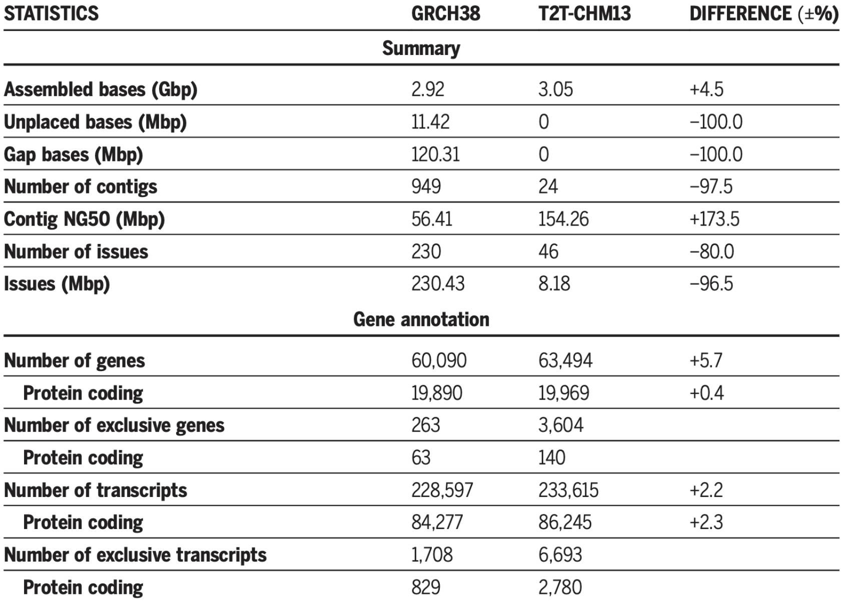
BIM3007-Assignment1

Junyang Deng (120090791)

October 16, 2022

1 Human genome project

1.1 According to the most updated paper on the human genome released in April 2022 [1], the size of human genome is 3.055 billion base pairs (bp). The updated number of protein-coding region is around 19,969 protein coding genes.
On average, a Chinese individual carries 3,068,811 SNVs and 257,832 INDELs, including 9106 missense variants, 10 stop loss, 73 stop gain, and 190 frameshift or non-frameshift INDELs, as reported by Westlake Biobank for Chinese (WBBC) in 2022 [2].



STATISTICS	GRCH38	T2T-CHM13	DIFFERENCE (±%)
Summary			
Assembled bases (Gbp)	2.92	3.05	+4.5
Unplaced bases (Mbp)	11.42	0	−100.0
Gap bases (Mbp)	120.31	0	−100.0
Number of contigs	949	24	−97.5
Contig NG50 (Mbp)	56.41	154.26	+173.5
Number of issues	230	46	−80.0
Issues (Mbp)	230.43	8.18	−96.5
Gene annotation			
Number of genes	60,090	63,494	+5.7
Protein coding	19,890	19,969	+0.4
Number of exclusive genes	263	3,604	
Protein coding	63	140	
Number of transcripts	228,597	233,615	+2.2
Protein coding	84,277	86,245	+2.3
Number of exclusive transcripts	1,708	6,693	
Protein coding	829	2,780	

Figure 1: Comparison between GRCH38 and T2T-CHM13

1.2 The gene structure is illustrated in Figure 2. Usually, the transcription start site (TSS) denotes the beginning of gene. After TSS, there is a short segment called 5' untranslated region (UTR). The main body of gene primarily consists of two parts: exon and intron. Exons encode genes that will be translated to protein, while genes in intron region will be truncated during mRNA processing. At the end of the gene, there is a translation STOP codon, which is followed by 3' UTR.

Genetic variants in different parts of gene exhibit different effects.

- Some mutations can affect the **presence** of a gene product. For example, if mutation occurs in the **regulatory region**, transcription factors can no longer bind to the sequence, and this will change the transcription status of gene. Similar consequences happen when missense occurs in start codon (**start loss**). This can also make the gene product disappear.

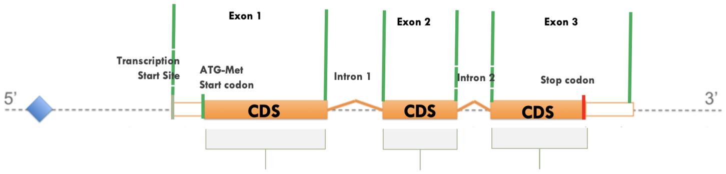


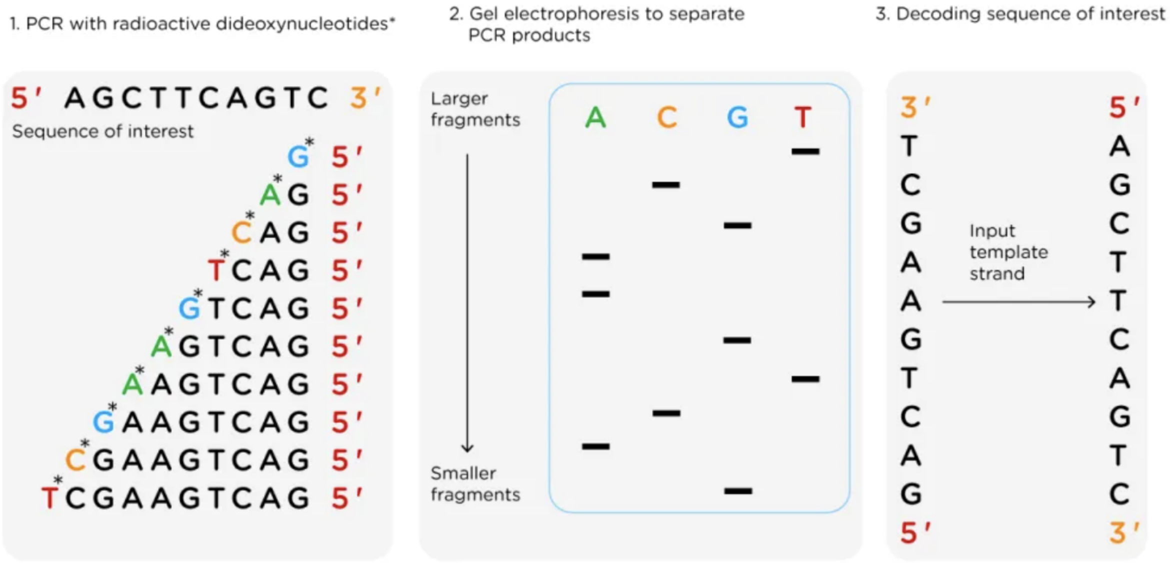
Figure 2: Typical gene structure

- Some mutations cause protein products to **malfunction**. Insertions, deletions, and nonsynonymous substitutions that happen in the genic area will change the composition of gene product. Whether or to what extend do genic mutation affect gene product depends on its location. If the mutation alters functional region of the gene product, the effect will be deleterious. If the mutation does not alter the functional region of the resulted protein, the effect will be more moderate.
- Some mutations, like synonymous mutations, have **no obvious effect** on gene products.

2 Evolution of Sequencing technology

2.1 Principles of first, second, third generation sequencing technologies.

First generation The first generation sequencing is also called Sanger sequencing. The procedure of Sanger sequencing is illustrated in Figure 3. First, dideoxynucleotide (ddNTP) are added to four tubes. When ddNTP is incorporated into the DNA, DNA polymerase can no longer add nucleotides to the chain, and the synthesis will be terminated. ddNTPs used in the experiment are labelled by fluorescent dyes. Therefore, the result product of the first step will be DNA fragments of various length with ddNTP at the 3' end. Next, these fragments will be sent to run electrophresis. In electrophoresis, fragments will be sorted according to their length. Combining the location of fragments and the fluorescent color they exhibit, we can decode the sequence of interest.



* Dideoxynucleotides are added to a normal PCR mixture (which includes regular nucleotides) because the addition of dideoxynucleotides to an elongating DNA sequence causes premature termination of that sequence. Radiolabeling of the dideoxynucleotides allows researchers to visualize the last nucleotide in each PCR product sequence.

Figure 3: Sanger Sequencing

Second generation The second generation sequencing (also called next generation sequencing) is a high-throughput, parallel sequencing method. The main difference between first and second generation is that, second generation sequencing adopts modified dNTPs with fluorescently labelled terminator (not ddNTP anymore). When the labelled dNTP is incorporated in the sequence, the synthesis will be paused. Then, the machine will use laser to excite the fluorescent label, and synthesis can continue again. Similar to Sanger sequencing, the nucleotide can also be inferred from the color of fluorescent signal. To control the quality of reads, NGS has a bridge amplification process. Overall, the next generation sequencing method is much faster and more accurate than Sanger sequencing.

Third generation The third generation is designed to read longer sequences in order to overcome the drawbacks of the first and second generation sequencing. Different companies are using fundementally different approaches to decode the DNA sequence. For example, Pacific Biosciences (PacBio) uses a Single-molecule real-time sequencing (SMRT-seq) method. This method detects the fluorescent signal right beside the DNA polymerase. Also, to ensure sequencing quality, circular consensus DNAs are formed. This allows the polymerase to sequence the fragments repeatedly [3]. Different from PacBio, Oxford Nanopore technology (ONT) detects the change of current in the pore of nanoprotein. Compared to PacBio, ONT reads longer sequences but with a higher error rate.

2.2 The pros and cons of different sequencing technologies.

- First generation
  - pros: High accuracy
  - cons: short read length (500 bp - 1,000 bp), time-consuming, labor-intensive, low throughput, cannot be used in highly repetitive region
- Second generation
  - pros: high-throughput, high accuracy (99.9%), higher coverage, higher sequence depth, fast speed.
  - cons: short read length (50 bp - 500 bp), cannot be used in highly repetitive region
- Third generation
  - pros: long read (usually tens of kbs fragments), can directly detect epigenetic modifications, suitable for de novo genome assembly [4].
  - cons: lower accuracy (ONT 96%, PacBio HiFi $>$ 99%)

3 Exome-sequencing data analysis

3.1 Exome includes all protein coding genes. Exome sequencing means to sequence the exome part of genome. The main goal of exome-sequencing technology is to figure out what loci are responsible for certain diseases at a relatively low cost.

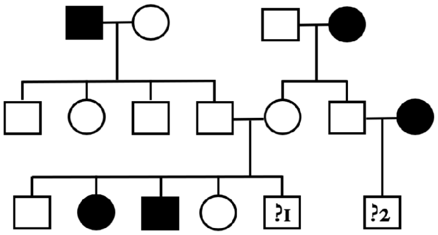
3.2 The process of Exome-sequencing data analysis includes three phases: raw data processing (denoising the data), variant calling, and integrative analysis.

- The first phase includes mapping, local realignment, duplicate marking, and base quality recalibration. The output will be analysis-ready reads.
  - Mapping: Raw sequencing data are many fragments of the sample DNA which includes replicates. So, the first step of analysing the reads is to map these short reads to a reference genome. (Software: BWA (Burrows-Wheeler Aligner))

- Local realignment: Some places are harder to alignment, for example indels. In order to improve the mapping quality, indel realignment is needed. In indel realignment, the software will first identify regions where realignments are needed. Then, reads in these regions will be realigned using a more accurate alignment strategy that takes all reads in the region into account.
- Duplicate marking: Fragments are duplicated during library construction. Before analysing the sequence, duplicates should be removed to avoid disturbance. (Software: Picard (by broadinstitude) (github codes, introduction))
- Base quality recalibration: Since sequencer might not procide the correct base quality score, we need to recalibrate it before variant calling. (Software: BQSR (Base Quality Score Recalibration): A machine learning approach to recalibrate base quality scores.)
- After all these processes, analysis-ready reads will be generated.

- The second phase, variant calling phase, analyzes the alignment files to get variants.
  - There are three main types of variants: single nucleotide polymorphism (SNP), indel (insertion and deletion) and structural variant (SV). (Software: we can use CoNIFER to find copy number variants)
- The third phase is integrative analysis. In this phase, external factors that can affect the variant calling result will be considered. With extra data, we can recalibrate variants and refine the genotype.
  - For example, variants that lie in regions with bad mapping quality are less trustful than others. So as variants that lie in regions with low depth. A software called VQSR (Variant Quality Score Recalibration) can do this job. It is a sophisticated filtering technique applied on the variant callset that uses machine learning to model the technical profile of variants in a training set and uses that to filter out probable artifacts from the callset.

4 Exome-sequencing data analysis (application)



4.1 The most likely inheritance is **autosomal recessive**. It is recessive because two patients appear in the third generation, but their parents are healthy individuals. It's also not X-linked recessive. If it is, II-6 must be ill because his mother passes a recessive gene to him.

4.2 The genotypes of child1's parents are Aa. So, the probability for them to have a diseased child is 1/4. As for the child2, his father's genotype must be Aa, so the probability of his disease is 1/2.

4.3 Experiment design:

1. The first step of the experiment is to conduct sequencing on ill individuals. I will choose to sequence III-2 and III-3 because 1) this combination only involves 2 people thus the cost is low, 2) this disease is autosomal recessive, which requires two copies of genes to be recessive to get the disease. To reduce the search space, I choose people with lowest P(IBD=2) to sequence. In this case, the probability of IBD=2 is 1/4 for III-2 and III-3, which is the lowest of all.

2. The second step is quality control. Data will go through the GATK pipeline (discussed in section 3.2). After that, a set of high quality variants are selected.
3. Then, I will do sex examination and familial relation examination on the data. This can avoid false labeling of the samples. Sex is determined by calculating the homozygous ratio in X-chromosome. Familial relationship is estimated by calculating the pairwise IBDs among samples.
4. Next, we will filter out some variants based on variant frequency. In this step, we need to know the prevalence of this disease in the population. For instance, assume this disease has 100% penetrance, if its prevalence is 0.01%, the allele frequency should not be higher than 0.1%.
5. After that, we will follow the ACMG guideline to interpret the candidate variants [5]. The ACMG guideline combines computational (predicted) data, functional knowledge and clinical knowledge to help people understand the pathogenic effect of variants. This procedure can be completed using InterVar.

I followed the ACMG guideline to define the pathogenic variant. The ACMG guideline uses typical types of variant evidence to classify variants into five categories −“pathogenic”, “likely pathogenic”, “uncertain significance”, “likely benign”,and “benign”. Typical types of variant evidence include population data (allele frequency), computational data (SIFT, PolyPhen, CADD), functional data (in vitro, in vivo), and clinical data (family history, segregation) etc.

The most important criterion for defining a pathogenic variant is the PVS1, which means a loss-of-function (LOF) mutation of a gene with known mechanism. Mutations that disrupt gene function by leading to a complete absence of gene products include nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single exon or multiexon deletion [5]. Once it is confirmed that the mutation is PVS1, the mutation will be classified as pathogenic or likely pathogenic, depending on whether it also meets other criteria.

After finding out all the evidence, I can decide which category the variant belongs to. As shown in the Appendix, the combination of certain types of evidence can help decide the pathogenicity of the variant.

5 GWAS

	BB	Bb	bb
AA	23	51	34
Aa	37	48	35
aa	13	19	11

5.1 Estimate the frquencies of the four haplotypes: AB, Ab, aB, ab using EM algorithm.

$$\mathrm{AB} = 23 \times 2 + 51 + 37 = 134$$
$$\mathrm{Ab} = 51 + 34 \times 2 + 35 = 154$$
$$\mathrm{aB} = 37 + 13 \times 2 + 19 = 82$$
$$\mathrm{ab} = 35 + 19 + 11 \times 2 = 76$$

1. First iteration:

$$\mathrm{P(AB\ and\ ab)} = 0.5$$
$$\mathrm{P(aB\ and\ Ab)} = 0.5$$

$$\begin{aligned} \mathrm{AB} &= 134 + 48 \times 0.5 = 158, & f_{\mathrm{AB}} &= 158/542 = 0.2915 \\ \mathrm{Ab} &= 154 + 48 \times 0.5 = 178, & f_{\mathrm{Ab}} &= 178/542 = 0.3284 \\ \mathrm{aB} &= 82 + 48 \times 0.5 = 106, & f_{\mathrm{aB}} &= 106/542 = 0.1956 \\ \mathrm{ab} &= 76 + 48 \times 0.5 = 100, & f_{\mathrm{ab}} &= 100/542 = 0.1845 \end{aligned}$$

$$\begin{aligned}&\text{P(AB and ab)} = 0.2915 \times 0.1845/(0.2915 \times 0.1845 + 0.3284 \times 0.1956) = 0.4557\\&\text{P(Ab and aB)} = 0.5443\end{aligned}$$

2. Second iteration:

$$\begin{aligned}\text{AB} &= 134 + 48 \times 0.4557 = 155.8736, & f_{\text{AB}} &= 0.2876\\ \text{Ab} &= 154 + 48 \times 0.5443 = 180.1264, & f_{\text{Ab}} &= 0.3323\\ \text{aB} &= 82 + 48 \times 0.5443 = 108.1264, & f_{\text{aB}} &= 0.1995\\ \text{ab} &= 76 + 48 \times 0.4557 = 97.8736, & f_{\text{ab}} &= 0.1806\end{aligned}$$

$$\begin{aligned}\text{P(AB and ab)} &= 0.4392\\ \text{P(Ab and aB)} &= 0.5608\end{aligned}$$

3. Third iteration:

$$\begin{aligned}\text{AB} &= 134 + 48 \times 0.4392 = 155.0816, & f_{\text{AB}} &= 0.2861\\ \text{Ab} &= 154 + 48 \times 0.5608 = 180.9184, & f_{\text{Ab}} &= 0.3338\\ \text{aB} &= 82 + 48 \times 0.5608 = 108.9184, & f_{\text{aB}} &= 0.2010\\ \text{ab} &= 76 + 48 \times 0.4392 = 97.0816, & f_{\text{ab}} &= 0.1791\end{aligned}$$

$$\begin{aligned}\text{P(AB and ab)} &= 0.4331\\ \text{P(Ab and aB)} &= 0.5669\end{aligned}$$

4. Fourth iteration:

$$\begin{aligned}\text{AB} &= 134 + 48 \times 0.4331 = 154.7888, & f_{\text{AB}} &= 0.2856\\ \text{Ab} &= 154 + 48 \times 0.5669 = 181.2112, & f_{\text{Ab}} &= 0.3343\\ \text{aB} &= 82 + 48 \times 0.5669 = 109.2112, & f_{\text{aB}} &= 0.2015\\ \text{ab} &= 76 + 48 \times 0.4331 = 96.7888, & f_{\text{ab}} &= 0.1786\end{aligned}$$

$$\begin{aligned}\text{P(AB and ab)} &= 0.4309\\ \text{P(Ab and aB)} &= 0.5691\end{aligned}$$

5. Fifth iteration:

$$\begin{aligned}\text{AB} &= 134 + 48 \times 0.4309 = 154.6832, & f_{\text{AB}} &= 0.2854\\ \text{Ab} &= 154 + 48 \times 0.5691 = 181.3168, & f_{\text{Ab}} &= 0.3345\\ \text{aB} &= 82 + 48 \times 0.5691 = 109.3168, & f_{\text{aB}} &= 0.2017\\ \text{ab} &= 76 + 48 \times 0.4309 = 96.6832, & f_{\text{ab}} &= 0.1784\end{aligned}$$

$$\begin{aligned}\text{P(AB and ab)} &= 0.4300\\ \text{P(Ab and aB)} &= 0.5700\end{aligned}$$

5.2 Describe whether the two SNPs are in linkage disequilibrium.

The allele frequencies and the estimated haplotype frequencies (from EM-algorithm) are shown in Table 1 and 2.

$$\begin{aligned}&D = P(AB) - P(A)P(B) = 0.2854 - 0.6199 \times 0.4871 = -0.0166\\&D_{\max} = -P_aP_b = -0.3801 \times 0.5129 = -0.1950\\&D' = D/D_{\max} = 0.0166/0.1950 = 0.0851\\&R^2 = D^2/(P_AP_BP_aP_b) = 0.0166^2/(0.3801 \times 0.5129 \times 0.6199 \times 0.4871) = 0.0047\end{aligned}$$

Since R^2 is less than 0.1, the two SNPs are likely independent (not in linkage disequilibrium).

haplotype	Frequency
AB	0.2854
Ab	0.3345
aB	0.2017
ab	0.1784

Table 1: Estimated haplotype frequencies

Allele	Frequency
A	0.6199
a	0.3801
B	0.4871
b	0.5129

Table 2: Estimated allele frequencies

Appendix

The ACMG guidelines for reporting of incidental findings in clinical exome and genome sequencing [5].

Pathogenic:

1. 1 Very strong (PSV1) AND
   (a) $\geq$ 1 Strong (PS1-PS4) OR
   (b) $\geq$ 2 Moderate (PM1-PM6) OR
   (c) 1 Moderate (PM1-PM6) and 1 supporting (PP1-PP5) OR
   (d) $\geq$ Supporting (PP1-PP5)
2. $\geq$ 2 Strong (PS1-PS4) OR
3. 1 Strong (PS1-PS4) AND
   (a) $\geq$ 3 Moderate (PM1-PM6) OR
   (b) 2 Moderate (PM1-PM6) AND $\geq$ Supporting (PP1-PP5) OR
   (c) 1 Moderate (PM1-PM6) AND $\geq$ 4 Supporting (PP1-PP5)

References

[1] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[2] Pei-Kuan Cong, Wei-Yang Bai, Jin-Chen Li, Meng-Yuan Yang, Saber Khederzadeh, Si-Rui Gai, Nan Li, Yu-Heng Liu, Shi-Hui Yu, Wei-Wei Zhao, et al. Genomic analyses of 10,376 individuals in the westlake biobank for chinese (wbbc) pilot project. *Nature Communications*, 13(1):1–15, 2022.

[3] Aaron M Wenger, Paul Peluso, and et al. Rowell. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10):1155–1162, 2019.

[4] PacBio. What are the benefits of hifi sequencing?, 2020.

[5] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*, 17(5):405–423, 2015.