Distribution Agreement

Signature:

Shuai Yuan　　　　Date

Applying Diploid Method to Improve Read-mapping and Analysis Based on NGS Data

By

Shuai Yuan
Doctor of Philosophy

Computer Science and Informatics

Zhaohui Qin
Advisor

Shun Yan Cheung
Co-advisor

Yijuan Hu
Committee Member

Yun Li
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Applying Diploid Method to Improve Read-mapping and Analysis Based on NGS Data

By

Shuai Yuan
M.S., Emory University, 2010

Advisor : Zhaohui Qin, Ph.D.

Abstract of
a dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2014

Abstract

Applying Diploid Method to Improve Read-mapping and Analysis Based on NGS Data

By Shuai Yuan

Next generation sequencing (NGS) technologies have been applied extensively in genetics and genomics research. A fundamental problem when it comes to analyzing NGS data is accurately mapping short sequencing reads back to the reference genome. This important issue affects the interpretation and downstream analysis of the NGS experiments. Although plenty of read mapping algorithms and software have been developed, the majority of them uses the universal reference genome as a scaffold and do not automatically take into consideration the possibility of genetic variants. Ignoring the genetic variants information will cause a proportion of unmapped or incorrectly mapped reads, which affects the calculation, interpretation and analysis in many studies. Issues caused include the significant bias when detecting Allele-Specific Expression (ASE) from RNA sequencing data, low genotype calling accuracy, low Single Nucleotide Polymorphisms (SNPs) discovery rate and so on. Given that genetic variants are ubiquitous, it would be highly desirable if they can be factored into the read mapping procedure.

In our study, we developed a method that produces a personalized diploid reference genome based on all known genetic variants of that particular individual. We show that using such a personalized diploid reference genome with existing mapping software can improve mapping accuracy and significantly reduce the bias toward reference allele in ASE analysis.

By combining the imputation technology with reference genome personalization method, our studies, using real data, indicate further improvement in read mapping rate as well as genotype calling and SNPs discovery. Because many whole genome sequencing (WGS) studies are conducted on cohorts that have been previously genotyped using array-based genotyping platforms, we believe the strategy introduced here will be of high practical value to investigators working on WGS.

Our open source software is implemented as a standalone C++ code and has been integrated into Galaxy, a data intensive biomedical research platform, for pipeline visualization and better usability.

Applying Diploid Method to Improve Read-mapping and Analysis Based on NGS Data

by

Shuai Yuan
M.S., Emory University, 2010

Advisor : Zhaohui Qin, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2014

Acknowledgements

I would like to acknowledge many people for the help during my doctoral work.

First, I want to give my most earnest acknowledgment to my advisor Dr. Zhaohui Qin, who patiently guided me through my PhD research and dissertation. His guidance and encouragement have broadened my knowledge in bioinformatical field and stimulated my desire of applying advanced computational algorithms to interdisciplinary research. I would also like to thank my thesis committee members: Dr. Shun Yan Cheung, Dr. Yijuan Hu and Dr. Yun Li, who gave me many comments and suggestions in experiment design, paper writing, qualification, proposal and teaching. I would also like to thank all previous and current Qin group members for their stimulating conversation and excellent tutelage.

Second, I want to give special thanks to Dr. James Lu, for offering me this precious opportunity to join Emory.

Third, I am also very grateful for Li Li, Matthieu Maitre, Sang Choe, Shafiq Rahman and Naveen Thumpudi, who were my managers and mentors during my three internships at Microsoft, for helping me develop myself as a researcher from a different perspective.

Last, my thanks go to my wife Guibo Zhu, my son Jesse Yuan, my sister Weihua Yuan and my parents for their love and support over all the past years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Next Generation Sequencing

Since Frederick Sanger developed the chain-termination based DNA sequencing technology in 1977, it has been selected as the principle technology for the “first generation” laboratory and commercial sequencing applications for nearly three decades [1, 2]. After years of improvement and automation, Sanger sequencing technology became the major tool for the completion of the human genome project (HGP) [3] which cost about $2.7 billion and involved 18 countries during 11 years. This project published progressively improved versions of human reference genomic assembly (hg4∼hg15) in FASTA format (`http://en.wikipedia.org/wiki/FASTA_format`). Although HGP was completed, the developing of human reference genome is still continuing. The latest version is hg38 which was published for downloading on UCSC website in December 2013 (`https://genome.ucsc.edu`).

Not long after HGP was completed in 2003, Next Generation Sequencing (NGS) technologies were published in 2005 [4]. These new technologies, in contrast to its predecessor, are low-cost high-throughput sequencing technologies which massively parallelize sequencing analysis. Therefore, NGS is also called Massively Parallel Sequencing (MPS). It is practically supplanting the Sanger sequencing due to its ultra-high sequencing speed and ultra-low cost [5]. Since the emergence of NGS, the price dropping pattern of DNA sequencing has been more remarkable than that in Moore's Law [6].

SOLiD/Ion Torrent PGM from Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq from Illumina, and 454/GS FLX Titanium/GS Junior from Roche are the major NGS sequencer platforms and commercial vendors. Those platforms can generate millions of reads in a single run within hours. The output reads types can be DNA-seq [7], RNA-seq [8, 9], ChIP-Seq [10] or BS-Seq [11]. The length of generated reads ranges from 30 base pairs (bp) to 400 bp [12] , usually stored in a *de facto* standard FASTQ format [13] or Sequence Read Archive (SRA) format [14] from National Center for Biotechnology Information (NCBI). Reads can be single-ended or paired-ended, and they can be tagged as from sense strand, antisense strand or unknown strand.

NGS's high-throughput power has been harnessed by many researchers to investigate and solve genetical, genomical and clinical problems [8, 15–19]. Despite their diverse and wide range of categories and applications, most of them share a crucial problem: how to correctly and efficiently identify the genomic locations (chromosome, offset and strand) of the huge amount of NGS reads. This is a fundamental issue since read mapping affects many downstaiream analyses and results.

Read mapping problem is essentially an approximate string matching problem. It matches the reads against a reference genome, with biological specialties. However, the reads generated by NGS technologies are much shorter than Sanger method (typically 30∼400 bp in length) with much higher error rate (0.5∼1.0% error per raw base) [4, 20]. Such high error rates entail redundant sequencing of each base to distinguish sequencing errors from true polymorphisms, which greatly increases the computational and storage complexity.

In order to complete this task correctly and efficiently, numerous read-mapping tools and algorithms have been developed. Each one has its advantages and disadvantages by providing different trade-offs between speed and quality of the mapping, and can be applied in different situations. Most of them divide the mapping procedure into two phases: indexing and alignment. Indexing is to construct auxiliary indices for reference sequence or read sequences, or both, in order to accelerate the strings comparison. Although there are many indexing techniques [21], most-widely used ones belong to two categories: hashing and suffix/prefix tree.

Once the index is created, the alignment stage will use the built indices to approximately match each read against the reference genome to decide its potential source. Besides the nucleotide string, alignment algorithms may selectively take into account many additional elements, including but not limited to:

- Quality score of each base
- The maximum number of allowed mismatches, and their locations in the read
- Different types of mismatches (indels, repeats, reverses, swaps)
- Gapped alignment: gap numbers and lengths
- Multi-mapped reads: number of optimal and sub-optimal locations
- Distance between paired-ended reads

1.2 Read mapping tools

Based on those indexing and alignment algorithms, a lot of tools have been devised. BLAST (Basic Local Alignment Search Tool) is the first hash table indexing based local alignment search tool. It keeps each k-mer subsequences of the reads in a hash table and uses k-mers in the reference as the key to find exact matches and then extends around [22, 23]. SOAP (Short Oligonucleotide Analysis Package) adopts almost the same strategy except that it indexes the genome rather than reads. It also uses a special hashing technique called spaced seed [24]. CloudBurst [25] and GNUMAP (Genomic Next-generation Universal MAPper) [26] also use spaced seed but with different templates for different genomes. SeqMap [27] and MAQ [28] extend the method to allow more mismatches, at the cost of exponentially increased number of templates. BFAST (BLAT-like Fast Accurate Search Tool) [29, 30] proposed a concept of multiple-level hash indexing to reduce RAM requirement. It is able to perform a color-aware Smith-Waterman alignment.

MUMmer [31] and OASIS [32] are based on suffix tree, Vmatch [33] and Segemehl [34] on enhanced suffix array, and Bowtie [35], BWA [36], SOAP2 [37], BWT-SW [38] and BWA-SW [39] on Burrows-Wheeler transform (BWT). BWT is most widely used mainly because of its small memory footprint. QPALMA [40] and TopHat [41] were developed to solve spliced reads alignment like RNA reads. BSMAP (Bisulfite Sequence Mapping Program) [42] maps BS-seq using hashing technologies. BS Seeker [43] is a three-letter bisulfite sequence mapping procedure relying on Bowtie [35]. There are still a lot of read alignment and realignment tools unpublished, for more information about the mapping tools, readers are referred to these wonderful survey papers: [12, 21, 44].

1.3 Motivation

The output from sequence alignment tools are usually in SAM (Sequence Alignment/Map) format or BAM (Binary version of SAM) format [45], which are widely supported by alignment views including GBrowse [46], LookSeq [47], Tablet [48], BamView [49]. These alignment results are often quantified and can be used for many important tasks, such as detecting allele-specific expression (ASE), which is of great biological importance and can be used for *cis*-regulatory variant discovery and epigenetic imprinted region discovery [50–52].

However, the majority of current read mapping algorithms uses the universal reference genome and does not take into consideration the possibility of genetic variants. Ignoring genetic variants can cause severe problems. For example, when conducting ASE study using RNA-Seq data, there was a significant bias at heterozygous sites toward higher mapping rates. Degner tried to use a masked reference genome created with all known SNPs (Single Nucleotide Polymorphisms) to reduce the ASE bias, but that led to less reliable mapping results [53]. Another crucial problem for the successful application of NGS is variant detection and genotype calling at detected variant loci [5]. Some genetic variant callers such as SAMtools [45], GATK (Genome Analysis Toolkit) [54], VarScan [55] and BreakDancer [56] have been designed to perform such tasks. They all heavily rely on the correctness of read mapping results. Ignoring the generic variants already known will cause power loss in mapping rate, genotype calling accuracy and SNP discovery rate. Pickrell tried to increase the accuracy of gene variants analysis by creating a set of ethnicity-specific reference genome, but this method ignored the differences between individuals among the same population [57].

Currently, how to correctly interpret NGS data and to mine useful information remain challenging and are active areas of research. In this dissertation we try to solve some of these problems by constructing personalized reference genomes. This method, in our studies, shows apparent improvement in multiple aspects and we anticipate that it is highly useful for many future studies in bioinformatics.

Chapter 2

Read-mapping using personalized diploid reference genome for RNA sequencing data reduces ASE bias

Abstract

Next generation sequencing (NGS) technologies have been applied extensively in many areas of genetics and genomics research. A fundamental problem when it comes to analyzing NGS data is mapping short sequencing reads back to the reference genome. Most of existing software packages rely on a single uniform reference genome and do not automatically take into the consideration of genetic variants. On the other hand, large proportions of incorrectly mapped reads affect the correct interpretation of the NGS experimental results. As an example, Degner et al. showed that detecting allele-specific expression from RNA sequencing data was biased toward the reference allele. In this study, we developed a method that utilize DirectX 11 enabled graphics processing unit (GPU)'s parallel computing power to

produce a personalized diploid reference genome based on all known genetic variants of that particular individual. We show that using such a personalized diploid reference genome can improve mapping accuracy and significantly reduce the bias toward reference allele in allele-specific expression analysis. Our method can be applied to any individual that has genotype information obtained either from array-based genotyping or resequencing. Besides the reference genome, no additional changes to alignment algorithm are needed for performing read mapping Therefore, one can utilize any of the existing read mapping tools and achieve the improved read mapping result. C++ and GPU compute shader source code of the software program is available at: `http://code.google.com/p/diploid-mapping/downloads/`

2.1 Introduction

For diploid eukaryotic organisms, the maternally and paternally derived copies of most genes are expressed at similar levels. However, for some genes, the two alleles of an individual are expressed at different rates. This phenomenon is termed allele-specific expression (ASE). In recent years, much and increasing effort has been made to identify ASE genes since they present unique opportunities to study *cis*-regulatory variation [51, 58–62]

The newly emerged next generation sequencing (NGS) technologies have been increasingly recognized as an important and powerful tool for identifying ASE genes genome-wide, which improves our understanding about *cis*-regulatory variation. To identify ASE, one can conduct RNA sequencing (RNA-Seq) experiment [8, 63] to map all generated reads to the reference genome for all exonic SNPs that are known to be heterozygous, and then quantify the magnitude of expression of each allele by counting the number of times each allele is observed in reads that mapped to that locus. Despite its simplicity, systematic bias for read mapping may affect the accuracy of identifying ASE genes. This has been pointed out recently by Degner et al. [53]

Mapping short reads onto the reference genome is a fundamental problem in analyzing next generation sequencing (NGS) data and has been an area of intensive research in the past years. A wealth of successful software programs have been developed and enjoyed wide-spread usage in many different NGS applications such as MAQ [28], SOAP [64], BOWTIE [35], BWA [36], BFAST [29], mrFAST [65] and mrsFAST [66]. The details of these algorithms and plenty of other commonly-used read mapping software can be found in an excellent review paper [21].

Almost all of the existing read-mapping software relies on a universal reference genome—the National Center for Biotechnology Information (NCBI) human reference genome [67] which is derived from a small number of anonymous donors.

Although carefully annotated and maintained, this single reference genome cannot represent all the variants found in the general population. We know that each individual possess a unique set of genetic variants in hundreds of thousands that differ from the universal reference genome that distinguish him or her from others. Such wide-spread genetic variants compounded with non-ignorable sequencing errors and short read length caused a large proportion of reads unmapped or mapped to incorrect genomic locations. These mapping errors affect the interpretation of the NGS experimental results. As an example, Degner et al. showed that detecting allele-specific expression (ASE) from RNA sequencing data was biased toward the reference alleles because reads containing alternative alleles have less probability to align than reads that contains the reference allele. Therefore genes with a large amount of alternative alleles may be underestimated [53].

To reduce the impact of these genetic variants, Dewey et al. proposed to use ethnically concordant major allele reference genome sequence for read mapping [68]. Using estimated allele frequency data from the 1000 genome project [69], the authors developed three ethnically-specific major allele references for European, African and East Asian. When applied to four individuals from a nuclear family, Dewey et al. reported increased number of reads that mapped uniquely to the major allele reference genome than to the NCBI reference genome.

While much improvement is achieved using reference genomes that tailored toward the ethnical groups, it is important to note that there are still plenty of genetic variations at the individual level within each ethnical group. With the efforts such as the international HapMap project [70], the 1000 Genome project [69] and many others, we have accumulated and cataloged millions of known genetic variants, most in the form of single nucleotide polymorphisms (SNPs). In the past five years, the cost of array-based genotyping has declined sharply. As a result, for individuals or cell lines that we want to run RNA-Seq on, the genotypes of almost all common SNPs (minor allele frequency greater than 5%) are already known. In light of this, we believe that such information, whenever available should be incorporated into the process of read mapping.

In this study, we propose a novel method that utilizes all known genetic variant information of a particular individual and combine it with the NCBI reference genome to produce a "personalized" and diploid reference genome. We show that mapping against this personalized diploid reference genome will improve mapping accuracy and significantly reduce the bias toward reference alleles in allele-specific expression analysis. Our method can be applied to any individual whose genotype is known either from array-based genotyping or resequencing. Besides the reference genome, no additional change to alignment algorithm is required for performing read mapping therefore one can continue using any of the existing read mapping tools they like and achieve the improved read mapping result.

2.2 Methods

The goal of this project is to construct a personalized diploid reference genome using known genetic variants of an individual to reduce ASE bias. This reference genome can then be used for mapping reads generated from any sequencing assay conducted on this individual to improve the read mapping accuracy. There is no need to modify the read mapping software. Since genotypes are increasingly available and readily available, we believe incorporating such information in the read mapping step is important and beneficial. We have developed a software package available for public download that is able to achieve this goal conveniently.

2.2.1 Constructing personalized, diploid reference genome

In this study, we only consider SNPs, however our method can handle indels in a similar fashion. For better comparison with existing research results, we download universal NCBI reference genome (hg18.fa) from NCBI, which was used by Degner et al. [53], although our method can be applied to any version of universal reference genome including hg19.fa. To add alternative alleles, we go through each genotype stored in the individual's genotype file (usually in the VCF format) in parallel. A typical DirectX11 enabled graphics processing unit (GPU) usually has thousands of "Stream Processors" running on gigahertz level frequency, which is very suitable to perform such large amount of parallel computation. For a SNP that is homozygous wild type allele (identical to the reference allele), no action is taken; for a SNPs that is homozygous mutant allele, we edit the corresponding nucleotide in the reference genome sequence file; for a heterozygous SNP, we add a "mini chromosome" that is $w \geq 2k - 1$ bp in length where k is the read length and w can be specified by users. When $w > 2k - 1$ indels can be better detected. Suggested value of w is $2k - 1 + 2m$, where m is the maximum mismatches allowed during reads mapping. BWA, for example, sets the default value of m to 2 when the read length is 35 bp. The sequence of this "mini chromosome" is identical to the corresponding reference genome except at the middle position in which the alternative allele of that SNP is placed in. We name these "mini chromosomes" in a way such that their genomic locations can be easily identified.

Admittedly, adding these "mini chromosomes" may result in additional multiple mapping, however, with careful bookkeeping, such multiple-mapping incidences can be resolved post-hoc. If two SNPs are located near each other, i.e., with distance of d bp, where $d < k$, we use a slightly longer "mini chromosome", $(w + d)$ bp that cover both SNPs, and adding "mini chromosomes" with all possible combinations of covered SNPs (see Figure 2.1). More than two nearby SNPs can be handled in similar fashion. After this step, the personalized diploid reference genome contains tens of thousands of such mini chromosomes.

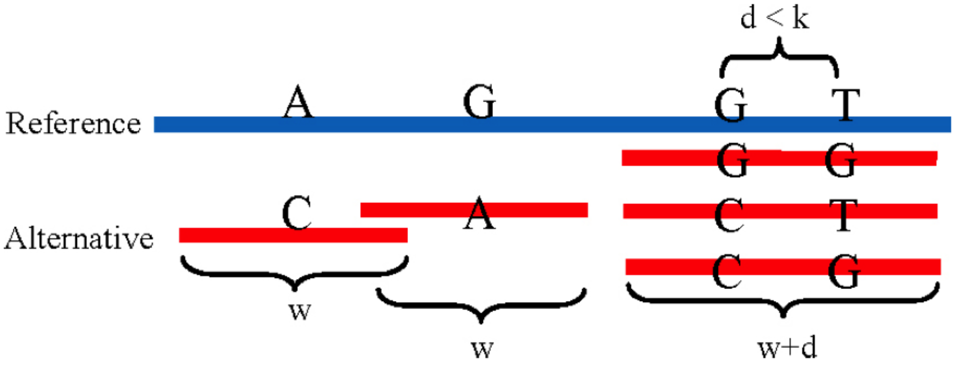


Figure 2.1: Examples of generated “mini chromosomes” at heterozygous SNPs. Most alternative chromosomes have the length of w. Heterozygous SNPs that are close to each other are merged into longer ones.

We choose not to simply add another set of whole chromosomes consisting with all the alternative alleles due to the following reasons: First, currently there is a limit of how large the reference genome can be handled by many existing read mapping software. Many mapping software have strict limitation on the length of the total reference sequence (mostly, 4G bp), because the data structure unsigned integer is defined in compilers as a 32-bit number ($2^{32} = 4G$) . Second, most of the genomic regions are homozygous, so it is not resource-efficient, and leads to many more multiple mapping incidences.

Our program can also accept an optional command line argument indicating the individual's gender. When this argument is set, for female individuals, we will exclude chromosome Y from the personalized reference genome, and chromosome X is treated the same as any other autosome.

2.2.2 Reads mapping

In this step, we perform read mapping using the personalized diploid reference genome instead of the universal NCBI reference genome. Although we use mapping software BWA v0.5.9 [36] with default parameters in this study, our pipeline scheme can accommodate any read mapping software

The raw output of the mapping step cannot be used directly because reads are mapped against a diploid reference genome that contains many "mini chromosomes". We take another step to process the mapping result such that reads mapped to "mini chromosomes" are correctly interpreted as mapped to the corresponding genomic location with the alternative allele present at the middle SNP. This step contains two parts: first, the correct genomic mapping locations are recovered; second, none-zero quality scores are assigned according to some confidence values. Figure 2.2 demonstrates the whole process of our pipeline.

Pipeline for mapping reads against diploid reference genome

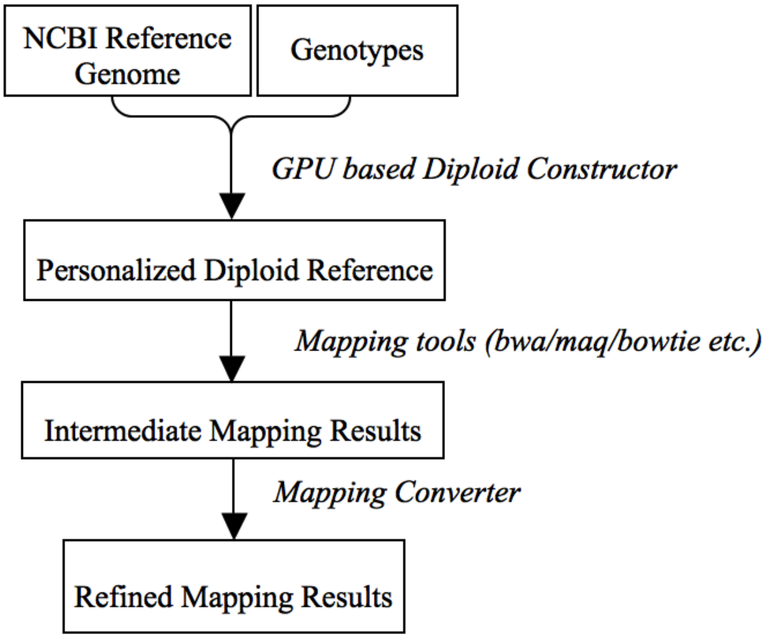

NCBI Reference Genome
Genotypes
GPU based Diploid Constructor
Personalized Diploid Reference
Mapping tools (bwa/maq/bowtie etc.)
Intermediate Mapping Results
Mapping Converter
Refined Mapping Results

Figure 2.2: 3-step pipeline for creating personalized diploid reference genome and mapping reads against it. *Diploid Constructor* takes NCBI reference genome and the individual's genotype file as input to create personalized diploid reference genome, which will then be used by multiple mapping tools to map reads. *Mapping Converter* converts intermediate mapping result to regular mapped file. For example, position "chr3b.l3843:5" is converted to "chr3:13847" (locations are 1-based).

2.2.3 An alternative method for reducing ASE bias

In addition to comparing with the common practice which is to use the universal reference genome for mapping, we also tested the masking strategy which has been used in the Degner et al. 2009 study. In this approach, all known SNP positions were "masked" prior to read-mapping. Masking was achieved by changing the nucleotide at each SNP location to one that differs from both the reference and alternative allele. The SNP locations were obtained by merging genotype files of 214 individuals defined in the 2007-03 version of the International HapMap Project (`http://hapmap.ncbi.nlm.nih.gov`). In order to prevent too strong binding of the bases on both alleles to a specific masking base, we randomly choose the masking base. For example, if the nucleotides at the SNP location on the reference and alternative allele are "A" and "G", the probability of using "C" or "T" as the mask are both 1/2.

2.2.4 Simulation studies

We conducted simulation studies to evaluate the impact on ASE bias and mapping quality when using the three competing mapping strategies: using the universal reference genome which is the *status quo*, using the masked universal reference genome which is introduced by Degner et al. 2009; and using the diploid personalized reference genome which we propose. In order to represent the diversity of human population and investigate its impact on the results, we selected three individuals from the HapMap panel, one Caucasian from CEPH (NA12865), one African from YRI (NA19238) and one Asian from CHB (NA18621). For each individual, we downloaded the individual's genotype information (2007-03 version) from the International HapMap Project. As Degner et al. did in their study, we randomly inserted sequencing errors on reads generated. We tested three different sequencing error rates: 0, 0.01 and 0.05. When simulating reads, we choose the sequencing read length to be 35 bp and 100 bp, and then randomly sample DNA fragments across the whole diploid reference genome except chromosome X and Y. We only keep reads that cover at least one heterozygous SNP. For each of the three sequencing error rates, 2 million reads were generated. Either reference or alternative allele was selected with equal probability thus assume balanced allele specific expression. To create the masked reference genome, all SNPs identified from the 214 individuals in the International HapMap Project (genotype information obtained from the 2007-03 version) are masked. In order to increase the precision with more mapped reads, we consider SNPs located in both exons and introns.

2.2.5 Real data studies

We analyzed two sets of RNA-Seq data: one is the one studied in Degner et al. 2009, the other is 68 individuals from Pickrell et al 2010 [57]. Just like in the simulation studies we also use three different read-mapping strategies. Here we only consider SNPs located within exons.

We analyzed two aspects of the performance of different mapping strategies: mapping bias towards reference alleles and total number of reads that are successfully mapped. For the first, at each SNP locus, we first decide the number of reads that cover the SNP. After filtering out SNPs with too shallow mapping depth (this step is optional). In this study, we use threshold of five reads. We then count the number of reads that match the reference allele and the alternative allele respectively. For the second, we want to maximize the number of RNA-seq reads that can be mapped successfully. Therefore, a mapping strategy that can produce more mapped reads with high accuracy is preferred. In the simulation study, because we know each read's true location, we can compare the number of reads that are correctly mapped back to their true locations; for real data, because reads' true locations are unknown, we compare the number of reads that are successfully mapped.

2.3 Results

2.3.1 ASE bias in simulation studies

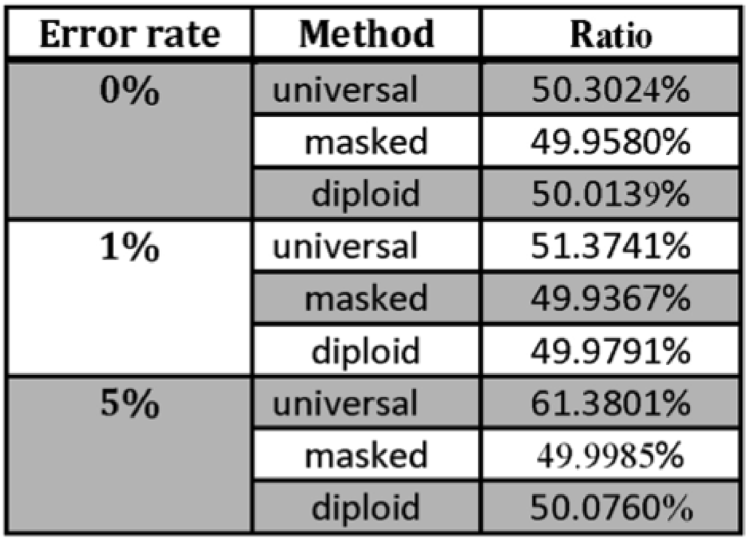
The most important statistic that measure ASE bias is the proportion of reads that mapped to the reference allele. Simulation studies showed that the ratios of reads mapped to reference alleles are very close to the theoretical value—50% for both diploid and masked genome methods, regardless of the error rate, whereas conventional method yielded upward bias towards the reference alleles and the bias increase with the error rate. This indicates that both methods yield much reduced bias at all error rates. Even assuming no sequencing error, universal reference genome method is still suffering from inherent bias. The same pattern was observed on all three HapMap samples that represent different ethnic groups. We also found that increasing read length resulted in more bias. Table 2.1 shows the results for individual YRI NA19238.

To better understand the magnitude of the bias and the impact of sequencing errors, we plotted the distribution of proportions of reference alleles obtained using different read mapping strategies (Figure 2.3). We found that when using universal reference genome method, the proportions of reference allele in majority of the SNPs are greater than 0.5. This asymmetry caused the mapping bias towards reference alleles. The asymmetry also increases dramatically as the error rate increases. However, neither diploid nor masked genome method shows apparent asymmetry.

2.3.2 Mapping accuracy

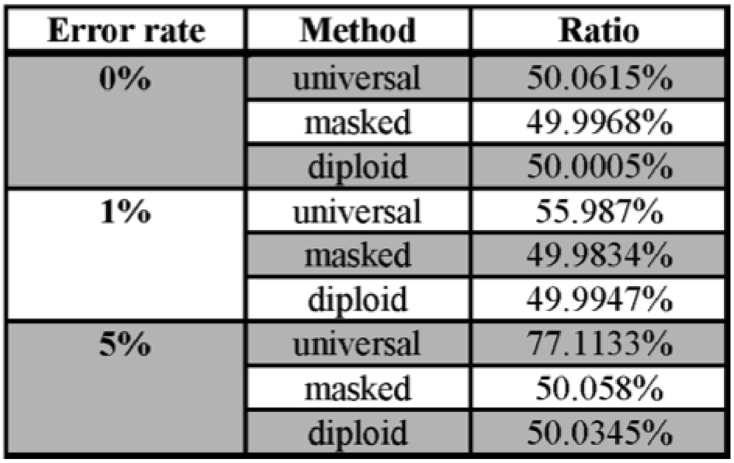
The percentage of correctly mapped reads is an important measure when evaluating mapping strategies. Although masking the reference allele in the universal

A ***Ratio of reads mapped to the reference alleles***
Read length=35 bp, max mismatch=2



Error rate	Method	Ratio
0%	universal	50.3024%
	masked	49.9580%
	diploid	50.0139%
1%	universal	51.3741%
	masked	49.9367%
	diploid	49.9791%
5%	universal	61.3801%
	masked	49.9985%
	diploid	50.0760%

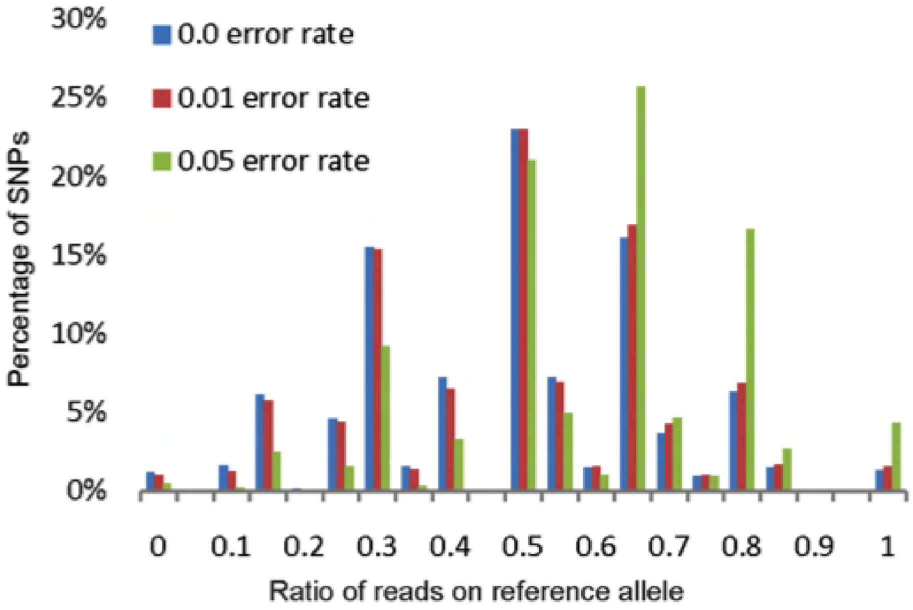
B ***Ratio of reads mapped to the reference alleles***
Read length=100 bp, max mismatch=2



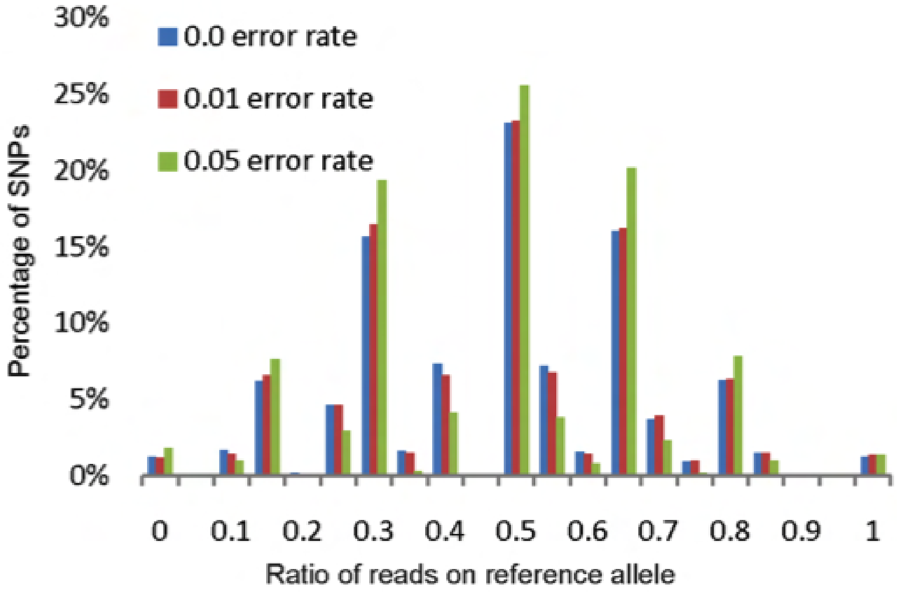
Error rate	Method	Ratio
0%	universal	50.0615%
	masked	49.9968%
	diploid	50.0005%
1%	universal	55.987%
	masked	49.9834%
	diploid	49.9947%
5%	universal	77.1133%
	masked	50.058%
	diploid	50.0345%

Table 2.1: Simulation results show that universal reference genome method suffers from serious bias towards reference alleles. This bias increases dramatically with the increment of error rate. However, the masked and diploid genome methods do not have apparent bias toward either reference or alternative alleles while error rate does not have influence on the ratio. (**A**) Read length is 35 bp and maximum mismatch is set to 2. (**B**) When the read length increases to 100 bp, ASE bias will also increase considerably given the same sequencing error rate.

A *Percentage of SNPs corresponding to the reference allele using universal reference genome*



B *Percentage of SNPs corresponding to the reference allele using masked reference genome*



C *Percentage of SNPs corresponding to the reference allele using personalized diploid reference genome*

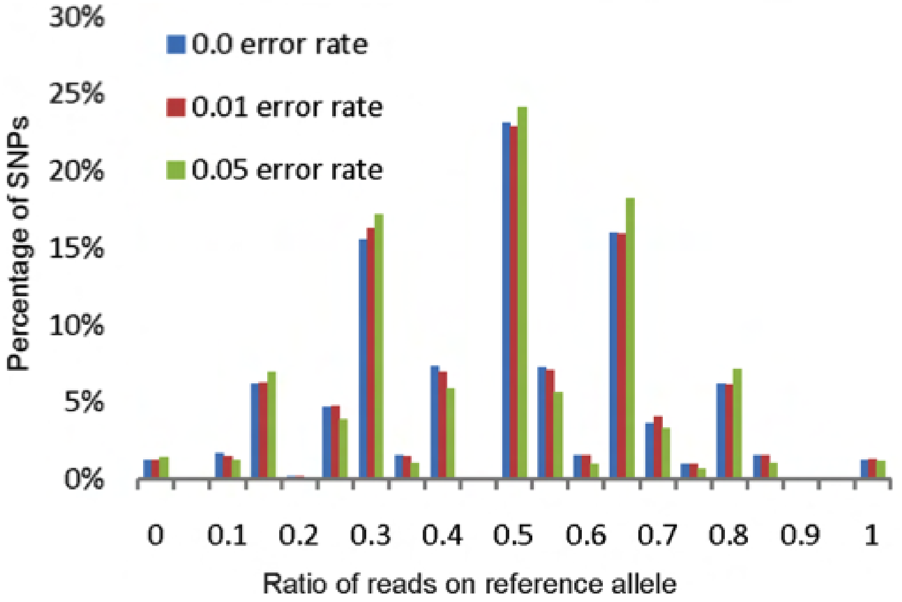


Figure 2.3: Distributions of reference allele proportions for SNPs tested in the simulation study (required read coverage depth > 5). The distributions spread across 0 to 1 due to randomness of sampling. (**A**) Using universal reference genome. (**B**) Using masked reference genome. (**C**) Using personalized diploid reference genome.

reference genome reduces ASE bias, we found this strategy produces less reliable mapping result [53] and significantly lower overall mapping success rate, especially when moderate sequencing error is present. Our simulation shows that when the sequencing error rate reaches 0.05, the diploid genome method can correctly map 25% more reads compare to masked genome method. This significant improvement suggests that the diploid genome method has a higher mapping success rate overall. Figure 2.4 shows the mapping success rates from the three methods when mapping 2 million reads with different error rates.

2.3.3 Real data analysis on ASE bias and mapped reads

We reanalyzed the real RNA-Seq data presented in Degner et al. 2009 to compare the levels of ASE detection bias resulted from different mapping strategies. We also compared the number of reads that were successfully mapped to cover heterozygous exon SNPs using the three read mapping strategies. Figure 2.5 shows the results. From the figure, we observe the same pattern as in the simulated data: using either masked reference genome or the personalized diploid reference genome vastly reduce ASE bias, while our method resulted in much higher success rate of read mapping than using the masked reference genome.

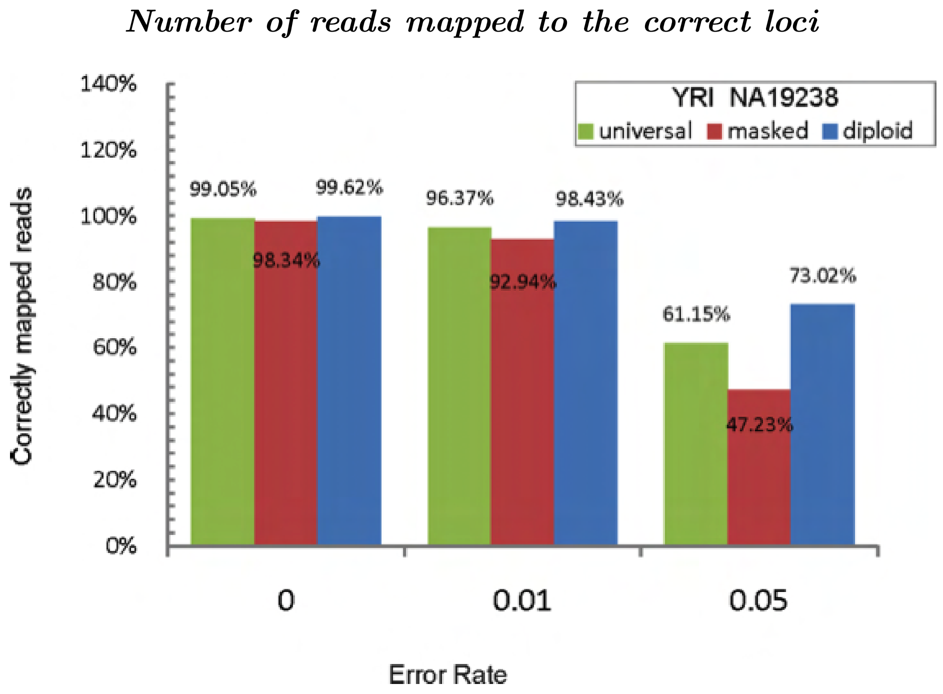
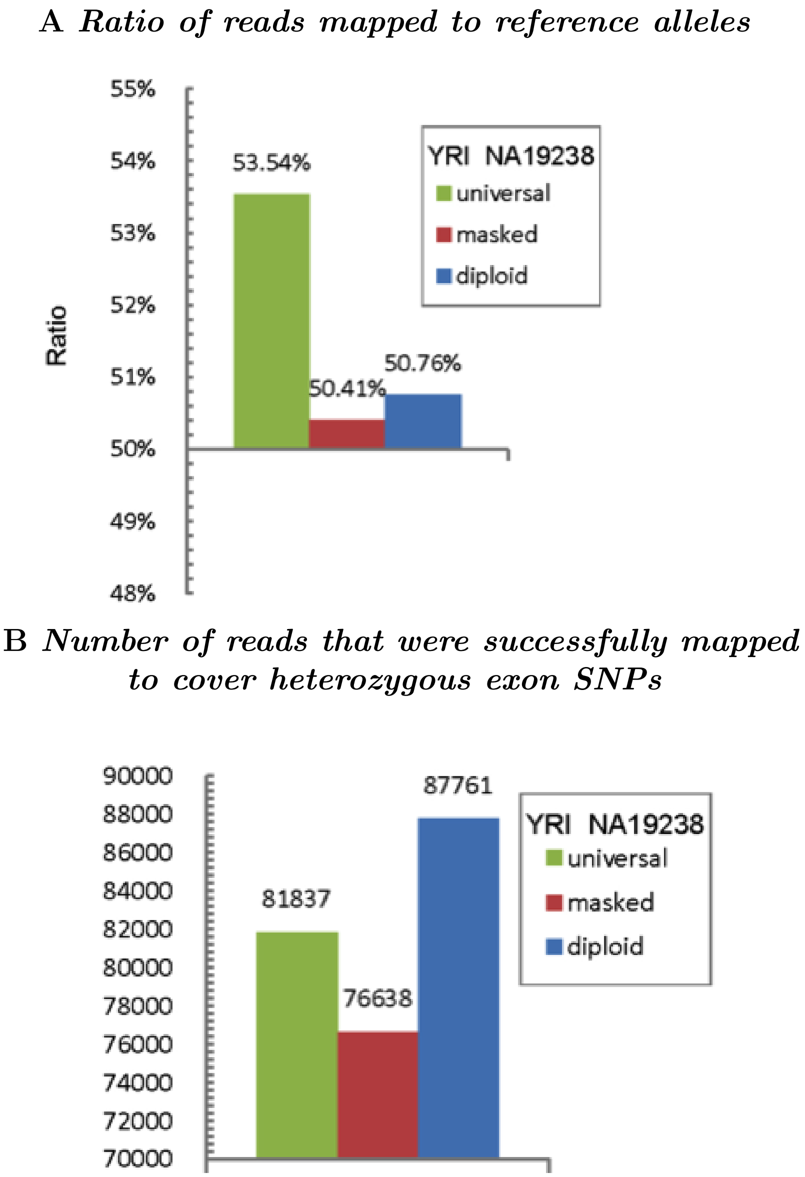


Figure 2.4: Simulation results of individual YRI NA19238 show that diploid genome method can improve mapping quality. The diploid genome method shows the highest correctness of mapping results among three methods. Universal reference genome method, although it has mapping bias, shows better mapping quality than masked genome method.



A ***Ratio of reads mapped to reference alleles***

B ***Number of reads that were successfully mapped to cover heterozygous exon SNPs***

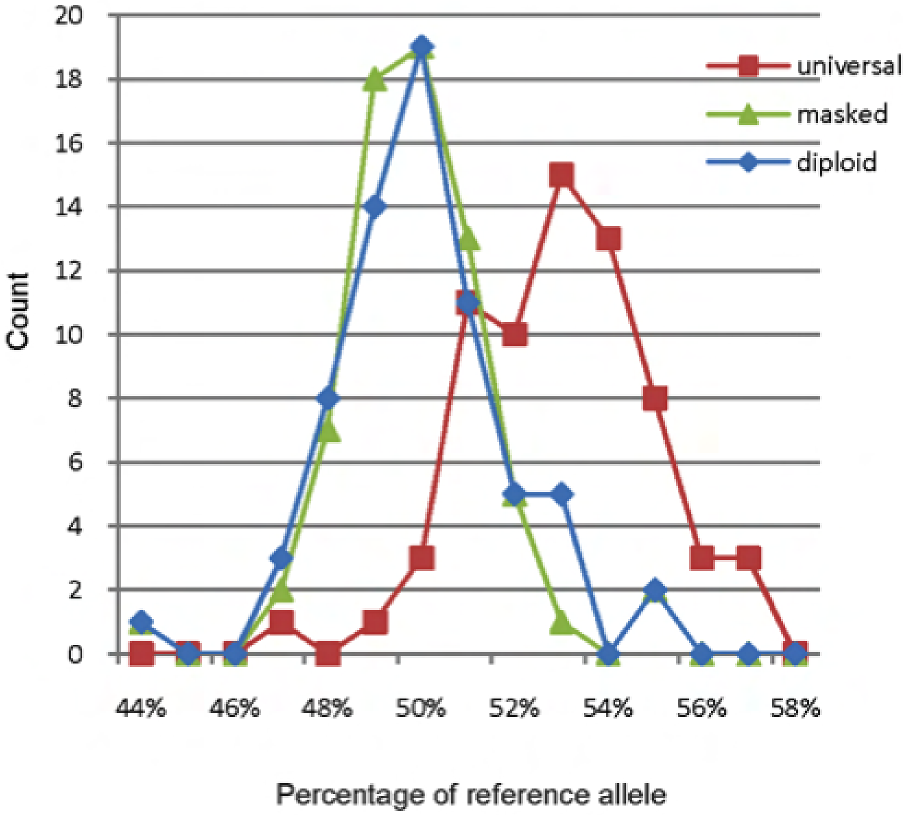
Figure 2.5: (**A**) Results derived from running three different methods on real data show that diploid genome method can effectively reduce bias towards reference alleles. Although masked genome method can also reduce such bias, it might be over-reduced because of its unreliable mapping result. (**B**) Number of reads that cover heterozygous exon SNPs. This figure shows that masked genome method loses 12.7% of successfully mapped reads compare to diploid genome method. Therefore, it can be inferred that masked genome method is problematic.

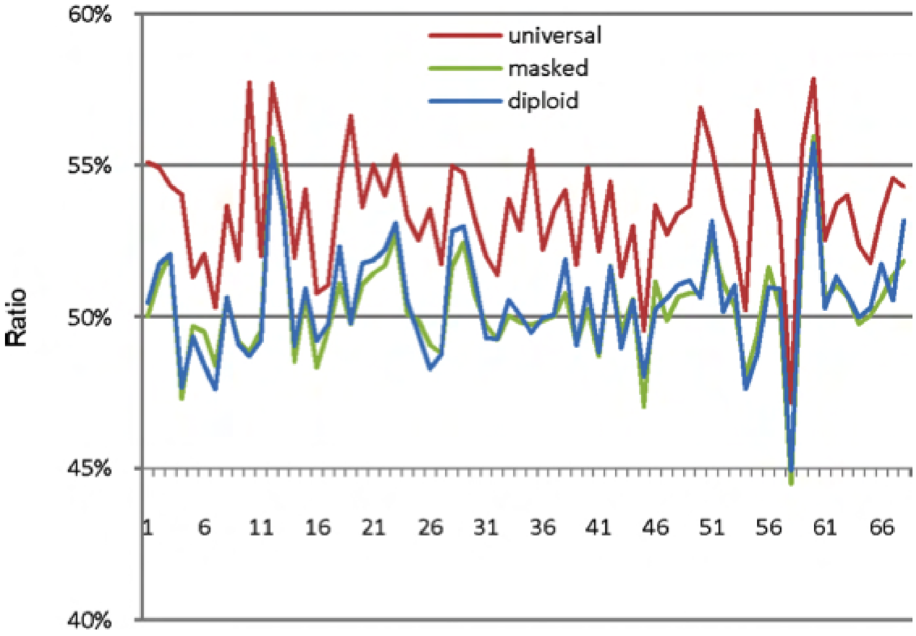
2.3.4 More real data results

To verify that our new method can reduce ASE bias in general population, i.e., individuals whose genotype is known, we conducted experiments on a set of individuals with real data from a recent study of Pickrell et al 2010 [57]. We download the entire dataset from `http://eqtl.uchicago.edu/RNA_Seq_data/`, and then select 68 individuals whose genotype can be found in the 2007-03 version of the International HapMap Project. Figure 2.6 shows the distribution of the ratio of mapped reference allele. Again we observe using masked reference genome or personalized diploid reference genome reduce bias towards reference allele.

As ASE is widespread across heterozygous SNPs the P-values yield from binomial test must also display this enrichment and its impact to the expression bias. Figure 2.7 shows the QQ-plot of the P-values across quantiles. When using universal reference genome, we see that the two curves representing “reference more” and “reference less” respectively are far apart, indicating bias. For the other two methods: using masked reference genome or personalized diploid reference genome, the bias essentially disappeared.

A ***Distribution of reference allele proportions***





B ***Reference allele proportion across 68 individuals***

Figure 2.6: The counts of individuals at each expression ratio and each individual's reference. (**A**) The reference allele ratios for all individuals are in the range [44%, 58%). For diploid and masked genome method, there are 19 individuals located in the ratio region [50%, 51%), which is the peak of their curves. The universal reference genome method, however, caused the shift of the peak to the ratio region [53%, 54%) for 15 individuals. (**B**) Using universal reference genome method always produces higher reference ratio compare to diploid and masked genome method, which indicates that universal reference genome method will inevitably introduce bias towards the reference alleles.

QQ-plot for binomial P-value across SNPs for NA18505 with mapping depth > 10

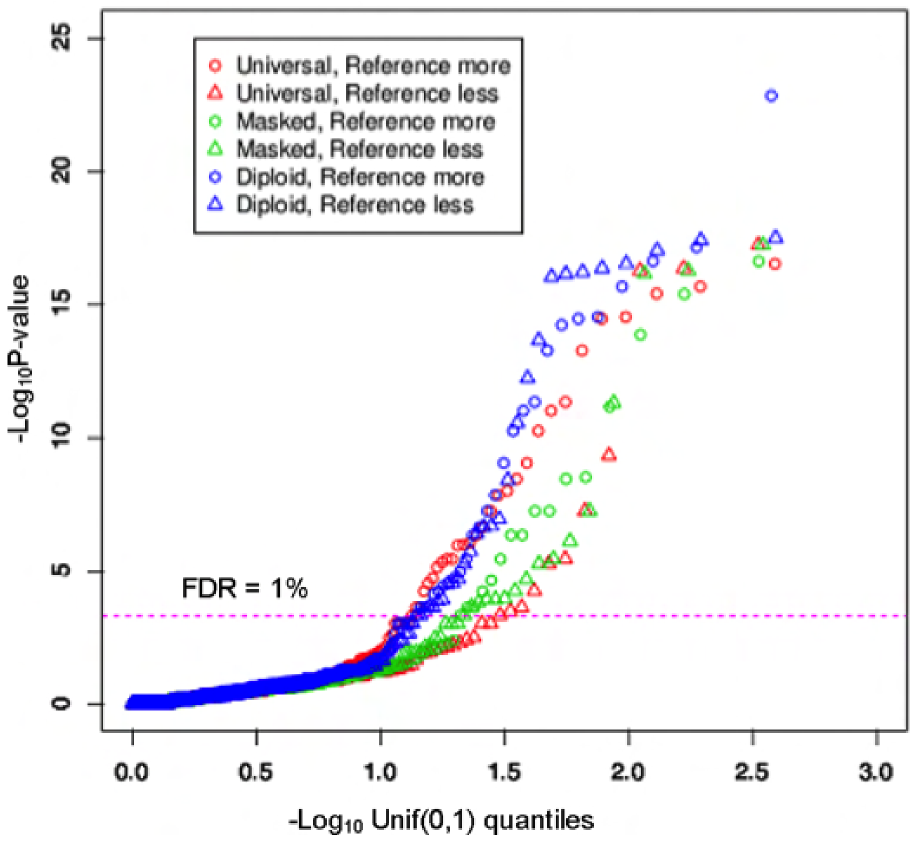


Figure 2.7: QQ-plot of P-values for one-sided binomial tests for heterozygous SNPs which are categorized into more or less expression for reference alleles than alternative alleles. The dotted line is the threshold FDR=1%.

2.4 Discussion

ASE offers biological insights from understanding transcription regulation to disease susceptibility. Detecting ASE from RNA-Seq data has become an increasingly important topic for genetics and genomics researchers. As pointed out by Degner et al., the current read mapping strategy produces "a significant bias toward higher mapping rates of the allele in the reference sequence, compared with the alternative allele." [53], therefore, it is of great importance to develop alternative strategy to reduce ASE bias. In this study, we proposed a novel strategy that utilizes known personal genotype information that is increasingly available in the post-genomic era. In our method, we first construct a personalized diploid reference genome using available genotype information, and then use the constructed reference genome with a regular existing read mapping software such as BWA to map reads generated from the RNA-Seq experiment. Using both simulated data and real data, we showed that our strategy can effectively reduce the ASE bias, and increase the success rate of read-mapping. We believe our method provides an attractive solution to ASE detection using RNA-Seq data.

The drawback of using the universal NCBI reference genome in read mapping has been noticed in the literature. Dewey et al. developed three ethnicity-specific major allele reference genomes for European, African and East Asian based on HapMap data and use that for read mapping. They reported improved genotyping accuracy using this synthetic reference genome [68]. In this study, we went further by constructing reference genome that is “personalized”, i.e., taking into account of known genotype information of that particular individual. With the rapid dissemination and declining cost of array-based genotyping technologies, genotypes of millions of SNPs are routinely available. Thus our method is widely applicable. Our strategy is developed independently of that of Vijaya Satya et al. 2012 [71]. Despite many similarities between the two methods, there are some notable differences: our personalized reference genome is able to accommodate indels in addition to SNP markers; we use a Mapping Converter to convert the reads mapped to alternative alleles to correct genomic positions. we have tested our strategies on a much larger datasets to examine the population-level of the performance improvement; we have tested the performance of our method on longer read (100 bp) and found even better result in reducing ASE; we also implemented the construction of the personalized diploid reference genome using GPU compute shader code to improve the computation speed.

Using our software program, constructing a personalized diploid reference genome from a dense genotyping file only takes about 10 minutes on a commodity computer (Intel Core 2 Duo CPU, AMD Radeon HD 6900 GPU, 8GB memory), which seems a small price to pay for enhanced read mapping result.

Chapter 3

Using personalized diploid reference genome to improve read mapping and genotype calling in DNA sequencing studies

Abstract

With rapid decline of the sequencing cost, researchers today rush to embrace the whole genome sequencing (WGS), or the whole exome sequencing (WES) approach as the next powerful tool for relating genetic variants to human diseases and phenotypes. A fundamental step in analyzing WGS and WES data is mapping short sequencing reads back to the reference genome. This is an important issue, because incorrectly mapped reads affect the downstream genotype calling and association analysis. Although many read mapping algorithms have been developed, the majority of them uses the universal reference genome and does not take into consideration the possibility of genetic variants. Given that genetic variants are ubiquitous, it is

highly desirable if they can be factored into the read mapping procedure. In this work, we developed a novel strategy that utilizes genotypes obtained *a priori* to customize the universal haploid reference genome into a personalized diploid reference genome. The new strategy is implemented in a program named RefEditor. When applying RefEditor to real data, we achieved encouraging improvements in read mapping, genotype calling and SNP discovery. Compared to standard approaches, RefEditor can increase genotype calling accuracy by 10-40% and reduce Mendelian inconsistency by 10-30% across various sequencing depths. Because many WGS and WES studies are conducted on cohorts that have been genotyped using array-based genotyping platforms previously or concurrently, we believe the proposed strategy will be of high value in practice, which can also be applied to the scenario where multiple NGS experiments are conducted on the same cohort.

3.1 Introduction

Mapping short reads onto the reference genome is a fundamental step in analyzing next generation sequencing (NGS) data and has been an area of intensive research in the past years. A wealth of successful software programs for mapping short reads, such as MAQ [28], SOAP [64], SOAP2 [37], BOWTIE [35], BOWTIE2 [72], BWA [36], BFAST [29], mrFAST [65], mrsFAST [66], NovoAlign (`http://novocraft.com`), SHRiMP [73], and STAR [74], have been developed and enjoyed wide-spread usage in many different NGS applications (e.g., whole genome sequencing (WGS) [75], whole exome sequencing (WES) [76], Chromatin Immunoprecipitation sequencing (ChIP-seq) [77–79] and transcriptome sequencing or RNA-seq [63]). The details of these programs can be found in excellent review articles [21,44]. Despite the vast differences in algorithms and indexing methods, almost all of the existing read-mapping programs rely on the universal haploid reference genome—the National Center for Biotechnology Information (NCBI) human reference genome [67] , which was derived from a small number of anonymous donors. At any multi-allelic position, a presumed consensus allele is used. Although carefully annotated and maintained, this single reference genome is not intended to represent all the variants found in the general population. Indeed, the human genome is diploid, and each individual possesses a unique set of genetic variants at millions of loci that distinguish him or her from others. Such wide-spread genetic variants, compounded with non-ignorable sequencing errors and short read length, cause a large proportion of reads to be unmapped or mapped to incorrect genomic locations. These mapping artifacts sometimes lead to misinterpretation of the NGS experimental results, such as the overstating the incidence of Allele Specific Expression [53] and affecting regulatory element identification at heterozygous variants [80].

Notably, genotype information is often available for samples that are undergoing NGS experiments. There are at least three scenarios in which the genotypes are available. First, many WGS or WES studies were conducted on samples that have been studied in the previous wave of genome-wide association studies (GWAS). These samples have already been genotyped by one of the array-based high-density genotyping platforms such as those from Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA) [81] . Comprehensive assessment of array-based genotyping platforms can be found in the review article [82] . Second, many NGS experiments were conducted on well-established cell lines such as HeLa and IMR90, whose genotypes have also been profiled using array-based genotyping platforms. Third, and more often, array-based genotyping and multiple NGS-based experiments such as RNA-seq, ChIP-seq and resequencing were conducted on the same samples in the same study [83].

Using array-based genotyping, we will be able to collect genotype information on a large proportion of common genetic variants. Aided by powerful genotype imputation techniques, such as MaCH [84, 85], MaCH-Admix [86] , IMPUTE [87] , IMPUTE2 [88] , Minimac [89] and BEAGLE [90] , we will gain substantial additional genotype information on genetic variants that are not found on the genotyping array but are included on one of the dense reference haplotype panels such as those from the 1000 Genomes Project [75] . All of the aforementioned imputation methods exploit the linkage disequilibrium between observed and unobserved SNPs to infer the genotype of unobserved SNPs.

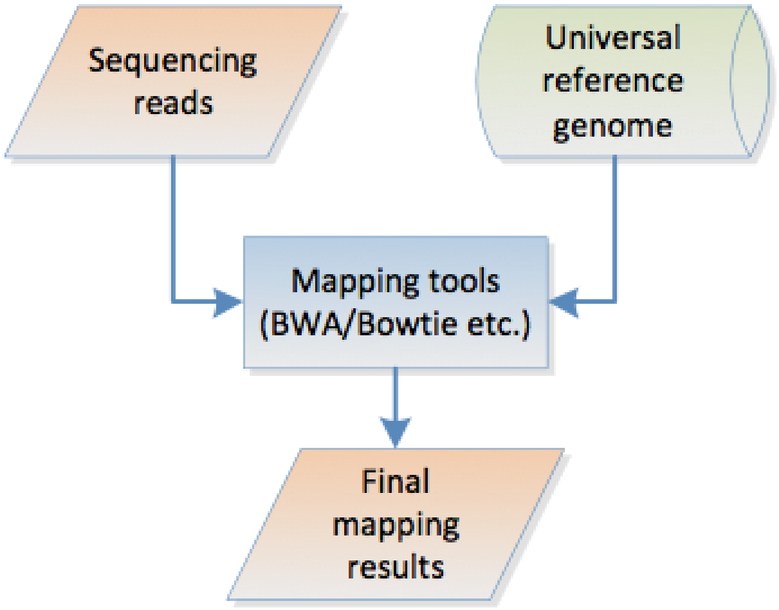
We believe that the substantial pre-existing genotype information, whether assayed or imputed, can be and should be utilized to fine tune the reference genome to reflect the unique features of each individual genome. An accurate reference genome sequence will lead to improved read mapping and consequently improved SNP discovery and genotype calling.

Here we present RefEditor, a software package developed to improve read mapping by customizing the universal haploid reference genome to reflect individual genetic variation. It contains two components, RefEdit and RefEdit+, both converting the universal reference genome into a personalized diploid reference genome. RefEdit uses the assayed genotypes only whereas RefEdit+ adopts an additional step to augment the assayed genotypes by imputation. The basic scheme of RefEdit and RefEdit+, as well as the comparison between standard read mapping process and the proposed read mapping process, are illustrated in Figure 3.1. Both RefEdit and RefEdit+ contain two main components: Diploid Constructor and Mapping Converter. Diploid Constructor converts the universal haploid reference genome to the personalized diploid reference genome by supplementing the universal reference chromosomes with short sequences containing alternative alleles (Figure 3.1[B] and Figure 3.1[C]). Mapping Converter modifies intermediate results of read alignment in SAM (Sequence Alignment/Map) format [45] by translating mapped locations on customized, diploid reference genome back to its genomic locations on the regular reference genome and reassigning mapping quality scores (Figure 3.1[B] and Figure 3.1[C]). Diploid Constructor and Mapping Converter are called upon before and after executing the read alignment tools, respectively.

3.2 Material and Methods

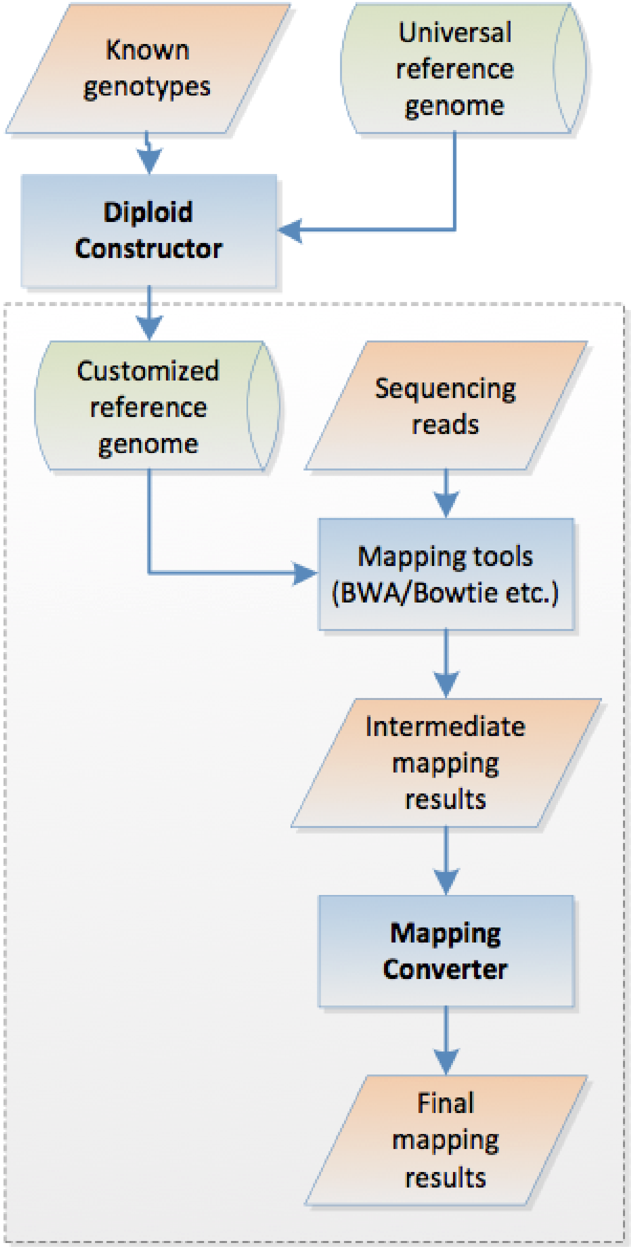
3.2.1 RefEdit+ Pipeline

The main objective of this project is to construct the personalized diploid reference genome using pre-existing genotype information of an individual, which is typically stored in a Variant Call Format (VCF) file (`https://github.com/samtools/hts-specs`). This reference genome can then be used for mapping reads generated from any sequencing assay conducted on this individual to improve the read mapping accuracy. There is no need to modify the read mapping software itself. Since genotype information is increasingly available from more and more array-based genotyping and sequencing experiments, we believe incorporating such information in the read-mapping step is important and beneficial. This goal can be conveniently achieved

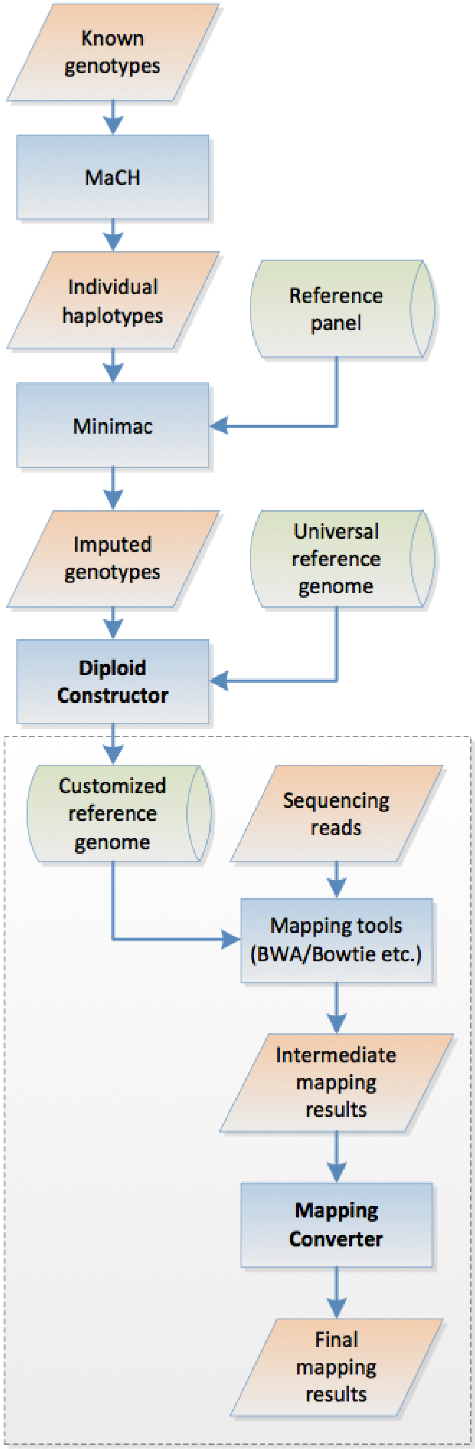


(A) ***Read mapping using universal reference genome***

Figure 3.1: The pipeline for imputation, diploid reference genome construction and read mapping. (**A**) Traditional read mapping method. (**B**) RefEdit read mapping strategy that incorporates known genotypes. (**C**) RefEdit+ read mapping strategy that incorporates both assayed and imputed genotypes. The parts inside dashed line boxes are identical for RefEdit and RefEdit+ methods


Known genotypes
Universal reference genome
Diploid Constructor
Customized reference genome
Sequencing reads
Mapping tools (BWA/Bowtie etc.)
Intermediate mapping results
Mapping Converter
Final mapping results

(B) *Read mapping using RefEdit.*



(C) *Read mapping using RefEdit+.*

with RefEdit and RefEdit+, with the later contains an additional imputation step to augment the existing genotypes set. The RefEdit+ pipeline consists of the following steps:

Step 1. Genotype imputation

In order to increase genotype information that can be used to customize the reference genome, we turn to the genotyping imputation techniques that have been developed in the past five years and showed great success in finding association of untyped SNPs and disease phenotype in many GWAS studies [91, 92]. In this study, we used MaCH version 1.0 [84] and Minimac [89] programs to perform genotype imputation. Default parameters are used for MaCH and Minimac throughout this pipeline. We use population-specific reference panels from the 1000 Genomes Project [69] which contains 25,802,094 SNPs for Yoruba in Ibadan, Nigeria (YRI) and 17,076,866 for Utah residents with ancestry from northern and western Europe (CEU). We use Rsq threshold of 0.7 for imputation quality control to balance the number of qualified genotypes and quality of imputation.

Step 2. Add alternative alleles (genotyped and imputed) to the reference genome

Next, we combine genotyped and imputed genotypes and use them to modify NCBI reference genome 37.1 (HG19 reference) to create a new personalized diploid reference genome. This step is achieved by using the program Diploid Constructor contained in the RefEditor software package. This new reference genome can be fed into any existing mapping tool in the exact same way as the universal reference genome. During the construction process, no action is taken at loci where genotypes are homozygous wild type (reference allele); at loci where genotypes are homozygous mutant alleles we edit the corresponding nucleotides in the reference genome sequence file; at heterozygous loci we add a mini chromosome of length $w \geq 2k - 1$ base pairs (bp) where k is the read length. Users can specify their own w. When $w > 2k - 1$ indels can be better detected at the cost of longer read mapping time. Suggested value of w is $2k - 1 + 2m$, where m is the maximum allowed indels during read mapping. In all studies presented here, read length k is 36, we set m to be 2 which is the default indel length used by BWA for read length 36. The sequence of this mini chromosome is identical to the corresponding segment of the universal reference genome except at the middle position in which the alternative allele of that SNP is placed. If two genotypes are located near each other, i.e., with distance of d bp, where $d < k + m$, we create mini chromosomes of all possible combinations of haplotypes that can possibly be covered by a read at the given read length. For other imputed variants like indels, we modify corresponding mini chromosomes to reflect such type of mutations. Those mini chromosomes are concatenated to the end of each traditional chromosome defined in the reference file, with a sequence of “N”s of

$m + 1$ in length to separate them. An auxiliary file is created to record the genomic location of these mini chromosomes. We could let these "mini chromosomes" to stand alone. The reason we choose to ligate them with the original ones is to ensure pair-end read mapping function to work properly because many mapping tools check whether the two ends map to the same chromosome.

The construction of the personalized reference genome is illustrated in Figure 3.1[B] and Figure 3.1[C]. RefEditor can also accept an optional command line argument indicating the individual's gender. When this argument is set for female individuals, chromosome Y will be excluded from the personalized reference genome. Using RefEdit, only non-ref/ref genotypes identified by the genotyping array will be incorporated, whereas using RefEdit+, all non-ref/ref genotypes identified from either the genotyping array or imputation will be incorporated.

Step 3. Read mapping using customized diploid reference genome

The customized diploid reference genome can be treated the same as the universal reference genome and used by almost all existing read mapping software. For this study, we use BWA v0.5.9 [36] with default parameters for its high performance on short reads mapping. The raw output of the mapping step needs to be post-processed such that reads mapped to those mini chromosomes are correctly interpreted as mapped to the corresponding genomic locations. The mapping quality scores will also be reassigned according to the Phred-scaled probability of mismatches between the read and reference [93]. This step is achieved by using the program Mapping Converter contained in RefEditor. The parts inside dashed line boxes in Figure 3.1[B] and Figure 3.1[C] illustrate the read mapping using customized diploid reference genome and post-processing using Mapping Converter.

Step 4. SNP finding and genotype calling

Genotypes are called from the reads successfully mapped with positive mapping quality found in sorted BAM format file. We use the Genome Analysis Toolkit (GATK) [54] to call genotypes. GATK is a widely used software package for detecting SNPs and calling genotypes from single or multiple samples. It takes into account the quality scores of each base in the mapped reads. The output from GATK will be filtered to only keep SNPs.

3.2.2 Competing read mapping strategies

Various strategies have been developed for dealing with sequence variants in read mapping. Here we briefly review other competing methods.

Ethnicity-specific major allele reference genome

In a recent study, Dewey et al. pointed out that the major alleles at many genomic loci are different among populations [68] . Given this, Dewey et al. developed a novel strategy that creates a set of ethnicity-specific reference genomes, including

European, African and East Asian. In these reference genomes, the allele that is most frequent among that particular population is used at polymorphic loci, resulting in around 1.5 million modifications in each population compare to the universal reference genome [68] . Read mapping is then performed against these ethnicity-specific major allele reference genomes. Dewey et al. showed that in real studies, using the ethnicity-specific reference genome results in improvement of genotype calling accuracy for disease-associated variant loci [68].

GSNAP

GSNAP (Genomic Short-read Nucleotide Alignment Program) uses universal reference genome and all SNPs from dbSNP in mapping. It also uses its own mapping algorithm based on hash tables generated from sampled k-mers from reference genome [94]. GSNAP considers all possible genotypes while still maintains running speed comparable to other existing read-mapping software, which impact the mapping results of 7-8% transcriptional reads although it does not significantly increase mapping success rates [94].

3.3 Results

3.3.1 An example

In Figure 3.2, we illustrate how including known genotypes improves the read mapping quality and SNP calling accuracy in a specific case using the sequence data from the 1000 Genomes Project. At the locus chr1:154568665, the reference allele is A. The sequencing read containing the alternative allele G at the locus can be successfully mapped to the personalized diploid reference genome with two mismatches. By contrast, this read fails to map to the universal reference genome because there are three mismatches, which exceeds the limit adopted by most mapping tools for this read length. Downstream 18 bp at the locus chr1:154568683, multiple mapped reads show the same type of mismatch, suggesting that there might be a new SNP at that locus. The alternative allele G is not known *a priori*. This new SNP is verified by gold standard genotype calls based on Complete Genomics Inc. (CGI) deep sequencing data.

3.3.2 Performance comparison study design

We conducted a series of studies using real data to evaluate the performance of RefEdit and RefEdit+ for read mapping, variant finding and genotype calling. In the first study, we focused on the mapping success rates, genotype calling accuracy and

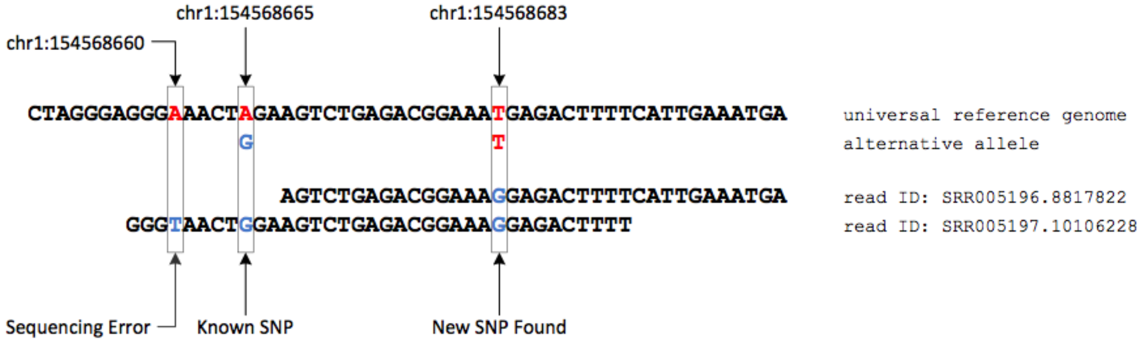

chr1:154568665
chr1:154568683
chr1:154568660
CTAGGGAGGGAAACTAGAAGTCTGAGACGGAAATGAGACTTTTCATTGAAATGA
universal reference genome
G
T
alternative allele
AGTCTGAGACGGAAAGGAGACTTTTCATTGAAATGA
read ID: SRR005196.8817822
GGGTAACTGGAAGTCTGAGACGGAAAGGAGACTTTT
read ID: SRR005197.10106228
Sequencing Error
Known SNP
New SNP Found

Figure 3.2: An example of how our RefEdit method can find new genotypes from known genotypes. The maximum mismatch threshold is set to 2 by default. The known genotype is A/G at chr1:154568665. The read with ID SRR005196.8817822 is mapped to chr1:154568668 with 1 mismatch at chr1:154568683. The read with ID SRR005197.10106228 is mapped to chr1:154568657 of the alternative allele in the customized reference genome with 2 mismatches (chr1:154568660 and chr1:154568683). RefEdit discovers the new SNP at chr1:154568683 because of multiple existences of non-reference alleles. The Universal method, however, fails to map the read with ID SRR005197.10106228 because it exceeds the maximum mismatch threshold, therefore the new SNP cannot be discovered with confidence

variant detection rates for two individuals from different populations. In the second study, we used the Mendelian inconsistency (MI) as the metric for performance.

3.3.3 Study samples

We selected samples from the HapMap [95, 96] and 1000 Genomes Project [69, 75] requiring that the samples have been genotyped by both array-based genotyping platform and deep sequencing. Specifically, the African descent NA19238 and the European descent NA12716 were qualified and chosen for the first study, and the African trio (NA19238 (mother), NA19239 (father) and NA19240 (child)) was chosen for the second study.

3.3.4 Genotypes from genotyping arrays

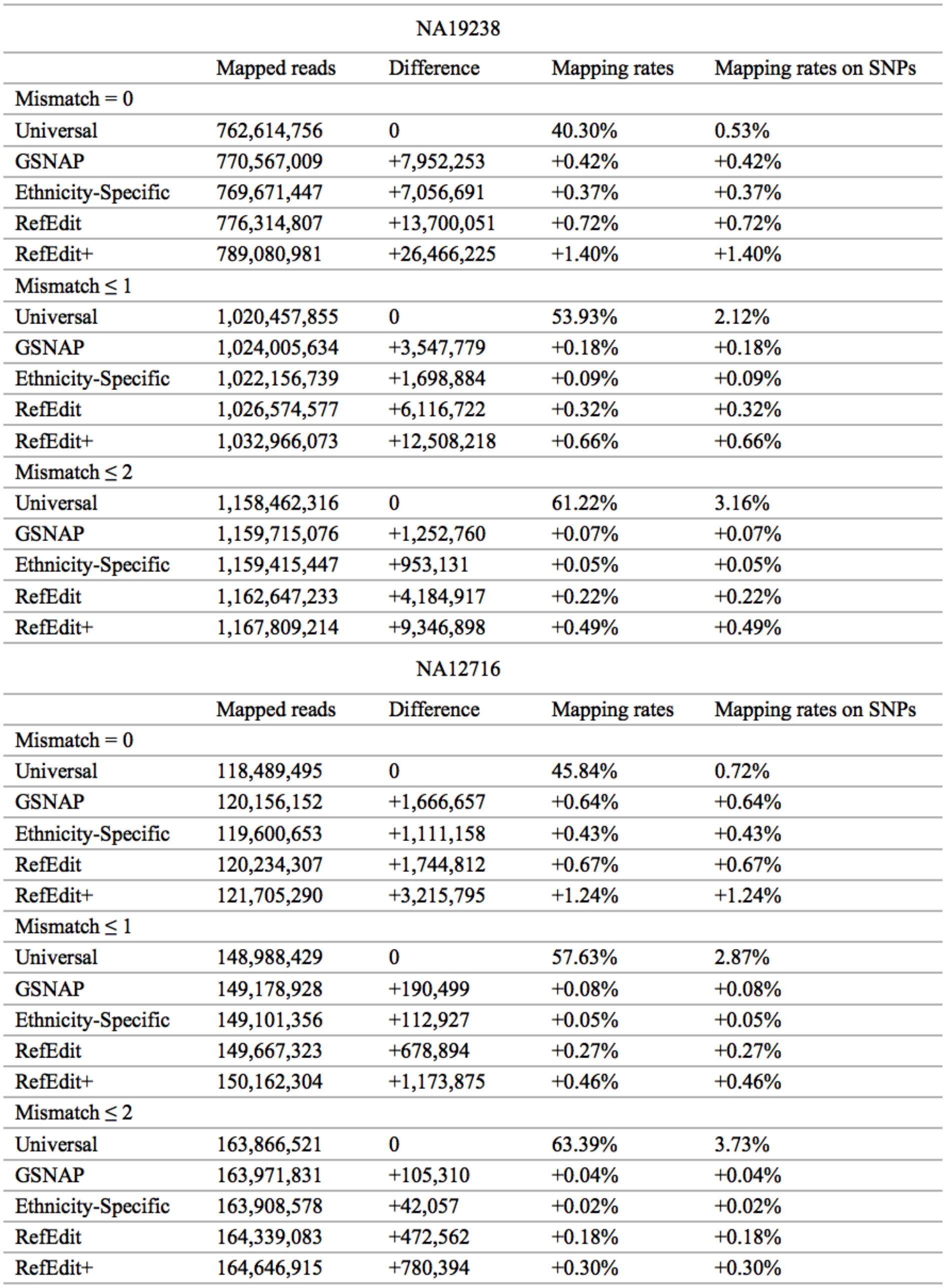
We chose the Affymetrix Axiom series array as the array-based genotyping platform in this study. This array contains about 6 million SNPs. We used the genotypes produced by the 1000 Genomes Project, which were called based on the CGI deep sequencing data, as the gold standard. This sequencing platform discovered about 41 million SNPs among 433 individuals. Both platforms produce high quality genotype calls and have been frequently used in other studies [97–99].

3.3.5 Genotype summary from genotyping array and imputation

We use population-specific reference panels from the 1000 Genomes Project [75] for imputation. To avoid biased results, the two haplotypes from the study sample are excluded from the panel during each run.

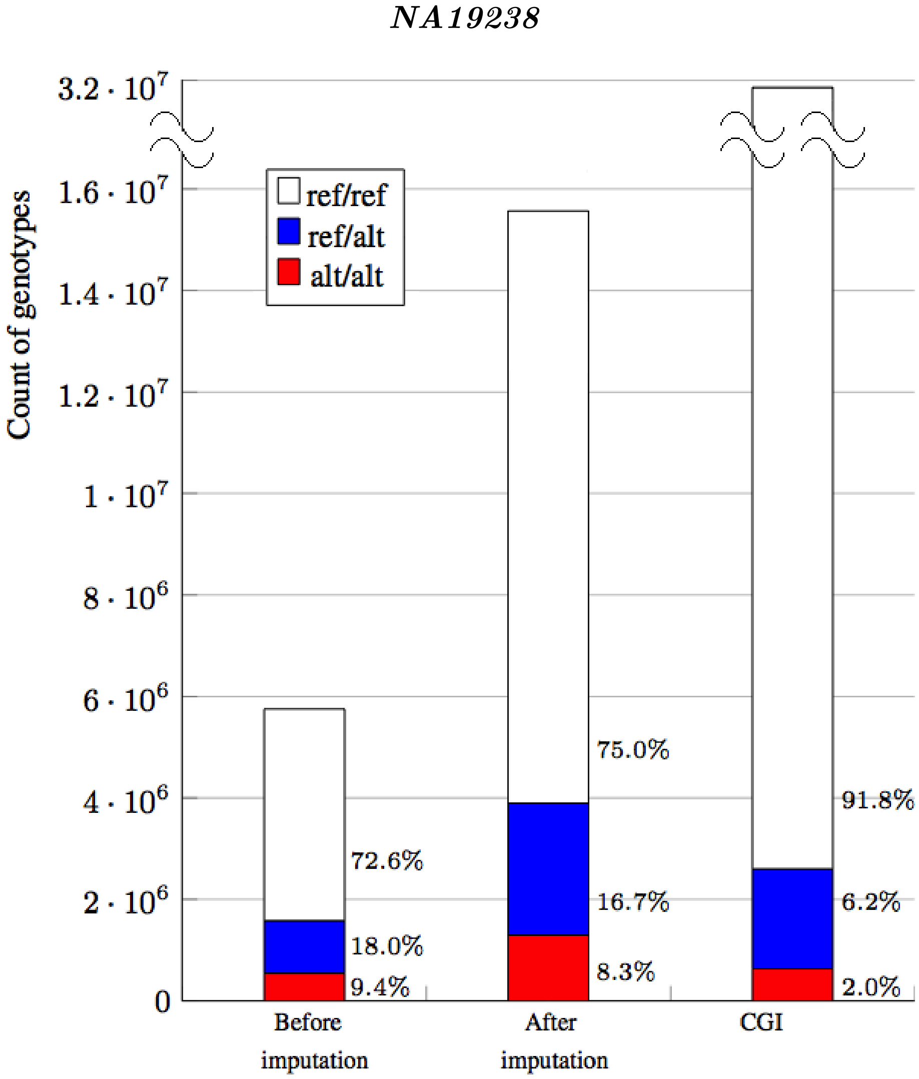
It is of interest to know, from the existing array-based genotype data, how many genotypes containing the alternative allele are identified and how many more can be added by genotype imputation. Genotype summaries (ref/ref, ref/alt, alt/alt proportions) for NA 19238 and NA 12716 are displayed in Figure 3.3 (also see Table 3.2 for numerical result). A Venn's diagram showing the overlaps between sets of assayed, imputed and the CGI gold standard genotypes can be found in Figure 3.4. The Axiom genotyping platform has very high concordance in the overlapping part with CGI genotypes (99.75% for NA19238 and 99.83% for NA12716) as shown in the Table 3.3., and hence is reliable. Details of the categorized consistencies between Affymetrix and CGI genotypes for individual NA19238 and NA12716 can be found in the Table 3.4.

The Rsq value is a good estimator of the correlation between the imputed and true genotypes, and thus is frequently used as a measure of imputation accuracy [85, 90, 100, 101]. By applying an appropriate Rsq threshold, we can achieve a reasonable

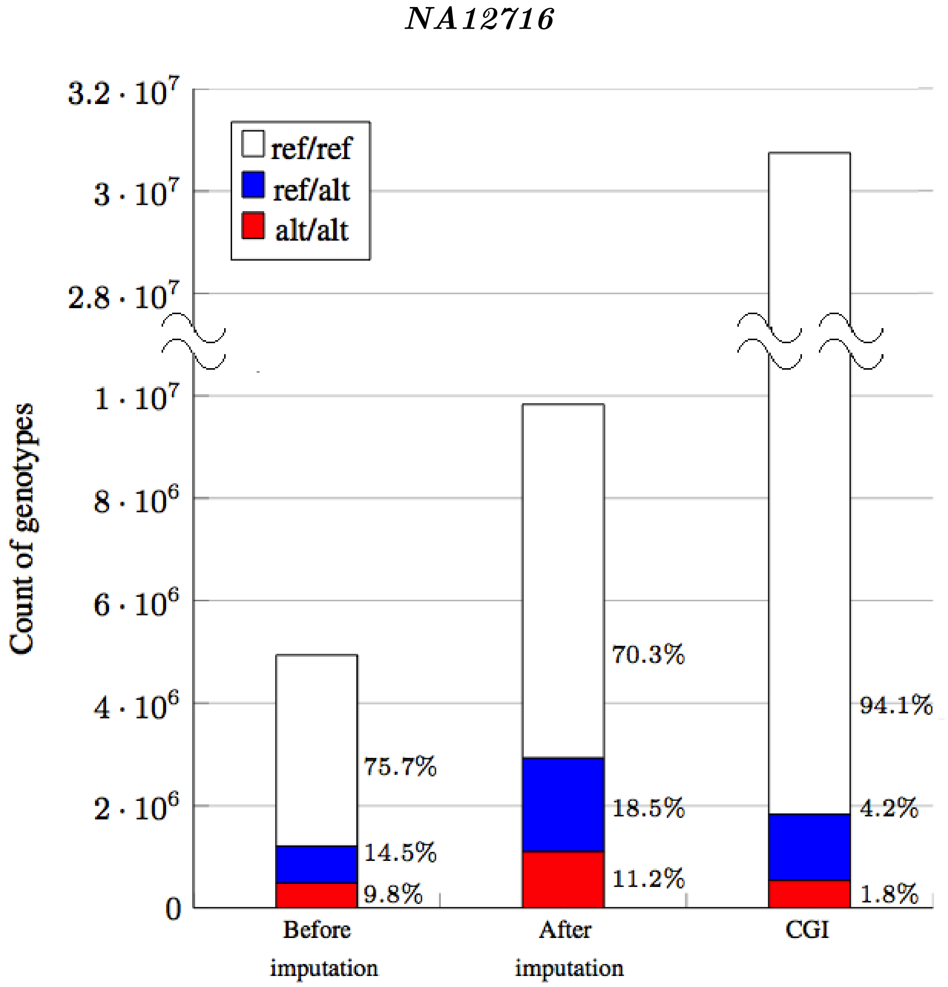


NA19238				
	Mapped reads	Difference	Mapping rates	Mapping rates on SNPs
Mismatch = 0				
Universal	762,614,756	0	40.30%	0.53%
GSNAP	770,567,009	+7,952,253	+0.42%	+0.42%
Ethnicity-Specific	769,671,447	+7,056,691	+0.37%	+0.37%
RefEdit	776,314,807	+13,700,051	+0.72%	+0.72%
RefEdit+	789,080,981	+26,466,225	+1.40%	+1.40%
Mismatch ≤ 1				
Universal	1,020,457,855	0	53.93%	2.12%
GSNAP	1,024,005,634	+3,547,779	+0.18%	+0.18%
Ethnicity-Specific	1,022,156,739	+1,698,884	+0.09%	+0.09%
RefEdit	1,026,574,577	+6,116,722	+0.32%	+0.32%
RefEdit+	1,032,966,073	+12,508,218	+0.66%	+0.66%
Mismatch ≤ 2				
Universal	1,158,462,316	0	61.22%	3.16%
GSNAP	1,159,715,076	+1,252,760	+0.07%	+0.07%
Ethnicity-Specific	1,159,415,447	+953,131	+0.05%	+0.05%
RefEdit	1,162,647,233	+4,184,917	+0.22%	+0.22%
RefEdit+	1,167,809,214	+9,346,898	+0.49%	+0.49%
NA12716				
	Mapped reads	Difference	Mapping rates	Mapping rates on SNPs
Mismatch = 0				
Universal	118,489,495	0	45.84%	0.72%
GSNAP	120,156,152	+1,666,657	+0.64%	+0.64%
Ethnicity-Specific	119,600,653	+1,111,158	+0.43%	+0.43%
RefEdit	120,234,307	+1,744,812	+0.67%	+0.67%
RefEdit+	121,705,290	+3,215,795	+1.24%	+1.24%
Mismatch ≤ 1				
Universal	148,988,429	0	57.63%	2.87%
GSNAP	149,178,928	+190,499	+0.08%	+0.08%
Ethnicity-Specific	149,101,356	+112,927	+0.05%	+0.05%
RefEdit	149,667,323	+678,894	+0.27%	+0.27%
RefEdit+	150,162,304	+1,173,875	+0.46%	+0.46%
Mismatch ≤ 2				
Universal	163,866,521	0	63.39%	3.73%
GSNAP	163,971,831	+105,310	+0.04%	+0.04%
Ethnicity-Specific	163,908,578	+42,057	+0.02%	+0.02%
RefEdit	164,339,083	+472,562	+0.18%	+0.18%
RefEdit+	164,646,915	+780,394	+0.30%	+0.30%

Table 3.1: Mapping rates of the five mapping strategies for individual NA19238 (1,892,304,208 reads) and NA12716 (258,507,654 reads).


NA19238
Count of genotypes
$3.2 \cdot 10^7$
$1.6 \cdot 10^7$
$1.4 \cdot 10^7$
$1.2 \cdot 10^7$
$1 \cdot 10^7$
$8 \cdot 10^6$
$6 \cdot 10^6$
$4 \cdot 10^6$
$2 \cdot 10^6$
0
ref/ref
ref/alt
alt/alt
72.6%
18.0%
9.4%
75.0%
16.7%
8.3%
91.8%
6.2%
2.0%
Before imputation
After imputation
CGI

(A)


NA12716
3.2 · 10^7
3 · 10^7
2.8 · 10^7
1 · 10^7
8 · 10^6
6 · 10^6
4 · 10^6
2 · 10^6
0
Count of genotypes
ref/ref
ref/alt
alt/alt
75.7%
14.5%
9.8%
70.3%
18.5%
11.2%
94.1%
4.2%
1.8%
Before imputation
After imputation
CGI

(B)

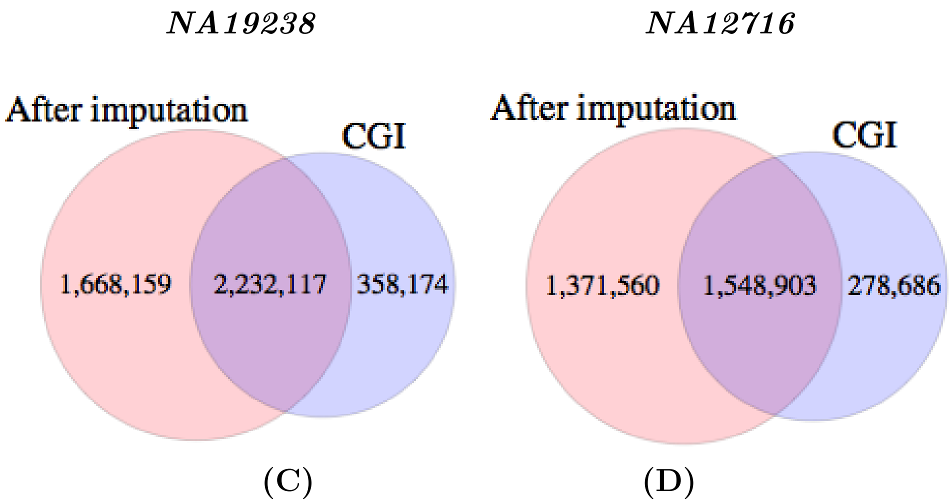


Figure 3.3: Comparison of genotype ratios before and after imputation for individuals NA19238 and NA12716. Imputation will considerably increase the amount of non-ref/ref genotypes. Non-ref/ref genotypes before and after imputation are incorporated into the customized reference genome construction for RefEdit and RefEdit+ methods respectively. (**A**) Genotype composition before/after imputation and CGI for sample NA19238. (**B**) Genotype composition before/after imputation and CGI for sample NA12716. (**C**) The overlapping of non-ref/ref genotypes between imputation and CGI for sample NA19238. Concordance is 98.94%. (**D**) The overlapping of non-ref/ref genotypes between imputation and CGI for sample NA12716. Concordance is 98.99%.

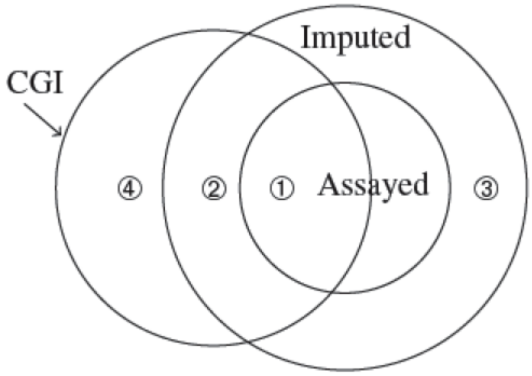


Figure 3.4: Venn's diagram illustrating SNPs with genotypes obtained from Affymetrix Axiom array, imputation and CGI sequencing for sample NA19238. (1) There are 4,611,084 overlapping SNPs between Affymetrix Axiom array and CGI with 99.75% concordant rate. (2) There are 6,851,861 overlapping SNPs between imputed and CGI with concordance rate 98.58%. (3) There are 2,965,053 SNPs with imputed genotype but not called by CGI sequencing. (4) There are 20,295,528 SNPs that called by CGI sequencing but not from Affymetrix Axiom array or imputation. Only 321,790 are non-ref/ref genotypes.

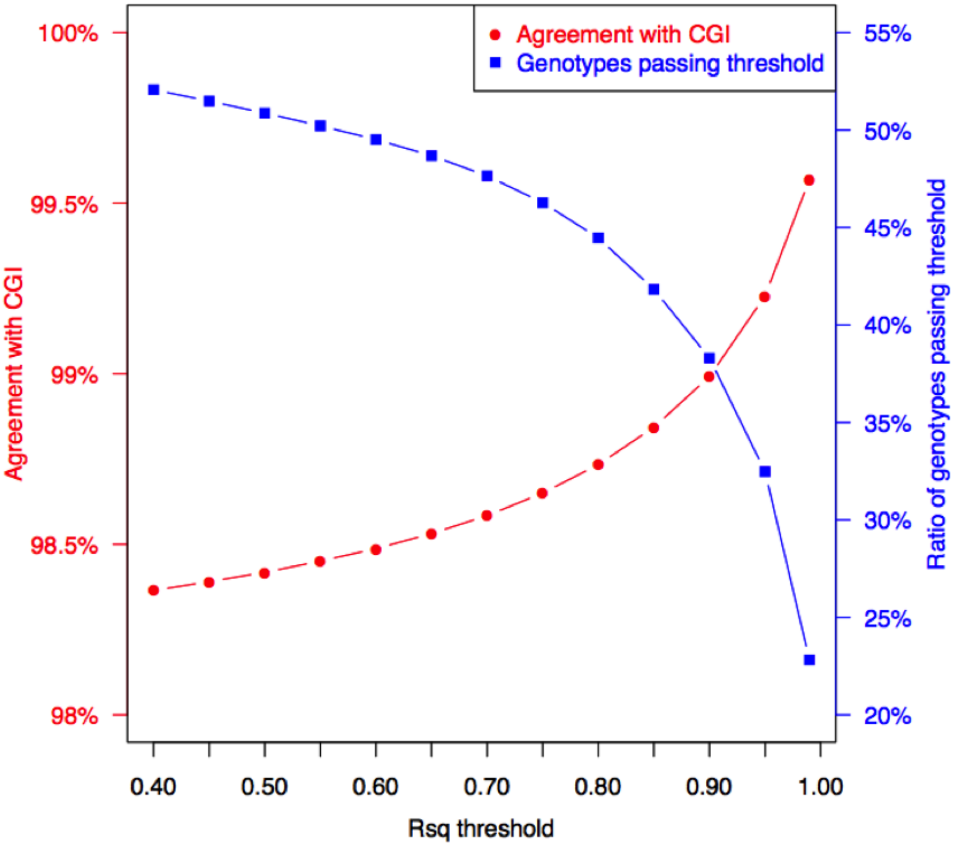
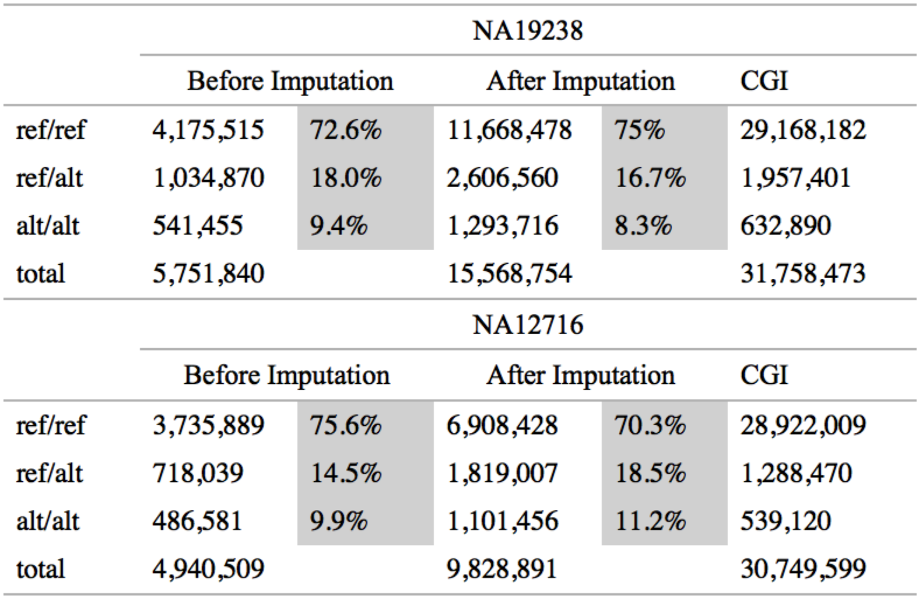


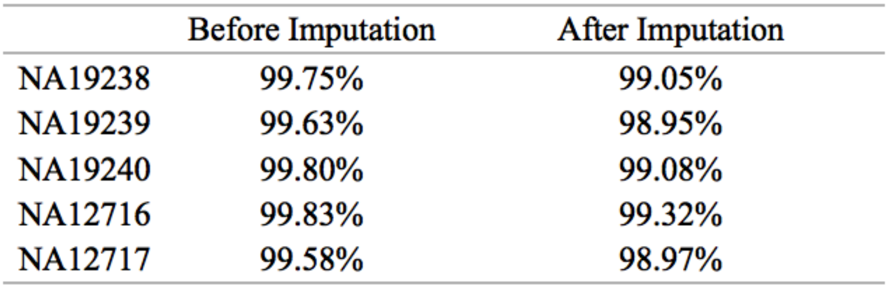
Figure 3.5: The proportions of imputed genotypes that passed the threshold and their accuracy compare to CGI gold standard across different Rsq value thresholds. The red curve indicates the concordance between imputed genotypes and CGI after applying the Rsq threshold. The blue curve indicates the proportions of the genotypes that pass the Rsq threshold.

Table 3.2: The total number and percentages of the three different types of genotypes that are being genotyped by the Affymetrix Axiom array, imputation and CGI sequencing.



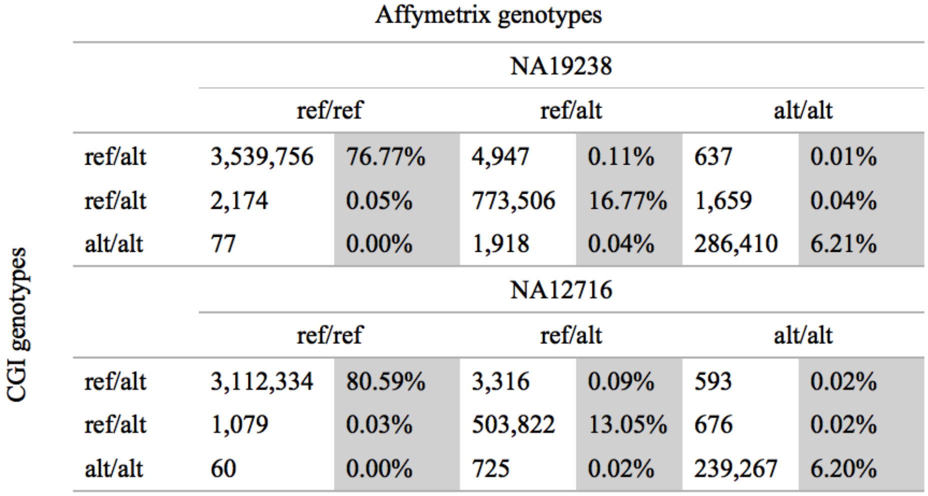
	NA19238				
	Before Imputation		After Imputation		CGI
ref/ref	4,175,515	72.6%	11,668,478	75%	29,168,182
ref/alt	1,034,870	18.0%	2,606,560	16.7%	1,957,401
alt/alt	541,455	9.4%	1,293,716	8.3%	632,890
total	5,751,840		15,568,754		31,758,473
	NA12716				
	Before Imputation		After Imputation		CGI
ref/ref	3,735,889	75.6%	6,908,428	70.3%	28,922,009
ref/alt	718,039	14.5%	1,819,007	18.5%	1,288,470
alt/alt	486,581	9.9%	1,101,456	11.2%	539,120
total	4,940,509		9,828,891		30,749,599

Table 3.3: Genotyping concordance rates between the Affymetrix Axiom array and CGI sequencing before and after imputation, ref/ref genotypes included.



	Before Imputation	After Imputation
NA19238	99.75%	99.05%
NA19239	99.63%	98.95%
NA19240	99.80%	99.08%
NA12716	99.83%	99.32%
NA12717	99.58%	98.97%

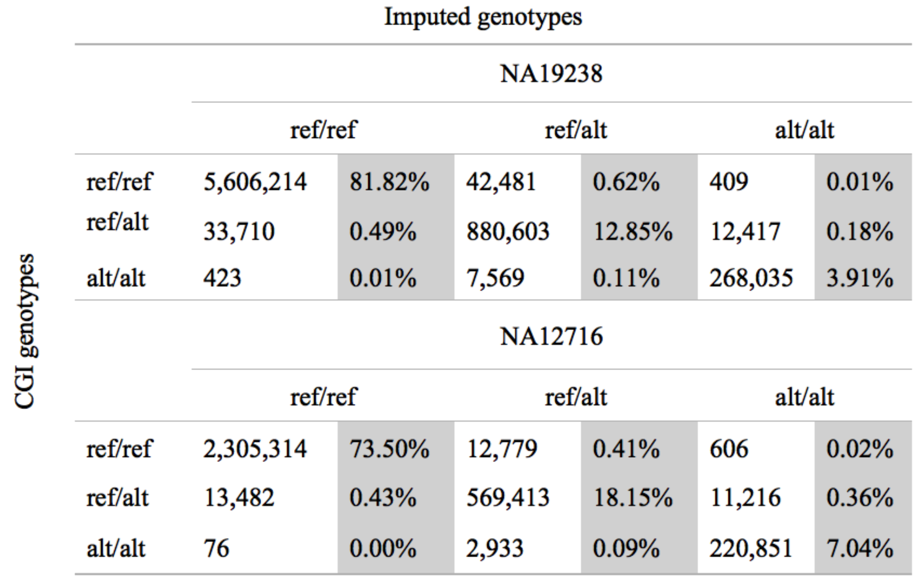
Table 3.4: The total numbers and percentages of the three different types of genotypes that are called by the Affymetrix Axiom array and the CGI sequencing.



Affymetrix genotypes

		NA19238					
		ref/ref		ref/alt		alt/alt	
CGI genotypes	ref/alt	3,539,756	76.77%	4,947	0.11%	637	0.01%
	ref/alt	2,174	0.05%	773,506	16.77%	1,659	0.04%
	alt/alt	77	0.00%	1,918	0.04%	286,410	6.21%
		NA12716					
		ref/ref		ref/alt		alt/alt	
	ref/alt	3,112,334	80.59%	3,316	0.09%	593	0.02%
	ref/alt	1,079	0.03%	503,822	13.05%	676	0.02%
	alt/alt	60	0.00%	725	0.02%	239,267	6.20%

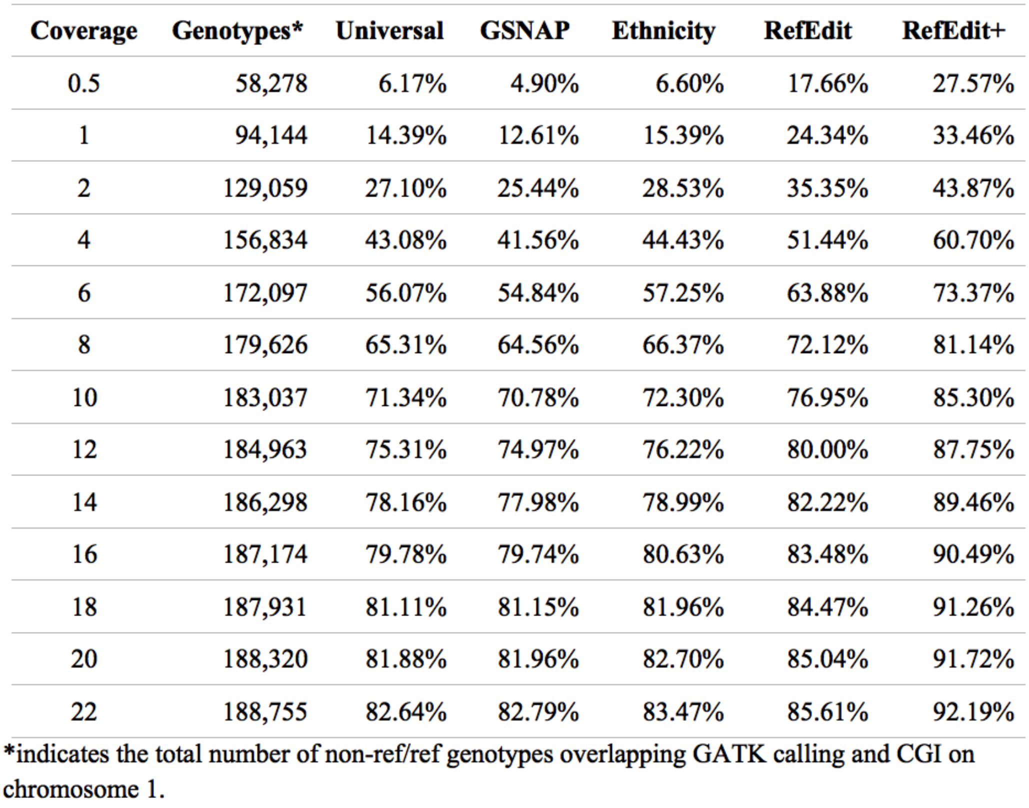
Table 3.5: The total numbers and percentages of the three different types of genotypes that are being genotyped by imputation and CGI sequencing.



Imputed genotypes

		NA19238					
		ref/ref		ref/alt		alt/alt	
CGI genotypes	ref/ref	5,606,214	81.82%	42,481	0.62%	409	0.01%
	ref/alt	33,710	0.49%	880,603	12.85%	12,417	0.18%
	alt/alt	423	0.01%	7,569	0.11%	268,035	3.91%
		NA12716					
		ref/ref		ref/alt		alt/alt	
	ref/ref	2,305,314	73.50%	12,779	0.41%	606	0.02%
	ref/alt	13,482	0.43%	569,413	18.15%	11,216	0.36%
	alt/alt	76	0.00%	2,933	0.09%	220,851	7.04%

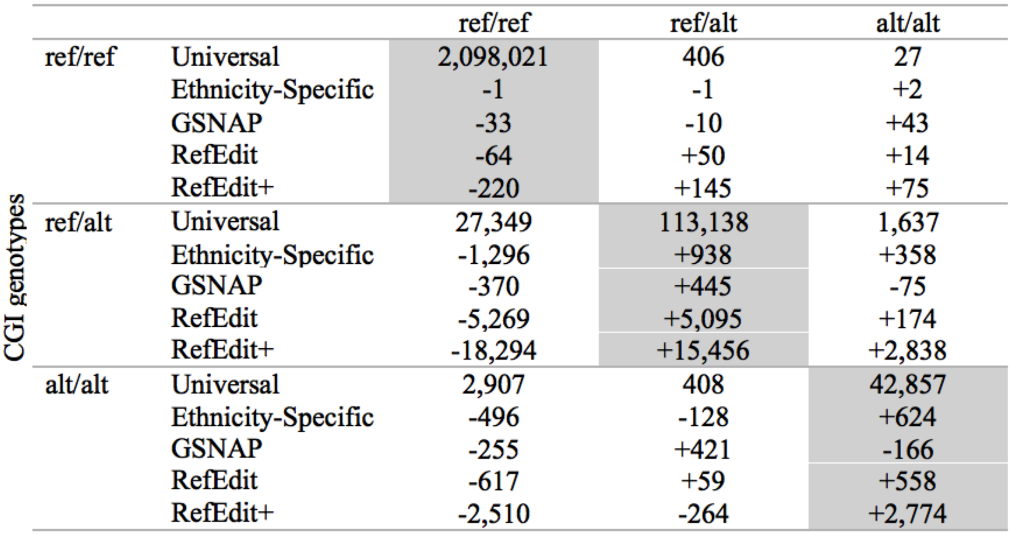
Table 3.6: S5 GATK genotype calling accuracy comparison of five methods for NA19238 on chromosome 1.



Coverage	Genotypes*	Universal	GSNAP	Ethnicity	RefEdit	RefEdit+
0.5	58,278	6.17%	4.90%	6.60%	17.66%	27.57%
1	94,144	14.39%	12.61%	15.39%	24.34%	33.46%
2	129,059	27.10%	25.44%	28.53%	35.35%	43.87%
4	156,834	43.08%	41.56%	44.43%	51.44%	60.70%
6	172,097	56.07%	54.84%	57.25%	63.88%	73.37%
8	179,626	65.31%	64.56%	66.37%	72.12%	81.14%
10	183,037	71.34%	70.78%	72.30%	76.95%	85.30%
12	184,963	75.31%	74.97%	76.22%	80.00%	87.75%
14	186,298	78.16%	77.98%	78.99%	82.22%	89.46%
16	187,174	79.78%	79.74%	80.63%	83.48%	90.49%
18	187,931	81.11%	81.15%	81.96%	84.47%	91.26%
20	188,320	81.88%	81.96%	82.70%	85.04%	91.72%
22	188,755	82.64%	82.79%	83.47%	85.61%	92.19%

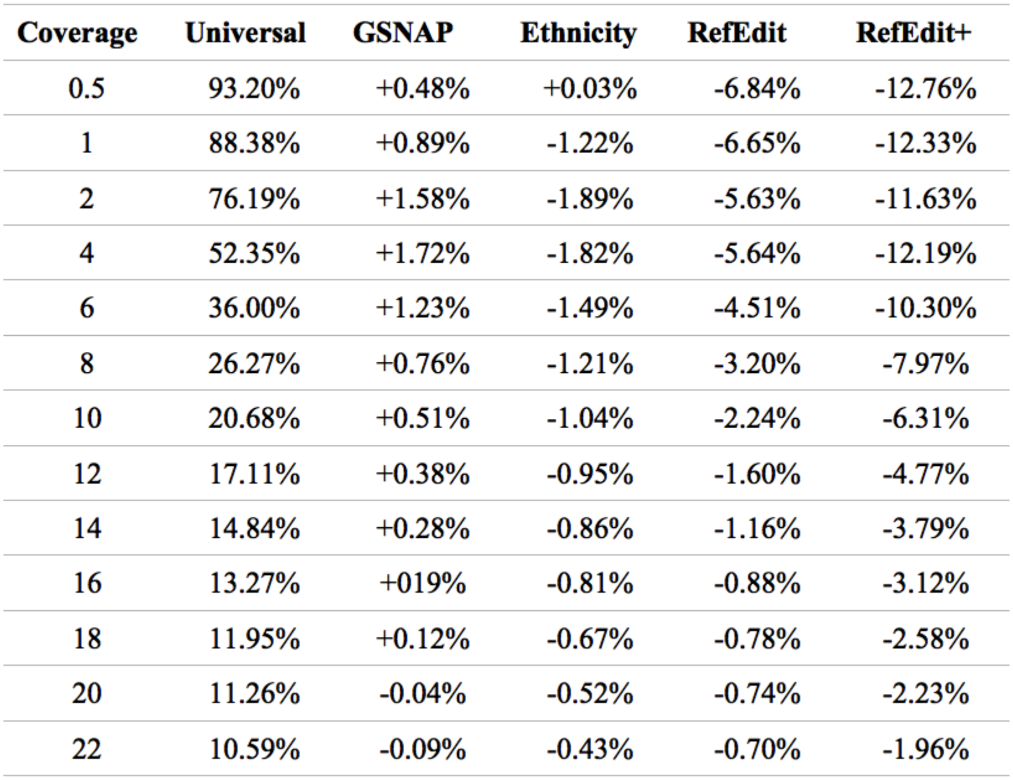
*indicates the total number of non-ref/ref genotypes overlapping GATK calling and CGI on chromosome 1.

Table 3.7: Comparison between GATK genotype calling results of five methods and CGI sequencing genotypes for NA19238 on chromosome 1. The sequencing depth is 22x. The differences (+/-) are from comparing to genotype calls using the universal reference genome method. The RefEdit and RefEdit+ methods increase the concordance (shaded parts) between genotype calls and the CGI gold standard genotypes.



CGI genotypes		ref/ref	ref/alt	alt/alt
ref/ref	Universal	2,098,021	406	27
	Ethnicity-Specific	-1	-1	+2
	GSNAP	-33	-10	+43
	RefEdit	-64	+50	+14
	RefEdit+	-220	+145	+75
ref/alt	Universal	27,349	113,138	1,637
	Ethnicity-Specific	-1,296	+938	+358
	GSNAP	-370	+445	-75
	RefEdit	-5,269	+5,095	+174
	RefEdit+	-18,294	+15,456	+2,838
alt/alt	Universal	2,907	408	42,857
	Ethnicity-Specific	-496	-128	+624
	GSNAP	-255	+421	-166
	RefEdit	-617	+59	+558
	RefEdit+	-2,510	-264	+2,774

Table 3.8: Mendelian Inconsistency comparison of five methods for YRI trio (NA19238, NA19239 and NA19240) on chromosome 1 using GATK genotype calling results. The differences (+/-) are from comparing to MI of using the universal reference genome method. The RefEdit and RefEdit+ methods consistently reduce the MI.



Coverage	Universal	GSNAP	Ethnicity	RefEdit	RefEdit+
0.5	93.20%	+0.48%	+0.03%	-6.84%	-12.76%
1	88.38%	+0.89%	-1.22%	-6.65%	-12.33%
2	76.19%	+1.58%	-1.89%	-5.63%	-11.63%
4	52.35%	+1.72%	-1.82%	-5.64%	-12.19%
6	36.00%	+1.23%	-1.49%	-4.51%	-10.30%
8	26.27%	+0.76%	-1.21%	-3.20%	-7.97%
10	20.68%	+0.51%	-1.04%	-2.24%	-6.31%
12	17.11%	+0.38%	-0.95%	-1.60%	-4.77%
14	14.84%	+0.28%	-0.86%	-1.16%	-3.79%
16	13.27%	+019%	-0.81%	-0.88%	-3.12%
18	11.95%	+0.12%	-0.67%	-0.78%	-2.58%
20	11.26%	-0.04%	-0.52%	-0.74%	-2.23%
22	10.59%	-0.09%	-0.43%	-0.70%	-1.96%

balance between the number and the quality of imputed genotypes. We used CGI as the gold standard to evaluate the genotype concordance. The ratio of imputed genotypes that passed the threshold and their accuracy compared to CGI genotype at different Rsq thresholds can be found in the Figure 3.5. We set the threshold at 0.7, which will retain 47.6% of the total imputed genotypes. The imputation accuracies for NA19238 and NA12716 are 99.05% and 99.32% respectively, as shown in the Table 3.3. Details of the categorized consistencies between the imputed genotypes and the CGI genotypes for individual NA19238 and NA12716 can be found in the Table 3.5. The numbers and proportions of newly imputed genotypes, along with those from the genotyping arrays, are shown in Figure 3.3.

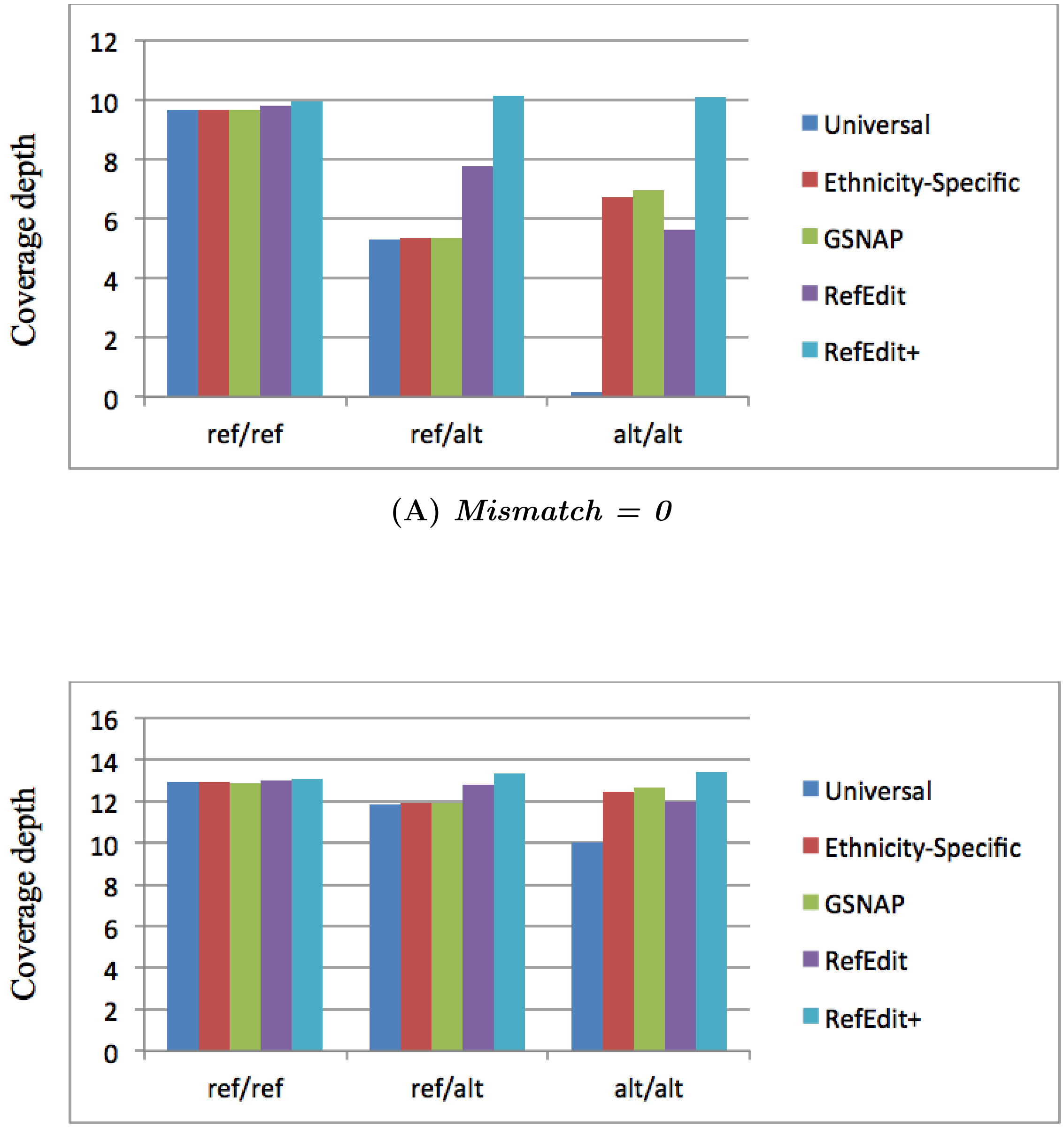
3.3.6 Read mapping rate

Since we do not know the true genomic location of a sequencing read generated from real sequencing experiments, we are unable to directly compare mapping accuracy. The proportion of successfully mapped reads among all sequenced reads is a reasonable alternative, which had been used in other studies [68, 94]. A successful mapping is defined as a unique mapping with no more than two mismatches. Here we compared the numbers and proportions of successfully mapped reads using different read mapping approaches. In addition to RefEdit and RefEdit+, we included three additional mapping strategies: standard read mapping with universal reference genome, read mapping with ethnicity-specific major allele reference genome [68], and mapping with GSNAP [94].

Our results indicate that RefEdit and RefEdit+ methods show consistent improvement in terms of the read-mapping rate. Table 3.1 summarizes the mapping rates of five methods under three mismatch thresholds. Figure 3.6 shows the average coverage depth of mapped reads from the five mapping strategies at different genotype categories for individual NA19238 (chr1∼chr22) using genotypes called from CGI sequencing data. Note that using the universal reference genome resulted in extremely low coverage depth at alt/alt loci when no mismatch is allowed, which is expected because only reads with sequencing errors happening to match the reference allele can be mapped to those loci.

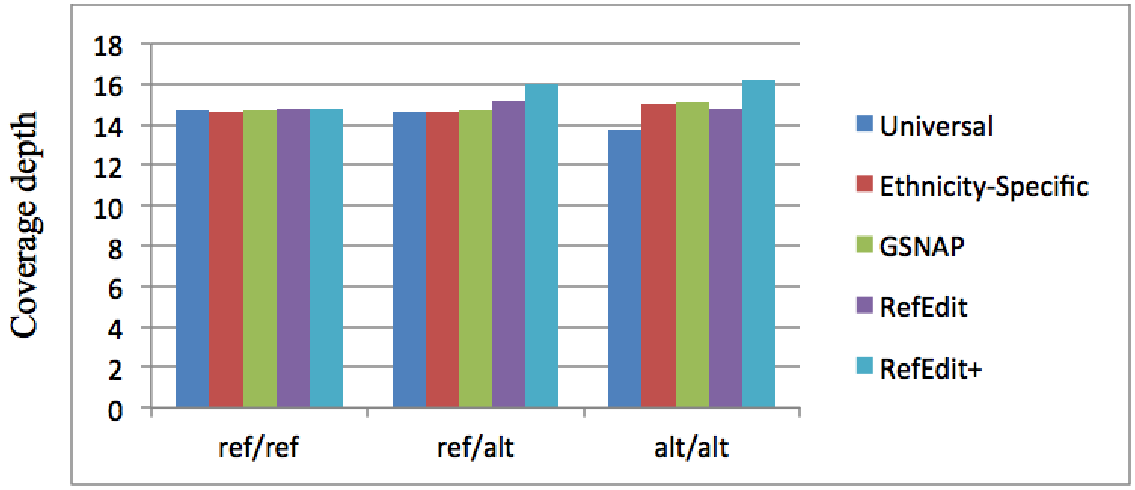
3.3.7 Genotype calling consistency

Using the CGI genotype calls as the gold standard, we evaluated the genotype calling accuracy of RefEdit/RefEdit+ with three competing methods at 13 different sequencing depths (0.5x, 1x, 2x, 4x, 6x, 8x, 10x, 12x, 14x, 16x, 18x, 20x, 22x) on individual NA19238. For each sequencing depth, performance comparison is conducted on the subset of non-ref/ref genotypes (according to CGI genotypes) that are called by GATK [54]. Figure 3.7[A] shows the concordance of the non-ref/ref genotypes for five different mapping methods. As expected, the genotype call consistency improves



(A) *Mismatch* = *0*

(B) *Mismatch* ≤ *1*



(C) $\boldsymbol{Mismatch \leq 2}$

Figure 3.6: Average depth of mapped reads from the five mapping strategies for individual NA19238 (chr1∼chr22), using CGI as gold standard for ref/ref, ref/alt and alt/alt loci. (**A**) Mismatch = 0. (**B**) Mismatch ≤ 1. (**C**) Mismatch ≤ 2. In the ref/ref loci group all methods have small differences in coverage depth; in ref/alt and alt/alt groups RefEdit+ method shows much higher coverage depth compare to other methods. The coverage depths increase when maximum allowed mismatches increase.

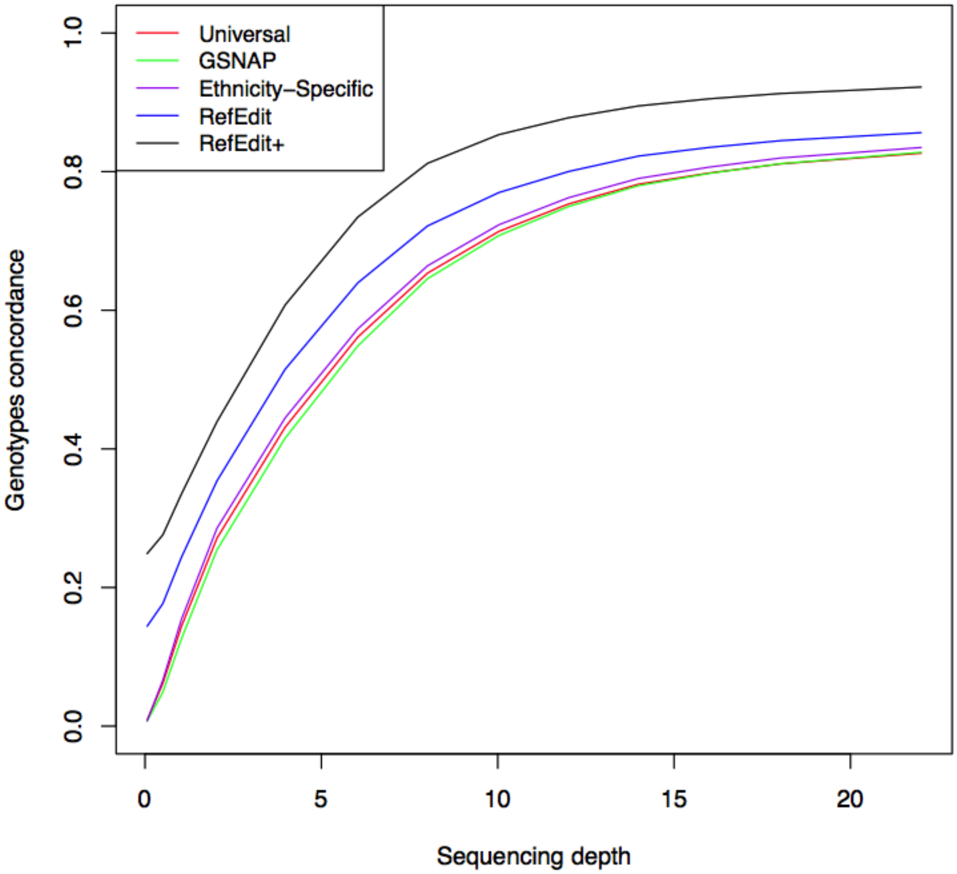
as the sequencing depth increases. Our RefEdit and RefEdit+ methods consistently outperformed the three competing methods in all read depths, with RefEdit+ performing the best. These results clearly demonstrate that incorporating genotype information of the individual into the read mapping process helps improving the accuracy of genotype calls. Note that the concordance rate is lower than reported elsewhere in the literature [5]. This is because here we chose a lower quality threshold in GATK to allow inclusion of more SNPs in the performance comparison study in light of the difference in sensitivity of different methods. Using the more commonly used threshold results in higher concordance across board and a similar pattern in terms of performance comparison (data not shown).

Remarkably, Figure 3.7[A] suggests that the read mapping using our RefEdit+ strategy can achieve the same level of accuracy as the read mapping using the universal reference genome, by using only a fraction of the reads required by the latter. Figure 3.7[A] shows that the method using the universal genome requires a sequencing depth of 22x to reach the same accuracy as RefEdit+ at a sequencing depth of ∼9x, albeit with about 4% fewer SNPs called by RefEdit+ at lower sequencing depth (Table 3.6). Given the cost associated with the sequencing depth, RefEdit+ provides a key benefit in terms of cost effectiveness. Compared to mapping using the universal reference genome, applying RefEditor can improve genotype concordance by 10% to 40% across different sequencing depth (from 22X to 4X).

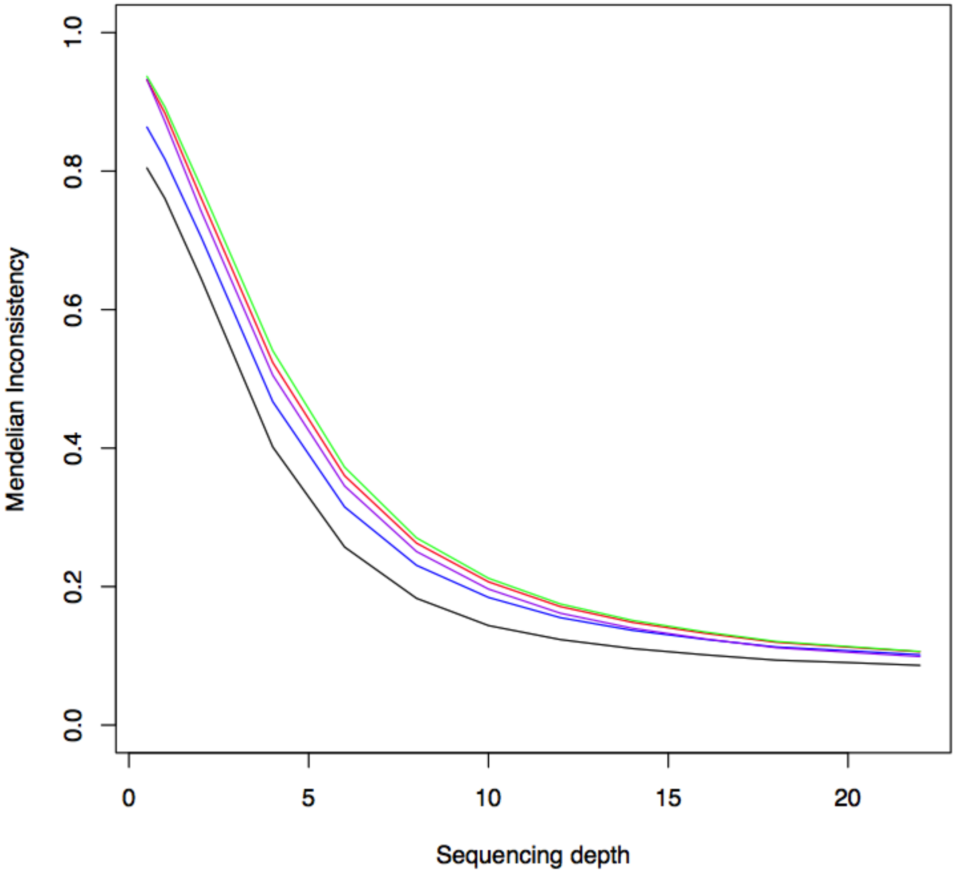
The detailed breakdown table of genotype concordance for five methods can be found in the Table 3.7, which shows that RefEdit+ moves a large proportion of genotypes that were previously incorrectly called as ref/ref by other methods to the correct genotypes of ref/alt or alt/alt, according to the CGI genotypes. The main reason for the incorrect ref/ref calls made by using the universal reference genome is that fewer reads that contain the alternative allele can be mapped to the correct locations compared to reads that contain the reference allele.

3.3.8 Mendelian inconsistency

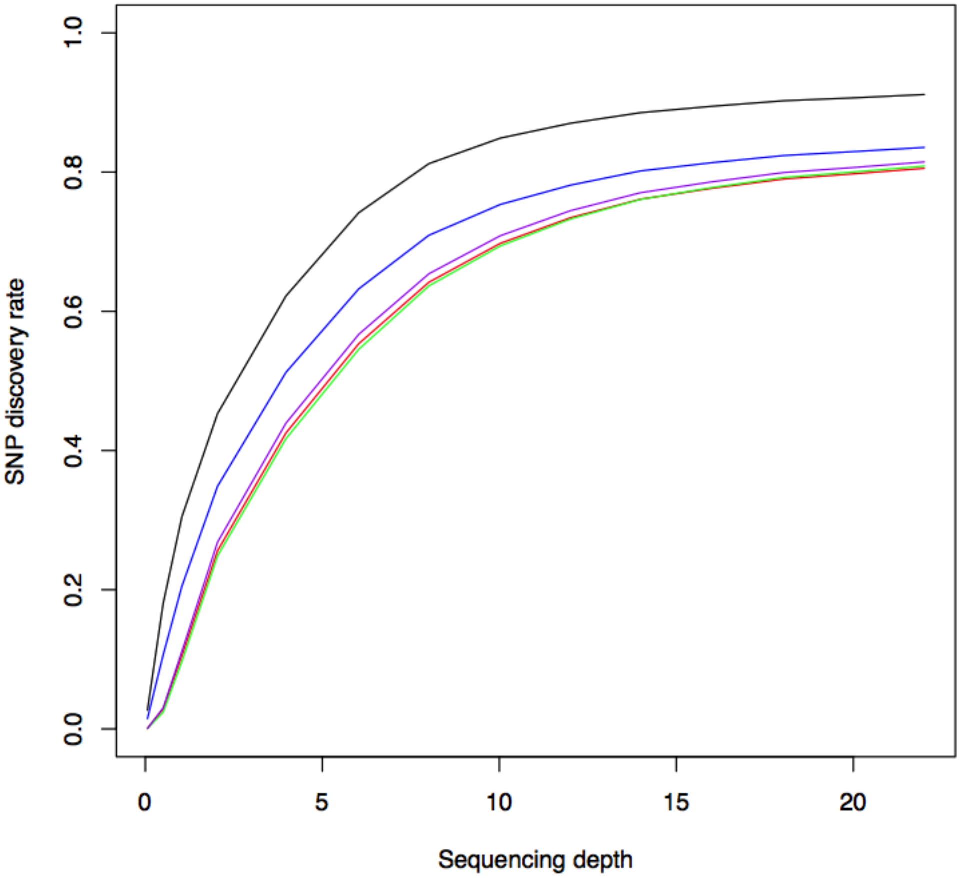
A drawback of evaluating performance using genotype concordance as above is that we need to designate a gold standard which may contain errors of its own, although error rate is rather low. Given that there are genotype data from two different platforms (array-based and sequencing-based) for parent-offspring trios from the International HapMap and 1000 Genomes Project, an alternative metric for performance evaluation is MI which counts the number of loci that show Mendelian errors within the trio. MI has been used in Dewey et al. to evaluate the performance of the ethnic-specific major allele reference genome approach [68]. For this study, we used data from chromosome 1 of an YRI trio (NA19238, NA19239 and NA19240) to calculate and compare MI at 13 different coverage depths (0.5x, 1x, 2x, 4x, 6x, 8x, 10x, 12x,


Universal
GSNAP
Ethnicity–Specific
RefEdit
RefEdit+
Genotypes concordance
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Sequencing depth

(A)


Mendelian Inconsistency
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Sequencing depth

(B)



(C)

Figure 3.7: (**A**) GATK genotype calling accuracy of five methods for NA19238 on chromosome 1 using CGI as a gold standard. (**B**) Mendelian Inconsistency of five methods for YRI trio (NA19238, NA19239 and NA19240) on chromosome 1 using GATK genotype calling results. (**C**) GATK SNP discovery rates of five methods for NA19238 on chromosome 1 compared to CGI SNPs.

14x, 16x, 18x, 20x, 22x). We only compared performance at loci where all three individuals made the genotype calls and not all of them have homozygous genotypes. MI rates are illustrated in Figure 3.7[B], which shows that the RefEdit+ method has the lowest MI values across all sequencing depths. A breakdown table of MI for all methods at different sequencing depths can be found in Table 3.8. Compared to mapping using the universal reference genome, applying RefEdit+ can significantly reduce MI by 10% to 30% across various sequencing depth.

3.3.9 SNP identification

Besides genotype calling accuracy at known SNP sites, when conducting WGS studies, it is also important to correctly identify novel SNP variants, as was illustrated in the previous example (Figure 3.2). Therefore, we assess whether RefEdit+ also improves SNP detection. To be specific, we compared the SNP detection rate when using different read mapping methods at different sequencing depths. For each read mapping strategy, we define the SNP detection rate as follows: among all SNPs identified by CGI sequencing, the proportion of SNPs that are also identified by GATK (non-ref/ref genotypes). As shown in Figure 3.7[C], RefEdit+ is able to identify the most number of SNPs, followed by RefEdit. The performance enhancement of RefEdit/RefEdit+ is maximized at about 10x coverage.

3.4 Discussion

With the price of DNA sequencing continuing its rapid decline, whole genome sequencing will likely to be performed *en masse* in research laboratories and perhaps clinics with the primary goal of identifying genetic variants. Mapping the sequencing reads to the human genome is an important early step to analyze data from all sequencing-based experiments including WGS. Multiple studies [68, 94] have demonstrated that genetic variants that occur in about 1% of the genome have a non-ignorable impact on the mapping accuracy, which in turn affects the accuracy of the genotype calls of these variants. Scientists have attempted to address this issue by either incorporating all known genetic variants [94] or ethnic-specific major variants [68] into the mapping process. In this study, we go one step further and propose a novel method that takes advantage of the increasingly available personal genotype information. The key of our approach is to customize the reference genome using known genotypes of that individual. Our extensive performance comparison studies demonstrate significant improvement in terms of read mapping, genotype calling and SNP identification.

The performance improvement of RefEditor over existing mapping strategies is easy to understand, because more information is being incorporated. Our work showed that the improvement could be achieved computationally efficiently and in a straightforward fashion using RefEditor. Because array-based genotyping technologies have matured and cost less than WGS, they have been the choice for most large-scale association studies to date [82] . A slew of special-design genotyping chips have also been developed or under-development to supplement the mundane GWAS genotyping chips [102, 103]. As a result, large amount of dense genotyping information is readily available for large cohorts of samples. Many WGS studies were conducted on these samples [104, 105]. Such a design makes our personalized reference genome strategy very attractive.

Figure 3.3[A] and [B] showed a surprising result that genotyping array augmented by imputation actually identified more SNPs (non-ref/ref genotypes) than deep WGS (3.9 million vs. 2.6 million for NA19238; 2.9 million vs. 1.8 million for NA12716). Although such results are atypical, the situation showed in Figure 3.3[C] and [D] is noteworthy which indicates that a large number of SNPs identified by genotyping array plus imputation were not called by WGS even with high coverage. We think there are two reasons behind it: first, the coverage depth of WGS is not uniform across the genome; second, loci harboring genetic variants tend to have relatively lower mapping coverage depth due to the fact that the reference genome does not contain any genetic variant information. The development of next generation of denser genotyping imputation panels promises to further improve the effectiveness of imputation of common and rare variants [106]. These suggest that one should consider supplementing WGS or WES by cheaper array-based genotyping and imputation in order to increase the quantity and quality of overall genotype calls; and a better read mapping strategy that can utilize existing genotypes such as RefEditor should be employed in the process.

It has been reported in the literature that multi-sample SNP calling strategy improves genotype calling in WGS studies [5]. Since that particular approach is carried out after the read mapping step, our strategy can also be applied during the read mapping step which we believe will further enhance the genotype calling downstream. Due to the requirement of a reasonable number of samples in the cohort to apply the multi-sample calling strategy, we are unable to evaluate the potential performance enhancement under that scenario in the current study.

Another important lesson we learned is that the genotype imputation strategy plays a key role in performance improvement for RefEdit+. Genotype imputation has been monumentally successful in GWAS analysis. We demonstrate that high quality imputed genotypes also improve the reference genome customization and therefore produce improved read mapping and genotype calling results.

An extension of our customized reference genome strategy is to apply RefEditor iteratively for multiple rounds. Specifically, after genotypes were called with the help of RefEditor, we can combine these genotypes with known genotypes that were used earlier to obtain an updated set of known genotypes, and then apply RefEditor to perform read mapping and genotype calling again. The same strategy can also be applied to WGS samples without existing genotype information.

Our performance comparison results demonstrate the importance and benefits of incorporating existing genotype information in read mapping, genotype calling and variants discovery in WGS studies. Admittedly, more work is required to perform read mapping with RefEditor: unlike using a single universal reference genome, one has to generate a reference genome for each individual sample in the cohort. A post-process step is also needed after read mapping. However, with our RefEditor package, the whole read mapping process can be automated using simple scripts, and therefore very little human time and intervention is needed in adopting our personalized read mapping strategy. As for computation time, in our experiment on a single core 1.4G Hz CPU and 8GB memory, Diploid Constructor took 4 minutes and 32 seconds to construct the diploid reference genome from hg19.fa and 15,568,754 genotypes (3,900,277 non-ref/ref). The reference genome size increased by 0.2 GB (from 3.0 GB to 3.2 GB) and indexing time increased 5 minutes and 30 seconds (from 87m8s to 92m38s). Reads mapping time increased 5 seconds (from 18m49 to 18m54s) to map 5,112,949 reads (read length is 36 bp). Mapping Converter took

49 seconds to convert the intermediate mapping results. Given the importance of accurately identifying genetic variants in WGS studies, which are often rare and have low sequencing coverage, we strongly advocate the new strategy of using personalized reference genome in read mapping.

3.5 Web Resources

The URLs for source code and data presented are as follows:

RefEditor source code, `http://code.google.com/p/refeditor/`

Universal hg19.fa, `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/`

MaCH version 1.0.18, `http://www.sph.umich.edu/csg/abecasis/MaCH/download/mach.1.0.18.source.tgz`

Minimac RELEASE STAMP 2012-11-16, `http://www.sph.umich.edu/csg/cfuchsb/minimac-beta-2012.11.16.tgz`

Ethnicity specific reference genome, `http://datadryad.org/bitstream/handle/10255/dryad.35120/YRIref.fasta.zip?sequence=3`

All FASTQ files, `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/`

Chapter 4

RefEditor-Galaxy, a Galaxy tool for enhancing read mapping as part of bioinformatics workflows

Abstract

RefEditor package and the Pipelines built on it (RefEdit and RefEdit+) have proved beneficial in many aspects of genomic research (Chapter 2, 3). We present RefEditor-Galaxy, a wrapper for using RefEditor within Galaxy. The functionality delivered by RefEditor (*i.e.* diploid reference genome construction) can be combined with the tools and workflows devised within the Galaxy framework or its repositories, resulting in an enhancement of RefEditor. A use case is provided in order to demonstrate RefEditor-Galaxy's capability for reference genome constructing, read mapping and format converting. Coupling RefEditor-Galaxy with other bioinformatics tools of the Galaxy framework results in a system that opens a new dimension of NGS experiments and analyses. RefEditor-Galaxy's source code can be downloaded

from: `http://toolshed.g2.bx.psu.edu/repos/superyuan/refeditor`

4.1 Introduction

4.1.1 RefEditor

RefEditor package is mainly designed to improve NGS read mapping and help downstream genomic studies for diploid organisms. Its capabilities have already been demonstrated as to be able to increase read mapping rate, to reduce ASE bias (Chapter 2), to increase genotype calling accuracy and SNP discovery rate (Chapter 3). However, it is a Command Line Interface (CLI) based software package. It requires researchers to have background knowledge about executing, debugging and organizing commands on Unix-family Operating Systems to be able to exploit RefEditor's functionalities to create versatile pipelines for typical bioinformatics analyses. Galaxy [107–110], an popular, web-based platform combining various genomic-oriented tools into workflows, offers an ideal Graphical User Interface (GUI) platform for making RefEditor part of bioinformatics analyses. Therefore, we have developed RefEditor-Galaxy, a tool to execute RefEditor programs from within Galaxy. RefEditor-Galaxy fully takes advantage of RefEditor's functionalities and can interact with other Galaxy tools in an integrated fashion. This chapter presents an overview of RefEditor-Galaxy's design and implementation, including a tested use cases that provides a basis for creating more complex analyses.

4.1.2 Galaxy

Galaxy offers an open, web-based unified workbench platform for performing genomic analyses [107–110]. It has established a significant community of users and developers [111]. Galaxy's public server (`http://usegalaxy.org`) makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist that has access to the Internet [112]. This server processes about 5,000 jobs per day. Individuals and groups have used Galaxy to perform many types of genomic research, including investigations of epigenomics [113, 114], chromatin profiling [115, 116], transcriptional enhancers [117], genome-environment interactions [118] and others [119, 120]. In addition to the public server, thousands of local Galaxy servers have been set up by downloading the Galaxy application and customizing it to meet particular needs.

Galaxy majorly features three parts:

- Making computation accessible. Several database resources have been integrated with the public Galaxy server and are included as part of the downloadable package. These resources include the UCSC Table Browser [121], BioMart Central Portal, InterMine, EpiGraph [113], EuPathDB [122] and Hb-

Var [123]. It has also integrated hundreds of tools from diverse categories including Text Manipulation, Filter and Sort, Formats Conversion, Alignment, Statistics, Motif detection, NGS processing and more. There are still hundreds of tools published in Galaxy's main tool shed and other repositories.

- Ensuring reproducibility. Reproducing experimental results is an essential facet of scientific inquiry, providing the foundation for understanding, integrating, and extending results toward new discoveries. Galaxy provides an environment to automatically generate metadata from each analysis step in order to record and repeat computational analyses history. It tracks the provenance of data and tool usage which enables users to selectively run and rerun particular analyses [124, 125]. From Galaxy users can extract workflows which are reusable templates analysis that can repeat on different data. Galaxy provides an environment to visualize, modify, annotate, save, delete and rerun workflows.

- Promoting transparency. Data, histories, workflows, and pages can be easily shared and published to Galaxy's public repositories. Those items can be import and immediately used for extensive studies inside Galaxy [110].

4.2 RefEditor-Galaxy

To add a set of tools to Galaxy, we first define the default sections for those tools in a file in Extensible Markup Language (XML) syntax. Those default sections can be changed when users install the tool set. We also write another XML configuration file for each new tool that describes how to run the tool, including detailed specification of input and output parameters. This specification allows the Galaxy framework to work with the tool abstractly and automatically generate layout of computational tool from XML descriptions to ensure a consistent look and feel. This XML file will invoke another interpreter to execute a program/script for data check. Those interpreters can be python, perl, bash or sh. After data checking, a final program will be invoked from those data checker. This final program can be any piece of software written in any language as long as it can be invoked from a console command. Figure 4.1 shows the call graph.

RefEditor-Galaxy consists of five components: tool_conf.xml, vcf2genotypes.∗, DiploidConstructor.∗, MappingConverter.∗ and test-data. They are organized in the directory tree shown in Figure 4.2

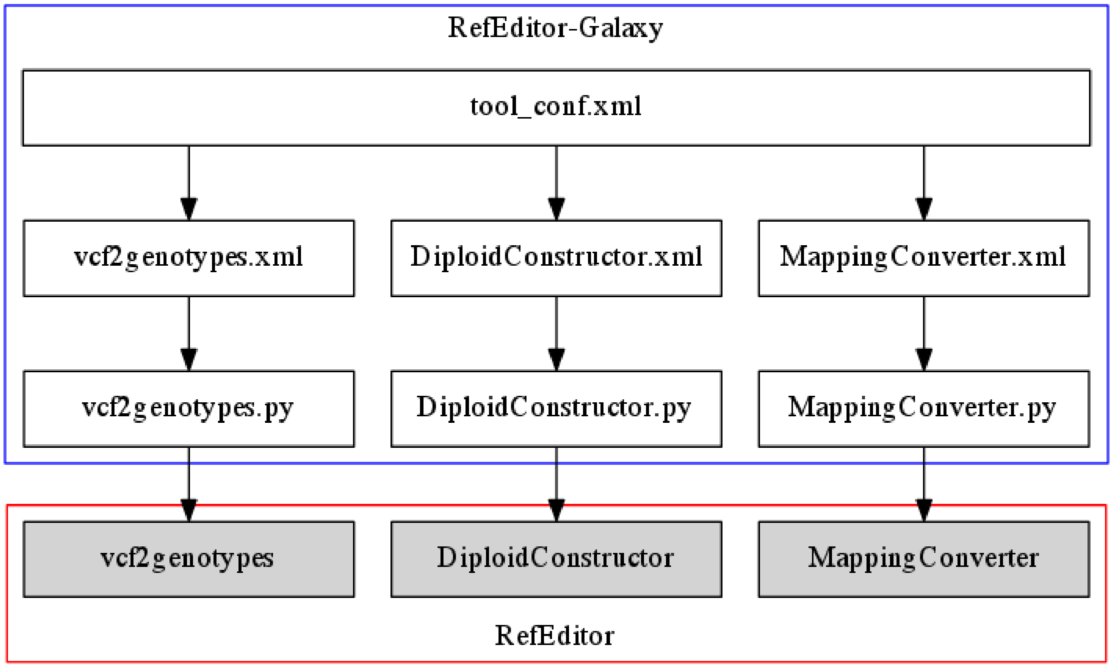


Figure 4.1: Calling graph of RefEditor-Galaxy. RefEditor-Galaxy (blue box) invokes RefEditor (red box). tool_conf.xml is the entry of other three independent tools (vcf2genotypes.∗, DiploidConstructor.∗ and MappingConverter.∗).

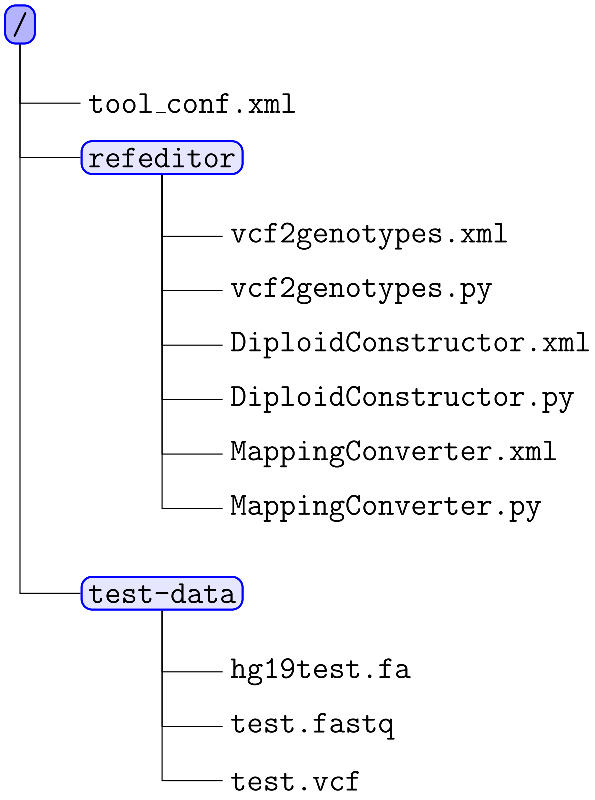


Figure 4.2: Directory tree for RefEditor-Galaxy. *tool_conf.xml* defines the sections the tools belong to. *refeditor* contains the definition of three tools. *test-data* contains all the test files.

4.2.1 tool_conf.xml

tool_conf.xml in the top directory describes the default section of other three tools (vcf2genotypes.∗, DiploidConstructor.∗ and MappingConverter.∗) shown in the Galaxy tools pane on the left side of the browser window. It also behaves as the entry for those tools.

4.2.2 vcf2genotypes.∗

vcf2genotypes.∗ is a tool that extracts the genotypes of a specified individual above the minimal quality score from a VCF file and store them in a standard genotypes format used by HapMap project. vcf2genotypes.xml accepts three parameters:

- A VCF file [List Data].
- Name of the target individual [String]. It must be a column name of an individual defined in the VCF file.
- Minimal Quality Score [Integer]. Default=0.

vcf2genotypes.xml will invoke vcf2genotypes.py to perform sanity check on parameters, and then execute vcf2genotypes.

4.2.3 DiploidConstructor.∗

DiploidConstructor.∗ takes the universal haploid reference genome and genotype file as the input, and then output the personalized diploid reference genome containing alternative alleles . DiploidConstructor.xml accepts five parameters:

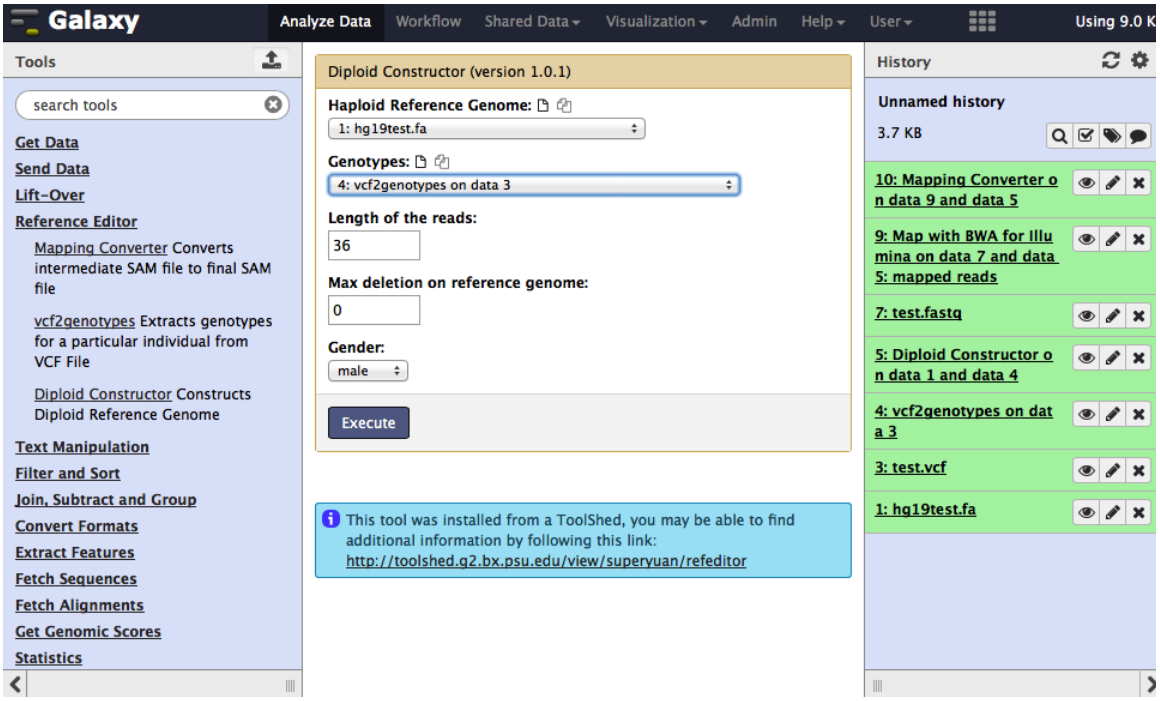
- A haploid reference genome file in FASTA format [List Data].
- A genotype file [List Data].
- Read length [Integer].
- Maximum length of deletions on mini-chromosomes [Integer]. Default=0.
- Gender [Binary Selection]. Default=male

DiploidConstructor.xml will invoke DiploidConstructor.py to perform sanity check on parameters, and then execute DiploidConstructor.

Figure 4.3 shows the web interface for DiploidConstructor.∗.

4.2.4 MappingConverter.∗

MappingConverter.∗ modifies intermediate results of read alignment in SAM format, and then output final mapping results that correspond to the universal reference genome. MappingConverter.xml accepts two parameters:



Diploid Constructor (version 1.0.1)

Haploid Reference Genome: 1: hg19test.fa

Genotypes: 4: vcf2genotypes on data 3

Length of the reads: 36

Max deletion on reference genome: 0

Gender: male

This tool was installed from a ToolShed, you may be able to find additional information by following this link: http://toolshed.g2.bx.psu.edu/view/superyuan/refeditor

Unnamed history

3.7 KB

10: Mapping Converter on data 9 and data 5

9: Map with BWA for Illumina on data 7 and data 5: mapped reads

7: test.fastq

5: Diploid Constructor on data 1 and data 4

4: vcf2genotypes on data 3

3: test.vcf

1: hg19test.fa

Figure 4.3: RefEditor-Galaxy Web interface. The RefEditor-Galaxy Web interface is displayed in the middle pane. In the left pane, a list of standard Galaxy tools and RefEditor-Galaxy are shown; in the right pane, a sample of a history of the executed tasks is shown.

- A SAM file that contains intermediate mapping results [List Data].
- A diploid reference genome file, based on which the intermediate mapping results are generated [List Data].

MappingConverter.xml will invoke MappingConverter.py to perform sanity check on parameters, and then execute MappingConverter.

4.2.5 test-data

Galaxy Dev team suggests that an optional test-data folder be included in the tool package. This folder contains all testing files so that automated tests can be conducted in the repository server. Users can also manually test RefEditor-Galaxy using provided test data.

RefEditor-Galaxy eventually invokes binary programs defined in RefEditor to fulfill the tasks as shown in Figure 4.1.

4.3 Installation

Galaxy Tool Shed is a *de facto* AppStore for Galaxy tools [126]. Galaxy Main Tool Shed (`http://toolshed.g2.bx.psu.edu/`) and Test Tool Shed (`http://testtoolshed.g2.bx.psu.edu/`) are already included with Galaxy's distribution (`tool_sheds_` `conf.xml`). Hundreds of tools have been published on those Sheds. Users can easily download and install a tool with just a few mouse clicks, which is another important benefit Galaxy provides.

RefEditor-Galaxy is published on the Galaxy Main Tool Shed. It can be downloaded by going to Galaxy Admin Interface and click *Search and browse tool sheds* → *Galaxy Main Tool Shed* → *Fasta Manipulation*, as shown in Figure 4.4 and Figure 4.5

The RefEditor, on which RefEditor-Galaxy relies, should also be installed from `http://code.google.com/p/refeditor/`

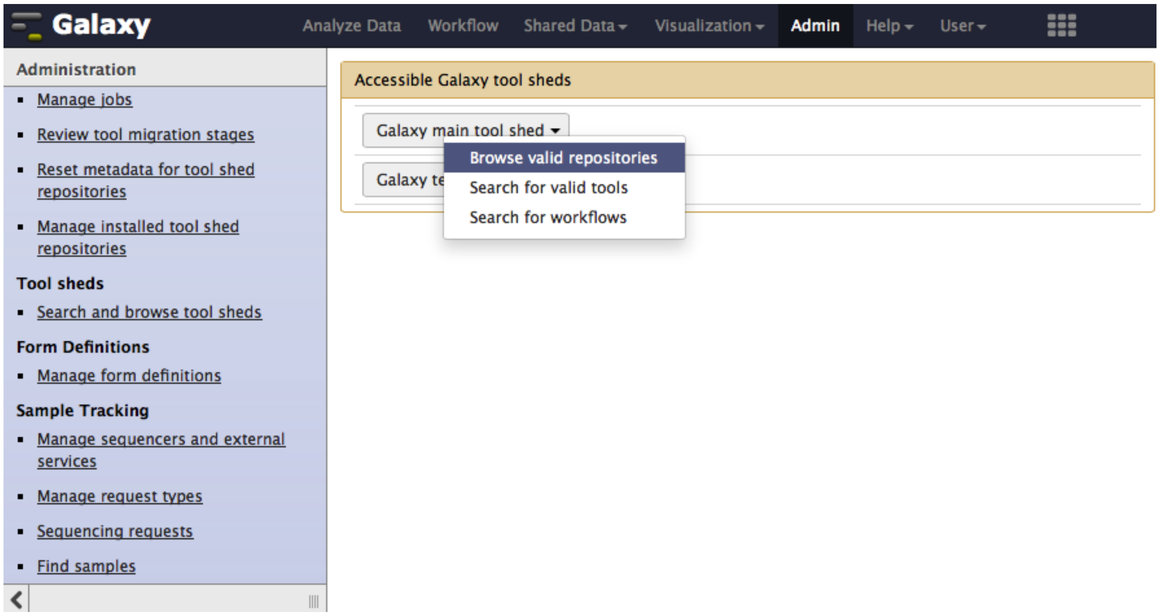
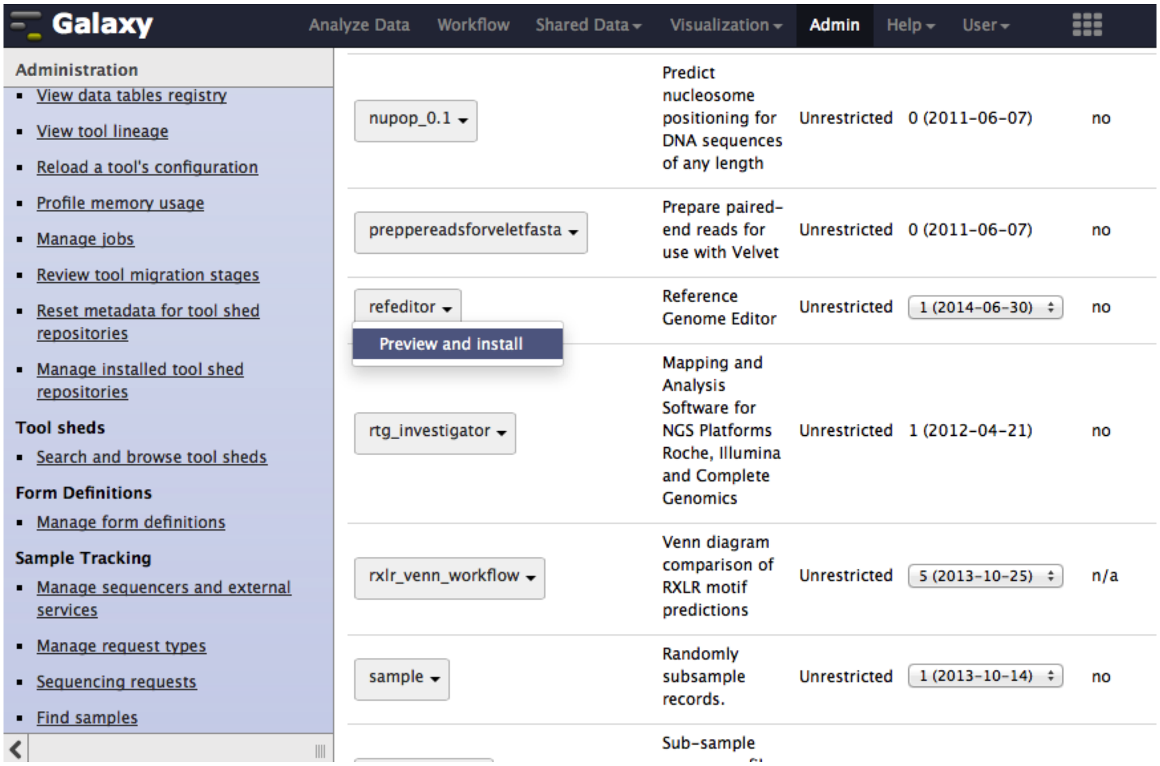


Figure 4.4: Galaxy Main Tool Shed and Test Tool Shed are already included with Galaxy's distribution. They can be found from Galaxy Admin Interface → *Search and browse tool sheds*

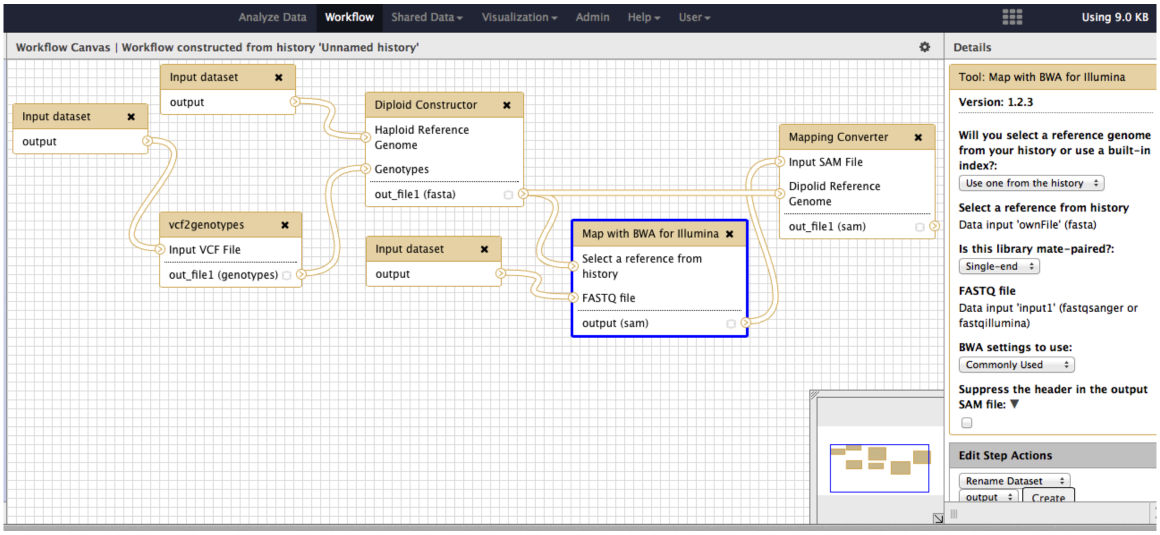


nupop_0.1	Predict nucleosome positioning for DNA sequences of any length	Unrestricted	0 (2011-06-07)	no
preppereadsforveletfasta	Prepare paired-end reads for use with Velvet	Unrestricted	0 (2011-06-07)	no
refeditor	Reference Genome Editor	Unrestricted	1 (2014-06-30)	no
rtg_investigator	Mapping and Analysis Software for NGS Platforms Roche, Illumina and Complete Genomics	Unrestricted	1 (2012-04-21)	no
rxlr_venn_workflow	Venn diagram comparison of RXLR motif predictions	Unrestricted	5 (2013-10-25)	n/a
sample	Randomly subsample records.	Unrestricted	1 (2013-10-14)	no
	Sub-sample			

Figure 4.5: RefEditor-Galaxy can be previewed and installed from *Galaxy Main Tool Shed* → *Fasta Manipulation* → *refeditor*

4.4 Use Case

Galaxy can automatically generate workflows from series of analyses steps. We use RefEditor-Galaxy and BWA_wrappers to conduct diploid read alignment based on the testing data included in RefEditor-Galaxy, and then extract the workflow from the operation history. This workflow, as shown in Figure 4.6, is the Galaxy version of RefEdit pipeline defined in Chapter 3. Users can manipulate the workflow to incorporate other tools (e.g. Filters).



Tool: Map with BWA for Illumina

Version: 1.2.3

Will you select a reference genome from your history or use a built-in index?: Use one from the history

Select a reference from history: Data input 'ownFile' (fasta)

Is this library mate-paired?: Single-end

FASTQ file: Data input 'input1' (fastqsanger or fastqillumina)

BWA settings to use: Commonly Used

Suppress the header in the output SAM file: ☐

Figure 4.6: RefEdit pipeline is implemented as a Galaxy workflow using RefEditor-Galaxy and BWA_wrappers. The latter can be installed from *Galaxy Main Tool Shed → Next Gen Mappers*

4.5 Discussion

Galaxy provides an integrated environment for bioinformatics analyses. It offers a general solution that enables a computational tool to be easily included in an analysis chain and run by scientists without programming experience. In this chapter we reviewed the implementation of RefEditor-Galaxy, a wrapper for RefEditor software package. A use case that implements RefEdit pipeline is also provided, and the equivalent workflow is extracted. To increase the accessibility of our tools, we put them into the Galaxy Main Tool Shed for easy downloading and installation.

We have not implemented the RefEdit+ pipeline in Galaxy platform since it entails that the imputation tools be integrated into Galaxy. However, once the imputation tools wrapper have been accomplished, RefEdit+ pipeline can be implemented in Galaxy without any change of RefEditor-Galaxy, since the RefEdit+ pipeline just requires additional invocations of genotype imputation tools.

Moving Forward, we want to conduct more experimental integrations on Galaxy platforms in order to make more meaningful bioinformatics discoveries.

Chapter 5

Conclusion

This dissertation has described a research effort and solution towards increasing read mapping ratio/accuracy, reducing ASE bias, improving genotype calling accuracy and SNP discovery ratio for diploid individuals through alternating the reference genome. We have developed an open source software package to achieve those goals conveniently. We also designed pipelines to utilize that software package.

5.1 Future Work

There still remain many unexplored areas worthy of investigation on this research topic. Some of them are listed here:

5.1.1 Dynamic reference genome

Currently, we only modify the reference genome once for each diploid individual. That strategy already showed benefits in multiple downstream analysis, including new SNP discovery. We wonder if those new SNPs have high enough quality to be incorporated into the next round of reference genome construction and further improvements on other aspects. If yes, then how to balance the quantity and quality of the SNPs used for diploid reference genome? How many rounds should we go through the iteration to create such a dynamic reference genome?

5.1.2 More formats support

The program *Mapping Converter* currently only supports SAM format. Although this is a very popular data format supported by many mapping tools, there are still some formats used by many mapping tools that worth to be supported, such as BAM, which is a compressed format mutually convertible with SAM. We can design *Mapping Converter* so that it automatically recognizes multiple file formats from the input, and then set the output format correspondingly, or use formats that users select. That requires, of course, RefEditor-Galaxy to be updated to allow additional parameters.

5.1.3 More studies

We did our study on one diploid species — human. We wonder if this strategy can also work effectively on other diploid species. We also want to conduct other WGS studies, such as ChIP-seq, BS-seq and imprinting studies. By integrating our methods with other tools in Galaxy, we are expecting more meaningful bioinformatics discoveries.

5.2 Summary

The benefits of using RefEditor to incorporate known genotypes and imputation technology have been apparently demonstrated in our studies. A strong integration of RefEditor and Galaxy platform has been accomplished by the implementation of RefEditor-Galaxy. The final result is a viable foundation for many future bioinformatical studies in NGS field.

Bibliography

[1] F Sanger, S Nicklen, and A R Coulson. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–7, Dec 1977.

[2] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012:11, 2012.

[3] Francis S Collins, Michael Morgan, and Aristides Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–90, Apr 2003.

[4] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, Sep 2005.

[5] Yun Li, Wei Chen, EricYi Liu, and Yi-Hui Zhou. Single nucleotide polymorphism (snp) detection and genotype calling from massively parallel sequencing (mps) data. *Statistics in Biosciences*, 5(1):3–25, 2013.

[6] G.E. Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, Jan 1998.

[7] Elaine R Mardis. Next-generation dna sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008.

[8] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009.

[9] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–17, Sep 2008.

[10] Peter J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, 2009.

[11] Shawn J Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215–9, Mar 2008.

[12] Ayat Hatem, Doruk Bozdag, Amanda Toland, and Umit Catalyurek. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1):184, 2013.

[13] Peter J A Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res*, 38(6):1767–71, Apr 2010.

[14] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 42(Database issue):D7–17, Jan 2014.

[15] Michael L. Metzker. Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, 11(1):31–46, 2010.

[16] David B. Goldstein, Andrew Allen, Jonathan Keebler, Elliott H. Margulies, Steven Petrou, Slave Petrovski, and Shamil Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet*, 14(7):460–470, 2013.

[17] L. G. Biesecker, W. Burke, I. Kohane, S. E. Plon, and R. Zimmern. Next-generation sequencing in the clinic: are we ready? *Nat Rev Genet*, 13(11):818–24, 2012.

[18] R. David Hawkins, Gary C. Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11(7):476–486, 2010.

[19] Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11(10):685–696, 2010.

[20] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, Nov 2008.

[21] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

[22] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[23] S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[24] Bin Ma, John Tromp, and Ming Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–5, Mar 2002.

[25] Michael C Schatz. Cloudburst: highly sensitive read mapping with mapreduce. *Bioinformatics*, 25(11):1363–9, Jun 2009.

[26] Nathan L Clement, Quinn Snell, Mark J Clement, Peter C Hollenhorst, Jahnvi Purwar, Barbara J Graves, Bradley R Cairns, and W Evan Johnson. The gnumap algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*, 26(1):38–45, Jan 2010.

[27] Hui Jiang and Wing Hung Wong. Seqmap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, 24(20):2395–6, Oct 2008.

[28] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, 2008.

[29] Nils Homer, Barry Merriman, and Stanley F. Nelson. Bfast: An alignment tool for large scale genome resequencing. *PLoS ONE*, 4(11):e7767, 2009.

[30] Nils Homer, Barry Merriman, and Stanley F Nelson. Local alignment of two-base encoded dna sequence. *BMC Bioinformatics*, 10:175, 2009.

[31] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome Biol*, 5(2):R12, 2004.

[32] Colin Meek, Jignesh M. Patel, and Shruti Kasetty. Oasis: An online and accurate technique for local-alignment searches on biological sequences. In *In VLDB*, pages 910–921, 2003.

[33] Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, 2(1):53 – 86, 2004.

[34] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, Sep 2009.

[35] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[36] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[37] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009.

[38] T W Lam, W K Sung, S L Tam, C K Wong, and S M Yiu. Compressed indexing and local alignment of dna. *Bioinformatics*, 24(6):791–7, Mar 2008.

[39] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–95, Mar 2010.

[40] Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Ratsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–80, Aug 2008.

[41] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–11, May 2009.

[42] Yuanxin Xi and Wei Li. Bsmap: whole genome bisulfite sequence mapping program. *BMC Bioinformatics*, 10:232, 2009.

[43] Pao-Yang Chen, Shawn J Cokus, and Matteo Pellegrini. Bs seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 11:203, 2010.

[44] Nuno A. Fonseca, Johan Rung, Alvis Brazma, and John C. Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 2012.

[45] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[46] Lincoln D Stein, Christopher Mungall, ShengQiang Shu, Michael Caudy, Marco Mangone, Allen Day, Elizabeth Nickerson, Jason E Stajich, Todd W Harris, Adrian Arva, and Suzanna Lewis. The generic genome browser: a building block for a model organism system database. *Genome Res*, 12(10):1599–610, Oct 2002.

[47] Heinrich Magnus Manske and Dominic P Kwiatkowski. Lookseq: a browser-based viewer for deep sequencing data. *Genome Res*, 19(11):2125–32, Nov 2009.

[48] Iain Milne, Micha Bayer, Linda Cardle, Paul Shaw, Gordon Stephen, Frank Wright, and David Marshall. Tablet–next generation sequence assembly visualization. *Bioinformatics*, 26(3):401–2, Feb 2010.

[49] Tim Carver, Ulrike Böhme, Thomas D Otto, Julian Parkhill, and Matthew Berriman. Bamview: viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics*, 26(5):676–7, Mar 2010.

[50] Tomas Babak, Brian Deveale, Christopher Armour, Christopher Raymond, Michele A Cleary, Derek van der Kooy, Jason M Johnson, and Lee P Lim. Global survey of genomic imprinting by transcriptome sequencing. *Curr Biol*, 18(22):1735–41, Nov 2008.

[51] D. Serre, S. Gurd, B. Ge, R. Sladek, D. Sinnett, E. Harmsen, M. Bibikova, E. Chudin, D. L. Barker, T. Dickinson, J. B. Fan, and T. J. Hudson. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet*, 4(2):e1000006, 2008.

[52] Xu Wang, Qi Sun, Sean D McGrath, Elaine R Mardis, Paul D Soloway, and Andrew G Clark. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One*, 3(12):e3839, 2008.

[53] Jacob F. Degner, John C. Marioni, Athma A. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics*, 25(24):3207–3212, 2009.

[54] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey

Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[55] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–5, Sep 2009.

[56] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and Elaine R Mardis. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 6(9):677–81, Sep 2009.

[57] Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[58] J. C. Knight. Allele-specific gene expression uncovered. *Trends Genet*, 20(3):113–6, 2004.

[59] L. Milani, A. Lundmark, J. Nordlund, A. Kiialainen, T. Flaegstad, G. Jonmundsson, J. Kanerva, K. Schmiegelow, K. L. Gunderson, G. Lonnerholm, and A. C. Syvanen. Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by cpg site methylation. *Genome Res*, 19(1):1–11, 2009.

[60] Patricia J. Wittkopp, Belinda K. Haerum, and Andrew G. Clark. Independent effects of cis- and trans-regulatory variation on gene expression in drosophila melanogaster. *Genetics*, 178(3):1831–1835, 2008.

[61] J. Ronald, J. M. Akey, J. Whittle, E. N. Smith, G. Yvert, and L. Kruglyak. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res*, 15(2):284–91, 2005.

[62] Hai Yan, Weishi Yuan, Victor E. Velculescu, Bert Vogelstein, and Kenneth W. Kinzler. Allelic variation in human gene expression. *Science*, 297(5584):1143, 2002.

[63] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Meth*, 5(7):621–628, 2008.

[64] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[65] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs, and E. E. Eichler. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*, 41(10):1061–7, 2009.

[66] Faraz Hach, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp. mrsfast: a cache-oblivious algorithm for short-read mapping. *Nat Meth*, 7(8):576–577, 2010.

[67] K. D. Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):D501–4, 2005.

[68] Frederick E. Dewey, Rong Chen, Sergio P. Cordero, Kelly E. Ormond, Colleen Caleshu, Konrad J. Karczewski, Michelle Whirl-Carrillo, Matthew T. Wheeler, Joel T. Dudley, Jake K. Byrnes, Omar E. Cornejo, Joshua W. Knowles,

Mark Woon, Katrin Sangkuhl, Li Gong, Caroline F. Thorn, Joan M. Hebert, Emidio Capriotti, Sean P. David, Aleksandra Pavlovic, Anne West, Joseph V. Thakuria, Madeleine P. Ball, Alexander W. Zaranek, Heidi L. Rehm, George M. Church, John S. West, Carlos D. Bustamante, Michael Snyder, Russ B. Altman, Teri E. Klein, Atul J. Butte, and Euan A. Ashley. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet*, 7(9):e1002280, 2011.

[69] A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[70] Consortium International HapMap, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, 2007.

[71] Ravi Vijaya Satya, Nela Zavaljevski, and Jaques Reifman. A new strategy to reduce allelic bias in rna-seq readmapping. *Nucleic Acids Research*, 2012.

[72] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Meth*, 9(4):357–359, 2012.

[73] Stephen M. Rumble, Phil Lacroute, Adrian V. Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: Accurate mapping of short color-space

reads. *PLoS Comput Biol*, 5(5):e1000386, 2009.

[74] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[75] Consortium Genomes Project, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[76] Jacob A Tennessen, Abigail W Bigham, Timothy D O'Connor, Wenqing Fu, Eimear E Kenny, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–9, Jul 2012.

[77] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, 2007.

[78] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–7, 2007.

[79] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007.

[80] M. Buchkovich, K.L. Mohlke, and T.S. Furey. Removal of mapping biases in sequence-based functional data improves regulatory element identification at heterozygous variants. 2012.

[81] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[82] Dalila Pinto, Katayoon Darvishi, Xinghua Shi, Diana Rajan, Diane Rigler, Tom Fitzgerald, Anath C. Lionel, Bhooma Thiruvahindrapuram, Jeffrey R. MacDonald, Ryan Mills, Aparna Prasad, Kristin Noonan, Susan Gribble, Elena Prigmore, Patricia K. Donahoe, Richard S. Smith, Ji Hyeon Park, Matthew E. Hurles, Nigel P. Carter, Charles Lee, Stephen W. Scherer, and Lars Feuk. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotech*, 29(6):512–520, 2011.

[83] Network The Cancer Genome Atlas Research, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle

Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–1120, 2013.

[84] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–34, 2010.

[85] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10(1):387–406, 2009.

[86] E. Y. Liu, M. Li, W. Wang, and Y. Li. Mach-admix: genotype imputation for admixed populations. *Genet Epidemiol*, 37(1):25–37, 2013.

[87] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913, 2007.

[88] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.

[89] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, 44(8):955–9, 2012.

[90] B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84(2):210–23, 2009.

[91] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, 2010.

[92] Eleonora A. M. Festen, Philippe Goyette, Todd Green, Gabrielle Boucher, Claudine Beauchamp, Gosia Trynka, Patrick C. Dubois, Caroline Lagacé, Pieter C. F. Stokkers, Daan W. Hommes, Donatella Barisani, Orazio Palmieri, Vito Annese, David A. van Heel, Rinse K. Weersma, Mark J. Daly, Cisca Wijmenga, and John D. Rioux. A meta-analysis of genome-wide association scans identifies il18rap, ptpn2, tagap, and pus10 as shared risk loci for crohn's disease and celiac disease. *PLoS Genet*, 7(1):e1001283, 2011.

[93] Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred.i. accuracyassessment. *Genome Research*, 8(3):175–185, 1998.

[94] Thomas D. Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

[95] The international hapmap project. *Nature*, 426(6968):789–96, 2003.

[96] Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[97] Madeleine P. Ball, Joseph V. Thakuria, Alexander Wait Zaranek, Tom Clegg, et al. A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences*, 109(30):11920–11927, 2012.

[98] Hong Li, Gustavo Glusman, Chad Huff, Juan Caballero, and Jared C. Roach. Accurate and robust prediction of genetic relationship from whole-genome sequences. *PLoS ONE*, 9(2):e85437, 2014.

[99] Jared C. Roach, Gustavo Glusman, Arian F. A. Smit, Chad D. Huff, Robert Hubley, Paul T. Shannon, Lee Rowen, Krishna P. Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Radoje Drmanac, Lynn B. Jorde, Leroy Hood, and David J. Galas. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.

[100] E. Y. Liu, S. Buyske, A. K. Aragaki, U. Peters, E. Boerwinkle, C. Carlson, C. Carty, D. C. Crawford, J. Haessler, L. A. Hindorff, L. L. Marchand, T. A.

Manolio, T. Matise, W. Wang, C. Kooperberg, K. E. North, and Y. Li. Genotype imputation of metabochip snps using a study-specific reference panel of 4,000 haplotypes in african americans from the women's health initiative. *Genet Epidemiol*, 36(2):107–17, 2012.

[101] P. L. Auer, J. M. Johnsen, A. D. Johnson, B. A. Logsdon, L. A. Lange, M. A. Nalls, G. Zhang, N. Franceschini, K. Fox, E. M. Lange, S. S. Rich, C. J. O'Donnell, R. D. Jackson, R. B. Wallace, Z. Chen, T. A. Graubert, J. G. Wilson, H. Tang, G. Lettre, A. P. Reiner, S. K. Ganesh, and Y. Li. Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in african americans: Nhlbi go exome sequencing project. *Am J Hum Genet*, 91(5):794–808, 2012.

[102] Chih-Cheng Su, Tzong-Zeng Wu, Li-Kuang Chen, Hui-Hua Yang, and Dar-Fu Tai. Development of immunochips for the detection of dengue viral antigens. *Analytica Chimica Acta*, 479(2):117–123, 2003.

[103] Benjamin F. Voight, Hyun Min Kang, Jun Ding, Cameron D. Palmer, Carlo Sidore, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*, 8(8):e1002793, 2012.

[104] David Gresham, Maitreya J. Dunham, and David Botstein. Comparing whole genomes using dna microarrays. *Nat Rev Genet*, 9(4):291–302, 2008.

[105] Neekesh V. Dharia, A. Taylor Bright, Scott J. Westenberger, S. Whitney Barnes, Serge Batalov, Kelli Kuhen, Rachel Borboa, Glenn C. Federe, Colleen M. McClean, Joseph M. Vinetz, Victor Neyra, Alejandro Llanos-Cuentas, John W. Barnwell, John R. Walker, and Elizabeth A. Winzeler. Whole-genome sequencing and microarray analysis of ex vivo plasmodium vivax reveal selective pressure on putative drug resistance genes. *Proceedings of the National Academy of Sciences*, 107(46):20045–20050, 2010.

[106] J. Marchini on behalf of the Haplotype Consortium. A haplotype map derived from whole genome low-coverage sequencing of over 25,000 individuals. 2013.

[107] Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb Miller, W James Kent, and Anton Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15(10):1451–5, Oct 2005.

[108] James Taylor, Ian Schenck, Dan Blankenberg, and Anton Nekrutenko. Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinfor-*

matics, Chapter 10:Unit 10.5, Sep 2007.

[109] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:Unit 19.10.1–21, Jan 2010.

[110] Jeremy Goecks, Anton Nekrutenko, James Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[111] Daniel Blankenberg, James Taylor, Ian Schenck, Jianbin He, Yi Zhang, Matthew Ghent, Narayanan Veeraraghavan, Istvan Albert, Webb Miller, Kateryna D Makova, Ross C Hardison, and Anton Nekrutenko. A framework for collaborative analysis of encode data: making large-scale analyses biologist-friendly. *Genome Res*, 17(6):960–4, Jun 2007.

[112] Public galaxy service. [http://usegalaxy.org].

[113] Christoph Bock, Greg Von Kuster, Konstantin Halachev, James Taylor, Anton Nekrutenko, and Thomas Lengauer. Web-based analysis of (epi-) genome data using epigraph and galaxy. In Michael R. Barnes and Gerome Breen, editors,

Genetic Variation, volume 628 of *Methods in Molecular Biology*, pages 275–296. Humana Press, 2010.

[114] Ryota Kikuchi, Shintaro Yagi, Hiroyuki Kusuhara, Satoki Imai, Yuichi Sugiyama, and Kunio Shiota. Genome-wide analysis of epigenetic signatures for kidney-specific transporters. *Kidney Int*, 78(6):569–77, Sep 2010.

[115] Kyle J Gaulton, Takao Nammo, Lorenzo Pasquali, Jeremy M Simon, Paul G Giresi, Marie P Fogarty, Tami M Panhuis, Piotr Mieczkowski, Antonio Secchi, Domenico Bosco, Thierry Berney, Eduard Montanya, Karen L Mohlke, Jason D Lieb, and Jorge Ferrer. A map of open chromatin in human pancreatic islets. *Nat Genet*, 42(3):255–9, Mar 2010.

[116] Thomas Hentrich, Julia M Schulze, Eldon Emberly, and Michael S Kobor. Chromatra: a galaxy tool for visualizing genome-wide chromatin signatures. *Bioinformatics*, 28(5):717–8, Mar 2012.

[117] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M Rubin, and Len A Pennacchio. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–8, Feb 2009.

[118] Shahaf Peleg, Farahnaz Sananbenesi, Athanasios Zovoilis, Susanne Burkhardt, Sanaz Bahari-Javan, Roberto Carlos Agis-Balboa, Perla Cota, Jessica Lee Wittnam, Andreas Gogol-Doering, Lennart Opitz, Gabriella Salinas-Riester, Markus Dettenhofer, Hui Kang, Laurent Farinelli, Wei Chen, and André Fischer. Altered histone acetylation is associated with age-dependent memory impairment in mice. *Science*, 328(5979):753–756, 2010.

[119] Daniel Blankenberg, Assaf Gordon, Gregory Von Kuster, Nathan Coraor, James Taylor, Anton Nekrutenko, and Galaxy Team. Manipulation of fastq data with galaxy. *Bioinformatics*, 26(14):1783–5, Jul 2010.

[120] Daniel Blankenberg, Nathan Coraor, Gregory Von Kuster, James Taylor, Anton Nekrutenko, and Galaxy Team. Integrating diverse databases into an unified analysis framework: a galaxy approach. *Database (Oxford)*, 2011:bar011, 2011.

[121] Donna Karolchik, Angie S Hinrichs, and W James Kent. The ucsc genome browser. *Curr Protoc Bioinformatics*, Chapter 1:Unit1.4, Dec 2012.

[122] Cristina Aurrecoechea, John Brestelli, Brian P Brunk, Steve Fischer, Bindu Gajria, et al. Eupathdb: a portal to eukaryotic pathogen databases. *Nucleic Acids Res*, 38(Database issue):D415–9, Jan 2010.

[123] Belinda Giardine, Sjozef van Baal, Polynikis Kaimakis, Cathy Riemer, Webb Miller, Maria Samara, Panagoula Kollia, Nicholas P Anagnou, David H K Chui, Henri Wajcman, Ross C Hardison, and George P Patrinos. Hbvar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Hum Mutat*, 28(2):206, Feb 2007.

[124] Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P Mesirov. Genepattern 2.0. *Nat Genet*, 38(5):500–1, May 2006.

[125] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22):3005–11, Nov 2009.

[126] Daniel Blankenberg, Gregory Von Kuster, Emil Bouvier, Dannon Baker, Enis Afgan, Nicholas Stoler, the Galaxy Team, James Taylor, and Anton Nekrutenko. Dissemination of scientific software with galaxy toolshed. *Genome Biology*, 15(2):403, 2014.