Evaluating and improving next generation sequencing performance in the absence of gold standard

Yue Zhai

THÈSE de DOCTORAT DE
L'UNIVERSITÉ CLAUDE BERNARD LYON 1

École Doctorale N° 341
Évolution, Écosystèmes, Microbiologie, Modélisation (E2M2)

Discipline: Biostatistiques

Soutenue publiquement le 11/07/2023, par :
Yue ZHAI

Évaluation et amélioration des performances du séquençage de nouvelle génération en absence de gold standard

Evaluating and improving next generation sequencing performance in the absence of gold standard

---

Devant le jury composé de :

PERDRY, Hervé	MCU	Université Paris-Saclay	Rapporteur
TEZENAS DU MONTCEL, Sophie	MCU-PH	Sorbonne Université	Rapporteure
JANNOT, Anne-Sophie	MCU-PH	Université de Paris Cité	Examinatrice
MAUCORT-BOULCH, Delphine	PU-PH	Université Lyon 1	Présidente, Examinatrice
ROY, Pascal	PU-PH	Université Lyon 1	Directeur de thèse
LESCA, Gaëtan	PU-PH	Université Lyon 1	Co-directeur de thèse
VIALLON, Vivian	Chercheur	CIRC Lyon	Invité

Abstract

Evaluating next-generation sequencing (NGS) performance suffers frequently from the absence of gold standard. Without gold standard, researchers often carry out replicates from the same individual and use concordance between replicates to evaluate NGS performance, whereas the appropriateness of that criterion is still debated. Furthermore, for a better performance, the replicates are often combined using various models to reconstruct a new high-performance callset.

This work aimed to investigate these two aspects of NGS performance evaluation and improvement in the absence of gold standard. In the first part, we examined the contributions and limitations of the concordance-discordance criterion. We analyzed the relationship between the probability of discordance and that of error using conditional probability under conditional independence and conditional dependence between two sequencing results. We compared the probabilities of discordance and error with various combinations of sensitivity, specificity, and correlation between replicates, then on real results of sequencing genome NA12878. We examined covariate effects on discordance and error using generalized additive models with smooth functions. The results showed that, with conditional independence of two sequencing results, the concordance-discordance criterion seems acceptable; however, it becomes questionable in presence of high correlation because of high percentages of false concordant results. Covariate effects' functional forms were close between discordance and error models, though the parts of covariate-explained deviance differed.

In the second part, we investigated the statistical methods able to combine callsets from replicates to reconstruct a new callset. Three technical replicates of genome NA12878 were considered and five model types were compared (consensus, latent class, Gaussian mixture, Kamila–adapted k-means, and random forest) regarding four performance indicators: sensitivity, precision, accuracy, and F1-score. We concluded that the compared non-supervised clustering models that combine multiple callsets are able to improve sequencing performance vs. supervised models previously tested elsewhere. Among the models compared, the Gaussian mixture model and Kamila offered non-negligible precision and F1-score improvements. These models may be recommended for callset reconstruction (from either biological or technical replicates) for diagnostic or precision medicine purposes.

Summary

Chapter 1 presents the general context of this work as well as some necessary background concepts. In Chapter 1, we first review briefly the history and principles of next-generation sequencing and then detail the steps of the sequencing stage and bioinformatics stage. An Illumina sequencing platform is used for the example that illustrate the sequencing process and the Burrows-Wheeler Aligners and GATK variant caller are used as main examples to describe the underlying statistical models of bioinformatics tools. We focus particularly on explaining the concepts and statistical principles as well as their evolution over time but not the implementation algorithms or computer programs. The generation and signification of output quality scores are also described as they are of great importance in quality analyses. Finally, recommended filtering strategies are discussed, including hard filters and soft filters implemented in the GATK workflow.

Chapter 2 introduces the research question of the work; that is, the methodology used to evaluate and improve the quality of NGS callsets. We first discuss the sources of errors in the NGS process; the main stemming either from the experimental steps or the bioinformatics analysis. Then we discuss the currently most widely used methods to evaluate the performance of a given NGS data set in situations with available "gold standard" set and absence of "gold-standard" set. We finish with a brief review of the attempts and researchers' findings regarding error detection and reduction in a variant calling output.

Chapter 3 includes theoretical work and real-data analyses aiming to evaluate the appropriateness of the concordance-discordance model, a model widely used in performance evaluation of NGS data. In the absence of gold standard, researchers are often compelled to use the concordance between several sequencing results as a substitution criterion, the discordance results are then interpreted as errors. However, whether the discordance rate corresponds to the error rate remains unclear. We first analyse the theoretical relationships between the error rate and the discordance rate under conditional independence and dependence and apply the principles of performance evaluation of diagnostic tests to the domain of NGS where the general concept of 'test' refers to a NGS process. This is then illustrated with simulations of various situations as well as data on three NA12878 genome

replicates. Finally, differences between estimates of covariate effects associated with error and discordance are examined.

In Chapter 3, we show that in case of conditional independence between two sequencing results, the overall probability of error in concordant results being negligible, the concordance-discordance method is acceptable. However, in settings with high correlation levels, the method becomes questionable because of a high proportion of false concordant results. With real data from NA12878 vs. GIAB benchmark set, discordance (as indicator of error) seemed acceptable but with caution in interpreting discordant or concordant results. Multivariate analyses showed substantial differences between error and discordance models; thus, caution is required in using the concordance criterion, especially in case of highly correlated results.

Chapter 4 looks into other models implemented to obtain a combined call sets from replicates. We focus on exploring non-supervised models instead of supervised models because the latter often require high-quality training data that are not always available. The literature about processing replicate sequencing results with non-supervised models is rather scanty and available models have been rarely objectively compared. We therefore explored the main models of dealing with several NGS results stemming from biological or technical replicates, investigated their properties, and compare their key performance indicators to help choosing the most performant among readily implementable methods able to improve sequencing performance. Section 4.1 aims to present the research context and provide a literature review of the methodology in related works. In section 4.2, we address the question in the statistical world as a clustering problem and give an overview of the major categories of clustering models. Then we apply representative models of each category to three technical replicates of the NA12878 genome and compare their performances in section 4.3 to 4.5. Precisely, we explore the consensus model, the latent class model, the mixture model, and random forest regarding their abilities to produce a callset with improved quality. We also compare the main performance indicators of these models; i.e., precision, recall, and F1-score.

In Chapter 4, we show that the non-supervised clustering models compared were all able to improve sequencing performance in terms of precision and F1-score, which is comparable to

what is reported about supervised models. Among the models compared, the Gaussian mixture model and Kamila offered improvements that made precision higher than 99% and F1-score close to 99%. Therefore, these models may be recommended to reconstruct new high-performance callsets from NGS replicates. This is of particular interest for diagnosis or precision medicine whenever DNA sequencing results stem from either biological replicates (more than one sample) or technological replicates (more than one sequencing platform or analysis pipeline).

Résumé

Les travaux de cette thèse ont porté sur la place du modèle de concordance-discordance dans l'évaluation de la performance du séquençage à haut débit et sur des comparaisons de performance entre modèles de classification dans la reconstitution de résultats de séquençage haut débit à partir de réplicats techniques.

Le premier travail a étudié le problème de l'évaluation du séquençage haut débit en l'absence de 'gold' standard et, dans ce cadre, la pertinence des critères de concordance-discordance. Il a examiné les relations entre le taux de discordance et le taux d'erreur dans diverses situations théoriques. Il a ensuite analysé les effets des covariables sur ces deux taux en utilisant un modèle additif généralisé avec des données réelles issues de réplicats de séquençage du génome NA12878.

Le second travail a étudié le problème de la fusion de résultats de séquençage haut débit de réplicats techniques en vue d'obtenir un nouveau jeu de données susceptible de comporter moins d'erreurs. Il a évalué et comparé les aptitudes des principaux modèles de partitionnement à améliorer la performance finale du séquençage à partir des résultats de trois séquençages du génome NA12878. L'étude fournit des arguments pour choisir le modèle le plus convenable et utiliser ces résultats en matière de diagnostic ou de médecine de précision.

Le Chapitre 1 présente le contexte général de ce travail, ainsi que certains indispensables concepts de base. Dans ce chapitre, nous commençons par rappeler brièvement l'histoire et les principes du séquençage de nouvelle génération, puis nous détaillons les étapes de la phase de séquençage et de la phase bioinformatique. Une plateforme de séquençage Illumina est utilisée pour illustrer le processus de séquençage, les aligneurs de Burrows-Wheeler (BWA) et l'appelant de variants GATK comme exemples principaux pour décrire les algorithmes et les modèles statistiques sous-jacents des outils bioinformatiques. Nous nous concentrons particulièrement sur l'explication des principes conceptuels et statistiques ainsi que sur leur évolution dans le temps, mais pas sur les algorithmes de mise en œuvre ou les programmes informatiques. La génération et la signification des scores de qualité des résultats sont

également décrites car elles sont d'une grande importance dans les analyses de la qualité. Enfin, les stratégies de filtrage recommandées sont également discutées, y compris les filtres durs et les filtres doux mis en œuvre dans GATK.

Le Chapitre 2 présente la question de la recherche ; à savoir, la méthodologie utilisée pour évaluer et améliorer la qualité des callsets NGS. Nous examinons d'abord les sources d'erreur qui interviennent dans le processus NGS et dont les principales proviennent soit des étapes expérimentales, soit de l'analyse bioinformatique. Nous examinons ensuite les méthodes les plus largement utilisées aujourd'hui pour évaluer les performances des séquençages NGS dans des situations d'existence et d'absence d'un 'gold standard'. Nous terminons par un bref examen des essais et des résultats des chercheurs concernant la détection et la réduction des erreurs inhérentes à un résultat d'appel de variants.

Le Chapitre 3 comprend un travail théorique et des analyses de données réelles visant à évaluer la pertinence du modèle de concordance-discordance, un modèle largement utilisé dans l'évaluation des performances de données NGS. En l'absence d'étalon-or, les chercheurs sont souvent obligés d'utiliser la concordance entre plusieurs résultats de séquençage comme critère de substitution, les résultats discordants étant alors interprétés comme des erreurs. Toutefois, il n'est pas certain que le taux de discordance corresponde au taux d'erreur. Nous analysons d'abord les relations théoriques entre le taux d'erreur et le taux de discordance en cas d'indépendance et de dépendance conditionnelles, puis appliquons les principes d'évaluation de performance des tests diagnostiques au domaine de la NGS où le concept général de 'test' désigne un processus de NGS. Ceci est ensuite illustré par diverses simulations de situations ainsi que par des données provenant de trois réplicats du génome NA12878. Enfin, nous examinons les différences entre les estimations des effets des covariables associées à l'erreur et à la discordance.

Dans le Chapitre 3, nous concluons qu'en cas d'indépendance conditionnelle entre deux résultats de séquençage, la méthode de concordance-discordance est acceptable parce que la probabilité globale d'erreur dans les résultats concordants est négligeable. Toutefois, lorsque les niveaux de corrélation sont élevés, la méthode devient discutable en raison de la proportion élevée de faux résultats concordants. Avec des données réelles sur NA12878 par

rapport à l'ensemble de référence GIAB, la discordance (en tant qu'indicateur d'erreur) semble acceptable, mais une prudence s'impose dans l'interprétation des résultats discordants ou concordants. Des analyses multivariées ont montré des différences substantielles entre les modèles d'erreur et les modèles de discordance ; il convient donc d'être prudent dans l'utilisation du critère de concordance, surtout en cas de résultats fortement corrélés.

Le Chapitre 4 examine d'autres modèles mis en œuvre pour combiner des ensembles d'appels provenant de réplicats. Nous nous concentrons sur l'exploration de modèles non supervisés plutôt que supervisés parce que ces derniers nécessitent souvent des données d'apprentissage de haute qualité qui ne sont pas toujours disponibles. La littérature sur le traitement des résultats de séquençage de réplicats à l'aide de modèles non supervisés est plutôt rare et les différents modèles utilisables ont été rarement objectivement comparés. Nous avons donc exploré les principaux modèles destinés au traitement de résultats NGS provenant de réplicats biologiques ou techniques, étudié leurs propriétés et de comparé leurs principaux indicateurs de performance pour aider à choisir les méthodes les plus performantes parmi celles qui peuvent être facilement mises en œuvre et qui sont capables d'améliorer les performances de séquençage. La section 4.1 vise à présenter le contexte de la recherche et à fournir une analyse documentaire de la méthodologie utilisée dans les travaux connexes. Dans la section 4.2, nous positionnons la question dans le monde statistique comme un problème de regroupement et nous donnons un aperçu des principales catégories de modèles de regroupement. Nous appliquons ensuite des modèles représentatifs de chaque catégorie à trois séquençages de trois répliques techniques du génome NA12878 puis comparons leurs performances dans les sections 4.3 à 4.5. Précisément, nous avons exploré le modèle de consensus, le modèle de classes latentes, le modèle de mélange et la forêt aléatoire et étudié leurs capacités à produire résultat de meilleure qualité. Nous avons aussi comparé leurs principaux indicateurs de performance : précision, rappel et score F1.

Dans le Chapitre 4, nous montrons que les modèles de partitionnement non supervisés comparés sont capables d'améliorer les performances de séquençage en termes de précision et de score F1, ce qui est comparable à ce qui est rapporté au sujet des modèles supervisés. Parmi les modèles comparés ici, le modèle de mélange gaussien et Kamila ont apporté des améliorations qui ont rendu la précision supérieure à 99 % et le score F1 proche de 99 %. Ces

modèles peuvent être donc recommandés pour reconstruire de nouveaux callsets performants à partir de réplicats NGS. Ceci est particulièrement intéressant pour le diagnostic ou la médecine de précision lorsque les résultats du séquençage de l'ADN proviennent soit de réplicats biologiques (plus d'un échantillon), soit de réplicats technologiques (plus d'une plateforme de séquençage ou d'un pipeline d'analyse).

Acknowledgements

First, I would like to thank my supervisor, Pr. Pascal Roy, for his enthusiasm, support, and mentorship; for giving me the confidence to undertake this thesis work; and for many pleasant conversations about science and culture. I am also particularly grateful to Pr. Gaëtan Lesca for fruitful exchanges and insights in genetics, to Dr. Claire Bardel for her constant help in bioinformatics, and to Dr. Jean Iwaz for his counselling in linguistics and science editing.

I also wish to thank the members of the jury for accepting my invitation to examine this thesis work, especially the reviewers, Dr. Hervé Perdry and Dr. Sophie Tezenas du Montcel, for the time and effort put into reviewing the manuscript and expressing valuable suggestions. I wish also to thank the members of the Thesis Committee, Pr.Jacques Benichou, Pr. Damien Sanlaville, Dr. Anamaria Necsulea, and Dr. Nicolas Parisot for their constructive discussions and helpful suggestions during the annual meetings.

I would like to thank all the colleagues in Service de Biostatistique of Hospices Civils de Lyon for their warm welcome. To those who taught the M2 B3S courses, thank you for introducing me to the world of biostatistics. To those in site Lacassagne, in particular Catherine, thank you for your encouragement during my adventure in French pastry. To the office neighbours, thank you for your kindness and patience with me. To fellow PhD students Alexandre and Corentin, thank you for the exchanges and advice.

This thesis work was made possible with a Scholarship granted by the Chinese Scholarship Council (CSC). I would like to thank Shanghai Jiao Tong University School of Medicine, as well, for the opportunity it gave me to participate in its exchange program. The day I took my first French lessons, in 2014, it was beyond my imagination that I would pursue a Master's then a PhD degree in France.

Finally, to my parents, thank you for your encouragement in a difficult COVID time and for the freedom and unconditional support you gave me to pursue whatever interests me in life. All these years of study would never have been possible without you.

Table of contents

1. Overview of Next-Generation Sequencing

1.1 Sanger sequencing and next-generation sequencing

DNA (deoxyribonucleic acid) is a double helix 'ladder' formed by base-pair 'steps'. There are four different bases (or nucleotide): adenine (A), guanine (G), cytosine (C), and thymine (T). These bases pair up this way: A with T and C with G. Genetic information is stored by the order of these bases, highlighting the importance of determining the exact sequence of bases along the DNA chain.

DNA sequencing consists in determining the order of these nucleotides or bases (A, T, C or G) in a molecule of DNA. An important property of DNA is that it can replicate. In a conceptually simplified form, DNA replication requires three types of molecules: a template strand, free bases, and a polymerase enzyme that links the free bases together one-at-a-time into a new strand that is complementary to the template.

The first-generation DNA sequencing method was Sanger sequencing (developed by Fredrick Sanger and colleagues in 1977); it was initially known as the chain-termination method. The 'Sanger' method relies on four separate polymerization reactions performed using tritium-radio-labelled primers, where each reaction is supplied with small amounts of one chain-terminating 2,3-dideoxynucleoside triphosphate (ddNTP) to produce fragments of different lengths. When the DNA polymerase incorporates a ddNTP at the 3'-end of the growing DNA strand, it lacks a 3'-hydroxyl group and chain elongation is terminated. The sequence is then deduced by comparing the sizes of the fragments.

Automated Sanger sequencing technologies have been implemented since the early nineties. In high-throughput production pipelines, the DNA to be sequenced is prepared and a PCR amplification is carried out with primers that flank the target. The output is an amplified template because many PCR amplicons are present within a single reaction volume. The sequencing biochemistry takes place in a 'cycle sequencing' reaction in which cycles of template denaturation, primer annealing, and primer extension are performed. The primer is complementary to a known sequence immediately flanking the region of interest. Each round of primer extension is terminated by the incorporation of fluorescently-labelled

dideoxynucleotides (ddNTPs). In the resulting mixture of end-labeled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. Sequence is then determined by high-resolution electrophoretic separation of the single-stranded, end-labelled extension products in a capillary-based polymer gel. An ensemble of DNA molecules—all originating from the same position on the template but having different size due to termination at different positions—are arranged in an electric field which separates them by size (because DNA is negatively charged). As the molecules migrate in the presence of the electric field, they flow past a detector that registers the fluorescence intensity and colour, yielding a series of peaks. A software translates these readouts into DNA sequence while also generating error probabilities for each base-call. After three decades of gradual improvement, the Sanger biochemistry can be applied to achieve read-lengths of up to ~1,000 bps (base pairs) and per-base 'raw' accuracies are as high as 99.999% (Shendure and Ji, 2008). Sanger sequencing led to a number of scientific breakthroughs, including the realization of the Human Genome Project in 2001.

A number of sequencing technologies emerged rapidly after 2005; they are commonly referred to as 'next-generation sequencing technologies'. Next-generation sequencing is characterized by large-scale massively parallel sequencing permitting the analysis of genome hundreds of times faster and over a thousand times cheaper than traditional Sanger sequencing (Metzker, 2010). Rather than exploit size separation to arrange the fluorescent molecules, NGS uses positional separation: millions of different template DNA strands bind to discrete positions on a glass slide and remain fixed at the same position throughout the sequencing reaction. Each template is then extended by a single modified base per cycle and a microscope captures an image that resolves both the position of each template on the glass as well as its fluorescent colour and intensity.

The two technologies share a common origin: both repurpose the DNA replication machinery that copies DNA during every cell division. The key innovation that transforms DNA replication into the DNA-sequencing strategy at the core of both Sanger and NGS is the use of unextendable, fluorescently-labelled modified bases. In both sequencing techniques, when polymerase incorporates a modified base into the copied strand, the extension of the new strand stops, and, critically, this newly-terminated strand is uniquely colored to reflect its

most recently added base (Muzzey et al., 2015). The critical difference between Sanger sequencing and NGS is the sequencing volume. While Sanger method sequences only a single DNA fragment at a time, the massively parallel NGS sequences simultaneously millions of fragments per run. The following figure in the review by Shendure and Ji (Shendure and Ji, 2008) illustrates well the similarities and differences between the two technologies.

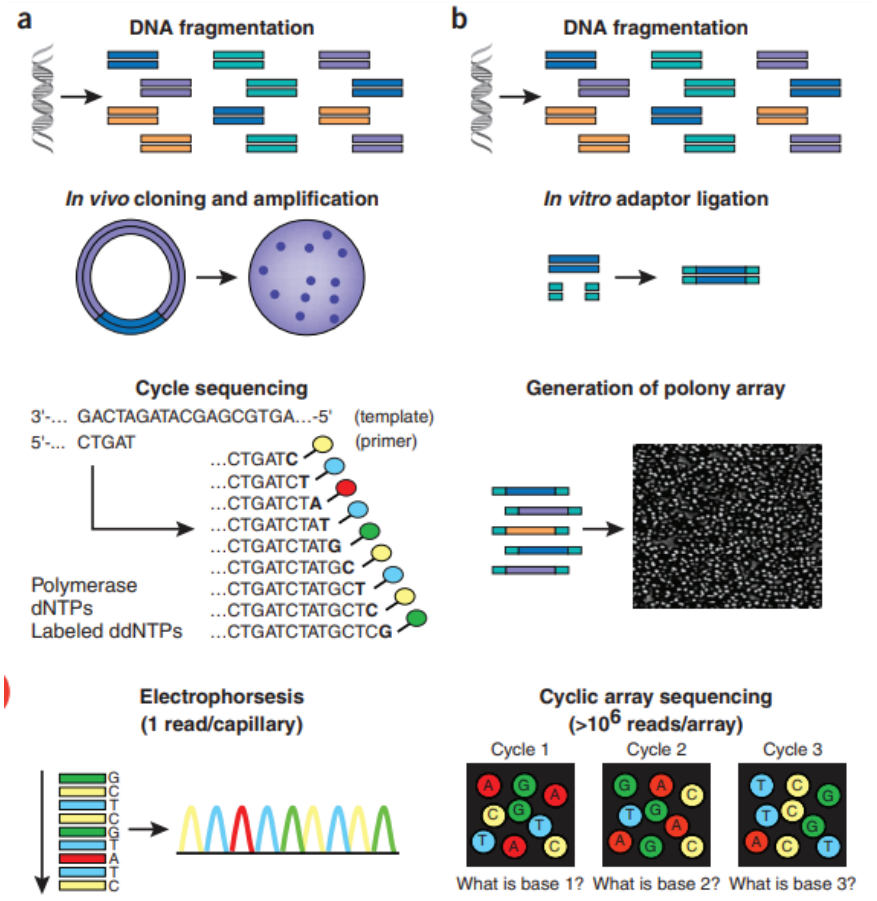

Figure 1 Work flow of conventional versus second-generation sequencing. (**a**) With high-throughput shotgun Sanger sequencing, genomic DNA is fragmented, then cloned to a plasmid vector and used to transform *E. coli*. For each sequencing reaction, a single bacterial colony is picked and plasmid DNA isolated. Each cycle sequencing reaction takes place within a microliter-scale volume, generating a ladder of ddNTP-terminated, dye-labeled products, which are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument. As fluorescently labeled fragments of discrete sizes pass a detector, the four-channel emission spectrum is used to generate a sequencing trace. (**b**) In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR colonies or 'polonies'[15]. Each polony consists of many copies of a single shotgun library fragment. As all polonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Similarly, imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature.

(Shendure and Ji 2008)

Next-generation sequencing technologies offer several advantages over Sanger sequencing, such as being faster and more sensitive due to the large number of reads generated. However, it also brings challenges. The short read length creates the need for sophisticated algorithms to determine the positions of each read (Metzker, 2010). The large amount of data with shorter read lengths, the higher per-base error rates, and the non-uniform coverage, together with platform-specific read error profiles and artefacts impose statistical and computational challenges for a reliable detection of variants from NGS data (Pfeifer, 2017).

1.2 The Sequencing Process - from sample to reads

There are several sequencing platforms (Illumina, Roche, Ion Torrent, etc.), among which Illumina platforms are the most widely used. According to their underlying biochemistries, platforms can be broadly divided into sequencing by ligation and sequencing by synthesis. The latter further divides into cyclic-reversible termination and single-nucleotide addition (Goodwin et al., 2016).

Although diverse in their sequencing biochemistries, the workflows of sequencing platforms are conceptually similar. A sequencing process includes a library preparation step and a sequencing step. Herein, we use an Illumina sequencing platform as example to illustrate the sequencing process.

1.2.1 Library preparation

All NGS approaches rely on a 'library' preparation using native or amplified DNA. The first step of template[1] generation is the fragmentation of the sample DNA into 200 to 500 bp short fragments followed by ligation to platform-specific 3' and 5' adapters at the end of each fragment. The adapters are recognized by the sequencing platform and used to distribute

[1] Template is a DNA fragment to be sequenced (Goodwin et al., 2016).

spatially the fragments by immobilizing them onto a solid surface. The fragments are then amplified in vitro by PCR (Ilumina platforms use a bridge PCR technique). This step results in producing clusters consisting of identical copies of a DNA sequence. The resulting sequencing library is loaded on a flow cell[2] and sequenced in Massive Parallel Sequencing reactions (Goodwin et al., 2016).

1.2.2 Sequencing

The sequencing technique used by Illumina is based on an optical readout of incorporated fluorescent nucleotides coupled to a reversible terminator by a DNA polymerase. During each sequencing cycle, a single fluorescently-labelled reversible terminator-bound dNTP is incorporated into each nucleic acid chain of the clustered fragments and the resulting fluorescence image of the flow cell is recorded. This image contains information on the type of nucleotide at a cycle for all spatially separated clusters in parallel. After imaging, the fluorophore attached to the freshly incorporated nucleotide is cleaved allowing a new cycle of synthesis and imaging to take place. Thus, after multiple cycles, the continuous sequence from each cluster can be obtained by translating each image taken at each cycle into a sequence of nucleotides.

1.2.3 Output and quality control

The output information of the sequencing step is stored in a FASTQ format file containing sequences of base calls along with the quality scores of each base. Sequencing quality scores are statistical measures of the probabilities that a base is incorrectly called, each base in a read is assigned a quality score by a phred-scaled algorithm. The quality score of a given base, Q, is defined by following equation:

$$Q = -10 * \log_{10} e$$

where *e* is the estimated probability of the base call being wrong. For example, a quality score of 20 (i.e, Q20), represents an estimated error rate of 1% (meaning every 100 bp sequencing read may contain an error) and a corresponding call accuracy of 99%.

[2] Flow cell is a surface with multiple lanes to adsorb and immobilize DNA fragments through attached adapters.

Quality scores are calculated for each base call in a two-step process: i) For each base call, a number of quality predictor values are computed. Quality predictor values are observable properties of clusters from which base calls are extracted, such as intensity profiles and signal-to-noise ratios that measure various aspects of base call reliability. These predictors have been empirically determined to correlate with the quality of the base call (Ewing and Green, 1998). ii) To estimate a new quality score, the quality predictor values are computed for a new base call and compared to values in the pre-calibrated quality table. The quality table, also known as Q-table, lists combinations of quality predictor values and relates them to corresponding quality scores, this relationship is determined by a calibration process using empirical data (Illumina, 2014). The percentage of base calls with a quality score of at least Q30 is often used as an indicator to assess the overall sequencing run quality on the Illumina platforms.

In addition to the above-presented quality scores, Illumina platforms also practice quality score binning in most cases; that is, the original quality scores may be compressed into fewer quality bins. For example, the original quality scores 20 to 24 may form one bin, and will be assigned a new value of 22. According to Illumina, the choice of bins is empirically optimized to minimize the loss of quality score resolution across the data, while minimizing the storage footprint. Moreover, the quality table that produces quality scores is often updated when significant characteristics of the sequencing platform change, such as new hardware, software, or chemistry versions. For example, in NovaSeq 6000 System, Quality scores are calculated through a process that is more streamlined than previous Illumina systems. Only three quality scores are possible and these quality scores represent the average error rate of a group. The three groups in the quality table correspond to marginal (< Q15), medium (~ Q20), and high-quality (> Q30) base calls and are assigned specific scores of 12, 23, and 37, respectively. Additionally, a null score of 2 is assigned to any no-calls. The simplification aimed to have more efficient data storage, which translates into lower storage costs and lower bandwidth requirements for sequencing data (Illumina, 2017).

Here, it is worth noting that different NGS technologies or platforms differ greatly in their specific characteristics due to different biochemistries, despite the commonalities in protocols. In fact, each platform is associated with unique biases introduced during library preparation

and sequencing; this results in strong differences between platforms regarding the average per-base error rates and the underlying reasons for the error (Pfeifer, 2017). These various error profiles motivate different error correction strategies in the bioinformatics analyses afterwards.

1.3 The Bioinformatics process – from reads to variant calls

NGS-based bioinformatics analysis can be broadly categorized into primary, secondary, and tertiary analyses. In brief, a primary analysis consists in processing raw sequencing instrument signals into nucleotide base and short-rad data. A secondary analysis involves mapping the short sequences of nucleotides (reads), to a reference sequence and determining variation from that reference. A tertiary analysis provides interpretation to the information generated during an NGS experiment by associating the sample-specific genomic profile with descriptive annotations (Oliver et al., 2015).

In this work, we focused mainly on the secondary analysis, which includes read alignments and variant calling as two main steps. Read alignment is the process of aligning reads against a human reference genome in order to determine the position of each read. Afterwards, the variant calling step aims to compare the aligned read to the reference genome to identify potential differences; i.e., variants. Various open-source or commercial bioinformatics tools are available for each of the two steps. In this section, we use Burrows-Wheeler Aligners (BWA) and GATK variant caller as main examples to describe the algorithms and the underlying statistical models of these bioinformatics tools. We particularly focus on explaining the conceptual and statistical principles as well as their evolution over time but not the implementation algorithms or computer programs.

1.3.1 Alignment

The first step of the secondary analysis is the alignment of reads to a reference genome. Output reads from sequencing platforms do not contain location information; therefore, the goal of alignment is to map individual reads to the position in the reference genome from which they most likely originated. As NGS technologies can generate hundreds of millions of

'short' reads per experiment, efficiency (= speed), scalability (= storage space), and accuracy are all required for an alignment algorithm (Reinert et al., 2015).

Read alignment is essentially a string match problem of large scale in which two strings[3] are compared and scored on the basis of dissimilarity (Robinson et al., 2021). The primary metric to measure dissimilarity between two sequences is called 'edit distance'. For example, Levenshtein distance is the minimal number of edit operations required to change one sequence into another (such edit could be deletion, insertion, or mismatch). The edit distance can then be used to calculate a similarity score with a predefined scoring system; i.e., different weights for matches, mismatches, insertion, or deletion. An algorithm (e.g., the Smith-Waterman algorithm (Smith and Waterman, 1981) is then applied to find the optimal local alignment(s) by either minimizing the distance or maximizing the similarity score. However, the expense of speed and storage challenge of this kind of direct exhaustive search makes it impossible to be directly applied to map sequences to large reference genomes. Over the past decade, various read alignment tools (i.e., aligners) have been developed employing different indexing and compression methods to optimize the speed and the memory footprint. Early generation of algorithms were mostly based on hash tables[4] and indexed either the query reads or the reference genome; later algorithms often used suffix-prefix tries or Burrow-Wheeler transform (BWT). The advantage of the latter is that multiple identical substrings in the reference genome are stored in a single path. Not having to align sequences that are identical makes the search process more efficient and less memory-intensive (Li and Homer, 2010).

Another important aspect in the alignment problem is that it requires inexact matching to be able to cope with sequencing errors as well as true differences between the sequenced genome and the reference genome (Nielsen et al., 2011). An inexact match problem can be regarded as finding string matches with no more than k differences including insertions, deletions, and mismatches. To solve this problem, most aligners use a seed-and-extend approach (Baeza-Yates and Perleberg, 1996) to find the inexact matches. The idea is that, based on pigeonhole lemma, an alignment of a sequence of length m with at most k differences must contain an

[3] A string is an ordered sequence of symbols that are selected from an alphabet.
[4] A hash table is a data structure that stores information about which reads or where in the reference genome a particular substring or subsequence occurs. (Nielsen et al., 2011)

exact match at least $s = m/(k+1)$ bp long because when a read is cut into $k + 1$ pieces, at least one piece would not contain a difference. Therefore, the algorithm searches first for an exact match seed with a predefined length then extends the seeds until the differences exceed a certain threshold. In the earlier years, alignment algorithms often required end-to-end alignment (i.e., every read base had to be aligned to the reference) and were developed for short reads that were about 36 bps in length (Li and Homer, 2010). With improved chemistry technologies, NGS reads became 100 bps or longer, the aligners began to allow long gaps and report multiple non-overlapping local hits potentially caused by structural variations or misassemblies in the reference genome (Li, 2013). To allow for gapped match, the algorithm usually assigns different weights for opening a gap, extending a gap, in addition to the weights for mismatches.

The most used aligners include BWA, Bowtie2, and BWA-MEM. For example, BWA is an aligner based on Burrows-Wheeler Transform; it was developed to align efficiently short sequencing reads against a large reference genome allowing mismatches and gaps. Given a read of length m, BWA tolerates a hit with at most k differences (mismatches or gaps) and k is chosen to be $< 4\%$ of m. However, long reads with 2% uniform base error rate may contain more differences. Later, the BWA-MEM algorithm (Li, 2013) was developed to align 100 bps or longer reads, allowing for mismatches and long gaps with improved speed. It initially seeds an alignment with supermaximal exact matches (SMEMs) using an algorithm which essentially finds, at each query position, the longest exact match covering the position. This algorithm is reported to be more robust to sequencing errors than BWA and applicable to a wide range of sequence lengths from 70 bps to a few million bases. Simulation suggests that BWA-MEM may work well given 2% error for an 100 bp alignment, 3% error for 200 bps, 5% for 500 bps, and 10% for 1000 bps or longer alignment (Li, 2013). Reinert et al. provide a comprehensive overview of the aligners' algorithms (Reinert et al., 2015), the goal of these computational algorithms is to optimize speed and storage, while still have a high performance (especially, sensitivity). The differences between different aligners lie essentially in whether or how they allow for gapped match during the seeding or extension step, which algorithm is used for seed extension, which indexing or compression method is used, etc.

The alignment data are usually stored in the sequence alignment/map format (SAM) or its binary composed version (BAM format) containing information about the location, orientation, and quality of each read alignment.

Mapping quality measures the reliability of the alignment, which may be interpreted as the likelihood of a read to be mapped to the correct position. Like the base quality score, the mapping quality score (MAPQ) is constructed as the phred-scaled probability that a read alignment may be wrong. For example, MAPQ = 30 implies there is a 1 in 1000 probability that the read is incorrectly mapped. The calculation of the mapping quality score was given in the following simplified form (Li et al., 2008):

"Suppose we have a reference sequence *x* and a read sequence *z*. On the assumption that sequencing errors are independent at different sites of the read, the probability $p(z|x,u)$ of z coming from the position *u* equals the product of the error probabilities of the mismatched bases at the aligned position. For example, if read *z* mapped to position *u* has two mismatches: one with phred base quality 20 and the other with 10, then $p(z|x,u) = 10^{-(20+10)/10} = 0.001$.

To calculate the posterior probability $p_s(u|x,z)$, we assume a uniform prior distribution $p(u|x)$ and, applying the Bayesian formula gives:

$$p_s(u|x,z) = \frac{p(z|x,u)}{\sum_{v=1}^{L-l+1} p(z|x,v)}$$

where $L = |x|$ is the length of x and $l = |z|$. Scaling p_s in the phred way, we get the mapping quality of the alignment:

$Q_s(u|x,z) = -10\ log_{10}[1 - p_s(u|x,z)]$ "

Solving this equation requires summing over all positions on the reference. It is impractical to calculate the sum given a human-sized genome. In practice, the Q_s were approximated by empirical formulas that differ between aligners.

One particular challenge in this step is the alignment of a short read from a repetitive or low-complexity genomic region that is longer than the read itself (Reinert et al., 2015). In this case, the reads often map equally well to multiple locations in the genome. Another challenge

is the alignment in regions with a higher level of diversity in the reference genome vs. the sequenced genome (e.g., the major histocompatibility complex: “a linked set of genetic loci encoding many of the proteins involved in antigen presentation to T cells”.

1.3.2 Post-alignment quality control and data pre-processing

Once reads have been aligned to the genome, several refinement steps are often performed. These steps include routinely i) flagging or filtering of duplicate reads likely to be PCR artefacts; ii) realignment, which leverages a collective view of reads around putative insertion or deletion (indels) sites to minimize erroneous alignment of reads; and, iii) base quality score recalibration, which aims to partially anticipate and correct certain platform-specific error profiles (DePristo et al., 2011). In this subsection, we describe the workflow recommended by the GATK best practices. The recommendations were subjected to regular updates because of the constant improvement of the bioinformatics algorithms and new research results; nevertheless, the general structure of the workflow remains similar.

1.3.2.1 Marking duplicates

Duplicate reads are reads that derive from the same physical DNA fragment in the sequencing library (Van der Auwera and O’Connor, 2020). Sequence duplications could be introduced during the PCR amplification step or during the sequencing step due to optical confusions (when a single cluster on the flowcell is called as two different reads). Duplications cause the reads to be a non-random sampling of the source genome and contain overrepresentation of certain sequences; they violate therefore the statistical assumptions of variant calling. They manifest as high coverage read support, often influence the coverage distribution and thus give rise to false positive variant calls. This is particularly tricky when a single molecule experiences a PCR error early in amplification because this error may be propagated and sampled many times during sequencing.

Tools do exist that detect and mark reads that are probable duplicates of one another. In the MarkDuplicates program in the Picard suite of tools which is implemented in the GATK framework, duplicate reads are identified as sets of read pairs that share the same alignment

start and end positions and have the exact same first five bases (for computational efficiency). The base quality scores of each read are summed up ignoring bases with quality scores below Q15 then the read with the highest sum of quality scores is retained.

Lastly, it is worth noting that this duplicate removal (de-duplication) strategy is not perfect. The implicit assumption is that it is unlikely (or sufficiently improbable) to sample the same exact molecule more than once from the source genome given that the sampling is truly random (Li, 2010). Thus, sequencings with very deep coverage (such as target enrichment sequencing) should not perform de-duplication. However, it is clear that this does occur even for whole genome sequencings, at various rates depending on the sequencing depth and the target regions. For example, for a 30× whole genome sequencing, true duplicate rates resulting from random sampling was estimated at 4.4%. The necessity of this de-duplication step has also been questioned using performance comparison of workflows with and without this step (Ebbert et al., 2016).

1.3.2.2 Local Realignment

Because alignment algorithms map reads individually to the reference genome, reads spanning insertions or deletions are often misaligned because most aligners have a tendency to introduce SNPs rather than structural variants in the mappings. Thus, at positions of unidentified indels, alignment artefacts result in false positive variant detections. To address this problem, some tools including GATK perform a realignment step to realign reads in suspicious regions to minimize the number of mismatching bases across all reads.

In the early versions of GATK tools (v. 1 and 2), the local realignment algorithm begins by identifying regions for realignment where i) at least one read contains an indel, ii) a cluster of mismatching bases exists, or iii) an already known indel segregates at the site (e.g., from the database dbSNP[5]). At each region, alternative haplotypes are "constructed from the reference sequence by incorporating know indels at the site, indels in reads spanning the site, or from Smith-Waterman alignment of all reads that do not perfectly match the reference sequence"

[5] The Single Nucleotide Polymorphism Database (dbSNP) is a free public archive for genetic variation within and across different species developed and hosted by the National Center for Biotechnology Information (NCBI).

(DePristo et al., 2011). For each resulting haplotype H_i, reads are aligned without gaps to H_i and the likelihood $L(H_i)$ calculated as the probability of observing all reads (see detailed formulas in DePristo et al., 2011). The haplotype that maximizes $L(H_i)$ is selected as the best alternative haplotype. Next, all reads are realigned against the best alternative haplotype H_1 and the reference H_0, each read R_j is assigned either to H_1 or H_0 whichever maximizes the probability of observing the read $L(R_j|H)$. If the log odds ratio of the two-haplotype model is better than the single reference haplotype by at least five log units, then the reads are realigned.

This realignment step has later evolved and been implemented into the haplotype caller (GATK version 3+), in which the steps are called "Identify Active Regions" and "Assemble plausible haplotypes" (Poplin et al., 2017). The active regions are first defined as regions where the aligned reads contain evidence of potential variants. Reads from these regions are reassembled into candidate haplotypes using a graph-based method. A pair Hidden Markov Model (pair-HMM) model (Durbin et al., 1998) is constructed to calculate a matrix of likelihoods for each read R_j to be sequenced from each haplotype H_i. In this pair-HMM model, the state transition probabilities (from a match "state" to an insertion or deletion "state") derived from the base qualities of read bases. Here, all reassembled haplotypes (as opposed to one retained haplotype in the earlier version) will be used to discover potential variants and derive an output file as an intermediate step of variant calling.

1.3.2.3 Base quality score recalibration

Most variant calling algorithms incorporate the phred-scaled base quality scores into their probabilistic framework; however, raw base quality scores are often systematically biased and convey inaccurately the true base-calling error rates (Nielson 2011). Therefore, quality scores allocated by the sequencing platforms are often recalibrated to be effectively used in the variant calling step.

One of the most widely applied base recalibration techniques has been implemented in the GATK (DePristo et al., 2011). Other recalibrations algorithms are used in other callers such as SOAPsnp (Li et al., 2009) and ReQON (Cabanski et al., 2012). The recalibration algorithm of

GATK takes into account several covariates such as the machine cycle and the dinucleotide context. For all sites that are not known to vary within a population, the bases that align to those sites are grouped into different categories with respect to several features: the reported base quality score, the position of the base (i.e., the machine cycle) in the read and the dinucleotide context (i.e., the two bases before the base of interest). For each category, the algorithm estimates an empirical quality score by using mismatches rate with respect to the reference genome. Recalibrated quality scores are then estimated by adding to the raw quality scores the residual differences between empirical quality scores and the raw quality scores (DePristo et al., 2011).

As described above, this algorithm uses a set of supposedly non-polymorphic sites. As a result, quality score recalibration depends strongly on the quality of previous polymorphism data; this restricts its usage to organisms with a public variant database. (Nielsen et al., 2011).

1.3.3 Variant calling (SNV discovery and genotyping)

One of the main objectives of a NGS bioinformatics pipeline is to detect differences between the sequenced genome and the reference genome. Such genomic differences, also called 'variants', include single nucleotide variants (SNVs), small insertion and deletions (indels), and larger alternations like structural variants (SVs), and copy number variants (CNV). In this work, we focus mainly on the detection of SNVs.

Variant calling is usually a multistep procedure: first, positions or regions where samples differ from the reference sequence are identified (variant calling) and then individual alleles at all variant sites estimated (genotyping).

Early methods for calling genotypes were based on counts. The analyses involve first a filtering step in which only high-confidence bases were kept, a commonly used cut-off is Phred-scaled quality score of Q20. Genotype calling would then proceed by counting the number of reads that supporting each allele and deciding genotypes with fixed cut-offs. For example, the algorithm would call a heterozygous genotype when the proportion of non-reference alleles is between 20% and 80%, otherwise a homozygous genotype. This procedure

works fairly well when the sequencing depth is high (>20×). For moderate or low sequencing depths, genotype calling based on fixed cut-offs will typically lead to under-calling of heterozygous genotypes, and the use of a simple filtering based on the quality score will lead to a loss of information. Additionally, this calling method does not provide measures of uncertainty in the genotype inference. For this reason, probabilistic methods have been developed to utilize the quality score to provide posterior probabilities of each genotype.

Most recent variant callers are based on different statistical approaches (Bayesian, maximum likelihood, or deep learning methods). Among these variant callers, a majority use Bayesian methods (for a summary table of implemented methods, see Pfeifer, 2017)).

The GATK variant callers employs a Bayesian probabilistic framework. The simple Bayesian genotyper in the first version of GATK (McKenna et al., 2010) computes the posterior probability of each of the possible 10 diploid genotypes, given the pileup of sequence reads that cover the locus. This computation is based on the Bayesian formulation (Shoemaker et al., 1999):

$$p(G|D) = \frac{p(G)p(D|G)}{p(D)}$$

where D represents the data (the read base pileup at this reference base) and G represents the given genotype. $p(G|D)$ is the posterior probability of the genotype, $p(G)$ is the prior probability of this genotype. The value of $p(D)$ is constant over all genotypes and

$$p(D|G) = \prod_{b \in pileup} p(b|G)$$

where *b* represents each base covering the locus. The probability of each base given the genotype $p(b|G)$ is calculated using the quality score of the read base, which is a phred-scaled score reflecting the error probability of each base, as presented in section 1.2. Finally, the assigned genotype at each site is the genotype with the highest posterior probability (McKenna et al., 2010).

The prior probability here is that of a genotype without incorporation of information from sequencing data. This prior genotype probability may be chosen to be equal across all

genotypes or can be based on external information, for example, from the reference sequence, SNP databases, or available population data. For example, a prior could be chosen based on the database dbSNP. When a certain polymorphism is reported in dbSNP, the prior probabilities at these sites are set to be high for the reported genotype and low for all other genotypes; otherwise a prior of 0.001 is applied to the other sites without known variation. Another note is that when the sequencing and alignment error is not incorporated into $p(D)$, the algorithm makes a significant assumption that any read present at a given site is actually located there. However, in reality, a certain percentage of reads are misaligned. Therefore, for a given read base to be used in the genotype likelihood calculation, several filters were applied including a base quality of at least Q20 and a mapping quality of its read of at east 20 (DePristo et al., 2011).

In the second version of GATK, multi-sample SNP callings were also incorporated. The likelihood of three genotype categories (homozygous reference, heterozygous variant, and homozygous variant) of each sample at each site were first estimated (DePristo et al., 2011) instead of the 10 genotypes in the first version. In a second stage, the genotype likelihoods of all samples were combined to determine the most likely alternate allele frequency in the cohort. Genotypes of each individual at that site were then estimated and assigned simultaneously through a heuristic algorithm, conditional on the estimated allele frequency (supplementary materials in DePristo et al., 2011).

In the third version of GATK with Haplotype caller, the algorithm was modified to perform joint genotyping across large numbers of samples. The variant caller performs first the local reassembly to construct haplotypes and assign potential variants for each sample. 'Raw' genotype likelihoods of each candidate variant are calculated using the pair-HMM model and stored in an intermediate variant calling file for each sample. The genotype likelihood across all samples is then used to perform the joint variant calling, including allele frequency estimation and genotype assignment.

This type of joint variant calling has been presented as having multiple advantages through sharing the information across multiple samples (Van der Auwera and O'Connor, 2020). For example, it could have higher sensitivity to call variants at sites where one sample has poor coverage but other samples provide enough reads of high quality.

The output of a variant calling dataset is usually stored in a Variant Calling Format (VCF) file. It is a tab-separated columnar text format in which each variant is represented on one line, indexed by a genomic location. Apart from the called genotype, variant-level information that describes the quality of the evidence supporting the variant call is also provided (Van der Auwera and O'Connor, 2020). In addition to other quality-related information resulting from previous steps such as read depth (DP), mapping quality (MAPQ), the GATK variant caller also computes variant quality score (QUAL) to reflect the confidence in the existence of a variant across samples and Genotype Quality (GQ) score to reflect the confidence in the called genotype. More specifically, the QUAL score is a phred-scaled transformation of the approximate posterior probability of a homozygous reference genotype. The GQ score is the phred-scaled probability of an incorrect genotype call, calculated as the difference between the phred-scaled likelihoods of the most likely genotype and the second most likely genotype.

1.3.4 Filtering

Initial variant calls often contain many false positive variants caused by sequencing or alignment errors. As a result, different filter criteria are often applied to reduce error rates (improve precision) in the data set. Here, we classify filtering strategies into two main categories: hard filtering and soft filtering.

Hard filtering is based on the assumption that false positive calls often show unusual properties. GATK recommends a set of filters in the best practice protocol. The hard filters for SNVs include Quality by Depth (QD) < 2.0, RMSMapping Quality[6] (MQ) < 40.0, StrandOddsRatio[7] (SOR) > 3.0, FisherStrand[8](FS) > 60.0 etc.

Soft filtering usually involve statistical modelling based on a set of known high-quality variant calls as well as a set of presumed false calls. The model is then used to predict the probability of each new variant call is correct.

Variant calling score recalibration (VQSR) (DePristo et al., 2011) is the first soft filtering model implemented in the GATK workflow. This algorithm is based on a Gaussian mixture

[6] RMSMappingQuality is the root mean square mapping quality over all the reads at a given site
[7] StrandOddsRatio is an estimation of strand bias using a test similar to the symmetric offs ratio test
[8] FisherStrand is the phred-scaled probability that there is strand bias at a given site

model with several covariates including genotype quality, strand bias, and mapping quality. The model is trained on a set of high-confidence variants considered as true set, then applied on the whole variant set to estimate the probability of each variant call being "true". The threshold of the filter can be modified by users according to the desired sensitivity. However, this recalibration algorithm requires well-curated training resources of know variants and is not suitable for small-sample-size experiments or exome sequencing (Van der Auwera and O'Connor, 2020). Some studies also showed that after applying VQSR, some "unvalidated variants" remain in the callset (O'Rawe et al., 2013).

Later, another machine learning model using convolutional neural networks (CNN) was developed and implemented in the GATK workflow as the "CNNScoreVariants" tool (Friedman et al., 2020). This model is able to use more information, including reference genome context and read data, and may be applied for callset with only one sample. The model proved having a higher performance than VQSR (for SNVs, precision 99.9%, recall 99.6% and F1-score 99.7%). The authors concluded that models trained on heterogenous data from various samples, truth sets, and sequencing platforms were found to have better performance and better generalizability across different genomes.

1.4 Quality control of NGS

The quality control can be performed for each of the three steps: sequencing, alignment and variant calling (Guo et al., 2014).

In examining sequencing data, the most important parameters to check for quality are the base quality, the nucleotide distribution, the GC content distribution, and the duplication rate. Sequencing data generated on Illumina platforms tend to have a median base quality score between 35 and 40 in the Phred scale. The nucleotide distribution of the four nucleotides (A, T, C, and G) across cycles should remain relatively stable, except for minor fluctuations at the end of the read. The percentage of GC in the exome regions is expected to be 49 to 51%, while for whole-genome sequencing, the GC content is around 38 or 39%. An abnormal GC

content percentage (say, more than 10% deviation from the normal range) can indicate sample contamination. (Guo et al., 2014)

For alignment quality control, the most important parameter for whole-genome sequencing is the average or median depth and the percentage of the genome covered by the sequencing at that depth. Illumina promises whole-genome sequencing with an average depth of 30 across 98% of the genome.

For checking the overall SNV quality in the variant calling set, the transition/transversion (Ti/Tv) ratio has been often used as a quality control parameter. The Ti/Tv ratio is the number of transition SNVs divided by the number of transversion SNVs. In substitution mutations, transitions are defined as the interchange of nucleotides of similar shapes: two-ring purine nucleobases (A ↔ G) or one-ring pyrimidine nucleobases (C ↔ T). Transversions are defined as interchanges of two-ring purine nucleobases and one-ring pyrimidine nucleobases (A ↔ C, A ↔ T, G ↔ T, G ↔ C) (Guo et al., 2014). When substitutions occur randomly, the Ti/Tv ratio is around 0.5 because there are two possible transitions and four possible transversions (Wang et al., 2015). However, in reality, transversions are more drastic than transitions because the former involve changes in the ring structure. Thus, in a human genome, the true Ti/Tv ratio is expected to be around 2.0 across the whole genome (Bainbridge et al., 2011), though the ratio differs by genomic regions (for example around 3.0 in exons). In the case of variant calling errors, this ratio should be close to 0.5 due to the equal probabilities of each type of substitution.

The heterozygosity to non-reference homozygosity ratio (het/ nonref-hom) is another quality control parameter for a variant callset. For a given human genome position, if A represent the reference base and B represent the variant base, then there are three genotype categories: AA, AB, and BB. The het/nonref-hom ratio is the number of SNVs with AB genotype divided by the number of SNVs with BB genotype. When the assumption of Hardy-Weinberg equilibrium is applied over a large set of SNVs in one individual, this het/ nonref-hom ratio is expected to be 2.0. However, in real sequencing data, this ratio is found to differ between individuals from different ancestry groups (Wang et al., 2015). The median het/ nonref-hom ratios among Africans, Asians, Americans, and Europeans are around 2.0, 1.4, 1.7, and 1.6, respectively.

2. Performance Evaluation of NGS Data

For a "test" with workflows as complicated as the NGS, one cannot properly evaluate or attempt to improve the performance without knowing the error sources. In this chapter, we first discuss the sources of error in the NGS process, stemming either from the experimental steps or the bioinformatics analysis. Then we discuss the current most widely used methods to evaluate the performance of a given NGS data set, in situations with available "gold standard" set and without "gold-standard" set. We finish with a brief review of the attempts and researchers' findings regarding detection and reduction of errors in a variant calling output.

2.1 Error sources and reproducibility of NGS

2.1.1 Source of errors in NGS

During sample preparation, errors can arise from a combination of human errors in sample handling which result in sample degradation, sample contamination, or low quantities of input DNA. During the preparation of sequence libraries, errors can occur when PCR amplification incorporates incorrect bases during synthesis cycles. Primer-mediated sequence amplification biases, barcode,[9] or adapter errors lead to cross-contamination of samples. Furthermore, machine failures are among the other sources of error that originate during sequence library preparation (Robasky et al., 2014).

During sequencing, user errors combined with the incorporation of additional bases during single sequence cycles, DNA damage, overlapping signals, strand biases, sequence complexity, and machine failures can contribute to sequence error. For most platforms, including Illumina, the number of errors increase towards the end of the read because of i) reductions in signal intensity caused by decreased enzyme activity (Kircher et al., 2009); ii) increased noise due to desynchronization between different copies of DNA templates in the

[9] A known DNA sequence appended to the ends of DNA fragments prior to sequencing for the purpose of pooling samples together to reduce cost (Robasky et al., 2014).

same cluster caused by incomplete read extension or non-reversible termination (Kircher and Kelso, 2010). With Illumina platforms, substitution errors can arise when incorrect bases are introduced during clonal amplification of templates. These errors show a bias toward certain substitutions such as A ↔ C and G ↔ T (Minoche et al., 2011). Moreover, random dispersion of clusters onto a surface (flowcell) coupled with limited sensor resolution may result in overlapping signals, where signals from nearby clusters interfere with the readout (Laehnemann et al., 2016). Certain DNA sequence characteristics, such as long homopolymer[10] or extremely high GC-content regions, may also increase read errors (Nakamura et al., 2011).

During bioinformatics process, short-read misalignment often arise around insertions and deletions as well as paralogs and other repetitive sequences. The incomplete reference genome is another important source of error that results in misaligned reads and variant calling errors. Other sources of error can arise from software algorithms limitations, including variant calling models and filtering strategies.

These sequencing errors could introduce bias in downstream analyses. For example, in genetic association studies, in the presence of genotype uncertainty, standard method for obtaining p-values using allelic test are not valid because of potential over-calling of heterozygotes or homozygotes. If the error structure is the same in cases and controls, tests will not suffer from excess of false positives. Nonetheless, they may suffer from reduced power because even a low level of genotyping errors can lead to a strong decrease in power (Huang et al., 2009).

2.1.2 Reproducibility of NGS

As described above, NGS is a multi-step process. Similar to error sources, one may want to access the reproducibility of each step, such as the reproducibility of the sequencing platform results, the analytical pipelines, or the overall sequencing process.

[10] a homopolymer is a sequence of consecutive identical bases

The reproducibility of NGS has been improving over time. Early researches showed relatively low concordance and high variability between sequencing platforms or analytical pipelines (Cornish and Guda, 2015). Lam et al. (Lam et al., 2012) compared the results of two platforms on biological replicates and found 88.1% concordance among all variants detected by at least one platform. O'Rawe et al. (O'Rawe et al., 2013) reported 57.4% of SNV concordance between the overall variant sets called by five different variant-calling pipelines using the same raw exome sequencing data. More recent studies have found a higher concordance rate (Patch et al., 2018). In a study using 15 combinations of sequencing platforms and variant callers, 90.39% of the SNPs were jointly identified by all 15 combinations and 94.22% were detected by at least 10 combinations (Chen et al., 2019). More recently, Pan et al (Pan et al., 2022) found that 91% of the SNVs were highly reproducible across six different variant callers. Most of the SNVs that were not highly reproducible were located in regions difficult to map with short reads and in segmental duplications.

The sources of variability and the impacts of each step in the process have been comprehensively assessed. Pan et al. (Pan et al., 2022) concluded that bioinformatics pipelines have a larger impact on variant reproducibility than sequencing platform or library preparation. More than 60% of the variance in sequencing results was attributed to callers as evaluated using gradient boosted classification trees. Aligners and sequencing platforms were the second and third contributor, respectively. This finding agrees with previous research works (Hwang et al., 2019) that attributed more variability to variant callers than to aligners. DNA samples of different individual genomes were found to have a limited impact on reproducibility; thus, Pan et al. suggested that any of the analysed publicly available DNA samples could be used for assessing reproducibility. Other researches led to similar conclusions, arguing that the reproducibility of bioinformatics tools depend primarily on the genomic context rather than on sample differences (Popitsch et al., 2017).

2.2 Evaluation method of NGS performance

2.2.1 Reference standard and benchmarking

For the performance evaluation of any test, a reference standard or "gold standard" is needed. Reference standards can be defined as control materials with known characteristics (for example, a known genotype) against which test performance can be measured (Hardwick et al., 2017). Given that reference standards can provide known "truths", the difference between the expected values and the measured values can provide an empirical estimate of test performance. This is otherwise difficult in the multi-step NGS process with different types and amounts of uncertainty. A reference standard can provide a cumulative measure of uncertainty associated with the final output. The original human reference genome does not provide a biological material to use as a reference standard because it derived from an assembly of multiple individuals' genome. Instead, various individual human genomes have been established as reference standards to benchmark NGS test performance. Stable gDNA from these individuals can be fairly, easily, and inexpensively sourced from transformed cell lines (Hardwick et al., 2017).

NA12878, the genome of a healthy female donor with European ancestry, the daughter in a father-mother-child 'trio' has become the foremost human genome reference standard. In 2014, the Genome In a Bottle Consortium (GIAB)[11] used a range of NGS technologies to characterize the NA12878 genome and provide a set of high-confidence genotypes that can be used to benchmark germline variant-calling pipelines (Zook et al., 2014). To minimize bias from any specific DNA sequencing method, the dataset was sequenced separately by 14 different sequencing experiments and 5 different platforms. This human WGS dataset is essentially the first near-complete human genome to have been extensively sequenced and re-sequenced by multiple techniques, with the results weighted and analysed to eliminate as much variation and errors as possible. Despite these efforts, a substantial proportion of the genome remains refractory to sequencing analysis due to extreme GC contents, low complexity, or repetitive sequences. The established high-confidence region of the benchmark

[11] The Genome In a Bottle Consortium was initiated in 2011 by the National Institute of Standards and Technology "to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice".

set covers approximately 90% of the reference genomes GRCh37 and GRCh38. Many clinical laboratories routinely sequence the NA12878 gDNA as a quality control for their NGS workflow (Linderman et al., 2014) and the identified variants can be benchmarked against high-confidence genotypes to assess performance. Other efforts such as the Platinum Genomes Project (PG) (Eberle et al., 2017) and Syndip (Li et al., 2018) have also produced benchmark sets using publicly available cell lines for the PG.

The diversity of human genetic variation has also motivated the development of reference genomes from different ancestries. Accordingly, NIST expanded its set of supported genome reference to include representatives from different ethnic populations (Zook et al., 2016). Reference genome banks and reference standards for specific countries or ethnics have also been developed (Gudbjartsson et al., 2015; Seo et al., 2016; Zhang et al., 2021).

2.2.2 Performance evaluation in the presence of gold standard

In the presence of a "truth set" or reference set, the overall performance of NGS can be evaluated using a contingency table (or confusion matrix) by comparison to a reference set (or gold standard).

Table 2.1: Performance table against reference set

	Test result	
	Positive	Negative
Gold standard		
Positive	True positives (TP)	False negatives (FN)
Negative	False positives (FP)	True negatives (TN)

Using this contingency table, one may extract the following indicators:

Sensitivity (or recall) = TP/ (TP + FN)

Specificity = TN/ (TN + FP)

Precision (or PPV[12]) = TP/ (TP + FP)

Accuracy = (TP + TN)/ Total

F1-score = 2TP/ (2TP + FP + FN)

Performance metrics that describe different aspects of the NGS test performance can be calculated using the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Performance metrics usually include sensitivity (Se) or recall, specificity (Sp), precision or positive predictive value (PPV), accuracy, and F1-score. In other medical diagnostic tests, sensitivity and specificity are the two most used metrics; however, in the case of NGS–here whole genome sequencing as an example–the real variants are relatively rare across the genome (~0.1%), i.e., the positive and negative class are unbalanced. Moreover, the number of true negative sites are hard to define for genomic variations particularly concerning structural variants, because an infinite number of potential variants could exist resulting in an infinite number of true negatives (Krusche et al., 2019). Thus, precision (also known as positive predictive value) is often used instead of specificity to describe the ability of a test to identify correctly the absence of variants or the absence of false positives. In this case, a precision-recall curve is used instead of a ROC curve to illustrate the trade-off between recall and precision.

The definitions of these performance metrics are not trivial when comparing variant calls. Due to the complexity of the human genome and the challenge that genotype comparisons do not fall in a binary classification model, TP, FP, and FN could have various definitions. The Global Alliance for Genomics and Health (GA4GH) published in 2019 a set of best practices aiming to standardize benchmarking methods and the definitions of performance metrics. In this guideline, three definitions were proposed to define the “true positives”, from the most to the least stringent: i) “genotype match”: sites with matching alleles and genotypes; ii) “allele match”: sites with matching alleles is counted as TP, even when genotypes differ; and iii) “local match”: sites in the query set with a nearby true variant within a pre-defined distance are counted as TPs, even when alleles and genotypes differ. (Krusche et al., 2019)

[12] PPV = Positive Predictive Value

It should be noted here that benchmarking against only high-confidence genotypes is likely to overestimate performance given the difficulty to identify variants that are in the omitted 'difficult' regions. For example, only 74.6% of exonic bases in ClinVar and OMIM genes belonged to high-confidence regions (Goldfeder et al., 2016). Nevertheless, with the constant improvement of the newer versions of the reference set, this bias should be less important in recent studies. Also, many variant-calling algorithms identify preferentially variants at known polymorphic sites. Therefore, novel or rare variants in patient samples may not be identified with the same sensitivity as with known variants that are considered as reference samples such as the NA12878 genome.

2.2.3 Performance evaluation in the absence of gold standard

In real individual data, evaluating the performance of sequencing result can be complicated by the lack of gold standard. In these situations, evaluating sensitivity requires often external data as source of complete variant set (Meynert et al., 2013). To assess precision (i.e., the positive predictive value) or accuracy, an experimental validation via Sanger sequencing (few sites) or 'target enrichment sequencing' (numerous sites) is often needed.

Sanger sequencing has been historically used as the reference technique when evaluating NGS data. However, validations with Sanger sequencing have limitations; e.g., the sensitivity of Sanger technique has been questioned in numerous researches (Beck et al., 2016). Other methods that use external datasets would also yield biased Se and Sp estimates, especially in easy-to-sequence genome regions (Li et al., 2018). Also, these reference datasets may contain sequencing and calling errors (Atkinson et al., 2022).

Evaluating performance may also be done by comparisons between results from different bioinformatics platforms or bioinformatics pipelines. Discordant variants imply errors, whereas concordant variants suggest a low error probability. For example, Reumers et al. (2012) defined each discordance between sequences of monozygotic twins as "error", each concordance as "truth", and carried out experiments to confirm assumptions of error or truth. Selected shared SNVs were all confirmed by genotyping, indicating a very low error rate among concordant SNVs. Later, Ratan et al. (2013) validated 92.7% of the concordant SNVs

and 60% of platform-specific variants providing more reliable concordant results; then, on three platforms, they showed a 64.7% concordance rate across platforms. This led to recommend evaluation by comparisons between samples from a parent-offspring trio or same individual (Li, 2014; Robasky et al., 2014).

2.2.4 The use of technical and biological replicates

Different types of replicates, including technical and biological replicates, are often used in NGS to mitigate user error, stochastic variability, and other experimental errors. Technical replicates are defined as the repeat analysis of the exact same sample whereas biological replicates are defined as the preparation and analysis of multiple biological samples from the same host under the same conditions (Robasky et al., 2014). The objective of using replicates is either to mitigate errors and improve accuracy or to evaluate performance metrics using the concordance between replicates as a substitute for the gold standard.

Indeed, the variation in the results of a test consists in imprecision (random error) and bias (systematic error), the latter is defined as "the difference between the expectation of measurement results and the true value of the measures quantity" (Fraser, 2001; Monach, 2012). By combining results from replicates or repeated test, one can reduce the impact of random errors but not that of a systematic error. In NGS, different sequencing platforms or bioinformatic tools use different methodologies and provide complementary information; no technique outcompetes others under all circumstances. Therefore, combining data sets from multiple platforms or bioinformatics tools is expected to achieve better performance.

In studies with replicates, loci are often called "concordant" when variant calls in all replicates agree and "discordant" when variant calls in at least one of the replicates differ from the others. Concordant loci then represent true positive variants, whereas discordant loci implicate false positive variants. A decade ago, Wall et al. (2014) used biological replicates (blood and saliva samples) sequenced by two different platforms to estimate genotype error rates. The authors assumed that a genotype call is correct when a majority concordance criterion was met; that is, when at least three out of the four replicate samples had the same genotype call with GQ $\geq$ 40. Other researchers used replicate discordance to characterize error

and provide lists of low reproducible regions (Atkinson et al., 2022) where variant calls in clinical analysis should be treated with caution.

2.3 Factors associated with NGS performance

2.3.1 Individual factors of NGS performance

Factors associated with the error rate have been explored by numerous studies. Here, we give a brief summary of existing studies regarding several site-level factors, other factors that are used to evaluate the overall performance of a callset were discussed in section 1.4 as "quality control" factors. These site-level factors can be broadly divided into factors related to the genome context (e.g., the first two factors in the following list) and factors related to a specific sequencing run, although there is no clear boundary between these two categories.

1) GC content

GC content is the percentage of C and G in a certain genome region, this percentage is well known to be correlated with the depth of coverage. Both GC-rich and GC-poor regions tend to be less well covered by sequencing platforms (Benjamini and Speed, 2012; Ross et al., 2013). This concerns mostly regions with GC content higher than 60% or lower than 25% (Rieber et al., 2013).

2) Difficult regions

Genome regions complexity and heterogeneity could result in erroneous alignment in low-complexity repetitive DNA regions, which was identified as one of the main sources of errors (Li, 2014; Popitsch et al., 2017; Treangen and Salzberg, 2012) These low-complexity regions (LCRs) account for 2% of the human genome.

3) Read depth or Depth of Coverage

Read depth is one of the most studied indicators of error. Generally, a low depth of coverage is associated with errors and a high coverage implies higher confidence in variant calling,

especially for heterozygous variants; however, the relationship between read depth and error rate is not linear.

Empirical models examined the relationship between read depth and sensitivity or error rate estimates (Ajay et al., 2011; Cornish and Guda, 2015; Meynert et al., 2013; Ratan et al., 2013; Reumers et al., 2012). A high sequencing coverage could overcome the error rate in easily sequenced regions. However, systematic sequencing errors due to sequencing artefacts and misalignment cannot be overcome by high coverage. Indeed, a read depth that is too high (for example $> d + 3\sqrt{d}$, where d is the average read depth) (Li, 2014) is also an indicator of false positives, which are often caused by CNVs or sequences not present in the human reference genome.

4) Allele balance

Allele balance is the proportion of reads supporting an alternative base in a given position (alternative read count divided by total read count at the site). In a diploid genome like the human genome, the expected allele balance for heterozygous genotypes would be around 0.5, for homozygous reference near zero, and for homozygous variant near one. A large deviation of observed to expected allele balance is thus an indicator of less confident variant calling. Although this proportion has been taken into account when variant callers compute genotype likelihoods, it remains an important factor in post-variant calling quality assessment. It can be used as a hard filter for variant discovery; for example, to consider called variants with allele balance < 0.1 as false positives. It has also been used to discover systematic errors and establish genotype confidence scores. For example, by solely considering allele balance in population-level sequencing data, Muyas and colleagues were able to develop a filter reflecting recurrent bias in allele balance and identify genome sites that require caution (Muyas et al., 2019)

5) Strand bias

Strand bias occurs when the genotypes inferred from the information presented by the forward strand and the reverse strand are significantly different. It has been found to be correlated with variant calling errors (Guo et al., 2012) and are frequently employed as variant filters (DePristo et al., 2011).

6) Quality score (QUAL)

Quality score refers to the variant quality score computed by sequencing platforms and often recalibrated by bioinformatics pipelines. QUAL has been considered as an important predictor of variant calling error (Bauer et al., 2019).

7) Presence of nearby indels or multiple variant calls

The proximity to other SNVs of indels or multiple other SNV calls is a factor that appears to have a significant importance in multiple studies. (Hofmann et al., 2017; Ratan et al., 2013; Reumers et al., 2012; Shringarpure et al., 2016). A high SNP frequency in a short region is an indication of false positives that may be caused by small insertions or deletions. GATK's protocols suggest that the likelihood of a false positive is high when there are two SNPs within 10 bps.

8) Genotype Quality (GQ)

Genotype quality is a phred-scaled score of the estimated error rate estimated by the variant caller; thus, it is not surprising that GQ is used as an indicator of error (Kumaran et al., 2019). However, the true error rates were found to be far higher than those estimated by the GQ score (Wall et al., 2014), suggesting that the GQ alone could not predict errors accurately.

2.3.2 Models combining multiple factors to improve NGS performance

Many existing studies have modelled the relationships between several covariates and NGS performance: i) a generalized linear model with 23 parameters was used to separate true positive (TP) from false positive (FP) calls in Sandmann et al. (2017); ii) variant-free simulated reads explored the relationship between FP calls number and seven covariates in Ribeiro et al. (2015); iii) the effects of twelve factors on the error rate were combined to better filter errors (Reumers et al. 2012), the procedure identifies optimized combinations of cumulative filters based on optimal balance of estimated sensitivity and specificity; iv) Hwang et al. (2019) trained a two-component mixture model on reference sets to separate true variants from calling errors. The model included concordance rate across callsets and six

other factors as covariates. Random forest models were also developed to distinguish true variant calls from false positives (Lek et al., 2016; Shringarpure et al., 2016).

3. Contribution and limit of the concordance-discordance model in performance evaluation of NGS

In the absence of gold standard, researchers are often obliged to use the concordance between multiple sequencing results as a substitute criterion. The discordance results are then interpreted as errors. However, whether the discordance rate corresponds to the error rate remains unclear.

In this chapter, we aim to examine the appropriateness of concordance as a substitute criterion. We first analyse the theoretical relationships between the error rate and the discordance rate under conditional independence and dependence. We then illustrate it with simulations of various situations as well as data on three NA12878 genome replicates. Finally, differences between estimates of covariate effects associated with error and discordance are examined.

3.1 Modelling the error rate and the discordance rate

In this section, we apply the principles of performance evaluation in diagnostic tests to the domain of NGS. Here we use the general concept "test" to refer to a NGS process. The statistical unit or subject is a base-pair position (i.e., site) in the sequenced genome. The outcome is considered as binary variables (variant or non-variant).

3.1.1 Modelling the response of one test against gold standard

As mentioned in section 2.2.2, the basic measures of test performance are sensitivity (Se) and specificity (Sp). Their definitions are illustrated by a contingency table as showed in table 2.1, where the rows summarize the data according to the true status, and the columns summarize the test results.

We denote the true status of the base pair by the indicator variable V, where V = 1 if it is a variant and V = 0 if it is a non-variant (i.e. reference). We denote test result by the variable T,

where T = 1 if it is called variant in the query variant calling set and T = 0 if it is called non-variant. Thus the sensitivity $Se = P(T = 1|V = 1)$, which is the probability that the site is called as variant given that the true status of the site is variant. Specificity $Sp = P(T = 0|V = 0)$, which is the probability that the site is called as non-variant given that the true status of the site is non-variant.

An error is either a variant in the gold standard set called as non-variant by the query set (i.e., a false negative, FN), or a non-variant called as variant (i.e. a false positive, FP). This recalls the "local match" "for which any site in the query with a nearby truth variant is counted as a TP (true positive), even if alleles and genotypes differ" as in Krusche et al. (2019).

The error rate for variants comes to the FN rate; i.e., (1–Se) = nr FNs/ nr variants. For non-variants, the error rate comes to the FP rate; i.e., (1–Sp) = nr FPs/ nr non-variants.

3.1.2 Modelling the joint response of two test

When results from two NGS process are available, whether from technical replicates or biological replicates, they can be regarded as two tests for a population with unobserved true status. Here the population refers to all sites in the sequenced genome.

We aim to model the contingency table between the two tests in the overall results (table 3.1c). This observed table is in fact the sum of two unobserved contingency table, for the "variant" population (table 3.1a) and "non-variant" population (table 3.1b), respectively. The performance metrics are related to these two unobserved tables.

Table 3.1 : Conditional probabilities for results of two sequencing tests

3.1a: Variant Table

V=1	B = 1	B = 0
A = 1	$\eta_{1.11}$	$\eta_{1.10}$
A = 0	$\eta_{1.01}$	$\eta_{1.00}$

$\eta_{1,ab}$: P (A = a, B = b | V = 1)

3.1b: Non-variant Table

V=0	B = 1	B = 0
A = 1	$\eta_{0.11}$	$\eta_{0.10}$
A = 0	$\eta_{0.01}$	$\eta_{0.00}$

3.1c: Total Table

	B = 1	B = 0
A = 1	P_{11}	P_{10}
A = 0	P_{01}	P_{00}

P_{ab} = P (A = a, B = b)

$P_{ab} = \pi \times \eta_{1.ab} + (1 - \pi) \times \eta_{0.ab}$ with π = P (V = 1)

Suppose we have two tests (here, sequencing process) A and B, $P(\mathrm{A} = \mathrm{a})$ denote the probability of outcome a for test A, $P(\mathrm{B} = \mathrm{b})$ denote the probability of outcome b for test B. let $\eta_{v.ab}$ denote the conditional probability of sequencing result $\mathrm{A} = \mathrm{a}$ ($\mathrm{a} = 0$ or 1), $\mathrm{B} = \mathrm{b}$ ($\mathrm{b} = 0$ or 1) with $\mathrm{v} = 1$ for variants and $\mathrm{v} = 0$ for non-variants. For any variant, the probability of correct classification by A and B is $\eta_{1.11}$, that of misclassification by A and B $\eta_{1.00}$, and that of misclassification by either A or B $\eta_{1.10} + \eta_{1.01}$ (i.e., discordance between A and B). For any non-variant, those probabilities may be written $\eta_{0.00}$, $\eta_{0.11}$, and $\eta_{0.10} + \eta_{0.01}$. Therefore, for all bps in the genome, π being the prevalence of a variant $\pi = P(V = 1)$, the expected probability P_{ab} may be written:

$$P_{ab} = P(A = a, B = b) = \pi * \eta_{1.ab} + (1 - \pi) * \eta_{0.ab} \quad [1]$$

The concept of this model is similar to the latent class model described by Goodman (Goodman, 1974), which summarizes probabilities of classification in the latent classes as

well as conditional probabilities of outcomes for each observed variables within each latent class. Here, the latent class variable is the unobserved true status of each site, and observed variables are the results from test A and B.

Consequently, considering all bp results (variants and non-variants), the probability of concordance will be $P_{11} + P_{00}$ and that of discordance $P_{10} + P_{01}$. A discordant pair of results means one is an error and concordant pair means both results are correct or both are errors. According to the proportion of 'false concordance' where concordant pairs are both errors, a criterion using pairwise agreement may be appropriate or not. If this proportion is low, the concordant pairs will be less concerned by errors and the pairwise agreement model will become appropriate.

With same notations and under an assumption of conditional independence given the true status of bps (either variants or non-variants), the conditional probabilities for any variant are $\eta_{1.11} = Se_A * Se_B$, $\eta_{1.00} = (1 - Se_A) * (1 - Se_B)$, and $\eta_{1.10} + \eta_{1.01} = Se_A * (1 - Se_B) + (1 - Se_A) * Se_B$. The conditional probabilities for any non-variant, the equations are the same but with Sp_B and Sp_B instead of Se_A and Se_B.

Therefore, for all bps (Table 3.1c), the expected probabilities for P_{11}, P_{00}, P_{10}, and P_{01} can be calculated using equation [1]:

$$P_{11} = \pi * Se_A * Se_B + (1 - \pi) * (1 - Sp_A) * (1 - Sp_B)$$
$$P_{00} = \pi * (1 - Se_A) * (1 - Se_B) + (1 - \pi) * Sp_A * Sp_B$$
$$P_{10} = \pi * Se_A * (1 - Se_B) + (1 - \pi) * (1 - Sp_A) * Sp_B$$
$$P_{01} = \pi * (1 - Se_A) * Se_B + (1 - \pi) * Sp_A * (1 - Sp_B)$$

3.1.3 Modelling correlation between two tests

The above-mentioned assumption of conditional independence does not correspond to real situations in many ways. At a certain base-pair position, because of the common covariates that influence the NGS accuracy, such as genomic context, the error of two NGS tests are expected to be correlated. In situation of non-independence, the distribution of probabilities $\eta_{1.ab}$ and $\eta_{0.ab}$ can be formulated using additional parameters that represent the correlation between the results obtained for two sequencing processes A and B.

Many parameters have been proposed to measure the association of two binary variables. Such as odds ratio, Cohen's kappa. However, these measures could be misleading if one of the outcomes is very dominant, for instance the concordant negative outcome in our example of genome sequencing. As the vast majority of pairs will be concordant negative given the small error rate of NGS, these measures might be very high even if there is only a small proportion of concordant positive pairs among all sites in the variant calling output file (which contains sites called as variant by at least one test). Therefore, although measures like odds ratio have nice mathematical properties (such as the absence of range restrictions, regardless of the marginal probabilities), it is sometimes not adapted to the characteristic of interest, due to its symmetry treating negative-negative concordance of equal importance as positive-positive concordance (Faes et al., 2008). In this study, we employ the correlation coefficient and conditional probability to model the dependency between two tests.

3.1.3.1 Modelling conditional dependence with correlation coefficients

In this study, two parameters of correlation were assigned to variants and non-variants. For a given combination of sequencer and variant caller, the correlation value should be stable for all samples or individuals. Let Cov_1 and Cov_2 denote the covariances of the two sequencing tests for $V = 1$ and $V = 0$, respectively. For variants, the conditional probabilities for the combination of the two test results are the following:

$$\eta_{1.11} = Se_A * Se_B + Cov_1 \qquad \eta_{1.00} = (1 - Se_A) * (1 - Se_B) + Cov_1$$

$$\eta_{1.10} = Se_A * (1 - Se_B) - Cov_1 \qquad \eta_{1.01} = (1 - Se_A) * Se_B - Cov_1$$

For non-variants :

$$\eta_{0.11} = (1 - Sp_A) * (1 - Sp_B) + Cov_2 \quad \eta_{0.00} = Sp_A * Sp_B + Cov_2$$

$$\eta_{0.10} = (1 - Sp_A) * Sp_B - Cov_2 \qquad \eta_{0.01} = Sp_A * (1 - Sp_B) - Cov_2$$

The ranges of covariance being:

$$0 < Cov_1 < min[(1 - Se_A) * Se_B,\ Se_A * (1 - Se_B)]$$

$$0 < Cov_2 < min[(1 - Sp_A) * Sp_B,\ Sp_A * (1 - Sp_B)]$$

The overall probabilities for all bps can be calculated using equation [1].

$$P_{11} = \pi * (Se_A * Se_B + Cov_1) + (1 - \pi) * [(1 - Sp_A) * (1 - Sp_B) + Cov_2]$$

$$P_{00} = \pi * [(1 - Se_A) * (1 - Se_B) + Cov_1] + (1 - \pi) * (Sp_A * Sp_B + Cov_2)$$
$$P_{10} = \pi * [Se_A * (1 - Se_B) - Cov_1] + (1 - \pi) * [(1 - Sp_A) * Sp_B - Cov_2]$$
$$P_{01} = \pi * [(1 - Se_A) * Se_B - Cov_1] + (1 - \pi) * (Sp_A * (1 - Sp_B) - Cov_2)$$

The relationships between correlation, covariance, and Se for variants and non-variants are then:

$$Cor_1 = \frac{Covariance_1}{\sqrt{Se_A(1-Se_A)} \times \sqrt{Se_B(1-Se_B)}} \text{ and } Cor_2 = \frac{Covariance_2}{\sqrt{Sp_A(1-Sp_A)} \times \sqrt{Sp_B(1-Sp_B)}}$$

It is clear that the probability of discordance ($P_{10} + P_{01}$) is lower in case of dependence than in the case of conditional independence, while the probability of concordance is higher. The probabilities of false concordance (η1.00 or η0.11) are also higher in the former than in the latter case

3.1.3.2 Modelling conditional dependence with conditional probability

Another way of presenting these probabilities is to use conditional probabilities rather than covariance or correlation, as correlation coefficients might be difficult to interpret in real experiment contexts. Conditional probability $\mathrm{P}(\mathrm{A} = 0 \mid \mathrm{B} = 0, \mathrm{V} = 1)$ is the probability that test A makes a mistake given test B has made a mistake for a variant. Conditional probability $\mathrm{P}(\mathrm{A} = 1 \mid \mathrm{B} = 1, \mathrm{V} = 0)$ is the probability that test A makes a mistake given test B has made a mistake for a non-variant. For example, the four conditional probabilities of calling results for variants may be written:

$$\eta_{1.11} = Se_A - (1 - Se_B) * (1 - \mathrm{P}(\mathrm{A} = 0 \mid \mathrm{B} = 0, \mathrm{V} = 1))$$
$$\eta_{1.01} = (1 - Se_A) - (1 - Se_B) * \mathrm{P}(\mathrm{A} = 0 \mid \mathrm{B} = 0, \mathrm{V} = 1)$$
$$\eta_{1.10} = (1 - Se_B) * (1 - \mathrm{P}(\mathrm{A} = 0 \mid \mathrm{B} = 0, \mathrm{V} = 1))$$
$$\eta_{1.00} = (1 - Se_B) * \mathrm{P}(\mathrm{A} = 0 \mid \mathrm{B} = 0, \mathrm{V} = 1)$$

The conditional probabilities for non-variants may be written the same way.
In case of technical replicates, assuming $Se_A = Se_B$ and $Sp_A = Sp_B$, the relationships between the conditional probabilities can be written:

$$\mathrm{P}(\mathrm{A}=0 \mid \mathrm{B}=0, \mathrm{V}=1)=\ \mathrm{P}(\mathrm{B}=0 \mid \mathrm{A}=0, \mathrm{V}=1)$$

$$\mathrm{P}(\mathrm{A}=1 \mid \mathrm{B}=1, \mathrm{V}=0)=\ \mathrm{P}(\mathrm{B}=1 \mid \mathrm{A}=1, \mathrm{V}=0)$$

Relation between Cov_1 and P(A = 0 | B = 0, V = 1) is:

$$P(A=0 \mid \mathrm{B}=0, \mathrm{V}=1)=\frac{\mathrm{P}(A=0, \mathrm{B}=0, \mathrm{V}=1)}{\mathrm{P}(\mathrm{B}=0, \mathrm{V}=1)}$$
$$=\frac{\mathrm{P}(\mathrm{V}=1) \times \mathrm{P}(A=0, \mathrm{B}=0 \mid \mathrm{V}=1)}{\mathrm{P}(\mathrm{V}=1) \times \mathrm{P}(\mathrm{B}=0 \mid \mathrm{V}=1)}=\frac{\eta_{1.00}}{1-Se_B}$$
$$=\frac{(1-Se_A)(1-Se_B)+Cov_1}{1-Se_B}\ =\ (1-Se_A)+\frac{Cov_1}{1-Se_B}$$

Under the assumption $\mathrm{Se_A} = \mathrm{Se_B} = \mathrm{Se}$, the relationship between correlation, covariance, and P (A = 0 | B = 0, V = 1) becomes:

$$\mathrm{Cor}_1=\frac{\mathrm{Cov}_1}{\mathrm{Se}\ \times(1-\mathrm{Se})}$$

$$P(A=0 \mid \mathrm{B}=0, \mathrm{V}=1)=(1-\mathrm{Se})+\frac{\mathrm{Cov}_1}{1-\mathrm{Se}}=\ (1\ -\ \mathrm{Se})\ +\ \mathrm{Cor}_1\ \times\ \mathrm{Se}$$

$$\frac{1-\mathrm{P}(A=0|\mathrm{B}=0, \mathrm{V}=1)}{1-\ \mathrm{Cor}_1}=\mathrm{Se}$$

3.2 Illustration with common NGS performance indicators

3.2.1 Scenario settings

For illustration, estimated performance values from the literature were used for the above-mentioned probabilities. The Se of detecting variants is estimated at 90 to 99% (95.4% in Goldfeder et al., 2016, 98.66% in Krishnan et al., 2021, 99.3% in Lam et al., 2012) and Sp at 99.9% to 99.99% in different situations (Li, 2014; Reumers et al., 2012). The total number of bps in the whole genome was set at 3×10^9.

In the analyses under conditional independence, sensitivity was set at two values 90% and 99% and specificity at 99% and 99.9%. The expected prevalence of any variant was set as 0.1%; i.e., the expected number of variant bps was 3 × 106. The conditional probabilities in Table 1 and the probabilities in Table 2 were calculated using the above-mentioned values.

The positive and negative predictive values (PPVs and NPVs) are shown for each Se-Sp combination with the overall results.

In the analyses with conditional dependence, adding two parameters for dependence leads to seven parameters in the probability model: Se_A, Se_B, Sp_A, Sp_B, Cor_1, Cor_2, and π. Here, sensitivity was set at 90% and 99%, specificity at 99% and 99.9%, and the expected prevalence of any variants at 0.1% and 0.2%. Various levels of correlation between replicates were considered; precisely, 30%, 50%, and 90% for low, intermediate, and high correlation levels, respectively.

3.2.2 Results under conditional independence

Here, as first example, assuming a 99% Se of calling results, for a given variant, the probability of discordance ($P_{10} + P_{01}$) would be 1.98% and the probability of false concordance (P_{00}) 0.1998% (Table 3.2, first row). For a given non-variant, assuming a 99.9% Sp, the probability of discordance ($P_{10} + P_{01}$) would be 0.1998% and the probability of false concordance (P_{11}) 0.0001% or $1/10^6$ (Table 3.2, third row).
For all bps, the probability of discordance would be nearly 0.1998%, with nearly 6×10^4 variants and 6×10^6 non-variants. The number of bps called as concordant variants would be 2.94×10^6, of which 3×10^3 bps would be actually non-variants. The number of bps called as concordant non-variants would be 3×10^9, of which 300 actual variants. With the above-shown values, the probability of error for a given pair of calls can be calculated as follows. For a positive concordant pair, $P (V = 0 \mid A = 1, B = 1) = 3 \times 10^3/3 \times 10^6 = 0.1\%$; i.e., PPV = 99.9%. The probability of error for a negative concordant pair, $P (V = 1 \mid A = 0, B = 0) = 3 \times 10^2/3 \times 10^9 = 1/10^7$; i.e., NPV = 99.99999% (Table 3.3, column 1). For a discordant pair, the PPV for the positive call was around $6 \times 10^4 /6 \times 10^6 = 1\%$ and the NPV for the negative call was around 99% (See Table 3.3 for more parameter value assumptions).

Table 3.2 - Conditional probabilities of test results for variants and non-variants under various sensitivity and specificity values.

	Concordance		Discordance
	P_{11}	P_{00}	$P_{10} + P_{01}$
Variants (V=1)			
99% sensitivity	98.01%	0.01%	1.98%
90% sensitivity	81%	1%	18%
Non-variants (V=0)			
99.9% specificity	0.0001%	99.8%	0.1998%
99% specificity	0.01%	98.01%	1.98%

Reading example: P_{11} is the probability of test A positive ($_1$) and test B positive ($_1$) for a given variant or non-variant.

Table 3.3 – Predictive values of test results according to the true base pair statuses in various sensitivity and specificity values.

Test results and true base pair status	99.0% Se & 99.9% Sp	99.0% Se & 99.0% Sp	90.0% Se & 99.9% Sp	90.0% Se & 99.0% Sp
A=0 and B=0				
V=0	99.99999%	99.999%	99.99999%	99.999%
V=1	0.00001%	0.001%	0.00001%	0.001%
A=1 and B=1				
V=0	0.1%	9.4%	0.1%	11.1%
V=1	99.9%	90.6%	99.9%	88.9%
A=1 and B=0				
V=0	99%	99.9%	91.7%	99.1%
V=1	1%	0.1%	8.3%	0.9%
A=0 and B=1				
V=0	99%	99.9%	91.7%	99.1%
V=1	1%	0.1%	8.3%	0.9%

Reading example: A=0 means test A negative and V=0 means non-variants. In all four conditions, prevalence π=0.1%.

3.2.3 Results under conditional dependence

Table 3.4 shows the probabilities of calling results conditional on variant/non-variant status with various combinations of Se and Sp values and various levels of conditional correlation (30%, 50%, or 90%).

For example, for a given variant, with Se = 99% and correlation = 90%, the probability of discordance (P_{10}+P_{01}) would be 0.2% and that of false concordance (P_{00}) 0.9%.; and, for a given non-variant, with Sp = 99.9% and correlation = 90%, the probability of discordance (P_{10}+P_{01}) would be 0.02% and the probability of false concordance (P_{11}) 0.09%. In this example, the probabilities of false concordance are much higher than those of discordance; this means that most errors are common to both calling results.

With 99% Se, 99.9% Sp, 0.1% prevalence π, and assuming conditional probabilities of error P(A = 0 | B = 0, V = 1) = 90% and P(A = 1 | B = 1, V = 0) = 90%, the probability of discordance (P_{01}+ P_{10}) would be 0.02% (i.e., number of discordant bps = 5.9×10^5). Among these discordant bps, 5.9×10^3 would be variants and the others non-variants. The number of positive concordant bps would be 5.7×10^6, of which 2.7×10^6 would be non-variants. The number of negative concordant bps would be 3×10^9, of which 2.7×10^4 would be variants.

With the above-shown values, the probability of error for a given pair of calls can be calculated as follows (First column of Table 3.5). For a positive concordant pair, the probability of error P (V = 0 | A = 1, B = 1) = 2.7×10^6 / 5.7×10^6 = 47.4%; i.e., the PPV of the pair = 52.6%. For a negative concordant pair, the probability of error P (V = 1 | A = 0, B = 0) = 2.7×10^4 / $3 \times 10^9 \approx 1/10^5$; i.e., the NPV of the pair ≈ 99.999%. For a discordant pair, the PPV for the positive call P (V=1 | A ≠ B) = 1% and the NPV for the negative call P (V=0 | A ≠ B) = 99%.

Table 3.4 - Theoretical probabilities of test results for variants and non-variants under different conditions of sensitivity, specificity, and degree of correlation.

Conditions	Correlation 90%	Correlation 50%	Correlation 30%
Variants (V = 1)			
Sensitivity 99%			
P_{11}	98.9%	98.5%	98.3%
$P_{01} + P_{10}$	0.2%	1.0%	1.4%
P_{00}	0.9%	0.5%	0.3%
Sensitivity 90%			
P_{11}	89%	86%	83%
$P_{01} + P_{10}$	2%	8%	14%
P_{00}	9%	6%	3%
Non-variants (V = 0)			
Specificity 99.9%			
P_{11}	0.09%	0.05%	0.03%
$P_{01} + P_{10}$	0.02%	0.1%	0.14%
P_{00}	99.89%	99.85%	99.83%
Specificity 99%			
P_{11}	0.91%	0.51%	0.3%
$P_{01} + P_{10}$	0.18%	0.98%	1.4%
P_{00}	99.01%	98.5%	98.3%

P_{11}: probability of positive concordance between two tests - P_{00}: probability of negative concordance - $P_{01} + P_{10}$: probability of discordance.

Table 3.5 - Predictive values of test results for variants and non-variants under different conditions of sensitivity, specificity, correlation and variant prevalence.

Test results and true base pair status	Joint conditions of sensitivity, specificity, correlation, and variant prevalence				
	Se = 99% Sp = 99.9% Cor_1 = 90% Cor_2 = 90% π = 0.1%	Se = 99% Sp = 99.9% Cor_1 = 50% Cor_2 = 50% π = 0.1%	Se = 99% Sp = 99.9% Cor_1 = 30% Cor_2 = 30% π = 0.1%	Se = 99% Sp = 99.9% Cor_1 = 50% Cor_2 = 50% π = 0.2%	Se = 90% Sp = 99% Cor_1 = 90% Cor_2 = 90% π = 0.1%
A=0 and B=0					
V=0	99.999%	99.9995%	99.9997%	99.999%	99.999%
V=1	0.001%	0.0005%	0.0003%	0.001%	0.001%
A=1 and B=1					
V=0	47.4%	34%	23.1%	20.3%	90%
V=1	52.6%	66%	76.9%	79.7%	10%
A=1 and B=0					
V=0	99.1%	99%	99%	98%	98.9%
V=1	0.9%	1%	1%	2%	1.1%
A=0 and B=1					
V=0	99.1%	99%	99%	98%	98.9%
V=1	0.9%	1%	1%	2%	1.1%

Cor_1: Correlation between test A and test B for variants - Cor_2: Correlation between test A and test B for non-variants - π: variant prevalence

3.3 Illustration with real data -- NA12878 replicates

3.3.1 Material and methods

The present study used calling results from sequencing three technical replicates of genome NA12878. All three procedures were carried out on Illumina NovaSeq 6000 system platform, samples were then aligned with Burrow-Wheeler Aligner (BWA-MEM) (Li, 2013) against GRCh37 version of the human reference genome. GATK duplicate marking, base quality score recalibration, and indel realignment were applied. The three replicates were joint-genotyped by GATK (McKenna et al., 2010). Variant calling was performed according to GATK Best Practices recommendations (DePristo et al., 2011; Van der Auwera and O'Connor, 2020) by joint genotyping. The latest version (v4.2.1) of GIAB variant calling benchmark set were used as 'gold standard'. This version has a higher coverage of the reference genome and includes more difficult-to-map regions than previous versions (Wagner et al., 2022; Zook et al., 2016).

Analyses on real WGS data were carried out only on the bps from the GIAB benchmark region. Each base-pair position was considered as a statistical unit and each GIAB benchmark result was considered as the true status of each bp. On these bps, two types of analysis were performed: i) performance analysis by comparing results from each replicate with the truth set (Table 2.1); and, ii) concordance analysis by comparing results between any two replicates using a two-by-two contingency table (Table 3.1). Only performance for SNVs was analyzed in this study.

Performance evaluation used the same above-provided definitions of TPs and TNs. When both the VCF file and the GIAB benchmark set considered a given bp as variant, that bp was classified as a TP. When the bp was identified as variant in the VCF file but not in the GIAB benchmark set, it was classified as a FP. When the bp was identified as variant by the GIAB benchmark set but a non-variant or a no-call in the VCF file, it was classified as a FN. Here, TPs + FNs = the number of variants in the Benchmark call set. When a homozygote reference bp in the GIAB benchmark set was considered as a non-variant or a no-call in the VCF file, it was classified as a TN. N being the number of homozygote reference bps in the GIAB benchmark call set, Se was calculated as TPs / (TPs + FNs), Sp as 1 – (FPs / N), and the PPV as TPs / (TPs + FPs).

For concordance analysis, the definition of "local match" was used. The concordance rate and the correlations were calculated for the real variant positions (positions called as variants in the truth set), the real non-variant positions (positions called 'homozygous reference' in the truth set), and all positions.

3.3.2 Results

The total number of bps in the GIAB benchmark region is around 2.5×10^9, of which 3,238,599 variants in the gold standard set. The number of called variants within the same region in the joint VCF file is 3,351,415 (precisely 3,311,321; 3,308,075; 3,305,948 for the three replicates, respectively).

Within the GIAB benchmark region, the estimated sensitivity for the three replicates ranged from 98.97 to 98.99% and specificity from 99.9958 to 99.9960%. For called variants, the PPV for the three replicates ranged from 96.82 to 96.95% (Table 3.6).

In the concordance analysis, for replicates 1 and 2, the proportion of concordant bps across all positions in the VCF file (called as variant in at least one replicate) was 98.38%. The proportion of concordant bps for the variants in the benchmark set was 99.85% and that for the non-variants 99.9980% (Table 3.7).

Based on the conditional two by two tables for replicates, the estimated correlation coefficients and the rates of false concordant bps are showed in Table 3.7. For variant positions, the correlation coefficients between any two replicates ranged from 92.37% to 93.28% (with 0.94% rate of false concordant bps), whereas, for non-variants, the correlation coefficients ranged from 74.62% to 76.04% (with a rate of false concordant bps of 0.0031% to 0.0032%).
Regarding all observed results, the PPV ranged from 97.57% to 97.66% for positive concordant results between any two replicates and from 8.87% to 9.57% for discordant results.

Table 3.6 - Performance in sequencing the three NA12878 replicates.

	Replicate 1	Replicate 2	Replicate 3
All observed positives	3,311,321	3,308,075	3,305,948
True Positives	3,205,932	3,205,240	3,205,142
False positives	105,389	102,835	100,806
False negatives	32,667	33,359	33,457
Sensitivity	98.99%	98.97%	98.97%
Specificity	99.9958%	99.9959%	99.9960%
Positive predictive value	96.82%	96.89%	96.95%

Number of variants in the GIAB benchmark call set: 3,238,599.

Number of base pairs in the GIAB benchmark region: 2,502,460,587.

Number of non-variants in the GIAB benchmark call set: 2,499,221,988 (= 2,502,460,587 - 3,238,599

Table 3.7 - Analyses of concordance between replicates

Replicate comparisons	Concordance rate	Correlation	False concordance	PPV for positive concordance ($_{11}$)	PPV for discordance ($_{01 \text{ or } 10}$)
For variants (N=3,238,599)					
Rep 1 vs. Rep 2	99.85%	93.28%	0.94%	-	-
Rep 1 vs. Rep 3	99.85%	92.87%	0.94%	-	-
Rep 2 vs. Rep 3	99.83%	92.37%	0.94%	-	-
For non-variants (N=2,502,460,587)					
Rep 1 vs. Rep 2	99.9980%	76.04%	0.0032%	-	-
Rep 1 vs. Rep 3	99.9980%	75.14%	0.0031%	-	-
Rep 2 vs. Rep 3	99.9980%	74.62%	0.0031%	-	-
For all positions in the VCF (N= 3,351,415*)*					
Rep 1 vs. Rep 2	98.38%	-	-	97.57%	8.89%
Rep 1 vs. Rep 3	98.34%	-	-	97.62%	8.87%
Rep 2 vs. Rep 3	98.33%	-	-	97.66%	9.57%

3.4 Covariable analysis

3.4.1 Methods

Using the GIAB benchmark set as gold standard, generalized additive models were built with a logistic link to estimate covariate effects on the probabilities of discordance and the probability of error. Mathematically, this model may be written:

$$Logit(P) = \beta_0 + \sum_{m=1}^{M} f_m(X_{ijm}) + \epsilon_{ij}$$

In this formulation, i is the number of replicates, j the bp position, X_m a covariate (m = 1,…, M), β_0 the intercept, ε_{ij} the Gaussian error, and P = P (Y = 1) with Y = 1 for discordance or error and Y = 0 for concordance or correct call. Here, the concordance for replicate 1 was defined as the concordance between replicate 1 and replicate 2, the concordance for replicate 2 as the concordance between replicate 2 and replicate 3, and the concordance for replicate 3 as the concordance between replicate 3 and replicate 1.

The modeling used function “bam” of R package mgcv adapted to large dataset analyses. The smoothing method used for this model was a natural cubic regression spline. The smoothing parameter estimation method was the (default) fast restricted maximum likelihood (REML) computation. This function uses penalized iteratively re-weighted least squares (PIRLS) with a single iteration in model fitting, a method similar to “performance-oriented iteration”. (Wood et al., 2017, 2015).

The covariates included in this study were:

1) The depth of coverage (DP); i.e., the number of informatics reads covering a given base pair. Sites with DP >100 were excluded due to high probability of mapping artefacts (Li, 2014).
2) The allele balance (AB); i.e., number of reads supporting the alternative allele (other than the reference) divided by the number of all informatics reads at a specific site.
3) The genotype quality (GQ); i.e., the Phred-scaled confidence for the called genotype (range: 0 to 99).
4) The QualByDepth (QD); i.e., the site-level Phred-scaled confidence for the existence of variant, QUAL score, normalized (divided) by the total number of reads supporting the alternative allele in variant samples.

5) The mapping quality (MQ); i.e., the root mean square of the mapping quality of reads across all samples.

Covariates DP, AB, and GQ were obtained from the VCF file for each bp in each sample (here, replicate). MQ and QD were obtained from the VCF file for each bp and had the same values across three samples.

Both univariate and multivariate analyses were conducted. In the multivariate analyses, the optimal model was selected on the basis of the Akaike Information Criterion (AIC). The proportions of deviance explained by the models were also estimated.

3.4.2 Results

For most covariates, the functional forms that describe each covariate effect in each discordance or error model were quite close (Figure 3.1). As expected, overall, GQ, MQ, and QD had decreasing trends; i.e., the higher was the score, the lower was the probability of discordance or error. The DP had a V-shape effect; the lowest probability of error or discordance corresponded to the DP value with the highest density (Figure 3.2).

In the error model, AB showed an M-shape effect; the three lowest probabilities of error corresponded to AB values close to 0, 0.5, and 1. In the discordance model, the main difference was that the probability of discordance did not show a minimum at AB values close to 0. This difference was also found in the density graphs, where the density of discordance increases as AB approaches 0. However, the contribution of each covariate to the deviance differed between discordance and error models (e.g., DP explained 11.6% of the deviance in discordance model vs. 22.2% in error model). GQ, AB, and QD were more correlated with discordance, whereas DP and MP were more correlated with error.

Figure 3.1 - Univariate regression models with smoothing.

The first row is for the relationship between probability of discordance and each covariate and the second for the relationships between probability of error and each covariate. Each curve shows the estimated function for each covariate. The top right corner of each graph shows the percentage of deviance explained by the covariate.

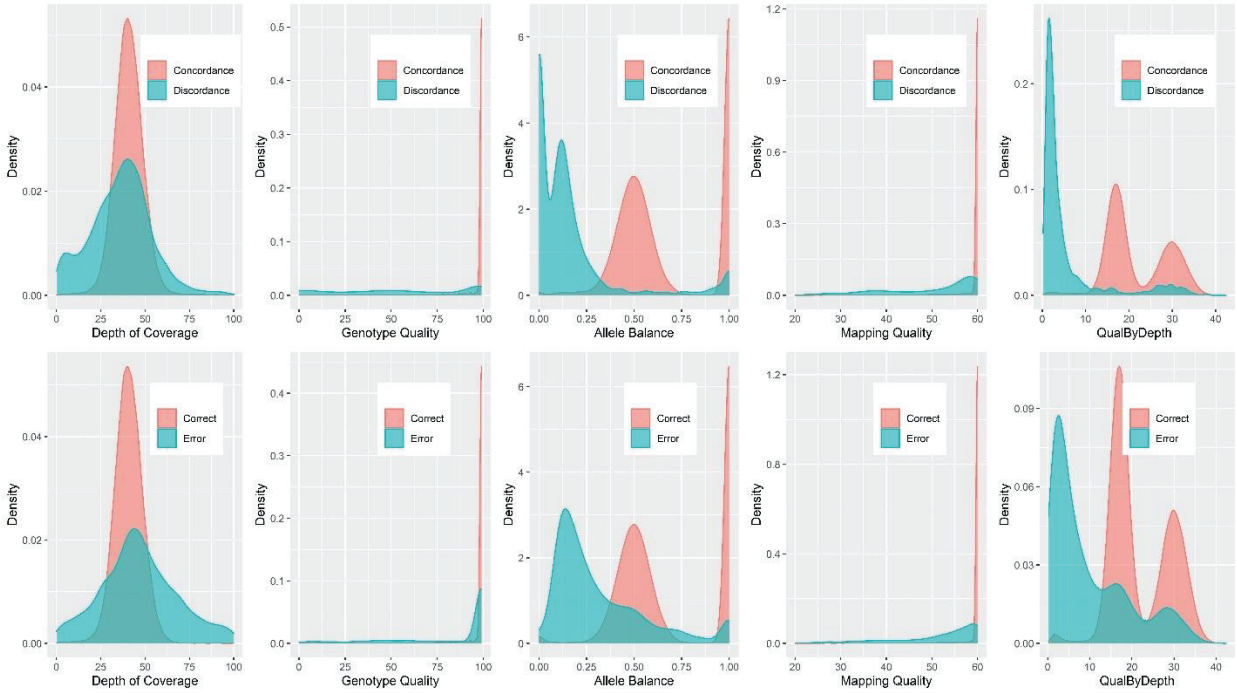

Figure 3.2 - Density functions of the covariates.

The first row presents density functions of the covariates for concordant and discordant calling results, the second row presents density functions for correct and error calling results.

In multivariate analyses, after model selection, the two optimal regression models for discordance and error were the models that included all five covariates. The deviance explained by the model was slightly higher with the error than with the discordance model (69.7% vs. 67.7%) (Figure 3.3). The shapes of the estimated functions of the adjusted covariate effect in multivariate models differed from the univariate models for most of the covariates (Figure 3.3 vs. Figure 3.1). The shapes of estimated adjusted functions differed between discordance model and error model for some covariates, e.g. DP, AB, and MQ (Figure 3). DP had the similar "V" shape function form in multivariate model and univariate model of error; however, in the multivariate model of discordance, it had a quasi-monotonically decreasing trend. MQ had little effect in the multivariate discordance model, but a similar form in both univariate and multivariate error models.

GQ had little effect in both multivariate models for discordance and error, despite having clear decreasing trend in both univariate models. AB had similar functional forms in the univariate and the multivariate discordance model but a smaller effect in the multivariate model. However, in the error model, after the first peak at around 0.15, the function did not show the second peak of the M shape as in the univariate model, instead, the estimated error probability decreased as AB increased. QD had similar functional forms in the univariate and both multivariate models

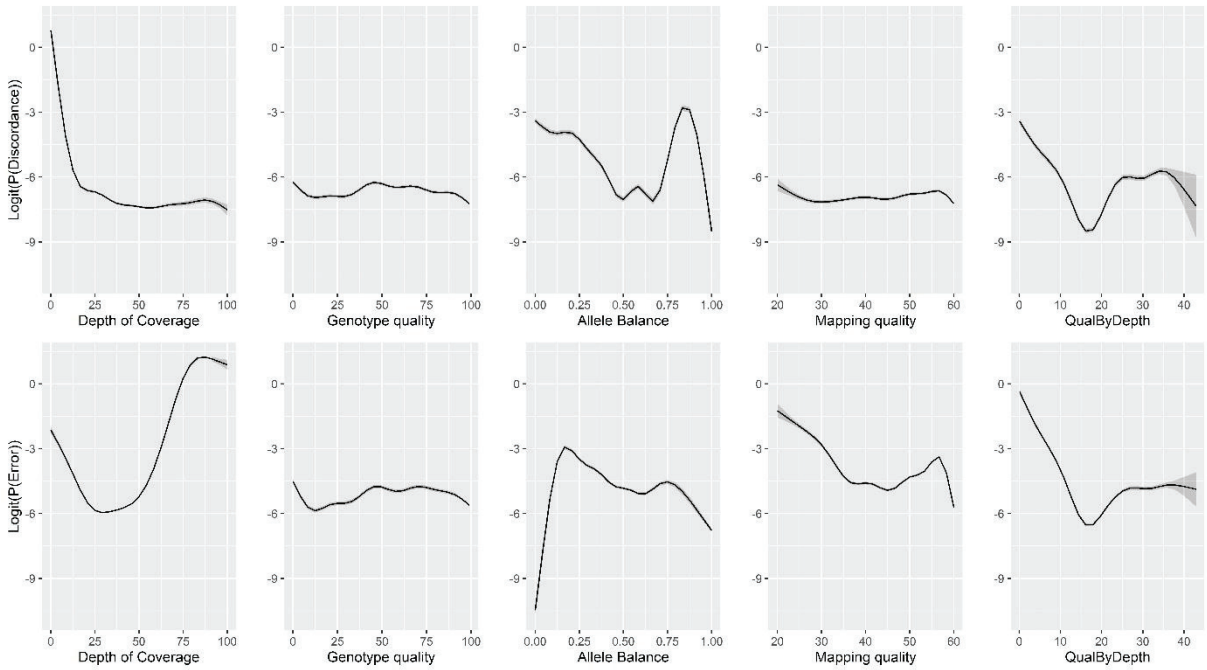

Figure 3.3 - Multivariate regression models with smoothing.

The first row uses a model for the relationships between probability of discordance and each covariate and the second another model for the relationships between probability of error and each covariate. Each curve shows the estimated function for each covariate when the other covariates in the model are set to their median values. The percentages of deviance explained by those models are 67.7% and 69.7%, respectively.

3.5 Discussion

In WGS studies on real data, concordance between replicates has been widely used as a substitute for a lacking gold standard. However, its appropriateness has been rarely questioned. The present study was motivated by the need to clarify the relationship between discordance and error in real data analyses. It used first theoretical analyses with conditional probabilities of error and discordance coupled with the most common WGS performance metrics. It analysed next real sequencing data to compare error rate and discordance rate. Lastly, generalized additive models with smooth functions were built to estimate the effects of sequencing and variant calling covariates on error and discordance.

In case of conditional independence between two sequencing results, the overall probability of error for concordant results was found almost negligible; the concordance-discordance method is then an acceptable substitute for a gold standard. However, in settings with high levels of correlation between sequencing results, a high proportion of false concordant results was found among all concordant results; the concordance methods becomes questionable. With the real sequencing on NA12878 genome, the probability of being a variant in case of discordant bps between replicates ranged from 8.87% to 9.57%, indicating that in pairs of discordant calls, the positive call had a high probability of being a false positive. However, the PPV for a concordant pair of positive calls ranged from 97.57% to 97.66%, which is not much higher than the PPV for a single positive call (96.82% to 96.95%); thus, the error rate for concordant calls was non-negligible.

In the univariate modelling analyses, the shapes of the estimated functions of most model covariates were quite similar; however, covariate contributions to the deviance differed between discordance models and error models. In the multivariate analyses, substantial differences were found between the error model and the discordance model. These results indicate that the concordance criterion should be used with caution, especially when the sequencing process and the calling process generate highly correlated results.

The data used in this study feature several peculiarities. The sequencing results were obtained from the same sequencer and the same variant caller. This common practice for inaccuracy detection using replicates generates usually high correlations between two sets of results (Naj et al., 2019; Robasky et al., 2014). Here, the three replicates were jointly

genotyped by GATK (Poplin et al., 2017); which is expected to generate even higher correlations. The more correlated are two tests, the higher is the probability of concordant results between them. Correlation values may not have a great influence on the number of concordant bps or the proportion of errors in variants or non-variants given that Se and Sp are already close to 1(here, the vast majority of the 3×10^9 bps remained correct). However, correlation values could have a dramatic influence on the number of discordant bps, thus on the PPV and NPV for concordant or discordant bps. Here, the correlation levels were about 92% among variants and about 78% among non-variants, which are high correlation levels versus those examined in the theoretical part of this work. These high correlation levels led naturally to a much higher number of concordant false negatives than the number of single false negatives (i.e., discordances) for variants. For true variants, the percentage of false negatives in each sequencing (= 1 – Se) ranged from 1.01 to 1.03%, and the percentage of concordant false negatives among false negatives was 0.94%; this means that nearly 92% of false negatives in one replicate are shared with another (0.94% of all variants). Similarly, for non-variants, the percentage of false positives in each sequencing (= 1 – Sp) was 0.0041% and the percentage of concordant false positives among false positives was 0.0031%; this means that nearly 76% of false positives would be concordant with another replicate.

Comparing performance analysis with concordance analysis required NA12878 genome sequencing because this genome has a gold standard. However, the sequencing performance metrics of genome NA12878 might not be sufficiently representative of those of real individual genomes because the former presents pipeline-linked trend toward overfitting. In addition, the use of 'local match' instead of genotype match to build a 2 × 2 contingency table for performance analysis led to high performance metrics estimates. Thus, the estimates of Sp with the real data were higher than those seen with the theoretical data (99.996% vs. 99.9%), and the estimates of Se reached the highest of two values in the theoretical data (99%). These performance metrics are comparable to those reported by other studies. For example, i) 97% Se and 98.6% PPV with GATK4 for SNVs in the NA12878 genome in Supernat et al. (2018); ii) 98.66% Se and 99.15% PPV for whole exome regions in Krishnan et al. (2021); and, iii) 97.17% Se and 99.999% Sp with BWA-MEM and GATK UnifiedGenotyper in Highnam et al. (2015)

In this work, using discordance as indicator of error (more specifically, of false positivity) seemed acceptable for all calling results (variants and non-variants) because the observed PPV for the positive call in a discordance pair was 9% and the PPV for a positive concordant

pair was 97.3%. Still, interpreting discordant or concordant results requires carefulness. In fact, the PPV of an individual positive call was around 96.6%; thus, there was no great difference between the PPV of a separate positive call and that of a concordance positive call due to the strong correlation.

Here, the concordance rate between replicates was 98.3 to 98.4% for all called variants and around 99.9980% for all positions in the benchmark region. These rates recall previous rates in whole genome sequences: i) 98.69% concordance rate among called SNVs in WGS (Adelson et al., 2019): ii) 99.49% average pairwise concordance rate between replicates sequenced in different centers and called by GATK pipeline (Naj et al., 2019); and, iii) 99.998% concordance rate among all callable positions across the whole genome (Ajay et al., 2011).

Regarding the analysis of the covariates' effects, the shapes of the estimated functions for discordance and error were comparable for most covariates, but the percentages of deviance explained by the models differed. GQ, AB, and QD were rather correlated with discordance (e.g., for GQ, 36.9% for discordance and 9.49% for error), whereas DP and MQ were rather correlated with error (e.g., for DP, 11.6% for discordance and 22.2% for error). This indicates that using discordance instead of error may lead to different model fits. In terms of functional forms, GQ, MQ, and QD had generally decreasing trends, whereas DP had a V shape and AB an M shape. The latter two forms are consistent with previous findings (Li, 2014; Muyas et al., 2019). For AB, different shapes appeared in the part where AB values was close to 0. When the AB approached 0, the estimated probability of discordance increased in the discordance model, whereas, it decreased significantly in the error model. This difference can be also found in the density graphs where the density of discordance increased when AB approached 0. Of note is that, in this study, covariate effect analyses were conducted on bps in the VCF file (i.e., called positive by at least one replicate); therefore, there were very few bps with AB close to 0.

In the univariate and multivariate analyses, AB and QD were the two covariates that contributed the most to the deviance. While the estimated functional forms of QD were comparable in both analyses, the functional forms of AB showed a substantial difference, especially in the error model in that the second peak of the M shape tended to disappear. The contribution of GQ in both multivariate models and that of MQ in the discordance model were small: the shapes of the functional forms were almost flat. These differences in the estimated

functions between univariate and multivariate analyses are probably due to different correlation levels between covariates. AB and QD had a high level of correlation (0.9), whereas GQ and MQ were moderately correlated (0.4). DP had very low correlation levels with the other covariates; this would explain its relatively high level of contribution to the deviance explained by the model in the multivariate models.

One limitation of the present study is the use of "local match" instead of a more accurate "genotype match" or "allele match". Further analysis are needed to check whether the same conclusions may be drawn with "genotype match" or "allele match". Nevertheless, the knowledge, the methods, and the applications derived from this study would still be valid. Another perspective would be establishing predictive models with multiple covariates to generate estimated error rates for individual base-pair positions.

3.6 Conclusions

In case of conditional independence between two sequencing results, the overall probability of error for concordant results being negligible, the concordance-discordance method is acceptable. However, in settings with high correlation levels, the method becomes questionable because of a high proportion of false concordant results. With real data from NA12878 vs. GIAB benchmark set, discordance (as indicator of error) seemed acceptable but with caution in interpreting discordant or concordant results. Multivariate analyses showed substantial differences between error and discordance models; thus, caution is required in using the concordance criterion, especially in case of highly correlated results.

4. Performance comparison of clustering models with NGS replicates

In chapter three, we evaluated the usefulness and limitations of the concordance-discordance model in dealing with the NGS replicates. We showed that the concordance-discordance model may not be the best choice when test results are highly correlated, which is often the case in NGS replicates. Therefore, we looked into the literature for other models implemented to combine callsets from replicates.

Indeed, many models, including sophisticated machine-learning models, have been used in this research field. However, some of them are supervised models that often require high-quality training data that are not always available. Moreover, the generalization of training to test data sets has always been a concern in performance evaluation of these models (Wang et al., 2020). In case of substantial differences in data structure between the training set and the test set; the models may not generalize well. Therefore, we focus on exploring the non-supervised models that do not require the use of training (or external) data and investigate whether they may give stable results and lead to performance improvement.

The literature on processing replicate sequencing results with non-supervised models is rather scanty and the different models used have been rarely objectively compared. This work intends to explore the main models that deal with multiple NGS results stemming from biological or technical replicates, investigate their properties, and compare their key performance indicators to help choosing the most performant among readily implementable methods able to improve sequencing performance. Precisely, this work explores the consensus model, the latent class model, the mixture model, and random forest regarding their abilities to produce a callset with improved quality. It compared their main performance indicators: precision, recall, and F1-score.

Section 4.1 aims to present the research context and provide a literature review of the methodology in related works. In section 4.2, we position the question in the statistical world as a clustering problem and give an overview of the major categories of clustering models. Then we apply representative models of each category to three technical replicates of the NA12878 genome and compare their performances in section 4.3 to 4.5.

4.1 Context – Combining multiple variant calling sets

As presented in section 2.2, the use of replicates (either technical or biological) to mitigate error and obtain a more accurate result is very common in studies involving NGS. However, the way of using the information coming from multiple callsets of a single underlying truth remains debatable. Several distinct approaches have been proposed for this purpose. Table 4.1 provides a short overview of the most relevant methods and studies designed to combine multiple callsets.

One category of methods is consensus-based. For example, the consensus genotyper for exome sequencing (CGES) employs a two-stage voting scheme between four bioinformatics tools by first voting for the variant positions then for the genotype of the variants (Trubetskoy et al., 2015). The level of concordance required for the voting (e.g., three out of four or four out of four) can be specified by the user. Later, a web-based automated interface consensus variant calling system (CoVaCS) was developed using a similar majority voting scheme between three variant callers. It demonstrated a similar performance improvement to that brought by CGES (Chiara et al., 2018).

Another category of algorithms is based on statistical models. For example, the BAYSIC (BAYeSian Integrated Caller) tool is based on a latent class analysis approach fitted with a Bayesian method (Cantarel et al., 2014). The threshold for posterior probability was 0.8 by default, but it can be modified by the user. A different combining indicator was proposed by Hofmann et al. (2017). The authors used the ranks of variants in terms of confidence score within each variant caller to form a combined ranking score on the same scale across different variant callers, taking into account the correlations between callers. This combined ranking score that ranges from 0 to 1 was interpreted as a probability of error. A threshold was then applied to this combined confidence score to obtain the final callset.

One advantage of statistical models is that they often estimate posterior probabilities reflecting the error rates or confidence. These probabilities are intuitive and useful for the interpretation of model output.

In recent years, more machine-learning or deep-learning models have been explored for NGS data as well. Different types of models such as random forest (RF), support vector machine (SVM), or convolutional neural networks (CNN) have been used for variant calling or callset filtration (Friedman et al., 2020; Lek et al., 2016; O'Fallon et al., 2013). These models developed to perform 'classification' tasks can be easily adapted for the classification of

‘errors’ and ‘true variants’ in a given callset or combining multiple callsets. For example, VariantMetaCaller (Gézsi et al., 2015) uses SVM to combine multiple variant callers and compute the estimated probabilities of called variants to be true variants. This SVM model is trained by using fully concordant variants as positive training examples and variants called by only one of the callsets (“singly-called variants”) as negative training examples. One problem pointed out by the authors was that a substantial proportion of the negative training set consisted in actually true variants. In this study, nearly 50% of the singly-called SNVs with above 30× coverage were true variants, whereas the percentage of variants called by all four methods was 99.83%. In fact, due to the sensitivity of the training data set, the drawbacks of machine learning models often lie in the choice of the training data, their quality, and the differences between the training data and the test data. In a comparison study of different methods that combine multiple somatic variant callers (Wang et al., 2020), the machine learning models showed very inconsistent performance metrics. The performances varied largely depending on the similarity between the training set and the test set and, in many cases, were not better than a simple consensus-based approach.

One interesting advantage of machine-learning or deep-learning models is that they are often able to include more information as input than statistical models. This capacity of including more quality-related information generated by the variant callers could result in improved performance, as demonstrated by Gézsi et al. (2015).

Table 4.1 – Overview of the most relevant methods and studies designed to combine multiple callsets

Authors	Algorithm	Model type	Reference
Trubetskoy et al., 2015	CGES	Consensus	Bioinformatics 2015;31(2):187. doi:10.1093/bioinformatics/btu591
Wang et al., 2020	SomaticCombiner	Consensus	Sci Rep 2020;10:12898 doi:10.1038/s41598-020-69772-8
Chiara et al., 2018	CoVaCS	Consensus	BMC Genomics 2018;19:120. doi: 10.1186/s12864-018-4508-1
Hwang et al., 2014	---	Consensus and logistic regression	Hum Mutat 2014;35(8):936. doi: 10.1002/humu.22587
Cantarel et al., 2014	BAYSIC	Bayesian latent class model	BMC Bioinformatics 2014;15:104. doi: 10.1186/1471-2105-15-104.
DePristo et al., 2011	VQSR	Gaussian mixture model	Nat Genet 2011;43(5):491. doi: 10.1038/ng.806.
Hwang et al., 2019	---	Gaussian-multinomial mixture model	Sci Rep 2019;9(1):3219 doi: 10.1038/s41598-019-39108-2
Huang et al., 2019	SMuRF	Random forest	Bioinformatics 2019; 35(17): 3157. doi: 10.1093/bioinformatics/btz018

4.2 Overview of clustering methods in statistics

In this section, the problem of clustering consists in partitioning a set of objects or data points into a fixed number of non-empty classes (clusters) that are as homogeneous as possible. Accoding to the indicators used to measure “homogeneity” or “similarity”, the clustering methods can be broadly divided into distance-based and model-based clustering models. Here, we did not include density-based models because they aim to separate clusters according to different densities and do not fit in our clustering objective.

Another way of categorizing clustering approaches is based on the assignment of a ‘membership’ to each data point. The methods can then be divided into hard clustering and soft clustering (sometimes called fuzzy clustering). In hard clustering, each observation is assigned to one cluster, whereas in soft clustering an observation could belong to several clusters with cluster-specific belonging probabilities. The approaches can also be categorized according to the data types that they are developed for (categorical data, continuous data, or a mixed of categorical and continuous data).

In this section, we first overview distance-based and model-based methods. Then we briefly summary the extensions of these models to mixed data (categorical and continuous variables). This is relevant to our application in NGS data because the variant calling results are categorical and the quality-related factors are mostly continuous. Finally, we present several criteria for the evaluation of clustering model performance.

4.2.1 Distance-based clustering

4.2.1.1 Hierarchical clustering

Hierarchical clustering methods can be divided into divisive (top-down) hierarchical clustering and agglomerative (bottom-up) hierarchical clustering. In agglomerative approaches, the clustering begins with each cluster containing one observation and then merging the two most similar atomic clusters regarding a certain similarity measure resulting in larger and larger clusters until all observations are included in a single cluster. Contrarily, divisive hierarchical clustering method first sets all data points into one initial cluster, divides the initial cluster into several smaller clusters, and iteratively partitions these clusters into

smaller ones until each cluster contains only one data point or data points within each cluster are similar enough (Saxena et al., 2017).

In both types, various similarity measures based on difference distance measures may be used. For example, single linkage clustering defines the similarity measure between two clusters as the shortest distance between data points from the two clusters; ii) complete linkage uses the largest distance between data points from two clusters to define the similarity measure; and iii) group average linkage uses the average distance between data points from two clusters to define the similarity measure; According to James et al. (James et al., 2013), average and complete linkage are generally preferred over single linkage as it tends to yield more balanced dendrograms.

One advantage of hierarchical clustering is that a complete hierarchy of clusters can be obtained; thus, the model can give multiple consistent partitions of the data by cutting at different levels. One common criticism of classical hierarchical clustering is that once an observation is assigned to a cluster, it can no more be considered in the following clustering steps. This means that the algorithms are not able to correct previous misclassifications. Besides, they are sensitive to outliers (Saxena et al., 2017). Some more advanced algorithms have been developed to address these disadvantages, such as Clustering Using Representatives (CURE) (Guha et al., 2001) and Balanced Iterative Reducing and Clustering Using Hierarchies (BIRTH) (Zhang et al., 1996).

4.2.1.2 K-means algorithm

Suppose X is an p-dimensional data set with n points and is divided into k clusters $C = \{C_1, C_2, \dots, C_K\}$. Let $z = \{z_1, z_2, \dots, z_K\}$ be the K cluster prototypes, where z_k is the mean of cluster C_k. The goal of k-means algorithm is to minimize the sum of the square error within clusters:

$$d(X, C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - z_k\|^2$$

The k-means algorithm operates by iterating the following steps:

1) Initialization: selects randomly a set of K data points as the initial cluster means;
2) Assignment: assigns each data point to its closest cluster mean;

3) Update: calculates the new cluster means according to the assignments;
4) Repeat of step 2 and 3 until there is no change of assignments.

One advantage of the k-means algorithm is that it is fast and guarantees convergence to a local minimum. Nevertheless, one disadvantage is that it requires that the variables be standardized to avoid the domination of the variables having the most variation.

4.2.1.3 Generalized k-means algorithms

The generalization of k-means algorithm relies largely on two aspects: the definition of a cluster center and the expression of the distance function.

The sum-of-squares type clustering criteria have been generalized in many ways in order to comply with different data types or cluster properties. In *k-medians algorithm*, the distance to minimize is the Mahalanobis-type one-norm distance ($\|x_i - z_k\|$) instead of the Euclidean distance, where z_k is the class prototype.

Concerning the definition of cluster center or cluster prototype, one extension of the k-means algorithm is the use of an actual point prototype (real data point) instead of a virtual point prototype such the mean. For example, in *k-medoids algorithm*, also referred to as *partitioning around medoids* (PAM), the algorithm searches for an optimal set of *K* data points as cluster prototypes to minimize the objective function (Kaufman and Rousseeuw, 1990). In *k-medians algorithm*, the median of each cluster is chosen to be the cluster prototype instead of the mean. This type of algorithm is more robust to outliers than the k-means algorithms, but it requires longer computation time. To adapt for large datasets, other algorithms combining sampling method and the k-medoids algorithm were also proposed to improve the computational efficacy, such as “Clustering LARge Application” (CLARA) program. It first creates multiple random samples of the data set and perform k-medoids clustering on each sample set. The resulting mediods are used to cluster the whole dataset and the solution with the minimum dissimilarity is selected. (Kaufman and Rousseeuw, 1990)

The classical k-means algorithm works only with numeric values. To deal with categorical data, Huang introduced the *K-modes* algorithm where a simple matching measure is used as dissimilarity measure. The mode of each cluster is used as cluster center and updated with a frequency-based method (Huang 1998).

4.2.2 Model based clustering

The model-based (or distribution-based) clustering approaches regard the observations as random samples from a finite mixture of distributions. By making assumptions about the forms of the distributions of mixture components, a statistical model, expressing usually the likelihood of observed data can be obtained. The parameters of each component distribution can be estimated using the maximum likelihood method. The conditional group-membership probabilities of each observation can then be used to obtain the clusters.

4.2.2.1 The general Finite mixture model

Suppose we have *n* observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. For each observation, *m* variables are available, denoted $\boldsymbol{x}_i = (x_{i1}, \dots, x_{im})$. The classical probability model of the K-component mixture distribution is a weighted average of K probability density functions (or probability mass functions in the case of discrete variables):

$$f(\boldsymbol{x}_i|\boldsymbol{p}, \boldsymbol{\alpha}) = \sum_{k=1}^{K} p_k f(\boldsymbol{x}_i|\boldsymbol{\alpha}_k)$$

In this equation, the mixing proportion p_k denotes the probability that variables of observation *i* were generated from the k[th] component, $\sum_{k=1}^{K} p_k = 1$. The parameters of the distribution of the k[th] component is $\boldsymbol{\alpha}_k$. The parameters of the model are $\boldsymbol{p}$ and $\boldsymbol{\alpha}$**.** The likelihood is:

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{p}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k f(\boldsymbol{x}_i|\boldsymbol{\alpha}_k)$$

The majority of model-based clustering applications use the EM algorithm (Dempster et al., 1977) for inference. The EM algorithm is an iterative algorithm where each iteration consists of an expectation step (E-step) and a maximization step (M-step). In this algorithm, the unobserved component membership of each observation is denoted $\boldsymbol{z}_i = (z_{i1}, \dots, z_{ik})$, where $z_{ik} = 1 = 1$ when observation i belongs to component k , and $z_{ik} = 0$ otherwise. $(\boldsymbol{x}, \boldsymbol{z})$ are regarded as the complete data. The EM algorithm then works with the likelihood function of the complete data. In the E-step, the conditional expectation of the complete data log-likelihood function is computed given the observed data and the current parameter estimates. In the M-step, the expected complete data log-likelihood function from the E-step is

maximized with respect to the model parameters. Iterating these E- and M-steps until convergence achieves at least a local maximum of the observed data likelihood function, under mild regularity conditions (Dempster et al., 1977). At convergence, the value $\widehat{z_{ik}}$, (i.e., the conditional expectation of z_{ik}) is the estimated conditional probability that observation i belongs to cluster k. The EM algorithm, as the k-means algorithm, is sensitive to the initial values and need to be run from a variety of start values to ensure that finding a global minimum is found.

In the case of clustering categorical variables, methods have been developed under the name of Latent Class Analysis (LCA); they are mathematical equivalent to binomial or multinomial mixture models. When the observed variables are continuous, the most popular model is the Gaussian mixture model (GMM), sometimes also referred to as latent profile analysis (Oberski, 2016).

One advantage of the model-based approach is that the covariables do not need to be scaled, and they are generally less sensitive to outliers than the distance-based models.

4.2.2.2 The latent class analysis (LCA)

The latent class analysis (LCA) model is the classical model used for clustering multivariate categorical data. The clustering problem can be naturally viewed as a latent variable problem where the cluster membership of each observation is unobserved or latent. A classical LCA assumes that the categorical variables are conditionally independent given the cluster memberships; this is known as the local independence assumption. (Goodman, 1974)

Several extensions of the model have been proposed to relax the local independence assumption by introducing a conditional dependence using difference measures.

4.2.2.3 The Gaussian mixture model (GMM)

A Gaussian mixture model is a finite mixture model where the distributions of each component are modelled as multivariate Gaussian distributions. In such a case, $\boldsymbol{\alpha}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean and variance matrix of the kth component.

Various constraints may be imposed upon the covariance structure (Banfield and Raftery, 1993) consider eigen-decomposition of the covariance matrices of the form

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \boldsymbol{\Delta}_k {\mathbf{D}_k}^T$$

where $\lambda_k = |\boldsymbol{\Sigma}_k|^{1/p}$ is the associated proportionality constant, $\mathbf{D}_k$ the matrix of eigenvectors of $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\Delta}_k$ a diagonal matrix such that $|\boldsymbol{\Delta}_k| = 1$, that contains the normalized eigenvalues of eigenvalues of $\boldsymbol{\Sigma}_k$ in descending order. "Geometrically, λ_k represents the volume of the ellipsoid, $\boldsymbol{\Delta}_k$ specifies the shape of the density contours, and $\mathbf{D}_k$ determines the orientation of the ellipsoid" (Banfield and Raftery, 1993). The volumes, shapes, and orientations of the cluster densities can be constrained to be equal (E) or variable (V) across clusters resulting thus in a family of fourteen models named "VEV", "EVE", etc. The first letter denotes whether the volumes are constrained to be equal (E) or variable (V) across clusters; the second letter denotes whether the shapes are constrained to be equal (E) or variable (V) across clusters or the clusters are spherical (I); and the final letter refers to the clusters' orientation: equal (E) or variable (V) across clusters or the clusters are axis-aligned (I). Studies showed that when all variables are continuous, a clustering based on the matrices of within-group sums of squares (e.g. k-means) corresponds to a clustering obtained from a multivariate Gaussian mixture distribution with constraints on the form of the Gaussian covariance matrix, i.e., equal within-cluster variance. (Banfield and Raftery, 1993)

This family of models is implemented in the widely use R package mclust. (Scrucca et al., 2016)

4.2.2.4 Extensions of Gaussian mixture models

In practice, the distributions of covariates within each component are often not Gaussian, to solve this problem, extensions of gaussian mixture models were introduced to robustly model the data and account for skewness, light or heavy tail, and dependency between covariates. The mixture of t distributions were introduced to model the heavy-tailed data (McLachlan and Peel, 2000); Mixtures of skew normal distributions and skew t distributions were proposed to model the asymmetrical data (Lee and McLachlan, 2013). Dang et al. (Dang et al., 2015) introduced mixture of power exponential distribution and skewed power exponential distribution to allow for model "components with varying levels of peakedness, skewness, and tail-weight". To model the dependencies between covariates, Gaussian mixture copulas

have been applied. (Kasa and Rajan, 2022) The hidden Markov model and mixtures of linear Mixed Models are other propositions for clustering correlated data. (McLachlan et al., 2019)

4.2.3 Clustering mixed dataset (categorical and continuous data)

The models introduced above were developed to cluster variables either all categorical or all continuous. When the dataset group both categorical and continuous variables, several approaches may be used to express the similarity.

For distance-based models, one approach is to dichotomize all of the variables and then use a dissimilarity measure for categorical data, or to convert the categorical variables into numeric variables and use a distance measure. Another approach is to construct a dissimilarity measure for each of the two types of variables and then use a weighting method to combine them into a single coefficient. For example, in k prototypes (Huang, 1998), the squared Euclidean distance is used as dissimilarity measure for continuous variables, the number of mismatches as dissimilarity measure for categorical variables, and a weighted sum of the two as the overall dissimilarity measure. However, the weight needs to be specified beforehand, which requires a prior knowledge of the data regarding the attribution of categorical variables compared to continuous variables. The algorithm was later extended to the W-K-prototypes algorithm to include the weight estimation in the model (Huang et al., 2005). Another model to estimate weights was proposed by Modha and Spangler (Modha and Spangler, 2003). In this model, the weight that minimizes the product of the continuous and categorical dispersion ratio is selected.

In model-based clustering approaches, the mixed type data are typically assumed to follow a Gaussian-multinomial distribution (McParland and Gormley, 2016). This Gaussian distribution assumption could be relaxed using kernel density methods (Li et al., 2007).

Based on a combination of generalized k-means algorithm and model-based clustering, a semi-parametric model termed Kamila (Kay-means for Mixed Large data sets) was proposed by Foss and colleagues (Foss and Markatou, 2018). In each iteration, the cluster parameters as well as the weight of continuous versus categorical variables are estimated, through maximizing the likelihood. The densities of continuous variables were estimated using kernel density estimation based on Euclidean distance under assumptions that the distributions are

spherically symmetric (Foss et al., 2016). This model showed a high performance in comparison study of clustering models (Preud'homme et al., 2021).

4.2.4 Model-selection criteria

The arguably most difficult methodological problem with clustering models is choosing the number of clusters. In this work, we fixed the number of desired clusters using biological prior knowledge and interpretability. Therefore, we do not discuss specifically the selection of an appropriate number of clusters, but only recall several general model selection criteria for model selection. Many indicators have been proposed for model evaluation, among which different information criteria.

The model selection criteria are mainly based on the likelihoods. One way to compare models is to perform a hypothesis test on the likelihood, called likelihood ratio test (LRT). However, hypothesis for the null distribution of the LRT does not always hold. Another way is considering a penalized log likelihood. As the likelihood is expected to increase with increased model parameters, penalized log likelihood could lead to a consistent selection of models. Various information criteria fall into the second category. (McLachlan et al., 2019)

The most popular information criterion for model selection is the Bayesian information criterion (BIC) that may be formulated as:

$$BIC = -2\ell\left(\widehat{\boldsymbol{\theta}}\right) + p \log n$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, $\ell\left(\widehat{\boldsymbol{\theta}}\right)$ the maximized log-likelihood, and p the number of free parameters. (Schwarz, 1978)

The theoretical assumption is based on the selection of the mostly likely model given the data. Using the Bayes' theorem, this probability for a given model M_j can be formulated as :

$$Pr\left(M_j \middle| data\right) = \frac{p\left(M_j\right)\lambda_j}{p(M_1)\lambda_1 + \cdots + p(M_k)\lambda_k}$$

Where λ_j is the marginal likelihoods of model M_j. Approximation for the log-likelihood with an increasing n gave the formula of BIC.

BIC is theoretically a consistent selector. Assuming one of the compared models is the true model, a consistent selector is one the selects the true model as the number of observations increasing.

Another information criterion is the Akaike's information criterion (AIC). (Akaike, 1973)

$$AIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p$$

It aim to choose the model that most accurately describes the true process that generate the data. (Dziak et al., 2020)

AIC is not consistent because it tend to choose a complex model as number of observations becomes large. Other modified AIC were later proposed such as the sample-size-adjusted AIC, or consistent AIC, using different penalty weights.

Some studies indicate that AIC tend to choose models with more parameters and BIC with less (Dziak et al., 2020). In the selection of the number of clusters, BIC is often observed to favour models with more components (McLachlan et al., 2019).

4.3 Material and methods

4.3.1 The study data

The present study used calling results from sequencing three technical replicates of genome NA12878. NA12878 is a human DNA sample that is "thought to represent the best-characterized diploid human genome in the world", is "considered as a 'reference material' by the National Institute of Standards and Technology (NIST)", and includes "near-perfect genome sequences for public use" as well as "truth sequences" established after repeated sequencings "using a wide variety of technologies and computational pipelines". Today, more than 80% the NA12878 cell line's genome is considered known with high confidence. This is why it is used as benchmark for assessing the performance of sequencing platforms or bioinformatic pipelines (Krol, 2015).

All three sequencing procedures were carried out on Illumina NovaSeq 6000 system platform. The samples were then aligned with Burrow-Wheeler Aligner (BWA-MEM) (Li, 2013) against the GRCh37 version of the human reference genome. Genome Analysis Toolkit (GATK) duplicate marking, base quality score recalibration, and indel realignment were

applied (McKenna et al., 2010). The resulting sequencing data were deposited in the European Nucleotide Archive.

Variant calling was performed by joint genotyping according to the GATK Best Practices recommendations (DePristo et al., 2011; Van der Auwera and O'Connor, 2020). Concordance rates between the calling results of the replicates were calculated. The concordance rate was defined as the number of sites called in the same category (see 4.3.2) by each replicate divided by the total number of sites called as variants by at least one of the replicates.

The latest version (v 4.2.1) of Genome in a Bottle (GIAB) variant calling benchmark set was used as 'gold standard' (Wagner et al., 2022; Zook et al., 2016). This version has a higher coverage of the GRCh37 reference genome and includes more difficult-to-map regions than the previous version (Wagner et al., 2022).

4.3.2 Basic definitions and main covariables

Performance considered only bps from the GIAB benchmark region, each bp position being a 'statistical unit' and each GIAB benchmark result a true status of each bp. Here, only performance in single nucleotide variant (SNV) analysis was considered.

In this analysis, the variant calling results in the VCF file and the GIAB benchmark callset (gold standard set) were considered to belong to one of three categories: homozygous reference, heterozygous variants, and homozygous variants. A true positive (TP) was defined as a variant call in the query callset that belongs to the same category as in the gold standard set; i.e., both are heterozygous variants or both homozygous variants despite potential allele or phasing differences. A false negative (FN) was defined as a variant in the gold standard set called as non-variant in the query callset. A false positive (FP) was defined as a non-variant in the gold standard set called as variant in the query callset or a variant in the gold standard set called as variant in a different category. A true negative (TN) was defined as a non-variant in the gold standard called as non-variant in the query callset. No-calls in the VCF file were considered as non-variants. This recalls the "genotype match, for which only sites with matching alleles and genotypes are counted as TPs" (Krusche et al., 2019), though, in this study, the criteria for true positivity were less stringent.

The covariables included in the models were:

1) The depth of coverage (DP); i.e., number of informatics reads covering a given base-pair. In this study, the mean DP value across the three replicates was circa 38 and the DP value ranged from 0 to 13,858.
2) The allele balance (AB; i.e., the number of reads supporting the alternative allele divided by the number of all informatics reads at a specific site) ranged from 0 to 1.
3) The QualByDepth (QD); i.e., the site-level Phred-scaled confidence for the existence of variant divided) by the number of reads supporting the alternative allele in variant samples. Here, the QD value ranged from 0.02 to 42.9.
4) The genotype quality (GQ); i.e., the Phred-scaled confidence for the called genotype (Ranged from 0 to 99).
5) The mapping quality (MQ); i.e., the root mean square of the MQ of reads across all samples. (Ranged from 20 to 60)

Covariates DP, AB, and GQ were obtained from the VCF file for each bp in each sample (here, replicate), and then the mean of each of the three values was calculated. MQ and QD were obtained from the VCF file for each bp and had the same values across three samples.

4.3.3 Clustering models used for NGS reconstruction

Five types of models were selected for reconstructing NGS result from technical replicates.

The consensus (or concordance-based) model

In this model, 'strict consensus' was considered whenever all variant calling results across all replicates agreed and 'majority consensus' whenever there was a majority of variant calling results across all replicates (Trubetskoy et al., 2015; Wang et al., 2020). Here, it is the majority consensus that was used. In case of no majority consensus, the sites were classified as homozygous variants.

The latent class model without covariables

This type of analysis was often used to evaluate the performance of diagnostic tests in the absence of gold standard. A latent class analysis is a mixture model where both the observed and unobserved variables are categorical. A classical LCA assumes conditional independence between observed variables (here, called genotype categories) given the latent class (here, the true genotype status).

Let *i* represent each site in the VCF file, *r* the latent classes 1 to 3. $\boldsymbol{Y_i}$ represents the calling results in replicates 1 to 3 for site *i* (Y_1, Y_2, and Y_3 are categorical variables with three categories that correspond to the three genotype categories). p_r denotes the prevalence of latent class *r*. $\pi_r(Y_1), \pi_r(Y_2)$, and $\pi_r(Y_3)$ are the probability mass functions of variables Y_1, Y_2, and $Y_3$for latent class *r*

The equation of this model may be written:

$$P(\boldsymbol{Y_i}|\boldsymbol{p}, \boldsymbol{\pi}) = \sum_{r=1}^{3} p_r \times \pi_r(Y_1) \times \pi_r(Y_2) \times \pi_r(Y_3) \quad [2]$$

The model parameters, namely $p_r$and π_{jrk} were estimated with an expectation-maximization (EM) algorithm using 50 sets of random initial values.

The latent class model with covariables

In this model, covariables' effects were put on the prior probability of class membership (P_r in equation [2]) and modelled using a logistic link. Covariables that are potentially correlated with the latent bp status were included; namely, Allele Balance (AB; the mean AB value of the three replicates), QualByDepth (QD), and Mapping Quality (MAPQ). Univariate models were first fitted for each covariable, then models were fitted with all possible pairs of covariables. Model parameters ($\boldsymbol{\pi}, \boldsymbol{p}$) were estimated using 100 sets of random initial values. Models with distinct covariables were compared with the Bayesian information criterion (BIC) as a measure of model fit.

The latent class model without covariables and the latent class model with covariables were fitted using package "poLCA" (v. 1.6.0.1) in R (v. 4.1.3) (Linzer and Lewis, 2011).

The Gaussian mixture model

The Gaussian mixture model assumes that the observed variables within each latent class follow a multivariate normal distribution. Here; it is the observed continuous covariables that were modelled, the calling results of each replicate were not included. The covariables included in the model were read depth (DP; the mean DP value of the three replicates), allele balance (AB; the mean AB value of the three replicates), and quality by depth (QD); and were assumed to be normally distributed.

Let $\boldsymbol{x}_i$ denote the vector of covariables for site *i*, p_r the prevalence of each latent class (*r*=1 to 3), $\boldsymbol{\alpha}_r$ the parameters of the multivariate normal distribution for latent class *r*. $h(x_i|\alpha_r)$ is the probability density function for latent class *r*, with parameters $\boldsymbol{\alpha}_r$. Thus, the probability density function for $\boldsymbol{x}_i$ can be written as:

$$f(\boldsymbol{x}_i|\boldsymbol{p},\boldsymbol{\alpha}) = \sum_{r=1}^{R} p_r h(\boldsymbol{x}_i|\boldsymbol{\alpha}_r)$$

The model parameters, namely $\boldsymbol{p}$ and $\boldsymbol{\alpha}$ were estimated with an expectation-maximization (EM) algorithm. This model was fitted using package "mclust" (v. 6.0.0) in R (v. 4.1.3) (Scrucca et al., 2016).

Kamila model (k-means for mixed large datasets)

Kamila is a model-based adaptation of the k-means clustering algorithm for heterogeneous variables (mix of categorical and continuous). It uses a kernel density estimation technique to model flexibly spherical clusters in the continuous domain and uses a multinomial model in the categorical domain (Foss et al., 2016). The model parameters were estimated with an iterative process similar to an EM algorithm. One advantage of this model is to include both types of variables at the same time without pre-specifying the weights of continuous versus categorical variables.

The categorical covariables included were: the calling results of the three replicates and a binary covariable to indicate whether a site is present in a 'difficult region' (Amemiya et al., 2019). The continuous covariables included were DP, AB, and QD. The algorithm is sensitive to outliers because it uses kernel density estimation and Euclidean distance for continuous covariables. Here, the maximum value of DP was set to 150.

This model was applied with package "Kamila" (v. 0.1.2) in R (v. 4.1.3) (Foss and Markatou, 2018).

The random forest

An unsupervised version of the random forest model for clustering was implemented (Shi and Horvath, 2006). The algorithm started with an unsupervised random forest model to generate a synthetic dataset without correlation between covariables, and then classified the observations into the synthetic or the original dataset using a classical random forest. This

generates a proximity matrix that represents the number of times observations were classified into the correct dataset. A hierarchical clustering was then applied using the proximity scores as dissimilarity measure between observations.

This model was applied with Package 'RandomForest' (v. 4.7-1.1) in R (v. 4.1.3) (Liaw and Wiener, 2002). Because this model is computationally expensive, only 10,000 sites from the VCF file were sampled for its use. The number of trees used was 1000.

4.3.4 Clustering choices

Among the six above-mentioned models, five generate clusters. As the purpose was identifying the three latent classes that correspond to the three genotype categories, the number of clusters in each model was fixed to three. The largest cluster had to correspond to the heterozygous variants, the intermediate cluster to the homozygous variants, and the smallest cluster to the homozygous reference. Also, any model that showed any cluster with < 0.1% of the observations was considered unable to identify three clusters, and therefore not retained. This choice was made according to a prior knowledge about the relatively stable proportions of the three categories in a VCF file of WGS. The ratio of heterozygous variants to homozygous variants in the VCF files is expected to be around 2 (Guo et al., 2014; Wang et al., 2015). The reference sites (i.e.; the false positives for at least one replicate) occupy usually 0.1 to 10% in WGS data (Zhao et al., 2020).

4.3.5 Model result comparisons

Each callset was compared against the GIAB gold standard set. This comparison used the above-provided definitions of TPs, FPs, FNs, and TNs as well as the following performance indicators:

i) Accuracy (or 1 – the overall classification error rate) was calculated as (TPs + TNs) / (TPs + FPs + FNs + TNs); i.e., over the total number of sites in the VCF file;

ii) Recall (or sensitivity) was calculated as TPs / (TPs + FNs);

iii) Precision (or positive predictive value, PPV) was calculated as TPs / (TPs + FPs);

iv) F1-score was calculated as 2 × recall × precision / (recall + precision).

All callsets (except the one generated from the random forest) included all sites in the VCF file. For the random forest callset, the total number of real variants was estimated as the number of variants in the gold standard set multiplied by the sampling proportion.

4.4 Results

4.4.1 Performance indicators for calling results of individual replicates

The precisions relative to the three replicates (1 to 3) had very close values (96.7 to 96.9%) and the sensitivities were nearly the same (~ 98.9%) (Table 4.2). The concordance rates of Replicate 1 vs. Replicates 2 and 3 were 98.4% and 98.3%, respectively; whereas the concordance rate of Replicate 1 vs. Replicate 3 was 98.2%. The concordance rate across the three replicates was 97.5%.

Thus, as expected, the three replicates had similar performance indicators and there were high concordance rates between replicates. However, given the number of total loci in the VCF file (n = 3,351,415), the number of discordant sites across replicates was not negligible (n = 84,753).

Among the concordant sites across the three replicates, precision differed for different genotype categories. For the concordant heterozygous variant sites (n = 1,993,116), the precision was 96.8%. For the concordant homozygous variant sites (n = 1,273,546), the precision was 99.6%. Among the discordant sites, 55.9% were homozygous references, 39.6% heterozygous variants, and 4.5% homozygous variants in the gold standard.

Table 4.2 – Performance indicators of the clustering models under study

Clustering model	Accuracy	Precision	Recall	F1-score
None	96.7 to 96.9%	96.7 to 96.9%	98.9%	97.8 to 97.9%
Majority consensus	97.0%	97.0%	98.9%	97.9%
Latent class analysis without covariables	97.8%	97.9%	98.8%	98.3%
Latent class analysis with covariables	98.0%	98.0%	98.9%	98.4%
Gaussian mixture model	98.5%	99.3%	98.2%	98.7%
Kamila	99.0%	99.2%	98.8%	99.0%
Random Forest	98.2%	99.5%	97.9%	98.7%

4.4.2 Comparison of model fits

In this study, the five types of models used neither the same amount of information nor the same type of covariables: i) the consensus model and the classical latent class model used the categorical variant calling results from the three replicates; ii) the Gaussian mixture model used continuous covariables; iii) the latent class model with covariables, Kamila model, and random forest used categorical variant calling results as well as categorical or continuous covariables. It was therefore difficult to compare directly model fits across model types. This section presents only comparisons within each model type.

With the latent class models with one covariable (AB, QD, or MAPQ), the effect of each covariable was significantly different from 0. The model with AB showed the smallest BIC and was therefore considered as the most fitted to the data.

With the latent class model with two covariables, among the three models relative to the three pairs of covariables, the model with AB and QD had the lowest BIC. Here, it is useful to note that, with some models, the estimations of the parameters of the latent class model with

covariables were not stable. With some models, the global maxima of the log-likelihood were reached in only 10% of estimation attempts. The most frequent local maxima were seldom the global maxima and the estimated proportions of heterozygous variant, homozygous variant, and homozygous reference sites were substantially different between estimation attempts. Therefore, a large number of sets of random initial values (100 rather than 50) were necessary to avoid local maxima. (table 4.3 and table 4.4).

With the Gaussian mixture model, the chosen model (the one with the lowest BIC) was the model with three covariables: DP, AB, and QD.

Table 4.3 -- Stability of the estimates obtained with the latent class analysis model with covariable QualByDepth. The model was fitted 1000 times using 1000 random initial values. The table shows the five most frequent maximum log-likelihood estimations.

Maximum log-likelihood	Number of occurrences	Estimated latent class proportion (%)		
-779028	818	61.626	38.090	0.28382
-447774*	36	59.171	38.197	2.6323
-794167	34	59.078	38.169	2.7520
-802348	28	61.746	38.154	0.099742
-805528	27	61.814	38.185	2.5314×10^{-7}

* Global maximum log-likehood

Table 4.4 -- Stability of the estimates obtained with the latent class analysis model with covariable Allele Balance. The model was fitted 1000 times using 1000 random initial values. The table shows the five most frequent maximum log-likelihood estimations.

Maximum log-likelihood	Number of occurrences	Estimated latent class proportion (%)		
-281366*	241	59.365	38.090	2.5436
-697800	205	61.910	38.090	3.3126×10^{-7}
-671834	189	61.868	37.972	0.15996
-282909	160	59.364	38.059	2.5764
-672082	66	61.831	37.979	0.18954

* Global maximum log-likehood

4.4.3 Performance comparisons

The performance indicators (accuracy, precision, recall, and F1-score) of the models are shown in Table 4.2. Figure 4.1 shows the precision and the recalls of callsets of individual replicates and clustering models. The consensus method improved the precision by 0.1% without much decrease of the recall. Among the five clustering models, the Gaussian mixture model showed the highest accuracy (98.5%). The random forest model showed the highest precision (99.6%) but the lowest recall (98.2%). The consensus model and the latent class model with covariables showed the highest recall (98.9%). The Gaussian mixture model and random forest had high F1-scores (98.7%). Kamila model showed the highest F1-score (99.0%).

The proportions of the three genotype categories in each callset, including the gold standard GIAB benchmark set, are shown in Table 4.5 (Total loci: 3,351,415 in the VCF file). The first row shows the "true" category proportions in the GIAB benchmark set for all sites in the VCF file. More than 4% were classified as reference sites in GIAB set, which corresponds to the marginal false positive rate in the VCF file. Rows 2 to 5 show the proportions in the three replicates and the consensus callset. With the model-based methods (rows 6 to 10), these proportions were the estimated latent-class proportions. The callsets generated by the clustering models grouped more sites into the smallest class (interpreted as reference; thus, false positives) than into the consensus callset; this explains the improved precision of these models. With the Gaussian mixture model, the highest proportion was found in the reference category, which explains its higher precision and lower recall versus the other models.

Table 4.5 - Proportions of the three genotype categories in each callset

	Callset	Homozygous reference	Heterozygous variants	Homozygous variants
1	Gold standard (GIAB)	4.241%	57.891%	37.868%
2	Calling results of Replicate 1	1.064%	60.800%	38.136%
3	Calling results of Replicate 2	1.230%	60.672%	38.098%
4	Calling results of Replicate 3	1.295%	60.618%	38.087%
5	Majority consensus	1.287%	60.586%	38.127%
6	Latent class analysis without covariables	2.283%	59.596%	38.121%
7	Latent class analysis with covariables	2.632%	59.171%	38.197%
8	Gaussian mixture model	4.426%	58.001%	37.573%
9	Kamila	3.586%	58.310%	38.104%
10	Random forest	5.560%	57.440%	37.000%

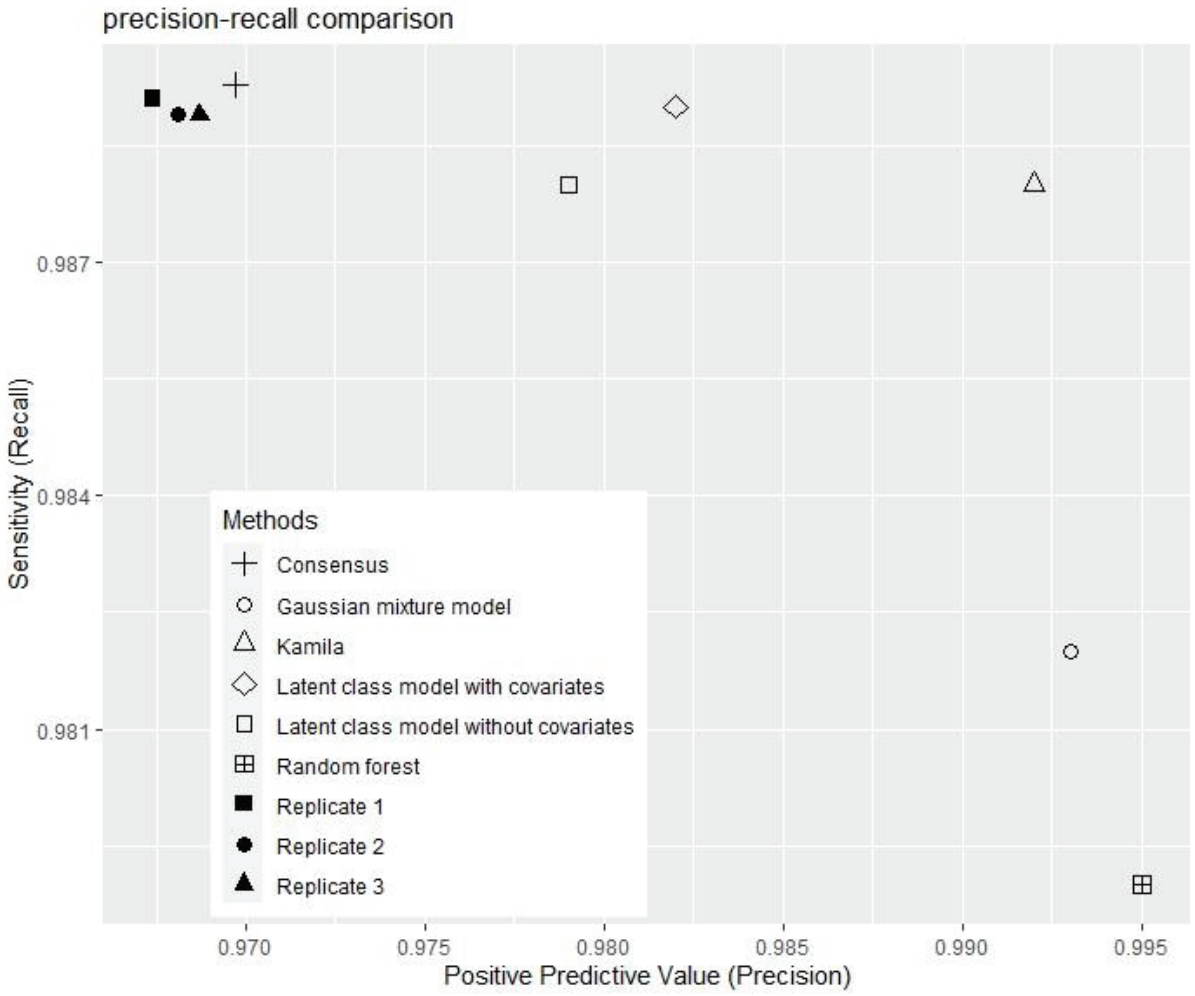

Figure 4.1 - Positive predictive values and sensitivities of callsets without and with selected clustering models

4.5 Discussion

In this study, six clustering algorithms were run on real sequencing replicates of the NA12878 genome to compare their abilities in allowing reconstruction of a new callset with improved performance: one consensus model, two latent-class models, a Gaussian mixture model, a Kamila (adapted k-means) model, and a random forest model. These models showed various advantages. For example, the consensus model improved slightly the precision (by 0.1%) whereas the latent class model provided a non-negligible 1% precision improvement (97% to 98%) without compromising recall (98.9%). In comparison with no use of a clustering model, all six models brought ≥ 1% gain in sensitivity, which is not negligible: i) the Gaussian mixture and the random forest models provided callsets with high precision (> 99%) but at the price of lower recall; ii) Kamila increased precision (99.2%) and kept a high recall (98.8%); it proved having the best overall performance.

In this work, the models were chosen to represent a range of major clustering models, from the most naïve (consensus) to the most sophisticated machine-learning type (random forest). One interest of this choice is that all models may be readily implemented with packages in R software. However, here, only non-supervised clustering models were compared and not supervised ones because the latter need high-quality training data (Sandmann et al., 2018) which are not usually available in clinical practice settings. The models dealt with by BAYSIC and SomaticCombiner or their equivalents were actually considered in the article as latent class model and consensus model, respectively. Indeed, in this work, the former algorithm was not considered because its results would be quite similar to those obtained with a classical latent class model and the latter is based on an approach that is close to the consensus model.

Most of the models considered here have been previously used for similar purposes; i.e., merging several either constitutional or somatic variant calling results to obtain a new callset with better performance indicators (precision or recall). Previous authors used: i) the consensus model (Chiara et al., 2018; Di Nanni et al., 2019; Hwang et al., 2014; Trubetskoy et al., 2015); ii) the Bayesian latent class model (Cantarel et al., 2014); iii) the Gaussian mixture model (DePristo et al., 2011; Hwang et al., 2019); iv) random forest (Huang et al., 2019; Wang et al., 2020). However, though usual, these models have been rarely compared, their comparison results often unclear, and the final conclusions controversial. For example,

the random-forest-based ensemble caller for somatic mutation has obtained higher F1-scores than the simple consensus approach (Huang et al., 2019); however, in a study by Wang et al. (Wang et al., 2020), the authors observed that the consensus method was more robust and stable than supervised machine-learning models. They suggested that the difference between the training data and the test data contributed to the poor generalizability of machine-leaning models. In another research on the NA12878 genome that used the GIAB benchmark set as gold standard, a two-component mixture-model-based method that considered results from 70 pipelines did not significantly improve performance in terms of precision at the highest analytical sensitivity achievable vs. the highest performance of a single pipeline. However, the method led to performance improvement with another gold standard set from the '1000 Genomes Project' (Hwang et al., 2019).

The models compared here did not include the same number of variables because of the hypotheses inherent to each model. Some require only continuous variables (e.g., the Gaussian mixture model), whereas others require only categorical variables (e.g., the latent class model). Thus, performance comparisons between new callsets generated by different models should be interpreted with this difference in mind. For example, Kamila and random forest models are able to include more covariables than the other models. In future works, comparisons between models with same covariables would be welcome. One current aim was to use information already available in a VCF file; however, the possibility of including more covariables may be interesting too.

In some previous research works, sites in the VCF file of presumably very low quality were filtered out before applying merging methods; i.e., a small number of sites were considered as false positives and thus excluded (Sandmann et al., 2018). Here, no sites were filtered out (all sites from the VCF file were included in the models); this allowed a more objective evaluation of the overall performance of each model. However, this choice introduced some difficulties due to the extreme values of certain variables. For example, DP has typically a long-tailed distribution and the presence of extremely high values is often an indicator of sequencing artefacts, alignment artefacts, or copy number variations (Guo et al., 2014; Li, 2014; O'Rawe et al., 2013). In common practice, the solution to extreme DP values is to exclude sites with values higher than a threshold defined according to various formulas that use the mean and standard deviation of DPs (Li et al., 2018; Pan et al., 2022); for example, a threshold 120 in the hard filters recommended by the GATK (Van der Auwera and O'Connor, 2020).

In the present work, the mean DP across the three replicates was circa 38 and its maximum 13,858 and, among the compared models, Kamila is known to be relatively sensitive to extreme values because it minimizes a dissimilarity measure that is partially based on Euclidean distance in the case of continuous variables. This might explain why it failed to identify the three clusters with acceptable proportions. Indeed, the model grouped a small number of sites with extremely high DP values into one cluster (n = 254; i.e., 0.008% of all sites) and, as stated in 4.3.4, models that led to any cluster with < 0.1% of the sites were considered unable to identify three clusters and thus not retained. One way to address this issue is to add one more cluster in the model (4 instead of 3). However, in this work, only three clusters were considered to allow model performance comparisons and allow each cluster to represent each genotype category. Therefore, with Kamila, the maximum DP value was set at 150 and higher values grouped together at 150. The other models that involved DP (i.e., the Gaussian mixture model and the random forest model) performed well despite the presence of high DP values (these were not then filtered out).

This study focused on the VCF file (i.e., on all sites called as variants in at least one replicate) and not on all three billion bp positions across the human genome. This is one reason for which the indicators of performance kept were only recall and precision (specificity was ignored). There are also two other practical reasons: i) negative sites are much more numerous (almost 1000 times the number of sites in the VCF file) and contain less information; thus, using them is computationally expensive and adds little information; ii) researchers, especially practitioners and lab professionals, usually use only the VCF file for routine analyses; thus, a model that requires information from the BAM file for sites called as 'reference' would not be practical.

One limitation of this study is that it evaluated only callsets' performance regarding SNVs. Further studies are worth being conducted to evaluate the performance of clustering models regarding copy number variations and structural variations. Also, except for Kamila, the study included only the most classical model from each clustering algorithm type. Some model features may prove more adapted to the distribution of the variables or have more convenient underlying hypotheses. For example, latent class models that relax the conditional independence between observed variables through correlation, random effects, or covariables with effects on the class-conditional probabilities.

The Gaussian mixture model used here showed good performance vs. the other five models. However, all components of a variable distribution might not be Gaussian. For example, i) the distribution of allele balance has been already modelled using a mixture of 0-inflated beta distribution, binomial distribution, and 1-inflated beta distribution for the homozygous reference, heterozygous variant, and homozygous reference categories, respectively (Muyas et al., 2019); ii) to take into account heavy-tails, read depth distributions have been modelled using a compound Poisson distribution, a negative binomial distribution, or a log-normal distribution (Daley and Smith, 2014; Deng et al., 2020; Robinson et al., 2010).

From a theoretical viewpoint, a very recent article by Dang et al. (Dang et al., 2023) reviewed a selection of "mixture models that can deal with varying cluster tail-weight, skewness and/or concentration, and kurtosis" (e.g., mixtures of multivariate t-distributions, mixtures of skew-t distributions, mixtures of normal inverse Gaussian distributions, etc.). Furthermore, these authors introduced a multivariate skewed power exponential distribution that "allow for robust mixture models for clustering with skewed or symmetric components" and "model components with varying levels of peakedness, skewness, and tail-weight (light, heavy, Gaussian)". In practice, the use of multivariate non-Gaussian mixture models is often difficult because of identifiability issues and the instability of parameter estimation. This might explain the rarity of applications on real data, which is worth being explored. We especially hope an exploration of the appropriateness of the above-mentioned models within the context of WGS data.

4.6 Conclusions

In this study, several clustering models were evaluated within the context of combining callsets from DNA sequencing replicates. These non-supervised clustering models proved able to improve sequencing performance in terms of precision and F1-score, which is comparable to what is reported about supervised models. Among the models compared here, the Gaussian mixture model and Kamila offered improvements that made precision higher than 99% and F1-score close to 99%. These models may then be recommended to reconstruct new high-performance callsets from NGS replicates. This is of particular interest for diagnosis or precision medicine whenever DNA sequencing results stem from either biological replicates (more than one sample) or technological replicates (more than one sequencing platform or analysis pipeline).

References

Adelson, R.P., Renton, A.E., Li, W., Barzilai, N., Atzmon, G., Goate, A.M., Davies, P., Freudenberg-Hua, Y., 2019. Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. Sci Rep 9, 16156. https://doi.org/10.1038/s41598-019-52614-7

Ajay, S.S., Parker, S.C.J., Ozel Abaan, H., Fuentes Fajardo, K.V., Margulies, E.H., 2011. Accurate and comprehensive sequencing of personal genomes. Genome Res. 21, 1498–1505. https://doi.org/10.1101/gr.123638.111

Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. Proceedings of the 2nd International Symposium on Information Theory 267–281.

Amemiya, H.M., Kundaje, A., Boyle, A.P., 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep 9, 9354. https://doi.org/10.1038/s41598-019-45839-z

Atkinson, E.G., Artomov, M., Karczewski, K.J., Loboda, A.A., Rehm, H.L., MacArthur, D.G., Neale, B.M., Daly, M.J., 2022. Discordant genotype calls across technology platforms elucidate variants with systematic errors in next-generation sequencing (preprint). Genomics. https://doi.org/10.1101/2022.03.24.485707

Baeza-Yates, R.A., Perleberg, C.H., 1996. Fast and practical approximate string matching. Information Processing Letters 59, 21–27. https://doi.org/10.1016/0020-0190(96)00083-X

Bainbridge, M.N., Wang, M., Wu, Y., Newsham, I., Muzny, D.M., Jefferies, J.L., Albert, T.J., Burgess, D.L., Gibbs, R.A., 2011. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol 12, R68. https://doi.org/10.1186/gb-2011-12-7-r68

Banfield, J.D., Raftery, A.E., 1993. Model-Based Gaussian and Non-Gaussian Clustering. Biometrics 49, 803. https://doi.org/10.2307/2532201

Bauer, P., Kandaswamy, K.K., Weiss, M.E.R., Paknia, O., Werber, M., Bertoli-Avella, A.M., Yüksel, Z., Bochinska, M., Oprea, G.E., Kishore, S., Weckesser, V., Karges, E., Rolfs, A., 2019. Development of an evidence-based algorithm that optimizes sensitivity and specificity in ES-based diagnostics of a clinically heterogeneous patient population. Genetics in Medicine 21, 53–61. https://doi.org/10.1038/s41436-018-0016-6

Beck, T.F., Mullikin, J.C., the NISC Comparative Sequencing Program, Biesecker, L.G., 2016. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clinical Chemistry 62, 647–654. https://doi.org/10.1373/clinchem.2015.249623

Benjamini, Y., Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research 40, e72–e72. https://doi.org/10.1093/nar/gks001

Cabanski, C.R., Cavin, K., Bizon, C., Wilkerson, M.D., Parker, J.S., Wilhelmsen, K.C., Perou, C.M., Marron, J., Hayes, D.N., 2012. ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data. BMC Bioinformatics 13, 221. https://doi.org/10.1186/1471-2105-13-221

Cantarel, B.L., Weaver, D., McNeill, N., Zhang, J., Mackey, A.J., Reese, J., 2014. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics 15, 104. https://doi.org/10.1186/1471-2105-15-104

Chen, J., Li, X., Zhong, H., Meng, Y., Du, H., 2019. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep 9, 9345. https://doi.org/10.1038/s41598-019-45835-3

Chiara, M., Gioiosa, S., Chillemi, G., D'Antonio, M., Flati, T., Picardi, E., Zambelli, F., Horner, D.S., Pesole, G., Castrignanò, T., 2018. CoVaCS: a consensus variant calling system. BMC Genomics 19, 120. https://doi.org/10.1186/s12864-018-4508-1

Cornish, A., Guda, C., 2015. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. BioMed Research International 2015, 1–11. https://doi.org/10.1155/2015/456479

Daley, T., Smith, A.D., 2014. Modeling genome coverage in single-cell sequencing. Bioinformatics 30, 3159–3165. https://doi.org/10.1093/bioinformatics/btu540

Dang, U.J., Browne, R.P., McNicholas, P.D., 2015. Mixtures of Multivariate Power Exponential Distributions.

Dang, U.J., Gallaugher, M.P.B., Browne, R.P., McNicholas, P.D., 2023. Model-Based Clustering and Classification Using Mixtures of Multivariate Skewed Power Exponential Distributions. J Classif. https://doi.org/10.1007/s00357-022-09427-7

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1).

Deng, C., Daley, T., Calabrese, P., Ren, J., Smith, A.D., 2020. Predicting the Number of Bases to Attain Sufficient Coverage in High-Throughput Sequencing Experiments. Journal of Computational Biology 27, 1130–1143. https://doi.org/10.1089/cmb.2019.0264

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–498. https://doi.org/10.1038/ng.806

Di Nanni, N., Moscatelli, M., Gnocchi, M., Milanesi, L., Mosca, E., 2019. isma: an R package for the integrative analysis of mutations detected by multiple pipelines. BMC Bioinformatics 20, 107. https://doi.org/10.1186/s12859-019-2701-0

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511790492

Dziak, J.J., Coffman, D.L., Lanza, S.T., Li, R., Jermiin, L.S., 2020. Sensitivity and specificity of information criteria. Briefings in Bioinformatics 21, 553–565. https://doi.org/10.1093/bib/bbz016

Ebbert, M.T.W., Wadsworth, M.E., Staley, L.A., Hoyt, K.L., Pickett, B., Miller, J., Duce, J., Kauwe, J.S.K., Ridge, P.G., 2016. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics 17, 239. https://doi.org/10.1186/s12859-016-1097-3

Eberle, M.A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B.L., Bekritsky, M.A., Iqbal, Z., Chuang, H.-Y., Humphray, S.J., Halpern, A.L., Kruglyak, S., Margulies, E.H., McVean, G., Bentley, D.R., 2017. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 27, 157–164. https://doi.org/10.1101/gr.210500.116

Ewing, B., Green, P., 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res. 8, 186–194. https://doi.org/10.1101/gr.8.3.186

Faes, C., Geys, H., Molenberghs, G., Aerts, M., Cadarso-Suárez, C., Acuña, C., Cano, M., 2008. A Flexible Method to Measure Synchrony in Neuronal Firing. Journal of the American Statistical Association 103, 149–161. https://doi.org/10.1198/016214507000000419

Foss, A., Markatou, M., Ray, B., Heching, A., 2016. A semiparametric method for clustering mixed data. Mach Learn 105, 419–458. https://doi.org/10.1007/s10994-016-5575-7

Foss, A., Markatou, M., 2018. kamila : Clustering Mixed-Type Data in R and Hadoop. J. Stat. Soft. 83. https://doi.org/10.18637/jss.v083.i13

Fraser, C.G., 2001. Biological variation: from principles to practice. AACC Press, Washington, DC.

Friedman, S., Gauthier, L., Farjoun, Y., Banks, E., 2020. Lean and deep models for more accurate filtering of SNP and INDEL variant calls. Bioinformatics 36, 2060–2067. https://doi.org/10.1093/bioinformatics/btz901

Gézsi, A., Bolgár, B., Marx, P., Sarkozy, P., Szalai, C., Antal, P., 2015. VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. BMC Genomics 16, 875. https://doi.org/10.1186/s12864-015-2050-y

Goldfeder, R.L., Priest, J.R., Zook, J.M., Grove, M.E., Waggott, D., Wheeler, M.T., Salit, M., Ashley, E.A., 2016. Medical implications of technical accuracy in genome sequencing. Genome Med 8, 24. https://doi.org/10.1186/s13073-016-0269-0

Goodman, L.A., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61, 215–231. https://doi.org/10.1093/biomet/61.2.215

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351. https://doi.org/10.1038/nrg.2016.49

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E., Sigurdsson, G.T., Stacey, S.N., Frigge, M.L., Holm, H., Saemundsdottir, J., Helgadottir, H.T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., Sverrisson, J.T., Gretarsdottir, S., Walters, G.B., Rafnar, T., Thjodleifsson, B., Bjornsson, E.S., Olafsson, S., Thorarinsdottir, H., Steingrimsdottir, T., Gudmundsdottir, T.S., Theodors, A., Jonasson, J.G., Sigurdsson, A., Bjornsdottir, G., Jonsson, J.J., Thorarensen, O., Ludvigsson, P., Gudbjartsson, H., Eyjolfsson, G.I., Sigurdardottir, O., Olafsson, I., Arnar, D.O., Magnusson, O.T., Kong, A., Masson, G., Thorsteinsdottir, U., Helgason, A., Sulem, P., Stefansson, K., 2015. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 47, 435–444. https://doi.org/10.1038/ng.3247

Guha, S., Rastogi, R., Shim, K., 2001. Cure: an efficient clustering algorithm for large databases. Information Systems 26, 35–58. https://doi.org/10.1016/S0306-4379(01)00008-4

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D.C., Shyr, Y., 2012. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics 13, 666. https://doi.org/10.1186/1471-2164-13-666

Guo, Y., Ye, F., Sheng, Q., Clark, T., Samuels, D.C., 2014. Three-stage quality control strategies for DNA re-sequencing data. Briefings in Bioinformatics 15, 879–889. https://doi.org/10.1093/bib/bbt069

Hardwick, S.A., Deveson, I.W., Mercer, T.R., 2017. Reference standards for next-generation sequencing. Nat Rev Genet 18, 473–484. https://doi.org/10.1038/nrg.2017.44

Highnam, G., Wang, J.J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N., Mittelman, D., 2015. An analytical framework for optimizing variant discovery from personal genomes. Nat Commun 6, 6275. https://doi.org/10.1038/ncomms7275

Hofmann, A.L., Behr, J., Singer, J., Kuipers, J., Beisel, C., Schraml, P., Moch, H., Beerenwinkel, N., 2017. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. BMC Bioinformatics 18, 8. https://doi.org/10.1186/s12859-016-1417-7

Huang, J.Z., Ng, M.K., Hongqiang Rong, Zichen Li, 2005. Automated variable weighting in k-means type clustering. IEEE Trans. Pattern Anal. Machine Intell. 27, 657–668. https://doi.org/10.1109/TPAMI.2005.95

Huang, L., Wang, C., Rosenberg, N.A., 2009. The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations. The American Journal of Human Genetics 85, 692–698. https://doi.org/10.1016/j.ajhg.2009.09.017

Huang, W., Guo, Y.A., Muthukumar, K., Baruah, P., Chang, M.M., Jacobsen Skanderup, A., 2019. SMuRF: portable and accurate ensemble prediction of somatic mutations. Bioinformatics 35, 3157–3159. https://doi.org/10.1093/bioinformatics/btz018

Huang, Z., 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283–304. https://doi.org/10.1023/A:1009769707641

Hwang, K.-B., Lee, I.-H., Li, H., Won, D.-G., Hernandez-Ferrer, C., Negron, J.A., Kong, S.W., 2019. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Sci Rep 9, 3219. https://doi.org/10.1038/s41598-019-39108-2

Hwang, K.-B., Lee, I.-H., Park, J.-H., Hambuch, T., Choe, Y., Kim, M., Lee, K., Song, T., Neu, M.B., Gupta, N., Kohane, I.S., Green, R.C., Kong, S.W., 2014. Reducing False-Positive Incidental Findings with Ensemble Genotyping and Logistic Regression Based Variant Filtering Methods. Human Mutation 35, 936–944. https://doi.org/10.1002/humu.22587

Illumina, 2017. NovaSeq™ 6000 System Quality Scores and RTA3 Software. https://emea.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/novaseq-hiseq-q30-app-note-770-2017-010.pdf.

Illumina, 2014. Understanding Illumina Quality Scores. https://emea.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf.

James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), 2013. An introduction to statistical learning: with applications in R, Springer texts in statistics. Springer, New York.

Kasa, S.R., Rajan, V., 2022. Improved Inference of Gaussian Mixture Copula Model for Clustering and Reproducibility Analysis using Automatic Differentiation. Econometrics and Statistics 22, 67–97. https://doi.org/10.1016/j.ecosta.2021.08.010

Kaufman, L., Rousseeuw, P.J. (Eds.), 1990. Finding Groups in Data, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/9780470316801

Kircher, M., Kelso, J., 2010. High-throughput DNA sequencing - concepts and limitations. Bioessays 32, 524–536. https://doi.org/10.1002/bies.200900181

Kircher, M., Stenzel, U., Kelso, J., 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10, R83. https://doi.org/10.1186/gb-2009-10-8-r83

Krishnan, V., Utiramerur, S., Ng, Z., Datta, S., Snyder, M.P., Ashley, E.A., 2021. Benchmarking workflows to assess performance and suitability of germline variant calling pipelines in clinical diagnostic assays. BMC Bioinformatics 22, 85. https://doi.org/10.1186/s12859-020-03934-3

Krol, A., 2015. Genome in a Bottle Uncapped.

Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., De La Vega, F.M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B.A., Salit, M., Zook, J.M., 2019. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol 37, 555–560. https://doi.org/10.1038/s41587-019-0054-x

Laehnemann, D., Borkhardt, A., McHardy, A.C., 2016. Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. Brief Bioinform 17, 154–179. https://doi.org/10.1093/bib/bbv029

Lam, H.Y.K., Clark, M.J., Chen, Rui, Chen, Rong, Natsoulis, G., O'Huallachain, M., Dewey, F.E., Habegger, L., Ashley, E.A., Gerstein, M.B., Butte, A.J., Ji, H.P., Snyder, M., 2012. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30, 78–82. https://doi.org/10.1038/nbt.2065

Lee, S.X., McLachlan, G.J., 2013. On mixtures of skew normal and skew t -distributions. Adv Data Anal Classif 7, 241–266. https://doi.org/10.1007/s11634-013-0132-8

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., Tukiainen, T., Birnbaum, D.P., Kosmicki, J.A., Duncan, L.E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D.N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M.I., Moonshine, A.L., Natarajan, P., Orozco, L., Peloso, G.M., Poplin, R., Rivas, M.A., Ruano-Rubio, V., Rose, S.A., Ruderfer, D.M., Shakir, K., Stenson, P.D., Stevens, C., Thomas, B.P., Tiao, G., Tusie-Luna, M.T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D.M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J.C., Gabriel, S.B., Getz, G., Glatt, S.J., Hultman, C.M., Kathiresan, S.,

Laakso, M., McCarroll, S., McCarthy, M.I., McGovern, D., McPherson, R., Neale, B.M., Palotie, A., Purcell, S.M., Saleheen, D., Scharf, J.M., Sklar, P., Sullivan, P.F., Tuomilehto, J., Tsuang, M.T., Watkins, H.C., Wilson, J.G., Daly, M.J., MacArthur, D.G., 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291. https://doi.org/10.1038/nature19057

Li, H., 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30, 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio].

Li, H., 2010. Mathematical Notes on SAMtools Algorithms.

Li, H., Bloom, J.M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., MacArthur, D., 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat Methods 15, 595–597. https://doi.org/10.1038/s41592-018-0054-7

Li, H., Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics 11, 473–483. https://doi.org/10.1093/bib/bbq015

Li, H., Ruan, J., Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18, 1851–1858. https://doi.org/10.1101/gr.078212.108

Li, J., Ray, S., Lindsay, B.G., 2007. A Nonparametric Statistical Approach to Clustering via Mode Identification. Journal of Machine Learning Research 8, 1687–1723.

Li, R., Li, Y., Fang, X., Yang, H., Wang, Jian, Kristiansen, K., Wang, Jun, 2009. SNP detection for massively parallel whole-genome resequencing. Genome Res. 19, 1124–1132. https://doi.org/10.1101/gr.088013.108

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Linderman, M.D., Brandt, T., Edelmann, L., Jabado, O., Kasai, Y., Kornreich, R., Mahajan, M., Shah, H., Kasarskis, A., Schadt, E.E., 2014. Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med Genomics 7, 20. https://doi.org/10.1186/1755-8794-7-20

Linzer, D.A., Lewis, J.B., 2011. poLCA : An R Package for Polytomous Variable Latent Class Analysis. J. Stat. Soft. 42. https://doi.org/10.18637/jss.v042.i10

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110

McLachlan, G., Peel, D., 2000. Finite Mixture Models: McLachlan/Finite Mixture Models, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/0471721182

McLachlan, G.J., Lee, S.X., Rathnayake, S.I., 2019. Finite Mixture Models. Annu. Rev. Stat. Appl. 6, 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

McParland, D., Gormley, I.C., 2016. Model based clustering for mixed data: clustMD. Adv Data Anal Classif 10, 155–169. https://doi.org/10.1007/s11634-016-0238-x

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat Rev Genet 11, 31–46. https://doi.org/10.1038/nrg2626

Meynert, A.M., Bicknell, L.S., Hurles, M.E., Jackson, A.P., Taylor, M.S., 2013. Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics 14, 195. https://doi.org/10.1186/1471-2105-14-195

Minoche, A.E., Dohm, J.C., Himmelbauer, H., 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol 12, R112. https://doi.org/10.1186/gb-2011-12-11-r112

Modha, D.S., Spangler, W., 2003. Feature Weighting in k-Means Clustering. Machine Learning 52, 217–237. https://doi.org/10.1023/A:1024016609528

Monach, P.A., 2012. Repeating tests: different roles in research studies and clinical medicine. Biomark Med 6, 691–703. https://doi.org/10.2217/bmm.12.57

Muyas, F., Bosio, M., Puig, A., Susak, H., Domènech, L., Escaramis, G., Zapata, L., Demidov, G., Estivill, X., Rabionet, R., Ossowski, S., 2019. Allele balance bias identifies systematic genotyping errors and false disease associations. Human Mutation 40, 115–126. https://doi.org/10.1002/humu.23674

Muzzey, D., Evans, E.A., Lieber, C., 2015. Understanding the Basics of NGS: From Mechanism to Variant Calling. Curr Genet Med Rep 3, 158–165. https://doi.org/10.1007/s40142-015-0076-8

Naj, A.C., Lin, H., Vardarajan, B.N., White, S., Lancour, D., Ma, Y., Schmidt, M., Sun, F., Butkiewicz, M., Bush, W.S., Kunkle, B.W., Malamon, J., Amin, N., Choi, S.H., Hamilton-Nelson, K.L., van der Lee, S.J., Gupta, N., Koboldt, D.C., Saad, M., Wang, B., Nato, A.Q., Sohi, H.K., Kuzma, A., Wang, L.-S., Cupples, L.A., van Duijn, C., Seshadri, S., Schellenberg, G.D., Boerwinkle, E., Bis, J.C., Dupuis, J., Salerno, W.J., Wijsman, E.M., Martin, E.R., DeStefano, A.L., 2019. Quality control and integration of genotypes from two calling pipelines for whole genome sequence data in the Alzheimer's disease sequencing project. Genomics 111, 808–818. https://doi.org/10.1016/j.ygeno.2018.05.004

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, Md., Ogasawara, N., Kanaya, S., 2011. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research 39, e90–e90. https://doi.org/10.1093/nar/gkr344

Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12, 443–451. https://doi.org/10.1038/nrg2986

Oberski, D., 2016. Mixture Models: Latent Profile and Latent Class Analysis, in: Robertson, J., Kaptein, M. (Eds.), Modern Statistical Methods for HCI. Springer International Publishing, Cham, pp. 275–287. https://doi.org/10.1007/978-3-319-26633-6_12

O'Fallon, B.D., Wooderchak-Donahue, W., Crockett, D.K., 2013. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. Bioinformatics 29, 1361–1366. https://doi.org/10.1093/bioinformatics/btt172

Oliver, G.R., Hart, S.N., Klee, E.W., 2015. Bioinformatics for Clinical Next Generation Sequencing. Clinical Chemistry 61, 124–135. https://doi.org/10.1373/clinchem.2014.224360

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., Wei, Z., Wang, K., Lyon, G.J., 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5, 28. https://doi.org/10.1186/gm432

Pan, B., Ren, L., Onuchic, V., Guan, M., Kusko, R., Bruinsma, S., Trigg, L., Scherer, A., Ning, B., Zhang, C., Glidewell-Kenney, C., Xiao, C., Donaldson, E., Sedlazeck, F.J., Schroth, G., Yavas, G., Grunenwald, H., Chen, H., Meinholz, H., Meehan, J., Wang, J., Yang, J., Foox, J., Shang, J., Miclaus, K., Dong, L., Shi, L., Mohiyuddin, M., Pirooznia, M., Gong, P., Golshani, R., Wolfinger, R., Lababidi, S., Sahraeian, S.M.E., Sherry, S., Han, T., Chen, T., Shi, T., Hou, W., Ge, W., Zou, W., Guo, W., Bao, W., Xiao, Wenzhong, Fan, X., Gondo, Y., Yu, Y., Zhao, Y., Su, Z., Liu, Z., Tong, W., Xiao, Wenming, Zook, J.M., Zheng, Y., Hong, H., 2022. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. Genome Biol 23, 2. https://doi.org/10.1186/s13059-021-02569-8

Patch, A.-M., Nones, K., Kazakoff, S.H., Newell, F., Wood, S., Leonard, C., Holmes, O., Xu, Q., Addala, V., Creaney, J., Robinson, B.W., Fu, S., Geng, C., Li, T., Zhang, W., Liang, X., Rao, J., Wang, J., Tian, M., Zhao, Y., Teng, F., Gou, H., Yang, B., Jiang, H., Mu, F., Pearson, J.V., Waddell, N., 2018. Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. PLoS ONE 13, e0190264. https://doi.org/10.1371/journal.pone.0190264

Pfeifer, S.P., 2017. From next-generation resequencing reads to a high-quality variant data set. Heredity 118, 111–124. https://doi.org/10.1038/hdy.2016.102

Popitsch, N., WGS500 Consortium, Schuh, A., Taylor, J.C., 2017. ReliableGenome: annotation of genomic regions with high/low variant calling concordance. Bioinformatics 33, 155–160. https://doi.org/10.1093/bioinformatics/btw587

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M.J., Neale, B., MacArthur, D.G., Banks, E., 2017. Scaling accurate genetic variant discovery to tens of thousands of samples (preprint). Genomics. https://doi.org/10.1101/201178

Preud'homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., Couceiro, M., Devignes, M.-D., Kobayashi, M., Huttin, O., Ferreira, J.P., Zannad, F., Rossignol, P., Girerd, N., 2021. Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Sci Rep 11, 4202. https://doi.org/10.1038/s41598-021-83340-8

Ratan, A., Miller, W., Guillory, J., Stinson, J., Seshagiri, S., Schuster, S.C., 2013. Comparison of Sequencing Platforms for Single Nucleotide Variant Calls in a Human Sample. PLoS ONE 8, e55089. https://doi.org/10.1371/journal.pone.0055089

Reinert, K., Langmead, B., Weese, D., Evers, D.J., 2015. Alignment of Next-Generation Sequencing Reads. Annu. Rev. Genom. Hum. Genet. 16, 133–151. https://doi.org/10.1146/annurev-genom-090413-025358

Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., Sabbe, B., Despierre, E., Vergote, I., Hilbush, B., Lambrechts, D., Del-Favero, J., 2012. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat Biotechnol 30, 61–68. https://doi.org/10.1038/nbt.2053

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., Jäger, N., Kool, M., Taylor, M., Lichter, P., Pfister, S., Wolf, S., Brors, B., Eils, R., 2013. Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. PLoS ONE 8, e66621. https://doi.org/10.1371/journal.pone.0066621

Robasky, K., Lewis, N.E., Church, G.M., 2014. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 15, 56–62. https://doi.org/10.1038/nrg3655

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, T., Harkin, J., Shukla, P., 2021. Hardware acceleration of genomics data analysis: challenges and opportunities. Bioinformatics 37, 1785–1795. https://doi.org/10.1093/bioinformatics/btab017

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., Jaffe, D.B., 2013. Characterizing and measuring bias in sequence data. Genome Biol 14, R51. https://doi.org/10.1186/gb-2013-14-5-r51

Sandmann, S., Karimi, M., de Graaf, A.O., Rohde, C., Göllner, S., Varghese, J., Ernsting, J., Walldin, G., van der Reijden, B.A., Müller-Tidow, C., Malcovati, L., Hellström-Lindberg, E., Jansen, J.H., Dugas, M., 2018. appreci8: a pipeline for precise variant calling integrating 8 tools. Bioinformatics 34, 4205–4212. https://doi.org/10.1093/bioinformatics/bty518

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.-T., 2017. A review of clustering techniques and developments. Neurocomputing 267, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Statist. 6. https://doi.org/10.1214/aos/1176344136

Scrucca, L., Fop, M., Murphy, T., Brendan, Raftery, A., E., 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. The R Journal 8, 289. https://doi.org/10.32614/RJ-2016-021

Seo, J.-S., Rhie, A., Kim, Junsoo, Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, Jihye, Kuk, J., Park, G.H., Kim, Juhyeok, Ryu, H., Kim, Jongbum, Roh, M., Baek, J., Hunkapiller, M.W., Korlach, J., Shin, J.-Y., Kim, C., 2016. De novo assembly and phasing of a Korean human genome. Nature 538, 243–247. https://doi.org/10.1038/nature20098

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat Biotechnol 26, 1135–1145. https://doi.org/10.1038/nbt1486

Shi, T., Horvath, S., 2006. Unsupervised Learning With Random Forest Predictors. Journal of Computational and Graphical Statistics 15, 118–138. https://doi.org/10.1198/106186006X94072

Shoemaker, J.S., Painter, I.S., Weir, B.S., 1999. Bayesian statistics in genetics: a guide for the uninitiated. Trends in Genetics 15, 354–358. https://doi.org/10.1016/S0168-9525(99)01751-5

Shringarpure, S.S., Mathias, R.A., Hernandez, R.D., O'Connor, T.D., Szpiech, Z.A., Torres, R., De La Vega, F.M., Bustamante, C.D., Barnes, K.C., Taub, M.A., 2016. Using genotype array data to compare multi- and single-sample variant calls and improve

variant call sets from deep coverage whole-genome sequencing data. Bioinformatics btw786. https://doi.org/10.1093/bioinformatics/btw786

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Supernat, A., Vidarsson, O.V., Steen, V.M., Stokowy, T., 2018. Comparison of three variant callers for human whole genome sequencing. Sci Rep 8, 17851. https://doi.org/10.1038/s41598-018-36177-7

Treangen, T.J., Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13, 36–46. https://doi.org/10.1038/nrg3117

Trubetskoy, V., Rodriguez, A., Dave, U., Campbell, N., Crawford, E.L., Cook, E.H., Sutcliffe, J.S., Foster, I., Madduri, R., Cox, N.J., Davis, L.K., 2015. Consensus Genotyper for Exome Sequencing (CGES): improving the quality of exome variant genotypes. Bioinformatics 31, 187–193. https://doi.org/10.1093/bioinformatics/btu591

Van der Auwera, G.A.V. de, O'Connor, B.D., 2020. Genomics in the cloud: using Docker, GATK, and WDL in Terra, First edition. ed. O'Reilly, Beijing Boston Farnham Sebastopol Tokyo.

Wagner, J., Olson, N.D., Harris, L., Khan, Z., Farek, J., Mahmoud, M., Stankovic, A., Kovacevic, V., Yoo, B., Miller, N., Rosenfeld, J.A., Ni, B., Zarate, S., Kirsche, M., Aganezov, S., Schatz, M.C., Narzisi, G., Byrska-Bishop, M., Clarke, W., Evani, U.S., Markello, C., Shafin, K., Zhou, X., Sidow, A., Bansal, V., Ebert, P., Marschall, T., Lansdorp, P., Hanlon, V., Mattsson, C.-A., Barrio, A.M., Fiddes, I.T., Xiao, C., Fungtammasan, A., Chin, C.-S., Wenger, A.M., Rowell, W.J., Sedlazeck, F.J., Carroll, A., Salit, M., Zook, J.M., 2022. Benchmarking challenging small variants with linked and long reads. Cell Genomics 2, 100128. https://doi.org/10.1016/j.xgen.2022.100128

Wall, J.D., Tang, L.F., Zerbe, B., Kvale, M.N., Kwok, P.-Y., Schaefer, C., Risch, N., 2014. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res. 24, 1734–1739. https://doi.org/10.1101/gr.168393.113

Wang, J., Raskin, L., Samuels, D.C., Shyr, Y., Guo, Y., 2015. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics 31, 318–323. https://doi.org/10.1093/bioinformatics/btu668

Wang, M., Luo, W., Jones, K., Bian, X., Williams, R., Higson, H., Wu, D., Hicks, B., Yeager, M., Zhu, B., 2020. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. Sci Rep 10, 12898. https://doi.org/10.1038/s41598-020-69772-8

Wood, S.N., Goude, Y., Shaw, S., 2015. Generalized additive models for large data sets. J. R. Stat. Soc. C 64, 139–155. https://doi.org/10.1111/rssc.12068

Wood, S.N., Li, Z., Shaddick, G., Augustin, N.H., 2017. Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data. Journal of the American Statistical Association 112, 1199–1210. https://doi.org/10.1080/01621459.2016.1195744

Zhang, P., Luo, H., Li, Y., Wang, Y., Wang, J., Zheng, Y., Niu, Y., Shi, Y., Zhou, H., Song, T., Kang, Q., Xu, T., He, S., 2021. NyuWa Genome resource: A deep whole-genome

sequencing-based variation profile and reference panel for the Chinese population. Cell Reports 37, 110017. https://doi.org/10.1016/j.celrep.2021.110017

Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: an efficient data clustering method for very large databases. SIGMOD Rec. 25, 103–114. https://doi.org/10.1145/235968.233324

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., Hovig, E., 2020. Accuracy and efficiency of germline variant calling pipelines for human genome data. Sci Rep 10, 20222. https://doi.org/10.1038/s41598-020-77218-4

Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., Henaff, E., McIntyre, A.B.R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Saghbini, M., Dzakula, Z., Hastie, A., Cao, H., Deikus, G., Schadt, E., Sebra, R., Bashir, A., Truty, R.M., Chang, C.C., Gulbahce, N., Zhao, K., Ghosh, S., Hyland, F., Fu, Y., Chaisson, M., Xiao, C., Trow, J., Sherry, S.T., Zaranek, A.W., Ball, M., Bobe, J., Estep, P., Church, G.M., Marks, P., Kyriazopoulou-Panagiotopoulou, S., Zheng, G.X.Y., Schnall-Levin, M., Ordonez, H.S., Mudivarti, P.A., Giorda, K., Sheng, Y., Rypdal, K.B., Salit, M., 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3, 160025. https://doi.org/10.1038/sdata.2016.25

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., Salit, M., 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32, 246–251. https://doi.org/10.1038/nbt.2835

List of figures

List of tables

Annex: Communications and publications

Oral communications:

July 2021 Evaluating DNA sequencing performance: concordance-discordance model and latent class model.

42nd Annual Conference of the International Society for Biostatistics (ISCB), Lyon, FR

May 2023 Comparaison des performances des modèles de partitionnement dans la reconstitution des résultats de séquençage ADN à partir de réplicats techniques

17ème conférence francophone d'EPIdémiologie CLINique (EPICLIN), Nancy, FR

Publications:

1. An article based on Chapter 3 is in preparation to be submitted.

Zhai Y, Bardel C, Vallée M, Iwaz J, Roy P. Place of concordance-discordance model in evaluating NGS performance. *In preparation*.

2. An article based on Chapter 4 is published in *Frontiers in Genetics*.

Zhai Y, Bardel C, Vallée M, Iwaz J, Roy P. Performance comparisons between clustering models for reconstructing NGS results from technical replicates. *Front Genet.* 2023 Mar 16;14:1148147. doi: 10.3389/fgene.2023.1148147. PMID: 37007945; PMCID: PMC10060969.

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 16 March 2023
DOI 10.3389/fgene.2023.1148147

OPEN ACCESS

EDITED BY
Li-Xuan Qin,
Memorial Sloan Kettering Cancer Center, United States

REVIEWED BY
Xiangyu Luo,
Renmin University of China, China
Jian Zou,
University of Pittsburgh, United States

*CORRESPONDENCE
Yue Zhai,
ext-yue.zhai@chu-lyon.fr

SPECIALTY SECTION
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

RECEIVED 19 January 2023
ACCEPTED 06 March 2023
PUBLISHED 16 March 2023

CITATION
Zhai Y, Bardel C, Vallée M, Iwaz J and Roy P (2023), Performance comparisons between clustering models for reconstructing NGS results from technical replicates.
Front. Genet. 14:1148147.
doi: 10.3389/fgene.2023.1148147

Performance comparisons between clustering models for reconstructing NGS results from technical replicates

Yue Zhai[1,2,3]*, Claire Bardel[1,2,3,4,5], Maxime Vallée[6], Jean Iwaz[1,2,3,4] and Pascal Roy[1,2,3,4]

[1]Université Lyon 1, Lyon, France, [2]Université de Lyon, Lyon, France, [3]Laboratoire de Biométrie et Biologie Évolutive, Villeurbanne, France, [4]Service de Biostatistique-Bioinformatique, Hospices Civils de Lyon, Lyon, France, [5]Service de Génétique, Hospices Civils de Lyon, Bron, France, [6]Cellule Bioinformatique de La Plateforme de Séquençage Haut Débit NGS-HCL, Hospices Civils de Lyon, Bron, France

To improve the performance of individual DNA sequencing results, researchers often use replicates from the same individual and various statistical clustering models to reconstruct a high-performance callset. Here, three technical replicates of genome NA12878 were considered and five model types were compared (consensus, latent class, Gaussian mixture, Kamila–adapted k-means, and random forest) regarding four performance indicators: sensitivity, precision, accuracy, and F1-score. In comparison with no use of a combination model, i) the consensus model improved precision by 0.1%; ii) the latent class model brought 1% precision improvement (97%–98%) without compromising sensitivity (= 98.9%); iii) the Gaussian mixture model and random forest provided callsets with higher precisions (both >99%) but lower sensitivities; iv) Kamila increased precision (>99%) and kept a high sensitivity (98.8%); it showed the best overall performance. According to precision and F1-score indicators, the compared non-supervised clustering models that combine multiple callsets are able to improve sequencing performance vs. previously used supervised models. Among the models compared, the Gaussian mixture model and Kamila offered non-negligible precision and F1-score improvements. These models may be thus recommended for callset reconstruction (from either biological or technical replicates) for diagnostic or precision medicine purposes.

KEYWORDS

next generating sequencing, performance evaluation, clustering model, replicate analysis, sensitivity

1 Introduction

Evaluating the performance of an individual's DNA sequencing results is often hampered by the lack of gold standard. A number of researchers use then replicates of DNA sequencing results from the same individual or from monozygotic twins to reconstruct a set of high-quality calls (Zook et al., 2014). Sequencing results obtained from two or more distinct samples from a same individual are called biological replicates, whereas sequencing results obtained from two or more distinct vials of a single sample are called technical replicates (Robasky et al., 2014). Technical replicates may stem from using different sequencing platforms, different bioinformatics analysis tools, or repeated sequencing

TABLE 1 Overview of the most relevant methods and studies designed to combine multiple callsets.

Authors	Algorithm	Model type	References
Trubetskoy et al., 2015	CGES	Consensus	Bioinformatics 2015; 31(2):187
Wang et al., 2020	SomaticCombiner	Consensus	Sci Rep 2020; 10:12898
Chiara et al., 2018	CoVaCS	Consensus	BMC Genomics 2018; 19:120
Hwang et al., 2014	---	Consensus and logistic regression	Hum Mutat 2014; 35(8):936
Cantarel et al., 2014	BAYSIC	Bayesian latent class model	BMC Bioinformatics 2014; 15:104
DePristo et al., 2011	VQSR	Gaussian mixture model	Nat Genet 2011; 43(5):491
Hwang et al., 2019	---	Gaussian-multinomial mixture model	Sci Rep 2019; 9(1):3219
Huang et al., 2019	SMuRF	Random forest	Bioinformatics 2019; 35 (17): 3157

with the same platform and same bioinformatics tool. With both types of sequencing replicates, several methods have been widely used to obtain more reliable sequencing results.

Among these methods, a simple one is the concordance-based model where a "consensus" can be defined according to various degrees of agreement between callsets (Trubetskoy et al., 2015). Although this model may seem "naïve", several investigations have suggested that its performance may not be worse than that of a machine-learning method (Wang et al., 2020).

Another method is latent class analysis (LCA) that is commonly used in biology and medicine to evaluate test performance without gold standard. In a classical latent class model, the latent variable and the observed variables are all categorical and there is a conditional independence between the observed variables within each latent class. Extensions of this classical model have been developed to account for local dependence, such as using random effects or correlation coefficients. Other extensions included covariables with effects on the latent variable or on the observed variables (Huang and Bandeen-Roche, 2004). Furthermore, Bayesian latent class analyses have been also used to provide combinations of callsets with improved performance indicators (Cantarel et al., 2014). A similar approach was the Gaussian mixture model in which the categorical latent variable is the class membership of the observations and where the observed continuous variables within each latent class follow hypothetically a Gaussian distribution. Finally, machine-learning methods (k-nearest neighbors, random forest, naïve Bayes classifier, or support vector machine) were also used to merge several callsets (Gézsi et al., 2015; Wang et al., 2020). Table 1 provides a short overview of the most relevant methods and studies designed to combine multiple callsets.

The literature on processing replicate sequencing results is rather scanty and a number of methods do not satisfy specific research needs. This work intended to explore the main ways of dealing with multiple NGS results stemming from biological or technical replicates, investigate their properties, and compare their key performance indicators to help choosing the most performing among readily implementable methods able to improve sequencing performance. It explored the consensus model, the latent class model, the mixture model, and random forest regarding their abilities to produce a callset with improved quality. It compared their main performance indicators: precision, recall, and F1-score.

2 Methods

2.1 The study data

The present study used calling results from sequencing three technical replicates of genome NA12878. NA12878 is a human DNA sample that is "thought to represent the best-characterized diploid human genome in the world", is "considered as a 'reference material' by the National Institute of Standards and Technology (NIST)", and includes "near-perfect genome sequences for public use" as well as "truth sequences" established after repeated sequencings "using a wide variety of technologies and computational pipelines". Today, more than 80% the NA12878 cell line's genome is considered known with high confidence. This is why it is used as benchmark for assessing the performance of sequencing platforms or bioinformatic pipelines (Krol, 2015).

All three sequencing procedures were carried out on Illumina NovaSeq 6000 system platform. The samples were then aligned with Burrow-Wheeler Aligner (BWA-MEM) (Li, 2013) against the GRCh37 version of the human reference genome. Genome Analysis Toolkit (GATK) duplicate marking, base quality score recalibration, and indel realignment were applied (McKenna et al., 2010). The resulting sequencing data were deposited in the European Nucleotide Archive.

Variant calling was performed by joint genotyping according to the GATK Best Practices recommendations (DePristo et al., 2011; van der Auwera and O'Connor, 2020). Concordance rates between the calling results of the replicates were calculated. The concordance rate was defined as the number of sites called in the same category (see 2.2) by each replicate divided by the total number of sites called as variants by at least one of the replicates.

The latest version (v 4.2.1) of Genome in a Bottle (GIAB) variant calling benchmark set was used as 'gold standard' (Zook et al., 2016; Wagner et al., 2022). This version has a higher coverage of the GRCh37 reference genome and includes more difficult-to-map regions than the previous version (Wagner et al., 2022).

2.2 Basic definitions and main covariables

Performance considered only bps from the GIAB benchmark region, each bp position being a statistical unit and each GIAB

benchmark result a true status of each bp. Here, only performance in single nucleotide variant (SNV) analysis was considered.

In this analysis, the variant calling results in the VCF file and the GIAB benchmark callset (gold standard set) were considered to belong to one of three categories: homozygous reference, heterozygous variants, and homozygous variants. A true positive (TP) was defined as a variant call in the query callset that belongs to the same category as in the gold standard set; i.e., both are heterozygous variants or both homozygous variants despite potential allele or phasing differences. A false negative (FN) was defined as a variant in the gold standard set called as non-variant in the query callset. A false positive (FP) was defined as a non-variant in the gold standard set called as variant in the query callset or a variant in the gold standard set called as variant in a different category. A true negative (TN) was defined as a non-variant in the gold standard set called as non-variant in the query callset. No-calls in the VCF file were considered as non-variants. This recalls the "genotype match, for which only sites with matching alleles and genotypes are counted as TPs" (Krusche et al., 2019), though, in this study, the criteria for true positivity were less stringent.

The covariables included in the models were:

1) The depth of coverage (DP); i.e., the number of informatics reads covering a given base-pair. In this study, the mean DP value across the three replicates was circa 38 and the DP value ranged from 0 to 13,858.
2) The allele balance (AB; i.e., the number of reads supporting the alternative allele divided by the number of all informatics reads at a specific site) ranged from 0 to 1.
3) The QualByDepth (QD); i.e., the site-level Phred-scaled confidence for the existence of variant divided by the number of reads supporting the alternative allele in variant samples. Here, the QD value ranged from 0.02 to 42.9.
4) The genotype quality (GQ); i.e., the Phred-scaled confidence for the called genotype (ranged from 0 to 99).
5) The mapping quality (MQ); i.e., the root mean square of the MQ of reads across all samples (ranged from 20 to 60).

Covariates DP, AB, and GQ were obtained from the VCF file for each bp in each sample (here, replicate), and then the mean of each of the three values was calculated. MQ and QD were obtained from the VCF file for each bp and had the same values across the three samples.

2.3 Clustering models used for NGS reconstruction

Five types of models were selected for reconstructing NGS result from technical replicates.

2.3.1 The consensus (or concordance-based) model

In this model, "strict consensus" was considered whenever all variant calling results across all replicates agreed and "majority consensus" whenever there was a majority of variant calling results across all replicates (Trubetskoy et al., 2015; Wang et al., 2020). Here, it is the majority consensus that was used. In case of no majority consensus, the sites were classified as homozygous variants.

2.3.2 The latent class model without covariables

This type of analysis was often used to evaluate the performance of diagnostic tests in the absence of gold standard. A latent class analysis is a mixture model where both the observed and unobserved variables are categorical. A classical LCA assumes conditional independence between observed variables (here, called genotype categories) given the latent class (here, the true genotype status).

Let i represent each site in the VCF file, r the latent classes 1 to 3. $\boldsymbol{Y_i}$ represents the calling results in replicates 1 to 3 for site i (Y_1, Y_2, and Y_3 are categorical variables with three categories that correspond to the three genotype categories). p_r denotes the prevalence of latent class r. $\pi_r(Y_1)$, $\pi_r(Y_2)$, and $\pi_r(Y_3)$ are the probability mass functions of variables Y_1, Y_2, and Y_3 for latent class r.

The equation of this model may be written:

$$P\left(\boldsymbol{Y}_i|\boldsymbol{p},\boldsymbol{\pi}\right) = \sum_{r=1}^{3} p_r \times \pi_r(Y_1) \times \pi_r(Y_2) \times \pi_r(Y_3) \qquad (1)$$

The model parameters, namely, p_r and π_{jrk}, were estimated with an expectation-maximization (EM) algorithm using 50 sets of random initial values.

2.3.3 The latent class model with covariables

In this model, covariables' effects were put on the prior probability of class membership (P_r in Eq. 1) and modelled using a logistic link. Covariables that are potentially correlated with the latent bp status were included; namely, Allele Balance (AB; the mean AB value of the three replicates), QualByDepth (QD), and Mapping Quality (MAPQ). Univariate models were first fitted for each covariable, then models were fitted with all possible pairs of covariables. Model parameters $(\boldsymbol{\pi}, \boldsymbol{p})$ were estimated using 100 sets of random initial values. Models with distinct covariables were compared with the Bayesian information criterion (BIC) as a measure of model fit.

The latent class model without covariables and the latent class model with covariables were fitted using package "poLCA" (v. 1.6.0.1) in R (v. 4.1.3) (Linzer and Lewis, 2011).

2.3.4 The Gaussian mixture model

The Gaussian mixture model assumes that the observed variables within each latent class follow a multivariate normal distribution. Here, it is the observed continuous covariables that were modelled; the calling results of each replicate were not included. The covariables included in the model were read depth (DP; the mean DP value of the three replicates), allele balance (AB; the mean AB value of the three replicates), and quality by depth (QD); all were assumed to be normally distributed.

Let $\boldsymbol{x}_i$ denote the vector of covariables for site i, p_r the prevalence of each latent class (r = 1, 2, or 3), $\boldsymbol{\alpha}_r$ the parameters of the multivariate normal distribution for latent class r. $h(x_i|\alpha_r)$ is the probability density function for latent class r, with parameters $\boldsymbol{\alpha}_r$. Thus, the probability density function for $\boldsymbol{x}_i$ can be written:

$$f\left(\boldsymbol{x}_i|\boldsymbol{p},\boldsymbol{\alpha}\right) = \sum_{r=1}^{R} p_r h(\boldsymbol{x}_i|\boldsymbol{\alpha}_r)$$

The model parameters, namely, $\boldsymbol{p}$ and $\boldsymbol{\alpha}$, were estimated with an expectation-maximization (EM) algorithm. This model was fitted using package "mclust" (v. 6.0.0) in R (v. 4.1.3) (Scrucca et al., 2016).

2.3.5 Kamila model (k-means for mixed large datasets)

Kamila is a model-based adaptation of the k-means clustering algorithm for heterogeneous variables (mix of categorical and continuous). It uses a kernel density estimation technique to model flexibly spherical clusters in the continuous domain and uses a multinomial model in the categorical domain (Foss et al., 2016). The model parameters were estimated with an iterative process similar to an EM algorithm. One advantage of this model is to include both types of variables at the same time without pre-specifying the weights of continuous *versus* categorical variables.

The categorical covariables included were: the calling results of the three replicates and a binary covariable to indicate whether a site is present in a "difficult region" (Amemiya et al., 2019). The continuous covariables included were DP, AB, and QD. The algorithm is sensitive to outliers because it uses kernel density estimation and Euclidean distance for continuous covariables. Here, the maximum value of DP was set to 150.

This model was applied with package "Kamila" (v. 0.1.2) in R (v. 4.1.3) (Foss and Markatou, 2018).

2.3.6 The random forest

An unsupervised version of the random forest model for clustering was implemented (Shi and Horvath, 2006). The algorithm started with an unsupervised random forest model to generate a synthetic dataset without correlation between covariables, and then classified the observations into the synthetic or the original dataset using a classical random forest. This generates a proximity matrix that represents the number of times observations were classified into the correct dataset. A hierarchical clustering was then applied using the proximity scores as dissimilarity measure between observations.

This model was applied with Package 'RandomForest' (v. 4.7-1.1) in R (v. 4.1.3) (Liaw and Wiener, 2002). Because this model is computationally expensive, only 10,000 sites from the VCF file were sampled for its use. The number of trees used was 1000.

2.4 Clustering choices

Among the six above-mentioned models, five generate clusters. As the purpose was identifying the three latent classes that correspond to the three genotype categories, the number of clusters in each model was fixed to three. The largest cluster had to correspond to the heterozygous variants, the intermediate cluster to the homozygous variants, and the smallest cluster to the homozygous reference. Also, any model that showed any cluster with <0.1% of the observations was considered unable to identify three clusters, and therefore not retained. This choice was made according to a prior knowledge about the relatively stable proportions of the three categories in a VCF file of WGS. The ratio of heterozygous variants to homozygous variants in the VCF files is expected to be around 2 (Guo et al., 2014; Wang et al., 2015). The reference sites (i.e., the false positives for at least one replicate) occupy usually 0.1%–10% in WGS data (Zhao et al., 2020).

2.5 Model result comparisons

Each callset was compared against the GIAB gold standard set. This comparison used the above-provided definitions of TPs, FPs, FNs, and TNs as well as the following performance indicators:

i) Accuracy (or 1−the overall classification error rate) was calculated as (TPs + TNs)/(TPs + FPs + FNs + TNs); i.e., over the total number of sites in the VCF file;
ii) Recall (or sensitivity) was calculated as TPs/(TPs + FNs);
iii) Precision (or positive predictive value, PPV) was calculated as TPs/(TPs + FPs);
iv) F1-score was calculated as 2 × recall × precision/(recall + precision).

All callsets (except the one generated from the random forest) included all sites in the VCF file. For the random forest callset, the total number of real variants was estimated as the number of variants in the gold standard set multiplied by the sampling proportion.

3 Results

3.1 Performance indicators for calling results of individual replicates

The precisions relative to the three replicates (1 to 3) had very close values (96.7%–96.9%) and the sensitivities were nearly the same (~98.9%) (Table 2). The concordance rates of Replicate 1 vs. Replicates 2 and 3 were 98.4% and 98.3%, respectively; whereas the concordance rate of Replicate 1 vs. Replicate 3 was 98.2%. The concordance rate across the three replicates was 97.5%.

Thus, as expected, the three replicates had similar performance indicators and there were high concordance rates between replicates. However, given the number of total loci in the VCF file (n = 3,351,415), the number of discordant sites across replicates was not negligible (n = 84,753).

Among the concordant sites across the three replicates, precision differed for different genotype categories. For the concordant heterozygous variant sites (n = 1,993,116), the precision was 96.8%. For the concordant homozygous variant sites (n = 1,273,546), the precision was 99.6%. Among the discordant sites, 55.9% were homozygous references, 39.6% heterozygous variants, and 4.5% homozygous variants in the gold standard.

3.2 Comparison of model fits

In this study, the five types of models used neither the same amount of information nor the same type of covariables: i) the consensus model and the classical latent class model used the categorical variant calling results from the three replicates; ii) the Gaussian mixture model used continuous covariables; iii) the latent class model with covariables, Kamila model, and random forest used categorical variant calling results as well as categorical or continuous covariables. It was therefore difficult to compare directly model fits across model types. This section presents only comparisons within each model type.

TABLE 2 Performance indicators of the clustering models under study.

Clustering model	Accuracy	Precision	Recall	F1-score
None	96.7%–96.9%	96.7%–96.9%	98.9%	97.8%–97.9%
Majority consensus	97.0%	97.0%	98.9%	97.9%
Latent class analysis without covariables	97.8%	97.9%	98.8%	98.3%
Latent class analysis with covariables	98.0%	98.0%	98.9%	98.4%
Gaussian mixture model	98.5%	99.3%	98.2%	98.7%
Kamila	99.0%	99.2%	98.8%	99.0%
Random Forest	98.2%	99.5%	97.9%	98.7%

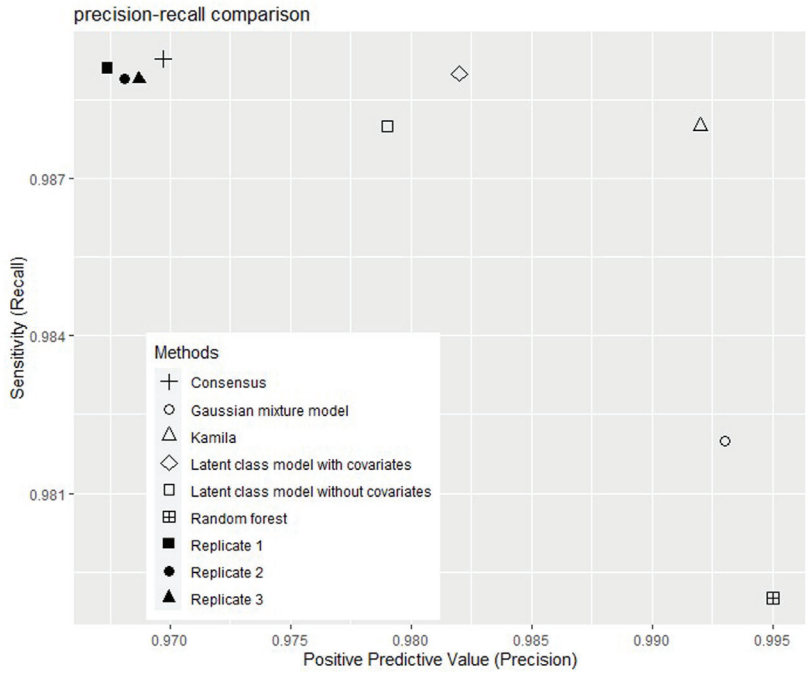

FIGURE 1
Positive predictive values and sensitivities of callsets without and with selected clustering models.

With the latent class models with one covariable (AB, QD, or MAPQ), the effect of each covariable was significantly different from 0. The model with AB showed the smallest BIC and was therefore considered as the most fitted to the data.

With the latent class model with two covariables, among the three models relative to the three pairs of covariables, the model with AB and QD had the lowest BIC. Here, it is useful to note that, with some models, the estimations of the parameters of the latent class model with covariables were not stable. With some models, the global maxima of the log-likelihood were reached in only 10% of estimation attempts. The most frequent local maxima were seldom the global maxima and the estimated proportions of heterozygous variant, homozygous variant, and homozygous reference sites were substantially different between estimation attempts. Therefore, a large number of sets of random initial values (100 rather than 50) were necessary to avoid local maxima. (Supplementary Table S1).

With the Gaussian mixture model, the chosen model (the one with the lowest BIC) was the model with three covariables: DP, AB, and QD.

3.3 Performance comparisons

The performance indicators (accuracy, precision, recall, and F1-score) of the models are shown in Table 2 and Figure 1 shows the precision and the recalls of callsets of individual replicates and clustering models. The consensus method improved the precision by 0.1% without much decrease of the recall. Among the five clustering

TABLE 3 Proportions of the three genotype categories in each callset.

	Callset	Homozygous References (%)	Heterozygous variants (%)	Homozygous variants (%)
1	Gold standard (GIAB)	4.241	57.891	37.868
2	Calling results of Replicate 1	1.064	60.800	38.136
3	Calling results of Replicate 2	1.230	60.672	38.098
4	Calling results of Replicate 3	1.295	60.618	38.087
5	Majority consensus	1.287	60.586	38.127
6	Latent class analysis without covariables	2.283	59.596	38.121
7	Latent class analysis with covariables	2.632	59.171	38.197
8	Gaussian mixture model	4.426	58.001	37.573
9	Kamila	3.586	58.310	38.104
10	Random forest	5.560	57.440	37.000

models, the Gaussian mixture model showed the highest accuracy (98.5%). The random forest model showed the highest precision (99.6%) but the lowest recall (98.2%). The consensus model and the latent class model with covariables showed the highest recall (98.9%). The Gaussian mixture model and random forest had high F1-scores (98.7%). Kamila model showed the highest F1-score (99.0%).

The proportions of the three genotype categories in each callset, including the gold standard GIAB benchmark set, are shown in Table 3 (Total loci: 3,351,415 in the VCF file). The first row shows the "true" category proportions in the GIAB benchmark set for all sites in the VCF file. More than 4% were classified as reference sites in GIAB set, which corresponds to the marginal false positive rate in the VCF file. Rows 2 to 5 show the proportions in the three replicates and the consensus callset. With the model-based methods (rows 6–10), these proportions were the estimated latent-class proportions. The callsets generated by the clustering models grouped more sites into the smallest class (interpreted as reference; thus, false positives) than into the consensus callset; this explains the improved precision of these models. With the Gaussian mixture model, the highest proportion was found in the reference category, which explains its higher precision and lower recall *versus* the other models.

4 Discussion

In this study, six clustering algorithms were run on real sequencing replicates of the NA12878 genome to compare their abilities in allowing reconstruction of a new callset with improved performance: one consensus model, two latent-class models, a Gaussian mixture model, a Kamila (adapted k-means) model, and a random forest model. These models showed various advantages. For example, the consensus model improved slightly the precision (by 0.1%) whereas the latent class model provided a non-negligible 1% precision improvement (97% to 98%) without compromising recall (98.9%). In comparison with no use of a clustering model, all six models brought ≥1% gain in sensitivity, which is not negligible: i) the Gaussian mixture and the random forest models provided callsets with high precision (>99%) but at the price of lower recall; ii) Kamila increased precision (99.2%) and kept a high recall (98.8%); it proved having the best overall performance.

In this work, the models were chosen to represent a range of major clustering models, from the most naïve (consensus) to the most sophisticated machine-learning type (random forest). One interest of this choice is that all models may be readily implemented with packages in R software. However, here, only non-supervised clustering models were compared and not supervised ones because the latter need high-quality training data (Sandmann et al., 2018) which are not usually available in clinical practice settings. The models dealt with by BAYSIC and SomaticCombiner or their equivalents were actually considered in this article as latent class model and consensus model, respectively. Indeed, in this work, the former algorithm was not considered because its results would be quite similar to those obtained with a classical latent class model and the latter is based on an approach that is close to the consensus model.

Most of the models considered here have been previously used for similar purposes; i.e., merging several either constitutional or somatic variant calling results to obtain a new callset with better performance indicators (precision or recall). Previous authors used: i) the consensus model (Hwang et al., 2014; Trubetskoy et al., 2015; Chiara et al., 2018; Di Nanni et al., 2019); ii) the Bayesian latent class model (Cantarel et al., 2014); iii) the Gaussian mixture model (DePristo et al., 2011; Hwang et al., 2019); iv) random forest (Huang et al., 2019; Wang et al., 2020). However, though usual, these models have been rarely compared, their comparison results often unclear, and the final conclusions controversial. For example, the random-forest-based ensemble caller for somatic mutation has obtained higher F1-scores than the simple consensus approach (Huang et al., 2019); however, in a study by Wang et al. (Wang et al., 2020), the authors observed that the consensus method was more robust and stable than supervised machine-learning models. They suggested that the difference between the training data and the test data contributed to the poor generalizability of machine-leaning models. In another research on the NA12878 genome that used the GIAB benchmark set as gold standard, a two-component mixture-model-based method that considered results from 70 pipelines did

not significantly improve performance in terms of precision at the highest analytical sensitivity achievable vs. the highest performance of a single pipeline. However, the method led to performance improvement with another gold standard set from the '1000 Genomes Project' (Hwang et al., 2019).

The models compared here did not include the same number of variables because of the hypotheses inherent to each model. Some require only continuous variables (e.g., the Gaussian mixture model), whereas others require only categorical variables (e.g., the latent class model). Thus, performance comparisons between new callsets generated by different models should be interpreted with this difference in mind. For example, Kamila and random forest models are able to include more covariables than the other models. In future works, comparisons between models with same covariables would be welcome. One current aim was to use information already available in a VCF file; however, the possibility of including more covariables may be interesting too.

In some previous research works, sites in the VCF file of presumably very low quality were filtered out before applying merging methods; i.e., a small number of sites were considered as false positives and thus excluded (Sandmann et al., 2018). Here, no sites were filtered out (all sites from the VCF file were included in the models); this allowed a more objective evaluation of the overall performance of each model. However, this choice introduced some difficulties due to the extreme values of certain variables. For example, DP has typically a long-tailed distribution and the presence of extremely high values is often an indicator of sequencing artifacts, alignment artifacts, or copy number variations (O'Rawe et al., 2013; Guo et al., 2014; Li, 2014). In common practice, the solution to extreme DP values is to exclude sites with values higher than a threshold defined according to various formulas that use the mean and standard deviation of DPs (Li et al., 2018; Pan et al., 2022); for example, a threshold 120 in the hard filters recommended by the GATK (van der Auwera and O'Connor, 2020).

In the present work, the mean DP across the three replicates was circa 38 and its maximum 13,858 and, among the compared models, Kamila is known to be relatively sensitive to extreme values because it minimizes a dissimilarity measure that is partially based on Euclidean distance in the case of continuous variables. This might explain why it failed to identify the three clusters with acceptable proportions. Indeed, the model grouped a small number of sites with extremely high DP values into one cluster ($n = 254$; i.e., 0.008% of all sites) and, as stated in 2.4, models that led to any cluster with <0.1% of the sites were considered unable to identify three clusters and thus not retained. One way to address this issue is to add one more cluster in the model (4 instead of 3). However, in this work, only three clusters were considered to allow model performance comparisons and allow each cluster to represent each genotype category. Therefore, with Kamila, the maximum DP value was set at 150 and higher values grouped together at 150. The other models that involved DP (i.e., the Gaussian mixture model and the random forest model) performed well despite the presence of high DP values (these were not then filtered out).

This study focused on the VCF file (i.e., on all sites called as variants in at least one replicate) and not on all three billion bp positions across the human genome. This is one reason for which the indicators of performance kept were only recall and precision (specificity was ignored). There are also two other practical reasons: i) negative sites are much more numerous (almost 1000 times the number of sites in the VCF file) and contain less information; thus, using them is computationally expensive and adds little information; ii) researchers, especially practitioners and lab professionals, usually use only the VCF file for routine analyses; thus, a model that requires information from the BAM file for sites called as 'reference' would not be practical.

One limitation of this study is that it evaluated only callsets' performance regarding SNVs. Further studies are worth being conducted to evaluate the performance of clustering models regarding copy number variations and structural variations. Also, except for Kamila, the study included only the most classical model from each clustering algorithm type. Some model features may prove more adapted to the distribution of the variables or have more convenient underlying hypotheses. For example, latent class models that relax the conditional independence between observed variables through correlation, random effects, or covariables with effects on the class-conditional probabilities.

The Gaussian mixture model used here showed good performance vs. the other five models. However, all components of a variable distribution might not be Gaussian. For example, i) the distribution of allele balance has been already modelled using a mixture of 0-inflated beta distribution, binomial distribution, and 1-inflated beta distribution for the homozygous reference, heterozygous variant, and homozygous reference categories, respectively (Muyas et al., 2019); ii) to take into account heavytails, read depth distributions have been modelled using a compound Poisson distribution, a negative binomial distribution, or a log-normal distribution (Robinson et al., 2010; Daley and Smith, 2014; Deng et al., 2020).

From a theoretical viewpoint, a very recent article by Dang et al. (Dang et al., 2023) reviewed a selection of "mixture models that can deal with varying cluster tail-weight, skewness and/or concentration, and kurtosis" (e.g., mixtures of multivariate t-distributions, mixtures of skew-t distributions, mixtures of normal inverse Gaussian distributions, etc.). Furthermore, these authors introduced a multivariate skewed power exponential distribution that "allow for robust mixture models for clustering with skewed or symmetric components" and "model components with varying levels of peakedness, skewness, and tail-weight (light, heavy, Gaussian)". In practice, the use of multivariate non-Gaussian mixture models is often difficult because of identifiability issues and the instability of parameter estimation. This might explain the rarity of applications on real data, which is worth being explored. We especially hope an exploration of the appropriateness of the above-mentioned models within the context of WGS data.

5 Conclusion

In this study, several clustering models were evaluated within the context of combining callsets from DNA sequencing replicates. These non-supervised clustering models proved able to improve sequencing performance in terms of precision and F1-score, which is comparable to what is reported about supervised models. Among the models compared here, the Gaussian mixture model and Kamila offered improvements that made precision higher than

99% and F1-score close to 99%. These models may then be recommended to reconstruct new high-performance callsets from NGS replicates. This is of particular interest for diagnosis or precision medicine whenever DNA sequencing results stem from either biological replicates (more than one sample) or technological replicates (more than one sequencing platform or analysis pipeline).

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://www.ebi.ac.uk/ena/browser/home. Accession number: PRJEB60499.

Author contributions

YZ and PR designed the study. YZ performed the statistical analysis and wrote the manuscript. CB contributed to data extraction and sequence alignment. MV did the variant calling. JI helped writing, commenting on, and editing the manuscript.

Funding

The first author was supported by a China Scholarship Council grant (Grant No. 201906230310).

Acknowledgments

The authors acknowledge the support by project SIRIC (LYRICAN, Grant INCa-DGOS-INSERM_12563) and by AURAGEN platform (France Médecine Génomique 2025 National Plan).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1148147/full#supplementary-material

References

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The ENCODE blacklist: Identification of problematic regions of the genome. *Sci. Rep.* 9 (1), 9354. doi:10.1038/s41598-019-45839-z

Cantarel, B. L., Weaver, D., McNeill, N., Zhang, J., Mackey, A. J., and Reese, J. (2014). Baysic: A bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinforma.* 15 (1), 104. doi:10.1186/1471-2105-15-104

Chiara, M., Gioiosa, S., Chillemi, G., D'Antonio, M., Flati, T., Picardi, E., et al. (2018). CoVaCS: A consensus variant calling system. *BMC Genomics* 19 (1), 120. doi:10.1186/s12864-018-4508-1

Daley, T., and Smith, A. D. (2014). Modeling genome coverage in single-cell sequencing. *Bioinformatics* 30 (22), 3159–3165. doi:10.1093/bioinformatics/btu540

Dang, U. J., Gallaugher, M. P. B., Browne, R. P., and McNicholas, P. D. (2023). Model-based clustering and classification using mixtures of multivariate skewed power exponential distributions. *J. Classif.* 2023. doi:10.1007/s00357-022-09427-7

Deng, C., Daley, T., Calabrese, P., Ren, J., and Smith, A. D. (2020). Predicting the number of bases to attain sufficient coverage in high-throughput sequencing experiments. *J. Comput Biol. July* 27 (7), 1130–1143. doi:10.1089/cmb.2019.0264

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43 (5), 491–498. doi:10.1038/ng.806

Di Nanni, N., Moscatelli, M., Gnocchi, M., Milanesi, L., and Mosca, E. (2019). isma: an R package for the integrative analysis of mutations detected by multiple pipelines. *BMC Bioinforma.* 20 (1), 107. doi:10.1186/s12859-019-2701-0

Foss, A., Markatou, M., Ray, B., and Heching, A. (2016). A semiparametric method for clustering mixed data. *Mach. Learn* 105 (3), 419–458. doi:10.1007/s10994-016-5575-7

Foss, A. H., and Markatou, M. (2018). Clustering mixed-type data in R and hadoop. *J. Stat. Soft* 83 (13), 1–44. doi:10.18637/jss.v083.i13

Gézsi, A., Bolgár, B., Marx, P., Sarkozy, P., Szalai, C., and Antal, P. (2015). VariantMetaCaller: Automated fusion of variant calling pipelines for quantitative, precision-based filtering. *BMC Genomics* 16 (1), 875. doi:10.1186/s12864-015-2050-y

Guo, Y., Ye, F., Sheng, Q., Clark, T., and Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. *Briefings Bioinform* 15 (6), 879–889. doi:10.1093/bib/bbt069

Huang, G. H., and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* 69 (1), 5–32. doi:10.1007/bf02295837

Huang, W., Guo, Y. A., Muthukumar, K., Baruah, P., Chang, M. M., and Jacobsen Skanderup, A. (2019). SMuRF: Portable and accurate ensemble prediction of somatic mutations. *Bioinformatics* 35 (17), 3157–3159. doi:10.1093/bioinformatics/btz018

Hwang, K. B., Lee, I. H., Li, H., Won, D. G., Hernandez-Ferrer, C., Negron, J. A., et al. (2019). Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Sci. Rep.* 9 (1), 3219. doi:10.1038/s41598-019-39108-2

Hwang, K. B., Lee, I. H., Park, J. H., Hambuch, T., Choe, Y., Kim, M., et al. (2014). Reducing false-positive incidental findings with ensemble genotyping and logistic regression based variant filtering methods. *Hum. Mutat.* 35 (8), 936–944. doi:10.1002/humu.22587

Krol, A. (2015). Genome in a Bottle uncapped. Available at: https://www.bio-itworld.com/news/2015/05/21/genome-in-a-bottle-uncapped.

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* 37 (5), 555–560. doi:10.1038/s41587-019-0054-x

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997 (accessed Dec 2022).

Li, H., Bloom, J. M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., et al. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* 15 (8), 595–597. doi:10.1038/s41592-018-0054-7

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30 (20), 2843–2851. doi:10.1093/bioinformatics/btu356

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R. News* 2 (3), 18–22.

Linzer, D. A., and Lewis, J. B. (2011). poLCA: An *R* package for polytomous variable latent class analysis. *J. Stat. Soft* 42 (10), 1–29. doi:10.18637/jss.v042.i10

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9), 1297–1303. doi:10.1101/gr.107524.110

Muyas, F., Bosio, M., Puig, A., Susak, H., Domènech, L., Escaramis, G., et al. (2019). Allele balance bias identifies systematic genotyping errors and false disease associations. *Hum. Mutat.* 40 (1), 115–126. doi:10.1002/humu.23674

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., et al. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. *Genome Med.* 5 (3), 28. doi:10.1186/gm432

Pan, B., Ren, L., Onuchic, V., Guan, M., Kusko, R., Bruinsma, S., et al. (2022). Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. *Genome Biol.* 23 (1), 2. doi:10.1186/s13059-021-02569-8

Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15 (1), 56–62. doi:10.1038/nrg3655

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Sandmann, S., Karimi, M., de Graaf, A. O., Rohde, C., Göllner, S., Varghese, J., et al. (2018). appreci8: a pipeline for precise variant calling integrating 8 tools. *Bioinformatics* 34 (24), 4205–4212. doi:10.1093/bioinformatics/bty518

Scrucca, L., Fop, M., MurphyBrendan, T., and Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R. J.* 8 (1), 289–317. doi:10.32614/rj-2016-021

Shi, T., and Horvath, S. (2006). Unsupervised learning with random forest predictors. *J. Comput. Graph. Statistics* 15 (1), 118–138. doi:10.1198/106186006x94072

Trubetskoy, V., Rodriguez, A., Dave, U., Campbell, N., Crawford, E. L., Cook, E. H., et al. (2015). Consensus genotyper for exome sequencing (CGES): Improving the quality of exome variant genotypes. *Bioinformatics* 31 (2), 187–193. doi:10.1093/bioinformatics/btu591

van der Auwera, G., and O'Connor, B. D. (2020). *Genomics in the cloud: Using docker, GATK, and WDL in terra*. First edition. Sebastopol, CA: O'Reilly Media.

Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Khan, Z., Farek, J., et al. (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genom* 2 (5), 100128. doi:10.1016/j.xgen.2022.100128

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., and Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* 31 (3), 318–323. doi:10.1093/bioinformatics/btu668

Wang, M., Luo, W., Jones, K., Bian, X., Williams, R., Higson, H., et al. (2020). SomaticCombiner: Improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Sci. Rep.* 10 (1), 12898. doi:10.1038/s41598-020-69772-8

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., and Hovig, E. (2020). Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci. Rep.* 10 (1), 20222. doi:10.1038/s41598-020-77218-4

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* 3 (1), 160025. doi:10.1038/sdata.2016.25

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., et al. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* 32 (3), 246–251. doi:10.1038/nbt.2835