Aspects of Penicillium genomics : Molecular combing genome assembly, genetic exchange in food and potential for secondary metabolite production

Kevin Cheeseman

▶ **To cite this version:**

Kevin Cheeseman. Aspects of Penicillium genomics : Molecular combing genome assembly, genetic exchange in food and potential for secondary metabolite production. Agricultural sciences. Université Paris Sud - Paris XI, 2013. English. NNT : 2013PA112280 . tel-01124167

HAL Id: tel-01124167

https://theses.hal.science/tel-01124167v1

Submitted on 6 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE PARIS-SUD

Ecole Doctorale Agriculture, Alimentation, Biologie, Environnement, Santé

INRA-Micalis & Genomic Vision

THÈSE DE DOCTORAT en Sciences de la Vie

soutenue le 20/11/2013
par

Kevin Cheeseman

Aspects of *Penicillium* Genomics: molecular combing genome assembly, genetic exchange in food and potential for secondary metabolite production.

Directeur de thèse :	**Pierre Renault**	Directeur de recherches, INRA, Jouy-en-Josas
Co-encadrant :	**Maurizio Ceppi**	Chef de Projet, Genomic Vision, Bagneux
Composition du jury :		
Président du jury :	**Armel Guyonvarch**	Professeur université Paris Sud XI, Orsay
Rapporteurs :	**Sylvie Dequin**	Directrice de recherches, INRA, Montpellier
	Ronald De Vries	Directeur de recherches, CBS-KNAWL, Utrecht
Examinateurs :	**Cecile Neuvéglise**	Directrice de recherches, INRA, Grignon
	Eric Record	Directeur de recherches, INRA, Marseille

Acknowledgements

After so many years, it is finally time to conclude… It dawns on me that I am indebted to so many people it would require more than this acknowledgment to express my gratitude to all the wonderful people I met during this adventure. Memories surge to my mind, and I will never forget them. I should start by these words to all of you: Thank you!

First of all, my thanks go to the members of my jury. I am indebted to Mrs. Sylvie Dequin and Mr. Ronald De Vries for accepting to act as reviewers of my work. Equally, my thanks go to Mrs Cécile Neuvéglise and Mr Eric Record for being my examiners. Thank you for giving up some of your precious time to read and comment my work. Mr Armel Guyonvarch thank you so much, for accepting to act as "President du Jury" and for being so kind and helpful in your link with the doctoral school and the university. To all of you, it is amazing as a student to have you in my jury!

Above all, and very special to me, I will never be able to thank my advisors enough for their guidance through these years of learning. Pierre, Maurizio: Thank you! These words do not rightfully describe my feelings for all you have done for me.

Pierre, you have been amazing in your guidance, unfaltering even in times of hardship.

Maurizio, you have always been there for me.

Christine, Anne-Laure, Eric, thank you so much for stepping in when I needed it the most. This manuscript could never have been without you.

Yves thank you so much for your work and support and cheerfulness during this thesis. So many results would not be there if you had not been around!

Aaron, Daniel, Erwan, thank you so much for giving me the opportunity to perform a PhD at Genomic Vision. Witnessing a start-up grow is an amazing experience.

Joelle, Rachel, Pierre, thank you so much for reading parts of my manuscript.

Thank you to the "Wallaby team" for all these discussions and meetings.

Emmanuel, thank you for our discussions and your advices.

A very special thanks to my colleagues and friends at Genomic Vision. A special mention to Charlotte, Emilie, Fanny, Stéphanie, Aurélie, Agnès, Solenne, Yannick, Fabrice, Patricia, Sébastien, Guillaume & Jun.

A very special thanks to my colleagues and friends I met at the INRA. A special mention to Victoria, Mathieu, Victor, Nicolas, Emmanuelle, Edi, Mihai, Fouad, Florian, James, Mathieu, Moez, Hela, Celine, Nicolas, Benoit… You made this place very special to me.

Thank you to my family. You are a constant source of inspiration and support. Thank you Mum, Dad, Greg, Carine... and Shadow! You did so much for me.

Thank you Lucie... for everything.

Contents

Foreword:

What is genomics, and why genomics?

Born from the encounter of various, diverse scientific fields and discoveries in the late XXth century, genomics is a recent discipline of biology studying the molecular organisation of nucleic acids, molecules belonging to the smallest level of organisation of life accessible. It aims at discovering how these molecules encode the necessary information for the cell to function and how it acts as a memory support when passed on over generations.

The work presented herein belongs to Genomics by several aspects, notably the acquisition of whole genome sequences and the structure and dynamics of genomes. The genomes under scrutiny are micro-organisms living in an anthropomorphic environment. To fully integrate the work here in the bigger picture, one should understand what genomics is and where does it sit in biological sciences and society. This foreword is here for this very purpose. It should not be seen as a history of genetics and genomics (diving into the ocean of discoveries and technological innovations that led to genomics would require more than just a foreword) nor as a philosophical digression but simply as what it is: a foreword aiming at giving the necessary primers to grasp why such studies exists.

Origin of life, who am I?

"Who am I?" has always been a recurrent question throughout Man's history. This question can only be addressed through understanding the origin of life and how it functions. Comprehend, or grasp what life and its complexity are, is a key step towards resolving the origin of life. Indeed, there are many different life forms, as exemplified by the striking morphological differences of multicellular organisms, be they wasp, birds, plants, mammals or worms. Microorganisms too, many of which are impossible to observe by eye, are also incredibly diverse. Biodiversity is the embodiment of life's complexity. Bacteria, archea, and unicellular eukaryotes, as well as microscopic fungi, each possess unique features. Yet their organisations share common features. This paradox between unity and difference is a dogma of biology.

In its constant quest to understand its origins, Man tried to classify life forms. Biology or the study of Life was historically, and is still today typically carried out at different levels of organisation. They each represent a coherent unity between structure and function. These levels of organisation reflect different orders of complexity: biomes, ecosystems, populations, organisms, organs (for multi-cellular organisms), cells, organites, molecules and finally atoms. One level is typically defined by

structural constraints from the underlying level, and selective constraints from the level above. Indeed, our body (organism) is made of many organs carrying different functions, themselves made of specialised cells. In turn cells are made of molecules. In a cell, these molecules can be classified in four main types namely lipids, glucids, proteins, and nucleic acids. Genomics aims at studying one part of the molecular level of organisation, the nucleic acids. It is at the present day the smallest level of organisation accessible to life science researchers to understand how life functions. The genome can be seen as the whole set of nucleic acids carrying the necessary instruction for the cell to function, as well as a memory support on which genetic information can be safeguarded.

Perhaps, however concepts and goals of genomics existed since the birth of civilisations. Natural History and Genomics share this common approach that is a cornerstone in life sciences – they compare several objects and classify them based on common features and differences. Indeed, early in the antiquity, Aristotle and many others were already classifying life forms. It seems such an endeavour, mainly to gain an understanding of who we are, has always been a constant quest of mankind.

A not so brief history

However, despite early classifications of living beings in the antiquity, and advances in "primitive" medicine during the middle age, it was only around the mid XIXth century that discoveries and progresses in natural sciences started to shape biology into what would be cellular biology, molecular biology and genomics. This led from natural history to modern biology. Perhaps - almost at the same time - when Charles Darwin was on board the H.M.S. Beagle and Mendel applying rigorous observation and statistical analysis to garden peas, were produced the sparks of innovation that would spur the development of novel methodologies and studies, leading to immense progress resulting in the advent of cellular and molecular biology, evolutionary theories, and indeed genetics.

Genomics is itself a field in which all these disciplines merge. Genomics could never have been without Schleiten and Schwann discovering the cell, or computer science unlocking powerful means to rapidly share data and ideas. Not to mention all the technological innovations (microscope, sequencing, computers, the internet, to name only a few) necessary to the study of nucleic acids.

Indeed in 1858, when Wallace and Darwin jointly announced their findings under the theory of evolution via natural selection, and a year later when "On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life" was published, our understanding of life took a great step forward. It is the first description of how new species arise via evolution and how natural selection uses natural variation to evolve new forms. This concept has

since then, despite being questioned and refined, never been proven wrong, and has been, and will certainly continue to be a keystone concept in biology, never far away from any experiment and thematic. Equally important, at that time, scientific observation in life sciences took another visage, when a scientific discovery transformed into a debate of great importance to society.

A few years later, in 1865, Gregor Mendel discovered the concept of particulate inheritance, or gene inheritance. Largely unnoticed, due to research at that time being carried out in isolation, and often published in different languages, his work will be rediscovered in the early 1900s, and when combined to Darwin's work, provide the foundation for modern biology. Noteworthy, his work outlines the importance of developing pure lines, and statistical analysis of the data (which other scientists further emphasized in these times). His crossing approach for following phenotypes over several generations is still the only approach used today to understand the genetic inheritance of a trait. During this century, the location of the heredity material to the nucleus was proposed by Haeckel, while Mieschler concluded the material in the nucleus is nucleic acid. Many other scientists made important observations on the behaviour of chromosomes (as testified by August Weismann's theory of chromosomal behaviour in 1887).

Early in the XXth century, Mendel's work will be rediscovered, confirmed, and extended. This period marks the beginning of the age of genetics. The chromosomal theory of inheritance was formalised, defining the physical properties of parental chromosomes and their separation during meiosis. Sutton and Bovery's work suggests chromosomes are the material on which heredity lies. Concomitantly, chromosomal aberrations, so important in cancer progression, are observed. In these years, groundbreaking research carried out by many scientists (Correns, De Vries, von Tschermak, Bateson, to name only a few...) will lead to the concepts underlying modern genomics. Indeed, in the 1900s, "gene", "genotype", "phenotype" were coined. The theory of genetic equilibrium was formulated by Hardy and Weinberg, and still today underlies population genomics. Morgan proposes that genes reside on chromosomes and establish the fruit fly drosophila as a model organism.

The mid XXth century witnessed the demonstration that DNA (or deoxyribonucleic acid) is the genetic material. Griffith's transformation experiment of non-lethal bacterium in a lethal one by other bacterium paved the way for Avery, McLeod and McCarthy to determine that DNA, not protein or RNA is responsible for the conversion. Hershey and Chase further confirmed this experiment. Watson and Crick discovered the structure of DNA in 1953, paving the way for the elucidation of replication, protein synthesis, gene expression, and recombination of genetic material. Soon after, the genetic code was deciphered. Exploitation of all these processes provided tools and technologies for molecular biology, such as recombinant DNA, molecular cloning and gene expression, new types of

enzymes (restriction, ligases, polymerases...). Quite importantly, this multitude of discoveries and progresses led to the creation of the biotechnology industry, which has a tremendous impact on society today.

The birth of genomics

Following the discoveries from the past decades, and notably the process of DNA replication, sequencing techniques are proposed by Maxam and Gilbert and by Sanger in the 1970s. These sequencing methodologies, as well as protein sequencing, enabled life science researchers to access another one of life's level of organisations. Species could now be compared at the molecular level, leading to an intensive era of phylogenetic studies that saw the emergence of many algorithms and software. Kimura proposed the neutral theory of evolution, which questioned Darwin's work. However both theories are not mutually exclusive, leading to our current knowledge of evolution.

Finally these decades are also the advent of computer and information technologies. More than crucial, they are an essential component of modern genomics. They allowed rapid data sharing, creation of databases, and computational analysis. 1969 saw the development of both Arpanet, a link to computers at four universities, which is thought by many as the beginning of the internet (which will be developed in 1974), and the release of the UNIX operating system, so important to scientific computing.

In 1980, the first complete genome sequence is published. It is the one of the bacteriophage ΦX174 of *E. coli*. Soon after (1986), the first automated DNA sequencing system is constructed by Leroy Hood and colleagues. The first genome sequence of a bacterium is released in 1995. A year later, the genome of *S. cerevisiae* is published. This is the first complete eukaryotic genome. These sequences will be followed by many others from model organisms (including *C. elegans*, *A. thaliana*, *D. melanogaster*...). In 2001, the human genome is published. Expected to be the key to curing many diseases and understanding many traits, its state – with many errors and regions of uncertainties-, and the current knowledge of the structural organisation of DNA partially prevents the rapid execution of such utopic goals. It also leads to controversy and debate about the use of genomic data, highlighted by the fierce competition between the publicly funded project and Celera Genomics.

The advances in automation and technology development have constantly decreased the prohibitive cost of genome sequencing. In 2004 and 2005, new sequencing technologies arrived on the market, which was until then dominated by automated Sanger sequencing. These technologies promised high-throughput, lower costs, and bigger amounts of data generated in a single experiment. This

brought genome sequencing, previously the hunting ground of large scale genome centres to smaller laboratories. It resulted in a deluge of a data, new applications of sequencing and many issues surrounding experiments, data storage and sharing, as well as ethical issues, raised by the fact that anyone can now basically sequence anything.

Multitude of applications

Modern Genomics uses more or less the same type of technologies and approaches to address many different questions and needs. Today, many different technologies with advantages and drawbacks are available to researchers and clinicians. Projects can study a single genome, many different genomes of the same species, many different genomes of the same environment, tumour genomes, single cell genomes, transcriptomes, exomes, some set of genes... Some projects aim at deciphering the human genome and its related disease or associated conditions, while others are sequencing species of industrial importance, for a particular trait (goat cashmere, for example). Crop sciences aim at deciphering traits of interest, while some scientist are aiming at understanding the genomic basis of host-pathogen interaction to gain a better pest control. Applications can be as diverse as biological warfare or disease management. Personalised medicine is on the verge of becoming a reality, while new sequencing technology could revolutionise patient care and our understanding of diseases.

Ethical and societal concerns

The rapid increase in theoretical and technological concepts in the latest 150 years or so dissociated science and philosophical thinking. Indeed life sciences have recently undergone a dramatic specialisation in each of its fields, leading researchers to focus on one area of research. However, Genomics is raising many issues and concerns around ethics and society. More and more species are sequenced each day, ecosystems can now be characterised through genomics (metagenomics), and comparative genomics enables new insights into the origin of modern man. Species can be engineered to be more productive or resistant. Genetics conditions can be detected prior to birth, leading to ethical questions. Genetically Modified Organisms -Despite representing an impressive solution to famine, and pest control- are feared to spread and take over new species; globalisation is leading to the encounter of diverse species, occasionally leading to the emergence of pathogens. Genomics also has applications in the food industry, especially as a quality control means. Recent scandals in the food industry are a testimony of society's expectancies around precise quality control.

The ever changing landscape of sequencing technologies, the accessibility of sequencing to any laboratory and the lack of standard analysis procedures, as well as the lack of policies concerning the quality of data and its storage, renders the need for as vast as possible a technological toolbox to be available. Indeed researchers in genomics need to be able to validate sequencing data using other tools since the quality of the data shared, as well as the quality of results, have a tremendous impact on society.

The work presented hereafter belongs to genomics in several ways. On the technological point of it, it reports the development of novel approach to validate and further improve sequencing data, by the use of a single DNA molecule technology, Molecular Combing, thus returning to in vitro experiments after *in-silico* predictions. On the fundamental point of view, it reports aspects of comparative genomics between several species of *penicillium*, a taxon in which whole genome sequences were lacking. It also reports exchange of genetic material between these sequenced species and discusses implications of such events from a scientific point of view as well as its potential impacts on the food industry and on ecosystems.

Context:

The work reported in this manuscript was funded by a CIFRE grant ("Convention Industrielles de Formation par la REcherche" – Industrial Convention of Training through REsearch") from the ANRT, the French national agency for technology and research. This type of funding is designed to foster industrial – academic relationships by jointly carrying a research project and through hosting a graduate student in the two environments. As a consequence, I was during the time of this thesis a member of two laboratories, "Genomic Vision", a biotechnology company and the "Food and Commensal Bacteria" team (Bactéries Alimentaires et Commensales-Bac) of the INRA-Micalis institute. The direct consequence of this is expected outcomes of the project of different nature: It must result in both the generation of knowledge and the generation of applications, intellectual property or know-how of industrial interest.

Genomic Vision is a biotechnology company founded in 2004, as a spin-off of the Pasteur Institute by Aaron Bensimon and Daniel Nerson. Its main mission and focus is the development and application of a proprietary technology - Molecular Combing- to Life Science in clinical diagnostics, pharmacogenomics and research. Its main focus is the development of diagnostic tests answering unmet medical needs in genetic and genomic diseases, as well as the assessment of replication dynamics and kinetics through the Replication Combing Assay. Molecular Combing is a molecular biology technology enabling to stretch single DNA molecules onto microscopy coverslips. Subsequent fluorescence *in-situ* hybridisation experiments enable the monitoring of genomic landscape at specific loci, on a single molecule, genome wide basis, over many genomes at the same time, thus providing a unique window of observation among DNA technologies. The essence of the work reported here sits in the exploration and development of a new application of Molecular Combing to provide answer to the genome assembly field, thereby providing a means to further improve genome assembly quality.

The INRA Micalis "Food and Commensal Bacteria" team focuses its activities on the study of microbial genetics in the context of human health and food safety. Its expertise lies within the commensal bacteria field, lactic acid bacteria, and within the metagenomics of food and food associated human environments. The team is involved in several high profile research projects in these areas. The project in which this work took place is the Food-Microbiomes project, coordinated by Dr. Pierre Renault of the food and commensal bacteria team. This project aims at generating a large body of knowledge through the application of conventional and innovative methodology to the cheese ecosystem. This will help in further characterising the cheese ecosystem and further assessing

the safe use and innocuity of the microorganisms occurring in the food environment. This project uses conventional microbiology technology and metagenomics to achieve these goals. For many of the micro-organisms occurring in the cheese environment, including the emblematic and traditionally used *Penicillium* species we do not possess a sufficient body of knowledge for an association to health and safety norms. One of the aim of the Food Microbiomes project was therefore to generate a sufficient knowledge of the cheese ecosystems and to further characterise these important microorganisms of the food supply chain One of the aim of the work reported here was therefore to provide a better overview of the genomics of *Penicillium* species with both academic outcomes in the genomics of a taxon of filamentous fungi, and to help in establishing their safe use in the food supply chain.

As a consequence, this manuscript reports several aspects of *Penicillium* genomics, in the improvement of the genome sequence of a *Penicillium* species, *Penicillium roqueforti*, in the observation of gene flux between *Penicillium* species in the cheese environment, and also provides a very early assessment of the genomic potential for secondary metabolite production.

Chapter One: Whole Genome Sequencing & Assembly or Unravelling a Genome's Sequence and Organisation.

Sequencing, and its ultimate goal, reconstructing the complete DNA sequence of an organism was until recently the hunting ground of large genome centres. The advent of second generation sequencing in the mid 2000s provided access to sequencing to many laboratories through reduced costs and higher throughputs. As a result, *de-novo* genome sequencing became a wider, less specialised business. This democratisation of sequencing and its subsequent assembly process came with an increase in data generation and knowledge, as well as quite often a decrease in quality due to lack of expertise, infrastructure and gold standards. The following chapter aims at describing how and why genome sequencing became what it is today. Starting with a brief history of sequencing, we introduce the main strategies for genome sequencing and subsequent assembly and discuss practical considerations affecting the quality of the final assembly. The aim is not to provide an in-depth review of technical points, but to highlight the complex landscape of genome sequencing, and how actual trends and techniques lead to more published genomes but a global reduction in quality.

1. Sequencing technologies: A brief history and overview

--Brief history of sequencing:

The first methods for DNA sequencing date back to the mid 70's. In 1975, Frederick Sanger announced a rapid method for determining sequences in DNA by primed synthesis using a DNA polymerase (Sanger 1975). Two years later, the DNA sequencing method involving chain terminating inhibitors from the Sanger lab and the Maxam and Gilbert method were published (Sanger et al. 1977, Maxam & Gilbert 1977). These two articles demonstrated the possibility of determining the DNA base sequence through biochemical reactions. The Sanger method became the most commonly used for the next decades. In 1986, the first automated DNA sequencer was announced (Smith et al. 1986; Ansorge et al. 1986). The need for tedious gel preparation disappeared in 1996 when ABI introduced the ABI Prism 310, an automated sequencer utilising gel slabs. Two years later capillary electrophoresis was added to sequencers, eliminating the need for gel preparation completely.

The result of a sequencing reaction has been called a read, because one deduces the order of the four bases in a sequential manner, thereby mimicking the reading of words made of an alphabet of four letters. Sanger sequencing typically produces reads of 600 to 900 base pairs long. Improvements and automation in technology drove the resolving power of sequencing experiments from a couple of reads, suitable for the investigation of the sequence of a small locus such as a gene, to massive sequencing able to produce reads all along a genome, in which large scale genome centres

specialised. They were until recently the only ones with the infrastructures able to support the cost, time and expertise required to sequence whole genomes.

This changed a few years ago when the first three "second generation technologies" were introduced in 2004 and 2005 (Margulies et al. 2005; Bennett 2004; Shendure et al. 2005). These technologies introduced different sequencing schemes, "sequencing by synthesis" (illumina), "sequencing by ligation" (SOLiD), and "pyrosequencing" (454). A complete description of the technical principles underlying the different sequencing technologies would be beyond the scope of this introduction. However, these second generation sequencing technologies and the now appearing third generation technologies (relying on single molecule sequencing) have had and continue to have a tremendous impact on genomics. By reducing the cost of sequencing by 100 to 1000 fold and through their ability to sequence DNA at an unprecedented pace, "Next Generation Sequencing" (NGS) unlocked novel biological applications and achievements. However, these dramatic changes in the way sequencing is carried out came with significant challenges and difficulties. New bioinformatics tools were and are still needed, new applications pop out at an unprecedented speed, and there are many concerns about issues surrounding data standardisation, management and storage.

-- The output of a sequencer is smaller than the object studied

The output of a sequencing reaction is a "read" of DNA with length ranging from tens of bases to several hundred bases depending on the sequencing technology used. Sequencing is usually carried out to unravel the base sequence of a particular locus, transcript, whole genome or transcriptome. Whatever the application intended, these reads of DNA sequences are smaller than the object studied. To overcome this limitation, a process called "*assembly*" is performed after the sequencing phase itself. Assembly is the process of stitching together several reads on the basis of shared sequence between two reads (fig 1). It enables the reconstruction of sequences theoretically up to complete chromosomes. Sequencing system manufacturers frequently improves read lengths by the use of new or refined chemistries and processes. Sanger Sequencing produces reads of size between approximately 600 to 900 bases, and second generation technologies produces reads ranging in sizes from 25 to 500 base pairs. Despite improvements in read lengths, these lengths are still short. Third generation sequencing technologies such as Pacific Biosciences sMRT now produces higher read length, from several hundreds to over 20 kilobases (Eid et al. 2009), but this increase in length comes at the price of an increased error rate in base calling (Koren et al. 2012; Nagarajan & Pop 2013). Downstream applications of sequencing are many, and usually aim at reconstructing a complete sequence of a locus of interest, a whole exome, transcriptome, genome, metagenome... Whichever the application intended, "assembly" is the computational process that uses overlap in sequences of

different reads to stitch them together, thereby reconstructing the information lost during the extraction and sequencing processes. Since producing a high quality genome sequence of a microbial genome was one of the goals of the work described in this manuscript, we focus here on the context of *de-novo* whole genome sequencing.

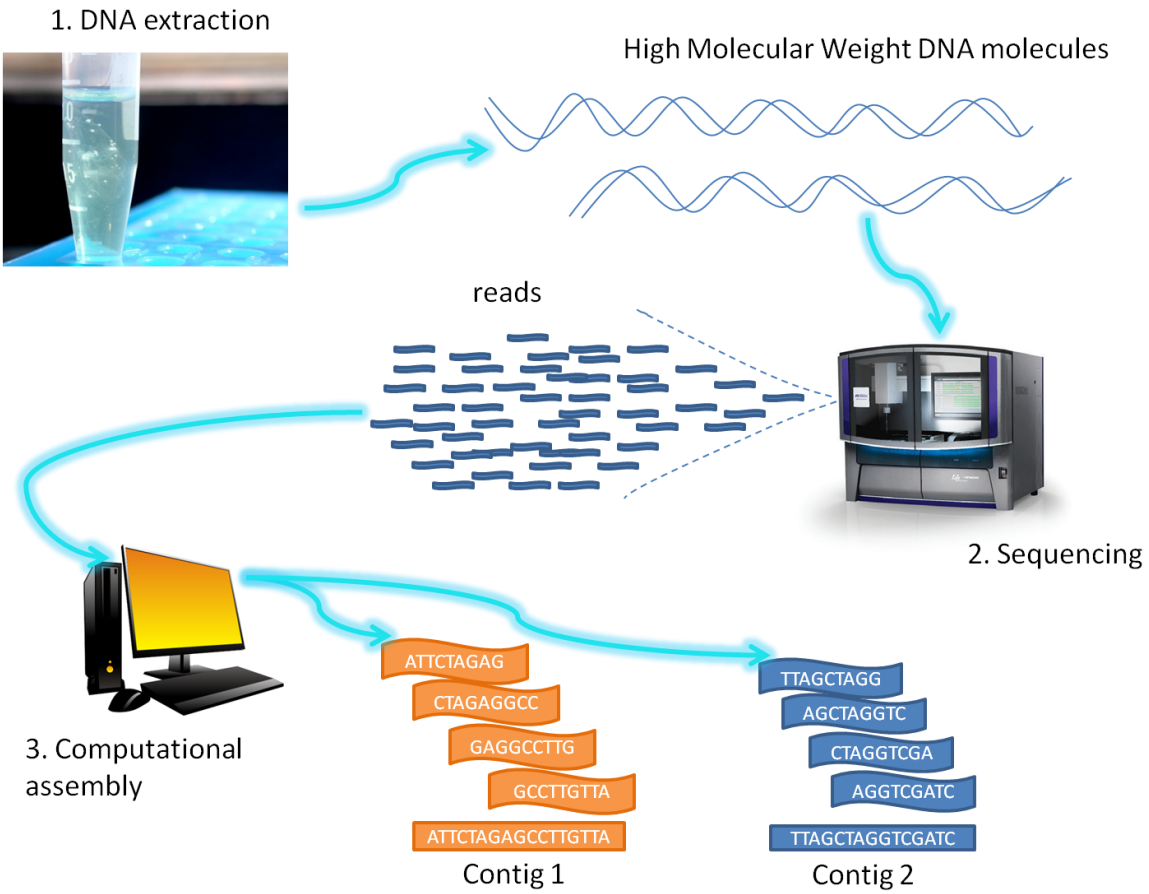


Figure 1: Sequencing & Assembly. High quality DNA extraction and preparation is achieved prior to the sequencing run (2). The sequencer outputs millions to billions of "reads", small strings of sequence. Assembly (3) is a computational process whereby reads sharing overlap are grouped together into contiguous sequences, or contigs.

2. Whole genome sequencing strategies:

Whole genome sequencing investigates the whole base content of a genome. Its ultimate goal is to produce completely resolved chromosomes of an organism. Genome structure and organisation, as well as gene content, regulatory elements, etc, can then be analysed. To do so, a sequencing project typically relies on one of two strategies, both of them coming with its own set of advantages, biases and limitations. The two main strategies used today are "hierarchical shotgun"; also called "BAC-by-BAC" approach and the "Whole Genome Shotgun" approach (Fig 2).

Hierarchical shotgun was historically the first approach used for whole genome sequencing. Hierarchical shotgun is performed in several steps. These include shearing genomic DNA and incorporating fragments into Bacterial Artificial Chromosomes (BACs), creating a library of such vector-insert structures representing as much of the genome as possible, defining by physical mapping a minimal tiling path through these BACs. This minimal tiling path enables sequencing and assembly through a chromosome walking approach, where each of the BACs from the minimal tiling path is sequenced and assembled into contigs corresponding to the sequence. These contigs are then assembled together according to the BAC physical map, thus representing as much of the genome as possible. This approach is usually employed today for large polyploid genomes including plant genomes, and large genomes with a lot of condensed chromatin (Szinay et al. 2008a; George et al. 2012). Perhaps the most emblematic example of this strategy is the sequencing of the human genome by the Human Genome Project consortium. The primary goal of this approach is to reduce the number of BACs to sequence to represent the whole chromosome, thereby diminishing the actual cost of sequencing. Main drawbacks include the costly and labour-intensive process of deriving a BAC library and defining a minimal tiling path. Besides, some regions of genomes are toxic to other organism, and as such, non clonable, leading to gaps in the minimal tiling path. Of note, different haplotypes, complex regions and copy number variations can render the mapping hazardous and despite finishing efforts lead to incomplete genomes.
(http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/, and see figure 8 later in this chapter; (Dolgin 2009)).

The second strategy, Whole Genome Shotgun (WGS), was first proposed in 1995 (Fleischmann et al. 1995), demonstrating what was previously thought to be an intractable feat –assembling a genome without the help of a physical map. A method in which no BAC library is used, WGS involves shearing the genomic DNA into many fragments, sequencing these fragments and then performing the assembly without any previous knowledge of the location of the genomic region where the read comes from. This approach, although obviously more challenging in terms of assembly (see below), and more likely to yield very fragmented genomes, was successfully applied to the human genome by Celera Genomics. Much controversy surrounded the competition which naturally took place between the Human genome Project public consortium and the private initiative of Celera Genomics. As a result, and with much fanfare, 2001 witnessed the publication of two complete human genome articles and assemblies (Venter et al. 2001; Lander et al. 2001). WGS produces far more fragmented genomes, where contigs and scaffolds are not easily anchored to chromosomes despite the fact that it is a faster and cheaper methodology. A scaffold is made of contigs linked together on the basis of external information, given by physical maps, or mate pair information. A mate pair or paired end

reads is the result of a particular DNA preparation where two extremities of a DNA fragment of a controlled length are sequenced. This strategy basically gives information about the distance separating two "mate" reads. As a consequence a scaffold is made of sequences of contiguous DNA and gaps parts of DNA that could not be sequenced, but of which the length is known due to the distance separating the two reads of the mate-pair).

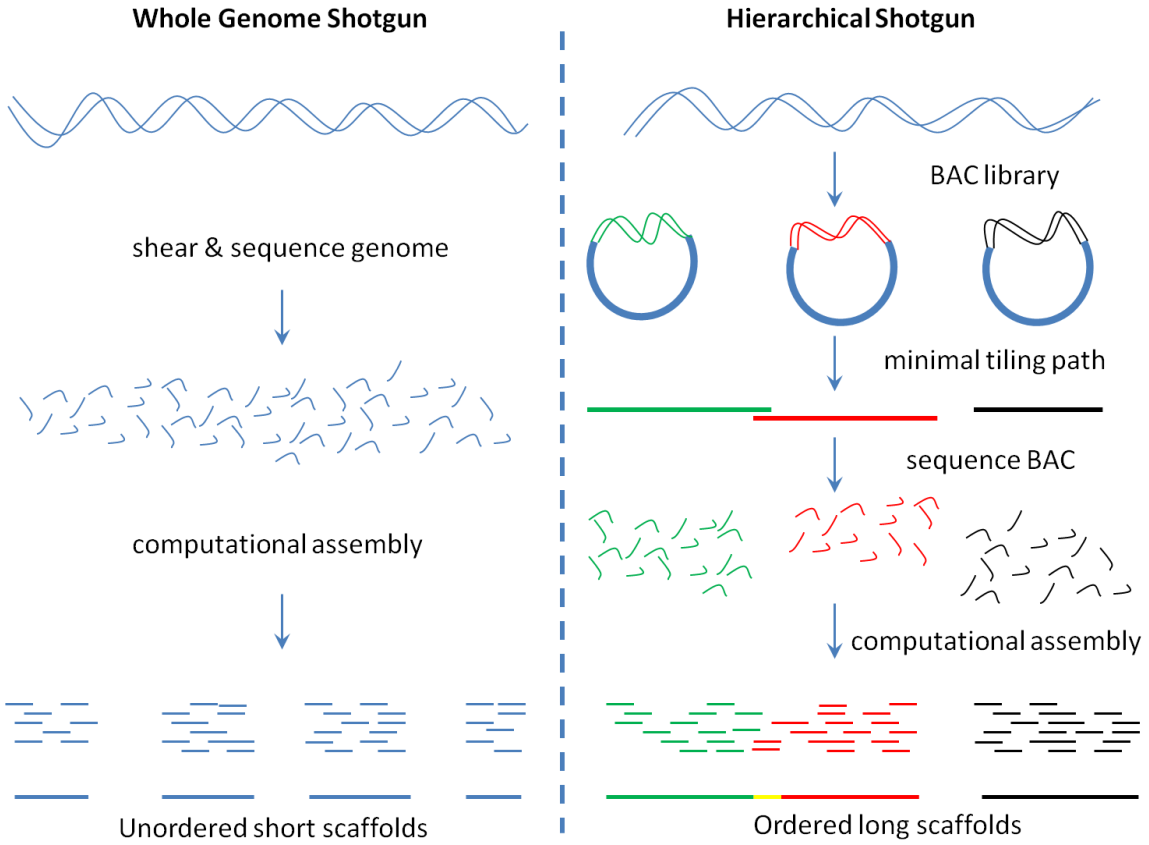


Figure 2: The two approaches to whole genome sequencing. "Whole genome shotgun" yields shorter and more numerous scaffolds, with the advantage of eliminating the BAC library and mapping of BAC clones required for a hierarchical approach. Scaffolds are longer and ordered in a hierarchical shotgun approach.

With the advent of second generation sequencing technologies, read lengths became shorter, but the cost and time to generate data diminished greatly. The considerably higher throughput of these technologies led to an increase in depth of sequencing, facilitating assembly from Whole Genome Shotgun sequencing (Paszkiewicz & Studholme 2010). Because of this, the WGS method is now the method of choice for microbial genome sequencing, and the hierarchical shotgun approach is mostly used in large and complex genomes.

The Assembly process represents a bottleneck for both approaches and for now is failing to completely reconstruct the chromosome or whole genome sequences of an organism. In the following section we present how assembly is performed and the current limitations of this computational process.

3. Assembly:

a. What is Genome Assembly?

Assembly is the process of piecing together the raw output of a sequencing machine. It is a necessary step of all but one application of sequencing (read mapping). Once the sequencing run is finished, the investigator is left with two possibilities. One can map the reads against a previously known sequence (a reference), as for instance is done when sequencing genes of clinical importance. This allows one to mine for differences in the base sequence (e.g. single nucleotide polymorphism) and missing stretches of a genome (e.g. genomic islands). One can also attempt at reconstructing a longer sequence by piecing together the reads, in a computational process called assembly. Assembly is necessary because the outputs of a sequencer are smaller than the object studied. Recent developments of sequencing technologies have had two kinds of impacts on this computational biology field, namely the development of new approaches compatible with the very short and reads of these technologies and the increased amount of reads to cope with, as well as the many application-oriented assembly strategies due to the massive pick up in biological questions addressed by sequencing (ChIP-seq, RNA-seq, *de novo* sequencing, re-sequencing, exome sequencing, etc.)

b. How is Assembly performed?

Assembly is always performed with the underlying assumption that reads with shared similarity are most certainly collinear and overlapping. An assembler uses one of several approaches to identify similarity between reads and stitch them together in an attempt to reconstruct the longer sequential string of nucleotides broken down during the sequencing process. The original DNA molecule is broken during the preparation step (extraction, mechanical breakage of the molecule) and because a sequencing system can only read a few bases of DNA (“technological” breakage of the molecule). As an information recovery process, assembly is sometimes impaired by the fact that the assumption of similarity going together with same location can be false. Genomes are full of repetitive sequences that confuse all existing assembly software. Moreover, genomes are sometimes diploid or polyploid. Haplotypes bear considerable similarity over long stretches while at the same time being different loci. For instance, the human reference genome assembly has been and is still today impaired by

regions of variability or hyper variability such as the MHC locus (http://infocus.nlm.nih.gov/2012/08/meet-ncbis-deanna-church-genom.html, see fig 8 for a current state of the human genome reference sequence). Adding to the complexity of resolving such repeats and similar regions, the error profile and shorter read length further augment this difficulty (and these are different for each sequencing technology, further complicating the problem). The more inaccurate a read is, and the more likely two different reads are to seem similar, when in reality they are not, leading the assembler to erroneously link them. Inversely, two reads from the same locus but with errors are likely to be placed at different loci by the assembler. To cope with this, several approaches are used by assembly software. The choice of one approach depends mainly on the read length and depth of sequencing (the average number of read per position), as well as the availability of "mate pairs" or "paired end" information (pairs of reads between which the distance and orientation are known). These approaches are embedded in a myriad of software released throughout the last three decades, developed according to the trends in sequencing at the time of development, each with specific applicability and advantages. Below we briefly describe the principles of these approaches, as underlying the assemblers they are partly responsible for the fragmented result. Importantly, these approaches and the software they are embedded in are the reflection of options and trade-offs between the ultimate goal of reconstructing a complete sequence and computational limitations and challenges.

The three main approaches used in assembly software are the Greedy approach, the Overlap-Layout-Consensus strategy and an approach relying on the construction of De Bruijn graphs.

-the greedy approach is an assembly paradigm whereby the assembler starts by grouping reads with the greatest immediate, local benefit. Reads with the best overlap will be assembled first; reads with fewer overlaps will be assembled afterwards. (See figure 3). This approach gives priority to local possibilities first, without taking into consideration any global relationship with other reads. This approach is not widely used anymore because it does not include any easy integration of long mate pair links. Examples of these assemblers are Phrap (Ewing & Green 1998), TIGR assembler (Pop & Kosack 2004) and VCAKE (Jeck et al. 2007).

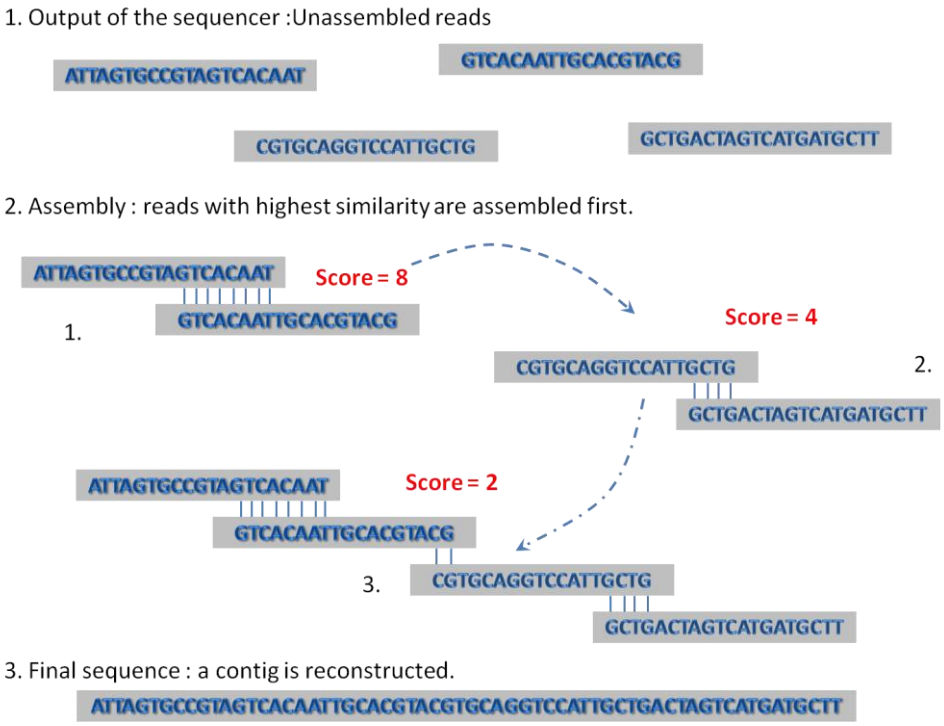


Figure 3: The greedy approach to assembly. A similarity score is used to link reads with the highest similarity first.

-The Overlap-Layout-Consensus (OLC) approach identifies all pairs of read overlapping sufficiently well, and regroups them in a graph (see fig 4) containing nodes representing the reads, and edges connecting each overlapping reads. This enables algorithms to take into account global relationships between reads. A variant of this approach, the string graph approach uses roughly the same process but removes any redundant information (*i.e.* bubbles in the graph cause by repeats). The Celera assembler (Huson et al. 2001) used to assemble the human genome was based on this approach. The OLC approach was abandoned when short read technologies appeared, due to problems in computational complexity. Shorter reads mean less overlap and higher throughput result in more reads to handle and thus more complex graphs. Recently, use of different and more memory efficient data structure led to a renewed interest in this approach for assembly software.

-The De Bruijn approach is a strategy where reads are subdivided into subsets of reads of length k. These subsets are called k-mers. In a similar fashion as in the OLC approach, a graph (fig4) is then constructed with paths in it representing possible assembly solutions. In a de Bruijn graph, the nodes represent each k-mer and the edges indicate adjacent k-mers overlapping by k-1 letters. Reads are thus implicitly modelled by paths through the graph. This approach can easily integrate external information such as mate-pairs to reduce the graph complexity (*i.e.* by removing theoretically possible paths between k-mers, which are known not to be present in the actual read set). One of the main drawbacks of this approach is that because it relies on k-mers to build the paths, it requires

a great base calling accuracy in the read. Because the third generation sequencing technologies tend to have longer reads and a lesser accuracy, this approach is most likely to be less frequently used in the near future. Examples of assemblers embedding this paradigm are EULER (Pevzner et al. 2001), VELVET(Zerbino & Birney 2008), SOAP de novo and ALLPATHS(Butler et al. 2008).

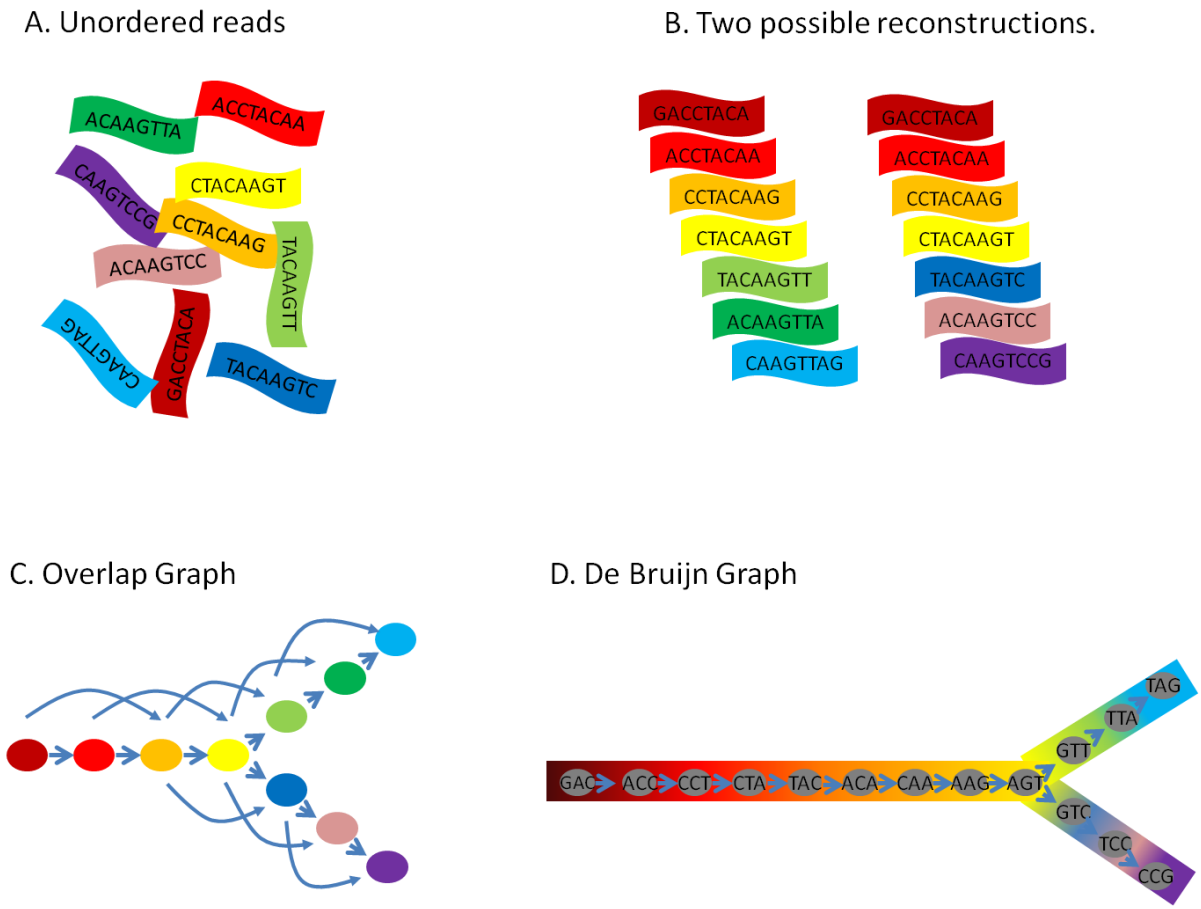


Figure 4: Different graphical representation of the assembly paradigm. An unordered set of reads (A) leading to two possible sequence reconstructions (B) can be abstracted as an Overlap graph, keeping into account relationships between reads (as shown by the colour code) or as a De Bruijn graph, where a set of k-mers (here k=3) and reads are implicitly depicted as paths through several nodes (as shown by the graduation in colour along the path). Of note a greedy algorithm would not have been able to resolve this assembly because of the same degree of difference at the fork. The fork illustrates the two sequences. Inspired from (Schatz et al. 2010).

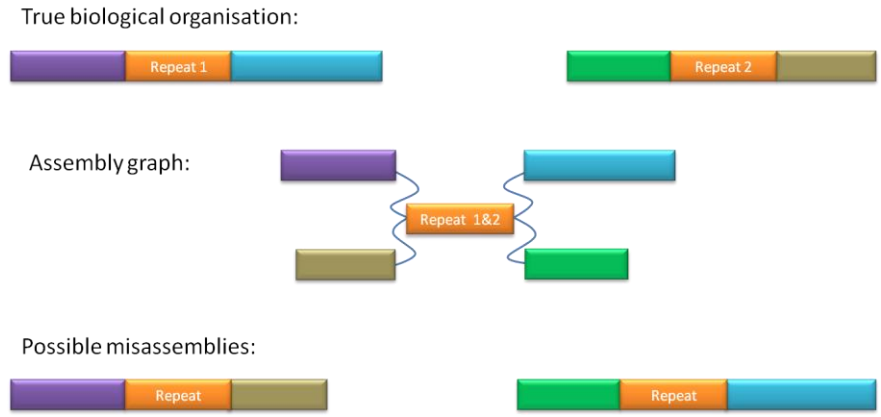
c. Pitfall and caveats

Genomes are full of repeats of different length with perfect, near perfect or imperfect identity. Repeats represent a major hurdle for assemblers (see fig 5 for examples). Because whatever the assembly paradigm underlying the software reads are grouped together by a shared overlap (or similarity) two reads overlapping two distinct repeats can be incorrectly merged by the assembler, thus creating a mis-assembly. The Greedy approach will “greedily” link reads with the best overlap with the repeat, and ignore the other possibilities. Repeats in graph-based approaches lead to the

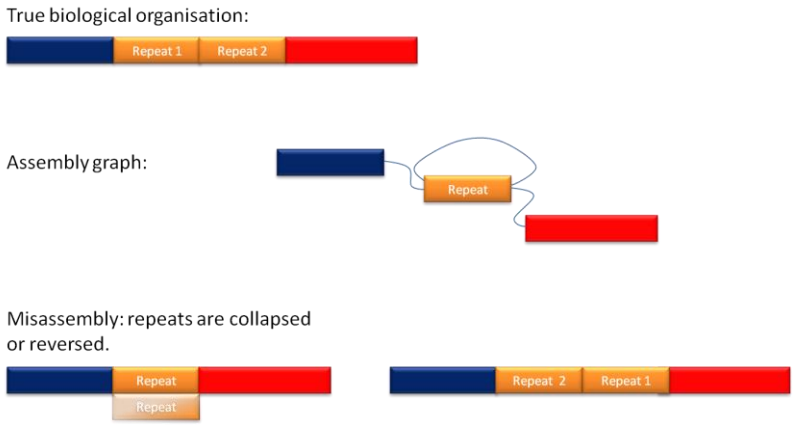
creations of bubbles, or loops in graphs, enabling an easy identification of the problem, but not an easy resolution. Theoretically, if the read length is longer than the repeat length, the repeat can be resolved and reads properly assembled (Nagarajan & Pop 2009). Further complicating the matter of repetitive element assembly, all reads are not error-free and this sometimes leads to the impossibility of assigning a read to its original locus, or worse, the assignment of this read to another locus. Because of this, there is a complex relationship between read length, base calling accuracy and genomic content in repetitive elements. Repetitive elements and other repetitive structures are numerous and of different kinds. They include transposable elements, which can be present in several copies with perfect or near perfect identity in different part of the genome. Other repeats exists too like for instance Short Interspersed Elements including alu sequences in primate genomes. These are known to be around 300 base pair long and represent a large fraction of the repetitive element content of the primate genomes, including *homo-sapiens*. Other repetitive structures of important biological significance confuse assemblers, such as Copy Number Variations (CNVs), which are iterative repetitions of a large portion of genomic content, often including genes or regulatory elements, present in different number on different alleles. These have been shown to account for major phenotypic differences among populations.

Repeats confuse assembly software and induce both interruptions of contigs at repetitive sequences and mis-assemblies by falsely joining two fragments of sequence ending with a similar repeat. Moreover, most of the assembly software collapse repeats and include procedures to eliminate reads originating from repeats before assembly and as a result only one copy of the element is present in the assembly, when several are present in the genome. Despite many strategies to overcome the major hurdle posed by repeats, be they in the graph treatment of the assembler or by using external information such as mate-pairs or paired end reads, all repeats cannot be easily resolved, and identification of mis-assemblies is a time-consuming and ungrateful job. If longer reads become more common with third generation sequencing, the accuracy of these reads is very low (15-40% of errors in Pacific Biosciences' sMRT technology for instance as well as a decreasing representation of longer reads (Nagarajan & Pop 2013). Assemblers then face another computationally difficult problem in the handling and alignment of reads having many errors. As such, it is unlikely that good quality assemblies come out only through bioinformatic treatment.

A.



B.



C.

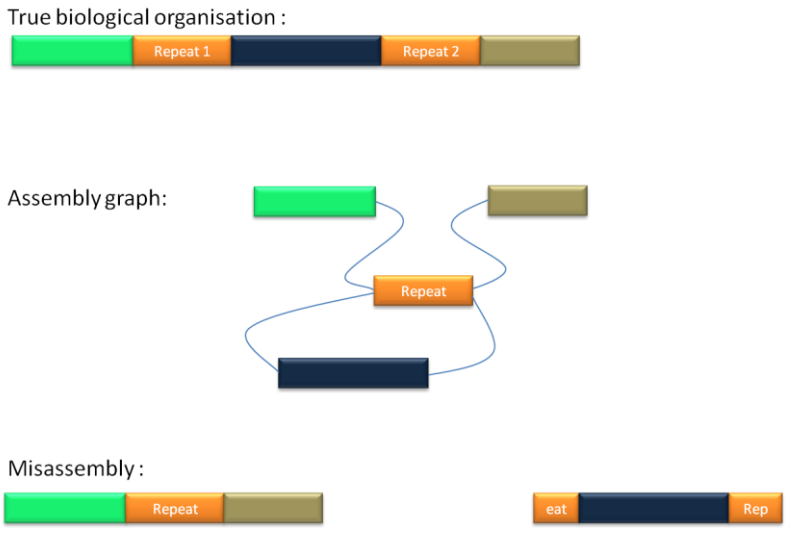


Figure 5: Examples of problems caused by repeats. Panel (A) shows possible misassemblies when two identical repeats are at different loci in the genome. Panel (B) displays the possibility of 2 tandem repeats collapsed or inversed. Panel (C) illustrates the possibility of a missassembly and a contig fragmentation when two identical repeats are close.

d. Trends and future, a practical matter

Another and last step of whole genome sequencing is the finishing step. This step aims at further linking contigs and scaffolds as well as correcting potential errors in the sequence. The end goal is to increase sequence accuracy, continuity and contiguity. These errors can be single base errors, small deletions or insertions, mis-assemblies, *i.e.* a stretch of sequence misplaced, or wrongly oriented, or a missing piece of sequence (called a "gap"). Once considered a mainstay in a genome project, critical to ensure high quality, gold standard sequences, in recent years it was deemed by many as too costly, labour intensive and not critical to address questions in some areas of research, particularly those looking at the global gene content of genomes, regardless of their location. By simply mapping a draft assembly or reads to a reference sequence, researchers can have a good idea about the gene content of a genome and of its missing parts. However, a closely related sequence is not always known, and if so is more often than not in the state of draft. Furthermore, strains have genomic differences, and the degree of difference bears on results deduced through comparative genomics. Draft assemblies already give a fairly good overview of the gene content of a newly sequenced species. However, draft assemblies are fragmented into hundreds to thousands of contigs and scaffolds, most of them containing mis-assemblies and errors. Despite many scientists advocating for high quality and gold standards in genomic sequences, the advent of second generation sequencing technologies, with their dramatically shorter reads increasing the complexity of the assembly process, as well as their reduced operating costs favoured the abandon of the finishing step in most genome projects (Lewin et al. 2009; Ricker et al. 2012). This combined to the democratisation and ever increasing ease of genome sequencing - Benchtop sequencers are now hitting the market! - to lead to a dramatic increase in the release of assemblies as drafts.

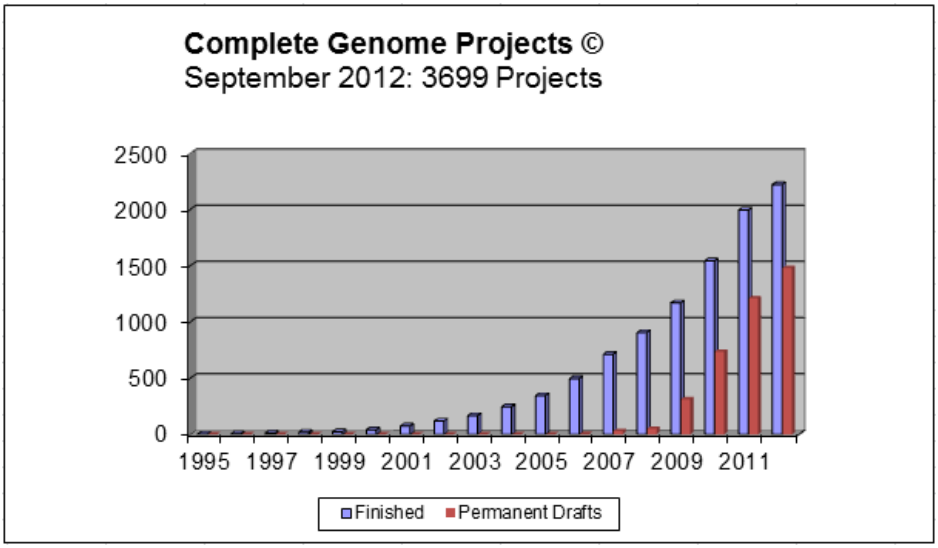


Figure 6: Assembly quality trend. Increasing release of permanent draft genome sequences in databases since the advent of next generation sequencing (around year 2007). Today, the number of permanent draft (4306, as of September 2013) largely outclass the number of finished genomes (2568, as of September 2013) - Numbers according to the Genome Online Database http://genomesonline.org).

In stark contrast, in the early years of genome sequencing, achieving a high quality reference level sequence was the end goal of every sequencing project with sequencing centres having whole teams dedicated to finishing (*e.g.* human, mouse and zebrafish genomes). The finishing phase of a genome project includes many physical mapping experiments, PCR reactions and re-sequencing leading to high costs. The democratisation of sequencing led to a decrease in quality of genome sequences released in database for several reasons. First, the decrease in sequencing cost was not accompanied by a decrease in the finishing phase costs and finishing thus represented a higher relative cost. Further, technological improvement in sequencing was not immediately followed by efforts in the assembly field rendering assemblies more fragmented as software were not tuned to the data produced. All this led researchers to publish draft genomes in databases rather than pursuing the aim of releasing high quality sequences. Worth mentioning, this trend is most likely imputable to the ever important pressure to publish quickly, with gene content being most of the time the main objective of sequencing. In addition, the cost of sequencing has recently been analysed (Sboner et al. 2011). The trends in cost fluctuations for the coming years have been estimated and described as higher than commonly thought. It is expected that the cost of the sequencing step itself will decrease, while the cost of downstream analyses will represent more and more of a sequencing project in the future.

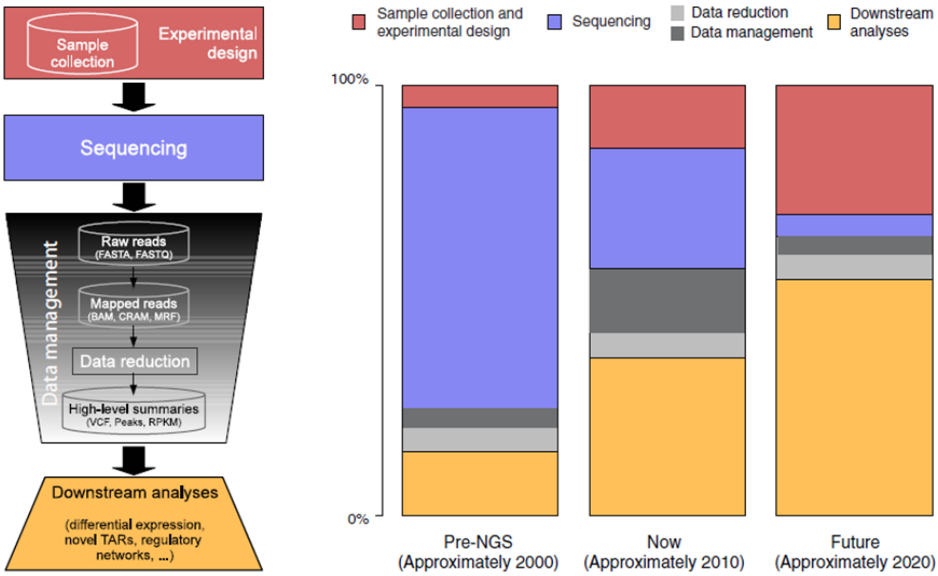


Figure 7: Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; CRAM, compression algorithm; MRF, Mapped Read Format; TAR, transcriptionally active region; VCF, Variant Call Format. Taken from (Sboner et al. 2011).

--The case for improving assembly quality and alternative technologies: practical considerations.

Important features to consider when discussing *de-novo* sequencing of whole genomes are the many practical considerations an investigator needs to take into account. The choice of a sequencing strategy, of one (or more) sequencing technology, of the assembly software to be used, the expertise available, the quality of the data required for subsequent experiments, the computational infrastructure and team (including bioinformaticians and bioanalysts) available for subsequent assembly and analyses are all parameters that will influence the quality of the final genome assembly. An investigator may be forced to make choices, and this often includes compromises with practical choices such as preferential prices by a specific platform proposing only some of the existing sequencing technologies. Even in the best of settings, the final result is usually rather error prone, with many missing sequences, gaps, unplaced reads and contigs, and even mis-oriented sequences. At the moment, the computational and technological complexities are too high for an "out-of-the-box", "off-the-shelf" easily usable solution for non specialists (Nagarajan & Pop 2013). Until such solutions exist, there is a need for assessing assembly quality and validating assemblies by means of other technologies than sequencing. When no reference or closely related genomes are available, assessment of what is missing or badly assembled is impossible, and more importantly, invisible to the eye of the biologist. Alkan and colleagues performed *de novo* sequencing of human genomes

using next generation sequencing technologies and found many differences with regard to the reference sequence assembly (Alkan et al. 2011). Their result highlights the need for complimentary efforts to experimentally validate genomes. Noteworthy, the human genome is certainly the genome having received the most attention, with dedicated finishing teams still trying to improve the quality of the sequence and resolve gaps and misassemblies (fig 8). Large "pseudo-molecule", high quality sequencing should not be abandoned until the balance between quantity and quality of genomes has been re-established in public databases. Furthermore, the availability of a closely related, high quality genome sequence can leverage many of the hurdles cited in the paragraphs above, justifying in the long term efforts to release as many high quality assemblies as possible in public databases.

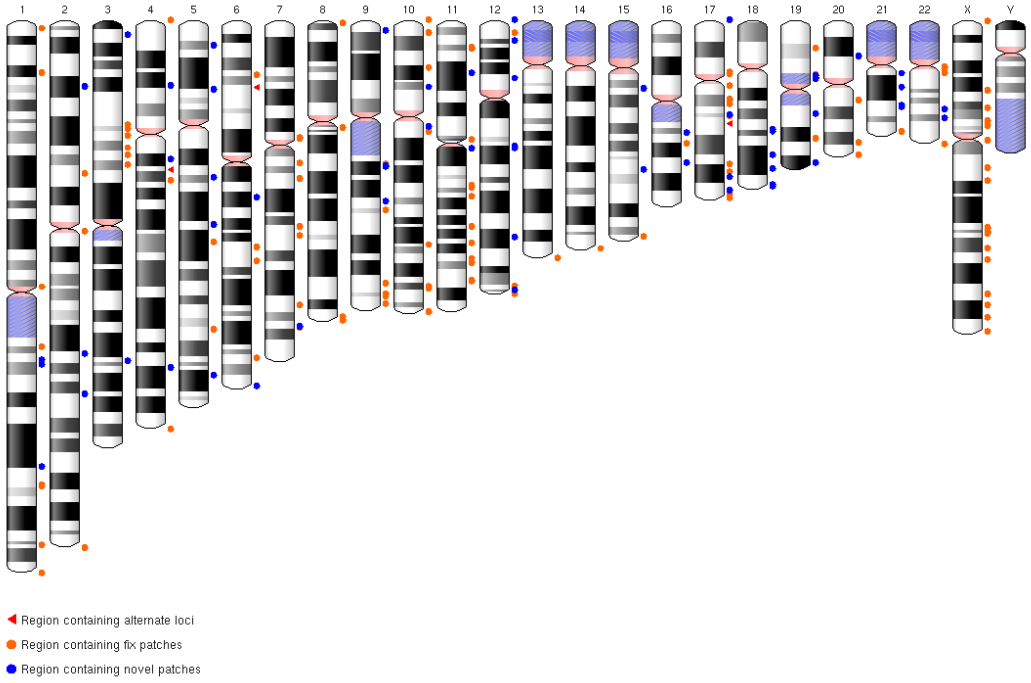


Figure 8: State of the human Genome Reference sequence as of 2013. Large regions of the genome are still not part of the assembly (blue and pink coloured regions) including the short arms of accrocentric chromosomes 13, 14, 15, 21 & 22. Centromeres are also not part of the assembly. Additionally a number of gaps, misassemblies and alternate loci are still present. From www.ncbi.nlm.nih.gov/projects/genome/.../human.

Worth mentioning, the "over-release" of draft genome assemblies in databases may lead us to unintentionally disregard important parts of prokaryotic and eukaryotic genomes (notwithstanding the fact that quite often, information regarding assembly quality is missing). This is especially true in the now-occurring era of massive high value sequencing projects like the insect 5000 genomes project (Robinson et al. 2011), the 1000 Fungal genomes project (http://1000.fungalgenomes.org), or the mammalian 10000 genomes project (Pennisi 2009). Assemblers are confused by repeats. Repeats promote rearrangements, and may be responsible for structural differences between strains and species. Repeats are present at a high proportion in genomic islands and horizontally transferred

regions (Ricker et al. 2012). These structural differences and presence or absence of genomic island and horizontally transferred region are quite often responsible for novel phenotypic traits and even adaptation to new environments (a detailed review of horizontal gene transfers is provided later in this introduction, see chapter four). With their high repetitive content, they are likely to be misassembled, represented by many very small contigs (which are quite often not used for comparative genomic studies) or even not assembled at all. Further, genomic context, whether in the case of chromatin condensation state (and thus transcription, and epigenetic regulation) or in the case of synteny, is a highly informative feature of genome sequences. Fragmented, draft assemblies do not allow access to this information. A good example of this is the "meso-synteny" mode of evolution of some groups of fungi. Meso-synteny is a mode of evolution where gene order among a chromosome is highly shuffled. Only micro-synteny and gene content on the chromosome remains conserved between species of the clade (Hane et al. 2011). For these reasons, many authors have advocated to pursue the goal of high quality sequences. Some have deemed draft assemblies limited for studying prokaryotic evolution (Ricker et al. 2012), while some others stated that without efforts to fully sequence complete genomes, the field of comparative genomics is likely to face a crisis (Alkan et al. 2011) and other advocates that every genome needs a good map (Lewin et al. 2009).

e. Assembly validation and physical mapping:

There are several ways in which an assembly can be improved. As described above, different assemblers generate different assemblies, and these can be compared for discrepancies between results. Additional rounds of sequencing might help bridge some gaps, by providing more coverage. As well, different sizes of libraries, or combining different sequencing technologies will also solve some gaps, correct errors and some mis-assemblies by providing different read length, solving assembly problems linked to error profiles of the sequencer (some technologies perform better in some situations, allow different type or insert sizes for mate pair generation and have different read length). However, despite these possibilities, some issues are greater than can be solved through sequencing alone. Further, these additional sequencing rounds are not always likely to be carried out because there are no easy ways to assess the extent of the improvement in sequence quality relative to the additional incurred cost. Noteworthy, for similar organisms, the results of a given sequencing strategy can vary widely (as exemplified by the difference in the number of scaffolds between *Penicillium roqueforti* (73) and *P. camemberti* (181) in the results parts, for instance. Additionally the same sequencing scheme on *Mucor* species, which have similar genome sizes, yielded thousands of scaffolds, because of a very high repeat content in these genomes. This was not expected prior to sequencing (Hermet A, personal communication).

Physical mapping has historically been linked to sequencing by being a key step in the hierarchical shotgun approach. BAC mapping has been performed to provide a means to order and orient BAC clone assembled sequences on genomes. This approach is still performed today. Several methodologies and technologies have been used towards this end, including BAC fingerprinting (Schein et al. 2004) and radiation hybrid mapping (Faraut et al. 2009), as well as FISH on metaphase chromosome (Szinay et al. 2008a; George et al. 2012). Physical maps can provide a different type of insight on the DNA sequence by providing a structural overview. Observing the genomic landscape on single DNA molecules (be they condensed chromosomes or single DNA fibres) provides an external and observed validation of the DNA molecule organisation. Optical maps are based on single molecule technologies. Some of these single molecule technologies have already proven their worth in improving or validating assemblies. At the present day, five main technologies exist to observe single DNA molecule and its organisation. They include:

-FISH on metaphase chromosome, with a resolution of 10 Mbp because of the condensed state of the chromosomes. Some specialised labelling and probes technologies can reduce the resolution down to 1 Mbp.

-Optical restriction mapping on flow stretched DNA, exemplified by the technique developed by David Schwartz (Schwartz & Samad 1997) and commercialised by OpGen Inc (http://www.opgen.com).

-Flow stretching of previously fluorescently labelled DNA, as exemplified and commercialized by BioNanoGenomics (http://www.bionanogenomics.com)

-Fibre-FISH, a FISH on decondensed DNA fibres (Szinay et al. 2008a)

-Molecular Combing, a DNA stretching technology initially invented by Aaron & David Bensimon (Allemand et al. 1997) at the Pasteur Institute, and further developed and commercialised by Genomic Vision. (http://www.genomicvision.com)

Molecular Combing has been suggested before as a powerful technology for improving assembly quality (Conti & Bensimon 2002), but has never been applied to a whole genome sequencing project. The molecular combing technology and its positioning within the single molecule toolkit is reviewed in the next chapter, which will be published later under a slightly revised form as a review.

Chapter Two: Molecular Combing or Bringing Together Single Molecule DNA Analysis & Multiple Genome Survey.

Molecular Combing represents an important aspect of this thesis. In this introductory chapter, we position combing with regards to other single DNA molecule technologies. Over the years, single molecule studies have become an important aspect of nucleic acid research. Nowadays, many technologies are at the disposition of the investigator. These technologies all rely on specific principles and have their own advantages and sets of limitations, rendering them more suited to one or another application. In this review, we briefly present some of the single molecule technologies available, with a specific emphasis on Molecular Combing and its applications. We also describe how other technologies have been applied to the assembly problem presented in the previous chapter. The aim of the chapter is to introduce the rationale of using Molecular Combing to the assembly problem.

1. Introduction*:*

This chapter aims at positioning Molecular Combing among single molecule technologies in nucleic acid research. Nucleic acid is one of the most important groups of biomolecules as it encodes all the information required at any time for the cell to function properly. Since the discovery of the structure of DNA by Watson and Crick, in 1953, DNA has been the object of many studies focusing on its properties and sequence. Most of these studies' results were based on an ensemble of molecules. Over the years, it has become clear that although studies of nucleic acid properties by observing a population or an ensemble of molecules unlock a large amount of knowledge, single molecule studies give different and complementary insights into the structure, behaviour and sequence of nucleic acids. Accurate and concise knowledge of nucleic acid simply cannot be gained through the sole observation of an average ensemble of molecule behaving in an average fashion. Nucleic acids are not only varying in sequence, structure and length, they also vary dynamically as a function of time. The highly regulated structural and conformational changes required for a proper functioning of the cell machinery are carefully regulated. It is obvious that these complex regulations cannot be achieved by average molecules with standard behaviours, and that some similar nucleic acid molecules must behave really differently. This cannot be highlighted using conventional bulk molecule studies, hence the need for another group of methodologies, allowing a glimpse into single molecule properties and behaviours (Efcavitch & Thompson 2010). The single molecule toolbox is now made of several technologies and among them is Molecular Combing.

One of the advantages of single molecule analysis is that by studying an "individual" molecule, one can observe "unique" or rare events as well as common events. The result is the reflection of one

single existing and witnessed event. Through these kind of studies, one gathers a different, more accurate kind of knowledge as it describes what really is. One other de facto advantage of some of the single DNA molecule technologies is that they enable the direct observation of the molecule.

2. Single Molecule Technologies: A brief overview

A few single molecule technologies are available, each with advantages and drawbacks making them more suited to some applications (see Table 1). Among them are optical and magnetic tweezers, allowing a glimpse into physical properties of the strands and DNA-protein interaction through twist and torque relations under differential strain and stress conditions (van Mameren et al. 2008), Single Molecule sequencing technologies (these have been addressed in the first chapter), such as Helicos' true Single Molecule Sequencing ™ (Harris et al. 2008) and Pacific Bioscience's Single Molecule Real Time ™ sequencing (Eid et al. 2009), and the soon-to-come nanopore technologies, Flow-based stretching of DNA molecules hinting at some partial sequence content information by restriction enzyme digestion (Latreille et al. 2007) as well as some physical properties by studying the behaviour of nucleic acid in electrophoretically controlled conditions giving insight into structure transition and folding of the molecule, FISH on interphase or metaphase-based technologies, enabling the study of single chromosomes and karyotypes, and surface tension-driven DNA stretching technologies, to which Molecular Combing belongs.

In the following section we give a brief account of the existing single molecule technologies and their respective applications and limitations.

a. Sequencing:

Now-generation sequencing technologies (Illumina Solexa, ABi SOLiD, roche 454, Sanger) are not single molecule technologies. However, the next (third) generation of sequencing technologies is shifting from amplification based to single molecule. An exhaustive review of their principles and promises is far beyond the scope of this review (for review see: Efcavitch & Thompson 2010). Single molecule sequencing can be split into two approaches: optical and non-optical methods. Optical methods rely on the detection of the emission of a fluorescence signal when a base is incorporated. An example of optical sequencing is embodied by Pacific Bioscience's approach, relying on the detection of a fluorescence signal in a zero-mode-waveguide (Eid et al. 2009). Non optical sequencing concepts are also under exploration. Whether they rely on the use of scanning tunnelling microscopy, or nanopores, they are based on the direct detection of the base sequence, or on the release of bases one after the other in a flow and through a nanopore where the base identity can be

deduced by the changing conductance in the pore. These methods both rely on the scanning of an intact DNA molecule one base at a time (Efcavitch & Thompson 2010; Branton et al. 2008; Tanaka & Kawai 2009). At the present day, no realistic assessments of the routine characteristics of these technologies are available, and their impact on the field of assembly is hard to assess.

b.Tweezers:

In a tweezers based experiment, a single DNA molecule with one extremity linked to a bead (or both extremities linked to two beads), is stretched. The stretching happens by moving the optically or magnetically trapped beads. Constant flow stretching can also be used in the case of a single molecule attached to only one bead. These types of methods allow for a precise control of the extension and the tension in the DNA, enabling an accurate measurement of the forces at play. These methods are particularly well suited to study DNA-Protein binding and protein movement along a molecule, and to probe at the effect of strain, stresses and torques on biochemical and biophysical processes intervening on the DNA (van Mameren et al. 2008). The other advantage of these studies is that they do not require any special surface treatment to specifically attach the DNA, making the unwanted blocking of binding proteins less likely.

Technology	Methodology	Single Molecule essence	Modification of physical / structural properties	Limitations	Resolution	Applications	References
Sequencing	sMRT™, tSMS™, nanopore, electronic microscopy, scanning tunnelling microscopy	Highly parallel	No	Sequence composition: homopolymer stretches, repetitive elements, base incorporation/reading errors. Difficulties in Sample preparation	Base pair analysis extended to whole genome.	Base sequence	Efcavitch 2010, Metzker 2010, Ansorge 2009
Tweezers	Optical, Magnetic.	One molecule is attached to a tweezer.	Yes – Strains, stresses and torques	Single assay, difficulty in reproducibility	Several hundreds of base pairs.	Protein-DNA interactions, structural modifications	Leuba 2009, Mameren 2008.
FISH on chromosomes	microscopy	chromosomes on a coverslip	No – Condensed DNA in Chromosome state	Condensed DNA	Megabase to Karyotype	Karyotyping, Low resolution physical mapping	Ohmido 2010.
Flow-based stretching	Electro/micro-fluidic & microscopy	DNA molecules stretched in nannochanels.	Yes – stretching.	Non uniform stretching due to inhomogeneous distribution of forces along the molecule, use of nannochannels, fluorescence ratio to define contour length of molecule	Tens to several hundreds of kilobases.	Optical restriction Mapping, Molecular responses to applied forces	Persson 2010, Latreille 2007.
Molecular Combing	Surface tension driven stretching & microscopy	DNA molecules stretched on a coverslip	Yes – Controlled Stretching	Stretched DNA is bound to the surface at multiple positions, High Molecular Weight genomic DNA.	One kilobase to whole genome.	Rearrangements, translocations, viral integration, replication, physical mapping	Lebofski 2003, Herrick 2009, Caburet 2005. Cheeseman 2012 Mahiet 2012 Levy-Sakin, 2013

Table 1: Single Molecule Toolkit

c. Fluorescent In-Situ Hybridisation:

In-situ hybridisation involves the binding of a specific nucleic acid probe to a substrate of genomic nature. The probe is labelled using a radioactive; chemical or fluorescent label to report the genomic location of the probe. Fluorescence *in-situ* hybridisation has had major contributions in the field of cytogenetics and genomics. A method of age (earliest reports on the use of FISH backdates to the mid 70's), FISH has been for a long time one of the only single molecule technology enabling a glimpse at the whole genome. FISH on chromosomes has resolved translocations, gross deletions, enabled physical mapping on chromosomes, at the megabase level, positioning of seed BACS prior to sequencing, etc. FISH on chromosome is a powerful tool to study condensed DNA, allowing the

discrimination between hetero- and euchromatin. More recently, FISH methodology has had and is continuing to have a tremendous impact on sequencing of crops and other plants of biotechnological interest (Szinay et al. 2008b; Ohmido et al. 2010). Fibre-FISH methods have also been developed. They involve extending a DNA fibre and performing hybridization of specifically designed probes. Fibre FISH enables FISH on uncondensed DNA, bringing down the resolution to the kilobase, albeit losing the whole genome advantage given by karyotypic studies, as the DNA fibres are broken and non-uniformly stretched. Perhaps one of the most striking and well-known example of Fibre FISH application is the correlation between the AMY1 copy number variant and different level of starch in diet (Perry et al. 2007).

d. Flow-based stretching:

The generation of well defined flow fields within micro- or nano-scale channels provides a means to stretch nucleic acid molecules. The flow usually relies on hydrodynamics or electric fields. These technologies are well suited for the study of DNA-protein interactions, as well as DNA behaviour under well defined constant forces, under several conditions defined by the buffers (buffers are easily changed, as fluid flow is responsible for the force generated.)

The Optical Mapping approach developed by D. Schwartz (Schwartz & Samad 1997), and commercialized by Opgen, involves stretching DNA by generating a flow with the buffer, and immobilising it as single molecule by electrostatic interactions on a charged substrate. A restriction enzyme digestion is then performed using a carefully selected nuclease. This allows for the generation of single molecule restriction maps, useful in the context of genome assembly and strain typing. The technology suffers from errors due to missing and false cuts, high variance of estimated fragment sizes and chimeric maps, resulting from artificially merged molecules. (Persson et al. 2009)

This technology has been helpful in the assembly of many bacterial genomes and widely used as a finishing tool (Latreille et al. 2007). Recent example includes the de novo assembly of the domestic goat genome (Dong et al. 2013) and the medicinal mushroom *Ganoderma lucidium* (Chen et al. 2002).

Another technology, Irys involves stretching of nick labelled DNA in nanochannels by the use of electric flows. This technology is being developed into an industrial application by BioNanoGenomics. It has recently been applied to the assembly of a 2.1Mbp genome of *Aegylops tauchii*, the suspected donor of the D genome of the hexaploid wheat (Hastie et al. 2013).

3. Molecular Combing:

Molecular Combing enables whole genome studies, over multiple genomes, on single DNA molecules. The Combing process results in homogeneously, uniformly stretched long DNA molecules on a single microscopy coverslip (fig 9). The first step of the process includes the obtention of high molecular weight DNA in solution by the gentle lysis of agarose plug embedded cells. A silanised microscopy coverslip is then dipped and incubated into the slightly acid DNA solution. DNA molecules in an acidic environment are in random coil conformation with hydrophobic domains of the bases exposed at the extremities because of the pH induced-denaturation (Allemand et al. 1997). This enables a strong interaction between the hydrophobic silanised surface and the DNA extremities. The coverslip is pulled out of the solution with a constant speed of 300µm.sec^{-1}. The nucleic acid molecule is stretched by the applied force of the hydrophobic receding meniscus. Because of the air-glass-solution triple interface configuration being the same at any given point of the microscopy coverslip, the surface tension force exerted by the meniscus on the DNA molecules is the same everywhere, and as a result, every surface-bound DNA molecule is stretched in the same manner. It is interesting to note that the degree of extension corresponds to a 65pN force, resulting in molecules stretched to 150 % of their initial contour length (Bensimon et al. 1995; Strick et al. 1996). The consequence of this is an array of long DNA molecules (with an average size between 500kb to 700kb when working with cell lines) stretched regardless of their sequence composition. Molecules are stretched with a physical distance to contour length correlation of 1µm equivalent to 2kb. The density of molecules is determined by the quantity of cells used during the preparation of the solution. Depending on the organism's genome size and initial cell concentration, a few hundred to several thousand genomes can be efficiently stretched on a single 22x22mm microscopy coverslip (fig9 A).

Hybridisation experiments, especially FISH experiments can then be applied to this array of stretched DNA molecules, thus enabling direct visualisation of the genomic organisation and content, at specific loci on single DNA molecule using epi-fluorescence microscopy. Probes length and distances separating them can be accurately determined by measures.

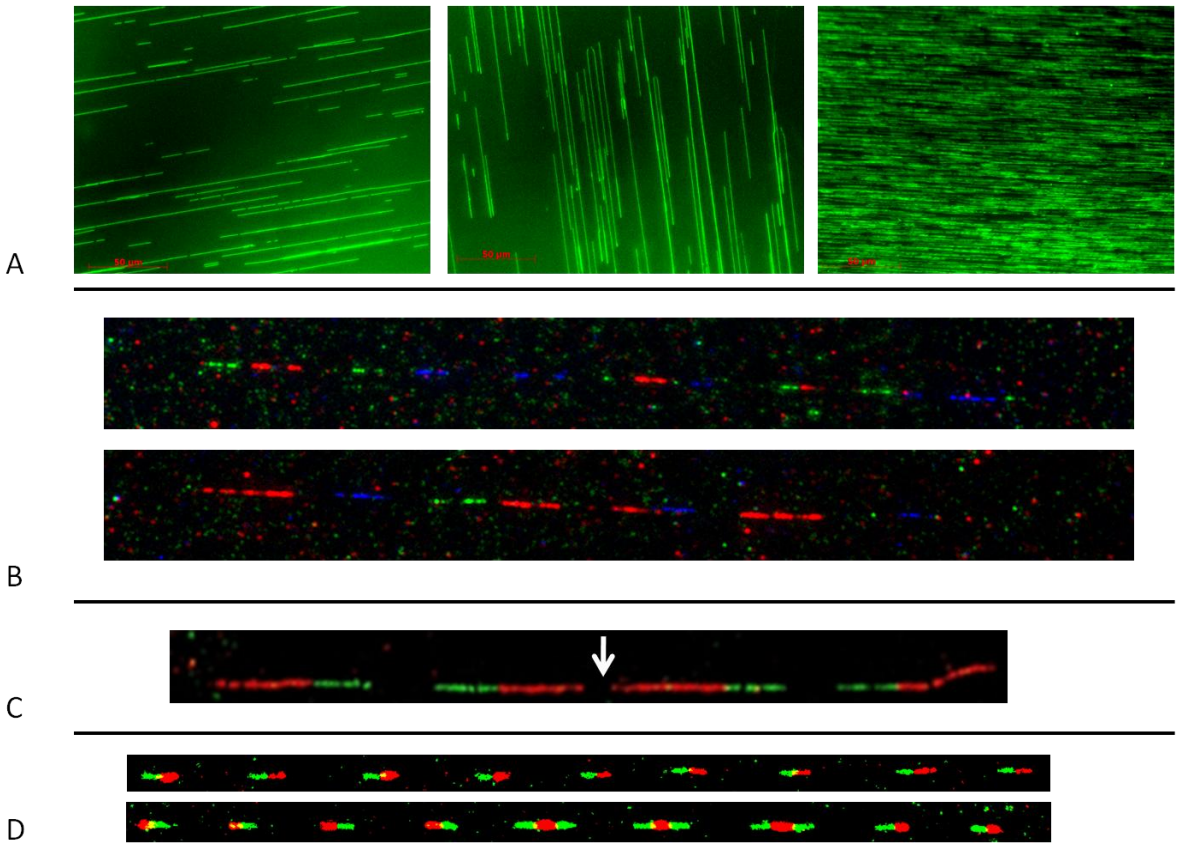


Figure 9: Molecular Combing

A) Combed DNA fibres counterstained with YOYO-1. Microscopy observation : 40x
B) Examples of two different Genomic Morse Codes on loci of interests. Microscopy observation : 40x
C) Replication Combing assay: Successive pulse chase of nucleotide analogues enables visualisation of DNA replication on single molecules, genome wide. Arrow indicates an origin firing.
D) Ribosomal DNA repeat organisation revealed using FISH on combed DNA. Two probes designed on the ribosomal DNA locus enable the orientation of the repeat.

Among the Single Molecule technologies available, Molecular Combing is a method of choice for studying many aspects of nucleic acids. Molecular Combing has been shown to have a broad range of applications, such as fast and accurate physical mapping (Conti & Bensimon 2002 and see below), viral integration (Mahiet et al. 2012), haplotyping, repeat organisation analysis (Fig 9D) (Caburet et al. 2005), DNA replication analysis (Fig 9C) (Palumbo et al. 2010; Técher et al. 2013) as well as the effects of chemical compounds of biotechnological or pharmaceutical interest on replication, and many other biological and non-biological applications , with a scope reaching research (Herrick & Bensimon 2009), diagnostics (Cheeseman et al. 2012, see annexes), and industry.

Advantages of Molecular combing over other conventional cytogenetic and molecular biology techniques are many. Molecular Combing's broad resolution range renders this technology useful to

study any event with a resolution from the kilobase by studying single fibres to a couple of Megabase, to whole genomes by studying many fibres. Not only is the resolution wide, but so are the kind of events studied. Changes in the sequence structure, like copy number variation and genomic rearrangements such as duplications and amplifications, deletions, inversions and translocations -whether balanced or not- are easily enlightened by this single molecule technology. One of the other advantages of Molecular Combing is that despite being a Single Molecule technology by nature, it also allows the study of many single nucleic acid molecules at the same time, as many genomes can be combed on a single microscopy coverslip, thus enabling an increase in throughput and multiple loci observation, as well as multiple genome survey.

a. Molecular Combing: a broad range of applications

A feature of any method, technique or technology can be an advantage in one context, while being a major hindrance in another context. The flexibility of Molecular Combing offers many context specific advantages over other methodologies making it a good alternative or complementary technology unlocking further information. In the following sections are briefly reviewed context specific advantages offered by Molecular Combing.

b. Detection of Large Genomic Rearrangements:

Molecular Combing enables the characterisation of genomic loci of interest, through FISH based physical mapping. By carefully designing probes, a Genomic Morse Code can be designed (Lebofsky et al. 2006) Fig 9B). A Genomic Morse Code consists of a set of differently labelled probes that enables the accurate cartography at the kilobase level of a locus of interest. The specific design of the Morse Code allows a clear and simple visualisation of any genomic rearrangement (balanced and unbalanced) whether it is a deletion, a duplication, an amplification of two or more copies, a copy number variant, a translocation, and even simply a discrimination between two alleles. Another powerful aspect of Molecular Combing is its ability to detect cellular mosaicism, which is defined by the presence of different alleles in different proportions among a sample. Because of all these features, Molecular Combing is a good alternative to current cytogenetic and molecular biology methods used today in research and diagnostics.

c. replication studies:

Molecular Combing has been shown to be specifically well suited for DNA replication studies. Its intrinsic features enable a direct observation of replication on single DNA molecules as well as the observation of replication over the whole genome. Using successive pulse-chases of labelling by iodo- or chloro-deoxyuridin compounds DNA replication patterns and kinetics are easily highlighted (fig 9C) (Lebofsky et al. 2006). Moreover, replication at a specific locus can be easily studied by adding FISH-probes on combed DNA. The replication programme can be extensively studied using different labelling strategies. Molecular Combing Replication Assay gives insight into genome wide replication parameters, such as spatial organisation of the origins one to another, amount of DNA replicated, replicon eye size with a resolution of 1 kb, inter-eye size, inter-origin distance, fork density and velocity, early and late firing origins, and density of fork per length of DNA over a specific portion of the genome (Caburet et al. 2005; Lebofsky & Bensimon 2003; Herrick & Bensimon 2009).

A direct application of DNA replication studies is the assessment of cellular perturbations induced by a chemical or biological compound of interest. Replications studies are also of interest in gaining insight in the loss of genomic stability often associated with tumorigenesis (genomic instability of tumour cells have been linked to DNA rearrangements and perturbation of the DNA replication programme).

d. Structural repeat organisation and copy number variation studies:

Repeats are a major hindrance of many cytogenetic and molecular biology technologies. They are hard to resolve using sequencing, as assembly or read mapping programs collapses them or misassemble specific loci by wrongly assigning one read to another, thus merging two separate loci. Microarray studies are limited in the resolution of duplications, as the result is underpinned by a fluorescence intensity ratio. Southern blot can be used to study variations in the number of repeats, but is not accurate and quite hard to set and limited for determining large number of repeats (Nguyen et al. 2011). Molecular Combing is a particularly powerful and simple technology to observe repeat organisation and copy number variations as it allows the clear visualisation of the molecule organisation at the loci of interest (Fig 9D) (Caburet et al. 2005). Molecular combing also allows for allele discrimination if there is any variation between alleles by using a carefully designed Genomic Morse Code.

f. Presence of viral DNA & viral DNA integration:

The presence of a virus within a cell can be detected through a wide arsenal of techniques belonging to immunology and molecular biology. The detection of viral DNA can be done by PCR, sequencing, microarrays, etc... Only Molecular Combing allows for a viral whole genome structure study, a feature particularly useful to discriminate between several malign and benign isomers or to study the presence of specific viral genes (Reisinger et al. 2006).

If the presence of viral DNA within a cell can be detected by a various array of techniques, viral DNA integration is often difficult to study, because of the small size of viruses related to the host genome which renders most technologies inapplicable. Targeted sequencing technologies can be used to assess viral DNA integration, but only to a certain extent as integration sites need to be well characterised. This is also true for many PCR based techniques. Microarrays and array-comparative genomic hybridisation cannot resolve viral DNA integration as even if they are able to identify the presence of viral DNA within an organism, they cannot discriminate the integration site. The broad resolution range of Molecular Combing, shown by the presence of hundreds of kilobases to a few megabase long DNA fibres, coupled to the high number of combed genomes by coverslip allows for an accurate DNA integration detection. Furthermore, the integration sites can be precisely mapped, even if numerous and widely distributed among the genome. Viral genome features can also be highlighted by the use of a carefully designed Genomic Morse Code. Recently, different viral structures and isomers generated during replication of the Herpes Simplex virus have been highlighted by molecular combing (Mahiet et al. 2012)

Concluding Remarks:

Single molecule analysis provides a different insight into nucleic acid research than conventional molecule studies. For more than half a century, nucleic acids have been studied at the population level. Bulk molecule studies have generated a tremendous amount of knowledge, and shaped biology as it is today. However, the development of the single molecule toolkit now allows for a complementary knowledge, enabling the life science community to further understand the complex role and functioning of one of the most important class of biomolecules. Single molecule analysis has moved from being a fascinating prodigy to a hallmark of nucleic acid research. Almost non-existent two decades ago, the single molecule toolkit is now wide, and will continue to develop further as detection methods, surface chemistry and surface manufacture and sample preparation improve.

Among the single molecule toolkit, Molecular Combing has proven to be a "all-track" technology, well suited to investigating many aspects of nucleic acid and genome research. In contrast with other single Molecule methods, Molecular Combing provides a unique window of analysis for probing into genome architecture and stability, through FISH on Combed DNA and the Replication Combing Assay. Molecular Combing is a method of choice to discover large genomic rearrangements, often implicated in genomic diseases as well as studying effects of molecules on the cell cycle.

The technology is most likely to improve in the near future as various works in the field of physics (development of new means to visualise DNA, improvement in microscopy and image analysis), biology, (labelling chemistry, sample preparation), computer science (analysis software and automation) and chemistry continue to be developed.

Aim of the project n°1: Exploration of the Molecular Combing technology as a means to improve and validate de novo genome assemblies:

In a sequencing project, Molecular Combing could be used as a complementary technology to sequencing. A sequencing project is made of a few steps, going from the sequencing reactions themselves, through a primary or draft assembly to a refined assembly sometimes reaching the reference level. Despite not being able to show base pair modifications, or to directly read sequences from combed fibres as a sequencing technology would, Molecular Combing has the potential to resolve short and long repeats, highlight bioinformatic contig mis-assemblies, and accurately locate and order scaffolds by the clear visualisation of the sequence structure on the DNA fibre. Molecular Combing's contribution to sequencing projects could possibly be at two stages: within the assembly and finishing steps by accurately positioning and orienting contigs and scaffolds on combed genomes, and within the selection of seed BACs by accurately positioning them on a genome enabling a non redundant sequencing by chromosome walking (This has already been done by positioning two BACs in a solanaceum genome project (Todesco 2008)), and a proof of principle positioning 8 BAC clones on 300 kb on the human genome (Conti & Bensimon 2002). Moreover, difficult-to-sequence loci could be resolved by molecular combing's powerful physical mapping potential. There are no cloning biases in a molecular combing experiment (whereas they are quite common in sequencing projects). Similarly GC-rich regions and homopolymer stretches are not a problem since mapping of loci or genomes are made by FISH and not by an enzyme driven approach. For these reasons, one of the objectives of the work reported herein is to conceive, develop and apply a methodology for improving a de novo whole genome sequence assembly. Such a proof of concept would constitute a first step towards producing a new way to improve genome assemblies, and could pave the way to subsequent benchmarking with other optical mapping technologies.

Chapter Three: Meet the Fungi

The following section gives a brief overview of the fungal kingdom, describing some aspects of their physiology and lifestyle, their ecology and how it impacts ecosystems, as well as describing some aspects of their impact on research and on our societies. The aim of the following sections is to provide and link enough material to highlight the importance of the fungal kingdom in natural ecosystems as well as in man-made environment. The first section will provide an overview of the general biology of fungi and their ecosystems. In a second section, their taxonomy will be introduced. An emphasis on their genomics and the associated landmark findings will be presented in the third section. The last sections will introduce their importance to Man and deal more in depth with the fungi studied in this thesis and its associated manmade environment, food and cheese.

Generalities. The fungal kingdom comprises various life forms striving in very diverse, different ecological niches. Although most of them are multicellular eukaryotes, some of them can be unicellular. Essential biomass decomposers, and playing key roles in biogeochemical cycles, they often are an essential component of the ecosystem they dwell in, the later most likely collapsing if fungal occupiers were removed. Fungi can be found in soil, marine water, rivers, cattle stomachs, plants... living as saprophytes, parasites, symbionts... Fungi, also known as Eumycota, include mushrooms, smuts, rusts, moulds, yeasts, and many other forms (21st century guide to the fungi, Moore et al. 2011). Although more than 99,000 species have been described to date, the vast majority of species have not been described or even encountered. Current estimates vary from 1.5 million of fungal species (Hawksworth 1991), to a more recent estimates of over 5.1 million species (Blackwell 2011). These estimates are mostly based on a fungal to plant species ratio. Organisms from the fungal kingdom can be living forms as small as a couple of micrometres, like the yeast *saccharomyces cerevisiae*, or span more than 2000 acres, as is believed to be the case for one specimen of *Armillaria ostoyae*, which is if not the biggest living organism on earth, among the biggest ones, rivalling with the biggest animals and plants. This individual spans a distance of 2,200 acres (roughly 9 km^2) with a lifetime of 2400 years to date, as well as a mass of several hundreds of tons (Smith et al. 1992). These considerable variations in forms and lifestyle are enabled by their nutrition mode.

Morphological structure. Most fungi lack flagella and have filamentous bodies with distinctive cell wall carbohydrates and haploid thalli as a result of zygotic meiosis. Most fungi are made of a thallus, an immobile structure made of filaments. These filaments, called hyphae are made of elongating

walled cells. They are interrupted by septa, but most of them possess a central pore, making the hyphae continuous and allowing the translocation of nutrients. Apart from yeasts and chitrids, for most fungi, the vegetative part of their organism is hidden around and/or inside the nutrient source. Hyphae organise in a network of filament, the mycelium. The cell wall surrounding the cytoplasm of fungi is generally made of chitin and glucans (Griffin 1996). Mycelium is a specialised structure that can be very long, as testified by the huge size of *Armillaria ostoyae* and other specimens. Mycelium can have modified forms, as for instance in some parasitic fungi, where a modified mycelium, the haustoria, penetrates the host's tissues. Perhaps the most convincing form of specialised mycelium is the modified mycelium of *Arthrobotrys* adapted to predation (fig 10). This mycelium bears several loops used to trap nematodes. Other structures formed by some mycelium include sticky and colourful knobs to catch and digest unwary insects.

1. Nutrition mode defines lifestyle

Nutrition. All Fungi are heterotrophic organisms feeding by absorption. Hydrolytic exoenzymes are secreted in the extracellular environment where they break down organic macromolecules into smaller compounds. These are in turn absorbed by the fungal cell, providing the nutrients required for cell survival and proliferation. Once a hyphae is in contact with a food supply source, internalised nutrients are translocated and relocated to other part of the colony where previous nutrient sources have been exhausted or where no food supply is available. Once nutrient is exhausted at a specific location, the hypha is extended and the mycelium network grows, reaching another nutrient rich environment. Conidia are also produced and dispersed to reach new nutrient rich environments.

The association between the immobility of fungi and their heterotrophic absorption of nutrients is well suited to the many different lifestyles of fungal species. Fungi can be found in soil, freshwater, marine water, in a wide combination of different temperatures, pH, mineral and hygrometry conditions. Some are parasitic, others are mutualists and some are free living... They derive their carbon, energy and electrons from a wide variety of organic sources. With regards to this, they fall in three main subgroups:

- The saprotrophs, which are the decomposers, and probably encompass the majority of fungi.
- The Necrotophs which invade and kill host tissues quickly. Usually relatively unspecialised pathogens, they attack the host's tissue surface when conditions are favourable. They are

responsible for diseases like foot rots, damping-off in seedlings, and leaf or stem blotch in many plants. Examples of these are *Fusarium*, *Septoria*, and others...

- The biotrophs: which are found on or inside their host, and do not kill their hosts. They are specialised pathogens or symbionts with a high host specificity and cause powdery mildew and rice blast among others (caused respectively by *Erysiphe graminis* and *Magnaporthe grisea*)

2. Biogeography and Ecology:

a. Where can Fungi be found?

While a large number of fungi (close to 100 000) have already been described (an exhaustive list can be found in the "dictionary of fungi"), the vast majority of fungal species remain unknown. Their potential impacts on society, as crop and human pathogens, biomass decomposers as well as their role in biogeochemical cycles or as biochemical compound factories and as an incredible reservoir for new drug discovery provides an important incentive to better characterise the vast biodiversity of fungi. Current estimates which range from 1.5 million species based on fungi to plant ratio (Hawksworth 1991) to between 3.1 and 5.1 million species (estimates based on soil community studies (O'Brien et al. 2005; Blackwell 2011) clearly highlight the hidden diversity to recover. This biodiversity lies in many different habitats. Because they can withstand extreme temperatures, water activity, and accommodate themselves of many different carbon sources, fungi can be found almost everywhere on earth (Raspor & Zupan 2006). The following sections highlight some of the habitats where fungi can be found, and some of their relationships to other organisms. Some have already been well studied and well sampled, whereas others are just being explored.

Even if tropical regions are usually considered to harbour the highest diversity for the majority of taxa (Hillebrand 2004), temperate and arctic regions are also known to harbour a high diversity of fungi, including important and well known species of drug producers (*P. chrysogenum*, etc...). Example of important secondary metabolites –compounds not necessary for their growth (in a lab), but responsible for other functions such as competition and extracellular signalling, see below) from temperate species includes the penicillin from a *Penicillium notatum* contaminant colony, discovered in 1929 by Alexander Fleming. The immune suppressant drug cyclosporine is synthesized by

Tolypocladium inflatum, from Norwegian soil. *Aspergillus terreus* also produces statins, used to treat high cholesterol levels.

From a drug discovery perspective, "original" or surprising habitats should not be overlooked, since they represent different ecosystems, and it is likely organisms striving in these particular conditions harbour novel compounds, which could have interesting applications. Even if not accounting for as vast a diversity as more tropical environments, they are host to some fungi with unique properties. Some yeast can be active under freezing conditions and lichens are frequently encountered within the Arctic and Antarctic polar circles.

b. Fungal associations with plant material.

Fungi are important symbionts and have been found associated with all groups of organisms. Perhaps the most well known association of fungal species to other organisms is their frequent association with plants. The relationships of fungi and plants can be divided into three major types of association: mycorrhizal, endophytic and plant pathogenic.

- Mycorrhizal associations are frequent, and often essential to the plant hosts, for which fungi take up water, nitrogen, phosphorus, and other nutrients from the soil, and transfer them to the roots, where the host plant can exploit them. While being essential to plants, this type of association is sometimes also obligate for some fungi. Plants implicated in these relationships with fungi can be flowering plants, conifers, bryophytes, ferns... (Pressel et al. 2010) Arbuscular mycorrhizal associations are found with at least 80% of plant species, and occur in 92% of the plant families. In addition, 600 ectomycorrhizal associations are known, most them being with Basidiomycotas.

- Fungi can also be found inside plant leaves and stems, living as endophytes. Almost all plants on earth are infected with endophytes. Endophytes are taxonomically diverse symbiotic fungi living between cells in the aerial parts of the plant host. They are not associated with major disease symptoms. Frequently the host also benefit from this association. Fungal endophytes of grasses for instance, form mutualistic symbiotic associations with temperate climate grasses and confer bio-protective benefits by producing fungal secondary metabolites and modifying the host metabolism. As a result, endophyte infection is responsible for enhanced host persistence by protecting from biotic and abiotic stresses (Tanaka et al. 2012). Dispersal can be mediated by insects. Some endophytes are known to protect the grasses from animal herbivory or insect feeding by secretion of secondary metabolites (Arnold et al. 2003).

- Another type of fungi-plant association is plant pathogenic. The difference with the endophytes is the triggering of pathogenicity in the plant host, rather than providing symbiotic or mutualistic benefits. Crop pathogens are responsible for economic losses worldwide. Many other pathogenic interactions with wild plants have numerous impacts on natural ecosystems. Indeed, the impact on society can be judged by the importance of the watch for and vigilance to avoid pest invasions and killing of naive native plant populations (Dutch Elm disease, Dogwood anthracnose, Redbay wilt, *Fusarium*-linked losses on crops, see the Importance of Fungi to Man section later in this chapter, (Zhang & Blackwell 2002; Harrington et al. 2008; Moore et al. 2011)). In light of globalisation, the introduction of new species of pathogenic fungi in new habitats can lead to considerable damage in plant populations, while damage in natural habitat can go almost unnoticed.

Further enhancing the importance of this group of fungi is the tendency of host shift within plant populations, leading to emerging diseases in plants. The concept of host specificity is also hard to assess.

c. Lichenicolous fungi.

Fungi can also be found associated to green algae and cyanobacteria, as lichens. Lichen is a mutualist symbiotic association between a mycobiont (fungus) and phycobiont or phytobiont (green algae or cyanobacteria) where the fungus surrounds the phyco/phytobiont cells, enclosing it in a complex fungal tissue. In some lichens the fungus forms haustauria in algal cells. The photosynthetic algal or cyanobacterial cell reduces atmospheric carbon dioxide into organic carbon sugars to feed both the mycobiont and the phytobiont, while the fungus protects and extends the collection surface of water and mineral nutrients for both organisms. The mycobiont also extracts nutrient from the lichen substrate. When the lichen is formed in conjunction with cyanobacteria, the latter is also able to provide nitrogen to the lichen. Lichens are found in extreme environments such as arctic tundra, deserts, rocky coasts, slag heaps, but also in less hostile niches, on wood substrates or on walls and rocks in temperate climates. The ability to colonize extreme environments is due to the poikilohydric nature of lichens. This association is considered by many as a highly successful nutritional strategy for both organisms (Honegger 2007; Wedin et al. 2004). Man has been using them as food and dyes, as well as producers of secondary metabolites with notable uses in sunburn protection or herbivory reduction. A serine protease of certain lichens has also been shown to degrade prions, leading to potential biomedical applications (Johnson et al. 2011).

Lichen represents 20% of all fungi and 40% of all Ascomycota (according to Answorth & Bisby's Dictionnary of the Fungi, 10th Edition, (Paul M. Kirk, J. A. Stalpers, David W. Minter 2011)). Most

lichens have a mycobiont belonging to the Ascomycota phylum, but some are Basidiomycota, and rare cases of Zygomycota have also been described. The fungus has a dominant role in forming lichen, as emphasized by the similar thalli and secondary metabolites secreted between different lichens having the same mycobiont and different phycobionts (Gauslaa 1997).

Other associations between fungi and other organisms have been described. An often omitted part of fungal diversity is the fungi associated with insects, arthropods and other invertebrate animals. These fungi are poorly known and often unculturable. They can be saprobes or necrotrophic parasites. Other fungi infect vertebrate animals too.

d. Fungal association with animals

Insects and invertebrates. Fungi can grow in or on both vertebrates and invertebrates. Fungi infecting insects and nematodes are sometimes believed to contribute to their population control. Among the important fungal interaction with animals are enthomopathogens. Mostly Ascomycota, Zygomycota and Chritidiomycota, they include famous examples like the Zombie ant fungus from brazil which upon infection changes the behaviour of the insect, making it grow up plants and biting plant tissue until it dies in a "death grip" (Evans et al. 2011; Hughes et al. 2011). Some members of the *Ophiocordiceps* phylum infect and consume caterpillars. *Ophiocordyceps sinensis* infects the body of caterpillars and eventually kill the caterpillar through growth. A compact mass of mycelium supporting fruiting bodies, the "stroma" then grows out of the head of the caterpillar. This has been used in traditional asian medicines for centuries to treat several conditions, including cardiac and kidney conditions. Cordycepin is also an example of antitumor compound produced by these fungi.

Some fungal species are so effective in killing insects that they are used as biological pest control. *Beauveria bassiana* for instance is used against termites and whiteflies, and is being assessed as potential pest control to counter malaria transmitting mosquitoes (Thomas & Read 2007).

Fungi can predate on nematodes too, as is the case of *Arthrobotrys oligospora* which entraps nematodes by using loops on its specialised mycelium. The nematode is immobilised, its body is invaded and digested (fig 10). *Pleurotus Ostreatus*, an edible fungus, is known to secrete toxic droplets to kill nematodes.

Evolutionary patterns of host shifting among plants, insects and fungi have been shown, highlighting the importance of further characterising fungi.

Vertebrates. Two famous and recent examples of fungi infecting vertebrates are responsible for decline of species. *Batrachochytrium dendrobatidis* is a chytrid responsible for the widespread

decline of frog populations worldwide. The fungus does not invade the frog's body but rather causes a skin disease as well as an imbalance in electrolytes leading to death through cardiac arrest (Voyles et al. 2009). North American bats are subject to the white nose disease (Blehert et al. 2009), when *Geomyces destructans* colonise their muzzle skin, ears and wing membranes.

Importantly, fungi from different taxa are responsible for many diseases in humans. Most of them are benign diseases caused by dermatophytes, such as some mycosis and ringworm disease. Immuno-compromised patients and babies are frequently infected by opportunistic fungal pathogens. Candida species are the perfect example. Aspergillosis is a severe lung condition caused by *Aspergillus fumigatus* and *Penicillium marneffei*. It is the leading cause of death in AIDS infected patients.

Not necessarily centred on a pathogenic or symbiotic association, soil, marine and freshwater environments are habitats where a large part of fungal biodiversity can be found.

Soil fungi. Soil is a habitat of high fungal diversity. Fungi in soil communities take part with bacteria in biogeochemical cycles (Vandenkoornhuyse et al. 2002). They are the most important biomass decomposers of all ecosystems (Moore et al. 2011). Mostly Ascomyceta and Zygomyceta, diversity is thought to vary locally, and is usually higher near organic material such as roots and root exudates. Estimates of 3150 known species have been proposed but the availability of metagenomic methods are increasing the rate of species acquisition. Some studies have proposed a dominance of fungi over bacteria in soil communities, as well as little overlap in composition from samples taken only a couple of metres away (Taylor & Ellison 2010). Fungi can also play stabilising roles in desert soils by forming crusts made of lichens or Ascomyceta.

Freshwater fungi. Over 3'000 species of Ascomyceta living in freshwater are known. Freshwater fungi can display adapted structures for life in freshwater including specialised spore dispersal structures. These ascomycetes display evanescent asci, ascopores with appendages and sticky spore sheaths for efficient fixation to substrate in the aquatic environment. Different dispersal strategies are exemplified by ingoldian or aero-aquatic spores. Ingoldian spores have a morphology allowing them to bind to plants and other decaying material in the water. They float on foam accumulated at the surface and are dispersed through the air when the bubble bursts. Aero aquatic spores have multicellular helical conidia with air inside so they can float on the surface of slow moving waters. Other fungi than Acomycota are present in freshwater, and this includes chytrids, a few Basidiomyceta, Bastocladiomycota and Monoblepharomycota (James et al. 2006).

Marine Fungi. They belong to a broad array of taxonomical group, and are often different from the freshwater species although again, most members of the marine fungi are Ascomycota and Basidiomycota (Nagahama et al. 2006). Some yeast degrade hydrocarbon compounds from seeps or spill, while others live on calcareous substrates like mollusc shells, cnidarians reefs, corals. Many antibiotic producers live in a symbiotic association with sponges. Some parasites of fish, porpoise and other vertebrates have also been reported.

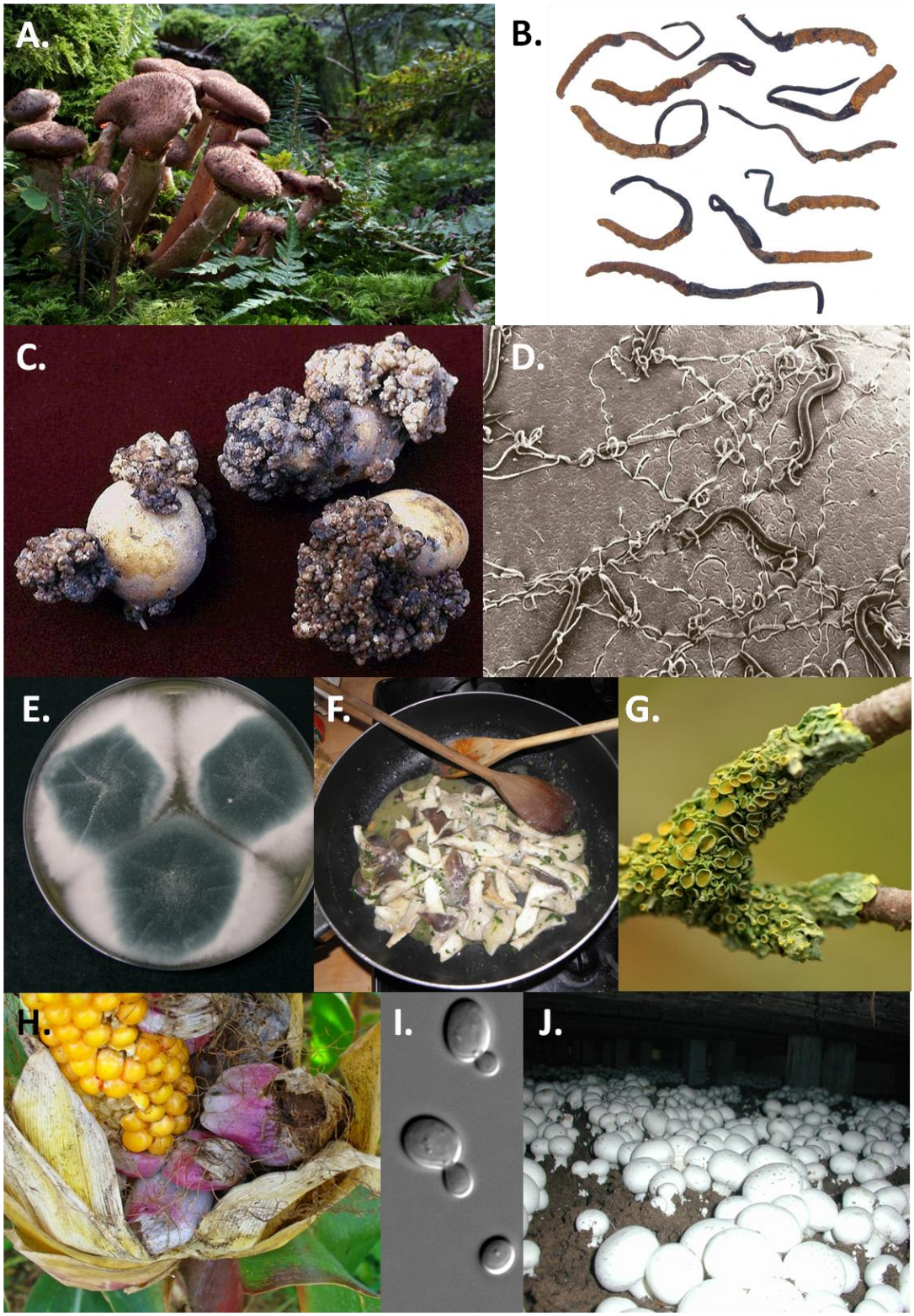


Fig 10: Meet the Fungi.. A. *Armillaria Ostoyae*, a parasitic basidiomycete in a southern Germany forest. B. *Ophiocordiceps sinensis*, a caterpillar parasite used in traditional Chinese medicine. C. *Synchytridium endobioticum* causing the black wart potato disease. D. Scaning electron microscopy of *Arthbotrys'* mycelium, with several nematodes entrapped. E. *Aspergillus fumigatus* on a petri dish. F. *Pleurotus ostreatus* are edible basidiomycotas. G. *Xanthoria parietina,* a lichen growing on a tree. H. *Ustilago Maidys* parasiting a Maize crop. I. The well known fermenter and laboratory workhorse *Saccharomyces cerevisiae*. J. *Agaricus bisporus* is intensively cultured as an edible mushroom.

3. Taxonomy

Generalities. The fungal kingdom is a kingdom where taxonomic relationships are not well defined and constantly evolving as new species are discovered. The concept of Fungi as a kingdom was first introduced by Jahn & Jahn in 1949, and further advocated by Whittaker in 1959. At first believed to be more closely related to plant than animals, this kingdom of heterotrophic organism is a sister group of animals, with which some characters are shared, like chitinous structures, glycogen storage, and mitochondrial tryptophan codon. In the following sections we will briefly discuss the major taxa of the fungal kingdom. The fungi are part of the Opisthokonts, based on the presence of a posterior flagellum, a character believed to have been lost several times in Fungi, and now found almost exclusively within the Chytridiomycota. The loss of the flagellum occurred at least four times, perhaps six (James et al. 2006). Previous classifications were based on morphological characters, and are being updated based on more accurate molecular phylogenies. For these reasons, as well as because of the increasing number of species being described, the taxonomic landscape of fungi is ever changing. At the present time, the fungal kingdom is made of a highly speciose subkingdom Dikarya, encompassing Ascomycota and Basidiomycota, and of several other phyla, some not monophyletic, such as the Chytridiomycota and the Zygomycota (see fig 11). Recently, Hibbett and colleagues published a revised and comprehensive taxonomy of the fungal kingdom, which is widely used these days (see fig). In this classification Hibbett and colleagues do not recognise the Zygomycota as a phylum, and Chytridiomycota are reduced to what James formerly described as “euchitrids” (Hibbett et al. 2007; James et al. 2006). Of note, *Rozella* species are believed to be the earliest diverging branch of fungi, and the cenancestor of fungi is most likely a nuclearia like organism (Steenkamp et al. 2006). The fungal kingdom is made of a subkingdom and seven phyla (Hibbett et al. 2007). Zygomycota are a polyphyletic group of fungi distributed in the phylum Glomeromycota, and four subphyla “incertae sedis”, or of unknown position (Mucoromycotina, Kickxellomycotina, Zoopagomycotina, Enthomophthoromycotina).

The following sections will describe the major groups of fungi without regards of phylogenetic relationships. Phylogenetic relationships are not within the scope of this thesis. The aim is to provide the reader with major trends in fungal research. However fungal phylogenetic relationships along with groups that were previously classified as Chytridiomycota or Zygomycota are presented in figure 11 (adapted from Hibbett, 2007 and Blackwell, 2011).

Chytridiomycota. Chytridiomycota, or chytrids are aquatic and terrestrial fungi. Approximately 900 species have been described, and are characterised by zoospores with a posterior flagellum. Mostly

saprotophs, some of these fungi represent important pathogens, as testified by the potato black wart disease caused by *Synchytrium endobioticum*, which has an important impact on the loss of potatoes. *Batrochochytrium dendrobatidis*, previously mentioned, is responsible for the worldwide decline of batracian populations (Wake & Vredenburg 2008).

Zygomycota. They are a polyphyletic group of fungi encompassing notably Mortierellales and Mucorales. Characterised by a caenocytic mycelium, zygospores and sporangiospores, they are mostly saprotrophs, decaying organic matters. They are also responsible for decay diseases of fruits and vegetables, and zygomycosis in immunocompromised patients. Some of them are parasites of insects and animals, or live in a mutualistic association in the guts of insects.

Glomeromycota. This monophyletic phylum was previously part of the Zygomycota (Schüβler et al. 2001). They represent an ancient group of arbuscular mycorrhizal fungi. Traces of Glomeromycota have been found to be as old as 400 million years in the fossil record. The relatively little amount of species known to date (169 - (Blackwell 2011)) live in obligate and mutualistic association with 80% of the vascular plant roots. The arbuscules formed in the plant roots have a role in carbohydrate exchange, from the plant to the fungus and in nitrogen, phosphorous and other mineral into the plant. It has been estimated that the fungus can absorb up to 40% of the photosynthetic metabolite produced by the plant.

Basidiomycota. Together with the Ascomycota, they form the subkingdom dikarya, based on the presence of a dikaryotic stage during the life cycle. With 31'515 species described as of 2011, Basidiomycota is the second largest phylum of Fungi. Basidiomycota are subdivised in three subphyla: Ustilaginomycotina, Pucciniomycotina, Agaricomycotina. While the first two are mostly plant parasitic pathogens, including smuts and rusts; the Agaricomycotina are a group of morphologically diverse fungi with basidia in various forms of fruiting bodies. These can be mushrooms, puffballs, shelf fungi, stinkhorns, jelly fungi, bird's nest fungi, etc. Many are saprotrophic and several representative of the basidiomycota are part of the edible fungi. Examples are *Agaricus bisporus*, the common mushroom, chanterelles, porcini... A few Agaromycotina are parasites, among them are *Amillaria* and *Rhizoctonia* species.

Ascomycota. Ascomyceta are part of the largest phylum of the fungal kingdom. With over 60'000 species described as of 2011, distributed among 3 major clades. Ascomycota represent a vast phylum with various impacts on ecosystems and man. Some Ascomycota alternate a sexual or meiotic reproduction characterised by the production of ascospores inside sac-like structures, called "ascus" (which gave the name Ascomycota) and an asexual reproduction through a mitotic process producing conidia. Conidia from different species have a wide variety of shapes, colours, septation. Many

species do not have described sexual state, and were until recently classified in an artificial phylum, the Deuteromyceta, or *fungi imperfecti*. However, with the advent of molecular phylogenies, this nomenclature is bound to disappear, as fungi can now be accurately positioned along the fungal tree of life among the species for which sexual states are known (Hibbett & Taylor 2013). In species for which both a sexual and an asexual phase are known, ascospores and conidia are usually formed, sometimes at different times of the year. In nature, vasexual phases seem to be more frequently encountered.

Ascomycota are further distributed in three subphyla: the Taphrinomycotina, the Saccharomycotina and the Pezizomycotina. The Taphrinomycotina are not producing any fruiting bodies. Saccharomycotina represent ascomyctous yeasts. Yeasts have naked asci, and reproduce by budding or fission. Mostly saprotrophic, some are pathogens, including of *Homo sapiens*, like the *candida* species. This phylum is important for industry as yeasts are used to produce a wide variety of compounds, and are used both as cell factories and as food fermenters. The last subgroup, Pezizomicotina is the largest subphylum of ascomycota, with over thirty two thousand species described. All lifestyles are encountered in this subphylum, which colonise a wide range of niches. They also bear different types of asci and fruiting bodies, the description of which is far beyond the scope of this introductory chapter. Of note, 40% of them are lichens.

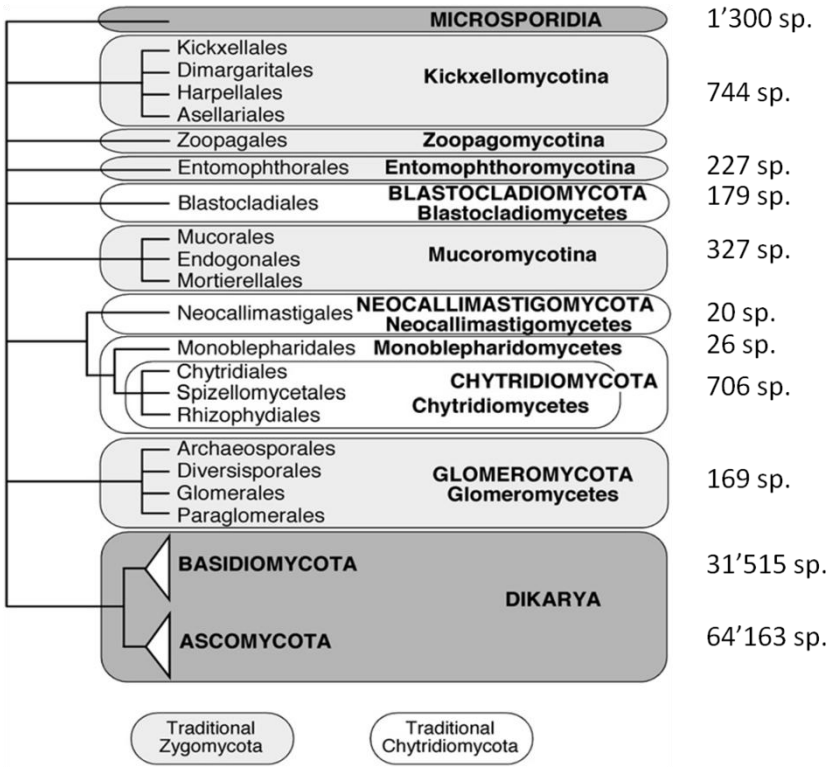


Fig 11: Phylogeny and classification of Fungi. Early diverging basal Fungi and subkingdom Dikarya. Branch lengths are not proportional to genetic distances. Approximate numbers of species in each group are derived from Kirk, 2008. Figure adapted from Hibbet, 2007, and Blackwell 2011.

4. On the Importance of Fungi to Man:

For centuries, the fungal kingdom has had tremendous positive and negative impacts on Man, ecosystems and their associated services. Despite some examples having been highlighted in the sections above, the following section further discusses this importance of Fungi to Man through three main aspects:

-Fungi are important pathogens of plants and animals, and there is a recent trend in emerging infectious diseases caused by Fungi which poses a threat to plant, animals and man alike,

-the important use of fungi as cell factories and producers of many useful organic compounds for the pharmaceutical, biotechnological and food industries,

-and finally their important role in food product manufactures.

Because the work carried out in this thesis is centred on *Penicillium* species from the cheese environment, a brief description of the cheese environment and its associated fungi is also given in this part.

a. Emerging fungal diseases in animal and plants lead to threats to ecosystem health.

The fungal kingdom comprises many pathogens of plants and animals. Despite relatively few fungi causing diseases in Man and these mainly being benign affections, fungi have a very important effect on our lifestyle by affecting our ecosystems in ways that have a great economical and environmental impact.

Plant pathogenic fungi are responsible for considerable agricultural losses. Fungal and fungal like diseases have had dramatic historical impact on man and if, in modern days fungal pathogenicity to crop do not include casualties and modification of civilisations, they still have a tremendous economical impact. One of the most striking historical examples of the effect of a fungal-like disease is perhaps the disease that led to the Irish potato famine. Despite being caused by an Oomycete (*Phytophtora infestans*), and not a true fungus, the disease is considered as fungal-like, and shows the same figures as fungal disease. In the mid XIXth century, the Irish population was heavily relying on potato culture to sustain its already starving population. When an Oomycete pathogen*, Phytophtora infestans* caused late blight disease all across Ireland, famine spread across the island, led to economic ruin, caused the deaths of one in eight of the Irish population, and drove many Irish to emigrate. It also led to the downfall of the English government. Almost 170 years later, while fungal and fungal like diseases do not cause this type of calamities anymore, they still have a very important economical and societal impact and today the effect of fungal pathogenicity in crops can be estimated through statistical economic losses. However, it is worth mentioning that behind these statistics are most likely personal tragedies for individual farmers and their families.

Fungal pathogens affect all the major crops used in modern agriculture. Wheat is infected by *Puccinia striiformis* and *P. graminis, Mycosphaerella graminicolla, Stagonospora nodorum*, *some Fusarium* species... Barley is infected by *Septoria passerinii*, and *Pyrenophora teres* as well as *Fusarium graminearum* among others. Mayze suffers losses to *Ustilago maydis* and *Sporisorium rilanum*. Examples are too many to be exhaustively cited or described here. It is estimated that today rice blast caused by the filamentous fungus *Magnaportha oryzae* destroys enough rice to feed 60 million people a year (Talbot 2007). In 1997, the effect of *Aspergillus* on the U.S. economy was estimated to total up to 45 billion $ (May & Adams 1997). Similarly, estimates from 2009-2010 data indicates for a lower estimate (taking into account only low level persistent diseases) a loss representing enough

food to feed 8.5% of the 7 billion humans alive. For a more severe scenario (taking into account severe epidemics in all five crops simultaneously), food for only 39% of the population would remain (Fisher et al. 2012). Indeed these are only estimates and are both unlikely to occur, with the yearly global fixtures fluctuating somewhere in the middle. Nevertheless, they clearly highlight the societal and economical importance of fungal diseases to agriculture and food security.

While plant pathogenic species have long been recognised as frequently occurring, fungal pathogens of animals and man are not traditionally recognised as a major threat. However several examples of species die-offs have recently been described, and there is a growing trend in the emergence of new pathogens. As described in the previous sections*, B. dendrobatydis* is responsible for the decline of amphibian population worldwide since the 1990s, and *G. destructans* raises concerns over the survival of bat species and population on the North American continent. Other examples are known, including terrestrial *Aspergillus sydowii* responsible for the decline of coral populations in the sea. This common terrestrial fungus is colonizing the sea environment through freshwater drainage and causes sea-fan aspergillosis. *Nosema* species causes colony disorder collapse and population decline in bee hives. The impact of fungal pathogen on man is mainly restricted to benign mycoses caused by opportunistic pathogens such as *Microsporum, Trychophyton, and Epidermophyton* strains. However, there is an increase in fungal pathogenicity associated to immunocompromised patients, and these fungal diseases can be lethal (examples are invasive aspergillosis, and fungal disease by *Penicillium marneffei* in AIDS infected patients).

If plant pathogens have been known for centuries, and their economical impact is easy to conceive, some other aspects of pathogenicity associated to fungi are harder to evaluate or perceive. Recent analyses of several reports highlight an increasing trend in the number of virulent infectious fungal diseases in the last twenty years or so (Fisher et al. 2012). It is believed that this unprecedented number of fungal and fungal like diseases have recently caused some of the most severe die-offs & extinctions ever witnessed in wild species and that they are jeopardizing food security. The data presented in these reports do indeed support the idea that fungi could pose a greater threat to plant and animal biodiversity relative to other taxonomic classes of pathogens and hosts, and that this trend is increasing (Fisher et al., 2012). Fungal lifestyle once again is probably responsible for the dynamics in fungal diseases leading to some extirpations (regional extinction of a species) or complete extinctions. Fungi can display a high virulence due to their high reproductive potential, long lived environmental stages, as well as the decoupling of some of their life stages from that of their host, enabling persistence in the environment and possible infections of other secondary hosts. But the most important factors driving the emergence of fungal infectious disease is Man associated activity. Transportation of individual, and transport and trade of infected goods leads to a

globalisation in fungi, favouring dispersal and encounter of species. As a consequence, non-pathogenic or mild pathogen species in their natural habitat can encounter naive susceptible hosts, leading to the emergence of disease. Furthermore, human activity also interacts with key fungal traits such as habitat flexibility, environmental persistence and multiple reproductive modes to cause the emergence of new diseases.

b. Fungi and Biotechnology industry:

Fungi are large scale producers of citric acid, industrial enzymes, amylases, proteases, lipases...Their ability to produce many primary and secondary metabolite (primary metabolites are fatty and organic acids, such as citric acid for instance) see section on secondary metabolites below). This type of biotechnology is often called whole organism biotechnology. It uses intact living organisms to produce commercially important products. Alcohol and citric acid are the world's most important fungal metabolites in terms of production volume. However, although most antibiotics used in medicine today originates from bacteria, fungi and the fungal biodiversity represents a largely untapped reservoir of interesting secondary metabolites with industrial or medicinal applications. The most widely known example of secondary metabolite produced by fungi is certainly the antibiotic penicillin, but other compounds with interesting properties are also of fungal origin. Cyclosporine is a secondary metabolite able to suppress the immune response in transplant patients to avoid organ rejection for instance. Three of the top ten selling pharmaceutical compounds in the late 1990s are of fungal origin. Mevinolin produced by *Aspergillus terreus*, is transformed into mevinolin, pravastatin and simvastatin, and these are used to reduce cholesterol levels, which is an important factor of stroke, cardiovascular diseases and several other illnesses. Ergot alkaloids, which have uses as vasodilatation drugs, steroid derivatives and antitumour compounds can be synthesised, but strain improvement and subsequent selection of high yield strains proved so efficient that fermentation is still the most cost-effective means of production of these important molecules.

Another aspect of fungal biotechnology is the bioconversion of chemical compounds by fungi. In this scheme, the fungi is not the original producer of the compound, but the enzymes produced by the fungi enable specific chemical transformation of the original compound, transforming it into a valuable chemical. Quite often, chemical transformation by other means would be ineffective, impossible, or simply too costly.

The main fungi used in industry are *Aspergillus niger* and *oryzae*, for the production of amylases and citric enzymes; *Trichoderma reesei*, useful in the production of cellulose, *Penicillium chrysogenum*, the main producer of penicillins, A*cremonium chrysogenum*, used in the production of cephalosporin and *Mortierella* species in the production of arachidonic acid.

Fungal species also have a tremendous role in the food industry, by being edible themselves, or by being instrumental in food production processes.

c. Fungi and the food industry:

The fermenting ability of fungi has made them an essential component of food production in every civilisation. Fermented beverages and products are a way to conserve food sources that would otherwise be wasted within weeks or even days. Various fungi are involved in many food fermenting processes. They include alcoholic beverages such as beer, wine and sake, breads and other bakery products and cheeses and fermented, cured meats.

Some yeasts and fungi are responsible for alcoholic fermentations, which is the most important fermentation worldwide and is used to make alcoholic drinks. Alcoholic fermentation has been incorporated into the way of life of every civilisation, with earliest record of alcoholic fermentation dating back to ancient Egypt, on mural and tomb ornaments, and scanning electron microscopy retrieved evidences of beer remains in a tomb pottery vessel dated from between 1550 and 1307 B.C. from the Valley of the Kings (Delwen 1996).

Another contribution of fungi to the food industry is, in the same way as it is useful to the biotechnology industry, the production of enzymes used in food manufacture processes. These are many and used in many processes. They include macerating enzymes like pectinases and cellulases, used in the improvement in pressing and extraction of fruit juices and oil for instance, pectins and polygalacturonases for the clarification of fruit juices, β-glucanases and mannanases used in food safety and preservation...

The last contribution of fungi to the food industry lies within the fermentation processes associated to cheese production and cured meats such as salami, and asian products. We take cheese as an example.

Cheese has been made by man for centuries. It was until recent conservation means (such as pasteurisation and refrigeration) the only method to preserve milk. Cheese is a solid or semi-solid food protein product made of milk. A basic outline of the cheese-making process would be the following: A first step involves adjusting milk pH between 5.8 and 6.4, and adding coagulating

enzymes in a cheese making vat. A subsequent reaction with casein solidifies the milk in the vat. In a second step, the initial coagulum is fragmented and heated to around 60°C, and stirred. This step drives the liquid whey out of the coagulum. This is possible because enzymes digests the casein and enables the formation of calcium bonds. Water is forced out of this structure (the curd) by the formation of hydrophobic region. In the following steps the curd will be cooked, scalded and processed into cheese. The differences in the end products (type of cheese) are due to differences in lactobacterial fermentation (and associated composition in species), temperature, pH and various additives, as well as differences in ripening time.

Contributions of fungi are at two stages in cheesemaking: first, fungal enzymes produced by *Mucor*, *Aspergillus* and others are sometimes used to enhance the coagulation of milk in a form of "microbial rennet" (rennet is the term for the traditional mixture used as the primary coagulant of milk proteins).

Secondly, and perhaps even more importantly, if considering the impact on the texture and flavour of the end product, fungi are important in cheese ripening. Cheese ripening relies on a complex interplay of a host of metabolic process and pathways. Flavour compounds are produced through proteolysis, lypolysis, glycolysis, citrate and lactate metabolisms. There are over 300 of these "flavour compounds" (Moore et al. 2011; Marilley & Casey 2004). Differences in production of these compounds (which is dependent itself on the different composition of the cheese microbiota) is responsible for the wide variety of cheese. Fungi take part in cheese ripening in the production of blue cheese like Roquefort, Stilton, Danish Blue and Blue Cheshire for example. In these cheeses, *Penicillium roqueforti* is inoculated into the cheese before storage by inoculation of spores into the cheese using metal combs. The fungus will grow throughout the cheese after inoculation and into the voids of between curd particles. During this growth are produced flavour and odour compounds (Moore et al. 2011). Holes and tracks made by the metal combs during inoculation are usually visible on the end product.

Soft cheese like camembert and brie are ripened by the mould *Penicillium camemberti* and *Geotrichum candidum*. *Penicillium camemberti* is responsible for changes in the texture of the cheese. It grows and the surface of the cheese and produces enzymes which digests the curd from the outside towards the centre. The white curd of these cheeses is in fact the mould itself. Other *penicillia* are used in the production of cured and fermented meet products, like for instance the emblematic *Penicillium nalgiovense*. It was used in the eighteenth century Italy and in other parts of Europe for the production of air dried sausages, for the same preservation reasons as for cheese and

milk. The fungus have several roles including imparting flavour and protecting the surface of spoilage by other microorganisms by occupying the surface.

The Food Microbiomes project which provides the frame and context for this PhD includes the sequencing of several *Penicillium* species involved in cheese making and other food production processes.

d. The *Penicillium* species:

Penicillium species are member of the Eurotiales order of ascomyceta. They are found in a wide range of habitats worldwide, including soil, marine environment, air and dust of indoor environments, food, silage... Their name comes from the resemblance of the conidiophores to a paint brush (*penicillus* in latin). They occur in food either as spoilage agents or as part of the food making process. Examples are many including the apple pathogen *P. expansum,* the citrus pathogen *P. digitatum*, the garlic pathogen *P. alii*. About 300 species have been described, with *Penicillium* spores found everywhere in the air and soil. They are the main cause of food spoilage.

Main species used in this work include:

-*Penicillium chrysogenum* and *Penicillium rubens*, of the *P. chrysogenum sensu lato* species complex. They are most common species in indoor environment. They are widely used as versatile cell factories to produce antiobiotics, including penicillin. Genome already sequenced in 2008 (Berg et al. 2008).

-*Penicillium roqueforti*, another widespread *Penicillium* occurring in soil, woods, silage and used to produce most of the blue cheeses (genome sequenced within the Food Microbiomes project).

-*Penicillium camemberti*, rarely found out of the dairy environment, responsible for the white rind of most camembert and brie like cheeses (genome sequenced during the Food Microbiomes project).

-*Penicillium digitatum*, a common cause of citrus fruit spoilage (two genomes sequenced in 2012 (Marcet-Houben et al. 2012a).

In addition, a number of other closely related *Penicillium* species were sequenced during the course of this project, and these includes strains of *P. roqueforti*, *P. fuscoglaucum, P. paneum, P. carneum, P. rubens, P.biforme, P. commune.*

Penicillia have been incriminated with producing many secondary metabolites (Frisvad et al. 2007), some of which may be toxic, others expected to have a role in the food making process itself. The availability of genomic sequences provides a unique window to investigate the secondary metabolite potential of these *Penicillia*. Work of carried out at the end of this thesis was done to provide through genome mining a catalogue of putative secondary metabolites of *Penicillia*. For this reason the following section describes secondary metabolite categories and synthesis.

e. Secondary metabolites:

Overview. Fungi produce many small products not necessary for their growth. These low molecular mass compounds are termed natural products or secondary metabolites. They have been proposed to play a role in fungal interaction with their ecosystems, as signalling molecules. They are believed to play roles in transcription and development too. Each fungus produces a wide and diverse array of secondary metabolites some of which toxic to other organisms, others enhancing growth of the host the fungus lives with, some responsible for the proper pigmentation and structure of conidia cell walls (Jørgensen et al. 2011). Man has been using them as antibiotics, immunosuppressant agents, anti-tumour compounds, pesticides and cholesterol lowering drugs. They are also responsible for the pathogenicity of some fungi on crops, or occur as hazardous mycotoxins in food products. The tremendous impact they have on society explains why genome mining efforts are carried out to retrieve or estimate the potentiality of fungi as secondary metabolites producers. Comparative genomics, with the advent of whole genome sequencing provides a good approach to identify the genes encoding secondary metabolite.

The vast majority of fungal secondary metabolites can be classified in three major groups (Hoffmeister & Keller 2007): Non ribosomal peptides (NRP), polyketides (PK) and hybrid non-ribosomal peptides and polyketides (NRPS-PKS hybrid). They are distinguished by the typical chemical backbone they are made of. Despite being different compounds, the way they are produced and the way the gene involved in their biosynthesis share many similar features. Genome mining efforts typically rely on the identification of the central genes required for their production, non-ribosomal peptide synthetases (NRPS), and polyketide synthases (PKS). Other genes involved together with the PKS or NRPS in the biosynthesis of secondary metabolites are usually organised in clusters. The following sections describe their organisation, as well as their regulations.

Structure. Secondary metabolites biosynthesis genes are organised in clusters (Brakhage 2013) (although a few exceptions have been described). These clusters consist of a couple to over a dozen genes. One or more central PKS or NRPS gene are encoding large multidomain enzymes responsible

for the construction of the PK or NRP backbone through iterative addition of amino acid (NRPS) or malonyl building blocks (PKS).

Non Ribosomal Peptide Synthetase structure. NRPS enzymes use an amino acid substrate, the selectivity of which is defined by 8 residues in the binding pocket of the enzyme (Rausch et al. 2005; Hansen et al. 2012). The minimal NRPS contains at least three domains, which are an adenylation domain, responsible for the activation of the amino acid, a peptydil carrier domain, which binds the cofactor to which the amino acid is covalently attached, and a condensation domain responsible for the peptide bond formation. This condensation domain may also have a role in amino acid selection (Hansen et al. 2012). Other domains can further complete these three minimal domains to further catalyse transformations on the extending amino acid backbone. Thioesterases domains are responsible for the cleavage or cyclisation of the amino acid backbone, reductases are involved in reducing the NRP. Epimerization domains are responsible for the change of the epimeric form of the amino acid substrate. Cyclisation domains modify serines, threonines or cysteines residues, and N-methylation domains have also been reported. The length of the Non Ribosomal Peptide produced is normally dependent on the number of adenylation domain present in the NRPS, although some exceptions have also been reported.

Polyketide synthase structure. PKS uses malonyl molecules as a substrate. Like the NRPS they are also made of three minimal domains, namely an acyl transferase, acyl carrier and ketoacyl synthase domain. The acyl transferase is responsible for the extender unit selection and transfer, the acyl carrier domain loads the extender unit and the ketoacyl synthase domain catalyses the decarboxtylative condensation of the extender unit with an acyl thioester. Numerous additional optional domains have also been reported, including: β-ketoacyl reductase responsible for reducing the ketone to a hydroxyl, dehydratase domains reducing hydroxyl groups to enoyl groups, enoyl reductase further reducing the enoyl to an alkylmethyltransferase, methyl transferase for C-methylation, and thioestherase, responsible for product release. Starter unit –ACP transacyclase selects the starter unit. The three minimal domains are used in each cycle, elongating the backbone with one single ketide unit every cycle. PKS can be further classified into reducing or non-reducing polyketide synthase based on which domain they harbour. The optional domains are not used in all iterations of the modules, providing a structural iterative module different from the functional iteration.

Both NRP and PK gene clusters also harbour additional tailoring enzymes encoded by other genes in the cluster providing another degree of secondary metabolites diversity.

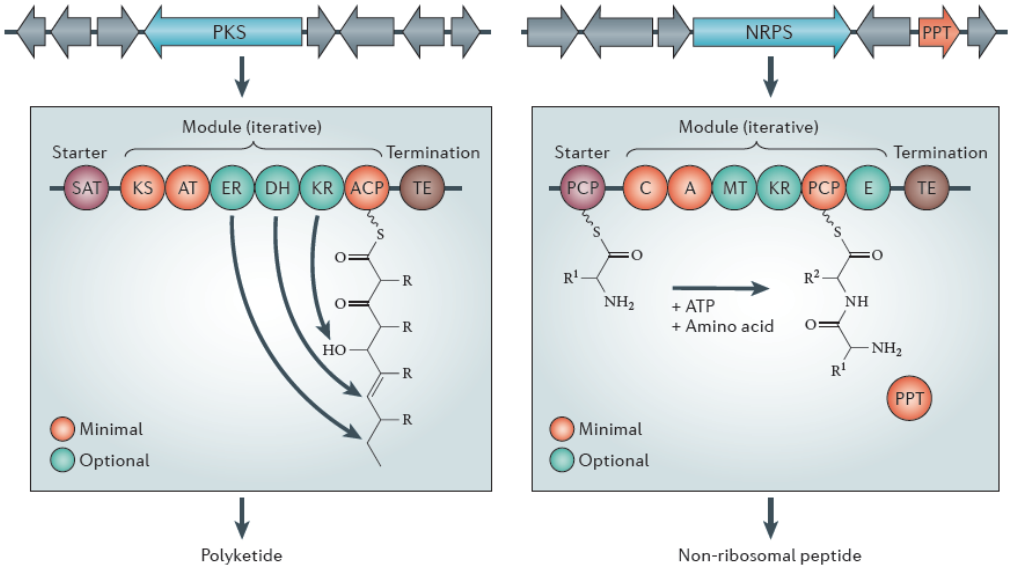


Figure 2 | **Gene clusters and enzymes for fungal secondary metabolism.** Gene clusters with a central non-ribosomal peptide synthetase (NRPS) gene (left) and polyketide synthase (PKS) gene (right). The domains of the encoded enzymes are indicated as spheres, with the minimal domains required to form a module of the enzyme shown in red. These enzymes are flanked by a starter ACP transacylase (SAT) domain and a termination domain. The acyltransferase (AT) domain selects extender units to add to the product and transfers them to the acyl-carrier protein (ACP) domain, which loads the units to extend the product. Several types of modification can take place on each product intermediate, as indicated by the various domains that can be present in the enzyme or encoded by other genes in the cluster: an adenylation (A) domain; a condensation (C) domain, which catalyses peptide bond formation; a dehydratase (DH) domain; an epimerization (E) domain; an enoyl reductase (ER) domain; a β-ketoacyl reductase (KR) domain; a ketoacyl synthase (KS) domain, for decarboxylative condensation of the extender unit (usually malonyl-CoA or methylmalonyl-CoA) with an acyl thioester; a methyltransferase (MT) domain; a peptidyl carrier protein (PCP) domain (also known as thiolation domain), which binds the cofactor 4′-phosphopantetheine (4′PP), to which the activated amino acid is covalently attached; a 4′PP transferase (PPT) domain, which is typically encoded by another gene in the cluster; and a thioesterase (TE) domain[6,16,18,22,23].

Figure 12: production of fungal secondary metabolites. Taken From Brakhage, *Nature* 2013.

Regulation. Fungal secondary metabolite expression is controlled and regulated at different levels by interactions between transcriptional, epigenetic and environmental and physiological factors. At the transcriptional level, the expression is controlled by pathway specific and global transcription factors.

Pathway specific regulators have been found for about 60% of the known clusters of both NRP and PK. These transcription factors are encoded within the clusters. Most of the PK factors belong to the zinc cluster family of transcription factors. They are more diverse for the NRP clusters. The signal inducing transcription of these factors is generally unknown, as well as the subsequent signal transduction cascade. Cases of crosstalk between two clusters where one cluster encodes a transcription factor responsible for the transcription of both the secondary metabolite encoded in its own cluster as well as the transcription factor of another cluster have been reported.

Transcription of secondary metabolite gene cluster is also mediated by globally acting transcription factors. Not part of the clusters they regulate, these global regulators control several other genes not involved in secondary metabolism. Famous examples are the CCAAT binding box complex (CBC) and

the pH response regulator pacC. PacC induces palD and acvA and ipnA in *A. Nidulans* at alkaline pH, and negatively regulates sterigmatocystin production, among other genes. The CBC has been shown to be involved in similar regulations under iron starvation conditions or particular red-ox physiological state. Similar regulators exist for carbon or nitrogen dependant environmental stimuli (Brakhage 2013). Global regulation represents an additional higher level of regulation of secondary metabolite clusters, integrating secondary metabolite clusters into larger regulation networks triggered by environmental and physiological or developmental conditions. They also provide a means to regulate those clusters not encoding their own pathway specific regulator.

Secondary metabolite regulation is also controlled by epigenetic mechanisms. Various histone modifications have been associated to expression of natural products including histone methylation and histone acetylation (Cichewicz 2010). These modifications are done by chromatin modifying enzymes which are also involved in other developmental and cellular processes, linking the expression of some secondary metabolites with developmental phases. Epigenetic control may be one of the reasons for the cluster organisation of secondary metabolites. Transcription can be restricted to a targeted area of the genome containing only the genes belonging to the cluster.

Chapter Four: Horizontal Gene Transfers: Rapid Acquisition of Novelties in the Face of Evolution?

Horizontal gene transfer, also known as lateral gene transfer is the non-sexual flow of genetic information between organisms. Called horizontal or lateral, as opposed to vertical flow through genetic inheritance from parent to offspring, this phenomenon was discovered during the mid-1990s. The existence of such a phenomenon, widely occurring in prokaryotic organisms, and still debated in eukaryotes affects many ideas pertaining to the tree of life, the neo Darwinian evolutionary synthesis, as well as practices in medicine (spread of antibiotic resistance) and food safety. Over the years, realisation of the extent and impact of this phenomenon led to an important and still on-going debate on the possibility of reconstructing the prokaryotic tree of life, as well as the adequacy of such a tool. More importantly, practices in healthcare and medicine and agriculture are being revised to avoid spread of antibiotic resistance, or contact between ecologically isolated organisms that could become superbugs.

1.prokaryotes: overview

The transmission of genetic material and information from parent to offspring, or the genetic inheritance is a well-known and well settled concept in biology. Experiments from the 1950's showed that the genetic material or the support of information was DNA. What was not realised at that time was that this material was transferred laterally from one organism to the other.

During the 1970's the question of how close and linked two species are, was addressed by phylogenetic association through comparison of molecular markers, be they nucleic acids or protein. At that time, the idea of using universally conserved markers or at least markers well conserved across many species was spread, and led to an increase in knowledge about the relatedness of many species.

A debate soon began to emerge, and still rages today on which parts of DNA, or phylogenetic markers are the most informative and should be used as a standard (Woese et al. 1975; Ciccarelli et al. 2006). This debate was alimented by the observation that some genes or genetic markers did not yield the same tree topology when used to construct a phylogeny.

The discrepancy between results yielded by comparison of genetic material originating from the same organism when compared to other organisms is in contradiction with the view that genetic material is transmitted from parent to offspring since the beginning of life and to the view that life evolves new species by gradually acquiring new characters or functions.

If combined, the demonstration that some life forms (bacteria) can pass on genetic material to co-existing organisms and the observation that part (fragments) of genetic information yields different relationship between species, the concept of lateral genetic exchange (transfer) emerge as the best explanation to a variation or contradiction between phylogenetic trees.

On this background, lateral genetic transfer emerged as a mechanism that could occur between unicellular organisms, mainly prokaryotes in the 80's. Few at that time considered the issues such a phenomenon could raise, if not anecdotal and limited. With the rise of genomics, more and more cases of phylogenetic incongruences were uncovered. These are most easily explained by horizontal gene transfer. Controversy arose through several aspects. Indeed, such a phenomenon questions several dogma of biology.

a. Prokaryotes: history

Early in the XXth century several experiments hinted at the fact that phenotypic traits could be passed from one bacterium to another, not necessarily related strain (Gurney-Dixon 1919; Griffith 1929; Avery et al. 1944). Over the last two decades, with the rise of genomic sequencing and comparative genomics, lateral genetic transfers have been repetitively found and described. It appears frequent in bacteria (Ochman et al. 2000; Woese 2000; Jain et al. 2003; Nakamura et al. 2004).

In prokaryotes, the exchange of DNA occurs through several well documented processes. This exchange needs to undergo several steps, all of them providing opportunities and barriers to fixation of the genetic material laterally transferred in the population. A successful horizontal genetic transfer event comprises the following steps:

-presentation of free, available DNA to a potential host cell,

-evasion of the host cell defence mechanisms,

-establishment as part of the cell genomic information, either as an extra-chromosomal element, or by integrating through recombination in the host genome,

-propagation in the population,

-amelioration of the sequence and integration into host cellular networks,

-long term fixation in the species' core or pan-genome.

DNA can be available to potential host cell through several ways. The first prerequisite for genetic material to be transferred is a physical proximity to the recipient cell. The genetic material to be transferred needs to be in contact with the host cell and thus to originate from the same habitat. Free DNA exists in all environments. DNA has been shown to exist in soil or marine ecological niches as free molecules (Levy-Booth et al. 2007). This DNA originates from living bacteria, fungi, or plant cells, decomposing biomass, dying bacteria, plant pollen or root cells. It has been shown to be up to fairly long sizes (up to tens of kilobases) and can persist for long times (Vlassov et al. 2007).

Apart from DNA in the environment, DNA can be mobilised though viruses which can mediate transfer both of viral DNA, but also of host DNA through subsequent replication and packaging after infection. Cross-specificity and viruses with several hosts from different species are known, and viruses can spread rapidly among environments and biomes, where they are known to persist for long times.

Living cells from the same or a different species can also act as DNA presenters through different mechanisms described below. These mechanisms are physiological processes frequently occurring in the life of the cell. Suffice to say that some bacteria have evolved quorum sensing strategies to stimulate competence and thus promote lateral gene transfer and intake of DNA. Of note, several features of DNA sequences can promote such transfers. Repetitive DNA and particularly mobile genetic elements such as transposons and insertion sequences can all act as mobilisers of lateral genetic transfer. Other prokaryotic strategies promoting horizontal transfers include the formation of biofilms where lateral genetic transfer is favoured by several means: the close physical proximity of different species, the availability of free RNA and double stranded DNA (part of the matrix complex forming the biofilm) released upon cell death and by type-III secretion systems.

Biofilms, as well as many other different environments are considered as specialised and promoting horizontal gene transfer. These lateral genetic transfer hotspots bring together potential donors, recipients, vector and transfer agents, as well as selective pressure promoting transfers. Examples are many and include rumen and digestive tracts, cytoplasm of some symbiotic cells, biofilms and their associated environments, highly decomposed organic environments, milk products...

DNA available for transfer can be presented to a potential host cell through several physical forms (free molecule, viruses and cells) in more or less transfer prone environments. The following sections describe the means by which this mobilisable DNA can enter its future recipient cell.

b. DNA intake

Members of the prokaryotic kingdoms display at least three different ways to internalise the environmental DNA into the host cytoplasm, thereby achieving another step of lateral genetic transfer. These three mechanisms are competence (also known as DNA transformation), phage transduction, and conjugation.

Competence is a naturally occurring mechanism in some bacterial species by which exogenous DNA from the surrounding is transported across the cell membrane (Johnsborg et al. 2007). This mechanism is in most case sequence independent but sometimes relies on short oligonucleotide motifs on the chromosomal DNA. In such cases, intake of DNA is biased towards the same composition, limiting donor spectrum as well as promoting integration into the host through homologous recombination.

Phages are also vectors for lateral genetic transfer by providing through infection and subsequent cycle a means for foreign DNA to enter a host cell. Phages infect bacteria; can integrate themselves in the recipient's chromosome (in the case of a lysogenic cycle) Once integrated, the now named prophage can readily spread vertically in the infected population until induction of a lytic cycle. Parts of the prophages' host can be packaged into replicating phages and further transferred to other strains and species in following infections. Other viruses not integrating in the genome can similarly package and spread host's DNA to other genome. Thus, viral infection mediates not only lateral transfer of viral sequences but may also mediate transfer of genetic material originating from a previous host.

Plasmid mediated conjugation is an important agent of lateral genetic transfer. Conjugation is a mechanism by which direct cell-to-cell contact enables the exchange of DNA from one cell to another. Plasmids are a class of mobile genetic elements that exist and replicate extra-chromosomally. They are transferred though conjugation between different hosts, enabling the transfer of non-essential bacterial genes. Recombination plays an important role in their structure, rendering them diverse and mosaic (see below). Conjugation functions are not always, though often, encoded on the plasmid itself. In cases where conjugation is not encoded by genes residing on the plasmid, conjugation is initiated by a co-occurring plasmid. Plasmids have been described as common vectors of phenotypic traits not essential to the bacterium they reside in (Slater et al. 2008). The ability of conjugation between distant, distinct bacteria or archaea enables lateral transfers of the phenotypic determinants between well separated taxonomic units. Of note, in some cases, Internal Chromosomal Elements (ICEs) are also mobilised by plasmid during conjugation (Burrus et al. 2006).

Conjugation is also known for mediating transfer of parts or whole chromosome if the bacterium to bacterium connection remains stable for long enough (Thomas & Nielsen 2005).

A barrier to rampant and generalised transfer by plasmids resides in the exclusion mechanisms which prevent conjugation between non-compatible bacteria, *i.e.* those already carrying plasmids encoding a similar conjugation mechanism. Conjugation between bacteria is also controlled through the expression of surface proteins enabling conjugation of some plasmids but not others. A further barrier is represented by incompatibility groups, from which two plasmids belonging to the same group cannot co-occur because they use the same mechanisms for replication and inheritance (Carattoli et al. 2005). Plasmids with a broad host range are not uncommon, even though their stability differs in different host species or strains. This further highlights their role as agents of lateral genetic exchange in prokaryotes.

c. Host defence: barriers to rampant DNA exchange.

Once environmental or invading foreign DNA has successfully been transferred into the cytoplasm, it needs to evade host defence systems prior to being able to successfully integrate into the genome or into an extra-chromosomal stable support.

DNA having entered the cell by one of the above depicted process is usually single stranded. As such it is available for integration into the host genome by recombination.

However, prokaryotic cells withstand viral predation, plasmid invasion and exposure to other invading DNA through different mechanisms. Two of these mechanisms are now well characterised: the restriction-modification system, and the more recently described CRISPR-cas system. These systems rely on self/non-self discrimination of nucleic acids and, as such, are often regarded as a form of prokaryotic immunity (Makarova et al. 2013). These immunity systems usually targets DNA.

The restriction-modification system is a defence mechanism targeting foreign DNA through the modification of a short oligonucleotidic sequence by methylation. This protects the self DNA from subsequent degradation by a restriction endonuclease which recognises specific non-methylated sites for binding and cleavage of DNA. Such systems are numerous, well characterised due to their extensive use in biotechnological applications. They are subdivided in four major groups on the basis of their composition (restriction, modification and energy usage encoded as part of a domain of the enzyme, or separately –whether ATP or GTP dependant.)

Another defence mechanism against plasmids and phages is the CRISPR-cas system. CRISPR stands for Clustered Regularly Interspersed Short Palindomic Repeats, and cas for CRISPR-associated genes

(Horvath & Barrangou 2010). Seen as an adaptive microbial immune system CRISPR are found in about 90% of archaea and at least 40% of prokaryotic genomes. These loci typically consist of multiple non-contigous direct repeats separated by spacers (which are stretches of DNA of variable length and base composition of plasmid and viral origins). Next to these repeats are cas genes. These genes encode proteins with various domains resembling nucleases, helicases polynucleotidic binding proteins and polymerases. These CRISPR-cas systems prevent the infection by phages and plasmids by specifying small RNAs targeting viral or plasmid complementary sequences of invading foreign DNA, which is then degraded. The sequences specifying targets are captured from previously encountered phages and plasmids, in the environment. As such the CRISPR-cas system represents a programmable barrier to lateral genetic exchange (Skippington & Ragan 2011; Horvath & Barrangou 2010). They have been proposed to "retain the memory of the local virus population" at a specific location (Sorokin et al. 2010). Both CRISPR and R-M systems can be laterally transferred providing another hint at arms race between invading genetic material and host identity.

d. Integration into the recipient genome:

Once foreign DNA has been physically transferred to a new host cell, and once it has successfully evaded the hosts defence mechanisms described above, to persist in the host, the invading DNA fragment needs to integrate into the recipient's chromosome or as an extra-chromosomal element.

Mechanisms for integration are homologous recombination, illegitimate (non-homologous) recombination or homeologous recombination (homology facilitated illegitimate recombination) or via end-joining repair mechanisms (Chayot et al. 2010).

-Homologous recombination occurs efficiently, at high frequency and enables alleles to spread in a population (Lawrence & Retchless 2009). Recombination by this process proceeds with a 25-100 base pair long sequence with high similarity to the host's genome. The efficiency of the process is dependent on the divergence of the sequence.

-Less similar foreign DNA can also be integrated by illegitimate recombination. Albeit less frequent and less efficient than homologous recombination, this process mediates integration into the recipient of foreign DNA originating from more distant taxon (Ochman et al. 2000; Vos 2009).

-homology facilitated illegitimate recombination is a recombination mechanism in which the integration of a non similar sequence is mediated by a short stretch of nucleotides with similarity to the host. This short stretch of sequence is responsible for initiation of the recombination by providing

an anchor enabling its adjacent non similar DNA sequence to integrate (de Vries & Wackernagel 2002).

Whichever the recombination scheme involved, it seems that efficiency or recombination correlates both with sequence divergence and length of the fragment transferred, (Touchon et al. 2009) and tends to occur at preferential sites in the genomes, or recombination hotspots.

The mismatch repair system represents a barrier to recombination at least between divergent species (Gogarten et al. 2002). Mobile elements can promote recombination by providing homologous sequences.

e. Integration in cell functioning.

Another step in the establishment of foreign DNA into a new host genome is its integration into hosts regulatory and molecular interaction networks. A key step to persistence in a host cell, newly integrated DNA needs to be expressed and that expressed material in turn needs to successfully integrate itself within the cell machinery.

Significant upfront costs could result from the non regulated expression of foreign genetic material, with a decrease in the host's fitness (Dorman 2007; Navarre et al. 2007). Some systems exist to control these costs by silencing foreign DNA, as is the case for example with the H-NS DNA binding proteins of enteric bacteria. These proteins silence genes with an atypical G+C content, and for instance, 90% of the silenced genes through the H-NS system in Salmonella are of lateral origin (Navarre et al. 2007). This seems to point for a barrier to lateral genetic transfer role for this system.

Recruitment of transcriptional regulators recognising the host's transcription and translation signals as well as changes in the sequence through codon optimisation and sequence amelioration are strategies by which the expression of the newly acquired DNA can be expressed, and goes with a reduction of costs imposed on fitness. However sequence optimisation has been shown to take several millions of years. A further crucial point in the integration into protein networks and complexes is the recognition of local signals for proper protein folding and assembly.

f. The complexity hypothesis:

It has been put forward that proteins involved in less complex networks or with less partners could be more easily transferred and retained through lateral genetic transfer because they would be functional right after transfer or with very little optimisation (Jain et al. 2003). The complexity hypothesis, formulated by Jain and colleagues (Jain et al. 1999) states an inverse correlation between

the successful integration into a network, and the numbers of partners involved. Indeed genes encoding proteins requiring few interaction partners, taking part into simpler networks are more easily transferred, needs a weaker integration, as exemplified by the poster-child of lateral genetic transfer: the spread of antibiotic resistance (see below). Some genes also encode their own genetic regulators or are integrated within the host's regulation systems by replacing part of gene or network already under regulation in the host. Also in line with an importance of regulation for genes of lateral origins is their tendency to have more regulators than native genes (Price et al. 2008) thus avoiding erratic expression.

The fate of the transferred material not only depends on its integration into the hosts regulatory networks and subsequent molecular interactions of the encoded products, but also on its selection by evolutionary mechanisms.

g. The cost-benefit equation: A selection for innovation?

The persistence of the transferred genetic material also relies on a complex interplay between the costs imposed on the host's fitness (which are originating from all the steps described above : capture, integration, maintenance, fine tuning and expression) as well as on other factors such as linkage of other phenotypic determinants on the transferred material, structure and dynamics of the population in which the event occurred, as well as other selective regimes imposed by environmental constraints. As such, fixation in the host of laterally inherited genetic material is best thought of as a complex cost-benefit equation to fitness.

This equation comprises several parameters of which a brief overview is given below. These parameters are mainly mutational load, metabolic costs and regulatory overhead.

Mutational load corresponds to the costs imposed through disruption of genes and other regulatory loci by insertion, as well as sub-optimal codon usage and other side effects of partial de-tuning of regulatory and interaction networks. This load can be seen as the cost of expressing a foreign gene. Perhaps the clearest example of this parameter is provided by the existence of unclonable genes in *E. coli* due to toxicity of the cloned product. As mentioned above, silencing systems represent a way to avoid such drastic upfront costs.

Metabolic costs are many and have different origins. They are mainly costs of substrate utilisation, replication as well as costs resulting from the activity or localisation foreign protein in the cellular environment.

Another parameter, regulatory overhead is directly linked to the integration of invading DNA in molecular networks. The recruitment of regulators previously not in use increases the costs. The further the regulation needs to be tweaked for proper cell functioning, the higher the cost. Indeed genes of lateral origins which are autonomously regulated are expected to impose a weaker cost. The same is true for genes not requiring complex regulation and integration.

Linkage of phenotypic traits or determinants also can have a significant impact on fixation of transferred material. Plasmids, for instance frequently carry different genes involved in different processes. Antibiotic resistances genes can be in physical proximity of resistance to heavy metal genes and detergents. This can lead to the spread of antibiotic resistance in the environment, even without the use of antibiotics because fixation is driven by other pressures on a linked d locus.

Selective pressure from different origins combined to linkage of different traits further complexifies the fixation and spread into populations of transferred material (Skippington & Ragan 2011).

h. Impacts of lateral genetic transfer on the prokaryotic kingdoms

Fundamental considerations:

Although potentially having a significant impact on the host cell by altering its phenotype, the impact of this phenomenon on prokaryotic evolution would be limited to various anecdotal innovations if not widespread and frequently occurring. Over the years, as more and more cases of lateral genetic transfer have been described, a shift of vision between horizontal gene transfers as a phenomenon occurring moderately in prokaryotes, to a frequently occurring phenomenon in all prokaryotic species arose. This led to the main question underlying all of today's controversies: How important is horizontal gene transfer in evolution of earth's biota? (Boto 2010). This question can only be addressed by answering the following one: How many genes in any given organism have been acquired through lateral genetic exchange? Indeed a single or a few genes acquired by lateral genetic transfer in prokaryotes can lead to new functions allowing the recipient to exploit new ecological niches (Fournier & Gogarten 2008) but this impact is of little global evolutionary significance if the phenomenon is limited throughout the history of life.

Several studies partly answered this question of the importance of horizontal gene transfer to bacterial and archaeal evolution. Despite controversial results depending on the methodology used, the conclusion of these studies is that horizontal gene transfer is by no means neglectable in the evolution of prokaryotes. It appears as frequent and seems to have occurred all along the evolutionary timeframe. Estimates in the percentage of genes of lateral origin in each microbial

genome ranges from 1.6 to 32.6 percent of the gene content (Koonin et al. 2001), and a more provocative estimates of 81±1.5 percent has been proposed by Dagan and colleagues by adding the cumulative impact of lateral gene transfer towards lineages (Dagan et al. 2008).

However widespread and however large the proportion of genome affected, if some types of genes are not affected, it is still possible to use markers to reconstruct tree-like relationships.

On this background, and in correlation with the integration in molecular and regulatory networks discussed above, the notion of core and pan genome is highly relevant. A core genome is the set of universally conserved gene within a taxonomic level. The pan genome represents the set of genes present in some but not all (at least one) members of the same taxonomic level. This means that "the genome complex that characterizes a bacterial species is much larger than can be contained within any single cell" (Syvanen 2012). For example, in 16 *E. coli* strains, the number of genes varies in each strain from 4200 to 5500, with only 2200 found in all strains. These 2200 genes represent the core genome, and the 13000 other genes represent the pan-genome. Core genes are believed to be good phylogenetic markers to reconstruct phylogenetic trees. However, if core genes are affected by horizontal gene transfer, phylogenetic reconstruction becomes impossible. A study by Ge and colleagues (Ge et al. 2005) concluded after a survey of forty microbial genomes that horizontal gene transfer frequency of core genes is relatively low (with an average of 2% of the core orthologous genes studied) although it remains unclear whether their result is underestimated. Lowering statistical thresholds raised this estimate to 13.1% and Novichkor (Novichkov et al. 2004) obtained estimates of 17 to 30%. As such, even if occurring at a lower rate than in accessory genes (pan genome), lateral genetic transfer also affects the core genome and led several authors to argue over the possibility of reconstructing prokaryotic history through a tree of life (Doolittle & Bapteste 2007; Ciccarelli et al. 2006; Ge et al. 2005; Sorek et al. 2007). They proposed a web of life (Doolittle 1999), a ring of life (Rivera & Lake 2004) or a cobweb hanging from tree branches (Ge et al. 2005) rather than a tree of life as metaphors to translate the huge impact of lateral genetic transfer on prokaryotic evolution.

All these considerations highlight the important impact of widespread rampant lateral genetic transfer on prokaryotes and their evolution. If lateral genetic transfer is so frequent, the notion of bacterial species becomes blurred at best, if not simply impractical, as such a phenomenon renders this notion dynamic and ever changing despite some genes conserving traces of inheritance (not considering amelioration also blurring or even erasing traces of ancient transfers.) It is worth mentioning that trees remain a sound framework to assess "recent" strain or species relatedness.

i. Societal impact: The example of *S. aureus*.

Environments with highly selective pressure promote lateral genetic transfer. Lateral genetic transfers have been shown to occur frequently in many environments and can be further promoted by the direct or indirect activity of man (Gillings & Stokes 2012). Some of these homogenous environments include milk products and other food-associated environments, crops, as well as clinical settings. One of the less fundamental, more practical and applied example of the impact of widespread lateral gene transfer in prokaryotes is the emergence of antibiotics resistant strains due to extensive use of antibiotics in the last 50 years. (Skippington & Ragan 2011) this example has become the poster child of genetic transfer, due to easy translation into headlines of the emergence of superbugs.

An example is the emergence since the 1940's of resistant strains of *Staphylococcus aureus*. Multidrug resistant strains have been described as the result of serial epidemic waves. Lateral genetic transfer is believed to have played a critical role in the emergence of penicillin resistant(1940, plasmid borne resistance), methicillin resistant (MRSA, 1960, only a few years after methicillin was introduced), community associated MRSA strains, vancomycin intermediate resistant strains and vancomycin resistant strains (arisen from conjugation with a vancomycin resistant *enterococcus faecalis*). Interestingly, despite this multiple emergence, *S. aureus* is known to limit horizontal gene transfer through several R-M systems and CRISPR and both within and inter species transfer barriers are believed to be at play. *S. aureus* both limits and promote lateral genetic transfer through barriers (R-M systems and CRISPRs) and promoting mechanisms (peptides inducing aggregative state in *Enterococcus faecalis*) facilitating transfers.

Many other examples of antibiotic resistance spread exist, but the practical impact of lateral gene transfer goes beyond the clinical setting. Lateral transfer is widespread in man induced environment such as the food supply chain (milk products (van Reenen & Dicks 2011)) waste and disposal sites (Rizzo et al. 2013)...

j. Modelling horizontal genetic transfer: Genetic exchange communities.

Attempts to control bacterial pathogens through the use of wide spectrum antibiotics and many more specialised compounds is believed to have led to an increase of resistant phenotypes to antibiotics, but also of other associated resistance to disinfectants (Baker-Austin et al. 2006). These resistances are now found in clinical and veterinary settings, but have propagated to natural environments via human waste streams (Storteboom et al. 2010). Their ability to replicate allows an

increase of anthropogenic associated resistance in natural environments. Indeed waste streams not only release resistance associated genetic material but also antibiotics, disinfectants and heavy metals in different concentrations, exposing bacteria to gradients of compounds and thus favouring selection.

An important mean to acquire antimicrobial resistance is via lateral genetic transfer. Under selective condition for innovation, stress response facilitates lateral genetic transfer (Prudhomme et al. 2002). Waste waters and effluents bring together pathogens, commensal organisms and environmental prokaryotes in areas with selective agents and resistance-carrying genetic material. As such, these sites represent hotspots for lateral genetic transfer and through the creation of genetic exchange communities, promote the dissemination of new combinations of resistance to diverse species (Schlüter et al. 2008). This has led some scientists to question the action of man on bacterial evolvability (Gillings & Stokes 2012) as well as drawing a worrisome picture of future emergence of increasingly more resistant pathogens.

The full extent and frequency of lateral genetic transfer in prokaryotes led to the realisation of the existence of genetic exchange communities (Jain et al. 2003; Skippington & Ragan 2011).

Genetic Exchange communities: Jain and colleagues described it in 2003, while first coining the expression, as “a collection of organisms that can share genes, but need not be in physical proximity”. Their study examined which of several factors both environmental (or external) and internal (or physiological/genomic) where the most associated with lateral gene transfer, and concluded that because lateral gene transfer is affected mostly by internal parameters, the community in which gene is exchanged laterally is not restricted by environmental or physical proximity, and can potentially cross phyletic and ecological barriers at the scale of an ecosystem.

In brief, a genetic exchange community as described by Jain and colleagues depicts gene flux as crossing locations and taxons by means of horizontal transfers.

Skippington and Ragan, in line with this conceptual definition, proposed a more detailed framework to describe and understand genetic exchange communities. Genetic Exchange Communities are the latest development of lateral genetic transfer research and provides a powerful tool to better assess and understand the impact of the genetic variation through gene flow in an ecosystem and evolutionary perspective. In their graph theory based framework, each node can be an entity carrying and potentially exchanging genetic material with another (or more) entity. The edges reaching each node provide the direction of transfer. Such a graph provides a snapshot of exchange relationships and potentialities in a complete ecosystem, but also across ecosystems.

Since exchange communities are constructed dynamically over time, the frequency of transfer can blurb the ancient exchange relationships, in a similar fashion as amelioration of codon usage sometimes prevents the identification of ancient events of horizontal gene transfer. As well, the dynamic nature of exchange communities is that certain nodes might not be represented on the graph (birth and death of plasmids, for instance).

Such a graphical network enables to identify, enumerate, and analyse GECs, as well as replacing them within a global map of lateral transfer that depicts the complete spectrum of exchange relationships, from mutual exchanges to one-off transformation by environmental DNA.

Depending on the scale and entity studied, GEC can link different genera across diverse hosts and/or environments (notably by the ability of plasmids to move from one species and from there to another.)

2. Eukaryotes:

While the importance and impact of lateral genetic exchange in prokaryotes is widely acknowledged, its contribution to eukaryotes has always been considered very limited and anecdotal. Despite some cases having been proposed in the early 90's (Syvanen 2012), very few examples existed before the beginning of 2000, and most of these examples were heavily hypothetical.

One of the reasons for this lack of examples was assumed to be due to the different nature of eukaryotes (especially differences in reproduction and accessibility of genomic material). However; examples are now piling up since the last ten-fifteen years or so, and are too many to be ignored.

a. Examples of lateral transfer in eukaryotes, and associated impacts:

The vast majority of lateral genetic transfer in eukaryotes is of prokaryotic origin. The comparison of eukaryotic genes to prokaryotes by Pisani et al (Pisani et al. 2007) identified many genes more closely related to prokaryotes, and only 36.6% of the 2 300 genes used for the study had no prokaryotic homolog.9.6% were of archeal origins, and 36.6% more closely related to bacteria than archea. This points out numerous contributions from prokaryotes to eukaryotic genomes.

Clear examples of bacterial or archeal transfer to eukaryotes are many: it is now known that some plant parasitic nematodes have gained or improved their plant cell wall degrading capacity by acquiring specific enzymes (Whiteman & Gloss 2010). Colonisation of land by plants is believed to have been promoted by the acquisition of bacterial gene through horizontal gene transfer (Yue et al.

2012). The hydra genome is thought to harbour 70 genes of bacterial prokaryotic origin (Chapman et al. 2010). An example of adaptive lateral gene transfer from prokaryotes to eukaryote led to parasitic properties of a beetle on coffee beans (Acuña et al. 2012). Obligate endosymbiontic *Wolbachia* has transferred many parts of its chromosome, including a case of almost complete chromosome to its host genome (Dunning Hotopp et al. 2007). *Leishmania* and trypanosomes are supposed to have evolved lineage specific biochemical properties via lateral transfer too. Prokaryotic contributions to functions related to anaerobic lifestyle in microaerophilic eukaryotes have also been described (Andersson 2009).

Unclear mechanisms:

Although as of yet quite unclear, several features can be responsible for transfer in eukaryotes. The steps of transfer are most likely similar to prokaryotes, with differences in the mechanism regulating these steps. For these reasons, among others, there is a bias in the extent of known genetic exchange in eukaryotic lineages. Some prokaryote to eukaryote transfers are easily explained by endosymbiontic relationships**.** Likewise, the phagotrophic ability of certain eukaryote is thought to help mediating entry of foreign DNA into the cell. One of the major innovations in the eukaryotic kingdom is the endoskeleton that allows some eukaryotes to engulf cells. These cells are sometimes retained as endosymbionts providing opportunity for transfer.

b. Prokaryote-to-eukaryote:

As discussed above, there seems to be a dominant pattern of prokaryote-to-eukaryote genetic exchange. Endosymbiotic transfer probably contributes to his bias, and is also considered to be responsible for the emergence of plastids and mitochondria. Transfers of genetic material from organelles to organelles or from organelles to the nucleus have also been described endo and ecto symbionts of prokaryotic nature can transfer large parts of their genome (e.g *Wolbacchia* and *buchnera*). Another reason for the observed trend in predominance of prokaryote to eukaryote gene transfer is most likely due to biases in approaches and experiments. The prokaryotic kingdoms are well sampled from a genome sequencing point of view, differences in sequence composition between eukaryotes and prokaryotes are usually marked. Other reasons also include the prokaryotic pool being bigger than the eukaryotic one, in terms of presence in ecosystems as well as most likely in closer physical proximity and interacting more than between eukaryotes (Keeling & Palmer 2008).

Also worth mentioning exclusion of bacterial sequence in *de novo* genome assembly of eukaryotic genomes is likely to lead to overlooked estimates of horizontal gene transfer of prokaryotic origins in

eukaryotes. However this has also led to high profile reports of bacterial gene transfer in human genomes, which was later proven wrong. This example has probably contributed to a decrease of interest in horizontal transfer searches in newly sequenced eukaryotic genomes.

Another reason explaining this prokaryotic origin bias could arise from the different nature of gene organisation. As prokaryotic genes are clustered in functional operons, a single transfer event of a relatively small fragment might result in an easier acquisition of a complete metabolic pathway.

c. Eukaryote-to-prokaryotes:

Very few cases of eukaryote to prokaryote have been described. These cases involve the acquisition of genes only found in eukaryotes. They are thought to provide the host with new functions. Examples include the transfer of a kinesin light chain encoding gene into an alpha & beta tubulin encoding operon in *prosthecobacter*, and genes encoding actin in *Mycrocystis aeruginosa*. In both cases it is likely that these proteins have structural roles in their prokaryotic host. Another example is the transfer of a fructose biphosphate aldolase (FBA) from red algal origin into *prochlorococcus* and *synechococcus*, inserted next to a FBA of non homologous prokaryotic origin. It is believed these genes play an important role in carbon fixation in their recipient.

Given the Impact of gene transfer in prokaryote, this low frequency of genes of eukaryotic origins is surprising. Horizontal transfer from eukaryote to prokaryote should be easy to detect. Potential reasons are differences in opportunities for transfer and the lack of conjugation or transduction between these kingdoms. The presence of introns may also play a major role. Of note, this tendency could also be due to a lower potential for innovations providing adaptive/evolutive advantages to prokaryotes.

d. Eukaryote-to-eukaryote:

Eukaryote to eukaryote gene transfer is not unheard of. Inter eukaryote transfer is often perceived as very limited because of the barriers a successful transfer event would need to overcome to become fixed in the population. These barriers include the lack of well established mechanisms for DNA intake (although phagotrophic feeding and hybridisation between species have been put forward).

Nevertheless, many examples have been described (for reviews, see Syvanen 2012; Keeling 2009; Keeling & Palmer 2008; Andersson 2009) including many examples of transposable elements (Price et al. 2008), some antifreeze proteins transferred between different species of fishes (Graham et al.

2008), carotenoid pathways from fungi to animals (Moran & Jarvik 2010; Altincicek et al. 2012) and a striking case involving bdelloid rotifers where a fourth of the genome arose from lateral gene transfer of eukaryotic and prokaryotic origin (Gladyshev et al. 2008). It has been proposed that in these strictly asexual species, horizontal gene transfer is a substitute for sexual reproduction, by providing an alternative to Muller's ratchet (Syvanen 2012).

Several biases have been put forward as limiting eukaryote to eukaryote gene transfer discovery, including an insufficient sampling of eukaryotic genes (this bias is being reduced with the advent of second and third generation sequencing), and confounding rates of gene duplication within some lineages.

3 Fungi

The following section focuses on the horizontal gene transfer in fungi, as they are the object of this thesis.

Fungi are eukaryotic organisms with relatively small genome sizes (from a few to several tens of megabases). Among the fungal kingdom resides many organism of fundamental or applied importance (See the chapter on fungi in this introduction). For these reasons, the fungal kingdom is the best sampled kingdom with regard to genome availability and therefore represents a good start to study horizontal gene transfer in eukaryotes. Many reports of gene transfer into fungi have been published (for reviews see Fitzpatrick 2012; Richards et al. 2011).

Fungi have evolved key adaptive features rendering them highly successful (see chapter three). Osmotrophic feeding combined to robust chitin-rich cell walls well suited to resist osmotrophic pressure and to drive extensive growth into their habitat made them an essential component of many ecosystems (especially soil) by being the most important biomass decomposers. Indeed high metabolic rate, resistant cells, and fast growth are most likely responsible their ecological success (Richards 2011).

However, as discussed above, their lifestyle excludes phagotrophy, which is suspected to be a major source of lateral genetic transfer among eukaryotes. Their robust cell wall is also expected to limit DNA exchange, suggesting key adaption to lifestyle in fungi provides barrier to horizontal transfer. Other barriers common to eukaryotic lineages are also expected to limit lateral exchange. These potential barriers include a membrane-bound nucleus, differential intron processing, the presence of incompatible promoters and the genome organisation into chromatin.

Despite all these potential barriers, many cases of horizontal gene transfer in fungi have been reported. The numbers of cases are now piling up, with transfers ranging in sizes from individual genes to whole chromosomes, some examples being clusters of genes or even larger clusters, including a striking case of acquisition of a whole metabolic pathway by acquisition of a large cluster (Slot & Rokas 2011).

Among the reported cases are found transfers of bacterial origin (Fitzpatrick et al. 2008) whether a single gene or more and transfers from fungi to fungi (see fig 13). It has been suggested that the transfer of individual genes, gene clusters or entire chromosomes can have a significant impact on niche specification, disease emergence (see chapter three) or shifts in metabolic capacities. Indeed, most of the reported examples bear significantly on this hypothesis, and notable cases include: the acquisition of thirteen prokaryotic genes in *Saccharomyces s288c*, enabling biotin synthesis, anaerobic growth and different sulphate assimilation (Hall et al. 2005); the acquisition of 34 genes at three different loci in *Saccharomyces* EC118 promoting wine fermentation and adaptation to the wine fermentation niche (Novo et al. 2009a). These genes are involved in stress response, nitrogen and carbon metabolism, cellular transport. Rumen fungi also acquired glycosyl hydrolases from prokaryotes (Garcia-Vallvé et al. 2000) and *Metarizhium anisoplae* has acquired a gene required for insect virulence from a prokaryotic donor (Duan et al. 2009). Transfer of a nitrate assimilation cluster is thought to have happened at least two times, every time improving fitness, and in one case originating from oomycetes (Slot & Hibbett 2007). 57kb of a secondary metabolite cluster was transferred from *Aspergillus nidulans* to *Podospora anserina* (Slot & Rokas 2011).

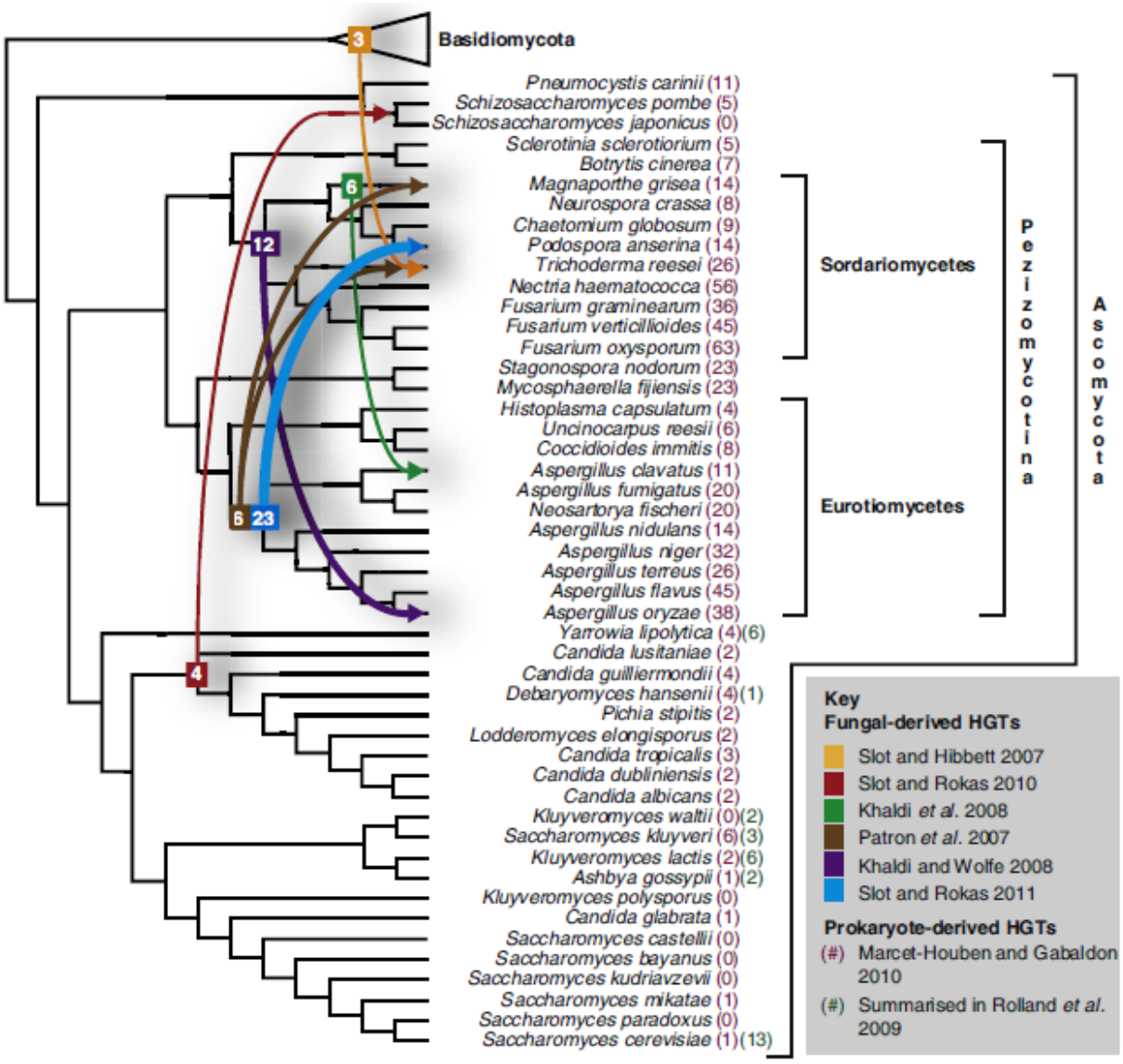


Figure 13: Gene Transfer between Fungi.
Examples of horizontal gene transfer (HGT) between fungi, supported with phylogenetic analysis and which match up to the taxon sampling in this tree, are illustrated using coloured arrows (see legend for source publications with the colour corresponding to the publication. For comparison, the numbers of currently identified prokaryote-derived HGTs are listed in parenthesis next to the species name. Taken from Richards, 2011.

Adaptation to niche, as exemplified by the cases above, is not the only consequence of horizontal gene transfer in fungi. Cases of virulence to plants of agricultural interest are known and these include the very recent transfer of a toxin to *Pyrenophora triciti-ripentis* from *Stagonospora nodorum* around the year 1941, which led to important agricultural losses in the United States (as described in the previous chapter about emerging fungal diseases, Friesen, 2006).A similar case of transfer between *Fusarium oxysporum* and *Nectria haematococca* enables *N. haematococca* to increase its virulence in pea by turning pisatin into a less toxic chemical (Matthews & Van Etten 1983). Another recent and striking case of whole chromosome transfer lies with the *Fusarium* species. Four of the 15

chromosomes of *F. Oxysporum* are of lateral origin by an unknown donor. Strikingly, one of these chromosomes mediates tomato pathogenicity and the authors of this study managed to reproduce this transfer in laboratory conditions by co-cultivation procedures, indicating that transfer could happen fairly easily in some environments (Ma et al. 2010)

The majority of transfers into the fungal kingdom are originating from prokaryotes. Cases of fungi to fungi transfers are numerous and more are being reported as new genomes emerge. Very few cases of transfers from other eukaryotic lineages are known. Richards and his colleagues reported four genes from various plant lineages and spider mites have acquired carotenoid biosynthesis from fungi (Richards et al. 2011; Altincicek et al. 2012). Despite several biases which could contribute to this trend (studies looking only for genes of prokaryotic origins, fewer fungal genomes than prokaryotic ones, genes originating from different kingdoms are easier to detect...) several facts argue in favour of this trend:

-If lateral genetic exchange is favoured by physical encounter, then the higher number of diverse prokaryotic populations constitutes a bigger pool for genetic exchange.

-prokaryotes organise their gene contents in operons, rendering innovation more likely in the case of exchange of a relatively small fragment of genetic material

-Also important intronless genes are more likely to be functional because they do not need recognition of splicing sites.

The number of reports of horizontal gene transfer in fungi today is not neglectable due and increasing. Despite many perceived barriers to horizontal gene transfer in fungi, several features of fungi may be promoting this phenomenon: bacteria to *Saccharomyces* conjugation and exchange of plasmids have been witnessed (Heinemann & Sprague 1989), transformation of yeasts have been achieved (Nevoigt et al. 2000) as well as *Agrobacterium* mediated transformation (Bundock et al. 1995) and it is also known that plasmid transfer can occur through cell lysis or cytoduction (Mentel et al. 2006).

Finally, potentially representing a more frequent mechanism and also occurring in non-yeast fungi, anastomosis, or somatic fusion between the same or different fungal species is likely to promote horizontal gene transfer including whole chromosome transfer (Ma et al. 2010; Friesen et al. 2006).

The fungal lifestyle is also likely to promote the encounter of many species, and ecological proximity could be responsible for the frequency of horizontal gene transfer.

At the present day, the evidence available suggests a non neglectable, but relatively low rate of horizontal gene transfer event in fungi (Fitzpatrick 2012). Still, as exemplified above, the impact of such transfers lead to niche shift or adaptation, as well as emergence of new pathogens, particularly in crops and other agricultural substrates.

Aim of the project n°2: Exploration of a case of multiple and frequent horizontal gene transfer between several *Penicillium* species from the food environment.

The availability of several *penicillium* species from the cheese environment provides a unique opportunity to look for horizontal gene transfer between these species. Horizontal gene transfer in food could potentially impact safety regulations and quality control procedures. The second part of this manuscript describes the occurrence of multiple and recent horizontal gene transfer of a large genomic island (over half a megabase) between *Penicillium* species from the cheese environment.

Chapter Five: How to Improve Penicillium roqueforti's genome assembly: development of a new methodology for improving assemblies.

The results described in this chapter have been produced with two aims in mind. The funding of the PhD being a CIFRE grant, a funding created to promote industrial and academic collaboration, the aim of this work was both to produce a high quality sequence of a *penicillium* genome for subsequent studies, and to develop a novel application of Genomic Vision's Molecular Combing technology. This chapter reports the development and application of a novel strategy for improving genome assemblies based on molecular combing and its application to the genome of a newly sequenced *penicillium* species.

1. Context and data availability :

Incentive for improving the assembly of a Penicillium species:

The work reported here was performed within the framework of the Food-Microbiomes project. This project aims at providing knowledge about the cheese ecosystems, in terms of species composition through conventional microbiology, metagenomic methods, as well as in terms of safe use of the microorganisms used in food processing (especially cheesemaking). In addition to the views on the necessity of releasing as good as possible sequences in public databases presented in the chapter one of the introduction, the necessity of providing a good quality genome arises from two main points:

-Metagenomics involves sequencing the DNA of a community rather than isolate genomes. Reads generated by this type of sequencing are then mapped to a reference database for the characterisation and quantification of the organisms present in the ecosystem under scrutiny. Prior to this study, no genomes of cheese filamentous fungi were available, rendering their identification through metagenomics impossible. Metagenomic screening of cheese ecosystems therefore require the creation of a database representative of the cheese environment. This is achieved through the collection of as many genomes as possible and the resolutive power of this type of methods also depends on the quality of genome sequences and annotation constituting the database. Several genomes of cheese filamentous fungi were sequenced by members of the Food Microbiomes consortium during this project, such as the yeast *Geotrichum candidum*, some *Mucor* species, a *sporendonema* species, as well as several *Penicillium* strains belonging to different species.

-One of the aims of the Food Microbiomes project is to provide sufficient genomic knowledge on some organisms frequently used in food processing. The availability of good quality genome sequences provides a starting point for many studies in which genomic content and context is required. Studies on evolution through genome dynamics and synteny relationships cannot be performed if the genomes are fragmented. Further, good quality assemblies provide a basis for improving other genomes as they are sequenced. Other studies requiring the observation of gene context, or cluster organisation like the detection of secondary metabolites for instance, can also be hindered by a fragmented assembly.

With in mind the aim of sequencing several *penicillium* strains from different species important for the food environment; we decided to improve the assembly of one genome to be used as a reference sequence.

Data available:

At the beginning of this project, the only *penicillium* genome available was the one of *Penicillium chrysogenum Wisconsin-1255* (later re-identified as *Penicillium rubens Wisconsin-1255*) which is the genome of the widely used versatile cell factory and industrial penicillin producer. The Food Microbiomes project sequenced over the course of four years over 10 *penicillium* strains belonging to various species, to provide a good overview of the cheese *penicilliums*, be they starters or contaminants. The characteristics of the assemblies of *Penicillium rubens Wisconsin 1255,* and *Penicillium roqueforti FM164* and *Penicillium camemberti FM041* (which were the first two genomes we sequenced) are shown in table 4, as well as those of *Penicillium paneum FM227* and *Penicillium carneum LCP5634*, which were sequenced later, but also used to generate these results. Of note, two alternate assemblies of *Penicillium roqueforti FM164* were performed, at the Genoscope and by the Laboratory of Plant Interactions with Microorganisms at the INRA Toulouse. These two assemblies yielded similar results with a fairly high number of scaffolds. The *Penicillium camemberti* assembly was even more fragmented, with 181 scaffolds for slightly higher, albeit similar genome size.

Strain and sequencing centre	Sequencing	Number of Scaffolds	N50	Genome size (Mb)
P. rubens Wisconsin 1255	Sanger	49	3'889'175	32
P. roqueforti FM164 Genoscope assembly Genoscope	454 – 8kkb mate pair library. Illumina 75 bp	81	-	28
P. roqueforti FM164 LIPM assembly Genoscope	454 – 8kkb mate pair library Illumina 75 bp	73	2'448'491	28
P. camemberi FM013 Genoscope	454 – 8kkb mate pair library Illumina 75 bp	181	947'716	34
P. paneum FM227 BGI-shenzen	Illumina	197	410'537 bp	26
P. carneum LCP 5634 BGI-shenzen	Illumina	224	416'003 bp	26

Table 2: Summary of the available assemblies during the assembly improvement part of the project.

Because *P. roqueforti* had a higher N50 and lower number of scaffolds, we choose to improve its assembly so it could be used as a reference in sequencing and working with the other *penicillium* sequences.

Worth mentioning, electrophortic karyotyping of *P. roqueforti FM164* was attempted during this Ph.D. but chromosome number could not be resolved due to too low yields of the protoplastisation method (this protoplastisation step is described in the material and methods of the next chapter). This prevented the determination of migration parameters for proper separation of the chromosomes and subsequent detection (10^5 to 10^6 protoplasts, where 10^8 to 10^9 protoplasts per plugs are usually required for conventional electrophoretic karyotyping). The precise number of chromosomes is therefore not known.

However, electrophoretic karyotypes of some *Penicillium* species show a number of chromosomes ranging from 4 to 8 (see table below).

Organism	Chromosome number	Chromosome size range (Mbp)	Estimated genome size (Mbp)
P. chrysogenum	4	6.8-10.4	34.1
P. notatum	4	5.4-10.8	32.1
P. nalgiovense	4	4.1-9.1	26.5
P. janthinellum	8	2.0-8.0	39.0-49.0
P. paxili	8	2.5-6.0	-
P. purpurogenum	5	2.3-7.1	21.2

Table 3: Chromosome number as determined by pulse field gel electrophoresis. Taken from (Chávez et al. 2001).

Penicillium roqueforti most likely has four to eight chromosomes covering its twenty eight megabase long genome. To improve the assembly, the number of scaffolds should be reduced by ordering and orienting these scaffolds into large superscaffolds representing most of the chromosomes.

2. Problematic: How to improve *Penicillium roqueforti FM164* assembly?

Given the sequencing technology used, it is unlikely that additional sequencing runs would significantly improve the assembly. The sequencing strategy used is already a hybrid approach with long reads in a large mate pair library (8kb inserts, 454 sequencing) and shorter high quality illumina reads. This hybrid strategy represents one of the best recipes for whole genome sequencing today, with long insert theoretically enabling repeat resolution and illumina reads correcting errors in homopolymers and filling in the gaps. Two different attempts at assembling the genome from the reads yielded by these sequencing technologies also resulted in assemblies of similar qualities (see table 2).

As a consequence, there is a need for an additional method to sequencing to improve the assembly. Molecular combing could represent a good technology to order and orient scaffolds on DNA molecules in *de novo* whole genome sequencing, thereby enabling an improvement of the assembly. However, despite having been proposed as a means to decrease complexity in hierarchical genome sequencing project in the past, at the physical mapping stage of these sequencing projects, and having been used to position two seed BACs in the tomato chromosome 6 sequencing project (Conti & Bensimon 2002; Todesco 2008)), no attempt at improving assemblies of *de novo* whole genome sequencing has ever been made. This is further highlighted in the very recent review of single molecule optical mapping technology (Levy-Sakin & Ebenstein 2013), which clearly states Genomic Vision's Molecular Combing technology as an optical mapping technology for "gene-scale" mapping. The application of molecular combing to the assembly problem would therefore require the

development of a new approach able to provide answers to the assembly problems. This translates into a strategy which could:

-resolve gaps in a (reference) sequence,

-position and order scaffolds separated by long distances,

-deal with several scaffolds at the same time,

-detect misassemblies in the bioinformatic based assembly.

Furthermore, Molecular combing is usually performed on bacterial cells, immortalised cell lines or blood samples, the cells of which are very different from the fungal cells. The latter have robust chitin-rich cell walls, are resilient to conventional DNA extraction procedures and usually require specific lysis protocol for DNA extraction. Molecular Combing is based on a specific extraction procedure in agarose plugs (similar to Pulse field Gel Electrophoresis extraction procedures) where the cells undergo a gentle lysis and protein digestion to protect the DNA fibres from shearing. A potential bottleneck is therefore the possibility of extracting high molecular weight, long DNA molecules from fungal cells prior to the combing step. From this context arise two main questions: "Can we extract and comb fungal DNA?" and "Can Molecular Combing provide answers to the assembly problems?"

3. Can we extract and comb fungal DNA?

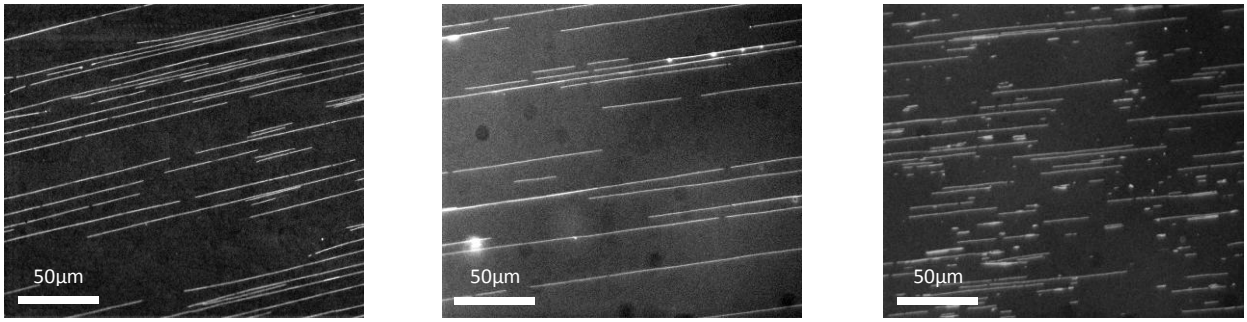
The conventional Molecular Combing procedure is based on several steps where a determined amount of cells is embedded into an agarose plug to protect DNA from mechanical shearing. Cells undergo a gentle lysis using a solution of sarkosyl, proteinase K and EDTA to break down cell membranes and for protein lysis. The agarose plug is then melted down in a combing buffer solution, and agarose molecules in the solution digested by a β-agarase digestion. Thus long DNA molecules are released in the combing solution with lengths as long as a megabase and the vast majority of fibres in the 500-700 kilobases range. The quantity of cells embedded in the plug determines the final DNA fibre density; the agarose plug step is responsible for protection of the DNA during the process until the final release of DNA in solution (more details on the procedure are available in the *BRCA1* article in the annexes of this thesis).

First extraction attempts at combing *Penicillium roqueforti FM164* DNA fibres from harvested spores did not yield any DNA using the conventional extraction procedure. This is most likely due to the protection provided by the spores' cell walls to cell and protein lysis by a solution of sarkosyl and proteinase K. As a consequence, for molecular combing, *P. roqueforti* DNA extraction requires a protoplastisation step prior to plug embedding. Such a protoplastisation protocol was kindly provided by P. Silar, F. Malagnac and A. Goarin (Genetics and Epigenetics of Fungi team, UMR8621, the protocol is detailed in the methods section of the next chapter). The inclusion of a protoplastisation step prior to plug embedding and cell lysis resulted in the obtention of combed DNA fibres. However DNA molecular combing quality was not as good as the expected "gold standard" of some bacterial cells or human cell lines (see figure 14). DNA fibres are shorter than expected, with many small fibres (10-20 kilobases) and some fibres being of longer length (300 to 500 kilobases). Several parameters which could potentially improve the quality of extraction and combing were tested, including increasing agarose concentration in the plug, which could protect the DNA from shearing during the extraction and release in solution, without any amelioration in combing quality, especially DNA fibre length and density.

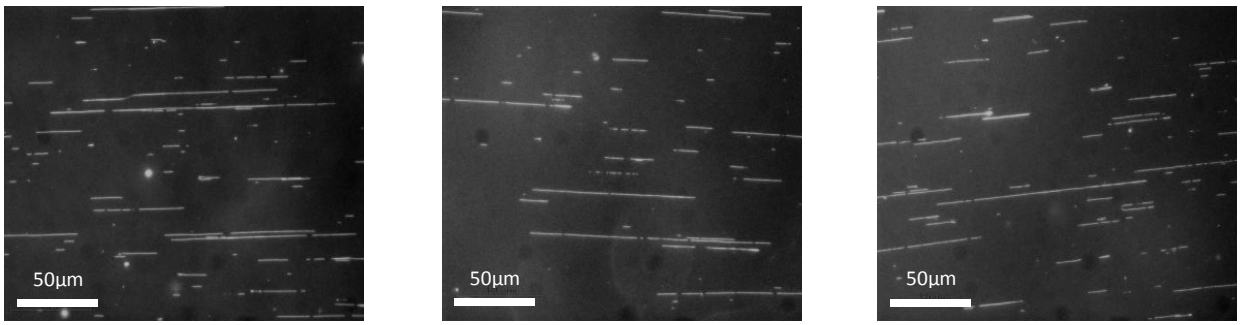
H. sapiens cell line
GM 17939

E. coli k12
MG1655

P. roqueforti
FM 164



P. roqueforti FM 164



0.8% [agarose]

1.5% [agarose]

2% [agarose]

Figure 14: Molecular Combing of *Penicillium roqueforti FM164.* top row shows example of good quality combed DNA fibres for *E.coli* and *H. Sapiens* cells. Combing quality for *P. roqueforti* is lower, as shown by the smaller fibres on the pictures. Bottom row: three different concentrations of agarose in the plug were tested, without improvement in overall combing quality and DNA fibre length. Images acquired at the 40x objective under a fluorescence microscope. DNA fibres are stained using YO-YO-1 intercalating dye. Scale bar: 50µm=100 kb.

In conclusion, *Penicillium roqueforti*'s DNA can be extracted and combed. DNA fibre length is not as good as the "gold standard", but some fibres are several hundreds of kilobases long, and may be long enough to physically map some scaffolds.

4. Can Molecular Combing provide answers to the assembly problems?

Development of a scaffolding methodology and application to the *Penicillium roqueforti FM164* assembly: A Proof of Concept.

DNA Molecular Combing has until now had two main types of applications: the study of DNA replication kinetics and the detection of large rearrangements on the DNA molecule for diagnostic purposes (see chapter two, and annexes for an example of large rearrangements detection in the Breast Cancer Associated Genes, *BRCA1* and *BRCA2*). When used for physical mapping, either to observe replication kinetics at a locus of interest, or to detect rearrangements, the approach relies on the detection of a previously designed Genomic Morse Code (Lebofsky et al. 2006) or of alterations of it. A Genomic Morse Code is a set of carefully designed probes of different sizes, separated by different length, and of different colours, generating a signature (or pattern) typical of the loci of interest. For Molecular Combing to provide answers to the assembly problems, the use of genomic morse codes would be of a different nature (with the exception of the detection of misassemblies, see below). Genomic Morse Codes could in theory be designed on scaffolds and the patterns to detect would be the combination of several Genomic Morse Codes, indicative of the order and orientation of the observed scaffolds. For such an approach to be realistically applicable, several parameters need to be taken into account, including the size of the genome and the size of the scaffold to be positioned on DNA molecules, the length of the combed DNA fibres, and the number of probes necessary to generate the required number of Genomic Morse Codes. Fibre length defines the upper limit of long distance scaffolding that can be achieved (that is, if the longest fibres are 200 kilobases long, for instance, then the maximal distance between two scaffolds that can be detected would be 200 kilobases minus the sizes of both genomic morse codes (which ranges from 4-5 kilobases to several tens of kilobases.)

There are two options to consider prior to designing a strategy for improving assemblies. The methodology can rely only on molecular combing and the scaffolds produced by the bio-informatics assembly, or it can rely on the use of external available information to reduce the complexity and the work required to improve the assembly. Given that the aim of this project is both to produce a proof of concept of the application of molecular combing to improve Next Generation Sequencing based assemblies and to produce a good quality assembly of a *Penicillium* species, and also given the fact that several other assemblies were available when designing the strategy, we chose to develop a first

proof of concept based on comparative genomics. Below is given an overview of the concepts and workflow of the methodology developed, as well as examples of the answer Molecular Combing provided to the assembly problems. Finally an overview of the improved genome is presented and discussed.

The availability of another good quality assembly (*P. rubens Wisconsin-1255*, previously published in 2008, Van den Berg et al.) as well as several other draft assemblies of other *Penicillium* species sequenced during the course of this project (*P. camemberti FM013, P. carneum LCP 5634, P. paneum FM227*, see table 2) enables the inference of hypothesis regarding the possible of some physical link between some of the *Penicillium roqueforti* scaffolds. These "link hypotheses" are all made under the underlying assumption of conservation of synteny between species: If a scaffold in one species is syntenic with the extremity of two scaffolds in the assembly to be improved, then these scaffolds are likely to be next to another. As genomes rearrange, synteny is sometimes lost, and as a consequence, this assumption is not always true. Software for comparative genomics like the Artemis Comparison Tool (Carver et al. 2005) and MUMmer (Kurtz et al. 2004) are useful in observing regions of conserved similarities and to infer link hypothesis. Long range PCR can then be used for a first fast screening of the possible junctions. Link between scaffolds can be confirmed by the presence of an amplicon. After this first rapid screening of the hypotheses, Genomic Morse Codes can be constructed at the extremities of the remaining scaffolds, and the molecular combing step performed to link more scaffolds.

Below is a summary of the workflow conceived and applied during this PhD:

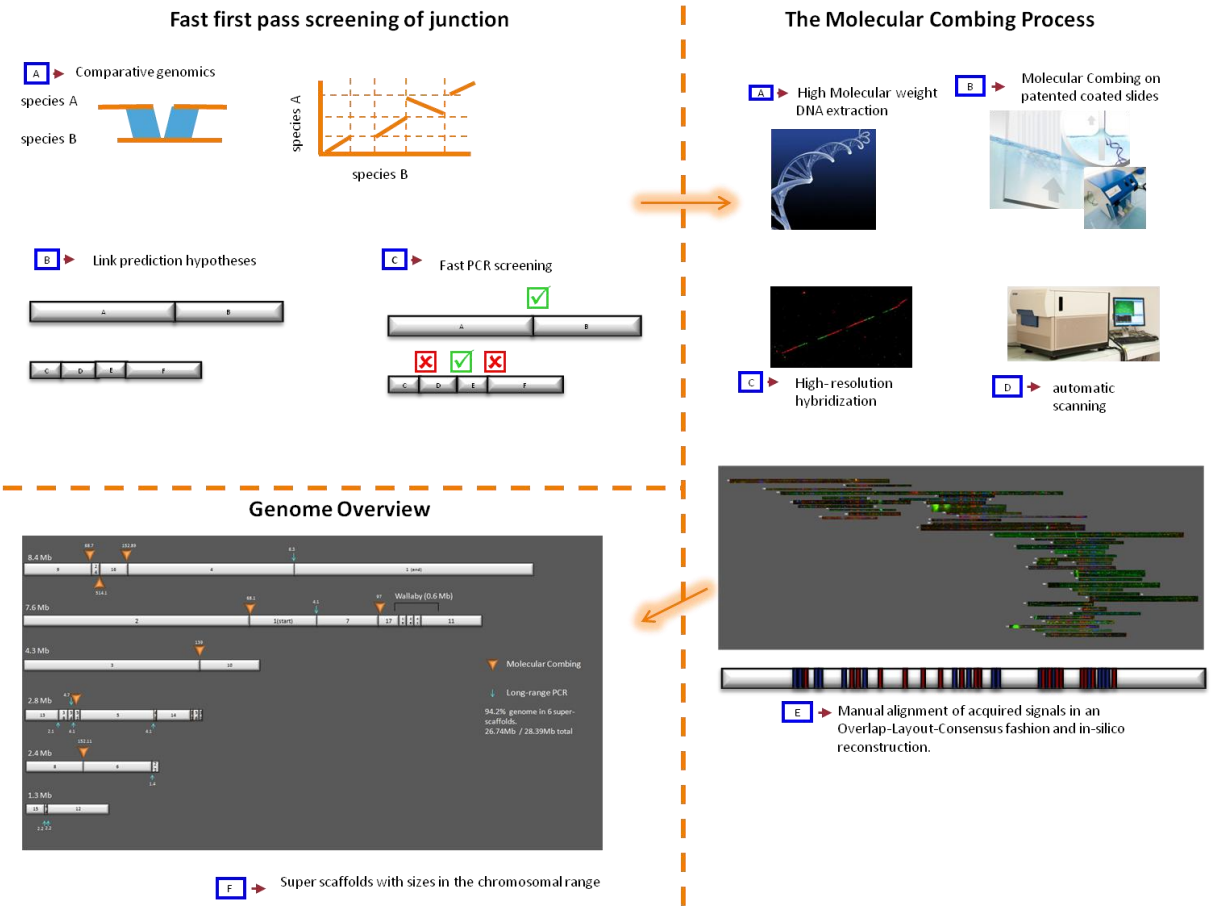


Figure 15: The comparative genomics based scaffolding procedure developed in this study. In a first step, comparative genomics allows the inference of "link prediction hypotheses" based on the comparison of scaffolds from different species and strains. A first pass screening of these hypotheses is performed using Long range PCR. In a third step, the remaining hypotheses are tested by designing Genomic Morse Codes at the extremities of scaffolds. Scaffolds for which no hypotheses have been made can also be included in this step. Finally, an improved assembly is produced, where scaffolds are ordered and oriented to reconstruct long pseudo molecule of chromosomal sizes.

a. Brief description of the key steps of the workflow:

Detection of conserved regions at the end of scaffolds:

The comparative genomics step is based on the pairwise comparison of genome sequences or scaffolds. Two programs are used, ACT and MUMmer. While ACT is a software enabling the visualisation of blast hits across scaffolds and enables to infer hypothesis based upon blast results, and provides a clear overview of where to design primers, MUMmer across whole genome sequences enables a global overview of synteny across genomes, and the inference of hypothesis at

several loci at the same time (fig 16). The objective of this step is simply to take advantage of possible synteny conservation at scaffold ends. These two programs were selected from the many comparative genomic software for both their ease of use, user friendly graphical output, rapid computing time, and the possibility of automating this step by simple shell scripts.

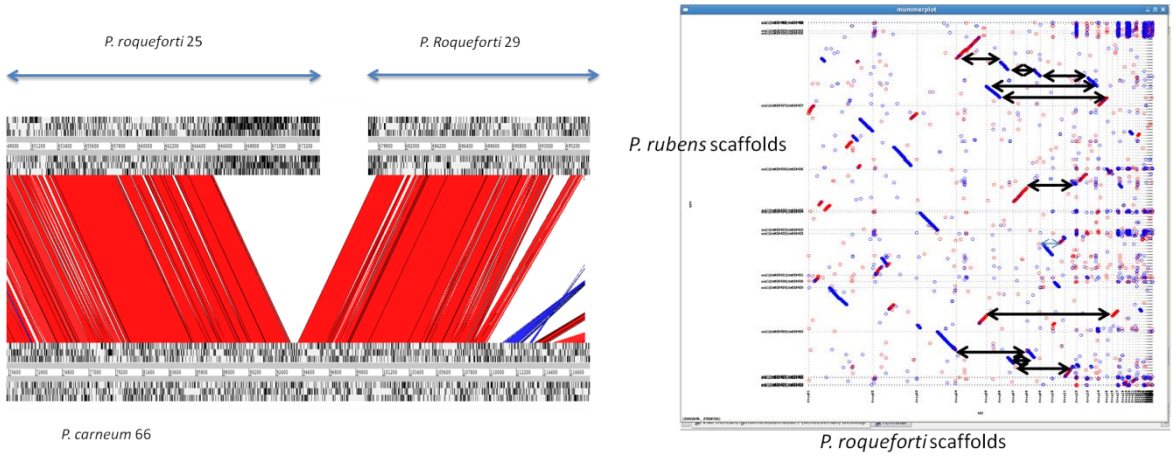


Figure 16: ACT & MUMmer are useful in inferring hypothesis. Left panel: local homology -ACT outputs of the comparison of *P. roqueforti* scaffolds 25 and 29 compared to scaffold 66 from *P. carneum* assembly. Synteny with *P.paneum* indicate these scaffold could lay one next to the other in *P.roqueforti.* Right Pannel : Genome wide homology -MUMmer outputs a dotplot of conserved regions across the scaffolds of two assemblies. Black arrows indicate possible link between scaffolds.

Fast PCR screening: Once hypotheses regarding possible link between scaffolds have been made, a Long range PCR screening is performed. Positive amplicons are indicative of scaffold junctions (and can be sequenced if complete gap filling is desired). The aim of this step is to reduce the number of scaffolds to be mapped by Molecular Combing, while avoiding the conventional combinatorial approach for finishing which requires many pcr reactions and is both costly and time consuming. Furthermore, A negative PCR result does not refute the link hypothesis between scaffolds: Scaffolds can be separated by more sequence than Long Range PCR can bridge (typically 20 -25 kilobases) and Long Range PCR heavily rely on proper reaction settings, including annealing temperature (a gradient has been used in this work to solve this problem), primer design, elongation time... Hypotheses which tested negative in this step are then tested by Molecular Combing, along with scaffolds for which no hypothesis could be made. The following section provides examples in which molecular combing provided answers to the assembly problems.

Molecular Combing:

As said above, four main types of issues need to be addressed when considering a physical mapping approach. These are the possibility of resolving gaps in already available sequences, the possibility of highlighting possible misassemblies in a newly sequenced genome, and the possibility of locating scaffolds separated by long distance, as well as locating several scaffolds at the same time. The following section provides examples in which molecular combing provided answers to the assembly problems.

b. Case n°1: Highlighting gaps in a reference sequence: the example of the *BRCA1* locus of the human genome.

While not involving the genome of *Penicillium roqueforti FM164*, this case is presented first, for two reasons. No reference sequence is available for *Pencillium roqueforti*, and genome finishing of reference sequence still represents an important issue in genomics. In addition, this result was obtained before starting the project, and is at the root of the idea of applying molecular combing to the assembly field (see annexes).

Molecular Combing can be used in a diagnostic setting to detect large gene rearrangements (i.e. duplication or amplification of part of a sequence, or deletion of a large fragment). These rearrangements are often missed by current sequencing technologies, and alternative technologies for their detection such as array-CGH or Multiplex Ligation Dependant Amplification (MLPA) have their drawbacks. In this context, I started a master internship at Genomic Vision with the task of developing a diagnostic test for the identification of large gene rearrangements in the Breast Cancer Associated Genes *BRCA1* and *BRCA2* (see annexes for more information). The detection of large genomic rearrangements by Molecular Combing relies on the observation of alteration in a previously designed Genomic Morse Code. The human reference genome sequence at the time of designing the Genomic Morse Code for *BRCA1* had a 100 kilobases long sequencing gap upstream of the locus. Since large rearrangements in *BRCA1* often overlap with adjacent genomic regions a Genomic Morse Code was designed to cover the gene and adjacent regions, including a set of probes located on the other side of the sequencing gap. Following Molecular Combing experiments (see figure 17), no gap could be observed, indicative of an error of the at the time latest available reference sequence assembly.

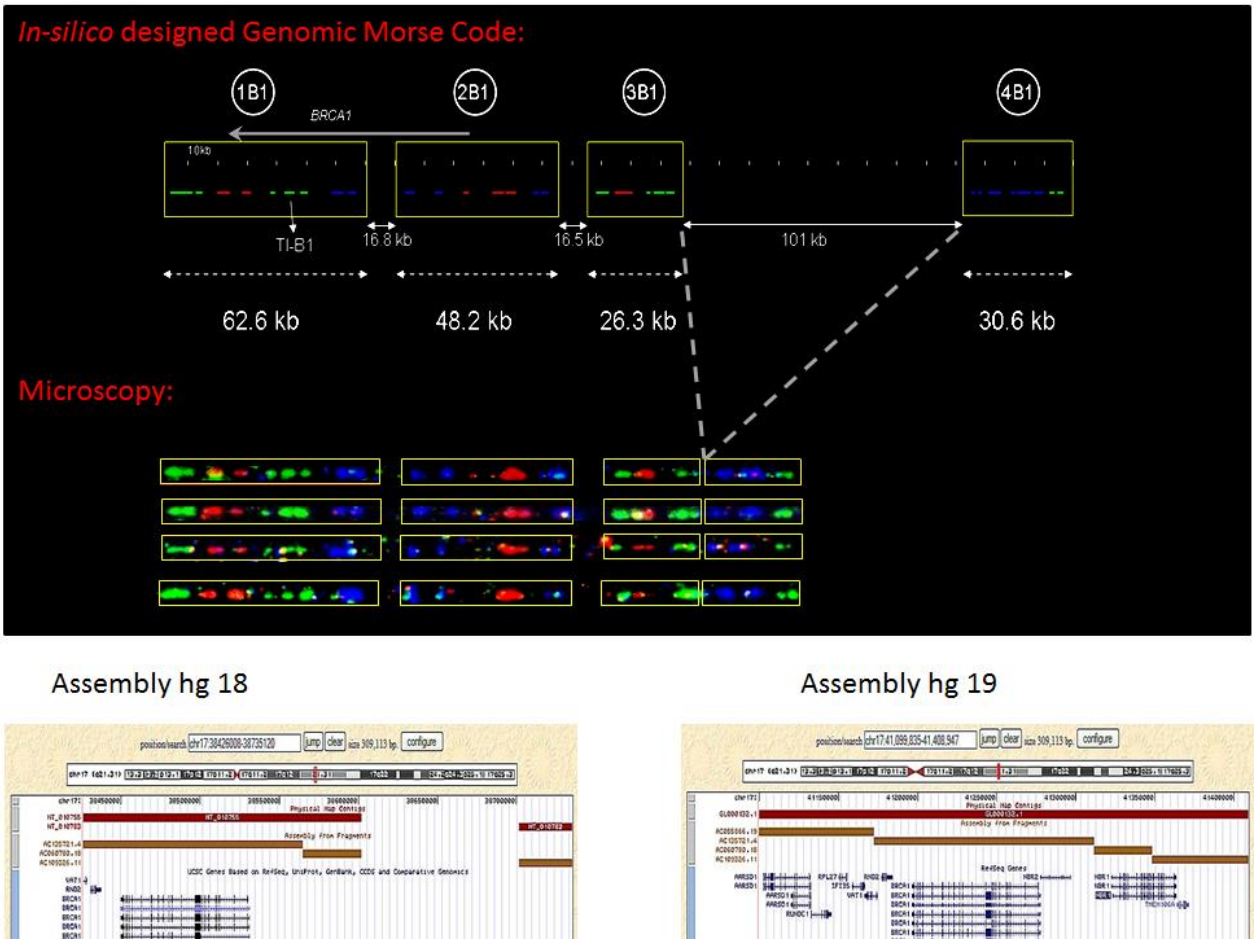


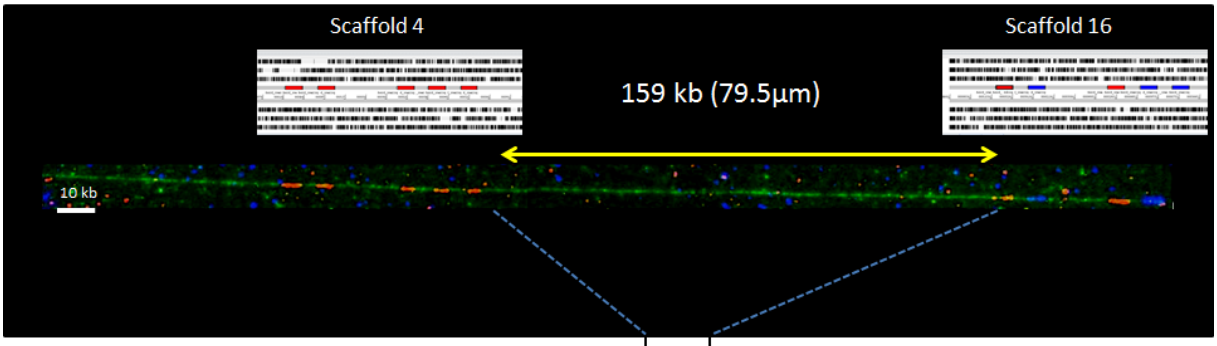
Figure 17: Solving gaps in reference sequence. Top panel show expected Genomic Morse Code as designed by *in-silico* prediction on Assembly hg18, the latest reference assembly at the time of design and obtained microscopy signals on Combed DNA. Absence of the 101 kilobase gap is clearly shown by the parts 3B1 and 4B1 of the genomic morse code being located next to the other. Bottom panel shows the difference between Assembly hg18, used for designed at the beginning of the *BRCA* project, and assembly hg19, released in 2009. This shows an independent validation by the human genome reference consortium of the molecular combing observation.

Molecular Combing is therefore a potent technology for correcting gaps in reference sequence. As such it could represent a good finishing and scaffolding technology. More details on the *BRCA* study can be found in the annexes.

c. Case n°2: Long distance scaffolding:

When no hypotheses can be made, or when these hypotheses cannot be confirmed by PCR, scaffolds may still be present on the same molecule but at a longer distance than can be bridged by mate pair information or long range PCR.

This case is illustrated below by the observation of the localisation of scaffold 4 and scaffold 16 on combed DNA fibres. Importantly, no hypothesis could be made regarding a potential link of these two scaffolds. The design of different Genomic Morse codes on their extremities and their subsequent hybridisation on combed DNA enabled to link them together. Measurements of the Genomic Morse Codes and gap between them indicated a distance of 159 kilobases between the two Genomic Morse Codes. Standard deviation between different measurements showed this measure to be within a 4kb precision range (see figure 18). Of note, by measuring several signals, it is possible to integrate in the measurements incomplete signals with partial genomic morse code. The size and colours of the probes, as well as the gaps between them enables to align signals and to use more measures in the case of too few complete signals. (For example, in the figure below, the GMC of the scaffold 16 is incomplete in four of the fibres measured. However, information about the gap could be recovered for 8 signals).



crop id		5,5	4,8	5,3	19,6	5,3	4,4	5,4	4,8	5,1	?	5,2	4,8	5,3	19,5	5,3	4,9	5,3	4,7	5,3	
Proq.0.005893.045303.007817.045460	185,47	5,15	5,49	7,09	20,96	5,81	3,22	7,1	3,86	6,12	152,89	5,48	6,44	5,17	20,31	3,22	7,41	5,8	5,15	4,83	105,45
Proq.0.027275.043251.028402.043383	42,89	5,15	8,71	4,51	22,25	4,83	6,12	5,81	4,83	3,87	164,17	2,57	10,32	3,56	9,02						
Proq.0.067415.003157.069198.003525	44,5	6,12	5,15	2,9	23,54	7,09	3,22	5,81	4,51	5,15	159,64	5,16	7,73	4,83	17,09	4,51	6,12	5,8			
Proq.0.000000.081554.002489.081824	13,54	5,15	6,12	6,12	19,67	7,09	4,84	4,83	4,51	5,15	157,39	6,45	4,19	7,41	19,35	5,16	4,51	3,54	6,78	3,86	437,04
Proq.0.029108.016988.031007.017119	134,16	7,09	4,83	5,81	19,35	7,09	4,51	8,71	4,51	5,8	158,67	6,44	3,54	5,48	21,28	6,45	4,83	5,48	4,19	6,45	115,78
Proq.0.056108.038596.058971.038969	587	6,13	4,51	5,15	22,25	3,86	6,78	6,12	4,51	5,48	165,45	5,8	6,12	3,54	21,92	3,22	5,8				
Proq.0.031793.003383.033512.003636	75,46	5,48	3,22	4,19	23,22	5,15	4,19	4,19	6,77	5,49	156,73	5,8	7,41	4,83	19,67	4,83	6,45	5,16	9,68	6,12	89,01
Proq.0.034766.003130.036706.003555	242,87	5,8	5,48	5,48	23,22	5,81	4,19	5,8	4,83	5,17	159,66	3,86	7,41	3,54	20,32	6,45	4,83	4,19			

mean meausurement	165,74	5,76	5,44	5,16	21,81	5,84	4,63	6,05	4,79	5,28	159,33	5,20	6,65	4,80	18,62	4,83	5,71	5,00	6,45	5,32	186,82
Standard deviation		0,68	1,58	1,29	1,63	1,20	1,27	1,38	0,85	0,66	4,03	1,34	2,13	1,32	4,14	1,33	1,05	0,93	2,40	1,20	
CV%		11,80	29,00	25,01	7,50	20,57	27,38	22,81	17,82	12,59	2,53	25,84	32,09	27,45	22,22	27,55	18,38	18,62	37,27	22,50	
delta		0,26	0,64	-0,14	2,21	0,54	0,23	0,65	-0,01	0,18		0,00	1,85	-0,51	-0,88	-0,47	0,81	-0,31	1,75	0,02	
Stretching Factor		1,91	1,77	2,06	1,80	1,81	1,90	1,79	2,00	1,93		2,00	1,44	2,21	2,09	2,19	1,72	2,12	1,46	1,99	

Figure 18: Molecular Combing Assembly can bridge long distance. Top panel indicate scaffold 4 and scaffold 16 lie on the same chromosome and are located 159 kb apart one from the other. A combed DNA fibre (Green) displays both Genomic Morse codes from scaffold 4 and 16 respectively. Square indicates *in-silico* mapping design of both GMC using the Artemis software. Table indicate measurements of probes and gaps between probes constituting the GMC. Measurements indicate a 159 ± 4 kb distance between the two GMCs. Crop id : image name, CV% variation coefficient, delta:

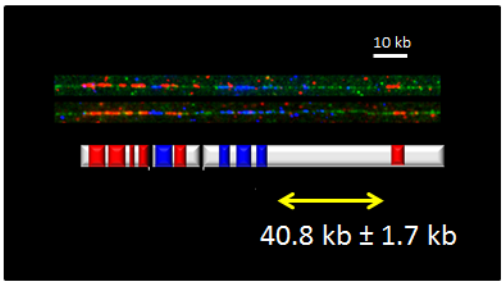
difference between mean measurement and expected measurements. Stretching factor serves as an indicator of combing quality (average stretching factor is around 2.0).

The use of Genomic Morse codes designed at the end of scaffolds enables to locate different scaffolds with long distances between them. Of note, distances recovered depends on the fibre length, and can be a lot bigger than what traditional mate pair libraries can bridge (long mate pair libraries are in the 40 kilobase range).

d. Case n°3: Highlighting misassemblies in the genome:

Assemblers are often confused by repeats. Genomes contain different type of repetitive sequences present in diverse quantities depending on the species considered, and as a result, genome assemblies contain misassemblies. Molecular Combing can detect misassemblies by detecting alterations in the expected Genomic Morse Code. Difference in measurements of observed genomic morse codes from what is expected by *in silico* design are indicative of misassemblies. Below is an example of such detected misassembly in *Penicillium roqueforti FM164*. Several signals were collected and measured (23). Measurement indicated a difference in the length between the last blue probe and the red probe (see figure 19). Measured distance was of 40.8 ± 1.78 kb. Expected distance as defined by the design of the GMC on the bioinformatic assembly indicated an expected distance of 48.2 kilobase. As a consequence, Molecular Combing has the potential of highlighting misassemblies, and these can then be corrected using conventional bioinformatic means (including mapping of the reads assembled and observation of mate pair constraints, comparison with alternative assemblies, or mapping of reads from other strains of the same species) or biological experiments (such as PCR amplification and sequencing of the amplicon for instance). In this case, the misassembly was caused by the presence of short LTR sequences, which lead to the insertion of two copies of sequences and four LTRs. This was confirmed using assemblies of other *Penicillium roqueforti* strains (issued from re-sequencing experiment using SOLiD technology, data not shown in this thesis.) A discussion of the methodology used to confirm the misassembly would be beyond the point of this chapter, as it aims at describing the answer molecular combing can provide to assembly.

a.



48.2 kb expected according to bioinformatic assembly.

b.

crop image id	4,9	48,2	4
14	2,28	40,72	3,24
199	4,84	40,4	4,51
293	4,2	39,75	4,52
294	3,88	41,38	3,88
448	3,92	41,56	3,92
457	2,97	43,64	3,24
499	4,84	40,78	4,84
520	3,56	44,47	4,84
645	2,6	41,4	3,88
689	5,17	41,32	4,84
743	4,52	41,63	5,8
757	1,93	42,02	4,52
795	4,19	35,8	4,51
886	6,16	42,8	4,24
1064	3,56	41,6	3,87
1091	4,83	39,35	5,48
1179	5,52	41,13	3,88
1185	5,84	41,43	4,52
1246	6,84	38,32	4,2
1257	5,52	40,17	5,17
1376	5,49	39,4	4,52
1386	6,45	39,37	5,17
1414	5,16	40	4,83

number measured	23	23	23
mean meausurement	4,53	40,80	4,45
Standard deviation	1,32	1,78	0,65
CV%	29,08	4,37	14,50
delta	-0,37	-7,40	0,45
Stretching Factor	2,16	2,36	1,80

c. Additional characterisation of the sequence

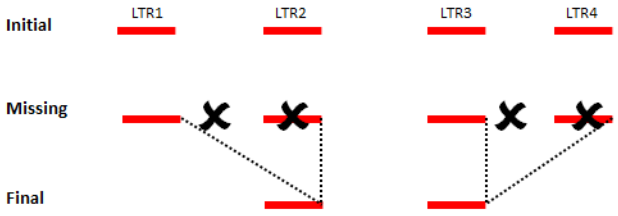


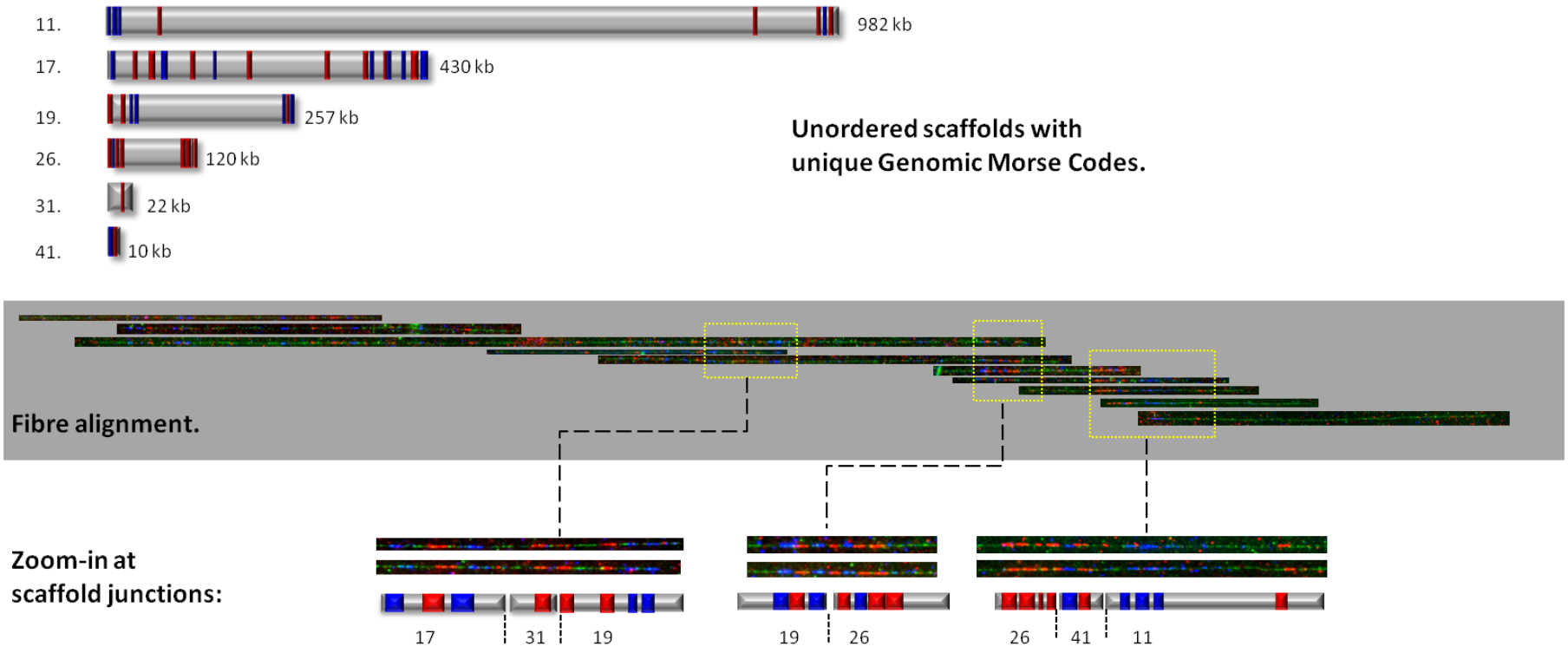
Figure 19: Molecular Combing can detect misassemblies: (a) examples of observed signals and expected genomic morse code, with measured distance between the last rightmost blue probe and the last red probe. (b) Measurements obtained for 23 similar signals. Expected length is indicated in the top row. The mean measurement of 40.8kb ±1.78 kb differs from the expected length by 7.40 kb (delta). (c) The misassembly was due to the insertion of two LTR repeats and some associated sequence between them (initial: initial assembly). Missing: black cross indicate non existing sequences. Final: structure of the final sequence. Only two LTR are really present.

e. Case n°4: Scaffolding several scaffolds at once:

While the previous examples have dealt with at most two scaffolds at the same time, to represent an efficient solution for further improving assemblies, the methodology presented here needs to deal with many scaffolds at the same time, thereby reducing the number of experiments to perform to improve the genome. The following example reports how, by combining unique GMCs and overlaps between them, a large locus can be characterised. Combed DNA fibres are broken in a random fashion, with different lengths. Fibres with hybridised GMCs can be aligned on the basis of shared patterns, in a process akin to the "overlap-layout-consensus" approach of the early assemblers. This enables to reconstruct a large locus from several combed DNA fibres. The longest fibres are indeed

the most informative, but any fibres having at least two probes should in theory be able to be positioned at a locus (unless the spacing and colour of these two probes are similar for two loci). Figure 20 displays an example of scaffolding by molecular combing using 6 different GMCs designed on six scaffolds of different sizes.

Figure 20: Scaffolding several scaffolds at once and reconstructing a large locus. Top panel shows the six GMCs designed on six scaffolds. Numbers indicate scaffold names, and theoretical lengths are displayed next to the scaffold. A fibre alignment (second panel) based on shared probes between overlapping fibres enables to link all six scaffolds together. Third panel shows zoom-ins at scaffold junctions. Bottom panel shows the large super-scaffold of 1.8 Mb reconstructed by aligning combed DNA fibres with GMCs.


11.
982 kb
17.
430 kb
19.
257 kb
26.
120 kb
31.
22 kb
41.
10 kb
Unordered scaffolds with
unique Genomic Morse Codes.
Fibre alignment.
Zoom-in at
scaffold junctions:
17
31
19
19
26
26
41
11
Final super-scaffold:
17
31
19
26
41
11
1'821kb long super scaffold

f. Overview of the results obtained on *Penicillium roqueforti FM164*'s assembly:

Using the strategy depicted above, the genome assembly of the *Penicillium roqueforti FM164* was improved by ordering and orienting most of the largest scaffolds into large super-scaffolds. The final assembly is now made of six large super-scaffolds with sizes ranging from 8.6 to 1.4 Megabase and a collection of forty-two shorter scaffolds. Molecular Combing identified fifteen junctions on these six super-scaffolds, while a first PCR screening identified eleven junctions between these scaffolds. The N50 was improved from 2'448'491 Mb to 7'752'559 Mb. The six large super-scaffolds are shown on figure 21. Of note, 5 large scaffolds and 1 superscaffold (with sizes between 300 and 100 kilobase could not be reunited with others, accounting for a large part of the remaining 8% of the genome that has not been incorporated into superscaffolds. These would require more experiments with new genomic morse codes to be positioned on the superscaffolds. Of note, comparative genomics derived hypothesis could either not be made, or were not confirmed by PCR. It could also be that these scaffolds are fragmented and inserted inside the super-scaffolds. It would be interesting to carry out further experiments to finish this assembly. Noteworthy, several large regions were observed by molecular combing. These do not seem to be of sizes in which the largest unplaced scaffolds could fit. It is possible that these represent large centromeres and rDNA loci. Another possible scenario is that they could also be indicative of misassemblies in the largest unplaced scaffolds were pieces of these unplaced scaffolds could fill in the gaps observed.

To conclude, Molecular Combing was successfully employed to improve the genome of *Penicillium roqueforti FM164*. While the resulting assembly is not complete, it still accounts for more than 92% of the genome on 6 superscaffolds. It is also interesting to note that large gaps observed by molecular combing could represent centromeres. If this is the case, given their location on four of the six superscaffolds, it could be that *Penicillium roqueforti* has only four chromosomes.

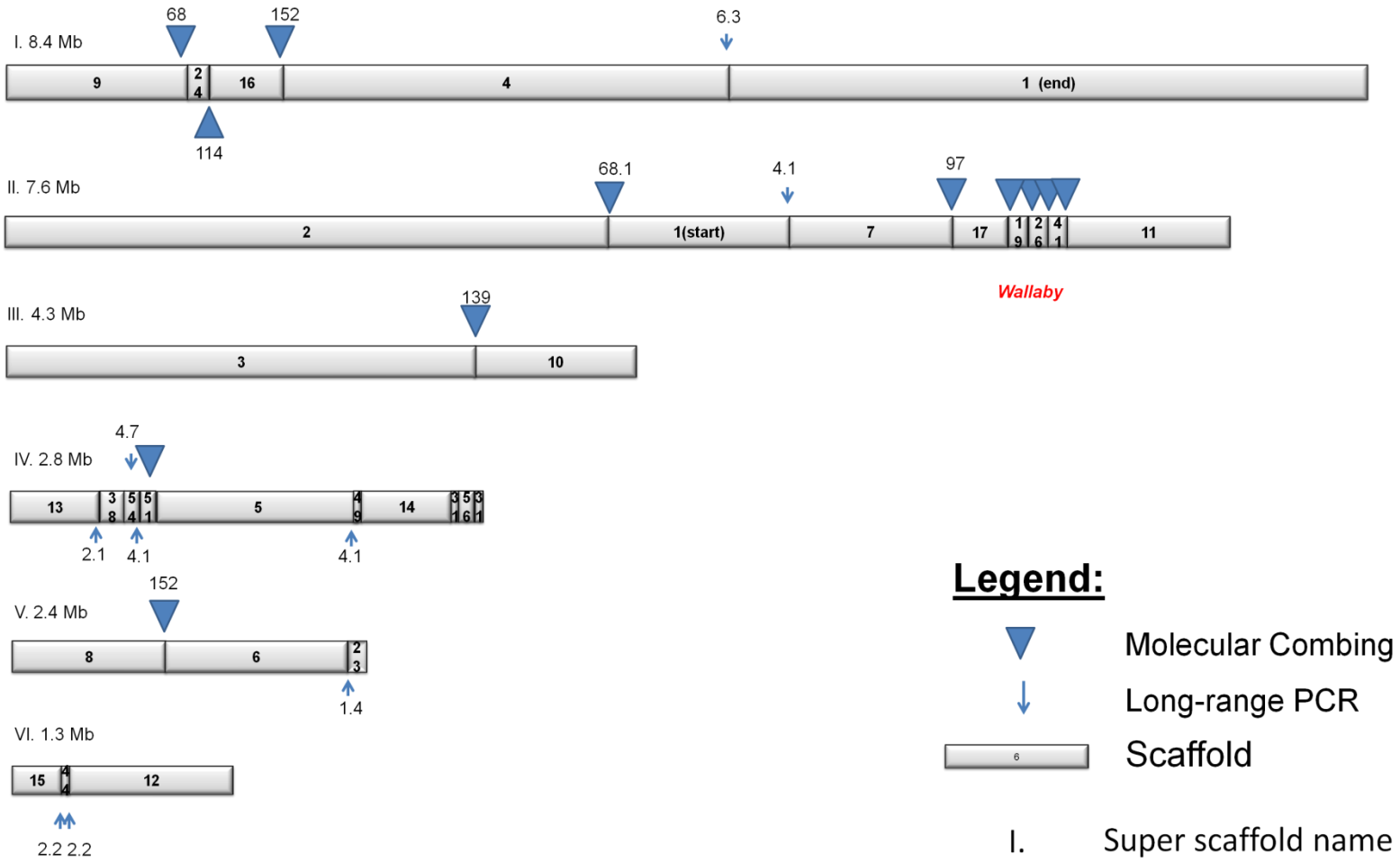


Figure 21: Overview of the final genome Assembly. The six largest superscaffolds are accounting for 92% of the genome, with sizes within the chromosomal range. Molecular Combing was particularly useful in linking many scaffolds with very long distance between them numbers indicate the distance between scaffolds in kb.

Chapter Six: Lateral Genetic Exchange in the Food Chain: Gene Flux in cheese.

The following chapter reports the multiple and recurrent occurrence of recent horizontal transfers of a very large genomic island between cheesemaking *penicillium species*. This article -which is at the time of writing, under revision at Nature Communications, deals with lateral genetic exchange in the food environment. Below is given a brief introduction prior to the article itself, which is included on the next page.

Context and broader picture

Horizontal gene transfer can be defined as the non-sexual movement of genetic information between two organisms. Horizontal transfer in prokaryotes is a widely acknowledged phenomenon, and recognised as ubiquitous with a tremendous impact on ecology and evolution. Horizontal gene transfers between eukaryotes have also been reported, with various ecological and functional impacts, but their extent and impact are still debated. They are often reported to be ancient, isolated or sporadic events and recent reviews in the field argue that "this flow of gene represents a form of genetic variation whose implications are not fully appreciated" (Syvanen 2012).

The findings presented in this chapter indicate a large genomic island has been repetitively horizontally transferred to several *Penicillium* species from the cheese environment. *Penicillium* species are ubiquitous filamentous fungi important for the production of many cheeses.

Compared to other cases of eukaryotic transfers our findings demonstrate:

- The transfer of a large region (0.5Mb, 250 genes)

- Its insertion within the chromosome of the recipient (as opposed to whole chromosome transfers which may occur in fungi)

- Its repetitive occurrence in a very recent timeframe (identical sequences, many species)

- Its likely triggering by man-made activity (food environment).

This has obvious implications for food safety assessment, as well as in other man-made environments such as crops. This case also argues for an under-evaluated occurrence and impact of horizontal gene transfer in eukaryotes, and our report could help in accelerating a re-thinking of the importance of eukaryotic gene transfer, both in man-made environments and in nature.

While the importance of lateral genetic exchange has been discussed in chapter four, the report of multiple and recent horizontal gene transfers in the dairy environment further adds to the emerging view of an understudied and most likely underestimated frequency of lateral genetic exchange in fungi. This work provides incentive for more screening of such events in food processing chains, especially when combined to other reports of horizontal gene transfer in the food chain, like for

instance between yeasts from the wine environment. Horizontal genetic transfer is a known process of adaptation to new environments or of niche shift. This has implications for the food industry, both in term of food safety assessment of the use of microorganisms and their associated product, as well as in term of innovation. The use of fungi as organisms to produce flavours or modify texture or appearance of food products could result in the encounter of isolated species, with impacts difficult to assess in the event of widespread transfers between strains.

In addition, in an age of globalisation and worldwide trade, as highlighted in the chapter three, New strains from the same or different species are more and more likely to be in contact, and could potentially exchange genetic material in a similar pattern than described in the next section...

Multiple recent horizontal transfers of a large genomic region - over 500 kb - in cheesemaking fungi

Kevin Cheeseman[1,2,3±], Jeanne Ropars[4,5,6±], Pierre Renault[1,3]*, Joëlle Dupont[4], Jérôme Gouzy[7,8,9], Antoine Branca[5,6], Anne-Laure Abraham[1,3], Maurizio Ceppi[2], Emmanuel Conseiller[2], Robert Debuchy[10,11], Fabienne Malagnac[10,12], Anne Goarin[10], Philippe Silar[10,12], Sandrine Lacoste[4], Erika Sallet[7,8,9], Aaron Bensimon[2], Tatiana Giraud[5,6]*, Yves Brygoo[13]

[±]These authors contributed equally to the work

[1] INRA, UMR1319 Micalis, F-78352 Jouy-en-Josas, France.
[2] Genomic Vision, 80-84 rue des Meuniers, 92220 Bagneux, France.
[3] AgroParisTech, UMR Micalis, F-78352 Jouy-en-Josas, France
[4] Origine, Structure, Evolution de la Biodiversité, UMR 7205 CNRS-MNHN, Muséum National d'Histoire Naturelle, CP39, 57 rue Cuvier, 75231 Paris Cedex 05, France
[5] Univ Paris-Sud, Ecologie, Systématique et Evolution, UMR8079, 91405 Orsay, France.
[6] CNRS, Ecologie, Systématique et Evolution, UMR8079, 91405 Orsay, France.
[7] LIMP Toulouse, INRA/CNRS, INRA, 24 Chemin de Borde Rouge – Auzeville, CS 52627, 31326 Castanet Tolosan Cedex, France
[8] INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, Castanet-Tolosan, F-31326, France.
[9] CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, Castanet-Tolosan, F-31326, France
[10] Univ Paris-Sud, Institut de Génétique et Microbiologie UMR8621, Orsay, France.
[11] CNRS, Institut de Génétique et Microbiologie UMR8621, Orsay, France.
[12] Univ Paris Diderot, Sorbonne Paris Cité, Institut des Energies de Demain (IED), 75205 Paris, France
[13] 13 ruelle d'Aigrefoin 78470 St Rémy-lès-Chevreuse

*Corresponding authors:

Pierre Renault: pierre.renault@jouy.inra.fr

Tatiana Giraud: Tatiana.Giraud@u-psud.fr

Abstract

While the extent and impact of horizontal transfers in prokaryotes are widely acknowledged, their importance to the eukaryotic kingdom is unclear and thought by many to be anecdotal, despite documented cases of gene transfers in eukaryotes reported over the last 20 years. We report here multiple recent transfers of a huge genomic island between *Penicillium spp.* found in the food environment. Sequencing of the two leading filamentous fungi used in cheese-making (*P. roqueforti* and *P. camemberti*) and comparison with the penicillin-producer *P. rubens* revealed a 575kb long genomic island — called *Wallaby* — in *P. roqueforti,* present as fragments at nonhomologous loci in *P. camemberti* and *P. rubens*, and with a nucleotide composition different from the rest of the genome. The screening of large *Penicillium* collections showed *Wallaby* to be present almost exclusively in strains associated with dairy environments. *Wallaby* encompasses about 250 predicted genes, some of which are probably involved in competition with microorganisms. The occurrence of multiple recent eukaryotic transfers in the food environment provides strong evidence for a historical and continuing evolutionary impact of this important, understudied and probably underestimated phenomenon in eukaryotes.

Humans have created many novel, nutrient-rich and homogeneous environments exerting strong selection pressures and inducing the rapid adaptation of microorganisms, such as fungal pathogens of crops, food spoilers and domesticated fungi used for fermentation in beverage or food (*e.g., Saccharomyces* for bread, beer and wine, *Aspergillus* for traditional fermented Asian foods, such sake, soy sauce and miso, and *Penicillium* for cheese and cured or fermented meat). These rapid adaptations in fungi provide excellent models for studying general processes of eukaryotic genome evolution, including the functional and ecological impact of horizontal gene transfer (Novo et al. 2009b) and changes in metabolism(Gibbons et al. 2012). Prokaryote-to-prokaryote transfers have been recognised as common and their associated impact important enough to raise questions about the possibility of reconstructing prokaryotic history through a tree of life or to change practices relating to antibiotic use in medicine. By contrast, horizontal transfers in eukaryotic species are still perceived by many to be isolated, sporadic events with a limited impact (for recent reviews, see refs(Keeling 2009)' (Andersson 2009)'(Syvanen 2012)). However, over the last 20 years or so, a number of cases of gene transfer in eukaryotes have been described. These documented cases include transfers of genetic material of prokaryotic origin into a eukaryote host (Alsmark et al. 2013) and transfers in man-made environments, *e.g.,* between yeasts used for wine fermentation(Novo et al. 2009b) or pathogenic fungi on crops (Mehrabi et al. 2011). Both the sizes and number of genes involved vary widely. Transfers of a single gene, a complete metabolic pathway (Slot & Rokas 2011), whole chromosomes (Ma et al. 2010) or even cases of the integration of almost complete genomes from bacterial endosymbionts into their eukaryotic hosts (Dunning Hotopp et al. 2007) have been described. Notable impacts of recently described horizontal transfers include key roles in land colonisation by plants (Yue et al. 2012), pigment production in spider mites through the acquisition of fungal carotenoid biosynthesis genes (Altincicek et al. 2012) and the emergence of plant diseases through transfers in fungi (Sanders 2006). Despite these examples, the lack of specific evolutionary trends in reported cases of lateral gene transfer in eukaryotes has led to the view of ancient, sporadic and isolated events with relatively little global impact on eukaryotic kingdoms, rather than a more frequently and widely occurring phenomenon. The frequency and importance of eukaryote-to-eukaryote gene transfer may, however, be underestimated.

Penicillium species are ubiquitous filamentous ascomycetes important to the biotechnology, biomedical and food industries. They commonly occur as food spoilage agents and opportunistic pathogens and are widely used as versatile cell factories. *Penicillium camemberti* and *P. roqueforti* are used as starter cultures for cheeses. *Penicillium camemberti,* used for the maturation of soft cheeses, such as Camembert, is the result of many selection programs aiming to improve the texture and colour of the conidia or physiological characteristics. *Penicillium camemberti* has never been

isolated from substrates other than dairy products. *Penicillium roqueforti* is widespread in food and also occurs in silage and natural environments. It is used as a starter culture in the production of most blue–veined cheeses (including Roquefort, Gorgonzola, Stilton and Danish Blue) and its abilities to tolerate cold temperatures, low oxygen concentrations, alkaline and weak acid preservatives, make it a common spoilage agent in refrigerated stored foods, meat products, rye bread and silage (Samson 2000; Pitt & Hocking 2009). To our knowledge, despite the importance of the filamentous fungi used in cheese-making, no genome sequence has yet been published for any of these sequences. The availability of these first two genome sequences will therefore provide a useful resource for improving our knowledge of edible cheese moulds, and for comparative genomics.

We sequenced and assembled the genomes of *P. roqueforti* and *P. camemberti* and compared them with two other available *Penicillium* genomes, for the penicillin producer and food spoilage agent *P. rubens*, previously known as *P. chrysogenum* (Berg et al. 2008), and the *Citrus* pathogen *P. digitatum* (Marcet-Houben et al. 2012a) (Table 1). We report a case of multiple, present-day horizontal transfers of a very large (over 500 kb) genomic island between several cheese fungi. These genomic island harbours about 250 genes, some of which are probably involved in competition with other micro-organisms. Beyond the potential conceptual and applied implications of recurrent horizontal transfers occurring in food, this finding indicates that horizontal gene transfer (HGT) may be more widespread and important than previously thought in eukaryotes.

Results

Detection and structural characterisation of a horizontally transferred genomic island

The global characteristics and comparisons of the genomes of *Penicillium camemberti*, *P. roqueforti*, *P. rubens* and *P. digitatum* are given in Table 1. The genomes have similar sizes and number of genes, with the exception of *P. digitatum*, which has fewer genes than the other three genomes. This smaller number of genes is thought to be the result of a streamlining process affecting the genome of *P. digitatum* due to its specialised plant pathogenic lifestyle (Marcet-Houben et al. 2012a). Assembly quality, as shown by the N50 metric and the number of scaffolds, is high for *P. roqueforti* and *P. rubens*. The initial *in silico* assembly for *P. roqueforti* has been experimentally validated and further improved (see below). The genome assemblies for *P. camemberti* and *P. digitatum* appear more fragmented.

Species	*P. camemberti*	*P. roqueforti*	*P. rubens*	*P. digitatum*
Strain	FM 013	FM 164	Wisconsin 54-1255	PHI26
Substrate	Cheese environment	Cheese environment	Mouldy cantaloupe	Citrus contaminant
Accession numbers EMBL	To be provided upon acceptance – deposition ongoing	To be provided upon acceptance – deposition ongoing	AM920416-AM920464	JH993687-JH993786
Authors	This study	This study	Van den Berg *et al.* 2008	Marcet-Houbet *et al.* 2012
Genome size (Mb)	34	28	32	26
Number of scaffolds	181	46 (combing assembly) 73 (initial assembly)	49	100
N50 (base pairs)	947,716	7,752,559 (combing) 2,448,491 (initial)	3,889,175	878,909
Maximum length (bp)	2,612,452	8,605,341 (combing) 5,324,254 (initial)	6,387,817	4,553,353
Minimum length (bp)	2,008	2,101	1,032	964
GC content	48.22	48.65	48.9	48.9
Number of protein coding genes	14,578	13,036	12,943	9,153
Mean gene length (bp)	1504.04	1812.71	1,515	1,387
Per cent genes with introns	77	81	83.5	67
Mean number of exons per gene	2.96	3.09	3	2.66

Table 1: Accession numbers and global characteristics of the genomes of *Penicillium camemberti*, *P. roqueforti*, *P. rubens* and *P. digitatum.*

Surprisingly, several scaffolds from *P. roqueforti, P. camemberti* and *P. rubens* had stretches of more than 5 kb, displaying 100% identity, in common (Supplementary Fig. S1), whereas mean pairwise identity was otherwise only 85 to 90% between genomes. *Penicillium digitatum* completely lacked these regions that were identical in the other *Penicillium* species. We investigated the nature of these shared sequences, by locating these regions accurately in the genome of *P. roqueforti* by improving assembly quality. The availability of a high-quality genome sequence in this clade will also improve the resolving power of comparative genome analysis in subsequent studies. For this purpose, we used a combination of PCR and molecular combing, a powerful FISH-based technique for the direct visualisation of single DNA molecules (Conti & Bensimon 2002; Lebofsky & Bensimon 2003). Molecular combing resulted in the successful mapping of scaffolds, including some separated by more than a hundred kilobases, onto single DNA molecules, thus constituting a new means of improving or experimentally validating *de novo* genome assemblies (Supplementary Fig. S2). It yielded an assembly in which 92% of the *P. roqueforti* genome was clustered into six superscaffolds of chromosomal size (Supplementary Fig. S3).

In the *P. roqueforti* FM164 strain, all the sequences found identical to both *P. rubens* and *P. camemberti* clustered together within a single 575 kb region accounting for 2% of the genome, which we called "*Wallaby*" (Figs. 1 and 2). This region lies within a 7.8Mb chromosome (Supplementary Fig. S3). The genomes of additional strains of *P. roqueforti* were examined by SOLiD® resequencing. Three

lacked the entire *Wallaby* region, whereas the fourth carried the very same *Wallaby* sequence as FM164.

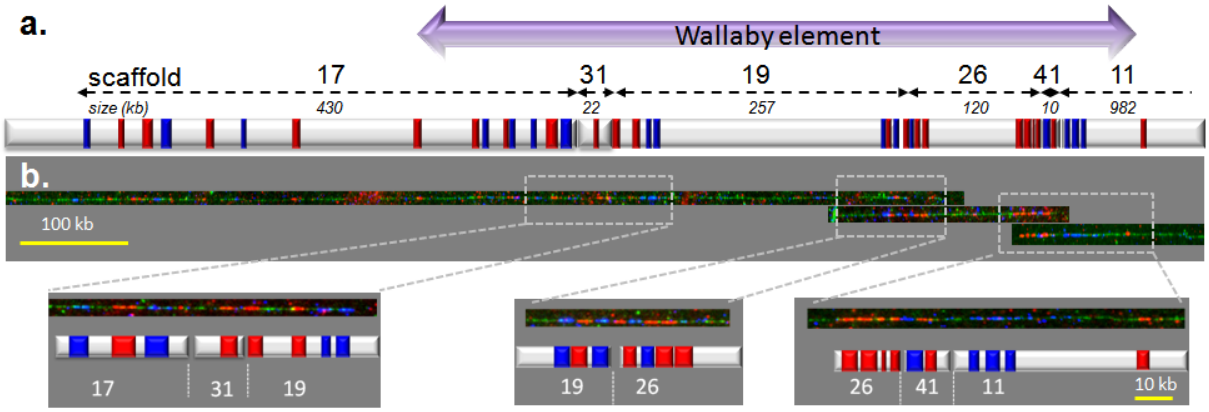


Figure 1: Structural characterisation of the *Wallaby* region in *Penicillium roqueforti*. A: *In silico* genomic Morse code for physical mapping, by molecular combing, of the six scaffolds bearing the *Wallaby* element (purple arrow). The numbers at the top indicate the scaffold names, with sizes in kilobases (kb) indicated below. B: Visualisation of the scaffolds mapped to the *Wallaby* locus. Grey boxes show a higher magnification of scaffold junctions.

These sequences were aligned for precise mapping of the *Wallaby* insertion point (Fig. 3). Some of the regions flanking *Wallaby* could be characterised in *P. rubens* and *P. camemberti* and were found to share 85 to 90% identity in the three species. In *P. roqueforti* strain FM164, all the sequences found identical to both *P. rubens* and *P. camemberti* sequences clustered together within a single 575 kb region accounting for 2% of the genome, which we called "*Wallaby*" (Figs. 1 and 2).

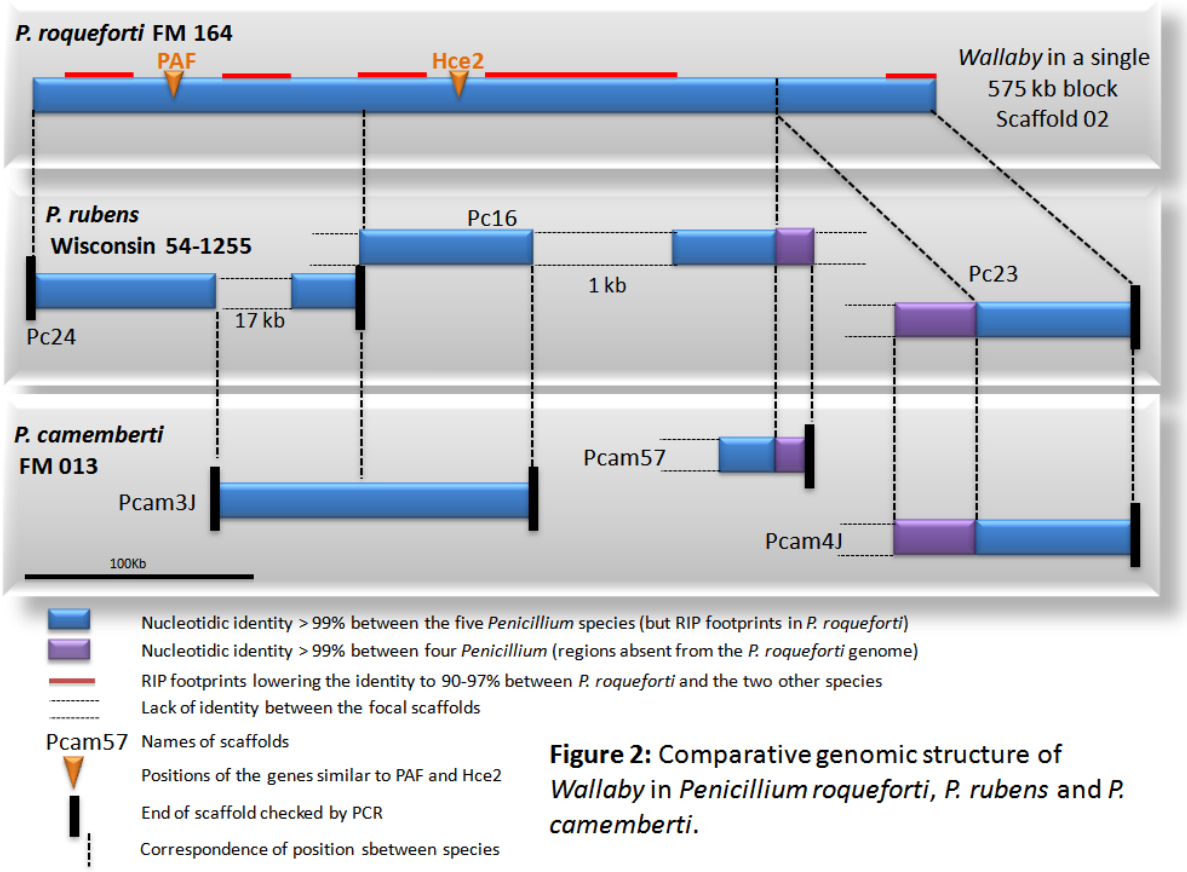


Figure 2: Comparative genomic structure of *Wallaby* in *Penicillium roqueforti*, *P. rubens* and *P. camemberti*. The sequence shown at the top corresponds to the *Wallaby* locus in *P. roqueforti* FM 164, *i.e.* from positions 1,487,035 to 2,061,670.

This region lies within a 7.8 Mb chromosome (Supplementary Fig. S3). The genomes of additional strains of *P. roqueforti* were examined by SOLiD® resequencing. Three lacked the entire *Wallaby* region, whereas the fourth carried the very same *Wallaby* sequence as FM164. These sequences were aligned for precise mapping of the *Wallaby* insertion point (Fig. 3, next page). Some of the regions flanking *Wallaby* could be characterised in *P. rubens* and *P. camemberti* and were found to share 85 to 90% identity in the three species. However, these sequences were located on other scaffolds than those carrying the *Wallaby* fragments in the latter two species (Fig. 2), indicating nonhomologous locations of *Wallaby* in the three species. The possibility of misassembly yielding these different locations was excluded by the successful PCR amplification of junction fragments overlapping the edges of the identical sequences in the three genomes. Furthermore, rearrangement events appeared to have occurred after the transfers in all three species. Indeed, the *Wallaby* sequences of *P. rubens* and *P. camemberti* lacked various fragments present in *P. roqueforti,* and both contained a 86-kb fragment absent from *P. roqueforti* (Fig. 2).

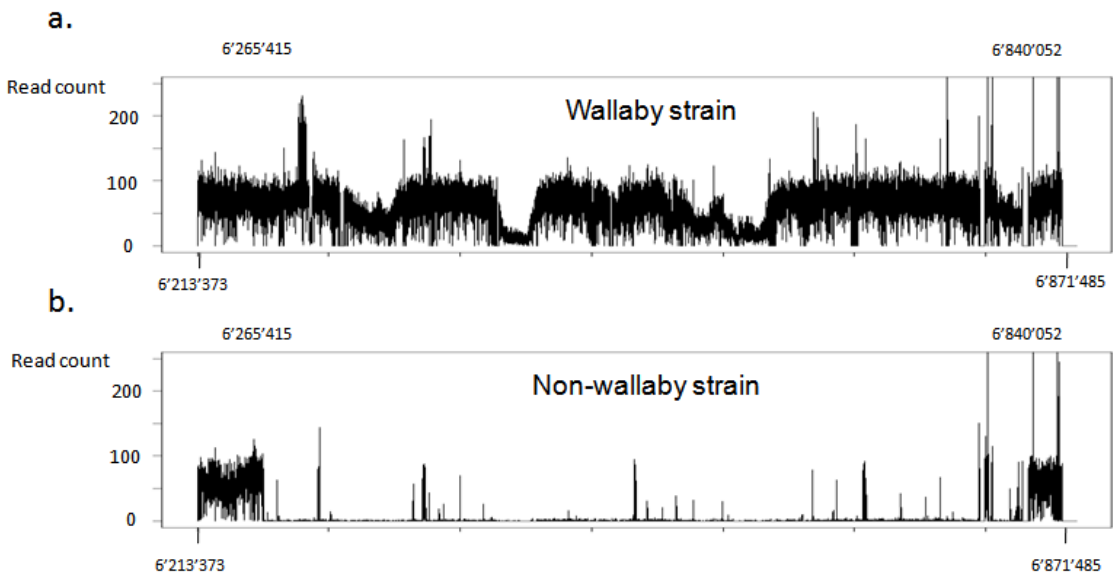


Figure 3: Read mapping of the genomes of two additional strains against the *Wallaby* locus of the *P. roqueforti* FM 164. Top panel: a strain bearing *Wallaby* isolated from cheese. Bottom panel: a strain not possessing Wallaby, isolated from an environment other than dairy products. This figure displays a count of the number of reads at each single base position. The X-axis represents genomic coordinates. The Y-axis scale represents the number of reads counted for every position. These are fixed dimensions and the scales are, therefore, the same for two different comparisons.

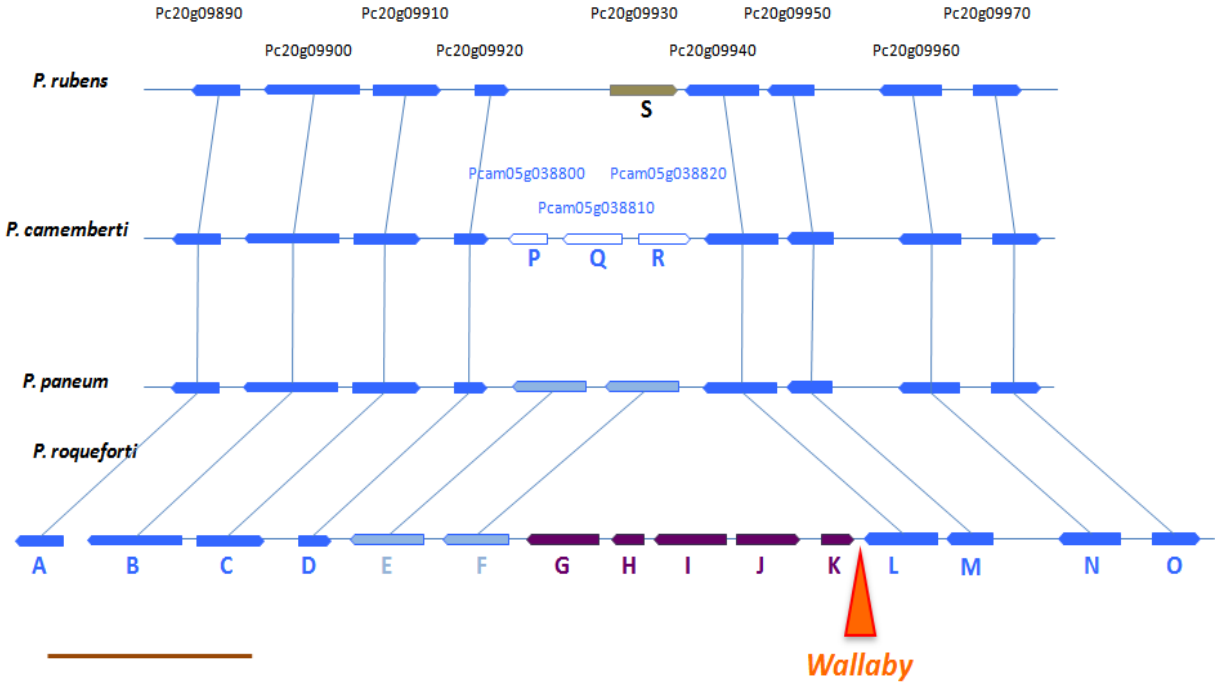
Wallaby displayed little similarity to sequences from public databases, even those of the well-studied *Aspergillus* genus, the closest relative of *Penicillium* (Berg et al. 2008). When Blast hits could be obtained, they matched fungal genomes (Supplementary Figure S4), indicating a fungal origin for *Wallaby*. A tetranucleotide composition analysis of *Wallaby* and its comparison with several fungal genomes revealed a nucleotide composition of *Wallaby* that seemed to be different from the rest of the genomes of *P. rubens, P. camemberti* and *P. roqueforti,* but still closer from these genomes than any other (Table 2 and Supplementary Fig. S5). The nucleotide composition of *Wallaby* also differed from that of other available fungal genomes, but was nevertheless closest to *Aspergillus* and *Penicillium* strains, suggesting that the donor species lie within this clade (Table 2, Supplementary Fig. S5).

.	Wallaby	Proqueforti WoWallaby	Prubens WoWallaby	Pcamemberti WoWallaby	Pdigitatum	Pcarneum	Aflavus	Akawachii	Aniger	Afumigatus
Proqueforti WoWallaby	0.917									
Prubens WoWallaby	0.921	0.989								
Pcamemberti WoWallaby	0.917	0.992	0.992							
Pdigitatum	0.776	0.914	0.913	0.927						
Pcarneum	0.876	0.994	0.979	0.983	0.920					
Aflavus	0.870	0.950	0.947	0.944	0.878	0.943				
Akawachii	0.869	0.934	0.930	0.931	0.857	0.922	0.967			
Aniger	0.848	0.930	0.922	0.922	0.860	0.923	0.970	0.995		
Afumigatus	0.885	0.936	0.932	0.929	0.849	0.922	0.963	0.960	0.959	
Aoryzae	0.872	0.950	0.948	0.945	0.880	0.943	1.000	0.967	0.969	0.963

Table 2: Pearson correlation coefficients calculated for the Z-scores of tetranucleotide frequency.

Overall, these results indicate that *Wallaby* has been recently and independently acquired, in at least some of these *Penicillium* species, via horizontal transfers. Other alternative explanations, such as introgression, can be excluded because of the non-homologous locations of the identical sequences in *P. rubens* and *P. camemberti*, precluding a parsimonious hybridisation hypothesis. The perfect identity of the sequences also argues for very recent transfer events. No synonymous mutations were found in *Wallaby*, except for some repeat induced point (RIP) mutation footprints in *P. roqueforti*, indicating that the presence of *Wallaby* is not an ancestral character. Pairwise comparison of the genome-wide distribution of 100% identical sequences between *P. roqueforti* and the three other genomes revealed no long stretches of sequences identical in all three genomes other than those in *Wallaby*. The only variation within *Wallaby* thus corresponded to repeat induced point (RIP) mutation footprints in the *P. roqueforti Wallaby* region (Fig. 2; Supplementary Fig. S6). In fungi, the RIP mechanism induces multiple C:G to T:A substitutions in repeated sequences during sex events (Braumann et al. 2008; J Ropars et al. 2012). As a consequence, sequence identity in *Wallaby* fragments with RIP footprints dropped to 90 - 97% between *P. roqueforti* and the other two *Penicillium* sequences, but this concerned exclusively RIP C:G to T:A substitutions. This is consistent with the occurrence of RIP after the transfer event in *P. roqueforti*.

The region flanking *Wallaby* in *P. roqueforti* may be a hotspot for DNA insertions, as other unrelated fragments appear to be inserted at this locus in other *Penicillium* genomes (Fig. 4). However, no footprints of duplication or transposable elements were found around the *Wallaby* insertion points for which flanking regions could be identified.



Genes	Protein length	IPR annotation
A	401aa	Protein of unknown function
B	757aa	Autocrine motility factor receptor
C	555aa	Transporter protein
D	236aa	Hypothetical protein
E	264aa	Hypothetical protein
F	176aa	Hypothetical protein
G	574aa	Phosphate transporter
H	133aa	Helix-loop-helix domain
I	564aa	Carboxypeptidase
J	768aa	Oligopeptide transporter
K	459aa	Nucleotidyltransferase
L	526aa	Cytochrome P450
M	353aa	Methyltransferase
N	489aa	Motif in proteasome subunits
O	313aa	Dehydrogenase
P	336aa	Alcohol dehydrogenase
Q	606aa	Unknown function
R	558aa	Cytochrome P450
S	480aa	Oxydase

Figure 4: Left: Schematic representation of the flanking regions of *Wallaby* in *Penicillium roqueforti* compared with homologous regions in other *Penicillium* species. Predicted genes are indicated, and those for which homology could be detected in this region in other species are linked and blue. White predicted genes have no detectable similarity to sequences in this region from other species. *Wallaby* in *P. roqueforti* is indicated by a red triangle. The *P. camemberti* and *P. rubens Wallaby* regions are fragmented and located in non-homologous regions and are therefore not shown on this figure. *Penicillium paneum* does not carry *Wallaby.* Bottom panel: Putative functions of the genes located in the flanking regions of *Wallaby* in *P. roqueforti.*

Wallaby has been transferred to several other strains and species

The presence of *Wallaby* was further investigated in 441 strains by PCR amplification with *Wallaby*-specific primers, designed from single-copy putative coding sequences, each amplifying 1 kb (Supplementary Table 1). We carried out an exhaustive screening of all the terverticillate *Penicillium* species, the clade to which *P. roqueforti, P. camemberti* and *P. rubens* belong, from the public collection of the MNHN, as well as subsets of additional collections, including strains isolated from various ecological niches, corresponding to 51 morphospecies (Fig. 5, Supplementary Table S2). Amplicons were obtained for all *P. camemberti* strains and a fraction of the strains from *P. roqueforti* and from the *P. chrysogenum* - *P. rubens* clade. The amplicons were sequenced and all were identical to *Wallaby* sequences. The lack of non-synonymous substitutions in all these species further indicates that the presence of *Wallaby* is not an ancestral character. Several strains from species closely related to *P. camemberti* and present in dairy environments or occurring as food contaminants, such as *P. caseifulvum, P. biforme, P. fuscoglaucum, P. palitans, P. solitum, P. nordicum* and *P. polonicum,* for example, also gave amplicons (Fig. 5; Supplementary Table S2). No amplicons were obtained from *P. roqueforti s*trains isolated from other environmental niches or from *Penicillium* species not associated with cheese environments (*e.g., P. carneum, P. expansum, P. cavernicola*), other than *P. chrysogenum s. l., i.e.,* the species complex encompassing both *P. chrysogenum* and *P. rubens* (Fig. 5; Supplementary Table S2).

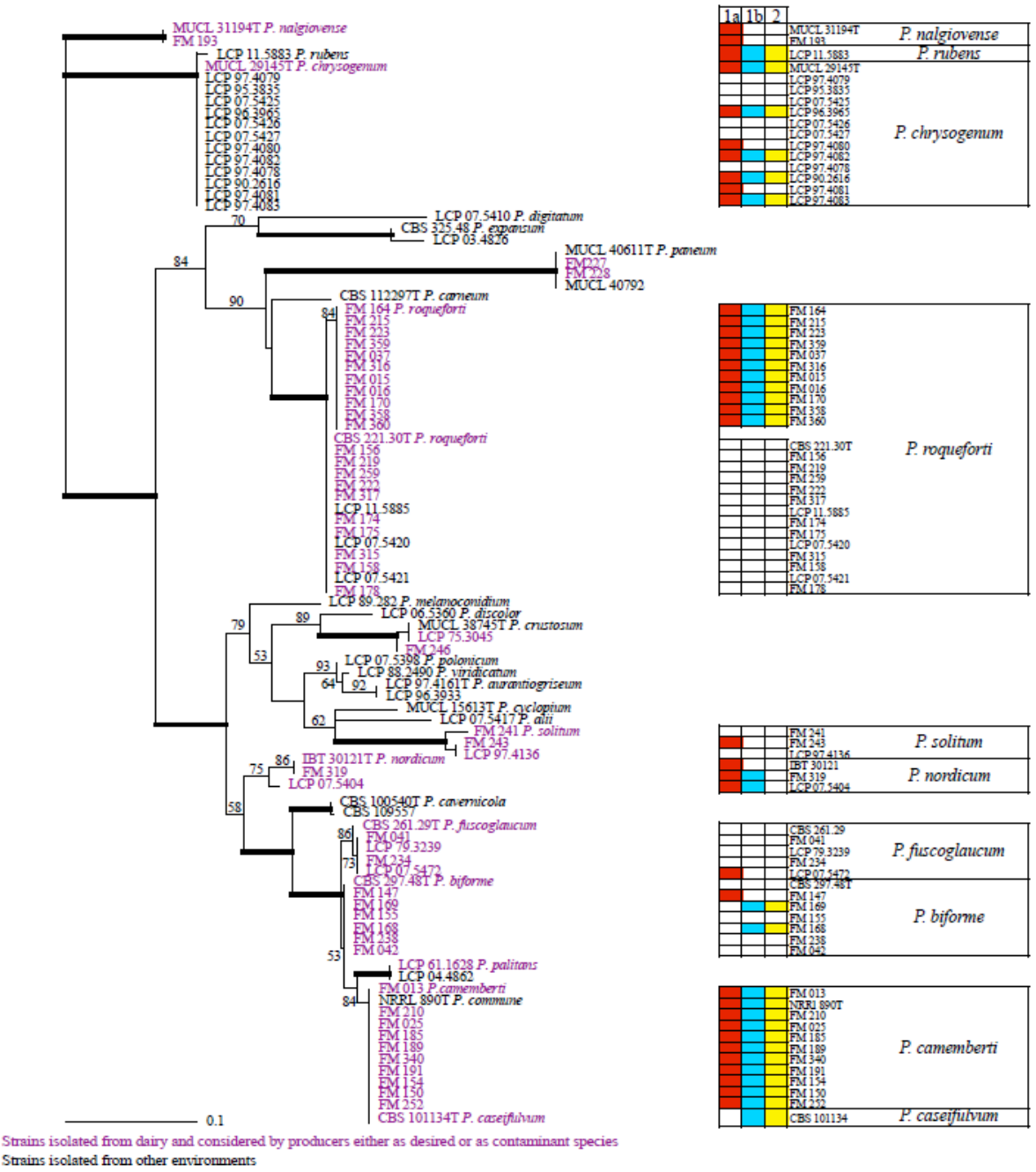


Figure 5: **Gene genealogy for the *Penicillium* strains screened by PCR for the presence of *Wallaby* and for which DNA sequences of the markers could be obtained.** *Penicillium* species associated with the cheese environment are indicated in purple. Colours in the pseudotable indicate positive amplification for the three primer pairs (red, blue and yellow, respectively) for single-copy predicted genes in *Wallaby*. See Supplementary Table S2 for the complete screening results of the collection.

The transferred region is probably involved in competition

In *P. roqueforti, Wallaby* was predicted to contain 248 genes, 149 of which were covered by expressed sequence tags. No expanded gene families were detected within *Wallaby.* Few genes could be annotated (see Supplementary Figure S7 for the gene ontologies recovered) and Fisher's exact tests indicated no significant enrichment of any particular function. Interestingly, some of the annotated genes were predicted to be involved in the regulation of conidiation (spore production) or in antimicrobial activities, suggesting functional advantages for *Wallaby*-bearing strains associated with competition in the cheese, which contain many other micro-organisms.

In particular, among expressed genes within *Wallaby*, afp (gene ID : ProqFM164S02.2755) encodes a protein identical to PAF *(Penicillium* antifungal protein*),* with 100% nucleotide sequence identity to the *paf* gene of *P. rubens*; PAF has been experimentally shown to be cytotoxic to fungi(Kaiserer et al. 2003) and to regulate spore production (Hegedüs et al. 2011). Another expressed protein with putative antimicrobial activities, ProqFM164S02g002866 (gene ID : ProqFM164S02.2870), matched to Hce2 (Stergiopoulos et al. 2012), which has an Ecp2 domain — originally identified as a virulence factor in *Cladosporium fulvum* — fused to a GH18 chitinase domain very similar to the α subunit of the yeast killer toxin zymocin from the dairy yeast *Kluyveromyces lactis* (Jablonowski & Schaffrath 2007). Hce2 proteins with such an architecture are thought to play a role in antagonistic interactions with other microorganisms (Stergiopoulos et al. 2012).

Discussion

The sequencing of the first two genomes of food *Penicillium* strains will provide an invaluable resource for comparative genomics of fungi, nicely complementing the 1000 fungal genomes project and the JGI fungal genome initiative, in which no *Penicillium* from the food environment is being sequenced (Anon n.d.)′ (Anon n.d.). We also report here the development and application of a methodology for improving and validating genome assemblies based on an original single-DNA molecule technology, molecular combing. Improving genome assembly, notably by the use of physical maps, has been largely advocated (Lewin et al. 2009; Ricker et al. 2012), and is even deemed crucial for comparative genomics (Alkan et al. 2011).

Our data indicate that food *Penicillium* strains contain regions, grouped at a single location and called *Wallaby* in *P. roqueforti,* that have very recently undergone multiple horizontal transfers. Several lines of evidence support the acquisition of *Wallaby* by horizontal transfer and rule out alternative hypotheses: 1) An ancestral presence of *Wallaby* is refuted by the complete identity (except for RIP mutation footprints in some fragments in *P. roqueforti*), even in non-coding regions, over more than 500 kb, between distant species with genome sequence identities otherwise of about 90%, and by the absence of *Wallaby* from all fungal species other than those in which it is identical, 2) the non-homologous locations rule out introgression, 3) the presence of *Wallaby* almost exclusively in species from the food environment and the different nucleotide composition of *Wallaby* from the rest of the recipient genomes provide further support for horizontal transfer. The nucleotide composition of *Wallaby* suggests that the donor species may belong to the *Aspergillus* or *Penicillium* clade. The function of one of the genes contained in *Wallaby,* experimentally demonstrated in *P. rubens* (Kaiserer et al. 2003)*,* suggests that it may confer adaptation associated with competition with other micro-organisms.

The occurrence of gene transfers in fungi has been reported before (Rosewich & Kistler 2000; Fitzpatrick 2012; Novo et al. 2009b; Ma et al. 2010), but this case of HGT is exceptional in several ways. The *Wallaby* genomic islands are unusually large, covering 2% of the host genomes. Furthermore, these sequences were found to be 100% identical in the species screened, with the exception of the RIP mutation footprints in *P. roqueforti*. The complete identity of the transferred sequences and the lack of corresponding sequences in databases prevent phylogenetic reconstruction of the history of transfers. However, we can hypothesize that the original donor had a complete island in a single block, as in *P. roqueforti,* and that this block was then transferred to other acceptor genomes, in which it was fragmented and may then have been transferred to other species. Some of the *P. rubens* scaffolds containing *Wallaby,* Pc23 and Pc24 have already been shown to be absent from *P. digitatum* (Marcet-Houben et al. 2012a). This study also showed that several fragments of these scaffolds matched other dispersed elements in the genome of *P. rubens*, with an identity of about 90%. It has been suggested that this level of similarity indicates gene family expansion rather than horizontal transfer (Marcet-Houben et al. 2012a). Our analysis of additional sequenced genomes revealed that these dispersed elements were specific to *P. rubens* and that horizontal transfers occurred in addition to possible gene family expansions.

The widespread occurrence of *Wallaby* in cheese species, including species never found in other environments, such as *P. camemberti,* and the identity of the sequences indicate that these transfers occurred in cheese and were, therefore, very recently promoted by human activities. Such horizontal transfer may be facilitated by the ability of fungi to form anastomoses — somatic fusions between mycelia known to occur in *Penicillium* species (Haas et al. 1956; Gabriela Roca et al. 2005) — and may facilitate the transfer of whole chromosomes in fungi grown in laboratory environment (Ma et al. 2010). Functions brought by the transferred material may, therefore, be beneficial in the food environment, as suggested by the functions identified in *Wallaby,* involved in antagonistic interactions with other micro-organisms (Stergiopoulos et al. 2012).

This study provides strong support for the view that horizontal transfers of large genomic regions play an important role in fungal adaptation to new environments (Neafsey et al. 2010; Ellison et al. 2011), including those created by humans. The example reported here is, however, unique to date among transfers, in terms of the size of the region transferred, its total identity between species and the number of species involved. Horizontal gene transfer in fungi may be beneficial, particularly in cheese-making strains or wine yeasts, in which large transfers of genes promoting fermentation have been detected (Novo et al. 2009b). The data presented here, particularly when considered in the light of the growing list of reports of horizontal gene transfer between eukaryotes, suggest that this phenomenon may occur frequently and may have a wider impact than previously thought (as already advocated (Syvanen 2012; Keeling 2009)). We foresee major implications for the management of pathogen species in the face of current changes, including globalisation, which may bring previously isolated species into contact. Rapid evolution by horizontal gene transfer may also have a major impact through the emergence of new diseases caused by fungi acquiring new virulence genes horizontally (Slot & Hibbett 2007; Friesen et al. 2006; Ma et al. 2010; Slot & Rokas 2011; Sanders 2006). Furthermore, in itself, this finding indicates that horizontal gene transfer can occur within the food chain. The frequency of this phenomenon should be investigated more thoroughly, as it may have a potential impact on food, agricultural and biotechnological practices.

Materials and Methods

Genome sequencing

DNA was extracted with the cetyl-trimethyl-ammonium bromide (CTAB) extraction procedure (Rogers & Bendich 1985; Rogers & Bendich 1988; Rogers & Bendich 1994). Mycelium (10 g) was recovered from ~ 10 Petri dishes and frozen with liquid nitrogen in a mortar. The mycelium was ground to a powder and 10 ml of 2xCTAB buffer was added. The frozen mixture was returned to room temperature and processed as previously described (Rogers & Bendich 1985; Rogers & Bendich 1988; Rogers & Bendich 1994). The DNA pellet was dried, and resuspended in 4 ml of 10 mM tris-hydroxy-methyl-aminomethane (TRIS) and 1 mM ethylene-diamine-tetra-acetic acid (EDTA) (pH 8.0). Nuclear DNA was purified and excess mitochondrial DNA was eliminated by centrifugation on a caesium gradient with 4', 6-diamidino-2-phenyl-indole (DAPI) (Kapuscinski 1995). We added 1.15 g solid caesium chloride and 4 µl of DAPI solution (1 mg/ml) per ml of DNA solution (final density: 1.65 g/ml). The DNA/CsCl/DAPI solution was centrifuged in a Quick-Seal centrifuge tube (part # 342412, Beckman, Palo Alto, USA) at 50,000 rpm, for 12-15 hours, in a Beckman VTi65 vertical rotor. Two bands of DNA were visible under UV light. The lower band (nuclear DNA; the upper band corresponds to mitochondrial DNA) was collected with a syringe (1.2 x 40 mm needle); DAPI was extracted with isoamyl alcohol saturated with CsCl (1.15 g/ml) and the DNA solution was dialyzed for 4 days against 10 mM TRIS, 1 mM EDTA (pH 8.0) with twice-daily replacement of the dialysis solution.

Sequencing and annotation

Penicillium camemberti FM 013 and *P. roqueforti* FM 164 were sequenced by Genoscope (Evry, France), via the 454 sequencing of an 8 kb mate pair library (793786 and 661034 cleaned read pairs, respectively) combined with Illumina Solexa sequencing (34126183 and 34253031 76 nt single-end reads, respectively). Assembly was performed with SOAPdenovo v1.05 and velvet v1.1.04 (Zerbino & Birney 2008). SOAPdenovo was run (kmer values 43 to 67) to generate contigs. Velvet was then run with combined raw reads and SOAPdenovo contigs (parameters "-cov_cutoff 5 -min_contig_lgth 100 -max_divergence 0.05 -long_mult_cutoff 1 -exp_cov auto"). The range of kmer values for Velvet was 41 to 57. The short-reads assembly with the maximum N50 value was used as input for the scaffolding process (SOAPprepare and SOAPdenovo). For *P. roqueforti*, maximum N50 values were obtained with a kmer value of 37 and a minimum of six links between contigs (parameter (default=5) pair_num_cutoff=6). For *P. camemberti*, a kmer value of 27 and a minimum of 14

links gave the highest N50 value.

Gene models were predicted with EuGene (Schiex et al. 2001; Foissac et al. 2008), a highly integrative eukaryotic protein-coding gene prediction platform. This gene predictor requires training: a dataset of 442 curated gene models was built and split into three homogeneous, independent datasets. The first was used for training, the second for parameter optimisation and the third for performance evaluation. We used 780,051 ESTs from *P. roqueforti* for genome annotation (unpublished data), together with the Swiss Protein database (February 2011), the proteomes of *Saccharomyces cerevisiae* and *Penicillium chrysogenum* Wisconsin 54-1255, a database of all Eurotiales proteins in GenBank (February 2011). Transposable elements were identified with REPET (Quesneville et al. 2005). InterPro was used to identify protein domains and families.

The *Penicillium roqueforti* FM164 draft genome was assembled into 73 scaffolds of over 2 kb, spanning 28 Mb. The *Penicillium camemberti* FM013 draft genome spans 34 Mb, assembled into 140 scaffolds of over 2 kb.

Fungal isolates, single-spore isolation, DNA extraction, PCR amplification and sequencing for *Wallaby* and phylogenetic analysis

We used 441 isolates in total (Supplementary Table 2). The 124 isolates provided by producers of starter cultures and cheeses were labelled FM and their origins are confidential. All were isolated from the cheese environment and belong to four distinct ascomycete genera: *Penicillium, Fusarium, Scopulariopsis* and *Sporendonema* (Jeanne Ropars et al. 2012), We analysed all 241 terverticillate penicillium strains from the public collection of the MNHN, and some *Eupenicillium* species. We also analysed the LUBEM-Brest collection encompassing 76 *P. roqueforti* isolates, labelled with an F. These isolates were obtained from blue cheeses from 14 countries. The numbers after the “F” correspond to individual cheeses. For the 39 cheeses, morphologically different strains were treated as different strains, labelled by a number following that identifying the cheese (e.g. F17.1 and F17.2 are two strains from cheese 17). Single-spore isolation was systematically performed by the dilution method, after growth for 3-5 days at 25°C on malt agar. The “F” strains were also obtained by spore dilution.

Genomic DNA was extracted from fresh mycelium of the isolates listed in Supplementary Table 2. Mycelium was obtained after 3–5 days on malt agar for *Penicillium*, *Scopulariopsis* and *Fusarium* strains and on a confidential medium provided by starter producers for *Sporendonema casei*. The Qiagen DNeasy Plant Mini Kit (Qiagen, Ltd. Crawley, UK) was used for DNA extraction.

PCR was performed in a volume of 50 µl, containing 25 µl template DNA, 1.25 U AmpliTaq DNA polymerase (Roche Molecular Systems, Inc., Branchburg, NJ, USA), 5 µl 10 x *Taq* DNA polymerase buffer, 5 µl 50% glycerol, 2 µl 5 mM dNTPs, 2 µl of each 10 µM primer and 50–100 ng template DNA. Strains were identified with the 5' end of the β-tubulin gene, with primers Bt2a/Bt2b(Glass & Donaldson 1995). The three primer pairs designed to detect *Wallaby* are shown in Supplementary Table 1. DNA fragments PC4 and PC13(Giraud et al. 2010) were used for the phylogenetic analysis. Amplifications were performed with 30 cycles of 30 s at 95°C, 30 s at 58°C for the three *Wallaby* primers, amplifying 1 kb each, and 2 min at 72°C. For microsatellite loci, the thermal regime was 35 cycles of 30 s at 94°C, 30 s at 50°C and 30 s at 72°C. All PCR programs had a final 5 min extension step at 72°C. PCR products were purified and sequenced by Genoscope (Évry, France).

Protoplast preparation for molecular combing

A *P. roqueforti* conidial suspension was plated on M3 medium (0.25 g/l KH_2PO_4, 0.3 g/l K_2HPO_4, 0.25 g/l $MgSO_4$, 0.5 g/l urea, 0.05 mg/l thiamine, 0.25 µg/l biotin, 2.5 mg/l citric acid, 2.5 mg/l $ZnSO_4$, 0.5 mg/l $CuSO_4$, 125 µg/l $MnSO_4$, 25 µg/l boric acid, 25 µg/l sodium molybdate, 25 µg/l iron alum, 5 g/l glucose, 25 mg/l chloramphenicol in seven Petri dishes and incubated four days at 25°C. Conidia were harvested and resuspended in 20 ml M3 medium. Two 50 ml M3 medium liquid cultures were inoculated with the conidial suspension. Germination occurred after 18 hours of incubation at 25°C, with shaking at 90 rpm. The germinating conidia were harvested by centrifugation (10 min, 2640 x *g*). Dried mycelium (1 g) was suspended in the protoplast isolation solution (400 mg Filtrozym (Laffort), 20 mg bovine serum albumin in 10 ml of TRF1 solution (1.2 M $MgSO_4$, 10 mM orthophosphate, pH 5.8)). The suspension was incubated at 30°C for 120 -150 minutes, with shaking at 90 rpm. Once the cell walls had been lysed, we transferred the 10 ml suspension to a 30 ml glass tube and overlaid it with 10 ml TRF2 (0.6 M sorbitol, 100 mM Tris-HCl pH 7.5). After centrifugation for 10 minutes at 2869 x *g* at room temperature, the protoplasts formed a layer at the interface. The protoplast layer was removed and washed with an equivalent volume of TRF3 (1 M sorbitol, 10 mM Tris-HCl pH 7.5), by centrifugation at 2640 x *g* for 10 minutes. The protoplast pellet was suspended in 1 ml TRF4 solution (1 M sorbitol, 10 mM Tris-HCl pH 7.5, 10 mM $CaCl_2$).

Molecular combing

Molecular combing and suitable hybridisation, detection and scanning procedures were performed essentially as previously described (Cheeseman et al. 2012). The major modifications made were the addition of a protoplast isolation step before plug embedding (see above). The PCR amplicons were purified and used directly as templates for the labelling reaction, rather than after subcloning (primer sequences available on request). The combed DNA was counterstained after the fluorochrome detection step, by incubating the coverslip in 30 µL of a 1:1000 YoYo-1 solution in milliQ water for 30 seconds and then washing three times, for three minutes each, in milli-Q water.

Genomic Morse codes (GMC) were designed at the extremities of scaffolds. A GMC is a signal generated by a set of specifically designed probes. A GMC consists of an alternating series of dots, dashes (probes) and gaps (region between two probes) of different sizes and colours, designed to generate an unambiguous pattern at a specific locus, thus physically mapping the region (Lebofsky et al. 2006). GMCs at the extremities of scaffolds generate unambiguous patterns making it possible to locate and orient scaffolds accurately on single-stretch DNA molecules, by combining two Morse codes in cases of neighbouring scaffolds. In the case of distant scaffolds, the distance separating the scaffolds is measured based on the distance separating the two genomic Morse codes (Supplementary Fig. 2).

Phylogenetic analysis

Genealogical relationships were inferred using the flanking regions of the two microsatellite loci PC4 and PC13 (classical loci like the β-tubulin gene are not variable enough). Sequences were manually aligned, with BioEdit (Hall 1999). Independent phylogenetic trees were built with TreeFinder (Jobb et al. 2004) under a maximum likelihood framework. A nucleotide substitution model (GTR+G) was inferred with jmodeltest (Posada 2008). As the two topologies were congruent, we concatenated the datasets to obtain a single phylogenetic tree based on 363 bp. Branch support was determined from a bootstrap analysis of 1000 resampled datasets.

Repeat induced point mutation analysis

Using fragments of the *Wallaby* sequences of *Penicillium roqueforti* FM 164, *P. camemberti* FM 013 and *P. rubens* Wisconsin 54-1255, we searched for RIP-like footprints (Galagan & Selker 2004). Multiple sequence alignments were built, using ClustalW with default settings (Thompson et al. 2002). RIP mutation-like footprints were sought with RIPCAL software (Hane & Oliver 2008) (Supplementary Fig. S6).

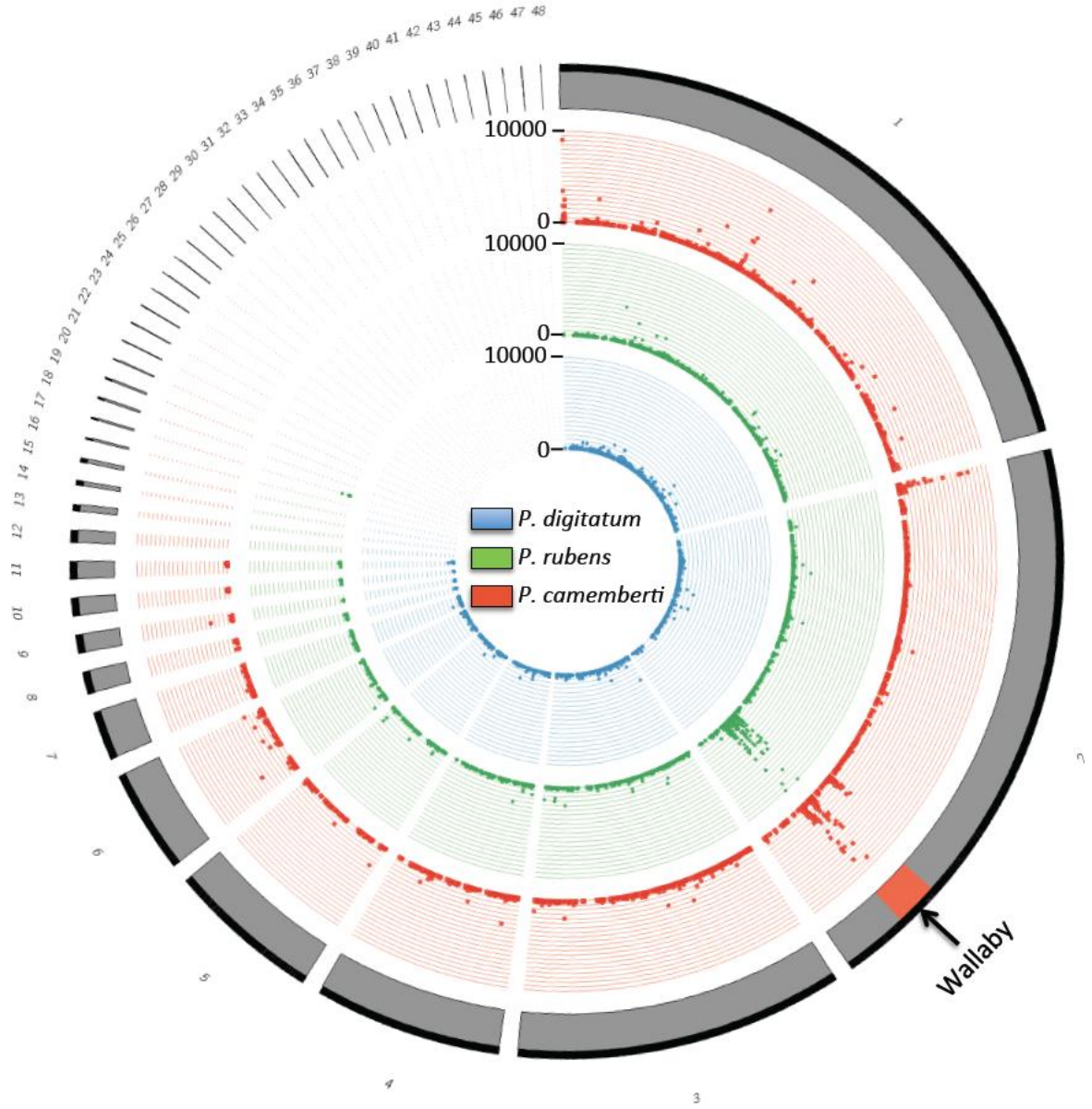
Tetranucleotide composition

Correlation between Z-score of tetranucleotide composition were assessed using jpsecies version 1.2.1[61].

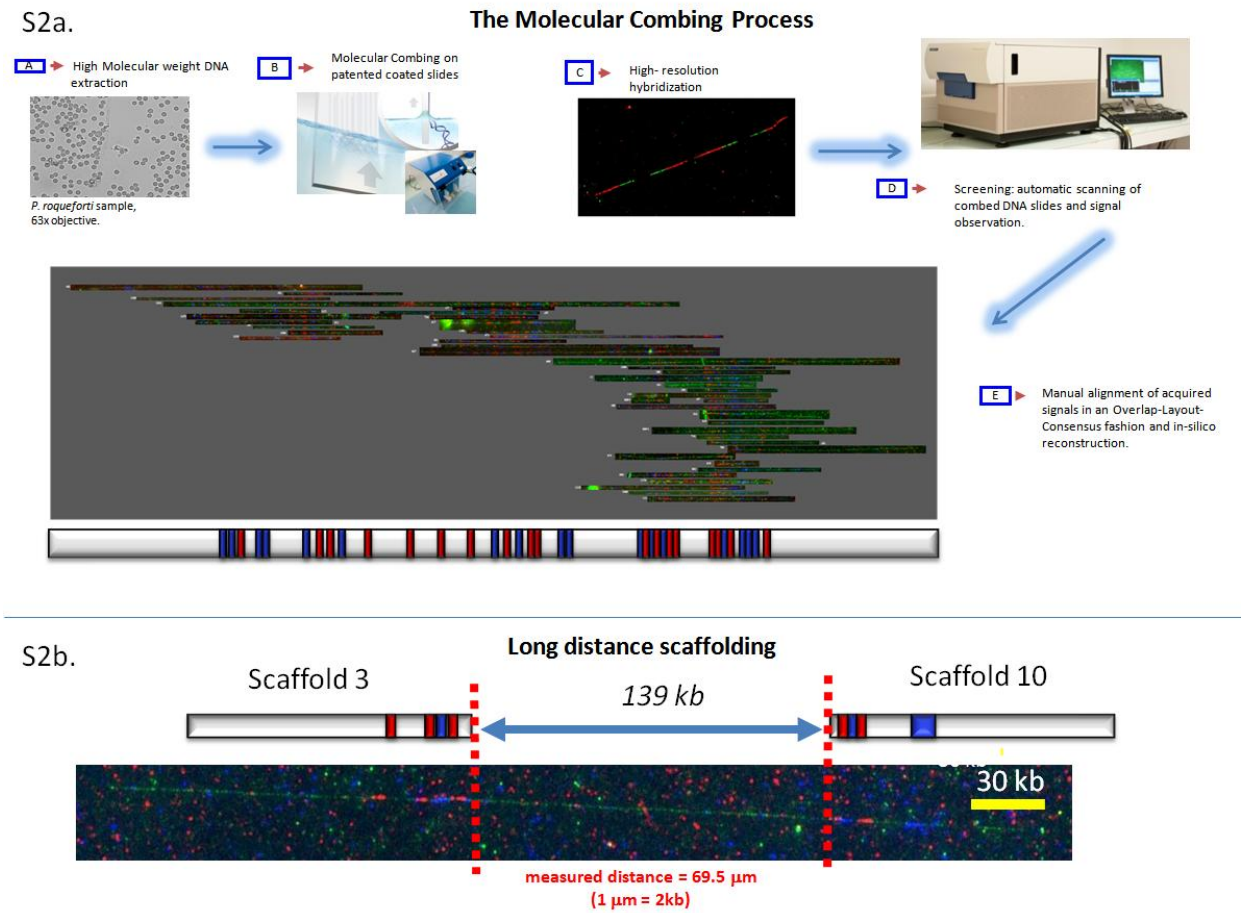
Acknowledgements This work was supported by the ANR grant “Food Microbiomes” (ANR-08-ALIA-007-02) coordinated by P.R. Molecular combing experiments were supported by GenomicVision. K.C. and J.R. received joint CIFRE grants from ANRT and GenomicVision, and ANRT and SPPAIL (*Syndicat Professionnel des Producteurs d'Auxiliaires pour l'Industrie Laitière*), respectively. A.B. received an ANR FROMA-GEN grant ANR-12-PDOC-0030. T.G. received an ERC starting grant, GenomeFun 309403. We thank the LUBEM laboratory of Brest (in particular, Emmanuel Coton and Monika Coton) for permission to use their strains (Supplementary Table S2c). We thank Ricardo Rodriguez de la Vega for invaluable help with the genomic analyses.

Supplementary Material to Chapter6:

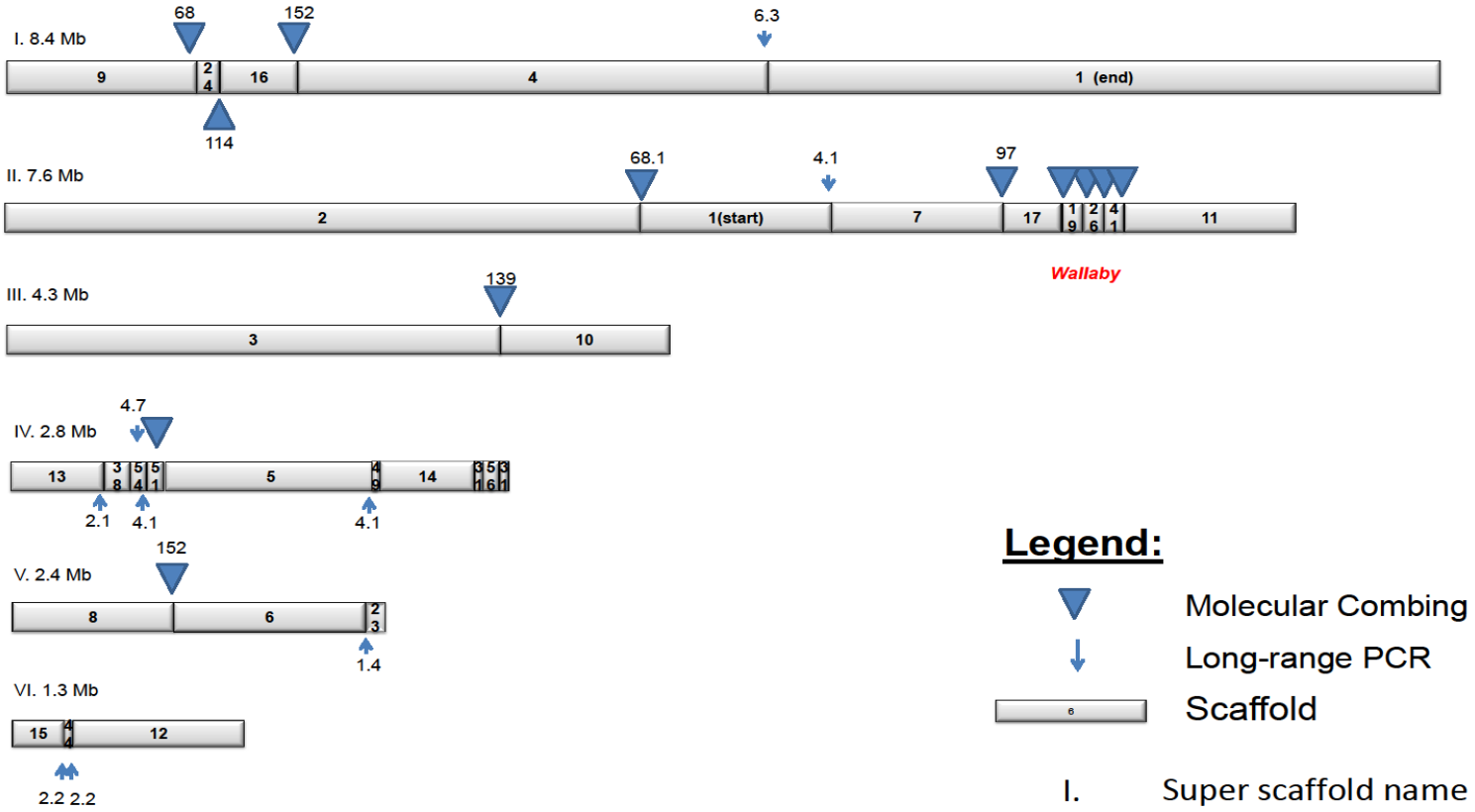
The following pages presents the supplementary data of the paper.



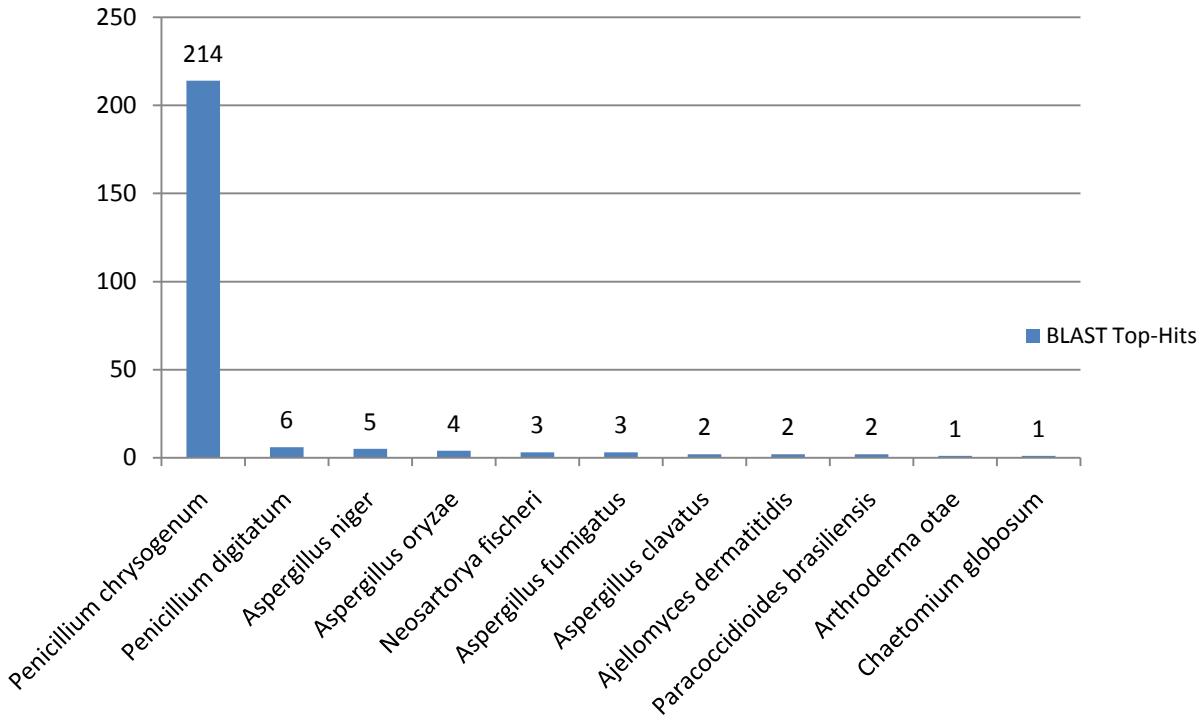
Supplementary Figure S1: Length of consecutive identical sequence (Y axis) between the *Penicillium roqueforti* genome and other *Penicillium* genomes positioned on each *Penicillium roqueforti* scaffold (X circular axis). Red: identity between *P. camemberti* and *P. roqueforti*; green: identity between *P. rubens* and *P. roqueforti*; blue: identity between *P. digitatum* and *P. roqueforti*). Wallaby contigs (11, 17, 19, 26, 31 and 41) are shown in red and other contigs in grey.



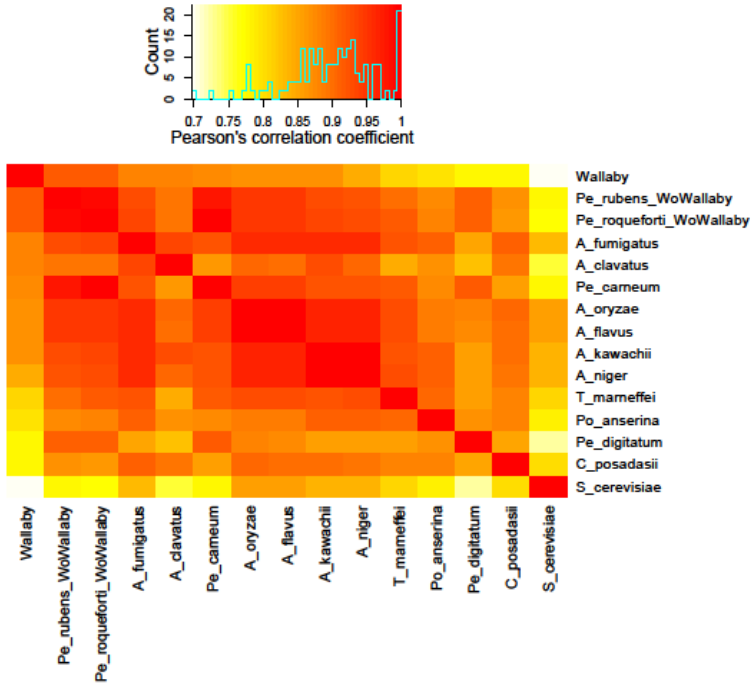
Supplementary Figure S2: Molecular combing assisted assembly method: a. Physical mapping of entire scaffolds by molecular combing: after a specific high-molecular weight DNA extraction (A), DNA is combed on specially treated glass surfaces (B). Hybridisation of genomic Morse codes (C) and subsequent automatic scanning (D) results in a collection of combed DNA fibers with genomic Morse code signals. These signals are aligned in an overlap layout consensus approach to construct a supermolecule spanning several scaffolds (E). b. Long-distance scaffolding. Example of molecular combing assisted mapping covering a long distance; 69.5 µm, corresponding to 139 kb, separates scaffold 10 and scaffold 3. Molecular combing allows broader scaffolding than long-range PCR.



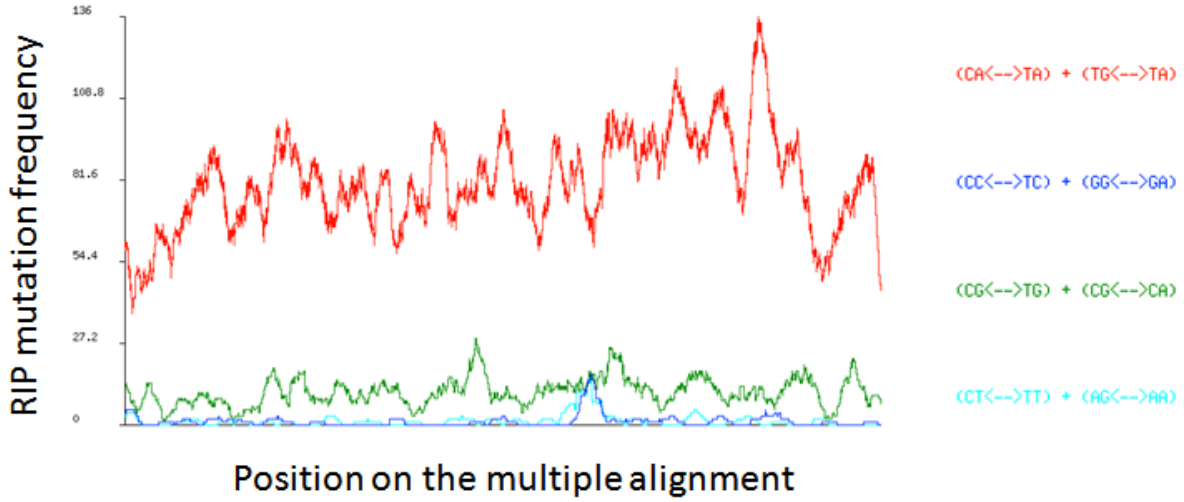
Supplementary Figure S3: Genomic overview of *Penicillium roqueforti*. Molecular combing and PCR experiments allowed the construction of six superscaffolds with sizes of between 1.3 and 8.4 Mb, similar to chromosomes found in other *Penicillium* species. Eventually, 92% of the total genome could be clustered into six superscaffolds, through 11 scaffold junctions confirmed by PCR and 15 observed by molecular combing.



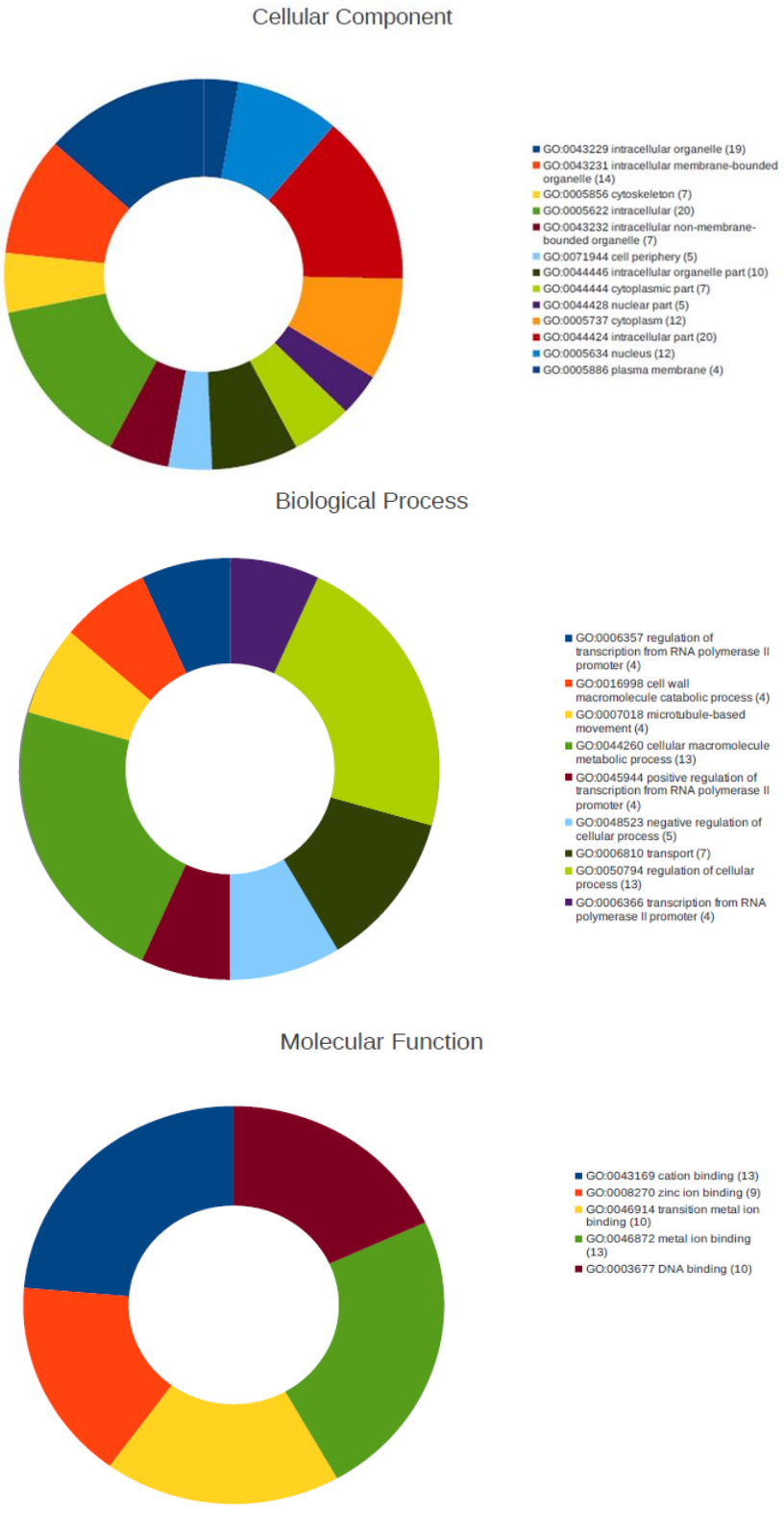
Supplementary Figure S4: Number of BLAST top-hits for the 248 *Wallaby* genes by species.



Supplementary Figure S5: Heatmap of pairwise Pearson's coefficients for the correlation between tetranucleotide Z-scores of different fungal genomes. The scores are indicated by colour, as in the legend. The matrix is sorted according to the correlation with the tetranucleotide content of Wallaby. Pe: *Penicillium*; A_: *Aspergillus*; T_: T*alaromyces*; Po_: *Podospora;* C_: *Coccidioides*; S_: *Saccharomyces*; WoWallaby: without the Wallaby sequence.



Supplementary Figure S6: RIP mutation frequency plotted over a sliding sequence window, corresponding to the alignment (not shown). Nucleotide polymorphisms (against the alignment consensus, which is also the sequence with the highest GC-content) correspond to CpA<- ->TpA or TpG<- ->TpA (red curve), as expected for RIP substitutions.



Supplementary Figure S7 Gene ontologies obtained for the predicted genes (out of 248) in Wallaby for which reliable results were obtained. The pie chart and legend show the number of genes (in brackets in the key) corresponding to a given ontology; a) Cellular Component annotation (24 annotated genes); b, Biological Process (35 annotated genes); c, Molecular function (43 annotated genes).

Supplementary Table 2 (next pages): Presence/absence of *Wallaby* in various isolates of diverse origins and different morphospecies, based on three 1-kb amplicons in single-copy predicted genes. In red: strains isolated from the cheese environment; highlighted lines: strains carrying at least a fragment of *Wallaby.* 2a: Screening of the whole terverticillate *Penicillium* group from a public collection of the MNHN and some additional *Eupenicillium* species (one of the teleomorphic states of *Penicillium*). Type strains are indicated with a "T" following the strain number; 2b: Screening of the FM collection, provided by French anonymous stakeholders; 2c: Screening of the LUBEM-Brest collection, containing only *Penicillium roqueforti* strains directly isolated from cheeses from around the world (the numbers after the "F", from 1 to 39, correspond to individual cheeses; Individual strains are labelled by a number following that identifying the cheese, e.g. F17.1 and F17.2 are two strains isolated from the same cheese, labelled 17).

2a: Screening of the whole terverticillate *Penicillium* group from a public collection of the MNHN and some additional *Eupenicillium* species (one of the teleomorphic states of *Penicillium*). Type strains are indicated with a "T" following the strain number

Species	Substrate, Country	Year of acquisition by the MNHN	LCP	Other public collection number	Primer 1	Primer 2	Primer 3
E. baarnense	Soil, Brazil	2003	4723		-	-	-
E. brefeldianum	Soil, Zaire	1990	2643		-	-	-
	Soil, Zaire	1990	2644		-	-	-
E. crustaceum	Hare dejection, Algeria	1990	2693		-	-	-
	Marine sediment, Spain	2003	4719		-	-	-
	Sediment, France	2003	4734		-	-	-
	Soil, Israel	2003	4758		-	-	-
E. javanicum	Unknown	2000	4449		-	-	-
E. osmophilum	Ice moraine soil, Norway	2006	5326		-	-	-
P. adametzii	Unknown	1988	2491		-	-	-
	Germany	2007	5436T	MUCL29173T	-	-	-
P. allii	Inner fridge wall, France	2007	5417		-	-	-
P. atramentosum	Soil, France (Jardin des Plantes)	1996	3983		-	-	-
P. aurantiogriseum	Cosmetic cream, France	1996	3933		-	-	-
	Unknown	1997	4161T	IMI195060T	-	-	-
P. bialowiezense	Inner fridge wall, France	2006	5372		-	-	-
	Inner fridge wall, France	2006	5373		-	-	-
	Inner fridge wall, France	2007	5465		-	-	-
	Inner fridge wall, France	2007	5467		-	-	-
	Inner fridge wall, France	2007	5469		-	-	-
	Inner fridge wall, France	2007	5470		-	-	-
P. biforme	Unknown	1988	2487		-	-	-
	Unknown	1975	2621		-	+	+
	French cheese, USA	1948	5529T	CBS297.48T	-	-	-
P. brevicompactum	Laboratory contaminant, France	1952	699		-	-	-
	Unknown	1997	4171		-	-	-
	Sea water, France	1998	4277		-	-	-
	Nut seed, France	2003	4798		-	-	-
	Inner fridge wall, France	2006	5364		-	-	-
	Inner fridge wall, France	2006	5365		-	-	-
	Inner fridge wall, France	2006	5439		-	-	-
	Inner fridge wall, France	2006	5363		-	-	-
	Inner fridge wall, France	2007	5432		-	-	-
	Inner fridge wall, France	2007	5461		-	-	-
	Inner fridge wall, France	2007	5462		-	-	-
	Inner fridge wall, France	2007	5463		-	-	-
	Inner fridge wall, France	2007	5464		-	-	-
P. camemberti	French Camembert cheese, USA	1966	584	CBS299.48	+	+	+

Species	Substrate, Country	Year of acquisition by the MNHN	LCP	Other public collection number	Primer 1	Primer 2	Primer 3
	Saint Marcellin cheese, France	1966	1920		-	+	+
	Bleu de Bresse cheese, France	2003	4810	CMPG30	+	+	+
	Camembert cheese, Belgium	1929	5527exT	CBS303.48exT	-	+	+
	Camembert cheese, France	1908	5528exT	CBS123.08exT	+	+	+
	Leaf litter, Netherlands	1930	5531	CBS216.30	+	+	+
P. carneum	Mouldy rye bread, Denmark	2009	5634T	CBS112297T	-	-	-
P. caseifulvum	Danablue cheese, Denmark	1998	5630T	CBS101134T	-	+	+
P. cavernicola	Wall of Lechuguilla cave, USA	1998	5631T	CBS100540T	-	-	-
	Guacharo cave, Venezuela	2001	5632	CBS109558	-	-	-
	Butter, Japan	2001	5633	CBS109557	-	-	-
P. chrysogenum	Seed, France (Grignon city)	1951	456		+	+	+
	Humus soil, France	1951	487		-	+	+
	Root of *Elaeis guineesis* , Cameroun	1947	673		+	-	+
	Forest soil contaminated by petroleum	1988	2522		+	+	+
	Wood, Bahamas	1990	2616		+	+	+
	Unknown	1997	3830		+	+	+
	Branch of *Hyssopus sp.* , Norway	1995	3835	CBS355.48	-	-	-
	Cream cosmetics, France	1996	3936		-	-	-
	Archive paper, France	1996	3965		-	+	+
	Spring water, Italia	1997	4047		+	+	+
	Leather boots, France	1997	4075		-	-	-
	Unknown	1997	4080	FRR4247	+	-	-
	Unknown	1997	4081	FRR4911	-	-	-
	Dried fish food, Indonesia	1997	4082	FRR2902	+	+	+
	Dried sausage skin, France (Puy de Dôme)	1997	4114		+	+	+
	Mould contaminant of a *P. roqueforti* culture, France (Roquefort city)	1998	4222		+	+	+
	Rind of cheese, Belgium	2000	4382	MUCL31462	-	-	-
	Unknown	2000	4383	IBT5217	+	-	+
	Unknown	2000	4384	IBT5222	+	+	+
	Shelf wood from a cheese dairy, Denmark	2000	4387	IBT16B3D	+	-	-
	Cheese, Denmark	2000	4389	IBT21D9D	+	-	-
	Soil, USA	2004	5010	CMPG874	-	-	-
	Inner fridge wall, France	2007	5425		-	-	-
	Inner fridge wall, France	2007	5427		-	-	-
P. clavigeum	Sunflower seed, Unknown	1961	1842		-	-	-
P. commune	Cheese, USA	2007	5380T	NRRL890T	+	+	+
P. crustosum	Unknown	1970	2069		-	-	-

Species	Substrate, Country	Year of acquisition by the MNHN	LCP	Other public collection number	Primer 1	Primer 2	Primer 3
	Unknown	1988	2525		-	-	-
	Soil, Argentina	1991	2777		-	-	-
	Food for animals	1975	3045		-	-	-
	Chocolate cream, France	1986	3439		-	-	-
	Food-processing industry atmosphere, France	1987	3488		-	-	-
	MNHN mycology laboratory atmosphere, France	1996	3950		-	-	-
	Naturalized wildebeest, France	1996	3955		-	-	-
	Spruce wood, France	1997	3993		-	-	-
	Soil, Argentina	2002	4540		-	-	-
	Inner fridge wall, France	2006	5350		-	-	-
	Inner fridge wall, France	2006	5351		-	-	-
	Inner fridge wall, France	2006	5352		-	-	-
	Inner fridge wall, France	2006	5353		-	-	-
	Citrus sp. , United Kingdom	2007	5438T	MUCL38745T	-	-	-
	Inner fridge wall, France	2007	5447		-	-	-
	Inner fridge wall, France	2007	5480		+	+	-
	Inner fridge wall, France	2007	5481		+	+	-
	Inner fridge wall, France	2007	5482		-	-	-
	Inner fridge wall, France	2007	5483		-	-	-
	Inner fridge wall, France	2007	5484		-	-	-
P. cyclopium	Unknown	1988	2489		-	-	-
	Unknown	2000	4366T	MUCL15613T	-	-	-
P. digitatum	Lemon fruit	1976	3112		-	-	-
	Lemon fruit	1979	3236		-	-	-
	Lemon fruit	1998	4263		-	-	-
	France	2004	5233	CMPG406	-	-	-
	Inner fridge wall, France	2007	5408		-	-	-
	Inner fridge wall, France	2007	5409		-	-	-
	Inner fridge wall, France	2007	5410		-	-	-
P. dipodomyis	Unknown	1997	4101	IBT12701	-	-	-
	Denmark	1997	4102	NRRL13485	-	-	-
	Unknown	1997	4103	IBT3353	+	-	-
P. expansum	Apple fruit, France	1951	897		-	-	-
	Orange fruit	1961	1199		-	-	-
	Fig fruit	1976	3120		-	-	-
	Peach fruit	1984	3384		-	-	-
	Apple fruit	1987	3492		-	-	-
	Grape juice, France	1996	3954		-	-	-
	Taxus sp. Seed, France	1997	3994		-	-	-
	Apple fruit, USA	1997	4160	CBS325.48	-	-	-
	Unknown	1997	4167		-	-	-
	Peach sorbet, France	1997	4190		-	-	-

Species	Substrate, Country	Year of acquisition by the MNHN	LCP	Other public collection number	Primer 1	Primer 2	Primer 3
	Yogurt tub, France	1998	4207		-	-	-
	Fruit yogurt, France	1998	4257		-	-	-
	Bird nest, unknown	Unknown	4826	CMPG911	-	-	-
	Grape, France	2006	5304		-	-	-
	Inner fridge wall, France	2006	5374		-	-	-
	Inner fridge wall, France	2007	5411		-	-	-
	Inner fridge wall, France	2007	5412		-	-	-
	Inner fridge wall, France	2007	5413		-	-	-
	Inner fridge wall, France	2007	5414		-	-	-
P. flavigenum	Unknown	1997	4104	IBT16616	-	-	-
	Wheat flour, Denmark	1997	4106	CBS419.89	-	-	-
P. freii	Henhouse, France	2003	4855		-	-	-
P. fuscoglaucum	Rubber, Indochina	1950	218		-	-	-
	Unknown, Belgium	1991	2798T	CBS261.29T	-	-	-
	Wood, unknown	1979	3239		-	-	-
	Inner fridge wall, France	2007	5472		+	-	-
P. glandicola	*Polyporus sp.* Fungus, unknown	1968	1991		-	-	-
P. griseofulvum	Dead stem of Utricaceae	1952	899	CBS384.48	-	-	-
	Grain elevator, South Africa	1979	3237	CBS315.63	-	-	-
	Unknown, Belgium	1979	3245	CBS185.27	-	-	-
	Soil contaminated by fuel, France	1985	3434		-	-	-
	Soil, France (Jardin des Plantes)	1996	3982		-	-	-
	Chestnut cream, France	2003	4831	CMPG832	-	-	-
	Nut seed, France	2004	4859	CMPG564	-	-	-
P. hordei	Plastic material phone, Vietnam	1994	3758		-	-	-
P. italicum	Inner fridge wall, France	2006	5348		-	-	-
	Inner fridge wall, France	2007	5405		-	-	-
	Inner fridge wall, France	2007	5406		-	-	-
	Inner fridge wall, France	2007	5407		-	-	-
P. melanoconidium	Unknown	Unknown	282	IP1129.75	-	-	-
P. nalgiovense	Dried sausage skin, France	1985	3435		+	-	-
	Cheese, Czechoslovakia	1995	3833	MUCL31194T	+	-	-
	Unknown	1996	3915		+	-	-
	Unknown	2000	4385	IBT12679	+	-	-
	Cheese, Danmark	2000	4388		+	-	-
	Cheese, Denmark	2000	4390		+	-	-
	Unknown	2000	4392		+	-	-
	Unknown	2000	4393		+	-	-
	Unknown	2003	4804	CMPG246	-	-	-
	Soil, France	2003	4809	CMPG517	+	-	-
	Lemon fruit, France	2003	4819		-	-	-
	Soil, France	2003	4823	CMPG315	+	-	-

Species	Substrate, Country	Year of acquisition by the MNHN	LCP	Other public collection number	Primer 1	Primer 2	Primer 3
	Mummy, France	2004	4863	CMPG1309	-	-	-
	Inner fridge wall, France	2006	5347		-	-	-
P. nordicum	Dried sausage skin, France	1997	4117T	IBT30121T	+	-	-
	Inner fridge wall, France	2006	5349		-	-	-
	Inner fridge wall, France	2007	5404		+	+	-
P. olsonii	Inner fridge wall, France	2006	5354		-	-	-
	Inner fridge wall, France	2006	5355		-	-	-
	Inner fridge wall, France	2006	5356		-	-	-
	Inner fridge wall, France	2006	5357		-	-	-
	Inner fridge wall, France	2006	5358		-	-	-
	Inner fridge wall, France	2007	5402		-	-	-
	Inner fridge wall, France	2007	5403		-	-	-
P. palitans	Netherlands cheese, France	1961	1628		-	-	-
	Unknown	1988	2485		-	-	-
	Rice flour, honey, France	2004	4862	CMPG557	-	-	-
	Inner fridge wall, France	2007	5446		-	-	-
P. paneum	Mouldy rye bread, Denmark	2009	5616	MUCL40611T	-	-	-
	Silage, *Zea mays*, Belgium	2009	5617	MUCL40792	-	-	-
	Cacao, France	2009	5618	MUCL47608	-	-	-
P. polonicum	Wool textile, unknown	1956	1501		-	-	-
	Grey amber, unknown	1952	1502		-	-	-
	Cheese, Denmark	2000	4386		+	-	-
	Crab shell, Sweden	2003	4811	CMPG819	-	-	-
	Soil, France	2003	4856		-	-	-
	Crab shell, Sweden	2003	4857		+	-	-
	Soil, USA	2004	4905	CMPG858	-	-	-
	Soil, France	2004	4998	CMPG176	-	-	-
	France	2004	5009	CMPG524	-	-	-
	Inner fridge wall, France	2006	5361		-	-	-
	Inner fridge wall, France	2007	5398		-	-	-
	Inner fridge wall, France	2007	5399		-	-	-
	Inner fridge wall, France	2007	5400		-	-	-
	Inner fridge wall, France	2007	5401		-	-	-
P. roqueforti	Roquefort cheese, France	1975	146		+	+	+
	Atmosphere brasserie, unknown	1950	148		-	-	-
	Cheese, Unknown	1964	1883		-	-	-
	Unknown	1988	2492		-	-	-
	Brioche under plastic wrap	1993	2939		-	-	-
	Fruit compote, France	1993	3676		-	-	-
	Fruit compote, France	1996	3914		-	-	-
	Fruit compote, France	1997	3969		-	-	-
	Wood, France	1997	4111		-	-	-
	French Roquefort cheese, USA	1997	4157T	CBS221.30T	-	-	-
	Strawberry sorbet, France	1997	4180		-	-	-
	Inner fridge wall, France	2007	5419		-	-	-

Species	Substrate, Country	Year of acquisition by the MNHN	LCP	Other public collection number	Primer 1	Primer 2	Primer 3
	Inner fridge wall, France	2007	5420		-	-	-
	Inner fridge wall, France	2007	5421		-	-	-
	Silage, France	2011	5885	LMSA1.11.033	-	-	-
	Cheddar cheese, unknown	2009	5629	CBS449.78	-	-	-
P. rubens	Unknown	Unknown	284		-	-	-
	Unknown	1946	612		-	+	+
	Sandy soil, Algeria	1956	1117		+	+	+
	Sand, Algeria	1956	1177		+	+	+
	Water of the Seine river, France	1989	2574		+	+	+
	Unknown	1990	2601		+	-	-
	Automobile surfacing material, France	1996	3913		-	-	-
	Compact comestics powder, France	1996	3932		+	+	+
	Soil, Australia	1997	4078		-	-	-
	Sorghum malt beverage, South Africa	1997	4079		-	-	-
	Rice imported from Iran, South Africa	1997	4083	FRR2700	-	+	+
	Dried sausage, France (Auvergne)	1997	4135		+	-	-
	Laboratory contaminant, Belgium	1997	4142	MUCL29142	-	+	+
	Isolated by Fleming in 1945	1997	4143	MUCL30169	-	+	+
	Unknown	2003	4838	CMPG1441	+	+	+
	Nut oil cake, France	2003	4854	CMGP138	-	+	+
	Sandy rock, Sahara	2004	4885		+	+	+
	Inner fridge wall, France	2007	5426		+	-	-
	Mouldy cantaloupe, USA	2007	5883	DSM Wisonsin, 54-1255	+	+	+
	Contaminant of the *P. chrysogenum* type strain CBS306.48	2013	6033		+	+	+
P. solitum	Dairy product, France	1997	3966		-	-	-
	Unknown	1997	4136		-	-	-
	Inner fridge wall, France	2006	5359		-	-	-
	Inner fridge wall, France	2006	5360		-	-	-
	Inner fridge wall, France	2007	5422		+	+	-
	Cheese environment, France	2008	5509		+	-	-
P. venetum	Asparagus, Grenoble	2003	4858		-	-	-
P. verrucosum	Unknown, Belgium	2000	4362	MUCL29089	-	-	-
	Unknown, Belgium	2000	4363	MUCL29186	-	-	-

Species	Substrate, Country	Year of acquisition by the MNHN	LCP	Other public collection number	Primer 1	Primer 2	Primer 3
	Unknown, Belgium	2000	4364	MUCL28674	-	-	-
P. viridicatum	Unknown	1988	2490		-	-	-
	Atmosphere, USA	2007	5435	MUCL39358	-	-	-

2b: Screening of the FM collection, provided by French anonymous stakeholders:

Species	FM number	Primer 1	Primer 2	Primer 3
Fusarium domesticum	FM005	-	-	-
	FM024	-	-	-
	FM038	-	-	-
	FM044	-	-	-
	FM088	-	-	-
	FM254	-	-	-
	FM255	-	-	-
	FM256	-	-	-
	FM277	-	-	-
Geotrichum candidum	FM074	-	-	-
Lecanicillium lecanii	FM017	-	-	-
P. biforme	FM042	-	-	-
	FM071	-	-	-
	FM147	+	-	-
	FM148	-	+	+
	FM155	-	-	-
	FM168	-	+	+
	FM169	-	+	+
	FM229	-	-	-
	FM230	-	-	-
	FM231	-	-	-
	FM232	-	-	-
	FM233	-	-	-
	FM236	-	-	-
	FM237	-	-	-
	FM238	-	-	-
	FM239	-	-	-
	FM261	-	-	-
P. camemberti	FM013	+	+	+
	FM014	+	+	+
	FM025	+	+	+
	FM026	+	+	+
	FM027	+	+	+
	FM150	+	+	+
	FM152	-	+	-
	FM154	+	+	+
	FM185	+	+	+
	FM186	+	+	+
	FM188	+	+	+
	FM189	+	+	+
	FM190	+	+	+
	FM191	+	+	+
	FM192	+	+	+
	FM210	+	+	+
	FM252	+	+	+
	FM253	+	+	+
	FM153	+	+	+
P. crustosum	FM245	+		
	FM246	-	-	-

Species	FM number	Primer 1	Primer 2	Primer 3
P. fuscoglaucum	FM041	-	-	-
	FM234	-	-	-
P. nalgiovense	FM151	+	-	-
	FM187	+	-	-
	FM193	+	-	-
	FM194	+	-	-
	FM195	+	-	-
P. paneum	FM227	-	-	-
	FM228	-	-	-
P.roqueforti	FM015	+	+	+
	FM016	+	+	+
	FM037	+	+	+
	FM156	-	-	-
	FM157	-	-	-
	FM158	-	-	-
	FM159	-	-	-
	FM160	-	-	-
	FM162	+	+	+
	FM163	+	+	+
	FM164	+	+	+
	FM165	-	-	-
	FM167	-	-	-
	FM170	+	+	+
	FM171	-	-	-
	FM172	-	-	-
	FM173	-	-	-
	FM174	-	-	-
	FM175	-	-	-
	FM176	-	-	-
	FM177	-	-	-
	FM178	-	-	-
	FM179	+	+	+
	FM211	+	+	+
	FM215	+	+	+
	FM216	-	-	-
	FM217	-	-	-
	FM218	+	+	+
	FM219	-	-	-
	FM220	+	+	+
	FM221	+	+	+
	FM222	-	-	-
	FM223	+	+	+
	FM224	+	+	+
	FM225	+	+	+
	FM259	-	-	-
	FM263	+	+	+
	FM315	-	-	-
	FM316	+	+	+
	FM317	-	-	-

Species	FM number	Primer 1	Primer 2	Primer 3
	FM358	+	+	+
	FM359	+	+	+
	FM360	+	+	+
P. solitum	FM070	+	-	-
	FM241	-	-	-
	FM242	-	-	-
	FM243	+	-	-
	FM244	-	-	-
P. verrucosum	FM072	+	+	+
Scopulariopsis candida	FM285	-	-	-
	FM286	-	-	-
Scopulariopsis flava	FM006	-	-	-
	FM068	-	-	-
	FM069	-	-	-
	FM281	-	-	-
	FM282	-	-	-
	FM283	-	-	-
	FM284	-	-	-
	FM287	-	-	-
	FM290	-	-	-
	FM291	-	-	-
Scopulariopsis fusca	FM280	-	-	-
	FM288	-	-	-
Sporendonema casei	FM313	-	-	-
	FM314	-	-	-
	FM318	-	-	-

2c: Screening of the LUBEM-Brest collection, containing only *Penicillium roqueforti* strains directly isolated from cheeses from around the world (the numbers after the "F", from 1 to 39, correspond to individual cheeses; Individual strains are labelled by a number following that identifying the cheese, e.g. F17.1 and F17.2 are two strains isolated from the same cheese, labelled 17).

P. roqueforti strains - LUBEM collection	Cheese substrate	Country of cheeses origin	Presence/absence *Wallaby*
F1.1	Blue cheese	Canada	+
F2.1	Blue cheese	Canada	+
F3.1	Blue cheese	Canada	+
F4.7	Blue cheese	Canada	-
F5.2	Fourme d'Ambert	France	+
F5.3	Fourme d'Ambert	France	+
F6.1	Gorgonzola	Italy	+
F6.3	Gorgonzola	Italy	+
F7.1	Gorgonzola	Italy	-
F7.3	Gorgonzola	Italy	+
F8.1	Gorgonzola	Italy	+
F9.1	Fourme d'Ambert	France	+
F9.4	Fourme d'Ambert	France	+
F9.5	Fourme d'Ambert	France	-
F10.1	Bleu d'Auvergne	France	-
F10.2	Bleu d'Auvergne	France	-
F10.3	Bleu d'Auvergne	France	-
F10.5	Bleu d'Auvergne	France	-
F11.1	Carré Aurillac	France	+
F11.3	Carré Aurillac	France	+
F11.5	Carré Aurillac	France	+
F12.1	Pigne	France	+
F12.2	Pigne	France	+
F12.5	Pigne	France	+
F13.1	Carré Auvergne	France	-
F13.2	Carré Auvergne	France	+
F13.3	Carré Auvergne	France	+
F13.4	Carré Auvergne	France	+
F14.1	Gorgonzola	Argentina	+
F14.3	Gorgonzola	Argentina	+
F14.5	Gorgonzola	Argentina	+
F14.6	Gorgonzola	Argentina	+
F15.1	Blue cheese	Brazil	+
F15.3	Blue cheese	Brazil	+
F15.4	Blue cheese	Brazil	+
F16.1	Picon Hoja	Spain	+
F16.2	Picon Hoja	Spain	+
F16.6	Picon Hoja	Spain	-
F17.1	Peña Santa	Spain	+
F17.2	Peña Santa	Spain	-
F18.1	Tresviso	Spain	+
F18.2	Tresviso	Spain	+
F18.6	Tresviso	Spain	-
F19.1	Peral	Spain	+
F20.1	Cabrales	Spain	+
F20.4	Cabrales	Spain	+

P. roqueforti strains - LUBEM collection	Cheese substrate	Country of cheeses origin	Presence/absence *Wallaby*
F21.1	Blue cheese	Spain	+
F22.1	Blue cheese	Netherlands	+
F22.2	Blue cheese	Netherlands	+
F22.5	Blue cheese	Netherlands	-
F23.1	Blue cheese	Netherlands	+
F23.3	Blue cheese	Netherlands	+
F24.2	Blue cheese	Netherlands	+
F25.1	Blue cheese	Netherlands	+
F25.6	Blue cheese	Netherlands	+
F26.2	Blue cheese	Netherlands	+
F26.3	Blue cheese	Netherlands	+
F27.1	Blue cheese	USA	-
F28.1	Blue cheese	Latvia	+
F28.2	Blue cheese	Latvia	+
F28.3	Blue cheese	Latvia	+
F29.1	Blue cheese	Denmark	+
F29.3	Blue cheese	Denmark	+
F30.1	Blue cheese	Poland	+
F31.1	Blue cheese	Latvia	+
F31.2	Blue cheese	Latvia	+
F32.1	Blue cheese	Denmark	+
F33.1	Blue cheese	Germany	+
F34.1	Blue cheese	Estonia	+
F35.1	Blue cheese	Germany	+
F35.5	Blue cheese	Germany	+
F36.1	Blue cheese	Germany	+
F37.1	Blue cheese	Germany	+
F37.4	Blue cheese	Germany	+
F38.1	Blue cheese	Germany	+
F39.1	Blue cheese	Germany	+

Chapter Seven: Early results towards assessing the genomic potential for secondary metabolite production in Penicillium species and its impact on food safety.

This chapter presents preliminary work carried out at the end of this thesis. It describes early results on the ability of *Penicillium* species to produce secondary metabolites. The aim is to generate a sufficient body of knowledge to serve as a comprehensive resource in further genetic and biochemistry experiments. This resource could take the form of a review on the genomic potential for secondary metabolite production in *Penicillium* species, and more work will be carried towards this aim. The study uses genomes generated during the Food Microbiomes project and therefore has a particular emphasis on species from the food environment. The study also includes the genome of the industrial *P. rubens* strain, as well as two strains of a postharvest citrus pathogen, *P. digitatum*. A large body of work remains to be carried out to fully achieve this goal; however, the data presented here provides an early assessment of the potential for secondary metabolite production in *Penicillium*.

Introduction:

Secondary metabolites are small bioactive molecules produced by many organisms, including bacteria, plants and fungi. Fungi produce many of these small compounds not necessary for their growth. These compounds have been found in the air, soil and crops, as well as in food substrates, such as cheese (Larsen et al. 2002). They play a significant role in fungal interaction with ecosystems, as signalling molecules, and as antibiotic or antifungal compounds. It has been estimated that over fifty percent of fungal secondary metabolites have antibiotic or antifungal or antitumour properties (Peláez 2006). Most of these compounds are unknown or uncharacterized, and as such they represent a large untapped reservoir for biotechnological innovation. Scientific interest in these compounds is relatively recent, dating back to the 1940's, with the realisation of the massive impact of penicillin on human health. The vast majority of literature reference on the fungal secondary metabolites has been published between 1995 and today. Man has been using them as antibiotics, immunosuppressant agents, anti-tumour compounds, pesticides and cholesterol lowering drugs, among other possible uses (Hoffmeister & Keller 2007). Part of the wide array of fungal secondary metabolites is commonly termed mycotoxins, and these include aflatoxin, fumonisin, trichosthecene and zearalone. Examples of mycotoxins with effects detrimental to human health are penitrems and cyclopiazonic acids, with neurotoxic effects, ochratoxin, with both carcinogenic and nephrotoxic effects, or the carcinogenic family of aflatoxins. Despite these detrimental effects, knowledge concerning mycotoxin production by species from the food environment is scarce and often only

constituted of biosynthetic pathway elucidation by biochemistry experiments. The genomic basis for the production of such compounds (whether beneficial or detrimental to human health or uses) is largely unknown, with to date and to our knowledge only the *Aspergillus* and *Fusarium* species having received significant attention at the genomic level (Inglis et al. 2013; Hansen et al. 2012).

Penicillium species *a*re known to produce several secondary metabolites. *Penicillium rubens* produces an array of metabolites, including penicillin, *Penicillium roqueforti, Penicillium paneum* and *Penicillium carneum* are known to produce roqefortine C and andrastin A. As well, both *P. paneum* and *P. carneum* produces the mycotoxin patulin while *P. roqueforti* does not. *Penicillium roqueforti* is also known to produce the PR toxin (O'Brien et al. 2006). Importantly, many of these secondary metabolites are not expressed under laboratory conditions, but could be produced in other environments. This is a major source of concern for food safety authorities, and is partly responsible for some widely used traditional food filamentous fungi not being awarded a Qualified Presumption of Safety status by the European Food Safety Association (EFSA).

Some of the mycotoxins are known to occur in food products and notably cheese. They include mycophenolic acid, cyclopiazonic acid, roquefortin C, penitrems, andrastins, and other compounds with unknown structure and toxicities (Larsen et al. 2002). Secondary metabolites in cheeses arise from milk contamination, by production of mycotoxins (or other secondary metabolites wich could have interesting technological properties) by food spoilage agents or by technological filamentous fungi used in the process (Ueno 1985). Many *Penicillium species* with an unknown potential for mycotoxin production occur in the cheese environment, both as fungal starters or contaminants. It is therefore important to assess the genomic potential of these species and their safe use in food processes. The availability of several *Penicillium* genomes sequenced by the Food Microbiomes project therefore represent a unique opportunity to evaluate the potential for production of secondary metabolites in cheese.

The first step in the biosynthesis of any secondary metabolite is catalysed by one or more main protein, herein referred to as "backbone". These are multidomain enzymes classified into polyketide synthases (PKS), non ribosomal peptide synthases (NRPS), hybrids of both, or dimethylallyl tryptophan synthases (DMATS) on the basis of their domain composition. (See chapter three in the introduction for a detailed overview of domain architecture of secondary metabolites in fungi). Additional enzymes are involved in further catalysing tailoring reactions to the intermediate backbone product synthesized by the backbone protein. All the genes encoding these tailoring proteins are usually found in clusters around the backbone gene.

Genome sequences used in this study:

-*Penicillium rubens Wisconsin-1255*, *Penicillium digitatum PHI26* and *Penicillium digitatum Pd1*: genomes and proteomes downloaded from the EMBL database. Full report on the genomes previously published (Berg et al. 2008; Marcet-Houben et al. 2012b).

-*Penicillium roqueforti FM164* and *Penicillium camemberti FM013*: sequenced at the genoscope through a mix of paired end 454 and illumina single read sequencing. Annotation performed using the Eugene platform by the INRA-LIMP, Toulouse. Publication in preparation, full report is given in chapter 6.

-*Penicillium paneum FM227* and *Penicillium carneum LCP5634* and *Penicillium nalgiovense FM193*: sequenced at the BGI-Shenzhen using illumine paired end libraries. Independant annotation performed at the BGI-Shenzhen (unpublished data).

-*Penicillium fuscoglaucum FM041, Penicillium biforme FM169:* sequenced at the BGI using illumine paired end sequencing. Annotations based on homology with *Penicillium camemberti.* Annotation was performed at the BGI-Shenzhen using the *Penicillium camemberti FM013* gene set. (unpublished data).

Phylogenetic relationships between species and strains can be seen on the tree in chapter six.

Secondary metabolite backbone genes (and cluster) identification:

Secondary metabolite backbone genes were detected using the SMURF software (Secondary Metabolite Unknown Region Finder, hosted at the John Craig Venter Institute (Khaldi et al. 2010); http://www.jcvi.org/smurf/index.php). SMURF systematically predicts clustered secondary metabolite genes based on their genomic context and domain content. Backbone genes and associated clusters are then classified into PKS, PKS-like, NRPS, NRPS-like, hybrids of PKS-NRPS and DMATS based on their domain composition. PKS-like and NRPS-like proteins are enzymes with at least two domains of the minimal combination of domains the NRPS or PKS encodes.

Orthology relationships between the identified backbone genes was then searched for using reciprocal protein blast hits with a threshold of over sixty percent identity over sixty percent of the query using the stand-alone NCBI blast program (Altschul et al. 1990).

Early results:

Identification of the secondary metabolite repertoire of *Penicillium species:*

Secondary metabolite backbone genes were present in all species screened. Between 24 and 58 total secondary metabolite backbone genes by strain were found (see table 4), indicating a large potential for secondary metabolite production in the *Penicillium genus* (419 total backbone genes identified). *Penicillium nalgiovense* possesses the smallest array of secondary metabolite backbone genes both in terms of number (22) and in terms of proportion relative to the whole protein coding gene content (0.22%), while the highest potential seems to reside within the *Penicillium camemberti* genome with 58 backbone genes identified. *P. biforme*, *P. camemberti* and *P. fuscoglaucum*, which are closely related species, all possess more PKS and PKS-like enzymes than the other *penicillium* species examined. This is in agreement with their seemingly high number of backbone genes which lies between 46 and 58. The two *Penicillium digitatum* strains possess only 32 backbone genes, which is, (with the exception of *Pencillium nalgiovense*), fewer than the other *Penicillium* species screened. However all the species examined possess a wide array of NRPS and NRPS like enzyme. A difference also lies in the number of DMATs encoding genes, with the four earliest diverging species having only one putative DMAT encoding gene. *Penicillium rubens Wisconsin-1255* also encodes a large number of secondary metabolite genes. This strain has been engineered towards production of pharmaceutical compounds.

strain	*P. nalgiovense FM193*	*P. rubens wisconsin-1255*	*P digitatum PHI26*	*P. digitatum Pd1*	*P. paneum FM227*	*P. carneum LCP5634*	*P. roqueforti FM164*	*P. fuscoglaucum FM041*	*P. biforme FM169*	*P. camemberti FM013*
Environment	cheese (present strain) sausage	industrial production	citrus pathogen	citrus pathogen	cheese	cheese	cheese	cheese	cheese	cheese
PKS-like	1	1	3	4	2	5	3	1	2	2
PKS	4	21	11	10	16	13	12	23	24	26
NRPS-like	11	14	4	4	5	14	9	9	9	11
NRPS	5	10	10	11	11	8	8	8	9	14
Hybrid	2	2	3	2	2	1	3	3	3	2
DMAT	1	1	1	1	5	7	4	2	3	3
TOTAL	24	49	32	32	41	48	39	46	50	58
protein coding genes	11104	12943	9133	8961	10002	10179	13036	13252	13482	14440
percentage	0,22	0,38	0,35	0,36	0,41	0,47	0,30	0,35	0,37	0,40

Table 4: Detected secondary metabolite backbone genes in the *Penicillium* species.

Possible orthologous genes were searched using a reciprocal blast hit approach between all the identified backbone genes (fig 22). Few genes are shared by all strains (2%, four groups of orthologs)

and only 15% of the backbone genes show orthology relationships between four strains or more representing 26 secondary metabolites. The vast majority of backbone genes are present in only one strain (49%). This trend is even higher when considering species rather than strains: among the 24% of orthologous genes between only two strains, 16% is due to genes present in both *Penicillium digitatum* strains, indicating the actual unique backbone genes between species is higher than 49% (the percentage of unique backbone genes between species is in fact of 65%). Both strains of *Penicillium digitatum* each possess a backbone gene the other strain does not have indicating that differences in secondary metabolite production exists between strains of the same species.

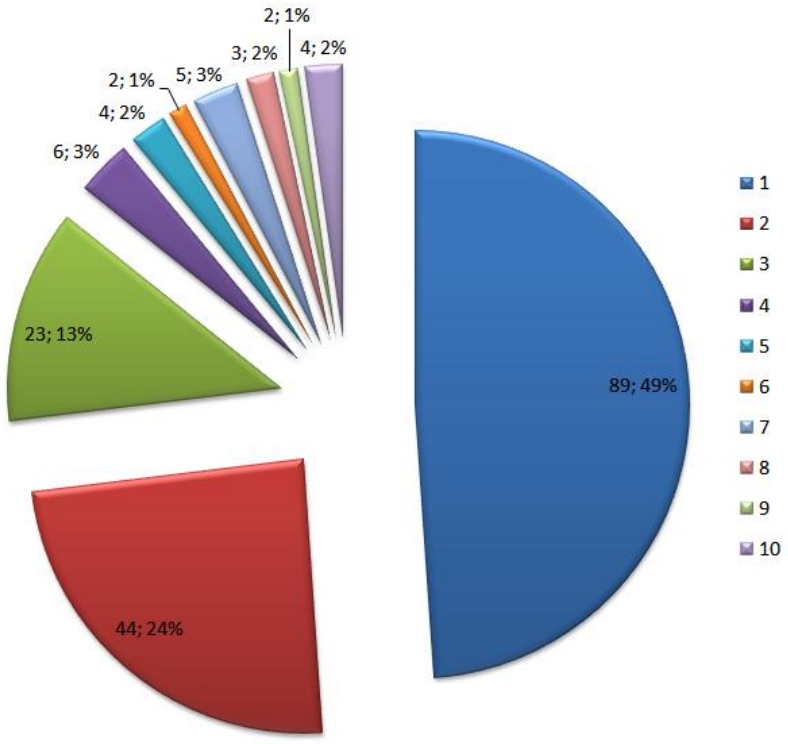


Figure 22: Percentage of orthologous backbone genes between species. Colours indicate the number of species in which the genes could be found as indicated on the right. Gene count and percentage are indicated inside or next to each portion of the pie chart.

Early conclusions and future work:

Screening for secondary metabolite production in ten P*enicillium* strains belonging to different species reveals a wide array of secondary metabolite backbone genes, potentially indicative of a high potential for the production of many compounds. Most of these secondary metabolites are species

specific, and it may be possible identify species on the basis of their secondary metabolite profile, as suggested previously (Frisvad et al. 2007). Analysis of the clusters of genes surrounding these backbone genes and their synteny between different species with characterized secondary metabolites may help in identifying the compounds these polyketide synthase, non ribosomal peptide synthases and dimethylallyl tryptophan synthases encodes. As well, extensive literature and database mining will provide useful insights for identifying already characterised secondary metabolites in closely related species. Biochemistry experiments in selected strains for which the genomic content is known can also be an interesting approach to further link the compounds to their associated gene clusters. Secondary metabolites clusters have been shown to be expressed only in certain unknown conditions and under epigenetic control (Brakhage 2013). Strains with mutated genes in histone deacetylases, H3K9 methyltransferases and other genes involved in histone methylation could be constructed to further characterise secondary metabolites at the biochemical level. As well, strains with mutation in the secondary metabolite backbone genes identified can be constructed, and differences in secondary metabolite production can be assessed and used to identify the gene cluster responsible for the production of some secondary metabolites. These two approaches should be combined, to identify as many compounds and confirm as many clusters as possible. Finally assessment of the toxicity of these compounds in some strains or species should be carried out, and the genomic detection of similar clusters in other food associated strains can be used to further assess their safe use and potential for secondary metabolite production.

Chapter 8: General discussions, perspectives and conclusions.

This chapter sums up the main results of this PhD, and present future perspectives. It also discusses the potential implications of the work carried out during this project. We focus here on the two main aims of the project, presented in chapter 5 and chapter 6. We begin the discussion by commenting on the development of a methodology for improving assemblies using an original single molecule technology, Molecular Combing and, continue with its application to the genome of *Penicillium roqueforti.* Possible perspectives in the further development of such an application are presented, as well as how they could be realised. In a second time, we briefly get back at the results of chapter 6, the discovery of a multiple and recurrent exchange of a very large genomic island, in a food environment. We then discuss how important is the assessment of the frequency of such transfers, not only in man-associated environments, but also in nature. The potential impacts such transfers can have on ecosystems is also presented, by considering aspects of fungal interactions with other organisms in ecosystems.

1. On the development of a new methodology for scaffolding and improving assemblies and its application to *Penicillium roqueforti FM164:*

Objectives:

One of the aims of the project was to explore the potential contributions of Molecular Combing to whole genome *de novo* assembly. This work was funded under a CIFRE grant, under the form of a joint partnership between Genomic Vision and the BAC team, and while the end results of the work were similar, the expected outcomes of the work are different.

-If considered in the framework of the Food-Microbiomes project, the aim was to produce a good quality assembly of a *Penicillium* species to serve as a reference sequence of these important food-related fungal species. This will help in comparative genomic studies, to gain a better knowledge of the genome dynamics, evolution and content of these species, as well as to assess their metabolic properties through genome mining, with applications in the biotechnology and food industries. The production of *Penicillium* genomes is also an important addition to databases used for metagenomic studies, to characterise ecosystems.

-From a technological perspective, one of the aims of this project was to carry exploratory work to assess the potential for a novel application of molecular combing. The goal was to obtain a proof of concept through the production of a good quality assembly of a newly sequenced genome.

This was achieved by the conception, development and application of a methodology for improving assemblies with as expected outcome the positioning of the technology in the sequencing field.

Results:

The approach developed at the beginning of this work yielded encouraging results. *Penicillium roqueforti*'s genome sequence was improved by locating many long scaffolds on combed DNA molecules and the resulting assembly is made of a collection of six large super-scaffolds and 42 smaller scaffolds. The six largest super-scaffolds are accounting for more than ninety two percent of the genome and represents chromosomes or large chromosomal fragments. The resulting assembly is not hundred percent complete but could still be improved. A couple of large scaffolds remain to be located. The four largest unplaced scaffolds have sizes between 300kb and 150kb. Attempts to position them during the project were unsuccessful, but the Genomic Morse Codes for these scaffolds were made of very few probes, separated by long distances. Because many of the most informative fibres are around 300 kb long or shorter it is likely that these Genomic Morse Codes were fragmented and that long enough fibres at this locus were not observed to enable scaffolding. New Genomic Morse Code with more probes and less spacing between them would most likely allow the location of these scaffolds. Another possibility is that these scaffolds are misassembled so that the expected pattern yielded by the GMCs could not be detected. This would also be solved by the design of new, better suited Genomic Morse Codes. These experiments have not been carried out for a question of time and other objectives in the project.

Many new *Penicillium* genomes were sequenced later during the project, including re-sequencing of four *Penicillium roqueforti* strains by SOLiD tehcnology and these could be used to position the remaining scaffolds. Despite SOLiD not being the best technology for assembly, assemblies of these strains were performed and while yielding very short scaffolds (N50 around 20 kb), they could still be useful to position some of the smallest FM164 scaffolds by bioinformatic analyses and PCR amplification and sequencing.

Altogether, the workflow developed allowed to significantly improve the assembly of *Penicillium roqueforti*'s genome sequence. As such this already represent an interesting result and a good proof-of-concept, and while, in its present state the methodology is relatively time consuming and labour intensive, it may still be useful to some researchers. Noteworthy, Molecular Combing does not always necessarily require the expertise and technical know-how of Genomic Vision for simple experiments (small genomes, single locus etc.). As such, and for simple cases, it is accessible to any

molecular biology scientist. For more complex settings (large or difficult to extract genomes) the technical expertise and technological power of Genomic Vision might be required.

Furthermore, the approach was designed to be as flexible as possible, and many steps could be improved and automated. Future steps in improving the workflow are described below.

Technological considerations:

The process developed here works well and has been designed to be flexible, giving the possibility to use any external information available. Adding several already existing programs into a pipeline could improve the workflow, and contribute to a more complete comparative genomic step, at several levels. Many such programs exist, and I will try to give below an overview of what might be achieved through examples. Other programs could be used in combination, or as alternatives:

-A first round of error detection, gap filling and scaffolding software could be used in an attempt to improve assembly by bioinformatic methods. The PAGIT (Post Assembly Genome Improvement Toolkit) suite of tools developed by the Welcome Trust Sanger Institute seems a good candidate (Swain et al. 2012). It proposes a tool (ABACAS) for positioning contigs against a closely related reference (if available); another tool using an iterative approach for closing gaps in assembled genomes using mate pair information (even when using the same data sets as used by the original assembler); ICORN, a program for correcting errors in the sequence by mapping reads to the assembly, and also important, RATT, a tool for transferring annotations from an earlier assembly onto the latest assembly (which is an important practical feature, as the final assembly will need to display the annotations to be of any use). REAPR could also be used to break scaffolds and contigs at misassemblies using mate pair information. This suite of tools would be used as a first round of assembly improvement with the purpose of working with the best quality of scaffolds as possible, excluding misassemblies, gaps and errors in sequence. Then MUMmer (Kurtz et al. 2004) could be used to compare the resulting assembly with other available assemblies as was done during this project. MUMmer is fast program for comparing large genomes and output sets of coordinates for similar sequences across scaffolds under the form of a table. These kinds of outputs are easily parsable using simple scripts. Once hypothesis have been made, using the coordinates from MUMmer, primer3 (Untergasser et al. 2012) can be used to automatically design primers for the PCR screening of junctions (the many primers designed for PCR screening and probe production in this study were automatically designed using a simple script with this software). All these software would render the comparative genomics step almost transparent to the biologist.

This could go as far as automatically generating unique Genomic Morse Codes and primers to generate the probes -thereby rendering the first steps of the workflow completely user-free. However, this would require the development of an algorithm to generate unique patterns based on the use of probe length, gap length and labelling colours. While this is conceptually simple, such software does not exist at the moment.

If the previous steps were to be automated, much of the labour of this approach would come from "at the bench" work. Depending on the context in which the assemblies are sequenced, amelioration could take place too. Large genome sequencing projects like the i5k project (insect 5000 genomes project, (George et al. 2012) already use automates to perform FISH on chromosome to physically map BACs in hierarchical shotgun sequencing. In addition to this, Genomic Vision is constantly improving molecular combing and its associated processes, and complete automation is already in development. It could be envisioned to use robots for PCR and probe production and labelling. While at first, this seems a costly endeavour, given Genomic Vision's core business and the already existing automations (Microscopy scanner, image analysis software, etc.) and recent development and effort in automating all other steps of the molecular combing technology, more automation seems realistic and should naturally occur in the coming years. If this is the case, then I believe Molecular Combing improvement of genome assemblies could become a more mainstream application.

Regarding the scope of application, the projects in which genomes are sequenced vary widely, with some projects only sequencing one small genome and others aiming at sequencing thousands of different species (e.g.: insect 5000 genomes project, Thousand fungal genome project, Genome 10k project (aiming at sequencing 10 000 vertebrate genomes, etc), while other others aims at understanding tumour genomes, or sequencing polyploid, highly repetitive plant genomes. Today it is possible to generate prokaryotic genomes at a low cost and high throughput. For these genomes, most of the resulting assemblies are almost complete and easily finished by comparative genomics or PCR alone, and similar work as to what was performed here would not bring anything to such projects. The proof of concept was developed on a relatively small eukaryotic genome, and already brought improvement to the genome sequence but the application would be even more useful in large and complex genomes, where assemblers outputs hundreds to thousands of small scaffolds. Importantly, so long as a good extraction protocol can be used, any genome can be combed, regardless of its size, repetitive content or ploïdy. All these are important genomic features responsible for the complexity of genome projects, and the methodology developed here could potentially help reduce this complexity. As such, I believe this methodology could be very useful in multicellular eukaryotic genome sequencing and assemblies, as well as in plant genome sequencing projects.

In addition, sequencing centres are specialised in providing new genomes, and some are already using alternative technologies like Opgen's optical mapping, indicating complementary technologies for assemblies are in demand. A proper benchmarking with these technologies would be required to assess whether a competitive product or service is commercially viable. However, Molecular Combing has many advantages over these technologies, including flexibility in the probe design (as opposed to sequence dependent pre-existing restriction sites for optical mapping) and yields longer and more uniform fibres. Probes could also be designed to map and size repetitive element regions, turning repeats which are major hindrance of sequencing projects into an advantageous feature of the methodology.

In conclusion, while a lot of work remains to fully automate the workflow, and while it is at the present day too early to make a clear assessment of the cost of the technology, it seems that the technology has not yet reached its full potential in the genome assembly field. Further complexifying the matter, this field is also dependant on the fast and ever changing landscape of available sequencing technologies.

on the future of sequencing:

One might wonder at the future of *de novo* assembly in light of the constant evolution of sequencing technologies. While the common lane of thought is that single molecule sequencing or very long reads will render finishing technologies useless, this is not the case at the moment. The commercially available technology presenting by far the longest read length is Pacific Biosciences' sMRT, with some reads in the 20-25 kilobases long length. It is interesting to mention that at the moment read length is not uniform, with this type of sequencer producing many short reads and very few long reads (see introduction). Of note this technology, which was announced in 2009 (Eid et al. 2009) still today suffers from a very high error rate (estimates are of 70% base calling accuracy) and as a consequence, is used for whole genome *de novo* sequencing only in combination with another second generation sequencing technology and albeit producing high quality assemblies, do not produce 100% complete genome sequences. The field of sequencing is full of stories of promising technologies pioneered by companies now shut down (Helicos Biosciences, Halcyon Molecular, ZS genetics,...) and somewhat overselling announcement (Oxford Nanopore announced two promising devices commercially available within 6 months at the AGBT in February 2012...). Rendering prediction for the future even more hazardous, the field of sequencing is filled with blocking patents

and stories of companies suing one another for patent infringement. At the present day, it seems unlikely to witness a dramatic change in the way *de novo* whole genome sequencing is carried out for the next five to ten years (and this is already a short time for a good research proof of concept to mature into a robust commercial technology for use in routine settings).

Other applications:

In light of these considerations, one of the most likely scenario for the near future of sequencing is a field where many different sequencers are available, each with their advantages and drawbacks, and little room for improvement (most of the time, it is improvement in chemistry that drive a small increase of read length). Public database would be filled with sequences of different qualities, similar to what is the case today, and finishing technologies would still be handy for improving these sequences in studies where it is required. In its present state, molecular combing based finishing can also be used to provide answer at one specific problematic locus, in a pure physical mapping approach. Other applications or case in which molecular combing could provide a clear advantage are centromeric regions (which has already been demonstrated in our lab on a large centromeric human chromosomes, L. Cinque, personal communication, unpublished data), difficult to sequence genomes such as many plant genomes which are full of long repetitive elements and quite often polyploid, tumour genomes, which are known to rearrange a lot, and in which the representation of genomes in the different cells of the sample is often heterogeneous with different cells presenting different rearrangements depending on their location in the tumour and the tumoural progression state.

As a conclusion, today the approach developed during this work represents a means to improve assembly, on a genome wide scale or at particular loci of interest. Its future remains unclear as current technologies are improved and new sequencing technologies hit the market. The methodology developed here can be rendered more efficient through automation of all the steps. These automations could be easily developed (comparative genomic step) and others are already underway (Wetware and software automations of Molecular Combing). One of the perspectives of evolution of this work also lies in the application of the approach to different questions than *de novo* whole genome sequencing including surveying genome rearrangements in tumour genomes and mapping of centromeric regions or other complex repetitive regions such as the highly repetitive large pseudo-autosomal regions found in many eukaryotic genomes.

2. On Comparative genomics and lateral genetic exchange in Fungi:

Estimates of the total number of fungal species are between 3.1 and 5.1 million species (Blackwell 2011). Of these, only a small fraction have been described (about 100 000) and even fewer been characterised at the biological level (be it genomic or physiological). Due to the nutrition mode of its representatives, the fungal kingdom proposes a wide diversity of lifestyles and associations with other organisms from the same or different kingdoms. Because of these lifestyles, fungi are key components of most ecosystems. They are the major decomposers of biomass, feeding by osmotrophy through the degradation of complex polymers by fungal enzymes secreted in the environment. An important aspect of this mode of nutrition is the nature of the interactions fungi establish in ecosystems. By breaking down complex molecules, often inaccessible to other organisms, into smaller compounds like sugars, amino acids, and nutrients, they render them accessible not only for themselves, but also to other organisms. Such interactions are often described as the production of "public goods"(Richards & Talbot 2013). This osmotrophic feature of fungi means they are involved in the production, protection and acquisition of these public goods and position them as a pillar of ecosystems. Many other species also depends on the acquisition of such public goods and as a consequence, engage in cooperative or competitive behaviours, shaping the functional interactions in ecosystems the fungi are part of. Equally important, fungi have a wide array of secondary metabolites such as mycotoxins to protect the public goods produced and exclude competitors. Studies on the potential for secondary metabolites of fungal species clearly highlight the difference between the known, characterised and produced compounds and the genomic potential of the fungal cell. Most of the detected compounds have unknown functions, but it has been proposed that over half of the secondary metabolites produced by fungi are involved in antimicrobial or antagonistic interactions with other microorganism (Peláez 2006)

Lateral genetic transfer has very recently been proposed as a major factor in both shaping and driving these public goods interaction (Richards & Talbot 2013). This is achieved through the acquisition of novel functions and by increased selective pressure for lateral genetic exchange towards osmotrophy and competition (both to produce the public goods and to ensure they remain available to the fungus). As a consequence, the acquisition by lateral transfer of secreted depolymerising enzymes, transporter proteins, and metabolic pathway associated with toxin production or detoxification is likely to have been pivotal in creating the diversity of fungi and other osmotrophic organisms which is in turn important for the ecosystem. It is therefore likely that the acquisition of genes involved in these functions by horizontal gene transfer have been under strong selective pressure, as it allows recipients to spread to new environment and utilise new food sources. In agreement with this hypothesis, re-analysis of the published evidences of lateral gene transfer into

fungi and oomycetes indicate that 32% (out of 370 transfer events into these osmotrophic organisms) have functions involved in public good interactions including secreted enzymes, transporters and toxin biosynthesis systems(Richards & Talbot 2013). This seems in good agreement with the increasing list of reports linking mechanisms driving genome variation and the evolution of social traits in microorganisms.

This indicates the importance and pivotal role of lateral genetic exchange in fungi, not only for the fungus itself, but also to other organisms living in the ecosystem.

Our results presented in chapter six indicate, in a manmade environment:

-the multiple occurrence of horizontal gene transfers in *Penicillium* species.

-These transfers account for over two percent of the genome in *Penicillium roqueforti* at a single locus.

-A very large genomic island with 250 genes, some of them appearing to be involved in antimicrobial functions (250 genes, many transcripts identified).

-These transfers affected many *Penicillium* strains mostly from the food environment.

-The potential promotion of these transfers by the activity of Man.

In light of these results and the trends presented above, it seems very important to assess the impact of horizontal gene transfer in fungi, in the cheese environment, in other manmade environment such as crops, but also in nature. The importance of lateral gene transfer in shaping fungal genomes, as well as its importance in shaping fungal interaction with ecosystems should be further investigated. To get a clear picture of this phenomenon, one should investigate how many genes in any given organism have been transferred, and what is their contribution to protein interaction networks. The fate of the transferred sequence depends on interaction into hosts regulatory networks and subsequent molecular interactions of the encoded products, as well as by selection through evolutionary mechanisms (Skippington & Ragan 2011).

Our results indicate that horizontal gene transfers occurred frequently and recently in a manmade environment and that they brought a large number of genes into species of industrial importance. Food-making practices and processes could be responsible for the promotion of such transfers. This could have dramatic impacts on food safety procedures, especially when considering both the fact that transfers may be biased towards the production of secondary metabolites (some of which are

known to be toxic), and also considering the vast repertoire of secondary metabolites the fungal genomes already seem to possess. It seems important to assess the frequency of such horizontal gene transfers in the food environment. Indeed the rapid acquisition of novel functions leading to competitive advantages and metabolic innovation could have dramatic importance both for engineering new food-adapted organisms for consumption and regarding the safe use of microorganisms in food. If Man is indeed promoting these horizontal transfers and through this inadvertently contributes to shaping new genomes, then we should proceed with extra care and be well aware of the need for an urgent rethinking of the impact of this phenomenon and how it should or should not affect our practices in the food, agriculture, biotechnological and pharmaceutical industries.

The availability of over ten *Penicillium* genomes from diverse environments including many produced during the course of the Food Microbiomes project provides a unique opportunity to understand horizontal transfer in fungi and its contribution to genome dynamics. Since the high frequency of transfer seems likely in *Penicillium spp.* It should be investigated whether this is a *Penicillium* specific trend, or whether it is widespread and generalised. However, the growing body of data about horizontal gene transfer in fungi seems to point at a more widespread phenomenon. The assessment of the frequency of transfer could first be done in the *Penicillium* taxon, but should obviously be assessed in all fungi. Large genome sequencing projects like the 1000 fungal genome project provides a unique opportunity to perform this work.

The assessment of transfer frequency and importance should also be performed in other man controlled environments, such as crops for instance. Many examples of transfers in these environments exists. The emergence of pathogen species resulting in dramatic losses in agricultural yields, and the associated economic and social impact already provides a good incentive to further investigate this lead.

Even less is known about the frequency and occurrence of horizontal transfer in natural environments. The number of genes transferred varies widely from one report to another. While some published cases involve the transfer of a single gene, some other examples described the acquisition of a complete secondary metabolic pathway, whole chromosome transfers, or reports like ours in which many genes are transferred (and expressed in the recipient cell). This raises question about the extent to which horizontal genetic exchange has had and continues to have in shaping fungal genomes. Answering this question would greatly increase our knowledge on the evolution of fungal genomes.

A parallel can be drawn between prokaryotic and eukaryotic genetic exchange, especially when considering the impact of this phenomenon in the spread of antibiotic resistance and emergence of multi resistant strains. Despite not likely as frequent as in prokaryotes, the emergence of new functions in ecosystems by acquisition of genetic material in fungi could lead to shifts in ecosystem balance and perturbation or changes in the ecosystem itself. This could have important economical and societal impacts. In our present day era of globalisation and intensive industrialisation, transfers occurring between previously isolated species could come into contact and by horizontal transfer lead to the emergence of new pathogens. This could lead or contribute to the rapid emergence of new fungal disease (which has already been witnessed (Fisher et al. 2012)) of importance to health, agriculture or with impacts on species conservation and biodiversity. However high the frequency of transfer in fungi many cases of transfer with impacts on health or economy have been published and a worrisome scenario can easily be imagined. Furthermore, the occurrence of transfers in manmade environment is undeniable and may occur at a higher frequency than in nature. These highly selective environments could act as reservoirs for innovation by transfer. Were this to be the case, then new strains could acquire selective advantages by man-imposed selective pressure in such "reservoir" environments and be further transferred to natural environment. This could lead to changes in ecosystem interactions with potentially dramatic effects, such as extirpation or extinction of natural species. For these reasons, it seems of utmost importance to dig deeper towards confirming or refuting the hypothesis under which horizontal genetic exchange in manmade environment is frequent and leads to innovation of novel and potentially unwanted traits. The question of whether man actively promotes the emergence of more competitive organisms should also be investigated and a reflection on the industrial practices in the use of fungi by Man envisioned to anticipate potentially required changes in practices.

Bibliography

Acuña, R. et al., 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proceedings of the National Academy of Sciences of the United States of America*, 109(11), pp.4197–202.

Alkan, C., Sajjadian, S. & Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1), pp.61–65.

Allemand, J.F. et al., 1997. pH-dependent specific binding and combing of DNA. *Biophysical journal*, 73(4), pp.2064–70.

Alsmark, C. et al., 2013. Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biology*, 14(2), p.R19.

Altincicek, B., Kovacs, J.L. & Gerardo, N.M., 2012. Horizontally transferred fungal carotenoid genes in the two-spotted spider mite Tetranychus urticae. *Biology letters*, 8(2), pp.253–7.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–10.

Andersson, J.O., 2009. Gene transfer and diversification of microbial eukaryotes. *Annual review of microbiology*, 63, pp.177–93.

Anon, 1000 fungal genomes project.

Anon, Broad institute, Current FGI sequence projects.

Anon, Yeasts in Extreme Environments. Available at: http://libra.msra.cn/Publication/47884554/yeasts-in-extreme-environments.

Ansorge, W. et al., 1986. A non-radioactive automated method for DNA sequence determination. *Journal of biochemical and biophysical methods*, 13(6), pp.315–23.

Arnold, A.E. et al., 2003. Fungal endophytes limit pathogen damage in a tropical tree. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26), pp.15649–54.

Avery, O.T., Macleod, C.M. & McCarty, M., 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of experimental medicine*, 79(2), pp.137–58.

Baker-Austin, C. et al., 2006. Co-selection of antibiotic and metal resistance. *Trends in microbiology*, 14(4), pp.176–82.

Bennett, S., 2004. Solexa Ltd. *Pharmacogenomics*, 5(4), pp.433–8.

Bensimon, D. et al., 1995. Stretching DNA with a receding meniscus: Experiments and models. *Physical review letters*, 74(23), pp.4754–4757.

Berg, M.A. Van Den et al., 2008. Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. *Nature Biotechnology*, 26(10), pp.1161–1168.

Blackwell, M., 2011. The fungi: 1, 2, 3 ... 5.1 million species? *American journal of botany*, 98(3), pp.426–38.

Blehert, D.S. et al., 2009. Bat white-nose syndrome: an emerging fungal pathogen? *Science (New York, N.Y.)*, 323(5911), p.227.

Boto, L., 2010. Horizontal gene transfer in evolution: facts and challenges. *Proceedings. Biological sciences / The Royal Society*, 277(1683), pp.819–27.

Brakhage, A.A., 2013. Regulation of fungal secondary metabolism. *Nature reviews. Microbiology*, 11(1), pp.21–32.

Branton, D. et al., 2008. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10), pp.1146–53.

Braumann, I., van den Berg, M. & Kempken, F., 2008. Repeat induced point mutation in two asexual fungi, Aspergillus niger and Penicillium chrysogenum. *Current Genetics*, 53(5), pp.287–97.

Bundock, P. et al., 1995. Trans-kingdom T-DNA transfer from Agrobacterium tumefaciens to Saccharomyces cerevisiae. *The EMBO journal*, 14(13), pp.3206–14.

Burrus, V., Marrero, J. & Waldor, M.K., 2006. The current ICE age: biology and evolution of SXT-related integrating conjugative elements. *Plasmid*, 55(3), pp.173–83.

Butler, J. et al., 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5), pp.810–20.

Caburet, S. et al., 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures. *Genome research*, 15(8), pp.1079–85.

Carattoli, A. et al., 2005. Identification of plasmids by PCR-based replicon typing. *Journal of microbiological methods*, 63(3), pp.219–28.

Carver, T.J. et al., 2005. ACT: the Artemis Comparison Tool. *Bioinformatics (Oxford, England)*, 21(16), pp.3422–3.

Chapman, J. a et al., 2010. The dynamic genome of Hydra. *Nature*, 464(7288), pp.592–6.

Chávez, R. et al., 2001. Electrophoretic karyotype of the filamentous fungus Penicillium purpurogenum and chromosomal location of several xylanolytic genes. *FEMS microbiology letters*, 205(2), pp.379–83.

Chayot, R. et al., 2010. An end-joining repair mechanism in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 107(5), pp.2141–6.

Cheeseman, K. et al., 2012. A diagnostic genetic test for the physical mapping of germline rearrangements in the susceptibility breast cancer genes BRCA1 and BRCA2. *Human Mutation*, 33(6), pp.998–1009.

Chen, M. et al., 2002. An Integrated Physical and Genetic Map of the Rice Genome. *Society*, 14(March), pp.537–545.

Ciccarelli, F.D. et al., 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, 311(5765), pp.1283–7.

Cichewicz, R.H., 2010. Epigenome manipulation as a pathway to new natural product scaffolds and their congeners. *Natural product reports*, 27(1), pp.11–22.

Conti, C. & Bensimon, A., 2002. A Combinatorial Approach for Fast, High-Resolution Mapping. *Genomics*, 80(2), pp.135–137.

Dagan, T., Artzy-Randrup, Y. & Martin, W., 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29), pp.10039–44.

Delwen, S., 1996. Archaeology of Ancient Egyptian Beer. *J. am. soc. brew. chem*.

Dolgin, E., 2009. Human genomics: The genome finishers. *Nature*, 462(7275), pp.843–5.

Dong, Y. et al., 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nature biotechnology*, 31(2), pp.135–41.

Doolittle, W.F., 1999. Lateral genomics. *Trends in cell biology*, 9(12), pp.M5–8.

Doolittle, W.F. & Bapteste, E., 2007. Pattern pluralism and the Tree of Life hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), pp.2043–9.

Dorman, C.J., 2007. H-NS, the genome sentinel. *Nature reviews. Microbiology*, 5(2), pp.157–61.

Duan, Z. et al., 2009. A phosphoketolase Mpk1 of bacterial origin is adaptively required for full virulence in the insect-pathogenic fungus Metarhizium anisopliae. *Environmental microbiology*, 11(9), pp.2351–60.

Dunning Hotopp, J.C. et al., 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science (New York, N.Y.)*, 317(5845), pp.1753–6.

Efcavitch, J.W. & Thompson, J.F., 2010. Single-molecule DNA analysis. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 3, pp.109–28.

Eid, J. et al., 2009. Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, 323(5910), pp.133–8.

Ellison, C.E. et al., 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), pp.1–10.

Evans, H.C., Elliot, S.L. & Hughes, D.P., 2011. Hidden diversity behind the zombie-ant fungus Ophiocordyceps unilateralis: four new species described from carpenter ants in Minas Gerais, Brazil. *PloS one*, 6(3), p.e17024.

Ewing, B. & Green, P., 1998. *Base-calling of automated sequencer traces using phred. II. Error probabilities.*,

Faraut, T. et al., 2009. Contribution of radiation hybrids to genome mapping in domestic animals. *Cytogenetic and genome research*, 126(1-2), pp.21–33.

Fisher, M.C. et al., 2012. Emerging fungal threats to animal, plant and ecosystem health. *Nature*, 484(7393), pp.186–94.

Fitzpatrick, D. a, 2012. Horizontal gene transfer in fungi. *FEMS microbiology letters*, 329(1), pp.1–8.

Fitzpatrick, D.A., Logue, M.E. & Butler, G., 2008. Evidence of recent interkingdom horizontal gene transfer between bacteria and Candida parapsilosis. *BMC evolutionary biology*, 8, p.181.

Fleischmann, R.D. et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223), pp.496–512.

Foissac, S. et al., 2008. Genome annotation in plants and fungi : EuGène as a model platform. *Current Bioinformatics*, 3, pp.87–97.

Fournier, G.P. & Gogarten, J.P., 2008. Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. *Journal of bacteriology*, 190(3), pp.1124–7.

Friesen, T.L. et al., 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature genetics*, 38(8), pp.953–6.

Frisvad, J.C. et al., 2007. Secondary metabolite profiling, growth profiles and other tools for species recognition and important Aspergillus mycotoxins. *Studies in mycology*, 59, pp.31–7.

Gabriela Roca, M., Read, N.D. & Wheals, A.E., 2005. Conidial anastomosis tubes in filamentous fungi. *FEMS Microbiology Letters*, 249(2), pp.191–8.

Galagan, J.E. & Selker, E.U., 2004. RIP: the evolutionary cost of genome defense. *Trends in Genetics : TIG*, 20(9), pp.417–423.

Garcia-Vallvé, S., Romeu, A. & Palau, J., 2000. Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Molecular biology and evolution*, 17(3), pp.352–61.

Gauslaa, Y., 1997. Rikkinen, J. 1995. What's behind the pretty colours? A study on the photobiology of lichens. *Nordic Journal of Botany*, 17(5), pp.556–556.

Ge, F., Wang, L.-S. & Kim, J., 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS biology*, 3(10), p.e316.

George, P., Sharakhova, M. V & Sharakhov, I. V, 2012. High-throughput physical mapping of chromosomes using automated in situ hybridization. *Journal of visualized experiments : JoVE*, (64).

Gibbons, J.G. et al., 2012. The evolutionary imprint of domestication on genome variation and function of the filamentous fungus Aspergillus oryzae. *Current Biology : CB*, 22(15), pp.1403–9.

Gillings, M.R. & Stokes, H.W., 2012. Are humans increasing bacterial evolvability? *Trends in ecology & evolution*, 27(6), pp.346–52.

Giraud, F. et al., 2010. Microsatellite loci to recognize species for the cheese starter and contaminating strains associated with cheese manufacturing. *International Journal of Food Microbiology*, 137(2-3), pp.204–13.

Gladyshev, E.A., Meselson, M. & Arkhipova, I.R., 2008. Massive horizontal gene transfer in bdelloid rotifers. *Science (New York, N.Y.)*, 320(5880), pp.1210–3.

Glass, N. & Donaldson, G., 1995. Development of primer sets designed for use with the PCR to amplify conserved genes from filamentous ascomycetes. *Applied and Environmental Microbiology*, 61(4), pp.1323–1330.

Gogarten, J.P., Doolittle, W.F. & Lawrence, J.G., 2002. Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution*, 19(12), pp.2226–38.

Graham, L.A. et al., 2008. Lateral transfer of a lectin-like antifreeze protein gene in fishes. *PloS one*, 3(7), p.e2616.

Griffin, D.H., 1996. *Fungal Physiology*,

Griffith, F., 1929. The Significance of Pneumococcal Types. *Journal of Hygiene*, 27(02), p.113.

Gurney-Dixon, S., 1919. transmutation of bacteria. *cambridge university press*, (XVIII), p.179.

Haas, F.L. et al., 1956. Heterokaryosis as a cause of culture rundown in Penicillium. *Applied microbiology*, 4(4), pp.187–95.

Hall, C., Brachat, S. & Dietrich, F.S., 2005. Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. *Eukaryotic cell*, 4(6), pp.1102–15.

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.*, 41, pp.95–98.

Hane, J.K. et al., 2011. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome biology*, 12(5), p.R45.

Hane, J.K. & Oliver, R.P., 2008. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinformatics*, 9, p.478.

Hansen, F.T. et al., 2012. Quick guide to polyketide synthase and nonribosomal synthetase genes in Fusarium. *International journal of food microbiology*, 155(3), pp.128–36.

Harrington, T.C., Fraedrich, S.W. & Aghayeva, D.N., 2008. Raffaelea lauricola ,. , 104(June), pp.399–404.

Harris, T.D. et al., 2008. Single-molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, 320(5872), pp.106–9.

Hastie, A.R. et al., 2013. Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex Aegilops tauschii Genome. *PloS one*, 8(2), p.e55864.

Hawksworth, D.L., 1991. The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycological Research*, 95(6), pp.641–655.

Hegedüs, N. et al., 2011. The paf gene product modulates asexual development in Penicillium chrysogenum. *Journal of basic microbiology*, 51(3), pp.253–62.

Heinemann, J.A. & Sprague, G.F., 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature*, 340(6230), pp.205–9.

Herrick, J. & Bensimon, A., 2009. Introduction to molecular combing: genomics, DNA replication, and cancer. *Methods in molecular biology (Clifton, N.J.)*, 521, pp.71–101.

Hibbett, D.S. et al., 2007. A higher-level phylogenetic classification of the Fungi. *Mycological research*, 111(Pt 5), pp.509–47.

Hibbett, D.S. & Taylor, J.W., 2013. Fungal systematics: is a new age of enlightenment at hand? *Nature reviews. Microbiology*, 11(2), pp.129–33.

Hillebrand, H., 2004. On the generality of the latitudinal diversity gradient. *The American naturalist*, 163(2), pp.192–211.

Hoffmeister, D. & Keller, N.P., 2007. Natural products of filamentous fungi: enzymes, genes, and their regulation. *Natural product reports*, 24(2), pp.393–416.

Honegger, R., 2007. The Lichen Symbiosis—What is so Spectacular about it? *The Lichenologist*, 30(03), p.193.

Horvath, P. & Barrangou, R., 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science (New York, N.Y.)*, 327(5962), pp.167–70.

Hughes, D.P. et al., 2011. Behavioral mechanisms and morphological symptoms of zombie ants dying from fungal infection. *BMC ecology*, 11, p.13.

Huson, D.H. et al., 2001. Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics (Oxford, England)*, 17 Suppl 1, pp.S132–9.

Inglis, D.O. et al., 2013. Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of Aspergillus nidulans, A. fumigatus, A. niger and A. oryzae. *BMC microbiology*, 13, p.91.

Jablonowski, D. & Schaffrath, R., 2007. Zymocin, a composite chitinase and tRNase killer toxin from yeast. *Biochemical Society transactions*, 35(Pt 6), pp.1533–7.

Jain, R. et al., 2003. Horizontal gene transfer accelerates genome innovation and evolution. *Molecular biology and evolution*, 20(10), pp.1598–602.

Jain, R., Rivera, M.C. & Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), pp.3801–6.

James, T.Y. et al., 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443(7113), pp.818–22.

Jeck, W.R. et al., 2007. Extending assembly of short DNA sequences to handle error. *Bioinformatics (Oxford, England)*, 23(21), pp.2942–4.

Jobb, G., von Haeseler, A. & Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evolutionary Biology*, 4, p.18.

Johnsborg, O., Eldholm, V. & Håvarstein, L.S., 2007. Natural genetic transformation: prevalence, mechanisms and function. *Research in microbiology*, 158(10), pp.767–78.

Johnson, C.J. et al., 2011. Degradation of the disease-associated prion protein by a serine protease from lichens. *PloS one*, 6(5), p.e19836.

Jørgensen, T.R. et al., 2011. The molecular and genetic basis of conidial pigmentation in Aspergillus niger. *Fungal genetics and biology : FG & B*, 48(5), pp.544–53.

Kaiserer, L. et al., 2003. Characterization of the Penicillium chrysogenum antifungal protein PAF. *Archives of microbiology*, 180(3), pp.204–210.

Kapuscinski, J., 1995. DAPI: a DNA-specific fluorescent probe. *Biotechnic and Histochemistry*, 70(5), pp.220–233.

Keeling, P.J., 2009. Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Current opinion in genetics & development*, 19(6), pp.613–9.

Keeling, P.J. & Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics*, 9(8), pp.605–18.

Khaldi, N. et al., 2010. SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal genetics and biology : FG & B*, 47(9), pp.736–41.

Koonin, E. V, Makarova, K.S. & Aravind, L., 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annual review of microbiology*, 55, pp.709–42.

Koren, S. et al., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7), pp.693–700.

Kurtz, S. et al., 2004. Versatile and open software for comparing large genomes. *Genome biology*, 5(2), p.R12.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Larsen, T.O., Gareis, M. & Frisvad, J.C., 2002. Cell cytotoxicity and mycotoxin and secondary metabolite production by common penicillia on cheese agar. *Journal of agricultural and food chemistry*, 50(21), pp.6148–52.

Latreille, P. et al., 2007. Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC genomics*, 8, p.321.

Lawrence, J.G. & Retchless, A.C., 2009. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. *Methods in molecular biology (Clifton, N.J.)*, 532, pp.29–53.

Lebofsky, R. et al., 2006. DNA Replication Origin Interference Increases the Spacing between Initiation Events in Human Cells □. , 17(December), pp.5337–5345.

Lebofsky, R. & Bensimon, A., 2003. Single DNA molecule analysis: applications of molecular combing. *Briefings in functional genomics & proteomics*, 1(4), pp.385–96.

Levy-Booth, D.J. et al., 2007. Cycling of extracellular DNA in the soil environment. *soil biology and biochemistry*, 39(12), pp.2977–2991.

Levy-Sakin, M. & Ebenstein, Y., 2013. Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Current opinion in biotechnology*, 24(4), pp.690–8.

Lewin, H. a et al., 2009. Every genome sequence needs a good map. *Genome research*, 19(11), pp.1925–8.

Ma, L.-J. et al., 2010. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature*, 464(7287), pp.367–73.

Mahiet, C. et al., 2012. Structural variability of the herpes simplex virus 1 genome in vitro and in vivo. *Journal of virology*, 86(16), pp.8592–601.

Makarova, K.S., Wolf, Y.I. & Koonin, E. V, 2013. Comparative genomics of defense systems in archaea and bacteria. *Nucleic acids research*, 41(8), pp.4360–77.

Van Mameren, J., Peterman, E.J.G. & Wuite, G.J.L., 2008. See me, feel me: methods to concurrently visualize and manipulate single DNA molecules and associated proteins. *Nucleic acids research*, 36(13), pp.4381–9.

Marcet-Houben, M. et al., 2012a. Genome sequence of the necrotrophic fungus Penicillium digitatum, the main postharvest pathogen of citrus. *BMC Genomics*, 13, p.646.

Marcet-Houben, M. et al., 2012b. *Genome sequence of the necrotrophic fungus Penicillium digitatum, the main postharvest pathogen of citrus.*,

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–80.

Marilley, L. & Casey, M.G., 2004. Flavours of cheese products: metabolic pathways, analytical tools and identification of producing strains. *International journal of food microbiology*, 90(2), pp.139–59.

Matthews, D.E. & Van Etten, H.D., 1983. Detoxification of the phytoalexin pisatin by a fungal cytochrome P-450. *Archives of biochemistry and biophysics*, 224(2), pp.494–505.

Maxam, a M. & Gilbert, W., 1992. A new method for sequencing DNA. 1977. *Biotechnology (Reading, Mass.)*, 24(2), pp.99–103.

May, G.S. & Adams, T.H., 1997. The Importance of Fungi to Man. *Genome Res.*, 7(11), pp.1041–1044.

Mehrabi, R. et al., 2011. Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. *FEMS microbiology reviews*, 35(3), pp.542–54.

Mentel, M. et al., 2006. Transfer of genetic material between pathogenic and food-borne yeasts. *Applied and environmental microbiology*, 72(7), pp.5122–5.

Moore, D., Robson, G.D. & Trinci, A.P.J., 2011. *21st Century Guidebook to Fungi with CD*, Cambridge University Press.

Moran, N.A. & Jarvik, T., 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science (New York, N.Y.)*, 328(5978), pp.624–7.

Nagahama, T. et al., 2006. Phylogenetic relationship within the Erythrobasidium clade: molecular phylogenies, secondary structure, and intron positions inferred from partial sequences of ribosomal RNA and elongation factor-1alpha genes. *The Journal of general and applied microbiology*, 52(1), pp.37–45.

Nagarajan, N. & Pop, M., 2009. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 16(7), pp.897–908.

Nagarajan, N. & Pop, M., 2013. Sequence assembly demystified. *Nature reviews. Genetics*, 14(3), pp.157–67.

Nakamura, Y. et al., 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature genetics*, 36(7), pp.760–6.

Navarre, W.W. et al., 2007. Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes & development*, 21(12), pp.1456–71.

Neafsey, D.E. et al., 2010. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome research*, 20(7), pp.938–46.

Nevoigt, E., Fassbender, A. & Stahl, U., 2000. Cells of the yeast Saccharomyces cerevisiae are transformable by DNA under non-artificial conditions. *Yeast (Chichester, England)*, 16(12), pp.1107–10.

Nguyen, K. et al., 2011. Molecular combing reveals allelic combinations in facioscapulohumeral dystrophy. *Annals of neurology*, 70(4), pp.627–33.

Novichkov, P.S. et al., 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *Journal of bacteriology*, 186(19), pp.6575–85.

Novo, M. et al., 2009a. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38), pp.16333–8.

Novo, M. et al., 2009b. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38), pp.16333–8.

O'Brien, H.E. et al., 2005. Fungal community analysis by large-scale sequencing of environmental samples. *Applied and environmental microbiology*, 71(9), pp.5544–50.

O'Brien, M. et al., 2006. Mycotoxins and other secondary metabolites produced in vitro by Penicillium paneum Frisvad and Penicillium roqueforti Thom isolated from baled grass silage in Ireland. *Journal of agricultural and food chemistry*, 54(24), pp.9268–76.

Ochman, H., Lawrence, J.G. & Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784), pp.299–304.

Ohmido, N., Fukui, K. & Kinoshita, T., 2010. Recent advances in rice genome and chromosome structure research by fluorescence in situ hybridization (FISH). *Proceedings of the Japan Academy. Series B, Physical and biological sciences*, 86(2), pp.103–16.

Palumbo, E. et al., 2010. Replication dynamics at common fragile site FRA6E. *Chromosoma*, 119(6), pp.575–87.

Paszkiewicz, K. & Studholme, D.J., 2010. De novo assembly of short sequence reads. *Briefings in bioinformatics*, 11(5), pp.457–72.

Paul M. Kirk, J. A. Stalpers, David W. Minter, P.F.C., 2011. *Dictionnary of the Fungi*,

Peláez, F., 2006. The historical delivery of antibiotics from microbial natural products--can history repeat? *Biochemical pharmacology*, 71(7), pp.981–90.

Pennisi, E., 2009. DNA sequencing. No genome left behind. *Science (New York, N.Y.)*, 326(5954), pp.794–5.

Perry, G.H. et al., 2007. Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10), pp.1256–60.

Persson, F. et al., 2009. Local conformation of confined DNA studied using emission polarization anisotropy. *Small (Weinheim an der Bergstrasse, Germany)*, 5(2), pp.190–3.

Pevzner, P.A., Tang, H. & Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp.9748–53.

Pisani, D., Cotton, J.A. & McInerney, J.O., 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular biology and evolution*, 24(8), pp.1752–60.

Pitt, J.I. & Hocking, A.D., 2009. *Fungi and Food Spoilage*, New York: Springer.

Pop, M. & Kosack, D., 2004. Using the TIGR assembler in shotgun sequencing projects. *Methods in molecular biology (Clifton, N.J.)*, 255, pp.279–94.

Posada, D., 2008. jModelTest 0.1.1. *Systematic Biology*, 1(April), pp.1–23.

Pressel, S. et al., 2010. Fungal symbioses in bryophytes : New insights in the Twenty First Century. , 253, pp.238–253.

Price, M.N., Dehal, P.S. & Arkin, A.P., 2008. Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome biology*, 9(1), p.R4.

Prudhomme, M., Libante, V. & Claverys, J.-P., 2002. Homologous recombination at the border: insertion-deletions and the trapping of foreign DNA in Streptococcus pneumoniae. *Proceedings of the National Academy of Sciences of the United States of America*, 99(4), pp.2100–5.

Quesneville, H. et al., 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS computational biology*, 1(2), pp.166–75.

Rausch, C. et al., 2005. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic acids research*, 33(18), pp.5799–808.

Van Reenen, C.A. & Dicks, L.M.T., 2011. Horizontal gene transfer amongst probiotic lactic acid bacteria and other intestinal microbiota: what are the possibilities? A review. *Archives of microbiology*, 193(3), pp.157–68.

Reisinger, J. et al., 2006. Visualization of episomal and integrated Epstein-Barr virus DNA by fiber fluorescence in situ hybridization. *International journal of cancer. Journal international du cancer*, 118(7), pp.1603–8.

Richards, T. a. et al., 2011. Gene transfer into the fungi. *Fungal Biology Reviews*, 25(2), pp.98–110.

Richards, T.A., 2011. Genome evolution: horizontal movements in the fungi. *Current biology : CB*, 21(4), pp.R166–8.

Richards, T.A. & Talbot, N.J., 2013. Horizontal gene transfer in osmotrophs: playing with public goods. *Nature reviews. Microbiology*, 11(10), pp.720–7.

Ricker, N., Qian, H. & Fulthorpe, R.R., 2012. The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics*, 100(3), pp.167–75.

Rivera, M.C. & Lake, J.A., 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431(7005), pp.152–5.

Rizzo, L. et al., 2013. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review. *The Science of the total environment*, 447, pp.345–60.

Robinson, G.E. et al., 2011. Creating a buzz about insect genomes. *Science (New York, N.Y.)*, 331(6023), p.1386.

Rogers, S. & Bendich, A., 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Molecular Biology*, 5(2), pp.69–76.

Rogers, S. & Bendich, A., 1988. Extraction of DNA from plant tissues. In S. Gelvin & R. Schilperoot, eds. *Plant Molecular Biology Manual*. Boston, MA: Kluwer Academic Publishers, pp. A6:1–10.

Rogers, S. & Bendich, A., 1994. Extraction of DNA from plant, fungal and algal tissues. In S. Gelvin & R. Schilperoort, eds. *Plant Molecular Biology Manual*. Boston, MA: Kluwer Academic Publishers, pp. D1:1–8.

Ropars, Jeanne et al., 2012. A taxonomic and ecological overview of cheese fungi. *International Journal of Food Microbiology*, 155(3), pp.199–210.

Ropars, J et al., 2012. Sex in cheese: evidence for sexuality in the fungus Penicillium roqueforti. *PloS one*, 7(11), p.e49665.

Rosewich, U.L. & Kistler, H.C., 2000. Role of Horizontal Gene Transfer in the Evolution of Fungi. *Annual review of phytopathology*, 38, pp.325–363.

Samson, R.A., 2000. *Introduction to food-and airborne fungi* Centraalbureau voor Schimmelcultures, ed.,

Sanders, I.R., 2006. Rapid disease emergence through horizontal gene transfer between eukaryotes. *Trends in ecology & evolution*, 21(12), pp.656–8.

Sanger, F., 1975. The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, 191(1104), pp.317–33.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–7.

Sboner, A. et al., 2011. The real cost of sequencing: higher than you think! *Genome biology*, 12(8), p.125.

Schatz, M.C., Delcher, A.L. & Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9), pp.1165–73.

Schein, J. et al., 2004. High-throughput BAC fingerprinting. *Methods in molecular biology (Clifton, N.J.)*, 255, pp.143–56.

Schiex, T., Moisan, A. & Rouzé, P., 2001. EuGène: an eukaryotic gene finder that combines several sources of evidence. *Lect Notes Comput Sci*, 2066, pp.111–125.

Schlüter, A. et al., 2008. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *Journal of biotechnology*, 136(1-2), pp.77–90.

Schüβler, A., Schwarzott, D. & Walker, C., 2001. A new fungal phylum, the Glomeromycota: phylogeny and evolution* *Dedicated to Manfred Kluge (Technische Universität Darmstadt) on the occasion of his retirement. *Mycological Research*, 105(12), pp.1413–1421.

Schwartz, D.C. & Samad, A., 1997. Optical mapping approaches to molecular genomics. *Current opinion in biotechnology*, 8(1), pp.70–4.

Shendure, J. et al., 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, 309(5741), pp.1728–32.

Skippington, E. & Ragan, M.A., 2011. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS microbiology reviews*, 35(5), pp.707–35.

Slater, F.R. et al., 2008. Progress towards understanding the fate of plasmids in bacterial communities. *FEMS microbiology ecology*, 66(1), pp.3–13.

Slot, J.C. & Hibbett, D.S., 2007. Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: a phylogenetic study. *PloS one*, 2(10), p.e1097.

Slot, J.C. & Rokas, A., 2011. Horizontal Transfer of a Large and Highly Toxic Secondary Metabolic Gene Cluster between Fungi. *Current biology : CB*, 21(2), pp.134–9.

Smith, L.M. et al., Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071), pp.674–9.

Smith, M.L., Bruhn, J.N. & Anderson, J.B., 1992. The fungus Armillaria bulbosa is among the largest and oldest living organisms. *Nature*, 356(6368), pp.428–431.

Sorek, R. et al., 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science (New York, N.Y.)*, 318(5855), pp.1449–52.

Sorokin, V.A., Gelfand, M.S. & Artamonova, I.I., 2010. Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Applied and environmental microbiology*, 76(7), pp.2136–44.

Steenkamp, E.T., Wright, J. & Baldauf, S.L., 2006. The protistan origins of animals and fungi. *Molecular biology and evolution*, 23(1), pp.93–106.

Stergiopoulos, I. et al., 2012. In silico characterization and molecular evolutionary analysis of a novel superfamily of fungal effector proteins. *Molecular biology and evolution*, 29(11), pp.3371–84.

Storteboom, H. et al., 2010. Identification of antibiotic-resistance-gene molecular signatures suitable as tracers of pristine river, urban, and agricultural sources. *Environmental science & technology*, 44(6), pp.1947–53.

Strick, T.R. et al., 1996. The elasticity of a single supercoiled DNA molecule. *Science (New York, N.Y.)*, 271(5257), pp.1835–7.

Swain, M.T. et al., 2012. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nature protocols*, 7(7), pp.1260–84.

Syvanen, M., 2012. Evolutionary implications of horizontal gene transfer. *Annual review of genetics*, 46, pp.341–58.

Szinay, D. et al., 2008a. High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. *The Plant journal : for cell and molecular biology*, 56(4), pp.627–37.

Szinay, D. et al., 2008b. High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. *The Plant journal : for cell and molecular biology*, 56(4), pp.627–37.

Talbot, N.J., 2007. Plant pathology: deadly special deliveries. *Nature*, 450(7166), pp.41–3.

Tanaka, A. et al., 2012. Fungal endophytes of grasses. *Current Opinion in Plant Biology*, 15(4), pp.462–468.

Tanaka, H. & Kawai, T., 2009. Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. *Nature nanotechnology*, 4(8), pp.518–22.

Taylor, J.W. & Ellison, C.E., 2010. Mushrooms: morphological complexity in the fungi. *Proceedings of the National Academy of Sciences of the United States of America*, 107(26), pp.11655–6.

Técher, H. et al., 2013. Replication Dynamics: Biases and Robustness of DNA Fiber Analysis. *Journal of molecular biology*.

Temporini, E.D. & VanEtten, H.D., 2004. An analysis of the phylogenetic distribution of the pea pathogenicity genes of Nectria haematococca MPVI supports the hypothesis of their origin by horizontal transfer and uncovers a potentially new pathogen of garden pea: Neocosmospora boniensis. *Current genetics*, 46(1), pp.29–36.

Thomas, C.M. & Nielsen, K.M., 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology*, 3(9), pp.711–21.

Thomas, M.B. & Read, A.F., 2007. Fungal bioinsecticide with a sting. *Nature biotechnology*, 25(12), pp.1367–8.

Thompson, J.D., Gibson, T.J. & Higgins, D.G., 2002. Multiple sequence alignment using clustalW and clustalX. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

Todesco, S., 2008. *Tomato sequencing project : sequencing and analysis of chromosome 12*.

Touchon, M. et al., 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS genetics*, 5(1), p.e1000344.

Ueno, Y., 1985. The toxicology of mycotoxins. *Critical reviews in toxicology*, 14(2), pp.99–132.

Untergasser, A. et al., 2012. Primer3--new capabilities and interfaces. *Nucleic acids research*, 40(15), p.e115.

Vandenkoornhuyse, P. et al., 2002. Extensive fungal diversity in plant roots. *Science (New York, N.Y.)*, 295(5562), p.2051.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51.

Vlassov, V. V, Laktionov, P.P. & Rykova, E.Y., 2007. Extracellular nucleic acids. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 29(7), pp.654–67.

Vos, M., 2009. Why do bacteria engage in homologous recombination? *Trends in microbiology*, 17(6), pp.226–32.

Voyles, J. et al., 2009. Pathogenesis of chytridiomycosis, a cause of catastrophic amphibian declines. *Science (New York, N.Y.)*, 326(5952), pp.582–5.

De Vries, J. & Wackernagel, W., 2002. Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 99(4), pp.2094–9.

Wake, D.B. & Vredenburg, V.T., 2008. Colloquium paper: are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proceedings of the National Academy of Sciences of the United States of America*, 105 Suppl , pp.11466–73.

Wedin, M., Döring, H. & Gilenstam, G., 2004. Saprotrophy and lichenization as options for the same fungal species on different substrata: environmental plasticity and fungal lifestyles in the Stictis-Conotrema complex. *New Phytologist*, 164(3), pp.459–465.

Whiteman, N.K. & Gloss, A.D., 2010. Parasitology: Nematode debt to bacteria. *Nature*, 468(7324), pp.641–2.

Woese, C.R. et al., 1975. Conservation of primary structure in 16S ribosomal RNA. *Nature*, 254(5495), pp.83–6.

Woese, C.R., 2000. Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences of the United States of America*, 97(15), pp.8392–6.

Yue, J. et al., 2012. Widespread impact of horizontal gene transfer on plant colonization of land. *Nature communications*, 3, p.1152.

Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5), pp.821–9.

Zhang, N. & Blackwell, M., 2002. Molecular phylogeny of Melanospora and similar pyrenomycetous fungi. *Mycological Research*, 106(2), pp.148–155.

Annexes:

Other scientific contributions realised during the thesis:

Despite not being part of the thesis project, the work presented here both led to the initial idea of exploring the potential contribution of Molecular Combing to the assembly problematic (which resulted in this project), and also represent a significant fraction of the time of this thesis. It also led me to deal with important aspects of a the scientific profession I would not have faced otherwise, namely the examination of patents and intellectual property rules, as the work described hereafter resulted in a publication, one patent application directly the result of this work, and provided the basis for another patent (of which I am only a minor contributor). Noteworthy, the *BRCA* case around the exploitation for diagnostic purposes by Myriad genetics of this sequence is a poster child of conflicts between intellectual property and ethics surrounding genomic data and its exploitation, where academic research, different governmental policies around the world and interest of patient are blending together, and the Molecular Combing technology has had a pivotal role in highlighting the incomplete testing of Myriad's BRCAnalysis diagnostic test, and ultimately led to the contestation by European researchers and institutions of the Myriad patents. More information can be found in the section "The BRCA story" in these annexes.

I started my master II traineeship at Genomic Vision with the task of developing a Molecular Combing diagnostic test for the detection of large genomic rearrangements in the Breast Cancer Associated genes *BRCA1* and *BRCA2.* As presented in chapter five as one of the case in which Molecular Combing can be useful to improve genome assemblies (See chapter five, case n°1), during the *BRCA* project, we realised the locus surrounding the *BRCA1* gene was not completely assembled. A gap of 101 kilobases was present upstream of this gene, and since a large number of genomic rearrangements were known to include both the gene and some additional sequence around the locus, I set up for a Genomic Morse Code also encompassing this 101 kilobase gap. Molecular Combing experiments revealed the at the time available human reference sequence was not correct at this locus, and no gap was present. This misassembly was independently corrected in the next release of the genome assembly by the human genome reference consortium. However the occurrence of such misassemblies next to a locus which is associated a high predisposition to cancer highlighted the importance of improving and validating assemblies. This lane of thought provided the foundation for this thesis project. The next section is made of the article that was published in 2012, and of the front pages of two patents applications filed. The work and problematic is explained in the article.

METHODS

Human Mutation

OFFICIAL JOURNAL

A Diagnostic Genetic Test for the Physical Mapping of Germline Rearrangements in the Susceptibility Breast Cancer Genes *BRCA1* and *BRCA2*

Kevin Cheeseman,[1†] Etienne Rouleau,[2†] Anne Vannier,[1] Aurélie Thomas,[1] Adrien Briaux,[2] Cedrick Lefol,[2] Pierre Walrafen,[1] Aaron Bensimon,[1] Rosette Lidereau,[2] Emmanuel Conseiller,[1] and Maurizio Ceppi[1*]

[1]*Genomic Vision, Bagneux, Paris, France;* [2]*Hôpital René Huguenin, Institut Curie, Laboratoire d'Oncogénétique, Saint-Cloud, France*

Communicated by David E. Goldgar

Received 29 August 2011; accepted revised manuscript 3 February 2012.
Published online 21 February 2012 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22080

ABSTRACT: The *BRCA1* and *BRCA2* genes are involved in breast and ovarian cancer susceptibility. About 2 to 4% of breast cancer patients with positive family history, negative for point mutations, can be expected to carry large rearrangements in one of these two genes. We developed a novel diagnostic genetic test for the physical mapping of large rearrangements, based on molecular combing (MC), a FISH-based technique for direct visualization of single DNA molecules at high resolution. We designed specific Genomic Morse Codes (GMCs), covering the exons, the noncoding regions, and large genomic portions flanking both genes. We validated our approach by testing 10 index cases with positive family history of breast cancer and 50 negative controls. Large rearrangements, corresponding to deletions and duplications with sizes ranging from 3 to 40 kb, were detected and characterized on both genes, including four novel mutations. The nature of all the identified mutations was confirmed by high-resolution array comparative genomic hybridization (aCGH) and breakpoints characterized by sequencing. The developed GMCs allowed to localize several tandem repeat duplications on both genes. We propose the developed genetic test as a valuable tool to screen large rearrangements in *BRCA1* and *BRCA2* to be combined in clinical settings with an assay capable of detecting small mutations.
Hum Mutat 33:998–1009, 2012. © 2012 Wiley Periodicals, Inc.

KEY WORDS: breast cancer; BRCA1; BRCA2; rearrangements; Genomic Morse Code; molecular combing

Introduction

Breast cancer is the most common malignancy in women, affecting approximately 10% of the female population. Incidence rates have increased dramatically for 50 years and it is estimated that about 1.4 million women will be diagnosed with breast cancer annually worldwide and about 460,000 will die from the disease [Jemal et al., 2011]. Germline mutations in the hereditary breast and ovarian cancer susceptibility genes *BRCA1* (MIM# 113705) and *BRCA2* (MIM# 600185) are highly penetrant and account for 5–10% of all breast or ovarian cancer cases [King et al., 2003; Nathanson et al., 2001]. A mutation in one of these two genes confers a 10–20 times increased relative risk of developing a breast cancer, translating into 70–80% risk of developing a breast cancer at age 70 [King et al., 2003]). Screening is important for genetic counseling of individuals with a positive family history and for early diagnosis or prevention in mutation carriers. Most common mutations have a small size, consisting of point mutations, nonsense/frameshifts (small insertions or deletions), missense mutations in conserved domains, or splice-site mutations resulting in aberrant transcript processing [Szabo et al., 2000]. Mutations also include large rearrangements, including deletions and duplications of large genomic regions that escape detection by traditional polymerase chain reaction (PCR)-based mutation screening combined with DNA sequencing [Mazoyer, 2005; Sluiter and van Rensburg, 2011]. It is estimated that between 10 and 15% of the hereditary breast and ovarian cancer cases are imputable to large rearrangements mainly in *BRCA1* but also in *BRCA2* genes. As a consequence, the screening of large rearrangements in both genes has become mandatory, and should be always performed in combination to the screening of point mutations [Puget et al., 1999b; Walsh et al., 2006].

Techniques adapted to detect large rearrangements for routine prescreening and predictive purposes are quantitative multiplex PCR of short fluorescent fragments (QMPSF) [Hofmann et al., 2002], real-time PCR [Barrois et al., 2004], fluorescent DNA microarray assays [Frolov et al., 2002], multiplex ligation-dependent probe amplification (MLPA) [Casilli et al., 2002; Hofmann et al., 2002], qPCR-HRM [Rouleau et al., 2009], and EMMA (enhanced mismatch mutation analysis). However, these routine techniques provide limited information to characterize the mutations. Techniques capable of detecting and characterizing large rearrangements in diagnostic settings include high-resolution oligonucleotide array comparative genomic hybridization (aCGH) [Rouleau et al., 2007; Staaf et al., 2008], followed by PCR and sequencing for the exact characterization of the breakpoints. Of notice, massively parallel sequencing combined with genomic capture has been recently proposed for simultaneous detection of small mutations and large rearrangements of 21 genes involved in breast and ovarian cancer [Walsh et al., 2010]. However, the limitation of NGS (next-generation sequencing) for the detection of large rearrangements has been recently highlighted because of the high content of hard-to-sequence repetitive sequences present in high percentage in both *BRCA1* and

Additional Supporting Information may be found in the online version of this article.

†Both authors contributed equally to this work.

*Correspondence to: Maurizio Ceppi, Genomic Vision, 80–84 rue des Meuniers, 92220 Bagneux, Paris, France. E-mail: m.ceppi@genomicvision.com

© 2012 WILEY PERIODICALS, INC.

BRCA2 [De Leeneer et al., 2011]. Therefore, there is a clear need for alternative technologies combined with the existing ones, capable of detecting efficiently the full spectrum of large rearrangements, including the often problematic tandem repeat duplications.

Molecular combing (MC) is a powerful FISH-based technique for direct visualization of single DNA molecules that are stretched and attached, uniformly and irreversibly, to specially treated glass surfaces [Herrick and Bensimon, 2009; Schurra and Bensimon, 2009]. This technology considerably improves the structural and functional analysis of DNA across the genome and is capable of visualizing multiple genomic regions at high resolution (in the kb range) in a single analysis. MC is particularly suited to the detection of structural variations such as copy number variations (CNVs), translocations, inversions and loss of heterozygosity (LOH) [Caburet et al., 2005], thus extending the spectrum of mutations potentially detectable in breast cancer genes. Of notice, MC has been recently employed in clinical settings to detect and measure the contraction of the repeat array *D4Z4*, associated with the Facioscapulohumeral dystrophy (FSHD), one of the most common hereditary neuromuscular disorders. The MC-based test enabled the accurate diagnosis of 32 FSHD patients and is becoming the reference routine diagnostic method, replacing the techniques employed so far [Nguyen et al., 2011].

MC has already been employed to detect large rearrangements in *BRCA1* [Gad et al., 2001, 2002a, 2003] and *BRCA2* [Gad et al., 2002b], using a first-generation low resolution "color bar coding" screening approach. The originally employed DNA probes (cosmids, PACs and long-range PCR products) also encompassed repetitive sequences particularly abundant at the two loci [Gad et al., 2001, 2002b]. This resulted in the superposition of individual colored signals after probe detection and in strong background noise, undermining the quality of the images and preventing robust measurement of the derived signals.

Here, we describe a substantial technical improvement of the original approach, based notably on the design of second-generation high-resolution *BRCA1* and *BRCA2* Genomic Morse Codes (GMCs). A GMC is a series of "dots or dashes" (corresponding to DNA probes with specific sizes and colors) and "gaps" (corresponding to uncolored gaps with specific sizes located between the DNA probes), designed to physically map and define with a specific "signature" a particular genomic region [Lebofsky et al., 2006]. For *BRCA1* and *BRCA2* GMCs, the majority of the repetitive sequences were eliminated from the DNA probes, thus reducing background noise and permitting robust measurement of the color signal lengths within the GMC. Both GMCs were statistically validated on samples from 50 controls and then tested on 10 patients (index cases) with a positive family history of breast cancer. Large rearrangements were detected on both genes, and the nature of all the identified mutations was confirmed by high-resolution aCGH. Four new large rearrangements in *BRCA1* and *BRCA2* were characterized, demonstrating the robustness of our approach, even for the detection and characterization of hard-to-detect mutations, such as tandem repeat duplications or mutations located in genomic regions rich of repetitive elements (e.g., 5′ region of *BRCA1*). The developed MC-based platform permits simultaneous detection of large rearrangements in *BRCA1* and *BRCA2*, and will be part of a novel diagnostic genetic test for breast and ovarian cancer.

Materials and Methods

Preliminary Patient Screening

The developed GMC were validated on samples from 50 negative controls with no deleterious mutations detected in *BRCA1* or *BRCA2*. The genetic test was validated on 10 samples from patients (index cases) with positive family history of breast cancer and known to bear large rearrangements affecting either *BRCA1* or *BRCA2*. Total human genomic DNA was obtained from PBMCs (peripheral blood monocytic cells) or EBV-immortalized lymphoblastoid cell lines. Preliminary screening for large rearrangements was performed with the QMPSF assay (Quantitative multiplex PCR of short fluorescent fragments) in the conditions described by Casilli et al. and Tournier et al. [Casilli et al., 2002, Tournier et al., 2004] or by means of MLPA (multiplex ligation-dependent probe amplification) using the SALSA MLPA kits P002 (MRC Holland, Amsterdam, The Netherlands) for *BRCA1* and P045 (MRC-Holland) for *BRCA2*. All 60 screened individuals gave their written consent for *BRCA1* and *BRCA2* analysis.

Molecular Combing Procedures

A detailed description of operation procedures is included as Supplementary Material and Methods. Briefly, EBV-immortalized lymphoblastoid cell lines or PBMCs were embedded in agarose plugs. DNA was purified by proteinase K and sarkosyl treatment overnight. Agarose was melted and digested by an overnight beta-agarose treatment. Purified DNA was diluted in MES buffer and combed on coverslips. Probes were initially subcloned by PCR in plasmids, which were used as templates for probe labeling by random priming. Hybridization was performed overnight with biotin-, digoxigenin-, or Alexa Fluor 488-labeled probes, detected using fluorophore-coupled antibody or streptavidin layers. The entire coverslip was scanned by an automated fluorescence microscope. Image analysis and signals measurement was performed using software developed in-house, and statistical analysis was performed as described in the Supp. Materials and Methods. As we are developing a series of MC-based assays for a CLIA labs, we have developed adequate QA/QC procedures (Supp. Fig. S1).

Dedicated Zoom-In CGH Array

A detailed description of operation procedures is included as Supp. Material and Methods. To confirm the large rearrangements detected by MC, a zoom-in CGH array was used as previously described [Rouleau et al., 2007]. For the interpretation of the oligonucleotide signal, the chosen threshold was deleted if the log2 ratio was <–0.4 and duplicated if >0.4. The analytical approach used for zoom-in CGH arrays has been described elsewhere [Rouleau et al., 2007].

Breakpoints Mapping

The breakpoints were characterized by using classical PCR amplification followed by sequencing. The PCR primers were selected in the genomic region surrounding the breakpoints and designed using the Oligo 6 software (Molecular Biology—Insights). We systematically chose two sets of primers (available on request) to nest the PCR. The PCR products were analyzed on agarose gel and then purified and sequenced in both directions by using each PCR primer with the BigDye Terminator Cycle Sequencing Reaction kit (Applied Biosystem) and an ABI Prism 3030 automated sequencer (Applied Biosystems, Foster City, CA).

Nomenclature

The present version of *BRCA1* and *BRCA2* is based on GenBank reference sequences NM_007294.2 (mRNA:U14680) and

NM_000059.3 (mRNA: U43746), respectively. For the *BRCA1* gene, all rearrangements were described with the same orientation than the *BRCA1* gene, telomeric to centromeric sense. So, the 5′ breaking point had a genomic position smaller than the 3′ breaking point. The nomenclature used was the genomic nomenclature based on HGVS recommendation and the build36/hg18 for the chromosome 17 (NC_000017.9). The first nucleotide of the ATG translation start site is +1. All mutations have been submitted to the UMD-*BRCA1* and *BRCA2* database which are fully and freely accessible [Caputo et al., 2012].

Results

Design of High-Resolution *BRCA1* and *BRCA2* GMCs

We have designed high-resolution GMCs covering the exons, the noncoding regions and large genomic portions flanking both genes. Importantly, all repetitive sequences were removed from the DNA probes. We identified 38 genomic regions in the *BRCA1* locus and 32 regions in the *BRCA2* locus that were devoid of repetitive sequences, and that were used to design and clone DNA hybridization probes compatible with the visualization process associated with MC (Supp. Figs. S2 and S3). The name, size, and color of the DNA hybridization probes, and the exons covered by the probes, are listed in Supp. Tables S1 (*BRCA1*) and S2 (*BRCA2*). Adjacent DNA probes of the same color form a signal. Thus, a GMC is composed of a series of colored signals distributed along a specific portion of the genomic DNA. Colors were chosen to create unique nonrepetitive sequences of signals, which differed between *BRCA1* and *BRCA2*. To facilitate GMC recognition and measurement, signals located on the genes were grouped together in specific patterns called "motifs." An electronic reconstruction of the designed *BRCA1* and *BRCA2* GMC is shown in Figure 1. The *BRCA1* GMC covers a region of 200 kb, including the upstream genes *NBR1*, *NBR2*, *LOC100133166*, and *TMEM106A*, as well as the pseudogene Ψ*BRCA1*. The complete *BRCA1* GMC is composed of 18 signals (S1B1–S18B), and the 8 *BRCA1*-specific signals are grouped together in 7 motifs (g1b1-g7b1) (Fig. 1A and B). The *BRCA2* GMC covers a genomic region of 172 kb composed of 14 signals (S1B2–S14B2), and the 7 *BRCA2*-specific signals are grouped together in 5 motifs (g1b2-g5b2) (Fig. 1C and D). Deletions or insertions, if present, are detected in the genomic regions covered by the motifs.

Validation of *BRCA1* and *BRCA2* GMC Signals in Negative Controls

The newly designed GMC were first validated on genomic DNA isolated from 50 negative controls. Typical visualized signals and measured motif lengths for one negative control are reported in Supp. Figure S4 and Supp. Table S3. Importantly, we never observed a "sequencing gap" in the *BRCA1* locus [Staaf et al., 2008] (*LOC100133166* was always located upstream and directly attached to *NBR1*), confirming the validity of the *BRCA1* sequence proposed in the GRChg19 genome assembly and underlining the utility of MC as a tool for physical mapping [Herrick and Bensimon, 2009; Schurra and Bensimon 2009]. For *BRCA1*, we obtained *delta* values (difference between μ and *calculated*) in the range of −0.2 and +0,6 kb, whereas *BRCA2 delta* values were in the range of −0.1 and +0.2 kb, underlining the precision of the developed measurement approach.

Within the 50 control negative samples, 1 false positive sample was found (Supp. Table S3, control nr. 27). This sample is supposed to be negative, since MLPA and aCGH analysis were also performed on it, without finding any mutation. The low number and the low quality of images derived after MC analysis on the same sample, suggested that for motif g2b2 on *BRCA2* the measurement could not be performed efficiently (Supp. Fig. S5A). Such an event is rather infrequent, and it can be verified by repeating the MC assay. In fact, a second MC analysis performed on the control sample nr.27, confirmed the absence of large rearrangements (Supp. Fig. S5B). Thus, based on the analysis performed on 50 control samples, test specificity resulted to be 98%, for rearrangements larger than 2 kb (Supp. Table S4). A larger set of negative samples may be necessary to consolidate our estimation. Calculations were performed according to standard mathematic formulas employed for diagnostics tests [Altman and Bland, 1994].

Characterization of Four Novel *BRCA1* and *BRCA2* Large Rearrangements in Familial Breast Cancer Patients

MC was then applied to 10 samples from patients with a severe family history of breast cancer and known to bear large rearrangements either on *BRCA1* or *BRCA2*. Importantly, the MC analysis was a blind test, meaning that for each patient the identity of the mutation was unknown before the test, since it was revealed to the operator only after having completed the test on all the samples. The *BRCA1* and *BRCA2* GMCs were measured in all 10 samples, and the size of all measured motifs, including the associated statistical analysis, is reported in Supp. Table S5 (see Supp. Materials and Methods for details). Ten different large rearrangements were identified, of which four were novel and six had already been described in the literature (see Table 1). Importantly, as mutations were detected and characterized in all analyzed samples, no false negative samples were found. Thus, within this limited set of samples, sensitivity resulted to be 100%, for rearrangements larger than 2 kb (Supp. Table S4). A larger set of positive samples may be necessary to consolidate our estimation. MC analysis of the four novel large rearrangements is reported in Figure 2. As the identified rearrangements have never been described before in the literature, they were also characterized by zoom-in high-resolution CGH array, and all four breakpoints were determined by sequencing.

Duplication from exon 17 to exon 20 of BRCA2

By visual inspection, the mutation appeared as a tandem duplication of the S6B2 red signal. After measurement, the mutation was estimated to have a size of 13.8 ± 2.0 kb related to the DNA probe *BRCA2-15* and a portion of the uncovered gap between signals S6B2 and S7B2, encoding exons 17 to 20 (Fig. 2A, top panel). This rearrangement was recently reported by Gaudet et al. [Gaudet et al., 2010] but was not fully characterized. The aCGH array gave a size between 8.3 and 12 kb and the mapping of the breakpoints for this rearrangement indicated that the duplicated region was 14,788 bp long (Fig. 2A, bottom panel).

Duplication from exon 5 to exon 7 of BRCA1

By visual inspection, the mutation appeared as a tandem duplication of the S8B1 blue signal. After measurement, the mutation was estimated to have a size of 12 ± 2.6 kb, restricted to a portion of the *BRCA1* gene that encodes exons 5 to 7 (Fig. 2B, top panel). The aCGH array gave a size between 8.4 and 9.4 kb and the mapping of the breakpoints indicated that the duplicated region was 8,128

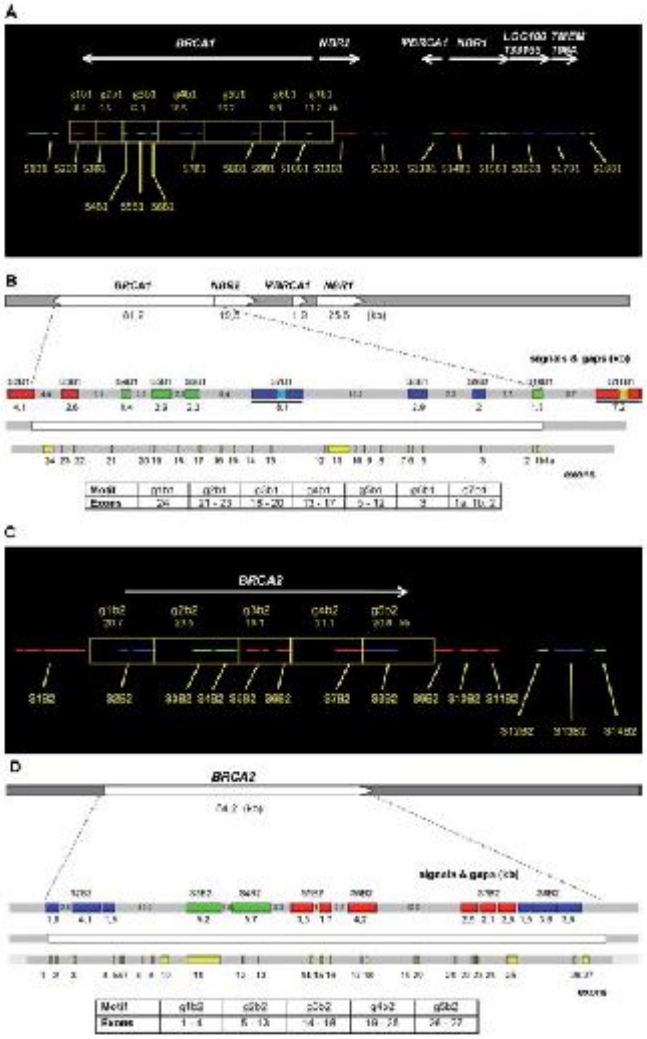


Figure 1. *In silico*-generated GMCs for high-resolution physical mapping of the *BRCA1* and *BRCA2* genomic regions. **A:** The complete *BRCA1* GMC covers a genomic region of 200 kb and is composed of 18 signals (S1B1–S18B) of a distinct color (green, red, or blue). Each signal is composed of one (e.g., S2B1) to three small horizontal bars (e.g., S15B1), each bar corresponding to a single DNA probe. The region encoding the *BRCA1* gene (81.2 kb) is composed of 7 "motifs" (g1b1-g7b1). Each motif is composed of one to three small horizontal bars and a black "gap" (no signal). **B:** Zoom-in on the *BRCA1* gene-specific signals and relative positions of the exons. **C:** The complete *BRCA2* GMC covers a genomic region of 172 kb and is composed of 14 signals (S1B2–S14B2) of a distinct color (green, red, or blue). Each signal is composed of one (e.g., S14B2) to five small horizontal bars (e.g., S1B2). The region encoding the *BRCA2* gene (84.2 kb) is composed of five motifs (g1b2-g5b2). Each motif is composed of two to four small horizontal bars and a black gap. **D:** Zoom-in on the *BRCA2* gene-specific signals and relative positions of the exons. Deletions or insertions, if present, will appear in the region covered by the motifs.

bp long (Fig. 2B, bottom panel). Of notice, the characterization of these two duplications indicate that the developed GMCs allow to unambiguously localize tandem repeats, providing information on the genome location, the nature and the size of the mutation, in one single experiment.

Deletion of exon 3 in BRCA1

By visual inspection, the mutation appeared as a deletion of the S9B1 blue signal. After measurement, the mutation was estimated to have a size of 9.1 ± 1.3 kb, on a portion of the *BRCA1* gene that

Table 1. Description of the 10 Characterized Large Rearrangements as Detected by MLPA, Molecular Combing, High-Resolution aCGH, and by Sequencing

Sample	Gene	MLPA status	Molecular Combing	Zoom-in aCGH	Breakpoints (bp)	Mechanism	Mutation name (HGVS)	Reference
Novel mutations								
BR07	*BRCA1*	Del ex 3	9.1 ± 1.3 kb/Del ex 3	8.3–12.2 kb/Del ex 3	38,526,903–38,515,491 (11,413 bp)	AluY–AluY	c.80+26.57_135-341.54del	Novel
BR10	*BRCA1*	Del ex 24	5.8 ± 0.6 kb/Del ex 24	5.5–5.9 kb/del ex 24	38,467,102–38,452,877 (5,776 bp)	AluJo–Non-Alu	c.5467+309_*1383+2736delinsAC	Novel
BR08	*BRCA1*	Dup ex 5–7	12 ± 2.6 kb/Dup ex 5–7	8.4–9.4 kb/Dup 5–7	38,509,535–38,514,662 (9,128 bp)	AluJb–AluSx	c.135-2586_442-1112dup	Novel
BR09	*BRCA2*	Dup ex 17–20	13.8 ± 2.0 kb/Dup ex 17–20	8.3–12 kb/Dup 17–20	31,831,434–31,846,221 (14,788 bp)	Non-Alu–MIET (C-int	c.7805+1368_8633-42386dup	Novel
Known mutations								
BR01	*BRCA1*	Dup ex 13	6.1 ± 1.6 kb/Dup ex 13	5–7.9 kb/Dup ex 13*	38,489,825–38,489,903 (6,080 bp)	AluSx–AluSx	c.4186-1785_4558-1667dup	Puget et al. (1999)
BR02	*BRCA1*	Del ex 2	40.8 ± 3.5 kb/Del ex 2	40.4–38.1 kb/Del ex 2*	38,525,493–38,562,427 (36,934 bp)	Pseudogene-gene	c.-232-631491_80+4067del	Puget et al. (2002)
BR03	*BRCA1*	Del ex 2	39.0 ± 2.6 kb/Del ex 2	40.4–38.1 kb/Del ex 2*	38,525,493–38,562,427 (36,934 bp)	Pseudogene-gene	c.-232-631491_80+4067del	Puget et al. (2002)
BR04	*BRCA1*	Dup ex 18–20	6.7 ± 1.2 kb/Dup ex 18–20	7.4–10.9 kb/Dup ex 18–20*	38,461,760–38,470,417 (8,658 bp)	AluY–AluJb	c.5075-923_5277+855dup	Gad et al. (2002)
BR05	*BRCA1*	Del ex 15	4.1 ± 1.2 kb/Del ex 15	1.9–5.0 kb/Del ex 15*	38,479,177–38,481,175 (2,998 bp)	AluSx/g–AluSp	c.4484+859_4676-1395del	Puget et al. (1999b)
BR06	*BRCA1*	Del ex 8–13	20.0 ± 2.8 kb/Del ex 8–13	20–24 kb / Del ex 8–13*	38,485,557—38,507,323 (25,767 bp)	AluSx–AluSp	c.442-1900_4358-1400del	Puget et al. (1999b)

*These patients were previously characterized by high resolution aCGH, and the reported values were originally described by Bordeau et al. (Bordeau, 2007).
Breakpoints positions relate to reference genome Hg18—Build 36—NC_000017.9.

encodes exon 3 (Fig. 2C, top panel). This variant had already been detected, but not characterized, by Rouleau et al. [Rouleau et al., 2007]. The aCGH array gave a size between 8.3 and 12.2 kb. Finally, the mapping of the breakpoints for this rearrangement indicated that the deleted region was 11,413 bp long (Fig. 2C, bottom panel). Such a large deletion in *BRCA1* exon 3 has never been previously reported.

Deletion of exon 24 of BRCA1

By visual inspection, the mutation appeared as a deletion of the genomic region located between the S2B1 and S3B1 red signals, including a portion of the DNA probe BRCA 1–3 on S2B1 (Fig. 2D, top panel). After measurement, the mutation was estimated to have a size of 5.8 ± 0.6 kb, including exon 24. The aCGH array gave a size between 5.5 and 5.9 kb. The mapping of the breakpoint for this rearrangement indicated that the deleted region was 5,776 bp long, never reported so far (Fig. 2D, bottom panel).

Detection of Known *BRCA1* Large Rearrangements in Breast Cancer Patients

All identified six known large rearrangements have been recently characterized by aCGH and breakpoints sequencing [Rouleau et al., 2007]. Complete characterization by MC of three selected known *BRCA1* large rearrangements is reported in Figure 3 and is described here below.

Deletion of the upstream 5′ region to exon 2

By visual inspection, the mutation appeared as a deletion of the S10B1 green signal, as well as a large genomic portion of the 5′ region upstream of *BRCA1*, including S11B1 and S12B1 (Fig. 3A). After measurement, the mutation was estimated to have a size of 40.8 ± 3.5 kb, encompassing the portion of the *BRCA1* gene that encodes exon 2, the entire *NBR2* gene (signal S11B1), the genomic region between *NBR2* and the pseudogene Ψ*BRCA1* (signal S12B1), and a portion of Ψ*BRCA1* (signal S13B1). Importantly, the reported size of the exon 2 deletion is highly variable, estimated to be in the range of 13.8–36.9 kb [Mazoyer, 2005]. Six different exon 1–2 deletions have been reported and there are 16 reports in the literature in various populations [Sluiter and van Rensburg, 2011]. The rearrangement reported here has already been described with an identical size (36,934 bp). The hotspot for recombination is explained by the presence of Ψ*BRCA1* [Puget et al., 2002]. MC proved capable of characterizing events even in this highly homologous region.

Deletion from exon 8 to exon 13

By visual inspection, the mutation appeared as a visible deletion of the S7B1 blue signal, including a large genomic portion between S7B1 and S8B1 signals (Fig. 3B). After measurement, the mutation was estimated to have a size of 20 ± 2.8 kb in a portion of the *BRCA1* gene that encodes from exon 8 to exon 13. The size reported in the literature and here is 23,767 bp [Puget et al., 1999b], and this is a recurrent mutation in the French population [Mazoyer, 2005; Rouleau et al., 2007].

Duplication of exon 13

By visual inspection, this mutation appears as a partial tandem duplication of the blue signal S7B1 (Fig. 3C, top panel). After measurement, the mutation was estimated to have a size of 6.1 ± 1.6 kb,

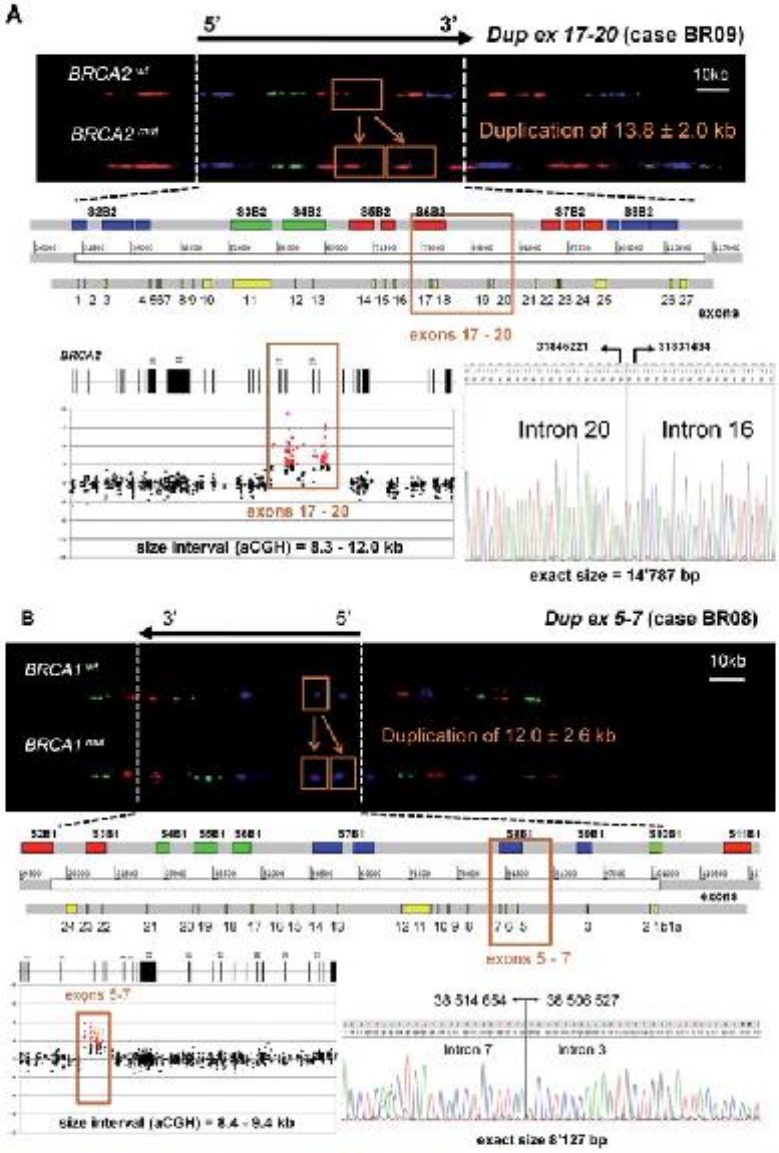


Figure 2. Four novel *BRCA1* and *BRCA2* large rearrangements detected in breast cancer patients. **A:** Dup ex 17–20 on *BRCA2* (case BR09), visible as a tandem repeat duplication of the red signal S6B2. To confirm the presence of the mutation, the motif *g4b2* (21.1 kb) was first measured on a mixed population of 40 images, comprising *wt* and *mt* alleles, and following values were obtained: μ (*BRCA2*+ *BRCA2*^{mt}) = 28.3 ± 7.3 kb and *delta* = 7.2 kb (≥2kb). To measure the mutation size, the images were then divided in two groups, 21 were *BRCA1*^{wt}, whereas 19 were *BRCA1*^{mt}: μ (*BRCA2*^{wt}) = 21.7 kb ± 2.3 kb, μ (*BRCA2*^{mt}) = 35.5 ± 2.1 kb, *mutation size* = μ (*BRCA2*^{mt}) − μ (*BRCA2*^{wt}) = 13.8 ± 2.0 kb. **B:** Dup ex 5–7 on *BRCA1* (case BR08), visible as a tandem repeat duplication of the blue signal S8B1. The motif *g5b1* (19.7 kb) was first measured on 29 images, yielding the following values: μ (*BRCA1*^{wt} + *BRCA1*^{mt}) = 23 ± 6.6 kb, *delta* = 3.3 kb (*delta* ≥ 2kb). The images were then divided in two groups, 16 images were *BRCA1*^{wt} and 13 images were *BRCA1*^{mt}: μ (*BRCA1*^{wt}) = 17.6 ± 1.9 kb, μ (*BRCA1*^{mt}) = 29.6 ± 3.0 kb, *mutation size* = μ (*BRCA1*^{mt}) − μ (*BRCA1*^{wt}) = 12 ± 2.6 kb. **C:** Del ex 3 on *BRCA1* (case BR07), visible as a deletion of the blue signal S9B1 and the downstream genomic region (motif g6b1). The motif g6B1 (9.3 kb) was first measured on a mixed population of 25 images, yielding the following values: μ (*BRCA1*^{wt} + *BRCA1*^{mt}) = 3.6 ± 4.6 kb,

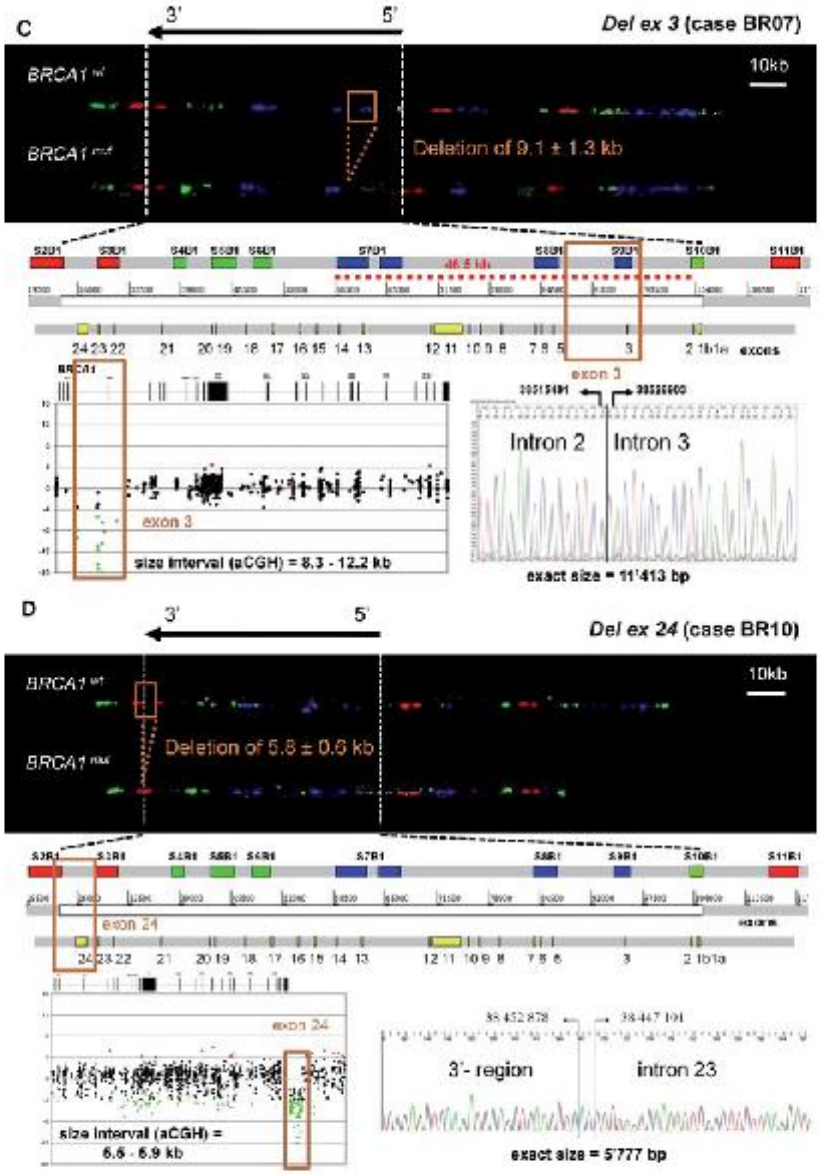


Figure 2. (Continued) *delta* = −5.7 kb (≤ −2kb). The images were divided in two groups, 10 images were *BRCA1*^{wt} and 15 images were *BRCA1*^{mt}: μ (*BRCA1*^{wt}) = 9.1 ± 2.1 kb, μ (*BRCA1*^{mt}) = 0 ± 0 kb, *mutation size* = μ (*BRCA1*^{wt}) − μ (*BRCA1*^{mt}) = −9.1 ± 1.3 kb. **D:** Del ex 24 on *BRCA1* (case BR10), visible as a deletion of the genomic regions between the two red signals S2B1 and S3B1, including a portion of S2B1 (motif *g1b1*). The deleted region encodes for exon 24. The motif *g1b1* (8.5 kb) was first measured on a mixed population of 35 images, yielding the following values: μ (*BRCA1*^{wt} + *BRCA1*^{mt}) = 5.8 ± 3.4 kb, *delta* = −2.7 kb, (≤−2kb). The images were then divided in two groups, 20 images were *BRCA1*^{wt} and 15 images were *BRCA1*^{mt}: μ (*BRCA1*^{wt}) = 9.0 ± 0.9 kb, μ (*BRCA1*^{mt}) = 3.2 ± 0.3 kb, *mutation size* = μ (*BRCA1*^{mt}) − μ (*BRCA1*^{wt}) = −5.8 ± 0.6 kb. The GIMC signals obtained after microscopic visualization are shown in the top panels. The zoom-in on the *BRCA1* or *BRCA2* gene-specific signals and the relative positions of the mutated exons are shown in the middle panels. The aCGH profiles including the mutation size interval and the identified breakpoints are shown in the lower panels. *mt* = mutated allele, *wt* = wild-type allele.

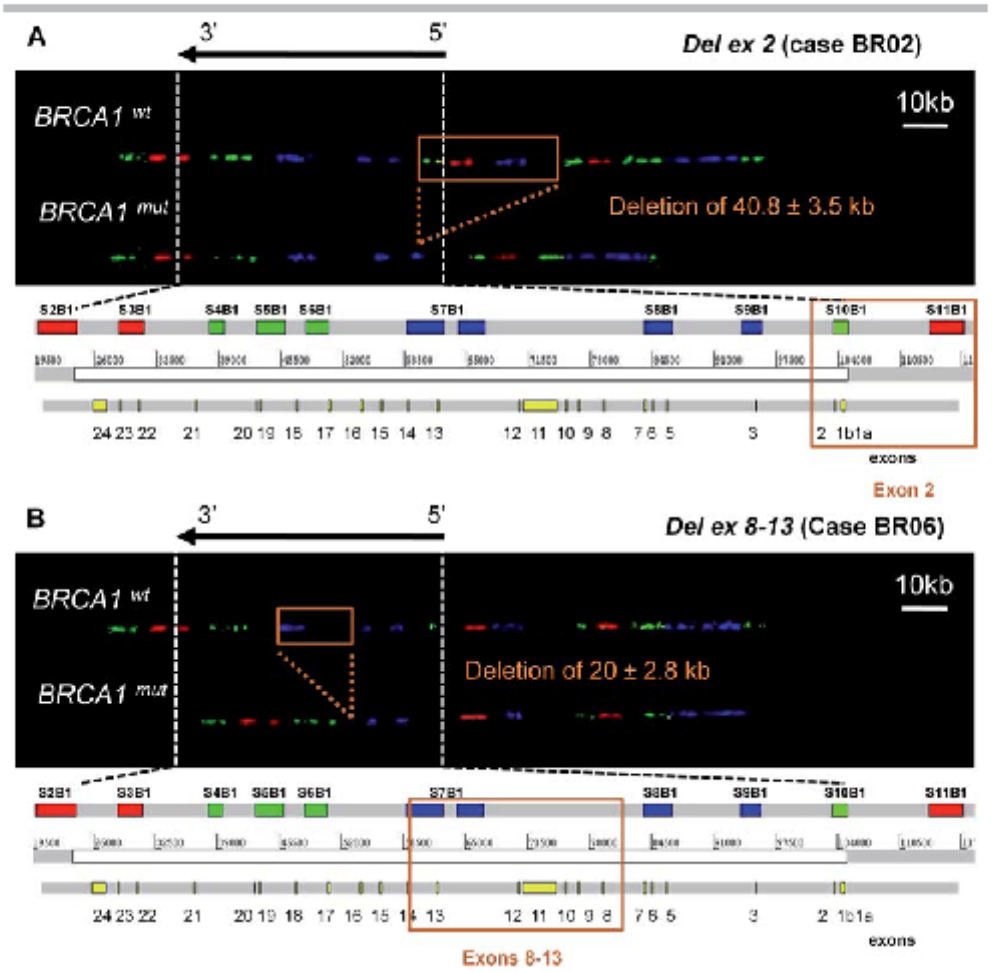


Figure 3. Known *BRCA1* large rearrangements detected in breast cancer patients. **A:** Del ex 2 (case BR02), visible as a deletion of the green signal S10B1, as well as a large genomic portion of the 5′ region upstream of *BRCA1*, including S11B1 and S12B1. To confirm the presence of the deletion in the *BRCA1* gene, the *g7B1* (17.7 kb) motif was first measured on a mixed population of 20 images, yielding following values: μ (*BRCA1*^{wt} + *BRCA1*^{wt}) = 12.3 ± 2.9 kb, *delta* = −5.4 kb (deletion is confirmed since *delta* ≤ −2kb). To measure mutations size within the *BRCA1*gene, 11 images were then classified as *BRCA1*^{wt} and 9 images as *BRCA1*^{wt}, yielding following values: μ (*BRCA1*^{wt}) = 18.1 ± 0.7 kb, μ (*BRCA1*^{wt}) = 8.1 ± 1.6 kb, *mutation size* = μ (*BRCA1*^{wt}) − μ (*BRCA1*^{wt}) = −10 ± 1.5 kb. To include the deleted genomic region upstream of *BRCA1*and determine the whole mutation size, we had to measure the genomic region between the signals S8B1 and S14B1 (89.9 kb). The S8B1-S12B1 region was first measured on 19 images, yielding following values: μ (*BRCA1*^{wt} + *BRCA1*^{wt}) = 62.3 ± 18.4 kb, *delta* = −27.6 kb. 11 images were then classified as *BRCA1*^{wt}, and 8 images as *BRCA1*^{wt}, yielding following values: μ (*BRCA1*^{wt}) = 92.2 ± 3.2 kb, μ (*BRCA1*^{wt}) = 51.4 ± 2.2 kb, *mutation size* = μ (*BRCA1*^{wt}) − μ (*BRCA1*^{wt}) = −40.8 ± 3.5 kb. **B:** Del ex 8–13 (case BR06), visible as a deletion of the blue signal S7B1, including a large genomic portion between signals S7B1 and S8B1. The *g4B1* (16.5 kb) and the *g5b1* (19.7 kb) motifs were first measured on a mixed population of 23 images, yielding following values. For *g4b1*: μ (*BRCA1*^{wt} + *BRCA1*^{wt}) = 17.5 ± 4.0 kb, *delta* = −2.2 kb (*delta* ≤ 2 kb); 13 images were then classified as *BRCA1*^{wt} and 10 images as *BRCA1*^{wt}: μ (*BRCA1*^{wt}) = 20.8 ± 1.6 kb, μ (*BRCA1*^{wt}) = 13.3 ± 1.1 kb, μ (*BRCA1*^{wt}) −μ (*BRCA1*^{wt}) = −7.5 ± 1.6 kb. For *g5b1*: μ (*BRCA1*^{wt} + *BRCA1*^{wt}) = 12.8 ± 5.5 kb, *delta* = −3.7 kb (*delta* ≤ −2kb); 13 images were then classified as *BRCA1*^{wt} and 10 images as *BRCA1*^{wt}: μ (*BRCA1*^{wt}) = 18.3 ± 1.3 kb, μ (*BRCA1*^{wt}) = 5.8 ± 0.5 kb, μ (*BRCA1*^{wt}) − μ (*BRCA1*^{wt}) = −12.5 ± 1.0 kb. Total *mutation size* = mutation size *g4B1* + mutation size *g5b1* = −20 ± 2.8 kb. **C:** Dup ex 13 (case BR01), visible as a tandem repeat duplication of the blue signal S7B1. The *g4B1* motif (16.5 kb) was first measured on a mixed population of 40 images, comprising wild type and mutated alleles, and following values were obtained: μ (*BRCA1*^{wt} + *BRCA1*^{wt} signals) = 19 kb ± 3.5 kb, *delta* = 2.5 kb (duplication is confirmed since *delta* ≥ 2kb). The images were then divided in two groups: 21 images were classified as *BRCA1*^{wt}, and 19 images were classified as *BRCA1*^{wt}. The size was then calculated as the difference between the motif mean sizes of the two alleles: μ (*BRCA1*^{wt}) = 16.1 ± 1.6 kb, μ (*BRCA1*^{wt}) = 22.2 ± 2.0 kb, *mutation size* = μ (*BRCA1*^{wt}) − μ (*BRCA1*^{wt}) = 6.1 ± 1.6 kb. The bottom panel shows the MLPA fragment display (left) and the normalized MLPA results (right), arrows indicating exons interpreted as duplicated. *mt*, mutated allele; *wt*, wild-type allele.

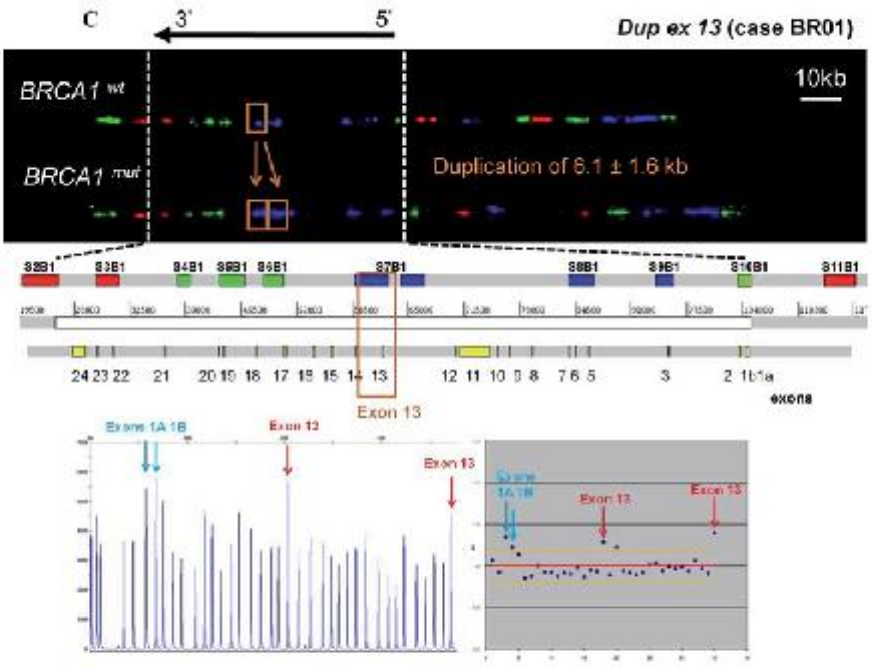


Figure 3. Continued.

restricted to a portion of the DNA probe *BRCA1-8* that encodes exon 13. The estimated mutation size is fully in line with the 6,081 bp reported in the literature [Puget et al., 1999a], and according to the Breast Cancer Information Core database, this mutation is one of the 10 most frequent mutations in *BRCA1* [Szabo et al., 2000]. The characterized patient was also analyzed by MLPA, and the duplication of exon 13 was confirmed. In addition, a duplication of exons 1A + 1B was also found by MLPA in the same individual (Fig. 3C, bottom panel), but this mutation could not be detected by MC analysis (a duplication of exon 1, if present, would yield two distinct S10B1 signals), QMPSF, or dedicated CGH array (data not shown). Therefore, we consider the exon 1A + 1B mutation detected by MLPA to be a false-positive signal.

Discussion

We describe a novel diagnostic genetic test for the detection of large rearrangements and other structural variations in the *BRCA1* and *BRCA2* genes. Large rearrangements represent 10–15% of deleterious germline mutations in the *BRCA1* gene and 1–7% in the *BRCA2* gene [Mazoyer, 2005; Sluiter and van Rensburg, 2011]. We designed specific high-resolution GMCs [Lebofsky et al., 2006] and tested them on a series of 60 biological samples. The robustness of the associated measurement strategy was statistically validated on 50 control samples, and 10 different large rearrangements (nine of *BRCA1* and one of *BRCA2*), initially detected with other techniques, were fully characterized by MC in samples from patients with a severe family history of breast cancer. The robustness of the newly designed GMCs, devoid of repetitive sequences, is endorsed by the fact that the nature of the identified mutations was confirmed by high-resolution zoom-in aCGH (11k), with a precision in the 1–2 kb range.

Four out of the 10 characterized large rearrangements have never been previously described: a 11.4 kb deletion of exon 3 (case BR07), a 5.8 kb deletion of exon 24 (case BR10), a 8.1 kb duplication of exons 5–7 (case BR08) of *BRCA1*, and a 14.8 kb duplication of exons 17–20 of *BRCA2* (case BR09).

Two deletions involving *BRCA1* exon 3 have been described in the literature, but with sizes significantly different (1,049 and 1,039 bp) [Payne et al., 2000; Walsh et al., 2006] from that reported in our study (11,413 bp). The rearrangement we describe is the largest reported to involve this exon. The mechanism is associated with Alu–Alu recombination involving the 5′ region of *BRCA1*. Five (6%) of 81 reported *BRCA1* deletions involve exon 24, which is localized in a known hotspot of deletion related to the 3′ region of *BRCA1*. Two deletions have been reported in the literature involving only the exon 24 and the 3′UTR, one of 4,427 bp [Armaou et al., 2007] and one of 1,506 bp [Engert et al., 2008]. The deletion we describe here measures 5,776 bp and involves another locus. As in the 4,427 bp deletion, there is an insertion of a few base pairs (5 and 2 bp, respectively) and an Alu-non-Alu mechanism: the 5′ breakpoint lies in a AluCo sequence and the 3′ breakpoint within a region without any Alu sequences. The impact of such an event is the deletion of the stop codon, the polyA tail region, and the 3′UTR of the *BRCA1* gene. No mRNA transcript production was detected so far. The

deletion in exon 24 is a recurrent event in the Greek population [Pertesi et al., 2011]. The deletion characterized in our study remains to be explored in the French population. Two 5 kb deletions including *BRCA1* exons 5–7 have been reported in the literature [de Juan et al., 2009; Hansen et al., 2009; Preisler-Adams et al., 2006], but this is the first observation of exons 5–7 duplication. The exact size of this event is 8,127 bp. A common region of 26 nucleotides was detected at the breakpoints with one lying in an AluSx and the other one in an AluJb. These events show the sensitivity of this particular genomic region, since it can contain different mutation types (deletions and duplications) with different sizes. Very few rearrangements at a specific locus include both deletions and mirror duplications. For instance, *BRCA1* exon 13 has been shown to exhibit a deletion of 3,835 bp [Petrij-Bosch et al., 1997] and three duplications of 6,081 bp (recurrent among people of English ancestry), 5,275 [Walsh et al., 2010] and 8,463 bp [Yap et al., 2006]. The distance between the breakpoints was between 144 bp and 3 kb. Another example is the rearrangement in exons 18 and 19, with deletions of 4,826 bp [Montagna et al., 2003], 1,940 bp [Foretova et al., 2006], and 7,245 bp [Engert et al., 2008] and a duplication of 5,923 bp [Walsh et al., 2006]. For these events the 5′ breakpoint was within 65–1,601 bp and the 3′ breakpoint was within 1 and 3 kb from the duplication. Finally, several deletions have been reported in *BRCA1* exon 20. A 7,029 bp deletion [Foretova et al., 2006] had a very similar breakpoint in the 3′-region (29 bp) to that of the 8,706 bp duplication [Agata et al., 2006]. Strangely, no duplications have been reported so far in the 5′-region involving at least exon 1 and 2, despite the wide variety of rearrangements, suggesting that such duplications might not be pathogenic. The composition of Alu sequences in those sensitive regions should be studied to understand the DNA-strand organization.

Sluiter et al. reported 17 large rearrangements in the *BRCA2* gene [Sluiter and van Rensburg, 2011]. A 10.8 kb deletion of exons 17–18 [Agata et al., 2006] and a 9.7-kb duplication of exons 19–20 [Walsh et al., 2006] have also been reported. The 14.8 kb *BRCA2* mutation reported here is the second biggest duplication, after the 16.2 kb duplication (exon 4 and part of exon 11) reported by Lim [Lim et al., 2007]. Another duplication, involving exons 19 and 20 has been reported in this region [Caux-Moncoutier et al., 2011]. This confirms the presence of a sensitive area between exon 17 and exon 20, with three different duplications. Only about 10% of large rearrangements in *BRCA1* and *BRCA2* are amplification events (duplication or triplication) [Sluiter and van Rensburg, 2011], and few of these events are associated with Alu–Alu recombination.

Duplications are the most difficult large rearrangements to detect and characterize. Contrary to other techniques, such as aCGH and MLPA, the capacity of MC to visualize hybridized DNA probes at high resolution permits precise mapping and characterization of tandem repeat duplications, as shown here in cases BR01 (*BRCA1* Dup Ex 13), BR09 (*BRCA2* Dup Ex 17–20), BR08 (*BRCA1* Dup Ex 5–7), and BR04 (*BRCA1* Dup Ex 18–20). aCGH can be used to determine the presence and size of duplications, but not their exact location and orientation on the genome (see cases BR08 and BR09). Only a combination of aCGH and sequencing would allow defining the identified duplications as tandem repeats. In contrast, MC allows to unambiguously localize tandem repeats, providing information on the genome location, the nature and the size of the mutation, in one single experiment and using one single technique.

In PCR-based techniques such as MLPA, duplications are considered to be present when the ratio between the number of duplicated exons in the sample carrying a mutation and the number of exons in the control sample is at least 1.5, reflecting the presence of three copies of a specific exon in the mutated sample and two copies in the wild-type sample. The ratio of 1.5 is difficult to demonstrate unambiguously by MLPA, which often gives false-positive signals, as observed in case BR01 (*BRCA1* Dup Ex 13). The limits of MLPA have been underlined in several recent studies [Cavalieri et al., 2008; Staaf et al., 2008]. MLPA is limited to coding sequences and can also give false-negative scores, due to the restricted coverage of the probes [Cavalieri et al., 2008]. Staaf et al. recently suggested that MLPA should be regarded as a predictive screening tool that needs to be complemented by other means of mutation characterization, such as aCGH [Staaf et al., 2008]. Another multiplex PCR assay similar to MLPA, is QMPSF [Charbonnier et al., 2000]. MLPA has the advantage over QMPSF that it allows the analysis of up to 40 loci in a single multiplex reaction, and, because of the required ligation step, is very specific, allowing copy number analysis of regions with high-sequence homology. Primer design, on the other hand, appears less critical for the QMPSF method and this approach may be more cost effective and suitable for rapid validation experiments of loci with unique sequences. The performance of both methods has not been compared, and therefore it remains speculative which technology is most suited for targeted high-throughput analysis of large rearrangements. We propose MC as a complementary technology for MLPA, QMPSF, or aCGH, as it unambiguously identifies and visualizes duplications.

Another advantage of MC is its capacity to cover noncoding regions, including the 5′-region of the *BRCA1* gene and the genomic region upstream of *BRCA1* that comprises the *NBR2* gene, the Ψ*BRCA1* pseudogene and the *NBR1* gene. Recent studies show that it is very difficult to design exploitable PCR or aCGH probes in this rearrangement-prone genomic region [Rouleau et al., 2007; Staaf et al., 2008], because of the presence of duplicated regions and the high density of Alu repeats, particularly *BRCA1* that contains more than 40% of Alu sequences. Genomic rearrangements typically arise from unequal homologous recombination between segmental duplication or Alu sequences [Hastings et al., 2009]. MC permits precise physical mapping within this difficult region, as shown here for three different large rearrangements. In cases BR07 (*BRCA1* Del Ex 3), BR02, and BR03 (*BRCA1* Del Ex 2), we were able to determine precisely the mutation sizes within the hard-to-sequence *BRCA1* 5′-region and confirm the result by aCGH / breakpoints mapping. In cases BR02 and BR03 (*BRCA1* Del Ex 2), we measured mutation sizes of 40.8 ± 3.5 kb and 39.0 ± 2.6 kb, respectively. The statistical error found by MC does not allow to state that these two mutations are different and aCGH analysis of these two samples delivered identical results (40.4–58.1 kb). The detected mutations are supposed to be identical since individuals BR02 and BR03 belong to the same biological family. This mutation was originally described by Puget et al., who determined a mutation size of 37 kb by breakpoint mapping [Puget et al., 2002]. The larger mutation size range estimated via aCGH is probably caused by the low density of exploitable oligonucleotide sequences in this genomic region and the reduced sensitivity of some oligonucleotides due to sequence homology [Rouleau et al., 2007].

We were able to demonstrate the absence of the 100-kb "sequencing gap" in the genomic region upstream of the *BRCA1* gene, thus confirming the *BRCA1* genomic structure proposed in the GRCh37 genome assembly. The zoom-in aCGH arrays designed by Staaf et al. were based on the NCBI build 35 genome assembly (release date May 2004), which still included the "sequencing gap." It follows that the 100-kb genomic gap that could not be covered by the arrays is indeed non-existent. Thus, the size of previously identified mutations that include this "sequencing gap" must be reduced by 100 kb. For example, the 300 kb *BRCA1* large rearrangement (case L1985)

must be reduced to 200 kb [Staaf et al., 2008]. MC can therefore be used for physical mapping of hard-to-sequence genomic regions that contain large numbers of repetitive elements. Here we demonstrate that the high concentration of Alu sequences in *BRCA1* does not represent an obstacle for MC.

From a practical point of view, the global turnaround time for the complete analysis of 10–20 patients via MC is 2 weeks (corresponding to 250–500 patients per year), which is close to the turnaround time of aCGH analysis, and compatible to the needs of clinical diagnostic testing. By comparison, MLPA can be used to process up to 100 samples in parallel (e.g., 96-well plates) in just 2–3 days, making this technology more suitable for routine high-throughput predictive testing [Schouten, 2002]. Automation of MC is being further improved, by developing all the necessary technological tools that will significantly increase the number of samples that can be analyzed in parallel. Another limitation of MC is the large amount of DNA required, in the range of 500–1,000 ng (typically corresponding to 5×10^5 to 10^6 cells). This is far more than required for MLPA (20–50 ng), but similar to what is required for aCGH analysis [Rouleau, 2007; Staff, 2008]. The smallest large rearrangement described in this work has a size of 3kb (case BR05), and the limit of resolution of MC is in the range of 1–2 kb [Herrick 2009; Lebofsky, 2003]. As a comparison the resolution of aGCH is in the range of 500 bp [Rouleau et al.]. Within the *BRCA1* large rearrangements reported so far in the literature, only 10% (9/81) have a size smaller than 2 kb and only 4% (3/81) have a size smaller than 500 bp [Sluiter et al., 2011]. According to our experience, the large rearrangements in the range of 500 bp (or smaller) should be also characterized with PCR-based technologies, since these mutations are close to detection limit of aCGH. Thus, the vast majority (90%) of large rearrangements identified so far are higher than our resolution limit. Generally speaking, if a potential mutation is detected with our genetic test, the whole MC assay is repeated. Measurements data in combination with the known position of probes can help to physically map the large rearrangement. For instance, primers can be placed on intact signals not involved in the mutations, which should get PCR product size smaller than 10 kb (applicable for classical or long-range PCR). In the cases of estimated PCR product sizes larger than 10kb, zoom-in CGH array can be employed to refine the mapping.

We propose the genetic test based on MC as a valuable tool for the detection and characterization of large rearrangements in *BRCA1* and *BRCA2*, to be combined in clinical diagnostic settings with an assay that allows the detection of small mutations (e.g., sequencing). This is particularly valuable for the 80% of patients with breast and ovarian cancer predisposition, for which no deleterious mutations is detected with MLPA/sequencing techniques. We estimate the price of the consumables related to our genetic test (DNA extraction kit, DNA probe set kit and coverslips) to be in the range of approximately $1,000 (€800). Other technologies currently employed in diagnostic testing of *BRCA1* and *BRCA2* are NGS and aCGH. As a comparison, the cost of the consumables related to the NGS, has been estimated to be in the order of $1,500 [Walsh et al., 2010]. Of notice, the limitation of NGS in clinical diagnostic settings was recently highlighted because of the high content of hard-to-sequence repetitive sequences present in high percentage in both *BRCA1* and *BRCA2* [De Leeneer et al., 2011]. Large rearrangements could be detected by NGS, only with a very high coverage of 1200x [Walsh et al., 2010], but a lower coverage of 120x did not allow the detection of large rearrangements in both genes [De Leeneer et al., 2011]. The needed high coverage could influence the overall cost of large rearrangements detection via NGS. The cost of CGH arrays has been described has varying between $250 and $800, depending on the array format, but no information has been provided on the global cost of consumables related to CGH analysis [Staaf et al., 2008]. Of notice, the standard commercial test (DNA sequencing of both genes and screening for five large deletions and duplications in *BRCA1*) proposed by Myriad Genetics costs $3,340, and comprehensive testing for gene rearrangements is offered as a separate test at an additional cost of $650.

We see the main application of the developed MC-based assay as a diagnostic genetic test. Once the overall throughput will be improved, we envisage to extended the application of the developed assay as a companion diagnostic test, for instance in the screening of *BRCA*-mutated cells in the context of the development of PARP-1 inhibitors. Thus, the genetic test may be applied not only to clinical blood samples, but also to circulating cells and heterogeneous cell populations, such as tumor tissues.

Conflict of Interest

All the authors declare that they have no Conflicts of Interest linked to the submitted manuscript.

Acknowledgments

This study is dedicated to Daniel Nerson. The authors would like to thank Sylvie Mazoyer, for critical reading of the manuscript and Clemence Thiberville for the provided inputs on QA/QC checkpoints analysis. The authors would also like to thank Jennifer Abscheidt and Solenne Guillon for the help provided in the analysis of the negative controls.

References

Agata S, Viel A, Della Puppa L, Cortesi L, Fersini G, Callegaro M, Dalla Palma M, Dolcetti R, Federico M, Venuta S, and others. 2006. Prevalence of BRCA1 genomic rearrangements in a large cohort of Italian breast and breast/ovarian cancer families without detectable BRCA1 and BRCA2 point mutations. Genes Chromosomes Cancer 45:791–797.

Altman DG, Bland JM. 1994. Diagnostic tests. 1: Sensitivity and specificity. BMJ 308:1552.

Armaou S, Konstantopoulou I, Anagnostopoulos T, Razis E, Boukovinas I, Xenidis N, Fountzilas G, Yannoukakos D. 2007. Novel genomic rearrangements in the BRCA1 gene detected in Greek breast/ovarian cancer patients. Eur J Cancer 43:443–453.

Barrois M, Bieche I, Mazoyer S, Champeme MH, Bressac-de Paillerets B, Lidereau R. 2004. Real-time PCR-based gene dosage assay for detecting BRCA1 rearrangements in breast-ovarian cancer families. Clin Genet 65:131–136.

Caburet S, Conti C, Schurra C, Lebofsky R, Edelstein SJ, Bensimon A. 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures. Genome Res 15:1079–1085.

Caputo S, Benboudjema L, Sinilnikova O, Rouleau E, Béroud C, Lidereau R: French BRCA GGC Consortium. 2012. Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. Nucleic Acids Res 40:D992-1002.

Casilli F, Di Rocco ZC, Gad S, Tournier I, Stoppa-Lyonnet D, Frebourg T, Tosi M. 2002. Rapid detection of novel BRCA1 rearrangements in high-risk breast-ovarian cancer families using multiplex PCR of short fluorescent fragments. Hum Mutat 20:218–226.

Caux-Moncoutier V, Castera L, Tirapo C, Michaux D, Remon MA, Lauge A, Rouleau E, De Pauw A, Buecher B, Gauthier-Villars M and others. 2011. EMMA, a cost- and time-effective diagnostic method for simultaneous detection of point mutations and large-scale genomic rearrangements: application to BRCA1 and BRCA2 in 1,525 patients. Hum Mutat 32:325–334.

Cavalieri S, Funaro A, Pappi P, Migone N, Gatti RA, Brusco A. 2008. Large genomic mutations within the ATM gene detected by MLPA, including a duplication of 41 kb from exon 4 to 20. Ann Hum Genet 72:10–18.

Charbonnier F, Raux G, Wang Q, Drouot N, Cordier F, Limacher JM, Saurin JC, Puisieux A, Olschwang S, Frebourg T. 2000. Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments. Cancer Res 60:2760–2763.

Patent application n°1:

de Juan I, Esteban E, Palanca S, Barragan E, Bolufer P. 2009. High-resolution melting analysis for rapid screening of BRCA1 and BRCA2 Spanish mutations. Breast Cancer Res Treat 115:405–414.
De Leeneer K, Hellemans J, De Schrijver J, Baetens M, Poppe B, Van Criekinge W, De Paepe A, Claes K. 2011. Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. Hum Mutat 32:335–344.
Engert S, Wappenschmidt B, Betz B, Kast K, Kutsche M, Hellebrand H, Goecke TO, Kiechle M, Niederacher D, Schmutzler RK and others. 2008. MLPA screening in the BRCA1 gene from 1,506 German hereditary breast cancer cases: novel deletions, frequent involvement of exon 17, and occurrence in single early-onset cases. Hum Mutat 29:948–958.
Foretova L, Petrakova K, Palacova M, Kalabova R, Navratilova M, Lukesova M, Vasickova P, Machackova E, Kleibl Z, Pohlreich P. 2006. Genetic and preventive services for hereditary breast and ovarian cancer in the czech republic. Hered Cancer Clin Pract 4:3–6.
Frolov A, Prowse AH, Vandervoer L, Bove B, Wu H, Godwin AK. 2002. DNA array-based method for detection of large rearrangements in the BRCA1 gene. Genes Chromosomes Cancer 35:232–241.
Gad S, Amrias A, Puget N, Mairal A, Schurra C, Montagna M, Pages S, Caux V, Mazoyer S, Bensimon A and others. 2001. Color bar coding the BRCA1 gene on combed DNA: a useful strategy for detecting large gene rearrangements. Genes Chromosomes Cancer 31:75–84.
Gad S, Bieche I, Barrois M, Caslli F, Pages-Berhouet S, Dehainault C, Gauthier-Villars M, Bensimon A, Amrias A, Lidereau R and others. 2003. Characterisation of a 161 kb deletion extending from the NBR1 to the BRCA1 genes in a French breast-ovarian cancer family. Hum Mutat 21:654.
Gad S, Caux-Moncoutier V, Pages-Berhouet S, Gauthier-Villars M, Coupier I, Pujol P, Frenay M, Gilbert B, Mangard C, Bignon YJ, and others. 2002a. Significant contribution of large BRCA1 gene rearrangements in 120 French breast and ovarian cancer families. Oncogene 21:6841–6847.
Gad S, Klinger M, Caux-Moncoutier V, Pages-Berhouet S, Gauthier-Villars M, Coupier I, Bensimon A, Amrias A, Sloppa-Lyonnet D. 2002b. Bar code screening on combed DNA for large rearrangements of the BRCA1 and BRCA2 genes in French breast cancer families. J Med Genet 39:817–821.
Gandet MM, Kirchhoff T, Green T, Vijai J, Korn JM, Gnirkucci C, Segre AV, McGee K, McGuffog L, Kartsonaki C and others. 2010. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. PLoS Genet 6:e1001183.
Hansen TO, Jonson L, Albrechtsen A, Andersen MK, Ejlertsen B, Nielsen FC. 2009. Large BRCA1 and BRCA2 genomic rearrangements in Danish high risk breast-ovarian cancer families. Breast Cancer Res Treat 115:315–323.
Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10:551–564.
Herrick J, Bensimon A. 2009. Introduction to molecular combing: genomics, DNA replication, and cancer. Methods Mol Biol 521:71–101.
Hofmann W, Wappenschmidt B, Berhane S, Schmutzler R, Scherneck S. 2002. Detection of large rearrangements of exons 13 and 22 in the BRCA1 gene in German families. J Med Genet 39:E36.
Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. 2011. Global cancer statistics. CA Cancer J Clin 61:69–90.
King MC, Marks JH, Mandell JB. 2003. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 302:643–646.
Lebofsky R, Heilig R, Sonnleitner M, Weissenbach J, Bensimon A. 2006. DNA replication origin interference increases the spacing between initiation events in human cells. Mol Biol Cell 17:5337–5345.
Lim YK, Lau PT, Ali AB, Lee SC, Wong JE, Putti TC, Sng JH. 2007. Identification of novel BRCA large genomic rearrangements in Singapore Asian breast and ovarian patients with cancer. Clin Genet 71:331–342.
Mazoyer S. 2005. Genomic rearrangements in the BRCA1 and BRCA2 genes. Hum Mutat 25:415–422.
Montagna M, Dalla Palma M, Menin C, Agata S, De Nicolo A, Chieco-Bianchi L, D'Andrea E. 2003. Genomic rearrangements account for more than one-third of the BRCA1 mutations in northern Italian breast/ovarian cancer families. Hum Mol Genet 12:1055–1061.
Nathanson KL, Wooster R, Weber BL. 2001. Breast cancer genetics: what we know and what we need. Nat Med 7:552–556.
Nguyen K, Walrafen P, Bernard R, Attarian S, Chaix C, Vovan C, Renard E, Dufrane N, Pouget J, Vannier A and others. 2011. Molecular combing reveals allelic combinations in facioscapulohumeral dystrophy. Ann Neurol 70:627–633.
Payne SR, Newman B, King MC. 2000. Complex germline rearrangement of BRCA1 associated with breast and ovarian cancer. Genes Chromosomes Cancer 29:58–62.
Pertesi M, Konstantopoulou I, Yannoukakos D. 2011. Haplotype analysis of two recurrent genomic rearrangements in the BRCA1 gene suggests they are founder mutations for the Greek population. Clin Genet 80:375–382.
Petrij-Bosch A, Peelen T, van Vliet M, van Eijk R, Olmer R, Drusedau M, Hogervorst FB, Hageman S, Arts PJ, Ligtenberg MJ and others. 1997. BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. Nat Genet 17:341–345.
Preisler-Adams S, Schonbrachner I, Hebig B, Welling B, Dworniczak B, Weber BH. 2006. Gross rearrangements in BRCA1 but not BRCA2 play a notable role in predisposition to breast and ovarian cancer in high-risk families of German origin. Cancer Genet Cytogenet 168:44–49.
Puget N, Gad S, Perrin-Vidoz L, Sinilnikova OM, Stoppa-Lyonnet D, Lenoir GM, Mazoyer S. 2002. Distinct BRCA1 rearrangements involving the BRCA1 pseudogene suggest the existence of a recombination hot spot. Am J Hum Genet 70:858–865.
Puget N, Sinilnikova OM, Stoppa-Lyonnet D, Audoynaud C, Pages S, Lynch HT, Goldgar D, Lenoir GM, Mazoyer S. 1999a. An Alu-mediated 6-kb duplication in the BRCA1 gene: a new founder mutation? Am J Hum Genet 64:300–302.
Puget N, Stoppa-Lyonnet D, Sinilnikova OM, Pages S, Lynch HT, Lenoir GM, Mazoyer S. 1999b. Screening for germ-line rearrangements and regulatory mutations in BRCA1 led to the identification of four new deletions. Cancer Res 59:455–461.
Renkeau E, Lefol C, Bourdon V, Coulet F, Noguchi T, Soubrier F, Bieche I, Olschwang S, Sobol H, Lidereau R. 2009. Quantitative PCR high-resolution melting (qPCR-HRM) curve analysis, a new approach to simultaneously screen point mutations and large rearrangements: application to MLH1 germline mutations in Lynch syndrome. Hum Mutat 30:867–875.
Renkeau E, Lefol C, Tozlu S, Andrieu C, Guy C, Copigny F, Nogues C, Bieche I, Lidereau R. 2007. High-resolution oligonucleotide array-CGH applied to the detection and characterization of large rearrangements in the hereditary breast cancer gene BRCA1. Clin Genet 72:199–207.
Schurra C, Bensimon A. 2009. Combing genomic DNA for structural and functional studies. Methods Mol Biol 464:71–90.
Shiller MD, van Rensburg EJ. 2011. Large genomic rearrangements of the BRCA1 and BRCA2 genes: review of the literature and report of a novel BRCA1 mutation. Breast Cancer Res Treat 125:325–349.
Staaf J, Tornqren T, Rambech E, Johansson U, Persson C, Sellberg G, Tellhed L, Nilbert M, Borg A. 2008. Detection and precise mapping of germline rearrangements in BRCA1, BRCA2, MSH2, and MLH1 using zoom-in array comparative genomic hybridization (aCGH). Hum Mutat 29:555–564.
Szabo C, Masiello A, Ryan JF, Brody LC. 2000. The breast cancer information core: database design, structure, and scope. Hum Mutat 16:123–131.
Tournier I, Paillerets BB, Sobol H, Stoppa-Lyonnet D, Lidereau R, Barrois M, Mazoyer S, Coulet F, Hardouin A, Chompret A and others. 2004. Significant contribution of germline BRCA2 rearrangements in male breast cancer families. Cancer Res 64:8143–8147.
Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, Roach KC, Mandell J, Lee MK, Ciernikova S and others. 2006. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA 295:1379–1388.
Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC. 2010. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci USA 107:12629–12633.
Yap KP, Ang P, Lim IH, Ho GH, Lee AS. 2006. Detection of a novel Alu-mediated BRCA1 exon 13 duplication in Chinese breast cancer patients and implications for genetic testing. Clin Genet 70:80–82.

The *BRCA* Story:

One in 9 women is expected to develop breast cancer during her lifetime, and it is estimated that one out of 27 will die of it. A mutation in one of two genes, *BRCA1* or *BRCA2*, accounts for 3 to 10% of all breast cancer cases. Among the many forms of breast or ovarian cancers hereditary breast and ovarian cancer syndrome (HBOC) accounts for 5 to 10% of breast or ovarian cancers. Most of these HBOC cases are due to a mutation in *BRCA1* or *BRCA2*. Briefly, a mutation in either *BRCA1* or *BRCA2* prevent the production of tumour suppressing proteins, which in turn leads to an accumulation of genomic rearrangements and inactivation of other tumour suppressor genes, resulting in an increased risk of contracting breast or ovarian cancer. To date, over 2000 different mutations in these genes have been characterized, varying from single point mutations (a change of one nucleotide in the sequence of the gene) to large scale gene rearrangements (large rearrangements are genomic events in which part of the gene or genome is either inverted, translocated to another part of the genome, deleted, or duplicated). A non-negligible proportion of HBOC cases are due to large gene rearrangements in *BRCA1* and *BRCA2* (Gad et al, 2002, Walsh, 2006). It is estimated that between 8 to 15% of the cases (depending on the study and the population studied) are imputable to large rearrangements in these two genes. A mutation in one of these two genes confers a 10 to 20 times increased relative risk of developing a breast cancer (Tan et al. 2007), this fact translating into a 70-80% chance of developing a breast cancer at age 70 (King *et al.*, 2003).

This document first briefly reviews how the international scientific community proceeded to identify these two genes and the associated intellectual property issues associated with their discovery. The second part describes the contribution of Molecular combing studies to the genetic knowledge of the disease and its potential contribution to testing. The third part will delve into the role of molecular combing in the diminution of the scope of *BRCA1* and *BRCA2* patents.

The discovery of breast cancer genes: a scientific race

In the late 80's and early 90's, several research groups were trying to identify breast cancer loci. At various times they collaborated, at other times they competed. Among them were groups led by Mary Claire King, or Mark Skolnick in the US, Michael Stratton, Bruce Ponder and Richard Woodster in the UK (mainly working on *BRCA2*). In France, Gilbert Lenoir and Dominique Stoppa-Lyonet were the main group leaders involved in *BRCA* research. Others were Yusuke Nakamura in Japan, and Stephen Narod (working with Gilbert Lenoir) or Jacques Simard (working alongside Myriad) in Canada.

All the teams above participated in the mapping of the two *BRCA* genes. They formed in 1988, despite fierce competition, the international Breast Cancer Linkage Consortium, reuniting teams from the United States, the United Kingdom, France, Canada, Japan, the Netherlands, Belgium, and others, to accelerate the identification of breast cancer genes.

In 1990 Mary Claire King's team, from the University of California at Berkeley, identified through linkage analysis the location of a breast cancer gene (*BRCA1*) on chromosome 17 (Hall *et al*., 1990). This was the first evidence that cancer predisposing genes existed. The team patented the marker used for the identification of the region, and licensed them to a US based company, OncorMed. These finding were further confirmed by G. Lenoir and S. Narod(Narod *et al.*, 1991). They published an article in which they also linked this locus to hereditary forms of Breast cancer, as well as to ovarian cancer.

In 1991, Skolnick and his group created a spin-off of the centre for genetic epidemiology of the University of Utah, Myriad genetics. Despite extensive work from the Breast cancer Linkage Consortium, it was Mark Skolnick's group who sequenced the *BRCA1* gene first. They published it in *Science* on October 1994 (Miki *et al*., 1994), just after filing a patent application covering the gene and its mutation (during august that year). Several additional patents were filed and granted by the United States Patent and Trademark Office (USPTO) in the following years (US 5 693 473, US 5 709 999, US 5 747 282, US 5 710 001, US 5 753 441, US 6 162 897). Through these patents, Myriad had acquired control over *BRCA1* mutation testing in the case of breast or ovarian cancers.

(It is noteworthy to mention that meanwhile, Oncormed had also filed patents over *BRCA1*, using the markers licensed from M.C. King, and filed infringements against myriad (who did the same.) Dispute was eventually settled by Myriad acquiring Oncormed's patents.

The *BRCA2* gene was located by Michael Stratton at the Wellcome Trust Sanger Institute (Wooster *et al.*, 1995). Stratton's group then published the sequence of *BRCA2* on December 1995, co-authored by 40 researchers from several countries. A patent was filed in the UK with the Intellectual Property Office (IPO).

However, the day before the article was published, Myriad announced that it too, had sequenced the *BRCA2* gene and deposited its sequence in GenBank, and filed for a patent in the US. They finally published the sequence in nature genetics during March 1996, claiming that Stratton's team had only reported a partial sequence and six mutations, whereas Myriad's team had supplied the complete sequence (Tavtigian *et al.*, 1996). Two patents were granted to Myriad covering *BRCA2* and its mutation testing (US 5 837 492 and US 6 124 104).

Intellectual property Concerns & patent exploitation: Myriad Genetics

Armed with 9 US patents in total, Myriad set out for the commercial exploitation of these patents, by opening in late 1996 a state-of-the-art laboratory. Myriad began with three tests, the Comprehensive Myriad BRACAnalysis (full sequencing of the *BRCA1* and *BRCA2* genes, at that time priced 2400 US$, the single site BRACAnalysis test (mutation specific targeted test, priced 395 US$), and the multisite three BRACAnalaysis, targeting three founder mutations in the Ashkenazi Jewish population.

Myriad's model was that the interested patient would present himself to a specifically trained physician or genetic counsellor (Myriad even proposed training for genetic counsellors , then to physicians directly when it realized that there was not enough genetic counsellors in the US area), which would determine if the patient should undergo genetic testing or not. If so, the healthcare provider would arrange for a blood sample to be sent to Myriad's lab for full sequencing of the gene. If a mutation was found, Myriad assumed other women in the patient's family would want testing, and these would undergo the targeted mutation specific test, at about a 10^{th} of the comprehensive BRACAnalysis cost.

However, by the time Myriad introduced its test, and got all its patents granted, many other clinical and academic laboratories had already begun *BRCA1* and *BRCA2* testing through other techniques. Myriad replied by sending cease and desist letters, sawing the seeds of the future controversy to come. Myriad's business model in the US heavily relied on a network of healthcare professionals, service providers and insurers, thus providing a good positioning in the United States.

Myriad applied somewhat the same international strategy, save that it intended to identify a single laboratory or company in each territory to license its test. Myriad expected to have the licensee send out the sample to its Utah laboratory for proband testing, and to let the licensee carry out the mutation testing among other members of the family (this also caused concern over the genetic information Myriad could privately gain access to). However, because of the hype surrounding genetics around the 90's, (the human genome project, the advent of sequencing technologies, etc …), at that time, everyone and especially the national health systems and policymakers anticipated the arrival of many new technologies and genetic tests. Their main concern and fears were how to integrate these new technologies and tests into pre-existing health systems. This context, in correlation with how the scientific race for the *BRCA1* and *BRCA2* gene discoveries happened, and Myriad's behaviour (which was perceived as aggressive by the scientific and clinical community, because of the patenting, and the cease and desist letters) is certainly partly responsible for the policy storm that was sparked by Myriad's patenting of the *BRCA1* and *BRCA2* genes. Public Healthcare administrators, clinicians, physicians and researchers were not used to a business model requiring them to give over the performance of the test and the sending of a sample to a private laboratory.

All these facts led to several bouts over Myriad's patent, in different areas of the globe. In Canada, because of a failure of Myriad to understand how the federal and provincial governments worked regarding healthcare policies, it ended up by Canada simply ignoring the patents delivered by the Canadian Intellectual Property Office (patent numbers : 2 196 797, 2 196 790 and 2 196 795), despite Myriad having identified and licensed a Canadian laboratory (MDS laboratories). In Australia, Myriad was forced to grant a Licence to GTG laboratories because of GTG laboratories suing Myriad for patent infringement. In return Myriad received a non-exclusive licence to GTG's patent over non-coding DNA, so that it could continue testing elsewhere. The Japanese clinical administrations required clinical trials to be held to demonstrate the effectiveness of *BRCA* testing on the Japanese population, and as a result proband testing is performed by Myriad's licensee (Falco biosystems Ltd).

Despite encountering difficulties elsewhere, it was in Europe that the patent contestation had the most results (for a review of the policy issues worldwide around Myriad Genetics, see Gold &

Carbone, 2010). With in mind fears that Myriad would enforce its patent rights against researchers carrying their own research on *BRCA1* and *BRCA2*, and the subsequent consequence that it would also prevent the development of better tests (whether in term of the spectrum of mutations that can be identified by the said test, or in term of cost efficiency, reproducibility or even quality assessment of Myriad's test), Dominique Stoppa Lyonnet (Institut Curie) in France, and later Mary Claire King in the US, among others, took a stand against Myriad, by claiming that Myriad's aggressive behaviour would hinder the development of better tests (Benowitz, 2003, Lecrubier, 2002). Dominique Stoppa-Lyonnet also claimed that Myriad's test could not detect all mutations of clinical importance and especially large scale disturbance in the gene (called large gene rearrangements). These large rearrangements, although previously known to be present in certain disease, have been observed in breast or ovarian cancer cases in the late 90's and studied more in-depth between 2000-2003. They played an important part in the opposition procedures' success.

Large gene Rearrangements in *BRCA1* & *BRCA2:* Molecular Combing's contribution

During the years 1997 to 1999, four rearrangements were described by three different teams. One of these rearrangements accounted for 36% of the Dutch population breast cancer cases Petrij Bosch *et al*., 1997), and was defined as a founder effect (a founder effect is a mutation transmitted widely through the evolutionary of one specific population and hence is present at a high frequency). The three other ones were found to be individually rare, and as such, did not constitute sound evidence to incorporate a screening for large scale rearrangements in *BRCA* mutation studies at that time.

In 1999, Nadine Puget (Stoppa Lyonnet was part of the authors) published a paper showing four new deletions. This paper was the first evidence of a fairly high frequency of large scale rearrangements in the *BRCA1* gene since it showed a 8% frequency of large scale rearrangements in a pool of 27 North American and 51 French families (Puget *et al.,* 1999). The study concluded that "the search for large rearrangement appears mandatory in *BRCA1* mutation screening studies".

In a study published in 2001, through collaboration between Stoppa Lyonnet's team at the Institut Curie and Dr Aaron Bensimon's Team at the Institut Pasteur, Sophie Gad *et al.*, used Molecular Combing to detect large rearrangements in the *BRCA1* gene in a family previously tested negative for point mutations, providing evidence that polymerase chain reaction (PCR) and sequencing based methods used for genetic testing missed large rearrangements. This study comprised four patients with rearrangements already characterised, as well as ten patients with no previous mutation characterized. It *led to the discovery of an unreported 17kb BRCA1* duplication encompassing exon 3 to 8 (Gad *et al.*, 2001). The study highlighted the ability of molecular combing to detect complex and unreported rearrangements such as inversions and deletions.

These findings were further amplified when in 2001, the same team studied by Molecular Combing a patient previously screened using Myriad BRACAnalysis test. Even though no mutation was detected by BRACAnalysis, the patient carried an 11.6 kb long deletion encompassing exon 13 to 15. This was the first evidence that Myriad did not catch every mutation with its full sequencing test of the *BRCA1* and *BRCA2* genes (Gad *et al.*, 2001). This study led another key opinion leader in the Breast cancer

field, Mary Claire King, to conduct a study on three hundred high risk cases tested negatively by Myriad. This study showed that thirty five of these women, (over twelve percent) did have mutations in *BRCA1* or *BRCA2* despite a negative result from Myriad. These studies are one of the reasons Myriad's patents were amended in Europe.

Molecular combing was also used to help characterize a recombination hotspot upstream of *BRCA1*. The presence of this recombination hotspot between a *BRCA1* pseudogene (*psi-BRCA1*) and the region encompassing *BRCA1* exon 1a, 1b and 2 is responsible for many rearrangements occurring in this part of the gene. (Puget N. *et al.*, 2002)

As it became clear that it was necessary to assess the frequency of rearrangements in populations, and to routinely screen for them when no point mutation was found among patients issued from high risk families, a clinical study searching for large rearrangement was carried out using Molecular Combing. This study, performed in 2002 by Sophie Gad *et al.* examined one hundred and twenty French breast and ovarian cancer families (Gad *et al.*, 2002). It was found that, in this series, deemed representative of the French population, rearrangements accounted for 3.3% of breast-ovarian cancer cases and for 9.5% of the *BRCA1* mutation spectrum. This study also characterized three new rearrangements (a recurrent 23.8kb deletion of exon 8 to 13, a 17.2kb duplication of exon 3 to 8, and an 8.6kb duplication of exon 18 to 20). The paper also stated that "*BRCA1* large rearrangements (...) screening is an important step that should be now systematically included in genetic testing surveys."

Molecular combing was subsequently applied to assess the frequency of *BRCA2* rearrangements. Twenty six high risk families with no point mutations were screened for rearrangements in the *BRCA1* and *BRCA2* genes. The study concluded that although the molecular combing approach could detect rearrangements in both genes, more studies were needed prior to the addition of a systematic screening for large rearrangements in *BRCA2* (Gad *et al.*, 2002).

In 2003, Gad S. and co-workers refined a mutation previously detected by quantitative PCR methods (QMPSF and real-time PCR assays) using molecular combing. It was at the time the largest mutation, a 161kb deletion on the *BRCA1* region, encompassing not only *BRCA1* exon 1 to 22, but also 3 neighbouring genes (*NBR1*, *psi-BRCA1* and *NBR2*) (Gad *et al.*, 2003).

These articles not only show that molecular combing can accurately detect large gene rearrangements in *BRCA1* and *BRCA2*, but also highlight the role it had in defining an accurate map of the region.

It is worth noting that albeit, to date, there are no other published molecular combing studies on the *BRCA1* and *BRCA2* genes, Genomic Vision improved the Genomic Morse Code approach used in the studies described above on these two genes, and carried out a study on 20 blood samples. This more recent study (Cheeseman *et al.*, 2012) discusses the use of molecular combing in a clinical setting.

Patent Opposition

Following the publication of many studies (including the ones described above) on different populations indicating the clinical importance of large rearrangements for genetic testing and because Myriad BRACAnalysis did not catch these, Dominique Stoppa Lyonnet, and others, led several French and European institutions to launch opposition procedures to challenge Myriad's European patents. The main claims were that the inventions lacked novelty, were not inventive, had no industrial application, or were not properly described. They also made the matter widely public by raising policy and ethic issues to the media. The main concern of Dominique Stoppa Lyonnet, was that because Myriad could enforce its patents, it would prevent the development of better test (that were required, in the light of the failure of Myriad's BRACAnalysis test to detect large rearrangements). During the years 2001-2003, tens of articles appeared in French national and regional newspapers, and a couple of them had a wider international reach. In the September 2001 issue of Nature, the Institut Curie "*argues that the patent draws on 'prior' art generated by public genome centres, and also that the sequence used in the initial patent application contained errors, limiting its usefulness*". Stoppa Lyonnet also claimed "*that the automated sequencing method used by the Myriad test to identify deletions and mutations detects only 10-20% of the expected mutations*" (butler & Goodman, 2001). In the Science issue of the same month, the Institut Curie expressed its fear that "*the Myriad patent is too broad and would block the use of other genetically based tests*." Seventeen French genetic testing laboratories discussed the possibility of challenging Myriad's patents before the European Patent Office (Balter, 2001). The Institut Curie test cost a third of the Myriad manufactured test (Financial times limited, September 2001.)

The opposition procedures were launched by the Institut Curie, together with the Assistance Publique-Hôpitaux de Paris, the Institut Gustave Roussy, the Belgian Human Genetics Society, as well as German, Dutch, Czech, Austrian, Swiss, British and Finnish genetic societies, and several patient associations. This unofficial consortium was successful in widely limiting the scope of Myriad's patents: after all the opposition and appeal procedures Myriad's patent EP 699 754 covering the diagnostic use of *BRCA1* was restricted to certain mutations of the genes and to diagnostic methods for their identification. The patent over *BRCA1* itself was maintained in an amended form. Patents over *BRCA2* also were amended (EPO press release, 2005).

Even if the EPO and its institutions focused on technical matters, the policy concerns over Myriad having the possibility to block research by enforcing its patent, as well as to gain a decisive advantage by building a database due to its mandatory in-house Utah laboratory sequencing certainly played an important role. It is worth mentioning that Myriad always claimed that it had no intention to enforce its patents on researchers conducting their own research on *BRCA1* and *BRCA2*. Moreover, Myriad contributed its detected mutation to the Breast Cancer Information Core database, making the data available for academic research. Myriad had the position that research contribution to public database could only benefit to their test, as it further emphasized the importance of *BRCA* testing.

Myriad also introduced as soon as it could a test to detect five common large rearrangements (part of the BRACAnalysis since august 12, 2002, BRCA- 5 site Large rearrangement Panel, Myriad website),

which still missed other large rearrangements, and by 2006 introduced BART, a BRACAnalaysis Rearrangement Test, which detects large rearrangements in the promoter, coding sequence (exons) and flanking regions of BRCA1 and 2. (The test is made of a set of eleven multiplex quantitative endpoint PCR (Wenstrup *et al*., 2007). however this test is part of the original analysis (and thus requires no additional fee) only when the patient meets myriad highly stringent criteria, and when the other BRACAnalysis test results were negative. In the US, many healthcare providers will not pay the additional pricing Myriad requires to have it performed (700$), rendering BART available to only a small subset of patient. Blue Cross/Blue Shield and Blue care Network even issued a Medical Policy in which it states that "The BRACAnalysis Rearrangement Test (BART) is experimental/investigational. It has not been scientifically demonstrated to have additional predictive value to the current BRCA testing."

Their rationale is that "there is no published medical literature about the efficacy of BARTtesting. What little information available can only be found on the Myriad Laboratories Website." Health Net Inc. has the same attitude as blue shield/blue cross ("Health Net Inc. considers BRACAnalysis Rearrangement Test (BART) for screening of the general population or re-testing previously tested high-risk members for large genomic re-arrangements not medically necessary due to lack of information in the peer review literature validating this utility of this test or its effect on patient outcomes."). In fact, Myriad is there victim of its own monopoly, and no institution has conducted a neutral study on BART's efficiency. The only neutral study that assessed BART testing has only been recently published and concerns which individuals need BART testing. Shannon and co-workers ordered BART testing on 257 patients, of which only 53 met Myriad's stringent criteria for BART testing. They found 5 large rearrangements using BART, 2 of which only would have been found if following Myriad's guidelines. (Shannon *et al.*, 2011) The study concludes (two other studies converge towards the same conclusion (Hansen *et al.*, 2009, Armaou *et al.*, 2009)) that "*when BRCA1 or BRCA2 genetic testing is performed, testing should always include large gene rearrangements testing so that the results are the most comprehensive and reliable.*" Of note, is the fact that the technology seems similar to what a normal research or genetic testing laboratory can perform (quantitative PCR, and its variant are common laboratory processes), and as such is not innovative. All these facts and especially the extra-cost of the BART test led to contestation against Myriad. Yale researchers addressed an open letter to Myriad asking them include BART as a full part of the BRACAnalysis test for instance (Yale cancer genetic counselling open letter, 2011).

The policy storm eventually reached Myriad's US patent, when in March 2010, the US federal court judge Robert W Sweet invalidated many of Myriad Genetics' patent claims on BRCA1 and BRCA2 (Matloff & Brierley, 2010)

Conclusion

The discovery of the *BRCA1* and *BRCA2* genes turned what is a real scientific and clinical advance in an inextricable case study in which policy, ethical issues, scientific competition, medical concerns, and differences in patent and public health laws and regulation worldwide melt together. What remains clear is the need for both point mutations testing by sequencing and large gene

rearrangement testing by as accurate a technology as possible. The later point raises questions as to which technology or array of technologies is best suited to meet the particular demands of clinical testing. To asses this question, one must take into account the fact that different countries possess different clinical structures, and as such, one or several centres will perform the testing thus changing the requirement of the technology used. To date, several technologies have been employed, in different academic or research settings. Of these technologies, Molecular Combing has been successfully employed in the better understanding of the structural organisation of the *BRCA1* and *BRCA2* regions, as well as in the identification of many different rearrangements, ranging widely in type (deletions and duplications), in sizes (a few kilobases to over one hundred kilobases) and clinical implications. As such Molecular Combing represents an all track alternative technology in the identification of large gene rearrangements in Breast and Ovarian cancers.

References:

1: Gad S, Klinger M, Caux-Moncoutier V, Pages-Berhouet S, Gauthier-Villars M, Coupier I, Bensimon A, Aurias A, Stoppa-Lyonnet D. Bar code screening on combed DNA for large rearrangements of the BRCA1 and BRCA2 genes in French breast cancer families. J Med Genet. 2002 Nov;39(11):817-21. PubMed PMID: 12414821; PubMed Central PMCID: PMC1735012.

2: Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, Roach KC, Mandell J, Lee MK, Ciernikova S, Foretova L, Soucek P, King MC. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA. 2006 Mar 22;295(12):1379-88. PubMed PMID: 16551709.

3: Tan DS, Marchiò C, Reis-Filho JS. Hereditary breast cancer: from molecular pathology to tailored therapies. J Clin Pathol. 2008 Oct;61(10):1073-82. Epub 2008 Aug 4. Review. PubMed PMID: 18682420.

4: King MC, Marks JH, Mandell JB; New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science. 2003 Oct 24;302(5645):643-6. PubMed PMID: 14576434.

5: Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 1990 Dec21;250(4988):1684-9. PubMed PMID: 2270482.

6: Narod SA, Feunteun J, Lynch HT, Watson P, Conway T, Lynch J, Lenoir GM. Familial breast-ovarian cancer locus on chromosome 17q12-q23. Lancet. 1991 Jul 13;338(8759):82-3. PubMed PMID: 1676470.

7: Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994 Oct 7;266(5182):66-71. PubMed PMID: 7545954.

8: Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995 Dec 21-28;378(6559):789-92. Erratum in: Nature 1996 Feb 22;379(6567):749. PubMed PMID: 8524414.

9: Tavtigian SV, Simard J, Rommens J, Couch F, Shattuck-Eidens D, Neuhausen S, Merajver S, Thorlacius S, Offit K, Stoppa-Lyonnet D, Belanger C, Bell R, Berry S,Bogden R, Chen Q, Davis T, Dumont M, Frye C, Hattier T, Jammulapati S, Janecki T, Jiang P, Kehrer R, Leblanc JF, Mitchell JT, McArthur-Morrison J, Nguyen K, Peng Y, Samson C, Schroeder M, Snyder SC, Steele L, Stringfellow M, Stroup C, Swedlund B, Swense J, Teng D, Thomas A, Tran T, Tranchant M, Weaver-Feldhaus J, Wong AK, Shizuya H, Eyfjord JE, Cannon-Albright L, Tranchant M, Labrie F, Skolnick MH, Weber B, Kamb A, Goldgar DE. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. Nat Genet. 1996 Mar;12(3):333-7. PubMed PMID: 8589730.

10: Gold ER, Carbone J. Myriad Genetics: In the eye of the policy storm. Genet Med. 2010 Apr;12(4 Suppl):S39-70. PubMed PMID: 20393310; PubMed Central PMCID: PMC3037261.

11: Benowitz S. European groups oppose Myriad's latest patent on BRCA1. J Natl Cancer Inst. 2003 Jan 1;95(1):8-9. PubMed PMID: 12509391.

12: Lecrubier A. Patents and public health. European institutions are challenging Myriad Genetics's patent monopoly on the brca1 gene. EMBO Rep. 2002 Dec;3(12):1120-2. PubMed PMID: 12475923; PubMed Central PMCID: PMC1308311.

13: Petrij-Bosch A, Peelen T, van Vliet M, van Eijk R, Olmer R, Drüsedau M, Hogervorst FB, Hageman S, Arts PJ, Ligtenberg MJ, Meijers-Heijboer H, Klijn JG, Vasen HF, Cornelisse CJ, van 't Veer LJ, Bakker E, van Ommen GJ, Devilee P. BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. Nat Genet. 1997 Nov;17(3):341-5. Erratum in: Nat Genet 1997 Dec;17(4):503. PubMed PMID: 9354803.

14: Puget N, Stoppa-Lyonnet D, Sinilnikova OM, Pagès S, Lynch HT, Lenoir GM, Mazoyer S. Screening for germ-line rearrangements and regulatory mutations in BRCA1 led to the identification of four new deletions. Cancer Res. 1999 Jan 15;59(2):455-61. PubMed PMID: 9927062.

15: Gad S, Aurias A, Puget N, Mairal A, Schurra C, Montagna M, Pages S, Caux V, Mazoyer S, Bensimon A, Stoppa-Lyonnet D. Color bar coding the BRCA1 gene on combed DNA: a useful strategy for detecting large gene rearrangements. Genes Chromosomes Cancer. 2001 May;31(1):75-84. PubMed PMID: 11284038.

16: Gad S, Scheuner MT, Pages-Berhouet S, Caux-Moncoutier V, Bensimon A, Aurias A, Pinto M, Stoppa-Lyonnet D. Identification of a large rearrangement of the BRCA1 gene using colour bar code on combed DNA in an American breast/ovarian cancer family previously studied by direct sequencing. J Med Genet. 2001 Jun;38(6):388-92. PubMed PMID: 11424920; PubMed Central PMCID: PMC1734901.

17: Puget N, Gad S, Perrin-Vidoz L, Sinilnikova OM, Stoppa-Lyonnet D, Lenoir GM, Mazoyer S. Distinct BRCA1 rearrangements involving the BRCA1 pseudogene suggest the existence of a recombination hot spot. Am J Hum Genet. 2002 Apr;70(4):858-65. Epub 2002 Mar 5. PubMed PMID: 11880951; PubMed Central PMCID: PMC379114. 18: Gad S, Caux-Moncoutier V, Pagès-Berhouet S, Gauthier-Villars M, Coupier I, Pujol P, Frénay M, Gilbert B, Maugard C, Bignon YJ, Chevrier A, Rossi A, Fricker JP, Nguyen TD, Demange L, Aurias A, Bensimon A, Stoppa-Lyonnet D. Significant contribution of large BRCA1 gene rearrangements in 120 French breast and ovarian cancer families. Oncogene. 2002 Oct 3;21(44):6841-7. PubMed PMID: 12360411.

19: Gad S, Bièche I, Barrois M, Casilli F, Pages-Berhouet S, Dehainault C, Gauthier-Villars M, Bensimon A, Aurias A, Lidereau R, Bressac-de Paillerets B, Tosi M, Mazoyer S, Stoppa-Lyonnet D. Characterisation of a 161 kb deletion extending from the NBR1 to the BRCA1 genes in a French breast-ovarian cancer family. Hum Mutat. 2003 Jun;21(6):654. PubMed PMID: 14961556.

20 : Cheeseman K, Rouleau E, Vannier A, Thomas A, Briaux A, Lefol C, Walrafen P, Bensimon A, Lidereau R, Conseiller C, Ceppi M. A predictive genetic test for the physical mapping of germline rearrangements in the susceptibility breast cancer genes BRCA1 and BRCA2, 2011, under submission.

21: Butler D, Goodman S. French researchers take a stand against cancer gene patent. Nature. 2001 Sep 13;413(6852):95-6. PubMed PMID: 11557932.

22: Balter M. Cancer research. Transatlantic war over BRCA1 patent. Science. 2001 Jun 8;292(5523):1818. PubMed PMID: 11397923.

23: http://www.timeshighereducation.co.uk/story.asp?storyCode=197099&sectioncode=26

24: https://www.myriadpro.com/BRAC_BART

25: http://meeting.ascopubs.org/cgi/content/abstract/25/18_suppl/10513

26: Blue Cross Blue/Shield, Blue Care Network of Michigan Medical Policy, “Genetic Testing – BRAC (breast cancer gene) Analysis Rearrangement Testing, 5/1/10.

27: Hhealth Net Inc. National Medical Policy, “genetic Testing for BRCA1 and BRCA2, Policy Number NMP136, updated July 2011.

28: Shannon KM, Rodgers LH, Chan-Smutko G, Patel D, Gabree M, Ryan PD. Which individuals undergoing BRACAnalysis need BART testing? Cancer Genet. 2011 Aug;204(8):416-22. PubMed PMID: 21962891.

29: Hansen TO, Jønson L, Albrechtsen A, Andersen MK, Ejlertsen B, Nielsen FC. Large BRCA1 and BRCA2 genomic rearrangements in Danish high risk breast-ovarian cancer families. Breast Cancer Res Treat. 2009 May;115(2):315-23. Epub 2008 Jun 12. PubMed PMID: 18546071.

30: Armaou S, Pertesi M, Fostira F, Thodi G, Athanasopoulos PS, Kamakari S, Athanasiou A, Gogas H, Yannoukakos D, Fountzilas G, Konstantopoulou I. Contribution of BRCA1 germ-line mutations to breast cancer in Greece: a hospital-based study of 987 unselected breast cancer cases. Br J Cancer. 2009 Jul 7;101(1):32-7. Epub 2009 Jun 2. PubMed PMID: 19491894; PubMed Central PMCID: PMC2713692.

31: http://yalecancergeneticcounseling.blogspot.com/2011/07/open-letter-to-myriad-genetics.html

32: Matloff ET, Brierley KL. The double-helix derailed: the story of the BRCA patent. Lancet. 2010 Jul 31;376(9738):314-5. PubMed PMID: 20674708.

Patent application n°1:

US 20130130246A1

(19) **United States**

(12) **Patent Application Publication**

BENSIMON et al.

(10) **Pub. No.: US 2013/0130246 A1**

(43) **Pub. Date: May 23, 2013**

(54) **METHODS FOR THE DETECTION, VISUALIZATION AND HIGH RESOLUTION PHYSICAL MAPPING OF GENOMIC REARRANGEMENTS IN BREAST AND OVARIAN CANCER GENES AND LOCI BRCA1 AND BRCA2 USING GENOMIC MORSE CODE IN CONJUNCTION WITH MOLECULAR COMBING**

(71) Applicants: **Aaron BENSIMON**, Anthony (FR); **Maurizio Ceppi**, Issy-Les-Moulineaux (FR); **Kevin Cheeseman**, Champigny-Sur-Marne (FR); **Emmanuel Conseiller**, Paris (FR); **Pierre Walrafen**, Montrouge (FR)

(72) Inventors: **Aaron BENSIMON**, Anthony (FR); **Maurizio Ceppi**, Issy-Les-Moulineaux (FR); **Kevin Cheeseman**, Champigny-Sur-Marne (FR); **Emmanuel Conseiller**, Paris (FR); **Pierre Walrafen**, Montrouge (FR)

(21) Appl. No.: **13/665,404**

(22) Filed: **Oct. 31, 2012**

Related U.S. Application Data

(60) Provisional application No. 61/553,906, filed on Oct. 31, 2011.

Publication Classification

(51) **Int. Cl.**
C12Q 1/68 (2006.01)

(52) **U.S. Cl.**
CPC ***C12Q 1/6886*** (2013.01)
USPC **435/6.11**

(57) **ABSTRACT**

Methods for detecting genomic rearrangements in BRCA1 and BRCA2 genes at high resolution using Molecular Combing and for determining a predisposition to a disease or disorder associated with these rearrangements including predisposition to ovarian cancer or breast cancer. Primers useful for producing probes for this method and kits for practicing the methods.

Patent application n°2:

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 December 2012 (06.12.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/164401 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/IB2012/001333

(22) International Filing Date:
1 June 2012 (01.06.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/493,010 3 June 2011 (03.06.2011) US

(71) Applicants *(for all designated States except US)*: **GENOMIC VISION** [FR/FR]; 80-84 rue des Meuniers, F-92220 Bagneux (FR). **CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE** [FR/FR]; 3, rue Michel-Ange, F-75794 Paris Cedex 16 (FR). **UNIVERSITE CLAUDE BERNARD DE LYON 1** [FR/FR]; 43 Bd du 11 Novembre 1918, F-69622 Villeurbanne Cedex (FR). **CENTRE DE LUTTE CONTRE LE CANCER LÉON BÉRARD** [FR/FR]; 28 Rue Laennec, F-69008 Lyon (FR).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **MAZOYER, Sylvie** [FR/FR]; 40 ter rue Seigneurmartin, F-69008 Lyon (FR). **TESSEREAU, Chloé** [FR/FR]; 72 rue Jaboulay, F-69007 Lyon (FR). **CEPPI, Maurizio** [CH/FR]; 2bis Henri Tariel, F-92130 Issy Les Moulineaux (FR). **CHEESEMAN, Kevin** [FR/FR]; 20 avenue Edmond, F-94500 Champigny Sur Marne (FR). **VANNIER, Anne** [FR/FR]; 17 rue Guy de Maupassant, F-76280 Saint-Jouin-Bruneval (FR).

(74) Agent: **GUTMANN, Ernest**; 3, rue Auber, F-75009 Paris (FR).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

WO 2012/164401 A1

(54) Title: ASSESSMENT OF CANCER RISK BASED ON *RNU2* CNV AND INTERPLAY BETWEEN *RNU2* CNV AND *BRCA1*

(57) Abstract: Polynucleotides useful for detecting copy number variation of *RNU2* sequences and methods of assessing risk of developing breast or ovarian cancer using molecular combing and/or detection or quantification of *BRCA1* expression.

Poster presented at the joint Welcome trust Sanger Institute & Cold Spring Harbor Laboratories "Genome Informatics" meeting in September 2010:

Molecular Combing Physical Maps: Improving Assemblies and studying genomic landscape/variability.

Kevin Cheeseman[1,2], GraceYao[1], Aaron Bensimon[1], Emmanuel Conseiller[1], Serge Casaregola[3], Pierre Renault[2], Maurizio Ceppi[1]

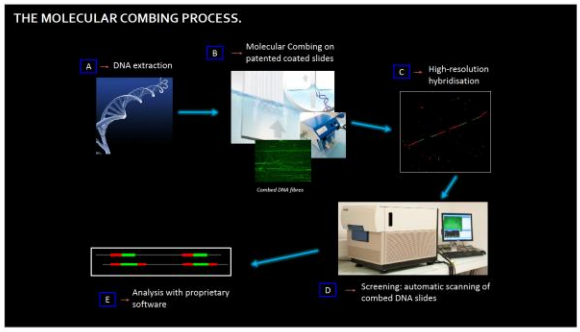
[1] Genomic Vision, Paris, France; [2] INRA, Jouy-en-Josas, France; [3] INRA, Thiverval-Grignon, France.

ABSTRACT

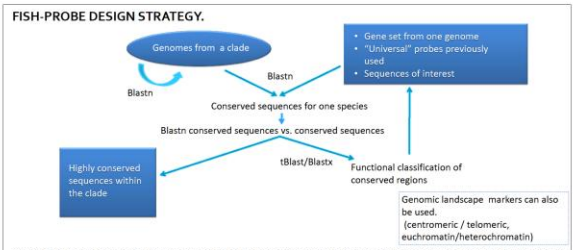
The arrival of new sequencing technologies has made sequencing easily accessible for small laboratories. However, the revolution brought by next-generation sequencing technologies has not come without drawbacks. The reads lengths, even if constantly increasing, are small, and the dedicated assemblers generate more fragmented assemblies, mostly due to difficulties in the resolution of repeats and sequencing biases. As a result, albeit considerably reduced sequencing costs and increased throughput, assembly validation and finishing steps remain labour-intensive and costly, and thus are often left aside sequencing projects, with contig- or scaffold-state sequences being the end product.

Molecular Combing represents an alternative and complementary technology, able to improve the quality of assemblies. Molecular Combing is a technology allowing to linearly and homogeneously stretch hundreds to thousands of genomes under the form of lone DNA fibres on a single microscopy coverslip. This technology enables the accurate positioning and measurement of genomic entities on the DNA molecule.
By providing a means to easily position, order and orientate all contigs from a *de novo* sequencing project on a whole genome Combing-based physical map, the effort required to obtain a high quality genome sequence is considerably reduced. A proof of concept on several genomes with different levels of complexity, and belonging to the genus of bacteria, yeasts and fungi is currently under development. The generation of high resolution, Molecular Combing-based physical maps does not only provide a means to improve assemblies, but also to study genomic structure and variability between strains or species. This methodology should help reduce experiments and time required to obtain high level assemblies or finished genomes.

THE MOLECULAR COMBING PROCESS.

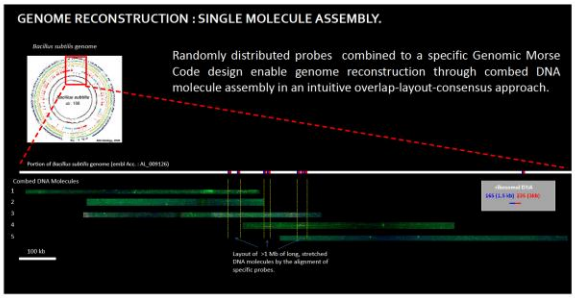


FISH-PROBE DESIGN STRATEGY.

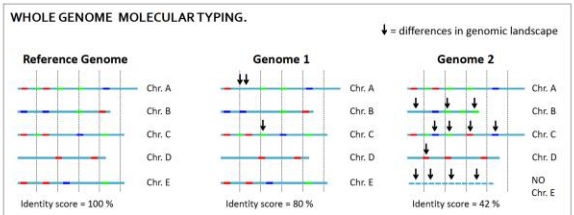


Organisms from one clade share conserved sequences. These evolutionarily conserved sequences can be used as probes distributed along the genome to create an assembly of Combed DNA fibres, thus generating a high resolution molecule-based physical map called "RefMap".

GENOME RECONSTRUCTION : SINGLE MOLECULE ASSEMBLY.

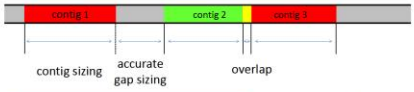


WHOLE GENOME MOLECULAR TYPING.



Genome 1 and Reference Genome are closely related strains belonging to the same species, as shown by the few differences between the Molecular Combing physical maps (Identity = 80%). The physical map of Genome 2 has fewer chromosomes and a completely different genomic landscape (Identity = 42%). Organism 2 belongs to a different and more distantly-related species. Molecular Combing physical maps could also be used to monitor genetic elements of importance, like trait-associated CNVs, pathogenicity islands or antibiotic resistance genes.

IMPROVING *DE NOVO* GENOME ASSEMBLIES BY ACCURATELY ORDERING AND SCAFFOLDING CONTIGS.

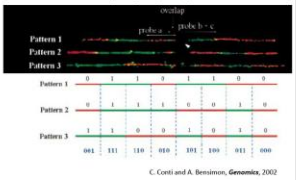


$$\#hybridisations = \frac{\ln(\#contigs)}{\ln(\#colours)}$$

According to the above equation, in a 3 colours setting, with only 5 hybridisations, 243 contigs could be mapped.

Theoretically, more than 2000 contigs could be accurately and unambiguously mapped using only 7 hybridisations, as a combinatorial strategy increases logarithmically the number of contigs that can be mapped.

A combinatorial approach successfully and accurately mapped 8 contigs spanning 300 kb of the human chromosome 11, requiring only 3 experiments.



C. Conti and A. Bensimon, *Genomics*, 2002

Mapping contigs on combed DNA should help decrease the parametric complexity of bioinformatic-based sequence assembly, as well as considerably reducing the finishing step, which today remains costly and labour intensive.

PROSPECTS & TECHNICAL CHALLENGES :

• The technology and methodology proposed here have the potential to improve *de-novo* genome assemblies. By lowering the complexity of sequence assembly and finishing, it should decrease cost and time required to obtain high quality genome sequences.

• The main challenges of the concept proposed here reside in the development of several algorithms for the assembly of combed DNA fibres as well as algorithms for the mapping and identification of probes on these assemblies of combed DNA molecules. An algorithm helping the combinatorial repartition of probes or contigs between colours will also be developed.

• The generation of genome wide, high resolution Molecular Combing physical maps also represent in itself an innovative genotyping method. To date, most of the genotyping methods are based on only a few biomarkers. Observing genomic variability at the whole genome level could provide a means to easily and reliably distinguish between strains or species.

e-mail: k.cheeseman@genomicvision.com

Further readings:

• Dujon B, *et al.*, **Genome evolution in yeasts**. *Nature* 2004.
• Herrick J, Bensimon A. **Introduction to molecular combing: genomics, DNA replication, and cancer.** *Methods Mol Biol*. 2009. Review.
• Latreille P, *et al.*, **Optical mapping as a routine tool for bacterial genome sequence finishing.** *BMC Genomics*. 2007
• Nagarajan N, Pop M. **Parametric complexity of sequence assembly: theory and applications to next generation sequencing.** *J Comput Biol*. 2009.
• Metzker ML. **Sequencing technologies - the next generation**. *Nat Rev Genet*. 2010. Review.
• Lebofsky R, Bensimon A. **Single DNA molecule analysis: applications of molecular combing.** *Brief Funct Genomic Proteomic*. 2003. Review.
• Conti C, Bensimon A., **A combinatorial approach for fast, high-resolution mapping.** *Genomics*. 2002.

Summary:

Penicillium species are filamentous fungi belonging to the Ascomycota genus. *Penicillium* species have been used by Man for centuries in food making processes. More recently they have also been used in the biotechnology industry for the production of compounds of pharmaceutical interest. Some *Penicillium* species are food spoilage agents, pathogens of plants including fruits. Aspects of their genomics are largely unknown. In this study, we analysed the genomes of two newly sequenced species*, Penicillium roqueforti and Penicillium camemberti.* Here we report the development of a new methodology for improving and validating genome assembly using an original single DNA molecule technology, Molecular Combing. Using this methodology we were able to produce a high quality genome assembly of *Penicillium roqueforti.* This work also reports the multiple and recurrent horizontal transfer of a large genomic island of over half a megabase between several *Penicillium* species. This horizontal transfer indicates a higher frequency of lateral genetic exchange between cheesemaking fungi than previously expected. Finally, we present an early assessment of the genomic potential for secondary metabolite production in these important food associated *Penicillia.*

Résumé:

Les *Penicillia* sont des champignons filamenteux appartenant au genre Ascomycota. Ces champignons ont été utilisés par l'homme pour la production de nourriture depuis des siècles. Plus récemment, ils ont aussi été utilisés dans l'industrie biotechnologique pour la production de composés chimiques d'intérêts pharmaceutiques. Certaines espèces de *Penicillium* sont par ailleurs des moisissures contaminants certains aliments, d'autres sont des pathogènes de plantes, y compris de certains fruits. Leur génomique est globalement peut connue. Dans cette étude, nous avons analysé les génomes de deux espèces nouvellement séquencées, *Penicillium roqueforti* et *Penicillium camemberti*. Nous reportons ici le développement d'une nouvelle méthodologie pour l'amélioration et la validation d'assemblage de génomes en utilisant une technologie permettant l'observation de molécules d'ADN unique, le Peignage Moléculaire. En utilisant cette méthode, nous avons amélioré l'assemblage de *Penicillium roqueforti*. Ce manuscrit décrit aussi de multiples occurrences d'un transfert horizontal d'un ilot génomique de plus de cinq cent kilobases entre plusieurs *Penicillium.* Ce cas de transfert horizontal indique une fréquence d'échange latéral de matériel génétique plus forte qu'attendue. Enfin nous présentons un inventaire préliminaire du potentiel génomique pour la production de métabolites secondaires dans ces importants *Penicillia* alimentaires.