BIM3007 Assignment 1

Student name: HU Linfeng

Student ID: 120090473

1. HGP

   1.1 Describe the size of human genome, the updated number of protein-coding genes identified in the human genome, as well as the average number of genetic variants each Chinese people could have.

   1.2 Please illustrate a gene structure, describe potential impact of genetic variants on a gene product.

Answer:

(1.1)

Size of human genome: in the haploid genome, human has approximately 3 billion base pairs (bps).

Updated number of protein-coding genes: 19370 (data from GENCODE latest release via https://www.gencodegenes.org/human/stats_41.html)

The average number of genetic variants each Chinese people could have: approximately 2500.

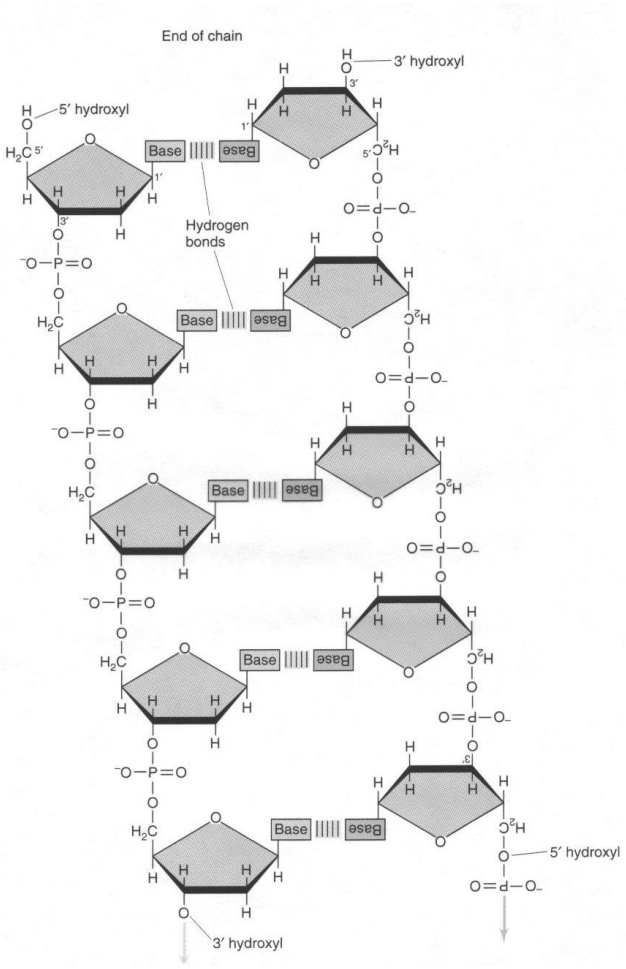
(1.2)

Gene structure:

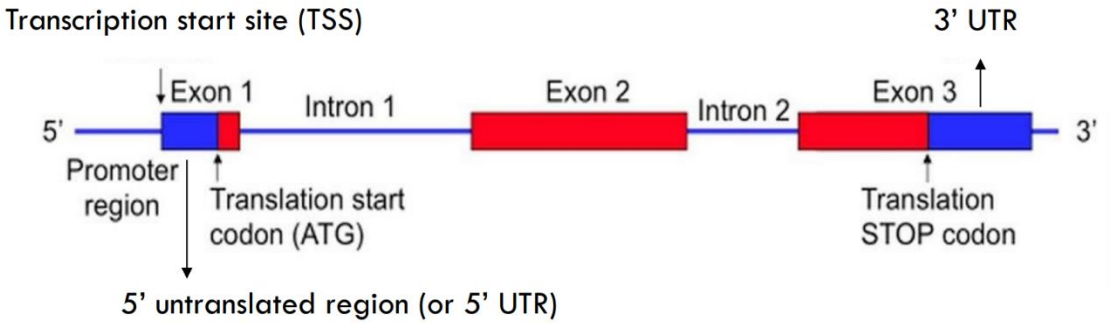
Genes are composed of deoxyribonucleic acid (DNA) in most of the cases. For some viruses, they have genes consisting of a closely related compound called ribonucleic acid (RNA). A DNA molecule is composed of two chains of nucleotides that wind about each other to resemble a twisted ladder, which is linked by phosphodiester (ribose to ribose). Base pairs match each other, forming hydrogen bonds that help shape the structure of genes. A fixed sequence of anti-parallel base pairs that code for a particular genetic code.

The schematic diagram can be shown as below:

Note: to make it clearly, the three dimensional helix is not shown here.



For a DNA, it consists of promoter region, exon, intron, gap DNA region. It can be shown as:



Potential impact of genetic variants:

Genetic variants can prevent one or more proteins from normal function. By changing a gene's instructions for making a protein, a variant can cause a protein to malfunction or to not be produced at all. When a variant alters a protein that plays a critical role in the body, it can disrupt normal development and cause anomaly.

Genetic variations that alter gene activity or protein function can introduce different

traits in an organism. If the variation is not suitable for the environment, it will pose threat to the organism.

2. Evolution of Sequencing Technology

2.1 Describe the principle of 1st, 2nd , and 3rd generation sequencing technologies.

2.2 Describe the pros and cons of these sequencing technologies.

Answer:

First generation sequencing (Sanger's sequencing):

Principles: A DNA primer complementary to the template DNA is used to be a starting point for DNA synthesis; in dNTPs (A, T, C, and G), the polymerase extends the primer by adding the complementary dNTP to the template DNA strand. To determine which nucleotide is incorporated into the chain of nucleotides, four ddNTPs (ddATP, ddGTP, ddCTP, and ddTTP) labeled with a distinct fluorescent dye are used to terminate the synthesis. After the synthesis, the reaction products are loaded into four lanes of a single gel depending on the diverse chain-terminating nucleotide and subjected to gel electrophoresis. According to their sizes, the sequence of the DNA is thus determined.

Pros: cost-efficiency for sequencing single genes, 99.99% accuracy, especially suitable for verification sequencing for site-directed mutagenesis or cloned inserts.

Cons: limited read length (500-1000 bp fragments), sometimes the quality of a Sanger sequence is not accurate enough in the first 15 to 40 bases since it is the position primer binds.

Second generation sequencing (Next-generation sequencing/NGS):

Principles: The next-generation sequencing process involves fragmenting DNA/RNA into multiple pieces, adding adapters, sequencing the libraries, and reassembling them to form a genomic sequence. NGS modified dNTPs such that they can be reversible terminators with distinct labels. Bridge amplification and sequencing by double-sided sequencing accelerates the whole sequencing process.

Pros: higher efficiency, higher sensitivity to detect low-frequency variants, high

throughput with parallelization of sequencing reaction.

Cons: cost, needs further error correction.

Third generation sequencing:

Principles: Utilizing the long-read sequencing techniques instead of PCR-based synthesis sequencing. PacBio developed the sequencing platform of single molecule real-time sequencing (SMRT), based on the properties of zero-mode waveguides. Signals are in the form of fluorescent light emission from each nucleotide incorporated by a DNA polymerase bound to the bottom of the zL well. Oxford Nanopore's technology involves passing a DNA molecule through a nanoscale pore structure and then measuring changes in electrical field surrounding the pore; while Quantapore has a different proprietary nanopore approach. Stratos Genomics spaces out the DNA bases with polymeric inserts, "Xpandomers", to circumvent the signal to noise challenge of nanopore ssDNA reading.

Pros: PCR-free, higher efficiency, longer reads (tens of kb fragments), excellent portability

Cons: relatively lower accuracy, difficulty in processing and analyzing data (computation challenge in blurred signals).

3. Exome-sequencing data analysis
   - 3.1 Describe the principle and main goal of Exome-sequencing technology.
   - 3.2 Describe the process of Exome-sequencing data analysis and the related software.

Answer:

(3.1)

Principle of exome-sequencing technology:

Exome-sequencing technology focusing on the sequencing of the protein-coding region in a genome. It first selects the protein coding region in DNA and then, sequences the exonic DNA using any high-throughput DNA sequencing technology.

Main goal of exome-sequencing technology:

The main goal is to identify genetic variants that alter protein sequences. Also, it can be cheaper than whole-genome sequencing.

(3.2)

The process of Exome-sequencing data analysis and the related software:

Three phases: NGS data processing, variant discovery and genotyping, and integrative analysis.

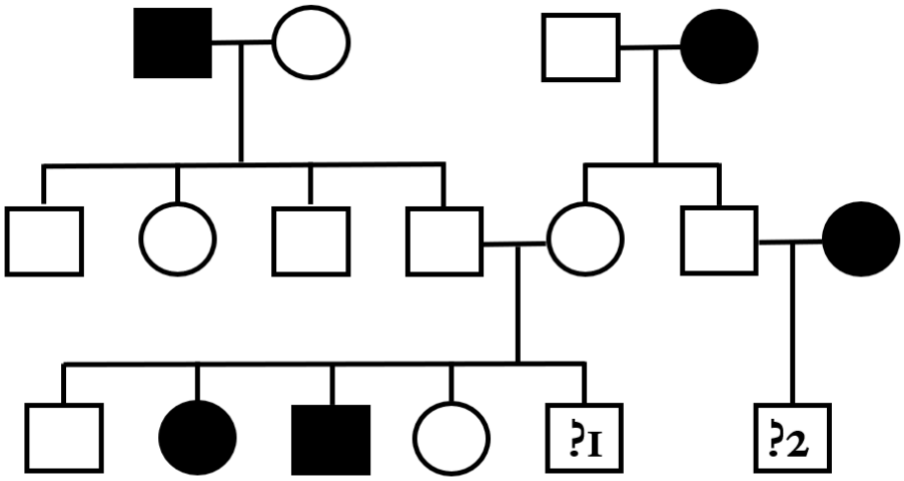
In NGS data processing, we can use QC, BWA, Samtools, and Picard to do the mapping and duplicate marking. For local realignment and analysis-ready read, GATK and Illumina/HiSeq2000 can be applied.

In variant discovery and genotyping, GATK can be used for SNPs and indels (variant calling). For structural variation (SV), CoNIFER can be used.

In last phase, integrative analysis, GATK and wANNOVAR are mainly utilized.

4. Exome-sequencing data analysis

Given the following pedigree,



4.1 What type of inheritance pattern would be the most likely explanation for the above pedigree? (5 points)

4.2 Based on the answer to 4.1, what is the probability that the two new-born males marked with ?1 and ?2 will be affected by the condition? (5 points)

4.3 Please design a study to identify the pathogenic variant for the condition and describe the criteria you used to define the pathogenic variant? (15 points)

Answer:

(4.1)

Autosomal recessive.

(4.2)

For number 1, we can infer that his parents are all carriers (Aa), and he has a probability of 25% to be affected (aa).

For number 2, we can infer that his father is a carrier (Aa) and mother is affected (aa), he has a probability of 50% to be affected (aa).

(4.3)

Raw sequence data examination: After collecting and loading the data from samples, use QC to check sequencing quality for each member, ensure the samples are not contaminated.

GATK pipeline and variant calling: then, we select a set of high-confident variants for sample quality control. Give certain parameters to Depth (DP), Mapping Quality (MQ), Genotype Quality (GP) and check how many variants are left.

Sex examination: also known as sex check. The process use X chromosome data to determine sex and flags individuals for whom the report sex in the PED file does not match the estimated sex. We will have a F value; if $F>0.8$, male; if $F<0.2$, label it female.

Familial relationship examination: by checking the P(IBD) to determine the relationship among family members. By the way, we can somehow determine the type of the chromosomal genetic disease.

Check mutations in the known genes and casual variant prioritization: after preparing raw data and GATK pipeline, we can have the QCed variants. Check the allele frequency, if the value is lower than 0.0001, we can say it is 100% penetrance. Autosomal dominant and LoF (Loss of function) variant should also be considered independently.

Summary: after validating the whole process, give out a clinical report and conduct therapy according to the report.

5. Genome-wide association study

Given the genotype frequency of SNP1 (A/a) and SNP2 (B/b) in a population as shown below:

SNP1 (A/a) \ SNP2 (B/b)	BB	Bb	bb
AA	23	51	34
Aa	37	48	35
aa	13	19	11

5.1 Please estimate the frequencies of the 4 haplotypes: AB, Ab, aB, ab, using EM algorithm (5 iterations). (10 points)

5.2 Please describe whether the two SNPs are in linkage disequilibrium and why? (10 points)

Answer:

The code is available at https://github.com/HULinfengHideki/BIM3007_HW1.

(5.1)

In [38]:
```
# 5th
E_AB=E_AB_root+E_AB_ab
E_Ab=E_Ab_root+E_Ab_aB
E_aB=E_aB_root+E_Ab_aB
E_ab=E_ab_root+E_AB_ab

f_AB = float(('%.4f ' % float(E_AB/total)))
f_Ab = float(('%.4f ' % float(E_Ab/total)))
f_aB = float(('%.4f ' % float(E_aB/total)))
f_ab = float(('%.4f ' % float(E_ab/total)))
P_AB_ab = float(('%.4f' %  float(f_AB*f_ab / (f_AB*f_ab+f_Ab*f_aB))))
P_Ab_aB = float(1-P_AB_ab)
E_AB_ab = float(('%.4f' % float(a_mid*P_AB_ab)))
E_Ab_aB = float(a_mid - E_AB_ab)

print('fifth iteration done')
print('-------------------------------------------')
# print('P_AB_ab in the fifth trail is:'+str(P_AB_ab))
# print('P_Ab_aB in the fifth trail is:'+str(P_Ab_aB))
# print('E_AB_ab in the fifth trail is:'+str(E_AB_ab))
# print('E_Ab_aB in the fifth trail is:'+str(E_Ab_aB))
print('f_AB in the fifth EM is:'+str(f_AB))
print('f_Ab in the fifth EM is:'+str(f_Ab))
print('f_aB in the fifth EM is:'+str(f_aB))
print('f_ab in the fifth EM is:'+str(f_ab))
```

```
fifth iteration done
-------------------------------------------
f_AB in the fifth EM is:0.2853
f_Ab in the fifth EM is:0.3346
f_aB in the fifth EM is:0.2018
f_ab in the fifth EM is:0.1783
```

From the screenshot of jupyter notebook, frequency of 4 haplotypes AB, Ab, aB, and ab are 0.2853, 0.3346, 0.2018, and 0.1783, respectively.

(5.2)

We can have:

	A	B	C	D	E	F	G
1	AB	x11	0.2853		A	p1	0.6199
2	Ab	x12	0.3346		a	p2	0.3801
3	aB	x21	0.2018		B	q1	0.4871
4	ab	x22	0.1783		b	q2	0.5129
5							
6	p1*q1	0.30195329					
7	D	-0.01665329					
8	D^2	0.000277332					
9	p1*p2*q1*q2	0.058866787					
10	R^2	0.004711181					

Since D != 0, it can be, somehow, considered as in linkage disequilibrium (LD).

However, since R^2 < 0.2, according to level of LD measured by R^2, they are likely independent.