Honglang Wang
Depart. of Stat. & Prob.
wangho16@msu.edu

The 1000 Genomes Project
Community Meeting @
University of Michigan

12th and 13th July 2012

Report for 1000 Genomes Project Community Meeting 2012

Abstract *This report is a conclusion and enforcing learning after the 1000 Genomes meeting. In this report, I will organize what I have learned during the meeting. In short, the 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype.*

Key Words*: Variant analysis, Imputation, Rare variants, Accuracy of variant calling.*

1 Concepts

1. Haplotype: A haplotype in genetics is a combination of alleles (DNA sequences) at adjacent locations (loci) on the chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome depending on the number of recombination events that have occurred between a given set of loci. In a second meaning, haplotype is a set of single-nucleotide polymorphisms (SNPs) on a single chromosome of a chromosome pair that are statistically associated. It is thought that these associations, and the identification of a few alleles of a haplotype block, can unambiguously identify all other polymorphic sites in its region. Such information is very valuable for investigating the genetics behind common diseases, and has been investigated in the human species by the International HapMap Project.

2. SNP: A single-nucleotide polymorphism (SNP, pronounced snip) is a DNA sequence variation occurring when a single nucleotide-A, T, C or G-in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in an individual. For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two alleles: C and T. Almost all common SNPs have only two alleles. The genomic distribution of SNPs is not homogenous, SNPs usually occur in non-coding regions more frequently than in coding regions or, in general, where natural selection is acting and fixating the allele of the SNP that constitutes the most favorable genetic adaptation.

3. Genetic recombination is the breaking and rejoining of DNA strands to form new molecules of DNA encoding a novel set of genetic information. Recombination can occur between similar molecules of DNA, as in the homologous recombination of chromosomal crossover, or dissimilar molecules, as in non-homologous end joining.

4. An allele is one of two or more forms of a gene or a genetic locus (generally a group of genes).

5. In the fields of genetics and genetic computation, a locus (plural loci) is the specific location of a gene or DNA sequence on a chromosome. A variant of the DNA sequence at a given locus is called an allele. The ordered list of loci known for a particular genome is called a genetic map. Gene mapping is the procession of determining the locus for a particular biological trait.

6. A biomarker, or biological marker, is in general a substance used as an indicator of a biological state. It is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. It is used in many scientific fields. In genetics, a biomarker (identified as genetic marker) is a DNA sequence that causes disease or is associated with susceptibility to disease.

7. Imputation, this is the estimation of missing genotype values by using the genotypes at nearby SNPs and the haplotype frequencies seen in other individuals.

8. Calling genotypes is estimating genotype values from raw data. Genotyping technology provides information about the underlying genotype, typically in the form of signal intensities or read counts of the two alleles. Statistical techniques are used to resolve this information into genotype calls. Typically, information across individuals is used, and correlation across SNPs (that is, haplotype phase) is also helpful.

2 Calling Variants

2.1 Source of Variations among Population

Since we know that the 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype, we first have to know what are the sources of variation among population.

1. A single-nucleotide polymorphism (**SNP**, pronounced snip) is a DNA sequence variation occurring when a single nucleotide-A, T, C or G-in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in an individual.

2. **Structural Variation** catches-all category includes insertions, duplications, deletions, inversions, recurring mobile elements, and other rearrangements, now usually defined as those covering 50 or more base pairs (Fig. 1). (The number is arbitrary; earlier definitions set the number at 1,000 base pairs until sequencing technologies capable of detecting smaller variants drove it down.) It is now recognized that, in

terms of the number of nucleotides, structural variation accounts for more differences between human genomes than the more extensively studied single-nucleotide differences. A 2010 study estimated that such "non-SNP variation" totaled about 50 megabases per human genome.

3. **loss-of-function (LoF) variants**-genetic changes that are predicted to be seriously disruptive to the function of protein-coding genes. These come in many forms, ranging from a single base change that creates a premature stop codon in the middle of a gene, all the way up to massive deletions that remove one or more genes completely. These types of DNA changes have long been of interest to geneticists, because they're known to play a major role in really serious diseases like cystic fibrosis and muscular dystrophy.
   But there's also another reason that they're interesting, which is more surprising: every complete human genome sequenced to date, including celebrities like James Watson and Craig Venter, has appeared to carry hundreds of these LoF variants. If those variants were all real, that would indicate a surprising degree of redundancy in the human genome. But the problem is we don't actually know how many of these variants are real—no-one has ever taken a really careful look at them on a genome-wide scale.

4. **De novo** is Latin for "from the beginning," and when describing genetic variation or mutation means that the variant has spontaneously arisen and was not inherited from either parent. In autism, de novo copy number variants are among the earliest clearly identified genetic risk factors. Given that these events are novel, natural selection has not acted on them, except for instances where the point mutation is lethal in early life. With next generation sequencing (NGS), we now have the opportunity to identify these events directly.

5. **Segregating variation** refers to the variation in the separation of paired alleles during meiosis so that members of each pair of alleles appear in different gametes.

2.2 Genotype calling & Haplotype phasing

Genotype calling is the key for the connection of NGS technology and the study of the association between genotype or haplotype and phenotype. Haplotype information is essential to the complete description and interpretation of genomes, genetic diversity and genetic ancestry. Although individual human genome sequencing is increasingly routine, nearly all such genomes are unresolved with respect to haplotype. Thus how to figure out the haplotype from the sequencing is a very important future work.

When people are doing genotype calling, So-called '**genotype likelihoods**'-which incorporate errors that may have been introduced in base calling, alignment and assembly-are coupled with prior information, such as allele frequencies and patterns of linkage disequilibrium (LD). The result is a SNP and genotype call and an associated measure of uncertainty (which is often described by a 'quality score'), both of which have a concrete statistical interpretation. An excellent paper [8] tells us one way how to do genotying.

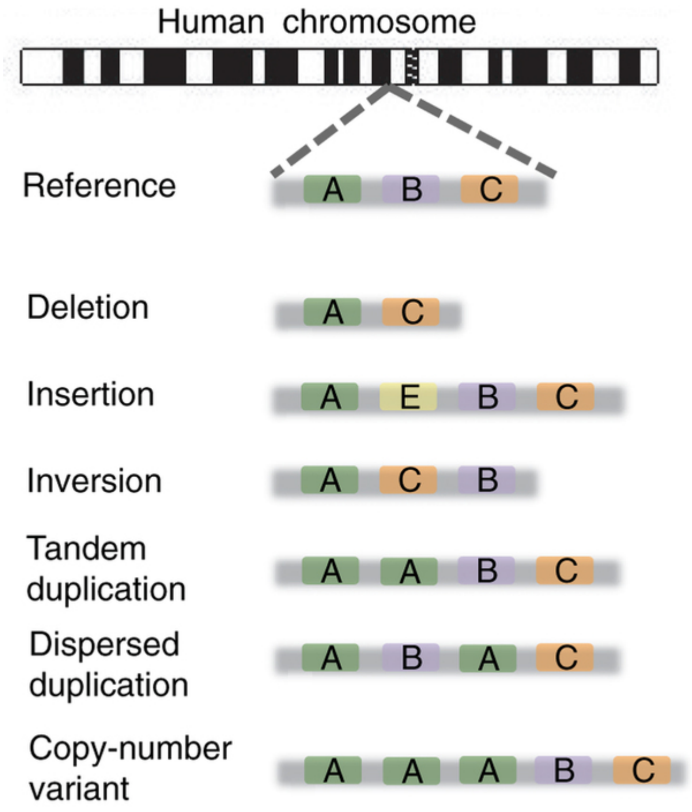


Figure 1: **Structural variation occurs in all forms and sizes.** Genome structural variation encompasses polymorphic rearrangements 50 base pairs to hundreds of kilobases in size and affects about 0.5% of the genome of a given individual.

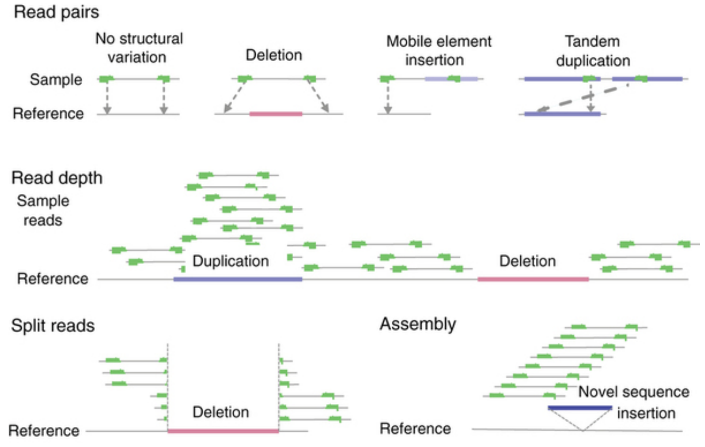
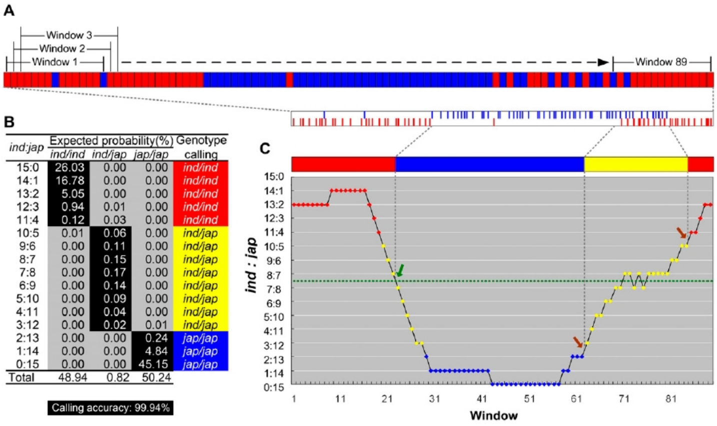


Figure 2: **Several analytic techniques are used to find structural variation.** Genome structural variation encompasses polymorphic rearrangements 50 base pairs to hundreds of kilobases in size and affects about 0.5% of the genome of a given individual.



ind:jap	Expected probability(%)			Genotype calling
	ind/ind	*ind/jap*	*jap/jap*	
15:0	26.03	0.00	0.00	*ind/ind*
14:1	16.78	0.00	0.00	*ind/ind*
13:2	5.05	0.00	0.00	*ind/ind*
12:3	0.94	0.01	0.00	*ind/ind*
11:4	0.12	0.03	0.00	*ind/ind*
10:5	0.01	0.06	0.00	*ind/jap*
9:6	0.00	0.11	0.00	*ind/jap*
8:7	0.00	0.15	0.00	*ind/jap*
7:8	0.00	0.17	0.00	*ind/jap*
6:9	0.00	0.14	0.00	*ind/jap*
5:10	0.00	0.09	0.00	*ind/jap*
4:11	0.00	0.04	0.00	*ind/jap*
3:12	0.00	0.02	0.01	*ind/jap*
2:13	0.00	0.00	0.24	*jap/jap*
1:14	0.00	0.00	4.84	*jap/jap*
0:15	0.00	0.00	45.15	*jap/jap*
Total	48.94	0.82	50.24	

Calling accuracy: 99.94%

Figure 3: **Sliding window approach for genotype calling and recombination breakpoint determination.** The table B is our rule for genotyping calling. And this rule guarantees that the calling accuracy is 99.94%. From C, we can see how to determine the recombination breakpoints.

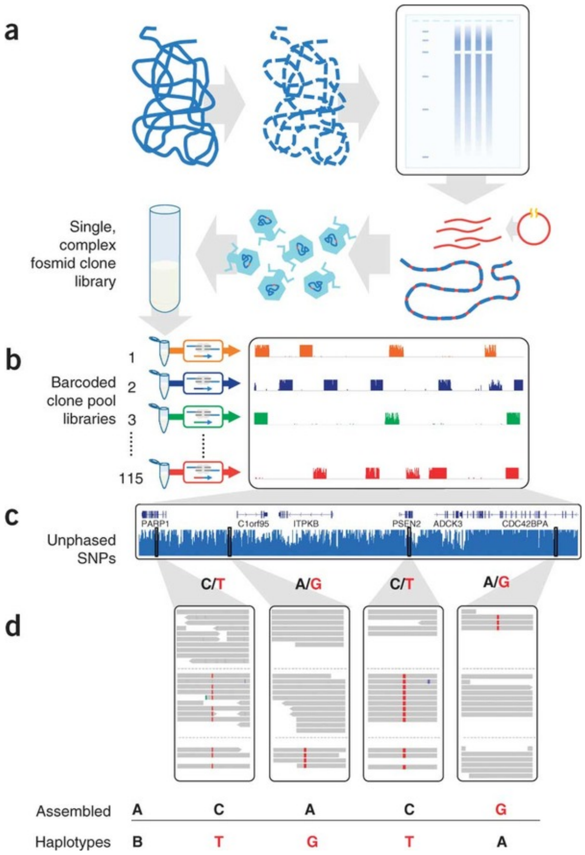


Figure 4: **Haplotype-resolved genome sequencing.** From d, Unphased variant calls were combined with haploid genotype calls to assemble haplotype blocks using a maximum parsimony approach[7]

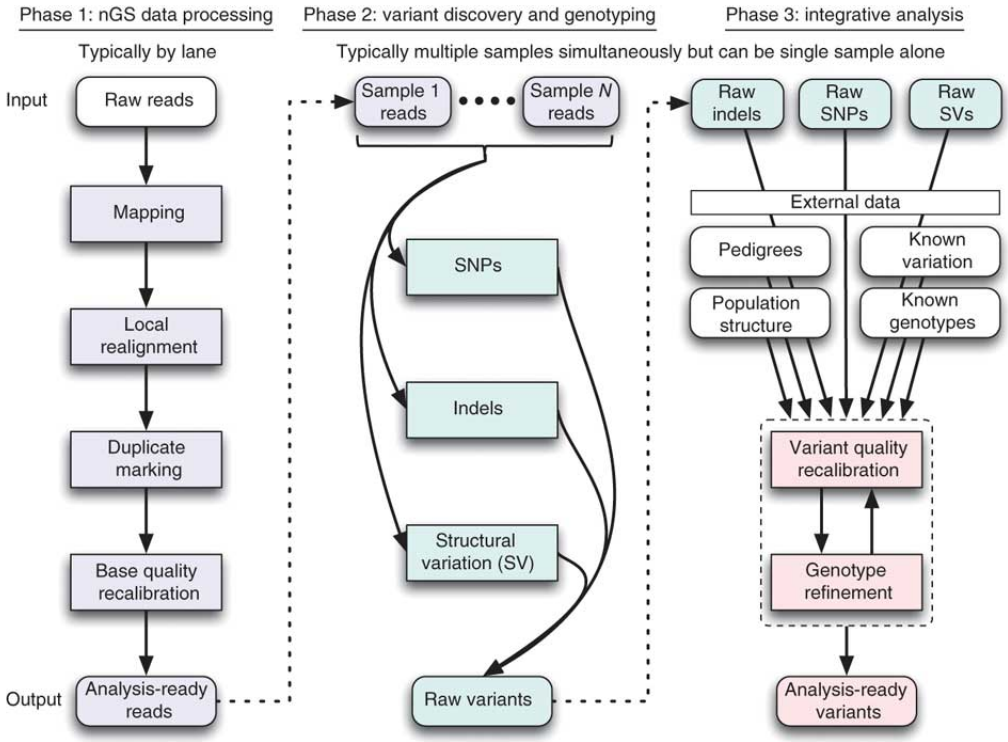


Figure 5: **Framework for variation discovery and genotyping from next-generation DNA sequencing.**

2.3 Assembly & Imputation

For high accuracy of the variant calling, sometimes we have to appeal to **de novo assembly**. Because the error of variant calling may due to the error in the reference genome, so de novo assembly is very important, although it is hard.

Genotype imputation is now an essential tool in the analysis of genomewide association scans. This technique allows geneticists to accurately evaluate the evidence for association at genetic markers that are not directly genotyped. Genotype imputation is particularly useful for combining results across studies that rely on different genotyping platforms but also increases the power of individual scans. Family samples constitute the most intuitive setting for genotype imputation.

3 Association Study

The fundamental problem for human geneticists is how to narrow to the single or few variants that are causal for a phenotype of interest. To date, nearly all successful studies applying exome sequencing to identify disease genes have adopted one of three paradigms for reducing search space. [9]

(1) For solving Mendelian disorders, a straightforward strategy involves exome sequencing of a small number of affected individuals, filtering of common variants by comparison to public SNP databases or unrelated controls, and prioritization of genes containing apparently rare, protein-altering variants in all or most affected individuals. The major advantage of this approach is that it can be independent of linkage analysis, that is, it enables the identification of the molecular basis of a Mendelian disorder without requiring access to pedigrees of sufficient size to properly map the locus, or any pedigrees, for that matter (though pedigree information can still be useful, especially for genetically heterogeneous disorders). For recessive disorders, particularly those occurring in consanguineous families, exome sequencing of just a single individual (that is, n = 2 in terms of affected chromosomes) followed by filtering of common variants may be sufficient to narrow to one or a few candidate genes.

(2) An alternative strategy involves exome sequencing of parent-child trios to identify the (approximately) one de novo coding mutation occurring per generation. This may be particularly effective for Mendelian disorders where a dominant mode of transmission is suspected and proband(s) with unaffected parents are available. More notably, however, this paradigm is being successfully applied to approach complex neuropsychiatric disorders, including intellectual disability, autism and schizophrenia. Although mutations in hundreds of genes may contribute to each of these genetically and phenotypically heterogeneous disorders, the fact that de novo, large-effect coding mutations appear to underlie a sizable proportion of sporadic cases provides a highly efficient means for identifying candidate genes.

(3) For cancer, a straightforward approach involves the pairwise comparison of exome

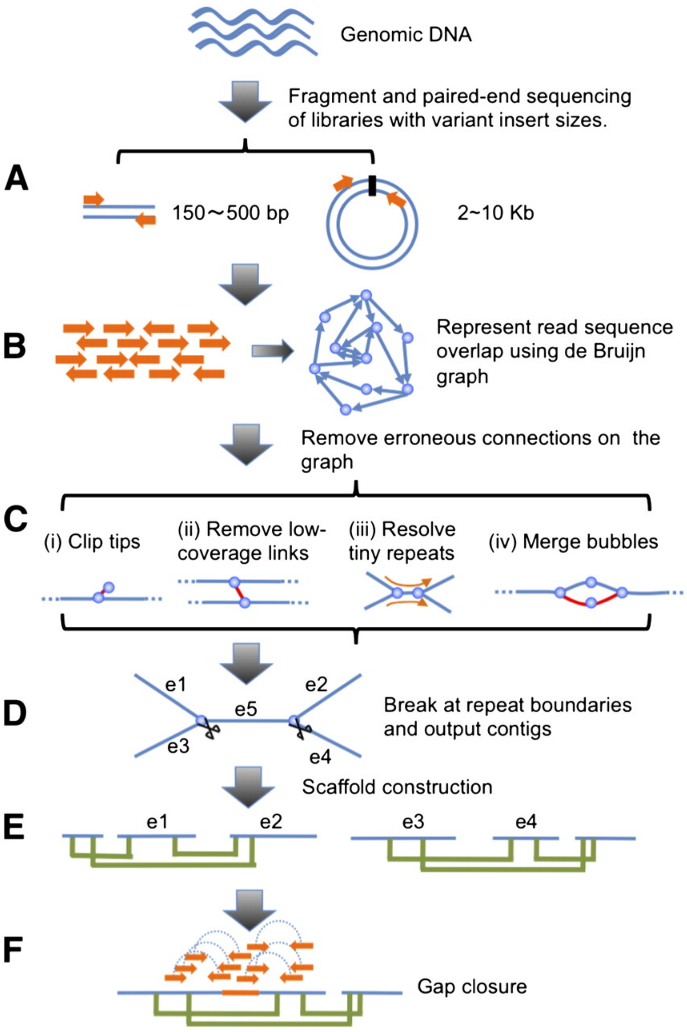
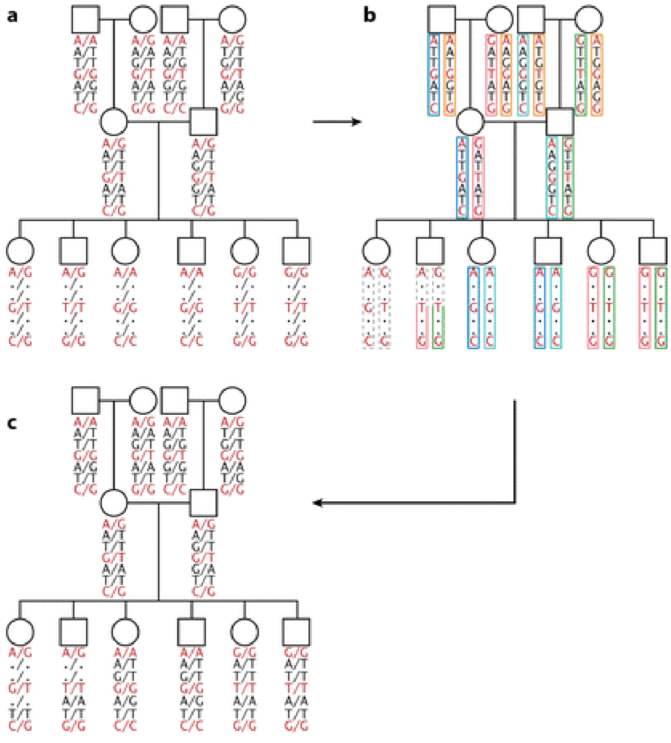

Genomic DNA
Fragment and paired-end sequencing of libraries with variant insert sizes.
A
150~500 bp
2~10 Kb
B
Represent read sequence overlap using de Bruijn graph
Remove erroneous connections on the graph
C
(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles
D
e1
e2
e5
e3
e4
Break at repeat boundaries and output contigs
Scaffold construction
E
e1
e2
e3
e4
F
Gap closure

Figure 6: **Schematic overview of the assembly algorithm.** (A) Genomic DNA was fragmented randomly and sequenced using paired-end technology. Short clones with sizes between 150 and 500 bp were amplified and sequenced directly; while long range (2-10 kb) paired-end libraries were constructed by circularizing DNA, fragmentation, and then purifying fragments with sizes in the range of 400-600 bp for cluster formation. (B) The raw or precorrected reads were then loaded into computer memory and de Bruijn graph data structure was used to represent the overlap among the reads. (C) The graph was simplified by removing erroneous connections (in red color on the graph) and solving tiny repeats by read path: (i) Clipping the short tips, (ii) removing low-coverage links, (iii) solving tiny repeats by read path, and (iv) merging the bubbles that were caused by repeats or heterozygotes of diploid chromosomes. (D) On the simplified graph, we broke the connections at repeat boundaries and output the unambiguous sequence fragments as contigs. (E) We realigned the reads onto the contigs and used the paired-end information to join the unique contigs into scaffolds. (F) Finally, we filled in the intrascaffold gaps, which were most likely comprised by repeats, using the paired-end extracted reads. [10]



Li Y, et al. 2009.
Annu. Rev. Genomics Hum. Genet. 10:387–406

Figure 7: **Genotype imputation within a sample of related individuals.** (a) The observed data, which consist of genotypes at a series of genetic markers. In this case, a subset of markers has been typed in all individuals (red), whereas the remaining markers have been typed in only a few individuals (black, in individuals in the top two generations of the pedigree). (b) The process of inferring information on identity-by-descent by examining markers for which genotypes are available in all individuals. Each IBD segment that appears in more than one individual is assigned a unique color. For example, a segment marked in blue is shared between the first individual in the grandparental generation at the top of the pedigree, the first individual in the parental generation, and individuals 3 and 4 in the offspring generation at the bottom of the pedigree. (c) Observed genotypes and IBD information have been combined to fill in a series of genotypes that were originally missing in the offspring generation. [11]

sequences of tumor and normal tissue from the same individual to distinguish the handful of somatic coding mutations from a large background of inherited variants. Exome sequencing of relatively modest numbers of matched tumor-normal pairs can yield the identification of novel, recurrent driver mutations for specific types of cancer.

3.1 Linkage Analysis-Family Based

Genetic linkage is the tendency of genes that are located proximal to each other on a chromosome to be inherited together during meiosis. Genes whose loci are nearer to each other are less likely to be separated onto different chromatids during chromosomal crossover, and are therefore said to be genetically linked.

The relative distance between two genes can be calculated by taking the offspring of an organism showing two linked genetic traits, and finding the percentage of the offspring where the two traits do not run together. The higher the percentage of descendants that do not show both traits, the farther apart on the chromosome the two genes are. Genes for which this percentage is lower than 50

Genetic linkage can also be understood by looking at the relationships among phenotypes. Among individuals of an experimental population or species, some phenotypes or traits can occur randomly with respect to one another, or with some correlation with respect to one another.

The former is known as independent assortment. Today, scientists understand that independent assortment occurs when the genes affecting the phenotypes are found on different chromosomes or separated by a great enough distance on the same chromosome that recombination occurs at least half of the time.

The latter is known as genetic linkage. This occurs as an exception to independent assortment, and develops when genes appear near one another on the same chromosome. This phenomenon causes the genes to usually be inherited as a single unit. Genes inherited in this way are said to be linked, and are referred to as ”linkage groups”. For example, in fruit flies, the genes affecting eye color and wing length are inherited together because they appear on the same chromosome.

A linkage map is a genetic map of a species or experimental population that shows the position of its known genes or genetic markers relative to each other in terms of recombination frequency, rather than a specific physical distance along each chromosome. Linkage mapping is critical for identifying the location of genes that cause genetic diseases.

A typical value for estimating recombination frequency is LOD score. The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance. A typical plot is in the Fig. 8

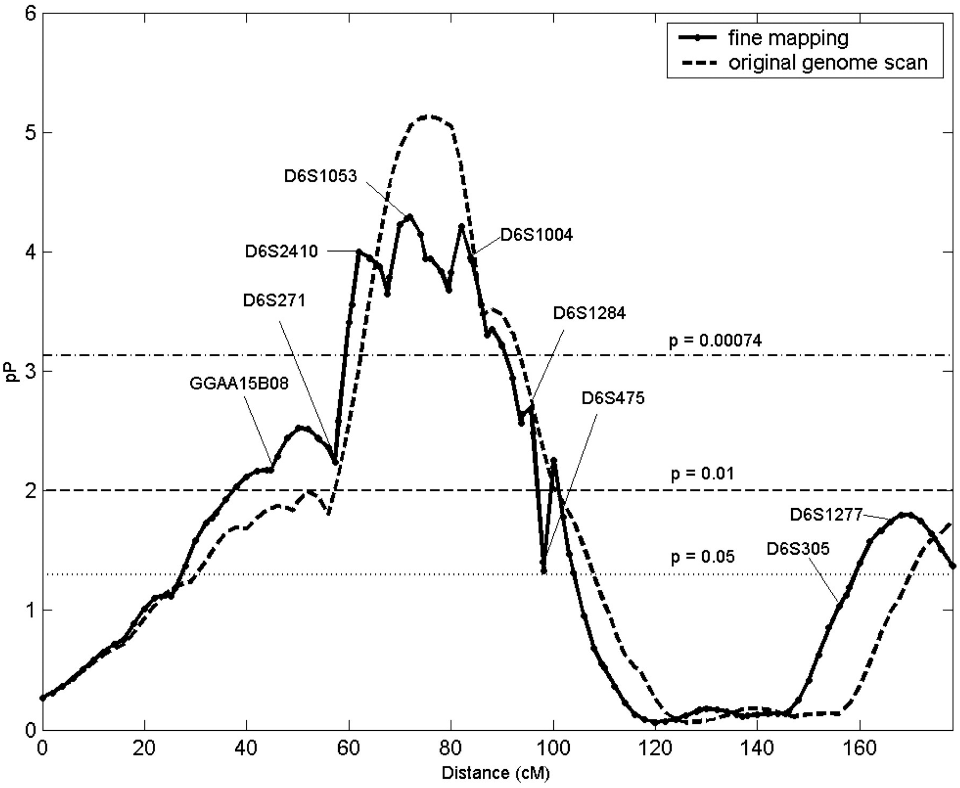


Figure 8: **A typical plot for genetic linkage analysis.** Genetic distance (cM) is plotted on the x axis against pP = -log10 (P value) on the y axis.

3.2 GWAS-Case and Control

In genetic epidemiology, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait. GWAS typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major diseases.

These studies normally compare the DNA of two groups of participants: people with the disease (cases) and similar people without (controls). Each person gives a sample of DNA, from which millions of genetic variants are read using SNP arrays. If one type of the variant (one allele) is more frequent in people with the disease, the SNP is said to be "associated" with the disease. The associated SNPs are then considered to mark a region of the human genome which influences the risk of disease. In contrast to methods which specifically test one or a few genetic regions, the GWA studies investigates the entire genome. The approach is therefore said to be non-candidate-driven in contrast to gene-specific candidate-driven studies. GWA studies identify SNPs and other variants in DNA which are associated with a disease, but cannot on their own specify which genes are causal.

And the results are often denoted by the following two pictures.

3.3 NGS-GWAS-Individual

Genome-wide association studies (GWAS), using tag SNPs in genome to analyze their association with diseases, follow a hypothesis-free approach and interrogate the majority of common SNPs across the human genome. It is designed to identify possible genetic variants that contribute to complex diseases. In the past five years, more than 100 complex diseases and traits have been studied by GWAS and numerous susceptibility genes/loci were identified.

However, the large-scale genome-wide association studies based on SNP genotyping have only identified a small fraction of the heritable variation of these diseases. One explanation is that many rare variants (a minor allele frequency, MAF¡5%), which are not included in the common genotyping platforms, contribute substantially to the genetic variation of these diseases. Recently, exome sequencing of 200 individuals from Denmark uncover more deleterious rare variants than expected, which also support that much of the heritable variation affecting fitness is caused by low-frequency mutations, which are often overlooked in the studies based on genotyping but not resequencing.

Next-generation GWAS is the next-generation sequencing based GWAS, which has the advantage of uncovering novel causative genetic mutations of human diseases through the combination of high-throughput sequencing and genotyping. Massively parallel sequencing of exome and targeted regions (which has been found by previous GWAS) are two promising and effective approaches to find missing heritability of complex diseases, by capturing more valuable data beyond common SNPs.

There are two novel strategies based on next-generation GWAS to discover novel and low-frequency causative genetic mutations associated with human complex diseases.

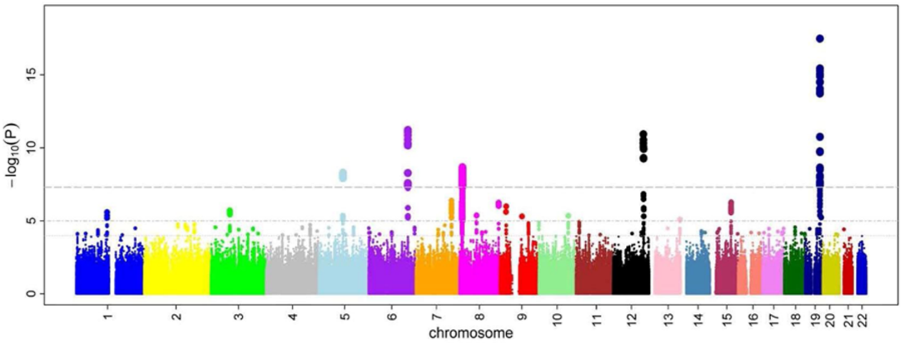


Figure 9: **Manhattan Plot for GWAS.** An illustration of a Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. This example is taken from a GWA study investigating microcirculation, so the tops indicates genetic variants that more often are found in individuals with constrictions in small vessels.

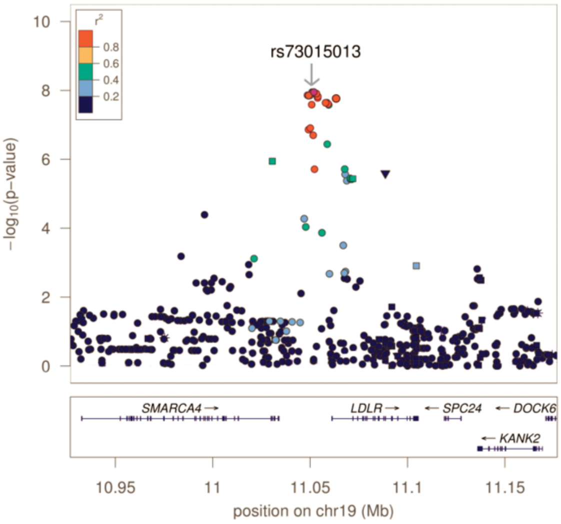


Figure 10: **Regional association plot for GWAS.** Regional association plot, showing individual SNPs in the LDL receptor region and their association to LDL-cholesterol levels. This type of plot is similar to the Manhattan plot in the lead section, but for a more limited section of the genome. The haploblock structure is visualized with colour scale and the association level is given by the left Y-axis. The dot representing the rs73015013 SNP (in the top-middle) has a high Y-axis location because this SNP explains some of the variation in LDL-cholesterol.

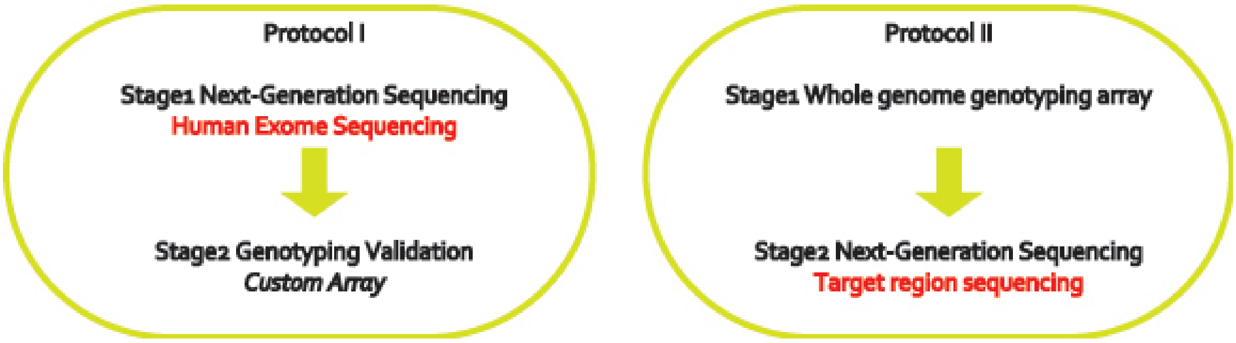

Protocol I
Stage1 Next-Generation Sequencing
Human Exome Sequencing
Stage2 Genotyping Validation
Custom Array
Protocol II
Stage1 Whole genome genotyping array
Stage2 Next-Generation Sequencing
Target region sequencing

Figure 11: **Finding the missing heritability by Next-Generation GWAS.**

Protocol I : Exome Sequencing & Genotyping validation

At the first stage of this two-stage design, we suggest applying exome sequencing of hundreds of cases and hundreds of controls to select the associated SNPs by allele frequency estimation. At the second stage, validate the best candidate SNPs selected from the first stage by genotyping in a larger set of individuals. This protocol is cost-effective and has the potential to detect rare SNPs that would not be captured by any of the major genotyping platforms.

Protocol II : Genome genotyping & Target region sequencing

At the first stage, a genome-wide genotyping is used to scan the case and control samples to obtain the candidate loci. At the second stage, using designed chip to capture these candidate loci or targeted regions, then sequencing the targeted regions in large-scale samples to verify these candidate loci, so as to identify disease-associated mutations.

4 Future Work

There are two things I want to mention as the future work:

1. The first one is the concept of ”a lattice of sequenced genome”, which is raised by Goncalo. It is saying that we should make the lattice of sequenced genome, whose lattice points are biomarkers, denser and denser by deep sequencing, broader sequencing, imputation and so on.

2. The second one is we need to appeal more on statistical genetics at every level, such as eQTL(mRNA level), mQTL(methylation), pQTL(protein level independent of mRNA), and miRNA QTL.

References

[1] Eric Banks, et,al *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature Genetics 43, 491-498 (2011)

[2] Hyun Min Kang, et,al, *A map of human genome variation from population-scale sequencing.* Nature 467, 1061-1073 (28 October 2010)

[3] Daniel G. MacArthur, et,tal *A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes.* Science 17 February 2012: Vol. 335 no. 6070 pp. 823-828

[4] Monya Baker, *Structural variation: the genome's hidden architecture.* Nature Methods 9, 133-137 (2012)

[5] Sharon R. Browning and Brian L. Browning, *Haplotype phasing: existing methods and new developments.* Nature Reviews Genetics 12, 703-714 (October 2011)

[6] Jay Shendure, et,al *Haplotype-resolved genome sequencing of a Gujarati Indian individual.* Nature Biotechnology 29, 59-63 (2011)

[7] Vikas Bansal and Vineet Bafna, *HapCUT: an efficient and accurate algorithm for the haplotype assembly problem.* Bioinformatics (2008) 24 (16): i153-i159.

[8] Xuehui Huang, Qi Feng, Qian Qian, et al, *High-throughput genotyping by whole-genome resequencing.* Genome Res. 2009 19: 1068-1076

[9] Jay Shendure, *Next-generation human genetics.* Genome Biology 2011, 12:408

[10] Yingrui Li, *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res. 2010. 20: 265-272

[11] Goncalo Abecasis, *Genotype Imputation.* Annual Review of Genomics and Human Genetics Vol. 10: 387-406

[12] Jay Shendure, et,al *Exome sequencing as a tool for Mendelian disease gene discovery.* Nature Reviews Genetics 12, 745-755 (November 2011)