because we used only one gene), and the BeadArray is limited to particular variations (able to be detected by the array). Ideally, because no gold standard data is available to estimate our specificity and sensitivity for multiple types of data in particular for somatic mutations, we would need to use a large number of samples and genes, and to validate each called mutation and each non-variant position, using an independent technology. This is particularly the case for very low VAF mutations (typically less than 1%), as (i) they are only beginning to be explored using methods such as needlestack, (ii) we understand very little about the dynamics of these mutations, (iii) it remains difficult to determine what are the true mutations and if there remains false positive mutations related to factors that we are, as yet, unable to control for. An extensive project that technically validate low VAF mutations, would be greatly informative and allow us to further refine methods like needlestack, as well as the variant filtering steps. These technical validations could be achieved by performing an independent NGS experiment (with independent library preparation and different sequencing technology) or ideally by performing an other type of DNA variant calling such as the Droplet digital PCR (ddPCR). Nevertheless, such extensive validation step is expensive, and can be unable to be undertaken due the lack of sufficient biological material the main drawback of this technique is that it requires high amount of DNA (typically more than 10 nanograms).

Extensions of needlestack usage

The primary scientific aim of developing our needlestack algorithm was to provide a high sensitivity to detect very low abundance mutations, in order to efficiently perform various projects such as our ctDNA projects. The primary advantage of using needlestack is to correct for systematic errors, as its core algorithm consists in the modelling of such errors.

Needlestack is relatively computationally intensive compared to state-of-the-art somatic and germline variant callers (as an example the variant calling with needlestack takes around 20 hours for around 50 WES on 100 CPUs whereas strelka2 [71] takes only 1 hour), and we believe that needlestack, while slower, is able to assist in identifying false positive variants that other methods would not detect. As such a promising extension of the way to use needlestack could be to launch it in conjunction with existing variant callers (that would be launched with relaxed filters to increase the sensitivity), as a subsequent step, only on a subset of positions in order to reduce then the false calls. Indeed, needlestack can identify systematic errors that are basically confused with low VAF variants by current methods, such as G to T transversion errors linked to DNA sonication (see needlestack paper and [77]). Typically, variant callers advise to filter on these low VAF to control the false discovery rate, but in some projects such as when studying tumor heterogeneity, such very low abundance variants can reflect the presence of small tumor subclones and therefore the are an important data to keep. The best strategy in this type of study would be to filter smartly on low abundance mutations, and for this needlestack could be a good candidate to filter only errors.

A second point to mention about possible extensions of needlestack is its ability to detect other types of DNA variations not addressed here, such as CNV or Structural Variation (SV). Two points should be raised here. First, needlestack idea is the modeling of the systematic errors to efficiently detect mutations as non-systematic DNA variations. This idea of using multiple samples to detect mutation has also been used by the CODEX [68] CNV variant caller. As we showed in the needlestack paper, the error rate of a DNA variation is negatively correlated with the length of the variation (supplementary figure 3 in the needlestack paper). If this observation is extrapolated to CNVs and SVs analysis, it would not be expected that CNV and SV harbour a high error rate, leading to an y-axis at zero for all the non-variant samples in our regression. In addition, the CNVs start and end positions slightly differ between samples. Even if this would not impact their detection with needlestack, it would be difficult to compare the results between samples.

The idea of our needlestack algorithm, *i.e.*, modeling systematic errors across samples using a negative binomial regression can be extended to other types of data. For example,

our algorithm could be used for ddPCR data. The main idea of the ddPCR technology is to realize a PCR individually on subsets of input DNA called "compartments" or "droplets", in order to reduce the competition between low and high abundance DNA during the PCR, such as in the case of ctDNA. At the end, for each sample is obtained a large number of droplets that are estimated as positive or negative for a target mutation using a fluorescence technique. We have observed that the number of positive droplets as a function of the total number of droplets per sample can be modeled using a robust negative binomial regression in order to detect outliers as true mutated samples, which is a comparable logic to our needlestack algorithm. We are currently estimating on real data the benefit of using the needlestack modeling approach compared to hard threshold on the number of positive droplets.

Global scientific value

This thesis should have multiple scientific impacts in the domain of computational cancer genomics. First, the variant calling is more than just a classical step in the analysis of NGS data. Indeed, one can consider that the major advantage of NGS emergence is actually the possibility to call variants with a larger spectra than the one reached with the Sanger sequencing, *i.e.* NGS is not restricted to high abundance variants. The fact that low abundance variants can be detected presents major improvements for cancer research, such as the study of tumor heterogeneity, the detection of somatic mutation in normal tissues to study cancer initiation and progression, or even the development of "liquid biopsies" that requires an accurate variant calling. Nevertheless, the detection of such genomic variations highly depends of the variant calling method (even though it also depends of other steps, such as the alignment or the sequencing itself). In this thesis we were interested in increasing the potential of NGS variant calling in order to be able to detect such low abundance variants (development of our needlestack method, see chapter 2) that are crucial to identify in studies like the one described above. The capacity to detect low abundance variant is crucial in such projects, but, in addition, the precision of the variant calling should also be controlled to remove remaining false discoveries. Indeed, minimizing the proportion of such remain-

ing errors is a key step when highly precise NGS data is required, as when developing genetic risk scores or performing burder tests, which are highly impacted by a small number of false discoveries. While scientific efforts were primary oriented toward the improvement of alignment and variant calling algorithms, a current enthusiasm about variant filtering and its promise to boost mutation detection precision is emerging. We were then interested in the development of variant filtering methods (see chapter 3) based on hard-thresholds on variant summary statistics and on machine learning algorithms to remove the need to choose arbitrary thresholds. We applied our hard-threshold method on our ctDNA project and our machine learning method on kidney cancer data. We host the source codes on GitHub in order to provide the scientific community with our variant filtering frameworks.

Conclusion

The methods presented in this thesis are based on the systematic nature of errors in NGS and provide the accurate detection of DNA variations from NGS data. We have presented two main methodologies, (i) a sensitive variant calling method that detect systematic errors and identify accurately the mutations, (ii) and variant filtering methodologies in order to boost the precision of the mutation detection. We have validated our methods on both real and simulated data in order to estimate their performance. We have also applied our methods on four distinct projects, in order to develop non-invasive biomarkers for cancer detection based on the detection of cancer mutations in body fluids.

Other applications that require an accurate detection of DNA variations could also be undertaken using the methods that we have developed in this thesis. For example, our variant detection methods, that have the advantage to also detect low abundance mutations, can be applied (i) to precisely identify tumor subclones, whereas current methods tend to rely on high VAF mutations, (ii) to detect somatic mutations in normal tissues, that are expected to be found with very low VAF.

Extensive technical validation using independent technologies will also allow the further refinement and optimization of our variant detection methods. Finally, we hope that these methods will allow the exploration and description of the entire spectrum of mutations. This will allow a more complete description of such mutations and how they contribute to disease development and can be used in secondary prevention.

Appendix A

Annexes

A.1 Supplementary work

A.1.1 TCGA germline variant calling for rare variant susceptibility project

In this project (application of the machine learning variant filtering, chapter 3), we used a gene prioritisation method in order to build a list of potential susceptibility genes, that were in a second step validated using a burden test in independent cohorts of cases and controls. To boost the precision of the calling of variants, in the second step I developed a germline variant filtering methodology based on a machine learning model. I was also implicated in the computation of variables used in the gene prioritisation part. Particularly, we have used the germline status of each candidate gene in the TCGA cohorts, *i.e.* for each gene we used the proportion of germline-mutated samples. This variable required to launch a germline variant calling on the whole TCGA dataset, that contains around 10,000 WES. We used the current state-of-the-art algorithm for germline variant calling, the haplotype-based variant caller Platypus [109]. To efficiently perform this task, I benefited from the emergence of cloud computing using the Seven Bridges Cancer Genomics Cloud (CGC). This work was divided into two parts:

- Platypus performance maximization using parameter variation
- TCGA germline variant calling using the CGC

Platypus performance maximization using parameter variation

The accurate benchmarking of a variant caller requires a well-defined catalogue of truth variants in order to precisely estimate the sensitivity of the method. In addition, the specificity estimation can only be achieved using non-variant high confidence genomic positions. To correctly estimate the performance of Platypus on WES data, we therefore benefited from the "Platinum" truth variant catalogue of the well-known NA12878 sample developed by Illumina [44]. Contrary to the National Institute of Standards and Technology (NIST) that developed the Genome In A Bottle (GIAB) similar catalogue, Illumina used a haplotype transmission information method from 17 individuals to generate the set of true called and non-variant positions. Illumina reports an high consistency rate with the NIST GIAB (more than 99.9%) with in addition 26% more SNVs and 45% more indels. Finally, they provide two joint datasets:

- ~1.2 billions of confident genomic positions
- ~5 millions of truth variants

From this, we firstly extracted the good-quality calls at exomic positions by providing a BED file of human exons and by filtering calls on coverage (reauired higher than 20) and on quality of call (required higher than 20 in Phred-scale). We finally obtained a set of around 30 millions high confidence exomic positions and around 180,000 truth exome variants.

When the Platypus variant calling was launched with default parameters, the sensitivity on this high-confidence regions was estimated as **0.855** and the specificity as **0.995**. Platypus reports in the VCF file all observed variants, with the *PASS* annotation if the variant is kept or it reports the filter that has removed the variant if not kept. To increase the sensitivity of Platypus, we used the filtered variants that were actually truth variants in the Platinum dataset to defined a set of input parameters that we then varied using *a priori* from these observations. Finally, we obtained a sensitivity of **0.970** and a specificity of **0.994** with the parameter combination presented in the table A.1. These efficient input parameters were then used to perform the germline variant on the whole TCGA dataset with Platypus.

Table A.1 – Details of Platypus input parameters and chosen values to increase the default performance computed on the Platinum truth set.

Parameter	Default	Efficient	Description
hapScoreThreshold	4	10	Maximum number of haplotypes supported in the calling window
scThreshold	0.95	0.99	Max fraction of the surrounding sequence which can be made of any 2 bases
rmsmqThreshold	40	20	Minimum root-mean-square mapping quality across region containing variant
qdThreshold	10	0	Minimum quality-by-read/depth
badReadsThreshold	15	0	Minimum median of base qualities around variant position (window=11pb)

TCGA germline variant calling using the Cancer Genomics Cloud

We benefited from the emergence of cloud computing to boost the efficiency of our computations. Indeed, cloud computing is based on the idea that "tool should be bring to the data" and not the opposite, by providing a network access to a shared pool of configurable computing resources. Cloud computing compared to traditional computing using local machine or shared cluster emphasizes on three major concepts ([82]):

- *elasticity*: user rent resources while paying for only what is used
- *reproducibility*: investigators can store multiple versions of data and analyses on the cloud without loss or modification
- *distributed collaboration*: analyses can be performed on the same data set by multiple investigators at multiple different geographic sites

We were particularly interested in using cloud computing for this germline variant calling due to the high number of TCGA samples that require to be downloaded if performing local computations (we have used the SevenBridges CGC platform). Finally, each analysed WES required around 0.05 dollars, which corresponds to around 500 dollars in total for the whole TCGA dataset. Around 10 minutes were needed for the analysis of one WES sample, and the CGC allows a parallelization by sets of 100 samples (*i.e.* 100 samples analyzed at the same time), which corresponds to a total of around 16 hours. The downloading of one par-

ticular BAM file was estimated as around 7 minutes, which corresponds to several hundreds of hours when downloading to whole set of samples.

The germline variant calling with Platypus was divided into three parts:

- Querying data (*i.e.* BAM file for each cancer type) using the CGC Application Programming Interface (API)
- Tool description using Common Workflow Language (CWL)
- Run the Platypus variant calling on the cloud using a loop on each sample locally with R

Examples of scripts used to perform these analyses are available on GitHub: https://github.com/tdelhomme/CancerGenomicsCloud_tutorial.

We performed the germline variant calling on a total of 32 TCGA cancer types, grouping around 10,000 samples.

Finally, I benefited from the acquired experience about cloud computing using the Seven Bridges CGC to give an internal course on this subject (with both theory and applications) at IARC in March 2018. The complete support of the course is freely available on GitHub: https://github.com/IARCbioinfo/SBG-CGC_course2018.

A.1.2 IARC bioinformatics pipeline homogenization

When developing our needlestack variant caller, I was concerned about multiple concepts defining "good" science, such as *reproducibility*, *efficiency of computations* and *user-friendliness*. To implement needlestack following these concepts, we used *nextflow* as the workflow manager coupled with *Docker*/*Singularity* to provide reproducible environments (which are hosted on DockerHub and SingularityHub) and *conda* to easily install the dependencies. We also versioned the source code on GitHub, that can communicate with *nextflow* when the user run the pipeline. To control for stability of the code when it is modified, we also used a continuous integration tool, *CircleCI*, that run pre-defined test when the pipeline is modified on GitHub. I was therefore involved in the homogenization of the IARC bioinformatics pipelines (https://github.com/IARCbioinfo/) following this implementation pattern. The figure A.1 is

presenting the schematic overview of the implementation pattern that we used for our IARC bioinformatics pipelines. The implementation process is defined as the following:

- source code is written with *nextflow* and pushed to GitHub when modified
- a particular git branching model is applied
- CircleCI manages a continuous integration by launching pre-defined tests
- When tests are ok, CircleCI updates containers on DockerHub and SingularityHub
- user can run the pipeline specifying a version and can use the containers to avoid installing manually the dependencies

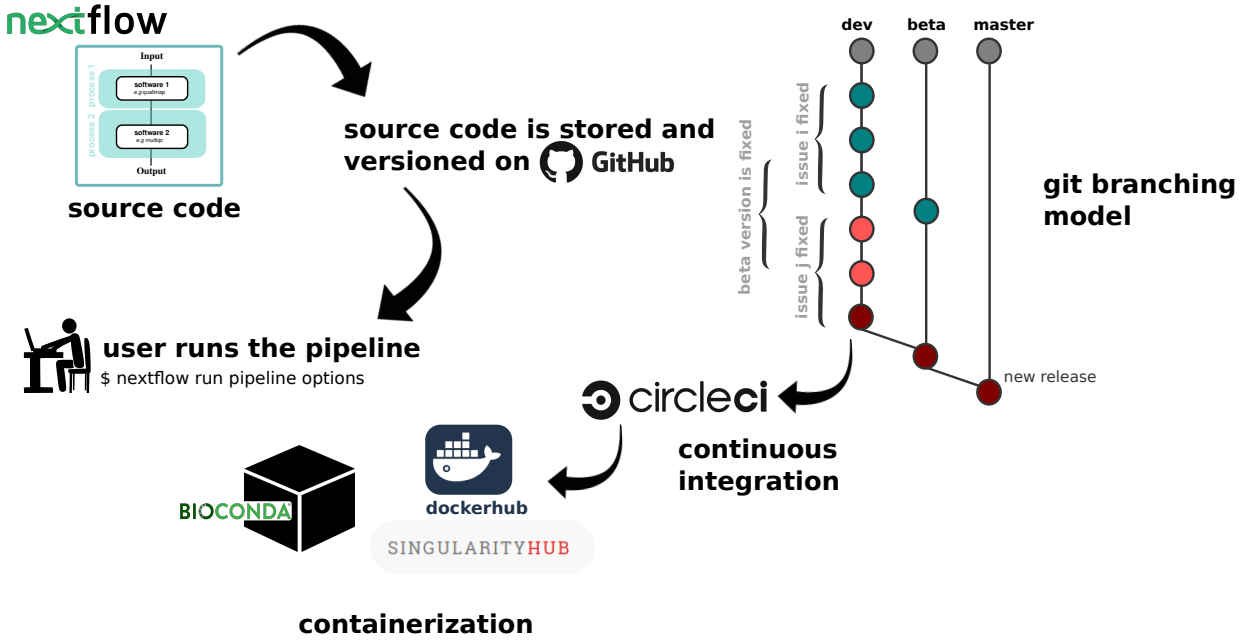


Figure A.1 – Schematic overview of the implementation pattern used for our IARC bioinformatics pipelines. The source code is written with nextflow, stored and versioned on GitHub, and we provide Docker and Singularity containers and continuous integration with CircleCI to control for code stability.

Table A.2 is presenting the catalogue of the bioinformatics pipelines developed following this implementation pattern, in collaboration with other bioinformaticians at IARC. Pipelines can be grouped in three main topics: quality control, DNA analysis and RNA analysis. A total of around 10 pipelines are currently available.

Finally, I was offered to present this collaborative work on the IARC bioinformatics pipeline development at the nextflow workshop in November 2018 at Barcelona, Spain. Details of the workshop is available on GitHub: https://github.com/nextflow-io/nf-hack18 as

well as the slide of my presentation: https://github.com/tdelhomme/Talks/nextflow_workshop2018.

A.2 Publications from other collaborations

A.2.1 Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors (George *et al.*, Nature Communications, 2018)

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-03099-x OPEN

Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors

Julie George et al.#

Pulmonary large-cell neuroendocrine carcinomas (LCNECs) have similarities with other lung cancers, but their precise relationship has remained unclear. Here we perform a comprehensive genomic ($n = 60$) and transcriptomic ($n = 69$) analysis of 75 LCNECs and identify two molecular subgroups: "type I LCNECs" with bi-allelic *TP53* and *STK11/KEAP1* alterations (37%), and "type II LCNECs" enriched for bi-allelic inactivation of *TP53* and *RB1* (42%). Despite sharing genomic alterations with adenocarcinomas and squamous cell carcinomas, no transcriptional relationship was found; instead LCNECs form distinct transcriptional subgroups with closest similarity to SCLC. While type I LCNECs and SCLCs exhibit a neuroendocrine profile with *ASCL1*^{high}/*DLL3*^{high}/*NOTCH*^{low}, type II LCNECs bear *TP53* and *RB1* alterations and differ from most SCLC tumors with reduced neuroendocrine markers, a pattern of *ASCL1*^{low}/*DLL3*^{low}/*NOTCH*^{high}, and an upregulation of immune-related pathways. In conclusion, LCNECs comprise two molecularly defined subgroups, and distinguishing them from SCLC may allow stratified targeted treatment of high-grade neuroendocrine lung tumors.

Correspondence and requests for materials should be addressed to J.G. (email: jgeorge@uni-koeln.de) or to E.B. (email: EBrambilla@chu-grenoble.fr) or to R.K.T. (email: roman.thomas@uni-koeln.de)
#A full list of authors and their affliations appears at the end of the paper

Molecular characterization studies have provided invaluable insight into the relationship between the major lung tumor subtypes[1–7]. These studies showed that morphologically defined lung adenocarcinomas, squamous cell carcinomas, and small cell carcinomas have distinct molecular phenotypes based upon their somatically altered genes[7]. Furthermore, global transcriptional analyses have revealed intra-group consistency, as well as substantial differences in the patterns of expressed genes, which led to the discovery of novel intra-group subtypes[2,3,8–11] and to the elimination of previous lung tumor categories (e.g., large-cell carcinoma)[7]. Of the remaining lung cancer subtypes, only large-cell neuroendocrine carcinomas (LCNECs) have so far not been characterized in depth using both transcriptomic, as well as genomic approaches.

LCNECs account for 2–3% of all resected lung cancers and belong to the category of neuroendocrine lung tumors, which also includes pulmonary carcinoids (PCa) and small cell lung cancer (SCLC)[12,13]. Contrary to pulmonary carcinoids, LCNEC and SCLC are clinically aggressive tumors presenting in elderly heavy-smokers with 5-year survival rates below 15–25% (LCNEC) and 5% (SCLC), respectively[12,13]. While therapy for both typical and atypical carcinoids and SCLC is primarily surgery and chemotherapy (in the case of SCLC), chemotherapy has limited efficacy in LCNEC patients and no standard treatment regimen exists for this tumor type[14]. Thus, LCNECs share both commonalities (e.g., neuroendocrine differentiation) and discrepancies (e.g., limited response to chemotherapy) with SCLC; however, the underlying molecular basis of these shared and distinct features is only poorly understood. Further complicating the histological classification, LCNECs are sometimes found combined with adenocarcinoma or squamous cell carcinoma and some SCLCs are combined with a component of LCNEC[12,13]. Thus, defining the molecular patterns of this tumor type presents the opportunity to not only reveal possible novel therapeutic targets, but also help clarifying the ontogeny and relationship of lung tumors in general.

Previous efforts in characterizing LCNECs through targeted sequencing of selected cancer-related genes[15–17] and through gene expression profiling[18] provided some first insights; however, global genomic studies combined with transcriptomic analyses have so far been lacking. Furthermore, given the lack of adequate therapeutic strategies in LCNECs, a precise delineation of the molecular boundaries between different neuroendocrine tumors is needed. We therefore aimed to comprehensively dissect both the mutational and the transcriptional patterns of this tumor type.

In this report, we show that LCNECs are composed of two mutually exclusive subgroups, which we categorize as “type I LCNECs” (with *STK11*/*KEAP1* alterations) and “type II LCNECs” (with *RB1* alterations). Despite sharing genomic alterations with lung adenocarcinomas and squamous cell carcinomas, type I LCNECs exhibit a neuroendocrine profile with closest similarity to SCLC tumors. While type II LCNECs reveal genetic resemblance to SCLC, these tumors are markedly different from SCLC with reduced levels of neuroendocrine markers and high activity of the *NOTCH* pathway. Conclusively, LCNECs represent a distinct subgroup within the spectrum of high-grade neuroendocrine tumors of the lung, and our findings emphasize the importance of distinguishing LCNECs from other lung cancers subtypes.

Results

Genomic alterations in LCNECs. We collected 75 fresh-frozen tumor specimens from patients diagnosed with LCNEC under institutional review board approval (Supplementary Data 1). All tumors were thoroughly analyzed, and the histological features of pulmonary LCNECs were confirmed by expert pathologists (E.B., W.T., R.B.) according to the 2015 WHO classification[13] (Supplementary Data 2). Most tumors were obtained from current or former heavy smokers, and enriched for stages I and II (68%). Nineteen of 75 LCNECs included in this study showed additional histological components of lung adenocarcinoma (ADC) ($n = 2$), squamous cell carcinoma (SqCC) ($n = 5$) or SCLC ($n = 12$) (Supplementary Data 1–2). In subsequent analyses nucleic acids were extracted only from pure LCNEC regions (Methods section).

Early genomic profiling studies employing targeted sequencing of selected cancer-related genes aided in the identification of some prevalent mutations in LCNECs[15–17]. In order to assess globally all genomic alterations in LCNECs and to compare them to those occurring in other lung tumors, we conducted whole-exome sequencing (WES) of 55 LCNEC tumor-normal pairs; we additionally performed whole-genome sequencing (WGS) in those cases where sufficient material was available ($n = 11$), thus amounting to sequencing data of 60 LCNECs in total (six tumors were both, genome- and exome-sequenced, Supplementary Fig. 1a). We furthermore performed Affymetrix 6.0 SNP array analyses of 60 and transcriptome sequencing of 69 tumors (Supplementary Data 1; Supplementary Fig. 1a). Despite initial review to include cases with a microscopic tumor content of >70%, sequencing data analysis revealed a median tumor purity of 59.5% and a median ploidy of 2.8 (Supplementary Data 1, Supplementary Fig. 1b, Methods section). On average, LCNECs exhibited an exonic mutation rate of 8.6 non-synonymous mutations per million base pairs and a C:G > A:T transversion rate of 38.7% (Fig. 1a, Supplementary Data 1), indicative of tobacco exposure[1–6]. We analyzed the signatures of mutational processes[19,20] in LCNECs, which confirmed a prominent smoking-related signature (signature 4[19,20]) that accounts for the majority of all somatic mutations, and which is in general comparable to most other lung tumors of heavy smokers (Supplementary Fig. 1c–f, Supplementary Data 3).

Analyses of chromosomal gene copy numbers revealed statistically significant amplifications of 1p34 (containing the *MYCL1* gene, 12%), 8p12 (containing *FGFR1*, 7%), 8q24.21 (containing *MYC*, 5%), 13q33 (containing *IRS2*, 3%), and 14q13 (containing *NKX2-1*, also known as TTF-1, 10%) ($Q < 0.01$, Supplementary Fig. 2a; Supplementary Data 4–5, Methods section). Statistically significant deletions affected *CDKN2A* (9p21, 8%) and a putative fragile site at *PTPRD* (9p24, 7%)[21]. While amplifications of *NKX2-1* and *FGFR1* frequently occur in lung adenocarcinomas[1,2,7,21] and squamous cell carcinomas[3,7,21,22], respectively, *MYCL1* amplifications are commonly found in SCLC[4–6,23]. Thus, LCNECs harbor significant copy-number alterations that occur in different lung cancer subtypes.

We next applied analytical filters to identify mutations with biological relevance in the context of a high-mutation rate and found eight significantly mutated genes ($Q < 0.01$, Methods section, Fig. 1a, Supplementary Data 6–7). *TP53* was the most frequently mutated gene (92%), followed by inactivating somatic events in *RB1* (42%); bi-allelic alterations in both genes, *TP53* and *RB1*—a hallmark of SCLC[4–6]—were found in 40% of the cases (Supplementary Fig. 2b, Supplementary Data 6–9). Notably, LCNECs with admixtures of other histological components mostly had *RB1* alterations (Fig. 1a). While genomic alterations in *RB1* resulted in loss-of-nuclear Rb1 expression ($P < 0.0001$, Fisher’s exact test, Supplementary Fig. 3a), immunohistochemistry revealed that absence of Rb1 was not only confined to the LCNEC component, but also evident in the combined other histological subtype (6/7 cases, Supplementary Fig. 3b, Supplementary Data 2). This may implicate shared genetic features between LCNECs and the admixtures of other histological carcinoma types.

We furthermore identified—frequently deleterious—somatic alterations in functionally relevant domains of *STK11* (30%) and

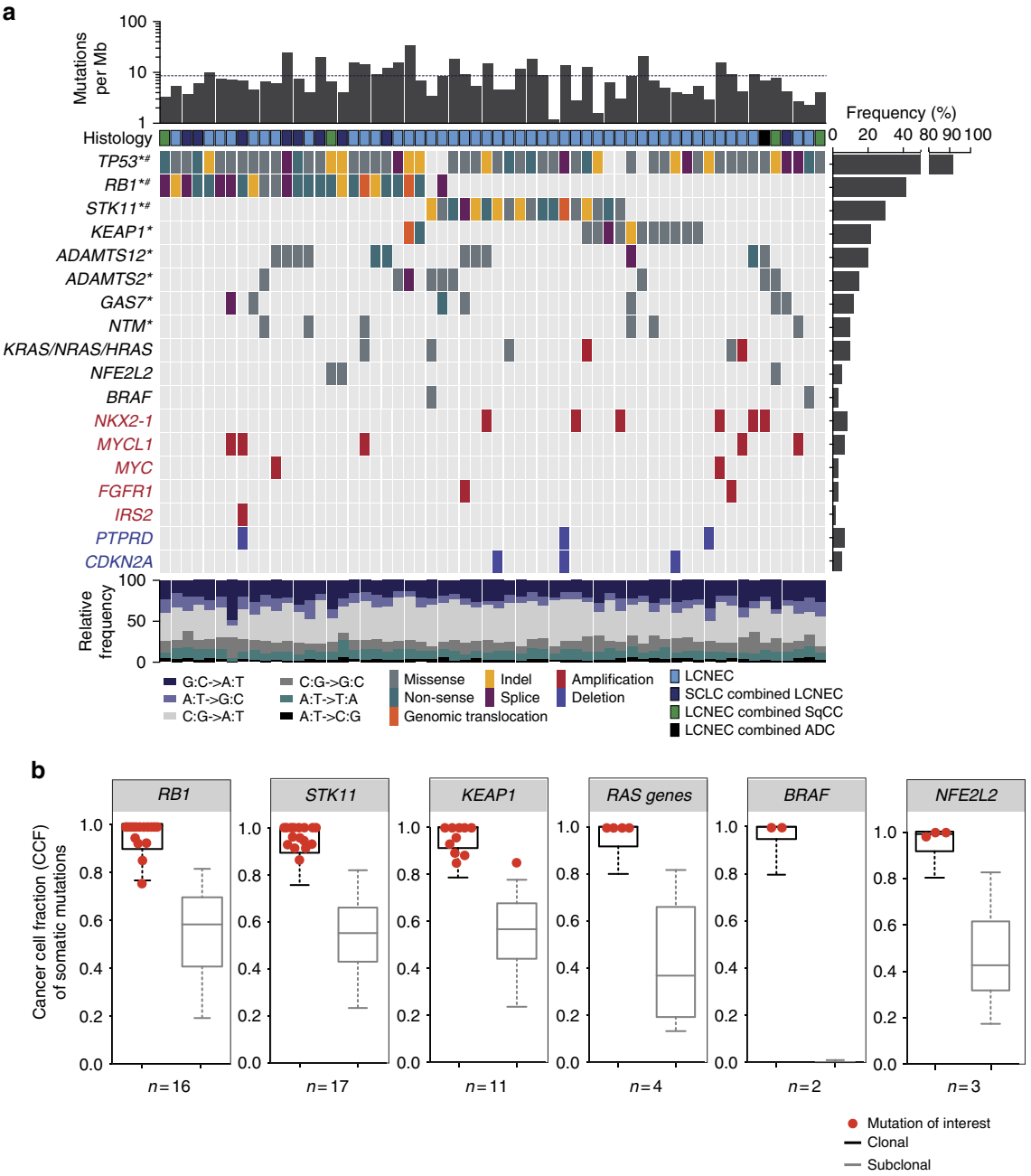


Fig. 1 Genomic alterations in pulmonary large-cell neuroendocrine carcinomas (LCNECs). **a** Tumor samples are arranged from left to right. Histological assignments and somatic alterations in candidate genes are annotated for each sample according to the color panel below the image. The somatic mutation frequencies for each candidate gene are plotted on the right panel. Mutation rates and the type of base-pair substitutions are displayed in the top and bottom panel, respectively; a dashed black line indicates the average value. Significantly mutated genes and genes with a significant enrichment of damaging mutations are denoted with * and #, respectively ($Q < 0.01$, Methods section). Genes with significant copy number (CN) amplifications (CN > 4) and deletions (CN < 1) (Supplementary Fig. 2a, Supplementary Dataset 5) are displayed in red and blue, respectively ($Q < 0.01$, Methods section). **b** The distribution of clonal and sub-clonal mutations was analyzed for tumor samples that harbored mutations in key candidate genes. The cancer cell fractions (CCF) of all mutations were determined, assigned to clonal or sub-clonal fractions (Methods section), and displayed as whiskers box-plot (median and interquartile range, whiskers: 5–95 percentile). The CCF of candidate gene mutations is highlighted in red

KEAP1 (22%)[1–3] (Fig. 1a, Supplementary Fig. 4a, Supplementary Data 6–9). Combined with loss-of-heterozygosity (LOH), bi-allelic alterations of *STK11* and *KEAP1* were found in 37% of the cases (Supplementary Fig. 2b, Supplementary Data 8). In those cases where WGS was performed, we were able to identify larger genomic rearrangements, which led to the inactivation of *RB1*, *STK11*, or *KEAP1* (Fig. 1a, Supplementary Fig. 4a, Supplementary Data 9). Altogether, somatic alterations of *RB1* and *STK11*/*KEAP1* were detected in 82% of the cases ($n = 49$) and occurred in a mutually exclusive fashion ($P < 0.0001$, Fisher's exact test, Fig. 1a). We furthermore observed a trend toward inferior outcome in patients with *RB1*-mutated tumors ($P = 0.126$, log-rank test, Supplementary Fig. 4b). The genomic profiling thus points to two distinct subgroups of LCNECs.

We additionally identified statistically significant mutations in the metalloproteinases *ADAMTS2* (15%) and *ADAMTS12* (20%),

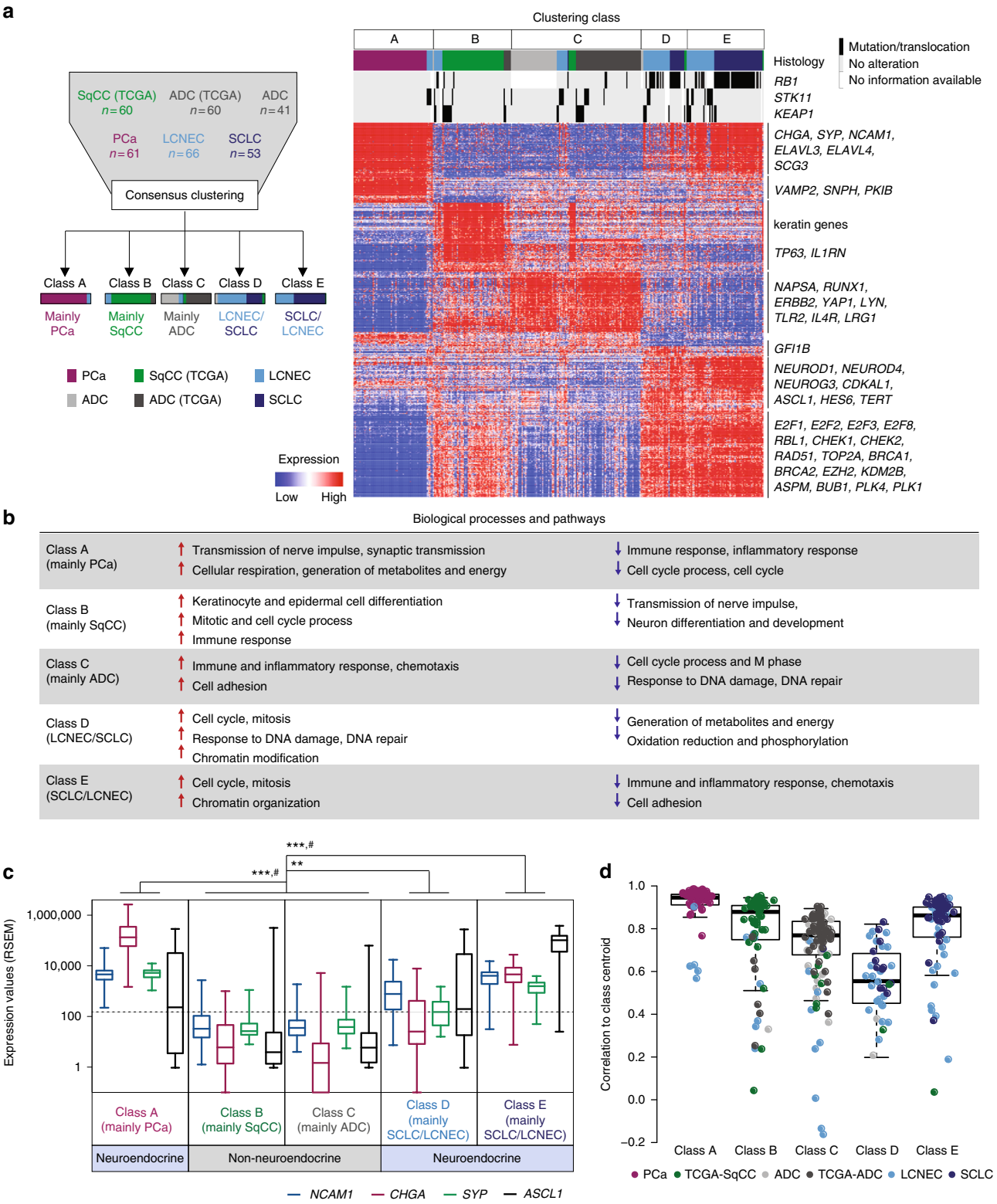


Fig. 2 Gene expression studies on lung cancer subtypes. **a** A schematic description of the unsupervised consensus clustering approach is provided on the left panel. The clustering results are displayed on the right panel as a heatmap, in which tumor samples are arranged in columns, grouped according to their expression clustering class, annotated for the histological subtype and for the somatic alteration status. Expression values of genes identified by ClaNC (Methods section) are represented as a heatmap; red and blue indicate high and low expression, respectively. Selected candidate genes are shown on the right. **b** Significant enrichment of differentially expressed genes in signaling pathways is displayed for all clustering classes ($P < 0.0001$, Methods section). **c** Expression values for key neuroendocrine differentiation markers are plotted for each clustering class as box-plots (median and interquartile range, whiskers: min–max values). Dashed black lines indicate the threshold for low expression (Methods section). $Q < 0.05$ (#), significance determined by SAM (Supplementary Dataset 12); $P < 0.001$ (***) Mann–Whitney *U*-test. **d** The correlation of each sample to the centroid of its clustering class was calculated and displayed as box-plot (median and interquartile range, whiskers 5–95 percentile)

and in *GAS7* (12%) and *NTM* (10%) ($Q < 0.01$, Methods section, Fig. 1a, Supplementary Fig. 4c, Supplementary Data 6–7), which so far have not been reported as significantly mutated in any other lung cancer subtype. The mutations affected functionally important protein domains, which may suggest a relevant role in the tumorigenesis of LCNECs (Supplementary Fig. 3c).

We also analyzed LCNECs for alterations in genes of known tumor-specific functions (e.g., *CREBBP*, *EP300*[3,4,6,21], *NOTCH*[3,6,21], *MEN1*[24], *ARID1A*[1–3,21,24]) (Supplementary Fig. 2b, Supplementary Fig. 4d, Supplementary Data 6) and found oncogenic mutations of *RAS* family genes (*KRAS*-G12V, -G12C, *NRAS*-D57E, *HRAS*-G13R), *NFE2L2* (2 cases with G31V and 1 case with E79Q) and *BRAF* (V600E, and G469V). Combined with focal amplifications, *RAS* genes were affected in 10% of the tumors (Fig. 1a; Supplementary Data 5–6). We also identified several private in-frame fusion events, e.g., involving the kinases *NTRK1* and *PTK6*, which were, however, not recurrent (Supplementary Fig. 5, Supplementary Data 10). Thus, LCNECs harbor alterations of oncogenes which are commonly found in lung adenocarcinomas, but usually absent in neuroendocrine tumors like SCLC.

The distinct mutational patterns in LCNECs and the presence of other histological components may suggest a high level of intra-tumor heterogeneity. We analyzed the clonal distribution of somatic alterations and determined the cancer cell fraction (CCF) of each somatic mutation call (Methods section). Despite the fact that some LCNECs were found with admixtures of other histological subtypes (Fig. 1a, Supplementary Data 1–2), our studies on the LCNEC component of such composite tumors pointed to little intra-tumor heterogeneity with a median of 7% sub-clonal mutations per sample (Supplementary Fig. 2b–c, Supplementary Data 1, Methods section). Furthermore, all relevant and significant mutations were found to be clonal within the tumor, thus suggesting these alterations as early events during tumorigenesis (Fig. 1b, Supplementary Data 6).

In summary, genome sequencing revealed distinct genomic profiles in LCNECs. While certain alterations (e.g., *RB1, MYCL1*) resemble patterns found in SCLC[4–6,23], others are typical of lung adenocarcinoma or squamous cell carcinomas (e.g., *STK11*, *KEAP1*, *NKX2-1*, *RAS*, *BRAF*, and *NFE2L2*)[1–3,7,21]. Thus, LCNECs appear to divide into molecularly defined subsets of tumors with genomic similarities to other major lung cancer subtypes.

Transcriptional profiles of LCNECs and other lung cancers. Our sequencing efforts have revealed genomic alterations in LCNECs that were previously known as canonical alterations in either, lung adenocarcinomas, squamous cell carcinomas[7,21], or SCLC[4–6]. In light of these distinct associations, it remained to be understood if these genomic correlates might reflect a relationship of LCNECs with these lung tumor subtypes on the level of gene expression. We therefore analyzed whether the transcriptional patterns in LCNECs are correlated with the expression profiles of other lung cancers.

We compared the expression data of LCNECs with lung adenocarcinomas[2,3,25–27], squamous cell carcinomas[3], SCLC[6] and pulmonary carcinoids[24] following extensive normalization of the transcriptome sequencing data (Fig. 2a, Methods section, Supplementary Data 11). Unsupervised consensus clustering yielded five consistent expression clusters, which correlated with the histological annotation of the tumors ($P < 0.0001$, Fig. 2a, Supplementary Fig. 6–7, Supplementary Data 12): pulmonary carcinoids, squamous cell carcinomas and adenocarcinomas formed distinct transcriptional classes (classes A, B, and C, respectively), with few LCNECs falling into these groups. However, the majority of SCLC and LCNECs clustered in two transcriptional subgroups (classes D and E) (Fig. 2a); a phenomenon that had previously been observed in other studies on high-grade neuroendocrine tumors[6,18]. While the majority of SCLC tumors formed consensus cluster E (75% of all SCLC cases analyzed), a fraction of SCLC tumors shared transcriptional similarities with LCNECs that predominantly formed cluster D. Thus, LCNECs appear to be more closely related to SCLCs than to adenocarcinomas or squamous cell carcinomas.

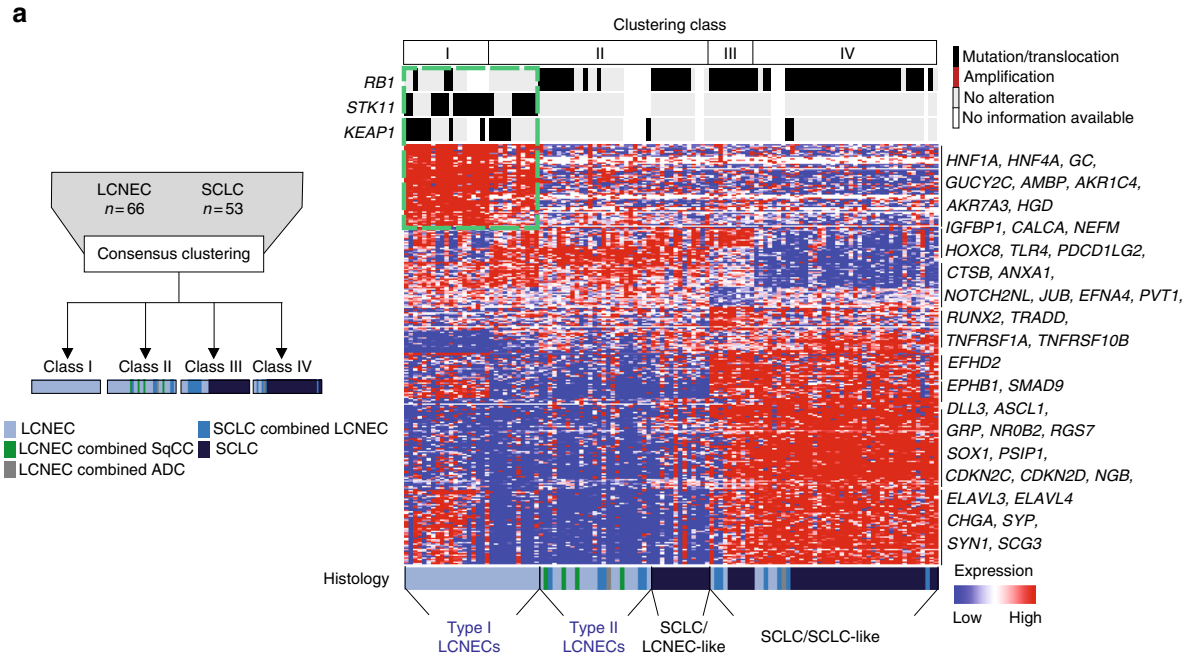
We next analyzed the transcriptome sequencing data for differentially expressed genes and their enrichment in biological pathways (Methods section). In line with previous observations[2,3,9–11,18,28], this analysis showed that both adenocarcinomas and squamous cell carcinomas exhibited upregulation of pathways controlling cell differentiation, adhesion and immune responses, along with higher expression of *ERBB2* and *TP63* (Fig. 2b, Supplementary Fig. 8a, Supplementary Data 13–14, $Q < 0.05$, Methods section). Lung neuroendocrine tumors, on the contrary, showed significantly higher expression of neuroendocrine and endocrine markers, Hu antigens (*ELAVL3* and *ELAVL4*) and the lineage transcription factor and oncogene *ASCL1*, which is in agreement with previous studies on lung cancer subtypes[11–13,18,29] ($Q < 0.05$, Methods section). Furthermore, particularly high expression of the neuronal and endocrine lineage transcription factors *NEUROD1*, *NEUROD4*, and *NEUROG3*[30,31] was found in SCLC and LCNECs of transcriptional class E (Fig. 2a, c, Supplementary Fig. 8b–e, Supplementary Data 13, $Q < 0.05$). While recent studies employing SCLC cell lines and mouse models indicated discordant expression patterns for *ASCL1* and *NEUROD1*[31], our sequencing data of human high-grade neuroendocrine lung tumors revealed expression of both neuroendocrine lineage factors in class E (Supplementary Fig. 8f).

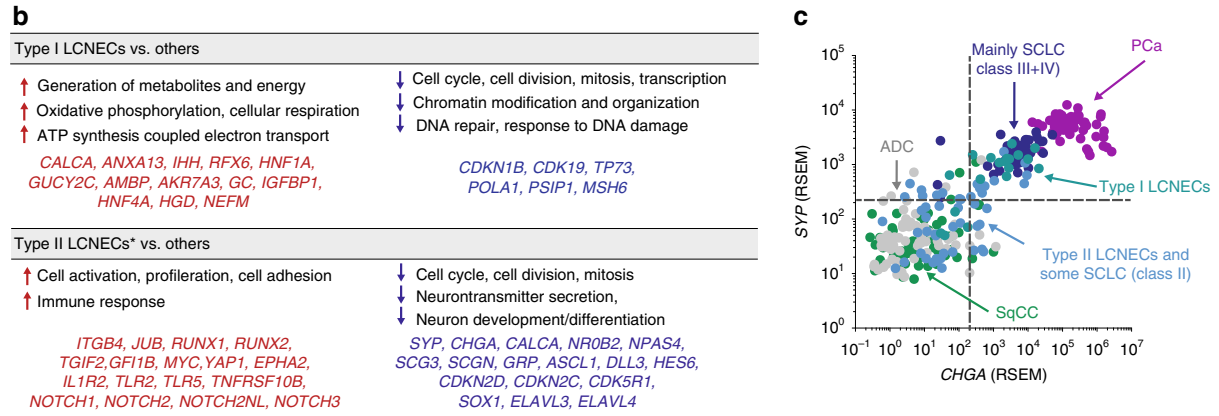
Within the spectrum of neuroendocrine lung tumors, pulmonary carcinoids formed a distinct subgroup with functional enrichment in pathways regulating cellular respiration and metabolism. LCNECs mostly shared similarities with SCLC, revealing upregulation of pathways and genes controlling cell cycle and mitosis (E2F transcription factors and checkpoint kinases), DNA damage response (*RAD51*, *TOP2A*, and *BRCA1*) and centrosomal functions (such as *BUB1*, *PLK1*, and *ASPM*); which, to some extent, were also found in squamous cell carcinomas (Fig. 2b; Supplementary Fig. 8g–i, Supplementary Data 13–14), and which is in agreement with previous studies[18]. Further supporting a molecular relationship of SCLC and LCNECs in a fraction of the cases, *RB1*-mutated LCNECs were enriched in classes D and E ($P < 0.05$, Fisher's exact test). Although, LCNECs also harbored alterations commonly observed in adenocarcinomas and squamous cell carcinomas, even LCNECs with such alterations in *KEAP1* or *STK11* were primarily found in transcriptional subclasses shared with SCLC (Fig. 2a, Supplementary Fig. 7c, Supplementary Data 12). Therefore, this observation supports the view that despite the similarity in oncogenic mutations, LCNECs rather constitute their own biological class; and may not be considered as neuroendocrine versions of adenocarcinomas or squamous cell carcinomas.

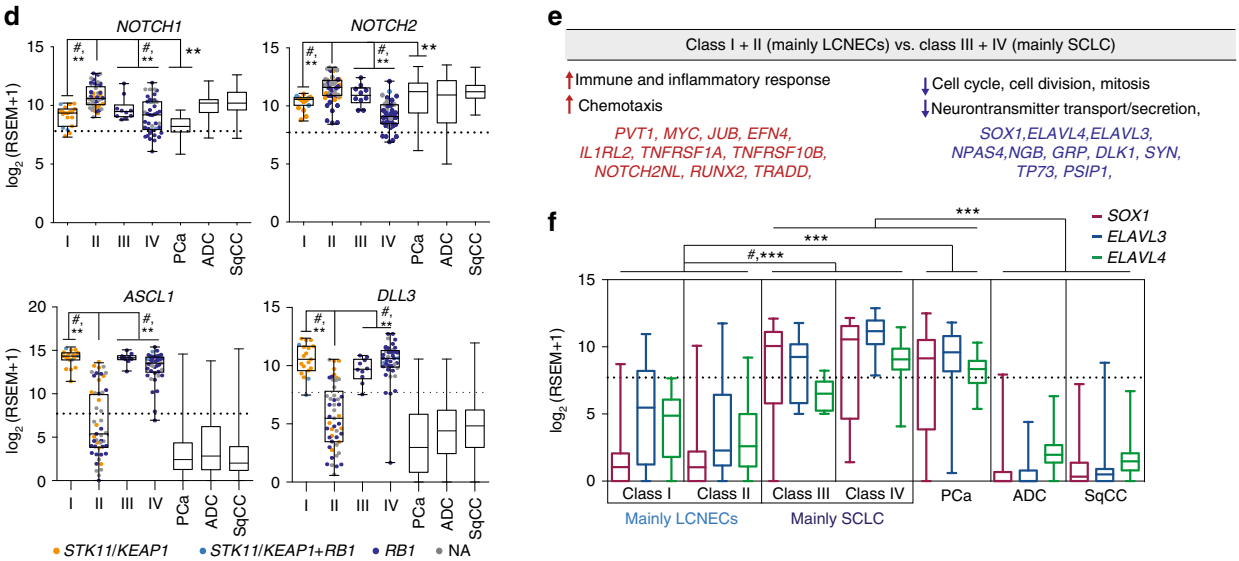
We also quantified the consistency of the expression profiles for each sample with respect to its clustering group. Again, this analysis revealed a strong correlation for most LCNECs clustering with SCLC tumors (classes D and E); on the other hand, expression profiles of those few LCNEC samples clustering with lung adenocarcinomas, squamous cell carcinomas, and pulmonary carcinoids were less consistent (Fig. 2d). Furthermore, we performed separate transcriptional clustering of LCNECs with adenocarcinomas and squamous cell carcinomas only (excluding SCLC), which did not suggest any unrecognized similarities between these lung cancer subtypes (Supplementary Fig. 9). Thus,

despite sharing somatic alterations with other tumor subtypes, such as adenocarcinomas and squamous cell carcinomas, LCNECs were transcriptionally dissimilar with all non-neuroendocrine lung tumors and showed closest similarities to SCLC.

The transcriptional relationship of LCNEC and SCLC. In the previous section, we sought for a global approach to identify common and distinct transcriptional profiles of LCNECs in relationship with other lung tumors, which showed that LCNEC and SCLC appear to share most transcriptional patterns.







However, strongly divergent tumors (e.g., carcinoids, adenocarcinomas) may drive these clusters and mask important differences between LCNECs and SCLC. We therefore sought to directly compare LCNECs and SCLC on the transcriptional level (Fig. 3a). The resulting unsupervised clustering analysis revealed four consensus clusters of LCNEC and SCLC that we termed classes I–IV in order to distinguish them from the above-mentioned classes A–E (Fig. 3a, Supplementary Fig. 10–11, Supplementary Data 12). Class I exclusively included LCNECs with *STK11* or *KEAP1* alterations; yet, a few cases with these alterations fell into class II that predominantly consisted of LCNECs with *RB1* loss (Fig. 3a). Some LCNECs, including tumors admixed with SCLC ("SCLC combined LCNECs")—clustered with the majority of SCLC tumors in the classes III and IV; similarly, some SCLC tumors were part of class II that included LCNECs bearing *RB1* alterations (Fig. 3a, Supplementary Fig. 11). Even though pathological review had been conducted to distinguish histological subtypes from one another, transcriptional clustering suggested high degrees of similarity for some LCNEC and SCLC cases; these tumors may therefore be considered as "SCLC-like" and "LCNEC-like" (Fig. 3a, Supplementary Fig. 11, Supplementary Data 11). Other major genome alterations (e.g., *NKX2-1*, *MYCL1*, *RAS* genes, *NFE2L2*, *BRAF*) did not segregate with the identified transcriptional subgroups (Supplementary Fig. 11). We further analyzed the consistency of the transcriptional subgroups by clustering LCNECs alone, which revealed high concordance with the transcriptional classes identified in Fig. 3a (62/66 cases, 94%, $P < 0.001$, Fisher's exact test, Supplementary Fig. 13, Supplementary Data 12). Thus, despite the similarities between LCNECs and SCLCs, subtypes of LCNECs exist with profound differences to SCLC.

The transcriptional clustering heatmap pointed to a strong gene expression pattern shared by all LCNECs bearing *STK11/KEAP1* alterations (Fig. 3a, Supplementary Fig. 12a, green box in upper left quadrant). We therefore conducted a supervised analysis of the gene expression data, in which LCNECs with *STK11/KEAP1* alterations were compared to tumors bearing *RB1* alterations. This analysis indicated specific expression profiles, which were similar to those observed in tumors constituting class I (Fig. 3b, Supplementary Fig. 12, Supplementary Data 13). We therefore assigned this genomic subset of tumors to one group, termed "type I LCNECs".

Type I LCNECs exhibited high levels of calcitonin A (*CALCA*), a known marker of pulmonary neuroendocrine cells[32–34] (Fig. 3a, Supplementary Fig. 12b, Supplementary Data 13). This subgroup furthermore displayed a pronounced upregulation of cellular metabolic pathways, which we also observed in pulmonary carcinoids (Fig. 2b), but which was less prominent in LCNECs and SCLC tumors with *RB1* alterations (Fig. 3a, b, Supplementary Data 12–13). Other genes found in type I LCNECs included gastrointestinal transcription factors (e.g., *HNF4A*, *HNF1A*, and *RFX6*), which were previously reported to play a role in de-differentiated lung tumors[35,36] (Fig. 3b, Supplementary Fig. 12c, d, Supplementary Data 13).

The most striking difference was found in the expression levels of neuroendocrine genes: while type I LCNECs and the majority of SCLC tumors (class III + IV) harbored high levels of neuroendocrine genes (*CHGA* and *SYP*; Fig. 3c; Supplementary Fig. 12e; Supplementary Data 12), most LCNECs and some SCLC tumors with *RB1* alterations in class II exhibited low levels of these genes (Fig. 3c, Supplementary Fig. 12e). By contrast, tumors in class II displayed elevated expression of genes associated with active Notch signaling (e.g., *NOTCH1*, *NOTCH2*, and *HES1*) and immune cell responses (e.g. *PDCD1LG2*, *TLR4*, and *CTSB*) (Fig. 3a, d, Supplementary Fig. 12f, Supplementary Data 12–13). Given the strong enrichment of LCNECs with *STK11* or *KEAP1* alterations in cluster I, and the prominent lack of expression of key neuroendocrine genes in most tumors of class II, we termed LCNECs within this transcriptional class as "type II LCNECs".

We have recently demonstrated that SCLC tumors usually harbor inactive Notch signaling and that activation of Notch reduces expression of neuroendocrine genes (e.g., *CHGA*, *SYP* and *NCAM1*) and Ascl1[6]. Consistent with this notion, we found that type II LCNECs and some SCLC within this transcriptional class exhibited signs of *NOTCH* upregulation and low expression of neuroendocrine markers, *ASCL1* and *DLL3*, an inhibitor of the Notch signaling pathway[37] (Fig. 3d, and Supplementary Fig. 12f). Conversely, type I LCNECs and the majority of the SCLC samples (class III and IV) showed higher levels of neuroendocrine genes, as well as of *ASCL1* and *DLL3*, and downregulation of *NOTCH* pathway genes (Fig. 3d, Supplementary Fig. 12f). Thus, despite the fact that type II LCNECs and some SCLCs harbor bi-allelic loss of *TP53* and *RB1*, their transcriptional signatures include low levels of neuroendocrine genes and a distinct profile of $NOTCH^{high}$ and $ASCL1^{low}/DLL3^{low}$, which differentiates these tumors from type I LCNECs and from the majority of SCLC cases. We did not identify any significant enrichment of somatic alterations in *NOTCH* pathway genes, which may explain these transcriptional differences (Supplementary Fig. 11). However, a recent study in a pre-clinical mouse model has established a central role of *REST* as a repressor of neuroendocrine markers in SCLC[38]. Compatible with these findings, type II LCNECs displayed significantly higher levels of *REST* (clustering class II, Supplementary Data 12, $Q < 0.05$), which may explain the low neuroendocrine phenotype in type II LCNECs marked by $ASCL1^{low}/DLL3^{low}/NOTCH^{high}$. Given the important role of NOTCH signaling and *ASCL1* in the decision of neuroendocrine fate and the development of neuroendocrine lung tumors[29,31,38], these findings provide further support for our distinction of type I and II LCNECs.

We next analyzed the relationship of the expression classes I–IV using hierarchical clustering, which revealed two major subgroups (Supplementary Fig. 11): one subgroup mainly consisting of LCNECs (type I and II LCNECs), and the other subgroup mainly containing SCLC tumors (classes III and IV). Thus, despite harboring distinct transcriptional subcategories, LCNEC and SCLC tumors largely followed their histological annotation and formed separate transcriptional subgroups. Differentially expressed genes included *SOX1* and the neuroendocrine Hu genes (*ELAVL3*,

Fig. 3 Gene expression studies on LCNEC and SCLC. **a** The expression profiles of LCNEC and SCLC tumors were analyzed following the annotation and approach described in Fig. 2a. Expression values of genes identified by ClaNC (Methods section) are represented as a heatmap in which red and blue indicate high and low expression, respectively. Selected candidate genes are shown on the right. Dashed green lines indicate an expression profile shared by LCNEC tumors with *STK11/KEAP1* alterations (type I LCNECs). **b** The significant enrichment of differentially expressed genes and signaling pathways are displayed for type I LCNECs and type II LCNECs. $P < 0.0001$ (Methods section); * some SCLC tumors that co-clustered with type II LCNECs were included in this analysis. Key candidate genes are highlighted in bold. **c**, **d** Expression values for **c** the key neuroendocrine differentiation markers *SYP* (synaptophysin) and *CHGA* (chromogranin A) (scatter plot), and **d** *NOTCH* pathways genes (box plots: median and interquartile range, whiskers: min–max values). **e** Significant enrichment of differentially expressed genes and signaling pathways was analyzed for class I and II vs class III and IV tumor samples; $P < 0.0001$ (Methods section). **f** Expression values of *SOX1*, *ELAVL3*, and *ELAVL4* are plotted for the clustering classes and other lung cancer subtypes (box plots: median and interquartile range, whiskers: min–max values). $Q < 0.05$ (#), SAM (Supplementary Dataset 12); $P < 0.01$ (**) Mann–Whitney *U*-test

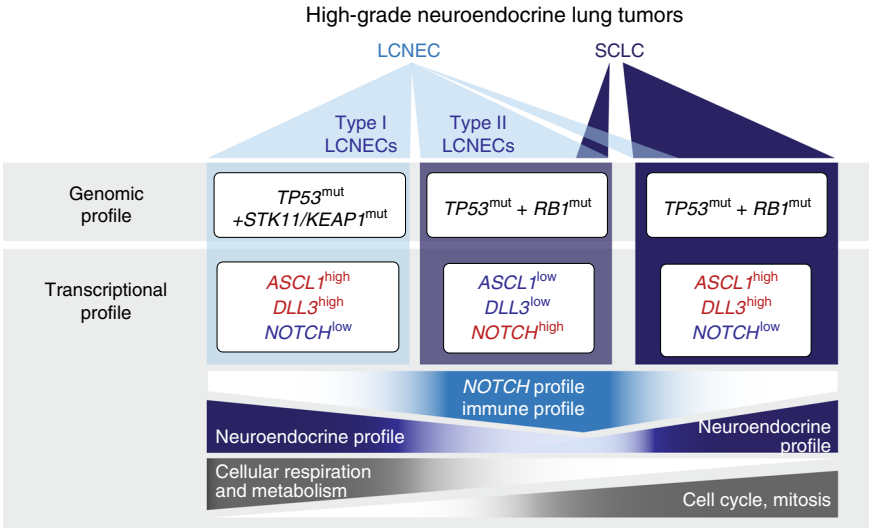


Fig. 4 Schematic overview of somatic alterations and expression profiles in high-grade neuroendocrine lung tumors. Significantly mutated genes are shown in black and differentially expressed genes are highlighted in red and blue, describing higher and lower expression, respectively. Upregulated expression profiles and signaling pathways are indicated by color gradients

ELAVL4), which were enriched in most SCLC samples (classes III and IV (Supplementary Data 13, $Q < 0.05$, Methods section) (Fig. 3f). This observation is in line with previous reports on auto-antibodies against Sox1 and Hu-proteins that are commonly found in SCLC patients[39]. While pulmonary carcinoids harbored similar expression levels, these genes were essentially absent or only moderately expressed in most LCNECs and other lung cancer subtypes (Fig. 3f).

We furthermore analyzed the impact of transcriptional subgroups on tumor stage and clinical outcome. While, we found no association of tumor stage with the molecular subsets found in high-grade neuroendocrine tumors (Supplementary Data 12), we observed a trend toward inferior survival in patients with SCLC (transcriptional profiles of classes III and IV; $P = 0.072$, log-rank test, Supplementary Fig. 14), which was similarly observed in previous studies on high-grade neuroendocrine lung tumors[18].

Conclusively, LCNECs exhibit a distinct expression profile within the spectrum of high-grade neuroendocrine lung tumors, which can further be divided into two subtypes: type I LCNECs with high neuroendocrine expression and, similar to SCLC, a profile of *ASCL1*^{high}/*DLL3*^{high}/*NOTCH*^{low}, and type II LCNECs with reduced expression of neuroendocrine genes and a pattern of *ASCL1*^{low}/*DLL3*^{low}/*NOTCH*^{high} (Fig. 4).

Discussion

Here we provide the first comprehensive molecular analysis of LCNECs, which allowed distinguishing between two genomic subgroups with specific transcriptional patterns, defined as “type I LCNECs” and “type II LCNECs” (Fig. 4).

Type I and II LCNECs harbor key genomic alterations and oncogenic mutations, which are commonly found in SCLC, lung adenocarcinoma or squamous cell carcinoma (e.g., in *RAS* genes, *BRAF*, *NFE2L2*, as well as in *STK11* and *KEAP1* in the case of type I LCNECS, and *RB1* losses in the case of type II LCNECs). One possible explanation for this observation might be a high level of intra-tumor heterogeneity, combined with occurrence of two tumor types in a single tumor. However, the key alterations that we found in LCNECs were mostly clonal, with limited genomic intra-tumor heterogeneity. Furthermore, thorough comparisons of gene expression profiles did not suggest similarities between LCNECs and lung adenocarcinomas or squamous cell carcinomas. Thus, the combinations of distinct sets of mutations with specific patterns of gene expression supports the view that LCNECs are not a variant of the other types of lung cancer, but represent a distinct subgroup within the spectrum of neuroendocrine lung tumors.

In a more focused comparison with the most frequent neuroendocrine type of lung cancer, SCLC, type I LCNECs with *STK11* and *KEAP1* alterations exhibited a high degree of similarity with these carcinomas, as well as high expression of neuroendocrine genes and a profile of *ASCL1*^{high}/*DLL3*^{high}/*NOTCH*^{low}. By contrast, type II LCNECs with *RB1* alterations revealed reduced expression of neuroendocrine genes and a pattern of *ASCL1*^{low}/*DLL3*^{low}/*NOTCH*^{high}. Notch family members play a multifaceted role in the development of neuroendocrine tumors with cell-type specific tumor suppressor and oncogenic functions[40]. We have shown in earlier studies that *NOTCH* serves as a tumor suppressor in SCLC[6], which mostly harbor high-level expression of the negative regulator of Notch, *DLL3*[6,37,41] (Fig. 4). A recent clinical trial with an antibody-drug conjugate targeting the non-canonical inhibitory NOTCH ligand, Dll3, has shown early signs of clinical activity in SCLC[37,41]. We now demonstrate shared neuroendocrine pathways between SCLC and type I LCNECs, which may be similarly susceptible to this agent. On the other hand, type II LCNECs with alterations in *RB1* exhibited active Notch signaling (Fig. 4). Clinical trials have assessed the efficacy of an antibody targeting Notch 2 and 3 in SCLC, but recently failed in demonstrating a clinical benefit[42,43]. Therefore, future clinical trials involving therapeutics, targeting activating or inhibitory members of the Notch pathway will—in our view—require clear assignment of the respective molecular subtype.

Perhaps another noteworthy finding, type II LCNECs exhibited a pattern of gene expression with upregulation of immune related pathways (Fig. 3b, Fig. 4), which has similarly been observed in various other tumor types[28] and which may impact the response of patients to immunotherapy. Taken together, the precise distinction of high-grade neuroendocrine tumors representing as type I LCNECs and as *RB1*-mutated SCLC or type II LCNECs, may be pivotal to assess the efficacy of targeted therapeutics, including Notch pathway and immune checkpoint inhibitors.

Our sequencing studies did not reveal any somatic events that may cause the transcriptional discrepancy observed in LCNEC and SCLC tumors with *TP53* and *RB1* alteration, which raises the

question if all neuroendocrine tumors share the same cell of origin. It remains to be understood whether distinct tumor-specific cell of origins or cellular processes allow for plasticity and trans-differentiation that consequently lead to distinct molecular phenotypes. Importantly, histological trans-differentiation from lung adenocarcinoma to SCLC has been observed, both spontaneously or as resistance mechanisms to kinase inhibitors[44,45]; in some cases these were linked with a loss of *RB1*[4,46]. Previous studies involving genetically engineered mouse models and human cell lines have emphasized the phenomenon of transcriptional heterogeneity in SCLC and pointed to discordant expression of key lineage factors (e.g. *ASCL1*, *NEUROD1*, *REST*)[31,38]. By contrast, human primary tumors revealed a more complex expression pattern with co-expression of these transcriptional regulators. As a limitation of bulk tumor sequencing, advances in single cell sequencing may further aid to resolve and study the level of transcriptional intra-tumor heterogeneity in high-grade neuroendocrine tumors. While our studies pointed to transcriptional correlates of genomically defined subsets in LCNECs (type I and type II LNCECs), additional analyses on a larger dataset are warranted to further interrogate subcategories of high-grade neuroendocrine tumors.

In summary, we provide the first comprehensive characterization of neuroendocrine lung tumors, which integrates the molecular phenotypes of less frequent lung tumor subtypes. Despite the fact that LCNEC and SCLC tumors share some common clinical and histological characteristics, our study emphasizes pronounced differences in the pattern of genomic alterations and in their transcriptome profiles. The precise distinction of type I and type II LCNECs from SCLC is consequently pivotal to evaluate the response of patients to treatment options and to further understand morphological trans-differentiation processes in lung cancer patients.

Methods

Human specimens. The institutional review board (IRB) of the University of Cologne approved this study. Patient samples were obtained under IRB-approved protocols following written informed consent from all human participants. We collected and analyzed fresh-frozen samples of 75 LCNEC patients, which were provided by multiple collaborating institutions; 42 tumors were previously subject of other studies conducted by Rousseaux et al.[47] ($n = 25$) and Seidel et al.[7] ($n = 37$) (Supplementary Data 1). Clinical data were available for most patients, who were predominantly male (approximate ratio of 4:1) and current or former heavy smokers (Supplementary Data 1). All tumor samples were reviewed and confirmed by independent expert pathologists (E.B., W.T., and R.B.), and the diagnosis of LCNEC and the assessment of combined histological components were confirmed by H&E staining and immunohistochemistry, including markers for chromogranin A, synaptophysin, CD56 and Ki67. All tumors were positive for at least one neuroendocrine differentiation marker (Supplementary Data 1–2). Specimens containing >70% of tumor cells were processed for DNA and RNA extractions. DNA was extracted from matching normal material that was provided in the form of blood or adjacent non-tumorigenic lung tissue, which through pathological evaluation was confirmed to be free of tumor contaminants.

Nucleic acid extraction. Total DNA and RNA were obtained from fresh-frozen tumor tissue and matched fresh-frozen normal tissue or blood. Depending on the size of the tissue, 15–30 sections, each 20 µm thick, were cut using a cryostat (Leica) at −20 °C. The matched normal sample obtained from frozen tissue was processed the same way. Nineteen LCNEC cases were identified with mixed histological components of SCLC, lung adenocarcinomas and squamous cell carcinomas (Supplementary Data 1); in these cases nucleic acids were extracted from pure LCNEC regions by only dissecting the LCNEC component. DNA was extracted with the Gentra Puregene DNA extraction kit (Qiagen) and diluted to a working concentration of 100 ng/µL. The DNA was analyzed by agarose gel electrophoresis and confirmed to be of high-molecular weight (>10 kb). The DNA of tumor and normal material was confirmed to originate from the same patient by short tandem repeat (STR) analysis which was conducted at the Institute of Legal Medicine at the University of Cologne (Cologne, Germany), or by subsequent Affymetrix 6.0 SNP array and sequencing analyses.

RNA was isolated from tumor tissues by first lysing and homogenizing tissue sections with the Tissue Lyzer (Qiagen). The RNA was then extracted with the Qiagen RNAeasy Mini Kit. The RNA quality was analyzed at the Bioanalyzer 2100 DNA Chip 7500 (Agilent Technologies) and cases with a RNA integrity number (RIN) of over seven were considered for RNA-seq experiments.

Next-generation sequencing (NGS). WES was performed by first fragmenting 1 µg of DNA (Bioruptor, diagenode, Liége, Belgium). The DNA fragments were then end-repaired and adaptor-ligated with sample index barcodes. Following size selection, the SeqCap EZ Human Exome Library version 2.0 kit (Roche NimbleGen, Madison, WI, USA) was used to enrich for the whole exome. The DNA libraries were then sequenced with a paired-end 2 × 100 bp protocol aiming for an average coverage of 90× and 120× for the normal and tumor DNA, respectively. The primary data were filtered for signal purity with the Illumina Realtime Analysis software.

WGS was performed with a read length of 2 × 100 bp. The samples were processed to provide 110 Gb of sequence, thus amounting to a mean coverage of 30× for both tumor and matched normal.

For RNA-seq, cDNA libraries were prepared from PolyA + RNA following the Illumina TruSeq protocol for mRNA (Illumina, San Diego, CA, USA). The libraries were sequenced with a paired-end 2 × 100 bp protocol resulting in 8.5 Gb per sample, and thus in a 30× mean coverage of the annotated transcriptome.

Whole genome, whole exome and transcriptome sequencing reactions were performed on an Illumina HiSeq 2000 sequencing instrument (Illumina, San Diego, CA, USA).

Copy-number analysis by Affymetrix SNP 6.0 arrays. Human DNA from fresh-frozen tumors was analyzed with Affymetrix Genome-Wide Human SNP 6.0 arrays to determine copy-number alterations. Raw copy number data were computed by dividing tumor-derived signals by the mean signal intensities obtained from a subset of normal samples which were hybridized to the array in the same batch. Circular binary segmentation was applied to obtain segmented raw copy numbers[48]. Significant copy-number alterations were assessed with CGARS[49] at a threshold of $Q < 0.01$ (Supplementary Data 4).

Data processing and analyses of DNA sequencing data. The sequencing reads were aligned to the human reference genome NCBI build 37 (NCBI37/hg19) with BWA (version 0.6.1-r104)[50]. Possible PCR-duplicates were masked and not included for subsequent studies. We applied our in-house analysis pipeline[4,6,51] to analyze the data for somatic mutations, copy number alterations and genomic rearrangements. In brief, the mutation calling algorithm considers local sequencing depth, forward-reverse bias, and global sequencing error, to thus determine the presence of a mutated allele. We determined the somatic status of these mutations by assessing the absence of these variants in the sequencing data of the matched normal.

We determined genomic rearrangements from WGS data of 11 human LCNECs following the procedure as previously described[6,51]. In brief, the sequencing data were analyzed for discordant read-pairs, which were not within the expected mapping distance (>600 base pairs) or which revealed an incorrect orientation. Discordant read-pairs were analyzed for breakpoint-spanning reads, in which one read-pair shows partial alignments to two distinct genomic loci. Rearranged genomic loci were then reported at instances where at least one breakpoint-spanning read was identified. The genomic rearrangements called from each tumor sample were further filtered against the sequencing data of a matched normal and additionally against a library of normal genomes to thus minimize the detection of false-positive rearrangements.

Significantly mutated genes were analyzed as previously described[4,6]. In brief, we first determined the overall background mutation rate of each gene by computing its expected number of mutations assuming that all mutations are uniformly distributed across the genome. We also considered the ratio of synonymous to non-synonymous mutations into a combined statistical model to determine significantly mutated genes. Since mutation rates in non-expressed genes are often higher than the genome-wide background rate, we furthermore filtered for the expression of genes by referring to the transcriptome sequencing data of LCNECs. Only genes with a median FPKM (Fragments Per Kilobase Million) value of >1 in at least 35 out of 60 samples were considered (Methods section: RNA sequencing data processing and analyses). The significance of recurrently mutated genes was determined at a Q-value of <0.01 (Supplementary Data 7). Following previously described methods, we furthermore analyzed the data for significant enrichment of damaging mutations (including splice site, non-sense, and frameshift mutations)[6] and for significant clustering of mutations in genomic hotspots following a re-sampling based approach[4]. Significance was determined at a Q-value of 0.01, if the gene was affected in >10% of the samples (Supplementary Data 7). The damaging impact of mutations was further assessed by Polyphen[52].

The clonal status of mutations was assessed by computing for every mutation the "cancer cell fraction" (CCF), which defines within a tumor the fraction of cancer cells harboring that particular mutation[53]. The CCF was computed following our previously described approach[6]. In brief, this method first estimates tumor purity, ploidy, and absolute copy numbers, and computes for each mutation in a given sample the expected allele frequency under the assumption of clonality. The CCF is the quotient of the observed allelic fraction and the expected allelic fraction of a mutation. The distribution of CCFs for every mutation in a sample allowed to further identify distinct clusters and to thus assign the mutations to clonal and subclonal populations. The analysis described in Supplementary Fig. 2c considers mutations, which were assigned to clonal and subclonal fractions with a

probability >90%. In consideration of the sequencing coverage and the overall distribution of CCFs of every mutation in a sample, we furthermore determined the significant enrichment of mutations in a subclone at a *P*-value of 0.01 (Fig. 1b).

Mutational signatures analyses. Mutational signatures were analyzed in lung cancer subtypes applying previously described methods[54,55] and referring to the datasets of 77 lung adenocarcinomas (50 heavy-smokers (hs) and 27 non-smokers (ns) from the TCGA project)[2,25], 52 lung squamous cell carcinomas (from the TCGA project)[3], 109 SCLC[6], and 60 LCNECs from this study. Tumor cases with at least 30 somatic variants were selected and the list of variants were either extracted from Supplementary Materials[6] or COSMIC v68 (for the TCGA data)[20]. Variants were annotated with Annovar (version 12 Nov 2014). Gene strand orientations were retrieved from the RefSeqGene database using a customized Perl script. Variants were included in the analyses only if they could be successfully annotated. Single-base substitutions were classified into 96 types determined by the six possible substitutions (C:G > A:T, C:G > G:C, C:G > T:A, A:T > C:G, A:T > G:C, A:T > T:A) in their tri-nucleotides sequence context (16 combinations for each type of substitution). For extracting mutational signatures, we used the non-negative matrix factorization (NMF) algorithm developed by Lee et al.[56] and implemented in the Welcome Trust Sanger Institute (WTSI) mutational signatures framework.

Di-deoxynucleotide sequencing. Somatic alterations of interest were determined and confirmed by two independent sequencing approaches, which included WGS, WES, RNA-seq or di-deoxynucleotide sequencing. Di-deoxynucleotide chain termination sequencing (Sanger sequencing) was performed to validate mutations, genomic rearrangements, and chimeric fusion transcripts. Primer pairs were designed to amplify the target region encompassing the somatic alteration. The PCR reactions were performed either with genomic DNA or cDNA. The amplified products were subjected to Sanger sequencing and the respective electropherograms were analyzed by visual inspection using 4 Peaks or Geneious.

Analysis of RNA sequencing data. In order to detect chimeric transcripts, RNA-seq data were processed using TRUP[4,27]. In brief, paired-end RNA-seq reads were aligned to the human reference genome (NCBI37/hg19). We used TRUP to identify potential chimeric transcripts. Gene expression levels were determined with Cufflinks v2.0.2 referring only to paired-end reads that uniquely mapped within the expected mapping distance. The expression was quantified as FPKM (Fragments Per Kilobase Million) and the expression values served as a filter for identifying significantly mutated genes (Methods section: Data processing and analyses of DNA sequencing data).

Gene expression profiling and clustering studies. We analyzed transcriptome sequencing data from a total of $n = 341$ lung cancer samples. $N = 221$ samples referred to the data generated at the University of Cologne, Department of Translational Genomics, which included 41 lung adenocarcinoma[26,27], 61 pulmonary carcinoids[24], 53 SCLC[6], and 66 LCNECs from this present study. $N = 120$ samples were randomly selected from both the TCGA lung squamous cell carcinoma ($n = 60$)[3] and TCGA lung adenocarcinoma ($n = 60$) cohorts[2,25] referring to the Genomics Data Commons Legacy Archive. Sequencing data of lung adenocarcinomas from two different platforms aided in controlling for potential batch effects in subsequent studies. The raw sequencing reads of the RNA-seq data were all similarly processed to analyze for gene expression profiles. Sequencing reads which passed the quality control were mapped to the human reference genome (hg19) using MapSplice[57]. Picard Tools v1.64 (http://broadinstitute.github.io/picard/) was used to assess the alignment profile. SAMtools was used to sort and index the mapped reads and to determine transcriptome coordinates. The aligned reads were further filtered for indels, large inserts, and zero mapping quality with UBU v1.0 (https://github.com/mozack/ubu). RSEM[58], an expectation-maximization algorithm that refers to UCSC gene transcript and definitions, was applied to estimate transcript abundance. In order to allow comparisons between all RNA-Seq samples, raw RSEM read counts were normalized to the overall upper quartile[59]. The expression was quantified for 20,500 genes in 341 tumor samples and the median expression value was determined at RSEM = 209, which served as a reference threshold to classify for low and high expression. The expression determined by RSEM is provided for LCNECs in Supplementary Data 11.

For clustering purposes a set of genes that were both highly expressed and had highly variable expression patterns was identified in all lung cancer subtypes. Quality control procedures performed prior to any clustering analysis did not detect any evidence of batch effects.

After median centering the $\log_2(\text{RSEM} + 1)$ values by gene, unsupervised consensus clustering was applied using the ConsensusClusterPlus R package[60,61] with partitioning around medioids and a Spearman correlation-based distance. Additional hierarchical clustering of the consensus clustering classes was performed, applying average linkage and a Pearson correlation-based distance.

The statistical significance of the differences in gene expression patterns present in the subtype was assessed with the SigClust R package[62] by referring to the clustering gene sets and by using 1000 permutations and the default covariance estimation method. ClaNC[63] was used to identify genes whose expression patterns characterize the subtypes. R 3.0.2[61] was used to perform all statistical analyses and create all figures.

We first conducted consensus clustering of all lung cancer subtypes. The expression data of all lung cancer subtypes ($n = 341$) was analyzed and the 0.75 quantile of all $\log_2(\text{mean(RSEM)})$ values was used to identify highly expressed genes, while the 0.9 quantile of $\log_2(\text{variance(RSEM)})$ was used as a threshold to identify clustering gene sets that have highly variable expression patterns, which yielded a set of 1854 genes (Supplementary Fig. 6a). The samples were clustered with ConsensusClusterPlus following partition around medoids (PAM), and the ConsensusClusterPlus output along with gene expression heatmaps, principal components analysis, and silhouette plots was analyzed. Manual review of ConsensusClusterPlus output suggested a possible clustering solution based on $k = 6$ groups. However, two of the six groups included mainly lung adenocarcinoma samples and the gene expression heatmaps and PCA plots showed that these groups were quite similar. Thus, we chose to collapse these groups, thereby producing a five-class solution. The consensus clusters highly correlated with the histological subtypes as determined by Fisher's exact test Monte Carlo version ($P < 0.001$, 10,000 permutations): class A ($n = 66$; enriched for pulmonary carcinoids), class B ($n = 65$, enriched for lung squamous cell carcinomas), class C ($n = 108$, enriched for lung adenocarcinomas; data generated by different institutes), class D ($n = 38$, enriched for LCNEC and SCLC cases), and class E ($n = 64$, enriched for SCLC and LCNEC cases) (Supplementary Fig. 6b, Supplementary Data 12). ClaNC led to the identification of 875 classifier genes, which are displayed in the expression heatmaps (Fig. 2, Supplementary Fig. 6–7, Supplementary Data 13).

We then conducted consensus clustering of LCNECs, SCLC, lung adenocarcinomas, and squamous cell carcinomas. The unsupervised clustering approach was repeated for a subset of lung cancer subtypes; here excluding pulmonary carcinoids. The feature selection of highly variable (0.75 quantile) and highly expressed (0.9 quantile) genes across these lung tumor subtypes ($n = 280$) involved a gene set of 1855 genes and the consensus clustering process through hierarchical clustering suggested the presence of three expression clusters (expression subtypes): class A ($n = 98$, enriched for lung adenocarcinomas), class B ($n = 115$, enriched for LCNEC and SCLC), and class C ($n = 67$, enriched for lung squamous cell carcinomas). ClaNC identified 300 classifier genes which are displayed in the respective expression heatmaps (Supplementary Fig. 9).

We performed consensus clustering of LCNEC and SCLC through unsupervised clustering of the expression data of LCNEC and SCLC tumors alone ($n = 119$). Exploratory analyses of the gene expression data suggested the use of the 0.9 quantile of both the $\log_2(\text{mean(RSEM)})$ and $\log_2(\text{variance(RSEM)})$ values as thresholds for highly expressed and highly variably expressed genes. This produced a set of 1416 clustering genes. The Consensus clustering approach included hierarchical clustering and yielded four gene expression subtypes: class I ($n = 19$, only LCNECs), class II ($n = 49$, LCNEC and some SCLC tumors), class III ($n = 10$, SCLC and some LCNECs), and class IV ($n = 41$, mainly SCLC and some LCNECs) (Fig. 3, Supplementary Fig. 10–11, Supplementary Data 12). Hierarchical clustering of these cases revealed two main subgroups: one mainly formed by class I and II (enriched for LCNECs) and one mainly formed by class III and IV (enriched for SCLC) (Supplementary Fig. 11). 300 classifier genes were identified by ClaNC and are displayed in the expression heatmaps (Fig. 3, Supplementary Fig. 11, Supplementary Data 13).

We also performed consensus clustering of LCNECs with lung adenocarcinomas or lung squamous cell carcinomas. A gene set of (a) 1335 and (b) 1338 highly variable (0.85 quantile) and expressed genes (0.925 quantile) was identified in subsets of lung cancer tumors, including (a) LCNECs and lung adenocarcinomas ($n = 167$) and (b) LCNECs and lung squamous cell carcinomas ($n = 126$). The consensus clustering approach through PAM (partitioning around medoids) suggested in both cases two transcriptional subclasses: for approach (a) class A ($n = 70$, mainly LCNECs) and class B ($n = 97$, mainly lung adenocarcinomas); and for approach (b) class A ($n = 58$, mainly LCNECs) and class B ($n = 68$, mainly lung squamous cell carcinomas). ClaNC identified 100 classifier genes in each approach, which were used for the expression heatmaps (Supplementary Fig. 9).

We furthermore performed consensus clustering of LCNECs alone. The transcriptional data on LCNECs was analyzed and hierarchical clustering referred to 475 very highly expressed (0.875 quantile) and very highly variable (0.975 quantile) genes. The consensus clustering approach yielded a $k = 4$ clustering solution: class 1 ($n = 11$), class 2 ($n = 21$), class 3 ($n = 24$), and class 4 ($n = 10$). ClaNC was then applied to the clustering solution, which further identified 540 classifier genes (Supplementary Fig. 13, Supplementary Data 13).

Differential expression analysis. The SAMR R package[64] was used to identify genes that were differentially expressed in the expression subtypes using 1000 permutations and a *Q*-value threshold of 0.05 (Supplementary Data 13). We then used the DAVID annotation database[65,66] to identify pathways that were enriched for differentially expressed genes at $P < 0.0001$ (Supplementary Data 14).

Immunohistochemistry. FFPE tissue sections of 3-µm thickness were stained for hematoxylin and eosin (H&E) and immunohistochemistry (IHC) was conducted for CD56 (*NCAM1*), Synaptophysin (*SYP*), Chromogranin A (*CHGA*, clone DAK-A3), TTF-1 (*NKX2-1*, clone 8G7G3/1), and Rb1 (*RB1*, clone 1F8 (ab81701; Abcam,

Cambridge, UK) (Supplementary Data 2, Supplementary Table 1). Hematoxylin and eosin (H&E) were scanned and can be viewed online or with the Pannoramic Viewer software (3D Histech) as specified in Supplementary Data 2 (for further information see "Data Availability").

Specifically, IHC for Rb1 was performed with the Novolink max polymer detection system (RE7280-CE, Leica Biosystems, Wetzlar, Germany) using EDTA buffer pH 8.0 (K038, Diagnostic BioSystems, Pleasanton, USA) antigen retrieval (4 × 5 min by microwave 700 W). The primary antibody was incubated overnight at 4 °C; the secondary antibody was incubated for 30 min at room temperature. The signal was visualized by diaminobenzidine after incubation for 5 min at room temperature. Sections were counter-stained with hematoxylin for 5 min. The *H*-score method was used for evaluating the immunostaining with Rb1 by multiplying the intensity of the staining (0: no staining, 1: weak, 2: moderate and 3: strong staining) with the percentage of the tumor or stroma stained. The minimum score was 0 and the maximum was 300 (Supplementary Data 2).

Fluorescence in situ hybridization assay. Genomic rearrangements of *PTK6* on chromosome 20 were assessed through a dual-color break-apart fluorescence in situ hybridization (FISH) assay following previous protocols[67]. In brief, the BAC clone RP11-939M14 labeled centromeres with biotin (red signal) and CTD-3228E10 labeled telomeric sites with digoxigenin (green signal). The samples were analyzed with a 63× oil immersion objective at a fluorescence microscope (Zeiss, Jena, Germany) equipped with appropriate filters, a charge-coupled device camera and the FISH imaging and capturing software Metafer 4 (Metasystems, Altlussheim, Germany). Two independent scientists analyzed the experiment (R.M. and S.P.). Translocations were derived from a split of a signal pair, resulting in a single red and green signal, single red or green signals resulting from signal loss, were referred to as a rearrangement through deletion. In cases where cells were wild type and displayed no rearrangements, a juxtaposed red and green signal (mostly forming a yellow signal) was observed.

NTRK1 break-apart FISH were performed with the ZytoLight SPEC *NTRK1* Dual Color Break Apart Probe (ZytoVision, Bremerhaven, Germany). According to previous protocols[68], 4 µm sections of FFPE tissue were treated with the Paraffin pretreatment reagent kit (Vysis, Abbott Molecular), and then stained with the probes following the instructions of the manufacturer. An *NTRK1* rearrangement was diagnosed when >15% of the nuclei showed either a split pattern with 3′ and 5′ signals separated by a distance superior to the diameter of the largest signal, or isolated 3′ (orange) signals.

Data availability. Sequencing data and Affymetrix 6.0 SNP array data are deposited at the European Genome-phenome Archive, which is hosted by the EBI (EGA, http://www.ebi.ac.uk/ega/), under accession number EGAS00001000708. Histological images of FFPE samples from LCNECs of this study are deposited as H&E images (domain 1: https://teleslide.chu-grenoble.fr/ > acces libre > recherche > recherche/TP/LCNEC-study > code access 1793) or as data files compatible with the Pannoramic Viewer software (3D Histech) (domain 2: https://uni-koeln.sciebo.de/index.php/s/xMjs4dqJpqbOVDn); an overview is provided in Supplementary Data 2.

Received: 10 April 2017 Accepted: 18 January 2018
Published online: 13 March 2018

References

1. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
2. Collisson, Ea et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
3. Hammerman, P. S. et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
4. Peifer, M. et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.* **44**, 1104–1110 (2012).
5. Rudin, C. M. et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.* **44**, 1111–1116 (2012).
6. George, J. et al. Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47–53 (2015).
7. Seidel, D. A genomics-based classification of human lung tumors. *Sci. Transl. Med.* **5**, 209ra153 (2013).
8. Bhattacharjee, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA* **98**, 13790–13795 (2001).
9. Hayes, D. N. et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J. Clin. Oncol.* **24**, 5079–5090 (2006).
10. Wilkerson, M. D. et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin. Cancer Res.* **16**, 4864–4875 (2010).
11. Chen, F. et al. Multiplatform-based molecular subtypes of non-small-cell lung cancer. *Oncogene* **36**, 1384–1393 (2017).
12. Travis, W. D. Advances in neuroendocrine lung tumors. *Ann. Oncol.* **21**, vii65–71 (2010).
13. Travis, W. D. et al. The 2015 World Health Organization Classification of lung tumors. *J. Thorac. Oncol.* **10**, 1243–1260 (2015).
14. Fasano, M. et al. Pulmonary large-cell neuroendocrine carcinoma: from epidemiology to therapy. *J. Thorac. Oncol.* **10**, 1133–1141 (2015).
15. Karlsson, A., Brunnström, H., Lindquist, K. E. & Jirström, K. Mutational and gene fusion analyses of primary large cell and large cell neuroendocrine lung cancer Patient material. *Oncotarget* **6**, 22028–22037 (2015).
16. Rekhtman, N. et al. Next-generation sequencing of pulmonary large cell neuroendocrine carcinoma reveals small cell carcinoma-like and non-small cell carcinoma-like subsets. *Clin. Cancer Res.* **22**, 3618–3629 (2016).
17. Miyoshi, T. et al. Genomic profiling of large-cell neuroendocrine carcinoma of the lung. *Clin. Cancer Res.* **23**, 757–765 (2017).
18. Jones, M. H. et al. Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet* **363**, 775–781 (2004).
19. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).
20. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2014).
21. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
22. Weiss, J. et al. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. *Sci. Transl. Med.* **2**, 62ra93 (2010).
23. Wistuba, I. I., Gazdar, A. F. & Minna, J. D. Molecular genetics of small cell lung carcinoma. *Semin. Oncol.* **28**, 3–13 (2001).
24. Fernandez-Cuesta, L. et al. Frequent mutations in chromatin-remodeling genes in pulmonary carcinoids. *Nat. Commun.* **5**, 3518 (2014).
25. Imielinski, M. et al. Mapping the Hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
26. Fernandez-Cuesta, L. et al. CD74-NRG1 fusions in lung adenocarcinoma. *Cancer Discov.* **4**, 415–422 (2014).
27. Fernandez-Cuesta, L. et al. Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol.* **16**, 7 (2015).
28. Rooney, M. S., Shukla, Sa, Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
29. Augustyn, A. et al. ASCL1 is a lineage oncogene providing therapeutic targets for high-grade neuroendocrine lung cancers. *Proc. Natl Acad. Sci. USA* **111**, 14788–14793 (2014).
30. Westerman, B. A. et al. Basic helix-loop-helix transcription factor profiling of lung tumors shows aberrant expression of the proneural gene atonal homolog 1 (ATOH1, HATH1, MATH1) in neuroendocrine tumors. *Int. J. Biol. Markers* **22**, 114–123 (2007).
31. Borromeo, M. D. et al. ASCL1 and NEUROD1 reveal heterogeneity in pulmonary neuroendocrine tumors and regulate distinct genetic programs. *Cell Rep.* **16**, 1259–1272 (2016).
32. Sutherland, K. D. et al. Cell of origin of small cell lung cancer: inactivation of Trp53 and Rb1 in distinct cell types of adult mouse lung. *Cancer Cell* **19**, 754–764 (2011).
33. Park, K. et al. Characterization of the cell of origin for small cell lung cancer. *Cell Cycle* **10**, 2806–2815 (2011).
34. Song, H. et al. Functional characterization of pulmonary neuroendocrine cells in lung development, injury, and tumorigenesis. *Proc. Natl Acad. Sci. USA* **109**, 17531–17536 (2012).
35. Sugano, M., Nagasaka, T. & Sasaki, E. HNF4 a as a marker for invasive mucinous adenocarcinoma of the lung. *Am. J. Surg. Pathol.* **37**, 211–218 (2013).
36. Snyder, E. L. et al. Article Nkx2-1 represses a latent gastric differentiation program in lung adenocarcinoma. *Mol. Cell* **50**, 185–199 (2013).
37. Saunders, L. R. et al. A DLL3-targeted antibody-drug conjugate eradicates high-grade pulmonary neuroendocrine tumor-initiating cells in vivo. *Sci. Transl. Med* **7**, 302ra136 (2015).
38. Lim, J. S. et al. Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. *Nature* **545**, 360–364 (2017).

39. Kazarian, M. & Laird-Offringa, Ia Small-cell lung cancer-associated autoantibodies: potential applications to cancer diagnosis, early detection, and therapy. *Mol. Cancer* **10**, 33 (2011).
40. Ranganathan, P., Weaver, K. L. & Capobianco, A. J. Notch signalling in solid tumours: a little bit of everything but not all the time. *Nat. Rev. Cancer* **11**, 338–351 (2011).
41. Pietanza, M. C. et al. Safety, activity, and response durability assessment of single agent rovalpituzumab tesirine, a delta-like protein 3 (DLL3)-targeted antibody drug conjugate (ADC), in small cell lung cancer (SCLC). *Eur. J. Cancer*. **51**, S712 (2015).
42. Yen, W. C. et al. Targeting notch signaling with a Notch2/Notch3 antagonist (Tarextumab) inhibits tumor growth and decreases tumor-initiating cell frequency. *Clin. Cancer Res.* **21**, 2084–2095 (2015).
43. Pietanza, M. C. et al. Final results of phase Ib of tarextumab (TRXT, OMP-59R5, anti-Notch2/3) in combination with etoposide and platinum (EP) in patients (pts) with untreated extensive-stage small-cell lung cancer (ED-SCLC). *J. Clin. Oncol.* 33, 7508 (2015).
44. Zakowski, M. F., Ladanyi, M. & Kris, M. G. EGFR mutations in small-cell lung cancers. *N. Engl. J. Med.* **355**, 213–215 (2006).
45. Morinaga, R. et al. Sequential occurrence of non-small cell and small cell lung cancer with the same EGFR mutation. *Lung Cancer* **58**, 411–413 (2007).
46. Niederst, M. J. et al. RB loss in resistant EGFR mutant lung adenocarcinomas that transform to small-cell lung cancer. *Nat. Commun.* **6**, 6377 (2015).
47. Rousseaux, S. et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung. *Cancers* **5**, 1–12 (2013).
48. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
49. Lu, X., Thomas, R. K. & Peifer, M. CGARS: cancer genome analysis by rank sums. *Bioinformatics* **30**, 1295–1296 (2014).
50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
51. Fernandez-Cuesta, L. et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat. Commun.* **5**, 3518 (2014).
52. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations a. *Nature* **7**, 248–249 (2010).
53. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
54. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
55. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
56. Lee, S. Y., Song, H. A. & Amari, S. I. A new discriminant NMF algorithm and its application to the extraction of subtle emotional differences in speech. *Cogn. Neurodyn.* **6**, 525–535 (2012).
57. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, 1–14 (2010).
58. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
59. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinforma.* **11**, 94 (2010).
60. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
61. R Core Team, R. F. for S. C. *R: A language and environment for statistical computing*. (2014). Available at http://www.r-project.org/
62. Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. Statistical significance of clustering for high-dimension, low–sample size dataset. *J. Am. Stat. Assoc.* **103**, 1281–1293 (2008).
63. Dabney, A. R. Classification of microarrays to nearest centroids. *Bioinformatics* **21**, 4148–4154 (2005).
64. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
65. Huang, D. W. & Lempicki, R. A. & Sherman, B. T. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
66. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
67. Menon, R. et al. Somatic copy number alterations by whole-exome sequencing implicates YWHAZ and PTK2 in castration-resistant prostate cancer. *J. Pathol.* **231**, 505–516 (2013).
68. McLeer-Florin, A. et al. Dual IHC and FISH testing for ALK gene rearrangement in lung adenocarcinomas in a routine practice. *J. Thorac. Oncol.* **7**, 348–354 (2012).

Acknowledgements

This work was supported by the German Cancer Aid (Deutsche Krebshilfe) as part of the small cell lung cancer genome sequencing consortium (grant ID: 109679 to R.K.T., M.P., R.B., P.N., M.V., and S.A.H.), by the German Ministry of Science and Education (BMBF) as part of the e:Med program (grant no. 01ZX1303A to R.K.T., R.B., U.L., M.P. and J. W, and grant no. 01ZX1406 to M.P.), by the EU-Framework program CURELUNG (HEALTH-F2-2010-258677 to R.K.T., J.W., E.B., and L.R.), by the Deutsche Forschungsgemeinschaft (DFG; through TH1386/3-1 to R.K.T.), by the Deutsche Krebshilfe as part of the Oncology Centers of Excellence funding program (R.K.T.), by the National Institute of Health (NIH U10CA181009 to D.N.H.), by the German Cancer Consortium (DKTK) Joint Funding program, by Associazione Italiana per la Ricerca sul Cancro (AIRC, IG 16847 to L.R.), and by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396 (L.B.A.). J. G. received funding as part of the IASLC Young Investigator award. L.B.A. is supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. We are indebted to the patients donating their tumor specimens as part of the Clinical Lung Cancer Genome Project initiative. We thank the regional computing center of the University of Cologne (RRZK) for providing the CPU time on the DFG-funded supercomputer 'CHEOPS', as well as the support. We would like to acknowledge that Australian specimens were provided with assistance of the Victorian Cancer Biobank. We furthermore thank Johannes Berg, Chau Nguyen, Philipp Lorimier, Elisabeth Kirst, and César Tejerina Álvarez for their technical assistance.

Author contributions

R.K.T., L.F.-C., J.G., and E.B. conceived and designed the project. J.G., V.W., S.P., S.A.H., M.F., D.N.H., W.D.T., L.F.C., E.B., and R.K.T. supervised the work and gave scientific input. J.G., V.W., L.B.A., L.M., T.M.D., M.A., No.L., M.P., G.B., R.S., A.D.R., M.G.S., M. D.W., S.A.H., M.O., Y.C., and L.F.-C. performed computational and statistical analyses. A.M.F., F.M., R.M., F.L., C.M., I.D., D.S., P.S., J.A., C.B., and M.B. performed experiments. R.B., W.D.T., and E.B. performed pathological review. D.M.S., C.G.B., S.L., A.S., W.W., V.T., O.T.B., M.L.I., A.H., S.S., S.A., G.W., B.S., L.R., U.P., I.P., J.H.C., J.S., R.B., Ni. L., and E.B. contributed with samples. D.S., G.B., F.L., L.M., Ni.L., V.A., U.L., P.N., P.M. S., J.D.M., J.W., M.V., and T.Z. helped with logistics. J.G., R.K.T., and L.F.-C. wrote the manuscript, which was reviewed by all the co-authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03099-x.

Competing interests: L.F.-C. and R.K.T. are inventors on a patent application related to findings described in this manuscript. R.K.T. is a founder of NEO New Oncology GmbH, now part of Siemens Healthcare. R.K.T. received consulting and lecture fees (Merck, Roche, Lilly, Boehringer Ingelheim, AstraZeneca, Daiichi-Sankyo, MSD, NEO New Oncology, Puma, Clovis). R.B. is a cofounder and owner of Targos Molecular Diagnostics and received honoraria for consulting and lecturing from AstraZeneca, Boehringer Ingelheim, Merck, Roche, Novartis, Lilly, and Pfizer. J.W. received consulting and lecture fees from Roche, Novartis, Boehringer Ingelheim, AstraZeneca, Bayer, Lilly, Merck, Amgen and research support from Roche, Bayer, Novartis, Boehringer Ingelheim. T.Z. received honoraria from Roche, Novartis, Boehringer Ingelheim, Lilly, Merck, Amgen and research support from Novartis. B.S. received consulting fees from AstraZeneca, Roche-Genentech, Pfizer, Novartis, Merck, and Bristol Myers Squibb. The remaining authors declare no competing financial interest.

Julie George[1], Vonn Walter[2,3], Martin Peifer[1,4], Ludmil B. Alexandrov[5], Danila Seidel[1], Frauke Leenders[1], Lukas Maas[1], Christian Müller[1], Ilona Dahmen[1], Tiffany M. Delhomme[6], Maude Ardin[7], Noemie Leblay[6], Graham Byrnes[8], Ruping Sun[9], Aurélien De Reynies[10], Anne McLeer-Florin[11], Graziella Bosco[1], Florian Malchers[1], Roopika Menon[12], Janine Altmüller[4,13,14], Christian Becker[13], Peter Nürnberg[4,13,15], Viktor Achter[16], Ulrich Lang[16,17], Peter M. Schneider[18], Magdalena Bogus[18], Matthew G. Soloway[2], Matthew D. Wilkerson[19], Yupeng Cun[1,4], James D. McKay[6], Denis Moro-Sibilot[20], Christian G. Brambilla[20], Sylvie Lantuejoul[21,22], Nicolas Lemaitre[21], Alex Soltermann[23], Walter Weder[24], Verena Tischler[23], Odd Terje Brustugun[25,26], Marius Lund-Iversen[27], Åslaug Helland[24,25], Steinar Solberg[28], Sascha Ansén[29], Gavin Wright[30], Benjamin Solomon[31], Luca Roz[32], Ugo Pastorino[33], Iver Petersen[34], Joachim H. Clement[35], Jörg Sänger[36], Jürgen Wolf[29], Martin Vingron[9], Thomas Zander[37,38], Sven Perner[39], William D. Travis[40], Stefan A. Haas[9], Magali Olivier[7], Matthieu Foll[6], Reinhard Büttner[38], David Neil Hayes[2], Elisabeth Brambilla[21], Lynnette Fernandez-Cuesta[1,6] & Roman K. Thomas[1,38,41]

[1]Department of Translational Genomics, Center of Integrated Oncology Cologne-Bonn, Medical Faculty, University of Cologne, Cologne, 50931, Germany. [2]UNC Lineberger Comprehensive Cancer Center School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7295, USA. [3]Department of Biochemistry and Molecular Biology, Penn State Milton S. Hershey Medical Center, 500 University Drive, Hershey, PA 17033, USA. [4]Center for Molecular Medicine Cologne (CMMC), University of Cologne, 50931 Cologne, Germany. [5]Department of Cellular and Molecular Medicine and Department of Bioengineering and Moores Cancer Center, University of California, La Jolla, San Diego, CA 92093, USA. [6]Genetic Cancer Susceptibility Group, Section of Genetics, International Agency for Research on Cancer (IARC-WHO), Lyon, 69008, France. [7]Molecular Mechanisms and Biomarkers Group, Section of Mechanisms of Carcinogenesis, International Agency for Research on Cancer (IARC-WHO), 69008 Lyon, France. [8]Section of Environment and Radiation, International Agency for Research on Cancer (IARC-WHO), 69008 Lyon, France. [9]Computational Molecular Biology Group, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany. [10]Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, 14 rue Corvisart, Paris, 75013, France. [11]CHU Grenoble Alpes, UGA/INSERM U1209/CNRS, Grenoble, France. [12]NEO New Oncology GmbH, 51105 Cologne, Germany. [13]Cologne Center for Genomics (CCG), University of Cologne, 50931 Cologne, Germany. [14]Institute of Human Genetics, University Hospital Cologne, 50931 Cologne, Germany. [15]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, 50931 Cologne, Germany. [16]Computing Center, University of Cologne, 50931 Cologne, Germany. [17]Department of Informatics, University of Cologne, 50931 Cologne, Germany. [18]Institute of Legal Medicine, University Hospital Cologne, 50823 Cologne, Germany. [19]Department of Genetics, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, NC, 27599-7295, USA. [20]CHUGA Grenoble, INSERM U 1209, University Grenoble Alpes, Institute of Advanced Biosciences (IAB), 38043, CS10217 Grenoble, France. [21]Department of Pathology, CHUGA, INSERM U 1209, University of Grenobles Alpes, Institute of Advanced Biosciences (IAB), 38043, CS10217 Grenoble, France. [22]Department of Biopathology, Centre Léon Bérard UNICANCER, 69008 Lyon, France. [23]Institute of Pathology and Molecular Pathology, University Hospital Zurich, 8091 Zurich, Switzerland. [24]Department of Thoracic Surgery, University Hospital Zurich, 8091 Zurich, Switzerland. [25]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, N-0424 Oslo, Norway. [26]Department of Oncology, Norwegian Radium Hospital, Oslo University Hospital, N-0310 Oslo, Norway. [27]Department of Pathology, Norwegian Radium Hospital, Oslo University Hospital, N-0310 Oslo, Norway. [28]Department of Thoracic Surgery, Rikshospitalet, Oslo University Hospital, N-0027 Oslo, Norway. [29]Department of Internal Medicine, Center of Integrated Oncology Cologne-Bonn, University Hospital Cologne, 50937 Cologne, Germany. [30]Department of Surgery, St. Vincent's Hospital, Peter MacCallum Cancer Centre, 3065 Melbourne, Victoria, Australia. [31]Department of Haematology and Medical Oncology, Peter MacCallum Cancer Centre, 3065 Melbourne, Victoria, Australia. [32]Tumor Genomics Unit, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS—Istituto Nazionale Tumori, Via Venezian 1, 20133 Milan, Italy. [33]Thoracic Surgery Unit, Department of Surgery, Fondazione IRCCS Istituto Nazionale Tumori, 20133 Milan, Italy. [34]Institute of Pathology, Jena University Hospital, Friedrich-Schiller-University, 07743 Jena, Germany. [35]Department of Internal Medicine II, Jena University Hospital, Friedrich-Schiller-University, 07743 Jena, Germany. [36]Institute for Pathology Bad Berka, 99438 Bad Berka, Germany. [37]Gastrointestinal Cancer Group Cologne, Center of Integrated Oncology Cologne-Bonn, Department I for Internal Medicine, University Hospital of Cologne, 50823 Cologne, Germany. [38]Department of Pathology, University Hospital Cologne, 50937 Cologne, Germany. [39]Pathology of the University Medical Center Schleswig-Holstein, Campus Luebeck and the Research Center Borstel, Leibniz Center for Medicine and Biosciences, 23538 Luebeck and 23845 Borstel, Borstel, Germany. [40]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA. [41]German Cancer Research Center, German Cancer Consortium (DKTK), 69120 Heidelberg, Germany. These authors contributed equally: Julie George, Vonn Walter, Lynnette Fernandez-Cuesta.

A.2.2 Integrative and comparative genomic analyses identify clinically relevant groups of pulmonary carcinoids and unveil the existence of supra-carcinoids (Alcala *et al.*, Nature Communications, 2019)

Integrative and comparative genomic analyses identify clinically relevant groups of pulmonary carcinoids and unveil the existence of supra-carcinoids

Alcala N[1*], Leblay N[1*], Gabriel AAG[1*], Mangiante L[1], Hervas D[2], Giffon T[1], Sertier AS[3], Ferrari A[3], Derks J[4], Ghantous A[5], Delhomme TM[1], Chabrier A[1], Cuenin C[5], Abedi-Ardekani B[1], Boland A[6], Olaso R[6], Meyer V[6], Altmuller J[7], Le Calvez-Kelm F[1], Durand G[1], Voegele C[1], Boyault S[8], Moonen L[4], Lemaitre N[9], Lorimier P[9], Toffart AC[9], Soltermann A[10], Clement JH[11], Saenger J[12], Field JK[13], Brevet M[14], Blanc-Fournier C[15], Galateau-Salle F[16], Le Stang N[16], Russell PA[17], Wright G[17], Sozzi G[18], Pastorino U[18], Lacomme S[19], Vignaud JM[19], Hofman V[20], Hofman P[20], Brustugun OT[21], Lund-Iversen M[22], Thomas de Montpreville V[23], Muscarella LA[24], Graziano P[24], Popper H[25], Stojsic J[26], Deleuze JF[6], Herceg Z[5], Viari A[3], Nuernberg P[7,27], Pelosi G[28], Dingemans AMC[4], Milione M[18], Roz L[18], Brcic L[25], Volante M[29], Papotti MG[29], Caux C[30], Sandoval J[2], Hernandez-Vargas H[31], Brambilla E[9], Speel EJM[4], Girard N[32,33], Lantuejoul S[3,8,16], McKay JD[1], Foll M[1#], Fernandez-Cuesta L[1#]

Affiliations

[1] Section of Genetics, International Agency for Research on Cancer (WHO), Lyon, France
[2] Health Research Institute La Fe, Valencia, Spain
[3] Synergie Lyon Cancer, Centre Léon Bérard, Lyon, France
[4] Maastricht University Medical Centre (MUMC), GROW School for Oncology and Developmental Biology, Maastricht, The Netherlands
[5] Section of Epigenetics, International Agency for Research on Cancer (WHO), Lyon, France
[6] Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France
[7] Cologne Centre for Genomics (CCG) and Centre for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany
[8] Translational Research and Innovation Platform, Cancer Research Centre of Lyon (CRCL), Lyon, France
[9] Grenoble Alpes University Hospital, CHU Grenoble Alpes, Grenoble, France
[10] Institute of Pathology and Molecular Pathology, University of Zurich, Zurich, Switzerland
[11] Jena University Hospital, Jena, Germany
[12] Bad Berka Institute of Pathology, Bad Berka, Germany
[13] Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, United Kingdom
[14] Pathology Institute, Hospices Civils de Lyon, University Claude Bernard Lyon 1, France
[15] Caen University Hospital, CHU Caen, Caen, France
[16] Department of Pathology, Centre Léon Bérard (CLB), Lyon, France
[17] St Vincent's Hospital and University of Melbourne, Melbourne, Australia
[18] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[19] Nancy Regional University Hospital, CHRU, CRB, INSERM U1256, Nancy, France
[20] Laboratory of Clinical and Experimental Pathology, FHU OncoAge, Nice Hospital, Biobank BB-0033-00025, IRCAN Inserm U1081 CNRS 7284, University Côte d'Azur, Nice, France
[21] Drammen Hospital, Vestre Viken Health Trust, Drammen, Norway
[22] Oslo University Hospital, Oslo, Norway
[23] Marie Lannelongue Hospital, Le Plessis Robinson, France
[24] Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy
[25] Diagnostic and Research Institute of Pathology, Medical University of Graz, Graz, Austria
[26] Department of Thoracopulmonary Pathology, Service of Pathology, Clinical Center of Serbia, Belgrade, Serbia
[27] Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany
[28] Department of Oncology and Hemato-Oncology, University of Milan, and Inter-Hospital Pathology Division, IRCCS Multimedica, Milan, Italy

[29] Department of Oncology, University of Turin, Turin, Italy
[30] Department of Immunity, Virus, and Inflammation, Cancer Research Centre of Lyon (CRCL), Lyon, France
[31] Cancer Research Centre of Lyon (CRCL), Inserm U 1052, CNRS UMR 5286, Centre Léon Bérard, Université de Lyon, Lyon, France
[32] University Lyon 1, Lyon, France ; INSERM U932, Paris, France ; Institut Curie, Paris, France
[33] European Reference Network (ENR-EURACAN)

*These authors contributed equally to this work
#These authors jointly supervised this work

Correspondence should be addressed to: fernandezcuestal@iarc.fr

Running title

Integrative analyses of lung neuroendocrine neoplasms

Abstract

We have generated the first multi-omic dataset for atypical pulmonary carcinoids and through machine learning and multi-omics factor analysis of newly generated and previously published data, we have compared and contrasted the genomic profiles of 116 pulmonary carcinoids (including 35 atypical), 75 large-cell neuroendocrine carcinomas (LCNEC), and 66 small-cell lung cancers. These integrative analyses on 257 lung neuroendocrine neoplasms stratified atypical carcinoids into two prognostic groups with a 10-year overall survival of 88% and 20%, respectively. We identified therapeutically relevant molecular groups of pulmonary carcinoids, suggesting DLL3 and the immune system as candidate therapeutic targets, we confirmed the value of *OTP* expression levels for the prognosis and diagnosis of these diseases, and we unveiled the group of supra-carcinoids. This group comprise samples with carcinoid-like morphology yet with molecular and clinical features of the deadly LCNEC, suggesting that these tumours might represent the lung analogous to the well-differentiated grade-3 gastroenteropancreatic tumours.

Introduction

According to the WHO classification from 2015[1] and a recent IARC-WHO expert consensus proposal[2], pulmonary carcinoids are low-grade typical and intermediate-grade atypical well-differentiated lung neuroendocrine tumours (LNETs) that belong to the group of lung neuroendocrine neoplasms (LNENs), which also include the high-grade and poorly differentiated small-cell lung cancer (SCLC) and large-cell neuroendocrine carcinomas (LCNEC). Pulmonary carcinoids are rare malignant lesions, which annual incidence has been increasing worldwide, especially at the advanced stages[3]. Pulmonary carcinoids account for 1–2% of all invasive lung malignancies: typical carcinoids exhibit good prognosis, although 10-23% metastasize to the regional lymph nodes, resulting in a 5-year overall survival rate of 82-100%. The prognosis is worse for atypical carcinoids, with 40-50% presenting metastasis, reducing the 5-year overall survival rate to 50%.

Contrary to pulmonary carcinoids, most of which are eligible to upfront surgery at the time of diagnosis[3], LCNEC and SCLC require aggressive, multimodal treatment upfront for most the patients. Due to these differences in clinical management and prognosis, the accurate diagnosis of these diseases is critical. However, there is still no consensus on the optimal approach for their differential diagnosis[2]; the current criteria, based on morphological features and immunohistochemistry, are imperfect and inter-observer variations are common, especially when separating typical and atypical carcinoids[4], and atypical carcinoids from LCNEC in small biopsies[5]. Ki67 protein immune-reactivity has been suggested as a good marker of prognosis in LNENs as a whole, and of differential diagnosis between carcinoids and SCLC[6,7], whereas this marker does not faithfully follow the defining histological criteria of typical and atypical carcinoids[4]. The difficulties in finding good markers to separate these diseases might be due to the limited amount of comprehensive genomic studies available for SCLC, LCNEC, and typical carcinoids, and the complete lack of such studies for atypical carcinoids[8]. In addition, such studies would also be needed to validate the recent proposed molecular link between pulmonary carcinoids and LCNEC[9,10].

In this study, we provide a comprehensive overview of the molecular traits of lung NENs—with a particular focus on the understudied atypical carcinoids—in order to identify the mechanisms underlying the clinical differences between typical and atypical carcinoids, to understand the suggested molecular link between pulmonary carcinoids and LCNEC, and to find new candidate for the diagnosis and treatment of these diseases.

Results

We have generated new data (genome, exome, transcriptome and methylome) for 63 pulmonary carcinoids (including 27 atypical) and 20 LCNEC. In order to perform comparative analyses, we have reanalysed published data for 74 pulmonary carcinoids[11], 75 LCNEC[12], and 66 SCLC[13,14] (**Online Methods**). Taken together, we have performed multi-omics integrative analyses on 116 pulmonary carcinoids (including 35 atypical), 75 LCNEC, and 66 SCLC (**Table S1 and Fig. S1**)**.** All new specimens were collected from surgically resected tumours, applying local regulations and rules at the collecting site, and including patient consent for molecular analyses as well as collection of de-identified data, with approval of the local Ethical Committees. These samples underwent an independent pathological review. For the typical carcinoids and LCNEC on which methylation analyses were performed, the DNA came from the samples included in already published studies[4,11-15], on which pathological review had already been done.

Molecular groups of pulmonary carcinoids and LCNEC

We performed an unsupervised analysis of the expression and methylation data of LNEN samples (i.e., 100 pulmonary carcinoids and 72 LCNEC) using the Multi-Omics Factor Analysis implementation of the group factor analysis statistical framework (Software MOFA)[16] (MOFA LNEN; **Fig 1A; Figs. S2-3; Online Methods**). We identified 3 latent factors that provided consistent groups of samples with similar expression and methylation profiles (i.e., clusters). Latent factors 1 and 2 explained a total of 45% and 34% of the variance in methylation and expression, respectively, and were both associated with survival (**Fig. S4**). Using consensus clustering on MOFA latent factors (**Figs. S5-7; Online Methods**), we identified three clusters, each of them enriched for samples of one of the three histopathological types (**Fig. 1A**). Cluster Carcinoid A was enriched for typical carcinoids (75%; Fisher's exact test p-value<$2.2x10^{-16}$); cluster Carcinoid B was enriched for atypical carcinoids (54%; Fisher's exact test p-value<$2.2x10^{-16}$) and male patients (80%; Fisher's exact test p-value=$1.6x10^{-9}$); and cluster LCNEC included 92% of the histopathological LCNEC (Fisher's exact test p-value<$2.2x10^{-16}$).

To assess whether the current histopathological classification could be improved by the combination of molecular and morphological characteristics, we undertook a machine-learning analysis with a random forest classifier but, in this instance, trained to predict the histopathological types based on the expression and methylation data (**Online Methods; Fig. 1B**). Ninety-five per cent of the carcinoids predicted as typical by the machine learning were in cluster Carcinoid A (**Fig. 1A**). Similarly, the majority of machine-learning atypical carcinoids (79%) belonged in cluster Carcinoid B. The machine learning stratified atypical carcinoids into two prognostic groups: the good-prognosis group with a 10-year overall survival similar to that of typical carcinoids (88%); and the bad-prognosis group with a 10-year overall survival similar to that of LCNEC (20%) (**Fig. 1C**). The molecular classification, trained on the histopathology, was, thus, able to separate the good- from the poor-prognosis pulmonary carcinoids much better than the histopathology alone. In fact, a model based ML-predictions yielded a 17-fold higher likelihood than a model based on histopathology (ΔAIC= 5.74; **Fig. S8A**).

A subgroup of atypical carcinoids presents molecular characteristics of LCNEC

Six atypical carcinoids, "supra-carcinoids", clustered with LCNEC in the MOFA LNEN (**Fig. 1A**). Consistent with this clustering, this group displayed a survival similar to LCNEC (10-year overall survival of 33% and 21%, respectively, Wald test *p*-value=0.574; **Fig. 2A**). The observed molecular link appear to be between supra-carcinoids and LCNEC rather than with SCLC, as shown by additional MOFA and PCA including expression data for 51 SCLC (**Fig. S9A-C** and **Fig. S6A-C**, respectively).

These samples originated from three different centres (two from each), and included two previously published samples (S01513 and S01522)[11], implying that this observation is unlikely to be the result of a batch effect. The limited number of supra-carcinoids did no allow to explore etiological links; however, it is of note that one of them (LNEN005) belonged to a patient with professional exposure to asbestos (which is known to cause mesothelioma[17]) (**Table 1**), and the tumour harboured a splicing *BAP1* mutation (a gene frequently altered in mesothelioma[18]). LNEN005 was the sample with the highest mutation load (37 damaging somatic mutations**; Table S2**). Gene Set Enrichment Analyses (GSEA) of mutations in the hallmarks of cancer gene sets (**Online Methods**)[19,20], showed a significant enrichment for hallmark "evading growth suppressor" (*q*-value=0.0213; **Fig. 2B, Table S3**), while "genome instability and mutation" and "activating invasion and motility" were almost significant (both with *q*-value=0.0647; **Fig. 2B**); however, the latter only included mutations detected in LNEN005. We had access to the Haematoxylin and Eosin (H&E) stain for three supra-carcinoids, on which the pathologists discarded misclassifications with LCNEC, SCLC, or mesothelioma in the case of the asbestos-exposed *BAP1*-mutated sample (**Fig. 2C; Table 1**).

While generally similar to LCNEC, and albeit based on small numbers, the supra-carcinoids appeared to have nonetheless some distinct genomic features based on genome-wide expression and methylation profiles (**Fig. 2D**). Supra-carcinoids displayed higher levels of immune checkpoint genes (both receptors and ligands; **Fig. 2E**), and also harboured generally higher expression levels of MHC class I and II genes (**Fig. 2E; Fig. S10**). Interestingly, the interferon-gamma gene—a prominent immune-stimulator, in particular of the MHC class I and II genes—also showed high expression levels in these samples (**Fig. S10**). The differences in immune checkpoint gene expression levels between groups were not explained by the amount of infiltrating cells, as estimated by deconvolution of gene expression data with software quanTIseq (**Fig. 2F**, **left panel**). However, supra-carcinoids contained the highest levels of neutrophils (greater than the 3rd quartile of the distributions of neutrophils in the other groups; **Fig. 2F, right panel**). Permutation tests showed that these levels were significantly higher than in other carcinoid groups and SCLC but not than in LCNEC (**Online Methods; Fig. S11**). Concordantly, GSEA showed that MOFA LNEN latent factor 1 (separating LCNEC and supra-carcinoids from the other carcinoids) was significantly associated with neutrophil chemotaxis and degranulation pathways (**Online Methods; Table S4**). By contrast no such association was observed in the MOFA performed only on carcinoids and SCLC samples (**Figs. S9C and S6C; Table S4).**

Mutational patterns of pulmonary carcinoids

In a previous study, mainly including typical carcinoids, we detected *MEN1*, *ARID1A*, and *EIF1AX* as significantly mutated genes[11]. We also found that covalent histone modifiers and subunits of the SWI/SNF complex were mutated in 40% and 22.2% of the cases, respectively. Genomic alterations in these genes and pathways were also seen in the new samples included in this study (**Fig. 3A; Table S2; Fig. S12**). Apart from the above-mention genes, *ATM, PSIP1,* and *ROBO1* also showed some evidence, among others, for recurrent mutations in pulmonary carcinoids (**Fig. 3A**). In addition to point mutations and small indels, the *ARID2*, *DOT1L*, and *ROBO1* genes were also altered by chimeric transcripts (**Fig. 3B**). *MEN1* was also inactivated by genomic rearrangement in a carcinoid sample with a chromothripsis pattern affecting chromosomes 11 and 20 (**Fig. 3C**). The full list of somatically altered genes, chimeric transcripts, and genomic rearrangements are presented in **Tables S2, S5, and S6**, respectively. Of note, *MEN1* mutations were significantly associated with the atypical carcinoid histopathological subtype (Fisher's exact test *p*-value=0.0096), as well as MOFA LNEN latent factor 2.

The immune system and the retinoid and xenobiotic metabolism pathways are altered in pulmonary carcinoids

The third latent factor from the MOFA LNEN accounted for 8% and 6% of the variance in expression and methylation, respectively, but unlike latent factors 1 and 2, latent factor 3 was not associated with patient survival (**Fig. S4**). The molecular variation explained by latent factor 3 appeared to capture different molecular profiles within cluster Carcinoid A (**Fig. S9B**). We therefore undertook an additional MOFA restricted to pulmonary carcinoid samples only (MOFA LNET; **Fig. 4A; Fig. S13**). As expected, the first two latent factors of the MOFA LNET were highly correlated with latent factors 2 and 3 from the MOFA LNEN, respectively (Pearson correlation greater than 0.96; **Fig. S9B**), and explained 41% and 35% of the variance in methylation and expression, respectively. Integrative consensus clustering identified three clusters (**Online Methods; Fig. S14**): cluster Carcinoid A1 and cluster Carcinoid A2, that together correspond to the samples in cluster Carcinoid A of the MOFA LNEN, plus the supra-carcinoids; and cluster Carcinoid B. Latent factor 2 was associated with age, with cluster Carcinoid A1 enriched for older patients ([60, 90) years old) and cluster Carcinoid A2 enriched for younger patients ([15, 60) years old).

We applied GSEA to identify the pathways associated with the different latent factors. We found significant associations with the immune system and the retinoid and xenobiotic metabolism pathways (**Table S4**). Numerous Gene Ontology (GO) terms and KEGG pathways were related to the immune system, immune cell migration, and infectious diseases. The GO terms and KEGG pathways related to immune cell migration included leukocyte migration, chemotaxis, cytokines, and interleukin 17 signalling. In particular, the expression of all β-chemokines (including CCL2, CCL7, CCL19, CCL21, CCL22, known to attract monocytes and dendritic cells)[21] (**Table S4**), and all CXC chemokines (such as IL8, CXCL1, CXCL3, and CXCL5, known to attract neutrophils)[22], were positively correlated with MOFA LNEN latent factor 1

(separating pulmonary carcinoids from LCNEC) and negatively correlated with MOFA LNET latent factor 2 (separating clusters Carcinoid A1 and A2).

The different LNET clusters did not differ in their total amounts of estimated proportions of immune cells, but they did differ in their composition (**Fig. S15**): cluster Carcinoid A (particularly A1) was significantly enriched in dendritic cells, and cluster Carcinoid B in monocytes (**Fig. 4B, upper panel**). As monocytes can differentiate into dendritic cells in a favourable environment[23], we assessed the levels of *LAMP3* and *CD1A* dendritic-cells markers[24], and found that samples in cluster Carcinoid A1 presented high expression levels of these genes (**Fig. 4B, lower panel**), implying that this cluster was indeed enriched for dendritic cells. We pursued this further by assessing the CD1A protein levels by immunohistochemistry (IHC) in an independent series of pulmonary carcinoids and found that 60% of them (12/20) were enriched in CDA1-positive dendritic cells, confirming the presence of dendritic cells in a subgroup of pulmonary carcinoids (**Fig. 4C; Table S7**).

Regarding the retinoid and xenobiotic metabolism pathways (e.g., elimination of drugs and environmental pollutants), the main genes driving the correlation with MOFA latent factors were the phase II enzymes involved in glucuronosyl-transferase activity (**Table S4**), but also the phase I cytochrome P450 (CYP) proteins. These pathways were positively correlated with MOFA LNEN latent factor 2 (separating LNEN clusters A and B) and negatively correlated with MOFA LNET latent factor 1 (separating LNET clusters A1 and A2 from cluster B). Indeed, we found that samples in cluster Carcinoid B were characterised by high levels of the CYP family of genes, and a very strong expression of several UDP glucuronosyl-transferases *UGT* genes (median FPKM=4.6 in *UGT2A3* and 28.1 in *UGT2B* genes; **Fig. 4D**), which contrasts with the low levels in other carcinoids (median FPKM=0 for both *UGT2A3* and *UGT2B*; **Fig. 4D**), LCNEC (median FPKM=0 and 1.2 for *UGT2A3* and *UGT2B*; **Fig. S16**) and SCLC (median FPKM=0 and 0.3 for *UGT2A3* and *UGT2B*; **Fig. S16**).

Molecular groups of pulmonary carcinoids

We explored the molecular characteristics of each cluster from the MOFA LNET based on their core differentially expressed genes (DEG) and corresponding promoter methylation profiles (**Fig. 5A; Table S8; Online Methods**), and their somatic mutational patterns (**Fig. 3A; Fig. 4A)**. We correlated the gene expression and promoter methylation data of the core DEG to identify genes, which expression could be mainly explained by their methylation patterns (**Fig. 5A**). One of the top correlations was found for *HNF1A* and *HNF4A* homeobox genes (**Fig. S17**), which expressions are almost completely silenced in cluster Carcinoid A1 (**Fig. S18**). In addition, these genes harboured core differentially methylated positions of cluster Carcinoid A1 in their promoter regions, indicating that their methylation profiles are specific of that cluster (**Table S9**). These two transcriptional regulators genes have been reported as having a role in the regulation of *ANGPTL3*, CYP, and UGT genes expression[25]. Samples in cluster Carcinoid A1 were also characterised by high expression levels of the delta like canonical Notch ligand 3 (*DLL3*, 75% with FPKM>1) and its activator the achaete-scute family bHLH transcription factor 1 (*ASCL1*)

(**Fig. 5A; Table S8**), with expression levels similar to SCLC and LCNEC (**Fig. 5B**); however, the expression levels of NOTCH genes did not differ between the different groups (**Fig. S19**). The supra-carcinoids were all negative for *DLL3* expression (**Fig. 5B**), and had generally high expression levels of *NOTCH1-3* (**Fig. S19**). We additionally tested the DLL3 protein levels in the aforementioned independent series of 20 pulmonary carcinoids and found 40% (8/20) with relatively high expression of DLL3 (**Fig. 4D; Table S7**), while in the other 12 samples DLL3 was strikingly absent (**Fig. 4D; Table S7**). Furthermore, we found a correlation between the protein levels of DLL3 and CD1A (Pearson test *p*-value=0.00034; **Fig. S20**), providing additional evidence for the existence of a DLL3+ CD1A+ subgroup of carcinoids. Core DEG in Cluster Carcinoid A2 included the low levels of *SLIT1* (slit guidance ligand 1; 97% with FPKM<0.01), and *ROBO1* (roundabout guidance receptor 1; 56% with FPKM<1) (**Fig. 5A-B; Table S8**). This cluster also contained the four samples with somatic mutations in the eukaryotic translation initiation factor 1A X-linked (*EIF1AX*) gene (**Fig. 4A**). Concordantly, samples with *EIF1AX* mutations had significantly higher coordinates on latent factor 2 (t-test *p*-value=0.0342).

As expected based on **Fig. 4D**, several UGT genes were core DEG of Cluster Carcinoid B (**Fig. 5A**). Also, accordingly with the worse survival of patients in this cluster (**Fig. 2A**), these samples were also characterised by the expression of angiopoietin like 3 (*ANGPTL3,* 90% with FPKM>1), and the erb-b2 receptor tyrosine kinase 4 (*ERBB4*, 67% with FPKM>1) (**Fig. 5B**). This cluster was also characterised by the universal downregulation of orthopedia homeobox (*OTP*; 90% with FPKM<1), and NK2 homeobox 1 (*NKX2-1*; 90% FPKM<1) (**Fig. 5B**). Interestingly, the SCLC-combined LCNEC sample (S00602) that clustered with the pulmonary carcinoids in the MOFA LNEN (**Fig. 1A**) was the only LCNEC in our series harbouring a high-expression level of *OTP* (290.26 FPKM *vs* 9.89 FPKM for the 2nd highest within LCNEC, the median for LCNEC being 0.22 FPKM). *UGT* genes, *ANGPTL3*, and *ERBB4* were also core genes of cluster B when compared to LNEN clusters Carcinoid A and LCNEC (**Table S10**), which indicates that their expression levels also significantly differed from that of LCNEC. Cluster Carcinoid B included all observed *MEN1* mutations, which is consistent with the fact that samples with *MEN1* mutations had significantly lower coordinates on LNET latent factor 1 (*t*-test *p*-value=7x10^{-6}; **Fig. 4A**). Nevertheless, mutations in this gene did not explain the poorer prognosis of this group of samples (logrank *p*-value=0.19; **Fig. S21**). To gain some insights into what might be driving the bad prognosis of cluster Carcinoid B samples, we performed a GSEA of mutations in hallmarks of cancer gene sets (**Online Methods**)[19,20]; while clusters Carcinoid A1 and A2 were not enriched for any hallmark of cancer, cluster Carcinoid B was significantly enriched for genes involved in "evading growth suppressor", "sustaining proliferative signalling", and "genome instability and mutation" (**Fig. 5C**). We also performed a Cox regression with elastic net regularisation based on the core DEG of this cluster (**Online Methods**). The model selected eight coding genes explaining the overall survival, *OTP* being one of these genes (**Fig. 5D; Table S11**). Further supporting their prognostic value, we found that the expression of six of these genes was significantly different between the good and the poor-prognosis atypical carcinoids based on the machine learning (**Fig. 1C, upper panel; Fig. S22**).

Finally, we also checked the *MKI67* expression levels in the different molecular groups and found relatively low levels in the Carcinoids A1, A2, and B groups (78% with FPKM<1 in 78%) and high levels in the supra-carcinoids (FPKM>1 in the three samples). As expected, LCNECs (99% with FPKM>1) and SCLCs (92% with FPKM>1) carried high levels of this gene. Although the levels of *MKI67* for each of the clusters were different, further analyses showed that *MKI67* expression levels alone were not able to accurately separate good- from poor-prognosis pulmonary carcinoids (**Figs. S8B-C**). An overview of the different molecular groups of pulmonary carcinoids and their most relevant characteristics is displayed in **Fig. 6**.

Discussion

Lung neuroendocrine neoplasms are a heterogeneous group of tumours with variable clinical outcomes. Here, we have characterised and contrasted their molecular profiles. For this, we have performed integrative analysis of transcriptome and methylome data, using both machine-learning (ML) techniques and multi-omics factor analyses (MOFA). ML analyses showed that the molecular profiles could distinguish survival outcomes within patients with atypical carcinoid histopathological features, splitting them into patients with good "typical carcinoid like" survival and patients with a clinical outcome similar to LCNEC. Overall, out of the 35 histopathological atypical carcinoids, ML reclassified 11 into the typical category.

Unsupervised MOFA and subsequent gene-set enrichment analyses unveiled the immune system and the retinoid and xenobiotic metabolism as key deregulated processes in pulmonary carcinoids, and identified three molecular groups—clusters—with clinical implications (**Fig. 6**). The first group (cluster A1) presented high infiltration by dendritic cells, which are believed to promote the recruitment of immune effector cells resulting in a strongly active immunity[26]. Samples in cluster A1 showed overexpression of *ASCL1* and *DLL3.* The transcription factor ASCL1 is a master regulator that induces neuronal and NE differentiation. It regulates the expression of *DLL3*, which encodes an inhibitor of the Notch pathway[27]. Overexpression of *ASCL1* and *DLL3* is a characteristic of the SCLC of the "classic" subtype[27] and the type-I LCNEC[12]. We validated the expression of DLL3 in an independent series of 20 pulmonary carcinoids assessed by IHC (40% positive). The fact that we found a correlation between the protein levels of DLL3 and CD1A (a marker of dendritic cells also assessed by IHC in this series, 60% positive) provides orthogonal evidence to support the existence of this molecular group. Phase I trials have provided evidence for clinical activity of the anti-DLL3 humanized monoclonal antibody in high–DLL3-expressing SCLCs and LCNECs[28], and additional clinical trials are ongoing in these and other cancer types.

The second group (cluster A2) harboured recurrent somatic mutations in *EIF1AX*, and showed down-regulation of the *SLIT1* and *ROBO1* genes. SLIT and ROBO proteins are known to be axon-guidance molecules involved in the development of the nervous system[29], but the SLIT/ROBO signalling has also been associated with cancer development, progression and metastasis. Pulmonary neuroendocrine cells (PNEC) represent 1% of the total lung epithelial cell population[30], they reside isolated (Kultchinsky cells) or in clusters named neuroepithelial bodies (NEBs), and are believed to be the cell of origin of most of the lung neuroendocrine neoplasms[31]. In the normal lung, it has been shown that *ROBO1/2* are expressed, exclusively, in the PNECs, and that the SLIT/ROBO signalling is required for PNEC assembly and maintenance in NEBs[32]. In cancer, this pathway mainly suppresses tumour progression by regulating invasion, migration, and apoptosis, and therefore, is often down-regulated in many cancer types[29]. More specifically, the SLIT1/ROBO1 interaction can inhibit cell invasion by inhibiting the SDF1/CXCR4 axis, and can attenuate cell cycle progression by destruction of β-catenin and CDC42[29]. Potential clinical avenues to this finding exist, especially the on-going development of CXCR4 inhibitors.

The third molecular group (cluster B) was enriched in monocytes and depleted of dendritic cells, and had the worst median survival. Even in the presence of T cell infiltration, this immune contexture suggests an inactive immune response, dominated by monocytes and macrophages with potent immunosuppressive functions, and almost devoid of the most potent antigen-presenting cells, dendritic cells, suggesting dendritic cell-based immunotherapy as a therapeutic option for this group of samples[33]. Cluster B was also characterised by recurrent somatic mutations in *MEN1*, the most frequently altered gene in pulmonary carcinoids and pancreatic NET[34], which is in line with the common embryologic origin of pancreas and lung. *MEN1* was inactivated by genomic rearrangement due to a chromothripsis event affecting chromosomes 11 and 20 in one of our samples. This observation, together with two additional reported cases involving chromosomes 2, 12, and 13[11], and chromosomes 2, 11, and 20[35], suggest that chromothripsis is a rare but recurrent event in pulmonary carcinoids. Interestingly, *MEN1* mutations did not have a clear prognostic value in our series. Regarding the above-mentioned deregulation of the retinoid and xenobiotic metabolism in pulmonary carcinoids, samples in cluster B presented high levels of UGT and CYP genes. In line with previous studies[15,36], these samples also harboured low levels of *OTP*, which gene expression levels were strongly correlated with survival in the ML predictions. High levels of *ANGPTL3* and *ERBB4* were also detected in this group of samples, representing novel candidate therapeutic opportunities. ANGPTL3 is involved in new blood vessel growth and stimulation of the MAPK pathway[37,38]. This protein has been found aberrantly expressed in several types of human cancers[39,40]. Similarly, overexpression of the epidermal growth factor receptor *ERBB4*, which induces a variety of cellular responses including mitogenesis and differentiation, has also been associated with several cancer types[41-44].

For many years, it has been widely accepted that the lung well-differentiated NETs (typical and atypical carcinoids) have unique clinico-histopathological traits with no apparent causative relationship or common genetic, epidemiologic, or clinical traits with the lung poorly-differentiated SCLC and LCNEC[3]. While molecular studies have sustained this belief for pulmonary carcinoids *versus* SCLC[11,13,14], the identification of a carcinoid-like group of LCNECs[10,12], the recent observation of LCNEC arising within a background of pre-existing atypical carcinoid[45], and a recent proof-of-concept study supporting the progression from pulmonary carcinoids to LCNEC and SCLC[9] suggest that the separation between pulmonary carcinoids and LCNEC might be more subtle than initially thought, at least for a subset of patients. Our study supports the suggested molecular link between pulmonary carcinoids and LCNEC, as we have identified a subgroup of atypical carcinoids (supra-carcinoids) with a clear carcinoid histopathological pattern but with molecular characteristics similar to LCNEC. In our series, the proportion of supra-carcinoids was in the order of 5.5%; however, considering the intermediate phenotypes observed in the MOFA, the exact proportion would need to be confirmed in larger series. We found high estimated levels of neutrophil infiltration in the supra-carcinoids. For both supra-carcinoids and LCNEC (but not SCLC), the pathways related to neutrophil chemotaxis and degranulation were also altered. Neutrophil infiltration may act as immunosuppressive cells, for example through PDL1 expression[46]. Indeed, supra-carcinoids also presented levels of immune

checkpoint receptors and ligands (including *PDL1* and *CTLA4*) similar—or higher—than those of LCNEC and SCLC, as well as up-regulation of other immunosuppressive genes such as HLA-G, and interferon gamma that is speculated to promote cancer immune-evasion in immunosuppressive environments[47,48]. If confirmed, this would point to a therapeutic opportunity for these tumours since strategies aiming at decreasing migration of neutrophils to tumoral areas, or decreasing the amount of neutrophils have shown efficacy in preclinical models[49]. Similarly, immune checkpoint inhibitors, currently being tested in clinical trials, might also be a therapeutic option for these patients.

The current classification only recognises the existence of grade-1 (typical) and grade-2 (atypical) well-differentiated lung NETs, while the grade-3 would only be associated with the poorly-differentiated SCLC and LCNEC; however, in the pancreas, stomach and colon, the group of well-differentiated grade-3 NETs are well known and broadly recognised[50]. Whether these supra-carcinoids constitute a separate entity that may be the equivalent in the lung of the gastroenteropancreatic, well-differentiated, grade-3 NETs will require further research.

In summary, this study provides new and comprehensive insights into the molecular characteristics of pulmonary carcinoids, especially of the understudied atypical carcinoids. We have identified three well-characterized molecular groups of pulmonary carcinoids with different prognoses and clinical implications. Finally, the identification of supra-carcinoids further supports the already suggested molecular link between pulmonary carcinoids and LCNEC that warrants further investigation.

Methods

Please, see supplementary methods.

Acknowledgements

We thank the patients donating their tumour specimens. We also thank Prof. Roman K Thomas, Dr Martin Peifer, Dr Julie George, Dr Paul Brennan, and Dr Ghislaine Scelo for their help with logistics. We also thank Dr Ricard Argelaguet for his advice in using MOFA.

Funding

This study is part of the lungNENomics project. This work has been funded by the US National Institutes of Health (NIH R03CA195253 to LFC and JDM), the French National Cancer Institute (INCa, PRT-K-17-047 to LFC and TABAC 17-022 to JDM), the Ligue Nationale contre le Cancer (LNCC 2016 to LFC), France Genomique (to JDM), and the Italian Association for Cancer Research (AIRC) (IG 19238 to MV) (Special Program 5X1000, ED No12162 to UP, LR and GS). JS is a Miguel Servet researcher (CP13/00055 and PI16/0295). TMD has a fellowship from the LNCC.

Conflict of interests

The authors declare no conflict of interest related to the work presented here.

Author contribution

LFC conceived and designed the study. LFC and MF supervised all the aspects of the study. AG, AB, JA, FLCK, SB, JS, NG, and SL supervised some aspects of the study. BAA, EB, and SL performed the histopathological review. NoL, TG, JD, AC, CCu, GD, and NiL did the lab work. NA, NoL, AAGG, LM, DH, ASS, AF, TMD, RO, VM, CV, and LM performed the computational and statistical analyses. PL, ACT, AML, AS, JHC, JSa, JSt, JKF, MB, CBF, FGS, NLS, PAR, GW, LR, GS, UP, MM, SL, JMV, VH, PH, OTB, MLI, VTM, LAM, PG, MV, MGP, LB, HP, AMCD, EB, EJMS, NG, and SL contributed with samples and the corresponding histopathological, epidemiological, and clinical data. JFD, ZH, AV, PN, and JDM helped with logistics. JD, BAA, CCa, LR, MM, MV, MGP, LB, HP, GP, JDM, HHV, EJMS, NG, and SL gave scientific input. NA, NoL, AAGG, LM, JDM, MF, and LFC wrote the manuscript, which was reviewed and commented by all the co-authors.

References

1. Travis, W.D. *et al.* The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J Thorac Oncol* **10**, 1243-1260 (2015).
2. Rindi, G. *et al.* A common classification framework for neuroendocrine neoplasms: an International Agency for Research on Cancer (IARC) and World Health Organization (WHO) expert consensus proposal. *Mod Pathol* (2018).
3. Caplin, M.E. *et al.* Pulmonary neuroendocrine (carcinoid) tumors: European Neuroendocrine Tumor Society expert consensus and recommendations for best practice for typical and atypical pulmonary carcinoids. *Ann Oncol* **26**, 1604-20 (2015).
4. Swarts, D.R. *et al.* Interobserver variability for the WHO classification of pulmonary carcinoids. *Am J Surg Pathol* **38**, 1429-36 (2014).
5. Thunnissen, E. *et al.* The Use of Immunohistochemistry Improves the Diagnosis of Small Cell Lung Cancer and Its Differential Diagnosis. An International Reproducibility Study in a Demanding Set of Cases. *J Thorac Oncol* **12**, 334-346 (2017).
6. Marchio, C. *et al.* Distinctive pathological and clinical features of lung carcinoids with high proliferation index. *Virchows Arch* **471**, 713-720 (2017).
7. Pelosi, G., Rindi, G., Travis, W.D. & Papotti, M. Ki-67 antigen in lung neuroendocrine tumors: unraveling a role in clinical practice. *J Thorac Oncol* **9**, 273-84 (2014).
8. Derks, J.L. *et al.* New Insights into the Molecular Characteristics of Pulmonary Carcinoids and Large Cell Neuroendocrine Carcinomas, and the Impact on Their Clinical Management. *J Thorac Oncol* **13**, 752-766 (2018).
9. Pelosi, G. *et al.* Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm. *Virchows Arch* **472**, 567-577 (2018).
10. Rekhtman, N. *et al.* Next-Generation Sequencing of Pulmonary Large Cell Neuroendocrine Carcinoma Reveals Small Cell Carcinoma-like and Non-Small Cell Carcinoma-like Subsets. *Clin Cancer Res* **22**, 3618-29 (2016).
11. Fernandez-Cuesta, L. *et al.* Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat Commun* **5**, 3518 (2014).
12. George, J. *et al.* Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. *Nat Commun* **9**, 1048 (2018).
13. Peifer, M. *et al.* Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat Genet* **44**, 1104-10 (2012).
14. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47-53 (2015).
15. Swarts, D.R. *et al.* CD44 and OTP are strong prognostic markers for pulmonary carcinoids. *Clin Cancer Res* **19**, 2197-207 (2013).
16. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, e8124 (2018).
17. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Asbestos (chrysotile, amosite, crocidolite, tremolite, actinolite, and anthophyllite). in *IARC Monographs on the evaluation of carcinogenic risks to humans: Arsenic, Metals, Fibres and Dusts*, Vol. 100C 219-309 (2012).
18. Carbone, M. *et al.* BAP1 and cancer. *Nat Rev Cancer* **13**, 153-9 (2013).
19. Kiefer, J. *et al.* Abstract 3589: A systematic approach toward gene annotation of the hallmarks of cancer. *Cancer Research* **77**, 3589-3589 (2017).
20. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-74 (2011).
21. Shi, C. & Pamer, E.G. Monocyte recruitment during infection and inflammation. *Nat Rev Immunol* **11**, 762-74 (2011).
22. Kolaczkowska, E. & Kubes, P. Neutrophil recruitment and function in health and inflammation. *Nat Rev Immunol* **13**, 159-75 (2013).
23. Jakubzick, C.V., Randolph, G.J. & Henson, P.M. Monocyte differentiation and antigen-presenting functions. *Nat Rev Immunol* **17**, 349-362 (2017).

24. Cernadas, M., Lu, J., Watts, G. & Brenner, M.B. CD1a expression defines an interleukin-12 producing population of human dendritic cells. *Clin Exp Immunol* **155**, 523-33 (2009).
25. Odom, D.T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378-81 (2004).
26. Tran Janco, J.M., Lamichhane, P., Karyampudi, L. & Knutson, K.L. Tumor-infiltrating dendritic cells in cancer pathogenesis. *J Immunol* **194**, 2985-91 (2015).
27. Gazdar, A.F., Bunn, P.A. & Minna, J.D. Small-cell lung cancer: what we know, what we need to know and the path forward. *Nat Rev Cancer* **17**, 765 (2017).
28. Rudin, C.M. *et al.* Rovalpituzumab tesirine, a DLL3-targeted antibody-drug conjugate, in recurrent small-cell lung cancer: a first-in-human, first-in-class, open-label, phase 1 study. *Lancet Oncol* **18**, 42-51 (2017).
29. Gara, R.K. *et al.* Slit/Robo pathway: a promising therapeutic target for cancer. *Drug Discov Today* **20**, 156-64 (2015).
30. Boers, J.E., den Brok, J.L., Koudstaal, J., Arends, J.W. & Thunnissen, F.B. Number and proliferation of neuroendocrine cells in normal human airway epithelium. *Am J Respir Crit Care Med* **154**, 758-63 (1996).
31. Sutherland, K.D. & Berns, A. Cell of origin of lung cancer. *Mol Oncol* **4**, 397-403 (2010).
32. Branchfield, K. *et al.* Pulmonary neuroendocrine cells function as airway sensors to control lung immune response. *Science* **351**, 707-10 (2016).
33. Kimura, H. *et al.* Randomized controlled phase III trial of adjuvant chemo-immunotherapy with activated killer T cells and dendritic cells in patients with resected primary lung cancer. *Cancer Immunol Immunother* **64**, 51-9 (2015).
34. Scarpa, A. *et al.* Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature* **543**, 65-71 (2017).
35. Simbolo, M. *et al.* Lung neuroendocrine tumours: deep sequencing of the four World Health Organization histotypes reveals chromatin-remodelling genes as major players and a prognostic role for TERT, RB1, MEN1 and KMT2D. *J Pathol* **241**, 488-500 (2017).
36. Papaxoinis, G. *et al.* Prognostic Significance of CD44 and Orthopedia Homeobox Protein (OTP) Expression in Pulmonary Carcinoid Tumours. *Endocr Pathol* **28**, 60-70 (2017).
37. Chambard, J.C., Lefloch, R., Pouyssegur, J. & Lenormand, P. ERK implication in cell cycle regulation. *Biochim Biophys Acta* **1773**, 1299-310 (2007).
38. Yu, H. *et al.* Effects of ANGPTL3 antisense oligodeoxynucleotides transfection on the cell growths and invasion of human hepatocellular carcinoma cells. *Hepatogastroenterology* **58**, 1742-6 (2011).
39. Koyama, T. *et al.* ANGPTL3 is a novel biomarker as it activates ERK/MAPK pathway in oral cancer. *Cancer Med* **4**, 759-69 (2015).
40. Zhu, L. *et al.* Angiopoietin-like protein 3 is an indicator of prognosis in esophageal cancer patients. *Int J Clin Exp Med* **8**, 16101-6 (2015).
41. Bae, J.A. *et al.* An unconventional KITENIN/ErbB4-mediated downstream signal of EGF upregulates c-Jun and the invasiveness of colorectal cancer cells. *Clin Cancer Res* **20**, 4115-28 (2014).
42. Davies, S. *et al.* High incidence of ErbB3, ErbB4, and MET expression in ovarian cancer. *Int J Gynecol Pathol* **33**, 402-10 (2014).
43. Kurppa, K.J., Denessiouk, K., Johnson, M.S. & Elenius, K. Activating ERBB4 mutations in non-small cell lung cancer. *Oncogene* **35**, 1283-91 (2016).
44. Williams, C.S. *et al.* ERBB4 is over-expressed in human colon cancer and enhances cellular transformation. *Carcinogenesis* **36**, 710-8 (2015).
45. Fabbri, A. *et al.* Thymus neuroendocrine tumors with CTNNB1 gene mutations, disarrayed ss-catenin expression, and dual intra-tumor Ki-67 labeling index compartmentalization challenge the concept of secondary high-grade neuroendocrine tumor: a paradigm shift. *Virchows Arch* **471**, 31-47 (2017).
46. Wang, T.T. *et al.* Tumour-activated neutrophils in gastric cancer foster immune suppression and disease progression through GM-CSF-PD-L1 pathway. *Gut* **66**, 1900-1911 (2017).
47. Mojic, M., Takeda, K. & Hayakawa, Y. The Dark Side of IFN-gamma: Its Role in Promoting Cancer Immunoevasion. *Int J Mol Sci* **19**(2017).
48. Zaidi, M.R. & Merlino, G. The two faces of interferon-gamma in cancer. *Clin Cancer Res* **17**, 6118-24 (2011).

49. Ocana, A., Nieto-Jimenez, C., Pandiella, A. & Templeton, A.J. Neutrophils in cancer: prognostic role and therapeutic strategies. *Mol Cancer* **16**, 137 (2017).
50. Tang, L.H., Basturk, O., Sue, J.J. & Klimstra, D.S. A Practical Approach to the Classification of WHO Grade 3 (G3) Well-differentiated Neuroendocrine Tumor (WD-NET) and Poorly Differentiated Neuroendocrine Carcinoma (PD-NEC) of the Pancreas. *Am J Surg Pathol* **40**, 1192-202 (2016).

	LNEN005	LNEN012	LNEN021	LNEN022	S01513	S01522
CLASSIFICATION						
Histopathology	Atypical	Atypical	Atypical	Atypical	Atypical	Atypical
Morphological characteristics	carcinoid morph. 2 mitoses/2mm^2 No necrosis	carcinoid morph. 2 mitoses/2mm^2 No necrosis	LCNEC morph. 4 mitoses/2mm^2 No necrosis	NA	NA	NA
Machine learning	LCNEC	LCNEC	Atypical	LCNEC	Atypical	LCNEC
CLINICAL DATA						
Sex	M	F	F	F	M	M
Age at diagnosis	80	70	83	58	58	63
TNM Stage	IB	IIIC	IA1	IIB	IIIA	IV
Overall survival (months)	144.6	111.7	29.8	36.1	59	7
EPIDEMIOLOGY						
Smoking status	Former	NA	NA	NA	Never	Current
Other known exposure	Asbestos	NA	NA	NA	NA	NA
MULTI-OMICS DATA						
Data available	WES, RNAseq, Epic 850K	RNAseq	Epic 850K	Epic 850K	WGS, RNAseq	WES, Epic 850K
Cluster MOFA LNEN	LCNEC	LCNEC	LCNEC	LCNEC	LCNEC	LCNEC
Cluster MOFA LNET	Carcinoid A1	Carcinoid A1	Carcinoid A1	Carcinoid A1	Carcinoid A1	Carcinoid A1
Mutated genes	*JMJD1C, KDM5C, BAP1*	NA	NA	NA	*DNAH17*	*TP53*
***MKI67* FPKM**	2.6	7.3	NA	NA	1.9	NA

Table 1. Histopathological, clinical, epidemiological, and molecular characteristics of the six supra carcinoids

Figure legends

Fig. 1. Multi-omics unsupervised and supervised analyses of lung neuroendocrine neoplasms. A) Multi-Omics Factor Analysis (MOFA) of transcriptomes and methylomes of LNEN samples (typical carcinoids, atypical carcinoids, and LCNEC). Point colours correspond to the histopathological types; coloured circles correspond to predictions of histopathological types by a machine learning (ML) algorithm (random forest classifier) outlined in Panel B; filled coloured shapes represent the three molecular clusters identified by consensus clustering. The density of clinical variables that are significantly associated with a latent factor (ANOVA q-value<0.05) are represented by kernel density plots next to each axis: histopathological type for latent factor 1, sex and histopathological type for latent factor 2. **B)** Confusion matrix associated with the ML predictions represented on Panel A. The different colours highlight the prediction groups considered in the survival analysis and the colours for machine learning are consistent between Panel B and upper Panel C. For the unclassified category, the most likely classes inferred from the ML algorithm are represented by coloured arcs (black for typical, red for atypical, and grey for discordant methylation-based and expression-based predictions). **C)** Kaplan-Meier curves of overall survival of the different ML-predictions groups (upper panel) and histopathological types (lower panel). Upper panel: colours of predicted groups match Panel B. Lower panel: black - typical, red - atypical, blue - LCNEC. Next to each Kaplan-Meier plot, matrix layouts represent pairwise Wald tests between the reference group and the other groups, and the associated p-values.

Fig. 2. Molecular characterization of supra-carcinoids. A) Forest plot of hazard ratios for overall survival of the supra-carcinoids, compared to carcinoid clusters A and B, and LCNEC. **B)** Enrichment of hallmarks of cancer for somatic mutations in supra-carcinoids. Dark colours highlight significantly enriched hallmarks at the 10% false discovery rate threshold; corresponding mutated genes are listed in the boxes, and enrichment q-values are reported below. **C)** Hematoxylin and Eosin (H&E) stains of three supra-carcinoids. In all cases, an organoid architecture with tumour cells arranged in lobules or nests, forming perivascular palisades and rosettes is observed; original magnification x200. Arrows indicate mitoses. **D)** Radar charts of expression and methylation levels. Each radius corresponds to a feature (gene or CpG site), with low values close to the centre and high values close to the edge. Coloured lines represent the mean of each group. Left panel: expression z-scores of genes differentially expressed between clusters Carcinoid A and LCNEC or between Carcinoid B and LCNEC. Right panel: methylation β-values of differentially methylated positions between Carcinoid A and LCNEC clusters or between Carcinoid B and LCNEC clusters. **E)** Radar chart of the expression z-scores of immune checkpoint inhibitor genes (ligands and receptors) of each group. **F)** Left panel: average proportion of immune cells in the tumour sample for each group, as estimated from transcriptomic data using software quanTIseq. Right panel: boxplot and beeswarm plot (coloured points) of the estimated proportion of neutrophils.

Fig. 3. Mutational patterns of pulmonary carcinoids. A) Recurrent and cancer-relevant altered genes found in pulmonary carcinoids by WGS and WES. **B)** Chimeric transcripts affecting the protein product of DOT1L (upper panel), ARID2 (middle panel), and ROBO1 (lower panel). For each chimeric transcript the DNA row represents genes with their genomic coordinates, the mRNA row represents the chimeric transcript, and the protein row represents the predicted fusion protein. **C)** Chromotripsis case LNEN041, including an inter-chromosomic rearrangement between genes *MEN1* and *SOX6*. Upper-panel: copy number as a function of the genomic coordinates on chromosomes 11 and 20; a solid line separates chromosomes 11 and 20. Blue and green lines depict intra and inter-chromosomic rearrangements, respectively. Lower panel: *MEN1* chromosomic rearrangement observed in this chromotripsis case.

Fig. 4. Multi-omics unsupervised analysis of lung neuroendocrine tumours. A) Multi-Omics Factor Analysis (MOFA) of transcriptomes and methylomes of restricted to LNET samples (pulmonary carcinoids). Design follow that of **Fig. 1A**; filled coloured shapes represent the three molecular clusters (Carcinoid A1, A2, and B) identified by consensus clustering. The position of samples harbouring mutations significantly associated with a latent factor (ANOVA q-value<0.05) are highlighted by coloured triangles on the axes. **B)** Upper panel: proportion of dendritic cells in the different molecular clusters (Carcinoid A1, A2, and B) and the supra-carcinoids, estimated from transcriptomic data using quanTIseq (Online methods). Lower panel: boxplots of the expression levels of *LAMP3* (CDLAMP) and *CD1A*. **C)** DLL3 and CD1A immunohistochemistry of two typical carcinoids: case 6 (DLL3+ and CD1A+), and case 10 (DLL3- and CD1A-). Upper panels: Hematoxylin Eosin Saffron (HE) stain. Middle panels: staining with CD1 rabbit monoclonal antibody (cl EP3622; VENTANA), where arrows show positive stainings. Lower panels: Staining with DLL3 assay (SP347; VENTANA). **D)** Expression levels of genes from the retinoid and xenobiotic metabolism pathway—the most significantly associated with MOFA latent factor 1—in the different molecular clusters. Upper panel: schematic representation of the phases of the pathway. Lower panel: boxplot of expression levels of *CYP2C8* and *CYP2C19* (both from the CYP2C gene cluster on chromosome 10, *UGT2A3* and the total expression of *UGT2B* genes (from the UGT2 gene cluster on chromosome 4, expressed in fragments per kilobase million (FPKM) units.

Fig. 5. Molecular groups of pulmonary carcinoids. A) Heatmaps of the expression of core differentially expressed genes of each molecular cluster, i.e., genes that are differentially expressed in all pairwise comparisons between a focal cluster and the other clusters. Green bars at the right of each heatmap indicate a significant negative correlation with the methylation level of at least one CpG site from the gene promoter region. **B)** Boxplots of the expression levels of selected cancer-relevant core genes, in fragment per kilobase million (FPKM) units. **C)** Characteristic hallmarks of cancer in each molecular cluster (Carcinoid A1 without the supra-carcinoids, A2, and B), LCNEC and SCLC. Coloured concentric circles correspond to the molecular clusters. For each cluster, dark colours highlight significantly enriched hallmarks (q-value <0.05).

The mutated genes contributing to a given hallmark are listed in the boxes. Recurrently mutated genes are indicated in brackets by the number of samples harbouring a mutation. **D)** Survival analysis of pulmonary carcinoids based on the expression level of eight core genes of cluster Carcinoid B. The genes were selected using a regularized GLM on expression data. For each gene, coloured lines correspond to the Kaplan-Meier curve of overall survival for individuals with a high (green) and low (orange) expression level of this gene. Cut-offs for the two groups were determined using maximally selected rank statistics (**Online Methods**). The percentage of samples in each group is represented above each Kaplan-Meier curve.

Fig. 6. Overview of the main molecular and clinical characteristics of lung neuroendocrine neoplasms. Left panel**:** Radar chart of the expression level (*z*-score) of the characteristic genes (*DLL3, ASCL1, ROBO1, SLIT1, ANGPTL3, ERBB4*, UGT genes family, *OTP, NKX2-1, PD-L1 (CD274)*, and other immune checkpoint genes) of each LNET molecular cluster (Carcinoid A1, Carcinoid A2, and B clusters), supra-Ca, LCNEC and SCLC. The coloured text lists relevant characteristics—additional molecular, histopathological, and clinical data—of each group. Right panel: heatmap of the expression level (*z*-score) of the characteristic genes of each group from the left panel, expressed in *z*-scores.

List of Supplementary Figures

Fig. S1 Overview of the multi-omic experimental design for LNEN samples

Fig. S2 Robustness of the MOFA latent factors presented in Fig. 1A

Fig. S3 Correlations between MOFA latent factors (Figs. 1A and 4A) and the principal components of the PCA of expression (Fig. S6) and methylation (Fig. S7)

Fig. S4 Forest plot of the survival analysis based on the first three MOFA latent factors (LFs) of LNEN samples from Fig. 1A

Fig. S5 Robustness of the consensus clustering of LNENs presented in Fig. 1A

Fig. S6 Principal Component Analysis (PCA) of transcriptome data

Fig. S7 Principal Component Analysis of the methylation data

Fig. S8 Comparison of overall survival based on different classifications

Fig. S9 Consistency of MOFA across analyses including different histopathological types

Fig. S10 Radar chart of the expression levels of HLA class I and related immunostimulatory genes as a function of their molecular group

Fig. S11 Estimation of the amount immune cells in the different pulmonary carcinoid groups from transcriptome data

Fig. S12 Cancer-relevant somatically altered pathways altered in typical and atypical carcinoids

Fig. S13 Robustness of the MOFA latent factors presented in Fig. 4A

Fig. S14 Robustness of the consensus clustering of pulmonary carcinoids presented in Fig. 4

Fig. S15 Estimation of the amount of immune cells in the different LNET clusters and Supra-Ca from transcriptome data

Fig. S16 Expression levels of genes involved in phase I and phase II (cytochrome P450) xenobiotic metabolism

Fig. S17 Correlations between DNA methylation and gene expression for core genes of LNET clusters

Fig. S18 DNA methylation and gene expression levels of *HNF1A* and *HNF4A* in LNET samples

Fig. S19 Expression levels of NOTCH genes in the different LNET clusters, Supra-Ca, LCNEC and SCLC

Fig. S20 Correlation between *DLL3* and *CDA1* expression based on immunohistochemistry in a validation series

Fig. S21 Survival (Kaplan-Meier curve) of MEN1 wild type compared to mutant cases.

Fig. S22 Expression levels of genes explaining the carcinoids survival in the good- and poor survival atypical carcinoids (Fig. 1B).

Fig. S23 Sex reclassification and multi-omic validation of reported clinical sex

Fig. S24 Associations between clinical variables

Fig. S25. Associations between clinical variables and expression profiles of LNET

Fig. S26. Supervised analysis of histological types

Fig. S27 Estimation of the amount of immune cells in the different histopathological types from transcriptome data

Fig. S28 Assessment of the batch effects in the EPIC 850K methylation array analysis

List of Supplementary Tables

Table S1 Sample overview

Table S2 Somatic mutations

Table S3 Hallmarks of cancer gene set enrichment analysis

Table S4 Gene set enrichment analysis for MOFA latent factors

Table S5 Chimeric transcripts

Table S6 Chromosomic rearrangements

Table S7 DLL3 and CD1A Immunohistochemistry

Table S8 Differentially expressed genes between MOFA LNET clusters

Table S9 Differentially methylated positions between MOFA LNET clusters.

Table S10 Differentially expressed genes between MOFA LNEN clusters

Table S11 Genes selected by the regularized Cox regression model based on the expression of core genes of LNET cluster B

Table S12 Differentially expressed genes between histological types

Table S13 Differentially methylated positions between histological types

Table S14 Differentially methylated positions between MOFA LNEN clusters

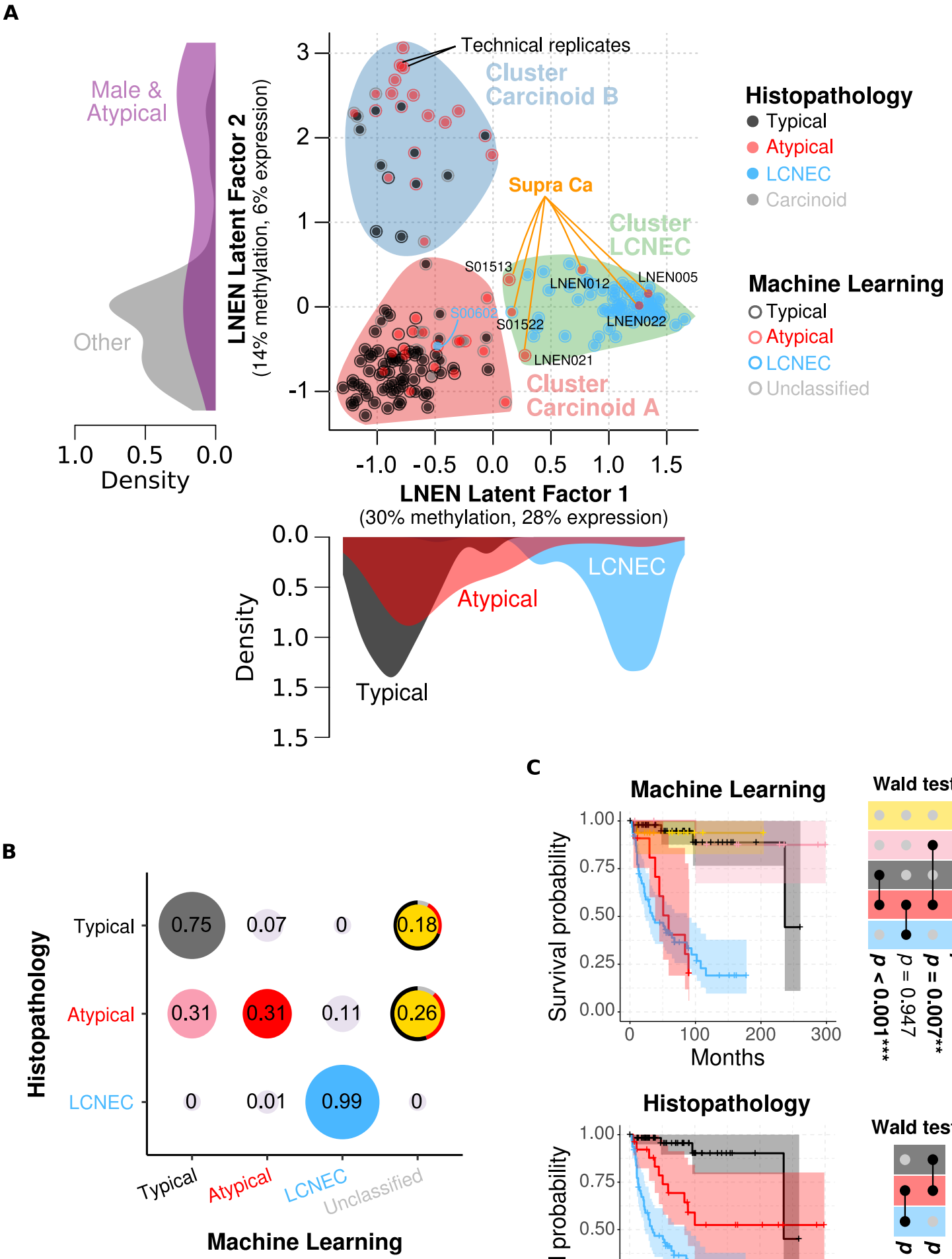

A
Male &
Atypical
Other
1.0
0.5
0.0
Density
LNEN Latent Factor 2
(14% methylation, 6% expression)
Technical replicates
Cluster
Carcinoid B
Supra Ca
Cluster
LCNEC
S01513
LNEN012
LNEN005
S00602
S01522
LNEN022
LNEN021
Cluster
Carcinoid A
3
2
1
0
-1
-1.0
-0.5
0.0
0.5
1.0
1.5
LNEN Latent Factor 1
(30% methylation, 28% expression)
Histopathology
Typical
Atypical
LCNEC
Carcinoid
Machine Learning
Typical
Atypical
LCNEC
Unclassified
Density
0.0
0.5
1.0
1.5
1.5
LCNEC
Atypical
Typical
B
Histopathology
Typical
Atypical
LCNEC
0.75
0.07
0
0.18
0.31
0.31
0.11
0.26
0
0.01
0.99
0
Typical
Atypical
LCNEC
Unclassified
Machine Learning
C
Machine Learning
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Months
Wald tests
p = 0.018*
p = 0.007**
p = 0.947
p < 0.001***
Histopathology
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Months
Wald tests
p = 0.008**
p = 0.004**

Figure 2

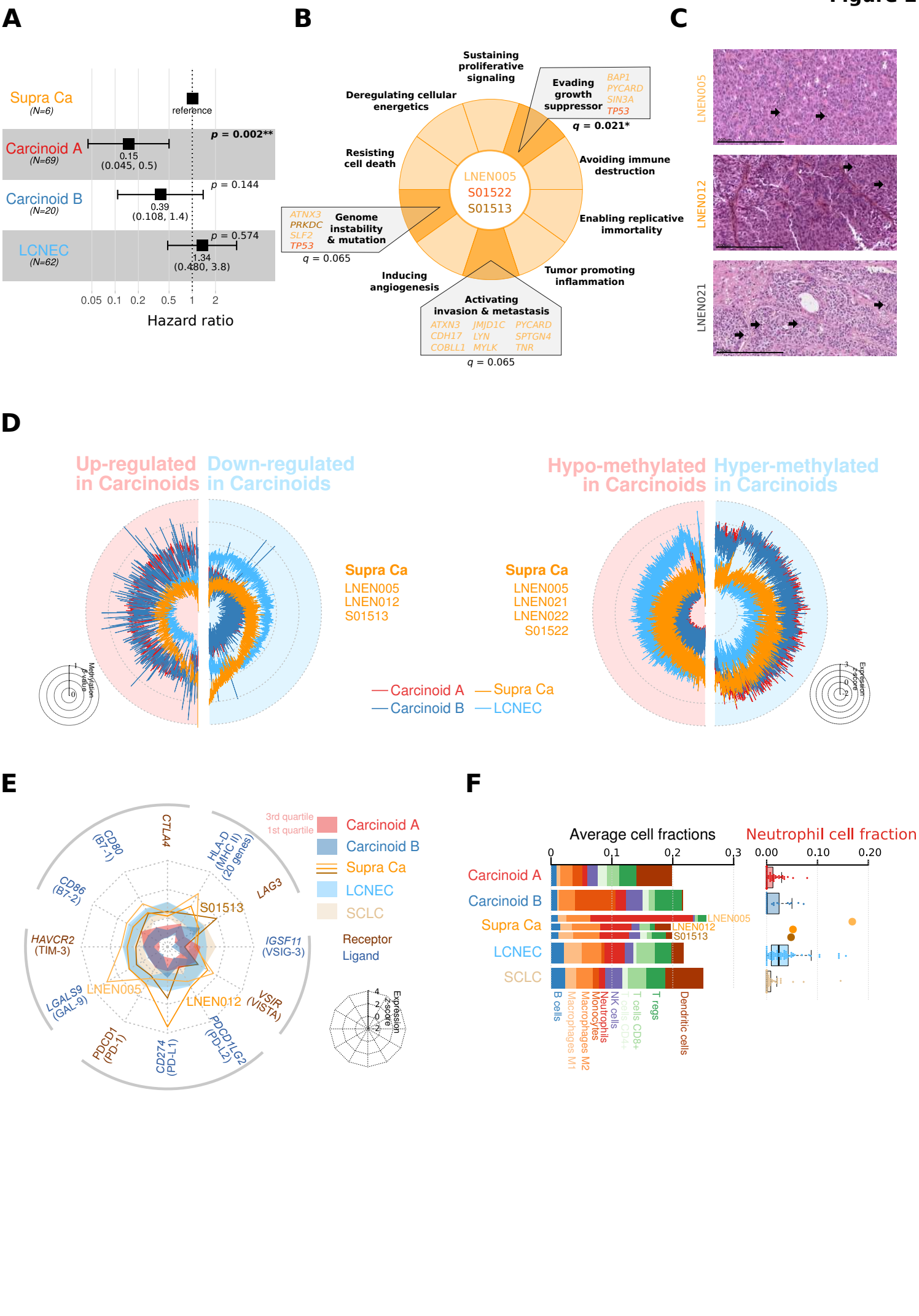

A
Supra Ca
(N=6)
reference
Carcinoid A
(N=69)
p = 0.002**
0.15
(0.045, 0.5)
Carcinoid B
(N=20)
p = 0.144
0.39
(0.108, 1.4)
LCNEC
(N=62)
p = 0.574
1.34
(0.480, 3.8)
0.05 0.1 0.2 0.5 1 2
Hazard ratio
B
Sustaining proliferative signaling
Evading growth suppressor
BAP1 PYCARD SIN3A TP53
q = 0.021*
Deregulating cellular energetics
Resisting cell death
Avoiding immune destruction
LNEN005
S01522
S01513
Enabling replicative immortality
ATNX3 PRKDC SLF2 TP53
Genome instability & mutation
q = 0.065
Tumor promoting inflammation
Inducing angiogenesis
Activating invasion & metastasis
ATXN3 JMJD1C PYCARD
CDH17 LYN SPTGN4
COBLL1 MYLK TNR
q = 0.065
C
LNEN005
LNEN012
LNEN021
D
Up-regulated in Carcinoids
Down-regulated in Carcinoids
Supra Ca
LNEN005
LNEN012
S01513
Hypo-methylated in Carcinoids
Hyper-methylated in Carcinoids
Supra Ca
LNEN005
LNEN021
LNEN022
S01522
Carcinoid A
Supra Ca
Carcinoid B
LCNEC
E
3rd quartile
1st quartile
Carcinoid A
Carcinoid B
Supra Ca
LCNEC
SCLC
Receptor
Ligand
CTLA4
HLA-D (MHC II) (20 genes)
LAG3
IGSF11 (VSIG-3)
VSIR (VISTA)
PDCD1LG2 (PD-L2)
CD274 (PD-L1)
PDCD1 (PD-1)
LGALS9 (GAL-9)
HAVCR2 (TIM-3)
CD86 (B7-2)
CD80 (B7-1)
S01513
LNEN005
LNEN012
F
Average cell fractions
Neutrophil cell fraction
Carcinoid A
Carcinoid B
Supra Ca
LCNEC
SCLC
LNEN005
LNEN012
S01513
B cells
Macrophages M1
Macrophages M2
Monocytes
Neutrophils
NK cells
T cells CD4+
T cells CD8+
T regs
Dendritic cells

Figure 3

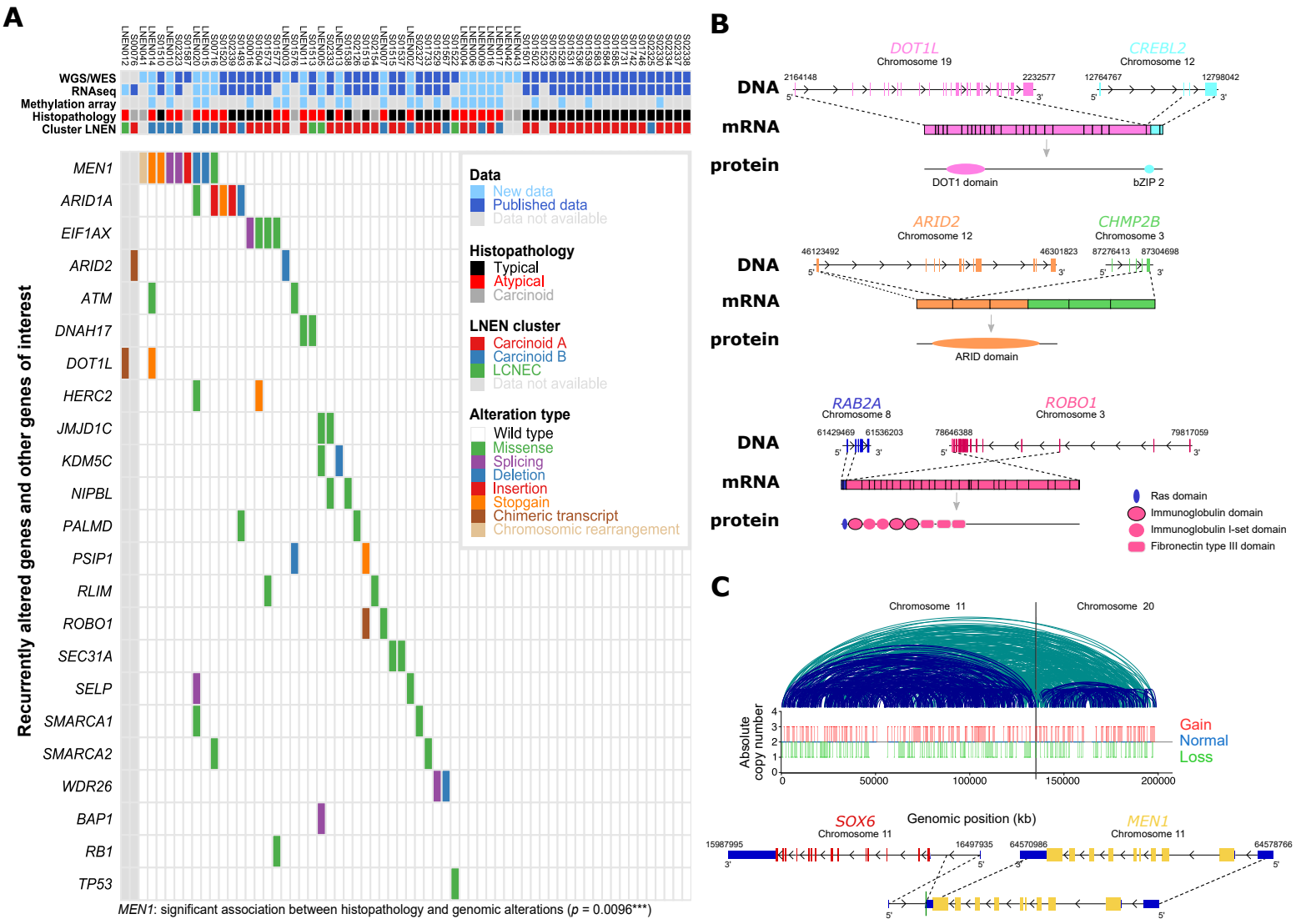
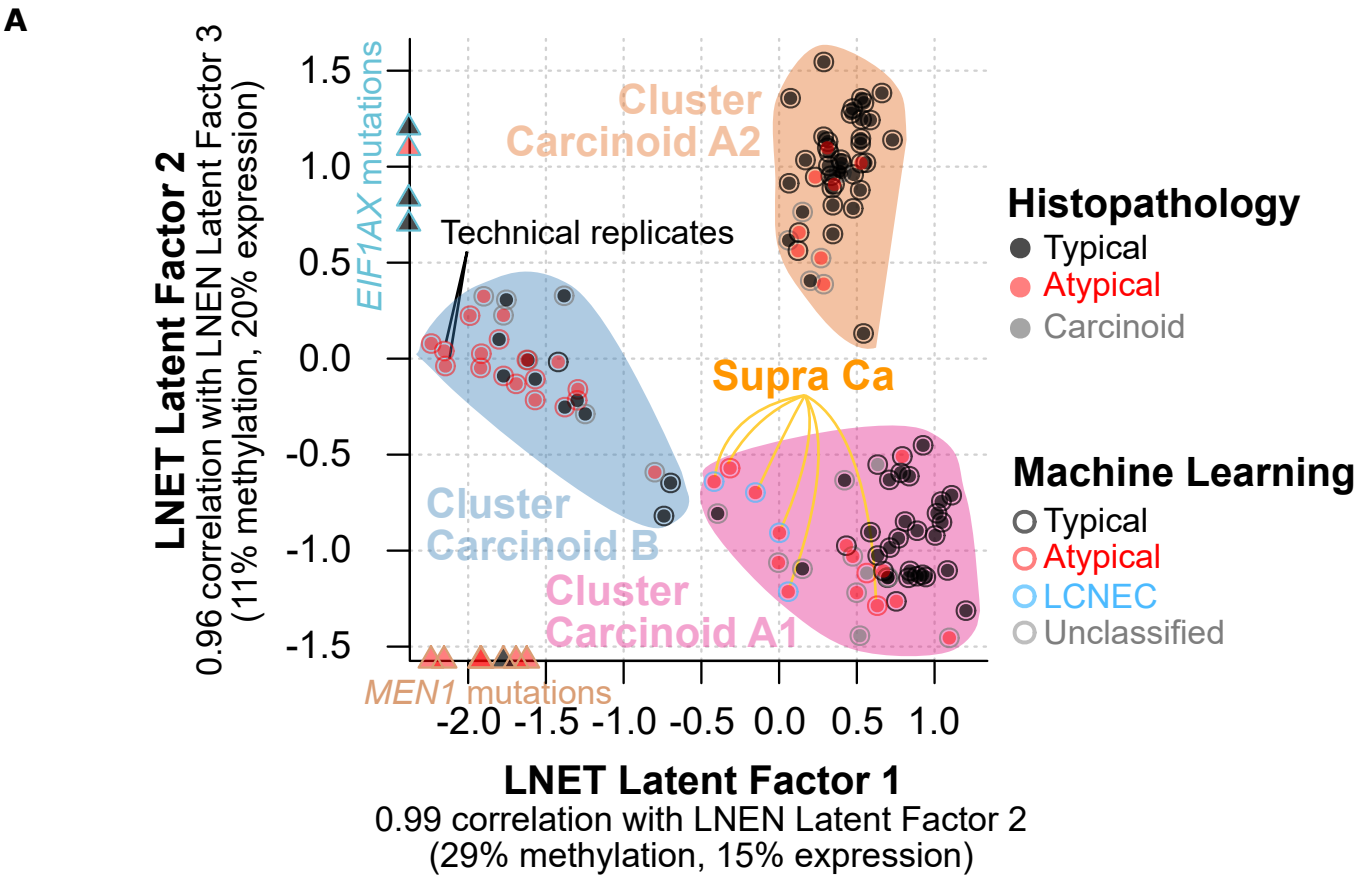
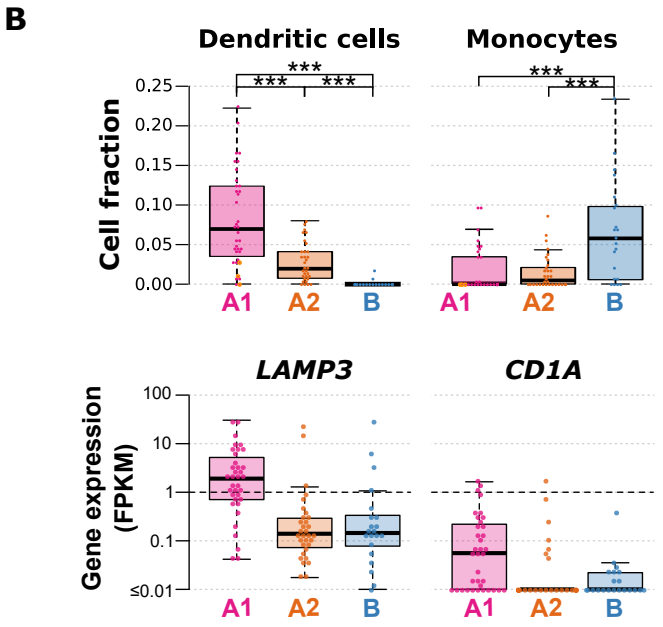
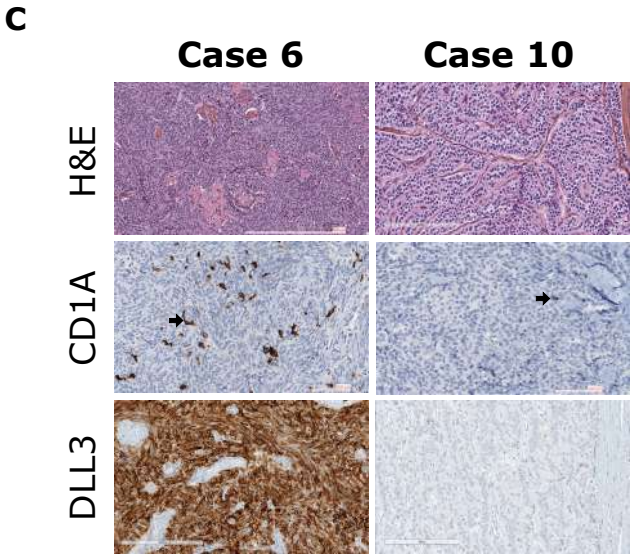

A
WGS/WES
RNAseq
Methylation array
Histopathology
Cluster LNEN
Recurrently altered genes and other genes of interest
MEN1
ARID1A
EIF1AX
ARID2
ATM
DNAH17
DOT1L
HERC2
JMJD1C
KDM5C
NIPBL
PALMD
PSIP1
RLIM
ROBO1
SEC31A
SELP
SMARCA1
SMARCA2
WDR26
BAP1
RB1
TP53
Data
New data
Published data
Data not available
Histopathology
Typical
Atypical
Carcinoid
LNEN cluster
Carcinoid A
Carcinoid B
LCNEC
Data not available
Alteration type
Wild type
Missense
Splicing
Deletion
Insertion
Stopgain
Chimeric transcript
Chromosomic rearrangement
MEN1: significant association between histopathology and genomic alterations (p = 0.0096***)
B
DOT1L
Chromosome 19
CREBL2
Chromosome 12
DNA
mRNA
protein
DOT1 domain
bZIP 2
ARID2
Chromosome 12
CHMP2B
Chromosome 3
ARID domain
RAB2A
Chromosome 8
ROBO1
Chromosome 3
Ras domain
Immunoglobulin domain
Immunoglobulin I-set domain
Fibronectin type III domain
C
Chromosome 11
Chromosome 20
Absolute copy number
Gain
Normal
Loss
Genomic position (kb)
SOX6
Chromosome 11
MEN1
Chromosome 11
Break point

Figure 4


A
LNET Latent Factor 2
0.96 correlation with LNEN Latent Factor 3
(11% methylation, 20% expression)
EIF1AX mutations
Technical replicates
Cluster Carcinoid A2
Supra Ca
Cluster Carcinoid B
Cluster Carcinoid A1
MEN1 mutations
Histopathology
Typical
Atypical
Carcinoid
Machine Learning
Typical
Atypical
LCNEC
Unclassified
LNET Latent Factor 1
0.99 correlation with LNEN Latent Factor 2
(29% methylation, 15% expression)


B
Dendritic cells
Monocytes
Cell fraction
LAMP3
CD1A
Gene expression (FPKM)
A1
A2
B


C
Case 6
Case 10
H&E
CD1A
DLL3

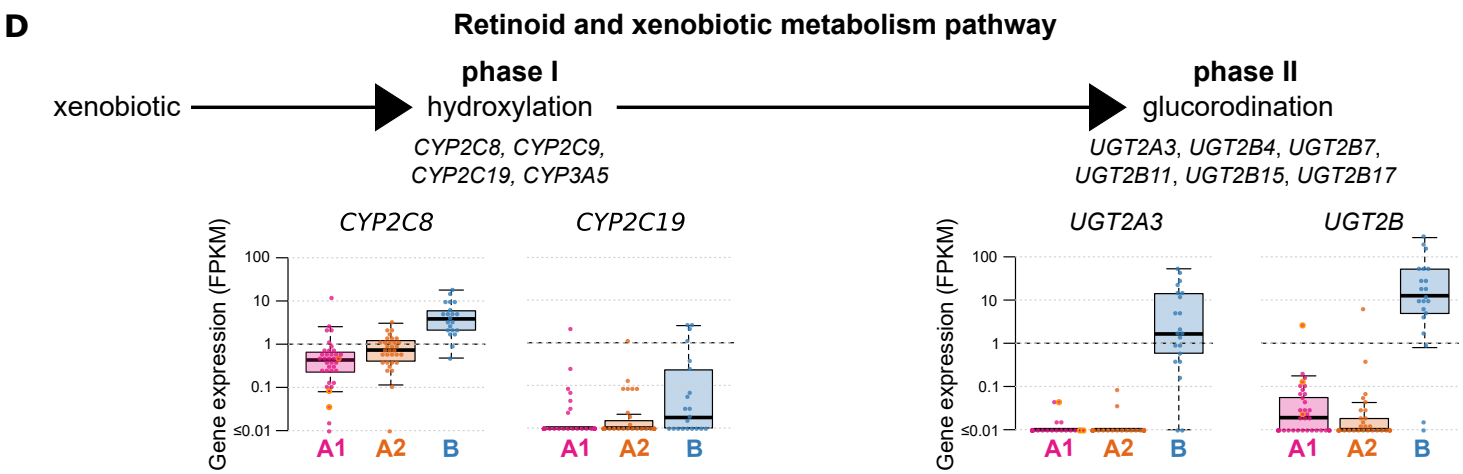

D
Retinoid and xenobiotic metabolism pathway
xenobiotic
phase I
hydroxylation
CYP2C8, CYP2C9, CYP2C19, CYP3A5
phase II
glucorodination
UGT2A3, UGT2B4, UGT2B7, UGT2B11, UGT2B15, UGT2B17
CYP2C8
CYP2C19
UGT2A3
UGT2B
Gene expression (FPKM)
A1
A2
B

Figure 5

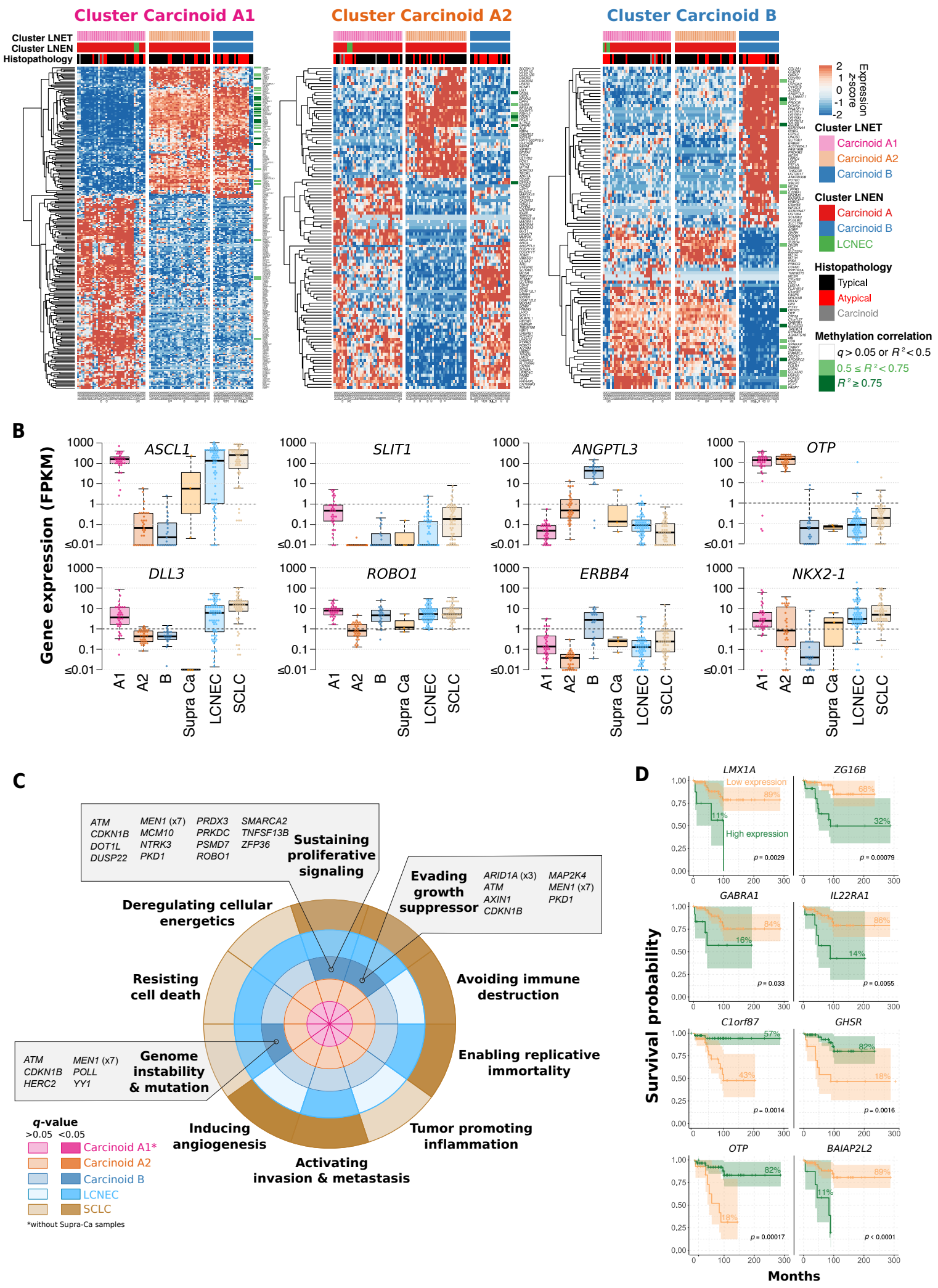
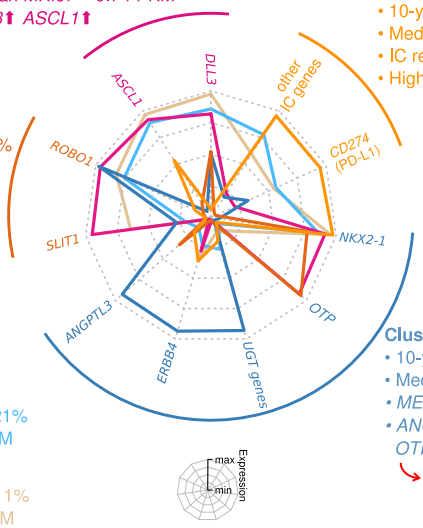

A
Cluster Carcinoid A1
Cluster Carcinoid A2
Cluster Carcinoid B
Cluster LNET
Cluster LNEN
Histopathology
Expression z-score
2
1
0
-1
-2
Cluster LNET
Carcinoid A1
Carcinoid A2
Carcinoid B
Cluster LNEN
Carcinoid A
Carcinoid B
LCNEC
Histopathology
Typical
Atypical
Carcinoid
Methylation correlation
q > 0.05 or R² < 0.5
0.5 ≤ R² < 0.75
R² ≥ 0.75
B
Gene expression (FPKM)
ASCL1
SLIT1
ANGPTL3
OTP
DLL3
ROBO1
ERBB4
NKX2-1
A1
A2
B
Supra Ca
LCNEC
SCLC
C
ATM
CDKN1B
DOT1L
DUSP22
MEN1 (x7)
MCM10
NTRK3
PKD1
PRDX3
PRKDC
PSMD7
ZFP36
SMARCA2
TNFSF13B
Sustaining proliferative signaling
Evading growth suppressor
ARID1A (x3)
ATM
AXIN1
CDKN1B
MAP2K4
MEN1 (x7)
PKD1
Deregulating cellular energetics
Resisting cell death
Avoiding immune destruction
Genome instability & mutation
ATM
CDKN1B
HERC2
MEN1 (x7)
POLL
YY1
Enabling replicative immortality
Inducing angiogenesis
Tumor promoting inflammation
Activating invasion & metastasis
q-value
>0.05
<0.05
Carcinoid A1*
Carcinoid A2
Carcinoid B
LCNEC
SCLC
*without Supra-Ca samples
D
LMX1A
ZG16B
GABRA1
IL22RA1
C1orf87
GHSR
OTP
BAIAP2L2
Low expression
High expression
Survival probability
Months

Figure 6

Cluster Carcinoid A1
- 10-year overall survival = 88%
- Median *MKI67* = 0.7 FPKM
- *DLL3*⬆ *ASCL1*⬆

Supra Ca
- 10-year overall survival = 33%
- Median *MKI67* = 2.6 FPKM
- IC receptors and ligands
- High neutrophil content

Cluster Carcinoid A2
- 10-year overall survival = 83%
- Median *MKI67* = 0.3 FPKM
- *EIF1AX* mutations
- *ROBO1*⬇ *SLIT1*⬇



LCNEC
- 10-year overall survival = 21%
- Median *MKI67* = 10.3 FPKM

SCLC
- 10-year overall survival = 11%
- Median *MKI67* = 11.2 FPKM

Cluster Carcinoid B
- 10-year overall survival = 60%
- Median *MKI67* = 0.4 FPKM
- *MEN1* mutations
- *ANGPTL3*⬆ *ERBB4*⬆ *UGT*⬆ *OTP*⬇ *NKX2-1*⬇

→ **Poor-survival subgroup**
- 38% of cluster B samples
- Atypical Carcinoids (histopathology & machine learning-predicted)
- 10-year overall survival = 20%

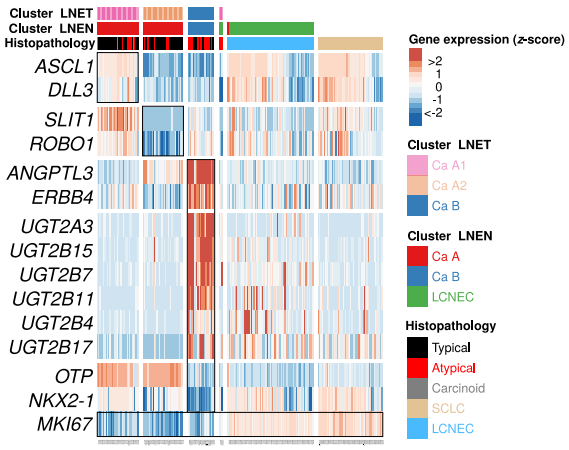


Table A.2 – Catalogue of our collaborative IARC bioinformatics pipelines developed using the previously described implementation pattern.

quality control		**DNA**		**RNA**	
type	*software*	*type*	*software*	*type*	*software*
BAM	qualimap	alignment	BWA	alignment	STAR
FASTQ	FastQC	somatic SNV/indel	strelka2	transcript identification and quantification	StringTie
		germline SNV/indel	strelka2	fusion-gene discovery	STAR-fusion
		CNV	Facets		
		structural variants	SvABA		
		low VAF	**needlestack**		

Appendix B

List of acronyms

API Application Programming Interface. IV

BLAST Basic Local Alignment Search Tool. 171

BQSR Base Quality Score Recalibration. 22

BWA Burrows-Wheeler Aligner. 9, 170, 171

cfDNA Circulating cell-Free DNA. 28–30, 38, 39, 116, 117, 132, 157, 167

CGC Cancer Genomics Cloud. I, III, IV

CNV Copy Number Variation. 5, 173

CSE Context-specific Error. 14

CT Computed Tomography. 28

ctDNA Circulating Tumor DNA. x, 28–30, 34, 84, 112, 116, 117, 165, 166, 168, 169, 172, 174, 175

CV Coefficient of variation. 68, 69

CWL Common Workflow Language. IV

ddPCR Droplet digital PCR. 172, 174

DNA deoxyribonucleic acid. 3, 5

GATK Genome Analysis ToolKit. 19

GIAB Genome In A Bottle. II

GRC Genome Reference Consortium. 5

HTS High-Throughput Sequencing. 8

IGV Integrative Genomics Viewer. 15, 23

indel Insertion or deletion. 5, 7, 38, 70, II

LCAP Low-Confidence Alteration Probability. 80–83, 85, 107, 110–112

LOD Logarithm of Odds. 21, 22

NCI National Cancer Institute. 3, 27

NGS Next-Generation Sequencing. 4, 8–10, 13, 16–18, 20, 24, 31, 34, 35, 70, 74, 78, 83, 87, 169, 170, 172, 174, 175, 177

NIST National Institute of Standards and Technology. II

NLST National Lung Screening Trial. 28

PairHMM Pair Hidden Markov Model. 19

PCC Pearson Correlation Coefficient. 69

PCR Polymerase Chain Reaction. 14, 75, 77, 174

RVSB Relative Variant Strand Bias. 15, 79, 110, 157

SCLC Small Cell Lung Cancer. 84, 116, 157, 165, 167

SER Sequencing Error Rate. 13, 16, 17, 19, 22, 34–36, 39, 67–70, 85

SNV Single Nucleotide Variation. 5, 38, 68, 70, 105, 110, II

SV Structural Variation. 173

TCGA The Cancer Genome Atlas. 13, I–IV

TSG Tumor Suppressor Gene. 7, 8

VAF Variant Allelic Fraction. 10–14, 17–19, 21–25, 29, 30, 34, 38, 39, 67, 68, 70, 83, 85, 105, 108, 110, 112, 116, 117, 132, 157, 167, 170, 172, 173, 177

VCF Variant Call Format. 37

VQSR Variant Quality Score Recalibration. 25

WES Whole-Exome sequencing. ix, 10, 25, 38, 70, 71, 104, 105, 173, I–III

WGS Whole-Genome Sequencing. 10, 25

WHO World Health Organization. 27

Appendix C

Bibliography

[1] National cancer institute website. [online]. available: https://www.cancer.gov. accessed: october 2018. 3, 27

[2] Thermo fisher company. iontorrent suite software, tmap aligner. [online]. available: https://github.com/iontorrent/ts/tree/master/analysis/tmap. 10, 85

[3] Abbosh, C., Birkbak, N.J., Wilson, G.A., Jamal-Hanjani, M., Constantin, T. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, 545(7655):446–451, apr 2017. 29, 30, 116

[4] Acuna-Hidalgo, R., Bo, T., Kwint, M.P., van de Vorst, M., Pinelli, M. et al. Post-zygotic point mutations are an underrecognized source of de novo genomic variation. *The American Journal of Human Genetics*, 97(1):67–74, jul 2015. 6

[5] Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, 76(1):7.20.1–7.20.41, jan 2013. 7, 157

[6] Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A. et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, apr 2010. 7

[7] Aeberhard, W.H., Cantoni, E. and Heritier, S. Robust inference in the negative binomial

regression model with an application to falls data. *Biometrics*, 70(4):920–931, aug 2014. 36

[8] Ainscough, B.J., Barnell, E.K., Ronning, P., Campbell, K.M., Wagner, A.H. et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nature Genetics*, 50(12):1735–1743, nov 2018. 25

[9] Alexandrov, L.B., , Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R. et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, aug 2013. 6

[10] Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, 6(1), dec 2015. 34

[11] Allhoff, M., Schönhuth, A., Martin, M., Costa, I.G., Rahmann, S. et al. Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(Suppl 5):S1, 2013. 14, 15

[12] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, oct 1990. 171

[13] Babayan, A. and Pantel, K. Advances in liquid biopsy approaches for early detection and monitoring of cancer. *Genome Medicine*, 10(1), mar 2018. 29

[14] Bahassi, E.M. and Stambrook, P.J. Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis*, 29(5):303–310, aug 2014. 16

[15] Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.e18, apr 2018. 7

[16] Baker, M. De novo genome assembly: what every biologist should know. *Nature Methods*, 9(4):333–337, apr 2012. 4

[17] Ballouz, S., Dobin, A. and Gillis, J. Is it time to change the reference genome? *bioRxiv*, jan 2019. 5, 170

[18] Beerenwinkel, N., Greenman, C.D. and Lagergren, J. Computational cancer biology: An evolutionary perspective. *PLOS Computational Biology*, 12(2):e1004717, feb 2016. 7

[19] Bettegowda, C., Sausen, M., Leary, R.J., Kinde, I., Wang, Y. et al. Detection of circulating tumor dna in early- and late-stage human malignancies. *Science Translational Medicine*, 6(224):224ra24–224ra24, feb 2014. 28, 29, 34, 116

[20] Boettiger, C. An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79, jan 2015. 37

[21] Bombard, Y., Robson, M. and Offit, K. Revealing the incidentalome when targeting the tumor genome. *JAMA*, 310(8):795, aug 2013. 6

[22] Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P. and Tyson, G.W. Shining a light on dark sequencing: Characterising errors in ion torrent PGM data. *PLoS Computational Biology*, 9(4):e1003031, apr 2013. 16

[23] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, sep 2018. 3

[24] Brierley, J., Gospodarowicz, M.K., Wittekind, C. and for International Cancer Control., U. *TNM classification of malignant tumours, Eighth edition.* Oxford: Wiley Blackwell, 2017. 27

[25] Calvez-Kelm, F.L., Foll, M., Wozniak, M.B., Delhomme, T.M., Durand, G. et al. Kras mutations in blood circulating cell-free dna: a pancreatic cancer case-control. *Oncotarget*, 7(48), oct 2016. 28

[26] Castle, J.C., Loewer, M., Boegel, S., Tadmor, A.D., Boisguerin, V. et al. Mutated tumor alleles are expressed according to their DNA frequency. *Scientific Reports*, 4(1), apr 2014. 11, 19

[27] Chaffer, C.L. and Weinberg, R.A. A perspective on cancer cell metastasis. *Science*, 331(6024):1559–1564, mar 2011. 3

[28] Chen, R. and Butte, A.J. The reference human genome demonstrates high risk of type 1 diabetes and other disorders. In *Biocomputing 2011*, pages 231–242. WORLD SCIENTIFIC, nov 2010. 170

[29] Cho, H., Mariotto, A.B., Schwartz, L.M., Luo, J. and Woloshin, S. When do changes in cancer survival mean progress? the insight from population incidence and mortality. *JNCI Monographs*, 2014(49):187–197, nov 2014. 27

[30] Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F. et al. Modernizing reference genome assemblies. *PLoS Biology*, 9(7):e1001091, jul 2011. 5

[31] Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, feb 2013. 21, 22

[32] Clark, W. Tumour progression and the nature of cancer. *British Journal of Cancer*, 64(4):631–644, oct 1991. 3

[33] Clarke, L.A. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Molecular Pathology*, 54(5):351–353, oct 2001. 14

[34] Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378):926–930, jan 2018. 28, 112

[35] Collisson, E.A., Campbell, J.D., Brooks, A.N., Berger, A.H., Lee, W. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, jul 2014. 13

[36] Cooper, G.M. *The Cell: A Molecular Approach. 2nd edition.* Sinauer Associates, 2000. 3

[37] Crowley, E., Nicolantonio, F.D., Loupakis, F. and Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nature Reviews Clinical Oncology*, 10(8):472–484, jul 2013. 28, 29, 116

[38] Cuykendall, T.N., Rubin, M.A. and Khurana, E. Non-coding genetic variation in cancer. *Current Opinion in Systems Biology*, 1:9–15, feb 2017. 5

[39] Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E. et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, jun 2011. 37

[40] der Auwera, G.A.V., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G. et al. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*, 11(10):1–33, nov 2014. 25

[41] Diehl, F., Schmidt, K., Choti, M.A., Romans, K., Goodman, S. et al. Circulating mutant dna to assess tumor dynamics. *Nature Medicine*, 14(9):985–990, may 2008. 28

[42] Do, H. and Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. *Clinical Chemistry*, 61(1):64–71, nov 2014. 76

[43] Druley, T.E., Vallania, F.L.M., Wegner, D.J., Varley, K.E., Knowles, O.L. et al. Quantification of rare allelic variants from pooled genomic DNA. *Nature Methods*, 6(4):263–265, mar 2009. 9

[44] Eberle, M.A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B.L. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, 27(1):157–164, nov 2016. 25, 113, II

[45] [editorial board]. Early detection: a long road ahead. *Nature Reviews Cancer*, 18(7):401–401, may 2018. 28

[46] Elmore, S. Apoptosis: A review of programmed cell death. *Toxicologic Pathology*, 35(4):495–516, jun 2007. 3

[47] Ewing, A.D., , Houlahan, K.E., Hu, Y., Ellrott, K. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, 12(7):623–630, may 2015. 39

[48] Fan, Y., Xi, L., Hughes, D.S.T., Zhang, J., Zhang, J. et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*, 17(1), aug 2016. 21

[49] Fernandez-Cuesta, L., Perdomo, S., Avogbe, P.H., Leblay, N., Delhomme, T.M. et al. Identification of circulating tumor DNA for the early detection of small-cell lung cancer. *EBioMedicine*, 10:117–123, aug 2016. 15, 28, 79, 84, 87

[50] Fisher, R.A. On the interpretation of x2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87, jan 1922. 15

[51] Fonseca, N.A., Rung, J., Brazma, A. and Marioni, J.C. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, oct 2012. 9

[52] Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, nov 2016. 157

[53] Francioli, L.C., , Polak, P.P., Koren, A., Menelaou, A. et al. Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, 47(7):822–826, may 2015. 6

[54] Gan, K.A., Pro, S.C., Sewell, J.A. and Bass, J.I.F. Identification of single nucleotide non-coding driver mutations in cancer. *Frontiers in Genetics*, 9, feb 2018. 5

[55] Garrison, E. and Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv*, jul 2012. 18

[56] George, J., Lim, J.S., Jang, S.J., Cun, Y., Ozretić, L. et al. Comprehensive genomic profiles of small cell lung cancer. *Nature*, 524(7563):47–53, jul 2015. 157

[57] Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P. et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3(1), jan 2012. 22, 39

[58] Gordon, D.J., Resio, B. and Pellman, D. Causes and consequences of aneuploidy in cancer. *Nature Reviews Genetics*, 13(3):189–203, jan 2012. 12

[59] Grann, V.R. and Jacobson, J.S. Population screening for cancer-related germline gene mutations. *The Lancet Oncology*, 3(6):341–348, jun 2002. 28

[60] Guo, Y., Li, J., Li, C.I., Long, J., Samuels, D.C. et al. The effect of strand bias in illumina short-read sequencing data. *BMC Genomics*, 13(1):666, 2012. 16, 78

[61] Guo, Y., Zhao, S., Sheng, Q., Ye, F., Li, J. et al. Multi-perspective quality control of illumina exome sequencing data using QC3. *Genomics*, 103(5-6):323–328, may 2014. 75

[62] Hashimoto, T.B., Edwards, M.D. and Gifford, D.K. Universal count correction for high-throughput sequencing. *PLoS Computational Biology*, 10(3):e1003494, mar 2014. 35

[63] Heitzer, E., Perakis, S., Geigl, J.B. and Speicher, M.R. The potential of liquid biopsies for the early detection of cancer. *npj Precision Oncology*, 1(1), oct 2017. 28

[64] Hench, I.B., Hench, J. and Tolnay, M. Liquid biopsy in clinical management of breast, lung, and colorectal cancer. *Frontiers in Medicine*, 5, jan 2018. 28

[65] Homer, N. and Nelson, S.F. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biology*, 11(10):R99, 2010. 14

[66] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001. 6

[67] Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K. et al. Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, oct 2016. 7

[68] Jiang, Y., Oldridge, D.A., Diskin, S.J. and Zhang, N.R. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Research*, 43(6):e39–e39, jan 2015. 173

[69] Jones, D., Raine, K.M., Davies, H., Tarpey, P.S., Butler, A.P. et al. cgpCaVEManWrapper: Simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Current Protocols in Bioinformatics*, 56(1):15.10.1–15.10.18, dec 2016. 20

[70] Kanda, M., Matthaei, H., Wu, J., Hong, S., Yu, J. et al. Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia. *Gastroenterology*, 142(4):730–733.e9, April 2012. 117

[71] Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15(8):591–594, jul 2018. 25, 173

[72] Kinde, I., Munari, E., Faraj, S.F., Hruban, R.H., Schoenberg, M. et al. TERT promoter mutations occur early in urothelial neoplasia and are biomarkers of early disease and disease recurrence in urine. *Cancer Research*, 73(24):7162–7167, oct 2013. 132

[73] Knudson, A.G. Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, apr 1971. 8

[74] Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, feb 2012. 20

[75] Krimmel, J.D., Schmitt, M.W., Harrell, M.I., Agnew, K.J., Kennedy, S.R. et al. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proceedings of the National Academy of Sciences*, 113(21):6005–6010, May 2016. 167

[76] Kurtzer, G.M., Sochat, V. and Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459, may 2017. 37

[77] Laehnemann, D., Borkhardt, A. and McHardy, A.C. Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, 17(1):154–179, may 2015. 173

[78] Lagerkvist, U. "two out of three": an alternative method for codon reading. *Proceedings of the National Academy of Sciences*, 75(4):1759–1762, apr 1978. 7

[79] Lahens, N.F., Ricciotti, E., Smirnova, O., Toorens, E., Kim, E.J. et al. A comparison of illumina and ion torrent sequencing platforms in the context of differential gene expression. *BMC Genomics*, 18(1), aug 2017. 10

[80] Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, 44(11):e108–e108, apr 2016. 20

[81] Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M. et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001. 8, 170

[82] Langmead, B. and Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, 19(5):325–325, feb 2018. III

[83] Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, dec 2011. 20

[84] Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, sep 2011. 37

[85] Li, H. and Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, may 2009. 10, 171

[86] Li, H. and Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, may 2010. 171

[87] Li, H., Bloom, J.M., Farjoun, Y., Fleharty, M., Gauthier, L. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature Methods*, 15(8):595–597, jul 2018. 113

[88] Li, H. and Durbin, R. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, jan 2010. 10, 171

[89] Liaw, A. and Wiener, M. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. 106, 108

[90] Libbrecht, M.W. and Noble, W.S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, may 2015. 25, 26

[91] Maley, C.C. and Greaves, M. *Frontiers in Cancer Research*. Springer New York, 2016. 7

[92] Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Loo, P.V. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886, may 2015. 22, 34, 39

[93] Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F. et al. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, oct 2018. 34

[94] Mattos-Arruda, L.D., Mayor, R., Ng, C.K.Y., Weigelt, B., Martínez-Ricarte, F. et al. Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nature Communications*, 6(1), nov 2015. 28

[95] McGranahan, N. and Swanton, C. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168(4):613–628, feb 2017. 12

[96] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K. et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, jul 2010. 19, 38

[97] Mendelsohn, J., Howley, P.M., Israel, M.A., Gray, J.W. and Thompson, C.B. *The colecular basis of cancer - third edition*. Elsevier, 2008. 8

[98] Meo, A.D., Bartlett, J., Cheng, Y., Pasic, M.D. and Yousef, G.M. Liquid biopsy: a step forward towards precision medicine in urologic malignancies. *Molecular Cancer*, 16(1), apr 2017. 28

[99] Mose, L.E., Wilkerson, M.D., Hayes, D.N., Perou, C.M. and Parker, J.S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*, 30(19):2813–2815, jun 2014. 14, 171

[100] Mouliere, F., Chandrananda, D., Piskorz, A.M., Moore, E.K., Morris, J. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Science Translational Medicine*, 10(466):eaat4921, nov 2018. 132

[101] Muzzey, D., Evans, E.A. and Lieber, C. Understanding the basics of NGS: From mechanism to variant calling. *Current Genetic Medicine Reports*, 3(4):158–165, sep 2015. 18

[102] Ng, P.C. and Henikoff, S. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, jul 2003. 7, 157

[103] Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, jun 2011. 18

[104] Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M. et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, 8(1), jul 2018. 16

[105] Pinsky, P.F. Lung cancer screening with low-dose CT: a world-wide view. *Translational Lung Cancer Research*, 7(3):234–242, jun 2018. 28

[106] Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, nov 2017. 18, 19, 20, 25, 38

[107] Quail, M., Smith, M.E., Coupland, P., Otto, T.D., Harris, S.R. et al. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13(1):341, 2012. 16

[108] Ravasio, V., Ritelli, M., Legati, A. and Giacopuzzi, E. GARFIELD-NGS: Genomic vARiants FIltering by dEep learning moDels in NGS. *Bioinformatics*, 34(17):3038–3040, apr 2018. 25

[109] Rimmer, A., , Phan, H., Mathieson, I., Iqbal, Z. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912–918, jul 2014. I

[110] Rivlin, N., Brosh, R., Oren, M. and Rotter, V. Mutations in the p53 tumor suppressor gene: Important milestones at the various steps of tumorigenesis. *Genes & Cancer*, 2(4):466–474, apr 2011. 157

[111] Robasky, K., Lewis, N.E. and Church, G.M. The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1):56–62, dec 2013. 14, 77

[112] Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A. and Mesirov, J.P. Variant review with the integrative genomics viewer. *Cancer Research*, 77(21):e31–e34, oct 2017. 23

[113] Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S. et al. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, jan 2011. 15, 23

[114] Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G. et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7):907–913, jan 2012. 20

[115] Sanger, F., Nicklen, S. and Coulson, A.R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, dec 1977. 8

[116] Sant, M., Allemani, C., Santaquilani, M., Knijn, A., Marchesi, F. et al. EUROCARE-4. survival of cancer patients diagnosed in 1995–1999. results and commentary. *European Journal of Cancer*, 45(6):931–991, April 2009. 117

[117] Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J. et al. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, may 2012. 18, 21

[118] Sausen, M., Phallen, J., Adleff, V., Jones, S., Leary, R.J. et al. Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. *Nature Communications*, 6(1), July 2015. 167

[119] Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C. et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, apr 2017. 5

[120] Shen, L., Shi, Q. and Wang, W. Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis*, 7(3), mar 2018. 7

[121] Shen, R., Fan, J.B., Campbell, D., Chang, W., Chen, J. et al. High-throughput SNP genotyping on universal bead arrays. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 573(1-2):70–82, jun 2005. 9

[122] Shen, R. and Seshan, V.E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, 44(16):e131–e131, jun 2016. 12

[123] Shendure, J. and Ji, H. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, oct 2008. 8, 9

[124] Shin, H.T., Choi, Y.L., Yun, J.W., Kim, N.K.D., Kim, S.Y. et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nature Communications*, 8(1), nov 2017. 14, 83

[125] Springer, S.U., Chen, C.H., Pena, M.D.C.R., Li, L., Douville, C. et al. Non-invasive detection of urothelial cancer through the analysis of driver gene mutations and aneuploidy. *eLife*, 7, March 2018. 167

[126] Spurr, L., Li, M., Alomran, N., Zhang, Q., Restrepo, P. et al. Systematic pan-cancer analysis of somatic allele frequency. *Scientific Reports*, 8(1), may 2018. 11, 19

[127] Su, X., Zhang, L., Zhang, J., Meric-Bernstam, F. and Weinstein, J.N. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, 28(17):2265–2266, jun 2012. 12

[128] Summa, S.D., Malerba, G., Pinto, R., Mori, A., Mijatovic, V. et al. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics*, 18(S5), mar 2017. 24, 25

[129] The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, aug 2011. 28

[130] Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, apr 2012. 15, 23

[131] Tian, R., Basu, M.K. and Capriotti, E. Computational methods and resources for the interpretation of genomic variants in cancer. *BMC Genomics*, 16(Suppl 8):S7, 2015. 13

[132] Tommaso, P.D., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. et al. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, apr 2017. 37

[133] Torracinta, R. and Campagne, F. Training genotype callers with neural networks. *bioRxiv*, dec 2016. 25

[134] Tukey, J.W. *Exploratory Data Analysis*. Addison-Wesley, 1977. 36

[135] van den Akker, J., Mishne, G., Zimmer, A.D. and Zhou, A.Y. A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing. *BMC Genomics*, 19(1), apr 2018. 25

[136] Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J. et al. Frequency of TERT promoter mutations in human cancers. *Nature Communications*, 4(1), jul 2013. 67

[137] Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. et al. Cancer genome landscapes. *Science*, 339(6127):1546–1558, mar 2013. 7

[138] Waddell, N., , Pajic, M., Patch, A.M., Chang, D.K. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540):495–501, feb 2015. 117

[139] Waters, A.M. and Der, C.J. KRAS: The critical driver and therapeutic target for pancreatic cancer. *Cold Spring Harbor Perspectives in Medicine*, 8(9):a031435, December 2017. 117

[140] Watson, J.D. and Crick, F.H.C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, apr 1953. 4

[141] Wetterstrand, K. Dna sequencing costs: Data from the nhgri genome sequencing program (gsp) [online]. available: www.genome.gov/sequencingcostsdata., April 2018. 8

[142] Williams, M.J., Werner, B., Heide, T., Curtis, C., Barnes, C.P. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50(6):895–903, may 2018. 12

[143] Winkler, H. *Verbreitung und Ursache der Parthenogenesis im Pflanzen - und Tierreiche*. 1920. 3

[144] Wishart, D.S. Is cancer a genetic disease or a metabolic disease? *EBioMedicine*, 2(6):478–479, jun 2015. 3

[145] Wong, S.Q., , Li, J., Tan, A.Y.C., Vedururu, R. et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Medical Genomics*, 7(1), may 2014. 14, 83

[146] Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24, 2018. 18, 34

[147] Ye, H., Meehan, J., Tong, W. and Hong, H. Alignment of short reads: A crucial step for application of next-generation sequencing data in precision medicine. *Pharmaceutics*, 7(4):523–541, nov 2015. 169

[148] Zhernakova, D.V., de Klerk, E., Westra, H.J., Mastrokolias, A., Amini, S. et al. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genetics*, 9(6):e1003594, jun 2013. 9, 10

[149] Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3):246–251, feb 2014. 113