IDENTIFICATION OF GENETIC VARIATION IN HIGHLY DIVERGENT REGIONS USING WHOLE EXOME SEQUENCING

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Shulan Tian

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DR. SUSAN SLAGER, ADVISOR
DR. CLAUDIA NEUHAUSER, CO-ADVISOR

December, 2016

Acknowledgements

First and foremost, I want to thank my advisor Dr. Susan Slager who fully supported me to work with this exciting project and spent enormous time to guide my thesis research. She also provided many inspiring ideas to ensure that the project went on smoothly. Without her encouragement and support, any of my accomplishments in this thesis would have not been possible.

I am also deeply thankful to my committee members, Drs. Claudia Neuhauser, James Cerhan and Nathan Pankratz, who provided invaluable insights and guidance. They have spent a lot of time to read my lengthy proposal and thesis despite of their busy schedules.

Working in a great environment within Mayo Clinic, I received many great suggestions from my collaborators and colleagues. In particular, I am grateful to my colleagues Mike Kalmbach, Neil Ihrke, who constantly assists me in tool installation and configuration, and Matt Bockol who kindly downloaded public data resources whenever needed. In addition, I thank my colleagues, Dr. Alexj Abyzov, Dr. Eric Klee, Dr. Zhifu Sun, Daniel O'Brien, Shannon McDonnell, Saurabh Baheti and Jared Evans for their helpful discussions. I also would like to thank the staff from RCF (Research Computing Facility) for their help.

Finally, I thank my family for their constant love and support. I thank my husband Dr. Huihuang Yan who gave me unconditional support and scientific guidance throughout the thesis, and my two great kids, Jenny Yan and Benjamin Yan, for spending their precious time in my dark office on sunshine weekends. I also thank my family in China who gave me a lot of encouragement.

Dedication

This thesis is dedicated to my husband Huihuang Yan for unconditional support and scientific guidance and to my two children, Benjamin Yan and Jenny Yan, who have made this accomplishment possible by sacrificing almost every weekend in my dark office.

Abstract

Whole exome sequencing is widely used for identifying disease-associated variants in both clinic and research settings. Using this technology to accurately identify genetic variants is essential, yet major challenges remain in highly divergent but medically important genomic regions, such as the human leukocyte antigen (HLA) region. We developed an analytical workflow enabling sensitive and accurate variant discovery for highly divergent genomic regions from whole exome sequencing data. Our workflow combines both mapping- and de novo assembly-based approaches, for which the tools were selected and optimized through extensive evaluation of their performance across different coverage depths and divergence levels, the two key factors profoundly impacting variant detection. We used simulated exome reads for an initial assessment and then public exome data from a well-studied CEPH individual NA12878 for more focused evaluations; the genetic variation in NA12878 has been deeply catalogued through multiple sequencing platforms and analytical strategies. Our analysis revealed that the 25 combinations between five mappers and five callers had comparable performance in the non-HLA regions as expected, which have approximately 0.1-0.4% divergence. However, they differed markedly in the HLA region in which different haplotypes can show up to 10-15% divergence. We also evaluated the effect of post-alignment processing and provide a practical guideline regarding the application of local realignment and base quality score recalibration in designing analytical workflows. Local realignment is not necessary for Novoalign and the three haplotype-based callers in both the HLA and non-HLA regions. Base quality score recalibration reduces the SNP calling sensitivity from the HLA region without noticeable gains in precision rate for most of the cases. We transferred our findings into a highly sensitive and computationally efficient workflow for mapping-based variant discovery. It excels in both sensitivity and speed through our two-tier mapping strategy, not only in regions of high divergence like HLA but also in lowly divergent regions. To utilize the local phasing information and identify transmitted variants, we also developed a de novo assembly-based variant calling workflow for whole exome data. It performs well over a wide range of coverage depths and divergence levels. In fact, for SNP detection from the HLA region, it is far more superior to all other existing de novo assembly-based methods and also performs better than the mapping-based variant detection methods based on both simulated and multiple benchmarked exome datasets. Finally, we incorporated the mapping- and de novo assembly-based approaches into a single pipeline, providing the flexibility of variant detection through executing either or both methods. Our pipeline should be particularly useful for WES projects focusing on diseases that are associated with HLA or other highly divergent regions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and outline of the chapters

1.1. Whole exome sequencing and its application in variation detection

Deciphering the genetic basis of human diseases is a fundamental undertaking in biomedical research [1]. The advent of next-generation sequencing (NGS) technologies has dramatically improved our capacity of interrogating the human genome. The information gained from this ongoing effort will contribute to disease prevention, diagnosis and treatment [2]. Whole genome sequencing (WGS) and whole exome sequencing (WES) are two main platforms for genome-wide variant discovery. The former aims to sequence the entire genome, which allows systematic identification of genetic variations like single nucleotide polymorphisms (SNPs) and insertion and deletions (INDELs), while the latter only targets protein coding and nearby regions representing ~1% of the genome [3].

In clinical settings, WES is currently preferred over WGS in identifying disease-associated variants for numerous reasons [4]. Most importantly, WES is cost-effective [3], as it requires only ~5% of the sequence data needed by WGS [5]. In addition, WES can reliably target >95% of the exons [6, 7], with the potential of identifying causative variants missed by WGS and resolving ambiguous calls more confidently [6]. In personalized medicine, mutations in the exons are more likely to be actionable [4, 6, 8, 9]. Currently, WES is used in the diagnosis of particularly difficult-to-diagnose patients [10], prenatal diagnosis [11, 12] and early diagnosis of debilitating disease [13, 14].

Several ongoing large-scale WES projects are making significant contributions to the field of biomedical research. For example, the Exome Aggregation Consortium (ExAC) currently hosts WES data from 60,706 unrelated individuals used in disease-specific or population genetic studies (http://exac.broadinstitute.org/). In fact, healthcare communities have started to combine genetic information from WES with patient electric records for disease prevention and treatment. For example, Regeneron and Geisinger Health System launched the “MyCode” project to conduct the world’s largest WES studies; curation of genomic data through such large-scale biobanking efforts may benefit personalized healthcare throughout the lifespan of patients.

1.2. Bioinformatics workflows for WES-based variation detection

1.2.1. Mapping-based variant calling workflows

Accurate identification of genetic variants in WES studies is essential to our understanding of disease etiology. Variant detection from NGS data requires the development of a series of software tools, which are more often integrated into pipelines or workflows for high efficiency and easy application. Mapping-based workflows are frequently used, often following the GATK (Genome Analysis Toolkit) Best Practices (https://www.broadinstitute.org/gatk/guide/best-practices) [15, 16]. In mapping-based approaches, raw reads are first mapped to a reference sequence using a short-read aligner (mapper), and the bases differing between mapped reads and the reference are identified by using variant calling tools (caller) [17, 18].

In recent years, many mappers have been developed to work with different sequencing platforms, read lengths and divergence levels [19]. There are two main types of mapping algorithms for NGS short reads: those based on a hash table, indexing either reference genome (like Novoalign) or reads (like ELAND), and others based on Burrows-Wheeler Transform (BWT) indexing [20-23]. The selection of an appropriate mapper impacts the accuracy of variant discovery. Current variant calling pipelines frequently implement BWT-based mappers like BWA (Burrows-Wheeler Aligner). They are much faster and more memory efficient, but somehow less sensitive compared with hash-based aligners [21, 24, 25]. More importantly, the BWT-based mappers were developed for mapping reads that perfectly match the reference or have only a few mismatches/INDELs

[24, 26]. A few aligners have been tailored to the mapping of more divergent reads. Among them, GSNAP [26], NextGenMap [25], Novoalign (http://www.novocraft.com/products/novoalign/) and Stampy [21] use a hash table representing the locations of k-mers (12- to 15-mer) in the reference sequence. Nevertheless, for these ‘variation-tolerant’ mappers, it remains less clear to what extent the difference in mapping will impact variant calling.

Identifying SNPs and INDELs is another key step in variant calling workflow. Several popular software packages are available for both single- and multi-sample variant calling. They can be largely grouped into two categories: those based on alignment and those based on haplotype or local de novo assembly. The former includes a few widely used callers like VarScan [27], SAMtools mpileup [28], and GATK UnifiedGenotyper (GATK UG) [15, 29] (https://www.broadinstitute.org/gatk/index.php). These algorithms first collect information about each allele from alignment profile, and use a Bayesian framework or heuristics to estimate the uncertainty in alignment and base calling [15, 30-32]. Based on the posterior probability (as the measurement of genotype confidence) or simply on empirical allele frequency (AF) cutoffs [33], each variant site is assigned a genotype, as homozygous variant, heterozygous variant or reference call. The alignment-based approaches generally have high SNP calling sensitivity and low requirement for computing resources. However, they identify the variation base by base from read alignments, so they do not identify different variant types (like SNPs and INDELs) simultaneously. On the other hand, haplotype and local de novo assembly based callers are capable of calling SNPs and INDELs simultaneously. Instead of calling variants base by base from read alignments, FreeBayes achieves haplotype-based variant discovery by identifying the most likely haplotype(s) based on alignments [34]. It can detect multiple nucleotide polymorphisms (MNPs) up to the read length in addition to SNPs and INDELs. GATK HaplotypeCaller [29] and Platypus [4] combine haplotype-based and local de novo assembly based calling. Such local de novo assemblers scan small individual windows (~ a few kb) for the presence of variations; for each informative window, they build a de Bruijn graph using all the pairs of reads with one or both ends mapped, where reads are first split into k-mers to increase graph continuity during the assembly. The programs then realign each assembled haplotype against the reference to identify potential variation sites. Candidate alleles are identified by following unique paths in the graph through a depth-first traversal algorithm. De novo assembly based approaches are generally more accurate since they completely assemble the reads in a region. However, these approaches have high computational requirements. Compared to alignment-based approaches, they have a lower sensitivity and are limited by repetitive sequences, as contiguity information is lost when the reads are broken into consecutive k-mers during graph construction.

In addition, GATK Best Practices workflows recommend two key post-alignment processing steps, i.e., local realignment around known INDELs and base quality score recalibration (BQSR) [15, 16]. Reads spanning INDELs have a high chance of being aligned incorrectly to the reference [32, 35]. A previous study revealed that BWA mapping generated misalignments for over 15% of the reads spanning known homozygous INDELs [15]. Without correction, those misaligned bases in reads can be easily called as spurious variants. Also, the confidence level in variant calling depends on the accuracy of both base calling and base quality score [33]. The latter measures the probability that a base is called incorrectly [36], with a Phred-scale quality score of Q ideally corresponding to an error rate of $10^{-Q/10}$ [33]. However, the raw quality score does not truly reflect the base-calling error rate; instead, it varies with multiple factors including sequencing platform, the number of machine cycles at which a base is sequenced, and the sequence composition [15, 33]. For example, the bases at the 3’ end of a read are typically more error prone than those at the 5’ end [31]. While these two steps can decrease false positive calls from BWA mapping and GATK UnifiedGenotyper calling [15], it is less clear whether the same benefit will be obtained from pipelines using other mappers and callers and for reads generated on newly developed sequencing platforms.

1.2.2. Challenges in variant detection from highly divergent regions

Human genetic variation exists both within and across populations. On average, interindividual sequence similarity is estimated at 99.5% based on diploid human genome reference HuRef [37]. Thus, the variation among individuals is currently estimated at 0.5% (100%-99.5%), which is five to seven-fold greater than previously estimated (~0.1%). Highly divergent regions refer to those with less than 99.5% sequence identity in the human genome; they are an important source of genetic variation often contributing to human disease [38]. The best example is the human leukocyte antigen (HLA) region on chromosome 6p21.3; this ~4-Mb region shows up to 10% or higher local sequence divergence between haplotypes [39-41]. Most importantly, the HLA region is associated with over 100 diseases, predominantly autoimmune diseases [42], and also with drug response to infectious agents [43]. Besides the HLA region, there are many other regions in the human genome that are also highly divergent [38]. McLure et al. identified another 107 highly divergent genomic regions using the genomic matching technique; some of those regions harbor medically important genes such as *KIR* and *HFE2* [38]. In addition, the Human Genome Project and the 1000 Genomes Project also identified large regions on chromosomes 8 (about 15 Mb) [44] and 16 (mostly large structural polymorphisms) [45] and subtelomeric regions on autosomal chromosomes that have high SNP density [46].

Several studies have compared the performance of multiple variant callers on WES data. However, those analyses did not focus on the highly divergent regions, and in most of cases, only BWA or ELAND2 were used as the mapper [47-49]. Both BWA and ELAND2 have low sensitivity in mapping reads that are more than 2% divergent from the reference. Moreover, those studies estimated variant detection sensitivity by using common SNP sites genotyped on SNP arrays. A more recent study compared GATK HaplotypeCaller, GATK

UnifiedGenotyper and SAMtools on Novoalign alignments [50]; however, it estimated sensitivity using the high-confidence call set that is known to have poor coverage in the HLA region [51].

Furthermore, a few studies have attempted to assess the overall concordance of different variant-calling algorithms on WES data [18, 52]. They often revealed low concordance among different pipelines, particularly for INDEL detection [52, 53]. Two analyses revealed only 27% [52] or 37% concordance [53] for INDELs among three methods and less than 60% for SNPs among five methods [52]. For WGS, variant calling methods generally agree well with one another in about 90% of the genome but show marked disagreement in the other ~10% ‘difficult regions’ of low-complexity and segmental duplications [54], which are especially prone to mutations [55]. In fact, existing variant calling methods could miss up to 25% of the variants in these difficult regions [54]. In another study, Abrandt et al. demonstrated the challenges of genotyping highly divergent regions in the 1000 Genomes Project. The authors compared the 1000G NGS data with the “gold standard” Sanger sequencing data across 930 samples, revealing that approximately 19% of SNP genotypes from the NGS data were incorrect in the five highly divergent HLA genes (*HLA-A*, *-B*, *-C*, *-DRB1* and *-DQB1*) [56]. Therefore, a comprehensive comparison of multiple mappers and callers together is needed to identify the most ideal combination(s) for variant discovery in highly divergent regions.

GATK Best Practices workflows generally work well for most of the genomic regions typically with only 0.1% divergence [24, 26]. The effect of local realignment and BQSR (base quality score recalibration) has been assessed using both WGS and WES data, mostly for BWA together with GATK UnifiedGenotyper, GATK HaplotypeCaller, and SAMtools [2, 15, 18, 32]. However, the results are inconsistent. In particular, it is not known whether these post-processing steps are needed for highly divergent regions like the HLA region. In addition, for applying BQSR, the algorithm needs to identify a list of supposedly non-polymorphic sites that do not overlap with any known polymorphic sites (like those in dbSNP) and that do not match the reference sequence either. It then builds a linear model taking into account the raw base quality score, base position in the reads (i.e., sequencing cycle), as well as the dinucleotide composition of those non-polymorphic sites, by which the recalibrated quality score is computed [15, 33]. While it is relatively straightforward to identify non-polymorphic sites in ordinary genomic regions where sequence diverges at the level of only about 0.1% [21], difficulty arises in regions of high divergence. In the latter case, a portion of the non-polymorphic sites likely represents true variants. In addition, local realignment and BQSR significantly increase the space for data storage. The storage of ever-increasing amounts of sequencing data presents a huge computational challenge for the genomics community [22]**.**

1.2.3. De novo assembly based variant calling

Despite the wide applications, mapping-based approaches have several limitations. For example, reference sequence is incomplete and contains misassemblies [17, 57]. It is estimated that 5-40Mb of euchromatic sequences are absent from a typical human genome [41] and additional 125–150 Mb of gene-rich regions of the genome are inaccessible to short read sequencing technologies [58]. Second, the human genome contains extensive segmental duplications, which causes the ambiguity of read alignment to the authentic versus to the paralogous regions [57]. Third, mapping-based approaches are less sensitive to INDELs [26] and their performance deteriorates in the presence of sequence variation [4, 21]. Given the inherent limitations of the standard mapping-based approaches, haplotype assembly of short reads has drawn more attention.

Sequence assembly, especially from short reads [59], has been challenging. Short read assembly often adopts graph-based algorithms including overlap-layout-consensus (OLC) and de Bruijn graph (DBG). OLC identifies all pairs of reads that overlap sufficiently and then organizes this overlap information into an assembly graph in which nodes represent reads and edges represent overlaps. The assembly generally follows three steps: the identification of pairwise overlaps among reads, layout of all reads (graph construction) and the construction of consensus sequence from overlap and layout. As it needs to compute all pairwise overlaps, the assembly process is computationally expensive, especially for large NGS datasets and large genomes. OLC graph is widely used for assembling long reads from Sanger sequencing, but is not applicable to mammalian-sized genomes. Gene Myers proposed the concept of ‘string graph’ in 2005 [60], which simplifies the global overlap graph by removing transitive edges. String graph is embodied in Celera Assembler that dominated sequence assembly until the release of next-generation short-read sequencing technologies [61].

In contrast, de Bruijn graph starts the assembly by breaking reads into consecutive strings of length k, called k-mers. In this assembly graph, vertices are k-mers from reads, and edges are overlaps of k-1 nucleotides. In De Bruijn graph, duplicated reads are collapsed and read overlap information is not needed; thus de Bruijn graph-based assembly is fast and can process large amounts of reads. Most short-read assemblers, like Velvet [62], ABySS [63], AllPath-LG [64] and SOAPdenovo [65], adopt the de Bruijn graph. However, de Bruijin graph is very sensitive to sequencing errors. Even though the Illumina platforms are producing long reads with less than 1% error rate on average, the probability of having error-free k-mers decreases geometrically [66].

The String Graph Assembler (SGA) is the first package to assemble mammalian-sized genomes through string graph [67, 68]. SGA represents the reads as a suffix array and finds overlaps using Ferragina Manzini index (FM-index) derived from BWT; the use of FM-index reduces the computational time from quadratic order (for overlap-layout-consensus and string graph without FM-index) to linear order of the number of reads. A comprehensive comparison rated SGA as one of the three most successful consensus sequence assemblers [69].

Methods for assembly-based variant discovery are still under active development without broad applications to large sequencing projects. They adopt three strategies: direct variant calling, microassembly and whole genome assembly. SGA was initially developed for the assembly of consensus sequence [68]. A new module called 'graph-diff' provides two options for direct variant calling, either from string graph or from de Bruijn graph. It looks for 'bubble' structures (variant bubbles) that are most likely to be caused by the polymorphisms, rather than by repeats or sequencing errors. Nevertheless, SGA graph-diff is rarely used in variant calling. Cortex applies colored de Bruijn graph, in which the nodes and edges are colored according to the samples having them; variants are directly called through the functions 'bubble-calling' and 'path-divergence calling' [57]. This package is particularly useful for joint analysis of multiple samples. However, the current version has a low sensitivity, with nearly 40% false negative rate [54]. Finally, Fermi is the only whole genome assembler, which outputs unitigs to preserve small variants and structural changes [17]. Variants are identified by mapping unitigs back to the reference genome, followed by variant calling from the alignments. Like SGA, it is also based on string graph and uses FM DNA index (FMD-index) for memory efficiency. This package is highly packaged, offering little flexibility for fine-tuning parameters.

These de novo assemblers represent an attractive alternative to mapping-based methods. They are not affected by the imperfection of the reference genome and are able to identify long INDELs and structural variations [17, 18, 53]. Nevertheless, assembly-based variant discovery is facing several challenges [17]. They often have lower sensitivity than the standard mapping based approaches [4, 17]. They are very sensitive to sequencing errors and are confronted by repetitive DNA [4], as it is difficult to differentiate 'bubbles' caused by true polymorphisms vs. those caused by repeats or sequencing errors. In de

Bruijn graph-based assembly, as the read information is lost after splitting reads into k-mers, some paths in the graph may be invalid without support by the actual reads [17, 70]. Erroneous calls may arise due to misassembly. They have high computational requirements [4, 17, 57]. Assembly of WES reads poses an additional challenge due to the much greater variation in coverage [4, 6]. Therefore, considerable efforts will be needed to make assembly-based variant detection a routine practice.

1.3. Quality metrics for assessing variant calling workflows

To evaluate the performance of variant calling methods, a ‘truth’ variant set needs to be defined for benchmarking. A variant call set can be compared to this ‘truth’ set from the same sample to identify overlaps, missing or extra calls. Reads simulation is commonly used in assessing read mapping, de novo assembly and variant calling, since the true variants are known. Simulation process can introduce variants randomly or preplace a list of pre-defined variants into the reference genome and then generate reads from the modified genome sequence. Many simulation tools have been developed and some were tailored to different sequencing platforms [71]. None of the tools are likely to capture the full spectrum of sequence characteristics and complexity. Rather, simulated data serve as a reasonable surrogate to evaluate key factors (like coverage and divergence) at multiple levels. The hypothesis generated in such a simplified scenario should be further tested using real sequencing data.

Several approaches can be used to evaluate the quality of a variant call set in real sequencing data. The most reliable method is to perform Sanger sequencing of the regions covering putative variants, which is costly, time-consuming and not scalable. Another popular method is to evaluate concordance between the variant call set from sequencing data and that from genotyping array for the same sample. While this approach is much more scalable than Sanger sequencing, the arrays usually only cover a subset of known variants, primarily common variants. The most common approach in benchmarking studies is to evaluate how a variant call set is concordant with a carefully curated ‘reference call set’. Such a reference set, typically generated from multiple sequencing platforms and variant calling algorithms, can be treated as a proxy for ‘truth set’ or ‘gold standard’. This method is not to judge the veracity of individual variants, but to estimate the overall quality of call sets from different variant calling algorithms.

In many of the studies, especially those focusing on the detection of rare mutations, it is not feasible to generate a reliable reference set from known variations. For sample NA12878, two public call sets are most notable. The Genome In A Bottle (GIAB) consortium has published a high-confidence variant call set [51]. To compile this call set, 14 whole genome and exome sequence datasets were generated using 5 sequencing technologies. Data were analyzed through 7 different read aligners and 3 variant callers. By its nature, this variant call set biases toward the variants that are located in ‘ordinary’ regions and that are relatively easy to characterize. Variant calling in the other ~10% ‘difficult’ regions is less reliable and complete; thus, benchmarking against the above high-confidence calls will lead to an overestimate of sensitivity for a method. In order to recover the variants from the difficult regions in NA12878, i.e., regions of low-complexity sequences and segmental duplications, 2x250bp WGS data were generated from PCR-free genomic libraries at high coverage [54]. Generated by the most advanced library preparation and Illumina sequencing platform, such WGS datasets are believed to be high in quality, which serve as a ‘gold standard’ in the 1000 Genomes Project. Variants were detected using both mapping-based (BWA-MEM+GATK HaplotypeCaller) and de novo assembly-based approaches (Cortex and DISCOVAR), where DISCOVAR, a method specifically developed to work with long (250bp) reads from PCR-free libaries, had a false positive and negative rate of 1.63% and 6%, respectively [54].

Once the ‘truth’ set is clearly defined, several quality metrics can be used to measure variant calling performance by comparing the called variants to the ‘truth’ set. We applied three widely used quality metrics in Chapters 2 to 4, including non-reference sensitivity (NRS), also known as true positive rate or recall, precision rate (PR, or positive predictive value) and non-reference (overall) genotype concordance. These metrics are defined as follows.

Non-reference sensitivity is the ratio of correctly identified variants over the total known variants in the reference set:

Non-reference sensitivity = true positive/(true positive + false negative)

Precision rate estimates the probability that a variant call is truly a reference variant and is defined as:

Precision rate = true positive/(true positive + false positive)

Non-reference genotype concordance is the number of genotype calls made by a benchmarking tool that are concordant with those in the ‘truth set’. It is normalized to the total number of genotypes in the ‘truth’ set.

Non-reference Genotype Concordance = Number of concordant genotypes in variant sites/total genotypes in true variant sites

We calculated all the three metrics for simulated data. True positive refers to the number of true (simulated) variants identified by a method, false negative represents the number of true variants missed by a method, and false positive represents the number of called variants that does not overlap the true variants. For real exome data, however, it is challenging to define false positive in a variant call set unless the whole genome is fully sequenced. Therefore, we can only confidently calculate non-reference sensitivity and genotype concordance. To measure precision indirectly, we randomly selected a subset of called variants and manually checked read pileup profile for signs of false positives, which could be caused by misalignments or sequencing errors.

1.4. Outline of the chapters

To address the challenges of variant calling in highly divergent regions, and ultimately to fully catalogue genetic variation genome-wide from WES experiments, my dissertation entails the development of a modulated pipeline for variant discovery and assessment. The pipeline contains two complementary components by implementing short read mapping- and de novo assembly-based strategies. The two approaches and their respective modules have been intensively tested and optimized for performance using both simulated and real exome data. In brief, Chapters 2 and 3 cover the key components in mapping-based variant discovery, especially in the HLA region; Chapter 4 elaborates the development of a de novo assembly-based variant detection workflow; and Chapter 5 illustrates a highly integrative yet flexible pipeline based on the findings described in Chapters 2 to 4. Below, we summarized the rationale of the design, key features and potential applications of the pipeline.

Chapter 2 provides a practical guideline regarding post-alignment processing in variant calling from exome data. We focused on two critical steps, i.e., local realignment around known INDELs and base quality score recalibration. Both steps are generally believed to improve variant calling accuracy but are computationally intensive. Though post-processing is widely used in mapping-based workflows, it is unknown whether this common practice contributes to variant discovery across all workflows, especially given that variant detection methods are updated regularly and the new sequencing platforms have greatly improved sequencing quality. For example, for a typical 2x75bp sequencing run on the Illumina HiSeq 4000 system, over 80% of the bases have a quality score greater than Q30. More importantly, it is unknown how these post-processing steps impact variant calling in highly divergent regions. Therefore, we comprehensively assessed the two steps across a panel of 25 mapper and caller combinations, using both simulated data and real exome data from NA12878. This analysis also considers other critical variables, such as coverage depth and genomic divergence. We found that local realignment contributes little to Novoalign, a commonly used mapper, and to three haplotype-based callers:

GATK HaplotypeCaller, FreeBayes and Platypus. On the other hand, realignment improves INDEL detection by SAMtools and GATK UnifiedGenotyper, the two callers that are known to be less suitable for INDEL detection. Our analyses indicated that base quality score recalibration is not needed. It leads to a loss of sensitivity, particularly in the highly divergent HLA region, with no or little gains in precision. Our findings provide critical information needed for further improvement of the current mapping-based variant discovery pipelines, especially for those used to study the disease-associated genetic variation in the HLA or other highly divergent regions.

Chapter 3 addresses the issue of how to choose the best mapper(s) and caller(s) in building mapping-based variant discovery pipelines. We assessed five popular mappers and five callers on simulated data, NA12878 exome data, and our internal exome sequencing data from CLL (chronic lymphocytic leukemia) patients. These mappers and callers are based on different algorithms. For example, the GATK Best Practice workflow that is widely used in large-scale sequencing projects uses BWA as the mapper. BWA is based on BWT and FM-index, requiring less memory at the cost of sensitivity. We found that these 25 methods are highly comparable in variant detection from low-divergence regions (i.e., non-HLA regions that show about 0.1-0.4% divergence). However, commonly used variant-detection methods missed up to 20% of the known variants in highly divergent regions (such as the HLA region that shows up to 10-15% divergence between different haplotypes), and even more in a few most highly divergent genes whose alleles are associated with human diseases. Therefore, at the end of this chapter, we proposed a workflow integrating two mappers and three callers, without base quality score recalibration (based on results in chapter 2). The two-step mapping strategy achieves both high speed (by BWA) and accuracy (by GSNAP), while sensitive detection of both SNPs and INDELs is accomplished through a joint calling strategy with three distinct variant callers (Platypus and the two GATK callers). We have implemented this workflow for large-scale variant calling from whole exome data.

Chapter 4 describes the development of a de novo assembly-based variant calling workflow for whole exome data. It performs de novo assembly and variant calling at the chromosome level, using both unmapped reads and reads mapped to a given chromosome. Other existing methods use only paired reads with at least one end mapped to a region and perform local de novo assembly at the kilobase (usually 2kb) level. Our novel approach is based on the notion that some read pairs simply could not be mapped using a short-read aligner due to the high sequence divergence. The exclusion of those reads from the assembly will miss the variants from unmappable and highly divergent regions. We began with the development of a haplotype-based assembly strategy which aims to preserves the heterozygosity between two haplotypes. We demonstrated its robustness over a wide range of divergence using simulated reads. We further optimized the pipeline using real Illumina sequencing reads and the eight publicly available HLA haplotypes. The workflow enables the detection of SNPs in both HLA and non-HLA regions at high sensitivity, as well as the detection of large INDELs. We systematically tested the workflow with 76, 100 and 150 bp reads from eight exome datasets in NA12878. Our workflow outperformed all 6 mapping-based and 5 de novo assembly-based approaches in SNP detection from the HLA region. Critically, the pipeline assembles reads into haplotype contigs, thus providing phase information needed for downstream HLA typing and the discovery of transmitted variants.

Chapter 5 describes a highly integrative yet flexible pipeline for both mapping- and de novo assembly-based variant discovery in highly divergent regions. It was developed based on the key findings discussed in Chapters 2 to 4.

Chapter 2

Impact of post-alignment processing in variant discovery from whole exome sequencing

2.1. Summary

GATK Best Practices workflows are widely used in large-scale sequencing projects. They recommend post-processing of alignments before variant calling. Two of the key steps are local realignment around known INDELs and base quality score recalibration, with the former being computationally demanding. Both are expected to reduce erroneous calls, an inference mainly supported by the initial assessment made on a pipeline that uses BWA mapping and GATK UnifiedGenotyper calling. Nevertheless, it is not known whether and to what extent post-processing might benefit analytical pipelines implementing other methods, given that both mappers and callers are typically updated from time to time. Also, sequencing platforms are upgraded regularly, which will likely provide better estimations of read quality scores. Finally, some medically critical regions in the human genome show high sequence divergence, and how post-processing might impact variant calling in those regions is not clear. We used both simulated and NA12878 exome sequencing data to comprehensively assess the impact of post-processing for five popular mappers together with five callers. Focusing on chromosome 6 that harbors the HLA region of high sequence divergence, we found that local realignment had little to no impact on SNP calling. Its role in INDEL calling, however, depended on the mapper and caller combination. Increased sensitivity was observed for a few methods, most notably for Stampy and GATK UnifiedGenotyper combination. However, a much less effect was detected on the three haplotype-based callers and no effect from local realignment for INDEL calling was detected using Novoalign. As for Base Quality Score Recalibration (BQSR), we found that it had a negligible effect on INDEL calling. It generally reduced SNP calling sensitivity, across factors including the caller used, the depth of sequencing coverage, and extent of divergence level. Specifically, for SAMtools and FreeBayes calling in the non-HLA regions, BQSR had reduced the sensitivity but in cases of insufficient coverage, precision improved. For the remaining callers in the non-HLA regions and all the callers in the HLA regions, BQSR had reduced the sensitivity with little to no gains in precision rate. **In conclusion,** we demonstrated that the effect of local realignment and BQSR is not universal; rather, it depends on mapper and caller, and is influenced further by the extent of coverage and genomic divergence. To support the generality of the observation, we captured the trends that are highly consistent over all the five NA12878 exome data. Overall, local realignment mainly improves GATK UnifiedGenotyper and SAMtools in INDEL detection; BQSR affects FreeBayes and SAMtools in SNP detection from low-coverage data, which reflected the trade-off between an increase in precision rate and decrease in sensitivity. Our analysis provides a practical guideline regarding post-alignment processing of Illumina exome data in variant discovery.

Keywords: Base quality score recalibration, Human leukocyte antigen, Local realignment, Variant calling, Whole exome sequencing

2.2. Introduction

Genetic variation is associated with the etiology of human disease and drug response [2, 58, 72]. Completely cataloguing the variants in individual genomes has been pursued largely through whole genome or whole exome sequencing efforts and is essential in disease diagnosis and pharmacogenomics studies [15, 54]. Whole exome sequencing is widely used in clinical settings, owing to the lower cost compared to whole genome sequencing, and has had remarkable success in identifying causative mutations underlying Mendelian diseases [4, 7, 73].

Variants (SNPs and short INDELs) in sequencing data are identified mainly through mapping-based approaches [17, 47, 74], in which raw reads are first mapped to a reference sequence, and the sites differing between reads and the reference are then identified by variant calling [17, 18]. Several variant calling algorithms have been developed, such as SAMtools [28], the Genome Analysis Toolkit (GATK) UnifiedGenotyper and HaplotypeCaller [29], FreeBayes [34], Platypus [4], Atlas2 Suite [75], and the SNP and INDEL callers in the Short Oligonucleotide Analysis Package (SOAP, http://soap.genomics.org.cn/) [31]. Of these, FreeBayes, GATK HaplotypeCaller, and Platypus are haplotype-based callers which implement De Bruijn graph-based local assembly [4, 29] or constructs haplotypes directly from mapped reads [34].

As outlined in the GATK Best Practices [16], some variant discovery pipelines perform post-processing of alignments prior to variant calling, with the expectation that this practice would improve variant calling accuracy [76]. The post-processing typically includes duplicate marking, local realignment around known INDELs, and base quality score recalibration (BQSR) [15].

Duplicates are reads or pairs of reads that are mapped to the same genomic location and the same strand. They are prevalent in both whole genome and exome sequencing data [18], believed to be artifacts from the polymerase chain reaction (PCR) amplification of the same DNA molecule during library preparation [15]. The inclusion of duplicates more likely gives rise to erroneous calls, for two reasons. First, the presence of duplicates would alter the ratio of reads supporting one allele versus the alternative one at heterozygous sites; also, some duplicates might carry errors introduced by the PCR amplification, thus complicating the variant calling [31]. Duplicates can be identified using Picard command-line tool MarkDuplicates (https://broadinstitute.github.io/picard/command-line-overview.html), such that only one of the duplicates will be used in the subsequent variant calling. Marking duplication is more effective for INDEL calling than for SNP calling [18], and in cases when the coverage is low since duplicated reads play more significant roles than in high coverage data [32].

Reads spanning INDELs have a high chance of being aligned incorrectly to the reference [32, 35]. A previous study revealed that BWA generated misalignment for over 15% of the reads spanning known homozygous INDELs [15]. Without correction, those misaligned bases in reads can be easily called as spurious variants. Finally, the confidence level in variant calling depends on the accuracy of both base calling and base quality score [33]. The latter measures the probability that a base is called incorrectly [36], with a Phred-scale quality score of Q ideally corresponding to an error rate of $10^{-Q/10}$ [33]. However, the raw quality score does not truly reflect the base-calling error rate; instead, it varies with multiple factors including sequencing platform, the number of machine cycles at which a base is sequenced, and the sequence composition [15, 33]. For example, the bases at the 3' end of a read are typically more error prone than those at the 5' end [31]. The GATK BaseRecalibrator and PrintReads commands can be used to calculate the recalibrated quality score, thereby to improve variant calling accuracy [33].

The effect of local realignment and BQSR has been assessed using both whole genome and whole exome sequencing data, mostly for BWA together with GATK UnifiedGenotyper, GATK HaplotypeCaller, or SAMtools [2, 15, 18, 32]. However, the results were inconsistent. The earliest study revealed that local realignment corrected misalignments at approximately 1.8 million sites in the whole genome data and over 100,000 sites in the whole exome data from NA12878 [15]. They found that duplicate marking, local realignment, and BQSR together removed about 2.5-6% of the raw SNP calls, of which the vast majority were false positives. Using exome data from breast cancer patients, a similar effect was detected for local realignment but not for BQSR in GATK UnifiedGenotyper and SAMtools calling [32]. Furthermore, using high (55-65x) coverage whole genome sequencing data in NA12878, Li (2014) found no effect from both local realignment and BQSR in the above three callers [18]. As local realignment and BQSR were evaluated only for a few variant discovery methods, their effectiveness for other methods remains unexplored. Local realignment is a computationally intensive process. In principle, haplotype-based callers apply local de novo assembly (like GATK HaplotypeCaller and Platypus) or build haplotypes directly from mapped reads (like FreeBayes), raising the question whether local realignment is indeed needed for these callers.

Finally, in BQSR, the algorithm needs to identify a list of supposedly non-polymorphic sites that do not overlap with any known polymorphic sites (like those in dbSNP) and that do not match the reference sequence either. It then builds a linear model taking into account the raw base quality score, base position in the reads (i.e., sequencing cycle), as well as the dinucleotide composition of those non-polymorphic sites, by which the recalibrated quality score is computed [15, 33]. While it is relatively straightforward to identify non-polymorphic sites in ordinary genomic regions where sequence diverges at the level of only about 0.1% [21], difficulty arises in regions of high divergence. In the latter case, a portion of the non-polymorphic sites likely represents true variants. In fact, there are over one hundred highly divergent regions in the human genome [38], and some reach the level of 10-15% divergence between haplotypes [41]. Many are clinically important, including the well-known human leukocyte antigen (HLA) region on chromosome 6p21 that are implicated in more than 100 diseases [42, 77, 78]. Thus, the effect of BQSR on variant calling in those highly divergent regions needs to be assessed separately.

In this study, we sought to examine the impact of local realignment and BQSR on a panel of 25 variant discovery methods, taking into account both coverage depth and divergence level. We began with simulated data and extended to five public exome-seq data in NA12878, which were generated using two exome capture kits on three Illumina HiSeq platforms. We found that local realignment mainly impacts INDEL calling and BQSR largely affects SNP calling. In the former, noticeable effect was restricted to only a few methods, with no gains from using Novoalign and limited gains on the three haplotype-based callers. As for BQSR, we found that it reduces SNP calling sensitivity for many of the methods. Thus, consideration should be given to the mapper and caller used, along with coverage depth, when deciding whether to apply the computationally intensive post-processing steps to exome data.

2.3. Methods

2.3.1. Simulation of exome-seq reads

Read simulation provides an ideal approach for initial assessment of individual methods in a controlled situation where the “true” variants are predefined [50]. We focused on chromosome 6, which contains the most highly divergent HLA region. To simulate exome-seq reads from this chromosome, we generated candidate regions by merging hg19 refGene exons (as of 12/10/2013, extended by +/-100 bp) with regions interrogated by any of the four Agilent SureSelectXT Human All Exon kits (All Exon 50Mbp, All Exon V4, All Exon V4+UTRs, and All Exon V5+UTRs, http://www.agilent.com). Simulation was done using Dwgsim (https://github.com/nh13/DWGSIM/wiki) at six mutation rates (0.05%, 0.01%, 0.5%, 1%, 5%, and 10%). Paired 100-base reads were simulated to an average per-base coverage of 100x, which were down-sampled to 40x and 5x coverage. Dwgsim was run with the following parameter settings: SNP-to-INDEL ratio of 9:1 per mutation rate, INDEL size of 1 to 20 base pairs (bp), inner distance of 200 bp between paired reads, no error, no random DNA, and a random seed of 123. To enable the assessment of BQSR, rather than relying on the dummy base quality score assigned by Dwgsim, we instead gave each read a string of empirical quality scores randomly taken from a pool of real Illumina sequencing data. They were from our internal 100-bp exome-seq data in chronic lymphocytic leukemia patients. A similar approach was adopted by the sequence simulation tool ART, which uses base quality scores from real sequencing data [79].

2.3.2. Mapping and post-processing of simulated reads

Reads were aligned to the hg19 human reference sequence using five mappers, including BWA [28], GSNAP [26], NextGenMap [25], Novoalign (http://www.novocraft.com/), and Stampy [21]. Unlike BWA, the other four mappers could map reads to highly divergent regions. The parameter settings for each mapper were as follows.

BWA (v0.5.9) [28] was run with the following parameters:

```
bwa aln ref.fa end1.fastq > end1.sai
bwa aln ref.fa end2.fastq > end2.sai
bwa sampe -a 1000 -f out.sam ref.fa end1.sai end2.sai end1.fastq end2.fastq
```

GSNAP (v2013-10-25) [26] was run with the following options:

```
gsnap -d ref.gmapdb -D ref.gmapdb -k 13 --orientation FR --max-mismatches 0.1 --maxsearch 1 --npaths 1 --ordered --show-refdiff -A sam end1.fastq end2.fastq > out.sam
```

NextGenMap (v0.4.9) [25] was run with the following options:

```
ngm -r ref.fa -1 end1.fastq -2 end2.fastq -k 15 -I 0 -X 800 -i 0.8 -n 1 -p -o out.sam
```

Novoalign (v3.01.011, www.novocraft.com) was run with the following options:

```
novoalign --hdrhd off -i PE 425,80 -r Random -F STDFQ -v 90 -x 5 -o SAM -d ref.nix -f end1.fastq end2.fastq > out.sam
```

Stampy (v1.0.21) [21] was run with the following options:

```
stampy.py -g ref -h ref --substitutionrate=0.1 -f sam --inputformat=fastq -o out.sam -M end1.fastq end2.fastq (for simulated data)
stampy.py -g ref -h ref --substitutionrate=0.1 -f sam --inputformat=fastq --xa-max=3 --xa-max-discordant=10 -o out.sam -M end1.fastq end2.fastq (for real data)
```

Alignments in the sequence alignment/map (SAM) format were converted into the binary alignment/map (BAM) format using SAMtools [80], and sorted by coordinate using the SortSam command in the Picard tools (http://picard.sourceforge.net/). Alignments from all chromosomes were processed sequentially through duplicate marking, local realignment, and BQSR.

Duplicates were marked using Picard MarkDuplicates command. We then performed six local realignments per dataset, designated LR-40 to LR-90, using GATK IndelRealigner command; they targeted random subsets representing 40% to 90% (with a 10% increment) of simulated INDELs. Subsequently, the GATK BaseRecalibrator and PrintReads commands were used in six runs of BQSR, referred to as BQSR-40 to BQSR-90, based on 40% to 90% (with a 10% increment) of simulated variants. In this process, LR-40 was followed by BQSR-40 and LR-50 by BQSR-50, and so on. To investigate how these two steps impact variant calling, variants were identified both before and after each step, as below.

2.3.3. Variant calling from simulated data

We selected five popular variant callers, including GATK UnifiedGenotyper and HaplotypeCaller [15, 29], FreeBayes [34], SAMtools mpileup [80], and Platypus [4]. The parameter settings for each of the callers were provided below.

GATK UnifiedGenotyper (v2.7-2) [15, 29]:

```
java –jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R ref.fa  -L chr6 -L
chr6.interval_list -isr intersection -glm BOTH -mbq 17 --dbsnp
dbsnp_135.hg19.vcf.gz -stand_call_conf 20 -stand_emit_conf 10 -o out.vcf -I
align.bam
```

Note: chr6.interval_list is in the format of chr:start-end.

GATK HaplotypeCaller (v2.7-2) [15, 29]:

```
java –jar GenomeAnalysisTK.jar -T HaplotypeCaller -R ref.fa  -L chr6 -L
chr6.interval_list -isr intersection --genotyping_mode DISCOVERY --dbsnp
dbsnp_135.hg19.vcf.gz -stand_call_conf 20 -stand_emit_conf 10 -o out.vcf -I
align.bam
```

FreeBayes (v9.9.2-27) [34]:

```
freebayes -f ref.fa -t chr6.bed -C 2 -3 40 -P 0.0001 -m 0 -q 17 -W 1,3 -S 4 -M 3 -B
25 -E 3 -v out.vcf -b in.bam
```

Note: chr6.bed is in the format of chr<TAB>start<TAB>end

Platypus (v0.5.2) [4]:

```
Platypus.py callVariants  --refFile=ref.fa --regions=chr6.interval_list --ploidy=2 --
maxVariants=15 --minBaseQual=17 --output=out.vcf --bamFiles=align.bam --
maxVariants=50 minMapQual=0 --rmsmqThreshold=0  --hapScoreThreshold=0
```

SAMtools mpileup (v0.1.19) [80]:

```
samtools mpileup -ugf ref.fa -l chr6.bed -q 0 -Q 17 -B -F 0.002 align.bam |
bcftools view -bvcg - > out.bcf
bcftools view out.bcf > out.vcf
```

To increase the comparability among these callers, the multiple-nucleotide polymorphisms (i.e., multiple SNPs within five bases that are reported as a single event) reported by Platypus and FreeBayes were decomposed into individual variants using GATK walker VariantsToAllelicPrimitives. Considering that the boundaries of INDELs are difficult to define precisely, overlap was inferred if an INDEL was called within five bases of a permuted INDEL.

Variant calling sensitivity and precision rate were estimated using GATK walker GenotypeConcordance, where sensitivity is estimated as “true positive/(true positive + false negative)” and precision rate as “true positive/(true positive + false positive)”. Here, true positive refers to the number of true (simulated) variants identified by a method, false negative represents the number of true variants missed by a method, and false positive represents the number of called variants that does not overlap the true variants. The effect of local realignment was measured as the change in sensitivity and precision rate before and after local realignment, and that of BQSR was measured similarly.

2.3.4. Variant calling from exome-seq data in NA12878

We further assessed local realignment and BQSR using exome-seq data from NA12878, in which several call sets have already been generated from whole genome and exome sequencing data [51, 54]. We first compiled a list of "true" variants (referred to as public call set) by combining the four variant lists below. The high-confident call set was generated through integrative analyses of 11 whole genome and 3 exome sequencing data (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/GIAB_integration/NIST_RTG_PlatGen_merged_highconfidence_v0.2.primitives.vcf.gz) [51]. To minimize bias toward a specific analytical pipeline, they used seven mappers together with three callers, Cortex [57], GATK UnifiedGenotyper, and GATK HaplotypeCaller. The other three call sets were generated using Cortex, DISCOVAR, and GATK HaplotypeCaller, from a PCR-free genomic library sequenced to 250 base pairs (http://genepattern.broadinstitute.org/ftp/distribution/crd/DiscovarManuscript/vcf/) [54].

We downloaded five paired-end exome-seq data in NA12878 that were not used to generate the above public call set (**Table 2.1**). Four of them had 66-100x coverage; SRR1919605 had 200x coverage, from which a maximum coverage of 100x was used. These data were chosen to represent two different capture kits and three Illumina sequencing platforms. The Roche SeqCap EZ Human Exome kit has 64 megabases of capture regions, versus only 37 megabases interrogated by Illumina Nextera Rapid Capture Exome kit. For the two 150-bp datasets, only the first 100 bases were used. To understand how the effect of post-processing varies across different coverage depths, each full (66-100x) dataset was down-sampled to 40x, 20x, and 10x coverage using SAMtools.

Reads mapping, post-processing, and variant calling followed the procedure used in the simulated data, with two modifications. First, local realignment was performed around the Mills and 1000G gold standard INDELs (Mills_and_1000G_gold_standard.indels.hg19.vcf.gz); and second, BQSR was conducted using known variants in dbSNP v135 together with the Mills and 1000G gold standard INDELs. While post-processing was performed genome-wide, variants were identified only from chromosome 6. Variants with a phred-scale quality of at least 20 were retained, which were separated into "known" and "novel" ones by intersecting with dbSNP v138 (ftp://ftp.broadinstitute.org/bundle/2.8/hg19/dbsnp_138.hg19.vcf.gz). By treating the above public call set as "true" positives, the sensitivity and precision rate were estimated, as done for the simulated data.

Table 2.1 Five public exome-seq data in NA12878

Sequence ID	Name	Approx. coverage	Length (bp)	Platform	Capture kit
NA12878-NGv3-LAB1360-A	NA12878-NGv3	66	100	HiSeq 2000	Roche
SRR1611181	SRR1611181	90	100	HiSeq 2000	Roche
FC1_NA12878_01	NA12878_01	100	150	HiSeq 2500	Illumina
FC1_NA12878_04	NA12878_04	100	150	HiSeq 2500	Illumina
SRR1919605	SRR1919605	100	76	HiSeq 4000	Illumina

The 76-bp and 150-bp reads are available at https://basespace.illumina.com/. SRR1611181 was downloaded from the National Center for Biotechnology Information Sequence Read Archive. NA12878-NGv3-LAB1360-A reads were downloaded from https://s3.amazonaws.com/bcbio_nextgen/. Capture kit: Roche, SeqCap EZ Human Exome kit v3.0 (with 64 Mb targets); Illumina, Nextera Rapid Capture Exome kit (with 37 Mb targets).

2.4. Results and discussion

We used both simulated reads from exonic regions of chromosome 6 and five exome data in NA12878. We assessed, for each of the mapper-caller combinations, how local realignment and BQSR impacted variant discovery across different divergence levels and coverage depths.

2.4.1. Impact of local realignment in simulated data

We assessed a total of 280 cases obtained from five mappers, five callers, two coverage depths (5x and 40x), and six divergence levels. For each case, six proportions (40-90%) of permuted INDELs (LR-40 to LR-90) were used in local realignment.

We first examined whether there is noticeable difference among the six local realignments. In SNP calling, both sensitivity and precision rate varied little (<=0.1% difference) across the six local realignments. Similar results were obtained in INDEL calling for the precision rate (<0.1% difference in all but one case). For three of the methods, however, there were modest differences in INDEL calling sensitivity at both coverage depths. They are SAMtools, Platypus, and GATK UnifiedGenotyper with NextGenMap, on which we observed a maximal difference of approximately 1.5% at 0.05-1% divergence and of approximately 0.5% at 5-10% divergence (**Figures 2.1M-2.1O** and **2.2M-2.2O**). For these three methods, using a higher proportion of permuted INDELs led to a bigger change (either increase or decrease) in sensitivity. In the other cases, there was nearly no difference (<0.2%) in INDEL calling sensitivity (**Figures 2.1A-2.1L, 2.1P-2.1Y, 2.2A-2.2L** and **2.2P-2.2Y**). We thus evaluated the effect of local realignment using 90% of the permuted INDELs, as below.

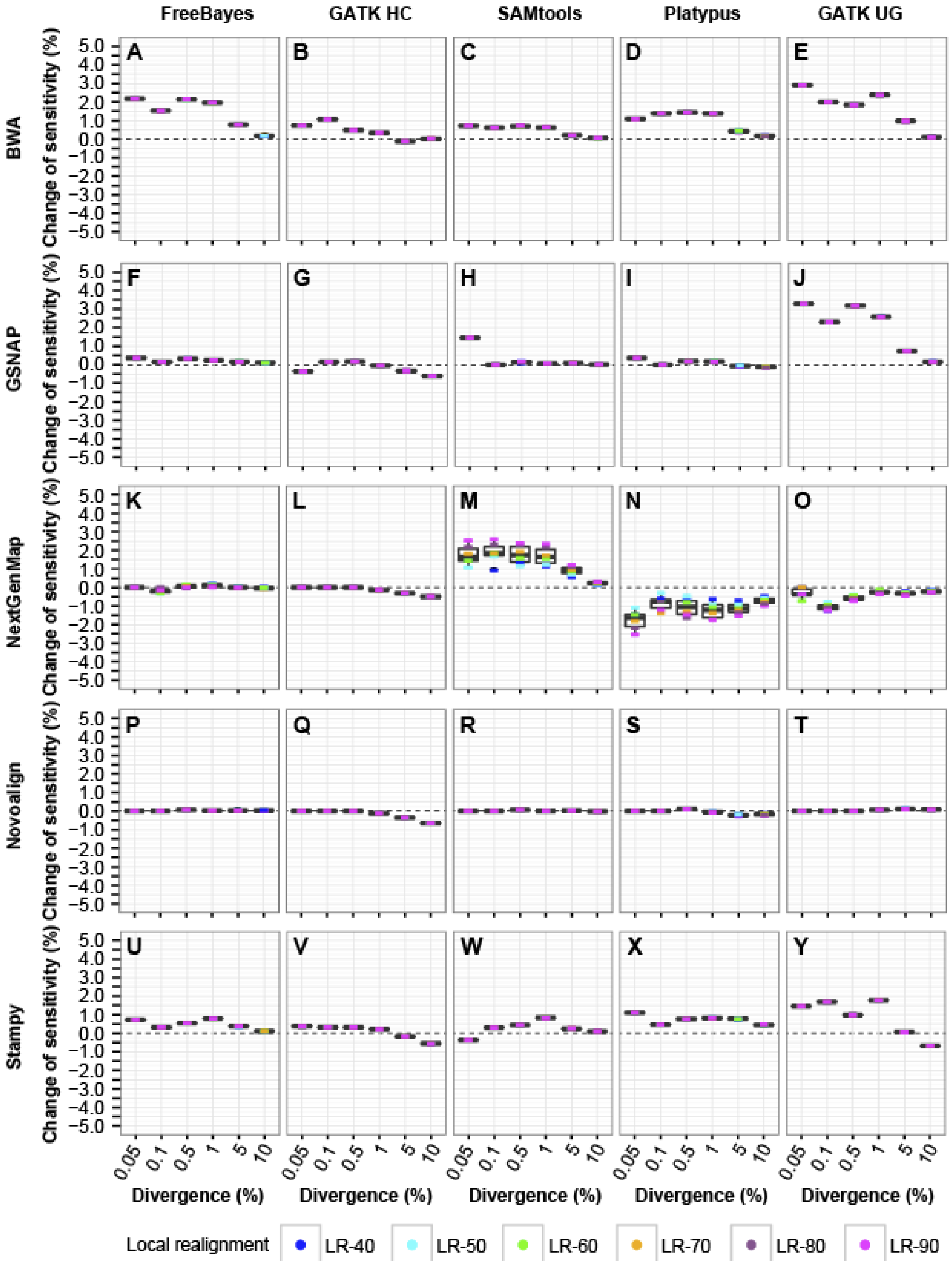


Figure 2.1 Change of INDEL calling sensitivity by local realignment at 5x coverage. Alignments from five mappers were subjected to duplicate marking and then to local realignment. The change in INDEL calling sensitivity (*y*-axis) is calculated as the sensitivity after local realignment, subtracted by that after duplicate marking. Each box plot displays measurements from six different proportions of preplaced INDELs (between 40% and 90%, with 10% increment, termed LR-40 to LR-90) used in the local realignment. For callers together with BWA, only datasets with 0.05-1% divergence are included. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

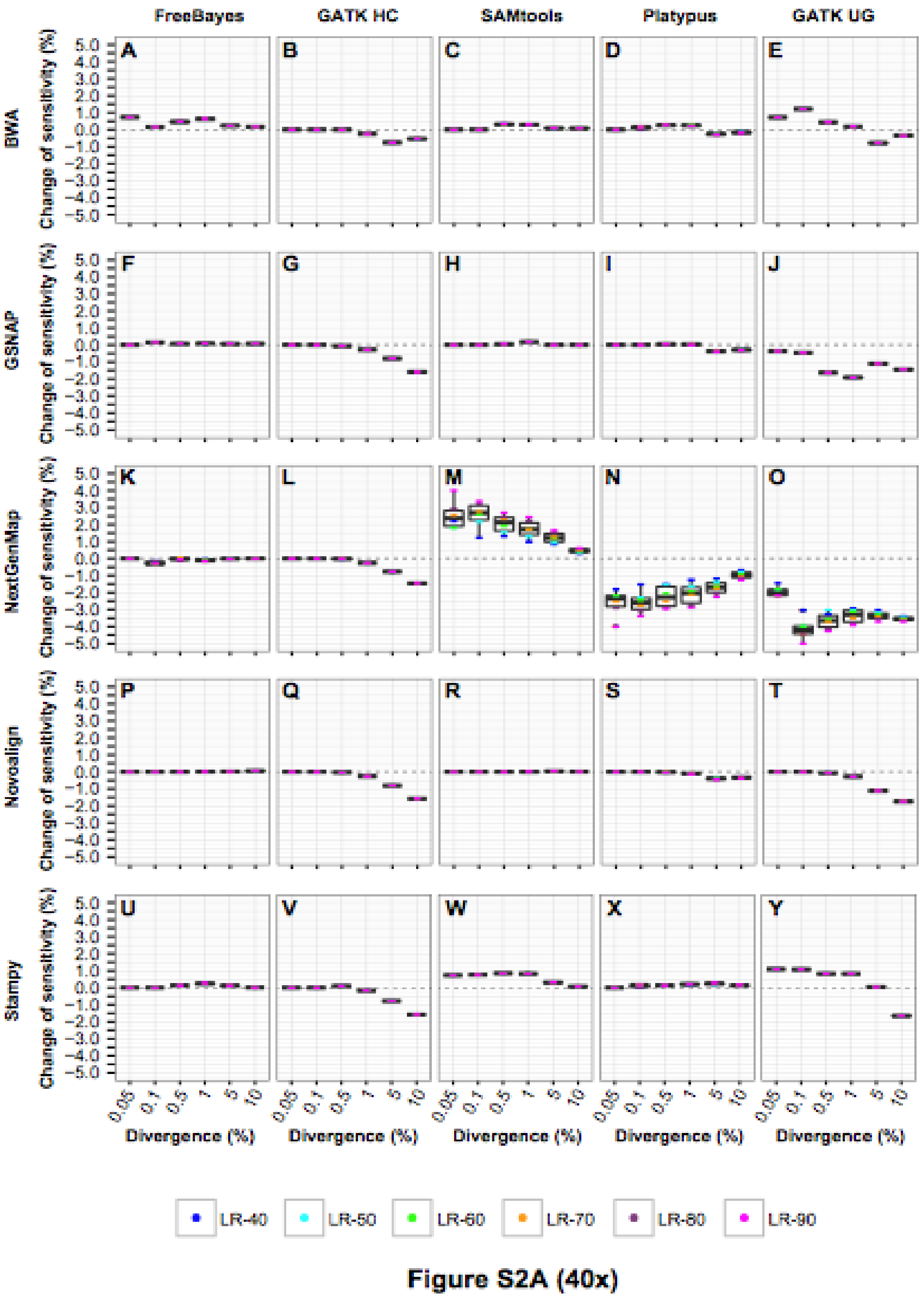


Figure 2.2 Change of INDEL calling sensitivity by local realignment at 40x coverage. See Figure 2.1 legend for details.

Table 2.2 Change of SNP calling sensitivity and precision rate after local realignment

Metrics	Change (%)	No. cases	% cases
Sensitivity	-0.1	4	1.43
Sensitivity	0	242	86.43
Sensitivity	0.1	33	11.79
Sensitivity	0.2	1	0.36
Precision rate	-0.1	1	0.36
Precision rate	0	224	80
Precision rate	0.1	44	15.71
Precision rate	0.2	10	3.57
Precision rate	0.3	1	0.36

In each case, the change of SNP calling sensitivity is calculated as the sensitivity after local realignment using 90% simulated variants, subtracted by that after duplicate marking. The change in precision rate is calculated in the same way. See Table 2.3 for details.

Table 2.3 Effect of local realignment in INDEL calling

Metrics	Change (%)	Cases	
		No.	%
Sensitivity	-5.1 − -4	2	0.71
Sensitivity	-4 − -3	5	1.79
Sensitivity	-3 − -2	5	1.79
Sensitivity	-2 − -1	17	6.07
Sensitivity	-1 − -0.2	34	12.14
Sensitivity	-0.2 − 0.2	129	46.07
Sensitivity	0.2 – 1	53	18.93
Sensitivity	1 – 2	19	6.79
Sensitivity	2 – 3	12	4.29
Sensitivity	3 – 4	4	1.43
Precision rate	-0.2 – 0	33	11.79
Precision rate	0 − 0.2	219	78.21
Precision rate	0.2 − 0.4	12	4.29
Precision rate	0.4 − 0.6	8	2.86
Precision rate	0.6 − 0.8	5	1.79
Precision rate	0.8 – 1	3	1.07

Alignments were subjected to duplicate marking and then to local realignment. A total of 280 cases were evaluated, which represent combinations among five mappers, five callers, six divergence levels, and two coverage depths (5x and 40x), excluding 20 cases with BWA mapping at 5-10%. In each case, the change of INDEL calling sensitivity is calculated as the sensitivity after local realignment using 90% simulated INDELs, subtracted by that after duplicate marking. The change in precision rate is calculated in the same way.

Local realignment had nearly no impact on SNP calling sensitivity (**Table 2.2**). Precision rate increased slightly (0.2-0.3%) in 11 cases (**Table 2.2**), involving GATK UnifiedGenotyper and Platypus. Next, we sought to understand to what extent local realignment might impact INDEL calling. We observed a 0.4-1% increase of precision rate in 16 (5.7%) cases (**Table 2.3**). They were from cases that used SAMtools and GATK UnifiedGenotyper together with Stampy (**Figure 2.3**). Little to no effect from local realignment was detected in the other cases, including all those that used any of the three haplotype-based callers (**Figure 2.3; Table 2.3**). On the other hand, local realignment led to obvious gain or loss of sensitivity (by 1-5%) in approximately one-fifth of the cases (**Table 2.3**), which depended on multiple factors as shown below (**Figure 2.4**).

In INDEL calling, the impact of local realignment on sensitivity is mapper-, caller-, and coverage-dependent. As for the mappers, NextGenMap was affected the most and Novoalign the least (**Figure 2.4**). For Novoalign, nearly no changes in sensitivity were detected at low divergence (**Figure 2.4**), likely reflecting the optimal alignments achieved by the underlying full Needleman-Wunsch algorithm. On the other hand, local realignment increased the sensitivity of NextGenMap with SAMtools (**Figures 2.4C** and **2.4H**) but decreased its sensitivity with Platypus (**Figures 2.4D** and **2.4I**) and GATK UnifiedGenotyper (**Figures 2.4E** and **2.4J**). A previous study showed that local realignment improved GATK UnifiedGenotyper calling accuracy on BWA alignment [15]. We indeed found that it generally increased sensitivity for BWA with all the callers at low (0.05-1%) divergence, more obvious at 5x coverage (**Figures 2.4A-2.4E**). For GSNAP, local realignment increased the sensitivity at 5x but decreased the sensitivity at 40x coverage in GATK UnifiedGenotyper calling (**Figures 2.4E** and **2.4J**). Of the callers, overall GATK HaplotypeCaller, Platypus and FreeBayes were less affected at 40x coverage and low divergence, compared to the other two callers (**Figures 2.4F-2.4J**). A similar finding was reported for GATK HaplotypeCaller in exome-seq data [81]. These haplotype-based callers are capable of alleviating alignment ambiguity around INDELs internally through local de novo assembly [4, 16] or direct construction of haplotypes from alignments [34]. For these callers, the benefit of applying local realignment in GATK Best Practices would be minimized.

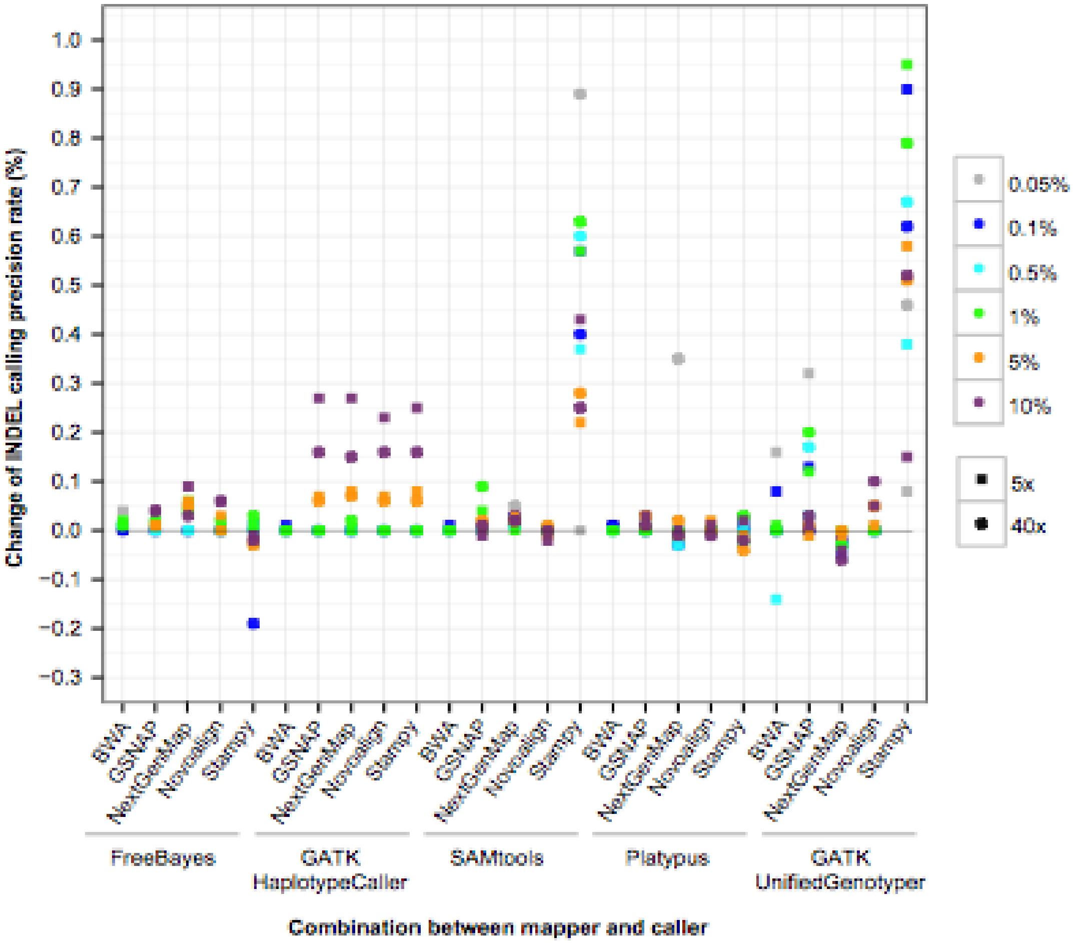


Figure S2

Figure 2.3 Change of INDEL calling precision rate following local realignment. The change in precision rate is calculated as the precision rate after local realignment, subtracted by that after duplicate marking. As precision rates are highly comparable among the six different proportions (40-90%) of preplaced INDELs used in local realignment, only local realignments using 90% INDELs (LR-90) are displayed. For each of the mapper-caller combinations (*x*-axis), the plot shows the distribution of changes in precision rate across six divergence levels (0.05-10%, by color) and two coverage depths (5x and 40x, by shape). For the combinations with BWA mapping, only 0.05-1% divergence levels are shown.

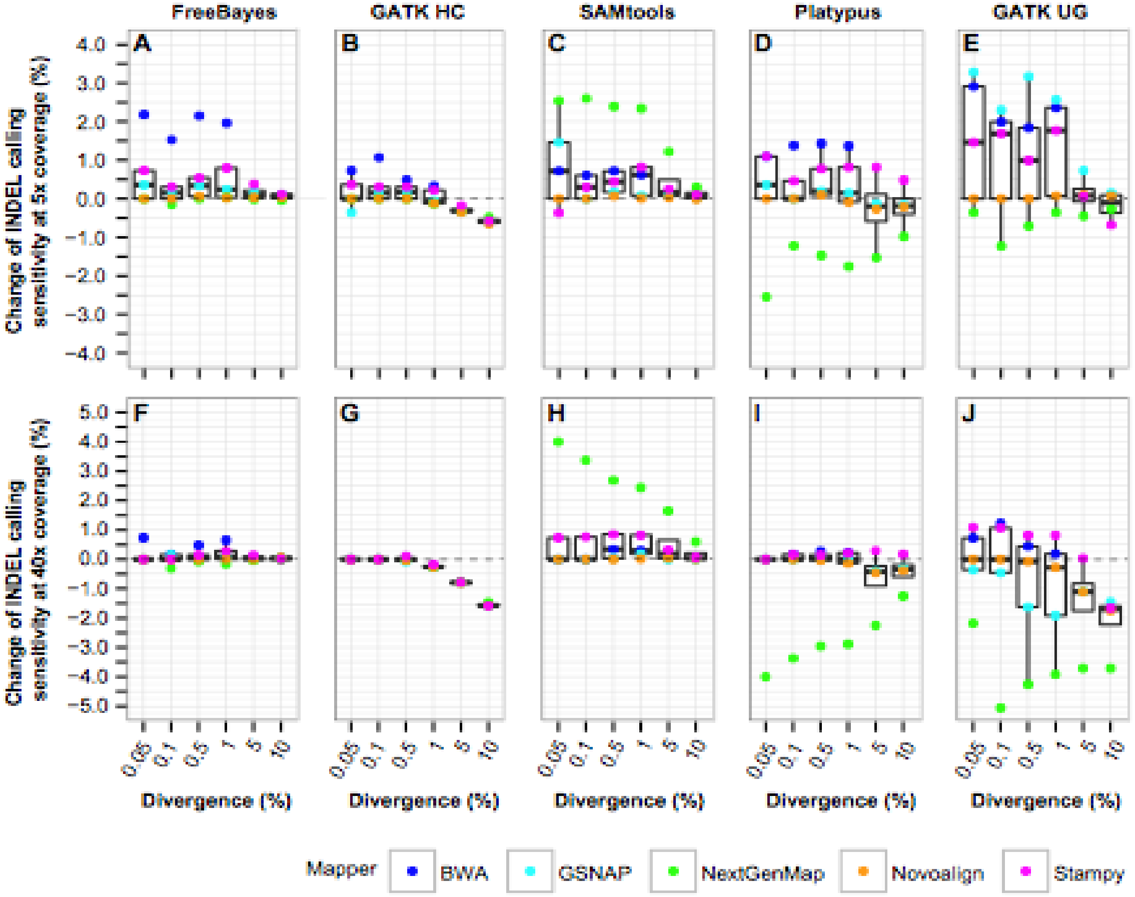


Figure 1

Figure 2.4 The change of INDEL calling sensitivity after local realignment. (**A-E**) Simulated datasets with 5x coverage. (**F-J**) Simulated datasets with 40x coverage. Local realignment was performed using 90% of the preplaced INDELs (LR-90). The change of INDEL calling sensitivity is calculated as the sensitivity after local realignment, subtracted by that after duplicate marking. For callers together with BWA, only datasets with 0.05-1% divergence were used. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

The role of local realignment in INDEL calling also depends on divergence. At 0.05-1% divergence, 33 cases showed increase of sensitivity (>=1%) after local realignment, versus only 2 (NextGenMap+SAMtools) at 5-10% divergence. Whereas, of the 29 cases whose sensitivity decreased by >=1%, they were equally from 0.05-1% and 5-10% divergence. In summary, local realignment mainly affects INDEL calling sensitivity, more conspicuous at low coverage. When the coverage and divergence are both low, local realignment tends to increase the sensitivity. At high coverage and low divergence, local realignment has nearly no impact on the haplotype-based callers. Of the five mappers, Novoalign is least affected, in contrast with BWA which generally shows an increase of sensitivity.

2.4.2. Impact of different BQSR in simulated data

To better understand the effect of BQSR on variant calling, we tested six BQSR (BQSR-40 to BQSR-90) using 40% to 90% of the permuted SNPs. We checked the difference among the six BQSR in SNP and INDEL calling sensitivity. For SNP calling at 5x coverage, we observed 0.2% and 2.3% (median) difference at 0.05-0.1% and 0.5-1% divergence, respectively, with the exception of FreeBayes that showed no variability in sensitivity. Nevertheless, the effect was much less obvious at 40x coverage. At 5-10% divergence, in seven cases the sensitivity dropped to less than 10% after any of the six BQSR; the remainder showed 34.3% (median) difference among the six BQSR (**Figures 2.5A-2.5D)**. Overall BQSR with 90% permuted SNPs performed better than the other options using less permuted SNPs.

For INDEL calling, we assessed the effect of different BQSR in 240 cases: 180 at 0.05-1% divergence and 60 at 5-10% divergence. The other 40 cases were excluded, since their sensitivity was low (<=35%) after local realignment (prior to BQSR). They are from GATK UnifiedGenotyper and SAMtools, the two callers that are less sensitive in INDEL detection (Tian *et al*., unpublished results). At 0.05-1% divergence, the difference in sensitivity was generally below 0.5%. However, at 5-10% divergence, the difference reached 9.6-80.5% in two-thirds (45/60) of the cases (**Figures 2.5E-2.5H**). As in SNP calling, BQSR using 90% permuted SNPs performs better than the other five options.

We therefore evaluated the effect of BQSR by measuring the change in sensitivity and precision rate between BQSR and local realignment that use 90% of the permuted variants (i.e., BQSR-90 – LR-90). To ensure meaningful comparison, we limited the analysis to a subset of the cases. In assessing the change in sensitivity, we required that the sensitivity prior to BQSR should exceed 35%, while in assessing the change in precision rate, we further required that the sensitivity should not decrease by more than 15% after BQSR.

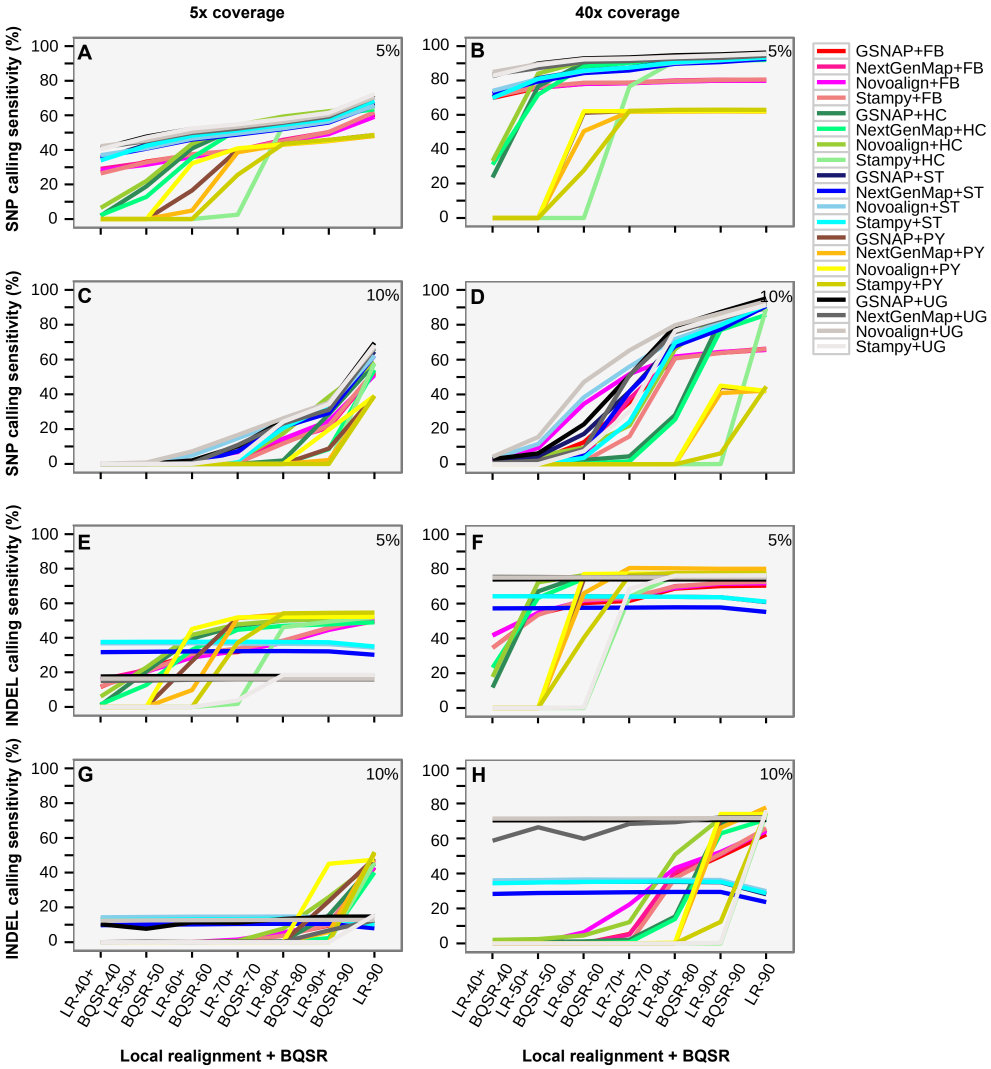


Figure 2.5 Effect of BQSR on variant calling sensitivity at high divergence. (**A-B**) SNP calling sensitivity at 5% divergence. (**C-D**) SNP calling sensitivity at 10% divergence. (**E-F**) INDEL calling sensitivity at 5% divergence. (**G-H**) INDEL calling sensitivity at 10% divergence. *X*-axis shows six trials of local realignment followed by BQSR, as well as local realignment using 90% of the preplaced INDELs (LR-90). *Y*-axis denotes the corresponding sensitivity in SNP or INDEL calling. BQSR-40 to BQSR-90 refers to base quality score recalibration (BQSR) using 40% to 90% (with a step size of 10%) of preplaced variants. LR-40 to LR-90 refers to local realignment using 40% to 90% (with a step size of 10%) of preplaced INDELs. BQSR was performed after local realignment, and both used the same proportion of preplaced variants; for example, BQSR-90 was performed after LR-90. LR-90 alone is used as the baseline to illustrate the effect of BQSR on variant calling. FB, FreeBayes; HC, GATK HaplotypeCaller; PY, Platypus; ST, SAMtools; UG, GATK UnifiedGenotyper.

2.4.3. Impact of BQSR in SNP calling of simulated data

The impact of BQSR on SNP calling is striking at low coverage and high divergence (**Figure 2.6**). We observed a trend of decrease in sensitivity by BQSR following the increase in divergence. At low divergence and 40x coverage, of the five callers, only GATK UnifiedGenotyper showed a small (0.3-0.5%) decrease in sensitivity at 0.5-1% divergence (**Figures 2.6D** and **2.6E**; **Table 2.4**). Nevertheless, at 5x coverage the effect was more obvious for several callers (**Figures 2.6A** and **2.6B**; **Table 2.4**). More specifically, BQSR largely increased the sensitivity at 0.05-0.1% divergence; at 0.5-1% divergence, it increased the sensitivity for GATK HaplotypeCaller but decreased the sensitivity for GATK UnifiedGenotyper and Platypus. At 5-10% divergence, BQSR reduced the sensitivity in vast majority of the cases (**Figures 2.6C** and **2.6F**; **Table 2.4**). At 5x coverage, sensitivity was decreased by 15.5% (median), versus 1.5% at 40x coverage (**Table 2.4**).

Table 2.4 Effect of BQSR in SNP and INDEL calling

Type	Metric	Div	5x coverage				40x coverage			
		(%)	Case	Min	Max	Mdn	Case	Min	Max	Mdn
SNP	Sensitivity	0.05-0.1	50	-0.1	1.6	0.6	50	0	0.2	0
SNP	Sensitivity	0.5-1	50	-4.5	0.9	-0.6	50	-0.5	0.1	-0.1
SNP	Sensitivity	5	20	-11.4	-2.6	-9.8	20	-1.8	0.2	-0.5
SNP	Sensitivity	10	20	-57.8	-19.5	-32.3	20	-89.4	3	-7.1
SNP	Precision rate	0.05-1	100	-0.1	0.3	0	100	-0.2	0.1	0
SNP	Precision rate	5	20	0	0.3	0.1	20	0	0.2	0
SNP	Precision rate	10	0	-	-	-	18	-0.3	1.5	0.2
INDEL	Sensitivity	0.05-0.1	40	-1.5	1.8	0.4	50	-0.2	0	0
INDEL	Sensitivity	0.5-1	40	-0.3	1	0.1	50	-0.1	0.1	0
INDEL	Sensitivity	5	12	-6.4	-0.1	-1.4	20	-0.9	2.8	0
INDEL	Sensitivity	10	12	-51.9	-2.4	-34.6	16	-75.3	0.1	-9.7
INDEL	Precision rate	0.05-1	80	-0.1	0	0	100	-0.1	0.2	0
INDEL	Precision rate	5	12	0	0.3	0	20	0	0.8	0
INDEL	Precision rate	10	0	-	-	-	13	-0.2	3.2	0

Between 5 to 25 mapper-caller combinations were assessed at each of the six divergence levels and two coverage depths in simulation, excluding combinations with BWA at 5-10% divergence. They were selected to have a sensitivity of at least 35%; to estimate the change in precision rate, we also required that sensitivity should not decrease by more than 15% after BQSR. Div, divergence; BQSR-90, base quality score recalibration using 90% of the permuted SNPs; LR-90, local realignment using 90% of the permuted INDELs; Mdn, median.

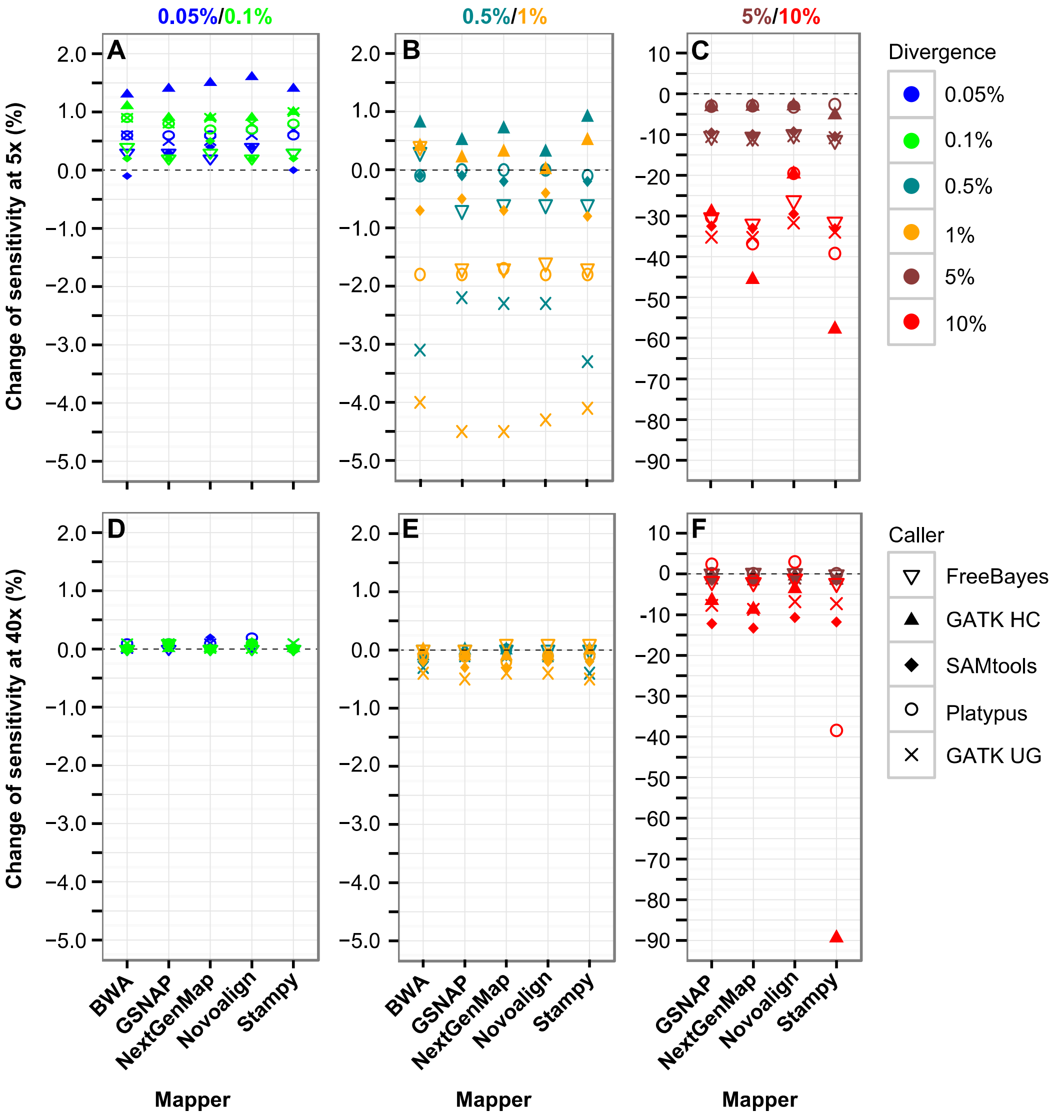


Figure 2.6 Change of SNP calling sensitivity after BQSR. (**A-C**) Datasets with 0.05-0.1%, 0.5-1%, and 5-10% divergence, respectively, at 5x coverage. (**D-F**) Datasets with 0.05-0.1%, 0.5-1%, and 5-10% divergence, respectively, at 40x

coverage. The change of sensitivity is calculated as the sensitivity after BQSR using 90% of the permuted SNPs, subtracted by that after local realignment using 90% of the preplaced INDELs. For callers together with BWA, only datasets with 0.05-1% divergence were used. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

There were only a few cases where BQSR altered the precision rate. Overall the effect was negligible at 0.05-5% divergence (**Table 2.4**). At 10% divergence, we excluded 22 cases whose sensitivity decreased by over 20% after BQSR. In the remaining 18 cases, all at 40x coverage, only SAMtools and GATK UnifiedGenotyper showed 0.2-0.5% increase and Platypus (with GSNAP and NextGenMap) showed 1-1.5% increase in precision rate.

2.4.4. Impact of BQSR in INDEL calling of simulated data

Compared to the three haplotype-based callers, GATK UnifiedGenotyper and SAMtools are less sensitive in INDEL calling (Tian *et al*., unpublished results). In assessing the impact of BQSR on sensitivity, we excluded 40 cases from the two callers due to low (<35%) sensitivity, mainly at 5x coverage. In two-thirds of the remainder, included all those at 0.05-1% divergence and 40x coverage, BQSR had little effect on sensitivity (<=0.3% change, **Table 2.4**). In the other one-third, BQSR either increased (39 cases, by 0.7-2.8%) or decreased (39 cases, by 0.5-75.3%) sensitivity (**Table 2.4**). The former was predominantly from 5x coverage and 0.05-1% divergence (**Figures 2.7A** and **2.7B**) and the latter from high divergence (**Figures 2.7C** and **2.7D**), which we discuss more below.

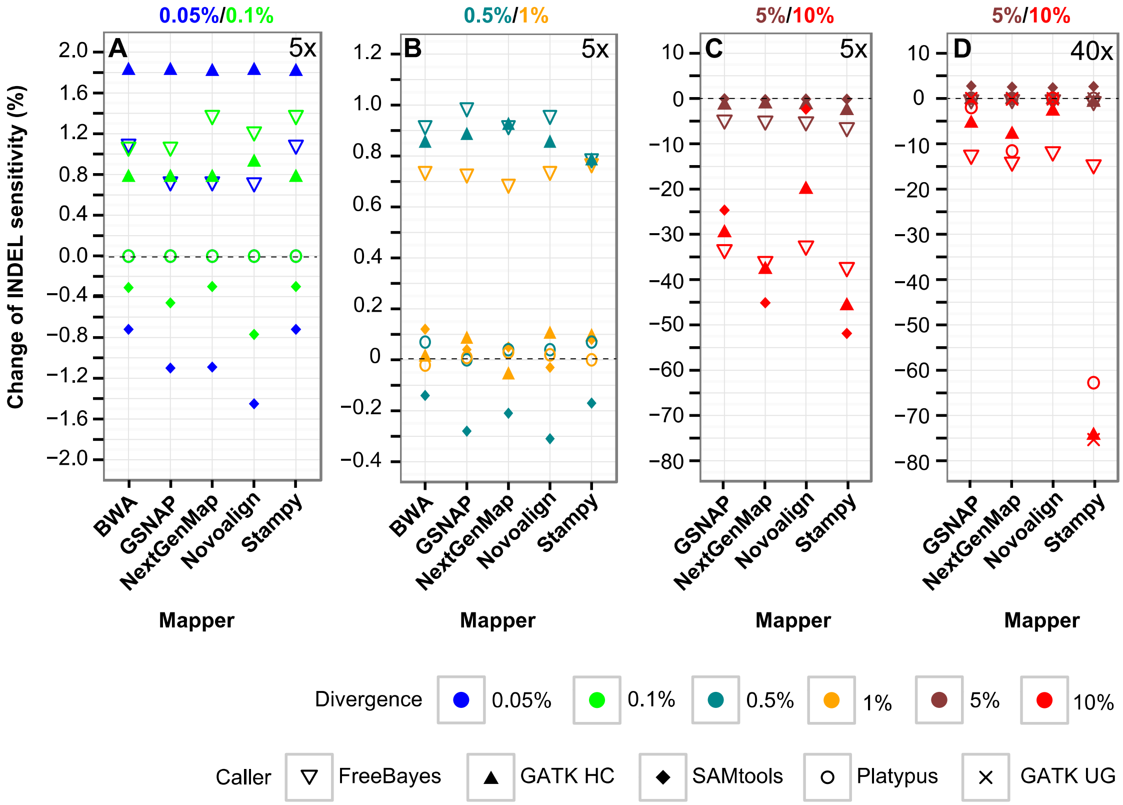


Figure 2.7 Change of INDEL calling sensitivity after BQSR. (**A-C**) Datasets with 0.05-0.1%, 0.5-1%, and 5-10% divergence, respectively, at 5x coverage. (**D**) Datasets with 5-10% divergence and 40x coverage. Datasets with 0.05-1% divergence and 40x coverage were not displayed, since there was nearly no change in sensitivity after BQSR. See Figure 2.6 legend for more information.

For GATK UnifiedGenotyper, the analysis was limited to the 40x coverage (see Methods). For GATK UnifiedGenotyper (with Stampy only) and Platypus, BQSR decreased the sensitivity only at 10% divergence (**Figures 2.7C** and **2.7D**). FreeBayes and GATK HaplotypeCaller were similarly impacted; BQSR increased the sensitivity at 5x and <=1% divergence (**Figures 2.7A** and **2.7B**), but decreased the sensitivity at 5x coverage and 5-10% divergence (**Figure 2.7C**) and at 40x coverage and 10% divergence (**Figure 2.7D**). Finally, for SAMtools, however, BQSR increased the sensitivity at 40x coverage and 5% divergence (**Figure 2.7D**) but generally reduced the sensitivity at 5x coverage and 0.05-0.1% divergence (**Figure 2.7A**). On the other hand, in 95% of the cases, there was nearly no (<0.1%) change in precision rate (**Table 2.4**).

2.4.5. Impact of local realignment in NA12878 exome data

Given the limitations in simulated data [18], we further assessed local realignment and BQSR over five exome-seq data in NA12878. To evaluate the impact of coverage in the post-processing, each full dataset was down-sampled into three subsets with 40x, 20x, and 10x coverage, respectively. To further take into account the sequence divergence, we limited the analysis to chromosome 6 and split it into the HLA region (4-Mb, 29,500-33,500 kb, hg19) of high sequence divergence and the flanking non-HLA regions with low sequence divergence.

In SNP calling, the impact of local realignment on sensitivity is only marginal (<=0.5% change), as observed in the simulated data (**Table 2.2**). In both HLA and non-HLA regions, over 97% of the cases showed <=0.2% change in sensitivity, with no change in about two-thirds of the cases. In INDEL calling, the effect of local realignment depends on mapper, caller, and coverage. Since the trend is highly comparable across all the five datasets, only the results from FC1_NA12878_04 are displayed (**Figures 2.8A** and **2.8B**). Of the five mappers, Novoalign was impacted the least. As for the callers, local realignment had little effect on GATK HaplotypeCaller in INDEL calling sensitivity but more obvious effect on SAMtools and GATK UnifiedGenotyper, consistent with our observation on simulated data. For these two callers, local realignment generally increases the sensitivity on the alignment from GSNAP and NextGenMap. However, the two callers are less effective in INDEL detection. INDELs are more difficult to detect and often under-reported [82]. On chromosome 6, the ratio of known INDELs to SNPs is roughly 1 to 10 in the non-HLA regions and 1 to 20 in the HLA region based on the public call set [51, 54]. Thus, there are only small changes in the number of INDEL calls, with the exception of Stampy+GATK UnifiedGenotyper. This method showed the most striking increase in sensitivity across the full range of coverage. Thus, the benefit of local realignment is not universal but mapper- and caller-dependent.

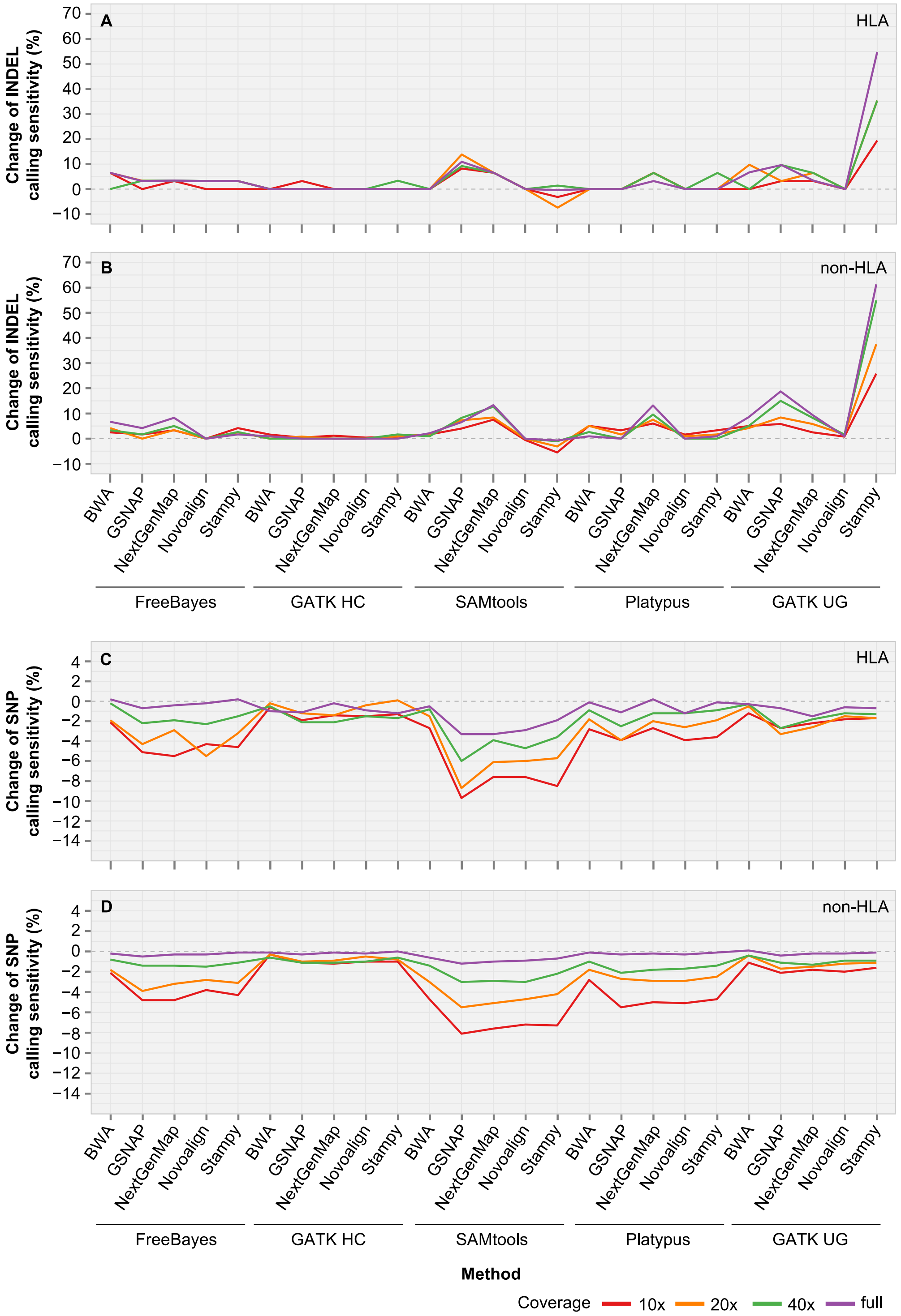


Figure 2.8 Change of variant calling sensitivity after local realignment and BQSR in NA12878. (**A-B**) Change of INDEL calling sensitivity after local realignment. (**C-D**) Change of SNP calling sensitivity after BQSR. See Figure 2.9 legend for more details.

2.4.6. Impact of BQSR in NA12878 exome data

In SNP calling from simulated data, we observed a general decrease of sensitivity at low coverage and/or high divergence (**Figure 2.6**). Using the NA12878 exome-seq data, we found that BQSR decreased SNP calling sensitivity for majority of the methods, especially in cases when the coverage is less sufficient (**Figures 2.8C, 2.8D and 2.9A-2.9H**). The pattern of variability in sensitivity was highly consistent across all five datasets, showing that BQSR had the greatest impact on SAMtools, followed by FreeBayes and Platypus. At full coverage, the effect of BQSR was much reduced in both HLA and non-HLA regions; however, SAMtools still showed a noticeable loss of sensitivity in the HLA region. Two previous studies also revealed little effect from BQSR in high-coverage whole genome and exome sequencing data [18, 32].

BQSR tended to reduce SNP calling sensitivity, i.e., to miss known SNPs annotated in the public call set. We then ask whether loss of sensitivity would be compensated by an increase in precision rate. Precision rate is difficult to estimate for real sequencing data as the false positives are not known. Assuming that some of the novel SNPs represent false positive calls, if BQSR reduces novel SNP calls, then it should be an indication of improvement in precision rate. Therefore, we assessed the impact of BQSR on the detection of known versus novel SNPs in all the five datasets. This analysis revealed two general trends, and we illustrated such trends on two of the datasets: FC1_NA12878_01 (**Figures 2.10A-2.10D**) and FC1_NA12878_04 (**Figures 2.11A-2.11D**).

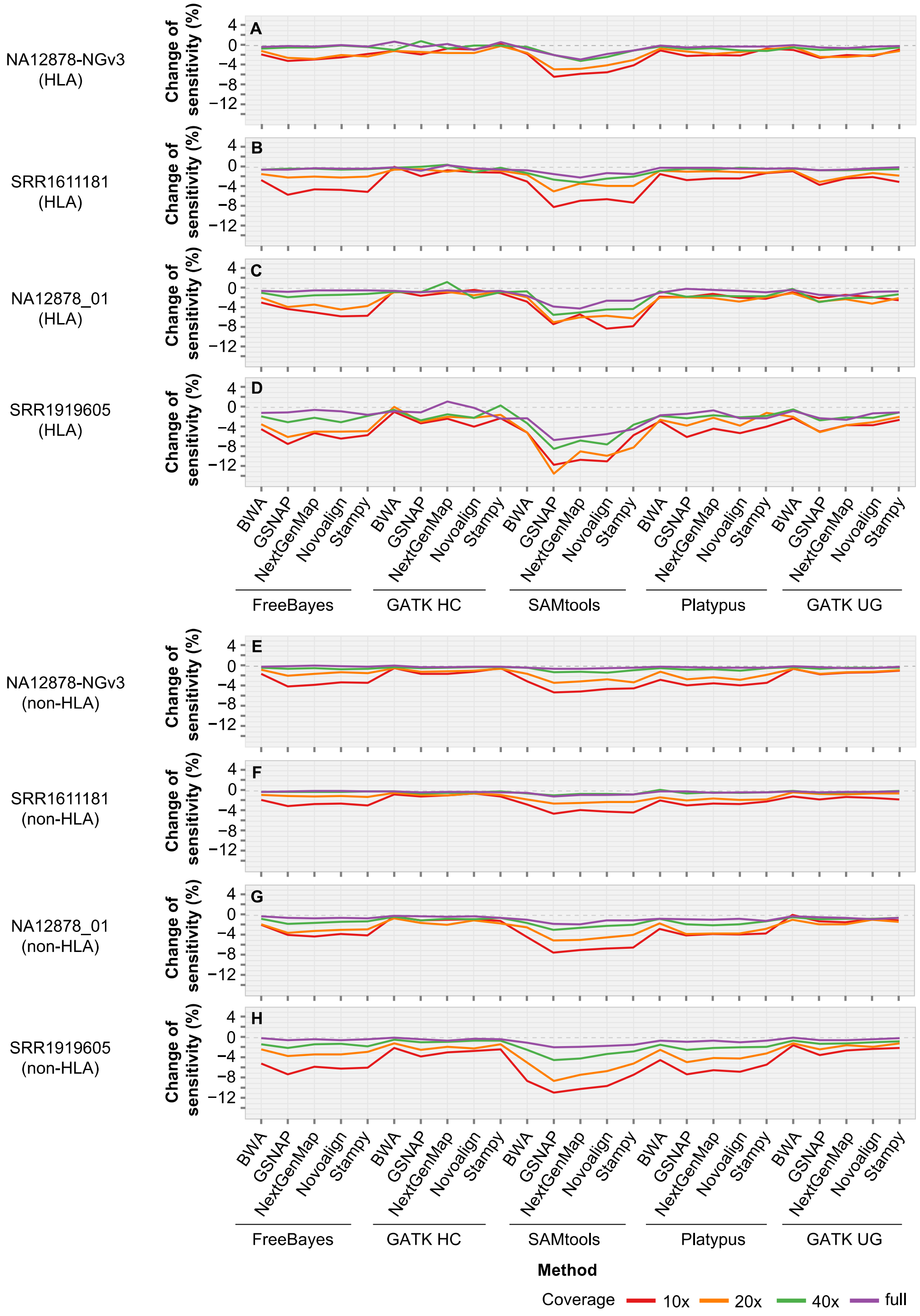


Figure 2.9 BQSR changed variant calling sensitivity in NA12878. NA12878 exome-seq data “FC1_NA12878_04” (~100x coverage, Table 2.1) was down-sampled into 40x, 20x and 10x coverage. Change of sensitivity is calculated as the sensitivity after BQSR, subtracted by that after local realignment.

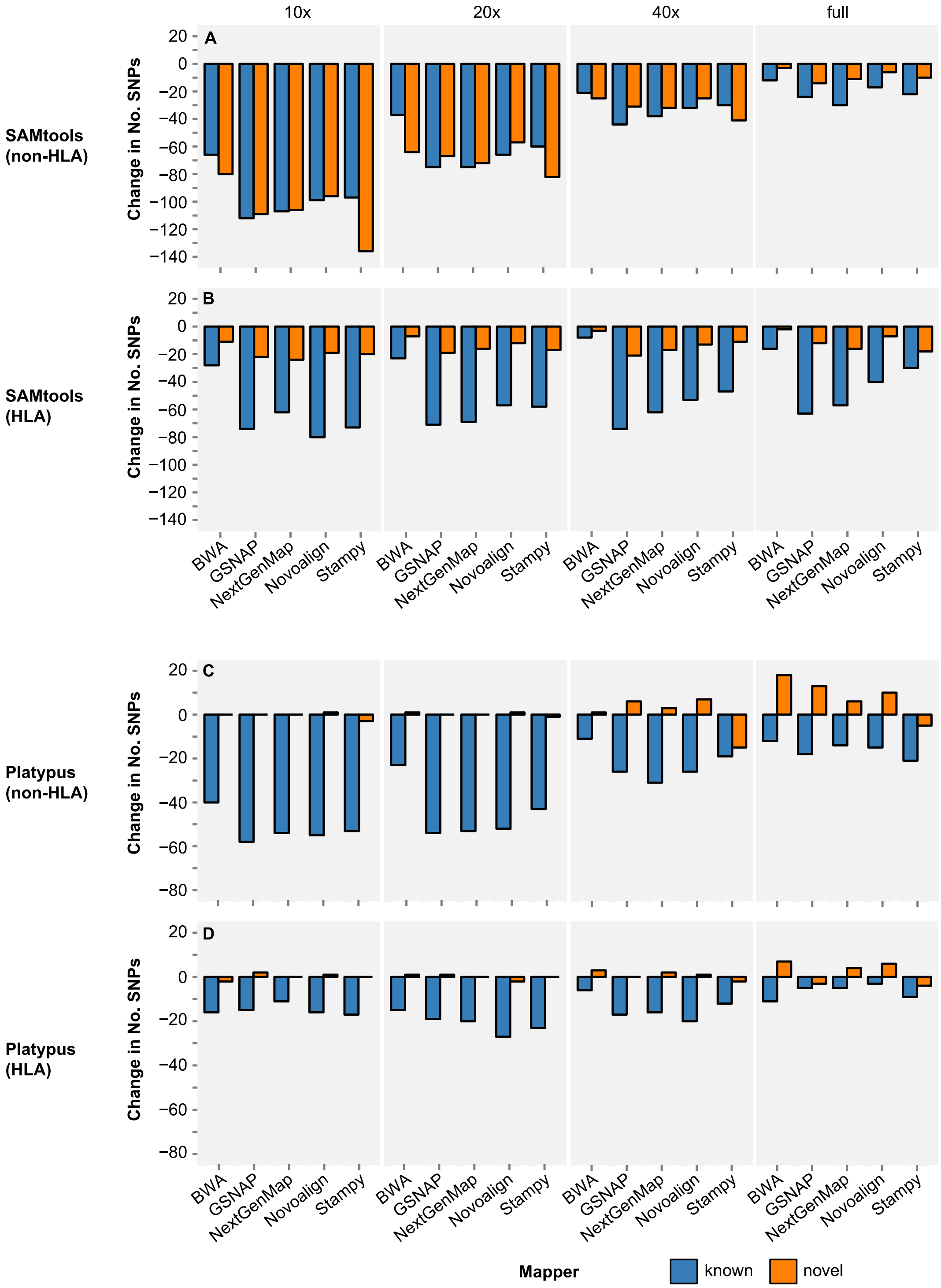


Figure 2.10 Change of known and novel SNPs in NA12878 after BQSR. NA12878 exome-seq data “FC1_NA12878_01” was used. See Figure 2.11 legend for details.

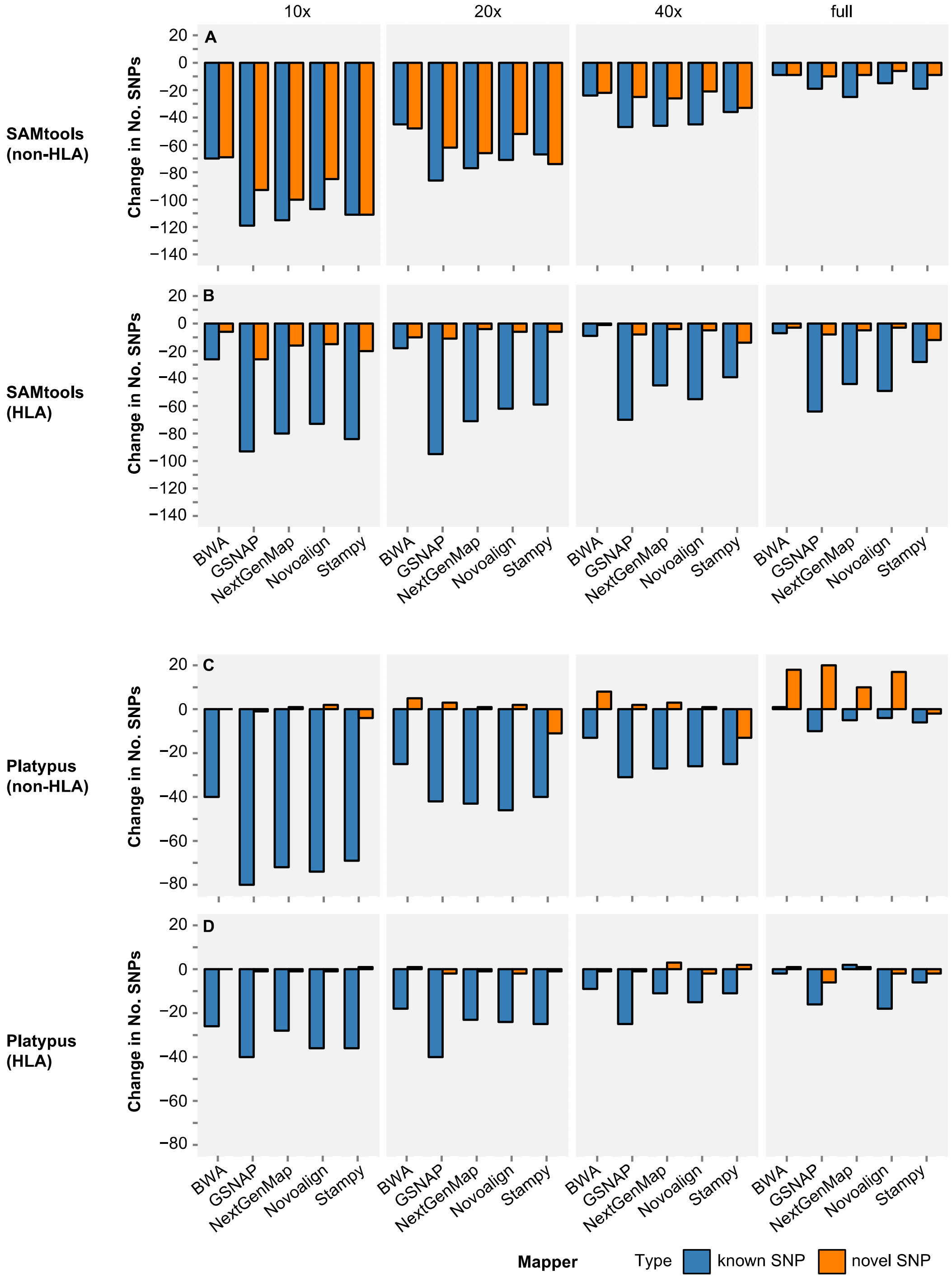


Figure 2.11 Change of known and novel SNP calls by BQSR in NA12878. Exome-seq data “FC1_NA12878_04” was used. *Y*-axis represents the difference in the number of SNPs before and after BQSR. Negative number indicates missed calls after BQSR, compared to local realignment. Known SNPs are those that match dbSNP v138.

For SAMtools (**Figures 2.10A and 2.11A**) and FreeBayes calling (data not shown) in the non-HLA regions, BQSR resulted in a marked loss of both known and novel SNPs at 10x and 20x coverage, but less so at 40x and full (66-100x) coverage. In the HLA region, however, the loss was well more obvious for known SNPs than for novel ones across all the coverage depths (**Figures 2.10B and 2.11B**). Over 90% (median) of the known SNPs missed by BQSR in the HLA and non-HLA regions overlapped the public call set, supporting the reliability of those calls. Thus, in the non-HLA regions, when the coverage is less optimal, BQSR increases the precision rate with sacrifice in sensitivity, indicating a trade-off between these two measurements. However, when the coverage is sufficient, the impact of BQSR becomes much less conspicuous. On the other hand, in the HLA region, BQSR reduces the sensitivity but with little gains in precision rate. On the other hand, for Platypus (**Figures 2.10C, 2.10D, 2.11C and 2.11D**), GATK UnifiedGenotyper and GATK HaplotypeCaller (data not shown), BQSR led to a much more reduction of known SNPs than novel SNPs, in particular at 10x to 20x coverage. For these three callers in SNP detection, the role of BQSR is even adverse.

We also investigated the impact of BQSR on INDEL calling. A total of 500 cases were assessed, which represent combinations among 5 exome data, 4 coverage depths, 5 mappers and 5 callers. INDELs were first grouped into known (in dbSNP v138) and novel, and both were then split into those that overlap the public call sets and others that are method-specific. For all four groups of INDELs from each case, we checked the change in the number of INDELs before and after BQSR (**Figures 2.12A-2.12D**). In INDEL calling by FreeBayes, GATK UnifiedGenotyper, Platypus, and SAMtools, the number of INDELs differed by no more than two in nearly all the cases in the HLA region and in 95-100% of the cases in the non-HLA regions. Slightly more changes were observed for GATK HaplotypeCaller in the non-HLA regions, more notably (up to 10 INDELs) for those that are unique to GATK HaplotypeCaller (missed in the public call set).

Overall, BQSR resulted in no or only small changes in INDEL calls in both the HLA and non-HLA regions.

Taken together, our analysis suggested that the effect of BQSR depends on caller, coverage, and sequence divergence. In both HLA and non-HLA regions, BQSR reduced SNP calling sensitivity for all the callers. However, it improved the precision rate only for FreeBayes and SAMtools in the non-HLA regions when the coverage is insufficient. On the other hand, INDEL calling appears to be less impacted.

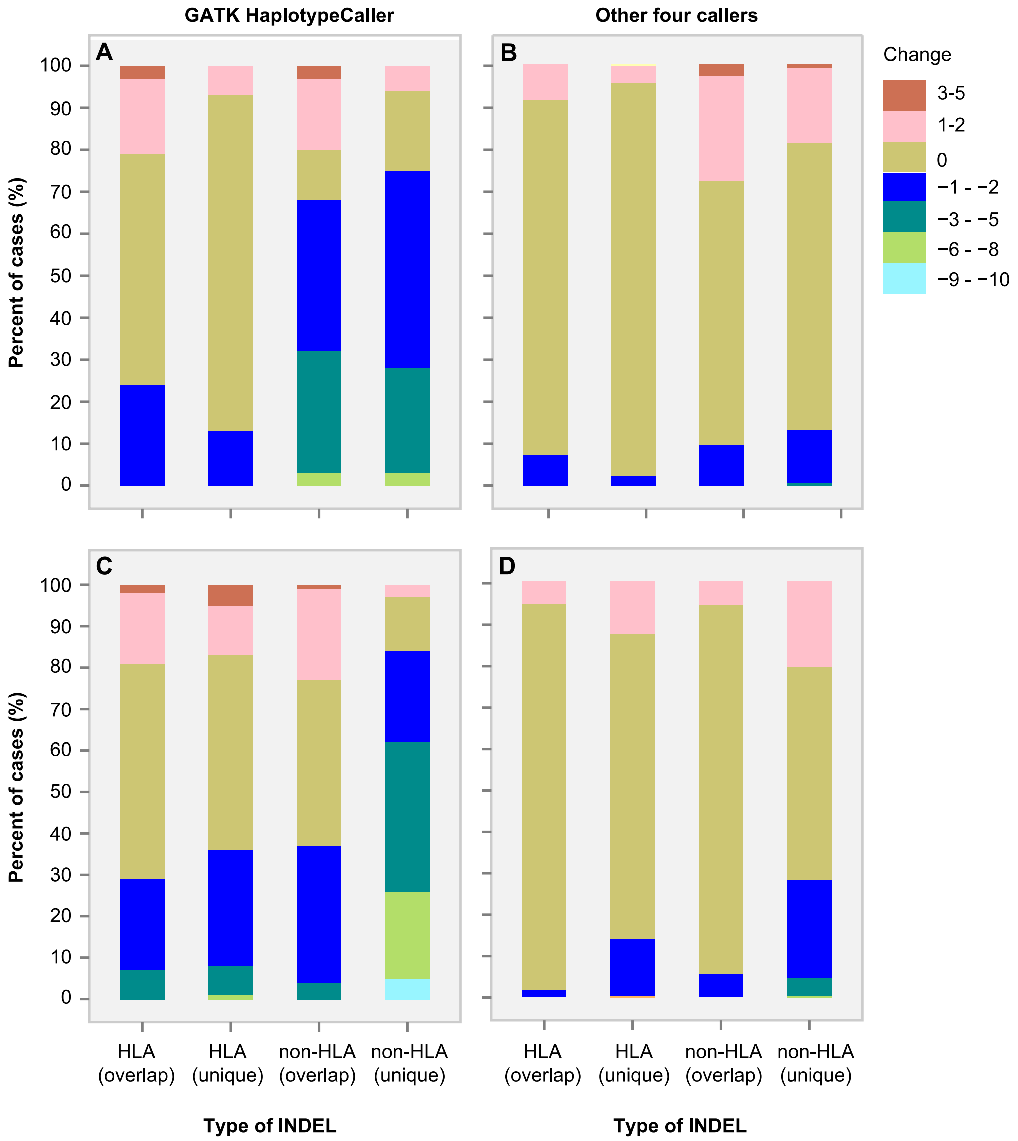


Figure 2.12 Change of known and novel INDELs in NA12878 by BQSR. INDELs were called from five exome-seq data in NA12878 both before and after BQSR. For each data, the full coverage dataset and three subsets of 10x, 20x and 40x coverage were used. INDELs were separated into known (**A** and **B**) and novel (**C** and **D**) by intersecting with dbSNP v138, and both were then compared to the public call set to identify overlapped and method-specific calls. A total of 500 cases, with 100 cases from GATK HaplotypeCaller and 400 cases from the other four callers, were assessed. These 500 cases were from five exome-seq data, four coverage depths, five mappers and five callers. Each case was assessed for change in the number of known and novel INDELs after BQSR, with positive number indicating increase and negative number indicating decrease of INDELs. Y-axis represents the proportion of cases with varying number of change in INDELs.

2.5. Conclusions

Post-alignment processing is frequently applied in current variant discovery pipelines. Using exome data we revealed that local realignment and BQSR did not always enhance variant detection as one would expect, and in fact, at times, it had a negative impact on variant discovery. The decision as to whether to use these post-processing steps are mapper and caller dependent, often varying with coverage depth and level of divergence.

Local realignment and BQSR mainly impacted INDEL and SNP calling, respectively. Local realignment overall increased INDEL calling sensitivity with NextGenMap alignment but had little impact on Novoalign. On the other hand, compared with the haplotype-based callers, the effect was more obvious on SAMtools and GATK UnifiedGenotyper, the two callers that are less effective in INDEL detection. For majority of the methods, BQSR reduced the SNP calling sensitivity, more obvious at lower coverage. In the non-HLA regions, when the coverage is not sufficient, SAMtools and FreeBayes showed decrease in sensitivity but increase in precision rate by BQSR. In other cases, loss of sensitivity was not associated with increase in precision rate, which argues against the application of BQSR in those instances. Our analysis offers a broad view about post-alignment processing in exome-based variant discovery. Thus, consideration should be given to both mapper and caller when deciding whether to apply post-processing to Illumina exome data.

Chapter 3

An analytical workflow for accurate variant discovery in highly divergent regions

3.1. Summary

Current variant discovery methods often start with the mapping of short reads to a reference genome; yet, their performance deteriorates in genomic regions where the reads are highly divergent from the reference sequence. This is particularly problematic for the human leukocyte antigen (HLA) region on chromosome 6p21.3. This region is associated with over 100 diseases, but variant calling is hindered by the extreme divergence across reference haplotypes. We simulated reads from chromosome 6 exonic regions over a wide range of sequence divergence and coverage depths. We systematically assessed combinations between five mappers and five callers for their performance on simulated data and exome-seq data from NA12878, a well-studied individual in which multiple public call sets have been generated. Among those combinations, the number of known SNPs called differed by about 5% in the non-HLA regions of chromosome 6 but over 20% in the HLA region. Notably, GSNAP mapping combined with GATK UnifiedGenotyper calling identified about 20% more known variants than most existing methods without a noticeable loss of precision, with 100% sensitivity in three highly polymorphic HLA genes examined. Much larger differences among these combinations were observed with INDEL calling in both non-HLA and HLA regions. We obtained similar results with our internal exome-seq data from a cohort of chronic lymphocytic leukemia patients. In summary, we have established an analytical workflow enabling variant detection, with high sensitivity and precision, over the full spectrum of divergence seen in the human genome. Comparison to public call sets from NA12878 has highlighted the overall superiority of GATK UnifiedGenotyper, followed by GATK HaplotypeCaller and SAMtools, in SNP calling, and of GATK

HaplotypeCaller and Platypus in INDEL calling, particularly in regions of high-diversity such as the HLA region. GSNAP and Novoalign are the ideal mappers in combination with the above callers. We expect that the proposed workflow should be applicable to variant discovery in other highly divergent regions.

Keywords: Alignment algorithm, Chronic lymphocytic leukemia, Exome sequencing, Human leukocyte antigen, Variant calling

3.2. Introduction

Genetic variations in protein-coding genes underlie the susceptibility to many human diseases [7, 83] and are also associated with the response to drug treatments [2]. Whole exome sequencing (WES) targets >95% of the exons or approximately 1% of the human genome [7, 84]. It has been widely used to identify causal variants [85], uncovering about 85% of the causative mutations identified in Mendelian diseases [7, 73].

Multiple bioinformatic methods have been developed to identify variants from whole genome or exome sequencing data. The most dominant ones are based on the mapping of reads to a reference genome [17, 18], which often follow the framework in the GATK (Genome Analysis Tool Kit) Best Practices [15, 16, 47]. This framework recommends read mapping by Burrows-Wheeler Aligner (BWA), alignment processing and then GATK variant calling. Compared to the hash-based mappers, the Burrows-Wheeler transform (BWT)-based mappers like BWA are faster but tend to be less sensitive [21, 24, 25]. They are developed for mapping reads to less divergent regions [24, 26]. However, divergence level varies markedly across the human genome [86-88], a factor that has a profound impact on the outcome of variant calling. While the bulk of the human genome has only ~0.1% divergence [21, 40], some regions are highly polymorphic [38, 87, 89]. The best example is the human leukocyte antigen (HLA) region on chromosome 6p21.3; this ~4-Mb region shows up to 10% or higher local divergence [39-41]. Most importantly, HLA region is associated with over 100 diseases, predominantly autoimmune diseases [42], and also with drug response [43]. At such high divergence, the BWA mapping rate drops to a few percent [21, 25]. Thus, accurate identification of sequence variation in this region is clinically important but currently hindered by the extreme polymorphism.

A few mappers have been tailored to aligning reads to more divergent regions, such as GSNAP [26], NextGenMap [25], Novoalign (http://www.novocraft.com/) and Stampy [21]. They use 11- to 15-mer hash tables generated from the reference sequence. Among them, Novoalign and Stampy are more accurate than BWA over a wide range of divergence [21, 33]. Stampy performs similarly as NextGenMap at 10% divergence [25] but is superior to Novoalign at higher (10-15%) divergence [21]. GSNAP is capable of mapping reads with multiple mismatches and/or long INDELs [26]. The choice of an appropriate mapper has a major impact on variant calling [15, 90]. However, previous studies often used simulated data that did not consider divergence level; their primary goal was to evaluate overall sensitivity and accuracy of different mappers. Therefore, for these 'variation-tolerant' mappers, it remains less clear which one(s) may strongly enhance variant detection in these divergent regions.

Several popular software packages are available for both single- and multi-sample variant calling, such as SAMtools [80], FreeBayes [34], GATK UnifiedGenotyper and HaplotypeCaller [15, 29]. In addition, packages like GATK HaplotypeCaller, Scalpel [53] and Platypus [4] combine mapping and local assembly, which are particularly attractive for INDEL detection. A few groups have attempted to assess the relative performance of numerous variant-calling algorithms [18, 23, 52]. Two analyses revealed low concordance on WES data, being only 27% [52] or 37% [53] for INDELs among three methods and less than 60% for SNPs among five methods [52]. While in one of the two cases the majority of the SNPs and about half of the INDELs specific to individual methods were confirmed by sequencing PCR amplicons on the Illumina MiSeq platform

[52], in the second case there was over threefold variability in the validation rate (22-77%) for caller-specific INDELs [53]. These two studies highlight the difficulties in obtaining high-quality variant calls from WES [4]. For whole-genome sequencing, variant calling methods generally agree well with one another in about 90% of the genome and show marked disagreement in the other ~10% ‘difficult regions’ of low-complexity and segmental duplications [54]. Obviously, a better understanding of the factors leading to the low concordance among different approaches is critical for further optimizing variant discovery. Furthermore, genome-wide performance of a variant detection method may not truly reflect the local scenario in highly divergent regions.

We seek to develop an analytical workflow for more accurate variant discovery from WES data, especially in highly divergent regions. By simulating reads from chromosome 6 exonic regions, we systematically evaluated five popular callers together with five mappers over a wide range of divergence level and coverage depths. Taking advantage of the existing call sets generated by both whole exome and whole genome sequencing in NA12878, we verified the findings on two exome-seq data in this well-studied CEU sample. Our analysis revealed key factors impacting variant discovery. We identified the best mapper-caller combinations for variant detection in both HLA and non-HLA regions, and further demonstrated their excellence on exome-seq data from a cohort of chronic lymphocytic leukemia (CLL) patients. Our strategies are particularly effective for exome-seq and should be applicable to whole genome sequencing data as well.

3.3. Methods

3.3.1. Simulation of exome-seq reads

The variation level in the human genome is typically ~0.1% [21] but can reach over 10% in some extremely divergent loci like those located in the HLA region [39, 40]. Therefore, in simulation we defined seven divergence levels between

0.05% and 15% and a control with a zero percent divergence (**Figure 3.1**). Here, divergence level represents the ratio of the number of permuted SNPs and INDELs over the total bases from regions included in simulation. To identify regions from chromosome 6 (Chr6) for simulation, we used exons from hg19 refGene annotation (as of 12/10/2013) together with capture regions included in any of the four Agilent SureSelectXT Human All Exon kits: All Exon 50Mbp, All Exon V4, All Exon V4+UTRs and All Exon V5+UTRs (http://www.agilent.com). To ensure full coverage for short exons and exon edges, the refGene exons were each extended by +/-100 bp and the extended coordinates were retrieved from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables). We merged the above five lists to generate 10,768 non-overlapping regions, from which 100-base paired-end reads were simulated to an average coverage of 100x using Dwgsim, a whole genome simulation tool (https://github.com/nh13/DWGSIM/wiki).

Dwgsim was run separately at each of the seven divergence levels (between 0.05% and 15%) and for the control, using the same parameter settings: 90% SNPs, 10% INDELs, inner distance of 200 bp, no random DNA read, and random seed of 123. We did not specify error rate in Dwgsim simulation (options –e and –E). Dwgsim outputs a VCF file, which shows the Chr6 positions, strand information and genotypes of all the permuted (preplaced) SNPs and INDELs, and a FASTQ file with 2.43 million pairs of simulated reads. Dwgsim assigns a constant quality score to all bases in simulated reads. To mimic the non-random distribution of base quality in real sequencing reads, we assigned empirical per-base quality scores to simulated reads by replacing the quality-score lines in the FASTQ files with those randomly taken from our CLL exome-seq data. From each of the eight full (100x) coverage datasets we randomly sampled six subsets with coverage depth of 80x, 60x, 40x, 20x, 10x and 5x, respectively.

3.3.2. Mapping simulated reads to the reference sequence

Five mappers were selected to align simulated reads against the hg19 reference sequence, using the parameter settings in **Table 3.1**. BWA is effective in mapping reads with relatively low divergence from the reference (<2%) [28]. The other four mappers are capable of aligning reads to more divergent regions, including GSNAP [26], NextGenMap [25], Novoalign (http://www.novocraft.com/), and Stampy [21]. The alignment in the sequence alignment map (SAM) format was converted into the binary alignment map (BAM) format using SAMtools [80] and sorted by coordinates using Picard SortSam command (http://picard.sourceforge.net/).

Table 3.1. Five mappers and five variant callers used in the study

Tool	Version	Command line	Reference
BWA	v0.5.9	1) bwa index ref.fa; 2) bwa aln ref.fa end1.fastq > end1.sai; 3) bwa aln ref.fa end2.fastq > end2.sai; 4) bwa sampe -a 1000 -n 1 -f out.sam ref.fa end1.sai end2.sai end1.fastq end2.fastq	[28]
Novoalign	v3.01.01	1) novoindex ref.nix ref.fa; 2) novoalign --hdrhd off -i PE 425,80 -r Random -F STDFQ -v 90 -x 5 -o SAM -d ref.nix -f end1.fastq end2.fastq > out.sam	www.novocraft.com
Stampy	v1.0.21	1) stampy.py --species=human --assembly=hg19_ncbi37 -G ref.fa; 2) stampy.py -g ref -H ref; 3a) stampy.py -g ref -h ref --substitutionrate=0.1 -f sam --inputformat=fastq -o out.sam -M end1.fastq end2.fastq (for simulated data); 3b) stampy.py -g ref -h ref --substitutionrate=0.1 -f sam --inputformat=fastq --solexa --xa-max=3 --xa-max-discordant=10 -o out.sam -M end1.fastq end2.fastq (for real data)	[21]
GSNAP	v2013-10-25	1) gmap_build -d ref.gmapdb -k 13 ref.fa; 2) gsnap -d ref.gmapdb -D ref.gmapdb -k 13 --orientation FR --max-mismatches 0.1 --maxsearch 1 --npaths 1 --ordered --show-refdiff -A sam end1.fastq end2.fastq > out.sam	[26]

NextGenMap	v0.4.9	1) ngm -r ref.fa -o ref.ngm; 2) ngm -r ref.fa -1 end1.fastq -2 end2.fastq -k 15 -I 0 -X 800 -i 0.8 -n 1 -p -o out.sam	[25]
GATK UnifiedGenotyper	v2.7-2	GenomeAnalysisTK.jar -T UnifiedGenotyper -L chr6 -isr intersection -glm BOTH -mbq 17 --dbsnp dbsnp_135.hg19.vcf.gz -stand_call_conf 20 -stand_emit_conf 10	[15, 29]
GATK HaplotypeCaller	v2.7-2	GenomeAnalysisTK.jar -T HaplotypeCaller -L chr6 -isr intersection --genotyping_mode DISCOVERY --dbsnp dbsnp_135.hg19.vcf.gz -stand_call_conf 20 -stand_emit_conf 10	[15, 29]
FreeBayes	v9.9.2-27	freebayes -b in.bam -f ref.fa -t target.bed -C 2 -3 40 -P 0.0001 -m 0 -q 17 -W 1,3 -S 4 -M 3 -B 25 -E 3 -v out.vcf	[34]
SAMtools mpileup	v0.1.19	samtools mpileup -q 0 -Q 17 -B -ugf ref.fa -l target.bed -F 0.002 in.bam | bcftools view -bvcg - > out.bcf; 2) bcftools view out.bcf >out.vcf	[80]
Platypus	v0.5.2	Platypus.py --bamFiles=in.bam --refFile=ref.fa --maxVariants=50 minMapQual=0 --rmsmqThreshold=0 --hapScoreThreshold=0 --minBaseQual=17 --regions=target.bed --output=out.vcf	[4]

ref.fa, hg19 reference sequence in fasta format.
target.bed, a list of non-overlapping Chr6 regions used in simulation or a list of ‘on-target’ regions for real exome-seq data, see Materials and Methods section for details.
in.bam, mapping output BAM file which has gone through coordinate-based sorting, duplicate marking and local realignment.

The performance was measured on the basis of mapping rate and accuracy. The former was defined as the ratio of mapped reads over the total number of simulated reads, and the latter as the ratio of reads mapping back to their original locations over the total. To simplify both calculations, the two reads from each pair were treated as single-end reads without considering the pairing information. Given that four of the mappers (except Stampy) use ‘soft-clipping’, which masks 5’ and 3’ unalignable termini of a read and report the chromosome position at which the actual alignment starts. As some of the simulated variants can be located at 5’ and/or 3’ in reads, soft-clipping will cause discrepancy in coordinates even when a read is indeed mapped back to its original location. Therefore, in estimating mapping accuracy we added the number of soft-clipped bases shown in the CIGAR string in the BAM files back to the reported mapping location. This adjusted mapping location was compared to the original location from where a read was simulated. Mapping accuracy was then estimated as the ratio of reads mapping back to the original start position over the total simulated reads.

3.3.3 Variant calling from simulated data

To reduce erroneous calls, alignments were subjected to duplicate marking and local realignment by following the published procedure [15], but without base quality score recalibration. The base quality score recalibration step is expected to improve variant calling by providing more accurate base quality scores. We skipped this step, since we found that it reduced the variant calling sensitivity in highly divergent regions (see Chapter 2 for details). Duplicates in the coordinates-sorted BAM files were marked by the Picard MarkDuplicates command. As bases spanning INDELs have a high chance of being incorrectly aligned to the reference, we used GATK IndelRealigner command to perform local realignment around 90%, rather than all of the preplaced (permuted) INDELs. This selection is based on the notion that, in variant calling from real exome-seq data, local realignment is performed around known INDELs, which are a subset of all the INDELs in a sample. Processed alignments were then used in variant calling from the regions used in simulation.

Five callers were selected for variant detection, including GATK UnifiedGenotyper and HaplotypeCaller [15, 29], FreeBayes [34], SAMtools mpileup [80] and Platypus [4]. The command and parameter settings for each caller can be found in **Table 3.1**. Parameters are selected to ensure comparability among different callers. Among these callers, Platypus and

FreeBayes report a single event for multiple SNPs within a stretch of <= 5 bp, called multiple-nucleotide polymorphism (MNP) [4]. To increase the comparability among different callers, multiple-nucleotide polymorphisms were decomposed into individual events using GATK walker VariantsToAllelicPrimitives. The performance of individual methods was evaluated on the basis of sensitivity, precision rate, and overall genotype concordance reported by the GATK walker GenotypeConcordance. For INDELs of lengths >= 2 bp, all callers report the most left position based on left-normalization. Considering this, in estimating levels of overlap the original positions of permuted INDELs were extended by +/- 5 bp before intersecting with the positions of called INDELs using BEDTools [91].

Sensitivity is estimated using the formula:
Sensitivity = true positive/(true positive + false negative)
Where true positive refers to the number of true variants identified by a caller, while false negative represents the number of true variants missed by a caller. Here, the true variants are preplaced (simulated) ones reported in the dwgsim output VCF file.

Precision rate is estimated as:
Precision rate = true positive/(true positive + false positive)
Where false positive represents the number of called variants that does not overlap the true (simulated) variants.

3.3.4. Variant calling from NA12878 exome-seq data

To confirm the generality of the findings made from simulated reads, we tested the same mappers and callers on exome-seq data from DNA sample NA12878. NA12878 is the first genome for which the reference genotype calls (‘high confident call set’) were generated by the Genome in a Bottle Consortium from 11 whole genome and three exome sequencing datasets [51]. The process integrates seven mappers and three callers. It is available at ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/GIAB_integration/

NIST_RTG_PlatGen_merged_highconfidence_v0.2.primitives.vcf.gz. Three other call sets were generated by GATK HaplotypeCaller and two de novo assembly-based callers, Cortex and DISCOVAR, from 250-base paired sequencing of a PCR-free genomic library (http://genepattern.broadinstitute.org/ftp/distribution/crd/DiscovarManuscript/vcf/) [54]. The union of these four public call sets was treated as the 'true' variants to assess the performance of individual mappers and callers**.**

Illumina, Inc. generated 12 replicates of 150 base-pair (bp) paired-end exome data in NA12878 (https://basespace.illumina.com/). We downloaded two replicates (FC1_NA12878_01 and FC1_NA12878_04), each with ~100x coverage on average. Considering that the simulated reads have a length of 100 bases, pairs with one or both reads being shorter than 100 bases were filtered out, which excluded 5.1 and 5.6% of the reads. From the remainder the first 100 bases were extracted for mapping (**Table 3.1**). Post-alignment processing was performed as described in the simulated reads, except that the Mills and 1000G gold standard INDELs (Mills_and_1000G_gold_standard.indels.hg19.vcf.gz, included in the GATK resource bundle for hg19) were used in local realignment. Variants were identified from the capture regions of Chr6 using the five callers in single-sample calling mode, quality-filtered (>=Q20) and those matching dbSNP v138 (ftp://ftp.broadinstitute.org/bundle/2.8/hg19/dbsnp_138.hg19.vcf.gz) were classified as known variants. We required exact match in genomic position for an INDEL to be classified as 'known'. Given that SNP clusters are prevalent in the HLA region, we did not filter out SNP clusters in the comparison. We compared the called variants from these two replicates with that of the public call set**.** The sensitivity, precision rate and overall genotype concordance of known variants were estimated, as described in the simulated data.

3.3.5. Variant calling from Chronic Lymphocytic Leukemia (CLL) exome data

Finally, we selected four of the mappers (exclude NextGenMap) and three of the callers (exclude FreeBayes and SAMtools mpileup) and applied them to WES data from 22 CLL patients. Buccal cell DNA was collected with written consent from the patients and approval by the institutional review board at Mayo Clinic. Exome capture was carried out using Agilent library capture kit V2 or V4. DNA was sequenced from both ends to 100 bases on a HiSeq2000 machine through the Mayo Clinic Medical Genome Facility.

Mapping, post-alignment processing, variant calling and the classification of variants into known versus novel followed the procedure used for NA12878 exome-seq data. We calculated the transition-to-transversion (Ti/Tv) ratios for known SNPs with the PLINK software (http://pngu.mgh.harvard.edu/~purcell/plink/). To assess the performance of individual mapper-caller combinations in complex genomic regions, we annotated variants regarding their sequence contexts. Specifically, variants were labeled with 'LCR' if they were located in low complexity regions (http://figshare.com/articles/Low_complexity_regions_in_hs37d5/969685) [18] and with 'SD' if located in regions of segmental duplications (>=95% sequence identity over at least 1 kb, http://humanparalogy.gs.washington.edu/build37/build37.htm). In addition, if three or more SNPs clustered within a window of 20 bp, they were flagged as 'SnpCluster'.

3.4. Results

3.4.1. Mapping of simulated reads

We simulated 100-bp paired reads at eight divergence levels (0-15%) and seven coverage depths (5-100x) from exonic regions of chromosome 6. Five mappers were selected to align the simulated reads to the reference sequence hg19 (**Table 3.1**). The mapping performance was assessed on the 100x coverage datasets in terms of the proportion of accurately mapped reads and that of unmapped reads (**Figures 3.1 and 3.2**).

At divergence levels of 1% or lower, the five mappers showed highly comparable mapping accuracy, except that BWA was about 0.5% lower in accuracy at 1% divergence (**Figure 3.1**). At 5-15% divergence, BWA had over 20% unmapped rate and was excluded from further comparisons (**Figure 3.1**). Of the other four, Stampy and NextGenMap had relative higher mapping accuracy. On the other hand, Novoalign implements a more aggressive soft-clipping than the others, which became most evidence at 15% divergence (**Figure 3.2**).

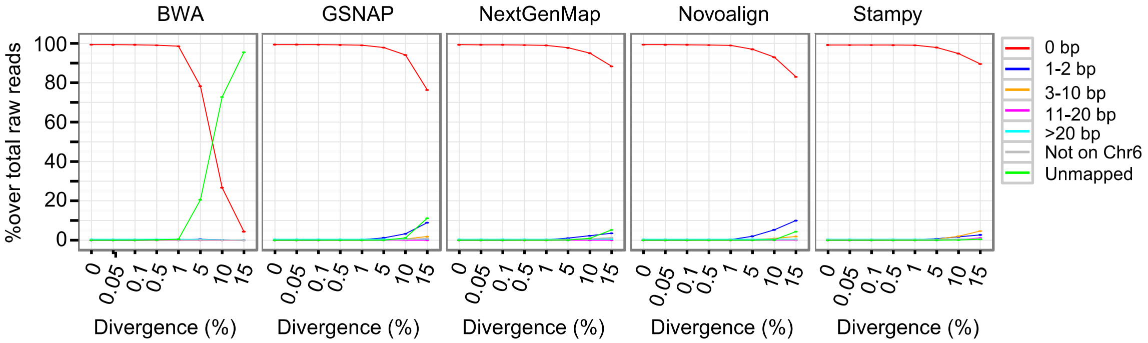


Figure 3.1 Mapping status of simulated reads. Eight datasets were simulated to 100x per-base coverage at seven divergence levels (0.05-15%) or without introducing sequence variation (control). After mapping, the total reads were first broken into those mapping to chromosome 6, those mapping to other chromosomes (‘Not on Chr6’) and unmapped reads (‘Unmapped’). The reads mapping to chromosome 6 were then grouped into five clusters based on the distance (0, 1-2, 3-10, 11-20 and >20 bp) from their original locations to the mapping locations reported by a mapper.

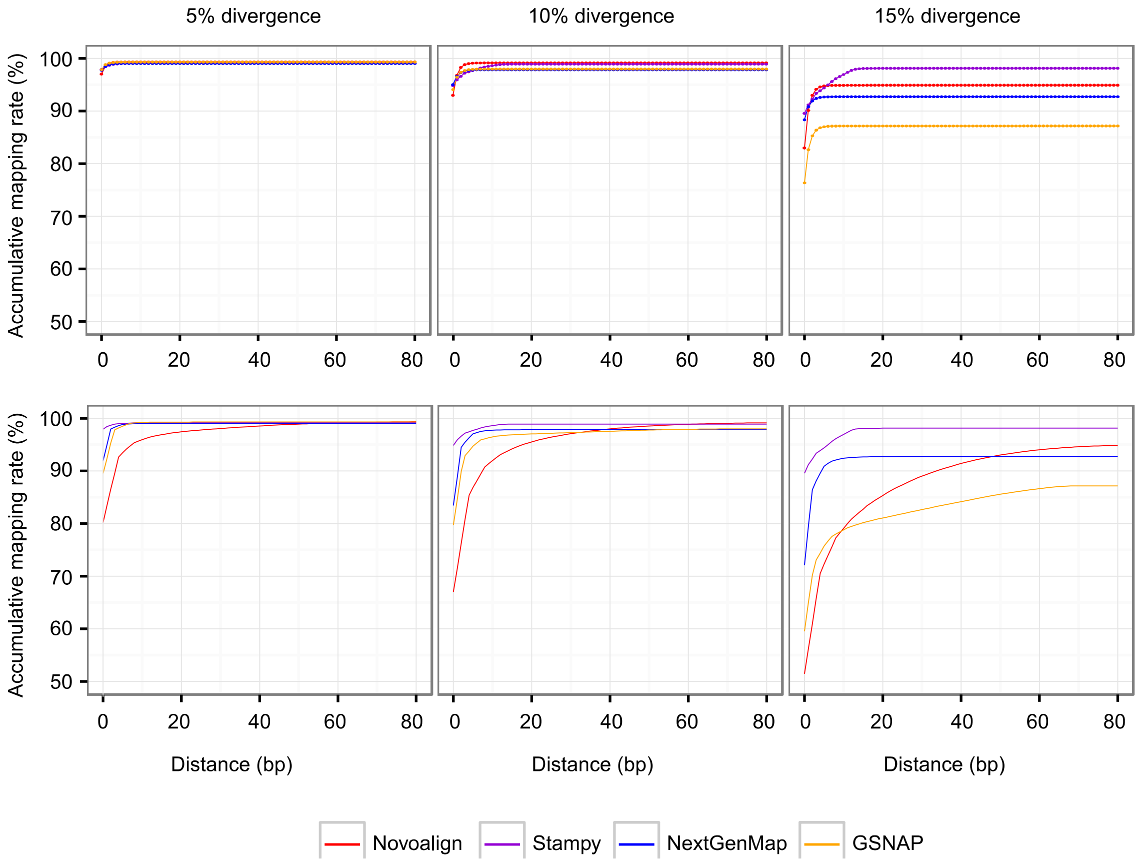


Figure 3.2 Accumulative mapping rate of four aligners. Accumulative mapping rate (*y*-axis) is plotted as a function of the distance (bp, *x*-axis) from the original position to the mapping position on chromosome 6. Upper panels: the number of soft-clipped bases was estimated from the CIGAR string in the BAM file and added back to the reported mapping position to generate the final mapping position; lower panels: the soft-clipped bases were not included. Only the 100x simulation datasets with 5-15% divergence were used. BWA has poor performance at high divergence and is not shown here.

3.4.2. Impact of coverage depth on variant calling

Coverage is another key factor in variant calling. We sought to know how coverage might impact different callers at low versus high divergence. A total of 160 cases per coverage were evaluated, representing combinations among seven divergence levels, four (BWA was not included at 5-15% divergence) or five mappers and five callers. Not surprisingly, in 94% (150/160) of the cases, the full (100x) coverage datasets showed the highest sensitivity in both SNP and INDEL calling.

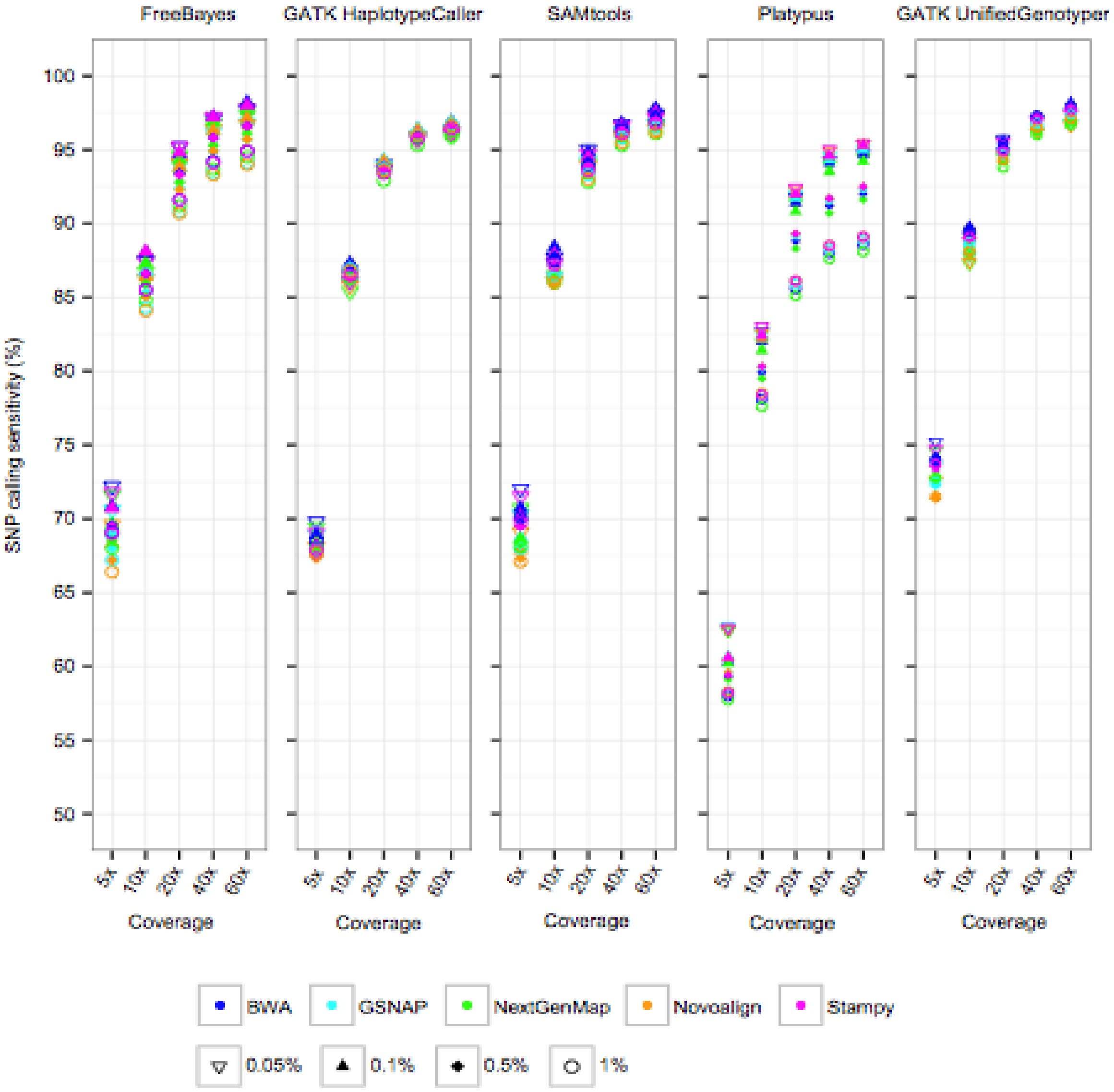


Figure S10

Figure 3.3 SNP calling sensitivity at 0.05-1% divergence. Sensitivity is calculated after local realignment with 90% of the simulated INDELs**.** For each caller at each coverage depth, data from the associated five mappers (by color) and four divergence levels (0.05-1%, by shape) were plotted together. Only coverage depths (*x*-axis) from 5x to 60x are displayed.

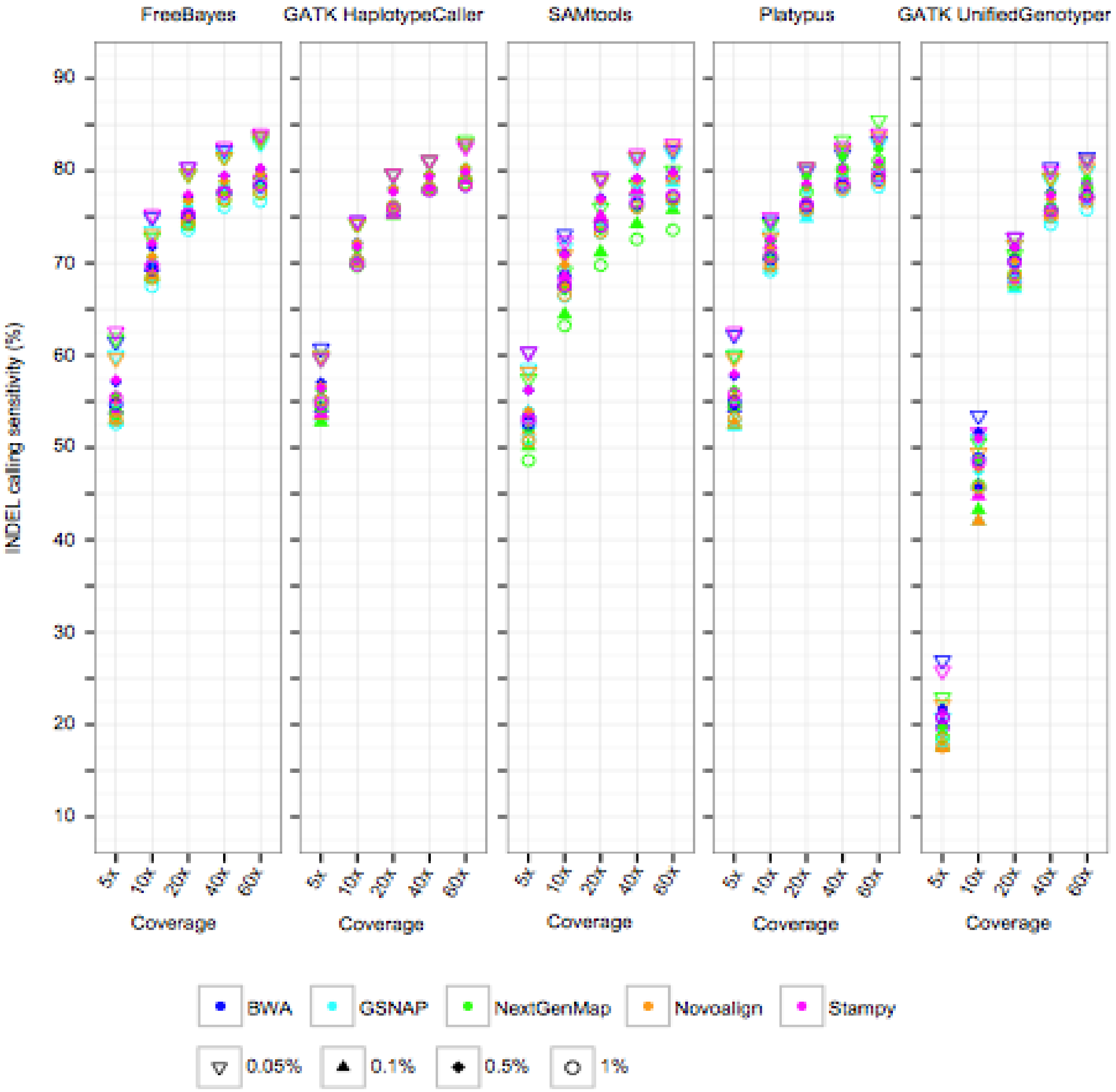


Figure S10

Figure 3.4 INDEL calling sensitivity at 0.05-1% divergence. See Figure 3.3 legend for details.

Profiling sensitivity as a function of coverage should reveal the optimal coverage depth for individual callers; beyond which there would be much less gains in sensitivity. To quantify the effect of coverage, we estimated the change in sensitivity between any two adjacent coverage depths. At <=1% divergence, increasing coverage from 5x to 10x led to the biggest increase in sensitivity in both SNP calling (**Figure 3.3**) and INDEL calling (**Figure 3.4**). In addition, Platypus in SNP calling and GATK UnifiedGenotyper in INDEL calling required higher coverage compared to the others. Overall, at 40x coverage, all callers (except GATK UnifiedGenotyper in INDEL calling) reached 98.4%-99.2% of the SNP calling sensitivity and 97.4-99.1% of the INDEL calling sensitivity at 100x coverage (data not shown). Therefore, 40x coverage is sufficient for both SNP and INDEL calling at low divergence.

At high (>=5%) divergence, the sensitivity of Platypus and FreeBayes was much lower than that of the other three callers in SNP calling (**Figures 3.5A** and **3.5B**). FreeBayes is known to be less sensitive to highly divergent regions [34]. SAMtools had much lower INDEL calling sensitivity compared to the other callers (**Figures 3.5C** and **3.5D**). As previously described with <=1% divergence (**Figures 3.3** and **3.4**), a similar trend of sensitivity versus coverage was observed at >=5% divergence. The sensitivity at 40x reached 97.3-100% of that at 100x in SNP calling and 96.5-98.7% (only 91.1-96.1% for GATK UnifiedGenotyper) in INDEL calling. Therefore, 40x coverage seems sufficient for most of the callers across the full range of divergence.

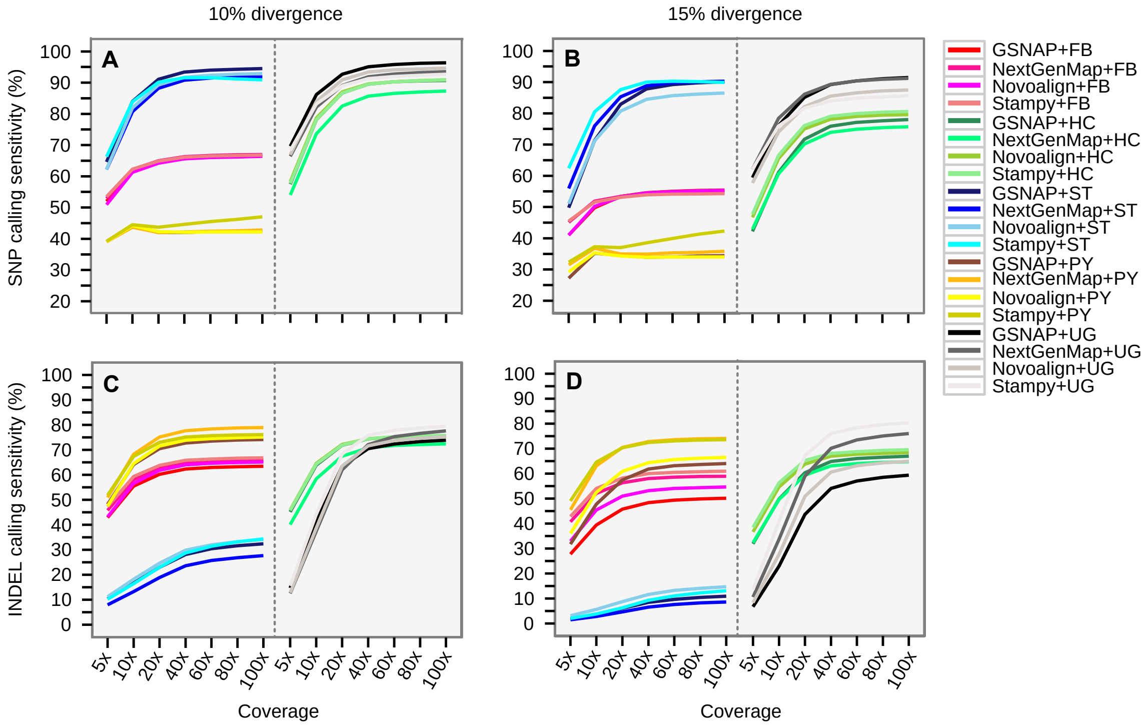


Figure 3.5 SNP and INDEL calling sensitivity at 10% and 15% divergence. *X*-axis indicates coverage depth. *Y*-axis denotes SNP (**A** and **B**) and INDEL calling sensitivity (**C** and **D**). Sensitivity is calculated after local realignment with 90% of the simulated INDELs**.** FB, FreeBayes; HC, GATK HaplotypeCaller; PY, Platypus; ST, SAMtools; UG, GATK UnifiedGenotyper.

3.4.3. Performance of the five variant callers

We tried to identify the callers that perform well at low and high coverage depths, especially at high divergence. We used 10x to represent low and 40x and 100x to represent high coverage depth, respectively. As shown in the SNP and INDEL sensitivity-versus-coverage plots (**Figures 3.3, 3.4** and **3.5A-3.5D**), the change of sensitivity from 40x to 100x was very similar among the different combinations. Therefore, the 60x and 80x datasets were not analyzed here. We also excluded the datasets with 15% divergence since regions with such high divergence are rare in the genome.

We analyzed each caller by considering all the mappers together, separately at low (0.05-1%) and high (5-10%) divergence (**Table 3.2**). In SNP calling, GATK

UnifiedGenotyper had the highest sensitivity over all the divergence levels, followed by SAMtools and GATK HaplotypeCaller (**Table 3.2**). FreeBayes was comparable to the latter two at low but was less sensitive (8-28% lower) at high divergence. Finally, Platypus had the lowest sensitivity in all the datasets. The three highly sensitive callers had similar precision rates (97.8-100%), which were comparable to that of FreeBayes and higher than that of Platypus (**Table 3.3**).

Table 3.2 Percent of SNP and INDEL calling sensitivity in simulated data

Type	Div	Cov			Caller		
	(%)		GATK UG	Platypus	SAMtools	GATK HC	FreeBayes
SNP	<=1	10	87.3-89.8 (A)	77.6-83 (C)	85.9-88.4 (B)	85.4-87.3 (B)	84.1-88.1 (B)
SNP	<=1	>=40	96-98.5 (A)	87.6-96 (D)	95.3-98.2 (B)	95.3-97.4 (C)	93.3-98.6 (C)
SNP	5-10	10	82.2-88 (A)	43.7-59.4 (E)	80.7-85.5 (B)	73.6-84 (C)	61.3-74.3 (D)
SNP	5-10	>=40	91.3-97.4 (A)	42-64.1 (E)	90.8-95.1 (B)	85.7-95.3 (C)	65.6-81.6 (D)
INDEL	<=1	10	42-53.5 (D)	69.1-74.9 (A)	63.2-73.1 (C)	69.7-74.5 (A)	67.5-75.3 (B)
INDEL	<=1	>=40	74.2-82.9 (C)	77.8-85.5 (A)	72.6-82.9 (C)	77.9-83.3 (B)	76-84 (B)
INDEL	5-10	10	37.4-47.4 (D)	64.1-70.9 (A)	13.3-47.7 (E)	58.4-69 (B)	55.5-65.5 (C)
INDEL	5-10	>=40	70.5-79.7 (B)	72.8-81.7 (A)	23.6-64.1 (D)	70.8-78.6 (B)	62.4-73.7 (C)

Individual datasets are binned into four groups based on coverage (10x or >=40x) and divergence (<=1% or 5-10%). The values are the range of sensitivity, calculated per caller from the associated mappers and divergence levels. The five callers within each group are ranked (given in parentheses), with "A" indicating the caller with the highest overall sensitivity. GATK UG, GATK UnifiedGenotyper; GATK HC, GATK HaplotypeCaller; Div, divergence; Cov, coverage.

Table 3.3. SNP and INDEL calling precision rate in simulated data

Type	Div (%)	Cov	Caller				
			GATK UG	Platypus	SAMtools	GATK HC	FreeBayes
SNP	<=1	10	99-100	98.2-100	99.3-100	99.1-100	98.3-99.6
SNP	<=1	>=40	98-100	96.6-100	98.9-100	97.8-100	98.1-99.6
SNP	5-10	10	98.9-99.7	97.5-99.3	98.8-99.8	99.3-99.6	99-99.3
SNP	5-10	>=40	98.5-99.7	92.3-98.9	98.7-99.8	99.2-99.6	98.9-99.3
INDEL	<=1	10	99.3-100	98.5-99.9	99.5-100	99.3-99.9	98-99.8
INDEL	<=1	>=40	99-100	98.3-99.9	99-100	98.7-99.9	98.7-99.9
INDEL	5-10	10	93.5-99.7	89.1-99.1	95.2-99.7	92.6-97.8	94.1-99.1
INDEL	5-10	>=40	87.9-99.5	89.7-99	78.4-99.5	92.5-97.8	93.8-99

The values are the range of the percent of precision rate for each caller, calculated from the associated mappers and divergence levels. GATK UG, GATK UnifiedGenotyper; GATK HC, GATK HaplotypeCaller; Div, divergence.

In INDEL calling, overall Platypus and GATK HaplotypeCaller were more sensitive across all the divergence levels. Compared to the above two, FreeBayes and SAMtools were less sensitive at high divergence (**Figures 3.5C** and **3.5D**); GATK UnifiedGenotyper, which requires a higher coverage (**Figures 3.5C** and **3.5D**), was over 20% less sensitive at 10x coverage (**Table 3.2**). In addition, the five callers had roughly similar precision rates (98-100%) at low divergence, though at 5-10% divergence Platypus and GATK HaplotypeCaller had reduced precision rates in some cases (**Table 3.3**). Below we analyzed the ideal SNP and INDEL callers individually to infer the best mapper(s).

3.4.4. Performance of different mapper-caller combinations

The performance of a caller often varies over alignments generated by different mappers. Thus, we attempted to identify the best mapper(s) for each of the ideal callers identified above, i.e., GATK UnifiedGenotyper, GATK HaplotypeCaller and SAMtools in SNP calling and Platypus and GATK HaplotypeCaller in INDEL calling. At 0.05-1% divergence, BWA works the best in most of the SNP and INDEL calling. The other four mappers also perform well in some cases, particularly at high coverage (**Tables 3.4** and **3.5**). For example, in SNP calling at

>=40x coverage, GSNAP and Novoalign performed similarly as BWA for GATK HaplotypeCaller (95.9%-98.5% sensitivity); Stampy had roughly the same sensitivity as BWA for GATK UnifiedGenotyper and SAMtools (96.1-98.5%), while the other three mappers had slightly reduced sensitivities. In addition, the five mappers had similar precision rates, except that Stampy alignment was about 0.5-2% lower in SNP calling.

Table 3.4. Percent of SNP calling sensitivity for three callers in simulated data

Caller	Div	Cov	Mapper				
	(%)		BWA	GSNAP	NGM	Novoalign	Stampy
SAMtools	<=1	10	87.5-88.4	86.4-87.4	86-87	85.9-86.8	87.2-87.9
SAMtools	<=1	40	96.3-96.8	95.8-96.5	95.3-96.1	95.5-96.1	96.1-96.6
SAMtools	<=1	100	97.7-98.2	97.2-97.9	96.7-97.5	96.9-97.7	97.4-98.1
SAMtools	5-10	10	-	84-85.2	80.7-83.4	81.9-84.3	83.9-85.5
SAMtools	5-10	40	-	93.4-93.9	90.8-92.5	91.7-93.3	91.6-93.5
SAMtools	5-10	100	-	94.5-95.1	91.9-93.8	92.9-94.5	90.9-94.1
SAMtools	<=1	10	A	C	C	C	B
SAMtools	<=1	40, 100	A	B	B	B	A
SAMtools	5-10	10	-	A	C	B	A
SAMtools	5-10	40, 100	-	A	C	B	B
GATK HC	<=1	10	86.5-87.3	86-87	85.4-86.6	85.8-86.8	86.1-86.6
GATK HC	<=1	40	95.9-96.3	96-96.4	95.3-95.8	95.9-96.3	95.7-96
GATK HC	<=1	100	96.8-97.2	96.9-97.4	96.4-96.7	96.9-97.4	96.8-97
GATK HC	5-10	10	-	78.1-84	73.6-81.7	78.7-84	78.2-83.6
GATK HC	5-10	40	-	89.5-94	85.7-92	89.6-94	89.5-93.8
GATK HC	5-10	100	-	90.7-95.3	87.3-93.5	90.9-95.2	90.9-95.1
GATK HC	<=1	10	A	B	B	B	B
GATK HC	<=1	40, 100	A	A	C	A	B
GATK HC	5-10	10	-	A	B	A	A
GATK HC	5-10	40, 100	-	A	B	A	A
GATK UG	<=1	10	89.2-89.8	87.8-88.8	87.3-88.5	87.4-88.3	88.8-89.1
GATK UG	<=1	40	97.1-97.3	96.4-96.9	96-96.3	96.2-96.6	96.9-97.2
GATK UG	<=1	100	98.2-98.5	97.5-98.2	97.3-97.6	97.3-98	98.1-98.4
GATK UG	5-10	10	-	86.2-88	82.2-85.8	84-87	83.2-87.5
GATK UG	5-10	40	-	95.1-96.2	91.9-94.4	93.4-95.6	91.3-95.5
GATK UG	5-10	100	-	96.4-97.4	93.5-95.8	94.7-96.8	92.7-96.8
GATK UG	<=1	10	A	C	C	C	B
GATK UG	<=1	40, 100	A	B	B	B	A
GATK UG	5-10	10	-	A	C	B	B
GATK UG	5-10	40, 100	-	A	C	B	C

Shown are the range and rank (A to C) of the percent of SNP calling sensitivity for each mapper-caller combination. The datasets are clustered into four groups based on coverage and divergence, as in **Table 3.2**. For each caller within each group, the associated mappers are ranked based on the sensitivity of mapper-caller combination, with "A" indicating the mapper with the highest overall sensitivity together with a given caller. BWA mapping results at 5-10% divergence were not shown here. GATK UG, GATK UnifiedGenotyper; GATK HC, GATK HaplotypeCaller; Div, divergence; Cov, coverage.

Table 3.5. INDEL calling sensitivity for two callers in simulated data

Caller	Div (%)	Cov	Mapper				
			BWA	GSNAP	NextGenMap	Novoalign	Stampy
Platypus	<=1	10	70.5-74.5	69.1-73.1	70.7-74.2	69.7-72.7	70.8-74.9
Platypus	<=1	40	78.3-81.8	77.8-82.2	79.5-83.3	78.1-82.2	78.6-82.5
Platypus	<=1	100	79.5-83.3	78.8-83.3	81-85.5	79.2-83.6	79.7-84
Platypus	5-10	10	-	64.1-68.4	68.3-70.9	64.5-68.9	67.4-69.8
Platypus	5-10	40	-	72.8-76.3	77.7-80.1	73.9-77.2	75.2-77.5
Platypus	5-10	100	-	74.2-77.5	79-81.7	75.1-78.3	76.2-78.6
Platypus	<=1	10	B	C	B	C	A
Platypus	<=1	>=40	B	B	A	B	B
Platypus	5-10	10	-	C	A	C	B
Platypus	5-10	>=40	-	D	A	C	B
GATK HC	<=1	10	70.1-74.5	70.2-74.2	69.7-74.2	70.2-74.2	69.8-74.5
GATK HC	<=1	40	77.9-81.1	78-81.1	77.9-81.1	78-81.1	77.9-81.1
GATK HC	<=1	100	78.7-82.9	78.7-82.9	79-83.3	78.7-82.9	78.7-82.9
GATK HC	5-10	10	-	63.9-69	58.4-66	64.6-68.9	64.3-68.6
GATK HC	5-10	40	-	74.4-77.4	70.8-75.8	74.4-77.4	74.5-77.2
GATK HC	5-10	100	-	75.5-78.6	72.4-77.5	75.6-78.5	75.8-78.5
GATK HC	<=1	10	A	A	B	A	A
GATK HC	<=1	>=40	A	A	A	A	A
GATK HC	5-10	10	-	A	B	A	A
GATK HC	5-10	>=40	-	A	B	A	A

See **Table 3.4** for details. GATK HC, GATK HaplotypeCaller; Div, divergence; Cov, coverage.

For SNP calling at 5-10% divergence, overall the three callers performed best with GSNAP, and in some cases with Novoalign and Stampy as well (**Table 3.4**).

In contrast, NextGenMap had the lowest sensitivity (about 1-3% lower) in nearly all the cases. Specifically, GATK HaplotypeCaller had a similar sensitivity on GSNAP, Novoalign and Stampy alignments. For GATK UnifiedGenotyper (the most sensitive caller) and SAMtools, GSNAP is the best mapper. Of the combinations, GSNAP+GATK UnifiedGenotyper is most sensitive, being 0.5-6.3% higher in sensitivity at 5% divergence and 1.7-12.6% higher at 10% divergence. The precision rate varied slightly, with GSNAP being 0.4-0.9% lower than that of Novoalign in GATK UnifiedGenotyper and SAMtools calling.

For INDEL calling at 5-10% divergence, as noticed in SNP calling, GATK HaplotypeCaller achieved similar sensitivities together with three of the mappers, but lower sensitivity with NextGenMap (**Table 3.5**). In addition, GATK HaplotypeCaller had roughly the same precision rate over the four mappers (<0.5% difference); however, its precision rate at 10% divergence was ~5% lower than that at 5% divergence. In contrast, Platypus with NextGenMap had the highest sensitivity at both low and high coverage (**Table 3.5**). For Platypus, GSNAP alignment had the highest precision rate; however, Stampy alignment was over 6% lower at 10% divergence. We ranked the performance of individual methods based on the sensitivity in SNP and INDEL calling (**Tables 3.4** and **3.5**).

Simulated and real exome-seq data are different in several key aspects. Simulated data has a relative uniform distribution of coverage and divergence. However, in the real exome-seq data both features vary widely over different target regions. In addition, the variant type and distribution are much more complicated in real exome data. Considering the limitations in simulation, below we used NA12878 exome-seq data to assess the same mapper-caller combinations.

3.4.5. Evaluating SNP calling on NA12878 exome-seq data

In NA12878, the high-confident call set covers ‘ordinary’ variants which are more readily to be identified, while the ‘difficult’ ones are largely excluded, mostly in regions of low-complexity, segmental duplications and structural variations [51]. On the other hand, the other three call sets identified through 250-bp paired sequencing focus more on difficult variants in regions of low-complexity and segmental duplications [54]. The union of the four call sets, termed public call set, was treated as the ‘true’ variants. Since this public call set is compiled from variants identified through distinct calling algorithms, multiple library preparation protocols and sequencing platforms, systematic bias toward a particular method in the following comparisons will be minimized.

The mapping results suggested that replicates 1 and 2 of NA12878 had approximately 62- and 70-fold coverage of the capture regions, respectively, with 85-88% of the bases having at least 20x coverage. In the assessment, variant discovery methods showing higher levels of overlap with known variants can be reasonably assumed to have a higher sensitivity [15]. To assess the 25 methods, both known (in dbSNP v138) and novel variants were compared to the above public call set. The assessment was done mainly on replicate 1, and that on replicate 2 or both replicates was explicitly pointed out. We aimed to sort out the methods that are effective in both highly divergent and typical genomic regions. Toward this, chromosome 6 was split into two entities: the 4-Mb HLA region (29,500-33,500 kb) with the most extreme divergence and the non-HLA regions which represent typical genomic regions.

In the non-HLA regions, 91.3-96.1% of the SNPs called from Stampy alignments and 96-99.4% from alignments by the other four aligners matched known SNPs. The vast majority of the known SNPs overlapped the public call set (**Figure 3.6A**). Apparently, these methods only varied slightly in the number of known calls, with a maximal difference of 1.8-2.5% among the five mappers with the same caller and of 3-3.4% among the five callers with the same mapper (**Figure 3.6A**). The results generally agree with the assessment made on the simulation data (<=1% divergence and >=40x coverage, **Table 3.2**).

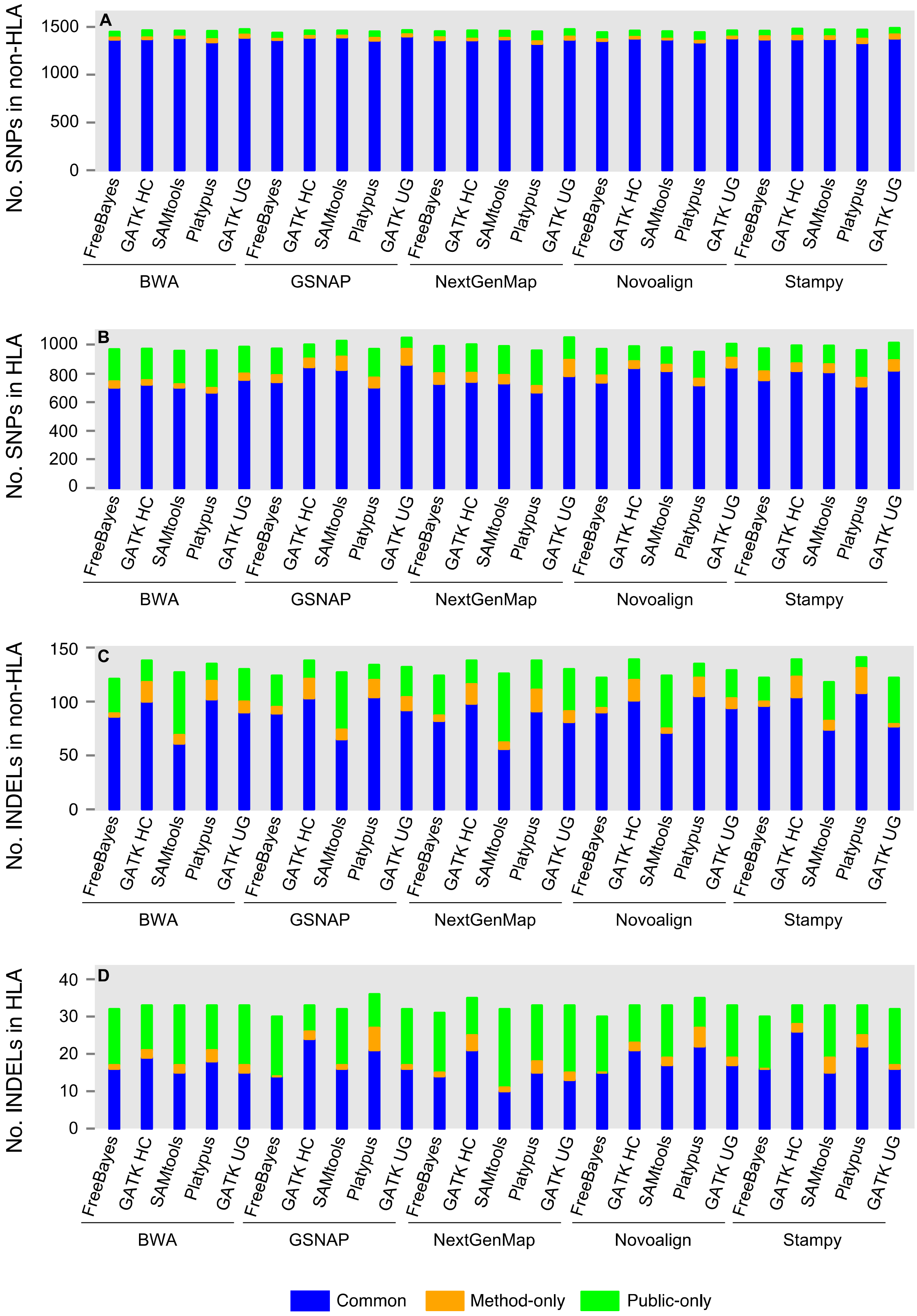

A
No. SNPs in non-HLA
1500
1000
500
0
B
No. SNPs in HLA
1000
800
600
400
200
0
C
No. INDELs in non-HLA
150
100
50
0
D
No. INDELs in HLA
40
30
20
10
0
FreeBayes
GATK HC
SAMtools
Platypus
GATK UG
BWA
GSNAP
NextGenMap
Novoalign
Stampy
Common
Method-only
Public-only

Figure 3.6 Bar plots showing number of known variants in the HLA and non-HLA regions on Chr6. The first 100 bases from the 150-base exome-seq data in NA12878 (FC1_NA12878_01) were mapped by five aligners and variants were identified using five callers (*x*-axis). *Y*-axis shows the number of known SNPs (**A-B**) and INDELs (**C-D**) that match dbSNP v138. For both HLA (**B** and **D**) and non-HLA regions (**A** and **C**), known INDELs and SNPs are broken into: a common portion shared with the 'public call set' ('common', see below), and two portions unique to a given mapper-caller combination ('Method-only') and to the public call set ('Public-only'), respectively. Exome data FC1_NA12878_01 is one of the twelve replicates that are available at https://basespace.illumina.com/analyses/6847907/inputs. The public call set in NA12878 is the union of four call sets below. The high confident call set was generated from 11 whole genome and three exome sequencing datasets and is available at ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/GIAB_integration/NIST_RTG_PlatGen_merged_highconfidence_v0.2.primitives.vcf.gz. The other three lists of variants were called by three packages, Cortex, GATK HaplotypeCaller and DISCOVAR, from 250-bp paired-end reads generated from a PCR-free genome sequencing library. (http://genepattern.broadinstitute.org/ftp/distribution/crd/DiscovarManuscript/vcf/). NA12878 exome-seq data refer to FC1_NA12878_01 (replicate 1) in all figures, unless stated otherwise. HLA region, 29,500,000-33,500,000 bp on Chr6; GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

Not surprisingly, compared to the non-HLA regions, the HLA region showed a 3- to 4-fold enrichment and the five genes (*HLA-A*, *-B*, *-C*, *-DQB1*, and *-DRB1*) showed 36-62 fold enrichment in the known SNPs. Compared to the non-HLA regions, in the HLA region there was much larger between-method variability in SNP calling (**Figures 3.6B, 3.7** and **3.8**). Sixteen methods that had relatively lower SNP calling sensitivity (**Table 3.6**). They used BWA or NextGenMap as the mapper and FreeBayes or Platypus as the caller. Of the other nine methods, GSNAP+GATK UnifiedGenotyper had the highest sensitivity, being about 3% higher than that of the next four combinations: GSNAP+GATK HaplotypeCaller, GSNAP+SAMtools, Novoalign+GATK UnifiedGenotyper and Novoalign+GATK HaplotypeCaller (**Table 3.6**). Of note, these three callers together with GSNAP also had the highest sensitivity in the non-HLA regions (**Table 3.6**). In addition, GSNAP+GATK UnifiedGenotyper covered up to 94.8-99.4% of the known SNPs identified by each of the other methods.

HLA-DRB1 had a SNP density (defined as the number of known SNPs per kb) of 87-88 in the two replicates, which is higher than that of the other four highly divergent HLA genes above. In *HLA-DRB1*, the SNP calling sensitivity was strikingly variable (**Table 3.6**). For example, GSNAP+GATK UnifiedGenotyper and BWA+GATK UnifiedGenotyper together identified 102 known SNPs; nevertheless, 32 were unique to the former but only 2 unique to the latter (**Figure 3.7**). Notably, the performance of the five callers on *HLA-DRB1* is highly comparable with that on the simulated data (5-10% divergence and >=40x coverage, **Table 3.2**). A similar pattern was observed in another highly polymorphic gene, *HLA-DQB1* (**Figure 3.8**). In both genes, exon 2 is excessively divergent, containing >40% of the known SNPs identified in the capture regions (**Figures 3.7** and **3.8**). BWA mapping missed 19 (out of 34) and 12 (out of 41) known SNPs in exon2 from *HLA-DQB1* and *HLA-DRB1*, respectively, arguing for the deployment of methods for more complete variant discovery. BWA-MEM performs local alignment based on seed extension and is more accurate than BWA [92].

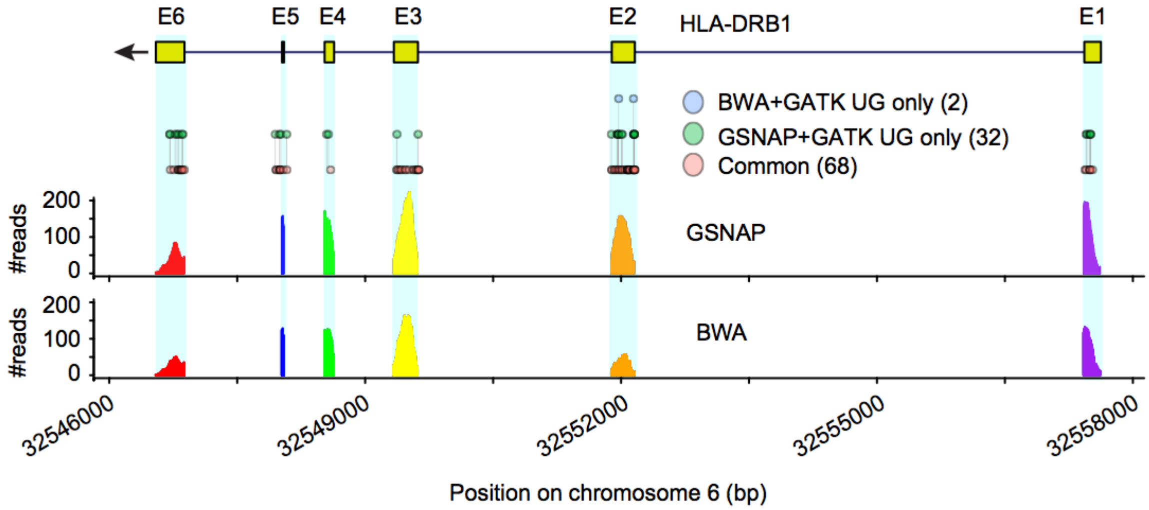


Figure 3.7 Known SNPs in *HLA-DRB1* from NA12878. *HLA-DRB1* structure is shown at the top, with filled boxes representing the six exons (E1 to E6) and arrow indicating transcription direction. The exome-seq reads from NA12878 were mapped by BWA and GSNAP, respectively, and SNPs were called by GATK UnifiedGenotyper (GATK UG). Known SNPs matching dbSNP v138 are showed as 'circles' and clustered into three groups. Number in parentheses indicates the number of known SNPs in each group. The two coverage plots depict BWA and GSNAP mapping coverage at base resolution.

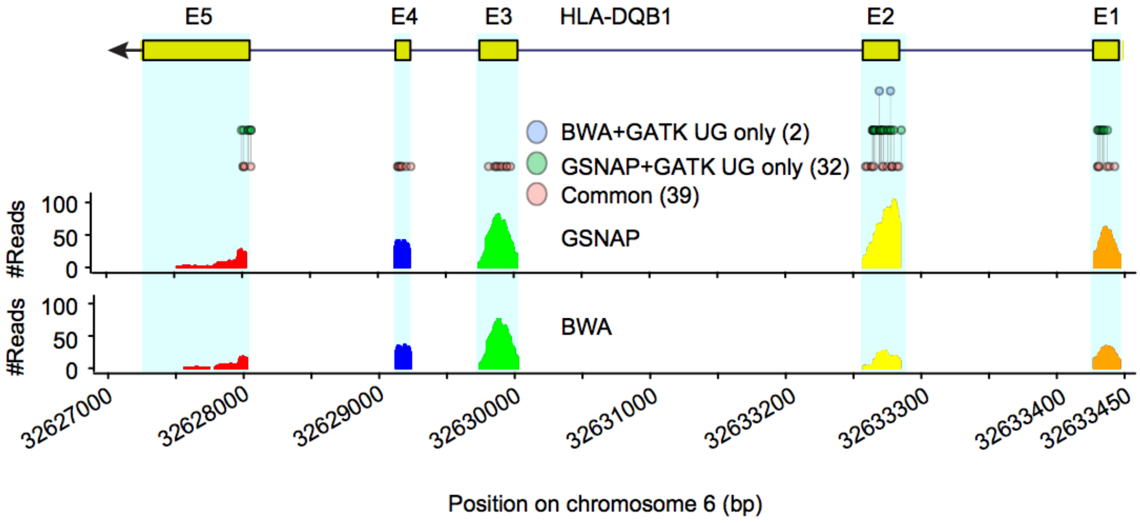


Figure 3.8 A snapshot of known SNPs in *HLA-DQB1* from NA12878. See Figure 3.7 legend for details.

Table 3.6 Percent of SNP and INDEL calling sensitivity in NA12878

Mapper	Caller	SNP			INDEL	
		Non_HLA	HLA	*HLA-DRB1*	Non_HLA	HLA
BWA	FreeBayes	95.8	76	49.3	72.9	50
BWA	GATK HC	94.9	76.9	58	83.3	61.3
BWA	GATK UG	96	80.5	78.3	74.4	46.7
BWA	Platypus	94	71.9	50.7	86.1	60
BWA	SAMtools	95.7	75.3	58.2	51.7	48.4
GSNAP	FreeBayes	95.6	80.3	59.4	75.2	46.7
GSNAP	GATK HC	96.1	90	72.5	85.7	77.4
GSNAP	GATK UG	97	91.7	100	76.3	51.6
GSNAP	Platypus	95.2	77.8	42.6	87.7	70
GSNAP	SAMtools	96.3	88.3	87.9	55.6	51.6
NextGenMap	FreeBayes	95.5	79.2	48.5	68.6	46.7
NextGenMap	GATK HC	94.1	79	62.3	81.5	67.7
NextGenMap	GATK UG	94.5	83.2	76.5	66.9	41.9
NextGenMap	Platypus	92.8	72.8	38.2	76.5	50
NextGenMap	SAMtools	94.8	78.6	60.6	47.1	32.3
Novoalign	FreeBayes	94.9	80.1	59.4	76.1	50
Novoalign	GATK HC	95.4	89.3	75.4	84	67.7
Novoalign	GATK UG	95.6	89.7	85.5	78.2	54.8
Novoalign	Platypus	93.8	79.3	53.6	88.6	73.3
Novoalign	SAMtools	94.7	87.4	83.8	60.2	54.8
Stampy	FreeBayes	96	82.7	65.2	81.4	53.3
Stampy	GATK HC	94.9	87.1	72.5	86.4	83.9
Stampy	GATK UG	95.4	87.4	82.6	63.9	51.6
Stampy	Platypus	93.5	78.8	53.7	91.2	73.3
Stampy	SAMtools	95.1	86.6	79.4	68.5	51.7

HLA region, 29,500,000-33,500,000 bp on Chr6; non-HLA region, other capture regions on Chr6; GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

3.4.6. Method-specific SNP calls in NA12878

We next examined the known SNPs that were identified by GSNAP+GATK UnifiedGenotyper but missed in the public call set (**Table 3.7**), and those that were only present in the public call set (**Table 3.8**). GSNAP+GATK UnifiedGenotyper identified 128 unique SNPs from the two replicates (**Table 3.7**). Ninety-four of them were shared between both replicates, including 29 in *HLA-*

DRB1. By blast search of representative reads spanning the 29 SNPs against the National Center for Biotechnology Information nucleotide collection (nt) database, we found that 16 of them showed 100% identity and another 11 showed 99% identity with *HLA-DRB1* sequences.

Table 3.7. GSNAP+GATK UnifiedGenotyper specific SNP calls

Replicate	Type	Mappers					No. SNP
		GSNAP	BWA	NextGenMap	Novoalign	Stampy	
Rep 1 only	Known	13	3	7	3	4	13 (9,1)
Rep 2 only	Known	21	2	8	5	4	21 (10,0)
Rep 1 & 2	Known	94	32	70	60	51	94 (75,0)
Rep 1 only	Novel	10	1	1	1	3	10 (5,3)
Rep 2 only	Novel	11	5	3	4	4	11 (2,0)
Rep 1 & 2	Novel	18	5	10	8	11	18 (12,1)

Shown is the number of SNPs in the HLA region of NA12878 that were called by GSNAP+GATK UnifiedGenotyper but missed in the public call set. Some of the SNPs were also identified by GATK UnifiedGenotyper together with the other four mappers. GSNAP+GATK UnifiedGenotyper specific SNPs from the two replicates were combined and then split into those shared between the two replicates and others that were unique to either replicate. The last column is the total SNPs, and those in SNP cluster and segmental duplications were shown in parentheses. Known, SNP matching dbSNP v138; novel, SNP not matching dbSNP v138; GATK UG, GATK UnifiedGenotyper; Rep, replicate; HLA, Chr6:29,500,000-33,500,000 bp.

Table 3.8. Public call set specific SNPs

Replicate	Type	Public call set				No. SNP
		Cortex	DISCOVAR	GATK HC	Conf	
Rep 1 only	Known	6	13	17	2	21 (11,0)
Rep 2 only	Known	3	5	11	4	11 (3,2)
Rep 1 & 2	Known	4	18	48	6	57 (33,10)
Rep 1 & 2	Novel	1	10	4	0	13 (2,5)

Shown is the number of SNPs in the HLA region of NA12878 that were missed by GSNAP+GATK UnifiedGenotyper but present in the public call set. The missed SNPs from the two replicates were combined and then split into those that were missed in both replicates and others that were missed in one of the replicates. The last column is the total SNPs, and those in SNP cluster and segmental duplicate were shown in parentheses. Cortex, DISCOVAR and GATK HaplotypeCaller (GATK HC) calls were from 250-bp paired sequencing of a PCR-free genomic library [54]. The high-confident call set ('Conf') was from [51]. Known, SNP matching dbSNP v138; novel, SNP not matching dbSNP v138; Rep, replicate; HLA, Chr6:29,500,000-33,500,000 bp.

The SNPs unique to GSNAP+GATK UnifiedGenotyper were enriched in SNP cluster. About three-fourths (94/128) of the unique calls were in SNP cluster, compared to only 2% of the total SNPs in the non-HLA regions and 38% in the HLA region. Misalignments around INDELs (within 10-bp) can lead to high false positives [15, 90]. However, 109 (85%) of the unique SNPs were at least 50 bp away from known INDELs, thus ruling out misalignments as being a major source of these unique calls. Also, 82 unique calls were identified by at least another two mappers. Finally, we manually checked GSNAP alignments in regions surrounding the 94 unique SNPs called in both replicates. Two of them were low-confident calls in both replicates (two out of four to five supporting reads were soft-clipped) and another two were each supported by only two reads (out of two in total) in replicate 1; all the others were fully supported by the alignments. We thus reason that the vast majority of the known SNPs called by GSNAP+GATK UnifiedGenotyper but missed in the public call set represent true variants.

On the other hand, with GATK UnifiedGenotyper calling, GSNAP missed a total of 89 known SNPs in the public call set (**Table 3.8**), including 66 missed by all the mappers and 13 identified by Stampy alone (the mapper with reduced precision in simulated data). Over half of the missed SNPs were in SNP cluster. Of the five HLA genes, only *HLA-DQB1* showed an obvious miss of SNP calls (12 SNPs, see below for manual inspection). We traced these public call set specific SNPs back to the methods identifying them. While all those SNPs were identified from the 250-bp genomic sequencing data [54], only 12 were in the high-confident call set [51].

We manually checked GSNAP alignments for reads spanning the 57 SNPs that GATK UnifiedGenotyper failed to call in both replicates. Two of them appeared to be false negatives, with one overlapping a 3-bp INDEL in *HLA-DQB1*. Twenty-eight miscalls were due to extremely low coverage (0-2 reads) or insufficient reads (0-2 out of 9-29 mapped reads) supporting the SNPs. Another 11 SNPs were at sites with high coverage (87-226x) in both replicates, but only 2-10% of the reads supported the alternative calls. Of the remaining 16 SNPs (57-2-28-11=16), five had 31-75x and 11 had 142-178x coverage; nevertheless, only between 0 and 4 reads supported the SNPs, raising the possibility that they are platform-specific calls or simply false positives in the public call set. Twelve SNPs were missed in *HLA-DQB1*; 11 of them had two or fewer supporting reads. We argue that a significant proportion of the known SNPs unique to the public call set is likely identifiable only through whole genomic sequencing, longer reads and/or by de novo assembly-based calling methods. Collectively, these results strongly support the findings from the simulated data (5-10% divergence, **Tables 3.2** and **3.4**).

In the HLA region, 3-7.1% of the SNPs were novel ones; this is particularly evident from Stampy alignment (5.4-7.1%), as already seen in the non-HLA regions. GSNAP+GATK UnifiedGenotyper identified 39 novel SNPs that were not in the public call set (**Table 3.7**), but missed 13 in the public call set (**Table 3.8**). Fifteen of the 39 SNPs were identified by at least three mappers, supporting the reliability of these calls. However, none of the 13 SNPs were detected by GATK UnifiedGenotyper with any of the mappers; they were all absent from the high-confident call set [51]. Thus, these public specific novel SNPs likely represent Illumina platform-specific calls.

3.4.7. Evaluating INDEL calling on NA12878 exome-seq data

As expected, it is more difficult to detect INDELs from the HLA regions than from the non-HLA region (**Table 3.6**). In assessing individual methods, we excluded NextGenMap, with which all five callers had the lowest sensitivity in nearly all the cases (**Table 3.6**). GATK HaplotypeCaller and Platypus [4] are haplotype-based callers implementing local de novo assembly, a feature that should contribute to INDEL detection. Indeed, both callers had higher sensitivity in HLA and non-HLA regions (**Figures 3.6C** and **3.6D**; **Tables 3.6 and 3.9**). In the non-HLA regions, FreeBayes and GATK UnifiedGenotyper had similar sensitivities, with SAMtools being least sensitive (**Table 3.6**); in the HLA region, these three callers had comparable sensitivity. These results largely agree with the findings from simulated data summarized in **Table 3.2**.

Table 3.9. Known INDELs in the HLA region of NA12878

Rep	Caller	No. INDELs from four mappers						Public call set		
		BWA	GSNAP	Novoalign	Stampy	Total		250-bp	Conf	Total
1	GATK HC	0	0	0	0	5		5	0	5
1	GATK HC	0	0	0	3	3		2	1	2
1	GATK HC	0	3	0	3	3		3	0	3
1	GATK HC	0	2	2	2	2		2	0	2
1	GATK HC	1	1	1	0	1		0	0	0
1	GATK HC	20	20	20	20	20		19	5	19
1	Sub total	21	26	23	28	34		31	6	31
2	GATK HC	0	0	0	0	4		4	0	4
2	GATK HC	0	0	0	3	3		3	0	3
2	GATK HC	0	2	0	0	2		0	0	0
2	GATK HC	0	2	2	2	2		2	0	2
2	GATK HC	26	26	26	26	26		22	6	22
2	Sub total	26	30	28	31	37		31	6	31
1	Platypus	0	0	0	0	7		7	0	7
1	Platypus	0	0	0	2	2		1	0	1
1	Platypus	0	1	0	0	1		0	0	0
1	Platypus	0	3	3	0	3		0	0	0
1	Platypus	0	4	4	4	4		4	0	4
1	Platypus	1	0	0	0	1		0	0	0
1	Platypus	1	0	1	0	1		1	0	1
1	Platypus	19	19	19	19	19		17	6	17
1	Sub total	21	27	27	25	38		30	6	30
2	Platypus	0	0	0	0	3		3	0	3
2	Platypus	0	0	0	3	3		2	0	2
2	Platypus	0	2	0	0	2		0	0	0
2	Platypus	0	1	1	0	1		0	0	0
2	Platypus	0	4	4	4	4		4	0	4
2	Platypus	3	0	0	0	3		1	0	1
2	Platypus	1	0	1	1	1		1	0	1
2	Platypus	1	1	0	1	1		0	0	0
2	Platypus	21	21	21	21	21		19	6	19
2	Sub total	26	29	27	30	39		30	6	30

Shown is the number of known INDELs in the HLA region that are identified by this study or present in the public call set. Public call set is combined from Cortex, DISCOVAR and GATK HaplotypeCaller calls from 250-bp paired sequencing of a PCR-free genomic library ('250-bp') [54] and a high-confident call set ('Conf') in NA12878 [51]. GATK HC, GATK HaplotypeCaller; HLA, Chr6:29,500,000-33,500,000 bp.

Table 3.10. Novel INDELs in the HLA region of NA12878

Rep	Caller	No. INDELs from four mappers						Public call set		
		BWA	GSNAP	Novoalign	Stampy	Total		250-bp	Conf	Total
1	GATK HC	0	0	0	0	19		18	1	19
1	GATK HC	0	0	0	3	3		2	0	2
1	GATK HC	0	2	0	0	2		0	0	0
1	GATK HC	0	1	0	1	1		0	0	0
1	GATK HC	0	3	3	0	3		3	0	3
1	GATK HC	0	8	8	8	8		8	0	8
1	GATK HC	2	0	0	0	2		1	0	1
1	GATK HC	2	2	2	0	2		1	0	1
1	GATK HC	24	24	24	24	24		19	5	19
1	Sub total	28	40	37	36	64		52	6	53
2	GATK HC	0	0	0	0	13		13	1	13
2	GATK HC	0	0	0	7	7		6	0	6
2	GATK HC	0	0	2	0	2		0	0	0
2	GATK HC	0	1	0	1	1		1	0	1
2	GATK HC	0	2	2	0	2		2	0	2
2	GATK HC	0	14	14	14	14		9	0	9
2	GATK HC	1	0	0	1	1		0	0	0
2	GATK HC	1	0	1	0	1		1	0	1
2	GATK HC	1	1	1	0	1		0	0	0
2	GATK HC	22	22	22	22	22		20	5	20
2	Sub total	25	40	42	45	64		52	6	52
1	Platypus	0	0	0	0	36		36	1	36
1	Platypus	0	0	0	14	14		5	0	5
1	Platypus	0	0	4	0	4		0	0	0
1	Platypus	0	0	1	1	1		0	0	0
1	Platypus	0	2	0	0	2		0	0	0
1	Platypus	0	2	2	2	2		1	0	1
1	Platypus	2	0	0	0	2		1	0	1
1	Platypus	1	0	1	0	1		0	0	0
1	Platypus	1	1	0	1	1		0	0	0
1	Platypus	11	11	11	11	11		10	5	10
1	Sub total	15	16	19	29	74		53	6	53
2	Platypus	0	0	0	0	25		25	1	25
2	Platypus	0	0	0	21	21		12	0	12
2	Platypus	0	0	5	5	5		1	0	1
2	Platypus	0	3	0	0	3		0	0	0
2	Platypus	0	1	0	1	1		1	0	1
2	Platypus	0	1	1	1	1		1	0	1
2	Platypus	2	0	0	0	2		1	0	1
2	Platypus	1	0	0	1	1		1	0	1
2	Platypus	1	0	1	1	1		0	0	0
2	Platypus	4	4	0	0	4		1	0	1
2	Platypus	12	12	12	12	12		10	5	10
2	Sub total	20	21	19	42	76		53	6	53

Shown is the number of novel INDELs (do not match dbSNP v138) in the HLA region. See Supplementary **Table 3.9** for details.

As for the novel INDELs, we focused on those that were identified by four of the mappers (exclude NextGenMap) in combination with Platypus and GATK HaplotypeCaller. Overall, the number of public call set- (26-31) and method-specific INDELs (22-44) was not markedly different in the non-HLA regions. In contrast, in the HLA region, there were two to ten times more novel INDELs unique to the public call set than those unique to the methods (**Table 3.10**). GATK HaplotypeCaller identified much more novel INDELs than Platypus in the HLA region, of which about 80% overlapped the public call set and were supported by at least two mappers. Thus, the majority of them should represent true variants.

3.4.8. Evaluating the robustness of variant discovery methods

We assessed the robustness of variant discovery methods by selecting a different highly divergent region and by choosing a different reference call set. Besides the HLA region, there are many other regions in the human genome that are also highly divergent. For example, the 1000 Genomes Project identified large regions on chromosomes 8 (about 15 Mb) and 16 and some subtelomeric regions on autosomal chromosomes that have increased SNP densities. To test the robustness of our conclusion, we extended our analysis to chromosome 8. We split it into the 15-Mb highly divergent region (chr8:1-15,000,000 bp) and the remaining regions of low sequence divergence. We only showed the results from SNP calling, as very few INDELs were detected in the 15-Mb region. These methods varied little in sensitivity in the low divergence regions but showed up to 17% difference in the 15-Mb region (**Figures 3.9A and 3.9B**), consistent with the findings from Chr6.

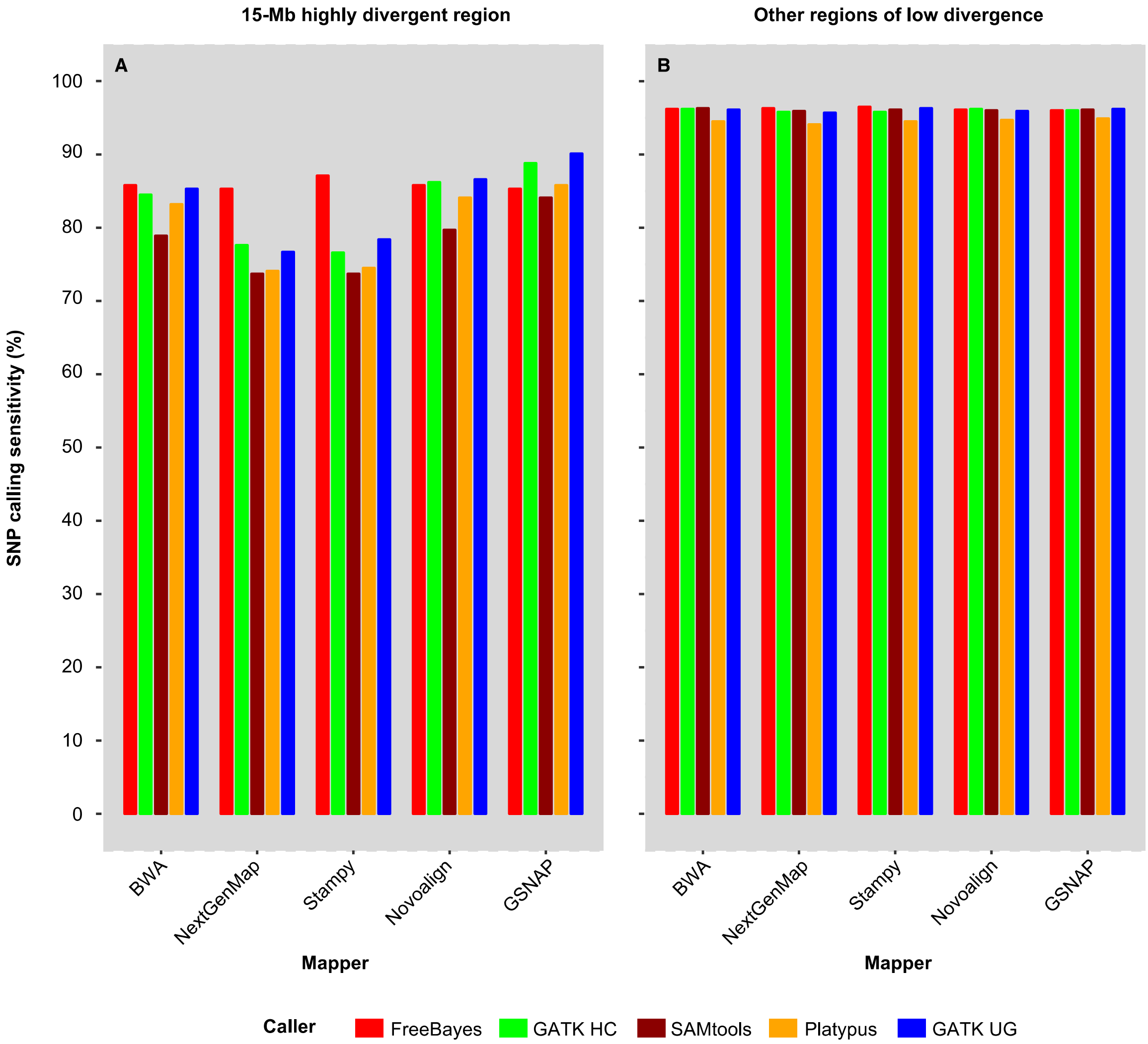


Figure 3.9 Sensitivity of SNP calling from chromosome 8 in NA12878. (**A**) The 15-Mb region of high sequence divergence. (**B**) Other regions of low sequence divergence. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

We compiled a variant reference set for NA12878 from two sources. One is the high-confident call set that was compiled by Zook et al. [51] and commonly used in benchmarking studies. The other was identified by DISCOVER, Cortex and BWA-MEM+GATK HaplotypeCaller from 2x250 bp sequencing of a PCR-free genomic library [54], which had a false positive rate of 1.39-1.64%. The compiled reference set has a good coverage of variants from both “ordinary” and “difficult” genome regions. Also, the variants in this reference set were identified through distinct analytical algorithms, multiple library preparation protocols and sequencing platforms, systematic bias toward a particular method in the assessment should have been minimized. To confirm whether similar findings will be obtained from using a different variant reference set, we chose the NA12878 variant call set from 1000G phase 3 high coverage release of a “Gold standard” trio (2x250 bp, PCR-free genomic library). In SNP calling, the 25 methods had similar sensitivities in the non-HLA regions of Chr6 (**Figure 3.10A**); in the HLA region (**Figure 3.10B**), particularly in the most highly divergent gene *HLA-DRB1* (**Figure 3.10C**), GATK UnifiedGenotyper generally had the highest sensitivity, followed by GATK HaplotypeCaller and SAMtools. In the INDEL calling from both HLA and non-HLA regions (**Figures 3.11A** and **3.11B**), GATK HaplotypeCaller and Platypus were more sensitive. Therefore, using the two reference call sets, we identified the same set of sensitive methods in both SNP and INDEL calling.

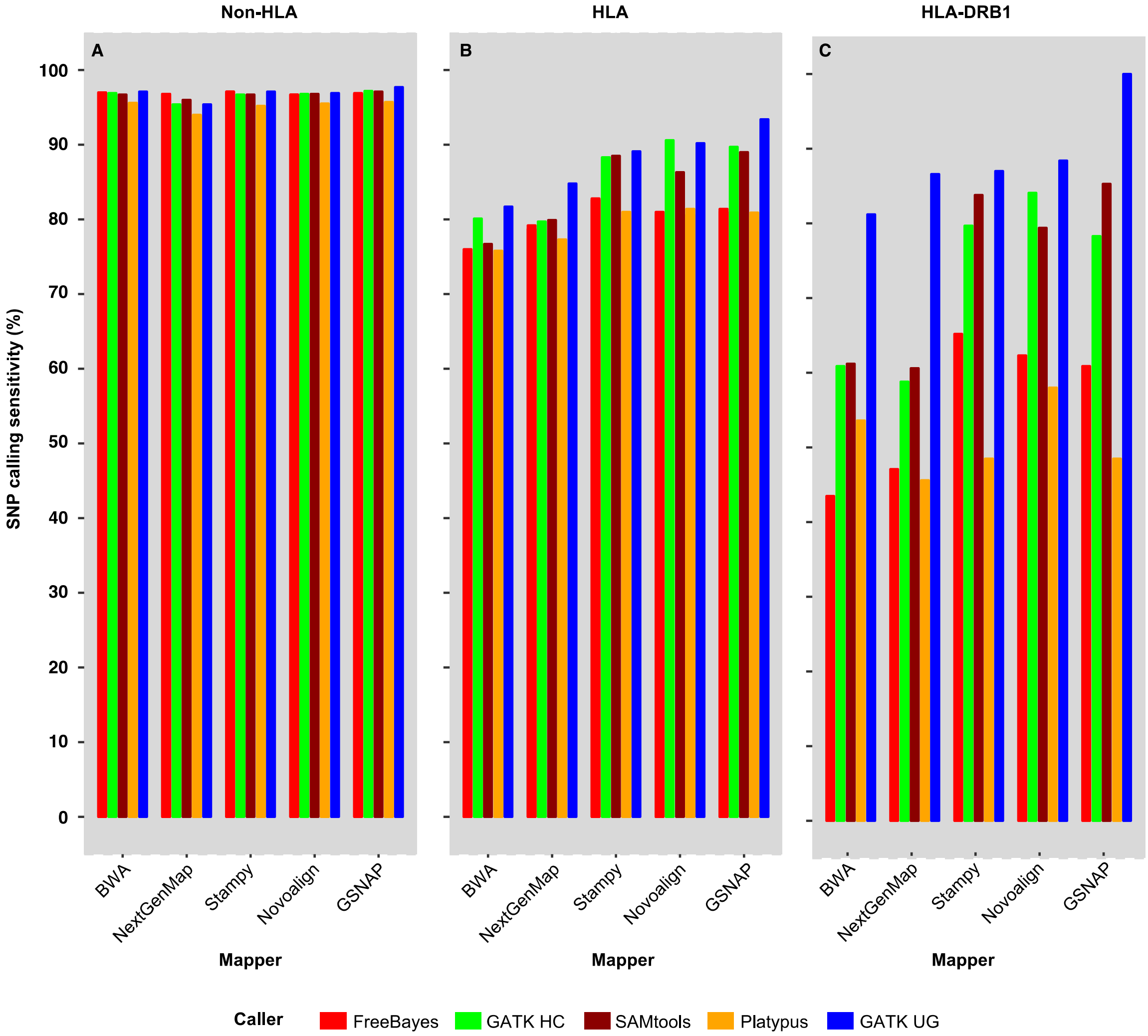


Figure 3.10 Sensitivity of SNP calling from chromosome 6 in NA12878. (**A**) Non-HLA regions. (**B**) HLA region. (**C**) *HLA-DRB1*. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

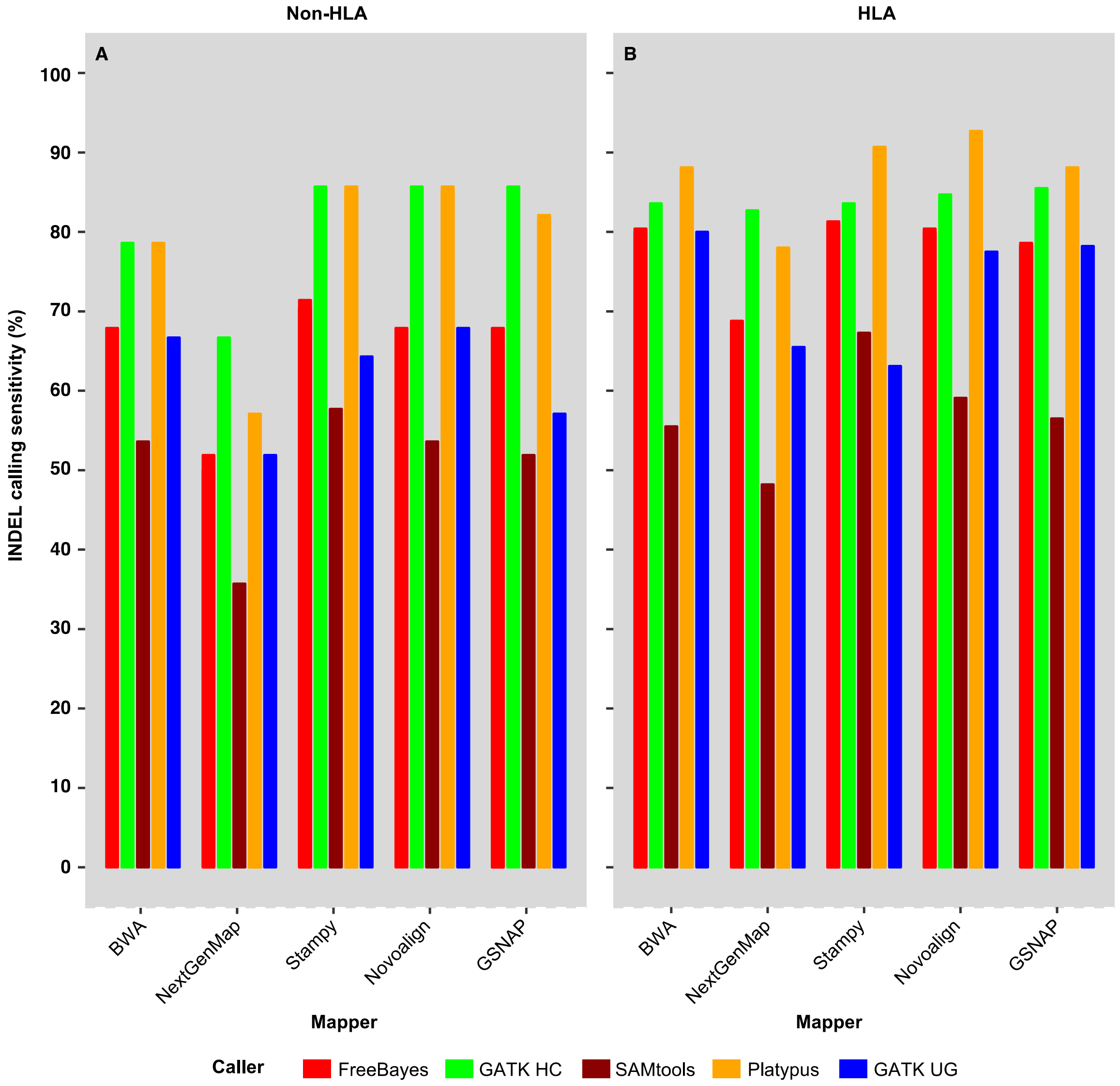


Figure 3.11 Sensitivity of INDEL calling from chromosome 6 in NA12878. (**A**) Non-HLA regions. (**B**) HLA region. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper.

3.4.9. Variant discovery from CLL exome-seq data

Finally, in order to ascertain the broad application of the selected variant calling methods (see below), we generated 100-bp exome-seq data from a cohort of 22 CLL patient samples. Reads mapping (exclude NextGenMap due to its low sensitivity) and variant calling followed the procedure used in NA12878. The average per-base coverage varied between 66.5x and 112.6x, with 82.4-94.9% of the bases having at least 20x coverage. As demonstrated in both simulated (**Table 3.2**) and NA12878 exome-seq data (**Table 3.6**), GATK UnifiedGenotyper and Platypus are effective for SNP and INDEL calling, respectively, while GATK HaplotypeCaller are ideal for both. Focusing on known variants, we ask whether these callers also outperform the others on the CLL exome data, particularly in the HLA region that contains CLL susceptibility loci [77, 93].

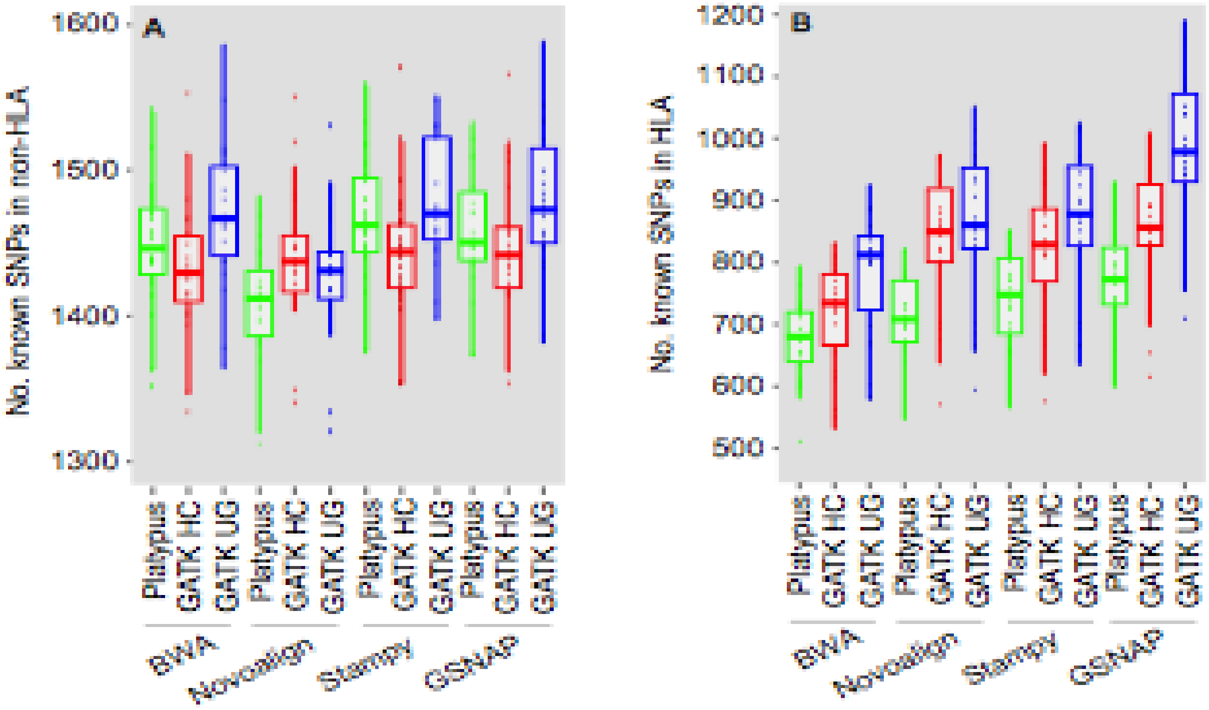


Figure S6

Figure 3.12 Box plots showing number of known SNPs in the HLA and non-HLA regions on Chr6. SNPs were called from 22 CLL samples using five mappers together with three callers. Known SNPs were identified by intersecting with dbSNP v138. (**A**) Known SNPs in the non-HLA regions from Chr6. (**B**) Known SNPs in HLA.

Indeed, in the non-HLA regions, the number of known SNPs per sample differed by 35-123 (2.4-8.8% of the total) among different methods (**Figure 3.12A**; **Table 3.11**). Overall GATK UnifiedGenotyper with Stampy, GSNAP and BWA identified slightly more known SNPs. In contrast, there was much larger difference (146-467 known SNPs or 19.3-70.3% of the total) in the HLA region, with GSNAP+GATK UnifiedGenotyper detecting the most and BWA+Platypus detecting the least or near the least number of known SNPs (**Figure 3.12B**; **Table 3.11**). Of the five highly polymorphic genes, GSNAP+GATK UnifiedGenotyper performed the best in *HLA-A*, *-B*, *-C* and *-DRB1*, and equally well with Stampy+GATK UnifiedGenotyper in *HLA-DQB1* (**Figures 3.13A-3.13E**). The majority of the SNPs within *HLA-DQB1* and *HLA-DRB1* were in SNP cluster (**Table 3.12**). GATK UnifiedGenotyper, which had high a precision rate (**Tables 3.3 and 3.7**), showed the highest sensitivity, followed by GATK HaplotypeCaller and Platypus, consistent with the inference made from both simulated (**Table 3.2**) and NA12878 data (**Table 3.6**).

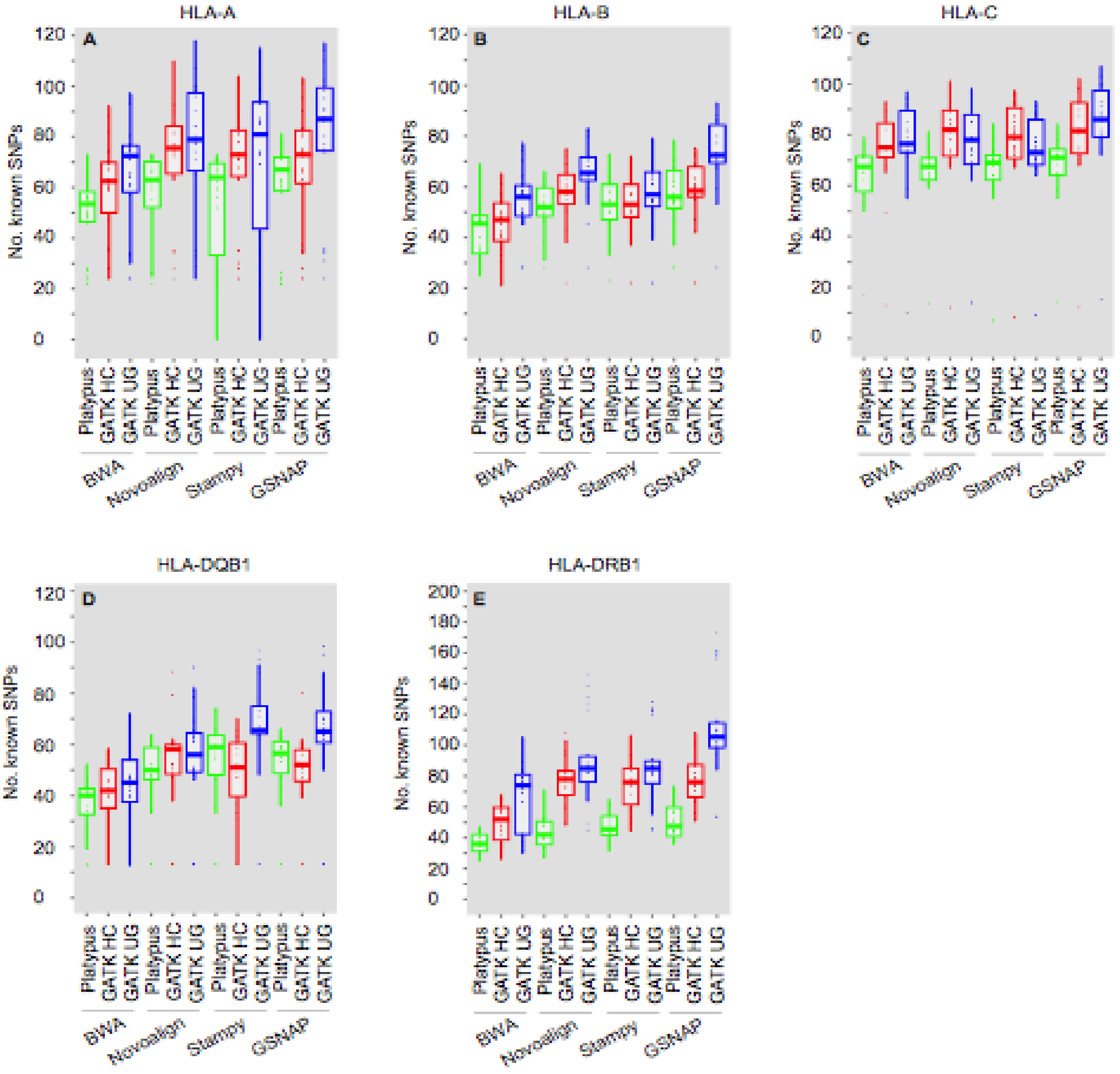


Figure S7

Figure 3.13 Number of known SNPs in five HLA genes. SNPs were called from 22 CLL samples using five mappers together with three callers. Known SNPs were identified by intersecting with dbSNP v138. (**A-E**) Known SNPs identified in *HLA-A* (**A**), *HLA-B* (**B**), *HLA-C* (**C**), *HLA-DQB1* (**D**) and *HLA-DRB1* (**E**).

Table 3.11. Number of SNPs called in 22 CLL samples

Mapper	Caller	Target			
		HLA (known)	HLA (novel)	Non-HLA (known)	Non-HLA (novel)
BWA	GATK HC	734 (532-831)	33 (17-48)	1430 (1334-1553)	31 (3-67)
GSNAP	GATK HC	856 (615-1009)	64 (21-99)	1442 (1354-1565)	46 (22-91)
Novoalign	GATK HC	850 (570-974)	42 (19-82)	1438 (1340-1550)	34 (6-80)
Stampy	GATK HC	828 (575-992)	55 (22-102)	1444 (1353-1569)	74 (43-108)
BWA	GATK UG	812 (579-925)	32 (20-59)	1467 (1364-1586)	36 (6-157)
GSNAP	GATK UG	978 (708-1189)	78 (30-107)	1473 (1381-1588)	46 (24-153)
Novoalign	GATK UG	860 (595-1050)	32 (19-42)	1431 (1321-1529)	11 (5-43)
Stampy	GATK UG	878 (635-1024)	62 (25-102)	1471 (1397-1605)	92 (45-239)
BWA	Platypus	680 (510-793)	32 (23-76)	1446 (1351-1540)	86 (23-195)
GSNAP	Platypus	772 (600-930)	54 (33-86)	1450 (1372-1531)	76 (34-211)
Novoalign	Platypus	708 (547-821)	31 (22-49)	1412 (1312-1482)	26 (9-69)
Stampy	Platypus	748 (566-852)	50 (30-93)	1462 (1376-1558)	156 (65-351)

Shown are the median and range (in parentheses) of called known and novel SNPs. Known, SNP matching dbSNP v138; novel, SNP not matching dbSNP v138. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper. HLA, Chr6:29,500,000-33,500,000 bp; Non-HLA, other capture regions from Chr6.

Table 3.12. Percent of known SNPs in SNP clusters in the 22 CLL samples

Mapper	Caller	Target			
		HLA	Non-HLA	*HLA-DQB1*	*HLA-DRB1*
BWA	GATK HC	28 (17-33)	1 (1-2)	48 (23-62)	64 (44-79)
GSNAP	GATK HC	36 (26-48)	1 (1-2)	57 (23-72)	76 (63-90)
Novoalign	GATK HC	35 (25-45)	1 (1-2)	57 (23-77)	76 (66-90)
Stampy	GATK HC	34 (24-44)	1 (1-2)	57 (23-68)	76 (55-89)
BWA	GATK UG	31 (20-41)	1 (1-2)	48 (0-69)	65 (26-84)
GSNAP	GATK UG	42 (33-55)	2 (1-3)	65 (23-84)	85 (77-98)
Novoalign	GATK UG	35 (23-49)	1 (1-2)	58 (23-74)	80 (67-97)
Stampy	GATK UG	35 (24-48)	1 (1-2)	65 (23-80)	76 (60-95)
BWA	Platypus	22 (16-29)	1 (1-2)	46 (0-56)	37 (0-60)
GSNAP	Platypus	30 (19-39)	2 (1-3)	62 (23-69)	50 (19-73)
Novoalign	Platypus	24 (18-32)	1 (0-1)	57 (23-67)	43 (0-76)
Stampy	Platypus	27 (19-36)	1 (1-2)	62 (23-69)	47 (10-73)

Shown are the median and range (in parentheses) of the percent of known and novel SNPs in SNP cluster. A SNP cluster is defined as the presence of three or more SNPs within a stretch of <=20 bp. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper. HLA, Chr6:29,500,000-33,500,000 bp; Non-HLA, other capture regions from Chr6.

Transition/transversion (Ti/Tv) ratio is sometimes used to assess SNP calling specificity, which, ideally, should approach the expected 3.0-3.3 for known SNPs in exome-seq data [15]. Since there were not sufficient novel SNPs in both HLA and non-HLA regions (**Table 3.11**), we only checked the Ti/Tv ratios for known SNPs, which were 1.6-1.92 (median) in the HLA and 2.38-2.62 (median) in the non-HLA regions (**Table 3.13**). Lower-than-expected Ti/Tv ratios (2.6-2.9) were reported for known variants in other exome data [51]. In addition, the Ti/Tv ratio, which is affected by the target regions [94], may be lower for difficult variants [51]. Thus, the apparently lower Ti/Tv ratio in the HLA region is more likely related to the high divergence and the difference in sequence composition rather than an indication of reduced precision rate.

For INDEL calling, Platypus was more sensitive than GATK HaplotypeCaller and GATK UnifiedGenotyper in the non-HLA regions (**Table 3.14**). Four of the mappers made nearly no difference, but stampy identified about 10% less INDELs in GATK UnifiedGenotyper calling. Conversely in the HLA region, with BWA excluded, GATK HaplotypeCaller identified more known INDELs than both Platypus and GATK UnifiedGenotyper (**Table 3.14**). In fact, GSNAP+GATK HaplotypeCaller identified the most known INDELs in the HLA region from 19 of the 22 samples, followed by Novoalign+GATK HaplotypeCaller and Stampy+GATK HaplotypeCaller that were overall highly comparable to each other. Also, GATK HaplotypeCaller was more sensitive to novel INDELs (**Table 3.14**), as previously revealed in NA12878 (**Table 3.10**).

Table 3.13. Transition/transversion ratio for known SNPs in 22 CLL samples

Mapper	Caller	Target	
		HLA	Non-HLA
BWA	GATK HC	1.88 (1.73-2.08)	2.58 (2.44-2.79)
GSNAP	GATK HC	1.73 (1.57-1.92)	2.54 (2.42-2.73)
Novoalign	GATK HC	1.72 (1.57-2.01)	2.53 (2.42-2.72)
Stampy	GATK HC	1.72 (1.56-1.89)	2.54 (2.4-2.71)
BWA	GATK UG	1.71 (1.6-2.04)	2.43 (2.23-2.62)
GSNAP	GATK UG	1.6 (1.47-1.9)	2.43 (2.28-2.63)
Novoalign	GATK UG	1.71 (1.58-2.03)	2.62 (2.48-2.81)
Stampy	GATK UG	1.7 (1.54-1.9)	2.38 (2.2-2.57)
BWA	Platypus	1.92 (1.57-2.18)	2.41 (2.29-2.58)
GSNAP	Platypus	1.85 (1.49-2.02)	2.41 (2.28-2.55)
Novoalign	Platypus	1.88 (1.58-2.05)	2.58 (2.46-2.76)
Stampy	Platypus	1.78 (1.52-2)	2.38 (2.23-2.5)

SNPs are split into known (matching dbSNP v138) and novel. Shown are the median and range of Ti/Tv ratios for known SNPs in 22 CLL samples. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper. HLA, Chr6:29,500,000-33,500,000 bp; Non-HLA, other capture regions from Chr6.

Table 3.14. Number of INDELs called in 22 CLL samples

Mapper	Caller	Target			
		HLA (known)	HLA (novel)	Non-HLA (known)	Non-HLA (novel)
BWA	GATK HC	28 (19-35)	26 (16-38)	98 (81-114)	26 (19-35)
GSNAP	GATK HC	36 (28-49)	42 (26-58)	99 (82-110)	27 (18-38)
Novoalign	GATK HC	36 (25-45)	37 (19-52)	98 (83-112)	27 (21-36)
Stampy	GATK HC	34 (25-47)	36 (23-53)	98 (80-110)	24 (18-34)
BWA	GATK UG	20 (12-27)	10 (5-14)	98 (83-110)	10 (3-18)
GSNAP	GATK UG	24 (16-33)	11 (5-14)	98 (84-113)	8 (2-13)
Novoalign	GATK UG	24 (18-30)	13 (7-19)	98 (85-111)	10 (5-16)
Stampy	GATK UG	19 (10-28)	32 (14-43)	86 (73-104)	21 (14-33)
BWA	Platypus	26 (21-37)	14 (10-19)	110 (95-129)	15 (7-35)
GSNAP	Platypus	28 (22-37)	16 (10-21)	111 (97-126)	11 (7-32)
Novoalign	Platypus	30 (21-38)	22 (11-31)	110 (97-126)	9 (4-14)
Stampy	Platypus	27 (18-36)	40 (25-53)	111 (98-128)	22 (11-236)

Shown are the median and range (in parentheses) of INDELs called from 22 CLL samples. Known, INDEL matching dbSNP v138; novel, INDEL not matching dbSNP v138. GATK HC, GATK HaplotypeCaller; GATK UG, GATK UnifiedGenotyper. HLA, Chr6:29,500,000-33,500,000 bp; Non-HLA, other capture regions from Chr6.

In summary, for SNP calling, GATK UnifiedGenotyper is more powerful than GATK HaplotypeCaller in the HLA region, especially with GSNAP (**Figure 3.6B**; **Tables 3.6 and 3.11**); in the non-HLA regions, these two callers are roughly comparable and both are better than Platypus (**Figure 3.6A**; **Tables 3.6 and 3.11**). For INDEL calling, GATK HaplotypeCaller and Platypus are the two best callers (**Figures 3.6C** and **3.6D**; **Tables 3.6, 3.9** and **3.14**). In the non-HLA region, Platypus is superior to GATK HaplotypeCaller; while in the HLA region, GATK HaplotypeCaller is often better than Platypus (**Table 3.14**). GSNAP and Novoalign are ideal mappers for both SNP and INDEL calling. One limitation with GSNAP (also Novoalign) is that it is over four times slower than BWA (data not shown). A possible solution would be to first map reads using BWA, extract pairs with unmapped read(s), and then re-map those using GSNAP. We tested its feasibility on the NA12878 data. Remarkably, in GATK UnifiedGenotyper and HaplotypeCaller calling, this two-step mapping approach recovered >99.5% of the known SNPs previously identified from chromosome 6 using GSNAP alone. On the other hand, Stampy has the flaw of being oversensitive, which would result in more false positive calls.

We next focused on one of the five CLL samples (ID 612703) with an average per-base coverage exceeding 100x (103-105x, depending on the mapper used). In the non-HLA regions, the methods showed 95-99% overlap between each other for known SNPs (**Figure 3.14A**).

There was considerable variability in SNP calling within the HLA region (**Figures 3.14B, 3.14C and 3.15A**). GATK UnifiedGenotyper performed better than GATK HaplotypeCaller and Platypus, with GSNAP+GATK UnifiedGenotyper being most sensitive, as observed in NA12878 (**Table 3.6**). GSNAP+GATK UnifiedGenotyper covered over 90% of the known SNPs identified by BWA+ GATK UnifiedGenotyper, BWA+GATK HaplotypeCaller, GSNAP+GATK UnifiedGenotyper, GSNAP+GATK HaplotypeCaller and GSNAP+Platypus together in HLA and *HLA-DRB1*, confirming its excellence in highly divergent regions.

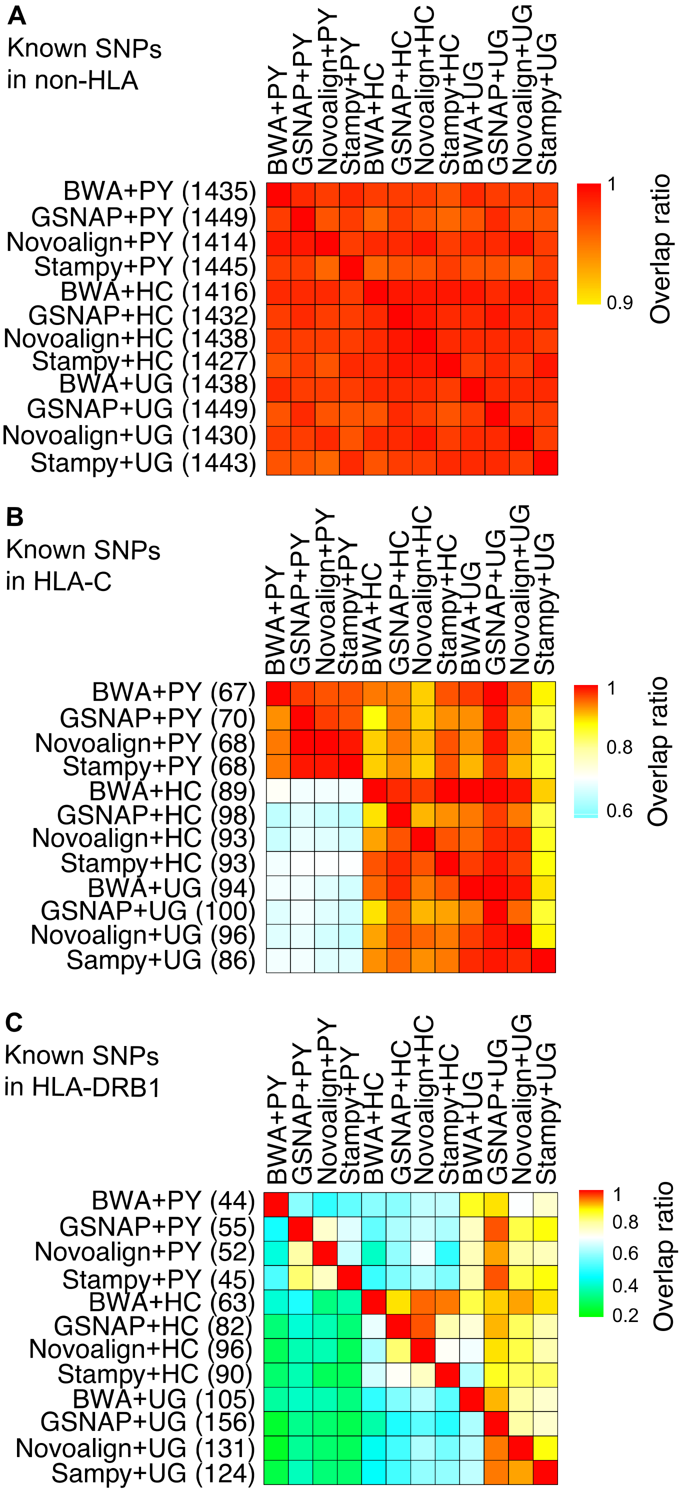


Figure 3.14 Heat maps illustrating the overlap of known SNPs. Variants were identified from a CLL sample 612703. (**A**) Known SNPs in the non-HLA regions of Chr6. (**B**) Known SNPs in *HLA-C*. (**C**) Known SNPs in *HLA-DRB1*. The 12 call

sets were generated by three callers together with four mappers. Number of known variants is shown in parentheses. Each non-triangle box is pseudo-colored to signify the proportion of the call set on the left that is overlapped by the call set showed on the top. HC, GATK HaplotypeCaller; PY, Platypus; UG, GATK UnifiedGenotyper.

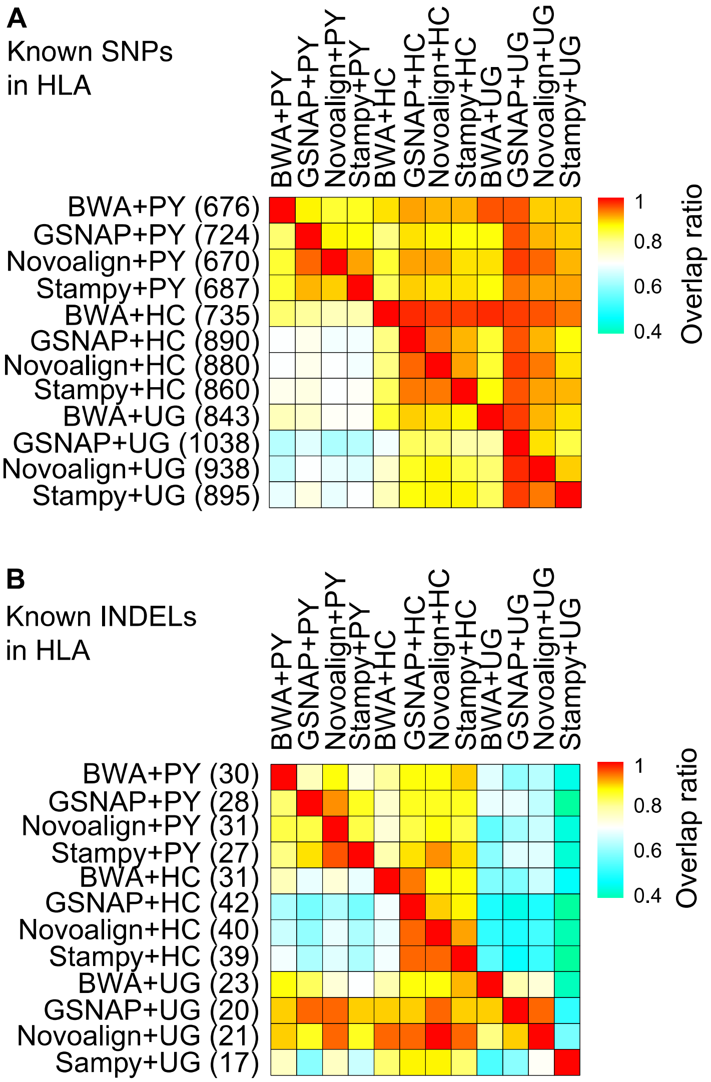


Figure 3.15 Overlap of known variants in the HLA region of the CLL sample 612703. **(A)** Overlap of known SNPs. (**B**) Overlap of known INDELs. The 12 call sets were generated by three callers together with four mappers. Number of known variants is shown in parentheses. Each non-triangle box is pseudo-colored to signify the proportion of the call set on the left that is overlapped by the call set showed on the top. HC, GATK HaplotypeCaller; PY, Platypus; UG, GATK UnifiedGenotyper.

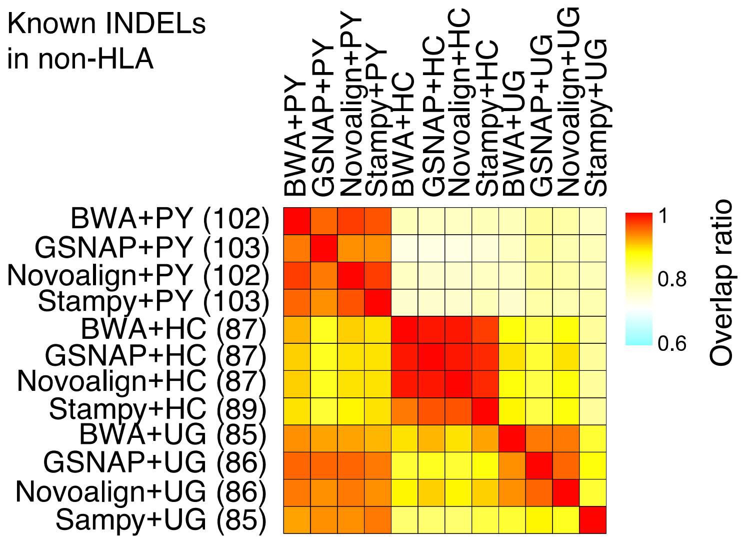


Figure 3.16 Overlap of known INDELs in the non-HLA regions of Chr6. INDELs were identified from the CLL sample 612703. Known INDELs are those that match dbSNP v138. See Figure 3.12 legend for more information.

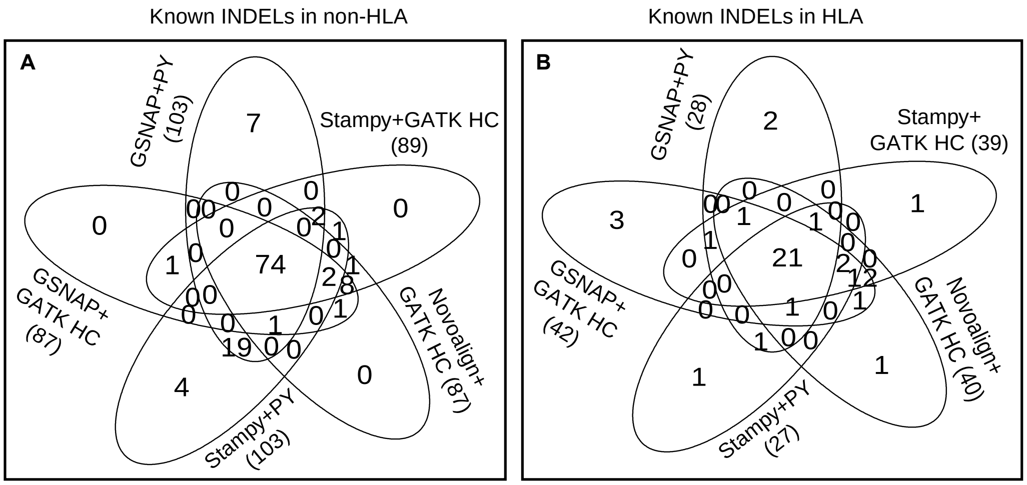


Figure 3.17 Venn diagrams depicting the overlap of known INDELs. INDELs were identified from the CLL sample 612703, using GATK HaplotypeCaller (GATK HC) and Platypus (PY) together with GSNAP, Novoalign and Stampy mapping. Known INDELs are those that match dbSNP v138 and are shown in parentheses. (**A**) Five call sets from the non-HLA regions of Chr6. (**B**) Five call sets from the HLA region.

For INDEL discovery, Platypus performed the best in the non-HLA regions (**Figure 3.16**) and GATK HaplotypeCaller in the HLA region (**Figure 3.15B**), as previously revealed across the CLL cohort (**Table 3.14**). GATK HaplotypeCaller together with Novoalign and GSNAP were the ideal methods for INDEL detection in the HLA region. We further checked overlap of known INDELs detected by five of the methods (**Figures 3.17A** and **3.17B**). The five methods together detected a total of 121 INDELs in the non-HLA regions (**Figure 3.17A**). Among them, GSNAP+Platypus and GSNAP+GATK HaplotypeCaller identified 103 and 87 of them, respectively, and 115 of them together. In the HLA region, the two methods identified 28 and 42 of the 49 known INDELs, and their union accounted for 46 of them (**Figure 3.17B**). These results suggest a need for implementing multiple methods toward more complete INDEL discovery.

3.5. Discussion

Accurate variant discovery is crucial for pinpointing the causal mutations underlying human diseases. Current computational methods are generally effective in detecting ordinary variants but less so to variants located in difficult regions [54]. One of those difficult regions is the HLA region, which is clinically important but extremely divergent. Focusing on chromosome 6 we comprehensively assessed five popular mappers together with five callers on both simulated and real exome-seq data from NA12878. We have developed an analytical workflow that allows more accurate variant discovery in the HLA region and across the genome.

Our analysis revealed marked difference among the five callers at high divergence. GATK UnifiedGenotyper performed the best in single-sample SNP calling, especially with GSNAP, on simulated data at 5-10% divergence. All but Platypus had similarly high precision rates. GATK UnifiedGenotyper was also about 1-6% higher in sensitivity than the other callers in vast majority of the cases with low divergence. For INDEL calling in simulated high divergence data, GATK HaplotypeCaller and Platypus were generally more sensitive, but at the cost of reduced precision. We revealed a similar trend of performance in the HLA and non-HLA regions in NA12878. Therefore, GATK UnifiedGenotyper, SAMtools and GATK HaplotypeCaller are ideal for SNP calling while GATK HaplotypeCaller and Platypus are more effective for INDEL calling.

The mapping accuracy is often calculated by considering only the alignment start position, which does not always reflect the true alignment status of individual bases. In addition, optimal pairwise alignments between individually mapped reads and the reference sequence may not guarantee high confidence in multiple alignments. Therefore, different mappers and callers need to be assessed together in order to identify the best combinations. The five mappers are known to vary in mapping highly divergent reads [21, 25, 26], which we also revealed in our simulated data. Even with a similar mapping rate, two mappers can perform quite differently in the context of variant calling. For example, GSNAP [26] and NextGenMap [25] are both designed to map highly divergent reads and we observed roughly comparable mapping accuracy at 5-10% divergence. However, for the above three sensitive callers in SNP detection, NextGenMap is obviously less suitable than GSNAP, evident by a loss of 8.5-11% sensitivity in the HLA region of NA12878. To support this, of the five GSNAP-mapped read pairs that contained six known SNPs in a 50-bp region in *HLA-DRB5* (32,489,626-32,489,675 bp), NextGenMap mapped only one of the ends correctly (to this 50-bp region) but mapped the other end to regions of 36-63 kb away. In INDEL calling by GATK HaplotypeCaller and Platypus, NextGenMap was 9.7-20% less sensitive than GSNAP in the HLA region of NA12878. On the other hand, there is also obvious difference across different callers given the same mapper. Using GSNAP as the mapper, GATK UnifiedGenotyper was 2-4.4% more sensitive than GATK HaplotypeCaller and SAMtools and 12-14% more sensitive than the other two in SNP calling from the HLA region. Together, our analysis has identified GSNAP+GATK UnifiedGenotyper as the most sensitive method for SNP detection in both HLA and non-HLA regions.

Strikingly, GSNAP+GATK UnifiedGenotyper achieved 100% sensitivity in *HLA-DRB1*, *HLA-A* and *HLA-C*, three highly polymorphic genes in NA12878. In *HLA-DRB1*, for example, this method identified all the 70 known SNPs annotated in the public call set [51, 54], plus an additional 30 unique known calls, of which 23 were in SNP cluster. In contrast, BWA+GATK UnifiedGenotyper, a widely used variant detection method, missed >30% of the known SNPs in this gene. We analyzed two additional NA12878 exome-seq datasets generated using Illumina Nextera Rapid Capture Exomes capture kit (SRR1919605) and Roche Life Science SeqCap EZ Human Exome Library v3.0 (SRR1611181). GSNAP+GATK UnifiedGenotyper also showed the highest sensitivity in SNP calling in the HLA region (data not shown). Comparable results were obtained by GATK HaplotypeCaller and Platypus in INDEL calling.

Traditionally, genotyping in the HLA region often relies on microarray hybridization [43] or sequencing of PCR amplicons targeting selected exons [42, 78] or entire genes [95], which are costly, time-consuming and low throughput. In addition, current studies often focus on known variants in the IMGT/HLA Database without considering novel variants [96, 97]. Lastly, though WES was used for HLA genotyping and variant discovery, the mapping-based approach did not work well even at very high coverage [98]. Given these limitations, our approach represents a more generalized methodology, which is effective for genome-wide variant detection but particularly sensitive in highly divergent regions like HLA. Though only tested on WES, it should be applicable to whole genome sequencing data as well.

3.6. Conclusions

We aimed to develop a strategy enabling more accurate variant discovery in highly divergent regions. Focusing on the HLA region which shows extreme divergence across different haplotypes, we have revealed marked differences among those methods in SNP and INDEL calling from NA12878 WES data. We capture a similar trend on WES data from a cohort of chronic lymphocytic leukemia patients. Specifically, GSNAP and Novoalign achieve high sensitivity in mapping divergent reads without losing the precision. The limitation in speed could be overcome through a two-step mapping approach, in which reads are first mapped by BWA and unmapped ones are then re-mapped by GSNAP or Novoalign. Together with these two mappers, GATK UnifiedGenotyper demonstrates its excellence in SNP calling, followed by GATK HaplotypeCaller and SAMtools; in INDEL calling, GATK HaplotypeCaller and Platypus outperform the others and a joint calling clearly enhances the outcome. Given that highly polymorphic regions distribute over many chromosomes and that they are often associated with human disease, our study brings additional options into the current variant calling practice.

Chapter 4

Toward *de novo* assembly-based variation discovery in personal genomes

4.1. Summary

Current variant discovery approaches often rely on an initial reads mapping to the reference sequence. Their effectiveness is limited by the presence of gaps, potential misassemblies, regions of duplicates with high sequence similarity and regions of high sequence divergence in the reference. Also, mapping-based approaches are less sensitive to large INDELs and complex structural variations and fail to provide long-range phase information in personal genomes. A few *de novo* assemblers have been developed to identify variants through direct variant calling from the assembly graph, micro-assembly and whole genome assembly, but mainly for whole genome sequencing data. Whole exome sequencing (WES) is preferred over whole genome sequencing in identifying causal variants. We developed SGA-haplotype, a *de novo* assembly workflow for haplotype-based variant discovery from WES. In addition to the error correction and string graph-based assembly model adopted from String Graph Assembler (SGA), SGA-haplotype provides other core functionalities for reads partition and quality-based preprocessing, contig filtering, as well as contig mapping to the reference and variant detection. Using simulated human exome data, we compared SGA-haplotype to five variation-aware *de novo* assemblers: Cortex, fermi, fermi2, and two direct variant callers in the SGA graph-diff module, and to BWA-MEM together with three haplotype-based callers, GATK HaplotypeCaller, FreeBayes and Platypus. SGA-haplotype outperforms the other assemblers in sensitivity and tolerance of sequencing errors. We recapitulated the findings on whole genome and exome data from a CEU trio, showing that SGA-haplotype had high sensitivity both in the highly divergent human leukocyte antigen (HLA) region and in non-HLA regions of chromosome 6. In particular, SGA-haplotype is more robust to sequencing error, k-mer selection, divergent level and coverage depth. Unlike mapping-based approaches, SGA-haplotype is capable of resolving long-range phase and identifying large INDELs from WES, more prominently from WGS as expected. We conclude that SGA-haplotype represents an ideal platform for WES-based variant discovery in highly divergent regions and across the entire genome.

Keywords: de Bruijn graph; Exome sequencing; Human leukocyte antigen; Micro-assembly; String graph; Structural variation

4.2. Introduction

Complete and accurate detection of sequence variations is a key prerequisite for deciphering the genetic etiology of disease [41, 54]. Mapping-based approaches currently dominate the field of variant discovery from whole genome and exome sequencing, but have limitations in several key aspects. First, the human reference sequence is not perfect [57], containing misassemblies [17] and gaps [99]. In addition, some regions show high sequence divergence or complex structural variations between haplotypes, or represent recent duplications with high sequence similarity, increasing the ambiguity in short reads mapping [4, 21]. The 1000 Genomes Project Consortium estimated 171 Mb (5.5%) of the human genome reference GRCh37 as being inaccessible to the short-read sequencing technologies [58], including 130 Mb of recent segmental duplications between which a reliable differentiation is often difficult [100]. Importantly, some of the missing, highly divergent and inaccessible regions are associated with human disease [41]. Current standard variation discovery practice is less effective in these instances [41]. For example, a recent study revealed that mapping bias led to 19% error rate in calling single nucleotide polymorphism (SNP) genotypes for genes in the HLA region that is associated with over 100 diseases [56].

Also, mapping efficiency often biases toward the reference allele in the presence of an INDEL [21], reducing the detection of INDELs, especially large ones and those located within microsatellites [26, 53]. Even though numerous tools have been developed specifically for INDEL detection, an initial read mapping is still required [101]. Finally, many downstream analyses require the determination of accurate haplotype structure to infer causal mutations. Current mapping-based approaches usually report un-phased genotypes or limited phasing information when combined with local *de novo* assembly based variant calling. Without trio or reference population data, it will be difficult to infer haplotype and discover transmitted variants in personal genomes.

The variation-aware *de novo* assemblers represent an attractive alternative. In principle, they are similar to the consensus sequence assemblers through the implementation of either de Bruijn graph [57] or string graph [17, 68]. In de Bruijn graph-based assembly paradigm, reads are split into k-mers that are often error-corrected before used to build contigs. Thus, the performance of an assembler is correlated with the k-mer coverage rather than the base coverage, which highlights the importance of k-mer selection, especially when long reads are used. In string graph-based assembly, contigs represent paths of string graph and are built from overlapped reads. Read coherence is fully retained in string graph but lost in de Bruijn graph. However, there is a key distinction between variation-aware and consensus sequence assemblers when applied to non-haploid organisms. While the former try to preserve heterozygotes at polymorphic sites, the latter collapse them into consensus bases [102]. The variation-aware assemblers are capable of identifying long INDELs and structural variations [17, 18, 41, 53], and assembling reads into haplotype contigs [103].

A few packages have been developed for variant discovery through direct variant calling from the graph [57], local assembly (micro-assembly) using mapped reads [53, 54], or whole genome *de novo* assembly [17, 104, 105]. In local and whole genome assembly, the resulting contigs or unitigs (uniquely assembled contigs) are mapped back to the reference before variant calling.

Cortex is the first *de novo* assembly-based algorithm for direct variant calling from short reads. Cortex implements colored de Bruijn graph and performs single- or multiple-sample variant calling, utilizing the reference sequence if available [57]. In the graph, the nodes and edges are colored by the samples having them, thus allowing population-based variant discovery. Variants are directly called from the graph through the functions “bubble-calling” or “path-divergence calling”. The bubble-calling algorithm is designed to identify simple variants appearing as “clean” bubbles in the graph. Such bubbles represent sequence divergence between samples, which are distinct from those caused by repeats that are expected to occur in both samples. On the other hand, the path-divergence algorithm targets complex (homozygous) variants by following the deviation of the reference path from the sample path. However, the current version has a low sensitivity, with nearly 40% false negative rate [54].

The String Graph Assembler (SGA) was originally developed for consensus sequence assembly of large genomes [68]. It creates overlapping graphs by first performing Burrows-Wheeler Transform and building FM (Ferragina Manzini) index from error-corrected reads [68]. With the implementation of such a data structure, SGA can efficiently compute the path of string graph, thus alleviating the requirement of high computing when applied to NGS datasets from large genomes, a limitation inherent to string graph [60]. The “graph-diff” module, a supplement to the SGA package, is still under active development. It provides two modes to call variants directly from string graph and de Bruijn graph. However, its efficiency has yet to be assessed.

Scalpel [53] and DISCOVAR [54] perform local assembly based on de Bruijn graph. Reads are first mapped to the genome reference and pairs with at least one mapped read are included. Scalpel was designed specifically for INDEL detection from whole genome or exome sequencing data [53]. On the other hand, DISCOVAR was developed for assembling long (250-bp) reads from whole genome sequencing of PCR-free libraries; its error-correction and variant detection algorithms might not work well on the typical shorter (76-150 bp) reads that are often sequenced from libraries involving PCR amplification [54]. Importantly, their performance in highly divergent regions remains unknown.

Finally, fermi and FermiKit (fermi2) are string graph-based whole genome assemblers. Fermi implements FM-DNA index for forward-backward extension of DNA sequences [17], a variety of FM index used in SGA. It outputs unitigs in the final assembly that preserve SNPs, short INDELs and structural variations, without providing the functionality for unitig mapping and variant calling. FermiKit is an updated version of fermi [104], which uses the BFC algorithm [106] for less greedy error correction compared to the one used by fermi. It also includes BWA-MEM [92] for mapping unitigs to the reference and HTSBox (https://github.com/lh3/htsbox) for variant detection. Nevertheless, both versions are highly packaged, lacking the flexibility needed for fine-tuning parameters when applied to WES or targeted gene panel sequencing data.

Except for Scalpel, the other packages described above were developed for whole genome sequencing data; their performance on WES has yet to be assessed. Scalpel was designed exclusively for INDEL detection in WES and WGS [53]. It implements de Bruijn graph, a data structure tending to be less effective compared to string graph when complex regions are assembled or long reads are used. *De novo* assembly of WES data is complicated by the considerable variability in coverage caused by capture, sequencing and mapping bias [4, 6, 74]. It is even more challenging for highly divergent regions.

We developed “SGA-haplotype”, a *de novo* assembly pipeline for haplotype-based variant discovery in WES. SGA-haplotype took advantage of the high-performing error-correction module for complex genomes and the efficient data structure implemented in SGA. Using simulated exome data and real whole genome and exome data, we assessed this workflow along with five *de novo* assembly-based and three mapping-based variant discovery methods. SGA-haplotype demonstrates excellence in both sensitivity and precision, which is largely independent of variability in coverage and divergence. SGA-haplotype can achieve long-range phasing over some of the most divergent HLA genes. It is powerful in detecting large INDELs, SNP clusters and complex structural variations from regions like HLA.

4.3. Methods

4.3.1. Overall analytical strategy

We developed SGA-haplotype (**Figure 4.1**) and compared it to five other variation-aware *de novo* assemblers in variant discovery from WES (**Figure 4.2; Table 4.1**). Cortex [57] and two methods in the graph-diff module of SGA [68] perform direct variant calling from the assembly graph. Fermi [17], fermiKit (fermi2) [104] and SGA-haplotype assemble reads into contigs, followed by contig mapping to the reference and variant calling. The initial assessment used simulated error-free reads. SGA paired de Bruijn direct calling (SGA-PDBG), SGA-haplotype and fermi2 that showed relatively better performance were then assessed using simulated reads with error; the first two were finally assessed on NA12878 whole genome and whole exome data. As controls, BWA-backtrack [28] (referred to as BWA) or BWA-MEM [92] together with three haplotype-based callers were also tested on simulated reads with error and NA12878 exome data (**Figure 4.2**).

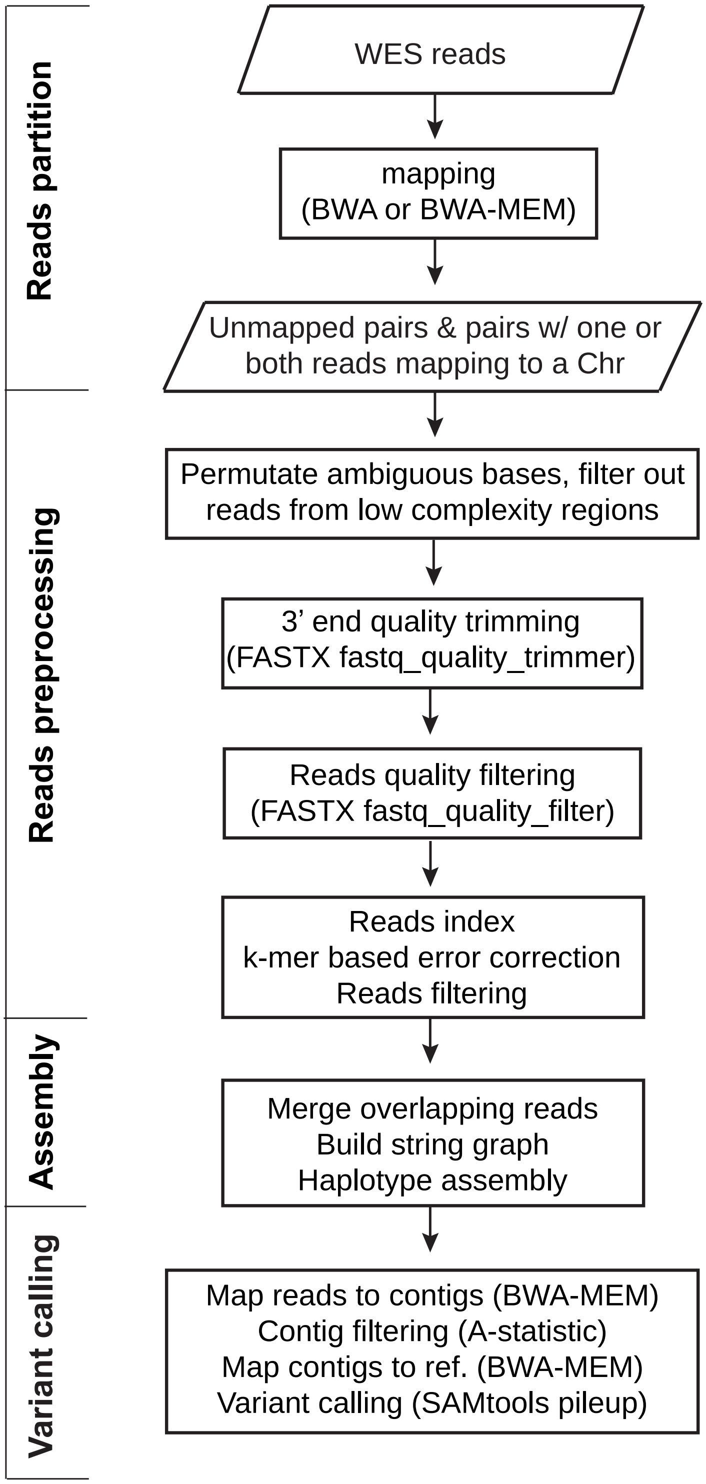

Reads partition
Reads preprocessing
Assembly
Variant calling
WES reads
mapping
(BWA or BWA-MEM)
Unmapped pairs & pairs w/ one or
both reads mapping to a Chr
Permutate ambiguous bases, filter out
reads from low complexity regions
3' end quality trimming
(FASTX fastq_quality_trimmer)
Reads quality filtering
(FASTX fastq_quality_filter)
Reads index
k-mer based error correction
Reads filtering
Merge overlapping reads
Build string graph
Haplotype assembly
Map reads to contigs (BWA-MEM)
Contig filtering (A-statistic)
Map contigs to ref. (BWA-MEM)
Variant calling (SAMtools pileup)

Figure 4.1 Key steps in SGA-hap variant discovery pipeline. Tools are provided in the parenthesis.

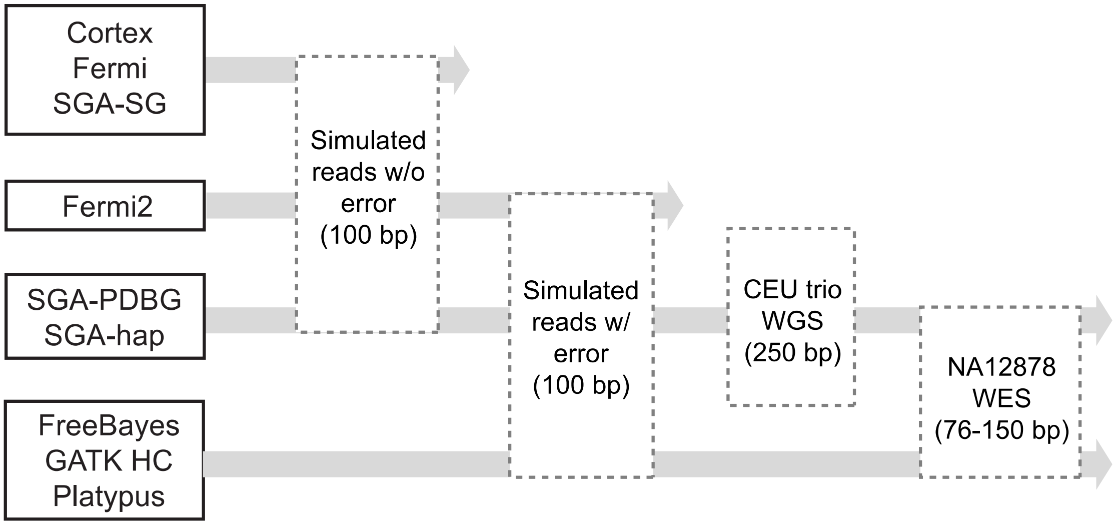


Figure 4.2 Flowchart illustrating variant discovery methods assessed in the study. GATK HC, GATK HaplotypeCaller; SGA-SG, SGA graph-diff string graph mode; SGA-PDBG, SGA graph-diff paired de Bruijn mode. All six *de novo* assemblers were first assessed on simulated reads without error; three of them, SGA-PDBG, SGA-haplotype and fermi2, showed relative better performance and were further assessed for their tolerance to sequencing error using reads simulated with error. The first two as representatives of de Bruijn graph- and string graph-based assemblers were finally assessed on the NA12878 WES and WGS data from a CEU trio. In parallel, three mapping-based approaches were also included as controls and tested on simulated reads with error and NA12878 WES data.

Table 4.1 Six *de novo* assembly-based variant calling methods tested in this study

Tool	Algorithm	Implementation	Application	Test datasets			Ref.
				A	B	C	
Cortex	De Bruijn	Direct calling	WGS	x			[57]
Fermi	String graph	Build unitigs, map unitigs and call variants	WGS	x			[17]
FermiKit	String graph	Build unitigs, map unitigs and call variants	WGS	x	x		[104]
SGA-PDBG	De Bruijn	Direct calling	WGS	x	x	x	[68]
SGA-SG	String graph	Direct calling	WGS	x			[68]
SGA-hap	String graph	Build and filter contigs, map contigs and call variants	WES	x	x	x	This study

WGS, whole genome sequencing; WES, whole exome sequencing.
A, simulated reads without error; B, simulated reads with error; C, CEU trio (NA12878, NA12891 and NA12892) WES data.

As *de novo* assembly based approaches are expected to be effective in regions with high divergence and heterozygosity, where high density of SNPs should help resolve haplotypes, we focused on the comparison between the highly divergent HLA region and the rest of chromosome 6. The HLA region encodes antigen-presenting molecules that play essential roles in the immune system [107]. We analyzed the impact of coverage and divergence on *de novo* assembly-based variant detection. We also sought to investigate how sequencing error impacts de Bruijn graph- versus string graph-based assemblers. For the data simulated with sequencing error, the dumpy base quality score does not fully reflect the characteristic error profile in real Illumina sequencing data. We thus performed a simple preprocessing by arbitrarily trimming 10 bases from the 3’ end, followed by k-mer based error correction. For whole genome and exome data from the CEU trio, a systematic, quality-based read filtering and end trimming was applied (see below).

4.3.2. Quality metrics

For simulated data, the performance was assessed based on sensitivity, precision rate and overall genotype concordance, where the pre-placed variants were treated as the “true positive” [108]. We estimated the sensitivity for NA12878 exome data, using the union of two public call lists below as a proxy for reference call set (“true positive”). The high-confidence call set was generated from 11 whole genome and 3 exome sequencing data using seven mappers and three callers [51], which is available at ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/GIAB_integration/NIST_RTG_PlatGen_merged_highconfidence_v0.2.primitives.vcf.gz. The second call set was merged from three separate variant lists, which were generated by Cortex, DISCOVAR and GATK HaplotypeCaller from 250-bp paired-end reads of a PCR-free genomic library [54]. The list is available at ftp://ftp.broadinstitute.org/pub/crd/DiscovarManuscript/vcf/. Considering the difficult in defining the exact boundary, for simulated data, INDELs identified within 10-bp of “true” ones were counted as matches. As the reference-quality sequence is not available for NA12878, we inferred the precision by i) manually checking alignments around a subset of randomly selected variants; and ii) performing regional assembly with the high coverage 2x250 bp whole genome sequencing data from PCR-free libraries in a CEU (Utah residents with Northern and Western European ancestry) trio (NA12878, NA12891 and NA12892).

4.3.3. Test datasets

We used simulated exome data and publicly available whole exome data from NA12878 and whole genome data from a CEU trio including NA12878, NA12891 and NA12892 (**Table 4.2**). The 250-bp whole genome sequencing data (~60x coverage) were generated using a TruSeq PCR-free protocol by the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk). The 100-bp NA12878 exome data were generated using Roche SeqCap EZ Human Exome kit (with 64 Mb targets) and the 76-bp and 150-bp exome data were generated using Illumina Nextera Rapid Capture Exome kit (with 37 Mb targets).

Table 4.2 Public exome data used in the study

Sequence ID	Length (bp)	Platform	Capture kit
NA12878-NGv3-LAB1360-A	100	HiSeq 2000	Roche
SRR1611181 (NA12878)	100	HiSeq 2000	Roche
SRR1611182 (NA12878)	100	HiSeq 2000	Roche
SRR1611183 (NA12878)	100	HiSeq 2000	Roche
NA12878-NRCE-N701_S2S13S4S37	76	HiSeq 4000	Illumina
NA12878-NRCE-N702_S1S2S4S38	76	HiSeq 4000	Illumina
FC1_NA12878_S1_S4	150	HiSeq 2500	Illumina
FC1_NA12878_S7_S10	150	HiSeq 2500	Illumina

The 76-bp and 150-bp reads are available at https://basespace.illumina.com/. The 100-bp reads from SRR1611181, SRR1611182 and SRR1611183 were downloaded from the National Center for Biotechnology Information Sequence Read Archive. NA12878-NGv3-LAB1360-A reads were downloaded from https://s3.amazonaws.com/bcbio_nextgen/. Roche, SeqCap EZ Human Exome kit v3.0 (with 64 Mb targets); Illumina, Nextera Rapid Capture Exome kit (with 37 Mb targets).

Simulated reads provide a simple system to evaluate the performance of variant discovery methods over key factors (like sequencing error, coverage and divergence level) since both "true" (pre-placed by simulation) and "false" variants are known [109]. We generated two simulated datasets, with and without sequencing error, from exonic regions of chromosome 6. These regions, totaling 6,396,915 bp, were compiled from hg19 refGene exon annotation and the capture regions interrogated by four Agilent SureSelectXT All Exon kits [108].

Dwgsim (v0.1.11, https://github.com/nh13/DWGSIM/wiki) was used to simulate 100-bp paired-end reads to a per-base coverage of 400x at each of the seven mutation rates between 0.05% and 15% (parameter -r). A control with 0% mutation rate was also generated. Dwgsim was run with the parameter settings "-C 400 -1 100 -2 100 -e $error_rate -E $error_rate -r $mutation_rate -d 200 -R 0.1 -I 1 -y 0 -n 0 -c 0 -S 2 -z 123 chr6_input.fa $output_prefix", where $error_rate takes 0 (error-free) or 0.0001-0.01 (0.01% error rate at the start and 1% error rate at the end of reads) and $mutation_rate takes 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, or 0.15. Coverage is a key factor in *de novo* assembly [110]. To examine the impact of coverage, the 400x coverage error-free reads were randomly down-sampled into 200x, 160x, 120x, 100x, 80x, 60x, 40x and 20x coverage using seqtk (https://github.com/lh3/seqtk). The simulated reads with error were down-sampled into 200x, 100x, 60x and 40x coverage.

4.3.4. *De novo* assembly-based variant detection

4.3.4.1. The development of SGA-haplotype

We built a workflow, termed “SGA-haplotype”, for haplotype-based variant discovery from WES (**Figure 4.1**). The workflow was designed to retain the heterozygosity between haplotypes, maximize the use of informative reads, minimize the impact of sequencing error, increase the robustness to divergence and coverage variability, and reduce computing requirements. It adopts the k-mer based error-correction and string graph assembly module from the SGA consensus assembly framework. We implemented other key functionalities for i) reads partition based on mapping status; ii) reads preprocessing; iii) local or chromosome-level haplotype assembly; iv) reads mapping to contigs and filtering of non-unique contigs; v) contigs mapping to the reference; and vi) variant detection from contig-reference alignments. We set up key rules to reinforce the haplotype assembly by requiring an exact overlap of at least length L when merging reads or sequences and using k-mer-based rather than overlap-based error correction. The key parameters were provided in **Table 4.3**.

Chr6 reference sequence was FM indexed using the commands “sga preprocess --permute chr6.fa”, “sga index”, and “sga gen-ssa”. The FM indexing of reads was repeated (via “sga index” command) each time when the input fastq file was changed. To extract reads for chromosome-level assembly, NA12878 reads were first mapped to the reference using BWA. We kept pairs with at least one end uniquely mapped to Chr6 and a mapping quality score of >=20. Considering that some of the unmapped reads might be from highly divergent regions or represent sample-specific novel sequences, unmapped pairs were also selected.

Table 4.3 Key parameters settings for SGA-hap

Type	Read length (bp)	K-mers	Minimal overlap (bp)		Minimal contig size (bp)	
			A	B	C	D
Simulated reads	100	31, 41, 51, 61	45	75	125	150
Real WES	76	27, 31, 41, 51	35	55	100	125
Real WES	100	27, 31, 41, 51, 61	45	75	125	150
Real WES	150	51, 61, 71, 81, 91	75	100	175	225
Real WGS	250	31, 51, 61, 91, 121	75	125	275	325

A, minimal overlap for merging reads into sequences; B, minimal overlap for assembling sequences into contigs; C, minimal size for contigs subjected to A-statistic test; D, minimal size for contigs not subjected to A-statistic test.

The retained reads were processed by trimming off low-quality bases and filtering out low-quality reads (**Figure 4.2**). In the first step, reads that were shorter than 65 bp, had uncalled bases (Ns), or contained low complexity sequence (dust score >4) were filtered out using “sga preprocess --pe-mode 1 --dust -m 65”. Next, tools from the FASTX-Toolkit package (v0.0.14, http://hannonlab.cshl.edu/fastx_toolkit/) were used for quality-based 3’ end trimming and read filtering. Specifically, “fastq_quality_trimmer” (options “-t 30 -l 65”) was used to trim low-quality bases from the 3’ end until the base with a quality score of >=30 was reached, discarding reads shorter than 65 bp after the 3’ end trimming; a further filtering was performed using “fastq_quality_filter” with the parameters “-q $base_quality -p 50”, which filtered out reads with a Phred-scale quality score lower than the threshold (20 for 150-bp reads and 30 for 76-bp and 100-bp reads) in at least half of the bases. The retained reads were error corrected based on the k-mer frequency, using “sga correct” with the options “-a kmer -k $kmer -x 2 -discard”. We tested four or five k-mers (parameter -k) between 27 and 91 in error correction, depending on read length (**Table 4.3**).

After error correction, duplicates with identical sequences were collapsed into a single one; low-complexity reads as well as reads covering one or more low-frequency k-mers (parameter -x, set to 2, see below) suggestive of potentially uncorrected bases were filtered out. All were done using the command “sga filter -k $kmer -x 2 --substring-only --low-complexity-check --homopolymer-check”.

Reads showing sufficient overlap (parameter -m) were merged together using “sga fm-merge -m $min_overlap”, where $min_overlap takes 35 (for 76-bp reads), 45 (for 100-bp reads), or 75 (for 150-bp reads). After merging, some reads (or merged sequences) may become substrings (subsequences) of other sequences. Subsequences were filtered out using “sga rmdup” as they are not allowed in the next step. Pairwise overlaps between all the retained sequences were computed using “sga overlap --exact -m $min_overlap”. We ran the command “sga assemble -m $min_overlap -r 7 -d 0 -g 0 -l $min_contig_length” to assemble sequences into contigs and remove short contigs below the minimum contig length cutoff, where $min_overlap takes 55 (for 76-bp reads), 75 (for 100-bp reads), or 100 (for 150-bp reads) and $min_contig_length takes 100 (for 76-bp reads), 125 (for 100-bp reads), or 175 (for 150-bp reads).

Considering that contigs assembled from highly repetitive regions are less reliable than those from unique regions, we assessed the contig uniqueness by mapping raw reads back to the contigs using BWA-MEM. An A-statistic was calculated for each contig using the command “sga-astat.py -m $min_contig_length -g $size read_contig.bam”. Given that roughly 60% (1/0.6=1.67) of exome reads are mapped to capture regions, $size is calculated as 1.67 times the total size of the capture regions. A-statistic approximates the log-odds ratio between a contig being unique versus being repetitive [111, 112]. Short contigs tend to be less unique and thus have lower A-statistic scores. Only contigs that either had A-statistic scores >=10 or met the minimum length threshold (125 bp for 76-bp reads, 150 bp for 100-bp reads, and 225 bp for 150-bp reads) were kept for further analysis.

BWA-MEM was used to map the retained contigs to the hg19 reference. SNPs were identified using SAMtools and those in the target regions were extracted using the Tabix tool [113]. INDELs were identified through SAMtools and left-aligned for easy comparison. To increase the sensitivity in INDEL discovery, we also tested a two-step contig mapping strategy with BWA-MEM followed by BLAT, a more permissive aligner. The contigs with soft-clipping in BWA-MEM alignments were re-mapped by BLAT, using the parameter settings "-tileSize=11 -minMatch=5 -maxIntron=100000 --top 1 --min_per_ID 85". The alignments were converted into BAM format and combined with the BWA-MEM alignments without soft-clipping. INDELs were identified as above.

In parallel, simulated reads were subject to the first step filtering described above. For simulated reads with error, fastx_trimmer was then used to arbitrarily trim 10 bp from the 3' end. Further processing of retained reads, reads assembly and contig filtering, as well as contig mapping and variant calling were performed following the procedure above, except that only BWA-MEM was used to map contig to the reference.

4.3.4.2. Direct variant calling with graph-diff

We used the graph-diff module from the SGA package to call variants directly from paired de Bruijn graph (SGA-PDBG, used for all data) and from string graph (SGA-SG, only for simulated reads without error). Graph-diff performs reference indexing, read preprocessing and indexing, and direct variant calling and filtering.

Reads were preprocessed as described in SGA-haplotype. We tested four k-mers (parameter -k, see below) for both methods, i.e., 31, 41, 51, and 61. Variants were identified from the string graph by the command "sga graph-diff -p $out_prefix -k $kmer -x 2 -t $NSLOTS -m 45 -r input.fastq --ref indexed_chr6.fa", requiring a minimal overlap (parameter -m) of 45 bp (default). For variant detection from the paired de Bruijn, we used the command "sga graph-diff -p $out_prefix -t $NSLOTS -k $kmer --min-discovery-count=2 --paired-debruijn --min-dbg-count=2 -r input.fastq --ref indexed_chr6.fa". In both calling modes, the provided script sga-variant-filters.pl was used to filter out raw calls from low complexity (dust score >4.0) and homopolymers (> 9 bp) regions, as well as those with strand bias (strand_cutoff=4). Multiple-nucleotide polymorphisms (MNPs), i.e. a single event consisted of two or more variants within 5 bp, were decomposed into individual SNPs for comparison.

At each polymorphic site on Chr6, genotype was then assigned as 0/0 (AF <0.2), 0/1 (0.2<=AF<=0.8), or 1/1 (AF >0.8), based on allele frequency (AF) cutoffs (0.2 and 0.8) that were commonly used in the early exome-seq-based variant discovery studies [33]. Heterozygous (0/1) and homozygous (1/1) variants were kept and split into SNPs and INDELs.

4.3.4.3. *De novo* assembly using fermi and fermiKit

Fermi is a variation-aware assembler for whole genome sequence. Internally, fermi performs quality-aware error correction based on under-represented k-mers but rarely removes "true" heterozygotes. Simply speaking, fermi first collects all 23-mers, each occurring at least 3 times in the dataset, and counts the occurrence of the different 24^{th} bases. It then scans each read using the 23-mer hash table and converts the low-quality and low-frequency 24^{th} base(s) to the most dominant base if a big difference exists in their frequencies. Sequencing error for INDELs is not corrected.

Fermi was only used to assemble simulated reads without error. We first ran "run-fermi.pl -e fermi -k $overlap -p $prefix end1.fastq.gz end2.fastq.gz" and then the "make" commands, which constructed the FMD (Ferragina-Manzini DNA) index from raw reads, performed error correction described above, and constructed the FMD index again using error-corrected reads. We tested 45, 55, 65 and 75 bp as the minimum overlap cutoffs (parameter -k) in merging reads. The "fermi unitig" command was then used to build unitigs (uniquely assembled contig) from the above error-corrected reads. These raw unitigs were cleaned by removing tips of less than 125 bp and weak overlaps supported by less than two reads, using the command "fermi clean -l 125 -e 4 -i 3 -o $overlap -n 6 -w 2 -r 0.15 -C -S". Bubbles could be caused by both polymorphisms and sequencing errors [114]. As no sequencing errors were introduced in simulation, bubble popping was disabled in the above step to preserve bubbles (by parameter -S). The mapping of cleaned unitigs to the reference and variant detection followed the procedure in SGA-haplotype.

We used fermiKit (fermi2) to call variants from simulated reads with or without error, following the online user manual (https://github.com/lh3/fermikit). For simulated reads with error, the last 10-bp from the 3' end were first trimmed off. Reads were assembled into unitigs using the two commands below:

```
fermi2.pl unitig -s3g -t $NSLOTS -l $read_length -p $prefix end1.end2.fastq.gz >
$prefix.mak
make -f $prefix.mak
```

The output unitig file $prefix.mag.gz was used to detect SNPs, small INDELs and structural variations, using "run-calling -t $NSLOTS BWA-MEM_indexed_ref.fa $prefix.mag.gz | sh". This command calls BWA-MEM for mapping unitigs to the reference and HTSBox for variant detection.

4.3.4.4. Direct variant calling with Cortex

We called variants from simulated reads without error following the instructions in the user manual (http://cortexassembler.sourceforge.net/cortex_var_user_manual.pdf). In brief, Chr6 reference binaries were built at four k-mers between 31 and 61 (step size 10), using commands “cortex_var_31_c1” (for 31-mer), “cortex_var_63_c1” (41-mer and 51-mer) and “cortex_var_95_c1” (61-mer). The use of different k-mers allows us to assess how k-mer selection might impact Cortex variant calling. The wrapper script run_calls.pl was used to call variants through both bubble-calling (parameter --bc) and path-divergence calling (parameter --pd).

```
run_calls.pl --first_kmer 31 --last_kmer 61 --kmer_step 10 --auto_cleaning yes --
bc yes --pd yes --ref CoordinatesAndInCalling --ploidy 2 --qthresh 1 --dups  --
do_union yes --homopol 9 --mem_height 25 --mem_width 100 --max_var_len 50
--genome_size 6396915 --stampy_hash chr6.fa --stampy_bin /path/to/stampy.py
--fastaq_index FILE_listing_fastq --refbindir /path/to/Chr6_reference_binary/ --
list_ref_fasta FILE_listing_ path_to_chr6.fa --vcftools_dir /path/to/ vcftools/ --
workflow joint.
```

This script calls the Stampy package to map the 5’-flanking sequence of a variant to the reference, by which the genomic coordinate of the variant itself can be inferred. For some of the variants, their 5’-flanking sequences failed to align to the reference or had a mapping quality below 40. These variants had the flag “MISMAPPED_UNPLACEABLE” in the resulting primitive VCF file and were filtered out. Cortex generated a variant list at each of the four k-mers. Multiple-nucleotide polymorphisms (MNPs) were decomposed into individual SNPs.

4.3.5. Mapping-based variant detection

As controls, we also performed mapping-based variant calling for simulated reads with error and NA12878 exome data. Reads were mapped using BWA (BWA-backtrack) or BWA-MEM. We previously found that local realignment and base quality score recalibration (BQSR) had little benefit on variant calling methods using these two mappers together with the three haplotype-based callers (**Figure 4.2**) [115]. Therefore, after duplicate marking, we only performed local realignment for NA12878 using the Mills and 1000G gold standard INDELs (Mills_and_1000G_gold_standard.indels.hg19.vcf.gz). Variants were identified with FreeBayes (v9.9.2-27) [34], GATK HaplotypeCaller (v2.7-2) [15, 29], and Platypus (v0.5.2) [4], following our previous study [115].

4.4. Results

4.4.1. The six *de novo* assemblers showed marked differences on simulated reads

In SGA-haplotype, we implemented key components and used highly stringent parameters to assembling reads into haplotype contigs and preventing the formation of chimeric contigs (see Methods for details). To validate our strategy, we estimated contig coverage by mapping reads back to the contigs. Following the increase of divergence, the degree of heterozygosity should increase, so does the number of haplotype contigs. When the two genomic copies are sufficiently divergent, as in the cases of highly divergent regions, there should be two dominant haplotype contigs and each should have about half of the sampled read depth. This is indeed the case in the simulated data (100x, **Figure 4.3**). We found that, at 10% and 15% divergence, the vast majority of the contigs had about half (50x) of the simulated 100x coverage. At the lowest divergence of only 0.05% and 0.1%, we observed a biomodal distribution of coverage depth, with the majority of the contigs having around 100x coverage or <50x coverage.

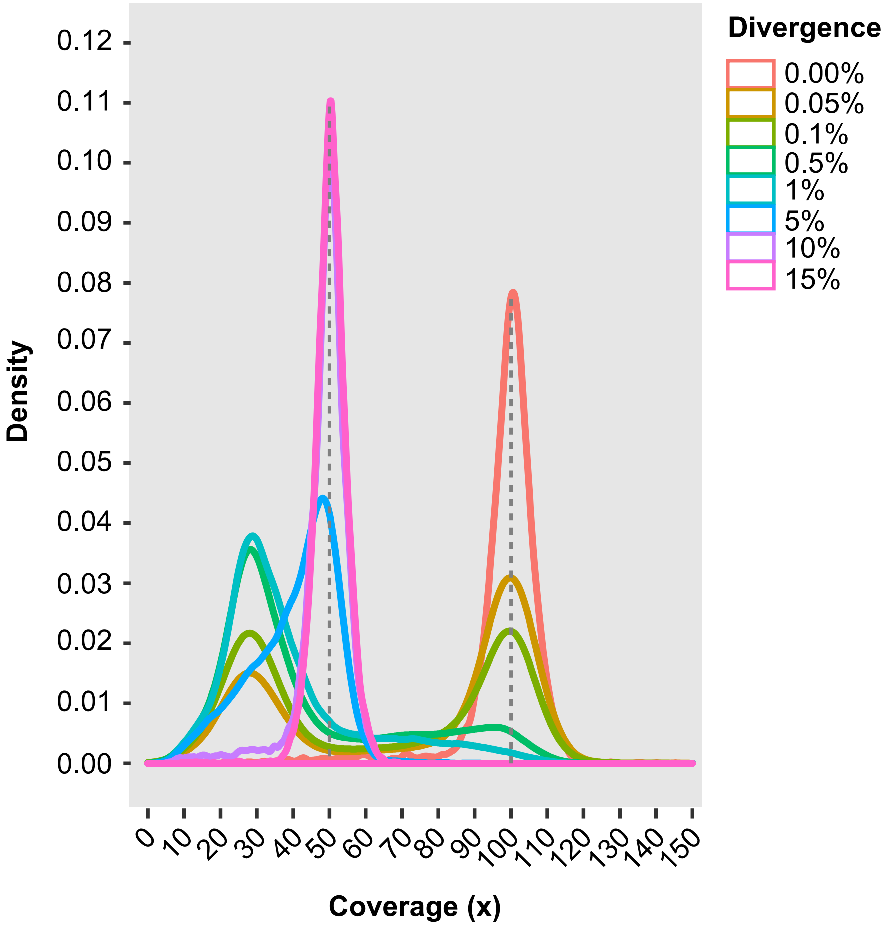


Figure 4.3 Coverage plot for contigs assembled by SGA-haplotype at different divergence levels. The 100x coverage error-free reads were used.

We first assessed SGA-haplotype with the other five *de novo* assembly-based variant callers (**Table 4.1**) on 56 (7 divergence levels and 8 coverage depths) simulated error-free datasets. Of those, SGA-haplotype, fermi and fermi2 first assemble reads into contigs or unitigs, map them to the reference genome, and call variants from the alignments. Cortex and SGA graph-diff, on the other hand, call variant directly from the graph. We plotted sensitivity (**Figures 4.4A-H**; **Figures 4.5A-E**), precision rate (**Figures 4.5F-J**) and genotype concordance (**Figures 4.5K-O**) over different coverage depths and divergence levels.

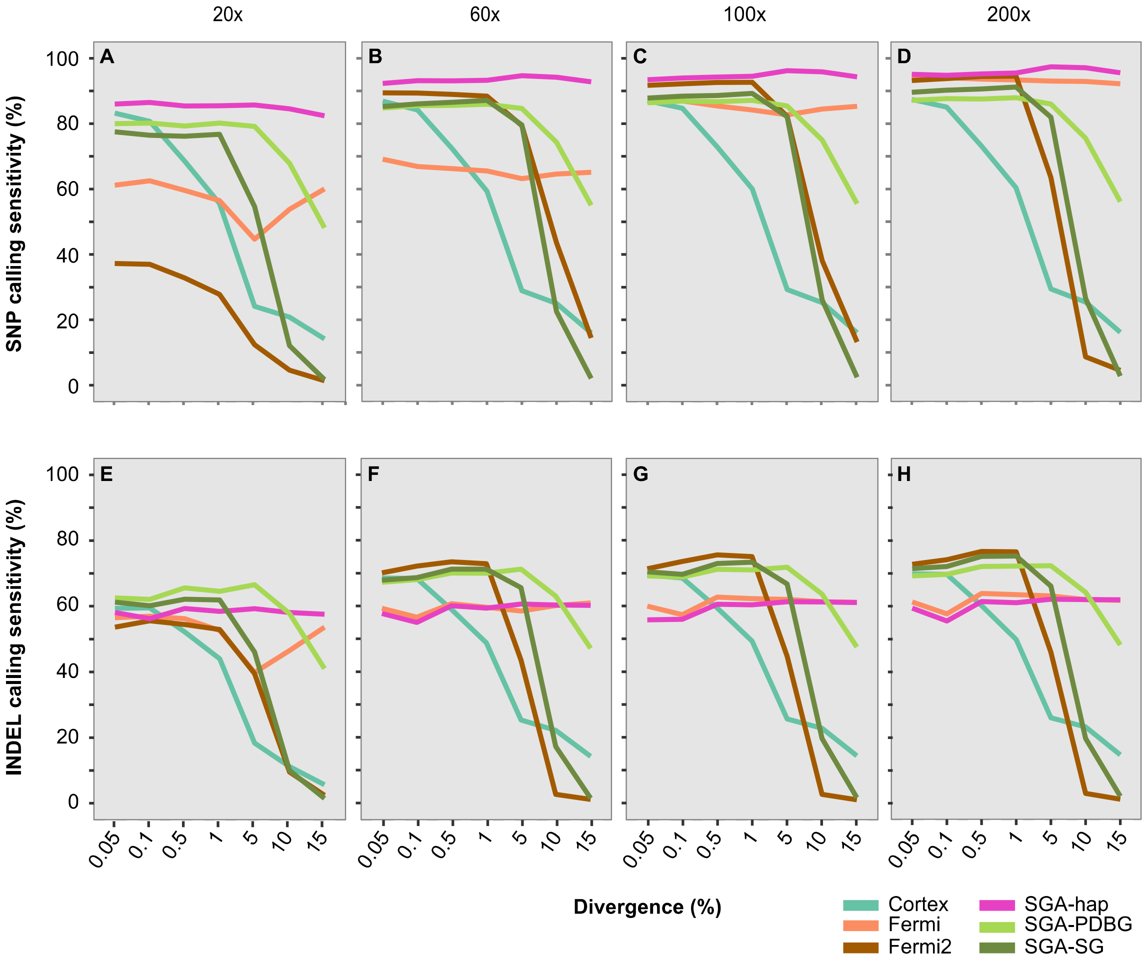


Figure 4.4 Sensitivity in SNP and INDEL detection with six assembly-based variant callers. (**A-D)** SNP sensitivity at four coverage depths. (**E-H)** INDEL sensitivity at four coverage depths. Reads were simulated without introducing error. In estimating the number of matches in INDELs, we extended the coordinates of the simulated INDELs by +/- 10 bp before intersection with those of the called INDELs. SGA, string graph assembler; SGA-hap, SGA-haplotype; SGA-PDBG, SGA paired de Bruijn graph; SGA-SG, SGA string graph.

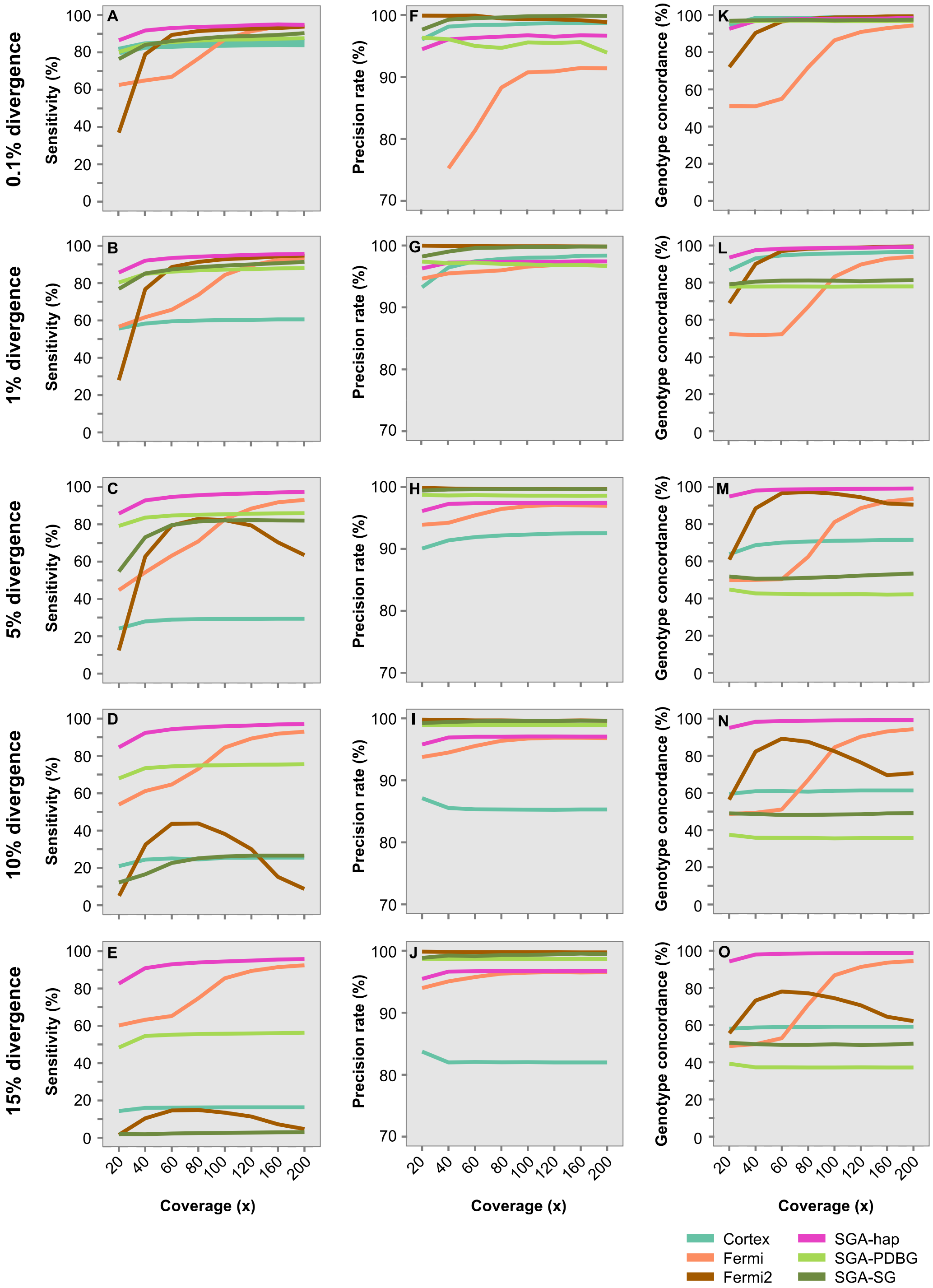


Figure 4.5 Quality metrics for SNP detection by six assembly-based variant callers. (**A-E)** SNP sensitivity. (**F-J)** SNP precision rate. (**K-O)** SNP genotype concordance. Reads were simulated without introducing error.

Overall, Cortex had the lowest SNP (**Figures 4.4A-D**) and INDEL (**Figures 4.4E-H**) calling sensitivity in a majority of the cases. It also had low precision rates (**Figures 4.5F-J**) and genotype concordance (**Figures 4.5K-O**) at high divergence**.** On the other hand, fermi and fermi2 required high coverage, especially in SNP calling (**Figures 4.4A-H**). Also, unlike SGA-haplotype and fermi that performed similarly across different divergence levels, fermi2, SGA-PDBG and SGA-SG showed a marked reduction of SNP and INDEL sensitivity at high divergence (**Figures 4.4A-H**). Remarkably, SGA-haplotype achieved the highest SNP calling sensitivity (median 94.3%, **Figures 4.5A-E**) and genotype concordance (median 98.6%, **Figures 4.5K-O**) across all coverage depths and divergence levels, although it had a slightly lower precision rate (median 96.8%, **Figures 4.5F-J**).

INDELs are more difficult to detect than SNPs. It is challenging to define the exact boundary of INDELs, especially for those located within microsatellites [53]. Considering this limitation, we counted the called INDELs as true positives if they were within ±10 bp of simulated INDELs. At <=5% divergence, SGA-haplotype and fermi are generally less sensitive than fermi2 and the two SGA graph-diff modes (**Figures 4.4E-H**). However, both determined the exact boundary for a larger proportion (54.7-66.6%) of the identified INDELs, compared to 31.7-52% by fermi2 and graph-diff (**Table 4.4**).

Error correction is a critical component in assembly-based variant calling. Based on the outcome from simulated error-free reads, we further assessed SGA-haplotype, SGA-PDBG and fermi2 on reads simulated with error. We performed a simple processing by arbitrarily trimming 10 bases from the 3' end of the reads that were simulated at a higher error rate than the 5' bases. The relative performance of these three methods is similar between error-free reads (**Figures 4.4A-H**) and reads simulated with error (**Figures 4.6A-H**) in terms of sensitivity. In SNP calling, overall SGA-hap had the highest sensitivity across all the divergent levels (**Figures 4.6A-D**). In INDEL calling, SGA-hap had the lowest sensitivity at low divergence but the highest sensitivity at high divergence (**Figures 4.6E-H**).

Table 4.4 Percentage of INDELs with accurately defined boundary

Methods	Data Type	Divergence (%)						
		0.05	0.10	0.50	1	5	10	15
Cortex	No error	50 (0.4)	49.1 (0.2)	50.2 (0.1)	na	na	na	na
Fermi	No error	63.6 (1.2)	65.9 (0.8)	61.9 (0.8)	59.1 (1.1)	54.7 (2.4)	49.4 (1.8)	44.4 (0.8)
Fermi2	No error	51.9 (0.5)	49.7 (0.3)	50.3 (0.1)	48.6 (0.1)	46 (0.1)	41.6 (0.7)	43.1 (1)
SGA-hap	No error	64.4 (0.5)	66.6 (0.8)	63 (0.7)	61.3 (0.4)	57.4 (0.3)	51 (0.2)	45.4 (0.1)
SGA-PDBG	No error	38 (0.4)	33.4 (0.4)	33.7 (0.1)	34.4 (0.1)	31.7 (0)	29.6 (0)	27.3 (0)
SGA-SG	No error	39 (0.3)	34.3 (0.2)	34.2 (0.1)	34.8 (0.1)	32.4 (0.1)	31.7 (0.7)	31.3 (1.7)
Fermi2	W/ error	48.1 (1.1)	48.3 (0.7)	49.8 (0.1)	49.2 (0.2)	46.7 (0.4)	44.9 (1.6)	43 (1.1)
SGA-hap	W/ error	63.2 (0.6)	62.8 (1.4)	63.4 (1.5)	61.9 (1.2)	56.8 (1.4)	50.9 (1.2)	45.3 (0.9)
SGA-PDBG	W/ error	33.2 (0.2)	33.6 (0.4)	34.6 (0.2)	34.5 (0)	32.1 (0.1)	29.5 (0.1)	27.5 (0.1)

Shown are the average and stand deviation (in the parenthesis) of the ratio between INDELs with the exact boundary and those within 10 bp of simulated INDELs. Simulated data with the coverage between 40x and 200x were used. SGA-SG, SGA graph-diff string graph mode; SGA-PDBG, SGA graph-diff paired de Bruijn mode.

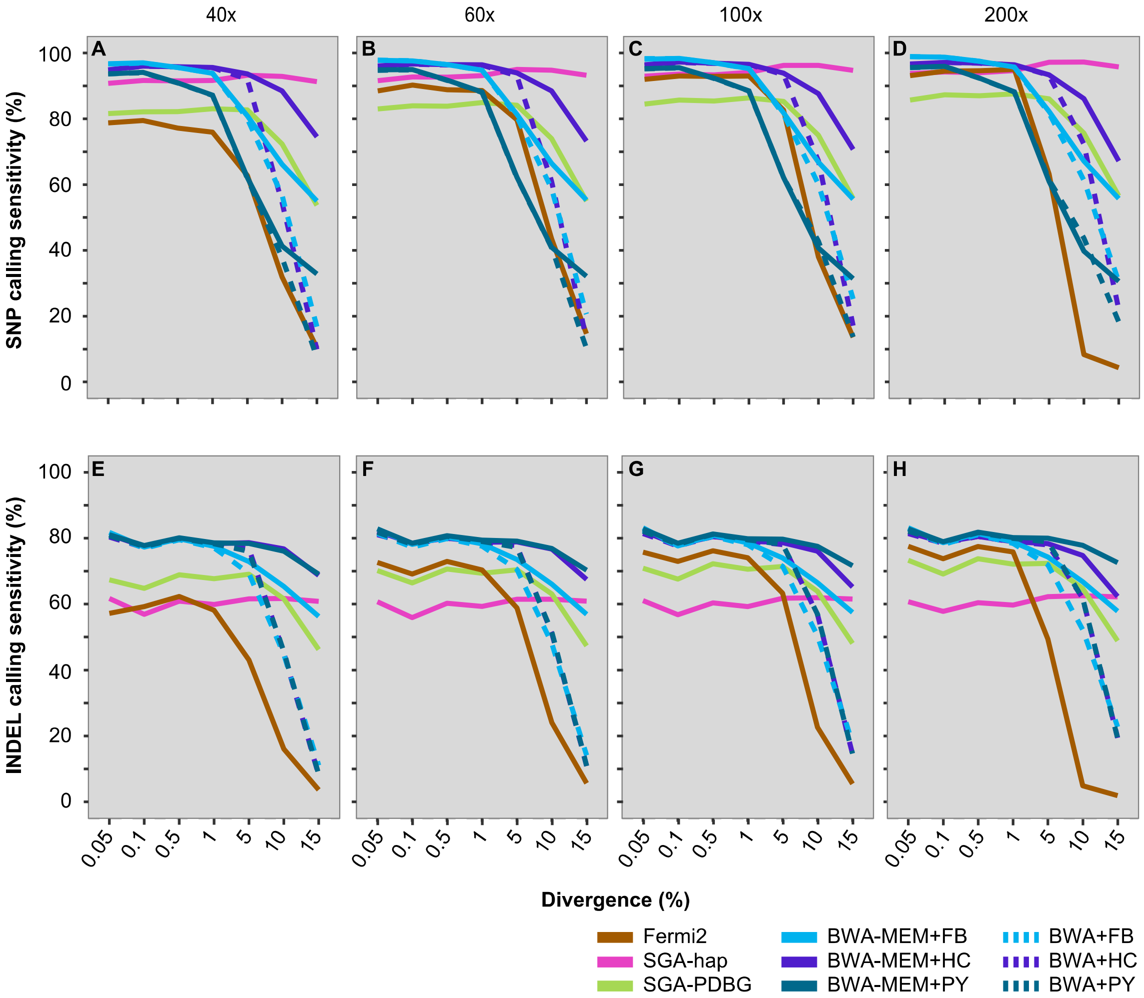


Figure 4.6 Sensitivity of SNP and INDEL detection with mapping-based approaches and *de novo* assembly-based variant callers. (**A-D)** SNP sensitivity at four coverage depths. (**E-H)** INDEL sensitivity at four coverage depths. Reads were simulated with error. For the three *de novo* assembly-based variant callers, 10 bases from the 3' end were trimmed off. In estimating the number of matches in INDELs, we extended the coordinates of the simulated INDELs by +/- 10 bp before intersection with those of the called INDELs. SGA-hap, SGA-haplotype; SGA-PDBG, SGA paired de Bruijn graph; FB, FreeBayes; HC, GATK HaplotypeCaller; PY, Platypus.

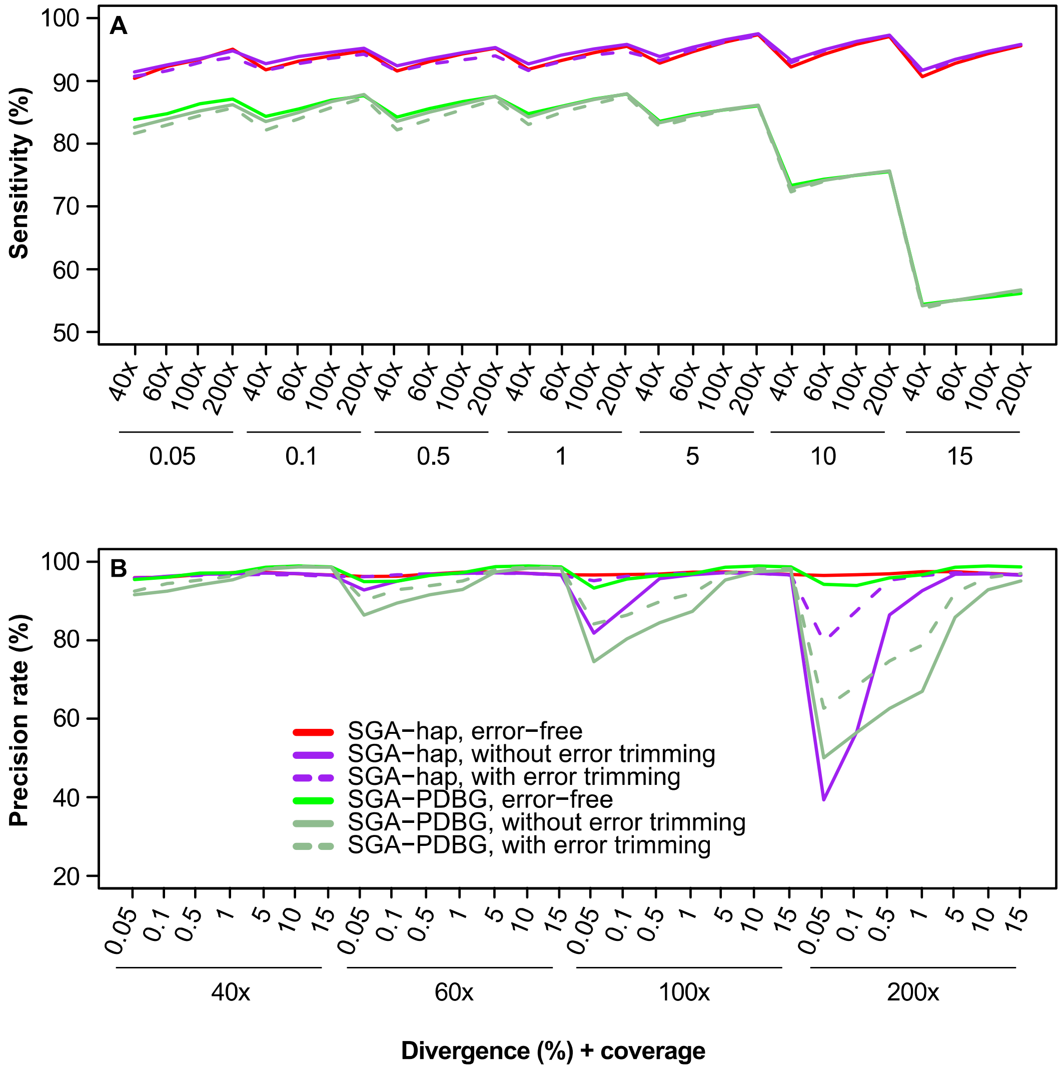


Figure 4.7 The effect of sequencing errors on SNP calling. (**A**) SNP calling sensitivity. (**B**) SNP calling precision rate. SGA-haplotype and SGA-PDBG were used to call SNPs from simulated reads without error, simulated reads with error but no trimming and simulated reads with error trimming. For error trimming, the last 10 bases were arbitrarily trimmed off. SGA-hap, SGA-haplotype; SGA-PDBG, SGA paired de Bruijn graph.

For both SGA-PDBG and SGA-hap, the SNP sensitivity was highly comparable among error-free reads, error-containing reads with and without 3’ end trimming (**Figure 4.7A**). However, sequencing errors reduced the precision rate in SNP calling, more obvious for SGA-PDBG at high coverage (**Figure**

4.7B). Without 3’ end trimming of reads with error, SGA-PDBG and SGA-haplotype lost 23.9% and 16.2% in the precision rate at 200x on average, versus 15.3% and 4.1% when 3’ end trimming was applied (**Figure 4.7B**). Thus, the 3’ end trimming is more effective for SGA-hap than for SGA-PDBG. This is consistent with the finding that de Bruijn graph is more sensitive to sequencing errors than string graph [116]. In de Bruijn graph-based assembly, errors lead to novel k-mers that do not exist in the genome, which complicate the building of k-mer graph. Also, current *de novo* assemblers could not properly handle ultra-deep coverage data, as suggested by a bigger loss of precision rate at 200x (compared to 40x to 100x) for both SGA-hap and SGA-PDBG ((**Figure 4.7B**). The assembly quality would decrease once the sequencing depth goes beyond a certain level [117-119], likely due to the accumulation of uncorrected sequencing errors.

In parallel, we also tested three haplotype (FreeBayes) or local *de novo* assembly (GATK HaplotypeCaller and Platypus) based callers on simulated error-containing reads **(Figure 4.2**). SGA-haplotype variant calling had a lower SNP sensitivity (92.9%) on average compared to FreeBayes and GATK HaplotypeCaller (96.4-96.5%) at <=1% divergence but showed superior performance (94.8% versus 44.8-84.3%) at >=5% divergence (**Figures 4.6A-D**).

4.4.2. Variant calling in real whole genome data

We have investigated six *de novo* assembly-based variant calling methods on simulated WES reads. However, real sequencing reads are far more complex. In particular, they contain platform- and sequence-specific bias (like GC bias) and error. Such bias may lead to low or no coverage in some regions [120]. In exome data, the actual coverage depends on the whole exome enrichment platforms [121]. In addition, though error rate is estimated to be less than 0.1% over at least 75-85% of the bases for illumina platforms (http://www.molecularecologist.com/next-gen-table-3c-2014/), *de novo* assembly is extremely sensitive to sequencing error. Without proper correction, sequencing error could dramatically reduce the assembly quality [114, 122]. Sequencing error, divergence level, coverage depth and read length are key factors in *de novo* assembly-based variant calling. It is critical to fully assess a method over these key factors.

The most comprehensive call set for NA12878 was generated from 250-bp paired-end reads of a PCR-free genomic library [54], which represents majority of the reference variants used in this study. This call set was identified using two assemblers (Cortex and DISCOVAR) and the state-of-the-art combination between BWA-MEM and GATK HaplotypeCaller. The same type of WGS data is available for a CEU trio (NA12878, NA12891 and NA12892) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data), which allows us to better assess SGA-hap and SGA-PDBG on WGS as well as the quality of the reference call set before applied to WES data.

We first assessed SGA-haplotype and SGA-PDBG on the 250-bp whole genome reads from NA12878. SGA-hap had a much higher sensitivity than SGA-PDBG in the HLA region, particularly in the six highly divergent genes (*HLA-A*, -*B*, -*C* -*DQA1*, -*DQB1*, and -*DRB1*) (**Table 4.5**). For example, in identifying known INDELs, SGA-PDBG had only ~10% sensitivity, compared to nearly 70% by SGA-hap. DISCOVAR is a de Bruijn-based assembler specifically developed for variant detection from the 250-bp paired-end reads [54]. To better understand how SGA-hap performs relative to DISCOVAR, we intersected the list of known variants identified by SGA-hap with the two public lists from DISCOVAR and BWA-MEM+GATK HaplotypeCaller [54]. In the HLA region, Nearly 40% (10266/26149) of the known variants identified by SGA-hap were not in the DISCOVAR list, compared to only 7.7% (2005/26141) not shared by GATK HaplotypeCaller (**Table 4.6**). In the six highly divergent HLA genes, up to 77% (2281/2964) of the SGA-hap calls were not shared by DISCOVAR, versus 13% (389/2958) by GATK HaplotypeCaller (**Table 4.6**). The results suggested that

DISCOVAR was much less sensitive than SGA-hap in the highly divergent HLA region.

Table 4.5 Percentage of variant calling sensitivity from NA12878 WGS

Type	SGA-hap		SGA-PDBG	
	HLA	6 genes	HLA	6 genes
Known SNP	94.4	97.2	88.5	85.8
Novel SNP	61.8	79.3	47.9	54.1
Known INDEL	69.7	67.6	8.9	14.5
Novel INDEL	39	25.7	8.6	10.8

Contigs were from 250-bp WGS and mapped to the hg19 reference using BWA-MEM. 6genes: *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1* and *-DRB1*.

Table 4.6 Shared and unique variants from NA12878 WGS and WES

Data	Type	Region	SGA-hap vs. DISCOVAR			SGA-hap vs. GATK HC		
			Shared	SGA-hap only	DISCOVAR only	Shared	SGA-hap only	GATK HC only
WGS	SNP	HLA	14,985	9,701	423	22,830	1,857	1,178
WGS	SNP	6 genes	660	2,169	7	2,458	369	64
WGS	INDEL	HLA	898	565	285	1,306	148	545
WGS	INDEL	6 genes	23	112	8	111	20	56
WES	SNP	HLA	614	346	28	860	98	53
WES	SNP	6 genes	159	248	6	361	45	26
WES	SNP	Non-HLA	1,320	105	73	1,358	67	66
WES	INDEL	HLA	15	11	5	21	5	11
WES	INDEL	6 genes	2	9	2	8	3	8
WES	INDEL	Non-HLA	79	27	32	80	26	28

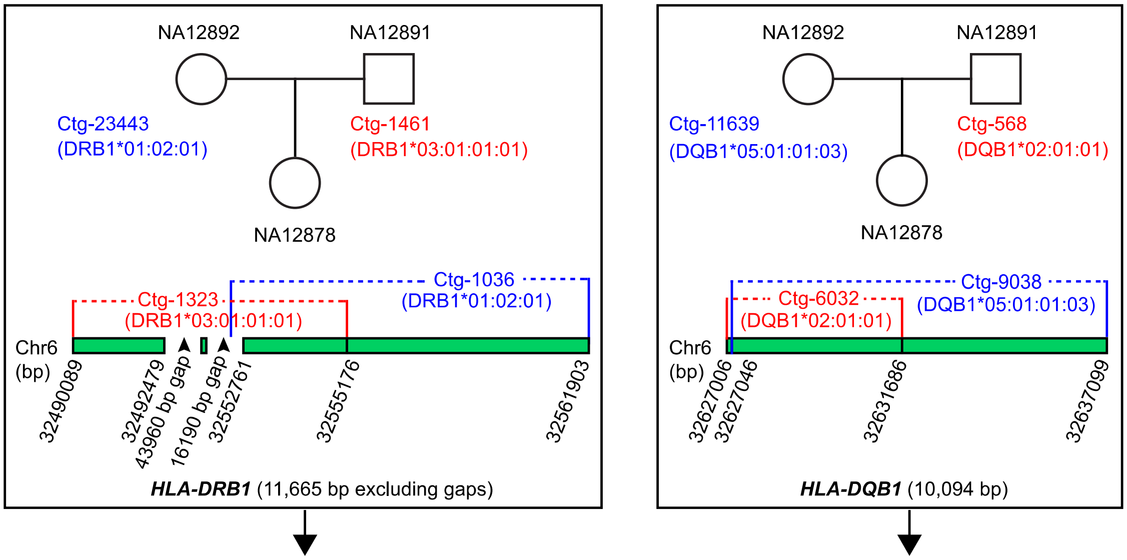
Known variants detected by SGA-hap were compared to two public call sets identified by DISCOVAR and BWA-MEM with GATK HaplotypeCaller (GATK HC) [54]. The 6 HLA genes (6 genes) are *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1* and *-DRB1*. For WES, the average number of variants from the two 150-bp datasets (**Table 4.1**) is shown. Known variants, variants in dbSNP v138.

We next assessed SGA-hap to see whether it can differentiate different alleles through long range phasing, using the 250-bp WGS data from the CEU trio. We focused on *HLA-DQB1* and *HLA-DRB1*, two of the most highly divergent genes. We first used BLAT to identify two longest contigs from NA12878 (daughter) that best matched each of the two genes in the IMGT/HLA database (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/). The four contigs (4,794-10,094 bp) from NA12878 were used to pull out the best hits from contigs assembled in NA12891 (father) and NA12892 (mother) via BLAT. For both genes, SGA-hap successfully separated the two alleles in NA12878 (**Table 4.7**; **Figure 4.8**). For *HLA-DRB1*, allele DRB1*03:01:01:01 (represented by contig Ctg-1323) is from NA12891 (Ctg-1461); both contigs showed 93% similarity with the hg19 reference but perfectly matched HLA haplotype 6_qbl_hap6. Allele DRB1*01:02:01 (Ctg-1036) is from NA12892 (Ctg-23443), with the first 7,439 bp from Ctg-1036 fully matching Ctg-23443. For *HLA-DQB1*, allele DQB1*02:01:01 (belongs to 6_qbl_hap6) is from NA12891. The second allele DQB1*05:01:01:03 (Ctg-9038) is from NA12892 (Ctg-11639), and contig Ctg-11639 covers the 10042-bp contig Ctg-9038. These results strongly suggest that SGA-hap can achieve long-range phasing across the highly divergent regions. A closer examination of an 11,665-bp region within *HLA-DRB1* of NA12878 revealed that DISCOVAR missed 69.4% (173) of the 249 phased SNPs covered by alleles DRB1*03:01:01:01 and DRB1*01:02:01 and BWA-MEM+GATK HaplotypeCaller missed 16.9% (42) of them (**Figure 4.8**).

Table 4.7 SGA-hap assembly of *HLA-DQB1* and *HLA-DRB*1 from WGS in a CEU trio

Sample	Gene	Contig	Length (bp)	Best hit in IMGT/HLA			Best hit in HLA haplotypes		
				Allele	Match (bp)	Mismatch (bp)	Best hit	Span (bp)	Sim (%)
NA12892	*HLA-DQB1*	Ctg-11639	14010	DQB1*05:01:01:03	7090	0	6	14020	97.4
NA12891	*HLA-DQB1*	Ctg-4703	5373	DQB1*06:02:01:01	4556	0	6	5373	100.0
NA12878	*HLA-DQB1*	Ctg-9038	10042	DQB1*05:01:01:03	7090	0	6	10054	96.5
NA12891	*HLA-DQB1*	Ctg-568	4724	DQB1*02:01:01	3963	0	6_qbl_hap6	4724	100.0
NA12878	*HLA-DQB1*	Ctg-6032	4794	DQB1*02:01:01	3985	0	6_qbl_hap6	4794	100.0
NA12892	*HLA-DRB1*	Ctg-23443	11257	DRB1*01:02:01	3654	1	6	13910	99.2
NA12891	*HLA-DRB1*	Ctg-14574	10605	DRB1*15:01:01:02	6292	0	6	10605	100.0
NA12878	*HLA-DRB1*	Ctg-1036	10094	DRB1*01:02:01	6305	5	6	10087	97.9
NA12892	*HLA-DRB1*	Ctg-23796	9266	RB1*01:02:01	7764	6	6	9615	96.6
NA12891	*HLA-DRB1*	Ctg-1461	5305	DRB1*03:01:01:01	5305	0	6_qbl_hap6	5305	100.0
NA12878	*HLA-DRB1*	Ctg-1323	5004	DRB1*03:01:01:01	5004	0	6_qbl_hap6	5004	100.0

Contigs were generated at 31-mer. BLAT was used to search contig sequences against the IMGT/HLA database and the eight different HLA haplotypes. Note that the first 7,439 bp from Ctg-1036 perfectly match Ctg-23443 in *HLA-DRB1* and Ctg-9038 perfectly matches a 10042-bp region from ctg-11639 in *HLA-DQB1*. IMGT, international ImMunoGeneTics Database; NA12878, daughter; NA12892, mother; NA12891, father; Sim, similarity.



Type	Phased	Missed SNP	
	SNP	DISCOVAR	GATK HC
Known	237	162 (68.4%)	39 (16.5%)
Novel	12	11 (91.7%)	3 (25%)
Total	249	173 (69.4%)	42 (16.9%)

Type	Phased	Missed SNP	
	SNP	DISCOVAR	GATK HC
Known	462	341 (73.8%)	21 (4.5%)
Novel	122	122 (100%)	31 (25.4%)
Total	584	463 (79.3%)	52 (8.9%)

Figure 4.8 Phasing of variants in *HLA-DRB1* and *HAL-DQB1* by SGA-haplotype. Contigs were assembled from 250-bp PCR-free WGS in a CEU trio and compared to the IMGT/HLA database and to hg19 reference sequence via BLAT (see **Table 4.7**), by which the correspondence was established for alleles in these two genes from NA12878 with those in NA12892 and NA12891. Phased variants from NA12878 in the indicated regions were intersected with two public lists identified by DISCOVAR and GATK HaplotypeCaller (GATK HC), also from 250-bp PCR-free WGS, to identify shared and method-specific calls.

Using the CEU trio WGS data, we found that some contigs from highly divergent regions and those from misassembly can only be mapped partially to the reference by BWA-MEM, which causes a stretch of hard- or soft-clipped bases at the contig ends (represented by “H” or “S” in the CIGAR string of the BAM file). Soft-clipped bases do not align to the reference but are kept in the output BAM file, while hard-clipped bases are excluded from the BAM file. We argued that at least some of the clipped bases represent extremely divergent bases, large INDELs or other structural variants. Based on extensive manual inspection of alignments between highly divergent contigs assembled from NA12878 and the human genome reference as well as the eight HLA haplotype sequences available in the UCSC (University of California, Santa Cruz) genome browser, we tested two-tier contig mapping strategy to improve alignment accuracy. All contigs were first mapped by BWA-MEM and those showing soft-clipping were re-mapped by BLAT. The known INDELs identified by BWA-MEM and BWA-MEM+BLAT were mostly <20 bp in size (**Figures 4.9A-B**). However, some large (>90 bp) novel INDELs missed by BWA-MEM were identified by BLAT (**Figures 4.9C-D**).

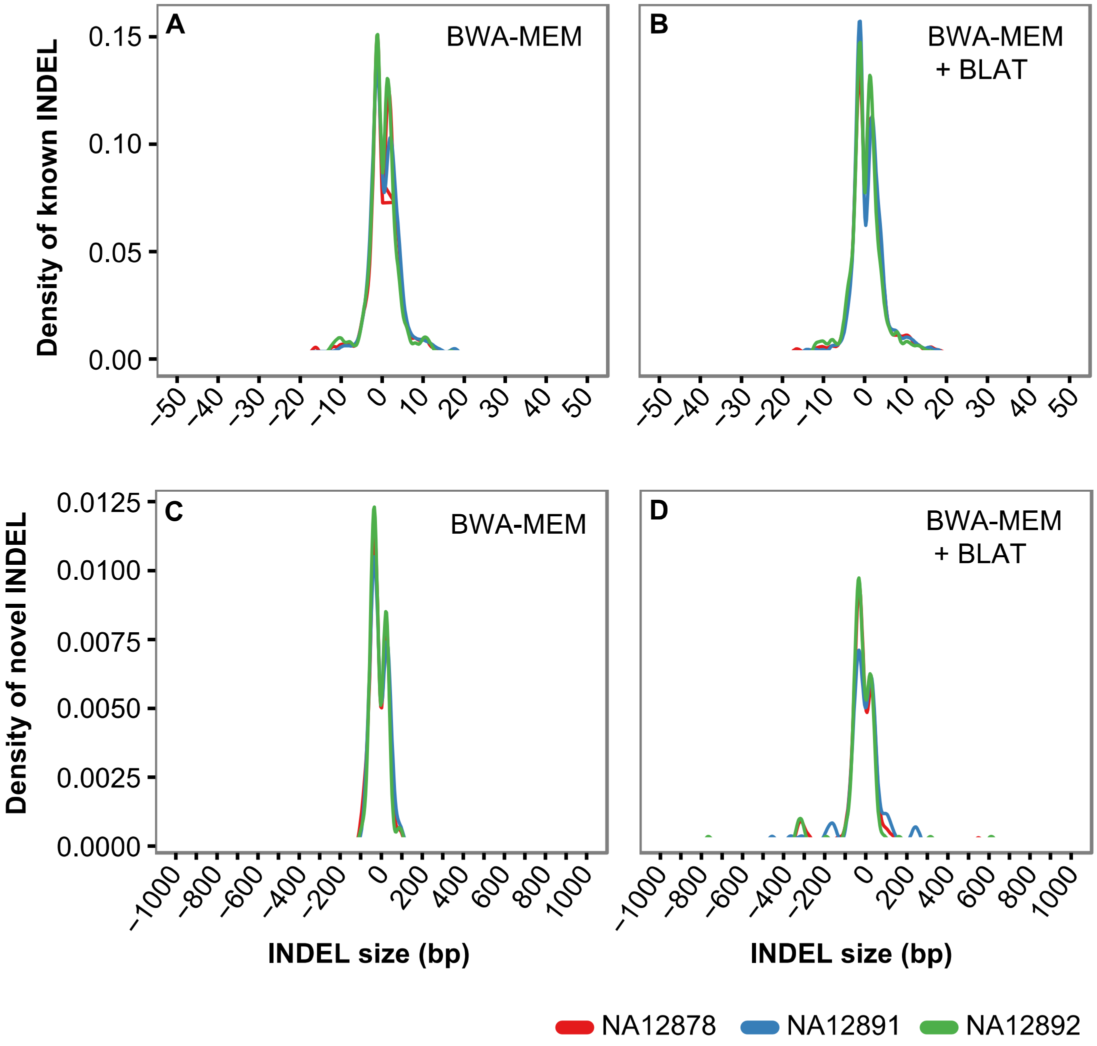


Figure 4.9 Size distribution of contigs assembled from 250-bp WGS data in a CEU trio. **(A-B)** known INDELs. **(C-D)** Novel INDELs. Contigs were assembled by SGA-hap and mapped to the hg19 reference by BWA-MEM or BWA-MEM+BLAT. Negative and positive values on the x-axis indicate size of deletion and insertion, respectively.

4.4.3. Variant calling in real whole exome data

To develop a robust workflow with high efficiency across different capture platforms, read lengths and genomic regions, we finally assessed SGA-haplotype, SGA-PDBG, together with the three mapping-based pipelines on eight NA12878 WES datasets. These datasets were generated from Roche (100 bp) and Illumina (76 and 150 bp) exome capture platforms (**Table 4.2**). As found in the whole genome sequencing data, analysis of the 150-bp NA12878 exome data further indicated that DISCOVAR missed a significant portion of known variants in the HLA region (**Table 4.6**). In contrast, in the non-HLA region, over 90% of the known variants were shared between DISCOVAR (93%) and SGA-hap (91.4%). Unlike SGA-hap that performs similarly well across a wide range of divergence (**Figures 4.4A-H**; **Figures 4.6A-H**; see below for HLA region in real exome data), DISCOVAR is much less sensitive in the HLA region.

SGA-hap generated extremely long contigs from all NA12878 samples (**Table 4.8**). By blasting the longest contig from each of the samples against the National Center for Biotechnology Information (NCBI) non-redundant database, we found that they perfectly aligned to the virus sequences, such as human papillomavirus and herpesvirus. For example, the longest contig (42,595 bp) from sample FC1_NA12878_S7_S10 (**Table 4.2**) represents 25% of human herpesvirus 4 DNA. This finding highlights the potential application of SGA-haplotype in detecting viral integration sites in the cancer genomes, an active research area in The Cancer Genome Atlas (TCGA) project.

Table 4.8 Summary of Chr6 contigs assembled NA12878 WES by SGA-hap

Capture platforms	Read length (bp)	No. contigs	Total size (Mb)	Contig size (bp)			
				Mean	Median	75% Quantile	Maximum
Illumina (38 Mb)	76	32,744	7.72	238	148	297	26,800
Roche (64 Mb)	100	24,958	8.87	355	228	487	21,740
Illumina (38 Mb)	150	37,995	12.99	342	267	383	42,595

Illumina Nextera kit contains ~3.8 Mb capture regions and Roche SeqCap EZ v3 kit contains ~6.4 Mb capture regions on Chr6. “Total size” refers to the total non-overlapping bases from all retained contigs.

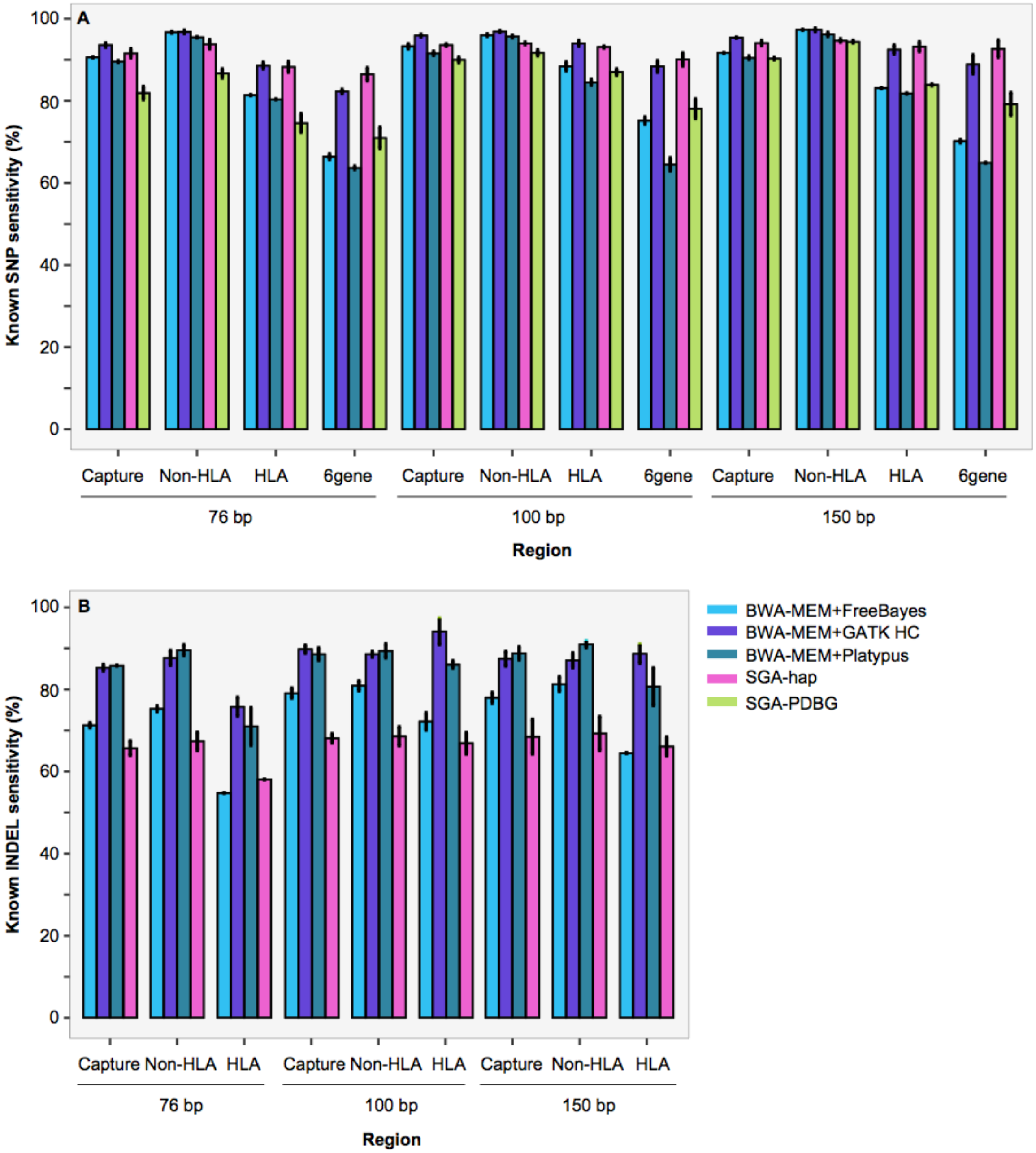


Figure 4.10 Sensitivity of detecting known variants in NA12878 WES. (**A**) Known SNP sensitivity. (**B**) Known INDEL sensitivity. The mean and standard deviation were plotted. Chr6 variants were identified by SGA-hap, SGA-PDBG and three mapping-based approaches and separated into known (in dbSNP v138) and novel. In mapping-based approaches, reads were mapped to hg19 reference using BWA-MEM. See **Table 4.2** for more information about the exome datasets. In calculating INDEL sensitivity, exact match in genomic coordinates is required. The six genes (“6gene”) are *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1*.

SGA-haplotype was more sensitive than SGA-PDBG ((**Figures 4.10A-B**). In the capture region, SGA-PDBG had only 1.3-6% sensitivity in INDEL detection, compared to 65.7-68.5% by SGA-hap. In identifying known SNPs, SGA-haplotype was slightly less sensitive (1.5-3% lower) than the three haplotype-based callers, FreeBayes, GATK HaplotypeCaller and Platypus, in the non-HLA regions (**Figure 4.10A**). However, SGA-haplotype generally had higher sensitivities in the HLA region, more obvious in the six highly divergent HLA genes, where SGA-haplotype was 14.9-27.8% more sensitive than FreeBayes and Platypus. On the other hand, SGA-haplotype was about 5% more sensitive on the 100-bp and 150-bp datasets, compared to the 76-bp datasets in the HLA region (**Figure 4.10A**). The trend was much less obvious in the non-HLA regions, indicating that longer reads favor SGA-haplotype in highly divergent regions.

Compared to the public call set, SGA-haplotype had 63 unique known SNPs but missed 75 in the non-HLA region from the 150-bp sample FC1_NA12878_S1_S4, versus 80 and 56 in the HLA region. Next we sought to determine the precision rate for SGA-haplotype by examining its unique known SNPs missed in the public call set. Since our focus is on the HLA region, below we first compared the above method-specific known SNPs with those identified from the 250-bp WGS data in the HLA region. We then manually checked the 80 known SNPs unique to SGA-haplotype and the 56 only present in the public call set using the Integrative Genomics Viewer [123].

Of the 80 known SNPs unique to SGA-haplotype, 41 were also identified by SGA-haplotype from both NA12878 and one the parents in the WGS data, supporting the fidelity of these calls. They were from long haplotypes of 2.3-10 kb assembled in four of the highly divergent genes, including *HLA-DQA1* (9), *HLA-DQB1* (4), *HLA-DRB1* (15) and *HLA-DRB5* (13). The other 39 SNPs were missed in NA12878 WGS, including 7 that were fully supported by the alignments in WES. We thus believed that 60% (48/80) of the SGA-haplotype specific known SNPs represent true variants. The remaining 32 SNPs are likely low-confidence calls or false positives. Nine of them were supported by only 2 reads (1), flanked by INDELs in low complexity regions (2), or called from one erroneously mapped contig (6). The other 23 (32-9) were supported by reads with low base quality at the polymorphic sites, with 16 in gene *MUC21*. These low quality bases were located in the middle of the reads and had >10x coverage, remaining unchanged after quality-based 3' end trimming and error correction. Apparently, the mapping-based approaches would not identify these SNPs as they only consider bases with Q>=17. These calls may represent sequence-specific errors in Illumina reads [124] or reflect non-specific capture.

We further checked the 56 known SNPs in the HLA region that are present in the public call set but missed by SGA-haplotype. Twenty-nine likely represent false positives in the public call set, as SGA-haplotype also missed them in NA12878 WGS. The other 27 may be false negatives from SGA-haplotype. Of the 27, 12 were identified with k-mer=31, which we didn't include in the final SGA-haplotype call list that was generated using 51-mer to 91-mer. Nine SNPs were in regions with no or only 1-6x coverage, likely reflecting low capture affinity. The other seven were in regions with reasonable (>=4x) coverage for the alternative alleles.

We previously demonstrated that GATK HaplotypeCaller and Platypus are ideal callers for INDEL detection (see **section 3.4.7** for details). We found that SGA-haplotype was less sensitive (13-27.2% lower) than these two methods in both HLA and non-HLA regions. SGA-haplotype was also less sensitive than FreeBayes in the non-HLA regions (**Figure 4.10B**). SGA-haplotype had 11 known (**Figure 4.11A**) and 16 novel (**Figure 4.11B**) INDELs that were overlapped by GATK HaplotypeCaller only, and 4 known and 6 novel INDELs that were overlapped by Platypus only. It had 23 known and 57 novel INDELs missed by both. Here we limited the analysis to the exome capture regions, which likely underestimates the sensitivity of SGA-haplotype, as the starting positions for many of the INDELs detected by SGA-haplotype are outside of the capture regions. By calling INDELs from the entire chromosome 6 rather than from the capture regions only, SGA-hap identified 10 large INDELs, ranging between 98 and 801 bp (mean size 238 bp), that were missed by both mapping-based callers, supporting the notion that *de novo* assembly-based approach should be able to identify larger (>=30 bp) INDELs [125].

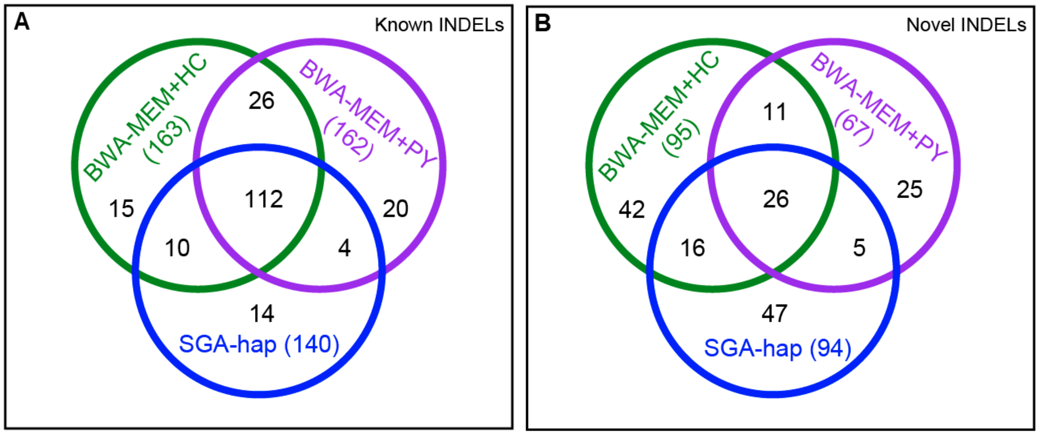


Figure 4.11 Overlap of INDELs called by two mapping-based methods and SGA-hap in NA12878. (**A**) Known INDELs in the capture regions. (**B**) Novel INDELs in the capture regions. The 150-bp WES data “FC1_NA12878_S1_S4” were used (**Table 4.2**). INDELs were identified from Chr6 and split into known and novel by intersecting with dbSNP v138. HC, GATK HaplotypeCaller; PY, Platypus.

For INDEL calling in the capture regions from the 150-bp sample FC1_NA12878_S1_S4, SGA-haplotype had 32 unique known and 72 unique novel calls missed in the NA12878 reference call set. We manually checked the alignments around each of the 32 SGA-haplotype specific known INDELs. Twenty-two were located in low complexity regions but were each supported by at least 20 reads and another 4 by 14-19 reads. Our analysis further supports the power of *de novo* assembly in the detection of INDELs [125], especially those in complex genomic regions. The other six INDELs were supported by only 3-9 reads. Still, SGA-haplotype successfully assembled contigs (190-260 bp) in these regions.

4.4.4. SGA-haplotype is less impacted by the k-mer selection

K-mer is a key factor for *de novo* assembly. In de Bruijn-based assembly, k-mer represents the error correction and overlap parameter. In string graph-based assembly, k-mer is only used for error correction. Therefore, the selection of k-mer may impact the two approaches differently in variant discovery. To confirm this hypothesis, we tested different k-mers in SGA-PDBG and SGA-hap assembly, using 250-bp WGS data and 150-bp WES data (FC1_NA12878_S1_S4) from NA12878 (**Figures 4.12A-D**).

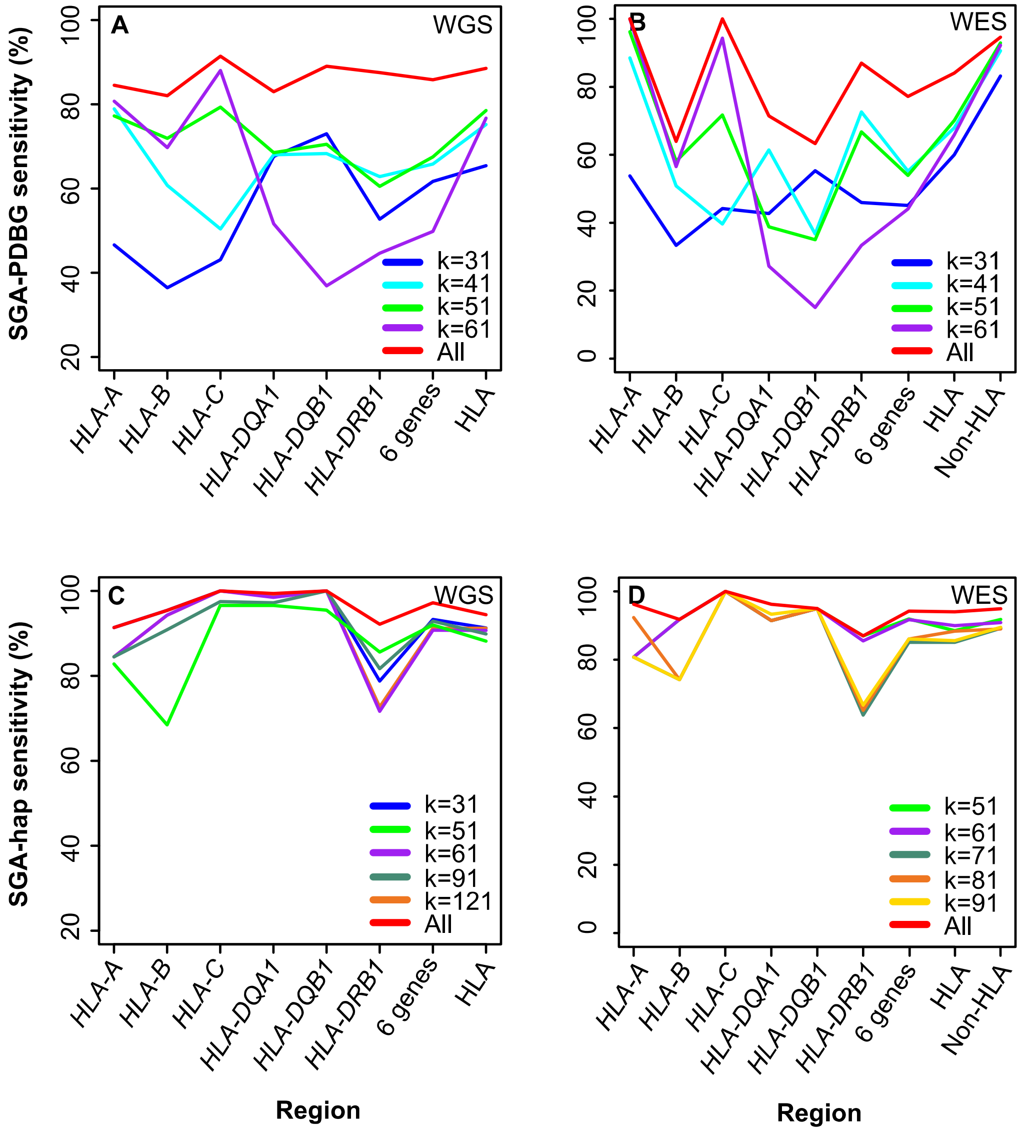


Figure 4.12 Known SNP sensitivity by SGA-hap and SGA-PDBG at different k-mers. Both 250-bp WGS and 150-bp WES data (FC1_NA12878_S1_S4, **Table 4.2**) from NA12878 were used. (**A**) SGA-PDBG on WGS. (**B**) SGA-PDBG on WES. (**C**) SGA-hap on WGS. (**D**) SGA-hap on WES. 6genes: *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1* and *-DRB1*; All: the union of known SNPs identified from all k-mers.

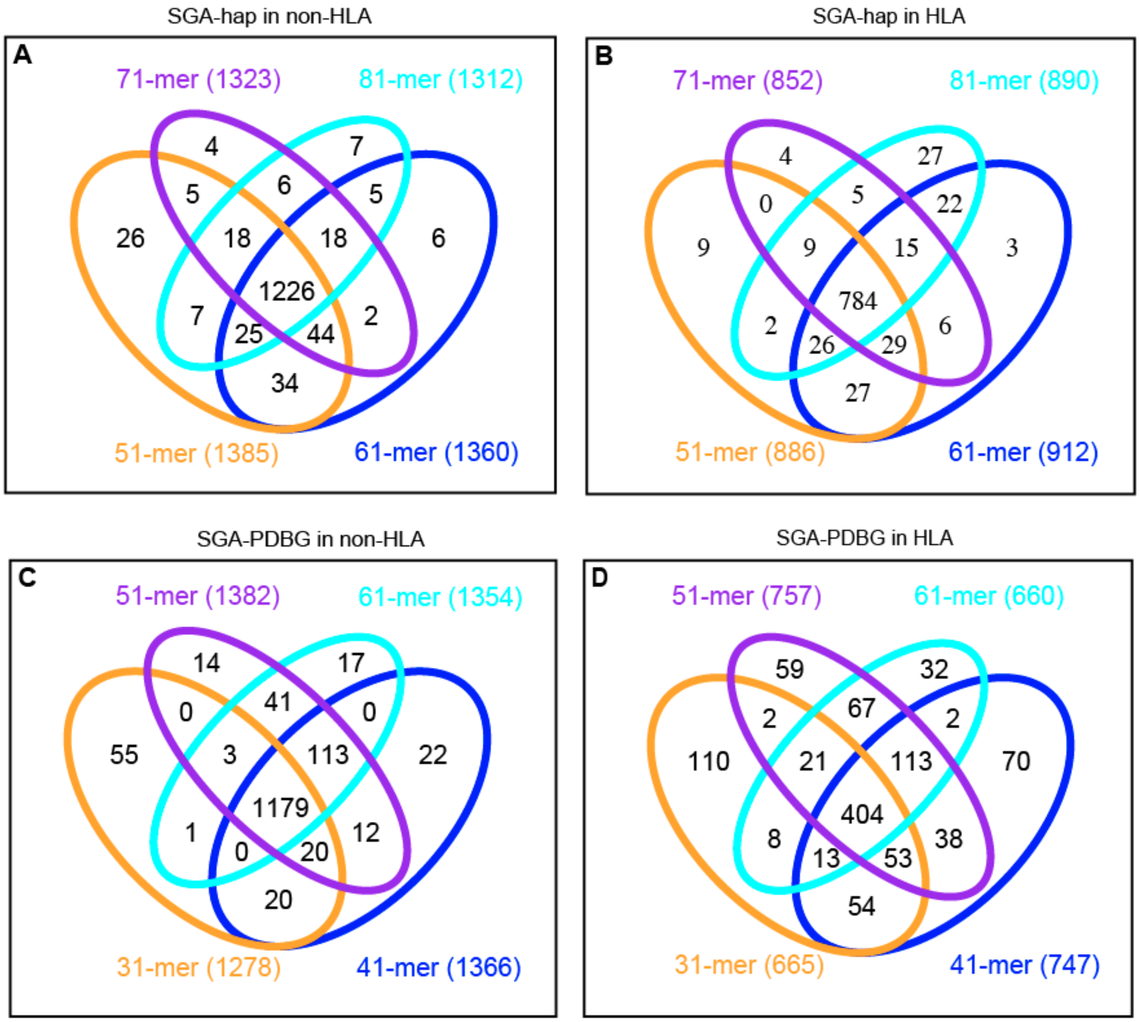


Figure 4.13 Overlap of known SNPs detected at four different k-mers. SGA-hap and SGA-PDBG were used to identify known SNPs in the 150-bp exome data "FC1_NA12878_S1_S4". (**A**) SGA-hap in the non-HLA regions. (**B**) SGA-hap in the HLA region. (**C**) SGA-PDBG in the non-HLA regions. (**D**) SGA-PDBG in the HLA region.

In SGA-PDBG assembly from both WGS and WES data, we observed large variation among the four k-mers in known SNP sensitivity (**Figures 4.12A-B**), suggesting the necessity of using multiple k-mers to maximize sensitivity. For SGA-hap, the SNP sensitivity was less variable among the five k-mers, except in *HLA-DRB1* (**Figures 12C-D).** Accordingly, in WES data FC1_NA12878_S1_S4 we observed a much larger overlap of known SNPs identified by SGA-haplotype at different k-mers (**Figures 4.13A-B**) than SGA-PDBG (**Figures 4.13C-D**). Eighty-one percent (784/968) of the known SNPs were shared by all k-mers in SGA-hap, compared to only 38.6% (404/1046) in SGA-PDBG. The less dependence on k-mer selection makes SGA-haplotype particularly appealing, since it will save both computing resources and turnaround time.

4.5. Discussion

We developed SGA-haplotype, a string graph-based, chromosome-scale *de novo* assembly pipeline*.* Utilizing reads mapped to a given chromosome as well as unmapped reads, SGA-haplotype achieved high sensitivity and precision in variant discovery from WES. *De novo* assembly is increasingly used in haplotype construction and the detection of large INDELs and complex variants, mostly from WGS. In these instances mapping-based approaches are less effective. The power of *de novo* assembly depends on read length, read quality, sequencing depth, assembly algorithms and parameter settings [41, 126]. Current methods used for short reads assembly often implement de Bruijn graph, a data structure operating on k-mers rather than on reads [127]. We found that SGA-haplotype outperformed all tested de Bruijn graph- and string graph-based assemblers, particularly in highly divergent regions such as the HLA region.

The performance of SGA-haplotype is generally independent of divergence and coverage, which is superior to DISCOVAR, Cortex, SGA graph-diff and fermi2 that have a much lower sensitivity at high divergence, and to fermi that requires high sequencing depth. Critically, SGA-haplotype could achieve long-range phasing up to a few kilobases and detect large INDELs in WES, more notably in WGS owing to the continuity in read coverage. In WGS from NA12878, a 10-kb region representing part of *HLA-DQB1* harbors nearly 600 phased SNPs; another 12-kb region from *HLA-DRB1* contains two large deletions of over 60 kb and 249 phased SNPs. Phasing provides pertinent information as to whether any two or more deleterious variants are located on the same allele [128]. In organ transplantation, phase information is critical for predicting the donor-recipient match. Although short reads may carry limited phase information if they contain multiple variation sites [129], *de novo* assembly can accomplish long-range phasing, increasing the chance of identifying linking variants that are associated with complex diseases [128, 130, 131]. In this aspect, the incorporation of mate-pair or chromosome conformation capture (3C)-based sequencing data into the assembly will help resolve haplotypes over large regions [128]. Obviously, haplotype construction has been and will continue to be essential for genomic-based disease diagnosis [131].

SGA-haplotype has demonstrated the high efficiency in variant discovery from both WGS and WES, despite the great variability of coverage in WES due to bias in base composition, exome capture and mapping efficiency. There are over 100 highly divergent regions in the human genome [38, 41], from which variant discovery has been challenging. Our study confirmed that SGA-haplotype represents the most sensitive variation-aware *de novo* assembler, not just in ordinary regions but also in highly divergent regions.

Current *de novo* assemblers perform either whole genome assembly or local assembly within a few kilobases. In the latter, only reads mapped to particular regions will be included, missing the variants covered by unmapped reads. SGA-haplotype performs chromosome-level assembly, which takes pairs with one or both reads mapped to a given chromosome and unmapped pairs. We argue that at least some of the unmapped pairs are derived from regions of complex variation or large INDELs, or regions not represented in the reference. It is estimated that 5-40 Mb euchromatic sequences are missed from human reference genome [41]. Some of them are associated with complex disease [132-134]. The inclusion of unmapped pairs should be particularly helpful for *de novo* assembly in highly divergent regions to which current short-read aligners are less sensitive. To ensure haplotype assembly and minimize noise, we have also implemented highly stringent reads filtering, reads trimming, k-mer-based error correction and contig filtering. Also, the chromosome-level assembly approach provides the options to target one or a few chromosomes or to run whole genome assembly much faster through parallel computing.

Nevertheless, detection of INDELs, especially large ones, has been difficult for most methods [53, 135]. At low divergence (like non-HLA regions), SGA-haplotype is generally less sensitive than fermi2, SGA-SG and the three mapping-based approaches. We reason that SGA-haplotype sensitivity is likely underestimated, partially due to the relatively lower accuracy and less complete annotation of INDELs in the reference call set. Existing evidence indicates that some of the INDELs in the reference call set represent false positives. Also, the vast majority of the INDELs in the reference were generated by BWA-MEM together with GATK HaplotypeCaller, which likely introduces bias into the comparison, since the three mapping-based approaches also used BWA-MEM as the mapper. In addition, we found that re-mapping of contigs with soft-clipping by BLAT improves the detection of large INDELs. Further effort will be needed to improve SGA-haplotype in the detection of INDELs. The strategy of joint calling with both mapping-based approaches and SGA-haplotype will certainly enhance the detection of INDELs.

The SGA-haplotype pipeline is designed for chromosome-level assembly. However, some studies may focus on some specific regions or genes, like the six HLA genes we analyzed here. Also, targeted resequencing of clinical samples is widely used in biomedical studies. Thus, implementing a separate module for regional assembly will be beneficial.

Parameter settings have a major impact on the outcome of *de novo* assembly. In de Bruijn graph, k-mer represents the overlap parameter. A longer k-mer improves the specificity at the cost of sensitivity, as the use of longer k-mer will reduce the coverage. Conversely, a shorter k-mer would create more fragmented contigs, resulting in higher false positives. It is recommended that de Bruijn graph assemblers should run multiple k-mers to increase the sensitivity, as implemented in Cortex [57]. However, extra efforts are required to consolidate the multiple variant sets into a unified list. In string graph based assemblers such as SGA-haplotype, the k-mer is only used in error correction and not a key parameter in the assembly step. Despite the importance, the choice of k-mers and other key parameters like minimal overlap length is often arbitrary. A more systematic assessment of various parameters is needed to develop a guideline for end users, which should enhance the SGA-haplotype performance and be instructive for other assemblers as well.

Great effort has been taken to establish public call sets from NA12878 WES and WGS through mapping- and *de novo* assembly-based approaches. While these lists are likely of high quality in ordinary regions, we found that they are much less complete in the HLA region. For NA12878, SGA-haplotype assembled two “error-free” haplotypes in both *HLA-DRB1* and *HA-DQB1* from 250-bp WGS that perfectly matched the alleles in one of the parents. There are 833 phased SNPs between these alleles and the hg19 reference sequence. However, the public DISCOVAR and GATK HaplotypeCaller call sets, the so-called “golden standard” [54], missed 76.4% and 11.3% of these variants, respectively. As the assessment of variant discovery methods strongly depends on the quality of the reference call set, the generation of a more complete and accurate reference call set for the HLA region should be one of the major tasks in this field.

4.6. Conclusions

Complete and accurate variant detection is critical as we move to precision-based medical practice. We have developed the SGA-haplotype pipeline for *de novo* assembly-based variant discovery from WES data. SGA-haplotype took advantage of the SGA data structure and implemented key modules toward haplotype assembly. By splitting mapped reads over individual chromosomes and including unmapped reads in the assembly, SGA-haplotype outperforms the other *de novo* assembly based approaches on WES data, especially in regions of high genomic complexity like HLA. Specifically, SGA-haplotype is robust against sequencing errors, coverage and divergence variability, and k-mer selection, and has much less memory usage. Though limited by the WES design to target only exons and some important genomic regions, SGA-haplotype can detect large INDELs and achieve long-range phasing, a feature more obvious in WGS. Further effort is needed to optimize SGA-haplotype in the detection of INDELs. Also, it is necessary to develop modules for local assembly of selected regions, such as those in the targeted gene sequencing panels.

Chapter 5

Conclusions

Next-generation DNA sequencing has enabled the sequencing of thousands of human genomes and exomes, from both healthy individuals (The 1000 Genomes Project Consortium) and many of the diseases (The Cancer Genome Atlas), for examining population variation and identifying causal mutations. The HLA region is associated with more than 100 diseases, and some specific alleles in HLA genes play critical roles in drug hypersensitivity [42, 136]. For example, strong associations have been identified between carbamazepine-induced Stevens-Johnson syndrome (SJS) or toxic epidermal necrolysis (TEN) and HLA-B*15:02 in multiple populations [137, 138]. Therefore, fully cataloguing variants from such highly divergent yet medically important regions should provide insight into disease etiology and contribute to pharmacogenomics studies. This dissertation addresses the challenges by providing a practical guideline for variant discovery in highly divergent regions. We recommend first using mapping-based approaches to identify variants from exome data without base quality score recalibration. The initial mapping can be done using BWA, followed by re-mapping of unmapped reads using a more sensitive mapper like GSNAP. To achieve high sensitivity for both SNP and INDEL detection, Platypus and the two GATK callers can be used in combination. SGA-haplotype can then be used for de novo assembly of specific chromosomes, such as chromosomes 6, 8 and 16 that contain regions of high divergence. SGA-haplotype can provide phase information of the identified variants, and can identify large INDELs and structural variants in WES data on which mapping-based approaches have low sensitivity. We developed an integrated pipeline for both mapping- and de novo assembly-based variant discovery from exome data.

Mapping-based approach represents the most sensitive method for determining genetic variation in majority of the genomic regions. Based on the extensive assessment with both simulated and real exome data (see **Chapter 3**), we proposed a variant discovery workflow that integrates two mappers and three callers, without base quality score recalibration (see **Chapter 2**). Specifically, we implemented a two-step mapping strategy that achieves both high speed (by BWA) and accuracy (by GSNAP), and a joint calling strategy through three distinct variant callers (Platypus and two GATK callers) which enables highly sensitive detection of both SNPs and INDELs. The workflow is designed with high flexibility; it can be run in distributed mode across different samples and chromosomes in parallel.

De novo assembly-based variant calling represents an attractive alternative but is not widely used yet in WES studies. It suffers from short read length, sequencing errors, and high demand on computing resources. To address these challenges, we developed a variant calling workflow, termed SGA-haplotype, for whole exome data on top of the SGA data structure (see **Chapter 4**). We demonstrated its high sensitivity and robustness over a wide range of divergence with both simulated and real exome reads. The workflow utilizes pairs with uniquely mapped read(s) and unmapped read pairs for chromosome-level de novo assembly, thus maximizing the detection sensitivity in highly divergent regions. Its operation at chromosome-level (rather than at whole genome-level) alleviates the limitation in computing. Besides the standard Variant Call Format (VCF) files, the workflow also outputs the assembled haplotype sequences that can be further used for HLA typing through the international ImMunoGeneTics (IMGT)/HLA database [97].

We developed a highly integrative yet flexible pipeline by incorporating the mapping- and de novo assembly-based approaches discussed above. For simplicity, we illustrated the two approaches separately (**Figures 5.1** and **5.2)**. The pipeline runs with a configure file specifying the mapping and assembling parameter settings and a sample information file listing the sequence files. It provides three running modes: mapped-based, assembly-based and combined. The pipeline was implemented as distributed computing and configured to run in computer clusters. However, it can be easily customized to run in standalone fashion.

We used both simulated and real exome datasets to comprehensively assess the performance of 25 methods in variant discovery. We proposed a strategy for post-alignment processing and developed workflows for both mapping- and *de novo* assembly-based variant detection, particularly in highly divergent regions. Of the mapping-based methods, GATK UnifiedGenotyper is ideal for SNP calling from highly divergent regions like HLA, followed by GATK HaplotypeCaller and SAMtools, while GATK HaplotypeCaller and Platypus are more effective in INDEL detection. Meanwhile, SGA-haplotype can be used for HLA typing and neoantigen discovery, as well as for the identification of virus integration sites and fusion genes in cancer genomes.

This project has a few limitations. First, our study focused on the detection of germline mutations without considering genomic complexity in cancer genomes, such as copy number changes. Second, our methods are intended for variant detection in personal genomes, rather than across multiple samples, thus for some cohort studies, further work is needed to genotype samples together based on the discovered variant sites from personal genome. Third, our approaches were developed for data from Illumina platforms, and additional tests will be needed to confirm their applicability for other platforms such as Pacific Bioscience where the sequencing errors are usually much higher than that of Illumina. Fourth, to extend the applications of SGA-haplotype, it is necessary to develop a separate module for regional assembly in specific regions or genes, such as those from the targeted resequencing panels; further effort is also needed to optimize SGA-haplotype in areas of contig mapping, variant detection and INDEL reporting in complex regions. Finally, following the advancement in long read sequencing technologies, variant discovery will likely be less dependent on reference sequence; as such, *de novo* assemblers, especially those based on overlap-layout-consensus or string graph, are expected to be the focus of future development.

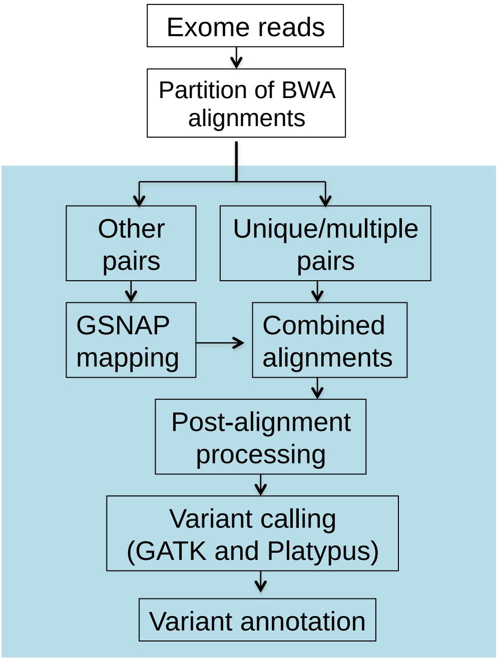


Figure 5.1. Workflow of mapping-based variant discovery in exome data. GATK refers to GATK UnifiedGenotyper and GATK HaplotypeCaller.

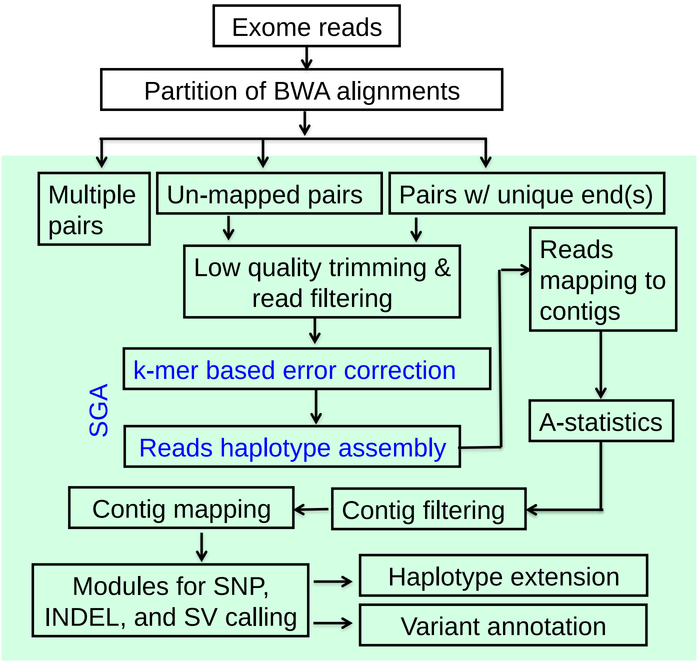


Figure 5.2. Workflow of de novo assembly-based variant discovery from exome data

Appendix

1. Dwgsim read simulation

1.1. Development of dwgsim

We used dwgsim to simulate 100-bp paired-end reads by introducing random mutations into the human genome reference. The tool was originally developed as part of the MAQ (Mapping and Assembly with Qualities) software package [30]. This tool uses two haplotypes to represent the reference genome and randomly generates a set of polymorphic sites (2/3 as heterozygotes and 1/3 as homozygotes) based on a pre-defined mutation rate. At a heterozygous site, dwgsim randomly selects one of the two haplotypes and then converts its base into either one of the other three bases; at a homozygous site, the two bases from both haplotypes will be converted into a different one. However, this module can only simulate single base pair INDEL. In 2011, Heng Li forked the module out as a standalone project (wgsim became part of the SAMtools software package) by dropping dependencies on other source code in the MAQ package. With the incorporation of patches from Colin Hercus, a lead developer of Novoalign, wgsim can simulate INDELs of longer than 1bp. Nils Homer (the Manager of software engineering in Genomics Platform at Broad Institute) further improved the package by adding functionalities, including that of simulating platform-specific sequencing errors, and made it part of the DNAA package (https://sourceforge.net/projects/dnaa/).

1.2. Key parameters

Among the 28 command line parameters in dwgsim, the most important ones are: “-r”, “-R” and “-X”. The -r option specifies the mutation rate in simulation (say 0.001, i.e., 0.1%) and -R specifies the fraction of mutated positions that are INDELs. If the mutation rate is 0.1% and the INDEL fraction is 0.1, then the INDEL mutation rate will be 0.01%. The option -X controls the probability distribution of INDEL length (default 0.3) with an equal likelihood in the number of insertions and deletions.

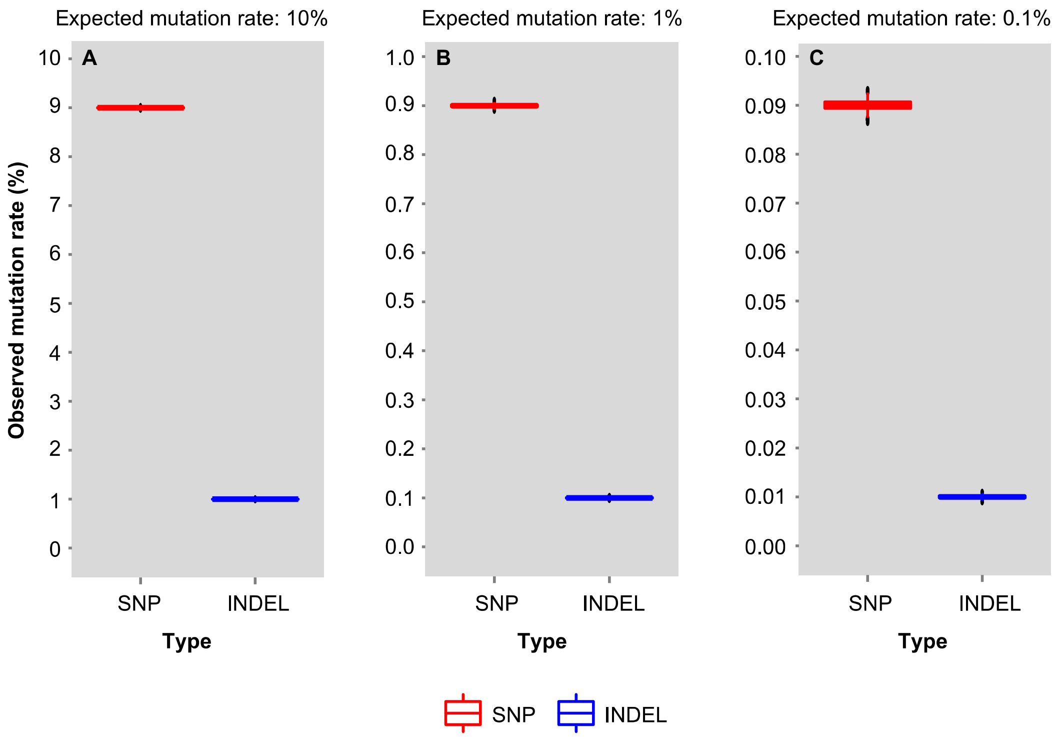
1.3. Random mutations

To introduce random mutations, the software checks base by base along a DNA sequence template for positions to be mutated. The process is controlled by a pseudo random number generator (srand48(time(0)); double r = drand48()), which returns a pseudo-random number in the range of 0.0 to 1.0 for each of the bases. The random number generator is the linear congruential algorithm and 48-bit integer arithmetic. Before the first call to "drand48", srand48(time(0) is called to initialize the random "seed". Here is the process:

For each base, check whether r <= mutation rate; if yes, mutate the base. The chance of getting a random number that is smaller than the mutation rate is mutation rate. Therefore, each base has the same chance of being mutated.

1.4. Validation of mutation rate in simulated reads

To validate the reliability of dwgsim in generating random mutations, we extracted a 10-Mb sequence from human reference (chr1:4,000,000-14,000,000) and performed 1,000 simulations for each of the 3 pre-defined mutation rates (r=0.1%, 1% and 10%, with a SNP-to-INDEL ratio of 9 to 1). We then counted the number of mutations in the output text file. The boxplots below showed the distributions of the mutation rates calculated from the 1,000 simulations, separately for SNPs and INDELs. For the three mutation rates, the mean was very close to the pre-defined mutation rate (Figures A-C below).



2. Variant calling quality metrics

If there is a reference call set, then all the (variation and non-variation) sites can be split into true positive (TP), false positive (FP), true negative (TN) and false negative (FN) sites based on the intersection between variants called by a method and the reference call set. This 2x2 table will allow one to estimate five common variant calling quality metrics. In our studies, as the true negatives are not known for real exome data, we only calculated sensitivity and precision rate.

		Reference variants set	
		Positive	Negative
List of called variants	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Metric	Formula	Interpretation
Accuracy	(TP+TN)/(TP+FP+TN+FN)	Ratio of correctly called variation and non-variation sites over total calls
Sensitivity	TP/(TP+FN)	Ratio of called variants overlapping reference (true) variants over total reference variants
Specificity	TN/(TN+FP)	Ratio of non-variation sites not called as variants over total non-variation sites
Precision rate	TP/(TP+FP)	Ratio of called variants overlapping reference (true) variants over total called variants
False positive rate	FP/(TN+FP)	Ratio of non-variation sites called as variants over total non-variation sites

Bibliography

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nat Rev Genet* 2008, **9**(5)**:**356-369.
2. Yu X, Sun S: **Comparing a few SNP calling algorithms using low-coverage sequencing data.** *BMC Bioinformatics* 2013, **14:**274.
3. Kaname T, Yanagi K, Naritomi K: **A commentary on the promise of whole-exome sequencing in medical genetics.** *J Hum Genet* 2014, **59**(3)**:**117-118.
4. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, Wilkie AO, McVean G, Lunter G: **Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications.** *Nat Genet* 2014, **46**(8)**:**912-918.
5. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet* 2010, **42**(1)**:**30-35.
6. Lupski JR, Gonzaga-Jauregui C, Yang Y, Bainbridge MN, Jhangiani S, Buhay CJ, Kovar CL, Wang M, Hawes AC, Reid JG, et al: **Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy.** *Genome Med* 2013, **5**(6)**:**57.
7. Rabbani B, Tekin M, Mahdieh N: **The promise of whole-exome sequencing in medical genetics.** *J Hum Genet* 2014, **59**(1)**:**5-15.
8. Berg JS, Khoury MJ, Evans JP: **Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time.** *Genet Med* 2011, **13**(6)**:**499-504.
9. Khurana E, Fu Y, Chen J, Gerstein M: **Interpretation of genomic variants using a unified biological network approach.** *PLoS Comput Biol* 2013, **9**(3)**:**e1002886.
10. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M: **Exome Sequencing: Current and Future Perspectives.** *G3 (Bethesda)* 2015, **5**(8)**:**1543-1550.
11. Iglesias A, Anyane-Yeboa K, Wynn J, Wilson A, Truitt Cho M, Guzman E, Sisson R, Egan C, Chung WK: **The usefulness of whole-exome sequencing in routine clinical practice.** *Genetics in medicine* 2014, **16**(12)**:**922-931.
12. Xu Y, Xiao B, Jiang WT, Wang L, Gen HQ, Chen YW, Sun Y, Ji X: **A novel mutation identified in PKHD1 by targeted exome sequencing: guiding prenatal diagnosis for an ARPKD family.** *Gene* 2014, **551**(1)**:**33-38.
13. Bras JM, Singleton AB: **Exome sequencing in Parkinson's disease.** *Clin Genet* 2011, **80**(2)**:**104-109.

14. Sassi C, Guerreiro R, Gibbs R, Ding J, Lupton MK, Troakes C, Lunnon K, Al-Sarraj S, Brown KS, Medway C, et al: **Exome sequencing identifies 2 novel presenilin 1 mutations (p.L166V and p.S230R) in British early-onset Alzheimer's disease.** *Neurobiol Aging* 2014, **35**(10)**:**2422 e2413-2426.
15. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43**(5)**:**491-498.
16. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al: **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.** *Curr Protoc Bioinformatics* 2013, **11**(1110)**:**11 10 11-11 10 33.
17. Li H: **Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly.** *Bioinformatics* 2012, **28**(14)**:**1838-1844.
18. Li H: **Toward better understanding of artifacts in variant calling from high-coverage samples.** *Bioinformatics* 2014, **30**(20)**:**2843-2851.
19. Caboche S, Audebert C, Lemoine Y, Hot D: **Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data.** *BMC Genomics* 2014, **15:**264.
20. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing.** *Brief Bioinform* 2010, **11**(5)**:**473-483.
21. Lunter G, Goodson M: **Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.** *Genome Res* 2011, **21**(6)**:**936-939.
22. Berger B, Peng J, Singh M: **Computational solutions for omics data.** *Nat Rev Genet* 2013, **14**(5)**:**333-346.
23. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: **A survey of tools for variant analysis of next-generation genome sequencing data.** *Brief Bioinform* 2014, **15**(2)**:**256-278.
24. Flicek P, Birney E: **Sense from sequence reads: methods for alignment and assembly.** *Nat Methods* 2009, **6**(11 Suppl)**:**S6-S12.
25. Sedlazeck FJ, Rescheneder P, von Haeseler A: **NextGenMap: fast and accurate read mapping in highly polymorphic genomes.** *Bioinformatics* 2013, **29**(21)**:**2790-2791.
26. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads.** *Bioinformatics* 2010, **26**(7)**:**873-881.
27. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples.** *Bioinformatics* 2009, **25**(17)**:**2283-2285.

28. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14)**:**1754-1760.
29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20**(9)**:**1297-1303.
30. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**(11)**:**1851-1858.
31. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J: **SNP detection for massively parallel whole-genome resequencing.** *Genome Res* 2009, **19**(6)**:**1124-1132.
32. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y: **Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data.** *BMC Genomics* 2012, **13 Suppl 8:**S8.
33. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet* 2011, **12**(6)**:**443-451.
34. Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** http://arxiv.org/abs/1207.3907v2. 2012.
35. Li H: **Improving SNP discovery by base alignment quality.** *Bioinformatics* 2011, **27**(8)**:**1157-1158.
36. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**(3)**:**186-194.
37. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al: **The diploid genome sequence of an individual human.** *PLoS biology* 2007, **5**(10)**:**e254.
38. McLure CA, Hinchliffe P, Lester S, Williamson JF, Millman JA, Keating PJ, Stewart BJ, Dawkins RL: **Genomic evolution and polymorphism: segmental duplications and haplotypes at 108 regions on 21 chromosomes.** *Genomics* 2013, **102**(1)**:**15-26.
39. The MHC sequencing consortium: **Complete sequence and gene map of a human major histocompatibility complex.** *Nature* 1999, **401**(6756)**:**921-923.
40. Raymond CK, Kas A, Paddock M, Qiu R, Zhou Y, Subramanian S, Chang J, Palmieri A, Haugen E, Kaul R, Olson MV: **Ancient haplotypes of the HLA Class II region.** *Genome Res* 2005, **15**(9)**:**1250-1257.
41. Chaisson MJ, Wilson RK, Eichler EE: **Genetic variation and the de novo assembly of human genomes.** *Nat Rev Genet* 2015, **16**(11)**:**627-640.
42. Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I: **Phase-defined complete sequencing of the HLA genes by next-generation sequencing.** *BMC Genomics* 2013, **14:**355.
43. Pillai NE, Okada Y, Saw WY, Ong RT, Wang X, Tantoso E, Xu W, Peterson TA, Bielawny T, Ali M, et al: **Predicting HLA alleles from high-resolution**

SNP data in three Southeast Asian populations. *Hum Mol Genet* 2014, **23**(16)**:**4443-4451.

44. Nusbaum C, Mikkelsen TS, Zody MC, Asakawa S, Taudien S, Garber M, Kodira CD, Schueler MG, Shimizu A, Whittaker CA, et al: **DNA sequence and analysis of human chromosome 8.** *Nature* 2006, **439**(7074)**:**331-335.
45. Martin J, Han C, Gordon LA, Terry A, Prabhakar S, She X, Xie G, Hellsten U, Chan YM, Altherr M, et al: **The sequence and analysis of duplication-rich human chromosome 16.** *Nature* 2004, **432**(7020)**:**988-994.
46. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**(7319)**:**1061-1073.
47. Liu X, Han S, Wang Z, Gelernter J, Yang BZ: **Variant callers for next-generation sequencing data: a comparison study.** *PLoS One* 2013, **8**(9)**:**e75619.
48. Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM: **Performance comparison of SNP detection tools with illumina exome sequencing data-an assessment using both family pedigree information and sample-matched SNP array data.** *Nucleic Acids Res* 2014, **42**(12)**:**e101.
49. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP: **Validation and assessment of variant calling pipelines for next-generation sequencing.** *Hum Genomics* 2014, **8:**14.
50. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, Mittelman D: **An analytical framework for optimizing variant discovery from personal genomes.** *Nat Commun* 2015, **6:**6275.
51. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M: **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.** *Nat Biotechnol* 2014, **32**(3)**:**246-251.
52. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, et al: **Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.** *Genome Med* 2013, **5**(3)**:**28.
53. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC: **Accurate de novo and transmitted indel detection in exome-capture data using microassembly.** *Nat Methods* 2014, **11**(10)**:**1033-1036.
54. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, et al: **Comprehensive variation discovery in single human genomes.** *Nat Genet* 2014, **46**(12)**:**1350-1355.
55. Campbell CD, Eichler EE: **Properties and rates of germline mutations in humans.** *Trends Genet* 2013, **29**(10)**:**575-584.
56. Brandt DY, Aguiar VR, Bitarello BD, Nunes K, Goudet J, Meyer D: **Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in**

the 1000 Genomes Project Phase I Data. *G3 (Bethesda)* 2015, **5**(5)**:**931-941.

57. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs.** *Nat Genet* 2012, **44**(2)**:**226-232.
58. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491**(7422)**:**56-65.
59. Pop M: **Genome assembly reborn: recent computational challenges.** *Brief Bioinform* 2009, **10**(4)**:**354-366.
60. Myers EW: **The fragment assembly string graph.** *Bioinformatics* 2005, **21 Suppl 2:**ii79-85.
61. Nagarajan N, Pop M: **Sequence assembly demystified.** *Nat Rev Genet* 2013, **14**(3)**:**157-167.
62. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5)**:**821-829.
63. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19**(6)**:**1117-1123.
64. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proc Natl Acad Sci U S A* 2011, **108**(4)**:**1513-1518.
65. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20**(2)**:**265-272.
66. Birol I, Chu J, Mohamadi H, Jackman SD, Raghavan K, Vandervalk BP, Raymond A, Warren RL: **Spaced Seed Data Structures for De Novo Assembly.** *Int J Genomics* 2015, **2015:**196591.
67. Simpson JT, Durbin R: **Efficient construction of an assembly string graph using the FM-index.** *Bioinformatics* 2010, **26**(12)**:**i367-373.
68. Simpson JT, Durbin R: **Efficient de novo assembly of large genomes using compressed data structures.** *Genome Res* 2012, **22**(3)**:**549-556.
69. Henson J, Tischler G, Ning Z: **Next-generation sequencing and large genome assemblies.** *Pharmacogenomics* 2012, **13**(8)**:**901-915.
70. Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing.** *Genome Res* 2010, **20**(9)**:**1165-1173.
71. Escalona M, Rocha S, Posada D: **A comparison of tools for the simulation of genomic next-generation sequencing data.** *Nat Rev Genet* 2016, **17**(8)**:**459-469.
72. Xu Q, Wu X, Li M, Huang H, Minica C, Yi Z, Wang G, Shen L, Xing Q, Shi Y, et al: **Association studies of genomic variants with treatment response to risperidone, clozapine, quetiapine and chlorpromazine in the Chinese Han population.** *Pharmacogenomics J* 2016, **16**(4)**:**357-365.

73. Wang Z, Liu X, Yang BZ, Gelernter J: **The role and challenges of exome sequencing in studies of human diseases.** *Front Genet* 2013, **4:**160.

74. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS: **Variant detection sensitivity and biases in whole genome and exome sequencing.** *BMC Bioinformatics* 2014, **15:**247.

75. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F: **An integrative variant analysis suite for whole exome next-generation sequencing data.** *BMC Bioinformatics* 2012, **13:**8.

76. Hwang S, Kim E, Lee I, Marcotte EM: **Systematic comparison of variant calling pipelines using gold standard personal exome variants.** *Sci Rep* 2015, **5:**17875.

77. Slager SL, Skibola CF, Di Bernardo MC, Conde L, Broderick P, McDonnell SK, Goldin LR, Croft N, Holroyd A, Harris S, et al: **Common variation at 6p21.31 (BAK1) influences the risk of chronic lymphocytic leukemia.** *Blood* 2012, **120**(4)**:**843-846.

78. Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, Levinson D, Fernandez-Vina MA, Davis RW, Davis MM, Mindrinos M: **High-throughput, high-fidelity HLA genotyping with deep sequencing.** *Proc Natl Acad Sci U S A* 2012, **109**(22)**:**8676-8681.

79. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator.** *Bioinformatics* 2012, **28**(4)**:**593-594.

80. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16)**:**2078-2079.

81. Warden CD, Adamson AW, Neuhausen SL, Wu X: **Detailed comparison of two popular variant calling packages for exome and targeted exon studies.** *PeerJ* 2014, **2:**e600.

82. Jiang Y, Turinsky AL, Brudno M: **The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection.** *Nucleic Acids Res* 2015, **43**(15)**:**7217-7228.

83. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al: **Mutational heterogeneity in cancer and the search for new cancer-associated genes.** *Nature* 2013, **499**(7457)**:**214-218.

84. Robinson PN, Krawitz P, Mundlos S: **Strategies for exome and genome sequence data analysis in disease-gene discovery projects.** *Clin Genet* 2011, **80**(2)**:**127-132.

85. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, et al: **Molecular findings among patients referred for clinical whole-exome sequencing.** *JAMA* 2014, **312**(18)**:**1870-1879.

86. Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunham I, Forbes S, Halls K, Howson JM, et al: **Complete MHC haplotype**

sequencing for common disease gene mapping. *Genome Res* 2004, **14**(6)**:**1176-1187.

87. Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, Palmieri A, Subramanian S, Zhou Y, Kaul R, et al: **Scan of human genome reveals no new Loci under ancient balancing selection.** *Genetics* 2006, **173**(4)**:**2165-2177.
88. Hodgkinson A, Eyre-Walker A: **Variation in the mutation rate across mammalian genomes.** *Nat Rev Genet* 2011, **12**(11)**:**756-766.
89. Middleton D, Gonzelez F: **The extensive polymorphism of KIR genes.** *Immunology* 2010, **129**(1)**:**8-19.
90. Cheng AY, Teo YY, Ong RT: **Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals.** *Bioinformatics* 2014, **30**(12)**:**1707-1713.
91. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**(6)**:**841-842.
92. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** *arXiv:13033997.* 2013.
93. Slager SL, Rabe KG, Achenbach SJ, Vachon CM, Goldin LR, Strom SS, Lanasa MC, Spector LG, Rassenti LZ, Leis JF, et al: **Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL.** *Blood* 2011, **117**(6)**:**1911-1916.
94. Heinrich V, Kamphans T, Stange J, Parkhomchuk D, Hecht J, Dickhaus T, Robinson PN, Krawitz PM: **Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects.** *Genome Med* 2013, **5**(7)**:**69.
95. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, Oka A, Umemura T, Joshita S, Takahashi O, et al: **Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers.** *Tissue Antigens* 2012, **80**(4)**:**305-316.
96. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG: **The IMGT/HLA database.** *Nucleic Acids Res* 2013, **41**(Database issue)**:**D1222-1227.
97. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG: **The IPD and IMGT/HLA database: allele variant databases.** *Nucleic Acids Res* 2015, **43**(Database issue)**:**D423-431.
98. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, Holt RA: **Derivation of HLA types from shotgun sequence datasets.** *Genome Med* 2012, **4**(12)**:**95.
99. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al: **Resolving the complexity of the human genome using single-molecule sequencing.** *Nature* 2015, **517**(7536)**:**608-611.

100. Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Newburger DE, West R, Sidow A, Batzoglou S: **Read clouds uncover variation in complex regions of the human genome.** *Genome Res* 2015, **25**(10)**:**1570-1580.
101. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al: **An integrated map of structural variation in 2,504 human genomes.** *Nature* 2015, **526**(7571)**:**75-81.
102. Bodily PM, Fujimoto M, Ortega C, Okuda N, Price JC, Clement MJ, Snell Q: **Heterozygous genome assembly via binary classification of homologous sequence.** *BMC Bioinformatics* 2015, **16 Suppl 7:**S5.
103. Yang WY, Hormozdiari F, Wang Z, He D, Pasaniuc B, Eskin E: **Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data.** *Bioinformatics* 2013, **29**(18)**:**2245-2252.
104. Li H: **FermiKit: assembly-based variant calling for Illumina resequencing data.** *Bioinformatics* 2015, **31**(22)**:**3694-3696.
105. Simpson JT, Pop M: **The Theory and Practice of Genome Sequence Assembly.** *Annu Rev Genomics Hum Genet* 2015, **16:**153-172.
106. Li H: **BFC: correcting Illumina sequencing errors.** *Bioinformatics* 2015, **31**(17)**:**2885-2887.
107. Gough SC, Simmonds MJ: **The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action.** *Curr Genomics* 2007, **8**(7)**:**453-465.
108. Tian S, Yan H, Neuhauser C, Slager SL: **An analytical workflow for accurate variant discovery in highly divergent regions.** *BMC Genomics* 2016, **17:**703.
109. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM: **Best practices for evaluating single nucleotide variant calling methods for microbial genomics.** *Front Genet* 2015, **6:**235.
110. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al: **GAGE: A critical evaluation of genome assemblies and assembly algorithms.** *Genome Res* 2012, **22**(3)**:**557-567.
111. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al: **A whole-genome assembly of Drosophila.** *Science* 2000, **287**(5461)**:**2196-2204.
112. Kelley DR, Salzberg SL: **Detection and correction of false segmental duplications caused by genome mis-assembly.** *Genome Biol* 2010, **11**(3)**:**R28.
113. Li H: **Tabix: fast retrieval of sequence features from generic TAB-delimited files.** *Bioinformatics* 2011, **27**(5)**:**718-719.
114. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data.** *Genomics* 2010, **95**(6)**:**315-327.

115. Tian S, Yan H, Kalmbach M, Slager SL: **Impact of post-alignment processing in variant discovery from whole exome data.** *BMC Bioinformatics* 2016, **17**(1)**:**403.
116. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, et al: **Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph.** *Brief Funct Genomics* 2012, **11**(1)**:**25-37.
117. Mirebrahim H, Close TJ, Lonardi S: **De novo meta-assembly of ultra-deep sequencing data.** *Bioinformatics* 2015, **31**(12)**:**i9-16.
118. Lonardi S, Mirebrahim H, Wanamaker S, Alpert M, Ciardo G, Duma D, Close TJ: **When less is more: 'slicing' sequencing data improves read decoding accuracy and de novo assembly quality.** *Bioinformatics* 2015, **31**(18)**:**2972-2980.
119. Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, Kulkarni V, Jere A: **Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data.** *PLoS One* 2013, **8**(4)**:**e60204.
120. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC: **Effects of GC bias in next-generation-sequencing data on de novo genome assembly.** *PLoS One* 2013, **8**(4)**:**e62856.
121. Meienberg J, Zerjavic K, Keller I, Okoniewski M, Patrignani A, Ludin K, Xu Z, Steinmann B, Carrel T, Rothlisberger B, et al: **New insights into the performance of human whole-exome capture platforms.** *Nucleic Acids Res* 2015, **43**(11)**:**e76.
122. Paszkiewicz K, Studholme DJ: **De novo assembly of short sequence reads.** *Brief Bioinform* 2010, **11**(5)**:**457-472.
123. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29**(1)**:**24-26.
124. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, et al: **Sequence-specific error profile of Illumina sequencers.** *Nucleic Acids Res* 2011, **39**(13)**:**e90.
125. Narzisi G, Schatz MC: **The challenge of small-scale repeats for indel discovery.** *Front Bioeng Biotechnol* 2015, **3:**8.
126. Compeau PE, Pevzner PA, Tesler G: **How to apply de Bruijn graphs to genome assembly.** *Nat Biotechnol* 2011, **29**(11)**:**987-991.
127. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proc Natl Acad Sci U S A* 2001, **98**(17)**:**9748-9753.
128. Selvaraj S, J RD, Bansal V, Ren B: **Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing.** *Nat Biotechnol* 2013, **31**(12)**:**1111-1118.
129. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J: **Haplotype estimation using sequencing reads.** *Am J Hum Genet* 2013, **93**(4)**:**687-696.

130. Fujimoto M, Bodily PM, Okuda N, Clement MJ, Snell Q: **Effects of error-correction of heterozygous next-generation sequencing data.** *BMC Bioinformatics* 2014, **15 Suppl 7:**S3.
131. Glusman G, Cox HC, Roach JC: **Whole-genome haplotyping approaches and genomic medicine.** *Genome Med* 2014, **6**(9)**:**73.
132. Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, Sudmant PH, Dorajoo R, Al-Shafai MN, Bottolo L, et al: **Low copy number of the salivary amylase gene predisposes to obesity.** *Nat Genet* 2014, **46**(5)**:**492-497.
133. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, Hebert M, Jones KN, Shu Y, Kitzmiller K, et al: **Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans.** *Am J Hum Genet* 2007, **80**(6)**:**1037-1054.
134. Shen S, Pyo CW, Vu Q, Wang R, Geraghty DE: **The essential detail: the genetics and genomics of the primate immune response.** *ILAR J* 2013, **54**(2)**:**181-195.
135. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS: **ABRA: improved coding indel detection via assembly-based realignment.** *Bioinformatics* 2014, **30**(19)**:**2813-2815.
136. Cheng CY, Su SC, Chen CH, Chen WL, Deng ST, Chung WH: **HLA associations and clinical implications in T-cell mediated drug hypersensitivity reactions: an updated review.** *J Immunol Res* 2014, **2014:**565320.
137. Alfirevic A, Jorgensen AL, Williamson PR, Chadwick DW, Park BK, Pirmohamed M: **HLA-B locus in Caucasian patients with carbamazepine hypersensitivity.** *Pharmacogenomics* 2006, **7**(6)**:**813-818.
138. Hung SI, Chung WH, Jee SH, Chen WC, Chang YT, Lee WR, Hu SL, Wu MT, Chen GS, Wong TW, et al: **Genetic susceptibility to carbamazepine-induced cutaneous adverse drug reactions.** *Pharmacogenet Genomics* 2006, **16**(4)**:**297-306.