DNA comparisons in genomics
a reference-based perspective

Mokveld, T.O.

DOI
10.4233/uuid:ba0ca195-949f-4f61-9daf-5e339fbae727

Publication date
2023

Document Version
Final published version

Citation (APA)
Mokveld, T. O. (2023). *DNA comparisons in genomics: a reference-based perspective*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ba0ca195-949f-4f61-9daf-5e339fbae727

DNA comparisons in genomics

A reference-based perspective

DNA comparisons in genomics
A reference-based perspective

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
maandag 22 mei 2023 om 15:00 uur

door

Tom Onno MOKVELD

Master of Science in Computer Science,
Universiteit Leiden, Nederland,
geboren te Leiden, Nederland.

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. M.J.T. Reinders,	Technische Universiteit Delft, promotor
Dr. ir. Z. Al-Ars,	Technische Universiteit Delft, promotor
Onafhankelijke leden:	
Prof. dr. J.B.J. van Meurs,	Erasmus Medisch Centrum
Prof. dr. C.F.H.A. Gilissen,	Radboud Universiteit Medisch Centrum
Prof. dr. ir. B.P.F. Lelieveldt,	Technische Universiteit Delft en Universiteit Leiden
Prof. dr. ir. D. de Ridder,	Universiteit Wageningen
Prof. dr. A. Hanjalic,	Technische Universiteit Delft, reservelid
Overig lid:	
Prof. dr. E.A. Sistermans,	Amsterdam Universiteit Medisch Centrum

The research presented in this dissertation was funded by Delft Data Science (DDS).

Keywords: bioinformatics, population genetics, population graphs, prenatal testing, metagenomics

Printed by: ProefschriftMaken

Cover: Stable Diffusion + LoRA

Style: TU Delft House Style, modifications by Moritz Beller and Tom Mokveld
`github.com/tomokveld/phd-thesis-template`

The author set this thesis in LaTeX using the Libertinus and Inconsolata fonts.

ISBN 978-94-6366-687-9

An electronic version of this dissertation is available at
`repository.tudelft.nl`

Contents

List of Figures

List of Tables

Summary

Genomics is a field devoted to understanding the differences in genetics between populations, individuals, and even within individuals. By constantly comparing and contrasting data from diverse sources, genomics can refine our understanding of life and identify new ways to improve our lives. However, this often presents technical and biological challenges that require careful consideration of what is compared, in what context, and what might be present. In this thesis I contribute to resolving these challenges in three different domains:

1. In genomic data analysis, analysts often compare and contrast new genomic data to an established reference to reduce costs. However, this approach biases comparisons in favor of population-specific genetics since such references encode only a fraction of the genetics of a given population. To address this bias, I propose a method that accounts for population variability in a way that integrates it directly into the comparison process. This integration ensures that the contrast between sample and reference becomes smaller and closer to personalized, so they are treated the same way regardless of the underlying population. The method improves genome characterization and simplifies downstream analyses that rely on these comparisons. As a result, a more accurate portrayal of the genetics of a given population as a whole is obtained.
2. In non-invasive sequencing-based prenatal testing, we rely on circulating cell-free DNA from maternal plasma to detect pathogenic variants that may affect the fetus. A healthy baseline, which describes the normative state, is generally required to determine the presence of such variants. However, because this DNA is a mixture of maternal and much lower fetal proportions, it remains difficult to disentangle the two, primarily because of biological and technical biases. While this bias can partially be mitigated by changing the baseline and thus contrasting within the individual DNA mixture rather than to a divergent population of mixtures, further improvements are still needed. I present a generalized framework in which the signal-to-noise ratio can be further improved by fully exploiting the information in sequencing data, allowing for more robust predictions at even earlier stages of pregnancy.
3. The composition of the gut ecosystem can have short- and long-term effects on our health. It is therefore important to understand how it is formed and how a healthy balance can be maintained for as long as possible to preserve our health. To do this, ecosystems must be stratified and compared based on health indices. I show in extremely contrasting Dutch subpopulations that we can obtain valuable characteristics of divergent health states by comparing the gut ecosystems of centenarians with those of Alzheimer's patients. However, significant efforts are required to enable these comparisons due to the many organisms present and the technological limitations in measuring them, introducing bias at all levels.

Samenvatting

Genomica is het vakgebied dat zich richt op het begrijpen van genetische verschillen tussen populaties, individuen en zelfs binnen individuen. Door voortdurend gegevens uit verschillende bronnen te vergelijken en te contrasteren, kan genomica ons begrip van het leven verfijnen en nieuwe manieren identificeren om ons leven te verbeteren. Dit brengt echter vaak technische en biologische uitdagingen met zich mee die zorgvuldige overweging vereisen van wat wordt vergeleken, in welke context, en wat er mogelijk aanwezig is. In dit proefschrift draag ik bij aan het oplossen van deze uitdagingen in drie verschillende domeinen:

1. Bij de analyse van genomische gegevens vergelijken analisten vaak nieuwe gegevens met een gevestigde referentie om kosten te besparen. Deze aanpak vertekent echter de vergelijkingen in het voordeel van populatie-specifieke genetica, aangezien dergelijke referenties slechts een fractie van de genetica van een bepaalde populatie coderen. Om deze vertekening aan te pakken, stel ik een methode voor die rekening houdt met populatievariatie op een manier die deze direct in het vergelijkingsproces integreert. Deze integratie zorgt ervoor dat het contrast tussen monster en referentie kleiner en persoonlijker wordt, zodat ze op dezelfde manier worden behandeld, ongeacht de onderliggende populatie. De methode verbetert de genoomkarakterisatie en vereenvoudigt downstream analyses die afhankelijk zijn van deze vergelijkingen.
2. Bij niet-invasieve sequencing-gebaseerde prenatale tests vertrouwen we op circulerend celvrij DNA uit maternaal plasma om pathogene varianten op te sporen die de foetus kunnen beïnvloeden. Een gezonde basislijn, die de ‘normale’ toestand beschrijft, is over het algemeen nodig om de aanwezigheid van dergelijke varianten te bepalen. Omdat dit DNA echter een mengsel is van maternaal en veel lagere foetale proporties, blijft het moeilijk om de twee te ontrafelen, voornamelijk vanwege biologische en technische vertekeningen. Hoewel deze vertekeningen deels verminderd kunnen worden door binnen het individuele DNA-mengsel te vergelijken in plaats van met een diverse populatie van DNA-mengsels, zijn verdere verbeteringen nodig. Ik introduceer een algemeen kader welke alle informatie uit sequencing data benut, wat leidt tot betrouwbaardere voorspellingen in eerdere stadia van de zwangerschap.
3. De samenstelling van het ecosysteem van de darm kan op korte en lange termijn effecten hebben op onze gezondheid. Het is daarom belangrijk om te begrijpen hoe het wordt gevormd en hoe een gezond balans zo lang mogelijk kan worden gehandhaafd. Om dit te doen, moeten ecosystemen gestratificeerd en vergeleken worden op basis van gezondheidsindices. Ik toon in extreem contrasterende Nederlandse subpopulaties aan dat we kenmerken van uiteenlopende gezondheidstoestanden kunnen verkrijgen door de darmecosystemen van honderdjarigen te vergelijken met die van patiënten met de ziekte van Alzheimer.

Prologue

In this thesis, I explored the different roles of DNA sequencing and data analysis in the identification and study of genomic variation. I have demonstrated how the chosen frame of reference can significantly influence the results of sample comparisons in different application areas. Given the breadth and diversity of the topics covered, I have chosen to address them separately. Therefore, rather than presenting a unified narrative, my work is divided into self-contained parts that have their own introductions and conclusions. This allows for a more accurate description of how my contributions helped to overcome specific challenges. I hope this approach will improve readability and guide the reader to the details most relevant to each particular topic.

Thesis outline

Part I: **Population Graphs**, describes the CHOP method, which creates a haplotype specific index of a genome reference representation encoding population-wide variation, *i.e.*, a graph structure known as a pangenome or variation graph. I show how an index that implicitly considers such population variation can improve genome alignment and downstream analyses.

Part II: **Non-Invasive Prenatal Testing**, shows an overview and benchmarks of commonly used methods in the context of sequencing-based non-invasive prenatal testing aimed at detecting congenital disabilities before delivery. In addition, I present the WisecondorFF method, which improves the detection of fetal copy number variation by extending the prevalent within-sample detection technique to include fragment size estimates from paired-end sequencing data.

Part III: **The Human Gut Microbiome**, presents a cross-sectional analysis using 16S rRNA sequencing data. I derive the gut microbial signatures from cohorts of centenarians and Alzheimer's disease patients in the Dutch population and demonstrate each cohort's overlapping and distinguishing features, focusing on trends in taxonomic composition in relation to health indices such as longevity and healthy aging.

I

Population Graphs

1

Introduction

Life has flourished and spread to almost every corner of our planet, yet we all share a common ancestry encoded in our DNA. Adaptations driven by natural selection, genetic drift, and isolation have led to differentiation and the creation of the complete collection of taxonomic diversity that we know today, which fans out from domain to class, to species, and ultimately to differences between *and* within individuals.

To understand life's mechanisms, one must have an in-depth understanding of its constituent parts and interactions. Life can be understood as an, at times, bi-directional flow of information between different media, with the transcription of DNA to RNA and translation to proteins at its heart [1]. The importance of DNA in this context means that enormous efforts have been made to develop technologies to measure DNA, in order to deepen our understanding not only of DNA but of life itself [2, 3]. Whereas it would once have taken years of experimentation to obtain a short sequence from a single organism [4], many millions of sequences can now be determined with great accuracy in a few hours, thanks to advances in sequencing technology [5]. Various platforms are now available for myriad of specialized purposes, ranging from RNA-Seq and Hi-C to ChIP-Seq [6]. These technological advances have directly directed research efforts toward the analysis of sequencing data [2].

Although taxonomic diversity among organisms is generally recognized and understood, it is usually necessary to introduce simplifications in this model to make any meaningful analysis possible. For instance, the canonical method of genomics analysis relies on comparing DNA sequences; nearly always, this occurs by first selecting a single representative haploid genome from a population of related genomes and then using it as a frame of reference, which means that any sequence comparisons made, use the chosen reference sequence as the point of comparison, regardless of variation within the population. Through an algorithmic process called alignment [7, 8], sub-sequences of a genome under study can consequently be mapped onto this reference genome, yielding a positional mapping of matching and differentiating segments between the two genomes. Such mappings form the basis for further downstream analyses, such as variant calling or haplotype phasing. These analyses enable the detection of similarities and differentiating sub-sequences that can then be attributed to specific genotypes or haplotypes, facilitating the answering of biological questions.

For several reasons, the use of a specific linear genome has become the status quo for alignment. While sequencing technologies have greatly improved since the days of the first human genome assemblies generated by Celera Genomics [9] and the Human Genome Project [10] (the latter now being the accepted reference). Over the years that followed, numerous revisions were made to this reference filling in gaps and making corrections [11]. However, shortcomings remained. Ultimately, obtaining a high-resolution genome remains prohibitively expensive, requiring considerable computational resources, integration of different sequencing technologies, and a significant workforce to achieve a truly complete reconstruction. For instance, finally completing the human genome by the Telomere-to-Telomere consortium took the combined effort of nearly 100 scientists over three years, utilizing the latest technological advances [12]. At the same time, high-throughput sequencing platforms have become increasingly ubiquitous and affordable [13], with a mature set of computational tools available to facilitate analysis. These factors combined have resulted in the most comprehensive reference genomes becoming embedded in the field, best exemplified by the human reference genome [14], which due to its enduring nature, is now the backbone of nearly all human DNA studies. Ultimately, this has resulted in alignment to these references using high-throughput sequencing data, becoming the most accessible form of sequence analysis.

Reference genomes impose a notion of consistency across studies, so that any knowledge gained from a specific reference can be generalized, such as annotations. However, this simultaneously reinforces its use in the field. Reference-specific annotations often describe regions of interest within the genome and may include, among many others: gene locations, conservation scores, or sequence uniqueness [15]. Such annotations lead to difficulties when a reference is updated — by introducing a new sequence or modifying an existing one — since results based on a specific genomic coordinate system may not be easily translatable from one reference version to another. This is one of the notable challenges of moving between human references, such as GRCh37 ⟶ GRCh38, and beyond [16]. While many annotations can be lifted from GRCh37 to work on GRCh38, in some cases, they cannot be lifted, and new annotations must be generated entirely [17]. Other changes to the genome assembly may also cause difficulties in specific applications, and tests must be performed to ensure that the new reference does not deviate from previously published results. For example, the new content of GRCh38 added to chromosome 21 contained sequences that replicated the sequence at locus U2AF1, resulting in a failure to detect variations in this region that were previously detectable in GRCh37 [18] and requiring intervention in the assembly to correct this failure [19]. In some applications, stable solutions rely on assembly-specific features, and in some cases, the changes to the reference are so substantial that the costs do not outweigh the benefits of switching to the new reference [20].

The utility of a reference genome is based on the assumption that any sequence must be of sufficient similarity for the reference to match it. This means that sufficient context must be available for a sequence to be matched, which becomes increasingly difficult as sequences become shorter [21], as is the case with high throughput sequencing. Therefore, the reference genome chosen directly constrains the space over which sequences can be aligned; when a sequence is too dissimilar or absent from the reference, it will either be misaligned or completely unaligned. Even if another more closely related reference is used, it will inevitably return a different set of misaligned and unaligned sequences. This effect is known as reference allele bias [22–24] and follows from the evident observation that sequences more similar to the reference are more likely to match. Dissimilar sequences are likely to have more alleles not present in the reference, which further complicates sequence alignment because it increases uncertainty about the origin of a sequence. Although this effect can be addressed by deeper sequencing, so that sufficient coverage is obtained for specific alleles, this is not sufficient if the dissimilar sections exceed the length of the sequenced fragments.

Reference bias also introduces a false sense of similarity, as a dissimilar organism may appear more similar to the reference than it actually is due to unaligned sequences. The amount of unaligned and partially aligned sequences can approximate this dissimilarity; however, sampling differences, such as sequencing yield, complicate this measure. The effects on downstream analysis are apparent, where variant calling will detect more variants for increasingly dissimilar samples. The 1000 Genomes Project shows an example of this reference bias [25, 26], given the human reference genome (GRCh37), whose base is predominantly European individuals [27, 28], fewer variants were detected from Europeans and more were found in Asians and Africans. More (complex) variations could be captured in Europeans because they are more similar to the reference. In Asians and Africans, on the other hand, more variation was found because of their dissimilarity, especially in the case of simple variations such as single nucleotide polymorphisms (SNPs). The problem with this dissimilarity is that it could prevent us from finding all the large variations within these latter groups. Reference bias increases proportionally to the size of the variant—and is even observed for small variations such as SNPs [29]. Technological advances in sequencing have enabled the generation of increasingly longer reads with lower error rates that span larger regions of the genome, providing more context for reliable alignment and effectively reducing reference bias. At the same time, the cost of high-throughput short-read sequencing has plummeted, and this trend is likely to continue, so this technology will be used for many years to come. Waiting until long-read sequencing becomes the norm would be impractical. Even if error-free long reads were generated that could cover every repetitive segment in a genome, any analysis would still be biased by the reference representation chosen. A single reference cannot account for all variation in the population, and any method that relies in some way on such a reference will be affected by reference bias, resulting in variation not being accounted for. Organisms with a high degree of ploidy are often represented only by a haploid reference, resulting in an often oversimplified reference system, and features such as natural variation and copy number variation will be missed. The human reference genome is an example of this. It is not representative of the entire human population and, as a result, the current reference genome includes alternative loci that diverge within the human population [11, 30]. However, there is no universally accepted form of unifying these alternative sequences in the context of a linear reference genome. Although some methods use the alternative loci as additional targets during alignment [31–33], mapping to the original coordinate system then becomes a challenge, especially in downstream analyses such as variant calling [34].

One method of combating reference bias is through reference-free de novo genome assembly, which incorporates only the sampled sequence into the assembled sequence. However, genome assembly itself presents many more challenges and is, in almost all cases, too expensive or impractical to adopt. Although genome assembly may be unrealistic for general purposes, some aspects are translatable to the alignment domain [35, 36]. During the assembly process, intermediate data structures capture the variation between sequences across the genome, capturing the ambiguity inherent in the sampled sequences before emitting an often haploid sequence. These data structures are typically graphical in nature, such as overlap graphs [37], de Bruijn graphs (DBGs) [35], and string graphs [36]. In most cases, such assembly graphs represent reads as nodes linked by (bi)-directional edges that derive from a weighting factor, for instance, the number of observations associating these nodes, established through pairwise read alignment [38], k-mer hashing [39], or min-hashing [40]. Ultimately, the task of assemblers is to reduce the complexity of a graph into a consensus sequence by collapsing connected nodes according to a set of rules. The principle of representing the initial variation between all reads in a graphical representation for genome assembly can readily be translated to the reference genome representation, which is now known as a population graph or pangenome [41, 42]. Instead of representing the variation between reads as nodes, each node in the graph relates to the variation observed in the population, and the edges describe how that variation is linked within the population by imposing specific orderings of node traversals *i.e.*, paths that spell different combinations of sequences through the graph. The graph structure substantially affects the implementation details within the alignment pipeline, especially regarding computational optimization. The choice particularly relates to how the graph represents certain types of variations. For instance, if a graph allows for cyclicity, repetitious sequences may be addressed non-redundantly by allowing edges to loop around. In contrast, such repeats would have to reappear in duplicated nodes within a directed acyclic graph. In (compressed) DBGs, this concept of non-redundancy is taken a step further by encoding shorter sub-sequences within the genome [43]. While reference compression enables more efficient memory usage, this also creates challenges in how to query the reference during alignment. *Indexing* is a process that builds a subsequence-to-location mapping on the reference, which is essential for efficiently querying sequences. Given an index, subsequences may be sampled from the to-be-aligned reads and matched to the index to thin out the number of potential alignment targets on the reference. Popular indexing methods based on sorting, such as the Burrows-Wheeler Transform [44], rely on making sequences uniquely addressable, which becomes challenging in such compressed graphs.

The increasing affordability of sequencing has expanded the depth of known variation in the population to the point where unification of this knowledge into a single substructure is warranted. The first step toward such a goal is to build a population-aware structure, which enables the aggregation of existing information while also providing a logical structure to support the integration of new information. If a genome is already assembled, it can be used as a reference genome and framed to simplify the problem of comparing existing and new genomic sequences. The significant costs associated with the need to sequence and assemble can then also be offloaded by using the assembly as a reference. This requires much lower sequencing coverage and shorter reads. However, doing so also exposes you to reference bias, which causes the sampled genome to be distorted from the chosen reference. By extending the reference genome to form a population graph, a population-aware alignment becomes possible, avoiding the reference bias inherent to linear reference genomes.

Part I of this thesis presents how a linear reference model can be augmented and transformed into a graphical representation, which can then account for population variation. The CHOP algorithm is described, which overcomes the challenge of exponential growth when indexing paths in graphs by restricting them to haplotypes, thus reducing the graphical representation to a quasi-linear representation retaining only observed haplotypes. The existing library of linear alignment tools can then use this reference representation as a target. Alignments to this representation reduce the effect of reference bias and allow access to larger portions of the genome, outperforming alignments to linear reference genomes. In benchmarks, our method is more scalable than others and is unaffected by problems such as exponential growth due to the determination of haplotype-constrained paths through the graph. Variants called from these graph alignments can also be reintegrated into the original graph, expanding it to access an even larger portion of the genome.

2

CHOP: Haplotype-aware path indexing in population graphs

This chapter is based on 🗎 *T. Mokveld, J. Linthorst, Z. Al-Ars, H. Holstege, and M. Reinders. CHOP: haplotype-aware path indexing in population graphs, Genome Biology'20* [45].

Abstract

The practical use of graph-based reference genomes depends on the ability to align reads to them. Performing substring queries on paths through these graphs is at the heart of this task. The combination of increasing pattern length and encoded variations inevitably leads to a combinatorial explosion of the search space. Instead of heuristic filtering or pruning steps to reduce the complexity, we propose CHOP, a method that constrains the search space by exploiting haplotype information, bounding the search space to the number of haplotypes encoded in order to avoid a combinatorial explosion. We show that CHOP can scale to large and complex datasets by applying it to a graph-based representation of the human genome, which contains the 80 million variants reported by the 1000 Genomes project.

2.1 Introduction

Pangenomes and their graphical representations have become widespread in the domain of sequencing analysis [42]. This adoption is partly driven by the increased characterization of genomic diversity within species. For instance, recent versions of the human reference genome (GRCh37 and up) include sequences of alternative loci representing highly polymorphic regions in the human population [14].

A pangenome can be constructed by integrating known variants into the linear reference genome. In this way, a pangenome can more accurately integrate the sequence diversity of the population than a typical linear reference genome. For example, aligning reads to a linear reference genome can result in an overrepresentation of the reference allele. This effect, known as reference allele bias, primarily influences highly polymorphic regions and regions that are absent from the reference [22, 23]. Incorporating variants into the alignment process can reduce this reference bias [46–48]. As a result, downstream analysis, such as variant calling, can be improved, with fewer misaligned variants induced by misalignments around indels and fewer missed variants [49].

Pangenomes can be represented intuitively in graphical data structures, often called population graphs [41, 42]. Population graphs can be understood as compressed representations of multiple genomes, with sequences (in some cases, both complements) typically represented on the nodes. These nodes are, in turn, connected by (bi)-directional edges so that the original sequence of any genome used to construct the graph can be determined by a specific path traversal in the graph. Alternatively, an arbitrary path traversal will result in a mixture of genomes.

A key application of reference genomes is read alignment. Most linear reference read aligners follow a seed-and-extent paradigm, wherein exact matching substrings (seeds) between the read and a reference are used to constrain a local alignment. Indexing data structures facilitate the efficient searching for exact matching substrings (seeding). The construction of these indexes is usually based on one of two methods: hashing-based indexing, which can be k-mer-based, where all substrings of length k of the reference are stored in a hash-map along with their positions [50, 51]; fingerprinting-based hashing that allows for finding candidate alignment positions as a nearest neighbor search approximating the Jaccard set similarity using MinHash [52, 53]; and sorting-based methods such as the Burrows-Wheeler Transform (BWT) [44, 54], where the reference sequence is transformed into a self-index that supports the lookup of exact-matching substrings of arbitrary length.

These existing indexing methods can be extended to population graphs, though this is a challenge. Graphs can encode a variable number of genomes, which is accompanied by an exponential growth in the number of paths in the graph as more variation is incorporated. Therefore, indexing sequences of arbitrary length is challenging, and indexing must generally be restricted to shorter substrings to minimize the combinatorial growth of the index. In addition, sorting-based indexing methods that rely on suffix determination and sorting are often infeasible in graphs since there will be multiple valid node orderings.

Several approaches have been developed that perform read alignment onto population graphs using indexes that report all k-length paths in the graph. Early examples of this include: GenomeMapper [55], which builds a joint k-mer hash-map that combines a collection of genomes in order to lookup seeds and subsequently align reads using banded dynamic programming; BWBBLE [56], which linearizes the population graph using IUPAC encoding for SNPs and describes indels with flanking sequences as alternate contigs, after which it applies the BWT for indexing; In Vijaya Satya et al. [57] an enhanced reference genome is generated from HapMap SNP-chip calls, wherein variants are encoded in read length segments used as alternative alignment targets alongside the reference genome. Yet, these methods are orders of magnitude slower than linear reference genome aligners or restricted to only small genomes. For instance, BWBBLE computes four times more suffix array intervals due to the expanded IUPAC alphabet. Moreover, these methods suffer from exponential growth in index space when variation density increases.

The increased scalability of population graph alignment has recently been demonstrated with Graphtyper [58], GraphAligner (designed for long reads) [59], the vg variation graph toolkit [60], and HiSat2 [61]. Graphtyper does this by first aligning reads to a linear reference sequence using BWA (as such, there remains an implicit reference bias), after which a graph-based alignment is performed on a much smaller set of unaligned or partially aligned reads. This graph-based alignment uses a k-mer hash-map of the population graph, in which exponential growth in variation-dense regions is reduced by removing k-mers that overlap too many alternative sequences. GraphAligner uses minimizers, maximal unique matches, or maximal exact matches to seed the read-to-graph alignments. Seeds are extended and aligned using a bit vector banded dynamic programming algorithm for arbitrary graphs. The vg toolkit provides general solutions for working with population graphs. To efficiently query substrings, it uses GCSA2 indexing [62], an extension of the BWT for population graphs, which supports exact query lengths up to 256 bp. Reads are aligned to graphs using a seed-and-extend strategy, returning subgraphs of the population graph (stored in a sparse graph index, xg), to which reads are then aligned using partial order alignment, a generalization of pairwise sequence alignment for directed acyclic graphs [63]. HiSat2 generates a global graph FM index plus an extensive collection of region-specific graph FM indexes, such that both local and global searching is possible. The indexing is based on the GCSA [64], the precursor to the GCSA2 index used by vg.

Graphtyper and vg index all possible paths in a population graph, in which they also cover complex regions where variation is dense. To avoid exponential growth, heuristics are used. These heuristics eliminate k-mers that cross more than a predefined number of edges or mask out subgraphs shorter than a defined number of bases. Although these techniques prevent exponential growth, they can altogether remove complex regions from the graph, resulting in a loss of sensitivity in alignment. Furthermore, they contradict one of the main aims of population graphs, namely, to address sequence variation in regions that are inaccessible through the application of a linear reference sequence. Similarly, HiSat2 filters out rare variants from the graph, thus effectively reducing the complexity of the graph, at the cost of addressing less sequence variation. An alternative solution that does not exclude complex regions would be to constrain indexing by haplotype, so only k-mers observed in the input genomes are encoded in the index. Although the above heuristics are also used in vg, the authors of vg have recently also proposed the use of haplotyping. In vg, such haplotyping is facilitated by the use of the GBWT [65, 66]. The GBWT is a graphical extension of the positional Burrows-Wheeler transform [67] that can store sample haplotypes as paths in the graph, thus allowing haplotype-constrained read alignment. However, the GBWT index must be constructed in parallel with the GCSA2 index, which will still require an enumeration of all k-paths in the graph (which will ultimately be pruned using the GBWT index). Therefore, although vg with the GBWT incorporates haplotype constraints in read alignment, during indexing the complexity is always dictated by the GCSA2 index, which explores all *k*-paths and thus grows exponentially with the amount of variation.

We present CHOP, an alternative path indexer for population graphs that uses haplotype level information to constrain the path indexing process without requiring heuristic filtering or pruning steps. This constraint eliminates the need to evaluate all *k*-paths and avoids the exponential growth of *k*-paths that other methods face. CHOP decomposes the graph into a set of linear sequences, similarly as in Vijaya Satya et al. [57] and Gunady et al. [68], so that reads can be aligned by long-established linear aligners, such as BWA or Bowtie2 [31, 69], which can then be followed up by typical downstream analysis. We show that the alignment performance of BWA when using CHOP is comparable to that of vg but that with CHOP, the alignment is faster and can scale more efficiently to very complex graphs such as those constructed from human genomes with variation data from the 1000 Genomes Project [26].

In summary, the contributions of our work are as follows: 1) CHOP decomposes a population graph into mappable sequences representing all observed haplotypes with which the population graph was constructed, 2) the haplotype constraint is implemented in a way that avoids exponential exploration of the graph, eliminating the need to filter or prune the graph in any way, so that complexity is limited by the number of coded haplotypes, instead of the number of variants or paths *k* in the graph, 3) the decomposition of the graph is time and memory efficient, and 4) by decomposing the graph into mappable sequences, it is possible to use linear aligners to map reads, with which one can benefit from fast alignment as well as build-up experience with parameter settings of these aligners.

2.2 Results

Throughout, we consider population graphs constructed from variations called per sample (haplotype) with respect to a linear reference genome or constructed from multiple sequence alignments. The graph encodes these variations so that nodes represent sequences and edges represent consecutive observed sequences. CHOP facilitates read-graph alignment, which is presented in detail in Section 2.4).

Briefly, CHOP transforms a population graph into a null graph (an edgeless graph) by a series of operations consisting of three steps: collapse, extend, and duplicate, such that nodes in the null graph contain every substring of length *k* originating from the encoded original haplotypes in the population graph. Established aligners (here, we used BWA) can then be used to align reads to these node sequences in the null graph. Subsequently, these alignments can be projected back onto the population graph, since the mapping of the node sequences in the null graph is known in the population graph (Figure 2.1).

2.2.1 Graph alignment evaluation

To evaluate CHOP and its applicability in population graph alignment, we first performed tests on *Mycobacterium Tuberculosis* (MTB) using the BWA-0.7.15-MEM read aligner [31]. MTB represents a good model for population graphs, given the high accuracy of available assemblies, the tractable genome size (4.4 Mb), and the limited degree of variation. 401

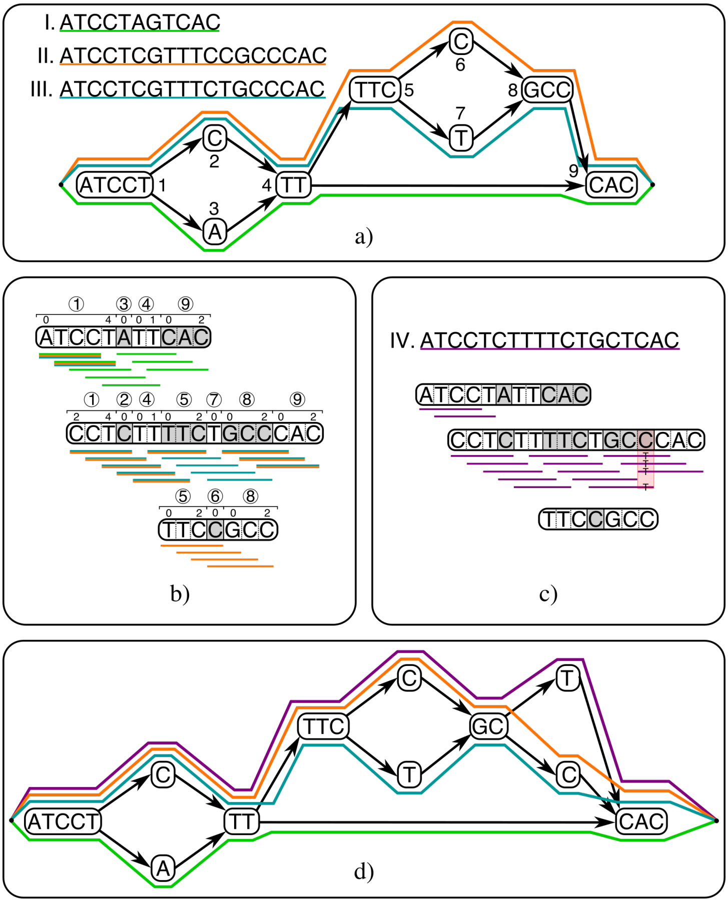


Figure. 2.1. Schematic overview of how CHOP facilitates reads alignment on a population graph. **a**) As input, CHOP accepts a graph representation of three distinct haplotypes (I, II, III). The colored paths through the graph identify the underlying haplotypes. **b**) CHOP decomposes the graph into a null graph (an edgeless graph) for substrings of length 4 (Supplemental Figure S2.1 gives all the details about the decomposition). The resulting null graph contains three nodes, and the sequence defined on these nodes covers all the substrings of length 4 that appear in the haplotypes encoded in the graph. The annotations above each node refer to intervals within nodes of the input graph. **c**) The reads (of length 4) from a new haplotype (IV) can be aligned to the null graph. Consequently, a mismatch can be called from the read pileup. **d**) With the interval definitions assigned to the null graph, the novel variant can be positioned on node 8 of the original graph and incorporated to generate a new graph.

variant call sets (VCF files) from different MTB strains (samples) were obtained from the KRITH1, and KRITH2 datasets [70, 71]. Variants were called with respect to the reference genome H37Rv, using Pilon-1.22 [72], and were filtered to exclude low-quality variants. For graph construction, we employed a leave-one-out strategy, wherein one sample was removed from the VCF file containing all 401 samples. The read set of the removed sample was then used for graph alignment. This was repeated with 10 randomly selected samples. The corresponding single-end read sets were obtained from EBI-ENA (Supplemental Section 2.5.2). To investigate the influence of introducing more variation on graph alignment, we progressively incorporated more samples from the full set into the constructed graphs, with up to 17,500 variants in the 400-sample graph (the rate of variation growth is shown in Supplemental Figure S2.2).

Because the ground truth of genomic positions in the read set data is unknown, we evaluated alignments based on the following criteria: the number of mismatches, insertions, deletions, clipped bases, unaligned reads, and perfectly aligned reads (definitions in Supplemental Section 2.5.4). These criteria allowed us to inspect the behavior of different read aligners. In order to avoid bias induced by multiple possible alignments of a single read, we considered only primary alignments.

To evaluate our haplotype-based approach, we compared CHOP to vg-1.12.1 without and with haplotyping (denoted vg+GBWT). The vg toolkit provides general solutions for population graphs, including graph construction, indexing, and read alignment. CHOP was set to index 101-length haplotyped paths (equivalent to read length) and used default parameters with BWA-MEM. To closely reflect the CHOP parameters, vg was set to index all 104-length paths (k = 13, 3 doubling steps).

Since CHOP uses BWA as an aligner while vg has its internal aligner, differences may occur based on the aligner and not the indexing algorithm. To understand the differences induced by the aligner and parameters, we first summarized the results of the 10 hold-out samples on the linear reference genome, presented in Table 2.1 for BWA and vg. Both aligners resulted in nearly the same number of perfectly aligned reads. However, alignments with vg resulted in fewer unaligned reads (−22.30%) and more mismatches (+4.01%) than BWA. We attribute this difference to the increased sensitivity with which vg aligns reads. This is reflected by the increase in clipped bases (+22.79%), inserted bases (+29.36%), and deleted bases (+34.53%), allowing vg to align shorter read fragments.

Using these alignment measurements as a baseline, the read-to-graph alignments were compared between CHOP/BWA and vg. The different graph construction methods of CHOP and vg were found to have minimal effect on the alignments as shown in Supplemental Figures S2.5 and S2.6. Figure 2.2 shows the increase in perfectly aligned reads using both CHOP/BWA and vg as more samples are incorporated into the graph (similar plots for the number of unaligned reads and mismatches can be found in Supplemental Figures S2.7 and S2.8). Table 2.1 shows the alignment results for the MTB graph with 400 samples.

Table. 2.1. Mean of alignment results across all 10 hold-out sample alignments to 1) the reference genome H37Rv (H37Rv columns) and 2) the 400 MTB genomes graph (Graph columns) for CHOP/BWA, vg with and without haplotyping, and HiSat2 to align the reads (note that when aligning only to H37Rv, CHOP is not used).

	All TB hold-out samples — Read length = 101										
Alignment criteria	BWA H37Rv	CHOP/BWA Graph (n=400)		vg H37Rv	vg Graph (n=400)		vg+GBWT Graph (n=400)		HiSat2 H37Rv	HiSat2 Graph (n=400)	
Reads aligned	6,160,920	6,162,033	(+0.018%)	6,241,270	6,245,907	(+0.074%)	6,244,004	(+0.044%)	5,536,194	5,489,149	(-0.850%)
Reads unaligned	360,236	359,123	(-0.309%)	279,886	275,249	(-1.657%)	277,152	(-0.977%)	984,962	1,032,007	(+4.776%)
Reads perfectly aligned	4,048,774	4,142,052	(+2.304%)	4,048,774	4,153,217	(+2.580%)	4,153,124	(+2.577%)	4,056,850	4,113,818	(+1.404%)
Bases aligned	596,380,132	596,611,260	(+0.039%)	599,244,753	599,601,399	(+0.060%)	599,528,267	(+0.047%)	553,338,901	548,757,802	(-0.828%)
Bases unaligned	62,191,423	61,960,355	(-0.372%)	59,307,655	58,949,429	(-0.604%)	59,023,102	(-0.480%)	105,271,191	109,852,201	(+4.352%)
Bases unaligned from clipped reads	22,349,569	22,380,472	(+0.138%)	27,442,533	27,690,552	(+0.904%)	27,575,464	(+0.484%)	3,625,336	3,589,573	(-0.986%)
Bases mismatched	3,458,029	3,308,480	(-4.325%)	3,596,667	3,458,707	(-3.836%)	3,455,296	(-3.931%)	2,164,703	2,029,911	(-6.227%)
Bases inserted	65,210	65,151	(-0.090%)	84,358	85,938	(+1.874%)	85,397	(+1.232%)	26,674	26,763	(+0.334%)
Bases deleted	52,272	51,165	(-2.118%)	70,324	72,082	(+2.500%)	70,347	(+0.033%)	11,793	11,756	(-0.317%)
Non-primary alignments	246,092	246,540	(+0.182%)	539,309	724,904	(+34.414%)	724,613	(+34.360%)	969,452	755,436	(-22.076%)
Time (s)	533	721		10,711	4,457		4,540		312	517	

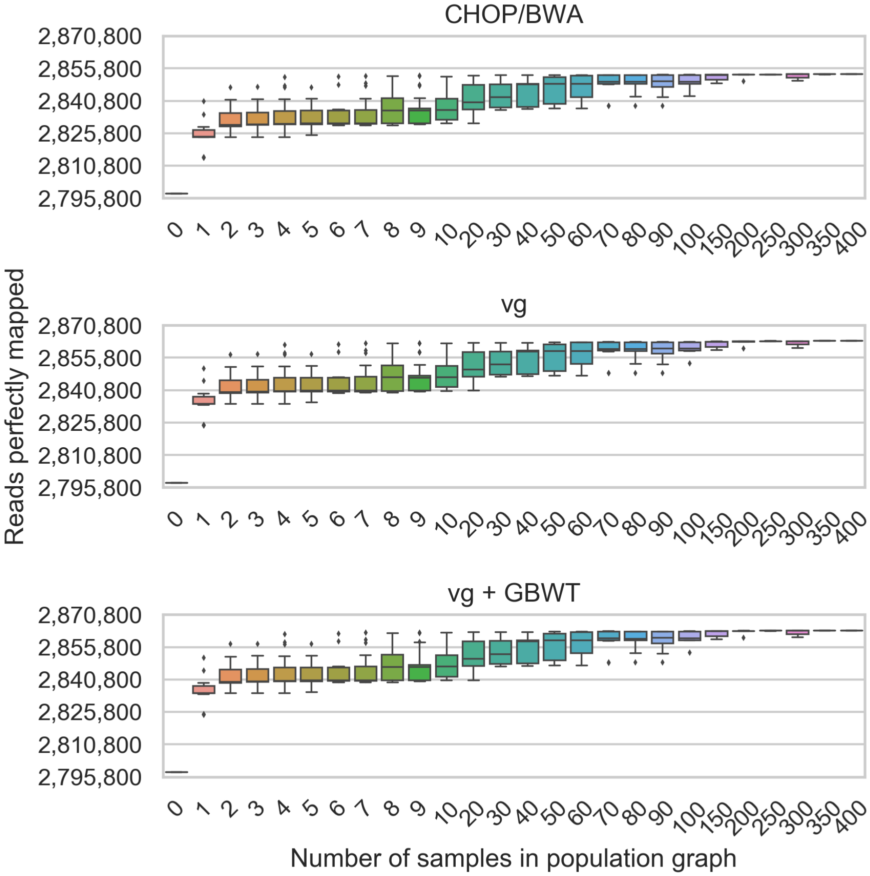


Figure. 2.2. Perfectly aligned read count for SRR833154 alignments to different sized population graphs, containing between 0 (only H37Rv, the linear reference) and 400 samples for both, when using CHOP/BWA and vg with and without haplotyping to align reads to the graph.

Figure 2.2 shows that incorporating more variation from samples into the population graphs increases the number of aligned bases, which is further demonstrated in Supplemental Figures S2.7 and S2.8. Spread is a consequence of sampling when constructing the population graphs, where samples closely related to the hold-out sample will give a more significant improvement than distantly related samples, which is demonstrated by the reduction in spread as the sample size increases.

We can observe the effects of haplotyping by comparing the number of aligned reads for vg and vg+GBWT. Since the indexing space was constrained to haplotypes only, a decrease in the number of aligned reads is expected.

The baseline alignments to H37Rv already highlighted that the aligners perform differently. However, throughout the course of the experiments, almost all alignment criteria show the same trend for both CHOP/BWA and vg. The exception is the number of unaligned reads, steadily decreasing with vg; this is not as pronounced when using CHOP/BWA. To better disentangle the aligner-specific differences of CHOP/BWA and vg, we directly compared CHOP and vg+GBWT by aligning to the CHOP null graphs using vg (denoted as CHOP/vg), as described in Supplemental Section 2.5.7. Although we observe differences in alignments between CHOP/BWA and vg+GBWT, these are merely due to differences in aligners. This was confirmed when comparing vg+GBWT with CHOP/vg which has shown nearly identical alignments (Supplemental Table S2.2). Alternatively, the alignment differences between CHOP/BWA and vg+GBWT could be minimized by optimizing the aligner parameters, as we only used the default settings for both.

In a similar setting we compared to HiSat2 (Supplemental Section 2.5.8), results presented in Table 2.1. Although HiSat2 aligns much faster than CHOP/BWA and vg(+GBWT), this can be attributed to its lower sensitivity, having many more unaligned reads in both baseline and graph alignments. Surprisingly the number of unaligned reads increases in the alignments involving the linear genome, while the number of non-primary alignments decreases. This may indicate that not all sequence in the graph is indexed.

Together these experiments show that as more and more genomes are populating a variation graph, 1) more reads can be aligned (with fewer mismatches), 2) constraining the alignment by haplotype does not negatively affect alignment, and 3) that as expected, the two haplotype constrained aligners (CHOP/BWA and vg+GBWT) have similar performance.

2.2.2 CHOP scales to *Homo sapiens*

To further evaluate the scalability and sensitivity of CHOP, we used chromosome 6 (170 Mb) of the GRC37 assembly in combination with the 1000 Genomes Phase 3 variation data [26]. The constructed graph has 14,744,119 nodes and 19,770,411 edges and encodes a total of 5,023,970 variants (4,800,102 SNPs, 97,923 insertions, and 125,945 deletions). Note that the variation set included diploid phasing of 2,504 individuals, which was incorporated into the graph as 5,008 paths (2 paths per sample) and one path representing the reference genome. Within the population, most variation (58.42%) is shared between at least two or more individuals (Supplemental Figure S2.9). For the graph alignments, we used 15 single-end read sets from 1000 Genomes Phase 3 (Supplemental Table S2.3) that were filtered to include only reads aligned to chromosome 6 or that could not be aligned anywhere on the genome (average read set size of 3,026,069).

CHOP was configured to index 100-length paths through the graph, matching the read length, which yielded 11,359,686 nodes in G^E. Memory utilization and indexing time were dominated by CHOP, with BWA indexing accounting for only 6.95% of the indexing time, requiring a fraction of the memory. We attempted indexing with vg and vg+GBWT for paths of up to 104 bp (k = 13, 3 doubling steps), but this was unsuccessful due to memory constraints (500 GB). Instead, doubling was lowered to 2, and paths up to 52 bp were indexed. By incorporating haplotyping in vg, the indexing requires significantly more time (6x longer) than indexing without haplotyping, while memory utilization remains constant. The read sets were aligned to both the linear reference of chromosome 6 and the graph representation, using either CHOP/BWA, vg, or vg+GBWT, which is summarized in Table 2.2.

We observed the same improvement when switching to a graph representation as in MTB, although more extensive, as more variants, including indels, are incorporated into the graph. Given the different path lengths, the time cannot be directly compared between

Table. 2.2. Mean of alignment results from 15 samples from the 1000 Genomes data when aligning to 1) the reference genome sequence of chromosome 6 (column GRC37) and 2) the population graph created from the 5,008 haplotypes, for both CHOP/BWA, vg with and without haplotyping, and GraphAligner.

	1000 Genomes samples — Read length = 100										
Alignment criteria	BWA	CHOP/BWA		vg	vg		vg+GBWT		GraphAligner	GraphAligner	
	GRC37	Graph (n=2504)		GRC37	Graph (n=2504)		Graph (n=2504)		GRC37	Graph (n=2504)	
Reads aligned	2,542,399	2,543,522	(+0.044%)	2,684,925	2,726,051	(+1.532%)	2,717,972	(+1.231%)	2,664,609	2,630,670	(-1.274%)
Reads unaligned	483,670	482,548	(-0.232%)	341,144	300,018	(-12.056%)	308,098	(-9.687%)	361,460	395,399	(+9.389%)
Reads perfectly aligned	1,794,564	1,977,952	(+10.219%)	1,807,158	1,993,967	(+10.337%)	1,993,469	(+10.310%)	1,789,327	1,950,435	(+9.004%)
Bases aligned	251,122,992	251,516,725	(+0.157%)	254,518,323	255,911,471	(+0.547%)	255,578,370	(+0.416%)	258,684,466	256,773,126	(-0.739%)
Bases unaligned	51,439,949	51,070,534	(-0.718%)	48,029,518	46,654,663	(-2.863%)	46,995,159	(-2.154%)	41,030,266	43,649,514	(+6.383%)
Bases unaligned from clipped reads	1,801,947	1,846,687	(+2.483%)	12,716,162	15,699,089	(+23.458%)	15,245,035	(+19.887%)	203,221	177,659	(-12.578%)
Bases mismatched	1,270,981	969,087	(-23.753%)	1,198,917	953,800	(-20.445%)	940,371	(-21.565%)	4,681,045	3,931,955	(-16.003%)
Bases inserted	43,979	19,661	(-55.296%)	59,078	40,786	(-30.962%)	33,391	(-43.480%)	2,541,719	1,925,093	(-24.260%)
Bases deleted	61,659	32,355	(-47.526%)	73,131	44,555	(-39.075%)	41,040	(-43.882%)	464,085	415,508	(-10.467%)
Time alignment (s)	544	1,807		19,996	10,436		10,871		916	2,102	
Memory alignment (MB)	412	5,534		837	3,296		4,389		3,047	14,446	
Time indexing (s)	186	CHOP: 43,625; BWA: 3,256		37	5,751		33,619		NA	NA	
Memory indexing (MB)	245	CHOP: 56,969; BWA: 3,813		269	45,670		45,868		NA	NA	

CHOP/BWA and vg. Nevertheless, it is unclear why vg took substantially more time to align than CHOP/BWA, especially when aligning to the linear reference. The differences (relative to MTB) between vg and vg+GBWT became more prominent as more samples are incorporated into the graph. Note that vg+GBWT is slower than vg in both indexing and alignment. This is because the GBWT index, used in vg+GBWT, is built and used alongside the GCSA2 and xg indexes already present in vg. Therefore, the gain of the GBWT index is mainly to correct the alignment process by adding haplotype constraints.

We observed substantial differences between CHOP/BWA, vg, and vg+GBWT regarding the decrease in unaligned reads −0.23% versus −12.06% and −9.69%, and the increase in read clipping +2.46% versus 23.46% and 19.89%, respectively. To evaluate this aligner-induced difference, we extracted all reads aligned exclusively onto the graph, which represents 21,661 reads in CHOP/BWA and 616,900 in vg. Supplemental Figure S2.10 displays the distribution of the number of aligned bases for each of those reads. Almost all (97.61%) of the newly vg-aligned reads had a length of 15–30 bp induced by clipping or extensive base insertion/deletion. However, at 30 bp and above, the aligners display very similar profiles, with a comparable number of newly aligned reads. At 69 bp, both aligners show a peak, and the newly aligned reads corresponding to this peak all align to the same region of the graph. This region closely resembles human mitochondrial DNA, which was excluded from the initial reference alignments. This has led to an increased number of unaligned mitochondrial sequencing reads in the graph-aligned dataset (Supplemental Section 2.5.11).

By simulating chromosome 6 read data, we measured the accuracy of alignments on graphs and linear references. We observed that by constructing graphs from subsets of available variants (selected based on allele frequency in the population), alignment performance could be improved (Supplemental Section 2.5.12). We observed similar improvements when aligning reads to a graph built from the multiple sequence alignment of alternate alleles from the MHC region of chromosome 6 (Supplemental Section 2.5.13).

In addition, we compared CHOP/BWA to Graphtyper (Supplemental Section 2.5.14). Since Graphtyper's primary purpose is genotyping and variant calling (and thus does not output alignments), we also called variants from the CHOP/BWA alignments. Although Graphtyper did not detect any new variants when aligning reads from sample HG00308 to the chromosome 6 1000G graph, it did genotype variants (144,800 out of 5M, after filtering). Contrarily, CHOP/BWA detected 1,212 variants, of which 57 remained after quality filtering. Note that calling variant from the CHOP/BWA output was more than two orders of magnitude faster than Graphtyper while using an order of magnitude less memory.

We failed to index the graph with HiSat2 due to extreme memory utility (>200GB within 709 seconds) and concluded that it does not scale to population graphs of this complexity (Supplemental Section 2.5.8). We also compared to the long read aligner GraphAligner (Supplemental Section 2.5.15) and the results are presented in Table 2.2. Note that GraphAligner is optimized for long reads and could generate suboptimal results when short reads are used. GraphAligner was able to index and align on the 1000G chromosome 6 graph, with alignment times close to those of CHOP/BWA. However, the alignment statistics, similar to the HiSat2 case for MTB, show a counter-intuitive decrease in the number of aligned reads when aligning to the 1000G graph instead of the linear genome.

To better grasp the practical limitations of CHOP, we indexed the previously introduced graphs for varying k values (Supplemental Section 2.5.16), where we note an approximately linear growth in indexing time and memory usage. We also compared CHOP and vg+GBWT using simulated variation graphs with different degrees of variation, number of encoded genomes, and shared variation between genomes under defined memory and time constraints (Supplemental Section 2.5.17). Figure 2.3 highlights the differences in indexing times of CHOP and vg+GBWT for simulated graphs with samples that encode 1,000 variants each. We show that CHOP indexes faster and more efficiently than vg+GBWT and can handle more complex graphs (CHOP could index 92.75% of all simulated graphs, while vg+GBWT was able to index 79.28%).

Finally, we performed alignments on the complete human genome. We constructed graphs of each chromosome encoded with the variants reported by the 1000 Genomes project Phase 3. Cumulatively, these graphs have 248,677,280 nodes, 33,3561,973 edges and encode a total of 84,745,123 variants (81,382,582 SNPs and 3,362,541 indels). We indexed the graphs with both CHOP for 100-length paths and vg+GBWT for 52-length paths; the peak memory utilization and time required for indexing are shown in Figure 2.4. Note that chromosomes 1, 2, 11, and X could not be indexed with vg+GBWT due to the complexity of the graphs (sometimes more than 50 variants in a 50 bp window), resulting to excessive memory (>500 GB) or disk (>6 TB) utilization, more details in Supplemental Section 2.5.18. For vg to be able to handle these chromosomes, the graphs would have to be simplified prior to indexing. Indexing with CHOP yielded 103,509,254 nodes in G^E, which increased the total sequence space by 14x. We again used BWA and aligned the sample ERR052836 to both the linear reference genome and the graph, where we noted a 2–3x (13,704 to 37,826 seconds) increase in read alignment time to the graph relative to the linear genome.

2.2.3 Variation integration

As graphs cover a larger search space, we investigated how this affects read alignment and variant calling. Theoretically, encoding more distinct sequences in a graph should enable the alignment of more reads and potentially allow the calling of new variants. To evaluate this, variants were integrated using a feedback loop. First, SRR833154 reads were aligned to H37Rv using BWA, and then variants were called using Pilon. Variants were quality filtered down to 838 SNPs and then used to construct a graph with H37Rv (now including two genomes). The same set of reads was then aligned onto the graph, and variants were called. We expected that the additional context offered by the graph would yield previously undiscovered variants. An example of this is schematically shown

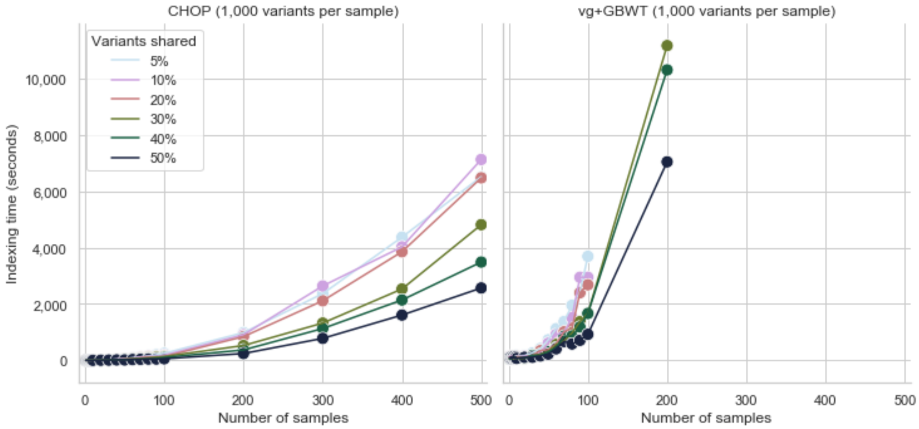


Figure. 2.3. CHOP and vg+GBWT indexing time (seconds) of graphs with increasing numbers of encoded samples, where each sample contributes 1,000 variants to the graphs. The coloring indicates different probabilities of sharing variants within the simulated population. For instance, with a probability of 5%, 95% of all sample variation will be unique to that particular sample, while the remainder is shared with one or more other samples. Missing dots in the plots indicate that the indexing failed by either exceeding 4 hours of computation time or the maximum memory of 80 GB. More details can be found in Supplemental Section 2.5.17.

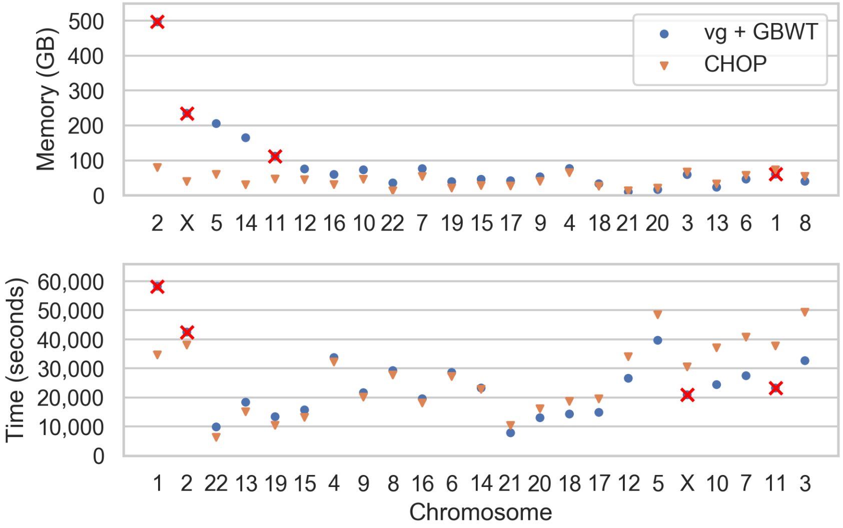


Figure. 2.4. Peak memory footprint and time required for indexing the human chromosomes using CHOP and vg+GBWT. Chromosomes are ordered according to the relative differences between CHOP and vg+GBWT. Chromosomes 1, 2, 11, and X are crossed out for vg, given that these exceeded the memory (>500 GB) or disk space (>6 TB) constraints.

in Figure 2.5a, in Supplemental Figure S2.19 we show an example of such newly aligned reads.

Integrating variants in a graph (Figure 2.5b) and realigning reads to the graph allowed

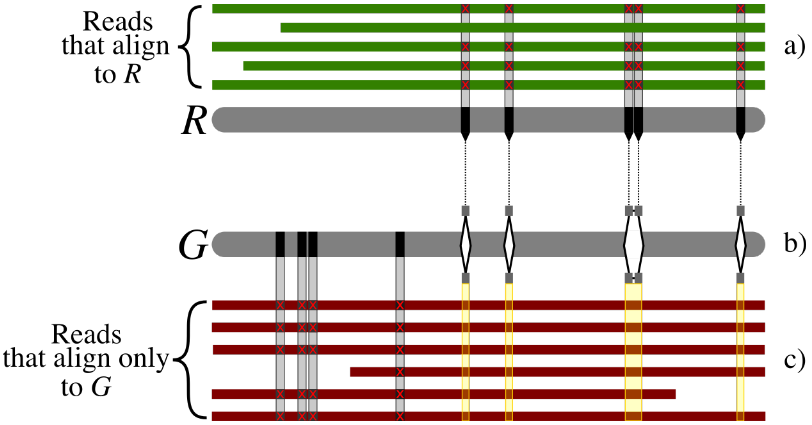


Figure. 2.5. a) Schematic alignment of SRR833154 reads to the reference R, H37Rv, with subsequent variant calling detecting 5 high-quality SNPs in this particular region. **b)** These and all other SNPs across the genome are integrated with the reference into graph G, followed by alignment of the same reads. **c)** Reads that previously did not align to R now align onto a haplotype of the graph G. The formation of a pileup allows for the detection of 4 new variants in this region.

reads to follow a path within the graph that best matches. This, in turn, allowed for reads that previously were not able to align now to be aligned. (Figure 2.5c). As a result, 19 $(+2,26\%)$ new high-quality variants could be called from these newly aligned reads.

2.3 Discussion

Population reference graphs that take into account within-species genetic diversity can potentially improve sequencing analyses by providing a more accurate alignment of sequencing reads. This, in turn, can improve various downstream analyses, like variant calling.

A challenge for efficient read-to-graph alignment is to find exact matching seeds of a fixed length k that can span the edges of the graph. Searching through an enumeration of all possible k-length paths in the graph is computationally challenging, as the exponential growth of paths negatively affects the memory footprint and indexing time, which limits the amount of variation that can be encoded in the population graph. We suggest the use of haplotype information to prevent this exponential growth. In doing so, the genetic linkage between neighboring variants can be exploited to counteract computational problems and the number of false positive matches that occur due to unobserved combinations of variants (variants encoded on different alleles). Recently Ghaffaari and Marschall [73] has proposed an alternative approach to circumvent the computational challenge of exponential path growth in graphs by combining a graph index with read chunk indexes, exploiting the limited k-mer space of reads relative to that of the graph. It will undoubtedly be interesting to see how a haplotyping approach can be combined with this method.

Here we introduced CHOP, a method that converts a haplotype-annotated population graph into a set of sequences that spans all haplotyped k-paths. It does this by transforming the population graph into a null graph (a graph with no edges) such that each observed *k*-path is represented by one of the resulting unconnected nodes. Since the resulting set of sequences (the null graph) is a compressed representation of all haplotyped *k*-paths through the graph, it becomes feasible to use values for *k* that are equal to the length of a typical NGS read (*e.g.*, 100 to 150). For this reason, an additional advantage of CHOP is that any NGS read aligner (*e.g.*, BWA or Bowtie) can be used to align reads onto the created null graph. Since each position in the null graph can be translated back to a position in the original population graph, we can effectively perform a scalable read-to-graph alignment. The advantage of CHOP over existing graph-based alignment approaches is that we propose a solution to incorporate the haplotype constraint throughout the whole procedure and thus truly do not suffer from a combinatorial explosion of possible paths since complexity is bounded by the number of haplotypes encoded in the graph. Through this solution, CHOP does not have to rely on filtering or pruning the graph to scale to complex population graphs.

With CHOP, we followed an approach more closely related to string/overlap graphs instead of a de Bruijn graph (DBG) approach, as they can better handle cycles induced by repetitive sequences. There is a crucial difference with a (compressed) DBG approach because a DBG is always constructed for a fixed value of *k*. For the index to remain manageable, this *k* value must be relatively small. However, for a small value of *k* (commonly ~15 is being used), the resulting DBG will contain cycles introduced by repeated *k*-mers. These cycles prevent parts of the genome/graph from being uniquely addressed. This, while the input data structure (the variation graph), is unambiguous. Representing the variation graph as a DBG, therefore, inevitably results in a loss of information. To further clarify this difference, we can state that with CHOP, each position in the variation graph corresponds to at least one unique position in the index (null graph), whereas with a DBG approach, multiple positions in the variation graph can correspond to the same position in the DBG, also exemplified in Supplemental Figure S2.20. The size argument for DBGs follows from the fact that these repeated *k*-mers are stored only once, which is exactly what introduces the ambiguity in the first place. Therefore, the higher the compression rate, *e.g.*, using bloom filters, the more ambiguity is introduced in the representation. Bloom filters are probabilistic data structures that balance the need to store these vast hash tables against the integrity of the resulting representation (as they allow for colliding hash functions, *e.g.*, edges in the DBG). Although these representations are a computational answer to the need to store and query vast hash tables (*e.g.*, DBGs with ‘large’ *k* values), they further impair the representation of the underlying variation graph by allowing for non-existent edges.

We have shown that read alignment using CHOP in combination with the aligner BWA (CHOP/BWA) easily scales to the entire human genome, encompassing the 84,745,123 variations reported by the 1000 Genomes project (2,504 individuals). The CHOP memory footprint per human chromosome when indexing is less than 80 GB and takes less than 50,000 seconds.

In addition, we have shown that graph indexing and alignment with CHOP/BWA resulted in more aligned bases than alignment to the linear reference genome. We also found that the number of aligned bases increased proportionally to the number of incorporated variants (samples). Interestingly, the amount of sequence required to store the resulting compressed *k*-paths grew faster than the time needed to perform the alignments. We attribute this to the increase in exact matching reads, which decreases the need to expand the initial seeds during alignment, a computationally demanding task.

We extensively compared CHOP/BWA to vg, the current state-of-the-art toolkit for working with population graphs. The GCSA2 indexing extension of the BWT supports exact query lengths of up to 256 bp, which enables vg to index and query population graphs. Recently, vg has been expanded to facilitate haplotype constrained alignment using the GBWT index, a graph extension of the positional Burrows-Wheeler transform. However, vg still requires the construction of the GCSA2 index with the GBWT to perform haplotype constrained alignment, which still risks the exponential path growth during indexing, an issue that does not occur with CHOP.

When comparing read alignments of CHOP/BWA with vg and vg with haplotyping (vg+GBWT) on population graphs of both *Mycobacterium tuberculosis* (MTB) and humans, we found very similar alignment results, as expected. However, compared to CHOP/BWA, alignment took 5–6 times longer with vg(+GBWT). Furthermore, CHOP scaled better with complex graphs, which we demonstrated by indexing and aligning to a graph of the complete human genome. Although vg and vg+GBWT were able to index most of the chromosomes, this was only possible when we adapted path lengths of $k = 52$, which was approximately half the length of CHOP k-paths. Then, still, for some complex chromosomes, indexing failed using vg. Moreover, we showed the scalability of CHOP for this particular graph with k-paths up to $k = 300$ (Supplemental Figure S2.13).

Our comparisons with HiSat2 and GraphAligner showed that HiSat2 does not scale to the human variation graph encoding all 1000G variants, whereas GraphAligner does. We should point out that HiSat2 has been shown [74] to scale to human by first pre-selecting which variants are included in the graph, which also reduces the number of false positive alignments. CHOP eliminates the need to pre-filter variants, leaving more freedom for users to make this decision or, alternatively, drastically increasing the number of genomes with high fidelity variants in the variation graph. Moreover, for HiSat2 and GraphAligner (perhaps attributable to its optimization for long reads), the alignment results do not agree with those of vg(+GBWT) and CHOP/BWA.

Interestingly, the read alignment results did not differ much between a haplotype-constrained aligner and a non-haplotype-constrained aligner. This can be best observed when comparing vg with vg+GBWT as they utilize the same aligner and parameters. Although, the number of aligned reads increases by 1.5% when considering all k-paths (vg) in the human graph relative to the linear reference genome, as opposed to an increase of 1.2% when considering haplotype-constrained k-paths (vg+GBWT). Inspection of the additionally aligned reads indicates that most of these alignments result from spurious matches induced by unsupported sequence combinations. Overall, this seems to suggest that indexing all possible k-paths does not add much value while increasing the risk of false positive alignments. Note that non-haplotype constrained alignment could still be helpful when the genome to be aligned is expected to be more distant from the encoded genomes in the variation graphs, and therefore recombined haplotypes could guide the alignment.

The advantage of limiting k-paths to observed haplotypes is further supported by our observation that population graph alignment improves with respect to a linear reference genome when not all observed variation is incorporated into the graph (Supplemental Section 2.5.12). Our simulations on a 1000G sample showed that improved read alignments (identified by a reduced number of false positive/negative alignments) can be obtained when the allele frequency of a variant is taken into account when building the population graph. Put simply, if the frequency of a variant increases, it is more beneficial for read alignment to incorporate such variants in the population graph at a minimal cost of introducing false positives. Note that rare variants within a sample can still be called after the read-to-graph alignment; they are just not used in constructing the population graph.

The graphs that serve as input to CHOP should encode phased variant calls. Although this information is typically not encoded in variant call formats, it is required at only short ranges (related to the value for *k*) and should be readily available from typical sequencing experiments. In most of our experiments, the complexity of incorporated variation was limited to SNPs and small indels. Therefore, the benefit of a population graph on increasing the number of aligned reads was limited since SNPs and small indels are well identifiable using a linear reference genome. However, CHOP is not restricted to graphs constructed from variant calls but can handle any acyclic sequence graph, *e.g.*, as generated from multi whole-genome alignments or haplotype-aware de-novo assembly algorithms [75, 76] (Supplemental Section 2.5.13). Therefore, short-range (SNPs/indels) and long-range (structural variants) haplotypes can be incorporated into the graph and the resulting index. The incorporation of larger structural variations will result in more substantial improvements. However, it should be realized that incorporating structural variation increases the amount of repeated sequence in the graph, *e.g.*, incorporating mobile element insertions and repeat expansions, which will increase ambiguously aligned reads.

CHOP does not directly support long reads or paired-end reads. For long reads, with *k* typically exceeding >10 Kb, this will still lead to an intractable number of haplotype-constrained *k*-paths. However, the alignment of long reads generally depends on detecting short seeds in the first place, which can easily be extracted from the compressed representation of *k*-paths generated by CHOP. Therefore, long-read alignments can be seeded, where a subgraph can be extracted (based on the seeds) and aligned with partial order alignment. Alternatively, the heaviest weighted sub-path can be extracted from the graph [77], followed by a typical sparse alignment on that linear sequence. For paired-end reads, reads are aligned to discrete *k*-paths, where an aligner such as BWA cannot directly measure the distance between any distinct *k*-path. Note that read pairing should be possible based on the haplotyped paths in the graph. Namely, the distance between any two nodes in the graph will follow a distribution of distances (from each reachable haplotype), which allows the evaluation of read pairs during alignment (in a stand-alone aligner) or as a post-processing step.

Compared to Graphtyper, we showed that by using CHOP/BWA; we could detect new variants when aligning reads to the 1000G variation graph, whereas Graphtyper can genotype variants in a large population. Finally, we showed that iterative integration of aligned sequencing reads derived from one genome to the linear reference genome using the graph representation improves variant calling. Aligning additional reads led to the additional calling of variants, which could then be merged with the original population graph, reiterating the whole process multiple times. This application of population graphs is similar to iterative realignment methods, such as ReviSeq [78], but is more generally solved when using population graphs as a starting point.

2.4 Methods

2.4.1 Population graph definition

Population graphs were constructed from existing reference genomes and variation sets, called from linear reference alignments (Supplemental Section 2.5.5) or by multiple sequence alignment (Supplemental Section 2.5.13). The nodes of the graphs are labeled, encoding genomic sequences that may be shared within multiple haplotypes, which are, in turn, connected by directed edges. Traversing a sequence of edges, *i.e.*, a path, will describe an observed haplotype within the graph.

2.4.2 Population graph specification

A population graph $G = (V,\ E)$ is defined as a set of nodes $V = \{v_0,\ \dots, v_N\}$, where $N = |V|$, and a set of edges E. Each of these edges is an ordered pair of nodes $(u, v) \in E$, where node $u \in V$ is connected to node $v \in V$. As G is a directed graph, it holds that for any edge $(u,\ v) \in E$, $(u,\ v) \neq (v,\ u)$.

For each node $v \in V$, the in-degree, $in(v)$, is defined as the number of incoming edges to that node; *i.e.*, the number of distinct edges $(u,\ v) \in E$ for any $u \in V$. Conversely the out-degree of node, v, $out(v)$, is defined as the number of outgoing edges from that node.

Every node, v, is assigned a sequence of characters, S, consisting of the alphabet $\Sigma = \{A,\ T,\ C,\ G\}$, such that $v_S = S[0,\ n-1]$, wherein $S[i] \in \Sigma$ for all i, and the length of the sequence, n, is defined as $n = |v_S|$. The range of any such sequence for any node, $v \in V$, lies between $1 \leq |v_S| \leq L$, where L is the length of the largest recorded sequence. Any substring of a sequence, S, is denoted as $S[i,\ j]$. Two types of substrings in particular are prefixes $S[0,\ j]$ and suffixes $S[i,\ n-1]$, which describe the left and right flanks of any sequence S, respectively.

A path, P, where $P = u_0 \cdots u_{q-1}$, is any consecutive series of nodes, $(u_i,\ u_{i+1}) \in E$ for all $i < q$, where $q = |P|$ is the total number of nodes on the path. If a path exists between any pair of nodes in a graph, it is a connected graph, *i.e.*, there are no unreachable nodes. The sequence, S, of a path, P_S, is the concatenation of sequences contained in the nodes, such that $P_S = u_{0S} \cdots u_{(q-1)S}$.

Given haplotyping information, the graph G is augmented with a set of haplotypes, H, where $H = \{H_0, \dots, H_{h-1}\}$, where $h = |H|$ is the number of observed haplotypes. Every edge $(u,\ v)$ is assigned a subset of H denoted as $(u,\ v)_H$, which describes the haplotypes that pass through the edge. Each encoded haplotype is represented by a path traversal through G, and may overlap other haplotypes.

Let G^E denote the null graph of G such that $G^E = \left(V', \emptyset\right)$, where V' originates from merging nodes in V (details of which are to follow in the subsequent section).

2.4.3 Constructing the null graph

The goal of indexing a population graph is to allow efficient substring querying on the paths that span the nodes and edges of the graph (Figure 2.6). For any non-trivial sized graph, enumerating all possible paths is often unfeasible, given the exponential nature of traversing all combinations of nodes and edges.

CHOP constrains queries through a graph to be part of a haplotype with which the population graph was built. To do this, CHOP transforms the graph G into a null graph

G^E such that each node in G^E represents a sequence of length k or longer, and that each substring of length k originating from the encoded haplotypes in G is also a substring in a node of G^E. If sequencing reads are true error-free samplings of an underlying haplotype and are the same length (or shorter) than the chosen value of k, they should correspond to a substring in a node of G^E. This, in turn, enables the application of any existing read aligner to place reads onto G^E. Through this transformation of G to G^E, all haplotyped paths of at least length k in the graph are considered. Three operators drive the transformation: collapse, extend, and duplicate (the pseudocode is given in Supplemental Algorithm 1), explained throughout the remainder of this section. Although the output of CHOP can depend on the order of these three operations, we did not observe any significant difference in runtime or indexing outcome for different orderings.

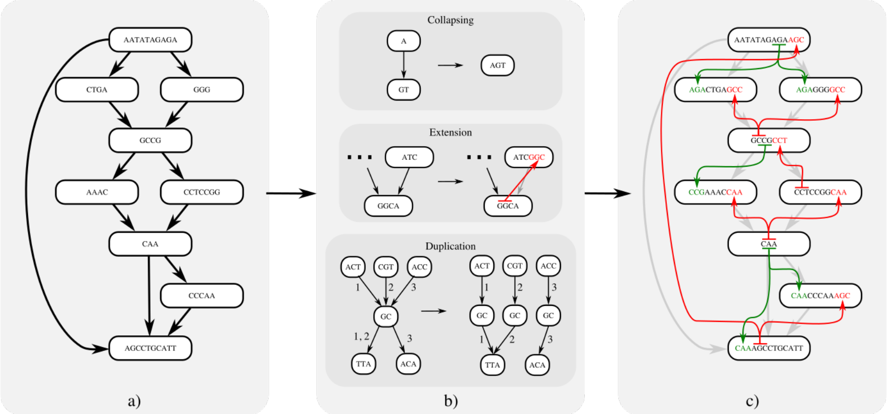


Figure. 2.6. Reporting the haplotyped k-paths in the population graph G transforms it into the null graph G^E, here $k = 4$. **a)** A population graph with sequence encodings on the nodes. **b)** Indexing of k-paths based on three operations; Collapsing, merging adjacent nodes. Extension, assigning k-length substrings as prefixes or suffixes between adjacent nodes. Duplication, copying of nodes, and redistribution of edges among copies. **c)** The null graph encodes all 4-length paths in the original graph, coloring of lines and text denote the origin of assigned prefixes (green) and suffixes (red) (note that colored lines are not edges in the graph).

Collapse

The first operation to transform G to G^E is collapse, which merges redundant traversals of nodes in the graph. If an edge $(u,\ v) \in E$ conforms to $out(u) = 1$ and $in(v) = 1$, then any path that traverses u, will be immediately followed by v. Therefore, it can be considered a redundant traversal such that the sequence on u and v can be merged without affecting the number of sequences the graph can spell out. To do this, the sequence and the corresponding intervals of u and v are merged, after which the outgoing edges of v are transferred to u, followed by the removal of v and the edge $(u,\ v)$. We denote this operation as collapsing, defined as $u||v$ for any edge $(u,\ v)$ (as shown in Figure 2.6b, the pseudocode is given in Supplemental Algorithm 2). The direction of collapsing is guided by minimizing the number of edge reassignments, such that when $in(u) > out(v)$, v is collapsed into u, joining the sequence $u_S = u_S \cdots v_S$. Alternatively u collapses into v, joining the sequence $v_S = u_S \cdots v_S$.

Extend

After collapsing redundant edges in the graph, a number of the remaining edges can be addressed with the extend operation. Extend is based on the observation that all k-length substrings that span a single edge $(u,\ v)$, *i.e.*, substrings that are defined by substrings of both the sequences of nodes u and v, can be accounted for by joining a $k-1$-length substring from one node and assigning it to the other. This extension of substrings may happen bidirectionally, namely the $k-1$-length right-hand flank of u is extended as a prefix of v, denoted as $u \twoheadrightarrow v$, provided that $in(v) = 1$ and $|u_S| \geq k-1$. Or vice versa, extending the $k-1$ length left-hand flank of v as a suffix of u, denoted as $u \twoheadleftarrow v$, provided that $out(u) = 1$ and $|v_S| \geq k-1$ (pseudocode is given in Supplemental Algorithm 3). To illustrate this operation consider the subgraph in Figure 2.7. Within this graph, both nodes u and v encode sufficient sequence to allow for extension between the two and report a k-length overlap, resolving the edge $(u,\ v)$. In Figure 2.8, a subgraph is shown in which extension is only possible for a subset of edges: $(u,\ w)$ and $(w,\ v)$. This does not apply for $(u,\ v)$, as $out(u) > 1$ and $in(v) > 1$. This shows a particular situation where only after resolving nearby edges, the subgraph can be sufficiently simplified to resolve all edges. Namely, $(u,\ w)$ and $(w,\ v)$ must first be resolved before $(u,\ v)$ can be solved by a collapse operation. Although the order in which substrings are extended may result in different null graphs, any of these will cover the same k-length substrings.

Since extension always concerns a $k-1$ length prefix or suffix, any substring of length k sampled from the underlying haplotypes will exclusively correspond to either the sequence in u or the prefixed sequence in v (or vice versa). In other words, by extending and subsequently removing edges in G, we introduce overlapping sequence as if we were converting G to the repeat-resolved string graph representation of a joint assembly of all genomes in G from all possible reads of length k [36].

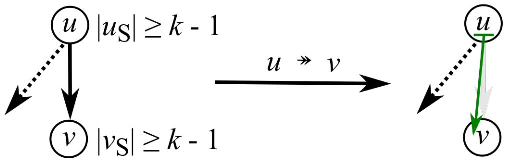


Figure. 2.7. A pair of nodes u and v where $|u_S| \geq k-1$ and $|v_S| \geq k-1$. Note that extension is only possible by prefixing v with the right-hand flanking substring of u, given that $out(u) > 1$. The extension operation denoted as $u \twoheadrightarrow v$ is defined as $v_S = u_S\,[|u_S| - k - 1, |u_S|] \cdots v_S$.

Duplicate

Sometimes neither collapse nor extend can be applied to any of the remaining edges in the graph without introducing path ambiguity, a situation in which there are several possible candidates to collapse or extend to/from, and choosing any candidate will block off paths to the remaining candidates. The graph topology must be simplified in these situations through the third operation, duplicate. The duplicate operation duplicates a node such that the set of incoming and outgoing edges are split between the duplicated nodes. (pseudocode is given in Supplemental Algorithm 4). Duplication allows consequent collapsing, enabling

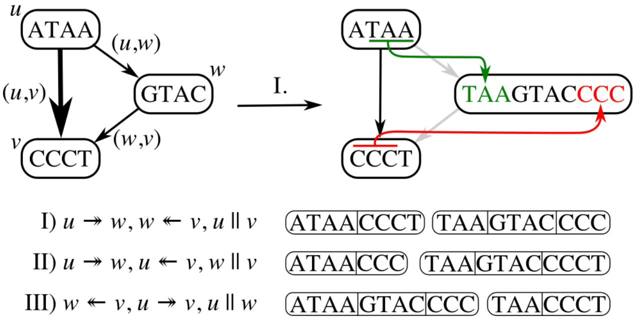


Figure. 2.8. Subgraph in which substring extension for $k = 4$ between (u, v) is not allowed unless either (u, w) or (w, v) are resolved first. Three distinct solutions can resolve this subgraph, and each solution is equivalent in k-path space.

substring extension, such that after a sufficient number of iterations, all edges in G can be resolved, either by extension or by collapsing.

Unlike methods that aim to track all possible paths through the graph, we propose using haplotype information modeled on the edges to constrain the number of node duplications needed from $in(u) * out(u)$ to δ. Where δ is the number of paired incoming and outgoing edges for u with at least one intersecting haplotype, note that δ is bounded by the number of haplotypes encoded in G and that there will never be more duplications than haplotypes in any region of the graph.

To illustrate this idea, Figure 2.9 shows a subgraph with haplotypes encoded on the edges. From the haplotyping, we can infer that not all paths through this graph are supported by the underlying haplotypes. For example, the path $u \longrightarrow d \longrightarrow f$ combines sequence segments that are unsupported (the haplotypes between (u, d) and (d, f) do not overlap). By excluding these unsupported paths through the graph, the number of duplications for node d can be constrained from 6 to 3. In this way, the search space for subsequent k-length substrings is greatly reduced compared to reporting all possible paths. Supplemental Figure S2.1 gives the full details about the transformation from Figure 2.1a to Figure 2.1b.

2.4.4 Aligning reads to CHOP's null graph

Established alignment tools can now directly align reads to the null graph representation, as long as reads are shorter or equal to $k+1$. Because the sequence modeled on the nodes in G^E is now a composition of sequences originating from adjacent nodes in G, the intervals that gave rise to these compositions need to be traced in order to convert the alignment of a read to a node in G^E to a path in G. For this reason, during the transformation from G to G^E, the originating node in G and corresponding offset for each prefixed, suffixed, or concatenated sequence is stored alongside the actual sequence. Note that, in theory, the defined operations can also be expressed purely in terms of interval operations, excluding any sequence. Given the intervals, a mapping between G^E and G is maintained, so any node in G^E can be traced back to the corresponding path of nodes in G. As a result, any alignment to a node in G^E can also be traced to a sub-path of this path, effectively enabling read alignment to graph G by using G^E as a proxy (Figure 2.1c).

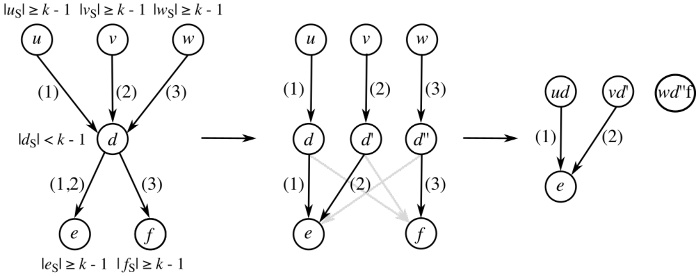


Figure. 2.9. Subgraph with haplotypes: $\{1,2,3\}$. Node d must be duplicated, as no more edges can be removed through extension or collapsing without introducing ambiguity. By grouping incoming and outgoing haplotypes on d, the number of duplications can be reduced. In the resulting graph, edges $(u,\ d)$, $\left(v,\ d'\right)$, and $\left(w,\ d''\right)$ can be collapsed. Finally, an extension can be applied to edges $(ud,\ e)$ and $\left(vd',\ e\right)$ which would lead to the null graph. Note that the introduction of grayed-out edges is prevented using haplotyping; hence the edge count is reduced from 6 to 3.

2.5 Supplementary materials

2.5.1 Transformation to a null graph

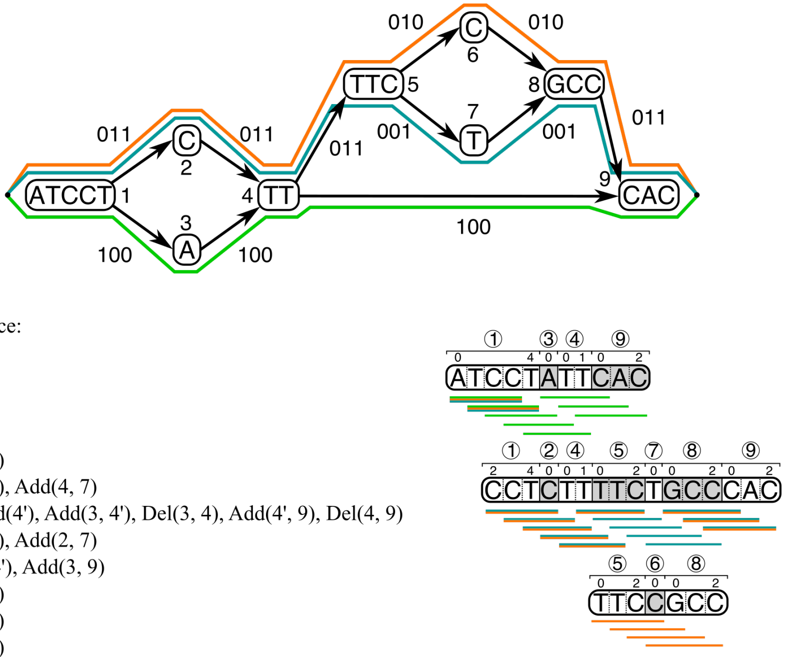
Through consecutive steps of extension, collapsing, and duplication (as described in the methods), CHOP can transform population graphs into null graphs. In this edgeless graph representation, each node now describes a haplotyped k-length path through the original graph. In Figure S2.1, we describe how the graph in Figure 2.1a is transformed into the null graph of Figure 2.1b.

2.5.2 *Mycobacterium tuberculosis* read sets

We obtained the 10 holdout samples from EBI-ENA, as shown in Table S2.1. All reads have a length of 101 bp and are single-end.

Table. S2.1. Samples used in MTB experiments, associated read sets are included, with KRITH1/2 accession numbers

Sample	Read set	KRITH1/2 ID	Read count
TKK-01-0053	SRR833154	G28639	5,263,942
TKK-04-0029	SRR1019154	G47382	5,481,779
TKK-02-0022	SRR1011463	G47310	4,301,550
TKK-01-0093	SRR958234	G38246	9,249,605
TKK-02-0066	SRR924236	G32253	5,168,899
TKK-01-0016	SRR832997	G27617	8,615,425
TKK-02-0051	SRR847783	G32041	7,571,230
TKK-01-0039	SRR833147	G27616	6,443,482
TKK-01-0047	SRR832984	G27644	5,532,779
TKK-01-0033	SRR833024	G27582	7,582,870



CHOP sequence:
1 ↠ 2
5 ↠ 6
6 ↞ 8
8 ↞ 9
1 || 3, Del(1)
5 || 7, Del(5), Add(4, 7)
Dup(4), Add(4'), Add(3, 4'), Del(3, 4), Add(4', 9), Del(4, 9)
4 || 2, Del(4), Add(2, 7)
4' || 3, Del(4'), Add(3, 9)
2 || 7, Del(2)
9 || 3, Del(9)
8 || 7, Del(8)

Figure. S2.1. The same graph as in Figure 2.1a is now shown with haplotypes on the edges encoded as bit vectors. Using CHOP, the input graph can be transformed into a null graph. Each of the steps performed by CHOP are shown in sequential order; Extension: $x \twoheadrightarrow y$ (y is prefixed by x), and $x \twoheadleftarrow y$ (x is suffixed by y). Collapsing: $x||y$ (x and y are collapsed into a single node). Duplication: Dup(x), (node x is duplicated).

2.5.3 Variation growth in MTB population graphs

When constructing graphs for the hold-out experiment, progressively more samples (from 1 to 400) are included in the constructed graphs. By including more samples, more variants are incorporated into the graphs, as can be seen in Figure S2.2.

2.5.4 Alignment criteria used for evaluation

To evaluate the behavior of the different aligners, we measure the following criteria: the number of mismatches, insertions, deletions, clipped bases, aligned reads/bases, unaligned reads/bases, perfectly aligned reads, and non-primary alignments. Base mismatches, insertions, deletions, and clipping may all be introduced to allow alignment of reads to the reference. To handle substitutions between reference and query, mismatches are introduced. Multiple base pair divergences are treated as insertions to the reference or as deletions from the reference. Base clipping masks portions of reads (from either end) that do not align end-to-end to the reference, meaning that shorter but contiguous read fragments are aligned. The extent of these operations in alignment can particularly characterize differences in alignments to linear references and population graphs, with the expectation that the incidence of these operations decreases in graph alignments (in proportion to the number bases aligned).

The number of aligned and unaligned reads indicates the proportion that aligns in a

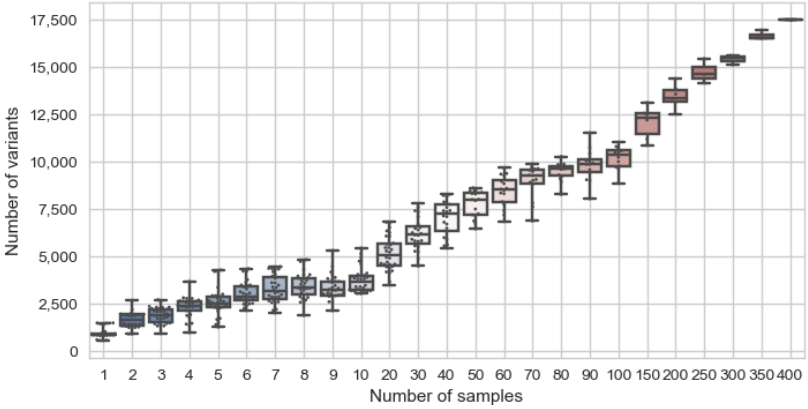


Figure. S2.2. Variable size variant sampling for VCF-based graph construction.

read set. For instance, reads may not be aligned due to insufficient sequence context on the reference or due to low-quality reads, random noise, or contamination. The number of aligned bases provides more detail, as not all reads are perfectly aligned. Perfectly aligned reads describe full-length alignments of reads for which no mismatches/insertions/deletions/clipping is introduced. The number of unaligned bases includes bases from unaligned reads, mismatches, insertions, and clipped bases. Reads for which multiple valid alignments score the same result in non-primary alignments, meaning that for each read, there will always be one primary alignment (or it is unaligned) and one or more non-primary alignments. The incidence of these non-primary alignments indicates the extent of alignment ambiguity, which is usually induced by the repetitiveness of the reference.

2.5.5 Graph construction from known variants

One way of constructing population graphs is to project sets of variants (from VCF files) called against a reference genome back onto this reference (similar to the construction in other methods such as vg and Graphtyper) (Figure S2.3a). Initially, a singleton graph is created, which encodes the reference sequence (Figure S2.3b). According to the order of their reference coordinates, the variants are iteratively inserted into the graph. For each variant, a minimum of three nodes are inserted into the graph. The reference node is first divided into two nodes, describing the sequence before and after the variation. Between these reference nodes, the reference and alternate alleles are introduced (Figure S2.3c). In the case of consecutive variants (variations no more than one base pair apart), the reference and variant alleles are connected to the preceding nodes and only then converge into a reference node (Figure S2.3d). The same procedure applies to indels and SNPs (Figure S2.3e). Haplotyping information is embedded on the edges, which includes the sample(s) and the reference.

The described graph construction strategy of CHOP differs from that of vg. The most notable change is the full combination of (consecutive) alleles and the treatment of SNPs and indels, as is shown in Figure S2.4 for CHOP and vg. Since the graph construction methods of CHOP and vg are similar but not entirely the same, this may also affect indexing (see

Figure S2.5), resulting in a different number of indexed paths in CHOP and vg, respectively.

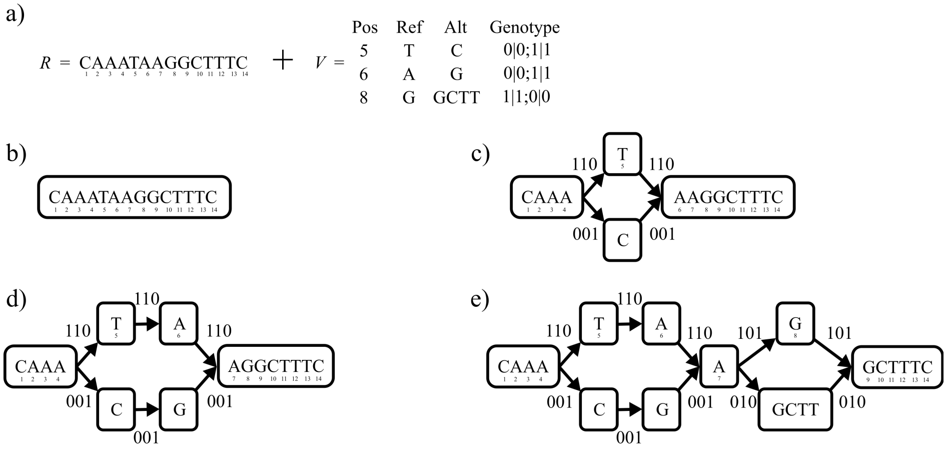


Figure. S2.3. **a)** The reference sequence and variant set used in graph construction. **b)** A graph is initialized with a single node encoding the reference sequence. **c)** In the order of the reference coordinate space, the variant T ⟶ C is introduced into the graph. **d)** A consecutive variant (A ⟶ G) is added to the graph. **e)** An insertion (G ⟶ GCTT) is added to the graph. The final population graph encodes three paths, one of which is the reference path.

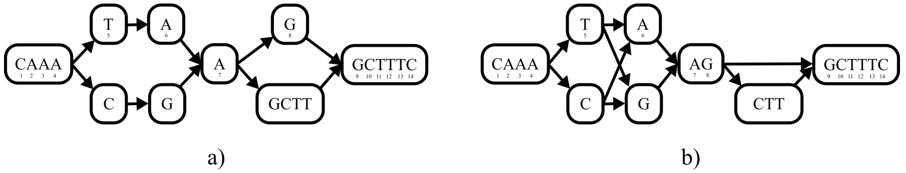


Figure. S2.4. Graph construction with the same input genome and variants as in Figure S2.3. **a)** Graph construction using CHOP. **b)** Graph construction with vg construct.

We evaluated whether read alignment is affected by the two differently constructed graphs. To do this, we aligned reads from sample SRR833154 with vg to graphs constructed by both vg and CHOP during the MTB hold-out experiment. We took the ratio from the number of perfectly aligned reads, unaligned reads, and mismatches for each of these alignments. The ratio is calculated by dividing (for example) the number of mismatches in the vg constructed graph alignment by those in the CHOP constructed graph alignment. Therefore, if there is no difference between the methods, the ratio should be equivalent to 1.0. The results in Figure S2.6 show that there is a minimal difference.

2.5.6 MTB read alignment to H37Rv and graph

In the hold-out experiment, 10 different single-end read sets are aligned to the reference genome, H37Rv, and population graphs that progressively include more samples (up to 400 excluding the hold-out). Alignments were evaluated on both a read and base-count basis. Shown for SRR833154 this includes the number of perfectly aligned reads (Figure 2.2), unaligned reads (Figure S2.7), and mismatched bases (Figure S2.8).

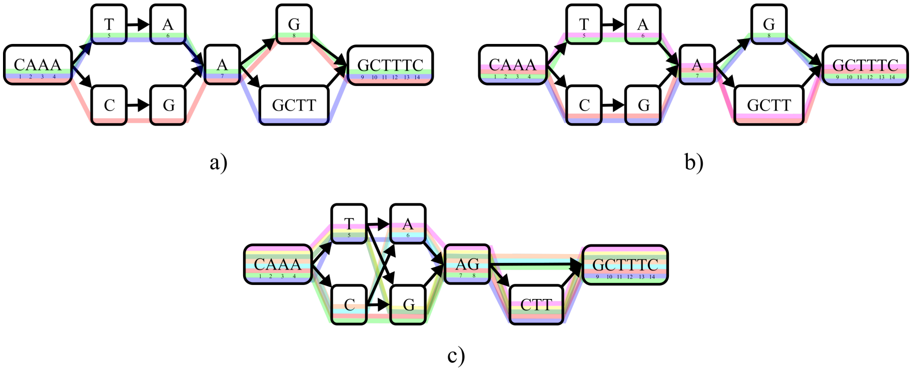


Figure. S2.5. The graph construction strategy of CHOP and vg (details in Figure S2.4) can affect the resultant paths that are indexed. **a)** CHOP extracts three paths from the CHOP constructed graph. **b)** If the same graph is indexed by vg, there will be four paths. **c)** The vg constructed graph indexed by vg has eight paths.

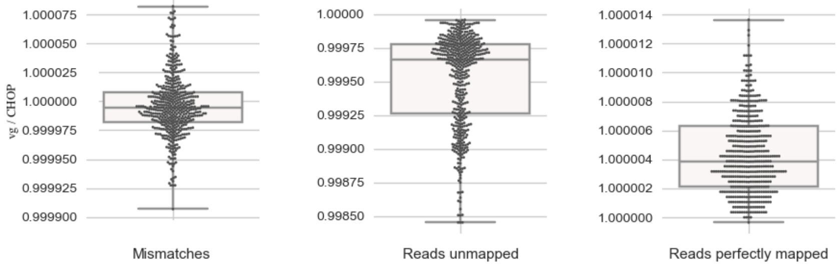


Figure. S2.6. SRR833154 graph alignments with vg using graphs constructed by either CHOP or vg. The y-axis represents a ratio (vg over CHOP) for the number of mismatches, unaligned reads, and perfectly aligned reads in the alignments.

2.5.7 Aligning to CHOP null graphs with vg

One strategy to directly compare CHOP to vg+GBWT and exclude any aligner-specific differences (BWA and vg) is to use the null graphs of CHOP with vg. When considering the null graph, each node within it can be understood as a path in the corresponding population graph; however, these paths can also be considered disjoint subgraphs. Therefore, interoperability of CHOP and vg should be possible if we consider the null graph as a collection of disjoint subgraphs. Since vg has subroutines for constructing graphs, we could directly construct vg graphs from the null graphs we generated.

We first evaluated this setup (CHOP/vg) using the same 10 MTB graphs (n=400, with one sample as a hold-out to align with). Here, CHOP was run for $k = 101$ on each of these graphs, with 16,909 paths being encoded on average in the null graphs. When converting the null graphs into the vg format, the graphs were ~9% larger on disk than when we constructed these graphs from a provided FASTA and VCF file (with approximately the same construction time). However, indexing the graphs with vg (GCSA2 and xg) took much longer (2,558 seconds on average), which is ~16x slower than indexing the graphs built with VCF files (164 seconds on average). For completeness, we also attempted to index the

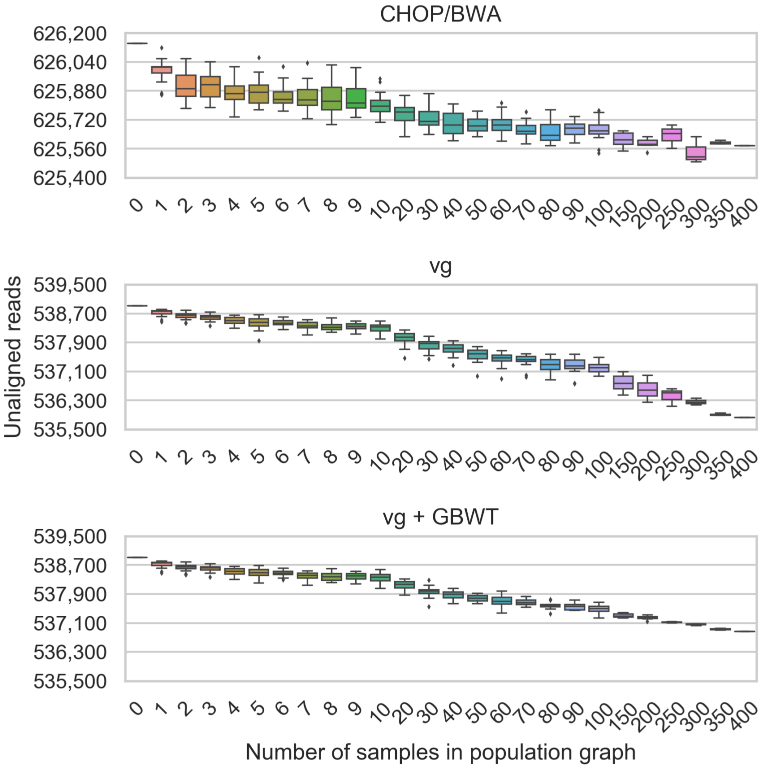


Figure. S2.7. Unaligned read count for SRR833154 alignments to different sized population graphs, containing between 0 (only H37Rv the linear reference) and 400 samples.

1000G chromosome 6 null graph with vg. However, we were unable to index the graph because the indexing time exceeded 7 days (for reference, this indexing took 5,751 and 33,619 seconds for vg and vg+GBWT, respectively).

In Table S2.2, we summarize the same alignments results as in Table 2.1, which now includes CHOP/vg alignment results. Overall, the results between CHOP/vg and vg+GBWT are similar. However, there is a 5x increase in non-primary alignments, which seems to result from increased redundancy in the index generated by CHOP. This may explain the slower index construction time and increased alignment times. These results further confirm that the alignment results of CHOP and vg are similar when the same aligner is used. Therefore, the differences between the two haplotype-aware aligners are mainly in the way haplotype constraints are implemented (fundamentally as in CHOP, or by the combined effort of GCSA2 and GBWT indexing in vg) as well as in the efficiency of the aligner (BWA or vg).

2.5.8 Comparing CHOP/BWA to HiSat2

We used the latest stable release of HiSat2 v2.1.0. Our attempts to index the 1000G graph of chromosome 6 with HiSat2 failed, with memory utilization exceeding 200 GB in 709

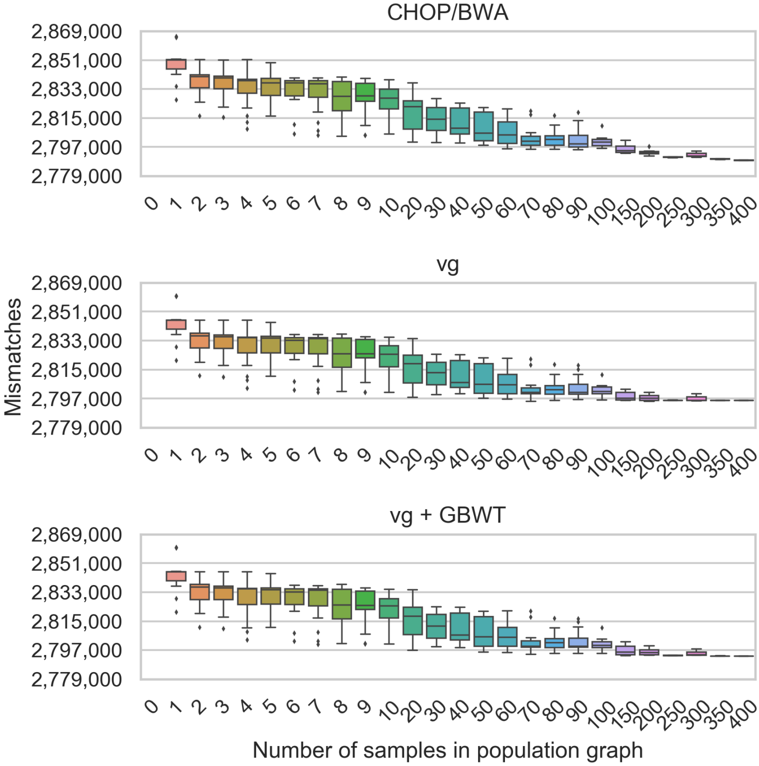


Figure. S2.8. Mismatch base count for SRR833154 alignments to different sized population graphs, containing between 0 (only H37Rv the linear reference) and 400 samples.

Table. S2.2. Mean of alignment results across all 10 hold-out sample alignments to 1) the reference genome H37Rv (H37Rv columns) and 2) the 400 MTB genomes graph (Graph columns) for CHOP/BWA, vg with and without haplotyping, and CHOP/vg to align the reads (note that when aligning only to H37Rv, CHOP is not used).

	All TB hold-out samples — Read length = 101									
Alignment criteria	BWA H37Rv	CHOP/BWA Graph (n=400)		vg H37Rv	vg Graph (n=400)		vg+GBWT Graph (n=400)		CHOP/vg Graph (n=400)	
Reads aligned	6,160,920	6,162,033	(+0.018%)	6,241,270	6,245,907	(+0.074%)	6,244,004	(+0.044%)	6,243,852	(+0.041%)
Reads unaligned	360,236	359,123	(-0.309%)	279,886	275,249	(-1.657%)	277,152	(-0.977%)	277,304	(-0.922%)
Reads perfectly aligned	4,048,774	4,142,052	(+2.304%)	4,048,774	4,153,217	(+2.580%)	4,153,124	(+2.577%)	4,152,800	(+2.569%)
Bases aligned	596,380,132	596,611,260	(+0.039%)	599,244,753	599,601,399	(+0.060%)	599,528,267	(+0.047%)	599,536,504	(+0.049%)
Bases unaligned	62,191,423	61,960,355	(-0.372%)	59,307,655	58,949,429	(-0.604%)	59,023,102	(-0.480%)	59,014,752	(-0.494%)
Bases unaligned from clipped reads	22,349,569	22,380,472	(+0.138%)	27,442,533	27,690,552	(+0.904%)	27,575,464	(+0.484%)	27,548,619	(+0.387%)
Bases mismatched	3,458,029	3,308,480	(-4.325%)	3,596,667	3,458,707	(-3.836%)	3,455,296	(-3.931%)	3,458,399	(-3.844%)
Bases inserted	65,210	65,151	(-0.090%)	84,358	85,938	(+1.874%)	85,397	(+1.232%)	85,510	(+1.366%)
Bases deleted	52,272	51,165	(-2.118%)	70,324	72,082	(+2.500%)	70,347	(+0.033%)	70,659	(+0.476%)
Non-primary alignments	246,092	246,540	(+0.182%)	539,309	724,904	(+34.414%)	724,613	(+34.360%)	**3,585,924**	**(+564.911%)**
Time (s)	533	721		10,711	4,457		4,540		**5,534**	

seconds. We believe this is due to an exponential growth in the number of *k*-paths in the graph. Unfortunately, there is no parameter (as in vg) to tune path lengths, so we could not evaluate this graph's alignment performance.

```
$ hisat2-build
  ↪ /.../hs37d5_chromosome_GRCh37_6_1_171115067_1.fa
  ↪ --large-index --snp /.../ALL.chr6...snp
↪ hs37d5_chromosome_GRCh37_6_1_171115067_1
```

Therefore, we focused our attention on the MTB graphs (n = 400 samples, plus the reference genome). Indexing these graphs took approximately 131 seconds and 12 GB of memory for each graph. HiSat2's memory footprint on these graphs is considerably higher than any of the tested methods.

In Table 2.1, we summarize the alignments results. Although indexing is more costly, HiSat2 aligns faster than BWA, CHOP/BWA, and vg. However, there are large differences in the alignment statistics. The HiSat2 aligner has many more unaligned reads in the baseline and graph alignments. This can be understood as an aligner-specific difference, *i.e.*, the aligner is less sensitive than BWA or vg. However, what is surprising is that the number of unaligned reads increases when considering the alignment of the graph with respect to the linear genome. If there is an increase in the number of non-primary alignments, this could be attributed to the multi-mapping reads. However, the number of non-primary alignments decreases (we did not observe this in CHOP or vg), suggesting a different cause. We speculate that gaps in the complete representation of all haplotypes may lead to missing sequences, which would explain both the increase in the number of unaligned reads and the decrease in non-primary alignments. From these experiments, we conclude that HiSat2 does not scale well and that the alignment results do not agree with those of CHOP/BWA and vg.

2.5.9 1000G variation linkage in chromosome 6

The 1000 Genomes Phase 3 variant set of chromosome 6 encodes 5,023,970 variants. To determine how much variation is shared between samples, we evaluated the genotyping of each variant, as is shown in Figure S2.9. This revealed that 41.58% of all variants are unique to its sample.

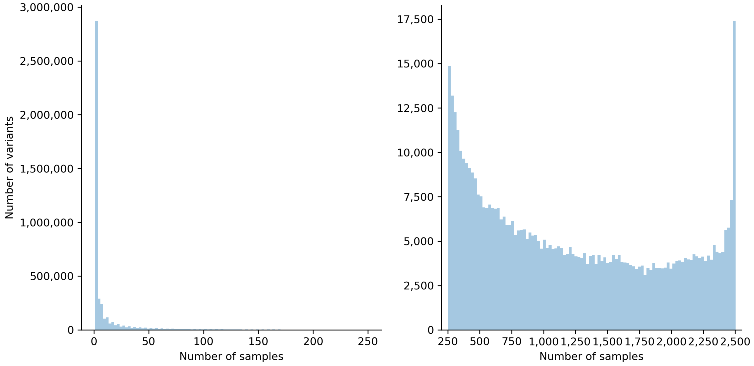


Figure. S2.9. Two histograms display the extent of shared variations among samples in the 1000 Genomes Phase 3 data of chromosome 6.

2.5.10 Filtered 1000G read sets

Since we only align to a chromosome 6 graph, the single-end read sets from 1000 Genomes Project Phase 3 (Table S2.3) were first filtered to exclude any reads aligning to other chromosomes. This was accomplished by aligning all 15 read sets to the human genome (excluding mitochondrial DNA) using BWA and then generating new read sets by extracting reads that were aligned to chromosome 6 or those that were unaligned. We chose random populations, and for each population chose two random samples (except for YRI, for which we chose one random sample).

Table. S2.3. The read sets from the 1000 Genomes Phase 3 used in alignments to chromosome 6.

			Filtered reads				
Population	Sample	Read set	Count	Mapped to chr6		Unmapped	
ESN	HG02938	ERR257960	6,238,375	5,572,661	(89.33%)	665,714	(10.67%)
ESN	HG03521	ERR257962	6,012,874	5,252,933	(87.36%)	759,941	(12.64%)
FIN	HG00308	ERR050084	3,882,577	2,814,118	(72.48%)	1,068,459	(27.52%)
FIN	HG00380	ERR050085	4,234,048	3,280,314	(77.47%)	953,734	(22.53%)
GBR	HG01791	ERR052834	3,066,482	2,358,096	(76.90%)	708,386	(23.10%)
GBR	HG01789	ERR052836	3,454,819	2,664,211	(77.12%)	790,608	(22.88%)
GIH	NA20881	ERR068420	2,278,409	2,015,089	(88.44%)	263,320	(11.56%)
GIH	NA20884	ERR068423	1,765,696	1,545,396	(87.52%)	220,300	(12.48%)
IBS	HG01670	ERR050090	1,115,839	859,298	(77.01%)	256,541	(22.99%)
IBS	HG02223	ERR056986	1,808,334	1,467,485	(81.15%)	340,849	(18.85%)
KHV	HG01595	ERR059932	1,217,726	1,059,586	(87.01%)	158,140	(12.99%)
KHV	HG02017	ERR059937	1,375,752	1,205,911	(87.65%)	169,841	(12.35%)
MSL	HG03054	ERR251326	4,720,003	4,293,293	(90.96%)	426,710	(9.04%)
MSL	HG03378	ERR251401	3,650,563	3,263,964	(89.41%)	386,599	(10.59%)
YRI	NA18517	ERR239432	569,541	478,520	(84.02%)	91,021	(15.98%)

2.5.11 Reads aligning to mitochondrial DNA

Figure S2.10 displays the distribution of the number of aligned bases for reads aligned exclusively on the graph. Of the reads corresponding to the peak at 69 bp, as shown in Figure S2.10, 97.54% of them aligned to the same fragment of a path in the graph. We used BLAST [79] to determine the origin of the path and found hits on chromosome 6 and mitochondrial DNA (corresponding to the path fragment). Realignment of the same mitochondrial DNA reads revealed that most of the reads aligned along the full length (CIGAR string is 100M), as partially shown in the pileup of Figure S2.11.

2.5.12 Alignment accuracy in chromosome 6

To measure the accuracy of CHOP/BWA-generated alignments, we compared the alignments of simulated reads to multiple linear and graph-based references. Reads were simulated using Mason 0.1.2 [81], which includes sequencing errors and base calling quality, as well as annotations indicating the ground truth location of each simulated read. Alignments were scored as correct if the aligned read was within 1 bp of the ground truth position. Only primary alignments were considered. We simulated 10,000,000 sequencing reads from a sequence on chromosome 6 that encoded variants of one sample with ID NA12878. Consequently, we generated a variation set encoding only SNPs and excluded variations and genotyping specific to NA12878 and family members.

Simulated reads were aligned to the linear chromosome 6 to provide a baseline mea-

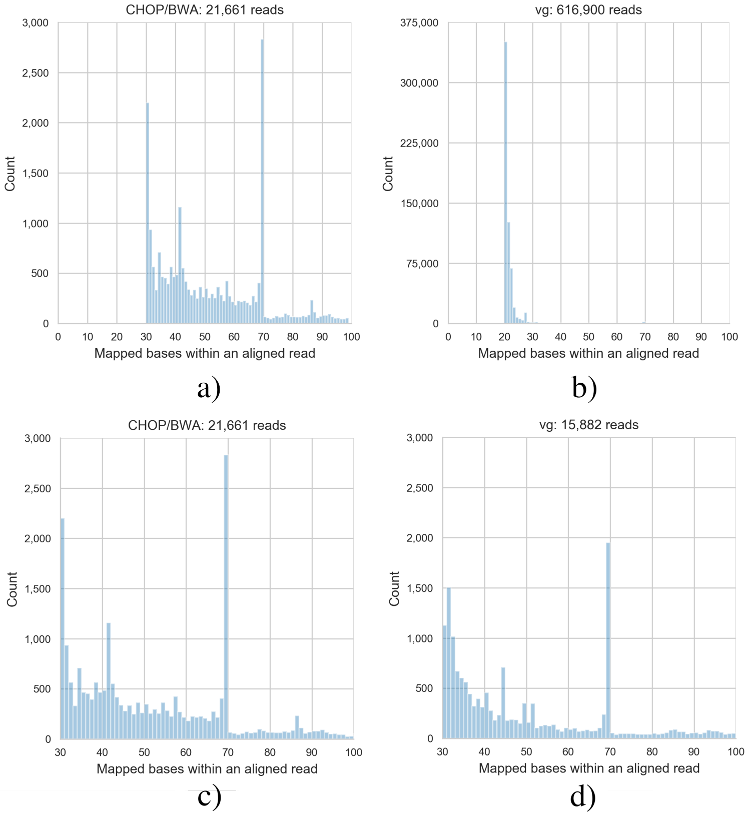


Figure. S2.10. The number of reads that have a particular number of bases aligned after their alignment onto the chromosome 6 population graph with CHOP/BWA (a) and vg (b), respectively. In c) and d), the same is shown in the range of 30 to 100 bases.

surement of the accuracy and to a personalized chromosome 6 (linear reference including all NA12878 SNPs) to obtain an idealized situation. Three graphs were constructed from the NA12878 filtered variation set: Full; graph encoding all 1000G variation in chromosome 6 (excluding NA12878), Min2; graph encoding only variations that were observed in at least two individuals; PopCov10+; graph encoding the top 10% scoring variations according to the FORGe [74] method, which weighs variants by the population allele frequency and minimizes the graph complexity. Figure S2.12a shows the fractions of reads that are correctly and incorrectly aligned to the different reference genomes. In Figure S2.12b, the sensitivity metrics of perfectly aligned reads and the number of mismatches are shown for the same alignments.

Although alignment sensitivity increases when more variants are introduced into population graphs, it also increases sequence repetitivity in the graph, which negatively influences alignment accuracy. This can be observed for both the Min2 and Full graphs, which are less accurate than the baseline, while they have comparable sensitivity with respect to the idealized reference genome. The trade-off between sensitivity and specificity is clearly

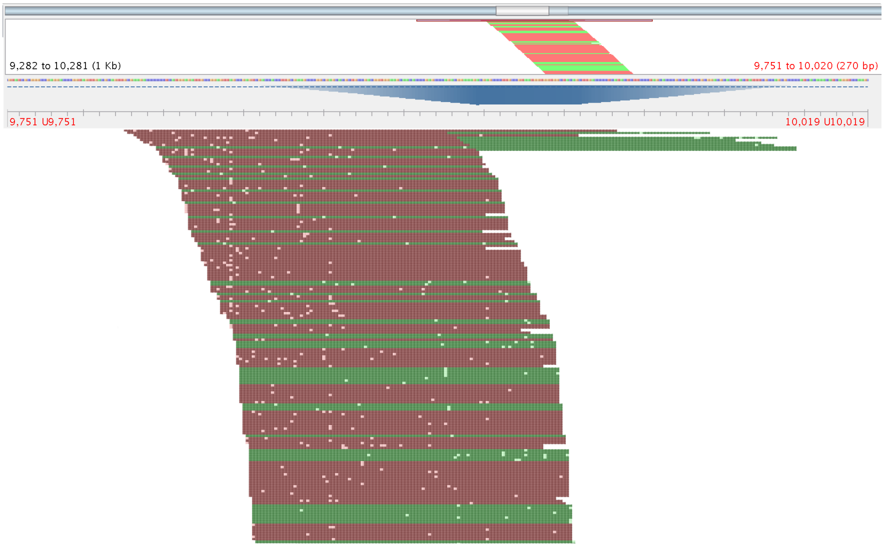


Figure. S2.11. Pileup visualization of mitochondrial DNA using Tablet [80] of reads that were previously unaligned on the linear reference genome (excluding mitochondrial DNA) but aligned on the graph corresponding to mitochondrial DNA.

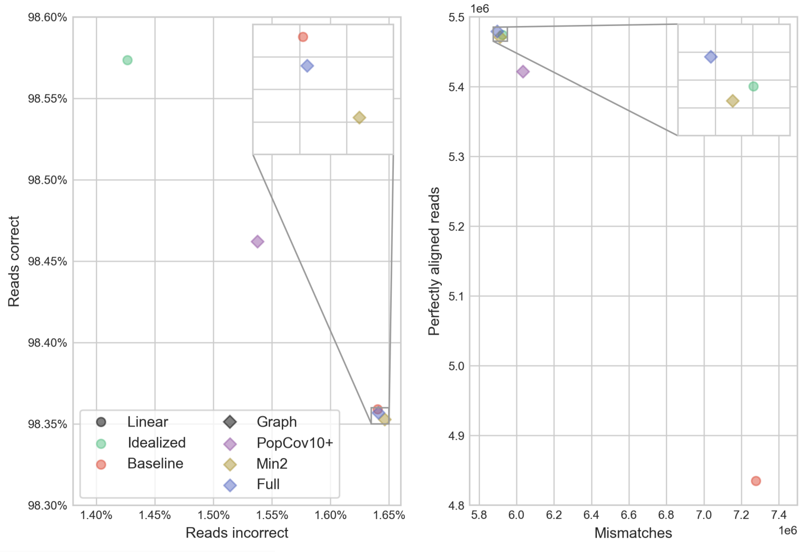
visible when variant selection is performed, as with the PopCov10+ graph, which improves accuracy at the expense of sensitivity.

2.5.13 Aligning to a MHC graph

So far, we have aligned to a variation graph (14,744,119 nodes and 19,770,411 edges) of chromosome 6 with all 1000G Phase 3 variants (5,023,970 variants). We made no exceptions to the variants included in this graph, regardless of their quality. Although we observed a clear improvement in alignment measures, we know that since MHC regions are more difficult to align, we can expect variants called in this region using the linear reference to be of lower quality than those found elsewhere.

To assess improvements in read alignment in the MHC region, we construct a graph from the GRC38 reference and alternate MHC alleles via multiple sequence alignment (MSA) using REVEAL [75]. This MHC graph is naturally much smaller than the previously constructed 1000G graph, with only 28,753 nodes and 40,032 edges. Since this graph was constructed using MSA, it also incorporates larger structural variations, which has not yet been explored using the variation graph.

Similarly, as described in Section 2.2.1, we ran CHOP on this graph for $k = 100$ and aligned the 15 read-sets onto the linear sequence of chromosome 6 (GRC38) and the null graph of the MHC graph using BWA. Since the graph only includes alternate sequences originating from the MHC region, any alignment improvements can be attributed to their inclusion. Our results are summarized in Table S2.4 along with our previous results when aligning to the GRC37 chromosome 6 linear sequence and the 1000G graph. There are clear advantages to switching from GRC37 to GRC38 in terms of improved alignability. We see



a) b)

Figure. S2.12. Alignment statistics of the NA12878 simulation. a) The fraction of correctly aligned reads and incorrectly aligned reads. b) Sensitivity metrics of perfectly aligned reads and the number of mismatches.

an improvement in aligned reads on this MHC graph compared to the linear reference genome (more aligned reads, fewer unaligned reads), showing the benefit of aligning reads on a population graph containing well-established haplotypes. The improvement on the MHC graph is less than for the 1000G variation graph, although we should note that the performance scores on the 1000G variation graph are averaged over the entirity of chromosome 6 and will be less beneficial for the MHC region.

Table. S2.4. Mean of alignment results from 15 samples from the 1000 Genomes data when aligning to 1) the reference genome sequence of chromosome 6 (column GRC37), 2) the 1000G graph created from the 5,008 haplotypes, 3) the reference genome sequence of chromosome 6 (column GRC38), 4) the MHC graph generated from a multiple sequence alignment with the reference sequence and the MHC alternate alleles.

1000 Genomes samples — Read length = 100						
Alignment criteria	BWA GRC37	CHOP/BWA Graph (n=2504)		BWA GRC38	CHOP/BWA MHC Graph)	
Reads aligned	2,542,399	2,543,522	(+0.044%)	2,551,410	2,551,774	(+0.014%)
Reads unaligned	483,670	482,548	(-0.232%)	474,659	474,296	(-0.077%)
Reads perfectly aligned	1,794,564	1,977,952	(+10.219%)	1,826,614	1,835,375	(+0.480%)
Bases aligned	251,122,992	251,516,725	(+0.157%)	252,080,501	252,142,201	(+0.024%)
Bases unaligned	51,439,949	51,070,534	(-0.718%)	50,483,271	50,423,549	(-0.118%)
Bases unaligned from clipped reads	1,801,947	1,846,687	(+2.483%)	1,812,049	1,814,726	(+0.148%)
Bases mismatched	1,270,981	969,087	(-23.753%)	1,205,282	1,179,256	(-2.159%)
Bases inserted	43,979	19,661	(-55.296%)	43,148	41,170	(-4.584%)
Bases deleted	61,659	32,355	(-47.526%)	60,870	58,445	(-3.983%)

2.5.14 Variant detection, CHOP and Graphtyper

Graphtyper's primary purpose is genotyping variants using a population graph and re-aligning reads in existing alignments. Unfortunately, Graphtyper does not output a graph realignment of the input linear alignment, which means we cannot evaluate its seeding and alignment capabilities. However, we can still compare its variant calling capability relative to CHOP/BWA. Since Graphtyper relies on a linear alignment (a BAM file from an aligner such as BWA), an implicit reference allele bias is introduced in the process. This bias should be reduced to some extent by realigning the reads to a given graph.

Note that Graphtyper has hard-coded limitations on the genomes that can be used with it. Because of this limitation, only human genomes (GRC37 and GRC38) can be used with the tool. We evaluated GraphTyper on human data with the variants from the 1000 Genomes Project Phase 3, using the latest stable release of Graphtyper v1.4. Our evaluations began with the 1000G chromosome 6 graph and the linear reference sequence of chromosome 6. Graph construction took 4,753 seconds (in the case of the graph) and 8 seconds (for the linear reference). Indexing required 16 GB memory, 8,248 seconds (graph), and 10 GB memory, 2,285 seconds (linear reference).

With this setup, we aligned, with BWA, the reads from ERR050084 (sample HG00308) — pre-filtered to only include reads that align to chromosome 6 or that were unaligned elsewhere on the genome (3,882,577 reads) — onto the linear sequence of chromosome 6. While variant calling a population graph already decreases the number of newly called variants, this is even lower in this particular graph since the variants (as called on a linear reference) of HG00308 are encoded in the graph. We called variants and genotypes using *graphtyper call*. This took 126,456 seconds (1.4 days) and 33 GB of peak memory with the linear graph. Next, we called variants on the 1000G graph (same command), ended after 382,329 seconds (4.4 days), with a peak memory of 93 GB.

```
$ graphtyper call /.../graphtyper.gt --sam=/.../in.bam
  ↪ --index=/.../.gt_gti 6
```

We followed the same steps for CHOP/BWA (note that in the linear case, this reduces to running BWA only), and variant calling was done using bcftools. The variant calling for the linear genome took 700 seconds and 200 MB of peak memory. The 1000G graph completed after 1,620 seconds, with peak memory of 9.5 GB.

```
$ bcftools mpileup --redo-BAQ --min-BQ 30
  ↪ --per-sample-mF --annotate DP,AD -f /.../ref.fa -O
  ↪ b /.../in.bam | bcftools call --multiallelic-caller
  ↪ --variants-only -Ob > /.../out.bcf
```

We focused on variant calling, as this would allow us to compare CHOP/BWA to GraphTyper, especially since GraphTyper can also call new variants at realigned positions. Variant calling the linear graph with GraphTyper resulted in no new variants (which was unexpected) or genotyped calls (as expected). It is not clearly defined how GraphTyper handles linear graphs. As such, we do not know what to expect in this scenario when calling new variants. The 1000G graph alignment yielded 0 new variant calls and 4,683,374 genotyped variant sites. The number of genotyped sites is exceptionally high relative to the number of variants (~5M) encoded in the graph. Clearly, GraphTyper is highly sensitive and reports many false positive variants; hence, filtering would be necessary to reduce this. The recommended procedure (as seen in the GraphTyper repository) was to utilize the vcffilter tool (part of vcflib: `https://github.com/vcflib/vcflib`). This reduced the number of genotyped variant sites to 144,800.

```
$ vcffilter -f "ABHet < 0.0 | ABHet > 0.30" -f "MQ >
 ↪ 30" -f "QD > 6.0" /.../.vcf
```

Variant calling of the CHOP/BWA alignment yielded 142,979 variant sites for the linear genome and 1,212 in the graph. As the utilized variant calling is highly sensitive, quality filtering reduced this number to 57 variants from previously unaligned reads on the linear genome. These results are more in line with expectations: 1) the number of variants detected on the graph is considerably lower than on a single linear reference genome, and 2) the variants detected on the graph are on top of the specific haplotypes, containing already variants with respect to the reference genome, so that reads from these regions have a smaller chance to align to the reference genome (due to large variation). Overall, we conclude from this experiment that CHOP/BWA is more time and memory efficient (the central claims we make); moreover, GraphTyper seems to generate unexpected results when calling variant.

2.5.15 Comparing CHOP/BWA to GraphAligner

GraphAligner is a long read aligner for graphs and supports similar input and output procedures as vg, meaning we could directly input our graph built by vg into GraphAligner. We aligned reads with GraphAligner (commit: `8e37ecbc832cca5538e8d142780 3e313089b17fb`) from the 15 samples on the 1000G chromosome 6 graph and linear chromosome 6 sequence.

```
        $ GraphAligner -g /.../.vg -f /.../.fq.gz -a
 ↪ /.../.json -t 1 -b 35 --try-all-seeds
 ↪ --seeds-mxm-cache-prefix ...

```

In Table 2.2 we summarize the results of the alignments. GraphAligner measurements are closer to CHOP/BWA and vg than to HiSat2. In addition, the alignment times are similar to those of CHOP/BWA. However, we note the same behavior as with HiSat2, where we see a decrease in aligned reads (and an increase in unaligned reads) when aligning to the graph instead of the linear reference genome, which may be attributed to the optimizations for seeding in long reads rather than the short reads used here.

2.5.16 Effects of varying *k* size in CHOP

First, we want to address that exponential path evaluations are eliminated in CHOP because it uses haplotype information. At any position in a graph, there can only be as many parallel paths as encoded haplotypes, and this worst-case scenario can only occur if none of the genomes in the graph share any linkage at that position. This was explored by changing the value of k in CHOP, *i.e.*, the minimal k-mer for which an exact match is required. The larger k is, the more variation is covered, which means that more paths need to be explored in the graph, which should match, at most, the number of haplotypes encoded in the graph.

We ran CHOP on a MTB graph ($n = 400$ samples, plus the reference genome), with values of k ranging from $[1, 10,000]$, and evaluated different metrics (Figure S2.13). We observe an approximately linear growth in processing time, peak memory, and the total number of bases in the resultant null graph as the value of k increases. The number of nodes in the null graph decreases as k increases and converges to the number of encoded genomes within the graph. From this, we see that for this variation graph choosing a k larger than the largest recorded genome (k > 4.4 MB) results in a null graph with 401 nodes (400 VCF genomes and a single reference genome). For completeness generating such a null graph takes 559 seconds with a peak memory of 6.5 GB and encodes ~1.76 gigabases of sequence, equivalent to the concatenation of all genomes in the graph.

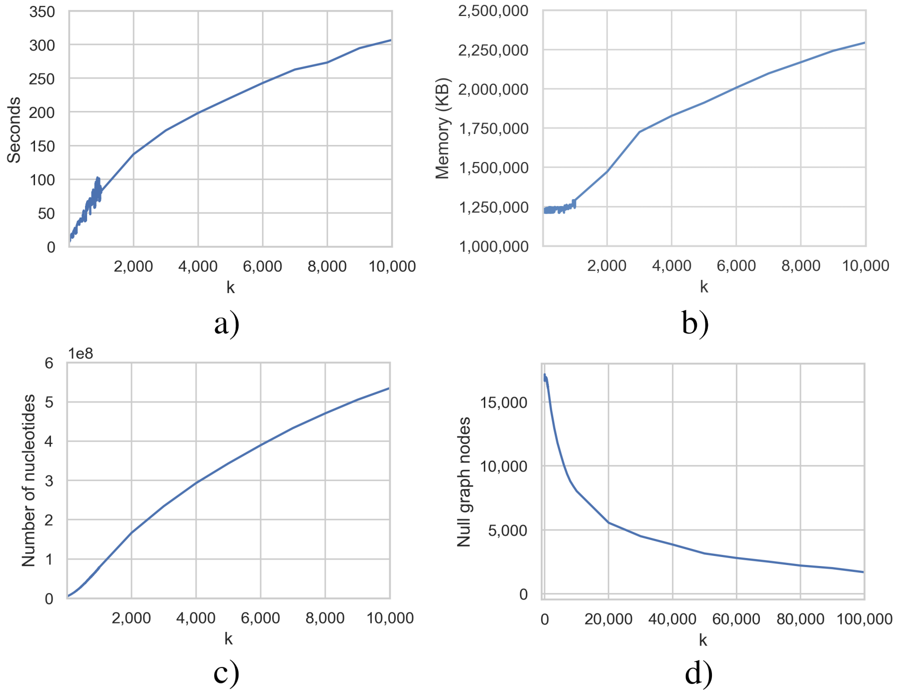


Figure. S2.13. Measurements of CHOP being run on an MTB graph (n=401) for increasingly larger values of k: a) indexing time; b) peak memory in KB; c) nucleotides encoded in the null graph; d) the number of nodes in the null graph.

In Section 2.5.18 we show a much higher variation density in the 1000G graph of chromosome 6. Because of this higher density, larger values for k will inevitably lead to an explosion of paths. We observed such growth when indexing with vg(+GBWT) because of GCSA2 (note that we had to reduce k from 104 to 52 to run vg(+GBWT)). This is prevented by CHOP. However, larger values for k will likely still lead to an intractable number of haplotype-constrained k-paths. To confirm this, we explored different settings of k for this graph, the results are presented in Figure S2.14. Indexing time grows — as expected — linearly. Peak memory, however, shows a decrease and then an increase, after which it evolves linearly. This erratic behavior could be caused by differences in local densities of variation within the graph.

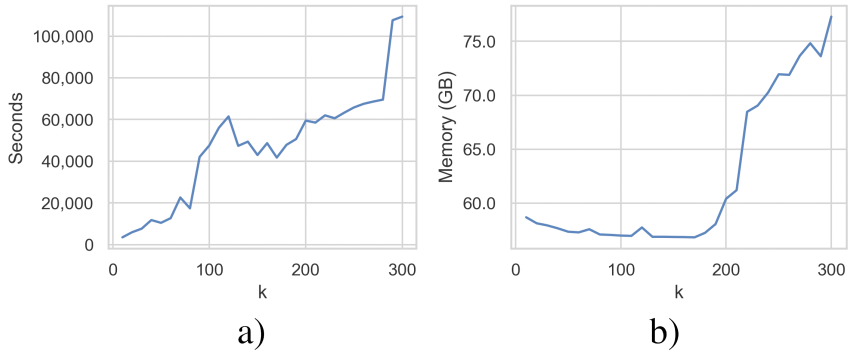


Figure. S2.14. CHOP is being measured on a population graph of human chromosome 6, for increasingly larger values of *k*: **a)** indexing time; **b)** peak memory in GB.

2.5.17 Simulated graph indexing

To better understand the behavior of CHOP under varying variation density, the number of samples, and between sample variation linkage, we set up the following experiment using the H37Rv reference genome (4.4 MB) as a starting point. Given this reference, we simulated sample VCFs with SNP variants uniformly distributed across the genome. In our simulation, we used different variation groupings, simulating samples with exactly: 100, 500, 1000, 5.000, or 10.000 variants. We also include variation linkage within the subsequent simulated sample for each of these groupings, given all previously observed samples. For example, starting from a single simulated sample *a*, the subsequent simulated sample *b* will share some variation of the previous sample *a*, and the next simulated sample *c* will share it with both *a* and *b*. The amount of variation linkage represents another grouping; these include 5%, 10%, 20%, 30%, 40%, and 50% shared variation between samples. For each grouping combination, we then generated merged VCF files with a varying number of samples; this included 1 to 10 (step = 1), 10 to 100 (step = 10), and 100 to 500 (step = 100), *i.e.*, 23 VCF files with different numbers of samples. Because these VCFs were generated for each combination of grouping, we created 690 VCF files in total; Figure S2.15 shows the variation distribution for each of these settings. We built graphs from each VCF file using the described VCF graph construction method. Next, we ran both CHOP and vg+GBWT with *k* = 104 on each of the graphs and measured runtime and peak memory as shown in Figures S2.16 and S2.17. Note that we allowed a maximum runtime of 4 hours and peak memory of 80 GB for each indexing method.

The complexity of the graphs varied considerably, and the majority could be indexed by both indexing methods within the set constraints (640 CHOP, 547 vg+GWBT). The remaining graphs were very complex and encoded at least 5,000 (500 for vg+GBWT) variants per sample, which considering the reference genome, results in regions very densely populated with variants. CHOP indexing is significantly faster, memory efficient, and can handle more complex graphs than vg+GBWT. Increasing the between-sample variation linkage simplifies the graph, with fewer unique variants being encoded, simplifying indexing for all approaches. Typically, variation can be expected to be shared at higher levels than shown here, especially when filtering out low allele frequency variants.

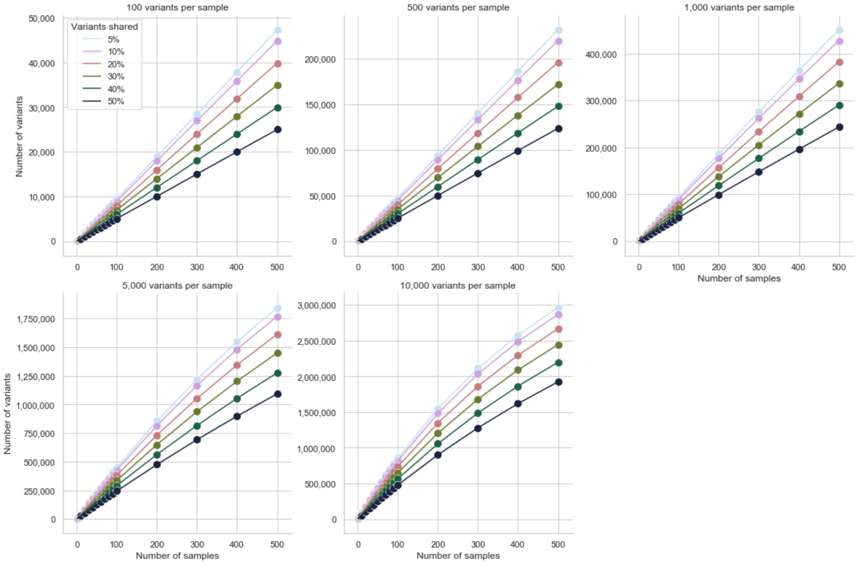


Figure. S2.15. The number of variants encoded in merged VCF files for all combinations of groupings: number of variants per sample (different plots) and probability of sharing variants with the population (different colors).

2.5.18 Variant density in human chromosomes

The computational costs required to index human chromosomes are highly dependent on the number of variants encoded in the graph, the density of those variants, and the size of the chromosome. For some human chromosomes, this this density of variants (possibly in combination with the size of the chromosome) can lead to explosive growth in memory or disk space required, which occurred with vg+GBWT for chromosomes 1, 2, 11, and X. To illustrate this, we quantified the number of variants across each chromosome in windows of 50 bp (note that we set vg to index $k = 52$ length paths) as is shown in Figure S2.18. Chromosomes 1, 2, 11, and X encode variants at higher densities than others, and chromosome 1 even exceeds 50 variants in a 50 bp window. Note that these measurements should also be put in the context of chromosome size.

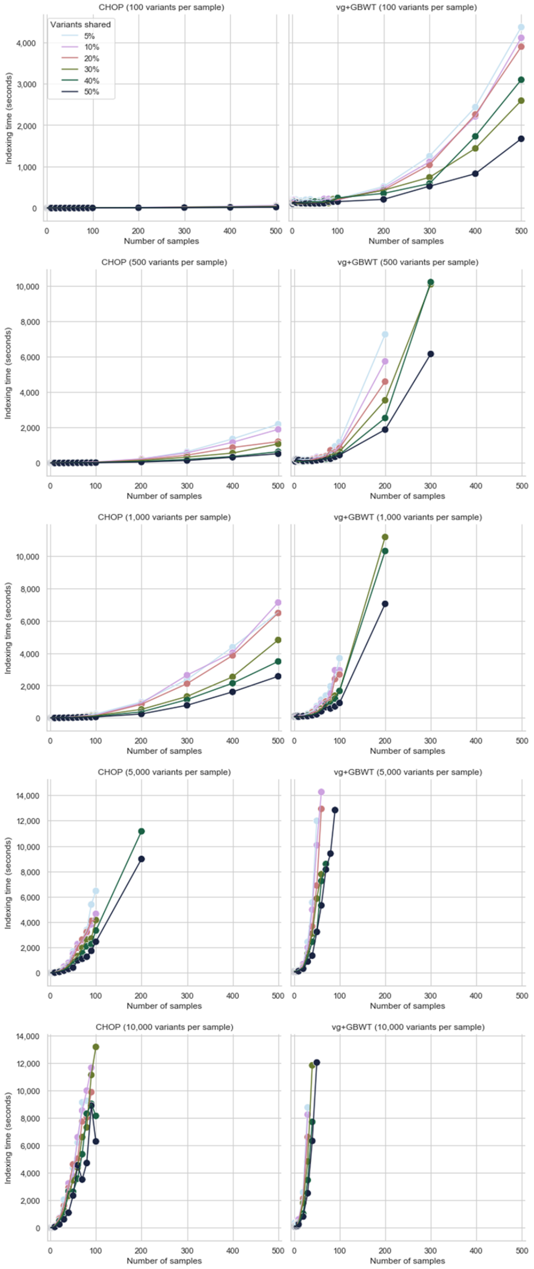


Figure. S2.16. CHOP and vg+GBWT indexing time (seconds) of the graphs for all combinations of groupings: number of variants per sample and probability of sharing variants with the population (different colors). Missing points in the plots indicate that indexing failed by exceeding one of the constraints.

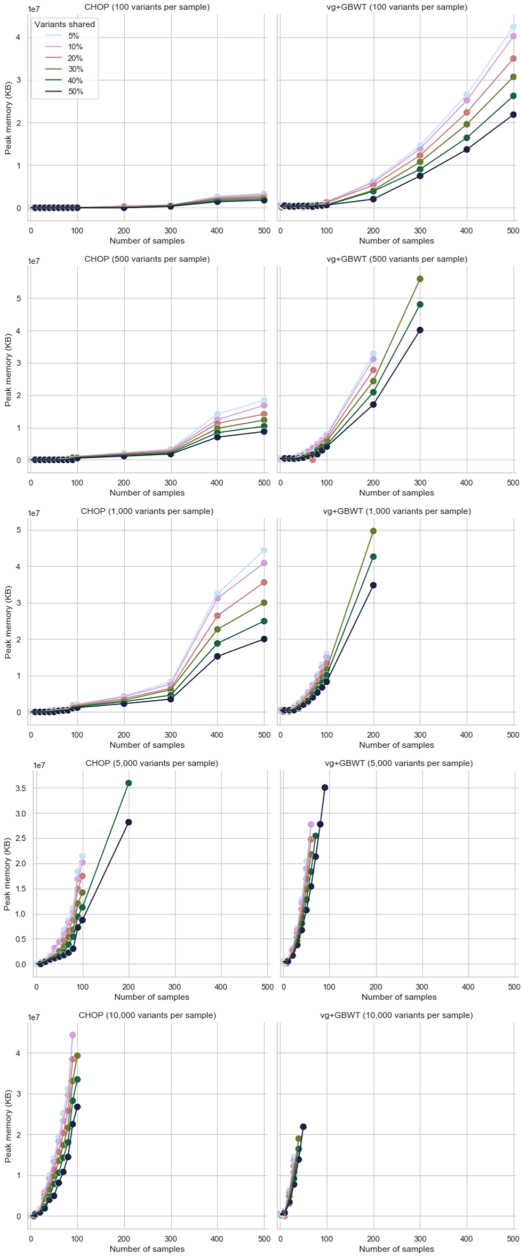


Figure. S2.17. CHOP and vg+GBWT peak memory (KB) during graph indexing for all combinations of groupings: number of variants per sample and probability of sharing variants with the population (different colors). Missing points in the plots indicate that indexing failed by exceeding one of the constraints.

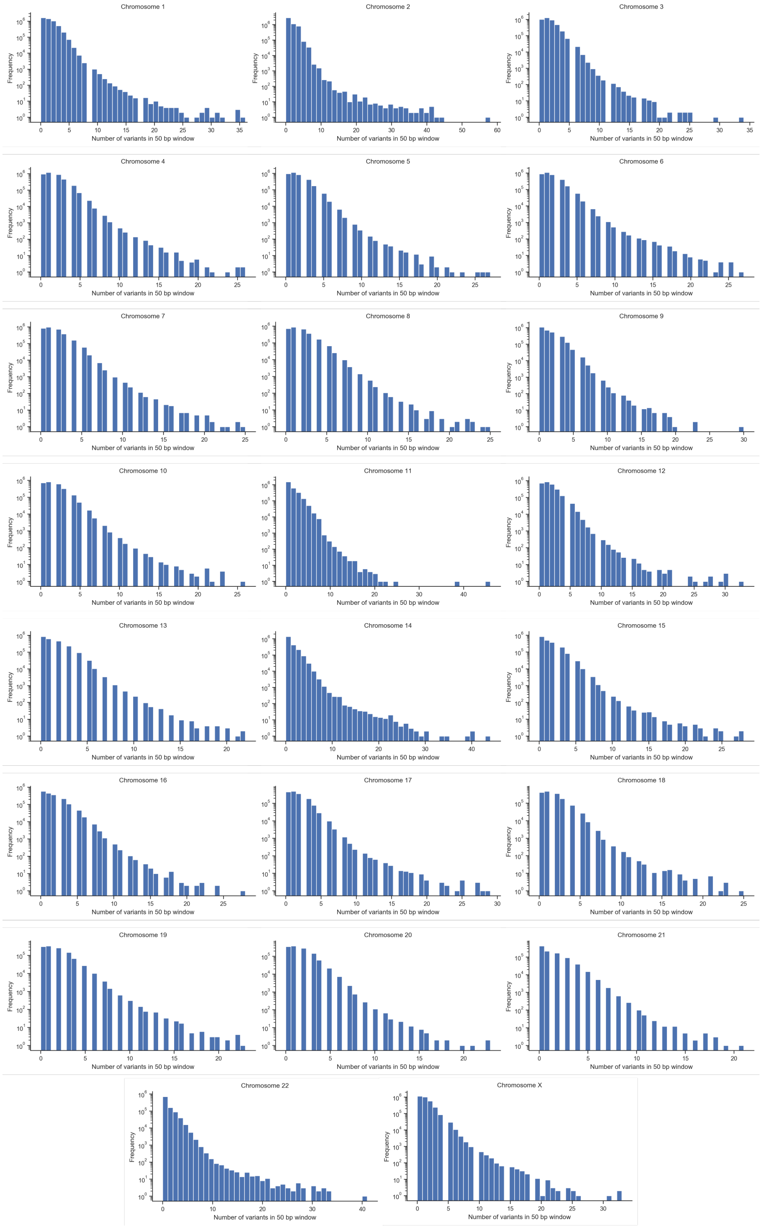


Figure. S2.18. Variant distributions for each human chromosome in 50 bp windows.

2.5.19 Variant integration

Realignment of SRR833154 reads on a graph representation of H37Rv with variants detected from the alignment of SRR833154 reads onto H37Rv using CHOP/BWA, allowing for the calling of novel variations using existing linear genome variant callers. Using Pilon, we called variants in the graph alignment, of which 19 remained after quality filtering. Since BWA, by default, outputs a SAM file, we can preprocess and prepare a BAM file that can be inspected using an alignment visualization tool such as Tablet [80]. In Figure S2.19 we show newly aligned reads from which variants can be called.

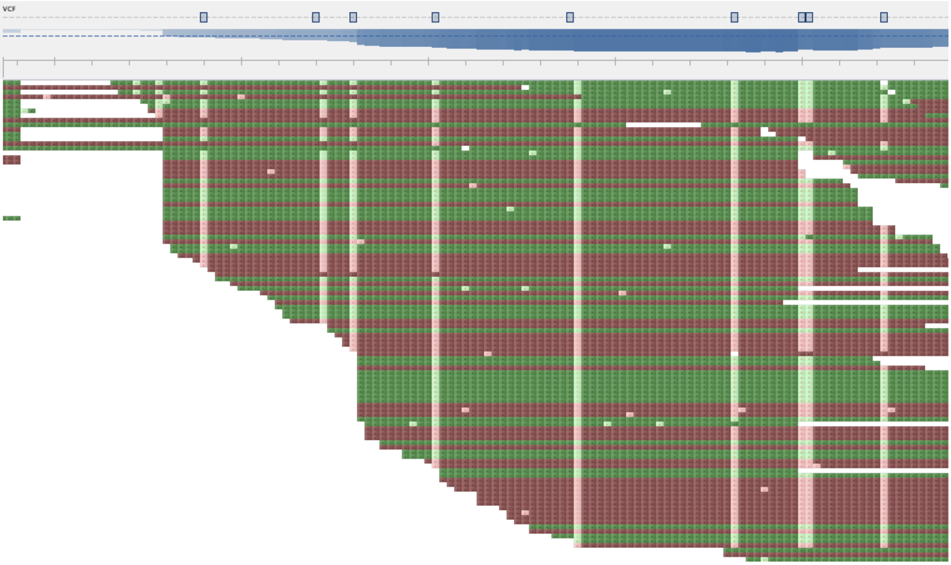


Figure. S2.19. Pileup visualization of SRR833154 reads aligned to a H37Rv graph using Tablet, with variant call annotations on top denoted as blue blocks.

2.5.20 CHOP, String graphs, and de Bruijn graphs

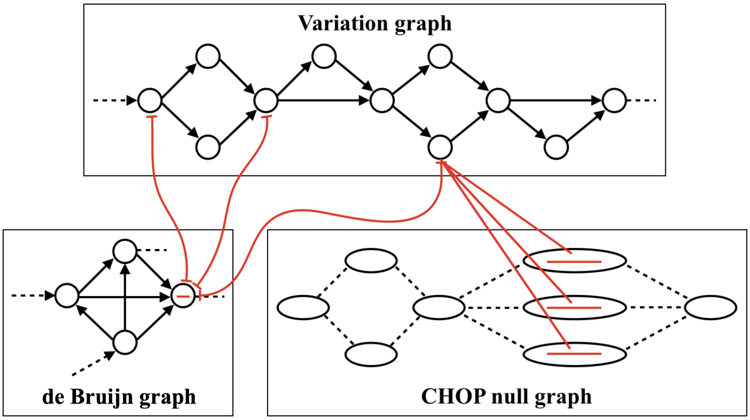


Figure. S2.20. Multiple positions in the variation graph can map to the same position in a de Bruijn graph index, while CHOP ensures a unique mapping for each.

2.5.21 Pseudocode CHOP procedures

Listing 1

```
procedure ChopGraph($G$)
    SimplifyGraph($G$)                          ▷ Extend and Collapse until exhaustion
    if $G_E \neq \emptyset$ then
        for $(u, v) \in G_E$ do
            if $deg(u) > deg(v)$ then
                Duplicate($G, u$)                ▷ Duplicate $u$
            else
                Duplicate($G, v$)                ▷ Duplicate $v$
            end if
        end for
        ChopGraph($G$)
    end if
end procedure
```

```
procedure SimplifyGraph($G$)
    modified ⟵ $True$
    while modified do
        modified ⟵ $False$
        for each edge $(u, v) \in G_E$ do
            if $out(u) = 1$ and $in(v) = 1$ then
                Collapse($G, u, v$)              ▷ Collapse $u||v$
                modified ⟵ $True$
            else if $in(v) = 1$ and $|u_S| \geq k - 1$ then
                Extend($G, u, v, 1$)             ▷ Prefix $u \twoheadrightarrow v$
                modified ⟵ $True$
            else if $out(u) = 1$ and $|v_S| \geq k - 1$ then
                Extend($G, u, v, 0$)             ▷ Suffix $v \twoheadleftarrow u$
                modified ⟵ $True$
            end if
        end for
    end while
end procedure
```

Listing 2

```
procedure Collapse($G$, $u$, $v$)
    if $in(u) > out(v)$ then                          ▷ $u \longleftarrow v$
        $u_S = u_S \cdots v_S$                        ▷ Concatenate sequence
        for $(v, x) \in G_E$ do                       ▷ Outgoing edges $u$
            add edge $(u, x)$
        end for
        delete node $v$
    else                                              ▷ $u \longrightarrow v$
        $v_S = u_S \cdots v_S$                        ▷ Concatenate sequence
        for $(x, u) \in G_E$ do                       ▷ Incoming edges $v$
            add edge $(x, v)$
        end for
        delete node $u$
    end if
end procedure
```

Listing 3

```
procedure Extend($G, u, v, isPrefix$)
    if $isPrefix$ then                                ▷ $u \twoheadrightarrow v$
        $v_S = u_S[|u_S| - k - 1, |u_S|] \cdots v_S$
    else                                              ▷ $v \twoheadleftarrow u$
        $u_S = u_S \cdots v_S[0, k - 1]$
    end if
    delete edge $(u,\ v)$
end procedure
```

Listing 4

```
procedure Duplicate($G, u$)
    for $(p, u) \in predecessors(G, u)$ do
        for $(u, s) \in successors(G, u)$ do
            group $\longleftarrow (x, u)_H \cap (u, x)_H$
            if group $\neq \emptyset$ then
                create node $i$                       ▷ $i \longleftarrow u$
                create edges $([(p, i), (i, s)])$
            end if
        end for
    end for
    delete node $u$
end procedure
```

3

Conclusion

The field of genomics plays a vital role in improving our understanding of life and its processes. Within genomics, sequence comparison plays a fundamental role, with alignment algorithms placing DNA sequencing data in the context of a chosen reference genome. Downstream analysis subsequently relies on finding (dis)similarity within this context. Such genome alignments are generally onto references wherein only a single copy is present of each allele. Current advances in computation have made it possible to abandon some of the trade-offs necessary in the early days of genomics. In addition, a host of population variations can be incorporated into the reference model to combat problems such as reference bias. The population graph is a natural representation to represent multiple different copies of each locus. Here, population graphs are presented as labeled directed acyclic graphs, which can encode any number of genomes (and their inherent variation) in a manner close to non-redundancy. I have shown how path indexing with CHOP decomposes labeled population graphs into haplotype-constrained sets of segments that allow existing linear genome aligners to index these segments, facilitating efficient proxy sequence alignment on the original graphs. Alignments onto haplotype-indexed population graphs reduce the issue of reference-based allele bias, improving characterization and access to complex regions of the genome. Graphical encoding of population variation can greatly simplify downstream analyses, such as genotyping and variant calling. For example, genotyping can be based on the identification of variants already embedded in the graph, and variant calling may not even be necessary unless the goal is to identify new variants.

The population graph should replace the linear reference genome. By placing the population in the reference context, solving any representation problems inherent in the linear representation becomes significantly easier. Furthermore, this adoption should be done in conjunction with an implicitly implemented haplotype-constrained path indexing, which not only limits the number of paths at any given position by the number of haplotypes encoded in the graph, but also exploits the compressibility of the variation linkage inherent in populations. Not only does this prevent the exponential growth of path space that might otherwise exhaust computational resources in complex graphs, but more importantly, haplotype path indexing prevents spurious alignments that would otherwise be introduced by arbitrary indexing.

There is still no clear rule about how to construct the perfect population graph or what variation to include. I have shown that including all observed variation in the population is counterproductive, as the graphs will become too complex and introduce alignment ambiguities even if haplotype constraints are applied. While several loose guidelines may be enforced, such as a minimum threshold of allelic frequency within the population, emphasizing larger variants that are more likely to be affected by reference bias, or considering variation proximity. Finding an objective measure to select variants for graph integration is essential. This would require measuring reference allele bias and understanding how variants may affect graph topology and alignment ambiguity. Then a trade-off can be made between balancing population variation and graph complexity. There is also contention in how a graph is built. In my work, graphs were both constructed based on detected variation against a reference genome and constructed from multiple sequence alignments. While the former approach is substantially more straightforward, it relies on the linearized representation to obtain the variants, which implicitly biases the graph reference. The least biased source for a population graph lies in multiple whole genome sequence alignment; this would likely also need to be non-progressive to avoid guide-tree bias. However, whole-genome alignment in extensive collections or large genomes remains a challenging prospect.

CHOP is compatible with existing linear aligners, which avoids the immediate need for the development of a stand-alone aligner using CHOP as its indexing basis. However, offloading the alignment task to a third party leads to its own challenges and drawbacks. Incorporating the nature of a graph structure using such aligners is nontrivial, and will require the indexing of excessively long paths through graphs. Aligners that utilize the Burrows-Wheeler transform or extensions thereof are typically more popular because they create a compact index. This was especially important in the early days of genomics when memory was at a premium. However, such indexes do result in slower querying speed. Nowadays, memory availability is becoming less of an issue, and in recent years there has been a resurgence of linear genome aligners utilizing hashing-based indexing, optimizing speed at the cost of space, such as SNAP, FSVA, Minimap2, and URMAP [82–85]. The hope was to build a speed-optimized graph aligner using a hash-based indexing scheme built from the haplotyped short paths obtained by CHOP. Such an aligner would follow the graph alignment paradigm. In this case, a path index and a sparse graph index are used to seed and extend candidate sites from the sequencing reads and perform an optimal local graph alignment on the seeded subgraphs extracted from the sparse graph index. This would solve many problems when relying on third-party linear aligners, such as supporting the alignment of long or paired-end reads, and simplifying downstream analysis.

While this work facilitates the indexing and alignment of complex population graphs, substantial effort is still required to create a complete solution for general purposes. Although existing linear methods can perform the alignment and downstream variant calling, no streamlined tools are available to relate findings made using the graph back to a linearized representation. While working on CHOP, this was ultimately left out of scope. However, it is important to highlight the challenge of moving from one reference model (linear) to another (population graph), which is challenging both in terms of algorithmic complexity and end-user adoption. Furthermore, while a population graph can represent all types of variation, there will be caveats depending on the graphical representation chosen. These may arise at different levels of the genomic pipeline and introduce their respective computational complexities and ambiguities. Ideally, interoperability of different graph types would enable the implicit or explicit conversion between them in population graph tools so that implementation idiosyncrasies do not affect the end users of the tools, thus allowing for a more seamless experience of using various graphs. Unfortunately, we have not yet reached this point. For now, the linear reference model remains entrenched in genomics research and will remain so for some time, even as graph-based approaches become more prominent in the field. However, it is unlikely that this transition will happen quickly. The community has not yet reached a consensus on the standard graph reference model and related data formats. Indeed, the scientific community has been using the linear reference model for decades, and the effort to reach a consensus on the linear model and all of its related data formats has taken a long time, and this is likely to be even more true for graph models. The scientific community will likely adopt a graph-based approach in the long run. However, in the meantime, the community must continue to move forward and make progress on creating a standardized reference model for graphs so that this transition to graphs can happen more quickly.

II

Non-Invasive Prenatal Testing

4

Introduction

Our motivations are clear when it comes to understanding illnesses and determining why someone becomes afflicted while another person remains unaffected — it is one of the critical drivers for medical research, with many aspects of bioinformatics being used in the pursuit of both fundamental research to understand the reasons why people get sick, and diagnostic research to apply this knowledge gained to diagnose and treat patients effectively [86, 87]. The burden of treatment is often much higher when disease runs to its symptomatic course [88, 89]. Therefore, in many cases, it is preferable to ascertain the likelihood of being affected in advance, such that where possible, preventive measures may be taken to minimize or even avert disease [90, 91]. The circumstances or markers that give rise to or betray disease can be challenging to identify. However, recognizing these factors and associating them with specific diseases is essential to advance our understanding [92, 93]. Accordingly, this is often the first step in detecting afflictions promptly.

One way of framing the concept of disease is by treating it as an abnormality from the norm. Since only a small fraction of people will be affected by any given disease, this perspective can help understand how a particular condition can meet the criteria for classification as a disease. However, it can be challenging to define the "norm" [94], and it may be easier to reason about disease as something outside of what would be likely to occur from a statistical standpoint. Keep in mind that even if something falls within the realm of normality, it does not necessarily mean it is healthy [95]. For example, there can be healthy conditions considered atypical (blood-type AB), just as there can be unhealthy conditions that adhere to general standards (being overweight in the United States) [95, 96]. The complexity of diseases can make studying the transition between healthy and unhealthy states challenging. Namely, some diseases may be chronic but manageable, such as diabetes. Additionally, even if a disease can be categorized as healthy or unhealthy, the distinction may not always be clear-cut. For example, someone with early-stage cancer may still be considered healthy because they are asymptomatic. Nevertheless, breaking disease down into these categories makes it easier to set up studies that determine what causes this transition. The feedback loop for such studies first involves stratifying individuals into groups that are unaffected, thus healthy, and those that are affected due to physical or mental indicators specific to the disease under study. Depending on the objective, measurements are taken at different levels, which may include: morphology [97, 98], mental assessments [99], biomarkers [100], or genomics [101] all of which may be integrated across domains for additional power. Such studies expand our understanding of disease and yield new markers that can differentiate healthy from disease states. The loop may then be repeated with follow-up studies that attempt to dig even deeper to better understand these diseases. In doing so, we not only improve our understanding of these diseases but also potentially find new ways to diagnose and treat them more effectively.

The markers discovered through research enable the creation of diagnostic tools for the early detection of diseases, which can then be rapidly responded to at any stage of life. In the aging population, common screens are for cancer [102] and other age-related illnesses such as diabetes type II [103] or dementia [104]. It is routine for women between 20 and 30 to have a Pap test and HPV screening to look for pre- or early cervical cancer and the HPV virus, respectively [105]. Newborns undergo neonatal screenings such as heel pricks to detect rare genetic, hormone-related, and metabolic conditions [106]. Finally, prenatal testing may detect potential congenital disabilities in the fetus [107]. The knowledge of these markers can have a profound effect, given that some may be genetic and heritable, potentially affecting future generations, especially in the case of prenatal testing. Prenatal testing can help parents prepare for developmental difficulties in their child or make an informed decision to possibly terminate the pregnancy, which has led to a drastic reduction in the incidence of children born with congenital disabilities to older mothers [108]. Knowing that their child may be born with a genetic disorder can greatly influence parents' choices about conception. In some cases, couples may choose to conceive through in vitro fertilization to select healthy cells and avoid transmitting the disease [109]. In other cases, mitochondrial replacement therapy may directly interfere with affected cell lines and prevent the disease from being passed on [110, 111]. These options give parents more control over their fertility and enable them to make informed decisions about their family planning.

Depending on the stage of fetal development, different methods are used to detect symptoms associated with specific disorders caused by genetic abnormalities, such as chromosomal aneuploidy. For example, echography, or ultrasound imaging, is a non-invasive method to screen for fetal morphological inconsistencies [98, 112] that may indicate abnormalities like excess nuchal translucency [113] or the absence of the nasal bone [114]. These abnormalities are often associated with genetic disorders, such as Downs syndrome. In maternal serum screening, biochemical analysis measures protein levels, where significant shifts in the abundance of specific proteins may indicate chromosomal disorders [100, 115, 116]. Because the specificity of such screens is low, they are often combined to increase statistical power [112, 117]. If the screen is positive, the fetus's cells are often collected using invasive methods such as chorionic villus sampling [118] or amniocentesis [119] to obtain a definitive diagnosis through cell culturing and karyotyping [119].

As sequencing technologies develop, it becomes easier to study the genetic factors underlying diseases [120]. DNA changes can occur at different scales and may have numerous effects, ranging from harmful or neutral to beneficial. Alignment to the human reference genome and subsequent variant calling can help identify such genetic factors, which has led to the development of various methods specialized in detecting different types of variation. Many types of genetic variants can be linked to specific diseases, including single nucleotide polymorphisms [121], insertions, and deletions [122]. Large variants such as copy number variants (CNVs) [123] can range from thousands of kilobases to whole chromosomal duplications [124], deletions [125], inversions [126], and translocations [127]. Repetitious variants such as short tandem [128], long terminal [129], and interspersed repeats [130] are also common.

Prenatal testing through karyotyping can detect some disorders caused by chromosomal abnormalities [119, 131], but it has several limitations. For example, it can only detect abnormalities involving sizable chunks of DNA (*i.e.*, CNVs from 5 megabases and up), and the required cell culturing takes substantial time [131]. In addition, obtaining the fetal cells for this testing usually requires invasive methods that carry risks to both mother and child [117]. DNA sequencing technologies can detect these same disorders more accurately and with greater resolution. CNV detection using WGS relies on sequence alignment against a reference genome, followed by identification of variant sequences that differ between the query genome and reference. However, DNA sequencing is imperfect because biological and technical biases introduce artifacts in the sequence [132]. This bias may be minimized by sequencing more copies of the genome. By taking the consensus of these copies, actual variations may be more easily distinguished from the artifacts [133]. Though high-yield sequencing allows for more accurate genome characterization, it may not always be feasible for diagnostic applications because of the associated costs. Yet, CNVs may still be detected without high yield sequencing since the amount of sequencing yield largely dictates detection resolution; as yield decreases, so does the desired level of resolution, confining CNVs to larger and larger sizes. CNV detection methods that use low-yield WGS based on coverage typically rely on an initial step of discretization to overcome the sparseness in the count data. In this process, genome-wide per nucleotide coverage counts are aggregated into larger predefined regions. In simple terms, CNV detection methods compare a queried sample to a reference set of healthy controls (the norm) to discern (ab)normalities. The mean coverage and standard deviation of discretized regions are compared between the two sets to obtain a corresponding Z-score for each region in the surveyed sample. Ultimately, these scores are used to determine the presence of chromosomal abnormalities in the query sample.

The described methodology, while effective, is not ideal as it requires invasive procedures to obtain fetal cells. Instead, an alternative approach may be developed that relies on cell-free DNA (cfDNA) fragments [134]. These cfDNA are the remnants of DNA molecules that circulate in the peripheral blood and originate primarily from cell death through apoptosis or necrosis [135]. Circulating cfDNA is constantly produced throughout the body and has a relatively short half-life (4–120 minutes) [136, 137]. By isolating cfDNA from blood, it is possible to develop applications for cancer diagnostics [138], COVID-19 diagnostics [139], or organ transplant monitoring [140]. Sources of circulating cfDNA are not only limited to the host organism but also extend into the realm of microorganisms such as bacteria. Microbial cfDNA may be detected during fulminant infections such as sepsis [141] and is a possible source of antibiotic resistance gene interchange [142]. In pregnant mothers, a fraction of the cfDNA corresponds to the placenta [143], which in almost all cases barring rare events such as fetoplacental mosaicism [144], shares the same DNA as the fetus [143]. This cell-free "fetal" DNA (cffDNA) can already be detected in the early stages of pregnancy and the proportion of cffDNA within cfDNA increases as gestation progresses (2% → 20%) [145]. However, factors such as maternal body mass [146] and gestation type (singleton vs. multiple) [147] can affect this proportion.

CNV detection from low-yield cfDNA sequencing data for non-invasive prenatal testing (NIPT) is more challenging, given that the cffDNA proportion relative to the cfDNA is small, which can cause previously unproblematic biological and technical bias to overshadow the fetal signal and introduce distortions. Hence, the conventional CNV detection methods must be adapted to this setting to account for these biases. Adaptations may include: re-sequencing the reference set samples alongside the query samples to eliminate technical bias [148]; applying LOWESS [149, 150] or principal component correction [151, 152] to minimize bias induced by GC content and other systemic noise. Although joint re-sequencing can reduce technical bias, it cannot solve the issue of different cffDNA proportions in cases and controls. Additionally, re-sequencing is not cost-effective and limits the size of the reference set, which would ideally be large to minimize variance. An alternative solution, Wisecondor [150], proposes a within-sample testing methodology. With this method, the query sample is compared to itself, eliminating technical bias and differences in cffDNA proportions between samples. Wisecondor uses the observation that regions on one chromosome can be judged relative to the behavior of similarly behaving regions on other chromosomes within the same sample. This is assuming that any potential aberration is limited to a subset of chromosomes, so sufficient “normal” regions remain to compare. Therefore, the reference set of healthy samples is only used to establish an initial mapping of each region to similarly behaving regions on other chromosomes.

The proportion of available cffDNA, *i.e.*, the fetal fraction, is an essential factor in determining the reliability of CNV detection. If fewer cffDNA fragments are available in the maternal serum, *i.e.*, the fetal signal, it becomes increasingly difficult to obtain robust CNV calls. Various methods are available to estimate the fetal fraction based on: differential methylation [153], quantification of unique fetal SNPs [154], read count distribution differentials on the Y-chromosome [155] and other chromosomes [156], or differences in fragment length [157]. Methods in the latter category rely on the observation that cffDNA fragments are relatively shorter than maternal cfDNA [157]. If the fragment size (or a proxy thereof [158]) can be determined, enrichment in shorter fragments concerning the baseline can be used to estimate the fetal fraction. This estimation may be based on genome-wide fragment size distributions [156], or the relative positioning and count of aligned reads to nucleosome sites [158].

Part II of this thesis substantially derives from the developments of Wisecondor and its derivatives. Here, I first evaluate the performance of various NIPT methods in benchmarks using experimental and simulated data. The findings show that Wisecondor outperforms other NIPT methods. Additionally, I show that the differences in fragment size between cfDNA and cffDNA can be used to detect the presence or absence of chromosomal abnormalities, as previous work has also shown [159]. This is done by detecting shifts in the fragment size distributions across chromosome regions. Namely, for a fetus affected by trisomy, an overall lower fragment size is expected on the affected chromosome compared to the unaffected chromosomes. Finally, I show how the within-sample testing methodology can be generalized such that multiple datatypes may be included. The integration of read count and fragment size data results in better performance than if either datatype is used in isolation.

5

A comprehensive performance analysis of sequence-based within-sample testing NIPT methods

This chapter is based on 🗎 *T. Mokveld, Z. Al-Ars, E. A. Sistermans, and M. Reinders. A comprehensive performance analysis of sequence-based within-sample testing NIPT methods, PloS One.*

Abstract

Background: *Non-Invasive Prenatal Testing is often performed by utilizing read coverage-based profiles obtained from shallow whole genome sequencing to detect fetal copy number variations. Such screening typically operates on a discretized region representation of the genome, where (ab)normality of regions of a set size is judged relative to a reference panel of healthy samples. In practice such approaches are too costly given that for each tested sample they require the re-sequencing of the reference panel to avoid technical bias. Within-sample testing methods utilize the observation that regions on one chromosome can be judged relative to the behavior of similarly behaving regions on other chromosomes, allowing the regions of a sample to be compared among themselves, avoiding technical bias.*

Results: *We present a comprehensive performance analysis of the within-sample testing method Wisecondor and its variants, using both experimental and simulated data. We introduced alterations to Wisecondor to explicitly address and exploit paired-end sequencing data. Wisecondor was found to yield the most stable results across different region size scales while producing more robust calls by assigning higher Z-scores at all fetal fraction ranges.*

Conclusions: *Our findings show that the most recent available version of Wisecondor performs best.*

5.1 Introduction

Non-Invasive Prenatal Testing (NIPT) is designed to detect significant genetic abnormalities of the fetus, such as chromosome aneuploidies, sub-chromosomal copy number variations (CNVs), and unbalanced translocations. The discovery of cell-free fetal DNA (cffDNA) in maternal peripheral blood [160], has enabled the development of NIPT methods for detecting genetic anomalies [161–163]. NIPT demonstrates high sensitivity and specificity for prevalent chromosomal aneuploidies, including trisomy 21, 18, and 13 [164, 165], and can also be employed for other autosomes [166, 167], and sub-chromosomal events [168–170]. However, detecting smaller CNVs requires increased depth of coverage, which can be prohibitively expensive in practice [171].

Moreover, limitations exist, such as support for only the most common trisomies and test failures due to complications like placental mosaicism or maternal copy number variation [163, 172]. In practice, NIPT methods utilizing whole genome sequencing (WGS) often depend on extremely low sequencing yield (~0.25x coverage) for cost-effective clinical applications [173, 174]. Alongside low coverage, the assumption that sufficient cffDNA is present in maternal plasma is another factor. The cffDNA typically contributes a small fraction (2–20%) to the total cfDNA sequenced, complicating accurate identification of CNVs [175, 176]. The available cffDNA in the sample, or fetal fraction, directly influences the reliability of detected CNVs, with a higher fetal fraction corresponding to increased confidence in CNV detection [177].

Coverage-based NIPT methods typically follow similar steps to detect CNVs [148, 178–181]. Initially, DNA is isolated from maternal plasma and subjected to low-coverage sequencing. Subsequently, the chromosomal origin of each read is identified through alignment with the human reference genome. These read alignments are counted and discretized into a coarse representation, discretizing the genome into equally sized regions. Lastly, the read coverage per region is statistically compared to a reference panel consisting of healthy samples, enabling the determination of significant deviations from the expected signal in each region. However, this described methodology has a major limitation: the need to re-sequence control samples for each test sample to mitigate technical biases.

An alternative that improves upon these methods is Wisecondor [150], which circumvents read frequency variation across different samples. In Wisecondor, each tested region is compared to a set of reference regions on other chromosomes within the same sample, exhibiting similar behavior. Such within-sample comparisons eliminate between-sample bias and differences within the fetal fraction, as regions with similar characteristics will behave similarly within the test sample, and all regions undergo the same experimental procedures. Wisecondor is freely available and can be modified and enhanced by the scientific community, as demonstrated with WisecondorX [152], designed as a general solution for WGS applications beyond NIPT. One limiting factor of Wisecondor is the use of Stouffer's Z-score sliding window method to segment and score events, which exhibits exponential computational complexity relative to decreasing region size. Generally, this is not an issue in shallow NIPT, given that only a fraction of the available DNA pertains to the fetus, necessitating the use of a larger region size.

Wisecondor was developed during a period when NIPT sequencing predominantly employed single-end technologies. Although Wisecondor can process paired-end data, it does not fully utilize the supplementary information that this technology provides. In this study, we introduce modifications to Wisecondor to incorporate pairing information and include this in a benchmark. We compare the modified Wisecondor versions to the original Wisecondor, WisecondorX, and CNVkit, an alternative approach for detecting CNVs, using both experimental and synthetic data.

5.2 Results

Initially, the detection of larger and common aneuploidies was investigated. A total of 526 samples were used (Section 5.4.1), of which 401 were confirmed negatives and served as controls, while the remaining 125 were confirmed trisomy 21 positive. The average estimated fetal fraction of all samples was approximately 7.5% (Section 5.4.1). All samples were aligned using BWA-mem to the hg19 human reference genome with an average depth of coverage of 0.257x across all samples. Wisecondor was modified to utilize full alignment read counting and read pairing (Section 5.4.2).

The different versions of Wisecondor are referred to as follows:

1. *WCR*: the baseline implementation of Wisecondor in which no read pairing is available and alignments are counted based on start positions.
2. *WCR+SE*: a modified implementation of Wisecondor utilizing the full but unpaired read alignments to determine region counts.
3. *WCR+PE* and *WCR+PEI*: both use the read pairing information when aligning reads before determining region counts, with WCR+PEI also contributing to region counts for the insert size of every paired read.
4. *WCRX*: WisecondorX.
5. *CNVkit*: a general-purpose CNV detector that showed the best competitive performance across different non-Wisecondor NIPT methods [152].

These versions were compared using the given dataset to assess their performance in detecting aneuploidies

5

5.2.1 T21 detection performance

The most direct measure of performance is the detection rate of expected aberrations in validated samples. A comparison of how the methods operate at different region sizes reveals that the majority of all expected T21s are detected by each method (Table 5.1) At the 250 kb scale, *WCR+PEI* performs the worst, while *WCR* and *WCR+SE* outperform the others, with overlapping calls. It is important to note that the originally published version of *WCR* performs significantly more poorly than the most recent version, for example, 118 calls at 250 kb; hence, remaining results of this version are omitted. In almost all cases, *WCR* outperforms the other methods, except for *WCR+PEI* at a 50 kb resolution. The non-Wisecondor method, CNVkit, is competitive with the Wisecondor methods; however, this performance is expected given the relative ease of detecting whole chromosome events.

Considering the scale of events in this context, it is unsurprising that the detection rate remains relatively stable as the region size increases. Performance only declines significantly at the 50 kb scale, indicating a failure to segment the events due to increasing variability within the normalized regions as they become smaller.

Table. 5.1. The number of ≥ 10 Mb events with Z-scores ≥ 5 on chromosome 21 detected by the methods in all T21 positive samples for varying region sizes. Note that *CNVkit* does not depend on region-sizes so only one performance measure is reported.

	50 kb	100 kb	250 kb	500 kb	750 kb	1 Mb	5 Mb	10 Mb
WCR	113	**122**	**123**	**124**	**123**	**123**	**122**	**121**
WCR+SE	114	121	122	122	**123**	121	121	**121**
WCR+PE	117	120	121	121	120	120	120	118
WCR+PEI	**118**	119	119	119	120	119	118	118
WCRX	78	118	122	122	120	121	120	103
CNVkit	120							

5.2.2 Region size relates to false positives

While the majority of all expected trisomies were detected by the methods, it is crucial to keep the number of false positives low. Since discretizing the genome into regions results in a signal/noise trade-off, we summarize the sensitivity across region sizes for all methods (Figure 5.1). As expected, a smaller region size increases the number of false positives detected by each method, except for *WCRX*. This stable performance might be explained by the segmentation algorithm that derives setting breakpoints from the variance across a segment, also resulting in a lower detection rate for small region sizes. The number of false positives for CNVkit is relatively large, giving an advantage to the Wisecondor methods with a larger region size that have better control of false positives.

Interestingly, *WCR+PE* and *WCR+PEI* become more sensitive compared to *WCR* at lower region sizes, whereas this relation reverses when the region size increases. This may be explained by the use of read pairing and insert size padding, which exaggerates large fluctuations in the signal while smoothing smaller fluctuations. If the region size becomes smaller, it is expected that the signal itself becomes noisier and therefore fluctuates more, causing the exaggerating effect to detect more events on the lower end. Conversely, the signal becomes increasingly smooth as the region size increases, which can lead to overly aggressive smoothing, potentially masking true fluctuations.

5.2.3 Per-region Z-score differentiation

To understand method-specific differences, the per region Z-scores were aggregated across all negative and T21 positive samples, as shown in Figure 5.2. *WCR* and *WCR+SE* perform identically at the 250 kb scale, a similar observation as in the previous comparison in Table 5.1. Furthermore, it is clear that *WCR+PE* and especially *WCR+PEI* yield overall lower Z-scores, which can eventually push a potential CNV call below the Z-score threshold. At this scale, the inclusion of read pairing (*WCR+PE*) smooths the read count signal, even more so when also including the insert size padding (*WCR+PEI*).

So far, evidence has shown that this smoothing adversely impacts the detection rate because of overall lower Z-scores at most resolution scales. However, an improvement can be noted with *WCR+PE* and *WCR+PEI*, at the cost of many additional false positives, when the region size becomes smaller, being the best performing at a 50 kb scale (Table 5.1). In fact, the average per region Z-scores at a 50 kb scale show that the scores of these two methods are overall higher than those of the others (Supplemental Figure S5.1).

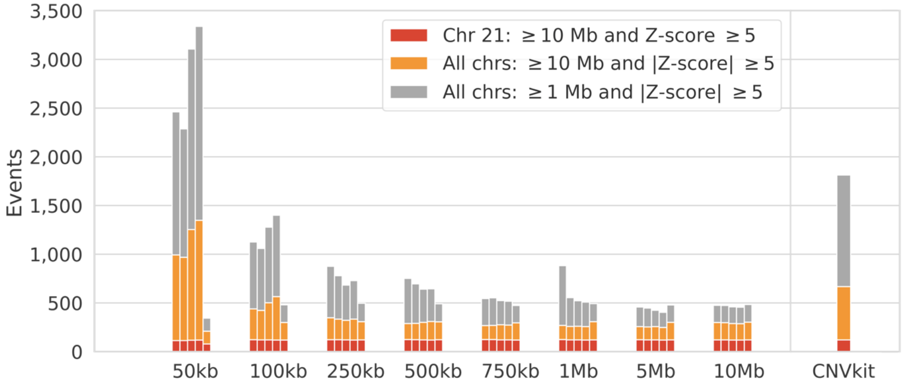


Figure. 5.1. Stacked counts of detected CNVs relative to region size and method, ordered from left to right as: *WCR*, *WCR+SE*, *WCR+PE*, *WCR+PEI*, *WCRX* (*CNVkit* results are displayed separately). Bar coloring denotes mutually exclusive filtering constraints. Red: filtering for ≥ 10 Mb CNVs with Z-score ≥ 5 (only including duplications) and only on chromosome 21. Orange: filtering for ≥ 10 Mb CNVs with an absolute Z-score ≥ 5 (thus also including deletions) and on all chromosomes. Gray: identical to the previous but for all CNVs ≥ 1 Mb.

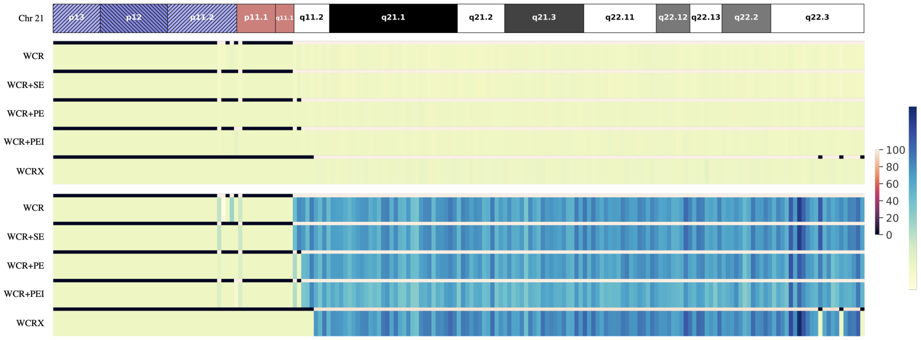


Figure. 5.2. Heatmaps of the summed per region Z-scores across all negative (top) and all T21 positive (bottom) samples at a 250 kb region scale for chromosome 21 and all different Wisecondor-based methods. The line above each method's heatmap corresponds to the average number of selected reference regions for each region of that method (black denoting that no similar reference regions are found and consequently these regions are excluded).

5.2.4 Detection power relative to fetal fraction

To further understand the differences in Z-scores between the methods, the Z-scores of individual events were examined in relation to the estimated fetal fraction of each sample. It is expected that samples with higher fetal fractions can be more reliably tested for CNVs, implying that the Z-scores of events are likely higher for calls made in samples with higher fetal fractions.

The analysis reveals a positive correlation between the fetal fraction and the Z-score of the detected events for all methods (Figure 5.3 and Supplemental Figure S5.3 for other region sizes). At a region size of 250 kb, WCR assigns higher Z-scores to detected events than any other method (with the exception of WCR+SE, which performs nearly identically).

This demonstrates that WCR has better power to detect expected T21s and thus detects events with greater confidence.

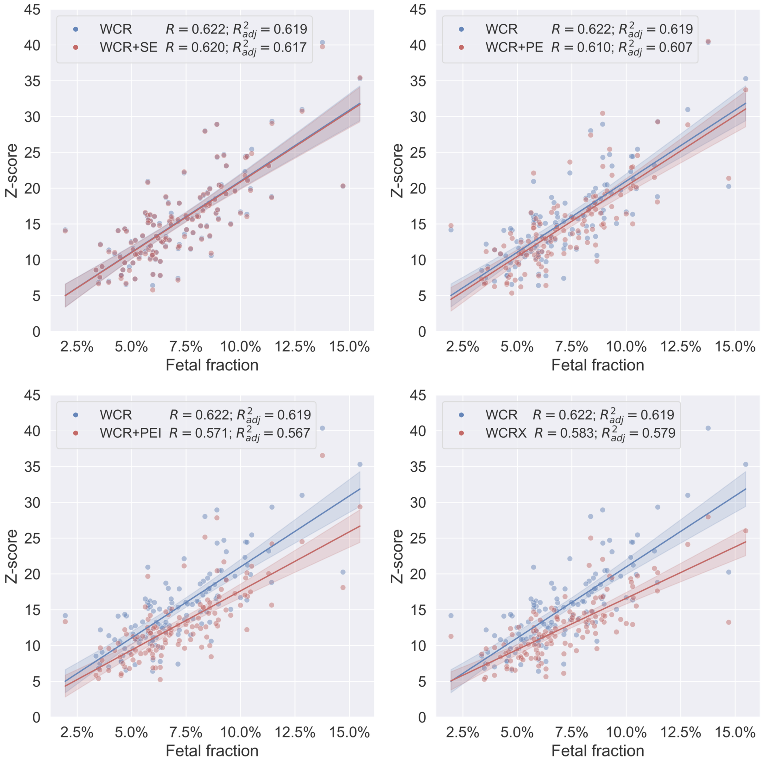


Figure. 5.3. All ≥ 10 Mb events with Z-scores ≥ 5 on chromosome 21, detected by the different methods for a 250 kb region size in the T21 positive samples, relative to the estimated fetal fractions of each sample. Each plot compares one of the methods *WCR+SE*, *WCR+PE*, *WCR+PEI*, and *WCRX* (all in red) with *WCR* (in blue), and each point corresponds to an event within a sample. The legend annotation relates each method with the respective R value and adjusted R squared value of the linear fit.

5

5.2.5 Performance in simulated data

Since there is no true ground truth available in the experimental data, performance could not be measured in concrete terms, such as on the breakpoints of events. To address this, synthetic read data was created (Section 5.4.3). A total of 400 positive samples were generated, equally split for duplications, deletions (both at varying scales), and sex (*WCRX* generates sex-specific reference sets), with a constant fetal fraction set at 7.5%. Additionally, 100 negative samples were generated, which were used to build reference sets. With the ground truth known, the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates can be determined directly from the intervals of the simulated events, the detected events, and the chromosomes (Section 5.4.4).

Figure 5.4 shows an overview of the F1 performance of the methods for different simulated events and event sizes. Table 5.2 shows the average F1 of duplications and deletions across all event scales and false positives calls. The overall performance of all methods is similar, with *WCR* performing best. Peculiar to *WCRX* is that the performance is poor in samples with 1 Mb events while stabilizing from 5 Mb and up until regressing again as the event size increases from 30 Mb and beyond. It is unclear what causes this behavior. However, in terms of false positive calls on other chromosomes (Table 5.2), *WCRX* does best with the fewest additional calls across all samples. *CNVkit*, increasingly struggles as the event size becomes smaller, calling many false positives.

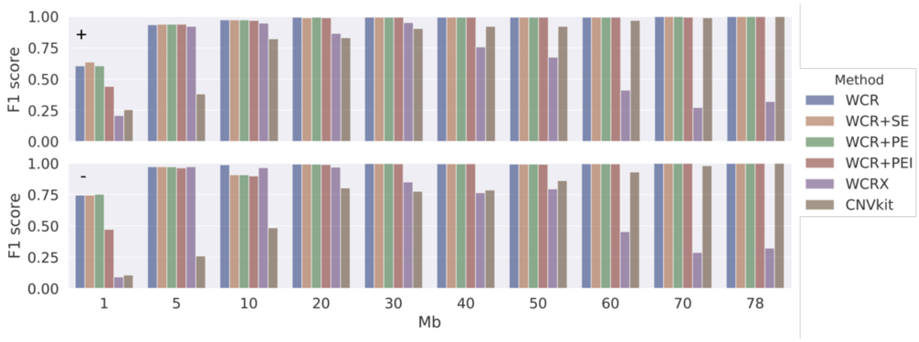


Figure. 5.4. F1 score (vertical axes) for simulated duplication (+, top panel) and deletion (-, bottom panel) events of varying sizes (1 to 78 Mb) at a fixed fetal fraction of 7.5% for the different methods using a 250 kb region size.

Table. 5.2. F1 score as averaged across all event sizes (1 to 78 Mb) for simulated duplications (+), deletions (-), and the number of false positive calls (FP) of the different methods using a 250 kb region size.

	F1(+)	F1(-)	FP
WCR	**0.996**	**0.995**	33
WCR+SE	0.995	**0.995**	32
WCR+PE	0.995	**0.995**	30
WCR+PEI	0.994	0.986	40
WCRX	0.577	0.605	2
CNVkit	0.929	0.845	556

5.2.6 Performance in challenging simulated data

The previous results highlight that, generally, events larger than 5 Mb can be detected by all methods, which signals that the real challenge lies within events that are smaller than this. These smaller events are also on the edge of what remains detectable with respect to the samples' read coverage and fetal fraction. To further investigate this, 480 additional samples were simulated, now with aberrations ranging from 250 kb to 5 Mb and fetal fractions ranging between 1–6%. Methods were again evaluated for a 250 kb region size, and the results are shown in Figure 5.5 for a fetal fraction of 6%, and the accumulated results for all fetal fraction ranges in Table 5.3.

As events become smaller than 1 Mb, the likelihood of detecting an event becomes extremely small under the given conditions. As expected, an increasing fetal fraction improves performance of every method. As the fetal fraction increases (>4%) *WCRX* performs closely to the other Wisecondor methods, but does worse otherwise. Note that *WCRX* does not exhibit the dramatic drop in performance at lower event scales as was observed when events became larger than 20 Mb (Figure 5.4). The performance of *CNVkit* is again significantly worse than the Wisecondor methods when events become increasingly smaller, even more so for deletions, which is not the case for the other methods which are balanced for both types of events.

When utilizing the insert size padding of *WCR+PEI*, the performance for these smaller events is always worse than with the other Wisecondor methods. This result is consistent with the previous findings when varying the region size. In terms of false positive calls on other chromosomes across all samples, *WCRX* again does best with the fewest additional calls (Table 5.3).

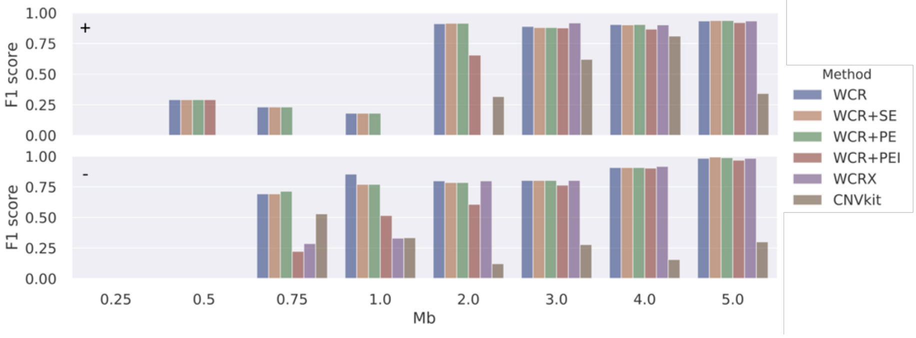


Figure. 5.5. F1 score for simulated duplication (+, top panel) and deletion (-, bottom panel) events of varying sizes (0.25 to 5 Mb) at a fixed fetal fraction of 6% for the different methods using a 250 kb region size.

Table. 5.3. F1 score as averaged across all event sizes (0.25 to 5 Mb) for simulated duplication (+) and deletion (-) events across varying fetal fractions (FF), and the number of false positive (FP) calls of the different methods at a 250 kb region size.

	FF	WCR	WCR+SE	WCR+PE	WCR+PEI	WCRX	CNVkit
+	1%	0	0	0	0	0	0
	2%	**0.108**	**0.108**	**0.108**	0	0	0
	3%	0.598	0.572	**0.609**	0.379	0.327	0.114
	4%	0.759	**0.763**	0.760	0.511	0.536	0.237
	5%	0.839	0.838	**0.840**	0.775	0.770	0.408
	6%	**0.837**	0.836	**0.837**	0.794	0.769	0.474
-	1%	0	0	0	0	0	0
	2%	0.227	**0.297**	0.247	0.090	0.156	0.125
	3%	0.522	**0.525**	0.524	0.295	0.380	0.155
	4%	0.665	**0.715**	0.710	0.658	0.696	0.157
	5%	**0.845**	0.842	0.842	0.781	0.808	0.190
	6%	**0.863**	0.862	0.862	0.799	0.835	0.210
	FP	125	117	126	134	1	515

5.3 Discussion

Detecting CNVs in highly imbalanced mixed samples continues to present significant challenges. These difficulties are intensified within the context of low-yield NIPT, where distinctions between fetal, maternal, and noise signals become increasingly indistinct. The signal-to-noise ratio must be enhanced to a level that permits confident calls, even at the cost of reduced resolution. This issue is particularly pronounced at lower fetal fractions. To address this trade-off, read count signals are commonly aggregated across larger genomic ranges, which effectively discretize the genome into regions. Consequently, the selected region size reflects a balance between resolution, noise, read coverage, and fetal fraction considerations.

Within-sample correction methods, such as Wisecondor, provide a solution to the challenge of re-sequencing control samples for each new set of samples requiring testing. Our work demonstrates that, in the context of non-invasive prenatal testing (NIPT) applications, Wisecondor (WCR) consistently outperforms other tested methods across both experimental and synthetic data. This superior performance is evidenced by more accurate detection and higher Z-scores assigned to identified CNVs Overall, all Wisecondor methods surpass the non-Wisecondor method, *CNVkit*, particularly when examining lower fetal fractions and smaller CNVs. This underscores the value of tailoring methods to specific domains, even when addressing a similar underlying issue, such as CNV detection. A notable advantage of WisecondorX lies in its ability to identify fewer false positives, resulting in a higher likelihood that called CNVs are genuine events, albeit with a lower detection rate. In an NIPT setting, reducing false positives is beneficial, as all findings are typically scrutinized, thereby enhancing workflow efficiency with WisecondorX

Several modifications were introduced to Wisecondor. Firstly, instead of counting the starting position of each read, the complete alignments were processed (*WCR+SE*). Despite expectations, this modification did not yield significant benefits in this particular setting due to the sparseness of read data, resulting in outcomes nearly identical to the original Wisecondor (*WCR*). However, such processing might prove advantageous with increased coverage and smaller region sizes, as the impact of this change would be amplified. Since Wisecondor does not utilize read pairings, we incorporated a modification that leverages paired-end reads (*WCR+PE*), and an adaptation that employs additional padding based on the insert size derived from these reads (*WCR+PEI*). Our experiments demonstrated that these modified versions allow for heightened sensitivity at smaller region sizes by accentuating fluctuations within the coverage signal. However, this improvement comes at the expense of excessive signal smoothing at larger region sizes, resulting in diminished sensitivity. Although signal smoothing proved beneficial, it was only advantageous when region sizes were significantly reduced, and this came with the trade-off of increased false positive calls.

The impact of incorporating read pairing in this study yielded only modest improvements in the performance of the methods. In some cases, using paired-end reads as single-end reads demonstrated superior results. Nevertheless, additional information within paired-end reads remains untapped. Potential enhancements for methods that utilize paired-end reads could exploit the fact that fetal DNA fragments exhibit shorter fragment sizes compared to maternal fragments [157]. This characteristic is attributed to the underlying mechanisms involved in DNA fragmentation, such as DNA methylation and its relationship to chromatin accessibility [182, 183]. The fragment size differences can be inferred from the insert size of aligned paired-end reads. Several methods have been proposed to leverage this size difference for detecting large chromosomal CNVs [159] or fetal de novo point mutations [184]. Moreover, combining fragment size differences with read count signals has enabled improved detection of fetal aneuploidies, as demonstrated by the COFFEE algorithm [185], a reference-free method that requires no control samples. Our group developed WisecondorFF [186] to extend the WISECONDOR within-sample testing framework and facilitate the combination of read count and fragment sizes for enhanced (sub)chromosomal CNV detection. We also note that method performance is significantly influenced by region size selection. Therefore, employing multiple region scales may prove advantageous, allowing a CNV to gain support across various scales rather than only one.

5.4 Methods

5.4.1 Sample specification and pre-processing

All samples in this study were obtained from the Dutch TRIDENT study [187]. DNA isolation, library preparation, and paired-end sequencing (36 bp) were conducted using the Illumina VeriSeq1 sequencing protocol, in compliance with the supplier's recommendations (Illumina, San Diego, USA). Both the Veriseq algorithm (which detects only trisomies 21, 13, and 18) and Wisecondor (which identifies other trisomies and smaller events) were used for analysis. A total of 526 samples were selected for this study, with 401 having no detected chromosomal aberrations, thus serving as negative controls. The remaining 125 samples tested positive for T21. All samples underwent similar pre-processing and were aligned to the hg19 human reference genome, excluding any decoy sequences, using BWA-

0.7.17 mem [31]. To measure both single-end and paired-end performance, read sets were aligned in both settings. In the single-end setting, each read pair was individually aligned before merging the output alignments, as opposed to the paired-end setting, in which all alignments were output simultaneously. The average coverage for the 401 negative samples was 0.258, while the 125 T21 positive samples had an average coverage of 0.256. SeqFF was employed to estimate the fetal fractions of all samples, with an average fetal fraction of 7.5% [156]. The final alignments were compressed and remained unfiltered, as all methods internally managed the quality of the alignments.

For each of the samples, initial genome wide region counts were generated at a 5 kb resolution, and subsequently scaled up to 50 kb, 100 kb, 250 kb, 500 kb, 750 kb, 1 Mb, 5 Mb, and 10 Mb for the construction of reference panels and/or sample testing against these references. All 401 control samples were used to build reference panels at every region scale for each of the five methods: *WCR*, *WCR+SE*, *WCR+PE*, *WCR+PEI*, and *WCRX*. Rather than employing a predefined blacklist to exclude genomic regions from the analysis, the methods determined such regions based on normalized region counts.

5.4.2 Wisecondor modification

We implemented three distinct modifications to Wisecondor (version commit: 9e95c75; note that this is not the published version of the algorithm but an updated variant). One modification leverages the complete read alignment instead of the starting position (*WCR+SE*), and the other two exploit the read pairings, one necessitating proper read pairing (*WCR+PE*) and the other utilizing the fragment size associated with paired reads (*WCR+PEI*).

1. *WCR+SE*: The method of counting aligned reads in the discretized regions was modified and is based on the full alignments, rather than the starting positions of the aligned reads. Doing so the full information content of an alignment can be utilized, as well as allow reads to partially contribute to multiple regions. All modified variants of Wisecondor, denoted as *WCR+*, employ this alternate read counting.
2. *WCR+PE*: Wisecondor does not utilize paired-end read information, but treats any aligned read as single-end. By requiring any read to be in a properly paired pair, an additional constraint is imposed for a read to be considered.
3. *WCR+PEI*: Additionally, the insert size distance between two properly paired fragments can be utilized to further pad the read coverage such that the fragment between the aligned read pairs also contributes to the total contribution of the reads. By using read pairing, the read count signal is smoothed and large fluctuations within the signal are exaggerated. Such signal smoothing and exaggeration is especially prominent in *WCR+PEI* (Supplemental Figure S5.2).

5.4.3 Data simulation

Chromosome 18 was selected to simulate duplications and deletions of size: 1, 5, 10, 20, 30, 40, 50, 60, 70, and 78 Mb. For each CNV size, 10 samples were simulated for both sexes and variation types, resulting in a total of 400 samples. Additionally, 100 negative samples were simulated, with 50 of each sex. To generate an aberrated sample, a random starting position was chosen for the event on chromosome 18, excluding highly repetitious regions, using a uniform distribution. From the starting position to the end position, the sequence is either deleted, leaving flanking segments of N's (larger than the fragment size of the simulated reads) at the site, or duplicated, leaving flanking segments of N's between the two sequences. For each aberrated sample, complete fetal reference sequences were generated by replacing the original chromosome 18 sequence with the modified one. The sex determined the inclusion or exclusion of the Y chromosome or one X chromosome.

For each sample, a total of 21,000,000 36 bp paired-end reads were simulated using Mason 2.0.9 [81]. The read sets were generated to yield a 7.5% fetal fraction when combined. Considering the shorter fetal fragment length compared to maternal fragments, the mean fragment length parameter of fetal samples was slightly reduced. A maternal read set (reads prefixed with M and totaling 19,425,000 reads) was then simulated, using the hg19 reference sequence without chromosome Y and decoy sequences, and a fetal read set (reads prefixed with F and totaling 1,575,000 reads) was created from the previously generated fetal reference sequences. The two disjoint paired-end read sets were ultimately merged and prefix sorted to eliminate any aligner bias.

5.4.4 Performance metrics

Within the simulated data the ground truth of a CNV, T, can be represented as an interval, denoted as $[T_s, T_e]$, which may be contained within a chromosome, G, *i.e.*, a larger interval, as $[G_s, G_e]$. For a CNV predicted by one of the methods, P, the same can be done, as the interval $[P_s, P_e]$, or the lack thereof as $[P_s = 0, P_e = 0]$. The true positives, can then be derived to be $TP = \max(0, \min(T_e, P_e))$; the false negatives, as $FN = T_e - T_s - TP$; the false positives, as $FN = P_e - P_e - TP$; the true negatives, as $TN = G_e - TP - FN - FP$. The F_1 performance score, *i.e.*, the harmonic mean of the precision and recall, may then be calculated as: $F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$.

5.5 Supplementary materials

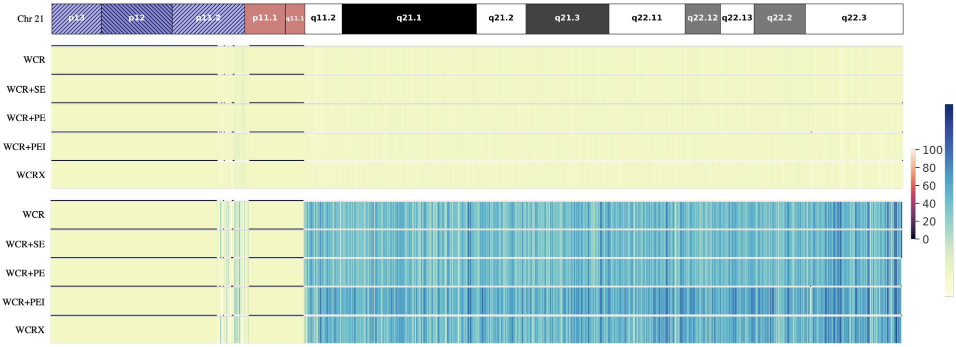


Figure. S5.1. Heatmap of the summed per region Z-scores across all negative (top) and all T21 positive (bottom) samples at a 50 kb region scale for chromosome 21 and all different Wisecondor-based methods. The line above each method's heatmap corresponds to the average number of selected reference regions for each region of that method (black denoting that no similar reference regions are found and consequently these regions are excluded).

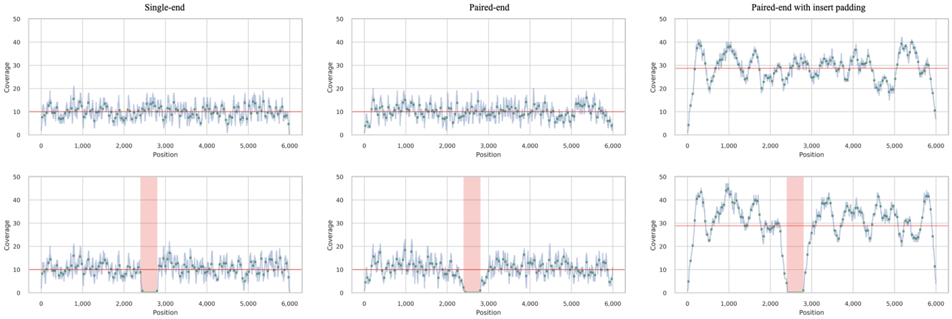


Figure. S5.2. Simulated read alignment counts across a small genome of single-end, paired-end, and paired-end with insert padding methods. On top this is shown for an unaffected sample, and on the bottom for a sample with a 400 bp deletion.

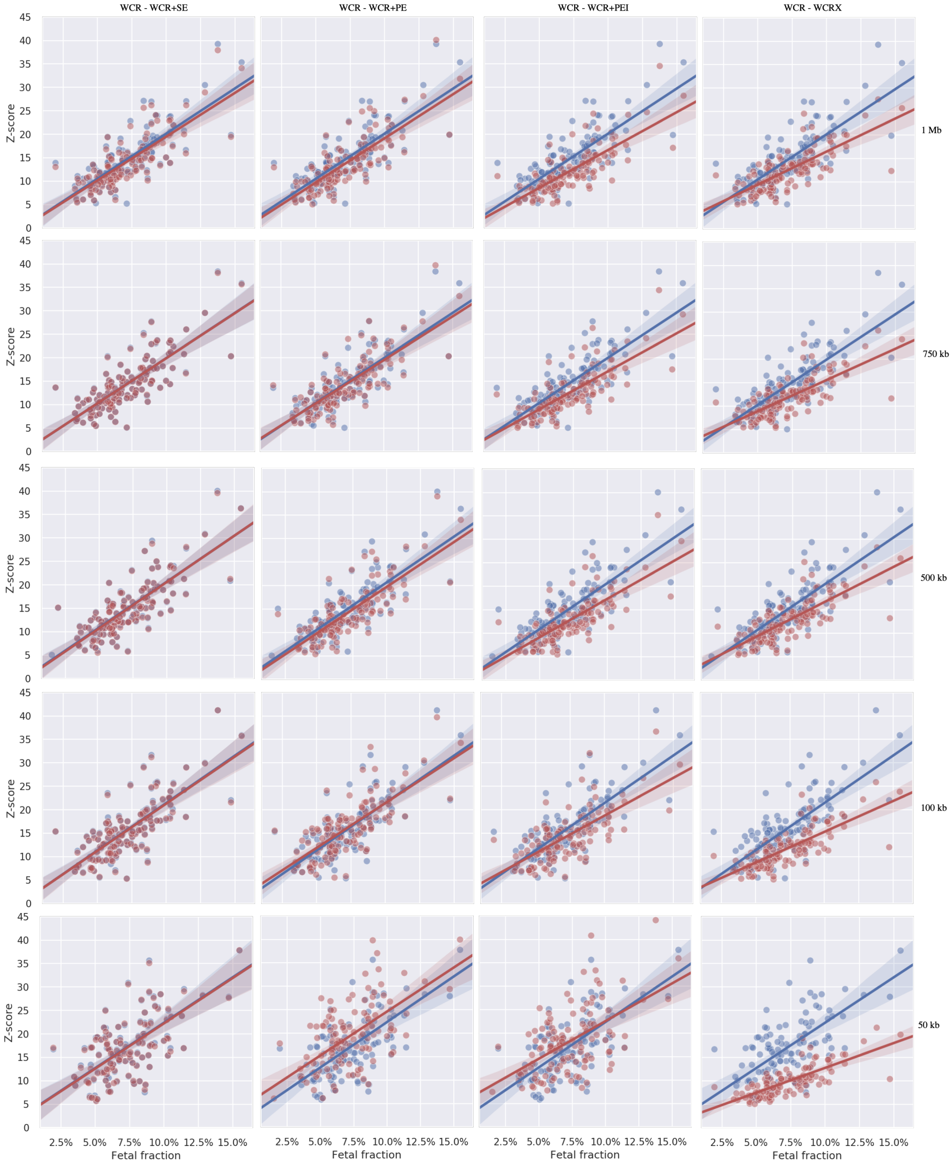


Figure. S5.3. All ≥ 10 Mb events with Z-scores ≥ 5 on chromosome 21 detected by the different methods in the 125 T21 positive samples relative to the estimated fetal fractions of each sample. Each plot displays one of the methods *WCR+SE*, *WCR+PE*, *WCR+PEI*, and *WCRX* (all in red) overlaid with *WCR* (shown in blue), for a set region-size resolution. Each point corresponds to a CNV within a sample.

6

WisecondorFF: Improved Fetal Aneuploidy Detection from Shallow WGS through Fragment Length Analysis

This chapter is based on *T. Mokveld, Z. Al-Ars, E. A. Sistermans, and M. Reinders. WisecondorFF: Improved Fetal Aneuploidy Detection from Shallow WGS through Fragment Length Analysis, MDPI Diagnostics'21* [188].

Abstract

In prenatal diagnostics, NIPT screening using read coverage-based profiles obtained from shallow WGS data is commonly used to detect fetal CNVs. From these same data, fragment size distributions of fetal and maternal DNA fragments can be derived, which are known to differ and are often used to infer fetal fractions. We argue that fragment size has the potential to aid in the detection of CNVs. By integrating, in parallel, fragment size and read coverage in a within-sample normalization approach, it is possible to construct a reference set encompassing both types of data. This reference then allows for the detection of CNVs in queried samples using both data sources. We present a new methodology, WisecondorFF, that improves sensitivity while maintaining specificity over existing approaches. WisecondorFF increases the robustness of detected CNVs, and can reliably detect even at lower fetal fractions (<2%).

6.1 Introduction

Prenatal screening is routinely used to measure and verify the fetus's health, including the detection of chromosomal CNVs (copy number variations), in a timely manner [107]. Most prenatal screening methods are now non-invasive. Invasive methods generally provide more conclusive results but confer a low risk of harming the mother, or fetus [118, 119, 189]. For example, verifying chromosomal aneuploidies through invasive genetic analysis carries a small but significant risk of causing fetal miscarriage [117]. Non-invasive methods assess fetal health indirectly, such as through morphological properties using ultrasound scans [98, 112], or biochemical markers by maternal serum sampling [100, 116]. In general, these screening methods must be performed at specific stages of pregnancy [190] and are usually deployed in parallel [112, 117], thus expanding the range of detectable conditions [190–192].

Since the discovery of cell-free fetal DNA (cffDNA) in maternal peripheral blood [160], it has become possible to measure fetal DNA without invasive procedures [161]. This discovery has opened up new possibilities for safely assessing fetal health using genetic markers [163] and can be used to detect a wide variety of pathologies caused by events such as chromosomal aneuploidy [161, 163], sub-chromosomal CNVs [193], and single-gene mutations [194]. One of the deciding factors for detecting smaller and smaller events is related mainly to sequencing yield, *i.e.*, the DNA coverage, which requires more sequencing to detect smaller events reliably. The other important factor is the proportion of cffDNA mixed with maternal blood, *i.e.*, the fetal fraction, which typically contributes 2–20% of the overall available DNA pool [177, 195] and increases as the pregnancy progresses. The combination of available coverage, fetal fraction, and CNV size determines the reliability of detection of a given aneuploidy.

6

In clinical practice, NGS-based non-invasive prenatal screening (NIPT) with cffDNA typically uses shallow whole-genome sequencing (WGS) to remain economical and accessible for mass screening [101]. Such low sequencing yield leads to practical limits when detecting events and often means that only high sensitivity and specificity are possible for larger events such as chromosomal aneuploidies. Therefore, NIPT typically tests for common trisomies, such as 21, 18, and 13 [164], since accuracy degrades too much for sub-chromosomal CNVs smaller than 5–10 megabases (Mb) [169, 170, 196]. Nevertheless, even for larger events, one must be careful of discordant results caused by biological phenomena such as placental mosaicism or maternal copy number variation [197, 198]. While the NIPT field is dominated by methods utilizing WGS, some use RNA-seq [199], methylation profiles [200], SNPs [201], or haplotyping [202]. Each of these can effectively detect specific events, especially when such data sources are integrated [199].

Most low sequencing yield NIPT methods use similar steps to detect events [148, 178, 203], starting with sequencing the DNA, aligning sequencing reads to a reference genome and finally detecting whether the observed read coverage exceeds expectations based on a reference baseline. Because coverage is extremely low, detection is not done at the level of individual nucleotides, but rather the genome is discretized into larger, equally sized regions or bins (often 250 kb to 1 Mb) in which read counts are aggregated to obtain a signal that can be compared to a baseline. Such a baseline signal, also known as a reference set, can be derived from a collection of healthy samples [174]. Although this method can be effective, there are some drawbacks. For example, the experimental conditions of the sample and healthy samples used to establish the baseline must be identical to eliminate any confounding technical bias in detection. Otherwise, false positives or negatives are easily obtained [150]. Re-sequencing the baseline with a new sample would protect against these biases but is extremely expensive.

An alternative approach to defining the baseline would be to compare the observed number of reads with expectations derived from the sample itself. In such a framework, a region should be compared to regions on other chromosomes that have been found to behave similarly in a healthy panel. By generating a map of similarly behaving regions from the healthy sample set, any NIPT sample can be tested by exploiting this map without (re)using the healthy sample set. This setup is effective when the number of events is typically limited to a subset of chromosomes at a time and is successfully adopted in clinical practice using methods such as Wisecondor and WisecondorX [150, 152].

It is known that fetal DNA fragments are shorter in size than maternal fragments [202]. Previous work has exploited this observation to predict the fetal fraction in cfDNA samples, *i.e.*, the relative proportion of maternal and fetal DNA present in a sample [154, 156]. The fragment size of cfDNA can be inferred from paired-end reads but also by other approaches, such as with methylation profiles [204] or based on read abundance approaches inside and outside nucleosomes [158]. Since fetal cfDNA fragments are shorter in size than maternal fragments, this can potentially be used to infer fetal read abundance in a given genomic region. Namely, if a fetus is affected by trisomy, an overall smaller fragment size is expected on the affected chromosome compared to unaffected chromosomes. We sought to enrich the current NIPT with-sample testing procedure with these available data, as it is currently standard to generate paired-end reads.

We present WisecondorFF, a methodology that detects fetal chromosomal aberrations from cfDNA by combining read coverage-based estimates and fragment size statistics. To control for variation, fragment size statistics are derived using a similar within-sample normalization approach operated by current read counting procedures such as Wisecondor. We show that chromosomal CNVs can indeed be detected from the inferred fragment sizes across the genome and that, combined with read coverage, this leads to improved accuracy and robustness of the NIPT procedure. As such, WisecondorFF is attractive to clinical practice because the data are readily available at most clinical diagnostic facilities since it relies only on paired-end sequencing of maternal serum DNA.

6.2 Methods

6.2.1 Sample specification and processing

Samples were generated as part of the Dutch TRIDENT study [187]. DNA isolation, library preparation, and paired-end sequencing (36 bp) were performed using the Illumina VeriSeq1 sequencing protocol, according to the vendor's recommendations (Illumina, San Diego, USA). Analysis was performed by the Veriseq algorithm (which detects only trisomies 21, 13, and 18) and Wisecondor, which also detects other trisomies and smaller events. For this study, we selected 526 samples, 401 of which had no chromosomal CNVs detected, and were used as negative controls. The remaining 125 samples all tested positive for T21. The average depth of coverage was 0.258 and 0.256, respectively, for negative and positive samples. All read data were similarly processed and aligned to the human reference genome hg19 (excluding decoy sequences) using BWA-0.7.17 mem [31]. Paired reads were filtered according to the following criteria: 1) reads must be in the correct position/orientation for pairing; 2) only primary alignments are considered; 3) alignments must exceed a minimum mapping quality of at least 1; 4) each read must have a unique starting location (Supplemental Figure S6.1). The alignments were compressed and left unfiltered, as all methods perform internal quality control on the alignments.

6.2.2 Fragment size estimation

When quantifying read coverage, reads are assigned to their respective regions based on their starting position, whereas for read-pairs, we adopt their midpoint, *i.e.*, the average of the starting positions. Fragment sizes were determined by the difference between the starting points of the paired reads. They were distributed with a mean of 173 bp and a standard deviation of 56 bp. Fragment sizes greater than 300 bp were ignored as they were found to be uninformative in distinguishing negative from positive samples. On a per-sample basis, regions of interest were filtered to allow for a more reliable estimation of the fragment size distribution by a lower bound on the minimum number of reads and an upper bound depending on the normalized read coverage (Supplemental Figure S6.2).

6.2.3 Reference set construction

Read coverage is normalized for all samples. The genome is divided into 5 kb regions and scaled to 250 kb, 500 kb, 750 kb, 1 Mb, 5 Mb, and 10 Mb to test for resolution-dependent differences. Region sizes below 250 kb were excluded, as the read count and fragment size signals become too noisy at the specified sequencing yield. We did not predefine a list of genomic regions to be excluded from the analysis. Instead, the within-sample methods define those based on normalized read counts during the construction of the reference set, with negligible differences between methods. These uninformative regions are masked and typically appear at centromeres or highly repetitive locations where an insufficient number of reads may be aligned. Subsequently, technical biases (*e.g.*, GC bias) are removed by training PCAs on the negative controls (Supplemental Figure S6.3). Note that this occurs in parallel. Therefore, two PCA mappings (one for each data type) are saved and applied to any sample with which the reference is queried.

6.2.4 WisecondorFF

WisecondorFF relies on the same within-sample testing method as Wisecondor and WisecondorX [150, 152]: construct a reference set of similarly behaving regions derived from control samples, and then process each region in a new sample against this reference set. The creation of the reference set is thus at the heart of the methodology, which is based on the observation that the (ab)normality of a region on one chromosome can be judged against the behavior of regions with similar behavior (the references) in control samples on other chromosomes. The similarity metric can depend on the type of data. For read count data, we follow Wisecondor, which uses the Euclidean distance between the two read count vectors in the two regions under consideration across the control samples. For each region, all regions in the other chromosomes are ranked according to the similarity metric, and the top K (here $K = 300$) regions are selected as the set of reference regions for the considered region. The regions are further weighted according to their reliability, based on the calculated distances. By doing this for all regions, the complete reference set is obtained.

For fragment size data, the similarity metric between two regions is the Euclidean distance between the two vectors of average fragment sizes in a region for each control sample. We experimented with different summaries of the fragment distributions in a region, such as the median (being less predictive, Supplemental Figure S6.4), or metrics that directly capture the difference between two distributions, such as the Jensen-Shannon divergence distance (symmetric Kullback-Leibler divergence). However, the latter was not feasible due to the amount of noise present in the distributions (full details in Supplemental Section 6.5.7). When searching for CNVs in a sample, a Z-score for each (query) region and data type is calculated separately. This Z-score can be calculated from the observed measure (either the number of reads or the fragment size) in the query region versus the mean and standard deviation of that measure calculated over the reference regions for the query region in that same sample. The data type-specific score (from each region) can be combined into a single score by Fisher's averaging. Then, the scores per region are used to find stretches of affected regions through segmentation. Here, we follow the methodology of WisecondorX [152] and use Circular Binary Segmentation (CBS) [205] to segment and finally obtain Z-scores of the detected events.

6.2.5 Fetal fraction estimation

The fetal fraction of the samples was estimated using SeqFF [156]; this method uses a pre-trained multivariate model that was trained on the number of autosomal reads stratified by region from WGS pairwise sequencing of maternal plasma cDNA. The average fetal fraction is 7.5% and ranged from 1.48% to 19.15% (Supplemental Figure S6.4).

6.3 Results

6.3.1 Fragment size distribution shifts

To determine whether the fragment size may indeed be indicative of samples with trisomy, we first examined the fragment size distribution within our cohort of 526 samples, across chromosome 21, for 125 samples with trisomy of chromosome 21 as well as for 401 samples with no trisomy (negative samples). Figure 6.1a shows that, indeed, on average, a distribution shift of ~1.52 bp towards shorter fragments can be observed for T21 samples. Note that the fragment size distributions for individual samples vary considerably, likely due to differences in fetal fraction and technical noise. While Figure 6.1a shows a shift in the fragment size distribution at the chromosome level, we can observe a similar shift in the distribution when we consider smaller regions across chromosome 21, as shown in Figure 6.1b. However, these differences become less noticeable as the region size decreases, as fewer reads fall within a region, resulting in noisier estimates of these distributions. With an average sequence coverage of 0.25×, we found that a minimum region size of 250 kb was required to estimate fragment size distributions with sufficient robustness (Supplemental Figure S6.6).

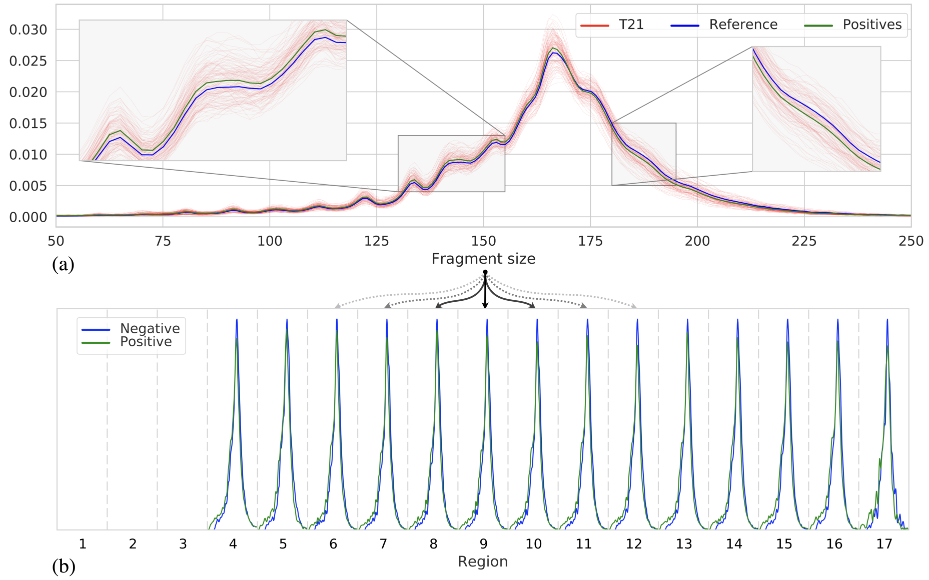


Figure. 6.1. (a) The fragment size distributions of chromosome 21 across samples: individual T21 positive samples (red), the mean of all 125 T21 positive samples (green), and the mean of all 401 negative samples (blue). (b) Discretized representation of chromosome 21 into 3 Mb sized regions, per region fragment size distributions are shown (as in (a)) of two samples with similar fetal fraction, one negative (blue) and one T21 positive (green).

6.3.2 T21 detection performance

Within our cohort of 526 samples, we investigated the detection of common whole chromosome aneuploidies using six different approaches. The first three make use of WisecondorFF (Section 6.2.4), denoted as WcrFF, and detect the presence of a CNV event from read count frequencies (WcrFF^{RC}), fragment size statistics (WcrFF^{FS}), or both (WcrFF$^{RC\&FS}$). Two other approaches are the latest versions of Wisecondor (Wcr) [150] and WisecondorX (WcrX) [152]. Finally, we included a method that does not use an in-sample testing approach to detect CNVs, CNVkit [206], a general-purpose CNV detector, which we use here as a baseline.

Almost all 125 T21s are detected by all methods, where we only consider events when segments are larger than 10 Mb with Z-scores ≥ 5, Table 6.1:I. The performance of all methods is relatively stable over the entire range of selected region sizes, except for WcrFF^{FS}, which improves continuously as the region size increases, and to a lesser extent WcrFF^{RC} and WcrX at 10 Mb, with a sharp drop in performance. The former can be attributed to the increase in noise in the fragment size signal, which we also encountered when we attempted to call events based on the fragment size distributions of each region rather than their averages (Supplemental Figure S6.7). The latter is probably due to the CBS algorithm used by both methods. Only WcrFF$^{RC\&FS}$ (at 750 kb) can detect all expected trisomies and shows near-optimal and stable performance in other region sizes. The baseline, CNVkit, is competitive with other methods, which is expected given the relative ease of detecting T21 events.

Although each method detected the majority of all expected trisomies, this result must be put in context with the additional results, *i.e.*, false positives. In Table 6.1, we also summarize the detection of events on chromosomes other than 21 in parentheses. Generally, sensitivity increases when a smaller region size is chosen, especially for Wcr. $\text{WcrFF}^{\text{RC\&FS}}$ makes far fewer false positive calls than Wcr at any resolution while detecting slightly more than WcrX. When we change the event acceptance constraints to segments larger than 1 Mb with |Z-score| ≥ 5, allowing smaller events to be called (Table 6.1:II). Wcr calls many more false positives (+247.17%) while still not detecting all expected T21 events, whereas $\text{WcrFF}^{\text{RC\&FS}}$ is now able to detect all expected events at nearly every resolution, with only a modest increase in false positives (+8.95%).

Table. 6.1. The number of detected events in the 125 T21 positive samples for the six different tested methods (rows). An event is detected when I: a segment is ≥ 10 Mb with a Z-score ≥ 5 or II: a segment ≥ 1 Mb with |Z-score| ≥ 5. The number of samples for which an event is detected on chromosome 21 is given for different reference region sizes (columns). We note the number of events detected on one of the other chromosomes between parentheses.

		250 kb		**500 kb**		**750 kb**		**1 Mb**		**5 Mb**		**10 Mb**	
WcrFF^{FS}	I	71	(0)	77	(0)	83	(0)	86	(6)	92	(14)	106	(5)
	II	71	(0)	77	(0)	83	(0)	87	(11)	92	(14)	106	(5)
WcrFF^{RC}	I	122	(0)	122	(1)	122	(1)	120	(2)	121	(0)	103	(0)
	II	122	(3)	122	(2)	122	(1)	120	(2)	121	(0)	103	(0)
$\text{WcrFF}^{\text{RC\&FS}}$	I	**124**	(17)	**124**	(14)	**125**	(13)	**124**	(14)	**124**	(6)	**122**	(3)
	II	**125**	(18)	**125**	(15)	**125**	(17)	**125**	(14)	**124**	(6)	**124**	(3)
Wcr	I	123	(162)	**124**	(78)	123	(34)	123	(27)	122	(9)	121	(8)
	II	124	(471)	124	(285)	123	(119)	123	(215)	122	(16)	121	(8)
WcrX	I	122	(1)	122	(0)	120	(1)	121	(1)	120	(1)	103	(0)
	II	122	(4)	122	(0)	120	(1)	121	(1)	120	(1)	103	(0)
CNVkit	I	120 (375)											
	II	124 (896)											

6.3.3 Per-region Z-score differentiation

Next, we examined the Z-scores generated by the different within-sample methods, as higher Z-scores indicate greater power to detect an event. Per region on chromosome 21, we calculated the average Z-scores across all T21 positive samples. From Figure 6.2a, we can observe that the highest Z-scores are found by the $\text{WcrFF}^{\text{RC\&FS}}$ method when considering region sizes of 750 kb. Note that the Z-scores for WcrFF^{RC} drop, behaving very similarly to WcrX, and drop even more dramatically with WcrFF^{FS}, which also performs worse than Wcr. Similar results are obtained across other region scales. The distribution of average Z-scores per region, as shown in Figure 6.2b, also shows the shift toward larger Z-scores for $\text{WcrFF}^{\text{RC\&FS}}$.

Next, we quantified the differences in Z-scores derived from fragment size or read coverage and the combined approach. Figure 6.3 shows the average Z-scores across chromosome 21 for each of the 526 samples in 1 Mb sized regions for the methods. As expected, the negative samples have mean Z-scores closely centered around zero. Overall, we can see that the Z-score magnitude of events detected by WcrFF^{RC} are larger than those detected by WcrFF^{FS}. Hence, the fragment size is less reliable and powerful than the read count.

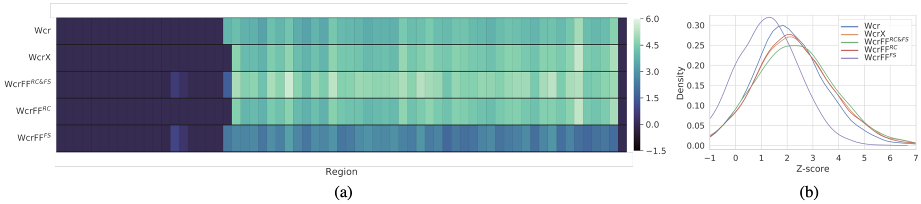


Figure. 6.2. (a) Heatmap of the Z-scores averaged across all T21 positive samples, shown per 750 kb sized region (columns) on chromosome 21 for the different methods (rows). (b) Z-score distributions of all T21 positive samples on chromosome 21, with zero value Z-scores filtered out.

However, combining the two measures, as in WcrFF$^{RC\&FS}$, makes it possible to separate all T21 negative and positive samples without detecting false positives. In addition, the magnitude of the Z-score for T21 samples is generally greater when combining the two inputs than when using read coverage alone.

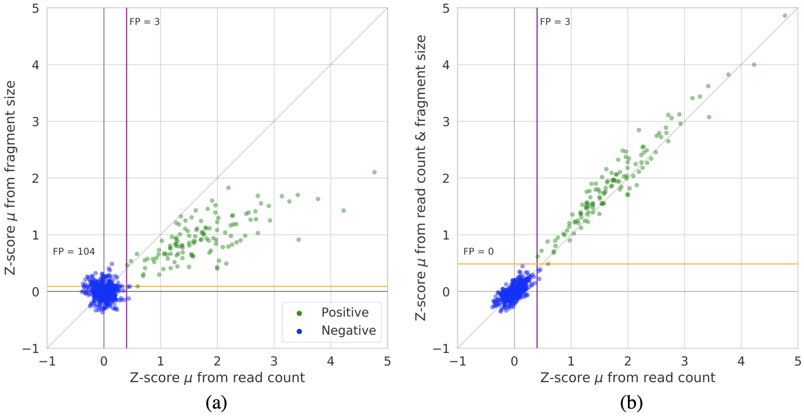


Figure. 6.3. The average Z-score of all 1 Mb sized regions on chromosome 21 for all 526 samples for: (a) WcrFF^{RC} compared to WcrFF^{FS}, and (b) WcrFF^{RC} compared to WcrFF$^{RC\&FS}$. The colored lines denote the Z-score cutoff boundaries that would capture all T21 positive samples for either method (purple for read count and orange for fragment size and fragment size with read count); annotations denote the number of false positives (FP) given these cutoffs.

6.3.4 Detection power relative to fetal fraction

Because a NIPT test can be performed at different stages of pregnancy, we were interested in the performance of methods at different cfDNA fractions available in maternal blood plasma. For this purpose, we estimated the fetal fractions of the 125 samples (Section 6.2.5) and compared those to the Z-scores of detected events on chromosome 21 (Figure 6.4). We show that WcrFF$^{RC\&FS}$ assigns greater Z-scores in almost all cases and ranges of fetal fractions. Of interest is the detection of a duplication event in a sample with a fetal fraction below 2%, which was undetectable by Wcr or WcrX.

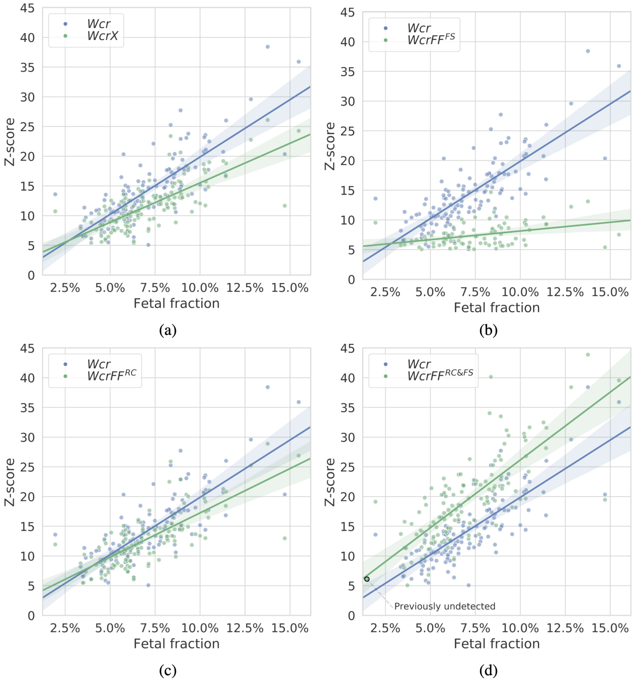


Figure. 6.4. Z-scores of duplication events on chromosome 21 (an event is detected when a segment is larger than 10 Mb with a Z-score ≥ 5) as detected by the different methods in the 125 T21 samples with respect to the estimated fetal fractions of each sample. In each panel one of the methods is compared with WCR (in blue): (a) WcrX, (b) WcrFF$^{\mathrm{FS}}$, (c) WcrFF$^{\mathrm{RC}}$, and (d) WcrFF$^{\mathrm{RC\&FS}}$ (all in green). The region size was 750 kb. Each point corresponds to an event within a sample.

6.4 Discussion

Detection of CNVs in heterogeneously mixed WGS data remains a challenge and is the subject of ongoing research in the field of NIPT and adjacent fields such as cancer diagnosis from cell-free tumor DNA [138]. With NIPT, one has to deal with a mixture of maternal and fetal DNA in which the fetal DNA, and thus the potentially affected part, is present at a much lower concentration. Moreover, different levels of the fetal fraction are encountered when collecting samples at different stages of pregnancy. In addition, sample preparation and sequencing differences can introduce enough noise to mask actual variations within the sample. It is challenging to deal with such noise when using reference-based methods that directly compare a sample's signature to a reference signal. A within-sample normalization method circumvents these problems of experimental noise.

In general, NIPT methods using WGS data exploit the relative frequencies of reads aligned to a reference genome. This is not without reason, as read coverage strongly predicts CNV presence. However, other data types are known to be informative for CNV detection. One source of information that can be easily derived from paired-end reads is the size of the cfDNA fragment. Although it is generally known that this size differs between maternal and fetal cfDNA, in NIPT applications of CNV detection, this information is often not used, even though most sequencing for NIPT is now performed using paired-end sequencing.

We have shown that at both the whole chromosome and subchromosomal levels, fragment size distributions are indeed detectably shifted when it comes to chromosomal CNVs in fetal DNA. To detect these events, we introduced WisecondorFF, a within-sample normalization method that uses both the relative frequencies of aligned reads and fragment size to detect chromosomal CNVs reliably.

We noted that the method we presented, WisecondorFF (WcrFF$^{RC\&FS}$), which uses both read count and fragment size statistics, provides several improvements over the other methods. Namely, WisecondorFF is more sensitive than Wisecondor and WisecondorX while being more selective and detecting fewer false positives. WisecondorFF is more robust than the others in assigning greater certainty to any event detected at any fetal fraction. This results in an advantage for WisecondorFF, where calls previously indistinguishable from the background signal are now more likely to be detected. By relating the Z-scores of detected events to (estimated) fetal fractions, we found that WisecondorFF can detect events at lower fetal fractions. Although WisecondorFF performed best with a region size of 750 kb, we showed that fragment size performs better at larger region scales. Fragment size and read coverage are currently integrated at the same scales. However, it would be possible to use asymmetric region sizes for different data types to achieve better performance. At this time, WisecondorFF does not build sex-specific reference sets as WisecondorX does and, therefore, cannot detect CNVs on sex chromosomes. We have limited the use of additional information exclusively to the fragment size. However, we believe that any data that can be discretized across genomic regions could be integrated within a parallel within-sample normalization framework, as we have shown. By fully exploiting the information available in current sequencing technology, we have shown that better performance in a low-coverage NIPT setting can be achieved by exploiting fragment size information. This opens the door to other applications related to fragment size differences, such as cancer diagnosis or other types of data integration within the NIPT process.

6.5 Supplementary materials

6.5.1 Processing and filtering paired-end reads

During pre-processing, WisecondorFF discretizes the genome and accumulates each sample's read coverage and fragment size statistics. During this accumulation process, paired-end reads are subjected to filtering based on the following criteria: 1) reads must be in the correct position/orientation for pairing; 2) only primary alignments are considered; 3) alignments must exceed a minimum mapping quality of at least 1; 4) every read should have a unique starting location. With 1), we consider that only proper pairing should be considered for further analysis, *i.e.*, when pairing fails, we exclude such reads. With 2) and 3), we consider and filter out reads that correspond to repetitive regions of the genome. As a result, there are multiple valid mappings of a read, which then receive a mapping quality score of 0. Finally, with 4), we exclude all reads that were found to map to previously aligned locations. In the context of low yield WGS (~0.25x coverage), while expecting an approximately uniform genome mapping distribution, it is unexpected for many reads to map to the same location. In Figure S6.1a, we show the coverage distribution of chromosome 1 within a single sample in which 4) filtering was not performed. Just at the centromere boundary, a large pileup is concentrated caused by (technical) duplication, which would adversely affect reference set construction. It is trivial to account for this duplication effect by using the 4) filtering, as can be seen in Figure S6.1b.

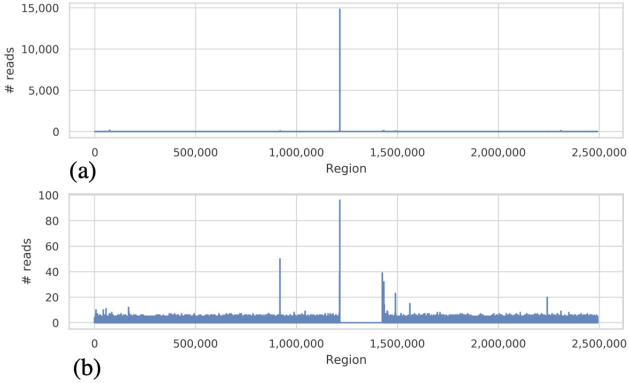


Figure. S6.1. 100 bp region stratified read coverage (~0.25x) of chromosome 1 in a single sample. (a) Prior to filtering out reads beginning at the same genomic location. (b) Post-filtering removes such duplicated reads.

6.5.2 Filtering regions by depth of coverage

Every sample processed by WisecondorFF is subjected to initial filtering. This filtering excludes regions (defined after discretization) that contain insufficient information, *i.e.*, too few reads are aligned in these genomic regions. In such cases, these regions cannot be used confidently as reference regions, meaning CNV calling is also infeasible at these locations. Although the fragment size is used, we base region filtering on the read coverage (fragment size is also derived from aligned reads). Our experiments have shown that any region must contain at least 500 reads to be sufficiently informative. This threshold depends directly on the average depth of coverage and the size of the chosen region, where a smaller chosen region size reduces the probability that a region will contain enough reads. In Figure S6.2a and Figure S6.2b we consider the genome-wide per-region coverage of two samples with different depths of coverage, with a different threshold for each distribution. Note that most regions contain enough reads, but a small proportion where (almost) no reads are aligned; these are the regions that need to be filtered. Constant size filtering is inadequate since each sample has a different depth of coverage, Figure S6.2c. A threshold may be too aggressive for a particular sample, *e.g.*, a threshold of 1,500 is appropriate for SAMPLE2 but too aggressive for SAMPLE1. To account for this, we use secondary filtering based on the normalized coverage, Figure S6.2d providing a uniform cutoff for all samples, regardless of coverage.

6.5.3 Fragment size, GC content, and read counts

Sample preprocessing includes correcting bias caused by effects such as GC content. The effects on the number of reads aligned due to GC content are well known (Figure S6.3) and can be corrected with LOWESS or PCA, we used the latter. Although the fragment/insert

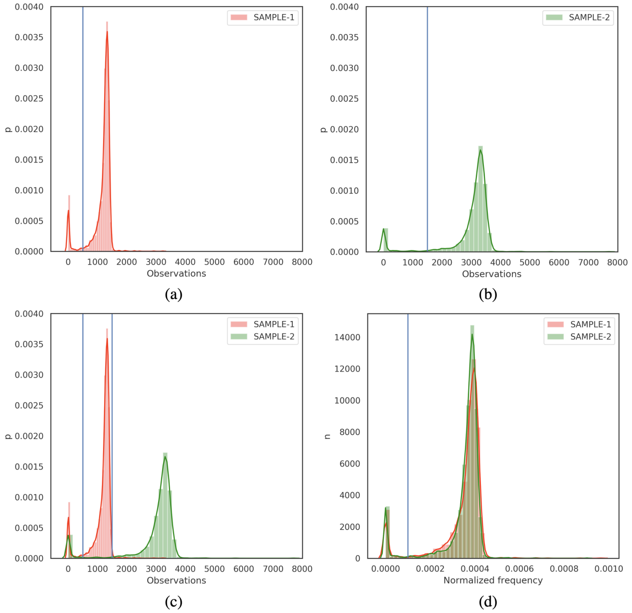


Figure. S6.2. Per-region coverage distribution plots: (a) Sample with an optimal cutoff at 500 reads. (b) Sample with an optimal cutoff at 1500 reads. (c) Both samples overlaid showing both cutoffs. (d) Both samples are overlaid with normalized frequencies showing a cutoff of 0.0001.

size is independent of GC content, we found that the PCA correction still accounted for other types of bias in the fragment size data.

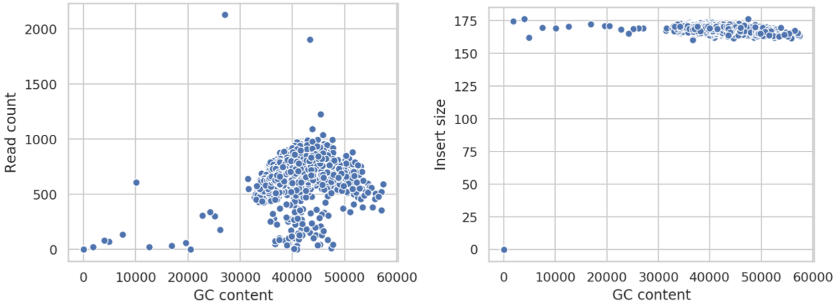


Figure. S6.3. Genome-wide GC content with respect to the read count and the fragment/insert size respectively.

6.5.4 Fragment size median is less predictive

Fragment size statistics by region can be used in several ways, *e.g.*, by comparing them directly by treating them as probability distributions as in Supplemental Section 6.5.7 or by further summarizing them using measures such as the mean or median. In practice, we have found that the mean of fragment size per region is the most predictive compared to the other summaries, with the median being the less effective secondary choice. Singular value metrics such as mean or median can be used interchangeably, assuming that a distance metric can be calculated between any two values. It was, therefore, trivial to experiment with different summarization metrics and evaluate their performance. Figure S6.4 shows the average Z-scores on chromosome 21 for each of the 526 samples for regions of 1 Mb size comparing WcrFF^{RC}, the mode of WisecondorFF that uses only the number of reads, and WcrFF$^{RC\&FSmedian}$, which combines the number of reads and the median of the fragment size. Compared to our results in Figure 6.4, the median fragment size, while predictive, is not competitive with the mean.

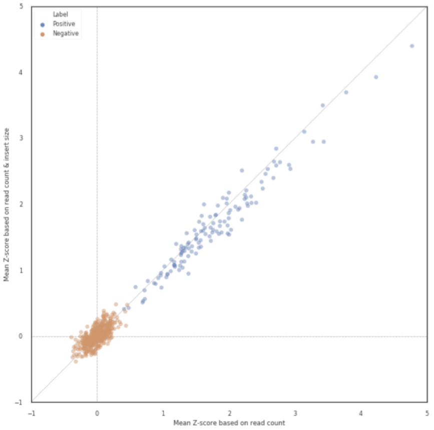


Figure. S6.4. The average Z-score of all regions (1 Mb size) on chromosome 21 for all 526 (positive and negative) samples comparing WcrFF^{RC} and WcrFF$^{RC\&FSmedian}$.

6.5.5 Fetal fraction estimation with SeqFF

The fetal fraction plays an essential role in NIPT screening. If the maternal serum contains too little fetal DNA, no reliable DNA-based test is possible. It is expected that the reliability of the test will improve with increasing fetal DNA concentration as the fetal signal becomes stronger. It is useful to relate the fetal fraction to the method's performance, *e.g.*, based on sensitivity or robustness, because any method must perform well for a given fetal fraction. It is, therefore, an approach for distinguishing and ranking the performance of methods. To determine the fetal fraction, we used SeqFF, which uses the number of aligned reads in specific autosomal regions by applying a weighting scheme derived from a pre-trained multivariate model. This model was trained on regionally stratified autosomal read counts from pairwise WGS sequencing of cfDNA in maternal plasma. In Figure S6.5 we summarize the stratified fetal fraction distributions of the 526 samples.

6.5.6 Fragment size distribution estimation

The fragment size statistics per sample can be discretized and processed at different scales. The choice of scale is a direct trade-off between resolution and noise, depending on the sample depth of coverage. With smaller region sizes, there may be too much noise or

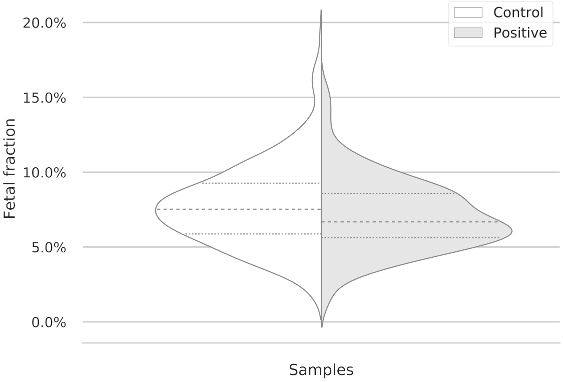


Figure. S6.5. Fetal fraction values for all 526 samples as estimated by SeqFF. Samples are grouped accordingly to 401 healthy controls and 125 T21 positives.

too few reads to reconstruct the fragment size distribution accurately. For within-sample testing, it is crucial that (almost) all regions have sufficient measurements to be reliably used as reference regions. Therefore, it is vital to determine the lower limit at which the fragment size distribution estimate becomes robust, meaning that there is no instability in any genomic region within a sample. In Figure S6.6, we summarize the fragment size distribution by region of chromosome 1 in terms of mean and standard deviation. Region sizes smaller than 250 kb show increased variability or missing values in the region fragment size distribution. This means that a minimum region size of 250 kb is required to reliably estimate fragment size distributions, which we further verified by testing the normality of the distribution and the distance from the expected mean and standard deviation of each region against a reference distribution and different samples.

6.5.7 Reference set from distributions

The methodology behind constructing a reference set for within-sample testing can be readily generalized for different data types. However, this assumes that it is possible to discretize these data in a manner analogous to read coverage. We noted that fragment size statistics could distinguish between positive and negative samples, even at a sub-chromosomal scale. The richest representation of the fragment size statistics is to consider it as a probability distribution of fragment sizes within a region, denoting a probability of observing a specific fragment size. Thus, when discretizing the genome into regions of a given size, each region maps to a corresponding probability vector. As described in Section 6.2.2, we have capped these distributions at 300 bp; only the range [0, 300] is considered.

Incorporating this data representation requires several changes within the within-sample testing methodology. When constructing the reference set, it is no longer possible to calculate the Euclidean distance between regions. Instead, an appropriate distance measure is needed to compare probability distributions; for this purpose, we chose the Jensen-Shannon divergence distance (JSD). The JSD is a symmetric generalization of the Kullback-

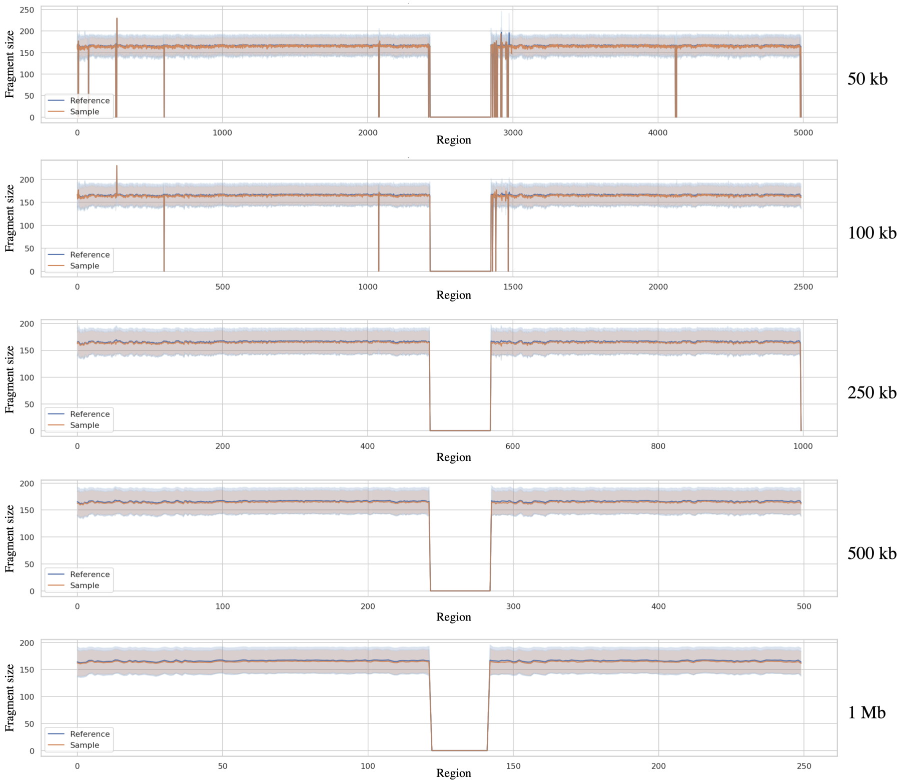


Figure. S6.6. Per region fragment size distribution of chromosome 1 as summarized by the mean and standard deviation across different region scales. The orange-colored line denotes a single sample, whereas the blue line is an aggregate of a collection of samples used here to compare as a baseline.

Leibler divergence (KLD) and can be used as a distance measure.

The KLD was not appropriate for this purpose, as all regions are compared in both directions (each region is compared to all other regions on the other chromosomes); as it is not symmetric, it would fail to generate a correct reference set. One caveat to consider when using JSD is that no null probability is present in the compared distributions. We have addressed this problem by uniformly padding the distributions so that the previous null values become tiny probabilities. From a computational perspective, this probability vector-based approach is much more expensive than using a single summary statistic per region.

Once the reference set is constructed, a sample can be processed like that described for the summary values. For example, to compute the Z-score of a query sample region, we first combine the reference region distributions into a single combined distribution and then compute the query region JSD for that combined distribution. Then, we obtain the mean and standard deviation by computing the JSD of each reference region distribution concerning the combined distribution. The region's Z-score can then be derived from these three values (JSD distance of the query region from the combined distribution, mean and standard deviation).

We found that this approach, WcrFF$^{\text{FS-JSD}}$, which directly compares fragment size distributions, was ineffective in predicting CNVs. In Figure S6.7 we show the average Z-scores across chromosome 21 for each of the 526 samples for regions of 1 Mb size, comparing WcrFF$^{\text{RC}}$ to WcrFF$^{\text{FS-JSD}}$. The divergence-based method is not able to distinguish positive from negative samples. We believe that the large amount of noise present in the distributions may explain this poor performance and that JSD is too sensitive to cope with it. We tried to reduce the influence of noise by capping the distributions to smaller ranges (*e.g.*, [125, 200]) or by smoothing the distributions, but this did not improve performance. Considering only the performance of the fragment size average, we observed that it improved as the region size increased (peaking at 10 Mb). The divergence-based approach could perform better with higher coverage (for more accurate distribution reconstruction) and larger region sizes. In addition, it might be interesting to investigate other distribution distance metrics, such as the Wasserstein distance (Earth's displacement distance), although initial experiments have not been promising.

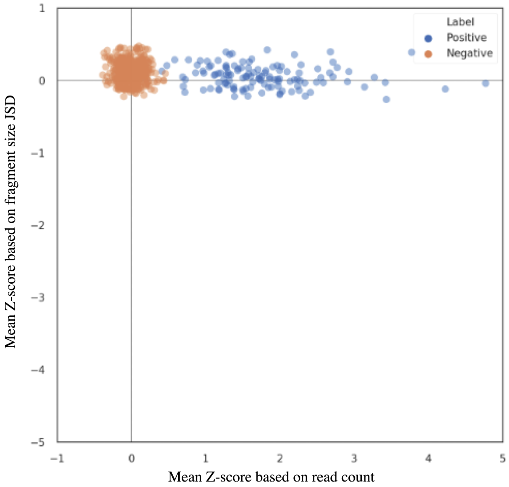


Figure. S6.7. The average Z-score of all regions (1 Mb region size) on chromosome 21 for all 526 samples for WcrFF$^{\text{RC}}$ compared to WcrFF$^{\text{FS-JSD}}$

7

Conclusion

Since the discovery of cell-free fetal DNA in 1997, non-invasive prenatal testing (NIPT) has catapulted from a laboratory curiosity to the global standard for fetal aneuploidy screening. It is now routinely offered in private and public sectors across America, Europe, and Asia, with complete coverage of publicly funded NIPT programs rapidly expanding geographically. As NIPT advances, health policy regulators and clinicians must consider how such testing might be delivered to populations ethically and affordably without compromising clinical effectiveness. The promise of ever-increasing detection resolution will inevitably also lead to the resolving of increasingly smaller copy number variations that may not always be known to be pathogenic—or even interpreted in a clinical context—yet may have substantial implications for the well-being of future families. This raises ethical questions about what clinicians should disclose to prospective parents since any finding may cause uncertainty and anxiety even if they are considered "benign" variants under current knowledge. Coupled with increasing misinformation presented on the internet, easily accessible by patients searching for self-diagnosis, this could unnecessarily influence parents' future pregnancy planning and result in abortions of fetuses with variants of uncertain significance that could be otherwise healthy. While addressing this ethical dilemma goes beyond this thesis's scope, I have some thoughts about this. One way to address this problem is to ensure that clinicians provide accurate and up-to-date information to prospective parents about the implications of NIPT results. Clinicians should be transparent about the limitations of current knowledge and make it clear that any interpretation of findings should be in the context of the individual patient's clinical situation. They should emphasize that any findings of uncertain significance may not indicate a health problem and that further testing may be needed to confirm a diagnosis. Parents should also be encouraged to seek out reliable sources; ideally, up-to-date information about prenatal testing is easily accessible to everyone, such as from their local health department, to help them make informed decisions about their pregnancies. However, it may also be argued that this would only add to the confusion and anxiety patients may already feel. Health policy regulators could provide more guidance on how NIPT should be delivered to the public by establishing standards for how NIPT should be offered and advertised; providing information on the risks of self-diagnosis and the importance of getting accurate information from a medical professional; creating educational materials for patients and clinicians on the use of NIPT.

The scope of NIPT methodologies is extensive and tailored to specific applications such as single gene disorders or aneuploidy detection. In addition, a range of options is available to address the same challenge in NIPT. For instance, aneuploidy detection using NGS may range from high-yield sequencing using heterozygous (tandem) single nucleotide polymorphisms to low-yield sequencing using coverage differentials across genomic regions. Alternatively, other technologies such as microarrays, methylation profiles, or fluorescence-based techniques may be used. The choice of a particular method depends on the question asked and the application required. However, an advantage of NGS is its flexibility to vary capturing kit or sequencing yield depending on the application requirements. For instance, establishing the copy number of a sizable genomic region (larger than any repeat) can be achieved with short reads and low coverage since 1) the actual sequence content is not relevant beyond establishing a genome alignment, and 2) the copy-number signal depends on global coverage distributions of reads rather than any local differences. If smaller genomic regions have to be surveyed genome-wide, coverage must increase to obtain sufficient signal to detect differences in the coverage distributions of smaller regions. If repetitious segments are analyzed, read length must increase, or in the case of paired-end reads, the insert size must increase such that these repeats may be spanned.

Here, I showed how the coverage-based within-sample testing method remains relevant as part of a NIPT application. It has the advantage of side-stepping issues inherent to methods that directly compare to reference panels. Furthermore, I demonstrate how the within-sample method can go beyond the coverage signal into a framework that integrates additional genomic signals, such as the fragment size estimated from paired-end reads. The proposed approach highlights the power of combining multiple signals and how such an approach can overcome some of the challenges in NIPT concerning fetal CNV detection. This highlights the importance of exploiting characteristics specific to a domain, which is, in this case, the difference in fragment size of maternal and fetal DNA fragments. Although the fragment size is not as predictive as the genome coverage, these signals, obtained from the same data, when combined, lead to a more robust prediction at lower fetal fractions, with no additional cost incurred. Currently, commonly used pairwise sequencing technologies can easily establish genome-wide fragment size distributions, a powerful signal to use alongside coverage. This could also be used in other applications, such as cancer diagnosis, using the NIPT methods discussed. However, this is much more difficult given the heterogeneous nature of cancer, where many more chromosomes may be affected than would be expected from NIPT. One of the assumptions used with within-sample testing is that only a subset of chromosomes is affected so chromosomal regions can be compared with each other. In cancer, the chromosomes may all have different (segments of) ploidy, so this scheme becomes difficult and may require adaptations to the reference set to account for these circumstances.

Although machine learning models such as deep learning can outperform traditional statistical methods in tasks such as image classification, object recognition, and segmentation, they are not as well suited for clinical applications such as NIPT. In this context, decision-making is not as straightforward as in computer vision tasks, and the risk of misclassifying results may be unacceptably high. The limitations of deep learning models in the clinic are due in part to their complex and opaque nature, which limits their ability to provide a transparent explanation of how the model made a prediction. NIPT methods require a high degree of interpretability, meaning that predictions must be explainable to clinicians and patients to make informed treatment decisions. Even if their performance improves, the models will be of limited use in the clinic if they cannot be interpreted. Therefore, additional work is needed to understand the complex decision-making process of deep learning models before clinical adoption can begin. Note that an interim solution could use them with statistical methods side by side for clinical decision-making, as they are not necessarily mutually exclusive.

With the continued development of sequencing technologies, the associated costs are simultaneously decreasing, and high-throughput sequencing is becoming more accessible. In the context of a within-sample testing method, which relies on the discretization of the genome into larger regions, an increase in available sequencing coverage translates into an easier transition to discretization into smaller regions so that more refined predictions can be made. Ultimately, this comes down to a trade-off between cost, availability, and required resolution, which depends on the weighting of the prevalence and severity of genomic disorders in our society. With the increasing prevalence of NIPT, the number of samples available to construct reference sets will increase. Because within-sample testing still relies on a reference set to map regions of similar behavior across the genome, an advantage of a larger reference is to reduce the effects of sample-specific biases. It may also open up opportunities to extend the test frame to be more specific, for example, to the fetal fraction of a query sample. NIPT using low-throughput sequencing is sensitive to different types of bias. While some types of bias are unavoidable, such as those from different sequencing runs or instruments used, innovative technological advances are addressing others, such as systemic sequencing errors, GC content, and PCR amplification bias. Although current methods attempt to correct these biases, it should be noted that they are imperfect and introduce their respective noise. Ultimately, it is in our best interest to reduce the amount of computational processing and rely on the raw sequencing data, as this also improves the interpretability of the NIPT.

NIPT techniques are effective, but there are still several rare conditions and factors to consider that can confound predictions. Examples of this include a vanishing or unreported twin (because the extra DNA from the vanished twin gives the impression of fetal aneuploidy) or maternal chromosomal aberrations (not found in the fetus), which can lead to false positive predictions. Somatic mutations are another problem because they can also be localized in subsets of cells, *i.e.*, mosaics. Since NIPT is based on the sequencing of circulating cfDNA, these subsets will only be represented by a small number of reads. Thus, if the subsets are small enough, variation may remain undetected. Dealing with these rare conditions is challenging in the context of the described NIPT method and may require condition-specific supervised models trained on expert knowledge and clinical data. Future studies will have to address these issues to improve the performance of NIPT in clinical practice, which could be imagined as an array of models used side-by-side.

III

The Human Gut Microbiome

8

Introduction

DNA, RNA, and protein measurements in tissue can elucidate a lot about someone's health and well-being. However, other important aspects may be overlooked if we also neglect to consider a person's interactions with their micro-environment. For hundreds of millions of years, animals have had a close relationship with microbes, during which time they have co-evolved [207]. Extensive research on the phylogenies of animal hosts and their microbiota suggests that there is specific selection based on co-adaptation [208]. Cooperative interactions between microbes and their hosts can benefit both partners, with the microbes participating in host functions such as defense and metabolism [209]. For example, when germ-free mice are compared to normal mice, it is evident that most of the metabolites detected in plasma originate from microbiota [210], which highlights the importance not only of a symbiotic relationship between animals and their microbial communities but also how vital these relationships are for our health. Humans share their livelihoods with many microbes, including bacteria, viruses, and archaea that potentially outnumber the number of human cells and coding genes [211]. Most of these microbes reside in the gut, with up to 2,000 bacterial species present at any given time [212].

The gut microbiota has a profound and far-reaching impact on our health; everything from lifespan [213], pathogen protectivity [214], and immune system regulation [215], to the development of inflammatory bowel disease [216] and neurodegenerative disorders, such as Alzheimer's disease [217, 218], may be influenced by it. Given the sheer diversity of conditions that may be impacted by the gut microbiota, measuring changes in its composition under varying conditions is crucial to better understanding diseases and developing more targeted treatments. However, a challenge here is a high degree of variation between populations and individuals in microbiota composition. This complicates the association of specific taxa with particular conditions and is further confounded by the fact that there is not a single universal configuration of a healthy microbiota [219, 220]. A great diversity of microorganisms are required to metabolize nutrients in adult diets, and low microbiota diversity has been associated with conditions such as autism [221], autoimmune disease [222], and obesity [223]. Hence, maintaining high bacterial richness and diversity is important because it provides functional redundancy, adaptability, and overall stability against environmental challenges [213].

Everyone experiences aging differently [224], with some individuals experiencing delayed aging or ‘healthy’ aging, where they remain vital and healthy despite advancing years [225]. Numerous factors may contribute to healthy aging, including genetics [226], environment [227], lifestyle choices such as diet [228], exercise habits [229], or, social interaction [230], and changes to the gut microbiota [213], all in turn affecting the rate of age-related decline. The microbiota is constantly changing during different stages of life [231] but is especially unstable during infancy [232, 233]. When infants switch to a more varied diet, their microbiota composition changes dramatically and becomes increasingly complex and stable, gradually resembling an adult’s microbiota composition [233]. Once established, the microbiota remains largely stable outside of intervention [234, 235]. The gut microbiota changes substantially during the transition from a relatively stable health state to deterioration of mental and physical health before death, even before any visible symptoms appear [236].

The composition of the gut microbiota affects an individual’s metabolism by altering concentrations of critical metabolites, such as short-chain fatty acids, vitamins, and lipids [237]. These compositional changes—which may occur through the addition, loss, or changes in the abundance of microbial species—may affect a person’s aging trajectory [213]. Previous studies have demonstrated that patients with inflammatory bowel diseases have a lower abundance of species that possess anti-inflammatory qualities [238]. Additionally, the uptake of microbial metabolites into the circulation has also been shown to affect behavior [239] and the development of neurodegenerative disorders [240].

Given the gut microbiota’s ability to influence various conditions, it would be advantageous to create customized probiotics as a form of intervention [241]. Utilizing gut microbiota-related signals associated with unhealthy aging may be feasible to achieve personalized microbial restoration through dietary intervention or fecal transplantation. Such that previously lost beneficial taxa may be replenished. In the context of the microbiota, healthy aging thus involves the maintenance of diversity and abundance of protective taxa as its deterioration can lead to age-related and disease-related problems. While these two types of issues are overlapping and interactive, it is essential to note that they are distinct from one another [213]. Identifying the taxa associated with (un)healthy aging can be achieved by longitudinal studies tracking individuals over an extensive period and relating the gut microbiota composition measured at intermediate time points to the final physiological or clinical status. However, this is challenging. A widely used alternative is to adopt a cross-sectional study design that stratifies older populations based on unhealthy and healthy aging indices and then identifies the corresponding taxonomic markers.

High-throughput sequencing is routinely used for microbiota analysis; popular methods include: shotgun sequencing and amplicon sequencing of taxonomic markers [242]. Shotgun sequencing provides broad coverage of genomic DNA, which may include bacteria, viruses, archaea, and fungi [243]. While it has a high resolution (allowing for species-level and even strains-level classification) [244], it requires more complex data analysis and higher sequencing yield for reliable classifications [243]. Amplicon sequencing is a type of taxonomic survey that only targets specific taxonomic markers pervasive in the domain under study. This limits both the resolution (to genus-level or species-level identification) [243], since only a tiny portion of the genome is surveyed, and the scope (for example, targeting the 16S ribosomal RNA (rRNA) gene covers only bacteria) of taxonomic surveys.

It may also be prone to amplification bias, resulting in an uneven representation of taxa [245]. While (deep) shotgun sequencing is considered the gold standard for microbiota studies, amplicon sequencing remains popular [246]. This is because it is substantially more affordable, and downstream amplicon data analysis is comparatively much more straightforward.

Bacterial diversity can be assessed by amplicon sequencing of the 16S rRNA gene. This gene is approximately 1500 bp long and contains nine hyper-variable regions interspersed throughout the otherwise conserved sequence [247]. Despite the increasing availability of long-read sequencers, most studies only partially sequence the 16S gene. This is because the Illumina sequencing platforms have a lower cost and higher throughput. However, this does limit the read length to shorter sequences (<300 bp). When targeting a specific region of the 16S, certain taxa are more likely to be identified than others, which introduces bias towards those particular taxa [247]. While long-read amplicon sequencing would enable more accurate identification of taxa by reducing sensitivity to copy number variation [248] and eliminating region-specific bias [247], other genomic regions may also vary between bacteria, so not all variation may be assessed [249].

The primary methods to estimate microbial diversity from 16S amplicon sequencing data are either by constructing operational taxonomic units (OTUs) by clustering sequences together using an arbitrary distance-based identity threshold [250], or by constructing exact biological sequences (ESVs) [251, 252]. In either case, a representation is obtained that relates an absolute abundance to each OTU/ESV, *i.e.*, the number of times a given OTU/ASV, or taxon, appears in a sample. While ESVs-based methods are thought to be superior to OTU-based methods because they are more specific and robust to sequencing errors, resulting in fewer spurious sequences than OTUs [253]. OTU-based approaches may be less likely to overestimate species richness [247] and deal better with (polymorphic) 16S copy number variants [254], especially with long reads [247], which can skew estimates of bacterial diversity if they are not accounted for. However, ESVs simplify comparative analysis between studies as they have an intrinsic biological meaning that does not depend on the reference database or study context – unlike OTUs [253].

A typical 16S ESV pipeline involves the following steps. Sequencing reads are first profiled for quality metrics, and accordingly, trimmed [255], de-duplicated [256], or error corrected [257]. Such (corrected) reads are then processed and merged into ESVs [251, 252, 258–261]; potentially performing chimera [262] or length-based filtering [263]. Taxonomic labels may be assigned to the ESVs in a classification process in which exact sequences are matched with labeled sequences obtained from taxonomic databases [264]. For nearly all ESVs, a classification may be obtained down to the genus level [247].

Before conducting diversity- and differential abundance-analysis, it is essential to normalize data [265, 266]. Different experimental conditions can introduce biases that complicate the comparison of absolute abundances across samples. Normalization helps to remove these biases by accounting for sampling variability. Without normalization, systematic bias may increase the false discovery rate and cause a loss in statistical power [267]. Sampling variability can arise from differences in library size (the total absolute abundance that was sequenced) between samples [268], which may be caused by for instance: differences in sequencing kit [269], samples preparation and handling [270], PCR amplification bias [271, 272], or sequencing center performance [273].

Several measures are used to quantify differences between samples or groups. Often such measures are applied top-down, each having a different level of granularity. For example, alpha diversity measures may measure diversity within a sample. Examples of such measures include the richness [274], evenness [275], Shannon diversity index [276], or Faith's phylogenetic diversity [277]. Richness is a measure of the number of taxa present in a sample, while evenness is a measure of how relatively abundant different taxa make up a sample's richness. The relative abundance, or proportion, is the absolute abundance of one taxon divided by the total absolute abundance for all taxa, which means that evenness measures the (in)balance between all taxa in terms of abundance. Measures that quantify both richness and evenness include the Shannon diversity index and Faith's phylogenetic diversity, with the latter being able to also account for the relatedness among species via their phylogeny.

Beta diversity measures may be used to compare samples in terms of distance or dissimilarity. This requires the construction of a distance/dissimilarity matrix by comparing pairs of samples. A non-phylogeny-based metric such as the Bray-Curtis dissimilarity measures the difference of absolute abundances between samples [278]. Phylogeny-based metrics such as the (un)weighted Unifrac distance use the distances between sequences and the proportion of shared and unique branches in samples within the reconstructed phylogenetic tree to quantify dissimilarity [279, 280]. The metric can be extended to include branch length weighting, using relative abundances [281]. Beta diversity is often ordinated using dimensionality reduction methods to inspect group differences. A frequently used method for testing whether microbiota composition is different given a meta-parameter (such as a grouping), based on beta diversity measures, is a permutational multivariate analysis of variance (PERMANOVA), which tests if the centroids of sample groupings differ [282].

Differential abundance tests are key in understanding which specific taxa differ between two or more groups and determining the biological processes and functions associated with such taxa. A differentially abundant taxon is one whose mean absolute abundance in the ecosystem is significantly different concerning a covariate of interest. As mentioned, normalizing the absolute abundances is crucial before performing any differential abundance test. A popular but controversial normalization method that receives a disproportionate amount of attention is rarefaction [283], which randomly subsamples samples to equalize library size across all samples using a set threshold—the chosen threshold results in a trade-off between sample and diversity inclusion. It is excluded if the sample has a count lower than the threshold. If the sample has a count higher than the threshold, it is down-sampled, resulting in lost diversity. Although rarefaction has been critiqued because it can result in a loss of statistical power [284]. It remains a useful normalization method when used before explorative analysis and differential abundance testing [285–287]. This is because its loss in statistical power is outweighed by its high control of false positives [285]. Given equalized data using normalization methods such as rarefaction, proportion transforms [288], or scaling with ranked subsampling [289]. Non-parametric tests, such as the Mann-Whitney or Kruskal-Wallis test, may be used to determine differential abundances of taxa. However, a problem with this method is that it does not consider the compositional structure of microbiota data [290] and assumes that all samples have been equally sampled; potentially leading to an inflated number of false positives [284]. Nevertheless, they are still widely used. A common problem with some (parametric) methods is the use of normalization/-

transformation, which may over-correct variance across and between samples, ultimately increasing the false discovery rate [291–293].

Many methods available have been successful in other domains, and these may also be effective for microbiota analysis. For instance, methods that determine differential expression of genes in RNA-Seq data, such as edgeR [291] or DESeq2 [293] can readily be translated to the microbiota domain to determine the differential abundance of taxa. Since both domains deal with similarly sparse data [294]. However, while such methods work well with gene expression data, they underperform with microbiota data, having inappropriately high false discovery rates [295, 296]. Microbiota data presents unique challenges that make RNAseq methods for differential expression analysis unsuitable [297]. This is because these methods assume that only a tiny fraction of genes are differentially expressed, which is often not the case for microbiota data [297], and a large proportion of taxa may be differentially abundant between two conditions. Additionally, the sparseness of microbiota data can exceed that of RNAseq data, such that specific methods cannot deal with the much higher sparsity without additional filtering steps. Given these drawbacks, there is a need for methods specifically tailored for microbiota analysis that also respect the compositional nature of this type of data, such as metagenomeSeq [298], ALDEx2 [299], and ANCOM2 [295, 300]. Of these methods, ANCOM2 was found to be most consistent across studies in controlling the false discovery rate and having the most substantial concordance with other differential abundance methods [296, 301]. ANCOM2 is an Aitchinson's log-ratio-based method that classifies and accounts for different sources of zeros in microbiota data [302].

Part III of this thesis places the human gut microbiota in the context of a cohort of extremely aged Dutch individuals, *i.e.*, centenarians, who remain cognitively healthy despite their advanced age and related them to a cohort of younger controls and individuals with Alzheimer's disease. The difference between the cohorts provides an opportunity to observe underlying differences in the composition of the gut microbiota, and we wanted to determine if and how it contributed to cognitive health, longevity, and healthy aging. Using 16S rRNA amplicon-based techniques, as discussed above, we determined the microbiota composition and analyzed which bacteria are present in these cohorts and what differences exist between these groups through differential abundance tests. In the process, we noted shared taxonomic patterns between the cohorts and differences that distinguish them, consistent with known patterns in other studies of centenarians and healthy aging, while also encountering microbiota characteristics that may be unique to the microbiota of Dutch centenarians.

9

A Cross-Sectional Study of Compositional Profiles of Gut Microbiota in Dutch Centenarians and patients with Alzheimer's disease

This chapter is based on 🗎 *T. Mokveld, B. J. H. Verhaar, A. Salazar, L. M. C. Lorenz, H. M. A. Hendriksen, F. de Leeuw, Y. Pijnenburg, K. van Vliet, M. Graat, R. Kraaij, C. E. Teunissen, Z. Al-Ars, W. M. van der Flier, H. Holstege, and M. Reinders. A Cross-Sectional Study of Compositional Profiles of Gut Microbiota in Dutch Centenarians and patients with Alzheimer's disease, Under review.*

Abstract

The composition of human gut microbiota changes during aging and can be altered in disease states, such as Alzheimer's. To explore the role that gut microbiota may play in healthy aging and longevity, we used metagenomic sequencing to determine the compositional differences in gut microbiota associated with subpopulations in the Netherlands including cognitively healthy centenarians (Age: 100.77 ± 1.0, *MMSE:* 25.38 ± 3.45*) and a memory clinic population of patients with AD dementia (Age:* 66.0 ± 8.0, *MMSE:* 20.27 ± 5.84*) and control subjects with subjective cognitive decline (SCD, Age:* 62.04 ± 7.5, *MMSE:* 28.72 ± 2.35*). The fecal samples from 199 subjects were analyzed by sequencing amplicons derived from the V3-V4 region of the 16S rRNA gene. Using a parallel taxonomic analysis, we found that while the gut microbiota of AD dementia patients and SCD subjects shared similar taxonomic profiles, a different pattern was found in centenarians. Centenarians had a higher diversity of core microbiota species than those in the AD and SCD groups. We found that compared to those of AD dementia patients or SCD subjects, centenarians displayed rearranged taxonomic patterns featuring significantly different relative abundances of phyla Firmicutes and Proteobacteria, as well as enrichment for certain bacterial species Ruthenibacterium lactatiformans, Bacteroides fragilis, and, Christensenellaceae. In addition, some microbiota species that are typically abundant in the guts of AD dementia patients or SCD subjects were depleted among centenarians, including Faecalibacterium prausnitzii, Agathobacter rectalis, Subdoligranulum variabile, and Roseburia.*

9.1 Introduction

The gut microbiota serves as an essential contributor to human health and disease [303], playing a crucial role in modulating host physiology [304]. The mutual reliance between hosts and their microbiota has been widely acknowledged [305]. Although the genome remains relatively stable throughout an individual's life [306], the gut microbiota exhibits a dynamic nature, adapting in response to various factors such as diet [228], stress [307], social interactions [308], and aging [213]. Such microbiota adaptations are affected by an array of individual-, population-, and environmental-specific determinants encountered in distinct geographical locations [309, 310]. The microbiota's impact is evident in myriad of conditions, such as neurodegenerative disorders [217, 218], irritable bowel syndrome (IBS) [216], or healthy aging [213]. Consequently, the prospect for therapeutic intervention targeting the intestinal microbiota is particularly promising [311], as this community of microorganisms is highly diverse [312], adaptable [220], and receptive to external stimuli [313, 314].

Improved living conditions and healthcare have resulted in an increasing number of individuals reaching old age, which has led to a rise in age-related diseases and frailty [315]. Consequently, there is a pressing need to investigate the mechanisms of longevity to enhance the health and quality of life of the elderly. Recent research has highlighted the significance of the gut microbiota in promoting longevity [213, 224], with numerous studies demonstrating the age-related differentiation of gut microbiota composition in humans across different populations and age groups [236, 316–320]. Although there is high inter-individual variance, a sequential trajectory has been found between age transitions [316]. In mice, healthy aging has been linked to the prolonged retention of microbiota specific to a younger age, with fecal microbiota transplantation from young mice reversing aging-linked deterioration in aged mice [240]. Additionally, the loss of critical taxa during aging that are essential for certain types of metabolism can be compensated for by the gain of alternative taxa that perform the same type of metabolism [316, 319].

Centenarians, who are at the extremes of human aging, provide valuable information about the mechanisms of longevity because they age most successfully and are well-adapted to the challenges of aging. Centenarians have a unique set of (non)-genetic factors that confer advantages [321], such as a lower incidence of chronic disease, lower mortality rates, and longer life spans [321–323]. People who age well are characterized with a more diverse gut microbiota is, whereas frail individuals have relatively less richness and diversity [213, 324]. These findings are recapitulated in the extreme cases of aging, where the microbiota of centenarians is comparatively more diverse than that of younger individuals from various populations [236, 316–320], more adaptable to opportunistic bacteria [325], and more efficient in lipid and amino acid metabolism [318].

Many factors influence the formation of the microbiota, resulting in high variability of the microbiota between different individuals and populations, especially when integrated over a lifetime. Yet, despite this heterogeneity, there is significant overlap in the gut microbiota composition that emerges in centenarians. Centenarians have an increased capacity for central metabolism production, including glycolysis and fermentation of crucial metabolites, such as short-chain fatty acids (SCFAs) [237, 326]. Depletion of SCFA-producing taxa has been associated with IBS [327] and neurodegenerative disorders such as Alzheimer's disease [328, 329]. Similarly, cross-sectional studies have shown that patients with Alzheimer's disease have a different gut microbiota than healthy individuals, suggesting that it may also play a role in the development or progression of the disease [328–330].

In this cross-sectional study, we examined, Dutch subpopulations of cognitively healthy centenarians [331] and a memory clinic cohort comprising individuals with Alzheimer's disease dementia or subjective cognitive decline, stratified based on indices of healthy and unhealthy aging [332]. Prior research on this specific memory clinic cohort, as reported in [329], has identified gut microbiota characteristics specific to Alzheimer's patients. Our objective was to juxtapose this group with cognitively healthy centenarians and middle-aged controls to discern distinct microbiota signatures indicative of healthy or unhealthy aging that extend beyond the scope of previous findings. We posited contrasting profiles associated to healthy and unhealthy aging: characterized by a more diverse or robust microbiota in successfully aging centenarians as opposed to a diminished diversity or resilience in the gut microbiota of Alzheimer's disease patients, as evidenced by the microbial alterations observed in these subjects.

9.2 Methods

9.2.1 Sample specification

We included participants with available fecal samples from several populations in the Netherlands. These were centenarians (CT) from the 100-plus Study cohort [331] and a memory clinic population from the Amsterdam Dementia Cohort (ADC) [332] of Alzheimer's disease (AD) dementia patients and subjects without cognitive impairment (SCD). We included 50 centenarians from the 100-plus study cohort, which selects participants for preserved cognitive function. The 100-plus Study cohort includes Dutch-speaking individuals who can provide official proof that they are 100 years of age or older, who report being cognitively healthy (which is confirmed by a proxy), who consent to give a blood sample, while optionally giving a stool sample, who consent to (at least) 2 home visits by a researcher, and who consent to an interview and a battery of neuropsychological tests. We included 149 ADC subjects. Note that the microbiota profiles of this particular population have previously been studied [329]. All ADC subjects underwent neuropsychological assessment and neurological examination as part of a standard dementia screening [333]. AD diagnoses were made by consensus at a multidisciplinary meeting according to the National Institute on Aging-Alzheimer's Association criteria [334]. Subjects with SCD presented with memory complaints but did not show objective cognitive impairment on neuropsychological testing nor meet criteria for dementia, psychiatric diagnoses, or other neurological diagnoses, and were used as a control group compared with the two other populations. Global cognitive functioning was assessed using the Mini-Mental State Examination (MMSE) [335]. The characteristics of the subjects enrolled in this study, which include only the common features found among all populations, are summarized in Table 9.1. Although measurements of Alzheimer's disease (AD) biomarkers were available, they were not used in the study due to differences in data availability. The centenarian cohort had biomarker data only from plasma, while the ADC has data from cerebrospinal fluid (CSF) measurements.

Subjects were asked to store their feces sample in a freezer, samples were transported to the hospital in a cooling bag. Samples were shipped to Erasmus Medical Center, Rotterdam, the Netherlands, for sequencing. Aliquots of ~300 mg feces were homogenized. DNA was isolated using bead-beating and the InviMag Stool DNA kit (Invitek Molecular GmbH, Berlin, Germany) on a KingFisher Flex robot (Thermo Fisher Scientific, Breda, Netherlands).

Table. 9.1. Characteristics of study subjects

	SCD	AD	CT
Subjects (n)	116	33	50
Sex (male/female)	65/51 (56/44%)	18/15 (54.5/45.5%)	20/30 (40/60%)
Age (yr)	62.0 ± 7.5 [44, 80]	66.0 ± 8.0 [48, 85]	100.8 ± 1.0 [100, 106]
MMSE (0, 30)	28.7 ± 2.4 [9, 30]	20.3 ± 5.8 [0, 29]	25.4 ± 3.5 [15, 30]

[a] The total number of subjects is 199, excluding individuals with unqualified stool samples or insufficient sequencing data. MMSE: mini-mental state examination. Value for parameters are shown as mean ± SD [range], for each group.

9.2.2 Gut microbiota determination

Fecal microbiota composition was determined by sequencing the V3-V4 hypervariable regions of the 16S rRNA gene on an Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA) using 319F (ACTCCTACGGGAGGCAGCAG) -806R (GGACTACHVGGG TWTCTAAT) primers and dual-indexing. Pre-processing and data analysis were matched with the procedures in [329]. The obtained sequence data consisted of high-quality filtered reads. Before further processing, samples were excluded with insufficient read count by setting a cut-off based on the read count distributions (Supplemental Figure S9.1). 5 ADC and 2 CT samples had insufficient counts and were excluded from the study. 16S rRNA primers were removed from the paired sequencing reads using TagCleaner v0.16 [336].

The reads were then processed into exact sequence variants (ESVs) using DADA2 v1.22.0 [260] as follows: after reviewing the read quality profiles, 50 bases were trimmed from the 5' end of the forward reads, and 60 bases from the 5' of the reverse reads. Reads were truncated at the first base with a Q score less than 4, quality filtered using 2 maximum expected error for forward reads, and 4 maximum expected errors for reverse reads — allowing no ambiguous bases. Reads were also de-duplicated and the remaining filtered reads were used to learn error rates and to infer ESVs separately for forward and reverse reads. The ESVs for the forward and reverse strands were merged by allowing no mismatched bases and requiring a minimum overlap of 20 bases. ESVs were deleted if they were chimeric or if their length was outside the [350, 500] range.

Taxonomy down to the species level was assigned using RDP to the remaining ESVs using DADA2, and the SILVA databases v.138 [337], allowing up to 3 multiple species-level assignments. It is possible for distinct ESVs to correspond to the same taxa, either due to ambiguous taxonomic assignment or genetic variation within strains. In such cases, the ESVs are considered separate biological features and are not merged. If a species for an ESV was not classified using the methodology described, BLAST was utilized to assign it, provided that an unambiguous match was identified. If no match was found, the taxa or ESVs were left unlabeled and explicitly tagged. To generate multiple sequence alignments of the ESVs, we used MAFFT v7.475 [338], and the phylogeny was reconstructed using

FastTreeDBL v2.1.11 [339]. The ESVs, taxonomy, phylogenetic tree, and metadata were integrated using the R package, phyloseq v1.38.0 [340].

9.2.3 Data normalization

Data were normalized to correct for variability between samples and between cohorts due to sampling differences (Supplemental Figure S9.1). Doing so enables modeling the actual abundance in the original samples from the read counts and ensures that the abundance distributions conform to the needs of statistical analysis. The high variability of results generated by differential analysis methods can be attributed to various factors, including: 1) violation of test assumptions due to data compositionality; 2) excessively high false discovery rates (FDR); 3) inadequate addressing of the high sparsity of microbiota data; and 4) differences in the number of ESVs that are identified as significantly differentially abundant [296].

Therefore, to ensure more robust biological interpretations in the differential analysis, two independent normalization methods were used in parallel. This approach aimed to mitigate the issues mentioned earlier and provide a more reliable result. The first choice was rarefaction, a method that randomly subsamples sequences and removes them from the sample library up to a defined library threshold, thus equalizing all samples [283]. Although rarefaction results in a loss of statistical power, this was found to be offset by its high control of false positives [285]. The library threshold for rarefaction was determined by generating rarefaction curves [341] (Supplemental Figure S9.2), and ultimately set conservatively so that no sample was excluded from either cohort, with 20,000 counts per sample (Supplemental Figure S9.3). The second normalization technique, ANCOM2 [295, 300], a compositional method, was chosen because it is conservative and has a high degree of concordance in benchmarks between differential analysis methods [296]. ANCOM2 uses the additive log-ratio transformation where the reference is the count abundance of a single taxon, which should be present with low variance in read counts across samples. In this case, the ratio between the reference taxon chosen and each taxon in that sample are compared across different sample groupings [295, 300].

9.2.4 Statistical analysis

The rarefied data were used to compare the composition of the microbiota between groups, and measures of α-diversity, including richness [274], evenness [275], and Shannon's diversity index [276] were determined. Groups were statistically tested using Wilcoxon rank-sum tests with Benjamini-Hochberg (BH) correction. The β-diversity of the groups was compared on the basis of the unweighted Unifrac distance [279, 280], visualized using non-metric multidimensional scaling (NMDS), and tested with permutational multivariate ANOVA (PERMANOVA) after conformation of heteroscedasticity between groups [282]. To determine the determinants contributing significantly to the NMDS loadings of specific groups, vector fitting permutation tests were used [342]. Kruskal-Wallis tests were performed to determine if there were significant differences in phyla between groups, followed by Dunn's post hoc multiple comparison test. Differential analysis of rarefied data was performed by determining the relative abundance of taxa and then pooling taxa at the genus level based on the median relative abundance across all samples. The threshold for pooling was set at ≥ 0.25%. Significance of taxa between groups was tested using Kruskal-

Wallis tests. Pairwise tests between significant taxon groups were then performed using Wilcoxon rank-sum tests; both tests were BH-corrected. ANCOM2 requires pre-processing prior to differential analysis by examining ESV abundances to identify different subtypes of zero so that they can be accounted for [300]. Taxa were filtered out if their proportions of zeros were greater than 90% in all samples. A pseudo-count of 1 was applied to the dataset to allow for log transformation. The significance of all additive log ratios for each taxon was tested using Wilcoxon rank-sum tests, and p-values were again corrected by BH. In ANCOM2, a detection threshold is applied as described in the original publication [295, 300], whereby a taxon is said to be differentially abundant if the corrected p-values reaching nominal significance for that taxon were greater than 70% of the maximum possible number of significant comparisons.

9.3 Results

9.3.1 α/β-diversity measurements

Per subject in the CT and ADC cohorts, respectively, we had ~214,236 (±89,463) and ~140,175 (±28,370) reads. From the 199 samples, we constructed 9,216 exact sequence variants (ESVs), of which 8,855 remained after rarefaction. When classified to the genus level, this yielded 284 groups (283 genera and one unidentified group). The CT group had the most distinct genus classifications (127), and the AD group had the fewest (34). The data set included 393 ESVs found in at least 20% of the samples in all cohorts. Based on the calculated α-diversity measures, as shown in Figure 9.1, there were significant differences in the CT group's richness and evenness measures with respect to the AD and SCD groups (and no difference between AD and SCD), but none given the Shannon diversity index.

The β-diversity of the microbial community was determined by calculating the unweighted UniFrac distances. The (dis)similarity, distribution, and clustering of subjects in the different groups were visualized by applying NMDS on the unweighted UniFrac distances of the ~8,855 ESVs (Figure 9.2). A slight shift was present between the clusters corresponding to the AD and SCD groups. The CT group, on the other hand, showed a pronounced shift regarding the AD and SCD. Pairwise PERMANOVA tests on the unweighted Unifrac showed that the CT group differed significantly from the AD and SCD groups ($p < 0.001$). At the same time, we observed no significant differences between the AD and SCD groups ($p > 0.05$). A few dominant genera largely drove the distribution of individuals in the NMDS. Genera contributing significantly to the ordination of CT samples, using vector fit permutation tests, $p < 0.01$ and $R^2 > 0.25$, included *Faecalibacterium*, *Flavonifractor*, *Eggerthella*, *Eisenbergiella*, *Anaerotruncus*, and *Family XIII AD3011 group*. Vector permutation tests on metadata variables showed that age was a significant factor in explaining variation between groups. In contrast, no sex- or MMSE-specific differences were observed.

9.3.2 Microbiota composition

At the phyla level, analysis of the taxonomy and composition of the three groups showed the presence of four predominant phyla that accounted for ~98.80% of all sequences detected in the samples: Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria (Figure 9.3a), which is consistent with previous results from human gut studies. Firmicutes was the most abundant phylum detected in the SCD, AD, and CT subjects, accounting for averages of

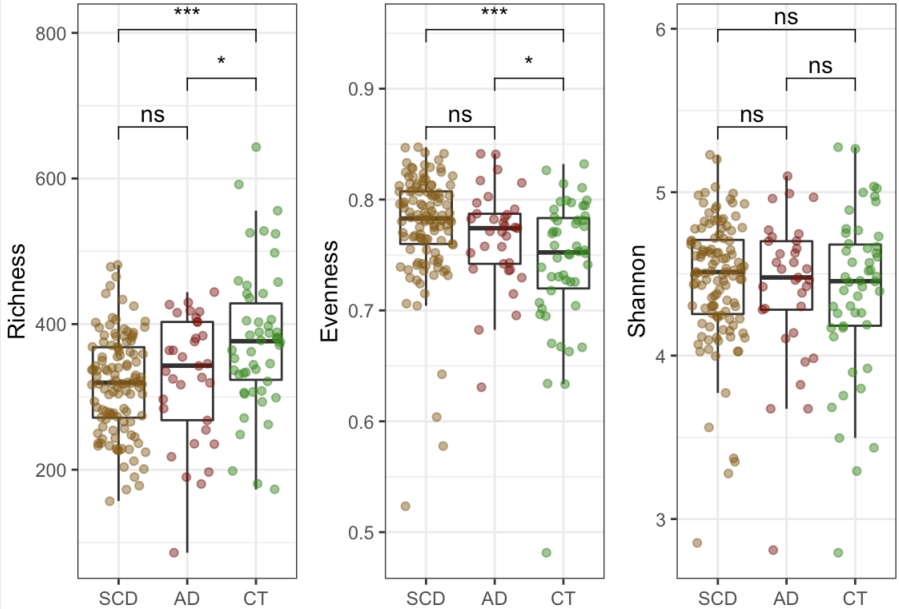


Figure. 9.1. Alpha measures: richness, evenness, and the Shannon diversity index, calculated on the rarefied data per group.

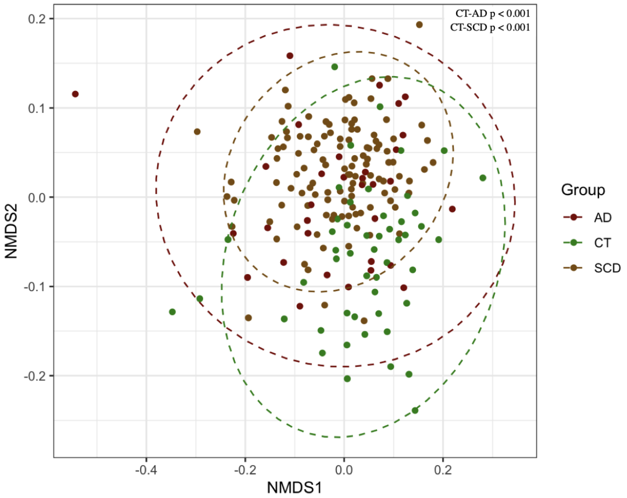
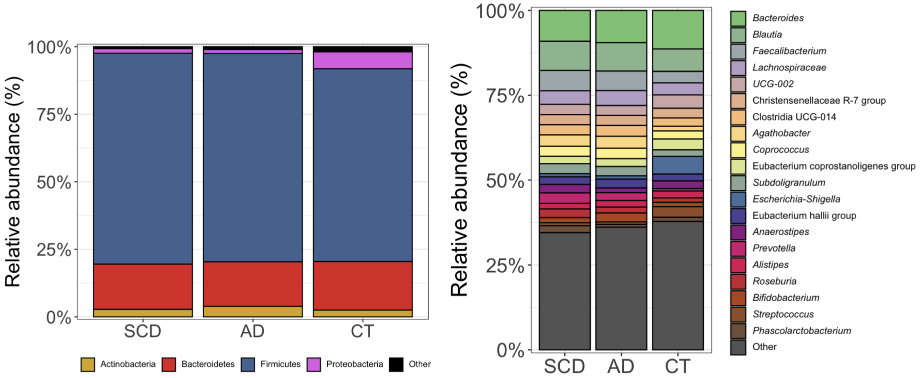


Figure. 9.2. NMDS plot of unweighted Unifrac distances per group. The pairwise PERMANOVA p-values are shown for the groups with significant differences.

~78.0%, ~77.1, and ~71.3% of total sequences, respectively. Between the AD and SCD control groups, no significant change was observed. However, the CT group had an increase in the proportion of *Proteobacteria* (6.26%), p = 0.00002, compared to the SCD (~1.73%) and AD

(~1.44%) groups. In addition, the CT group had a significant reduction in the proportion of Firmicutes, p = 0.004, compared to the SCD group. The other phyla Actinobacteria (~2.78%, ~3.89%, ~2.52%) and Bacteroidetes (~16.8%, ~16.5%, ~18.0%) were also present in the SCD, AD, and CT groups, respectively, but no significant differences between groups were found. As the two most abundant phyla in the gut, the Firmicutes/Bacteroidetes ratio is often used as an index of the structure of the gut microbiota [343]. However, no significant difference in the ratio, 5.58 : 5.26 : 4.51 (median of SCD : AD : CT), was found between the groups.



(a) **(b)**

Figure. 9.3. The microbiota composition of the SCD, AD and CT groups in terms of relative abundance was obtained from the rarefied data. a) Phylum-level composition. b) Genus-level composition of the top 20 genera.

9.3.3 Differential abundance analysis

The relative abundance of the 20 most abundant genera in the gut microbiota of the groups was determined to further explore the composition of the gut microbiota (Figure 9.3b). We observed lower abundance of genera belonging to *Blautia*, *Faecalibacterium*, *Agathobacter*, *Subdoligranulum*, and *Bifidobacterium* in centenarians compared to the AD and SCD groups, with genera *UCG-002*, *Bacteroides*, and *Escherichia-Shigella* enriched. To determine and test the differential abundance of genera/species in the rarefied data, ESVs were selected from the groups based on their median relative abundance (≥ 0.25%) and otherwise pooled, as shown in Figure 9.4, with significant differences annotated.

Pairwise tests were performed between the SCD, AD, and CT groups to determine differential abundance from the absolute count data using ANCOM2, as shown in Table 9.2. ANCOM2 filtered out ESVs for which count data were too sparse (≥ 90% zeros in all samples), resulting in ~800 tested ESVs. We considered ESVs to be differentially abundant if 1) the number of rejections of the null hypothesis (the average abundance of a given taxon in one group is equal to that of the other group) for a given ESV was ≥ 70% of the total ESVs tested. 2) the absolute mean difference in the additive log ratio between groups was ≥ 0.5. The differential abundance methods produced consistent results in four ESVs. These ESVs are *Escherichia-Shigella spp.*, *Ruminococcus torques*, *Faecalibacterium prausnitzii*, and *Lachnospiraceae ND3007 group*. In the case of *Escherichia-Shigella spp.* and

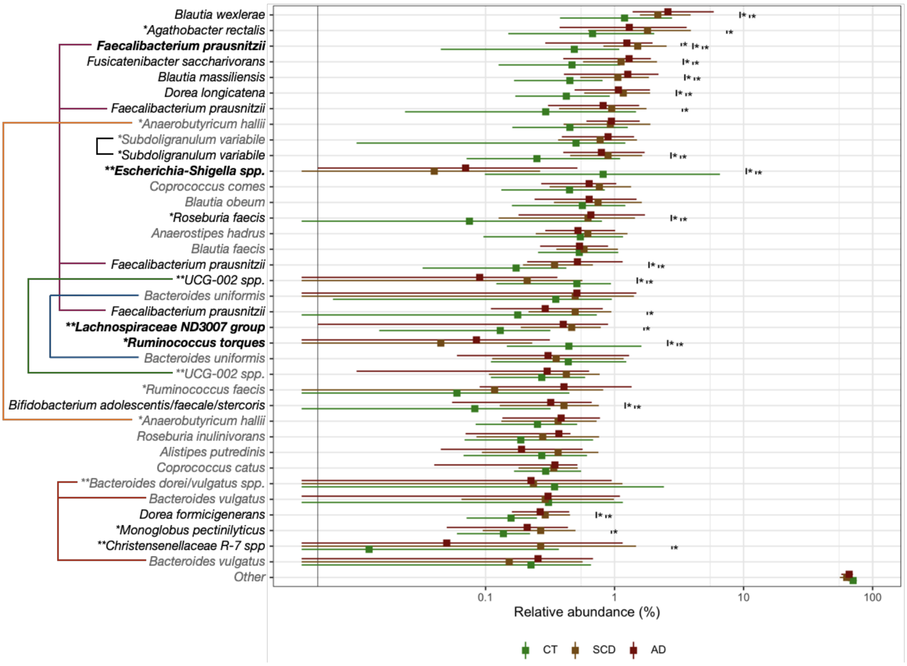


Figure. 9.4. Differential analysis of relative abundance of rarefied and median pooled (≥ 0.25%) data tested in SCD, AD, and CT groups using Kruskal-Wallis tests and pairwise testing between groups of significant ESVs (shown in black) using Wilcoxon rank-sum tests. Colored lines connect ESVs corresponding to the same taxa. ESVs in bold were also found to be significant in the ANCOM2 analysis (Table 9.2). For example, an ESV corresponding to *Blautia wexlerae* was found to be significantly less abundant within the CT group relative to both the AD and SCD groups, whereas no significant difference was found for the ESV corresponding to *Coprococcus comes* between any of the groups.

*ESVs in which species remained unclassified, assigned using BLAST without allowing for any ambiguity.

**ESVs where no unambiguous hits were found using BLAST.

Ruminococcus torques, their abundance increased, while for *Faecalibacterium prausnitzii* and *Lachnospiraceae ND3007 group*, their abundance decreased in CT compared to SCD or AD. One ESV: *Bacteroides uniformis*, was found to be more abundant in CT concerning SCD and AD only using ANCOM2. The other ESVs found to be differentially abundant with ANCOM2 all had a median relative abundance of < 0.25% in the rarefied data, where they were also differentially abundant. Several distinct ESVs, all corresponding to (potentially different strains of) the taxa *Christensenellaceae R-7 spp.*, were found to be both less abundant (rarefaction analysis) and more abundant (ANCOM2 analysis) in the CT group compared to the AD and SCD groups. ANCOM2 found no significant difference between the AD and SCD groups, and only a single ESV corresponding to *Faecalibacterium prausnitzii* was differentially abundant between the AD and SCD groups in the rarefied data.

Table. 9.2. Differential abundance analysis using pairwise ANCOM2 testing between the groups ranked according to the additive log ratio (ALR) mean difference in abundance given a grouping. For example, in CT - SCD, a positive ALR indicates that a taxon is more abundant in CT; conversely, a negative ALR indicates that a taxon is more abundant in SCD. Only differentially abundant ESVs exceeding the 70% threshold and those with a mean absolute ALR difference ≥ 0.5 are shown. ESVs marked in black or gray are also significant with the rarefaction approach with median relative abundance ≥ 0.25% (Figure 9.4) and < 0.25%, respectively. ESVs marked in bold are significant according to ANCOM2 but not with the rarefaction approach.

CT - SCD		CT - AD	
ESV	ALR	ESV	ALR
Escherichia-Shigella spp.**	1.93	_Ruminococcus torques_*	1.57
_Ruthenibacterium lactatiformans_*	1.80	_Escherichia-Shigella spp._**	1.53
_Ruminococcus torques_*	1.78	_Eubacterium coprostanoligenes_*	1.32
_Ruminococcus gnavus_*	1.08	**_Bacteroides uniformis_**	1.29
Christensenellaceae R-7 spp.**	1.07	_Ruthenibacterium lactatiformans_*	1.22
Faecalibacterium prausnitzii	-0.97	_Ruminococcus gnavus_*	1.01
_Eubacterium siraeum_*	0.93	_Faecalibacterium prausnitzii_	-0.71
_Bacteroides thetaiotaomicron_*	0.90	_Christensenellaceae R-7 spp._**	0.58
Bacteroides uniformis	0.81		
Roseburia intestinalis	-0.70		
Christensenellaceae R-7 spp.**	0.66		
Lachnospiraceae ND3007 group**	-0.59		
Bacteroides fragilis	0.59		
_Ruminococcus faecis_*	-0.54		
Alistipes finegoldii	0.53		
Lachnoclostridium spp.**	0.52		

* ESVs in which species remained unclassified, assigned using BLAST without allowing for any ambiguity.

** ESVs where no unambiguous hits were found using BLAST.

9.3.4 Core gut microbiota

We identified the core gut microbiota by classifying the top 15 bacterial taxa at the species level, or at the genus level if species identification was not possible. We determined the prevalence of these taxa among all subjects in the SCD, AD, and CT groups and ranked them accordingly, as shown in Table 9.3. Eight taxa belonging to the major phyla of Firmicutes were common to all groups. Five taxa were commonly found in the SCD and AD groups and may differentiate these from the CT group because these particular taxa are present at a much lower prevalence than in the other groups. One additional taxon *Lachnospiraceae ND3007 group*, one taxon *Collinsella aerofaciens*, and six taxa *Ruthenibacterium lactatiformans*, *Ruminococcus torques*, *Escherichia-Shigella spp.*, *Family XIII AD3011 group*, *Bacteroides uniformis*, and *Alistipes onderdonkii*, showed high prevalence only in the SCD, AD, and CT groups, respectively, and may allow for group differentiation.

Table. 9.3. The core gut microbiota of SCD, AD, and CT in terms of prevalence (the percentage of subjects in each group with a particular ESV). ESVs marked in bold have a higher prevalence in one group than in the other groups. ESVs marked in gray share a high prevalence across all groups.

SCD		AD		CT	
Taxa	Prev. (%)	Taxa	Prev. (%)	Taxa	Prev. (%)
Blautia wexlerae	100.0	_Blautia wexlerae_	100.0	_Blautia faecis_	100.0
Anaerostipes hadrus	100.0	_Subdoligranulum variabile_[*]	100.0	_Blautia wexlerae_	98.0
Blautia faecis	99.14	_Blautia obeum_	100.0	**_Ruthenibacterium lactatiformans_**[*]	98.0
Fusicatenibacter saccharivorans	99.14	_Blautia faecis_	100.0	_Blautia obeum_	98.0
Blautia massiliensis	99.14	_Blautia massiliensis_	100.0	_Blautia massiliensis_	98.0
Subdoligranulum variabile[*]	97.41	_Dorea longicatena_	96.97	**_Escherichia-Shigella spp._**[**]	94.0
Blautia obeum	97.41	_Faecalibacterium prausnitzii_	96.97	_Fusicatenibacter saccharivorans_	94.0
Agathobacter rectalis[*]	97.41	_Anaerostipes hadrus_	93.94	**_Ruminococcus torques_**[*]	92.0
Dorea formicigenerans	97.41	_Fusicatenibacter saccharivorans_	93.94	_Monoglobus pectinilyticus_[*]	92.0
Faecalibacterium prausnitzii	95.69	_Dorea formicigenerans_	93.94	_Anaerostipes hadrus_	90.0
Monoglobus pectinilyticus[*]	95.69	_Anaerobutyricum hallii_	93.94	_Dorea longicatena_	88.0
Lachnospiraceae ND3007 spp.[**]	94.83	_Monoglobus pectinilyticus_[*]	93.94	**_Family XIII AD3011 group_**[**]	88.0
Dorea longicatena	93.97	_Coprococcus comes_	90.91	**_Bacteroides uniformis_**	88.0
Anaerobutyricum hallii	93.97	_Agathobacter rectalis_[*]	90.91	**_Alistipes onderdonkii_**	88.0
Coprococcus catus	93.97	**_Collinsella aerofaciens_**	90.91	_Coprococcus catus_	88.0

[*] ESVs in which unclassified species were assigned using BLAST without allowing for any ambiguity.
[**] ESVs where no unambiguous hits were found using BLAST.

9.4 Discussion

In this work, the gut microbiota of centenarians was compared to that of memory clinic subjects to identify potential community structures associated with either Alzheimer's disease or healthy aging. By understanding the gut microbiota of these groups, it may be possible to contribute to a better understanding of both conditions. We found significant differences in the composition of the gut microbiota of centenarians. In agreement with previous findings in other cohorts of centenarians [236, 316–320], richness and diversity were higher compared to the comparatively younger AD and SCD groups.

Only slight differentiation was found within the memory clinic population, with the AD patients and SCD subjects clustering together and showing insignificant variance in the PERMANOVA analysis, as has also been shown previously using the same methodology [329]. Across the groups most of the identified microorganisms belonged to Firmicutes and Bacteroidetes at the phylum level, while Proteobacteria, Actinobacteria, and other phyla contributed less than 10% of the total. Therefore, the overall composition of the Dutch gut microbiota at the phyla level was similar to that observed in other populations [236, 316–320]. The association between changes at the ESV level and taxonomic rank becomes less obvious with higher taxonomic rank; however, large-scale trends can still be observed. ESVs classified at the phyla-level Bacteroidetes, and Actinobacteria were similar in all groups. In contrast, the abundance of Firmicutes and Proteobacteria revealed large-scale differences in centenarians regarding the memory clinic cohort, decreasing and increasing, respectively. Although the ratio of Firmicutes/Bacteroidetes in centenarians was lower than in younger groups, this was not a significant difference, as has been observed in other cohorts. The class *Clostridia*, belonging to the phylum Firmicutes as well as the phyla Bacteroidetes and Proteobacteria, dominated the gut microbiota of the Dutch subjects; approximately 90% of all subjects shared them. It should be noted that several genera were relatively consistent between groups, with members of the genus *Blautia* genus being the most consistent. In contrast, members of the genera *Monoglobus*, *Anaerostipes*, *Dorea*, and *Fusicatenibacter* were more variable. These genera may be part of a core microbiota for these cohorts and perhaps for residents of the Netherlands in general.

Among the inhabitants of the gut microbiota, SCFA-producing bacteria are some of the most important; major SCFA producers include the genera *Faecalibacterium*, *Roseburia*, *Eubacterium*, *Dorea*, *Coprococcus*, and *Blautia*, with their metabolites being vital to human health, particularly in the case of butyrate. The many properties of butyrate have been shown to prevent or delay age-related decline, including improving gut barrier function and insulin resistance, as well as delaying inflammation, cancer development, and cognitive decline [213]. In Dutch centenarians, several butyrate-producing bacteria are more abundant, including *Ruthenibacterium lactatiformans* and *Bacteroides fragilis*, indicating that centenarians may possess taxa associated with longevity. However, several other butyrate-producing bacteria usually associated with health in younger age groups are present in centenarians in lower abundance, including *Faecalibacterium prausnitzii*, *Agathobacter rectalis*, *Roseburia*, and *Subdoligranulum variabile*. In addition, the health-associated *Bifidobacterium* decreases in aging individuals, although, as with the reduced genera *Faecalibacterium* and *Fusicatenibacter*, this phenomenon is expected [344]. Note that the loss of these taxa is not as severe in healthy older individuals and that the gain of other health-related taxa may also offset the loss of these taxa [344, 345]. For example, we observe an inconsistent increase in the genus *Christensenellaceae*, a potential ecosystem signature in individuals with extreme longevity or a healthy aging trajectory [325], which is often lost when transitioning to a state of physiological decline, such as frailty [346] or cognitive decline [347]. In addition, age-related impairment of *Roseburia* shows variation between studies, where, as in the present study, a reduction is observed in other centenarian populations [318, 320]. Centenarians had a greater abundance of *UCG-002*, which is positively correlated with more active individuals and physical activity [348]. Furthermore, we noted an increase in potential bacterial pathogens such as *Ruminococcus* and *Escherichia-Shigella spp.*, both of which were associated with unhealthy aging [213]. Although other taxa may have suppressed the deleterious effects of these pathobionts, previous work has shown a spike in these taxa before death [236]. Finally, in another cohort of aging individuals studying physical frailty and cognitive health, unhealthy aging was characterized by an increase in the genera *Ruminococcus* and *Blautia*, with a decrease in *Christensenellaceae* among the other genera and an overall decrease in diversity [347].

Previous cross-sectional studies of differences in gut microbiota between AD patients and controls found that several microbes were less abundant, including *Faecalibacterium prausnitzii*, *Eubacterium*, *Anaerostipes*, *Ruminococcus*, and *Roseburia*. In contrast, other microbes were more abundant, such as *Bacteroides* and *Alistipes* [328, 330] and [329]. Of these taxa, we observed only a decrease in *Faecalibacterium prausnitzii* in AD patients relative to SCD subjects; note that this particular taxon is even less abundant in healthy centenarians. Furthermore, the taxa *Eubacterium*, *Ruminococcus*, *Bacteroides*, and *Alistipes* were more abundant in the centenarians, with *Roseburia* being less abundant. Overall, the observed decrease of *Faecalibacterium prausnitzii* in AD patients and distinct gut microbiota profile in healthy centenarians emphasize the intricate connection between gut microbiota, aging, and neurodegenerative conditions.

Because of the variability in results introduced by different normalization or differential abundance methods, we opted for a consensus of two independent approaches to analyze our data: rarefaction and ANCOM2. The reasoning behind this approach is that a robust result should appear across multiple independent methods. We found that the methods identified a similar set of differentially abundant taxa, with almost no disagreement. However, ANCOM2 is substantially more conservative than the rarefaction approach. This suggests that a consensus approach based on multiple differential abundance methods may help to achieve a more reliable result. Since discrepancies between different methods can be resolved by considering the results in a broader context, while concordant results can highlight robustness. Using multiple methods for normalization and differential abundance analysis introduces additional complexity into the process, as they may require different pre-processing or filtering steps. Interpretation of results can also become more difficult, especially if there are discrepancies between methods. Furthermore, there are no well-defined methods to unify the results of several independent methods. It is therefore essential to consider the limitations of each method and the impact they may have on the results. In addition to the cost of analysis, it is also difficult to determine which methods are most appropriate for a given data set. The optimal approach to this issue is currently unknown, and we recommend that the use of multiple methods be evaluated on a case-by-case basis or informed by benchmark studies.

Rarefaction, being a random process, offers only a snapshot of the microbial community at the smallest normalized library size, omitting a random subset of observed sequences and introducing artificial variation. Repeating rarefaction and downstream analysis with different seeds can help account for this data loss.

Although we found significant differences in the richness and evenness measures, the Shannon diversity index (a function of richness and evenness) was nearly identical between groups. The index's inherent weakness is that it can yield identical results for different combinations of richness and evenness. This emphasizes the importance of utilizing multiple measures when assessing diversity, in order to provide a more comprehensive and contextual understanding of the microbial community structure

In conclusion, the gut microbiota of centenarians exhibits several characteristics that appear to be universal across different geographic populations. These characteristics may be a consequence of adaptations to aging and allow the gut microbiota to better resist the effects of aging. Differentiation appears to be primarily age-related, and we did not observe sex-specific differences, which is consistent with previous findings that sex-specific characteristics decrease with age [349]. However, the characteristics of the shared microbiota and the impact on the health of centenarians remain unclear at this stage.

9.5 Supplementary materials

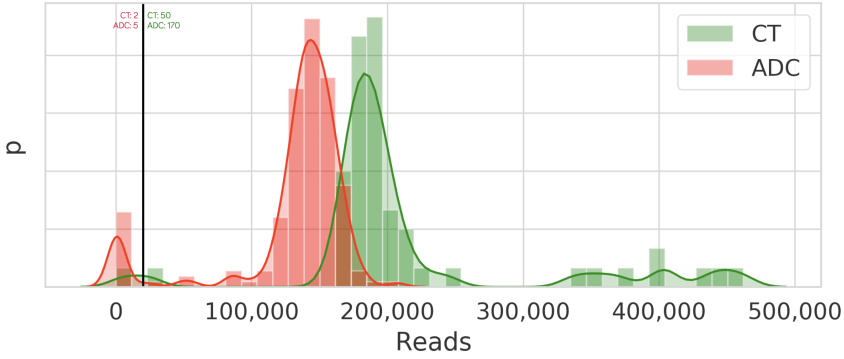


Figure. S9.1. Read distributions of the centenarian (CT) and memory clinic (ADC) populations.

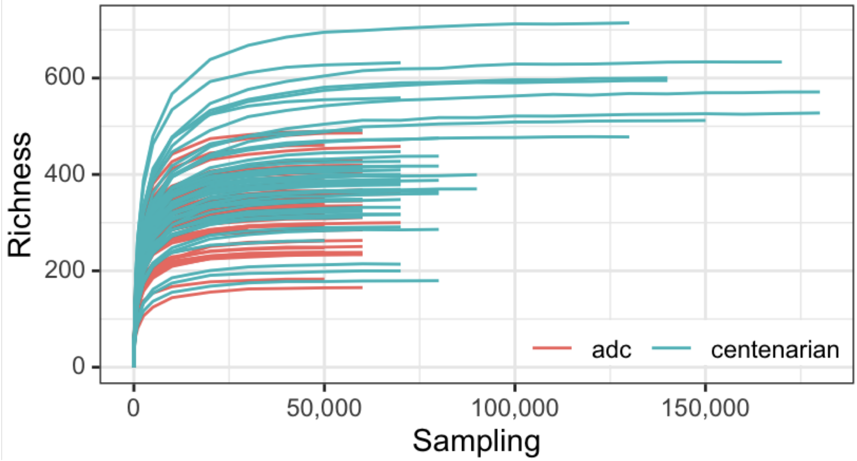


Figure. S9.2. Rarefaction curves.

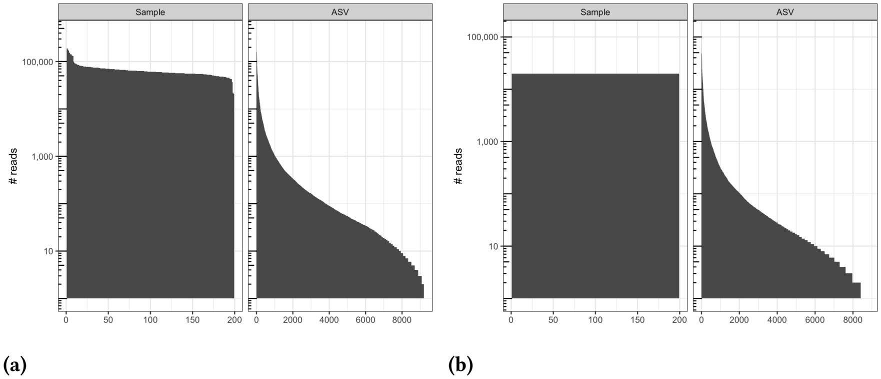


Figure. S9.3. Per sample read counts and per ESV counts. a) Un-normalized data. b) Rarefied data.

10

Conclusion

The gut microbiota has increasingly been shown to be involved in a wide range of human health conditions that have been studied to date. However, the microbiota is a complex ecosystem of many bacterial species of which the functional effects of the individual are often unknown or poorly understood, especially in the context of the broader community of bacteria that interact with one another. The microbiota's relationship with the host is further complicated since it interacts with multiple facets of host physiology. Nevertheless, it is vital to put the microbiota into the context of many different conditions to better understand its involvement in human health and disease. Yet, unlike the genome, the microbiota is in constant flux, with massive inter-individual and population-specific differentiation that can be affected by nearly any factor imaginable. Moreover, microbiota composition varies within a given individual's life span, making it difficult to study it holistically to link composition to function. While it is possible to determine the differentiating taxa between conditions, it remains challenging to link taxa to function, mainly because of the robust adaptability of the microbiota to compensate for any imbalance or missing taxa. Ultimately, there is no ideal or optimal microbiota configuration but rather a personalized bacterial signature that enters into a dynamic balance, given that its functional redundancy may lead to the same functional outcome regardless of the taxa present. Hence, investigating the microbiota composition should be regarded as a first step toward understanding the microbiota's effects on the host.

The metagenomics analysis in this work relies on partially sequencing the 16 rRNA gene 16S, a taxonomic marker that encodes approximately 1,500 bp of information. This approach makes a trade-off by sacrificing the accuracy of taxonomic assignment and phylogenetic resolution in exchange for more affordable deep sequencing than would otherwise be possible with WGS. It is essential to keep in mind the bias introduced by partial sequencing since different sub-regions of 16S result in classification variation across phyla and is inadequate for complete species level assignment. Such bias must be considered when comparing results from metagenomic analyses relying on 16S sequence classification, especially across studies that use different gene regions. The current availability of third-generation technologies means that high-throughput sequencing of the complete 16S gene is becoming more affordable and reliable due to fewer systemic sequencing errors and

PCR bias. Long reads facilitate more sensitive discrimination of different taxonomic levels, such as closely related strains, and identify multiple copies of 16S in species. Nevertheless, sequencing 16S alone will not provide a representative view of bacterial diversity in an ecosystem as it excludes all other discriminatory variations encoded within the genome, such that a shift to shotgun metagenomics will be essential to resolve species- and strain-specific variations.

Differential abundance analysis, which aims to detect taxa that differentiate ecosystems, is a complex problem in metagenomic studies because of the inaccessibility of the data needed to draw conclusions about microbial taxa that differ between two or more ecosystems. This problem stems partly from the difficulty of 1) obtaining samples from different ecosystems; 2) identifying and quantifying all the different taxa present in a sample; 3) accurately comparing samples from different ecosystems. An important unobservable parameter that impacts the differential abundance analysis is the random sampling of a sample from an ecosystem, which is affected by the size of the sequencing library, the total number of bacteria present in a sample, and the fraction of the sample obtained from the ecosystem. It is difficult to correct for this sampling because the number of taxa present in a sample is usually unknown or affected by sampling a subsystem in an ecosystem. Yet, as discussed, some methods attempt to correct these uncertainties when comparing samples. Another challenge is that the observed microbiota data are absolute abundances with a high proportion of zeros, making it difficult to determine which taxa are differentially abundant. It is difficult to assess whether and how this rarity should be addressed before differential analysis, although excluding taxa found in only some samples seems reasonable. The fundamental reason is that otherwise, the burden of correcting multiple tests becomes so great as to preclude the identification of any differentially abundant taxa. It should be recognized that more work is needed to establish an optimal threshold rather than arbitrarily selecting one. The microbiome field lacks consensus regarding standardization and differential analysis; this can make it difficult to transfer and generalize results from one study to another, as specific methods produce very different results. Therefore, using only one method in a study may be inadvisable, as this will likely introduce bias. Furthermore, it is necessary to use the same methodologies to eliminate this method-specific bias when comparing studies. A consensus approach based on multiple differential abundance methods might yield more robust result. In other words, if different methods produce similar results, we can have greater confidence in those results. However, discrepancies between methodologies can also highlight the robustness of specific results across independent studies.

This work investigated the gut microbiota composition of different subpopulations in the Netherlands. These subpopulations included those who were healthy, those with Alzheimer's disease, and those who were cognitively healthy centenarians. There is excellent potential to identify microbiome markers linked with longevity or healthy aging. Possible by studying the factors that differentiate centenarians from other, less long-lived individuals. However, only a limited and inconsistent alteration of microbiome composition was observed in the group with Alzheimer's disease. The microbiota composition of centenarians shows distinct qualities, with unexpectedly high diversity relative to younger individuals. Moreover, it shares striking similarities with other studies of extreme aging, despite the wide variation inherent to study populations' demographics and physiological status and methodological differences in sample collection, sample storage, DNA extraction protocol, sequencing platform, or sequencing center performance.

Despite a large body of research, many aspects of the gut microbiota and its relation to healthy aging are still unknown. Although, there are differentiating factors unique to extremely aged or Alzheimer disease-affected individuals. It is difficult to determine how these differences arise and assess their direct impact, especially in single-time-point cross-sectional studies. The microbiota composition and function are not static but the result of the interactions between the various taxa in the ecosystem. Thus, further study is needed to disentangle such interactions, which may be possible through longitudinal studies or by sampling a more fine-grained age range at single-time points. Furthermore, integrating metagenomics data with genomic variation, metabolomics, or biomarkers may obtain a clearer picture. Such studies will fulfill the goal of finding therapeutic interventions or optimized dietary patterns that can reestablish and maintain the diversity of the gut microbiota so that we can fully exploit its potential to impact healthy aging positively.

Epilogue

In this thesis, I have discussed the roles that DNA sequencing and data analysis play in identifying and studying variation in different domains. I have elucidated that a frame of reference is a common but crucial theme when making comparisons within and across different samples, across a number of different application fields.

For sequence alignment, I have shown that a more unbiased alignment can be obtained by incorporating population variation directly into the reference genome. In this case the reference represents a collection of genomes instead of what is now the custom: one genome as a fixed reference. In the context of NIPT, I have shown the importance of a reference when detecting chromosomal aberrations from low coverage sequencing data. Moreover, I have shown that the ability to differentiate chromosomal (ab)normality can be improved by incorporating additional data sources, such as for example the fragment length. For the microbiome analyses, I have stressed that cohorts should exhibit notably different characteristic microbial communities, which then can be detected by treating each cohort as a reference against all other cohorts, leading to a useful characterization of the diversity in each cohort.

I aimed to show how the concept and scope of a reference can have a deep impact on the outcome of a molecular analysis. Note that this is a recurrent phenomenon, even beyond the areas covered in this thesis. Consequently, we advocate a conscious consideration and selection of an appropriate frame of reference for an application at hand.

Bibliography

[1] Francis Crick. “Central dogma of molecular biology”. *Nature* 227, pp. 561–563, 1970. (See p. 3).

[2] Sara Goodwin et al. “Coming of age: ten years of next-generation sequencing technologies”. *Nature Reviews Genetics* 17, pp. 333–351, 2016. (See p. 3).

[3] Taishan Hu et al. “Next-generation sequencing technologies: An overview”. *Human Immunology* 82, pp. 801–811, 2021. (See p. 3).

[4] James M Heather and Benjamin Chain. “The sequence of sequencers: The history of sequencing DNA”. *Genomics* 107, pp. 1–8, 2016. (See p. 3).

[5] *Illumina sequencing platforms*. `https://www.illumina.com/systems/sequencing-platforms.html`. Accessed: 4-7-2022 (see p. 3).

[6] *For all you seq*. `https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/for-all-you-seq-single-cell.pdf`. Accessed: 4-7-2022 (see p. 3).

[7] Richard Bellman. “On the theory of dynamic programming”. *Proceedings of the national Academy of Sciences* 38, pp. 716–719, 1952. (See p. 3).

[8] Saul B Needleman and Christian D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. *Journal of molecular biology* 48, pp. 443–453, 1970. (See p. 3).

[9] J Craig Venter et al. “The sequence of the human genome”. *Science* 291, pp. 1304–1351, 2001. (See p. 4).

[10] Eric Lander et al. “Initial sequencing and analysis of the human genome”. *Nature* 409, pp. 860–921, 2001. (See p. 4).

[11] Deanna M Church et al. “Modernizing reference genome assemblies”. *PLoS biology* 9, e1001091, 2011. (See pp. 4, 5).

[12] Sergey Nurk et al. “The complete sequence of a human genome”. *Science* 376, pp. 44–53, 2022. (See p. 4).

[13] Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Available at: `www.genome.gov/sequencingcostsdata`. Accessed: 4-7-2022 (see p. 4).

[14] Valerie A Schneider et al. “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly”. *Genome research* 27, pp. 849–864, 2017. (See pp. 4, 11).

[15] W James Kent et al. “The human genome browser at UCSC”. *Genome research* 12, pp. 996–1006, 2002. (See p. 4).

[16] Deanna M Church. "A next-generation human genome sequence". *Science* 376, pp. 34–35, 2022. (See p. 4).

[17] James W MacDonald et al. "An updated map of GRCh38 linkage disequilibrium blocks based on European ancestry data". *BioRxiv*, 2022. (See p. 4).

[18] Christopher A Miller et al. "Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence". *The Journal of Molecular Diagnostics* 24, pp. 219–223, 2022. (See p. 4).

[19] *One of these things doesn't belong: efforts to exclude problematic sequences in GRCh38.* Available at: `http://genomeref.blogspot.com/2021/07/one-of-these-things-doest-belong.html`. Accessed: 5-7-2022 (see p. 4).

[20] Lisa A Lansdon et al. "Factors affecting migration to GRCh38 in laboratories performing clinical next-generation sequencing". *The Journal of Molecular Diagnostics* 23, pp. 651–657, 2021. (See p. 4).

[21] Mehran Karimzadeh et al. "Umap and Bismap: quantifying genome and methylome mappability". *Nucleic acids research* 46, e120–e120, 2018. (See p. 4).

[22] Jacob F Degner et al. "Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data". *Bioinformatics* 25, pp. 3207–3212, 2009. (See pp. 5, 11).

[23] Débora YC Brandt et al. "Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project Phase I data". *G3: Genes, Genomes, Genetics* 5, pp. 931–941, 2015. (See pp. 5, 11).

[24] Torsten Günther and Carl Nettelblad. "The presence and impact of reference bias on population genomic studies of prehistoric human populations". *PLoS genetics* 15, e1008302, 2019. (See p. 5).

[25] 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing". *Nature* 467, pp. 1061–1073, 2010. (See p. 5).

[26] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". *Nature* 526, pp. 68–74, 2015. (See pp. 5, 13, 17).

[27] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". *Nature* 431, pp. 931–945, 2004. (See p. 5).

[28] *NHGRI-DOE Guidance on Human Subjects Issues in Large-Scale DNA Sequencing.* Available at: `https://web.archive.org/web/20170512175857/https://www.genome.gov/10000921/`. Accessed: 5-7-2022 (see p. 5).

[29] Kraig R Stevenson et al. "Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome". *BMC genomics* 14, pp. 1–13, 2013. (See p. 5).

[30] *Human Genome Overview - Information about the continuing improvement of the human genome.* Available at: `https://www.ncbi.nlm.nih.gov/grc/human`. Accessed: 5-7-2022 (see p. 5).

[31] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". *arXiv preprint arXiv:1303.3997*, 2013. (See pp. 5, 13, 74, 83).

[32] *iBWA - Iterative Burrows-Wheeler Alignment.* Available at: `http://gmt.genome.wustl.edu/packages/ibwa/index.html`. Accessed: 5-7-2022 (see p. 5).

[33] Aleksandr Morgulis and Richa Agarwala. "SRPRISM (Single Read Paired Read Indel Substitution Minimizer): an efficient aligner for assemblies with explicit guarantees". *GigaScience* 9, giaa023, 2020. (See p. 5).

[34] Marten Jäger et al. "Alternate-locus aware variant calling in whole genome sequencing". *Genome Medicine* 8, pp. 1–15, 2016. (See p. 5).

[35] Pavel A Pevzner et al. "An Eulerian path approach to DNA fragment assembly". *Proceedings of the national academy of sciences* 98, pp. 9748–9753, 2001. (See p. 6).

[36] Eugene W Myers. "The fragment assembly string graph". *Bioinformatics* 21, pp. ii79–ii85, 2005. (See pp. 6, 27).

[37] Xiaoqiu Huang. "A contig assembly program based on sensitive detection of fragment overlaps". *Genomics* 14, pp. 18–25, 1992. (See p. 6).

[38] Humberto Carrillo and David Lipman. "The multiple sequence alignment problem in biology". *SIAM journal on applied mathematics* 48, pp. 1073–1082, 1988. (See p. 6).

[39] Daniel R Zerbino and Ewan Birney. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". *Genome research* 18, pp. 821–829, 2008. (See p. 6).

[40] Vladimír Boža et al. "GAML: genome assembly by maximum likelihood". *Algorithms for Molecular Biology* 10, pp. 1–10, 2015. (See p. 6).

[41] Deanna Church et al. "Extending reference assembly models". *Genome biology* 16, p. 13, 2015. (See pp. 6, 11).

[42] Benedict Paten et al. "Genome graphs and the evolution of genome inference". *Genome research* 27, pp. 665–676, 2017. (See pp. 6, 11).

[43] Timo Beller and Enno Ohlebusch. "Efficient construction of a compressed de Bruijn graph for pan-genome analysis". In: *Annual Symposium on Combinatorial Pattern Matching.* Springer. 2015. Pp. 40–51. (See p. 6).

[44] Ross A Lippert. "Space-efficient whole genome comparisons with Burrows–Wheeler transforms". *Journal of computational biology* 12, pp. 407–415, 2005. (See pp. 6, 11).

[45] Tom Mokveld et al. "CHOP: haplotype-aware path indexing in population graphs". *Genome biology* 21, pp. 1–16, 2020. (See p. 9).

[46] Heng Li and Richard Durbin. "Fast and accurate long-read alignment with Burrows-Wheeler transform". *Bioinformatics* 26, pp. 589–595, 2010. (See p. 11).

[47] Alexander Dilthey et al. "Improved genome inference in the MHC using a population reference graph". *Nature genetics* 47, pp. 682–688, 2015. (See p. 11).

[48] Yu Liu et al. "Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing". *BMC genomics* 15, p. 685, 2014. (See p. 11).

[49] Mark A DePristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. *Nature genetics* 43, p. 491, 2011. (See p. 11).

[50] Can Alkan et al. “Personalized copy number and segmental duplication maps using next-generation sequencing”. *Nature genetics* 41, p. 1061, 2009. (See p. 11).

[51] Stephen M Rumble et al. “SHRiMP: accurate mapping of short color-space reads”. *PLoS computational biology* 5, e1000386, 2009. (See p. 11).

[52] Andrei Z Broder. “On the resemblance and containment of documents”. In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171).* IEEE. 1997. Pp. 21–29. (See p. 11).

[53] Victoria Popic and Serafim Batzoglou. “A hybrid cloud read aligner based on MinHash and kmer voting that preserves privacy”. *Nature communications* 8, p. 15311, 2017. (See p. 11).

[54] Paolo Ferragina and Giovanni Manzini. “Opportunistic data structures with applications”. In: *Proceedings 41st Annual Symposium on Foundations of Computer Science.* IEEE. 2000. Pp. 390–398. (See p. 11).

[55] Korbinian Schneeberger et al. “Simultaneous alignment of short reads against multiple genomes”. *Genome Biol* 10, R98, 2009. (See p. 11).

[56] Lin Huang et al. “Short read alignment with populations of genomes”. *Bioinformatics* 29, pp. i361–i370, 2013. (See p. 11).

[57] Ravi Vijaya Satya et al. “A new strategy to reduce allelic bias in RNA-Seq readmapping”. *Nucleic acids research* 40, e127–e127, 2012. (See pp. 12, 13).

[58] H. P. Eggertsson et al. “Graphtyper enables population-scale genotyping using pangenome graphs”. *Nat. Genet.*, 2017. (See p. 12).

[59] Mikko Rautiainen et al. “Bit-parallel sequence-to-graph alignment”. *Bioinformatics*, 2019. (See p. 12).

[60] Erik Garrison et al. “Variation graph toolkit improves read mapping by representing genetic variation in the reference”. *Nature biotechnology*, 2018. (See p. 12).

[61] Daehwan Kim et al. “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. *Nature biotechnology* 37, pp. 907–915, 2019. (See p. 12).

[62] Jouni Sirén. “Indexing variation graphs”. In: *2017 Proceedings of the ninteenth workshop on algorithm engineering and experiments (ALENEX).* SIAM. 2017. Pp. 13–27. (See p. 12).

[63] Christopher Lee et al. “Multiple sequence alignment using partial order graphs”. *Bioinformatics* 18, pp. 452–464, 2002. (See p. 12).

[64] Jouni Sirén et al. “Indexing graphs for path queries with applications in genome research”. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 11, pp. 375–388, 2014. (See p. 12).

[65] Adam M Novak et al. “A graph extension of the positional Burrows–Wheeler transform and its applications”. *Algorithms for Molecular Biology* 12, p. 18, 2017. (See p. 12).

[66] Jouni Sirén et al. "Haplotype-aware graph indexes". *Bioinformatics* 36, pp. 400–407, 2020. (See p. 12).

[67] Richard Durbin. "Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT)". *Bioinformatics* 30, pp. 1266–1272, 2014. (See p. 12).

[68] Mohamed K Gunady et al. "Yanagi: Transcript Segment Library Construction for RNA-Seq Quantification". In: *LIPIcs-Leibniz International Proceedings in Informatics*. Vol. 88. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2017. (See p. 13).

[69] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". *Nature methods* 9, pp. 357–359, 2012. (See p. 13).

[70] Keira A Cohen et al. "Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of Mycobacterium tuberculosis isolates from KwaZulu-Natal". *PLoS medicine* 12, e1001880, 2015. (See p. 14).

[71] Abigail L Manson et al. "Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into emergence and spread of multidrug resistance". *Nature genetics* 49, p. 395, 2017. (See p. 14).

[72] Bruce J Walker et al. "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement". *PloS one* 9, e112963, 2014. (See p. 14).

[73] Ali Ghaffaari and Tobias Marschall. "Fully-sensitive seed finding in sequence graphs using a hybrid index". *Bioinformatics* 35, pp. i81–i89, 2019. (See p. 21).

[74] Jacob Pritt et al. "FORGe: prioritizing variants for graph genomes". *Genome biology* 19, p. 220, 2018. (See pp. 23, 38).

[75] Jasper Linthorst et al. "Scalable multi whole-genome alignment using recursive exact matching". *BioRxiv*, p. 022715, 2015. (See pp. 24, 39).

[76] Chen-Shan Chin et al. "Phased diploid genome assembly with single-molecule real-time sequencing". *Nature methods* 13, p. 1050, 2016. (See p. 24).

[77] Tom Mokveld. *Improving sequence alignment through population graph inference*. 2017. URL: `https://theses.liacs.nl/313` (visited on 03/06/2019) (see p. 24).

[78] Hongseok Tae et al. "Improved variation calling via an iterative backbone remapping and local assembly method for bacterial genomes". *Genomics* 100, pp. 271–276, 2012. (See p. 24).

[79] Stephen F Altschul et al. "Basic local alignment search tool". *Journal of molecular biology* 215, pp. 403–410, 1990. (See p. 37).

[80] Iain Milne et al. "Tablet-next generation sequence assembly visualization". *Bioinformatics* 26, pp. 401–402, 2009. (See pp. 39, 49).

[81] Manuel Holtgrewe. "Mason–a read simulator for second generation sequencing data". *Technical Report FU Berlin*, 2010. (See pp. 37, 75).

[82] Matei Zaharia et al. “Faster and more accurate sequence alignment with SNAP”. *arXiv preprint arXiv:1111.5572*, 2011. (See p. 54).

[83] Song Liu et al. “A fast read alignment method based on seed-and-vote for next generation sequencing”. *BMC bioinformatics* 17, pp. 193–203, 2016. (See p. 54).

[84] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. *Bioinformatics* 34, pp. 3094–3100, 2018. (See p. 54).

[85] Robert Edgar. “URMAP, an ultra-fast read mapper”. *PeerJ* 8, e9338, 2020. (See p. 54).

[86] Mónica Macías et al. “Liquid biopsy: from basic research to clinical practice”. *Advances in clinical chemistry* 83, pp. 73–119, 2018. (See p. 59).

[87] YM Dennis Lo. “Noninvasive prenatal testing: Advancing through a virtuous circle of science, technology and clinical applications”. *Prenatal Diagnosis* 41, pp. 1190–1192, 2021. (See p. 59).

[88] Milton C Weinstein. “The costs of prevention”. *Journal of General Internal Medicine* 5, S89–S92, 1990. (See p. 59).

[89] Frederick P Rivara et al. “Injury prevention”. *New England journal of medicine* 337, pp. 613–618, 1997. (See p. 59).

[90] Katarzyna Gajewska-Knapik and Lawrence Impey. “Congenital lung lesions: Prenatal diagnosis and intervention”. In: *Seminars in pediatric surgery*. Vol. 24. 4. Elsevier. 2015. Pp. 156–159. (See p. 59).

[91] Lavida RK Brooks and George I Mias. “Streptococcus pneumoniae’s virulence and host immunity: aging, diagnostics, and prevention”. *Frontiers in immunology* 9, p. 1366, 2018. (See p. 59).

[92] Wolfram Miekisch et al. “Analysis of volatile disease markers in blood”. *Clinical chemistry* 47, pp. 1053–1060, 2001. (See p. 59).

[93] Rainer Bischoff and Theo M Luider. “Methodological advances in the discovery of protein and peptide disease markers”. *Journal of Chromatography B* 803, pp. 27–40, 2004. (See p. 59).

[94] Lester S King. “What is disease?” *Philosophy of Science* 21, pp. 193–203, 1954. (See p. 59).

[95] Christopher Boorse. “Health as a theoretical concept”. *Philosophy of science* 44, pp. 542–573, 1977. (See p. 59).

[96] Martin Bunzl. “Comment on" health as a theoretical concept"”. *Philosophy of Science* 47, pp. 116–118, 1980. (See p. 59).

[97] Peter W Hamilton et al. “Do we see what we think we see? The complexities of morphological assessment”. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 218, pp. 285–291, 2009. (See p. 60).

[98] LJ Salomon et al. “ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth”. *Ultrasound in obstetrics & gynecology* 53, pp. 715–723, 2019. (See pp. 60, 81).

[99] Marshal F Folstein et al. ““Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician”. *Journal of psychiatric research* 12, pp. 189–198, 1975. (See p. 60).

[100] Brandon W Alleman et al. “A proposed method to predict preterm birth using clinical data, standard maternal serum screening, and cholesterol”. *American journal of obstetrics and gynecology* 208, 472–e1, 2013. (See pp. 60, 81).

[101] Huilin Wang et al. “Low-pass genome sequencing versus chromosomal microarray analysis: implementation in prenatal diagnosis”. *Genetics in Medicine* 22, pp. 500–510, 2020. (See pp. 60, 81).

[102] Robert A Smith et al. “Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening”. *CA: a cancer journal for clinicians* 69, pp. 184–210, 2019. (See p. 60).

[103] Jean-Marie Ekoe et al. “Screening for diabetes in adults”. *Canadian journal of diabetes* 42, S16–S19, 2018. (See p. 60).

[104] Peter K Panegyres et al. “Early dementia screening”. *Diagnostics* 6, p. 6, 2016. (See p. 60).

[105] Kathy L MacLaughlin et al. “Trends over time in Pap and Pap-HPV cotesting for cervical cancer screening”. *Journal of women’s health* 28, pp. 244–249, 2019. (See p. 60).

[106] Dekkers EHBM et al. *Uitvoeringstoets uitbreiding neonatale hielprikscreening*. 2017. doi: 10.21945/RIVM-2017-0041. url: http://hdl.handle.net/10029/620889 (see p. 60).

[107] L Nshimyumukiza et al. “Cell-free DNA noninvasive prenatal screening for aneuploidy versus conventional screening: a systematic review of economic evaluations”. *Clinical Genetics* 94, pp. 3–21, 2018. (See pp. 60, 81).

[108] Claude Stoll et al. “Impact of prenatal diagnosis on livebirth prevalence of children with congenital anomalies”. In: *Annales de genetique*. Vol. 45. 3. Elsevier. 2002. Pp. 115–121. (See p. 60).

[109] Tarun Jain et al. “30 years of data: impact of the United States in vitro fertilization data registry on advancing fertility care”. *Fertility and sterility* 111, pp. 477–488, 2019. (See p. 60).

[110] Paula Amato et al. “Three-parent in vitro fertilization: gene replacement for the prevention of inherited mitochondrial diseases”. *Fertility and sterility* 101, pp. 31–35, 2014. (See p. 60).

[111] Don P Wolf et al. “Mitochondrial replacement therapy in reproductive medicine”. *Trends in molecular medicine* 21, pp. 68–76, 2015. (See p. 60).

[112] Claire Colmant et al. “Non-invasive prenatal testing for fetal sex determination: is ultrasound still relevant?” *European Journal of Obstetrics & Gynecology and Reproductive Biology* 171, pp. 197–204, 2013. (See pp. 60, 81).

[113] Kypros H Nicolaides. “Nuchal translucency and other first-trimester sonographic markers of chromosomal abnormalities”. *American journal of obstetrics and gynecology* 191, pp. 45–67, 2004. (See p. 60).

[114] JD Sonek et al. “Nasal bone assessment in prenatal screening for trisomy 21”. *American journal of Obstetrics and Gynecology* 195, pp. 1219–1230, 2006. (See p. 60).

[115] Nicholas J Wald et al. “Maternal serum screening for Down’s syndrome in early pregnancy.” *British medical journal* 297, pp. 883–887, 1988. (See p. 60).

[116] Aida Catic et al. “Application of Neural Networks for classification of Patau, Edwards, Down, Turner and Klinefelter Syndrome based on first trimester maternal serum screening data, ultrasonographic findings and patient demographics”. *BMC medical genomics* 11, pp. 1–12, 2018. (See pp. 60, 81).

[117] Wybo Dondorp et al. “Non-invasive prenatal testing for aneuploidy and beyond: challenges of responsible innovation in prenatal screening”. *European Journal of Human Genetics* 23, pp. 1438–1450, 2015. (See pp. 60, 61, 81).

[118] R Saura et al. “Evaluation of chorion villus sampling”. *The Lancet* 338, pp. 449–450, 1991. (See pp. 60, 81).

[119] R Saura et al. “Early amniocentesis versus chorionic villus sampling for fetal karyotyping”. *The Lancet* 344, pp. 825–826, 1994. (See pp. 60, 61, 81).

[120] Hans-Hilger Ropers. “On the future of genetic risk assessment”. *Journal of community genetics* 3, pp. 229–236, 2012. (See p. 60).

[121] Peng Yue and John Moult. “Identification and analysis of deleterious human SNPs”. *Journal of molecular biology* 356, pp. 1263–1274, 2006. (See p. 61).

[122] Xinjun Zhang et al. “Impact of human pathogenic micro-insertions and micro-deletions on post-transcriptional regulation”. *Human molecular genetics* 23, pp. 3024–3034, 2014. (See p. 61).

[123] Fady M Mikhail. “Copy number variations and human genetic disease”. *Current opinion in pediatrics* 26, pp. 646–652, 2014. (See p. 61).

[124] David Patterson. “Molecular genetic analysis of Down syndrome”. *Human genetics* 126, pp. 195–214, 2009. (See p. 61).

[125] C Baccichetti et al. “Terminal deletion of the short arm of chromosome 5”. *Clinical genetics* 34, pp. 219–223, 1988. (See p. 61).

[126] Lars Feuk. “Inversion variants in the human genome: role in disease and genome architecture”. *Genome medicine* 2, pp. 1–8, 2010. (See p. 61).

[127] Delong Liu et al. “t (8; 14; 18): a 3-way chromosome translocation in two patients with Burkitt’s lymphoma/leukemia”. *Molecular Cancer* 6, pp. 1–5, 2007. (See p. 61).

[128] Karen Usdin. “The biological effects of simple tandem repeats: lessons from the repeat expansion diseases”. *Genome research* 18, pp. 1011–1019, 2008. (See p. 61).

[129] Iyoko Katoh and Shun-ichi Kurata. “Association of endogenous retroviruses and long terminal repeats with human disorders”. *Frontiers in oncology* 3, p. 234, 2013. (See p. 61).

[130] Masood A Shammas. “Repetitive sequences, genomic instability, and Barrett’s esophageal adenocarcinoma”. *Mobile Genetic Elements* 1, pp. 208–212, 2011. (See p. 61).

[131] C O'Connor. *Karyotyping for Chromosomal Abnormalities.* Available at: https://web.archive.org/web/20220324195245/https://www.nature.com/scitable/topicpage/karyotyping-for-chromosomal-abnormalities-298/. Accessed: 18-7-2022 (see p. 61).

[132] Michael L Metzker. "Sequencing technologies—the next generation". *Nature reviews genetics* 11, pp. 31–46, 2010. (See p. 61).

[133] Xiao Yang et al. "A survey of error-correction methods for next-generation sequencing". *Briefings in bioinformatics* 14, pp. 56–66, 2013. (See p. 61).

[134] P Mandel. "Les acides nucleiques du plasma sanguin chez 1 homme". *CR Seances Soc Biol Fil* 142, pp. 241–243, 1948. (See p. 61).

[135] Maniesh van der Vaart and Piet J Pretorius. "The origin of circulating free DNA". *Clinical chemistry* 53, pp. 2215–2215, 2007. (See p. 61).

[136] Peter Celec et al. "Cell-free DNA: the role in pathophysiology and as a biomarker in kidney diseases". *Expert reviews in molecular medicine* 20, 2018. (See p. 61).

[137] Vanessa Garcia Moreira et al. "Increase in and clearance of cell-free plasma DNA in hemodialysis quantified by real-time PCR". *Clinical Chemistry and Laboratory Medicine (CCLM)* 44, pp. 1410–1415, 2006. (See p. 61).

[138] Zhen Qin et al. "Cell-free circulating tumor DNA in cancer". *Chinese journal of cancer* 35, pp. 1–9, 2016. (See pp. 61, 88).

[139] Robert Stawski et al. "Cell-free DNA: potential application in COVID-19 diagnostics and management". *Viruses* 14, p. 321, 2022. (See p. 61).

[140] Simon Robert Knight et al. "Donor-specific cell-free DNA as a biomarker in solid organ transplantation. A systematic review". *Transplantation* 103, pp. 273–283, 2019. (See p. 61).

[141] Sara L Rassoulian Barrett et al. "Cell free DNA from respiratory pathogens is detectable in the blood plasma of cystic fibrosis patients". *Scientific reports* 10, pp. 1–6, 2020. (See p. 61).

[142] Markus Woegerbauer et al. "Cell-free DNA: an underestimated source of antibiotic resistance gene dissemination at the interface between human activities and downstream environments in the context of wastewater reuse". *Frontiers in Microbiology* 11, p. 671, 2020. (See p. 61).

[143] M Alberry et al. "Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast". *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis* 27, pp. 415–418, 2007. (See p. 61).

[144] April L Hall et al. "Positive cell-free fetal DNA testing for trisomy 13 reveals confined placental mosaicism". *Genetics in Medicine* 15, pp. 729–732, 2013. (See p. 61).

[145] Lisa Hui and Diana W Bianchi. "Fetal fraction and noninvasive prenatal testing: What clinicians need to know". *Prenatal Diagnosis* 40, pp. 155–163, 2020. (See p. 62).

[146] Tachjaree Panchalee et al. “The effect of maternal body mass index and gestational age on circulating trophoblast yield in cell-based noninvasive prenatal testing”. *Prenatal diagnosis* 40, pp. 1383–1389, 2020. (See p. 62).

[147] George Attilakos et al. “Quantification of free fetal DNA in multiple pregnancies and relationship with chorionicity”. *Prenatal diagnosis* 31, pp. 967–972, 2011. (See p. 62).

[148] Rossa WK Chiu et al. “Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma”. *Proceedings of the National Academy of Sciences* 105, pp. 20458–20463, 2008. (See pp. 62, 65, 81).

[149] William S Cleveland. “Robust locally weighted regression and smoothing scatterplots”. *Journal of the American statistical association* 74, pp. 829–836, 1979. (See p. 62).

[150] Roy Straver et al. “WISECONDOR: detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme”. *Nucleic acids research* 42, e31–e31, 2014. (See pp. 62, 65, 82, 83, 85).

[151] Kitty K Lo et al. “RAPIDR: an analysis package for non-invasive prenatal testing of aneuploidy”. *Bioinformatics* 30, pp. 2965–2967, 2014. (See p. 62).

[152] Lennart Raman et al. “WisecondorX: improved copy number detection for routine shallow whole-genome sequencing”. *Nucleic acids research* 47, pp. 1605–1614, 2019. (See pp. 62, 65, 66, 82–85).

[153] Anders OH Nygren et al. “Quantification of fetal DNA by use of methylation-based DNA discrimination”. *Clinical Chemistry* 56, pp. 1627–1635, 2010. (See p. 62).

[154] Peiyong Jiang et al. “FetalQuantSD: Accurate quantification of fetal DNA fraction by shallow-depth sequencing of maternal plasma DNA”. *NPJ genomic medicine* 1, pp. 1–7, 2016. (See pp. 62, 82).

[155] Amin R Mazloom et al. “Noninvasive prenatal detection of sex chromosomal aneuploidies by sequencing circulating cell-free DNA from maternal plasma”. *Prenatal diagnosis* 33, pp. 591–597, 2013. (See p. 62).

[156] Sung K Kim et al. “Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts”. *Prenatal diagnosis* 35, pp. 810–815, 2015. (See pp. 62, 74, 82, 84).

[157] KC Allen Chan et al. “Size distributions of maternal and fetal DNA in maternal plasma”. *Clinical chemistry* 50, pp. 88–92, 2004. (See pp. 62, 73).

[158] Roy Straver et al. “Calculating the fetal fraction for noninvasive prenatal testing based on genome-wide nucleosome profiles”. *Prenatal diagnosis* 36, pp. 614–621, 2016. (See pp. 62, 82).

[159] Stephanie CY Yu et al. “Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing”. *Proceedings of the National Academy of Sciences* 111, pp. 8583–8588, 2014. (See pp. 62, 73).

[160] YM Dennis Lo et al. “Presence of fetal DNA in maternal plasma and serum”. *The lancet* 350, pp. 485–487, 1997. (See pp. 65, 81).

[161] Amy J Sehnert et al. “Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood”. *Clinical chemistry* 57, pp. 1042–1049, 2011. (See pp. 65, 81).

[162] Peter Benn et al. “Prenatal Detection of Down Syndrome using Massively Parallel Sequencing (MPS): a rapid response statement from a committee on behalf of the Board of the International Society for Prenatal Diagnosis, 24 October 2011”. *Prenatal diagnosis* 32, pp. 1–2, 2012. (See p. 65).

[163] P Benn et al. “Non-invasive prenatal testing for aneuploidy: current status and future prospects”. *Ultrasound in Obstetrics & Gynecology* 42, pp. 15–33, 2013. (See pp. 65, 81).

[164] Diane Van Opstal et al. “False negative NIPT results: risk figures for chromosomes 13, 18 and 21 based on chorionic villi results in 5967 cases and literature review”. *PLoS One* 11, e0146794, 2016. (See pp. 65, 81).

[165] Dick Oepkes et al. “Trial by Dutch laboratories for evaluation of non-invasive prenatal testing. Part I—clinical impact”. *Prenatal Diagnosis* 36, pp. 1083–1090, 2016. (See p. 65).

[166] Shaowei Wang et al. “Maternal X chromosome copy number variations are associated with discordant fetal sex chromosome aneuploidies detected by noninvasive prenatal testing”. *Clinica Chimica Acta* 444, pp. 113–116, 2015. (See p. 65).

[167] Paul Brady et al. “Clinical implementation of NIPT–technical and biological challenges”. *Clinical Genetics* 89, pp. 523–530, 2016. (See p. 65).

[168] David Peters et al. “Noninvasive prenatal diagnosis of a fetal microdeletion syndrome”. *New England Journal of Medicine* 365, pp. 1847–1848, 2011. (See p. 65).

[169] Taylor J Jensen et al. “Detection of microdeletion 22q11. 2 in a fetus by next-generation sequencing of maternal plasma”. *Clinical chemistry* 58, pp. 1148–1151, 2012. (See pp. 65, 81).

[170] Yifang Jia et al. “Genetic effects of a 13q31. 1 microdeletion detected by noninvasive prenatal testing (NIPT)”. *International Journal of Clinical and Experimental Pathology* 7, p. 7003, 2014. (See pp. 65, 81).

[171] Henna V Advani et al. “Challenges in non-invasive prenatal screening for sub-chromosomal copy number variations using cell-free DNA”. *Prenatal diagnosis* 37, pp. 1067–1075, 2017. (See p. 65).

[172] Tanja Schlaikjær Hartwig et al. “Discordant non-invasive prenatal testing (NIPT)–a systematic review”. *Prenatal diagnosis* 37, pp. 527–539, 2017. (See p. 65).

[173] TK Lau et al. “Non-invasive prenatal testing for fetal chromosomal abnormalities by low-coverage whole-genome sequencing of maternal plasma DNA: review of 1982 consecutive cases in a single center”. *Ultrasound in Obstetrics & Gynecology* 43, pp. 254–264, 2014. (See p. 65).

[174] Hongtai Liu et al. “Performance evaluation of NIPT in detection of chromosomal copy number variants using low-coverage whole-genome sequencing of plasma DNA”. *PLoS One* 11, e0159233, 2016. (See pp. 65, 81).

[175] Maurizio Ferrari et al. “New trend in non-invasive prenatal diagnosis”. *Clinica chimica acta* 451, pp. 9–13, 2015. (See p. 65).

[176] Fergus Perry Scott et al. “Factors affecting cell-free DNA fetal fraction and the consequences for test accuracy”. *The Journal of Maternal-Fetal & Neonatal Medicine* 31, pp. 1865–1872, 2018. (See p. 65).

[177] Jacob A Canick et al. “The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies”. *Prenatal diagnosis* 33, pp. 667–674, 2013. (See pp. 65, 81).

[178] Hongyun Zhang et al. “Non-invasive prenatal testing for trisomies 21, 18 and 13: clinical experience from 146 958 pregnancies”. *Ultrasound in Obstetrics & Gynecology* 45, pp. 530–538, 2015. (See pp. 65, 81).

[179] Yan Du et al. “Detection of chromosome abnormalities using current noninvasive prenatal testing: A multi-center comparative study”. *BioScience Trends*, 2018. (See p. 65).

[180] Dongyi Yu et al. “Noninvasive prenatal testing for fetal subchromosomal copy number variations and chromosomal aneuploidy by low-pass whole-genome sequencing”. *Molecular genetics & genomic medicine* 7, e674, 2019. (See p. 65).

[181] Yuqin Luo et al. “Pilot study of a novel multi-functional noninvasive prenatal test on fetus aneuploidy, copy number variation, and single-gene disorder screening”. *Molecular genetics & genomic medicine* 7, e00597, 2019. (See p. 65).

[182] Kun Sun et al. “Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing”. *Proceedings of the National Academy of Sciences* 115, E5106–E5114, 2018. (See p. 73).

[183] Yunyun An et al. “DNA methylation analysis explores the molecular basis of plasma cell-free DNA fragmentation”. *Nature Communications* 14, p. 287, 2023. (See p. 73).

[184] KC Allen Chan et al. “Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends”. *Proceedings of the National Academy of Sciences* 113, E8159–E8168, 2016. (See p. 73).

[185] Kun Sun et al. “COFFEE: control-free noninvasive fetal chromosomal examination using maternal plasma DNA”. *Prenatal diagnosis* 37, pp. 336–340, 2017. (See p. 73).

[186] Tom Mokveld et al. “WisecondorFF: Improved fetal aneuploidy detection from shallow WGS through fragment length analysis”. *Diagnostics* 12, p. 59, 2022. (See p. 73).

[187] Karuna RM van der Meij et al. “TRIDENT-2: national implementation of genome-wide non-invasive prenatal testing as a first-tier screening test in the Netherlands”. *The American Journal of Human Genetics* 105, pp. 1091–1101, 2019. (See pp. 73, 82).

[188] Tom Mokveld et al. “WisecondorFF: Improved Fetal Aneuploidy Detection from Shallow WGS through Fragment Length Analysis”. *Diagnostics* 12, p. 59, 2021. (See p. 79).

[189] Csaba Papp and Zoltán Papp. "Chorionic villus sampling and amniocentesis: what are the risks in current practice?" *Current Opinion in Obstetrics and Gynecology* 15, pp. 159–165, 2003. (See p. 81).

[190] Anjali J Kaimal et al. "Prenatal testing in the genomic age: clinical outcomes, quality of life, and costs". *Obstetrics & Gynecology* 126, pp. 737–746, 2015. (See p. 81).

[191] Sebastian Larion et al. "Association of combined first-trimester screen and noninvasive prenatal testing on diagnostic procedures". *Obstetrics & Gynecology* 123, pp. 1303–1310, 2014. (See p. 81).

[192] Andrew McLennan et al. "Noninvasive prenatal testing in routine clinical practice–an audit of NIPT and combined first-trimester screening in an unselected Australian population". *Australian and New Zealand Journal of Obstetrics and Gynaecology* 56, pp. 22–28, 2016. (See p. 81).

[193] Maria C Neofytou et al. "Targeted capture enrichment assay for non-invasive prenatal testing of large and small size sub-chromosomal deletions and duplications". *PLoS One* 12, e0171319, 2017. (See p. 81).

[194] Hannah Skrzypek and Lisa Hui. "Noninvasive prenatal testing for fetal aneuploidy and single gene disorders". *Best Practice & Research Clinical Obstetrics & Gynaecology* 42, pp. 26–38, 2017. (See p. 81).

[195] G Ashoor et al. "Fetal fraction in maternal plasma cell-free DNA at 11–13 weeks' gestation: relation to maternal and fetal characteristics". *Ultrasound in Obstetrics & Gynecology* 41, pp. 26–32, 2013. (See p. 81).

[196] Maayke A de Koning et al. "From diagnostic yield to clinical impact: a pilot study on the implementation of prenatal exome sequencing in routine care". *Genetics in Medicine* 21, pp. 2303–2310, 2019. (See p. 81).

[197] Anne Mardy and Ronald J Wapner. "Confined placental mosaicism and its impact on confirmation of NIPT results". In: *American Journal of Medical Genetics Part C: Seminars in Medical Genetics.* Vol. 172. 2. Wiley Online Library. 2016. Pp. 118–122. (See p. 81).

[198] Xiya Zhou et al. "Contribution of maternal copy number variations to false-positive fetal trisomies detected by noninvasive prenatal testing". *Prenatal diagnosis* 37, pp. 318–322, 2017. (See p. 81).

[199] Saskia Tamminga et al. "Maternal plasma DNA and RNA sequencing for prenatal testing". *Advances in clinical chemistry* 74, pp. 63–102, 2016. (See p. 81).

[200] Yuqian Xiang et al. "DNA methylome profiling of maternal peripheral blood and placentas reveal potential fetal DNA markers for non-invasive prenatal testing". *Molecular human reproduction* 20, pp. 875–884, 2014. (See p. 81).

[201] K Martin et al. "Clinical experience with a single-nucleotide polymorphism-based non-invasive prenatal test for five clinically significant microdeletions". *Clinical genetics* 93, pp. 293–300, 2018. (See p. 81).

[202] Carlo Vermeulen et al. "Sensitive monogenic noninvasive prenatal diagnosis by targeted haplotyping". *The American Journal of Human Genetics* 101, pp. 326–339, 2017. (See pp. 81, 82).

[203] Min Zhao et al. “Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives”. *BMC bioinformatics* 14, pp. 1–16, 2013. (See p. 81).

[204] Xianlu Laura Peng and Peiyong Jiang. “Bioinformatics approaches for fetal DNA fraction estimation in noninvasive prenatal testing”. *International journal of molecular sciences* 18, p. 453, 2017. (See p. 82).

[205] Adam B Olshen et al. “Circular binary segmentation for the analysis of array-based DNA copy number data”. *Biostatistics* 5, pp. 557–572, 2004. (See p. 84).

[206] Eric Talevich et al. “CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing”. *PLoS computational biology* 12, e1004873, 2016. (See p. 85).

[207] Ruth E Ley et al. “Worlds within worlds: evolution of the vertebrate gut microbiota”. *Nature Reviews Microbiology* 6, pp. 776–788, 2008. (See p. 103).

[208] Howard Ochman et al. “Evolutionary relationships of wild hominids recapitulated by gut microbial communities”. *PLoS biology* 8, e1000546, 2010. (See p. 103).

[209] Andrew K Benson et al. “Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors”. *Proceedings of the National Academy of Sciences* 107, pp. 18933–18938, 2010. (See p. 103).

[210] William R Wikoff et al. “Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites”. *Proceedings of the national academy of sciences* 106, pp. 3698–3703, 2009. (See p. 103).

[211] Ron Sender et al. “Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans”. *Cell* 164, pp. 337–340, 2016. (See p. 103).

[212] Alexandre Almeida et al. “A new genomic blueprint of the human gut microbiota”. *Nature* 568, pp. 499–504, 2019. (See p. 103).

[213] Tarini Shankar Ghosh et al. “The gut microbiome as a modulator of healthy ageing”. *Nature Reviews Gastroenterology & Hepatology*, pp. 1–20, 2022. (See pp. 103, 104, 111, 121).

[214] Nobuhiko Kamada et al. “Control of pathogens and pathobionts by the gut microbiota”. *Nature immunology* 14, pp. 685–690, 2013. (See p. 103).

[215] Nobuhiko Kamada and Gabriel Núñez. “Regulation of the immune system by the resident intestinal bacteria”. *Gastroenterology* 146, pp. 1477–1488, 2014. (See p. 103).

[216] Atsushi Nishida et al. “Gut microbiota in the pathogenesis of inflammatory bowel disease”. *Clinical journal of gastroenterology* 11, pp. 1–10, 2018. (See pp. 103, 111).

[217] Eamonn MM Quigley. “Microbiota-brain-gut axis and neurodegenerative diseases”. *Current neurology and neuroscience reports* 17, pp. 1–9, 2017. (See pp. 103, 111).

[218] John P Haran et al. “Alzheimer’s disease microbiome is associated with dysregulation of the anti-inflammatory P-glycoprotein pathway”. *MBio* 10, e00632–19, 2019. (See pp. 103, 111).

[219] Fergus Shanahan et al. “The healthy microbiome—what is the definition of a healthy gut microbiome?” *Gastroenterology* 160, pp. 483–494, 2021. (See p. 103).

[220] Lianmin Chen et al. “The long-term genetic stability and individual specificity of the human gut microbiome”. *Cell* 184, pp. 2302–2315, 2021. (See pp. 103, 111).

[221] Joby Pulikkan et al. “Role of the gut microbiome in autism spectrum disorders”. *Reviews on Biomarker Studies in Psychiatric and Neurodegenerative Disorders*, pp. 253–269, 2019. (See p. 103).

[222] F De Luca and Y Shoenfeld. “The microbiome in autoimmune diseases”. *Clinical & Experimental Immunology* 195, pp. 74–85, 2019. (See p. 103).

[223] Cristina Menni et al. “Gut microbiome diversity and high-fibre intake are related to lower long-term weight gain”. *International journal of obesity* 41, pp. 1099–1105, 2017. (See p. 103).

[224] Carlos López-Otín et al. “The hallmarks of aging”. *Cell* 153, pp. 1194–1217, 2013. (See pp. 104, 111).

[225] Juan José Carmona and Shaday Michan. “Biology of healthy aging and longevity”. *Revista de investigacion clinica* 68, pp. 7–16, 2016. (See p. 104).

[226] Angela R Brooks-Wilson. “Genetics of healthy aging and longevity”. *Human genetics* 132, pp. 1323–1338, 2013. (See p. 104).

[227] Philippa Clarke and Els R Nieuwenhuijsen. “Environments for healthy ageing: A critical review”. *Maturitas* 64, pp. 14–19, 2009. (See p. 104).

[228] Hélio José Coelho-Júnior et al. “Cross-sectional and longitudinal associations between adherence to Mediterranean diet with physical performance and cognitive function in older adults: A systematic review and meta-analysis”. *Ageing research reviews* 70, p. 101395, 2021. (See pp. 104, 111).

[229] Gregory D Cartee et al. “Exercise promotes healthy aging of skeletal muscle”. *Cell metabolism* 23, pp. 1034–1047, 2016. (See p. 104).

[230] Katie E Cherry et al. “Social factors and healthy aging: findings from the Louisiana healthy aging study (LHAS)”. *Kinesiology review* 5, pp. 50–56, 2016. (See p. 104).

[231] Paul W O’Toole and Marcus J Claesson. “Gut microbiota: changes throughout the lifespan from infancy to elderly”. *International Dairy Journal* 20, pp. 281–291, 2010. (See p. 104).

[232] Chana Palmer et al. “Development of the human infant intestinal microbiota”. *PLoS biology* 5, e177, 2007. (See p. 104).

[233] Jeremy E Koenig et al. “Succession of microbial consortia in the developing infant gut microbiome”. *Proceedings of the National Academy of Sciences* 108, pp. 4578–4585, 2011. (See p. 104).

[234] Valeria D’Argenio and Francesco Salvatore. “The role of the gut microbiome in the healthy adult status”. *Clinica chimica acta* 451, pp. 97–102, 2015. (See p. 104).

[235] Raaj S Mehta et al. “Stability of the human faecal microbiome in a cohort of adult men”. *Nature microbiology* 3, pp. 347–355, 2018. (See p. 104).

[236] Zhe Luan et al. “Metagenomics study reveals changes in gut microbiota in centenarians: a cohort study of hainan centenarians”. *Frontiers in microbiology* 11, p. 1474, 2020. (See pp. 104, 111, 120, 121).

[237] Harry J Flint et al. “The role of the gut microbiota in nutrition and health”. *Nature reviews Gastroenterology & hepatology* 9, pp. 577–589, 2012. (See pp. 104, 111).

[238] Harry Sokol et al. “Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients”. *Proceedings of the National Academy of Sciences* 105, pp. 16731–16736, 2008. (See p. 104).

[239] Andrew J Tarr et al. “The prebiotics 3’ Sialyllactose and 6’ Sialyllactose diminish stressor-induced anxiety-like behavior and colonic microbiota alterations: Evidence for effects on the gut–brain axis”. *Brain, behavior, and immunity* 50, pp. 166–177, 2015. (See p. 104).

[240] Omar Mossad et al. “Microbiota-dependent increase in δ-valerobetaine alters neuronal function and is responsible for age-related cognitive decline”. *Nature Aging* 1, pp. 1127–1136, 2021. (See pp. 104, 111).

[241] Huiying Wang et al. “Bifidobacterium longum 1714™ strain modulates brain activity of healthy volunteers during social stress”. *The American journal of gastroenterology* 114, p. 1152, 2019. (See p. 104).

[242] Micah Hamady and Rob Knight. “Microbial community profiling for human microbiome projects: tools, techniques, and challenges”. *Genome research* 19, pp. 1141–1152, 2009. (See p. 104).

[243] Ravi Ranjan et al. “Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing”. *Biochemical and biophysical research communications* 469, pp. 967–977, 2016. (See p. 104).

[244] Chengwei Luo et al. “ConStrains identifies microbial strains in metagenomic datasets”. *Nature biotechnology* 33, pp. 1045–1052, 2015. (See p. 104).

[245] Patrick D Schloss et al. “Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies”. *PloS one* 6, e27310, 2011. (See p. 105).

[246] Benjamin Hillmann et al. “Evaluating the information content of shallow shotgun metagenomics”. *Msystems* 3, e00069–18, 2018. (See p. 105).

[247] Jethro S Johnson et al. “Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis”. *Nature communications* 10, pp. 1–11, 2019. (See p. 105).

[248] Jerald Conrad Ibal et al. “Information about variations in multiple copies of bacterial 16S rRNA genes may aid in species identification”. *PLoS One* 14, e0212090, 2019. (See p. 105).

[249] Pablo Yarza et al. “Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences”. *Nature Reviews Microbiology* 12, pp. 635–645, 2014. (See p. 105).

[250] Patrick D Schloss and Jo Handelsman. “Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness”. *Applied and environmental microbiology* 71, pp. 1501–1506, 2005. (See p. 105).

[251] Amnon Amir et al. “Deblur rapidly resolves single-nucleotide community sequence patterns”. *MSystems* 2, e00191–16, 2017. (See p. 105).

[252] Mikhail Tikhonov et al. “Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution”. *The ISME journal* 9, pp. 68–80, 2015. (See p. 105).

[253] Benjamin J Callahan et al. “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis”. *The ISME journal* 11, pp. 2639–2643, 2017. (See p. 105).

[254] Dong-Lei Sun et al. “Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity”. *Applied and environmental microbiology* 79, pp. 5962–5969, 2013. (See p. 105).

[255] Cristian Del Fabbro et al. “An extensive evaluation of read trimming effects on Illumina NGS data analysis”. *PloS one* 8, e85024, 2013. (See p. 105).

[256] Xiaofan Zhou and Antonis Rokas. “Prevention, diagnosis and treatment of high-throughput sequencing data pathologies”. *Molecular Ecology* 23, pp. 1679–1700, 2014. (See p. 105).

[257] Michael J Rosen et al. “Denoising PCR-amplified metagenome data”. *BMC bioinformatics* 13, pp. 1–16, 2012. (See p. 105).

[258] A Murat Eren et al. “Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data”. *Methods in ecology and evolution* 4, pp. 1111–1119, 2013. (See p. 105).

[259] A Murat Eren et al. “Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences”. *The ISME journal* 9, pp. 968–979, 2015. (See p. 105).

[260] Benjamin J Callahan et al. “DADA2: High-resolution sample inference from Illumina amplicon data”. *Nature methods* 13, pp. 581–583, 2016. (See pp. 105, 113).

[261] Robert C Edgar. “UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing”. *BioRxiv*, p. 081257, 2016. (See p. 105).

[262] Robert C Edgar. “UCHIME2: improved chimera prediction for amplicon sequencing”. *BioRxiv*, p. 074252, 2016. (See p. 105).

[263] Robert C Edgar and Henrik Flyvbjerg. “Error filtering, pair assembly and error correction for next-generation sequencing reads”. *Bioinformatics* 31, pp. 3476–3482, 2015. (See p. 105).

[264] Qiong Wang et al. “Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy”. *Applied and environmental microbiology* 73, pp. 5261–5267, 2007. (See p. 105).

[265] Alicia Oshlack and Matthew J Wakefield. “Transcript length bias in RNA-seq data confounds systems biology”. *Biology direct* 4, pp. 1–10, 2009. (See p. 105).

[266] Mark D Robinson and Alicia Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data”. *Genome biology* 11, pp. 1–9, 2010. (See p. 105).

[267] M Senthil Kumar et al. “Analysis and correction of compositional bias in sparse sequencing count data”. *BMC genomics* 19, pp. 1–23, 2018. (See p. 105).

[268] Jolinda Pollock et al. “The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies”. *Applied and environmental microbiology* 84, e02627–17, 2018. (See p. 105).

[269] Gemma Henderson et al. “Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities”. *PloS one* 8, e74787, 2013. (See p. 105).

[270] Perrine Cruaud et al. “Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems”. *Applied and environmental microbiology* 80, pp. 4626–4639, 2014. (See p. 105).

[271] Alan W Walker et al. “16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice”. *Microbiome* 3, pp. 1–11, 2015. (See p. 105).

[272] Daryl M Gohl et al. “Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies”. *Nature biotechnology* 34, pp. 942–949, 2016. (See p. 105).

[273] Andreas Hiergeist et al. “Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability”. *International Journal of Medical Microbiology* 306, pp. 334–342, 2016. (See p. 105).

[274] Andrew R Solow and Stephen Polasky. “Measuring biological diversity”. *Environmental and Ecological Statistics* 1, pp. 95–103, 1994. (See pp. 106, 114).

[275] Benjamin Smith and J Bastow Wilson. “A consumer’s guide to evenness indices”. *Oikos*, pp. 70–82, 1996. (See pp. 106, 114).

[276] Claude Elwood Shannon. “A mathematical theory of communication”. *The Bell system technical journal* 27, pp. 379–423, 1948. (See pp. 106, 114).

[277] Daniel P Faith. “Conservation evaluation and phylogenetic diversity”. *Biological conservation* 61, pp. 1–10, 1992. (See p. 106).

[278] J Roger Bray and John T Curtis. “An ordination of the upland forest communities of southern Wisconsin”. *Ecological monographs* 27, pp. 326–349, 1957. (See p. 106).

[279] Catherine Lozupone and Rob Knight. “UniFrac: a new phylogenetic method for comparing microbial communities”. *Applied and environmental microbiology* 71, pp. 8228–8235, 2005. (See pp. 106, 114).

[280] *UniFrac*. Available at: `https://web.archive.org/web/20090207102537/http://bmf2.colorado.edu/unifrac/index.psp/`. Accessed: 17-7-2022 (see pp. 106, 114).

[281] Catherine A Lozupone et al. “Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities”. *Applied and environmental microbiology* 73, pp. 1576–1585, 2007. (See p. 106).

[282] Marti J Anderson. “Permutational multivariate analysis of variance (PERMANOVA)”. *Wiley statsref: statistics reference online*, pp. 1–15, 2014. (See pp. 106, 114).

[283] Howard L Sanders. “Marine benthic diversity: a comparative study”. *The American Naturalist* 102, pp. 243–282, 1968. (See pp. 106, 114).

[284] Paul J McMurdie and Susan Holmes. “Waste not, want not: why rarefying microbiome data is inadmissible”. *PLoS computational biology* 10, e1003531, 2014. (See p. 106).

[285] Sophie Weiss et al. “Normalization and microbial differential abundance strategies depend upon data characteristics”. *Microbiome* 5, pp. 1–18, 2017. (See pp. 106, 114).

[286] Donald T McKnight et al. “Methods for normalizing microbiome data: an ecological perspective”. *Methods in Ecology and Evolution* 10, pp. 389–400, 2019. (See p. 106).

[287] Johnny Hong et al. “To rarefy or not to rarefy: robustness and efficiency trade-offs of rarefying microbiome data”. *Bioinformatics* 38, pp. 2389–2396, 2022. (See p. 106).

[288] David S Clausen and Amy D Willis. “Evaluating replicability in microbiome data”. *Biostatistics*, 2021. (See p. 106).

[289] Lukas Beule and Petr Karlovsky. “Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities”. *PeerJ* 8, e9593, 2020. (See p. 106).

[290] Gregory B Gloor et al. “Microbiome datasets are compositional: and this is not optional”. *Frontiers in microbiology* 8, p. 2224, 2017. (See p. 106).

[291] Mark D Robinson et al. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. *bioinformatics* 26, pp. 139–140, 2010. (See p. 107).

[292] Jonathan Friedman and Eric J Alm. “Inferring correlation networks from genomic survey data”. 2012. (See p. 107).

[293] Michael I Love et al. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. *Genome biology* 15, pp. 1–21, 2014. (See p. 107).

[294] Matthew CB Tsilimigras and Anthony A Fodor. “Compositional data analysis of the microbiome: fundamentals, tools, and challenges”. *Annals of epidemiology* 26, pp. 330–335, 2016. (See p. 107).

[295] Siddhartha Mandal et al. “Analysis of composition of microbiomes: a novel method for studying microbial composition”. *Microbial ecology in health and disease* 26, p. 27663, 2015. (See pp. 107, 114, 115).

[296] Jacob T Nearing et al. “Microbiome differential abundance methods produce different results across 38 datasets”. *Nature communications* 13, pp. 1–16, 2022. (See pp. 107, 114).

[297] Huang Lin. “Statistical Theory and Methodology for the Analysis of Microbial Compositions, with Applications”. 2020. URL: http://d-scholarship.pitt.edu/39165/ (see p. 107).

[298] Joseph Nathaniel Paulson et al. "metagenomeSeq: Statistical analysis for sparse high-throughput sequencing". *Bioconductor package* 1, p. 191, 2013. (See p. 107).

[299] Greg Gloor. "ALDEx2: ANOVA-Like Differential Expression tool for compositional data". *ALDEX manual modular* 20, pp. 1–11, 2015. (See p. 107).

[300] Abhishek Kaul et al. "Analysis of microbiome data in the presence of excess zeros". *Frontiers in microbiology* 8, p. 2114, 2017. (See pp. 107, 114, 115).

[301] James T Morton et al. "Establishing microbial composition measurement standards with reference frames". *Nature communications* 10, pp. 1–11, 2019. (See p. 107).

[302] Amy Y Pan. "Statistical analysis of microbiome data: The challenge of sparsity". *Current Opinion in Endocrine and Metabolic Research* 19, pp. 35–40, 2021. (See p. 107).

[303] John F Cryan et al. "The Microbiota-Gut-Brain Axis". *Physiological reviews*, 2019. (See p. 111).

[304] Jack A Gilbert et al. "Current understanding of the human microbiome". *Nature medicine* 24, pp. 392–400, 2018. (See p. 111).

[305] Thomas SB Schmidt et al. "The human gut microbiome: from association to modulation". *Cell* 172, pp. 1198–1215, 2018. (See p. 111).

[306] Michael Lynch. "Rate, molecular spectrum, and consequences of human mutation". *Proceedings of the National Academy of Sciences* 107, pp. 961–968, 2010. (See p. 111).

[307] Jane A Foster et al. "Stress & the gut-brain axis: regulation by the microbiome". *Neurobiology of stress* 7, pp. 124–136, 2017. (See p. 111).

[308] Katerina V-A Johnson. "Gut microbiome composition and diversity are related to human personality traits". *Human Microbiome Journal* 15, p. 100069, 2020. (See p. 111).

[309] Gwen Falony et al. "Population-level analysis of gut microbiome variation". *Science* 352, pp. 560–564, 2016. (See p. 111).

[310] Jason Lloyd-Price et al. "Strains, functions and dynamics in the expanded Human Microbiome Project". *Nature* 550, pp. 61–66, 2017. (See p. 111).

[311] Mark Mimee et al. "Microbiome therapeutics—advances and challenges". *Advanced drug delivery reviews* 105, pp. 44–54, 2016. (See p. 111).

[312] Alexis Mosca et al. "Gut microbiota diversity and human diseases: should we reintroduce key predators in our ecosystem?" *Frontiers in microbiology* 7, p. 455, 2016. (See p. 111).

[313] Maria G Dominguez-Bello et al. "Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns". *Proceedings of the National Academy of Sciences* 107, pp. 11971–11975, 2010. (See p. 111).

[314] Ilseung Cho et al. "Antibiotics in early life alter the murine colonic microbiome and adiposity". *Nature* 488, pp. 621–626, 2012. (See p. 111).

[315] Efraim Jaul and Jeremy Barron. "Age-related diseases and clinical and public health implications for the 85 years old and over population". *Frontiers in public health* 5, p. 335, 2017. (See p. 111).

[316] Toshitaka Odamaki et al. “Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study”. *BMC microbiology* 16, pp. 1–12, 2016. (See pp. 111, 120).

[317] Bong-Soo Kim et al. “Comparison of the gut microbiota of centenarians in longevity villages of South Korea with those of other age groups”. *J. Microbiol. Biotechnol*, 2019. (See pp. 111, 120).

[318] Lu Wu et al. “A cross-sectional study of compositional and functional profiles of gut microbiota in Sardinian centenarians”. *Msystems* 4, e00325–19, 2019. (See pp. 111, 120, 121).

[319] Ngangyola Tuikhar et al. “Comparative analysis of the gut microbiota in centenarians and young adults shows a common signature across genotypically non-related populations”. *Mechanisms of Ageing and Development* 179, pp. 23–35, 2019. (See pp. 111, 120).

[320] Orawan La-Ongkham et al. “Age-related changes in the gut microbiota and the core gut microbiome of healthy Thai humans”. *3 Biotech* 10, pp. 1–14, 2020. (See pp. 111, 120, 121).

[321] Diego Marcos-Pérez et al. “Centenarians as models of healthy aging: Example of REST”. *Ageing Research Reviews* 70, p. 101392, 2021. (See p. 111).

[322] Jessica Evert et al. “Morbidity profiles of centenarians: survivors, delayers, and escapers”. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 58, pp. M232–M237, 2003. (See p. 111).

[323] Mercedes Clerencia-Sierra et al. “Do centenarians die healthier than younger elders? A comparative epidemiological study in Spain”. *Journal of clinical medicine* 9, p. 1563, 2020. (See p. 111).

[324] Adriana Florinela Catoi et al. “Gut microbiota and aging-A focus on centenarians”. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1866, p. 165765, 2020. (See p. 111).

[325] Elena Biagi et al. “Gut microbiota and extreme longevity”. *Current Biology* 26, pp. 1480–1485, 2016. (See pp. 111, 121).

[326] Yuko Sato et al. “Novel bile acid biosynthetic pathways are enriched in the microbiome of centenarians”. *Nature* 599, pp. 458–464, 2021. (See p. 111).

[327] Qinghua Sun et al. “Alterations in fecal short-chain fatty acids in patients with irritable bowel syndrome: A systematic review and meta-analysis”. *Medicine* 98, 2019. (See p. 111).

[328] Nicholas M Vogt et al. “Gut microbiome alterations in Alzheimer’s disease”. *Scientific reports* 7, pp. 1–11, 2017. (See pp. 111, 112, 121).

[329] Barbara JH Verhaar et al. “Gut Microbiota Composition Is Related to AD Pathology”. *Frontiers in immunology* 12, 2021. (See pp. 111–113, 120, 121).

[330] Zhen-Qian Zhuang et al. “Gut microbiota is altered in patients with Alzheimer’s disease”. *Journal of Alzheimer’s disease* 63, pp. 1337–1346, 2018. (See pp. 112, 121).

[331] Henne Holstege et al. “The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description”. *European Journal of Epidemiology* 33, pp. 1229–1249, 2018. (See p. 112).

[332] Wiesje M Van Der Flier and Philip Scheltens. “Amsterdam dementia cohort: performing research to optimize care”. *Journal of Alzheimer’s Disease* 62, pp. 1091–1111, 2018. (See p. 112).

[333] Wiesje M van der Flier et al. “Optimizing patient care and research: the Amsterdam Dementia Cohort”. *Journal of Alzheimer’s disease* 41, pp. 313–327, 2014. (See p. 112).

[334] Guy McKhann et al. “Clinical diagnosis of Alzheimer’s disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer’s Disease”. *Neurology* 34, pp. 939–939, 1984. (See p. 112).

[335] Tom N Tombaugh and Nancy J McIntyre. “The mini-mental state examination: a comprehensive review”. *Journal of the American Geriatrics Society* 40, pp. 922–935, 1992. (See p. 112).

[336] Robert Schmieder et al. “TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets”. *BMC bioinformatics* 11, pp. 1–14, 2010. (See p. 113).

[337] Christian Quast et al. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. *Nucleic acids research* 41, pp. D590–D596, 2012. (See p. 113).

[338] Kazutaka Katoh and Daron M Standley. “MAFFT multiple sequence alignment software version 7: improvements in performance and usability”. *Molecular biology and evolution* 30, pp. 772–780, 2013. (See p. 113).

[339] Morgan N Price et al. “FastTree 2–approximately maximum-likelihood trees for large alignments”. *PloS one* 5, e9490, 2010. (See p. 114).

[340] Paul J McMurdie and Susan Holmes. “phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data”. *PloS one* 8, e61217, 2013. (See p. 114).

[341] Nicholas J Gotelli and Robert K Colwell. “Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness”. *Ecology letters* 4, pp. 379–391, 2001. (See p. 114).

[342] Jari Oksanen et al. “The vegan package”. *Community ecology package* 10, p. 719, 2007. (See p. 114).

[343] Spase Stojanov et al. “The influence of probiotics on the firmicutes/bacteroidetes ratio in the treatment of obesity and inflammatory bowel disease”. *Microorganisms* 8, p. 1715, 2020. (See p. 117).

[344] Ian B Jeffery et al. “Composition and temporal stability of the gut microbiota in older persons”. *The ISME journal* 10, pp. 170–182, 2016. (See p. 121).

[345] Tarini S Ghosh et al. “Adjusting for age improves identification of gut microbiome alterations in multiple diseases”. *Elife* 9, e50240, 2020. (See p. 121).

[346] Anna Picca et al. “Gut microbial, inflammatory and metabolic signatures in older people with physical frailty and sarcopenia: results from the BIOSPHERE study”. *Nutrients* 12, p. 65, 2019. (See p. 121).

[347] Serena Verdi et al. “An investigation into physical frailty as a link between the gut microbiome and cognitive health”. *Frontiers in Aging Neuroscience*, p. 398, 2018. (See p. 121).

[348] Faiga Magzal et al. “Increased physical activity improves gut microbiota composition and reduces short-chain fatty acid concentrations in older adults with insomnia”. *Scientific Reports* 12, pp. 1–14, 2022. (See p. 121).

[349] Jacobo de la Cuesta-Zuluaga et al. “Age-and sex-dependent patterns of gut microbial diversity in human adults”. *Msystems* 4, e00261–19, 2019. (See p. 122).

Acknowledgments

In these acknowledgments, I wish to commemorate the completion of my PhD. It has been quite a ride (seemingly endless) with its fair share of hurdles, including the unfortunate loss of a part of this work. Nonetheless, I may now finally be able to reminisce upon all the good memories and the support I received from so many individuals at Delft, Amsterdam, and beyond. Their contributions have played a crucial role in my journey and future path, and I take this opportunity to express my appreciation to all those who have assisted and encouraged me along the way.

First and foremost I'd like to express my gratitude to my family, my parents, Cees and Marjanka, and my sisters Marlijn and Winnie. Your support and willingness to listen to me vent made all the difference. Pap, our long, weekly walks through the forest and along the beach have been a source of comfort, helping me maintain my sanity.

I'd like to thank my promotors. Marcel, I recall our first encounter as I was wrapping up my bachelor thesis in Erik's group. Daphne, who had also been under your supervision, strongly encouraged me to pursue my master's in Delft, and nowhere else. Little did you know that you would be stuck with me for many years to come! Despite juggling a packed schedule, you were always approachable, and I am genuinely grateful for your patience, direction, and mentorship. You gave me space to explore my own research interests while providing me with guidance and advice every step of the way. Your enthusiasm for science and research is inspiring, and I am thankful that I had the opportunity to have worked with you. Under your supervision, I have gained invaluable knowledge and grown both personally and professionally. Zaid, many thanks for your technical expertise and valuable feedback throughout my research.

Jasper, I'd like to extend my thanks for your supervision during my Master's and your valuable input thereafter. I truly appreciated your straightforward attitude and gained a great deal from your critical approach towards research.

Many thanks to Saskia for all her assistance during my time at Delft and beyond. Additionally, a thank you to Bart, Ruud, and Robbert for your continued support in maintaining smooth operations and preventing us from blowing up the cluster.

I'd like to acknowledge some of the people I got to briefly overlap with in the early days at the Delft Bioinformatics Lab. Sjoerd, I really enjoyed your music. Thies, your impressive biking skills consistently amazed me. Marc, the best bio*statistician* I know. And Amin, who warned me never to mistreat an Iranian.

A shoutout to the triple OGs with whom I shared countless memorable moments in Partycity 5.920 - we've finally made it! Stavros, it's been great seeing your involvement in various projects since our Masters daysand attending your wedding in Greece was unforgettable. Alex, your barbecues were always enjoyable. Christian, THE German, your sense of humor never failed to brighten my day. Soufiane, we largely overlapped, and you were usually sitting right next to me if you weren't in Amsterdam, you were great company; it seems our days to "CHOPe" have come to an end. Lastly, Christine, thank you for remaining a great friend. Your lively presence added energy to our group, and you likely played a role in at least one person's transition to industry.

Our room enjoyed the company of many newcomers that I got to meet while I was still around, I'll mention just a few. Mostafa, a welcome addition to our room; your kind nature balanced your imposing presence. Lucas, your brief stay ended by escaping to the US, good for you! Duco, a person who actually worked on something interesting, was fun to help out a little in your work. Aysun, always up to try new restaurants, hopefully I have more time soon to join you. It was great to gossip with you and the others about everything that was going on in the building.

Tamim, I appreciated your football invitations; I stopped going because we averaged at least one injury per game and not because I lack the skill. Ahmed, you set a positive example for everyone, and I would have considered myself fortunate to work with you. Ramin, good thing I was warned in advance. I enjoyed our board game sessions with Aysun, Chirag, Amelia, Jasper, Alexander, Marcus, and Christine: Aysun, Chirag & Amelia, Jasper, Alexander, Marcus, Christine, and others. I hope to continue these gatherings in the future with all of you and any new faces. Henne, thank you for the warm welcome during our collaboration in the 100plus group and the retreat invitation. Meng, I am grateful for the exclusive opportunity to order lunch from the top-secret Chinese WhatsApp group.

Despite my tendency to be brief, I'd like to take a moment to acknowledge a few more individuals: Gerard, Tom, Erik, Nicco, Arlin, Sven, Linda, Mo, Wouter, Lieke, Colm, Laura, Stephanie, Osman, Joana, Thomas, Marco, David.

اخيراً وليس اخرا

Farah, my partner, you are a truly a treasure — the most driven and hard-working person I know. I am lucky to have some of that rub off on me. I am deeply grateful for the way you encouraged me to step out of my comfort zone and explore the world alongside you. I never could have imagined living in Iraq, yet it was there where I finalized my thesis, making it a uniquely special place for me. Thank you for everything!

Curriculum Vitæ

Tom Onno Mokveld

12/05/1991	Born in Leiden, The Netherlands

Education

2017-2022	PhD in Computer Science, Delft University of Technology
2014-2016	MSc in Computer Science Delft University of Technology & Leiden University
2009-2014	BSc Bioinformatics, University of Applied Sciences Leiden

Professional experience

2023-now	Bioinformatics Scientist, Pacific Biosciences of California, Menlo Park, USA
2016-2017	Scientific Programmer, Delft University of Technology, Delft, The Netherlands

List of Publications

1. **T. Mokveld**, J. Linthorst, Z. Al-Ars, H. Holstege, and M. Reinders. *CHOP: haplotype-aware path indexing in population graphs*, Genome Biology.

2. **T. Mokveld**, Z. Al-Ars, E. A. Sistermans, and M. Reinders. *A comprehensive performance analysis of sequence-based within-sample testing NIPT methods*, PloS One.

3. **T. Mokveld**, Z. Al-Ars, E. A. Sistermans, and M. Reinders. *WisecondorFF: Improved Fetal Aneuploidy Detection from Shallow WGS through Fragment Length Analysis*, MDPI Diagnostics.

4. **T. Mokveld**, B. J. H. Verhaar, A. Salazar, L. M. C. Lorenz, H. M. A. Hendriksen, F. de Leeuw, Y. Pijnenburg, K. van Vliet, M. Graat, R. Kraaij, C. E. Teunissen, Z. Al-Ars, W. M. van der Flier, H. Holstege, and M. Reinders. *A Cross-Sectional Study of Compositional Profiles of Gut Microbiota in Dutch Centenarians and patients with Alzheimer's disease*, Under review.

5. P. Lof, D. H. K. Gaillard, E. A. Sistermans, **T. Mokveld**, H. M. Horlings, M. Reinders, F. Amant, D. van den broek, L. F. A. Wessels, C. A. R. Lok. *Effective pre-operative diagnosis of ovarian cancer using minimally invasive Liquid Biopsies by combining HE4 and cfDNA in patients with an ovarian mass*, Intended for publication in European Journal of Cancer.