Joint Biotechnology Master Program

Palestine Polytechnic University
Deanship of Graduate Studies
and Scientific Research

Bethlehem University
Faculty of Science

PalHap, the first Palestinian haplotype exome panel: application in phasing compound heterozygous mutations

By

Reena Saeed

In Partial Fulfillment of the Requirements for the Degree

Master of Biotechnology

October 2020

The undersigned hereby certify that they have read and recommend to the Faculty of Scientific Research and Higher Studies at the Palestine Polytechnic University and the Faculty of Science at Bethlehem University for acceptance a thesis entitled:

PalHap, the first Palestinian haplotype exome panel: application in phasing compound heterozygous mutations

by
Reena Saeed

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in biotechnology.
Graduate Advisory Committee:

Committee Member (Student's Supervisor) *Prof. Dr. Fouad Zahdeh– Bethlehem University*	Date
Committee Member (Internal Examiner) *Prof. Dr. Moien Kanaan – Bethlehem University*	Date
Committee Member (External Examiner) *Dr. – Saad Lahham- Al-Najah National University*	Date

Approved for the Faculties

Dean of Graduate Studies and Scientific Research Palestine Polytechnic University	Dean of Faculty of Science Bethlehem University
Date	Date

PalHap, the first Palestinian haplotype exome panel: application in phasing compound heterozygous mutations

by Reena Saeed

ABSTRACT

Next Generation Sequencing (NGS) technology is widely used in clinical diagnosis to identify disease-causing variants. We have utilized NGS data and constructed parental haplotypes surrounding the mutation. Haplotype phasing would facilitate the identification of compound heterozygous mutation in which two distinct haplotypes each harbor a unique sequence variant. The cis- and trans-acting configurations of a compound heterozygous genotype influences gene expression and its potential functional outcome. Classical Mendelian methods are usually used to infer haplotypes phasing which requires genotyping, of all family members. Population-based phasing is currently a promising computational alternative and is shown to produce high phasing accuracy especially when ethnicity-matching haplotype reference panel is used. In this study we built the first Palestinian haplotype reference panel (PalHap) from 600 Palestinians Whole-Exomes sequenced at the Molecular Genetic Lab, Istishari Arab Hospital using SHAPEIT, a fast population-based phasing algorithm. We showed that PalHap with a sample number less than half of that in the international 1000Genomes reference panel, can outperform 1000Genomes in phasing compound heterozygous mutations in polymorphic CNVs regions. This study stands as the base for future works that may make use of PalHap in other applications such as Preimplantation Genetic Diagnosis (PGD).

أول لوحة أكسوم مرجعية فلسطينية للأنماط الفردانية: تطبيقات في استدلال الطفرات غير المتجانسة المركبة

PalHap

الملخص

تستخدم تقنية تسلسل الجيل التالي (NGS) على نطاق واسع في التشخيص السريري لتحديد المتغيرات أو الطفرات المسببة للأمراض. نستخدم بيانات NGS من أجل بناء الأنماط الفردية (haplotypes) الأبوية المحيطة بالطفرة. يسهل النمط الفرداني الكشف عن الطفرات غير المتجانسة المركبة التي تحتوي على نسختين مختلفتين كل منهما على أليلين مختلفين على نفس الجين. تتواجد هذه الطفرات أما بترتيب متقارب (in cis) أو بترتيب متباعد (in trans) و تؤثر على التعبير الجيني ونتائجه الوظيفية المحتملة. عادة، تُستخدم الطرق المندلية الكلاسيكية لاستنتاج النمط الفرداني و التي تتطلب النمط الجيني لجميع أفراد الأسرة. تعد الدراسات السكانية القائمة على النهج الحاسوبي هو حاليًا بديلاً حسابيًا واعداً. وتبين أنها تنتج الأنماط الفردية بدقة عالية، خاصة عند استخدام لوحة سكانية مرجعية مطابقة للأصل العرقي. في هذه الدراسة، قمنا ببناء أول لوحة مرجعية فلسطينية للأنماط الفردانية (PalHap) مكونة من عينات التسلسل الاكسومي (exome) الكامل من ٦٠٠ فرد فلسطيني في مختبر الوراثة الجزيئية- مستشفى الاستشاري العربي باستخدام SHAPEIT، و هي خوارزمية تنميط سريعة تعمل على مستوى السكاني. لقد أظهرنا أن PalHap، على الرغم من امتلاك أقل من نصف العينات في لوحة المرجعية 1000 Genomes العالمية، يمكنها أن تستنتج الطفرات غير المتجانسة المركبة بين الفلسطينيين في في بعض المناطق CVNs بدقة تتفوق على 1000 Genomes. تعتبر هذه الدراسة بمثابة الأساس للعديد من الأعمال المخبرية المستقبلية التي قد تستخدم PalHap في تطبيقات مثل التشخيص الوراثي قبل الزراعة (PGD).

DECLARATION

I declare that the Master Thesis entitled "dissertation title" is my own original work, and hereby certify that unless stated, all work contained within this thesis is my own independent research and has not been submitted for the award of any other degree at any institution, except where due acknowledgment is made in the text.

Name and signature: _____"Reena Saeed_____"____

Date_____

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for the joint master degree in biotechnology at Palestine Polytechnic University and Bethlehem University, I agree that the library shall make it available to borrowers under rules of the library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of the source is made.

Permission for extensive quotation from, reproduction, or publication of this thesis may be granted by my main supervisor, or in [his/her] absence, by the Dean of Higher Studies when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature: _____"Reena Saeed_____"_____

Date_____

Acknowledgments

I would like to express my deepest gratitude to Dr. Fouad Zahdeh for suggesting this work and his supervision. It is a great honor to work under his supervision. I am forever thankful to team of Molecular Genetic Lab at Istishari Arab Hospital and Hereditary Research laboratory personnel at Bethlehem University for without their contribution to this work and their support, I wouldn't be where I am. I would also like to thank Professor Dr. Moein Kanaan and Dr. Saad Lahham.

Abbreviations

NGS	Next Generation Sequencing
PGD	Preimplantation Diagnosis
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
CNV	Copy Number Variation
LD	Linkage Disequilibrium
HGP	Human Genome Project
EM	Expectation-Maximization Algorithm
HMM	Hidden Markov Model
IBD	Identical By Descent
HapCUT2	Haplotype Assembly Coverage Handling by Adapting Thresholds
CPU	Central Processing Unit
nPAR Genes	Non-pseudoautosomal Genes
WGS	Whole Genome Sequencing
WES	Whole Exome Sequencing
VCF file	Variant Call Format File
ALT	Alternate Allele
REF	reference allele
Qual	Quality Score
DP	Depth at Poisson Distribution (Depth of Coverage)
GQ	Genotype Quality
BQ	Base Quality

List of Figures

List of Tables

Table of Content

CHAPTER 1

Introduction to Population Genomics

1.1 Genetic Variations

Mendelian laws of inheritance describes the fundamental principles that govern transmission of inherited traits from parents to offspring. Genomic sequences are exposed to alterations and can be transmitted to offspring. The alterations include single nucleotide substitutions such as point mutations or Single nucleotide polymorphisms (SNPs) that are common variants in a population at a frequency of more than 1%. SNPs are utilized as genetic markers to mark and detect variations in close proximity that cause or may contribute to certain traits or diseases. In other words, SNPs help to determine the causative mutations in the DNA sequence whether being themselves or mutations co-linked to already identified SNPs. Alterations also include indels which are insertion or deletion of a small number of nucleotides. Structural variations however are alterations that involve large DNA segments (larger than 1 kilobases) such as copy number variations (CNVs), inversions, and translocations (Qi et al., 2014). CNV is a change in the normal two-copy number of a particular DNA segment in the diploid genome. Less than two copies is called a "loss" and more than two copies is called "gain". In the human genome, CNV lengths range from kilobases to several megabases. They cover cumulatively 35% of the genome (Database of Genomic Variants, DGV) and approximately 12% of the genome is subject to copy number variation (Redon et al., 2006). CNVs play a role in several human phenotypic traits, such as disease susceptibility (International HapMap Consortium et al., 2010) and neuropsychiatric disorders (Beroukhim et al., 2010). They also overlap and interfere with genes that will change the functional and phenotypic traits (e.x. alter coding regions, change DNA methylation, expression patterns; Henrichsen et al., 2009; Brahmachary et al., 2014). CNVs occur in the

population at different frequencies, if the frequency is fewer than 1%, then it is considered a rare CNV, and if the frequency is greater than 1%, it is considered a common or polymorphic CNV (Valsesia et al., 2013). Also, CNVs, as many SNPs, occur at different frequencies across populations due to so-called population stratification (difference in allele frequencies among populations of different genetic ancestry). Therefore, several causal CNVs are associated with continental ancestries (Khaja et al., 2006; Redon et al., 2006)

1.2 Recombination and Linkage Disequilibrium

Recombination is the exchange of genetic material between homologous chromosomes during the formation of the gametes producing a recombinant chromosome. Each parent passes gametes containing one copy of each chromosome to their child. If no recombination occurred along the haploid copy of the chromosome, all alleles remain linked on a single chromosome. If recombination occurred, segments exchange between both homologous chromosomes to the gametic recombinant chromosome. The yielded recombinant chromosome may possess allelic variations not seen in the parental genome (Kong et al., 2002).

Recombination rate across genomic regions on the chromosomes varies upon the distance correlation between co-linked variants and referred to as linkage disequilibrium (LD). The bigger the LD between these variants, the greater the probability of recombination to occur between two variants or set of variants (Kong et al., 2002). Thus, variants at close LD are more likely to remain co-linked compared to variants that are far apart in distance in which they likely to segregate in the two homologous chromosomes. Therefore, as recombination rate increases, LD decays in which recombination disrupts the physical linkage between allele segments of which

leads to the exchange of alleles between homologous chromosomes. Recombination is an evolutionary factor that increases the genetic diversity in the genome.

The concept of LD between variants (such as SNPs, see Section 1.1) is applied and has an important role in the current methods for detecting and mapping diseases associated variants. The associated mapping is reliable in identifying known SNPs at LD with unknown mutant alleles in disease susceptibility regions.

1.3 Mutation Detection and Whole-Exome Sequencing

Detection of mutation is of great significance for the diagnosis of genetic diseases, conformational diagnosis, pre-symptomatic testing as well as Preimplantation Genetic Diagnosis (PGD). The latter is a new genetic workflow that is gaining interest among low income countries used to determine genetic defects within embryos created manually in laboratories before implementation in case the parents carry a known genetic abnormality.

While cytogenetics methods capture mega structural variations, DNA sequencing methods identify alterations at the level of base pair. The classical Sanger sequencing utilizes oligonucleotide primers to elongate specific DNA regions of up to 900 base pairs (Morozova and Marra, 2008). While it shows successes in detecting point mutations and indel mutations (Totomoch-Serra et al., 2017), Sanger sequencing is weak in identification of low frequent mutations (up to 20%) due to limitations in the sensitivity (Hagemann, 2015). Moreover, it can be costly and labor intensive if used to sequence multiple regions. As a result, its usage is currently limited to sequence only specific regions. Next-generation sequencing (NGS) is a high-throughput sequencing technology that allows for rapid and parallel sequencing of genomes (i.e. whole genome sequencing WGS), and exomes (i.e. whole exome sequencing WES). The

massive and parallel high-throughput output of NGS reduces time and cost compared to classical Sanger sequencing.

WES is designed to capture sequences of exonic or protein-coding regions in the human genome. In spite of making only 1.5% of the whole genome, around 85% of disease-causing variants fall within the coding regions (Wang et al., 2013). Therefore, WES became a standard method in clinical genetic practice for identifying causative genetic variants of human diseases (Yang, et al., 2011). It is a powerful tool, cost-effective and labor-saving compared to WGS which rather attempts to sequence the entire genome including protein-coding and protein-non coding regions. WES is composed of several steps (Figure 1.2). It starts by shearing the purified samples of DNA into shorter fragments followed by library preparation which involves ligation of DNA fragments and specialized adaptor and followed by PCR amplification. Libraries are then hybridized to biotinylated probes specific to the target exon regions and fragments are enriched with capture beads. Fragments are finally amplified and selectively sequenced, producing sequence-ready targets. Sequence reads are then aligned to a reference genome through the bioinformatics pipeline for variant detection.

The Bioinformatics workflow starts with a basecalling step which identifies and calls individual bases on the reads so that read data are ready for mapping to a reference genome, which is a DNA sequence dataset assembled as a representative example of a human's set of genes. Then, assess the quality of raw reads to remove poor quality reads and base calling errors by using tools such as FASTQC. Next, align exome reads using read aligner (e.x. BWA, Bowite; Kumaran et al., 2019) followed by refining particular sites on the reads to corresponding sites in the reference genome to yield optimal alignment of reads. The aligned read results in Sequence Alignment Map (SAM) format and converts it to the Binary Alignment Map (BAM) format for

the reason of a more manageable file size (Li et al., 2009). Then a post-alignment processing step to remove duplicated aligned reads that could result from PCR artifacts. Variant calling then takes place where each site on the exome reads compared to the reference sequences using calling tools such as GATK, SAMTools and others (Kumaran et al., 2019), to detect variants, short indels or de-novo mutations through statistical models which help improve the reliability of variant predictions. Lastly, variants annotated to provide information about gene position, variant coordination and mutation type.

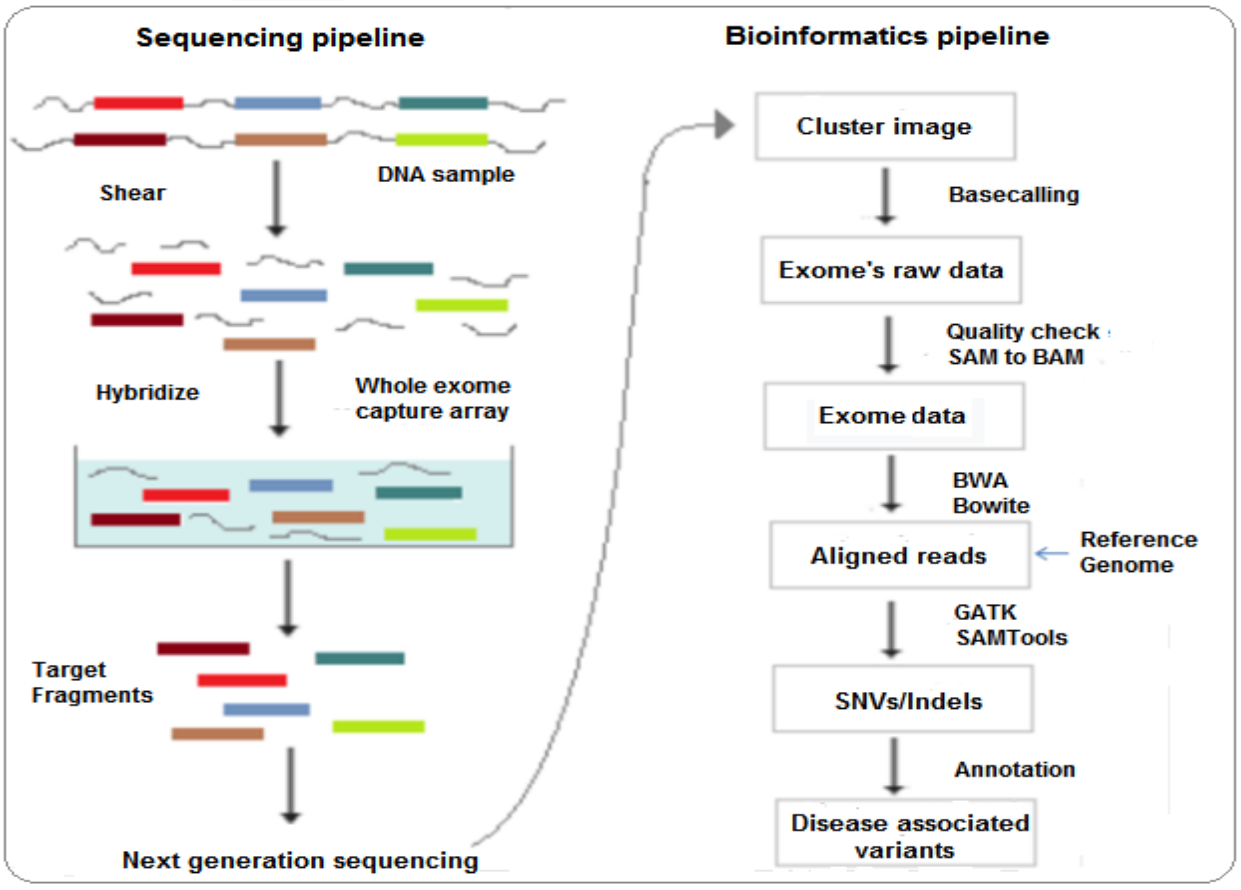

Figure 1.2. Overview of the whole exome sequencing pipeline. Illustrates the sequencing pipeline and bioinformatics pipeline to generate annotated variants of WES.

1.4 NGS Quality Control

NGS, just like Sanger sequencing, suffers from sequencing errors. The average NGS sequencing error rate is more than 0.1% per nucleotide (Mardis, 2013) in which they are falsely reported as single nucleotide substitutions (Fox and Reid-Bayliss, 2014). For example, if one exome produces 100,000 variants, then the 0.1% is 100 variants that are junk. Sequencing error occurs due to multiple limitations in sequencing technology and bioinformatic analysis. Technical errors occur during sample preparation, library preparation, or early synthesis cycles of PCR, may lead to the incorporation of incorrect bases and also by sequence amplification biases or failure of detecting the variant allele (i.e. allele dropout; Adey, 2017). Bioinformatic errors happen during sequence imaging where it fails to capture base incorporation due to base-colored signal decay or can misaligning sequences with respect to the reference sequence leads to penalize alternative alleles (i.e fewer alternative alleles would be aligned as reference alleles; Pereira et al., 2020). Failure to report incorporated bases or false variant calls affect the reliability and accuracy of aligned sequence reads and leads to an invalid analysis and diagnosis. Therefore, data filtering and quality check are of great importance to eliminate errors and biases in the sequence reads. Several tools for data quality checks are available, e.g. PLINK toolkit among them and most widely used. PLINK is an open source, user-friendly whole genome data analysis software. This toolkit is designed to perform large-scale genome-wide analyses such as per-SNP and per-individual, summary statistics, population stratification, and identity-by-descent estimation. The latter is important in inferring the degree of relatedness between individuals. In general, PLINK manipulates and analyzes large data sets of thousands of samples to make basic statistics of quality control for variants in genetic data in a computationally efficient manner (Purcell et al., 2007).

CHAPTER 2

Haplotype Phasing

2.1 Background

In 2001, Human Genome Project (HGP) was launched to sequence the first human genome (Venter et al, 2001). They used Sanger sequencing to determine sequences of relatively short fragments from human DNA and continued to sequence along these fragments by multiple rounds of sequencing, and they then aligned these fragments based on overlapping ends to assemble larger sequences of the DNA regions and eventually obtain the entire chromosomes (Levy et al., 2007). Sanger sequencing is extremely expensive and despite several efforts to reduce the cost, it was at the expense of providing incomplete haplotypes of the chromosomes (Suk et al., 2011). Later, cost effective sequencing technologies such as short reads NGS emerged.

NGS sequencing produces genetic sequences of the human diploid genome (two copies of the allele) in the form of unordered sets of alleles in which we cannot infer the chromosome of origin for each allele. In other words, NGS is suitable for inferring the genotypes (Figure 2.1). However, haplotyping or the identification of the ordered set of alleles on a single chromosome requires further assembling of the identified alleles based on their paternal origin (Druet et al., 2010). Genotype data reveal variants at homozygous and heterozygous states, but it fails to determine whether a heterozygous variant is inherited from the paternal or the maternal origin. In addition, heterozygous variants manifest themselves in the genetic mutations in various states and it is crucial to assign heterozygous variants observed in the offspring to the right parental origin for accurate diagnosis (section 2.2).

-------G----------A-------A----------C----------T------

-------G----------C-------A----------T----------T------

Figure 2.1 Two sequences on the same region of two homologous copies of a chromosome.
Variant sites are shown only and non variant sites are labeled with " -". Five variants are shown. In this example, the genotypes are {G,G} {A,C} {A,A} {C,T} {T,T} and the two haplotypes are GAACT and GCATT.

Haplotype phasing (i.e. identifying the haplotypes from genotyped sequences) requires the construction of a mendelian pedigree from family members who are affiliated with the disease. For this, Mendelian segregations can be used. Besides being an exhausting process in terms of cost, time and family members' availability or approval to participate, Mendelian methods cannot phase some compound heterozygous mutations in particular those that are heterozygous in both parents. Figure 2.2 shows different states for mutation heterozygosity on a gene. Mendelian pedigree cannot phase compound heterozygous mutations where two different mutations harbor two different gene loci. This type of complex compound heterozygous mutation can be revealed only if the haplotype of the parents were identified.

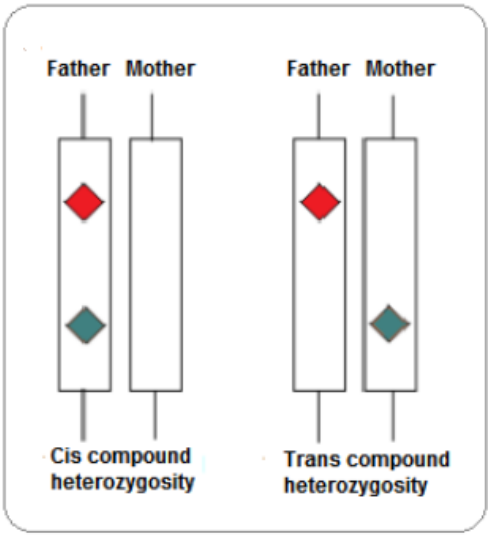

Figure 2.2 Different states of compound heterozygosity on a gene locus. The states of two different mutations that harbor two distinct loci on same transmitted allele (in cis configuration) or two different mutations harbor two distinct loci on both alleles (in trans configuration) remain unsolved. Mendelian phasing cannot phase these states of compound heterozygosity. Two triangles present two gene copies each on a homologous chromosome, and lozenge presents mutation and different colors for different mutations.

2.2 Clinical Importance of Haplotyping

Haplotype information provides accurate representation of the genetic variations and history of mutations in nuclear families as well as at the population level. Because haplotypes are blocks of variant alleles, they minimize the chance of allele dropout (allelic loss during PCR amplification step) compared to single targeted methods in clinical genetic tests. This advantage makes it valuable in disease diagnosis (Browning et al., 2011).

Phasing not only facilitates genetic diagnosis, but it also helps in identifying regions that recently underwent positive selection. This is because an allele which is under positive selection spreads quickly among individuals and reduces recombination by which alleles remain co-linked on the genome region. Therefore, haplotype blocks of co-linked allele that are abundant in a population is an indication of recent positive selection of beneficial alleles (Browning et al., 2011). The lactase persistence allele at the LCT gene that makes lactase enzyme is an example of recent positive selection. The allele acquires the ability to digest lactose in dairy products not only in childhood as in the ancient populations, but the allele activity persists into adulthood (Scrimshaw and Murray, 1988). Several studies have demonstrated that the allele lies on a long haplotype of more than 1 million base pairs (Bersaglieriet et al, 2004), which shows that it is under positive selection.

Haplotyping of Y-chromosome, since it is single and unique, provides delightful information about human history. Specific variants mark historic migrations events which helped in grouping Y-chromosome in several haplogroups (Narasimhan et al., 2017). For example, Palestinians belong to two main haplogroups: E and J, and a minority to haplogroup G (Fernandes et al., 2011). Haplogroups are groups of similar haplotypes at different chromosomal regions that descended from the same ancestry (Arora et al., 2015). Haplogroups of Y-chromosome are most

studied and utilized to define the genetic variations between populations (Semino et al., 2000). The Y-DNA is inherited solely from father to son, hence, changes due to recombination and mutations are by chance at each generation (Karafet et al., 2008) and thus it carries the largest non-recombining haplotype blocks and it is utilized in several applications. Such as genealogical reconstruction which seeks to construct the lineages of each ancestor by information of families history and trace their lineages, in evolutionary population genetics which studies the genetic differences and influences within and between populations, and also in medical genetics and forensics (Underhill and Kivisild, 2007).

Beside all benefits mentioned above, haplotyping is mainly used to provide better understanding of gene functionality, regulation, and association between genetic variants and diseases. One example is compound heterozygous mutations. Multiple mutations can alter gene copies on homologous chromosome pairs in various ways. If two mutations are located on the same chromosome or in other words the same haplotype (i.e. cis configuration) only one copy of the gene would be affected. However, if the two mutations are located on the two homologous chromosomes (i.e. trans configuration), both gene copies would be affected (figure 2.2). The trans configurations in clinical genetics have the same effect as homozygous mutations. Several disorders caused by pathogenic compound heterozygous mutations are seen among Palestinians, such as cerebellar ataxia, mental retardation, disequilibrium syndrome, sickle cell diseases (Samarah et al. 2018) and others. Some of these mutations manifest themselves in the same pathogenic traits or phenotype but different mutant genotypes. Haplotyping helps to determine the configuration of compound heterozygous mutations and to provide more accurate genetic consultation for patients. Compound heterozygosity is also reported in cancers. The tumorigenic effect of a cancer-susceptibility variant on one gene copy and then a somatic mutation (two hit

model) occurs which could affect the same copy or the other copy of the gene (Knudson, 1996), that implicates a precise detection for accurate diagnosis and treatment.

Finally, complete haplotype information of individuals has improved personalized medicine. Several studies associated specific haplotypes to disease susceptibility and to drug response, in particular diseases that are caused by a group of functionally-related genes that are spreaded on sparse genomic regions (Fan et al., 2011). An example is human leukocyte antigen (HLA) haplotypes and their possible associations to clinical outcomes in transplantations and autoimmune diseases. HLA loci are polymorphic and extend over 4 million base pairs, called the major-histocompatibility complex (MHC), on chromosome 6. Fan and his colleagues defined first the HLA allele at each locus on phased haplotypes of a European individual and then compared SNP haplotype at each HLA gene to those Caucasian (CUE, European) individuals from the international 1000Genomes reference panel whose HLA genes were phased already. Combination of all alleles at each loci that are determined and phased to the reference HLA genes yielded two haplotypes of the individual. They found that one of the haplotypes is a frequently observed haplotype among Caucasians and associated with immunopathological diseases (Fan et al., 2011). This finding implicates the need to improve the matching algorithms to find potential donors through HLA haplotypes in an effort to reduce the risk of clinical complications in organ transplantations (Petersdorf et al., 2007; Gragert et al., 2013).

CHAPTER 3

Methods for Resolving Haplotype

Phasing haplotypes of a set of genotyped sequences from an individual is also called *haplotype assembly* or *single individual haplotyping* (Rizzi et al., 2002). The aim is to build two haplotypes from variant information on aligned sequence reads. Many computational methods have been developed to solve the haplotype assembly problem. Below we will explain these in detail.

3.1 Computational Haplotype Phasing in Unrelated Individuals

Computational methods were utilized to find an alternative to the expensive experimental haplotyping and to overcome the limitations in mendelian phasing (Figure 2.2) (Windig and Meuwissen, 2004). The idea is to identify haploid sequences from a set of genotypes based on linkage disequilibrium (see section 1.2) of unrelated individuals or large cohorts.

One statistical approach for haplotype phasing of unrelated individuals is based on modeling of haplotype frequencies. It generates several candidate haplotype configurations for an individual's genotypes and estimates the probability of each candidate configuration through statistical modeling to pick the most likely haplotype configuration (Figure 3.1). Another approach is the rule-based approach which estimates the most likely haplotype configurations upon the assumption that those configurations are observed in other individuals (Browning et al., 2011). Below we will discuss the commonly used algorithms from both approaches.

Clark's Algorithm was the first phasing algorithm based on haplotype frequency for three or more tightly linked SNPs in unrelated individuals. It uses the parsimony criteria which seeks to infer the ambiguous SNP upon unambiguous haplotype of at most one heterozygous SNP in all genotypes and repeats it until all genotypes are resolved. Clark's algorithm estimates haplotypes

Genotypes	Possible haplotype A	Possible haplotype B	Possible haplotype C	Possible haplotype D
C A	C A	C A	C A	C A
T G	T G	T G	G T	G T
T A	T A	A T	T A	A T
Population haplotype frequency	45% 0.4%	2% 3%	25% 5%	20% 1%
Population frequency of haplotype pair	2 x 45% x 0.4% =0.36%	2 x 2% x 3% =0.12%	2 x 25% x 5% = 2.5%	2 x 20% x 1% =0.4%
Posterior probability of haplotype pair	0.36%/(0.36%+ 0.12%+2.5%+0.4%) = 10.6%	0.12%/(0.36%+ 0.12%+2.5%+0.4%) =3.5%	2.5%/(0.36%+ 0.12%+2.5%+0.4%) =73.9%	0.4%/(0.36%+ 0.12%+2.5%+0.4%) =11.8%

Figure 3.1. Haplotype frequencies to phase haplotype of unrelated individuals. Heterozygous genotype at each SNP site in a chromosome region of an individual and four possible haplotype configurations (A-D) for the given genotype. Population haplotype frequency of a configuration is the percentage of being observed in other individuals in a population. Population frequencies of haplotypes is the frequency of the two haplotypes of a pair and multiply by the factor 2 which accounts for possible assignments from both parental origins to the haplotype pair. The posterior probability obtained through statistical modeling of haplotype frequencies configured from a cohort of unrelated individuals for the given genotype to pick the most likely haplotype configuration. In this example, frequency of haplotypes in possible configuration C upon their existence in genotypes of the cohort are 25% and 5%. The frequency of haplotype pair multiplies by 2 and similarly to all possible configurations. The posterior probability is the frequency of haplotype pair to frequencies of all possible haplotype pairs. The possible configuration C (posterior probability is 73.9%) is the more likely haplotype for the given genotype.

using a minimum number of unique phased blocks. For variants that are not closely linked, the algorithm assigns several reasonable haplotypes but it remains inefficient. Later, Clark's algorithm integrated a significant advance to the phasing methods by introducing the so-called

Expectation-Maximization algorithm which can phase small numbers of distant variants (Clark, 1990). haplotype frequencies of the genotypes and then adjusts the parameters through iterative steps to generate the best haplotype estimation (Hawley and Kidd 1995). This algorithm is a good procedure for a small number of genetic markers but it does not consider assumptions of mutations and recombination events (Qin et al., 2002).

Coalescent-Based Methods were a breakthrough in modeling haplotype frequencies to infer new haplotypes originated from old haplotypes. These methods adapt stochastic models of a sequence from probabilistic states and each state produces an observed event, in which only the observed events are seen (McVean et al., 2005). A well known example of coalescent methods is the Hidden Markov Model (HMM) which similarly consists of an unseen sequence of hidden transition states (such as haplotype configuration) and a seen sequence of observed events (such as genotypes).

In Figure 3.2 illustrates HMM algorithm, it takes the first observed SNP on the genotype (SNP*g*) and tries to infer the possible underlying SNP of haplotype configuration (SNP*h*). It estimates the SNP*hs* on the haplotype configuration upon two probabilities. The emission probability (P*e*) which is the probability of the observed SNP*g* from the possible underlying SNP*h*, e.g. if the SNP*g1* is A, then what is the probability that the corresponding SNP*h1* is being one of the four nucleotide bases (G, C, A, T). It infers the SNP*h1* upon the observed SNP*g*s on the same position of other genotypes. The second probability is transition probability (P*t*) of SNP*h1* of being either nucleotides in LD relative to the phased SNP on the previous site of the ongoing estimated sequence of haplotype. HMM repeats iteratively this step until it generates the highest joint probability from transmission and emission probabilities of SNP*h1* to add into the haplotype sequence. SNP*h2* is phased upon P*e2* of the observed SNP*g2* and the Pt2 of the possible SNP*h2*

relative the previous SNP, SNP*h1*. HMM continues to calculate the highest joint probability at each SNP site which will determine the corresponding haplotype configuration from given genotype reads. The more number of overlapped genotype reads, the better HMM estimate and thus accurate haplotype phasing. HMM algorithm is the basis of several population-based phasing methods (chapter 4).

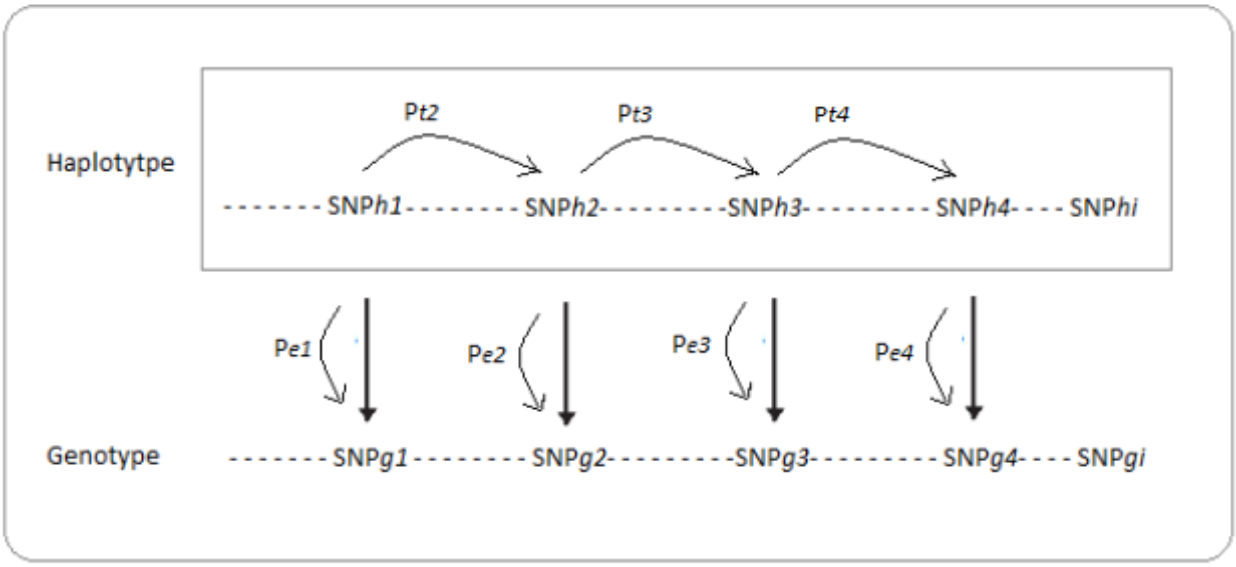

Figure 3.2 Spectral Algorithm for Hidden Markov Model (HMM) to phase haplotypes. Two chains of HMM; the observed genotypes (SNP*g*) and unknown haplotype configuration (SNP*h*). HMM infer SNP*h* upon the observation probability (P*e*) of SNP*g* and the transmission probability of SNP*h* relative to the previous phased SNP site on the haplotype sequence. The highest joint probability from both HMM probabilities at each SNP site determine the haplotype configuration.

Utilizing the Identity By Descent (IBD) is a rule-based approach that estimates haplotype configurations based on the idea that these configurations are seen in other individuals (Browning et al., 2010). IBD segment is a DNA sequence containing one or more genetic loci that are shared between two individuals or more from a common ancestor. Kong and his colleagues (Kong et al., 2008) were first to use IBD information to infer haplotypes by long-

range phasing algorithms. It was applied to the Icelandic population to identify long genomic segments containing shared alleles. IBD alleles must be ofcourse on the same haplotype. If a heterozygous genotype is located in an IBD region, the allele is phased relative to all other sites in that IBD region. However, IBD cannot be used to phase haplotype at sites where alleles are heterozygous in both individuals (Figure 3.3). This approach works well in isolated populations like the Icelandic. Additionally, a large proportion of the Icelandic population was genotyped which helped in accurately constructing IBD segments (Kong et al., 2008). IBD-based phasing method is also utilized to find haplotypes of related individuals. Combining IBD-based approaches and the previously described population-based models can enhance genome-wide haplotyping as IBD phase over long genomic regions while frequency-based approaches fine tune them at short scales. (Browning et al., 2011).

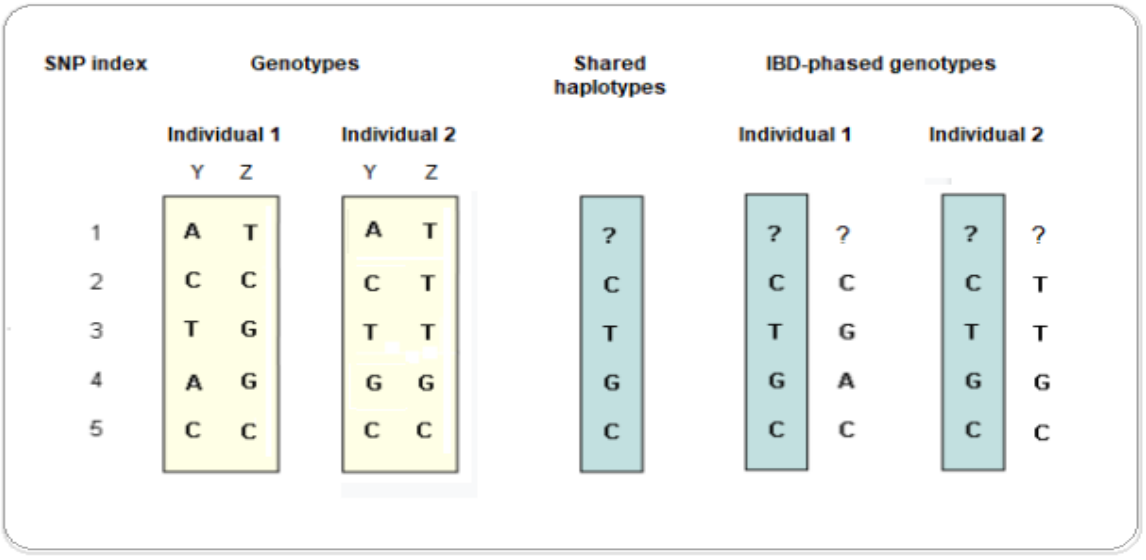

Figure 3.3. IBD-based phasing approach to determine haplotype. The use of IBD to phase genotypes that are Identical By Descent (the leftmost columns). Heterozygous alleles at SNP1 in both individuals and cannot be phased by IBD. SNP2 is heterozygous in individual 2 but alleles can be phased relative to all other sites in the IBD region. Similarly to SNP3 and SNP4 in individual 1. SNP5 is homozygous, thus phasing is trivial. In blue, the Haplotypes were phased by IBD; of note, SNP 1 remains unphased but can be phased by population-based techniques (see Chapter 4).

CHAPTER 4

Population-based Haplotype Phasing

Haplotypes cannot be observed directly from genotype data but can be inferred either directly from sequence data of an individual whole genome or whole exome NGS reads (read-based phasing) or indirectly using haplotype reference panel (population-based phasing).

In read-based phasing, sequence reads that contain several variants give partial information of haplotypes that can be assembled using computational methods into longer haplotypes (Bansal, 2019). HapCUT (Haplotype Assembly Coverage Handling by Adapting Thresholds) is a read-based phasing tool, also called as read-backed phasing, which utilizes a heuristic approach to assemble fragments of aligned sequence reads of identified variant sites to a pair of haplotypes with maximum consistency using likelihood-based function (Figure 4.1). HapCUT first converts the aligned sequence reads to custom haplotype-relevant fragments. It starts with two candidate haplotype fragments and searches for a set of variant sites to elongate the haplotype blocks. Then, it iteratively changes the phased variants on the given candidate fragments relative to the set of variants during elongation until it achieves a haplotype pair with greater likelihood (Edge et al., 2017). In spite of this approach applicability to construct haplotypes of the entire chromosome, only few variants per chromosome can be phased. Therefore, the completeness of the estimated haplotypes are limited (Selvaraj et al., 2013). However, it is highly accurate for the positions that it is capable to phase.

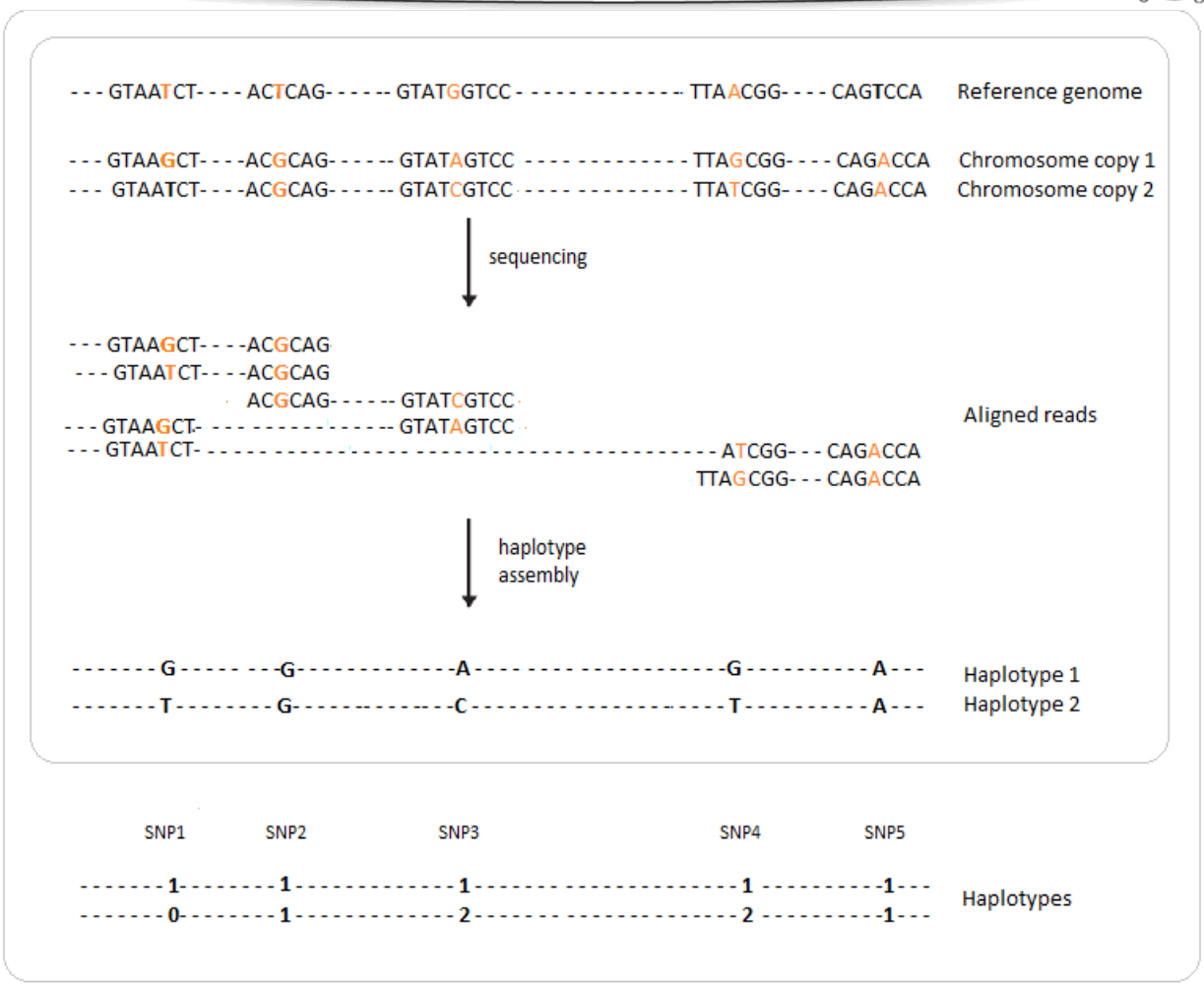

Figure 4.1 HapCUT Assembly. The custom data for HapCUT consist of aligned reads from a chromosome pair and variant genotypes. Alleles in orange color are heterozygous sites. Aligned reads that possess different alleles at a variant site are inferred to come from different chromosome copies. Reads that possess the same allele at a common site are inferred to come from the same chromosome copy and can be lined together to construct haplotypes. Two haplotypes are corresponding to two chromosome copies. Haplotypes determine the copy of the chromosome pair that has the variant allele (assigned by 1) such as at variant site 1. HapCUT phased mutation 2 on SNP 3 and 4 in cis configuration.

Population-based phasing is on the other hand based on leveraging haplotype information from population reference panels. Haplotype reference panels are built from a set of already phased haplotypes from a large cohort of unrelated individuals belonging to either the same population or mixed populations (Bansal, 2019). Example of population-specific reference panels are The

Genome of the Netherlands Consortium (Francioli et al., 2014), Ashkenazi reference panel (Carmi et al., 2014), UK-specific reference panel (Huang et al., 2015), Japanese population reference panel (Nagasaki et al., 2015), Estonian-specific panel (Mitt et al., 2017), Anabaptist Genome Reference Panel (AGRP) (Hou et al., 2017), Northeast Asian Reference Database (NARD; Yoo et al., 2019) and others (Gudbjartsson et al., 2015, Sidore et al., 2015, Bai et al., 2018). Examples of haplotype reference panels from a mix of several populations are the International 1000Genomes Project (The International HapMap Consortium 2003, The International HapMap Consortium et al., 2007), the International HapMap3 (HapMap; The International HapMap 3 Consortium, 2010) and the Haplotype Reference Consortium (HRC) (The Haplotype Reference Consortium et al., 2016).

Several algorithms utilize the population-based phasing from reference panels. Available tools that run population-based phasing are PHASE, fastPHASE, BEAGLE, IMPUTE2, MACH, and SHAPEIT. These tools use the same approach mentioned above but differ in some factors (see next).

PHASE is suitable for a small number of genetic markers and a small number of individuals. It uses all haplotypes other than those of the current individual being estimated as hidden states. This tool was a gold standard for accuracy among the other tools but very slow compared to the new ones (Marchini et al., 2006).

BEAGLE achieves haplotype phasing for large sample size, more than 1000 individuals, with high accuracy and significant phasing speed. This tool considers the iterative HMM at each locus to locally cluster haplotype segments and adjusts them to the given sample size and LD information, but does not consider HMM modelling of mutations and recombination events (Browning et al., 2011).

SHAPEIT (Segmented HAPlotype Estimation and Imputation Tool; shapeit.v2.r904 .glibcv2.12.linux.tar.gz) estimates short range haplotypes of genotype data from familial and population information. It builds HMM state at each variant site and hence it is able to modulate mutations at each position. Among all the available tools, only SHAPEIT can support multithreading which allows the central processing unit of the computer (CPU) to execute multiple processes in parallel and reduces the operating time (Williams et al., 2012). Thus, it is more efficient than all other aforementioned tools in which it can achieve high accurate haplotypes of large sets of genotype data in relatively short time (Delaneau et al., 2011).

Moreover, SHAPEIT has several model parameters to increase phasing accuracy. Such as conditioning state by which one can specify the number of conditioning reference haplotypes for the tool to search for the most similar subset of reference haplotypes to the given reads and help to achieve accurate corresponding haplotypes. Window size is a parameter to define the mean size of the conditioning haplotypes in megabases. SHAPEIT also runs different iteration sets to model HMM iterative algorithm. Running iterations re-estimate iteratively the probability until the best estimate is achieved, in which it uses transition probabilities to sample haplotypes and trim unlikely configurations. The burn-in iteration is the number of iterations for HMM algorithm to pick a good first conditioning haplotype to start phasing. SHAPEIT can also perform pruning iterations which use the transition probabilities to trim unlikely haplotype configurations. In addition, main iteration is to store transition probabilities and averaged them at the end of running iterations to obtain the final haplotype (Delaneau et al., 2011).

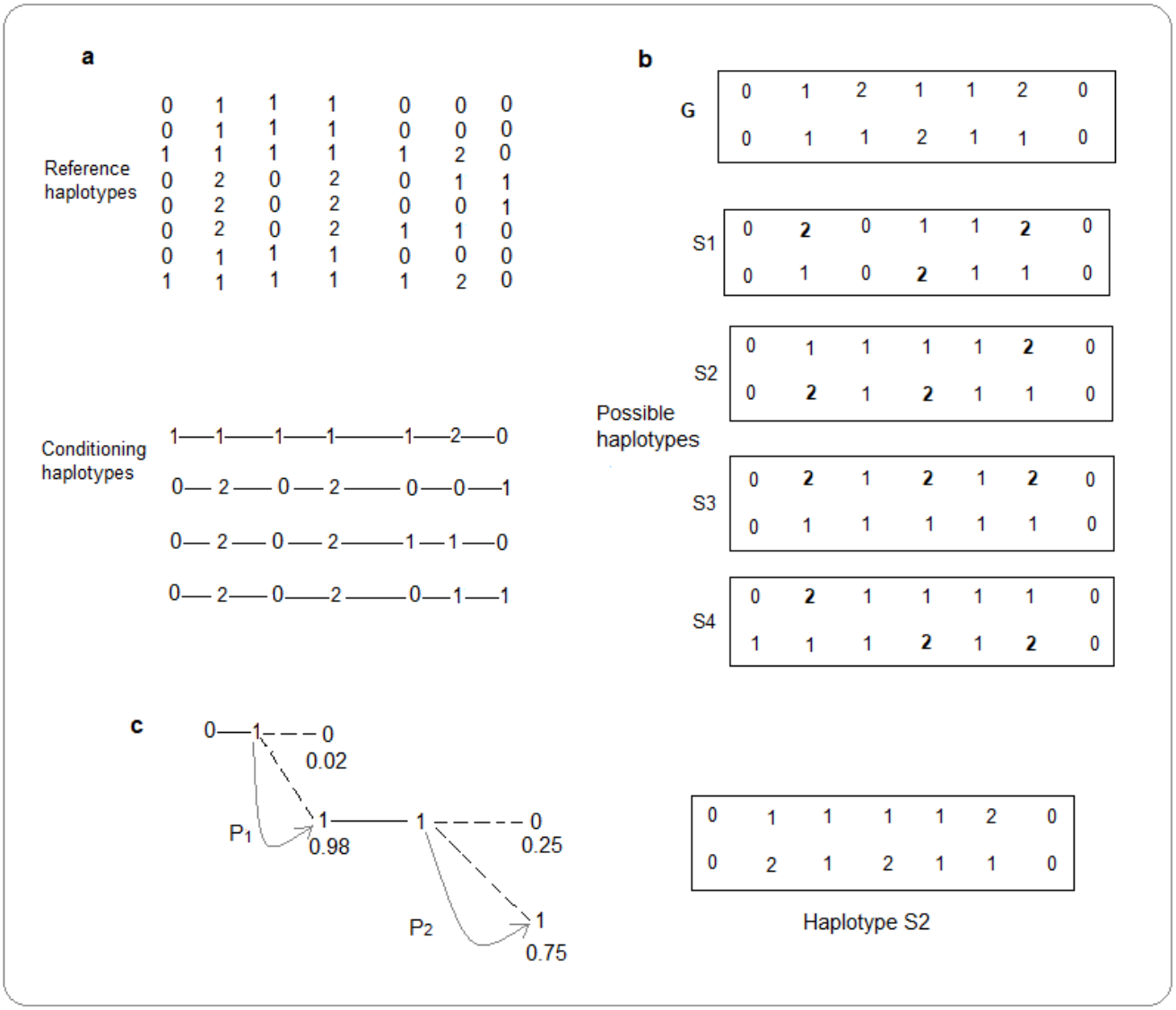

Figure 4.2. Schematic illustration of genotype phasing by SHAPEIT. Segments of an individual's genotype (G) include three heterozygous sites and the possible phase of the sites result in four possible haplotypes (S_i). (a) SHAPEIT specifies a number of conditioning haplotypes from reference haplotypes that are similar to G segments by conditioning state parameter . (b) Burn iterations of SHAPEIT re-estimate a good starting possible haplotypes for HMM to begin estimation by calculating the transition probabilities between sites of a segment (c) to decide the final haplotype configuration by the average of the main iterations. Here, the possible haplotype 2 is the corresponding haplotype for the given genotype reads.

CHAPTER 5

Phasing X Chromosome

The human X chromosome has unique features. Females inherit an X chromosome from each parent, while males inherit a single maternal X chromosome and a paternal Y chromosome. Gene expression is silenced on one of the female X chromosomes in the embryonic development (embryogenesis) and it continues inactive in the somatic tissues, but it reactivates and recombines with the other X chromosome during meiosis. In males, recombination on the X chromosome is restricted to a small region which carries equivalent genes on the Y chromosome. The genes that are shared between X and Y chromosomes are called pseudoautosomal genes and genes outside this region called non-pseudoautosomal (nPAR) genes (Darrow et al., 2016). Phasing X chromosomes are possible only after predicting the gender. Gender prediction relies on the heterozygosity of SNPs on genes at nPAR regions of chromosome X and chromosome Y (Bilton et al., 2019). Heterozygosity of chromosome X is informative as males possess only one copy of the X chromosome and hence cannot be heterozygous. Heterozygosity of SNPs on chromosome Y is informative also as females possess no copy of the Y chromosome and hence average of Y intensity should be lower in females (McClure et al., 2018). SNPs of nPAR region on the Y chromosome are easier to utilize for sex prediction. However, not all commercial SNP-detecting chips and NGS exome capture probes contain the Y SNPs, thus gender prediction often uses nPAR SNPs of X chromosome where females display higher heterozygosity rate (McClure et al., 2018). SHAPIT phsaes X chromosomes by inferring haplotype using nPAR SNPs after sex prediction of recruited samples (see Section 7.2.2).

CHAPTER 6

Motivation

The broad objective of this project is to build the first Palestinian haplotype exome panel which we named “PalHap”. Haplotype phasing will overcome the shortcomings of the current applied NGS technology and help in various clinical applications (see section 2.1). For instance, the identification of compound heterozygous mutations, which are observed frequently in the analysis, their correct configurations from whole exome or genome sequencing is challenging.

Population-based phasing is a promising approach that can provide accurate phasing of haplotypes and help in the detection of causative variants. The accuracy of this phasing method relies mainly on the ethnicity-matching of the used haplotype panel. Therefore, we wanted first to build a haplotype reference panel specific for Palestinians using variants from whole exomes. The performance of our panel in phasing compound heterozygous mutations among Palestinians is evaluated by comparing with read-backed phasing and 1000Genomes reference panel.

CHAPTER 7

Materials and Methods

7.1 Study Population

Whole exomes from 900 Palestinians were generated using Illumina Nextseq 500 at Molecular Genetic Lab-Istishari Arab Hospital. Participants include affected, healthy individuals, and five trios. Participants were asked to sign an informed consent form according to guidelines provided by Molecular Genetic Lab Institutional Review Board of Istishari Hospital.

7.2 variants Call

We identified the variants (SNPs and indels) from exome's raw data using BWA aligner (Li, et al., 2009) and GATK (Genome Analysis Toolkit) HaplotypeCaller according (McKenna et al., 2010) to GATK best practice guidelines (Van der Auwera et al., 2013). The variant calls including the allele change, chromosome, and position were stored in the standard format VCF file.

7.2.1 Variants filtration

The mitochondrial DNA and Y-chromosome were not included in the study because they are not subjective to haplotype phasing. The VCF file which stores the variant call includes a number of columns that describe the quality of the variant and the sample (Figure 8.1). We filtered the samples as follows. The combined read depth (DP) is the number of reads that cover a variant position at each sample. We filtered out any variant in which the average DP is less than four

reads or greater than 2000 reads. This is because below four there is not enough coverage to produce high genotype call and above 2000 the region is usually repetitive and that may indicate poor mapping.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	1/1:40:3

7.1. An example of VCF file format. VCF file is a text file format and each data line displays information about each site on the genome. It contains (from left to right) CHROM (chromosome), POS (position of variants), REF (reference allele), ALT (alternative allele), QUAL (quality), INFO (variant site-level quality), FORMAT (genotype-level qualities). Among sample quality, the parameter GQ (genotype quality) and DP (Read Depth) that we applied in the filtering criteria.

Genotype quality (GQ) is a quality metric which describes the confidence in the assigned genotype of a particular sample. We filtered out any variant which has GQ below 60 in which the probabilities of a genotype assigned incorrectly is 1 in million, thus we ensure that the assigned genotype of a base is correct. The filtering criteria of variants as follows. Base quality (QUAL) is the confidence of a variation detected at a site. We included variants that have QUAL greater or equal to 50Q which means the probability of the base to be incorrect is at least 1 in 100,000. In addition, we excluded positions which carry variants that are deviated from the mean of depth coverage with 2.5 standard deviation, since these positions may have aberrant variants due to low genotype quality or to unusual features such as insertion or deletion. We also used PLINK to filter out variants with a fraction of missing calls in the set of genotypes by

missing genotype rate, by which we removed variants with missing call rate exceeding the value of 5% missingness in all samples.

Variants that meet the thresholds of filtering criteria are then compressed and indexed using *tabix*, which helps to retrieve variant positions on overlapping regions by chromosome coordinates quickly (Li, 2011). Finally, we combined all VCF files (each correspond to one individual) into a single file using bcftools (Narasimhan et al., 2017).

Table 7.1. Quality Filters of variants

Quality filters	Corresponding range
Combined Read Depth (DP)	DP >= 4 & DP <= 2000
Conditional Genotype Quality (GQ)	GQ >= 60
Quality (QUAL)	QUAL >= 50
Filtering samples contain heterozygous calls with high standard deviation from the mean	het. calls ± 2.5SD
Missing genotype rate	geno 0.05 (5% missing)

7.2.2. Sex Prediction

Gender prediction of a given set of samples is an essential step to construct haplotypes for the sex chromosomes, regardless of the provided gender information in consent forms. PLINK tool implements the function of heterozygosity-based gender prediction and uses the coefficient of inbreeding *F* on X chromosomes. *F* coefficient means that an individual has two alleles at each locus that are identical by descent from the same ancestor of the two parents. We can infer that allele transmition on sex chromosomes to offsprings (inbreeding) is influenced differently in the two genders and the values of *F* differ among both genders (Ebel and Phillips, 2016).

Heterozygosity-based prediction (see Chapter 6) calculates the non-missing SNPs rate on X chromosome and considers the haploid heterozygous fraction as a gender error. PLINK makes a male call if the gender error with haploid heterozygous proportion is more than 80%, means that 80% of heterozygous sites on the X chromosome genotypes are found in one copy. Female call is assigned if the gender error or in other words the haploid heterozygous proportion is less than 20% (Qu et al., 2011).

7.3 Construction of PalHap Panel

We constructed the PalHap panel using a number of samples that passed the filtering criteria above. The merged VCF file was splitted by chromosomes using PLINK, in which each file contains variants of a single chromosome from all samples. The haplotypes were computationally phased using the population-based phasing tool SHAPEIT (see chapter 4). SHAPEIT takes as input the unphased per-chromosome VCF files or takes the PLINK format which consist of three files: the bed file contains variant genotypes, bim file contains variants' position, and fam file which describe the family relation (if any) among individuals. In order to ensure good accuracy of haplotypes, we altered the model parameters as the following: the number of conditioning states was set to 200, burn-in iterations to 70, pruning to 80 iterations and main iterations to 200 (for more details see Chapter 4)

7.4 Read-based Phasing (HapCUT2)

HapCUT2 takes as input unphased genotypes of the test samples in VCF format and the mapping file (BAM or Binary Alignment Map) which stores aligned reads. The tool first customizes these files into its custom input file "fragment file" by extractHAIRS (Extract Haplotype Informative Reads) tool which extracts aligned reads fragments from stored BAM file and variants from VCF

file to create informative haplotype fragments (Edge et al., 2017). Then, HapCUT2 takes the custom input file and iteratively changes the phased variants on the candidate fragments to assemble haplotype blocks with greater likelihood (see Figure 4.1). Because information from aligned reads may not be enough to construct the true phase configuration of some pairs of mutations (high heterozygosity at some sites), not all compound heterozygous pairs were phasable.

7.5 Phasing performance

The phasing accuracy score of the test samples is defined per position as the percentage of individuals that were correctly phased at a particular compound heterozygous pair. We will explain it in the following example. Let's say at chromosome 1 positions 1001 and 1007 we found two heterozygous mutations in a subset of the test samples (say 10 individuals). We found using HapCUT2 from the read-based phasing evidences that this compound heterozygous pair is "trans", in other words each mutation is located at a different haplotype. If SHAPEIT using our PalHap panel successfully phases this trans pair in 8 individuals, the phasing accuracy at this pair should be 80%. We similarly calculated the phasing accuracy score of the phased pairs per position using 1000Genomes panel.

7.6 Identification of Copy Number Variations (CNVs)

Calling copy number variations (CNVs) involves several steps. First, the coverage of mapped reads (i.e. copy number or CN) to each chromosome window from each sample were obtained. Second, normalization of the coverage was performed to take the between-samples variations

and between-regions variations into account. Third, the estimated CN values were transformed to segment mean value or score using Hidden Markov Model (HMM). HMM computes the scores based on the joint probabilities of emission probabilities which correspond to a form of coverage (CN intensity level) and transition probabilities which correspond to changes in CN, wherein the model estimates the CN as the most likely state of ploidy of the sequence. Lastly, scores are taken to the mean of CNVs for the positions on the segment to find the segment mean value that is equal to log2 (Oshlen et al., 2011; Halpern et al., 2014). Finally, a quality control is performed to remove variants called within telomeres and centromeres.

CHAPTER 8

Results

8.1 Data Description

The removal of whole exome samples that did not match the filtering criteria (Section 7.2.1) resulted in 696 whole exomes. We further handled the genotypes of X chromosome to predict the gender using Plink. The retained set of 696 whole exomes containing high quality genotypes of biallelic SNPs which include both common and rare variants, found in at least 10 individuals from autosomal and X chromosomes. Among the 696 samples we used 600 in constructing the haplotype panel (i.e. training samples), the rest 96 were used to test the performance of the phasing (i.e. test samples).

8.2 Construction of PalHap Panel

We built the PalHap panel using SHAPEIT (v2.12) which computationally estimates haplotype pairs at each chromosome. Shapiet's output is composed of two files: the Hap file which contains a set of phased haplotypes and a sample file which stores individual information. As recommended in Shapiet's documentation, we converted the haplotypes from the current format (Hap/Sample) to the standard format (Hap/Legend/Sample). The legend file in the standard format describes the positions of SNPs of the phased haplotypes. Up to this point, PalHap was constructed and can be used to phase any mutation which is already included in the panel.

8.3 Calling Copy Number Variations (CNVs)

Because the phasing accuracy might be affected by polymorphic structural events such as common losses and gains, we wanted to map the CNVs alongside the phased compound heterozygous pairs to compare.

We defined CNVs in all the 696 samples using XHMM tool (Fromer et al., 2014 and Methods) based on the segment mean log2 value (see Section 7.6). The diploid region will have a segment mean of zero, deletion will have negative value “loss” and duplicated regions will have positive value “ gain”.

8.4 Phasing Accuracy Performance of PalHap Panel

To evaluate the performance of PalHap panel relative to the commonly used 1000Genomes panel, we phased 96 exomes that were not included in the construction of the panel (which we call test samples) using three different methods: HapCUT2, PalHap panel, and 1000Genomes panel. HapCUT2 is the golden standard method which represents the true haplotypes as it constructs them using the actual NGS mapped reads. SHAPEIT was used to phase the test samples by PalHap and 1000Genomes. We plotted the phasing accuracy of each compound heterozygous pair per position (Figure 9.1). Next we mapped the polymorphic CNVs (frequency greater than 1%)- among Palestinains that were previously identified (Section 9.3 and Section 1.2)- on the top of the phasing accuracy per position plot for comparison. Results show that both PalHap and 1000Genomes can accurately phase compound heterozygous mutations. The majorities of the pairs were phased correctly (100%) in all the test samples that have them. However, with few exceptions PalHap panel achieve higher phasing accuracy of positions that are located within polymorphic CNVs compared to 1000Genomes.

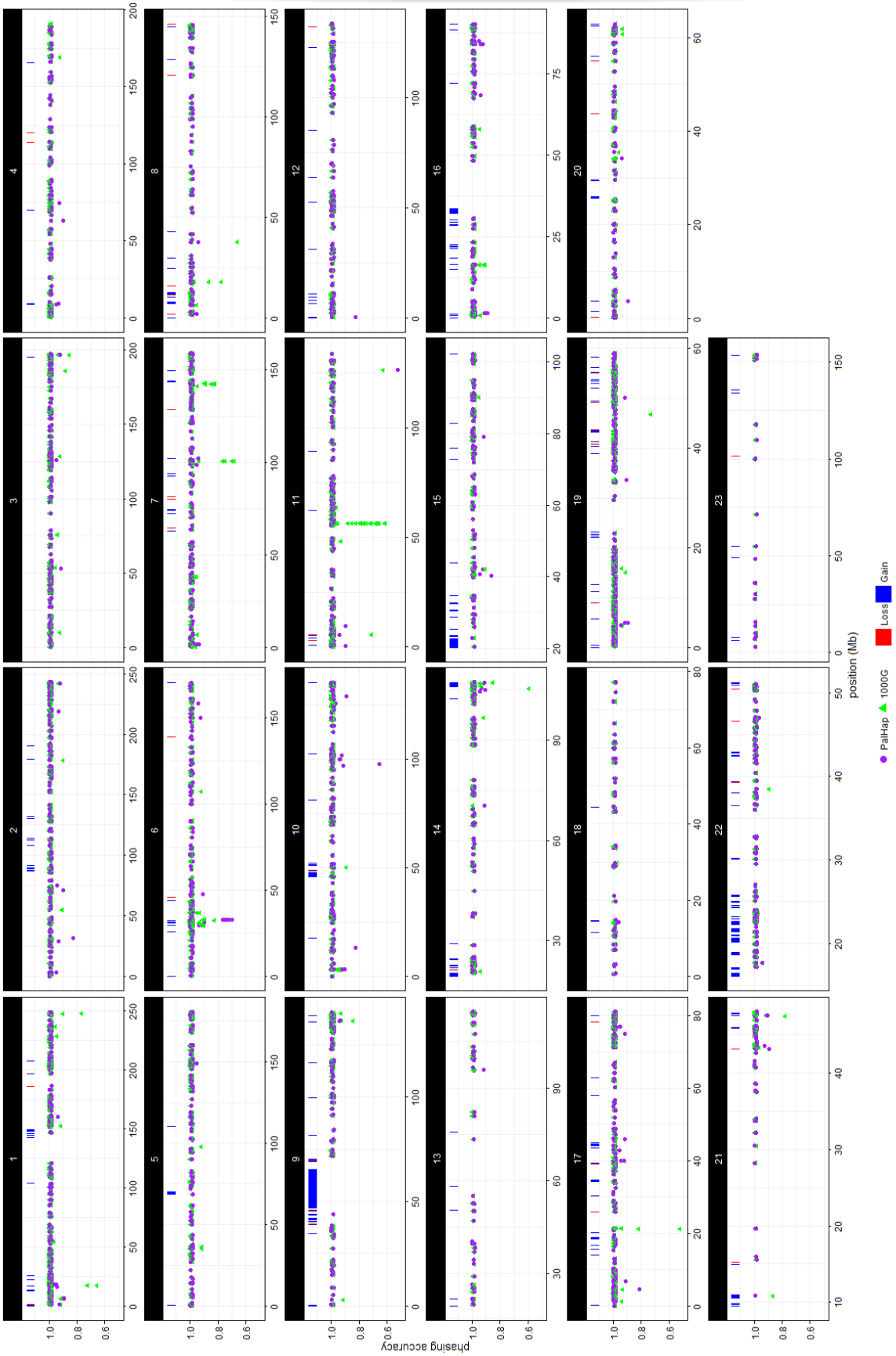
phasing accuracy
position (Mb)
PalHap
1000G
Loss
Gain

Figure 8.1: phasing accuracy of each compound heterozygous mutations in the 96 “test samples” using 1000Genomes (green triangles) and palhap (purple circles). The polymorphic copy number variations are mapped on each chromosome plot in red (low copy number variation) or blue (high copy number variation) segments. In general, the positions in which 1000Genomes show lower phasing accuracy overlap polymorphic CNVs regions.

CHAPTER 9

Discussion

The classical mendelian haplotype phasing method is widely used in clinical settings but it has a number of drawbacks. For example the recruitment of parents or family members, complex disease-associated mutations, and laborious and time-consuming process. Population-based phasing approaches overcome these shortcomings. However, it is crucial to apply the most reliable and accurate methods in the clinical settings. Recently, population-specific reference panels were established to enable a number of applications such as haplotypic association studies of common variants and imputation of missing common variants , as well as compare haplotype diversity in the evolutionary biology studies and population genetics. The goal of this project is to build the first Palestinian haplotype exome panel and examine its performance by determining whether population-specific reference panel can achieve accurate phasing of compound heterozygous mutations s.

To demonstrate the feasibility of PalHap panel, we compared its performance in phasing compound heterozygous mutations to 1000Genomes. CNVs contribute to rare variants that are involved in genetic disorders and mendelian diseases in families (Inoue et al., 2002; Lee et al., 2006). Number of studies reported association of polymorphic CNVs with complex diseases such as lupus glomerulonephritis, several autoimmune diseases, cases of schizophrenia and autism and others (Aitman et al., 2006; Fanciulli et al., 2007; Cook et al., 2008). In addition, a study identified loss-of-function variants through 1000Genomes data in CNVs regions (MacArthur et al., 2012). Therefore, years of SNP association studies have considered CNV associations as an equally susceptible factor for false associations between genetic variations and

phenotypes (McCarroll et al., 2007). Artificial CNV associations are mainly attributed to the limited ability of SNP studies to detect variations within CNVs (Lahita et al., 2010) and to insufficient knowledge of locations and frequencies of CNVs that segregate human populations (McCarroll et al. 2007). This makes it difficult to phase and assemble CNVs containing regions (Black et al., 2014). Therefore, we examined the phasing performance of PalHap panel to phase compound heterozygous mutations by taking CNVs into account. As expected, PalHap panel showed the capability to accurately phase compound heterozygous mutations compared to 1000Genomes especially in regions that overlap polymorphic CNVs among Palestinians. This is because PalHap panel has accurate representation of haplotypes and linkage disequilibrium blocks (LD) among Palestinians which level up the power of detecting causal variations (McCarroll et al., 2007; Conrad et al., 2010) than mixed ethnicity panels such as 1000Genomes.

This finding is compatible with other studies demonstrating that ethnicity matching reference panels improve phasing accuracy especially for rare and complex variants (Yasuda, J., et al. 2018; Hou et al. 2017). For instance, Hou et al showed that the Anabaptist Genome Reference Panel (AGRP; phased haplotypes and WGS from Amish and Mennonite) conferred better representation of rare Mendelian disorders-causing founder alleles and haplotypes, hence, improved imputation accuracy compared to Haplotype Reference Consortium reference panel (Hou et al., 2017). Moreover, PalHap panel showed significant phasing accuracy despite having less than half of the samples in the 1000Genomes panel (2,504 samples). The size of the reference panel was shown to correlate with its performance (Kong et al., 2008; Williams et al., 2012; Loh et al., 2016). For example, Choi and his colleagues show that population-based phasing accuracy improves significantly as the size of the reference panel increases (Choi et al., 2018). Also, Browning and his colleague concluded that a simple and effective strategy to

improve phasing accuracy is by raising the sample size via using a population-specific reference panel (Browning et al., 2011). Our results suggest that ethnicity-matching can overcome the small sample size as we show that 600 ethnicity-specific panel can be either as equal as or outperform 1300 mixed-ethnicity panel at specific regions.

Our study, as any other study, has some limitations. First, PalHap represents mainly individuals from the West Bank. Jerusalem residents and Gazinians are not well represented. Most Gazinians samples were filtered out because of low quality samples due to constant power cut at the clinical facilities and long inappropriate transportation conditions. Second, computational phasing cannot phase low frequency variants or those that are not well represented in the panel. Improving the aforementioned challenges will help to achieve accurate phasing of all variants.

CHAPTER 10

Conclusion

In this thesis, we first constructed PalHap, the first Palestinian haplotype exome panel and showed that population specific-reference panels provide accurate phasing for clinical haplotype-based applications such as phasing compound heterozygous mutations. The panel will pave the way for future work in adopting computational phasing in the clinical practices such as preimplantation diagnosis (PGD) which will require only genotypes from the embryo without the need to recruit the parents and other family members. It may also help in future genetic association studies and evolutionary population genetics.

CHAPTER 11

References

Adey, A.C. (2017) 'Haplotype Resolution at the Single-cell Level', *PNAS*, 114(47), pp. 12362 – 12364.

Aitman, T.J., et al (2006) 'Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans', *Nature*, *439*(7078), pp.851-855.

Bai H., et al. (2018) 'Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia', *Nat Genet*., 50, pp.1696–704.

Bansal, V., (2019) 'Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes', *Bioinformatics*, 35(14), pp. 242–248.

Beroukhim, R., et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, *463*(7283), pp.899-905.

Bersaglieri, T.,et al. (2004) 'Genetic signatures of strong recent positive selection at the lactase gene', *The American Journal of Human Genetics*, *74*(6), pp.1111-1120.

Bilton, T. P., et al. (2019) 'Using genotyping-by-sequencing to predict gender in animals', *Animal genetics*, *50*(3), pp. 307–310.

Black, H.A., et al. (2014) 'Inferring mechanisms of copy number change from haplotype structures at the human DEFA1A3 locus', *BMC genomics*, *15*(1), p.614.

Brahmachary, M., et al., (2014) 'Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats', *PLoS genetics*, *10*(6).

Browning, S. R., et al. (2010) 'High-resolution detection of identity by descent in unrelated individuals', *Am J Hum Genet*., 86(4), pp. 526-39.

Browning, B.L., et al. (2011) 'Haplotype phasing: Existing methods and new developments', *Nat Rev Genet*., 12(10), pp. 703–714.

Carmi, S., et al. (2014) 'Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins', *Nat Commun* 5, 4835.

Choi Y., et al. (2018) 'Comparison of phasing strategies for whole human genomes', *PLoS Genet*., 14(4).

Clark, A. G. (1990) 'Inference of haplotypes from PCR-amplified samples of diploid populations' *Mol BiolEvol.*, 7, pp.111–22.

Cook Jr, et al. (2008) 'Copy-number variations associated with neuropsychiatric conditions', *Nature, 455*(7215), pp.919-923.

Conrad, D.F., et al. (2010) 'Origins and functional impact of copy number variation in the human genome', *Nature, 464*(7289), pp.704-712.

Darrow, E. M., at el. (2016) 'Deletion of DXZ4on the human inactive X chromosome alters higher-order genome architecture', *Proceedings of the National Academy of Sciences,* 113(31), pp. 4504–4512.

Delaneau, O., et al. (2011) 'Shape-IT: new rapid and accurate algorithm for haplotype inference', *BMC bioinformatics,* 9(1), pp. 540.

Druet, T., et al. (2010) 'A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping', *Genetics,* 184, pp. 789–798.

Ebel, E. R., at al. (2016) 'Intrinsic differences between males and females determine sex-specific consequences of inbreeding', *BMC evolutionary biology, 16,* pp. 36.

Edge, P., et al. (2017) 'HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies', *Genome Res.* 27, 801–812.

Fan, H.C., et al. (2011) 'Whole-genome molecular haplotyping of single cells', *Nature Biotech,* 29(1), pp. 26–47.

Fanciulli, M., et al. (2007) 'FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity', *Nature genetics, 39*(6), pp.721-723.

Fernandez, A.F., et al. (2012), 'A DNA methylation fingerprint of 1628 human samples', *Genome research, 22*(2), pp.407-419.

Fox, E. J., et al. (2014) 'Accuracy of Next Generation Sequencing Platforms', *Next generation, sequencing & applications,* 1.

Francioli, L., et al. (2014) 'Whole-genome sequence variation, population structure and demographic history of the Dutch population', *Nat Genet* 46, 818–825.

Fromer, M. and Purcell, S.M., (2014) 'Using XHMM software to detect copy number variation in whole- exome sequencing data', *Current protocols in human genetics*, *81*(1), pp.7-23.

Gragert, L., et al. (2013) 'Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry', *ELSEVIER,* 74 (10), pp. 1313-1320.

Genome of the Netherlands Consortium (2014) 'Whole-genome sequence variation, population structure and demographic his-tory of the Dutch population.', *Nat Genet,* 46, pp. 818–825.

Glusman, G. et al. (2014) 'Whole-genome haplotyping approaches and genomic medicine', *Genome Med,* 6, pp. 73.

Gudbjartsson, D.F., et al. (2015) 'Large-scale whole-genome sequencing of the Icelandic population', *Nat Genet,* 47, pp. 435–444.

Inoue, K., et al. (2002) 'Molecular mechanisms for genomic disorders. *Annual review of genomics and human genetics'*, *3*(1), pp.199-242.

International Human Genome Sequencing Consortium (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409, PP. 860–921.

Hagemann, I. S. (2015) 'Clinical Genomics, Overview of Technical Aspects and Chemistries of Next-Generation Sequencing', *ScienceDirect,* pp 3-19.

Halpern, A., et al. (2014) '*Methods for estimating genome-wide copy number variations*', *Complete Genomics Inc*, U.S. Patent 8,725,422.

Hawley, M.E., et al. (1995) 'HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes', *J Hered*, 86, pp. 409–411.

Henrichsen, C.N., et al. (2009) 'Segmental copy number variation shapes tissue transcriptomes', *Nature genetics, 41*(4), p.424.

Hou L, et al. (2017) 'A population-specific reference panel empowers genetic studies of Anabaptist populations', *Scientific Reports*. 7(1).

Huang, J., et al. (2015) 'Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel', *Nat Commun* 6, 8111.

Karafet, T.M., et al. (2008) 'New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree', *Genome research*, *18*(5), pp.830-838.

Kitzman, J.O., et al. (2011) 'Haplotype-resolved genome sequencing of a Gujarati Indian individual', *Nat Biotechnol*, 29, PP. 59–63.

Khaja, R., et al. (2006) 'Genome assembly comparison identifies structural variants in the human genome', *Nature genetics*, *38*(12), pp.1413-1418.

Kong, A., et al. (2002) 'A high-resolution recombination map of the human genome', *Nature Genetics*, 31(3), pp. 241–247.

Kong A., et al. (2008) 'Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet*.,40(9), pp. 1068- 1075.

Knudson AG., (1996) 'Hereditary cancer: two hits revisited', *J Cancer Res Clin Oncol.*, 122(3), pp.135–140.

Kumaran, M., et al. (2019) 'Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data', *BMC bioinformatics*, *20*(1), pp.1-11.

Lahita R. G., et al. (2010) 'Systemic Lupus Erythematosus', fifth edition, pp. 17.

Lee, J.A. et al. (2006) 'Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders', *Neuron*, *52*(1), pp.103-121.

Levy, S., et al. (2007) 'The diploid genome sequence of an individual human', *PLoS Biology*, 5, pp. 254.

Li, N., et al. (2003) 'Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data', *Genetics*, 165(4), pp. 2213–2233.

Li, H. and Durbin, R., (2009) 'Fast and accurate short read alignment with Burrows–Wheeler transform', *bioinformatics, 25*(14), pp.1754-1760.

Loh, P.R., et al. (2016) 'Reference-based phasing using the Haplotype Reference Consortium panel', *Nature genetics, 48*(11), pp.1443.

MacArthur, D.G., et al. (2012) 'A systematic survey of loss-of-function variants in human protein-coding genes', *Science, 335*(6070), pp.823-828.

McKenna, A., (2010) 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome research, 20*(9), pp.1297-1303.

Mardis, E. R. (2013) 'Next-generation sequencing platforms. Annu Rev Anal Chem', *Palo Alto,* 6, pp. 287–303.

Marchini, J., et al (2006) 'A comparison of phasing algorithms for trios and unrelated individuals', *The American Journal of Human Genetics*, *78*(3), pp.437-450.

McCarroll, S.A., et al (2007) 'Copy-number variation and association studies of human disease', *Nature genetics, 39*(7), pp.S37-S42.

McClure, M.C., et al. (2018) 'SNP Data Quality Control in a National Beef and Dairy Cattle System and Highly Accurate SNP Based Parentage Verification and Identification', *Front Genet.*, 15 (9), pp. 84.

McVean, G.A. and Cardin, N.J., (2005), 'Approximating the coalescent with recombination', *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), pp.1387-1393.

Nagasaki, M., et al. (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals', *Nat Commun* 6, 8018.

Narasimhan, V. M., et al. (2017) 'Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes', *Nature Communications,* 8(1), pp. 303.

Mitt, M., et al. (2017) Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* 25, 869–876.

Morozova, O., et al. (2008) 'Applications of next-generation sequencing technologies in functional genomics', *Genomics,* 92: 255–264.

Olsen, H.L., (2011) 'A molecular genetic study investigating the role of maternal and placental laeverin gene mutation and fetal whole genome copy number variations in the pathophysiology of preeclampsia' (Master's thesis, Universitetet i Tromsø).

Petersdorf, E.W., et al. (2007) 'MHC haplotype matching for unrelated hematopoietic cell transplantation', *PLoS Med.*, 4, 8.

Purcell, S., et al. (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *American Journal of Human Genetics*, 81, pp. 559– 575.

Qi J, Chen Y, et al. (2014) 'Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping', *PNAS*, 111, pp.10007–12.

Qin, Z.S., et al. (2002) 'Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms', *Am J Hum Genet*, 71, pp. 1242–1247.

Qu, C., et al. (2011) 'Cost-effective prediction of gender-labeling errors and estimation of gender-labeling error rates in candidate-gene association studies', *Front Genet*, 2, pp. 31.

Redon, R., et al. (2006) Global variation in copy number in the human genome. *nature*, 444(7118), pp.444-454.

Roach, J.C., et al. (2010) 'Analysis of genetic inheritance in a family quartet by whole-genome sequencing', *Science*, 328, pp. 636–9.

Rizzi, R., et al. (2002) 'Practical Algorithms and Fixed-Parameter Tractability for the Single Individual SNP Haplotyping Problem', *Proceedings 133 of the second international workshop on algorithms in Bioinformatics WABI*. 02. London, UK, UK: Springer-Verlag, 29–43.

Samarah, F., et al. (2018) 'Frequency of Red Blood Cell Alloimmunization in Patients with Sickle Cell Disease in Palestine', *Advances in Hematology*, pp. 1–7.

Scrimshaw N, et al. (1988) 'The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance', *Am J Clin Nutr*, 48:1079–1159.

Semino, O., et al. (2000) 'The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective', *Science*, 290 (5494), pp. 1155–59.

Sidore, C., et al. (2015) 'Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers', *Nat Genet,* 47, pp. 1272–1281.

Suk, E.-K., et al. (2011) 'A comprehensively molecular haplotype-resolved genome of a European individual', *Genome Res.*, 21, pp. 1672-1685.

Snyder, M.W., et a.l (2015) 'Haplotype-resolved genome sequencing: experimental methods and applications', *Nature Reviews Genetics*, *16*(6), pp.344-358.

The 1000 Genomes Project Consortium, (2015) 'A global reference for human genetic variation', *Nature*, 526 (7571), pp. 68-74.

The Haplotype Reference Consortium, (2016) 'A reference panel of 64,976 haplotypes for genotype imputation', *Nature Genetics*, 48(10), pp. 1279-1283.

The International HapMap Consortium. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, pp. 52–8.

Totomoch-Serra, A., et al. (2017) 'Sanger sequencing as a first-line approach for molecular diagnosis of Andersen-Tawil syndrome', F1000Res, 6, pp. 1016.

Underhill, K.,et al. (2007) 'Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations', *Annu. Rev. Genet.*, 41 (1), pp. 539–64.

Valsesia, A., et al. (2013) 'The growing importance of CNVs: new insights for detection and clinical interpretation', *Frontiers in genetics*, 4, p.92.

Van der Auwera, et al. (2013) 'From FastQ data to high- confidence variant calls: the genome analysis toolkit best practices pipeline', *Current protocols in bioinformatics*, *43*(1), pp.11-10.

Selvaraj, S., et al. (2013) ' Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing', *Nature biotechnology*, *31*(12), pp.1111-1118.

Venter, J.C., et al. (2001) 'The sequence of the human genome', *science*, *291*(5507), pp.1304-1351

Wang, Z., Liu, X., Yang, Z. & Gelernter, J., (2013) 'The Role and Challenges of Exome Sequencing in Studies of Human Diseases', *Frontiers in Genetics,* Volume 4, pp. 1-8.

Williams A,L., et al. (2012) 'Phasing of many thousands of genotyped samples', *Am J Hum Genet*, 91(2), pp. 238- 251.

Windig, J. J., et al. (2004) 'Rapid haplotype reconstruction in pedigrees with dense marker maps', *J. Anim. Breed. Genetics*, 121, pp. 26–39.

Yang, H., et al. (2011) 'Completely phased genome sequencing through chromosome sorting', *Proc Natl Acad Sci U S A*, 108(1), pp. 12-7.

Yasuda, J., et al. (2018) 'Regional genetic differences among Japanese populations and performance of genotype imputation using whole-genome reference panel of the Tohoku Medical Megabank Project', *BMC Genomics*, 19, pp. 551.

Yoo, S., et al. (2019) 'NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants', *Genome Med* 11, 64.