New Gene Mutation Detection System for Sanger Sequencing Data

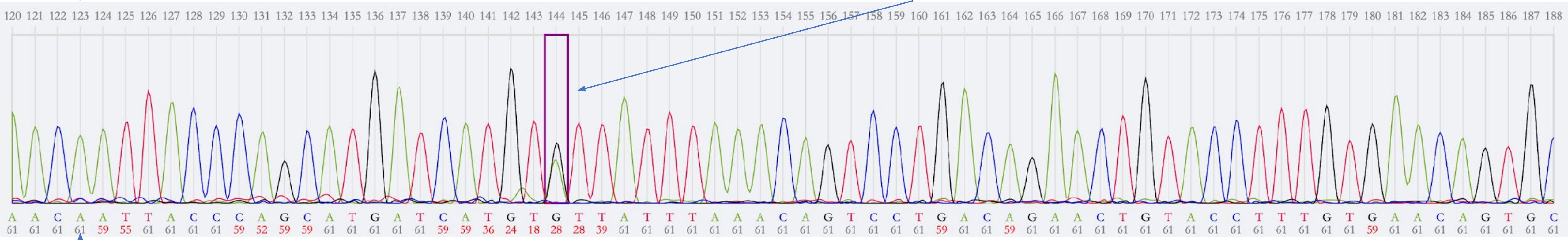
Chuanyang Jin
Yuting Wang
Tenghao Li

Base Fluorescence Signal Curves

The base fluorescence signals are weak and the sensor is easy to be interfered by electromagnetism. To improve the accuracy, Sanger sequencing equipment sequence each fragment thousands of times to avoid the white noise interference, thus forming a bell curve in accordance with the normal distribution at each normal test sequence site.

Base coordinates in the current sequencing result sequence

Heterozygous mutation

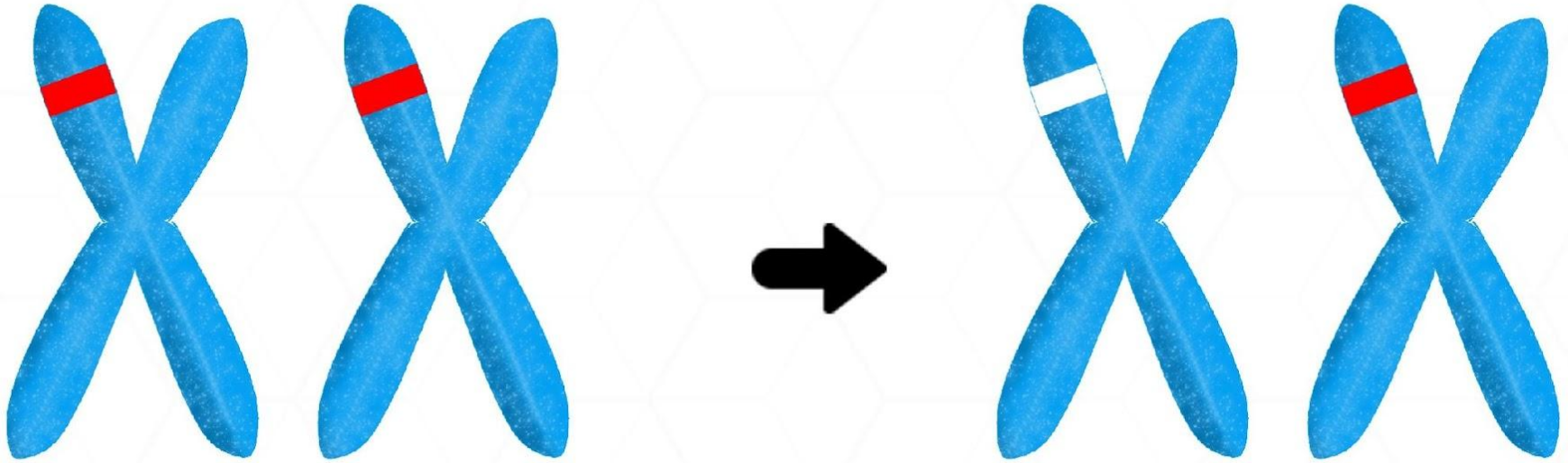


Phred score for base sequencing quality

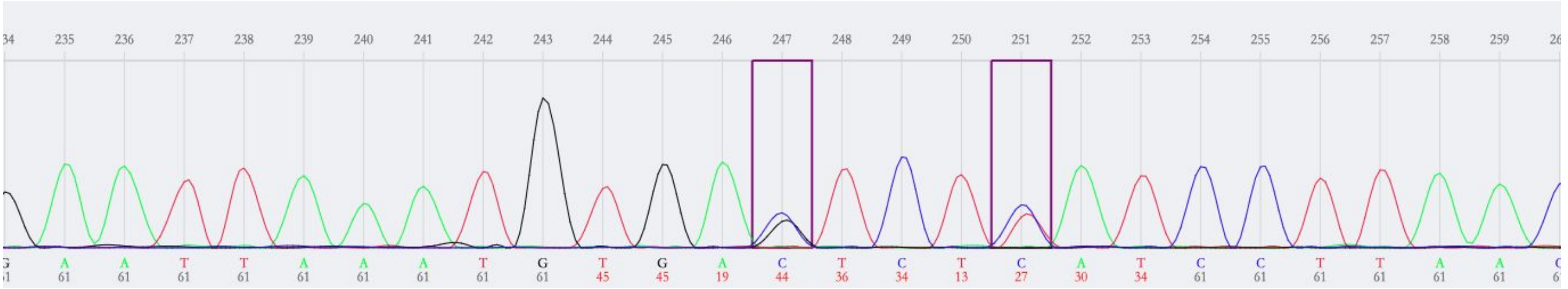
The ATCG at each integer point is recognized according to the value of the largest base fluorescence signal intensity

Heterozygous Mutation

Mutation occurs in only one of the two alleles on a pair of homologous chromosomes.

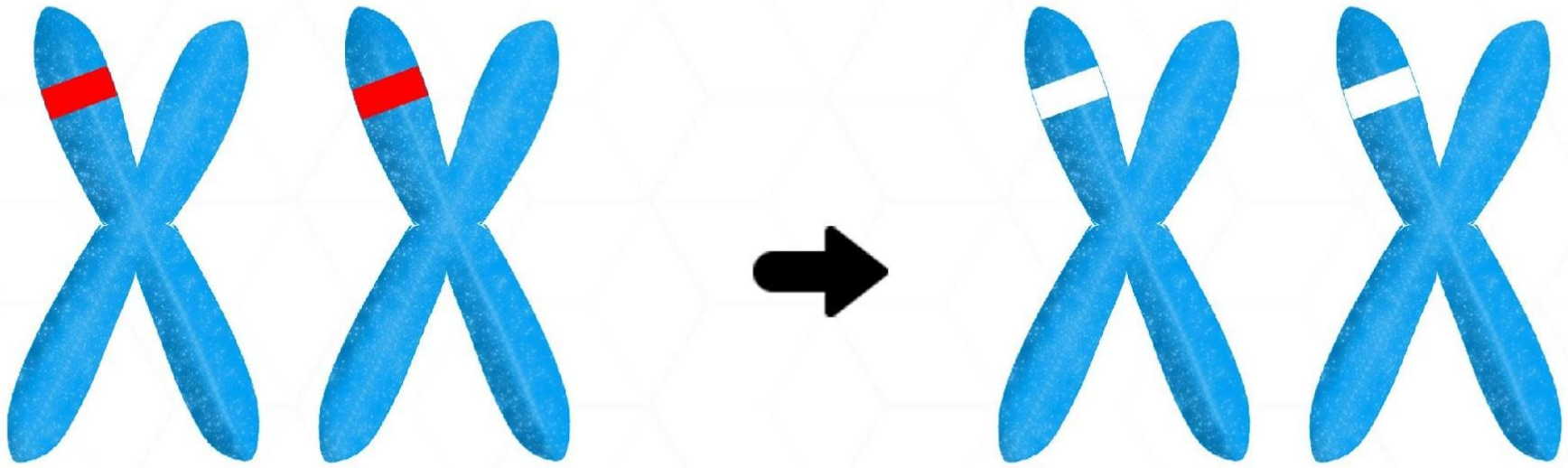


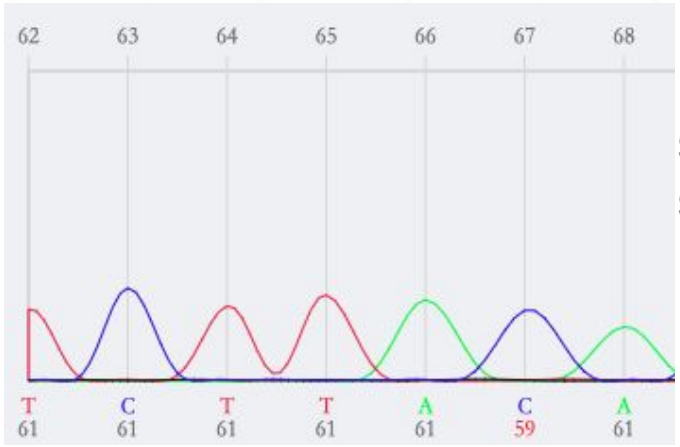
C or G C or T



Homozygous Mutation

Mutation occurs in both alleles on a pair of homologous chromosomes.

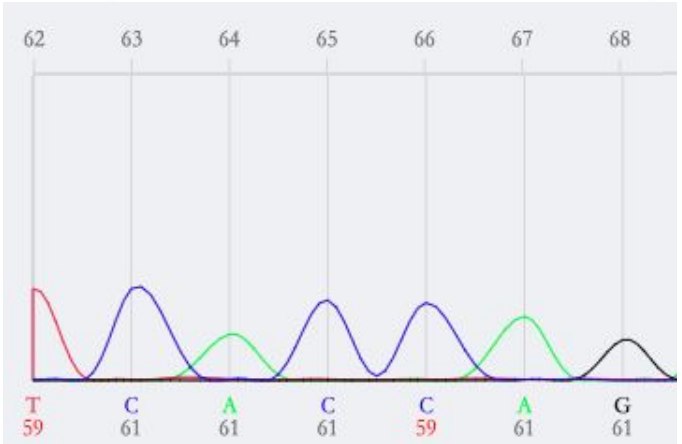




sequencing sequence →

TCTTACA
||
TCACCAG

← reference sequence



Single Nucleotide Polymorphism (SNP)

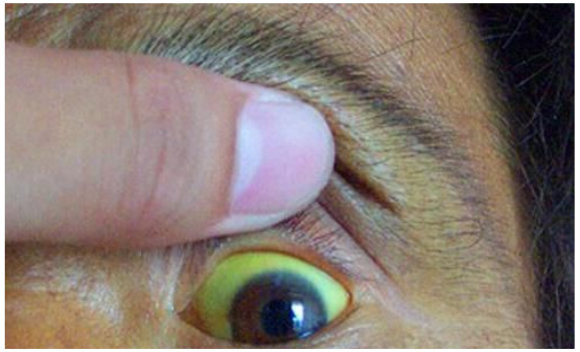
Dr. Yang Qi, biological science and medical engineering, Jinling Hospital

SNP = Heterozygous Mutation + Homozygous Mutation

It is the most common form of heritable variation. It accounts for more than 90% of all known polymorphisms. SNP exists widely in the human genome, with an average of one out of every 500-1000 base pairs. It is estimated that the total number of SNP can reach 3 million or more.

SNP affects biological properties and may cause diseases

Hearing loss (HL)

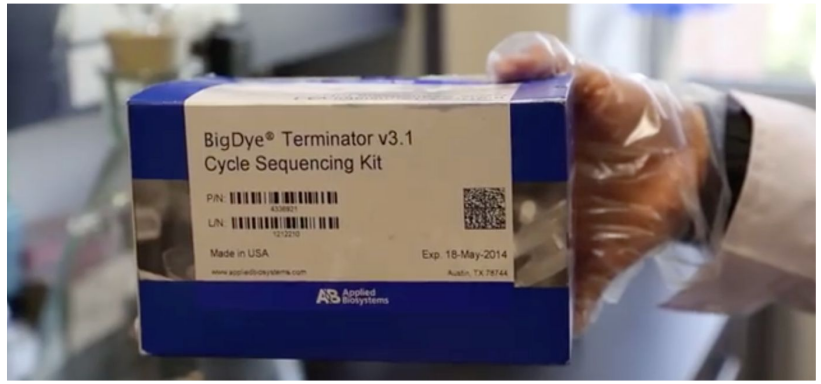


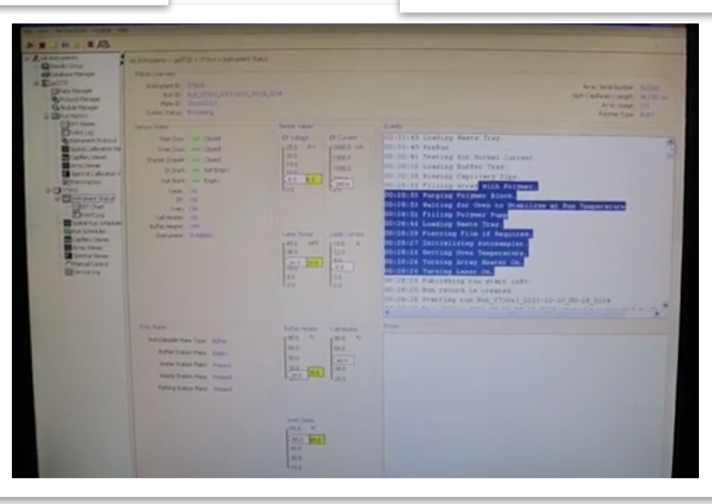
Hypertriglyceridemia (HTG)
Acute Pancreatitis (AP)

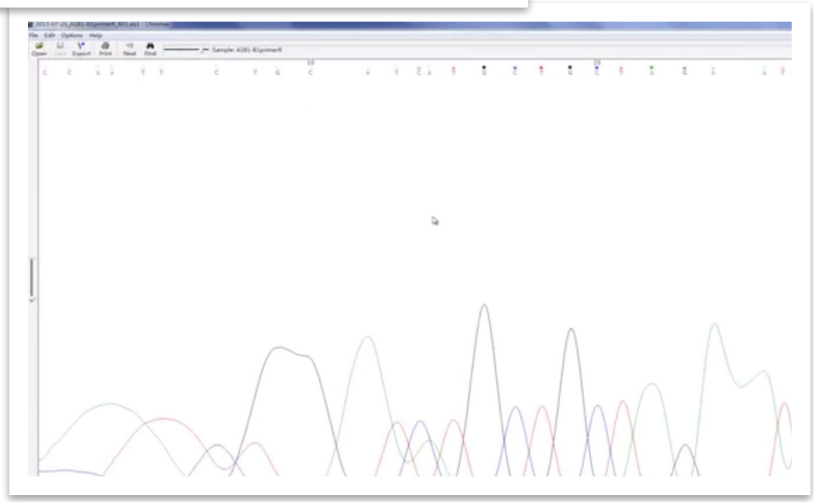
Sanger Sequencing Workflow

Dr. Yang Qi, biological science and medical engineering, Jinling Hospital

Blood Sampling - > DNA Extraction - > PCR Amplification Reaction - > Put into Sequencer - > Automatic Sequencing - > Analysis of Sequencing Results





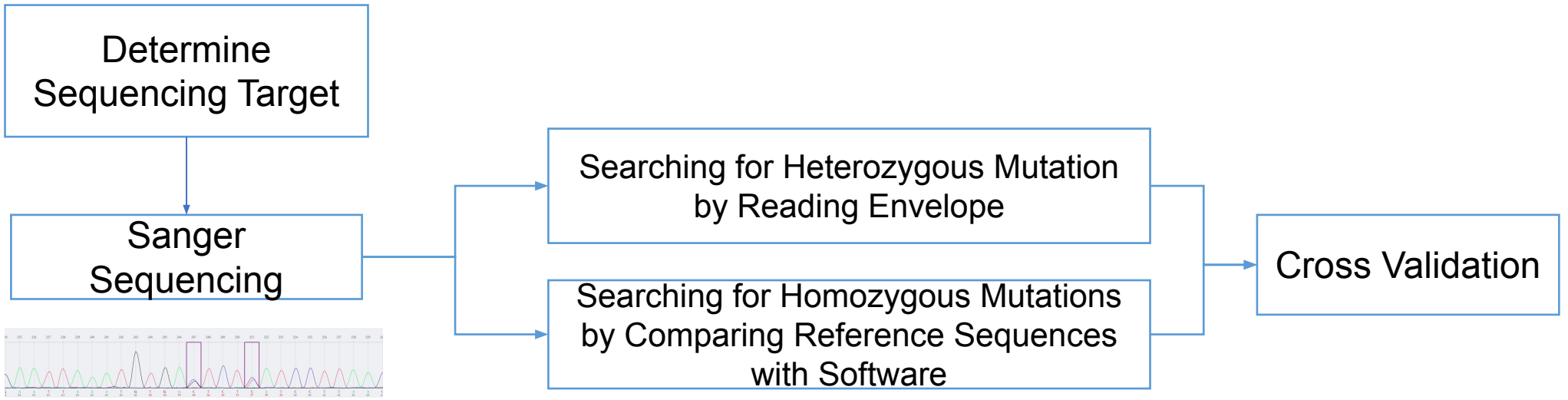


The experiment will be completed in about 2 weeks. Generally speaking, each subject will generate 50-100 sequencing files.

Sanger Sequencing - Clinical Gold Standard

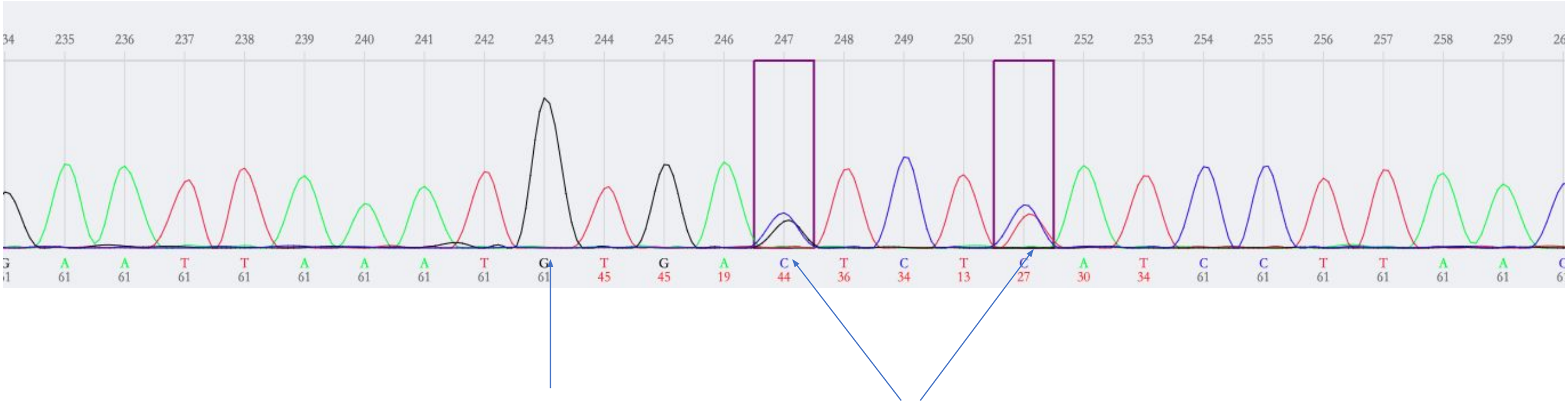
Dr. Yang Qi, Jinling Hospital

Sequencing technology category	Maximum flux of single sequencing	Output format	Sequencing accuracy
First generation Sanger sequencing	Short 200-1000 base sequences	Intensity curve of base fluorescence signal	Gold standard



Search for Ideas

PubMed ?



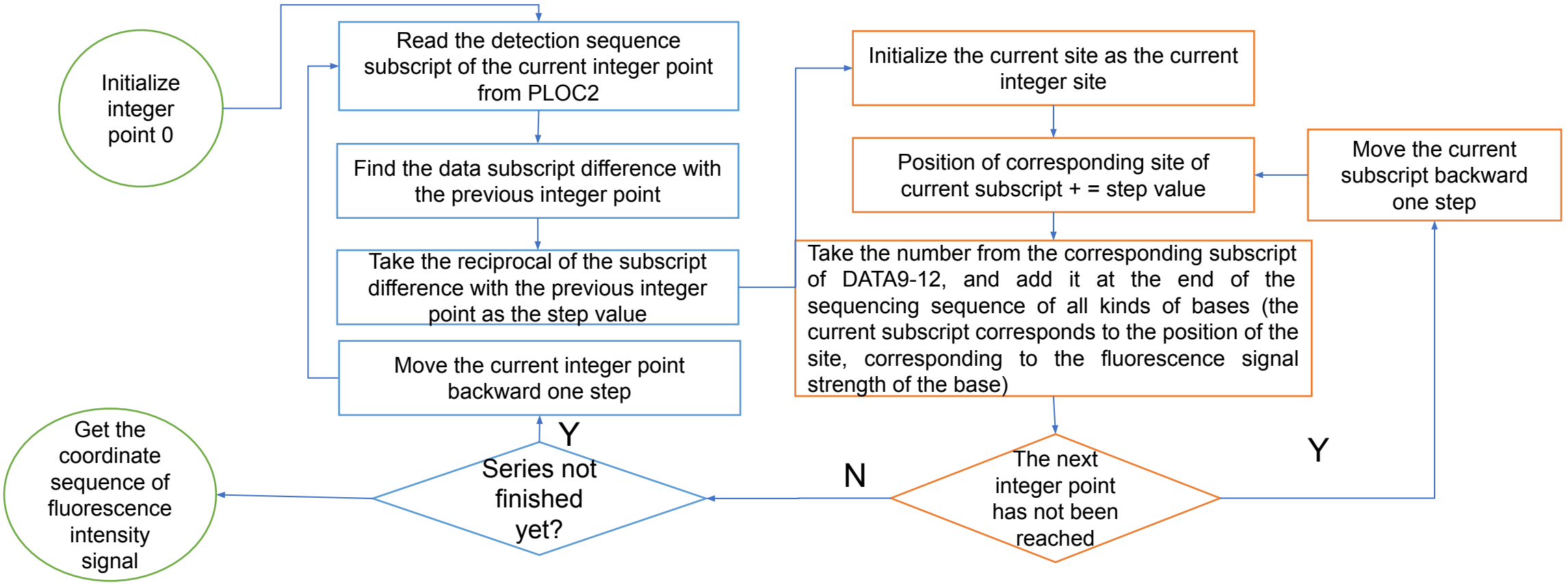
Visual perception:

Unimodal

Bimodal (Diploid)

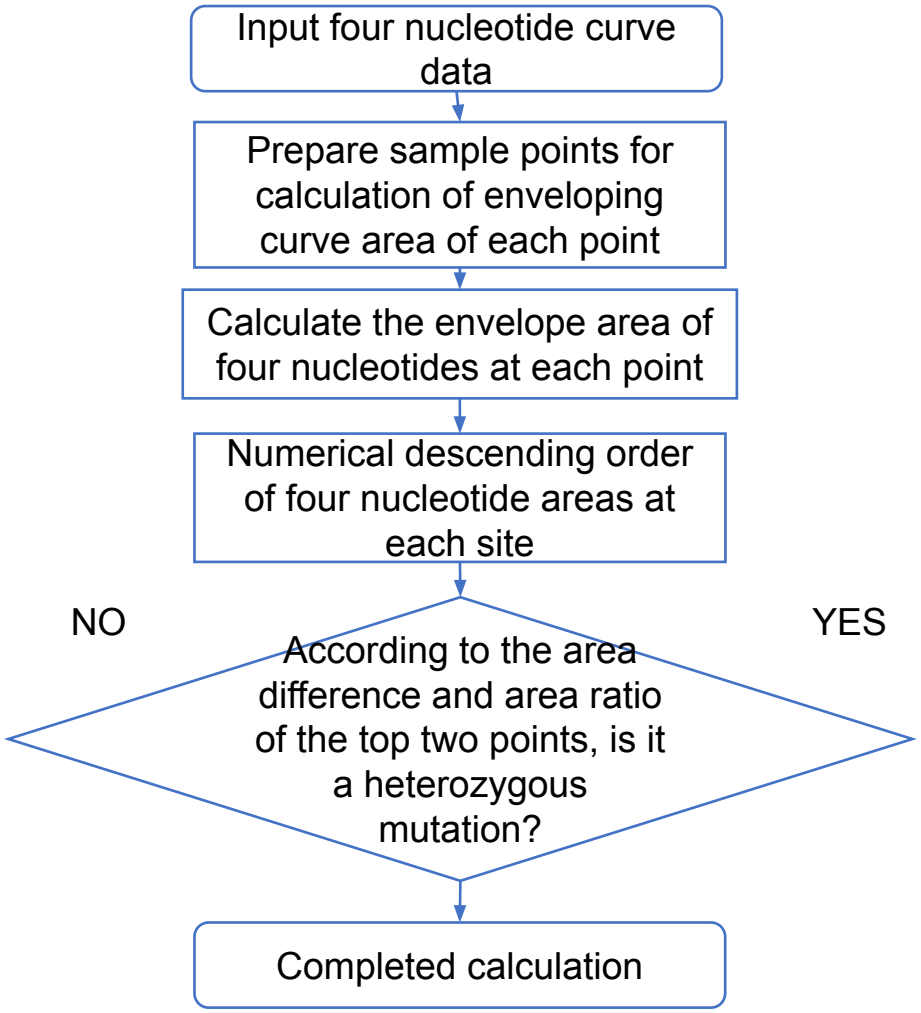
Coordinate Sequence of Fluorescence Intensity Signal Obtained from ABIF Format Sequencing Data

Name	Number	ABIF Type	Description
DATA	9-12	short[]	Short Array holding analyzed color data
PLOC	2	short[]	Array of peak locations as called by Baseceller

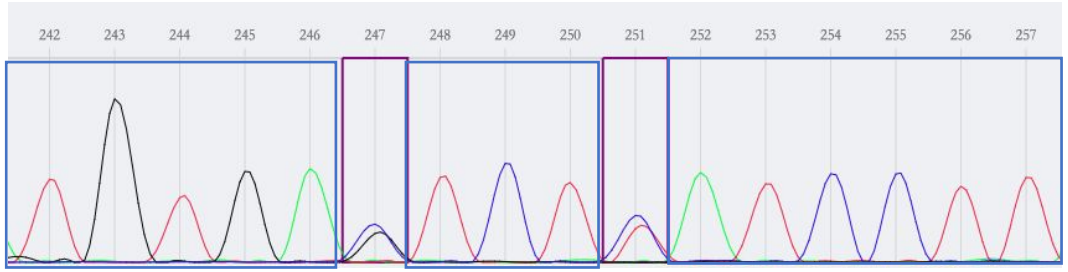


Heterozygous Mutation Detection by Computational Geometry

1. General process of heterozygous mutation identification



2. Discussion on the difference of area difference judgment

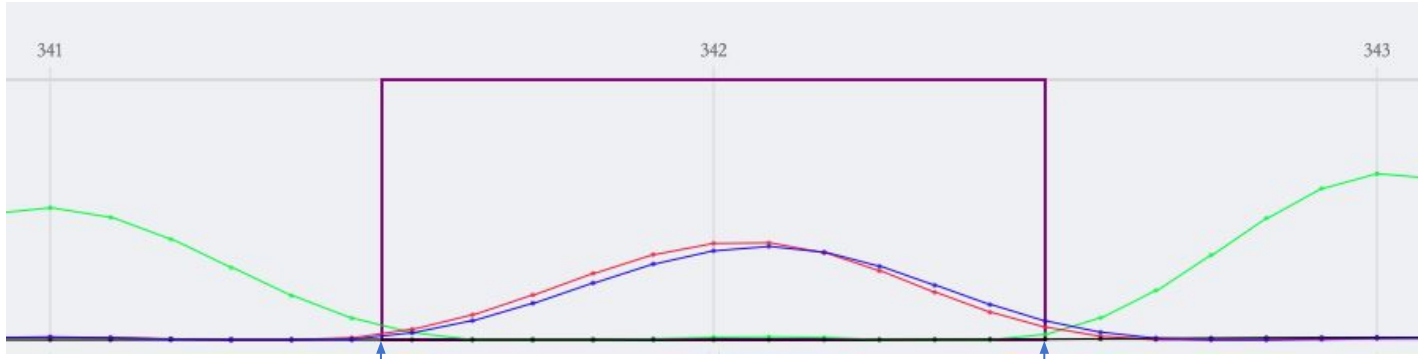


At each site, there are four envelope data of nucleotide fluorescence intensity signal. There are two types of situations:

A. For the non heterozygous mutation site, when a certain nucleotide is sequenced, only one nucleotide fluorescence intensity signal appears bell curve, and the other is almost 0, so its area difference value is large, and the area ratio is close to 1.

B. when the heterozygous mutation site is sequenced to this site, there will be two (or more, for n-ploid) nucleotide fluorescence intensity signals with bell shaped curves, which are similar in shape, so the area difference between the top two is smaller and the area is larger.

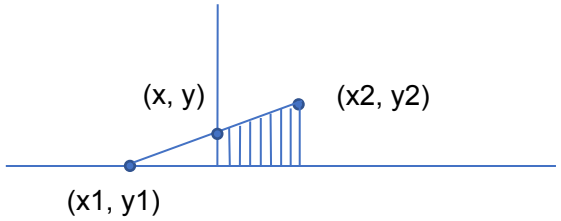
Prepare Sample Points for Calculation of Enveloping Curve Area of Each Site



There is no point on the cutting line here

There is a point on the cutting line here

Virtual points must be supplemented, otherwise the area of shadow area will be lost in area calculation



1. Use the linear slope equation to solve the vertical coordinate y on the cutting line.

If the former point is (x1, Y1) and the latter point is (X2, Y2), and the abscissa of the cutting point is x, then there are:

$$k = \frac{y2 - y1}{x2 - x1} \qquad y = y1 + (x - x1) * k$$

2. According to the four nucleotide fluorescence envelope signal curve samples at 0.5 position on each base left and right, cut them into the array.

Calculate the Envelope Area of Four Nucleotides at Each Site

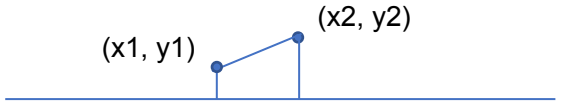
At each site, calculate the area of four kinds of nucleotide fluorescence envelope signal curve sample points respectively:

$$area = \sum_{i=1}^{n-1} t(p_i, p_{i+1})$$

Where n is the number of sample points at this point, $p(X_{p,}y_p)$ is the sample point, and t is the trapezoid area.

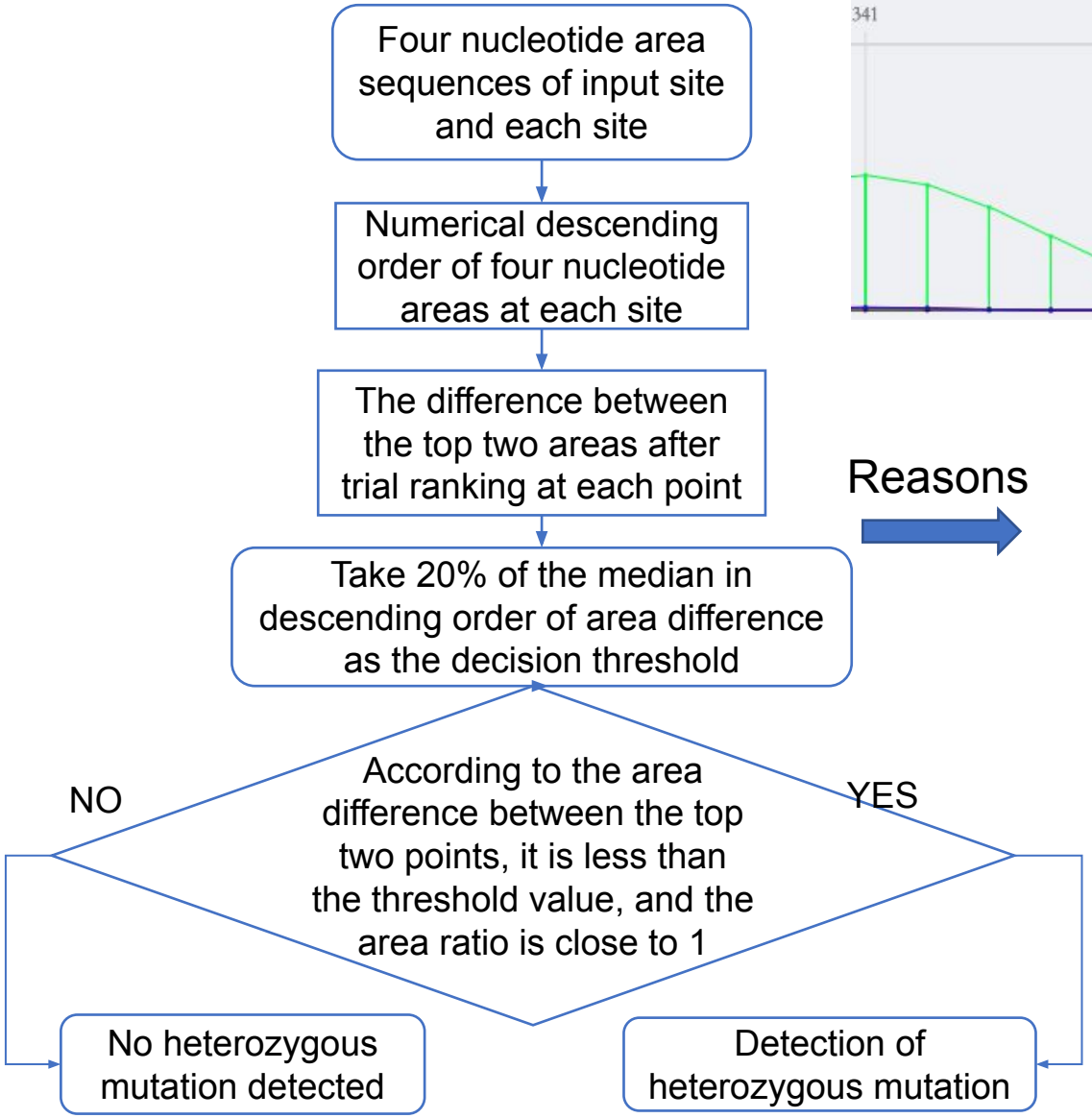
Calculation formula:

$$t = (y_{p+1} + y_p) * (x_{p+1} - x_p)/2$$

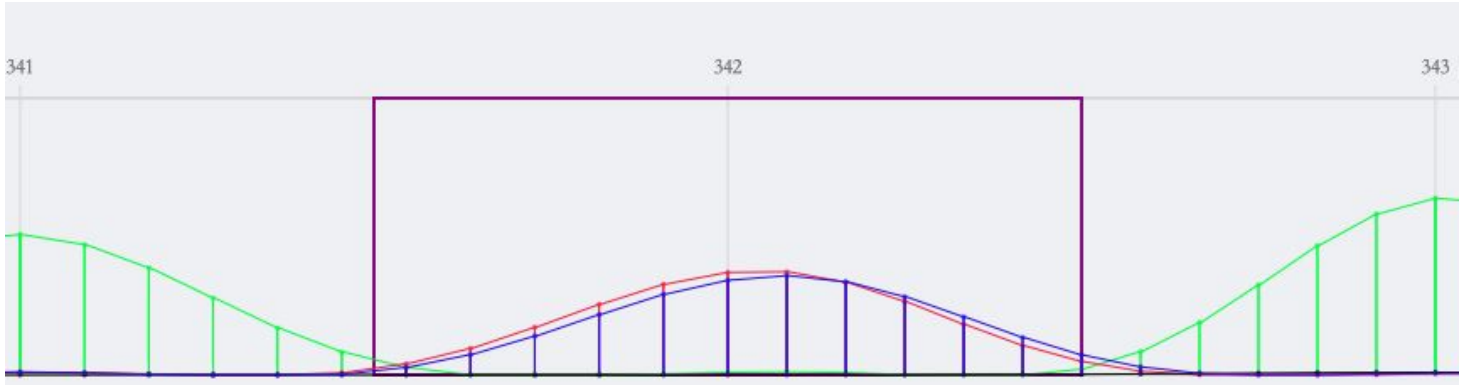


The corresponding envelope curve areas of four nucleotides at each point were obtained: areaA, areaT, areaG, areaC

Determine Whether it is a Heterozygous Mutation



Reasons



1. SNP exists widely in human genome, with an average of 1 in every 500-1000 base pairs;
2. The longest sequence of Sanger sequencing was 500-1000 bases;
3. SNP can be divided into heterozygous mutation and homozygous mutation.

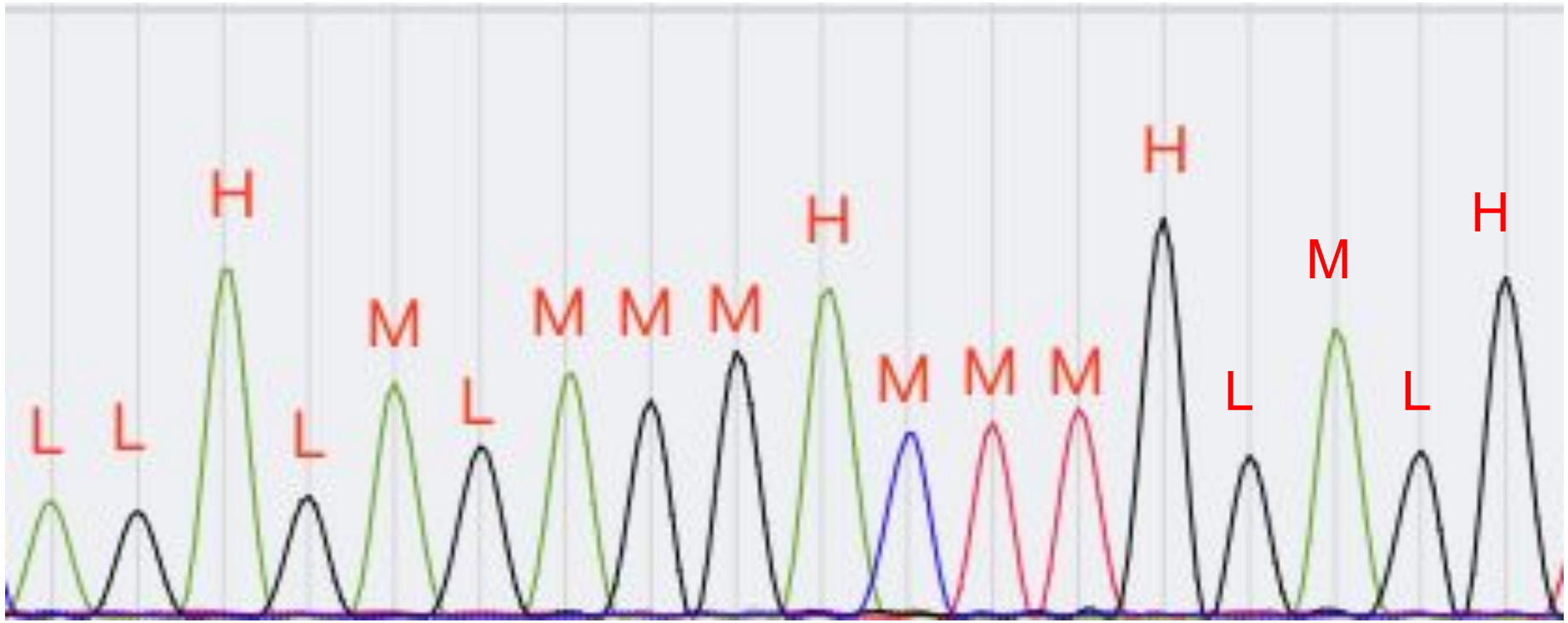
Therefore:

- In Sanger sequencing, the number of heterozygous mutations in each normal sequence should generally be less than 10, with a maximum probability of 1-2%

For those in the effective sequencing region:

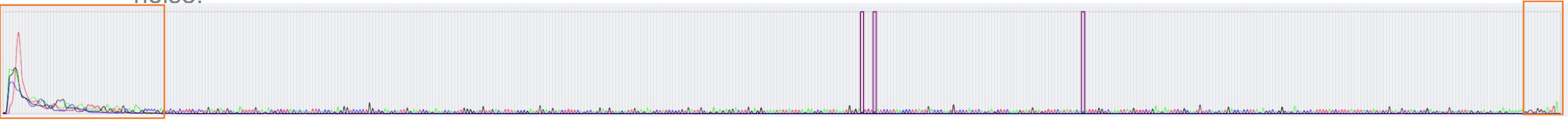
- The area difference of non heterozygous mutation points is mostly large, and the probability of occurrence is 98-99%;
- The area difference of heterozygous mutation point is almost 0, and the probability of occurrence is 1 ~ 2%.
- The median of the area difference descending sorting array should be the non heterozygous mutation point area is the larger value, 20% of the larger value is significantly greater than 0 and significantly less than the larger value, which is suitable for use as a threshold.

Use K-Means Clustering Algorithm to Increase the Detection Accuracy of Judgment Threshold



Head and Tail Noise

PCR amplification and equipment interference will introduce noise, which is characterized by disordered waveform overlap and high peak value.
It is generally distributed at the beginning and the end of the sequence. In this region, it is impossible to distinguish heterozygous mutation by area method.
When Sanger sequencing test is designed, the nucleic acid sequence that affects the coding will be placed in the middle of the sequencing target as much as possible, and redundancy will be increased to avoid being affected by the first noise.

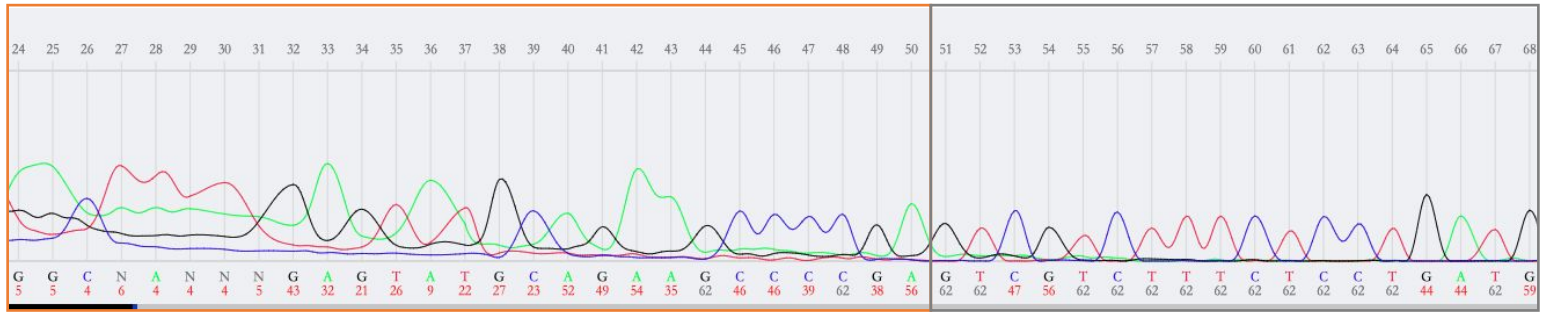


Phred score meaning

Phred quality score	Possibility of base detection error
10	1 for every 10
20	1 for every 100
30	1 for every 1000
40	1 for every 10000
50	1 for every 100000
60	1 for every 1000000

Noise area

Effective sequencing area

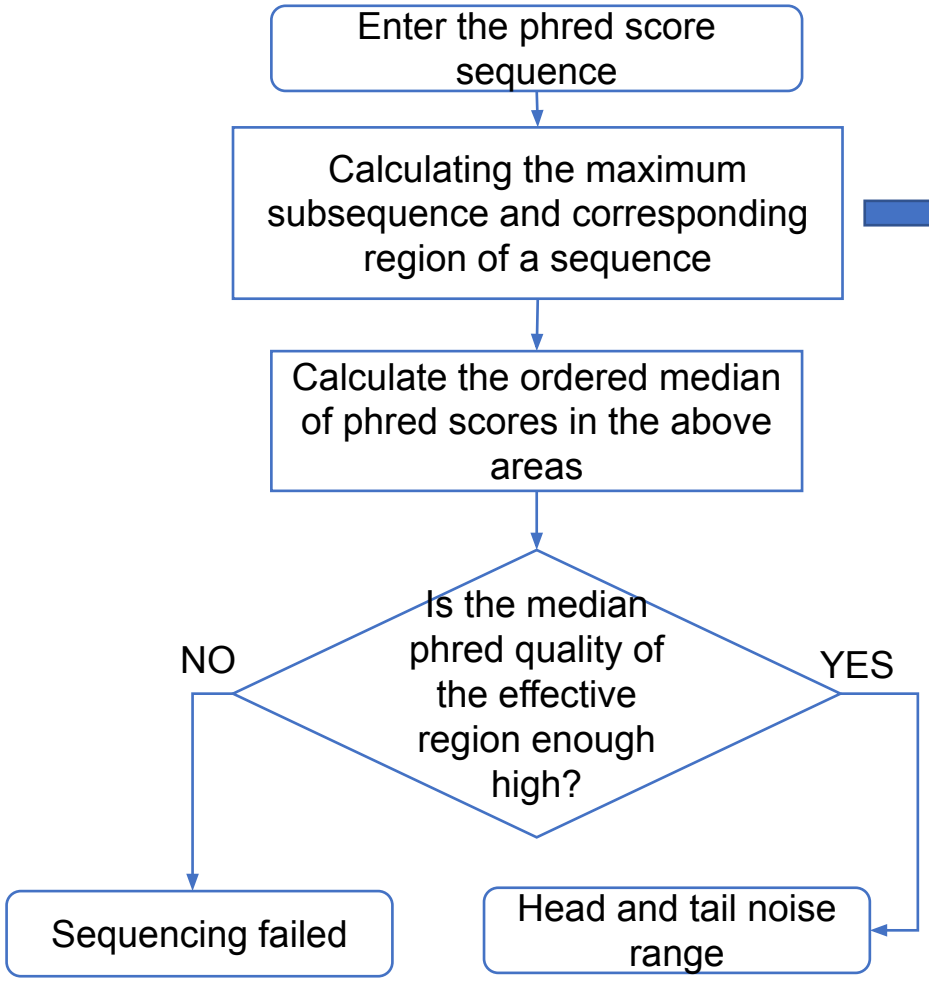


Lower Phred Score Average <30

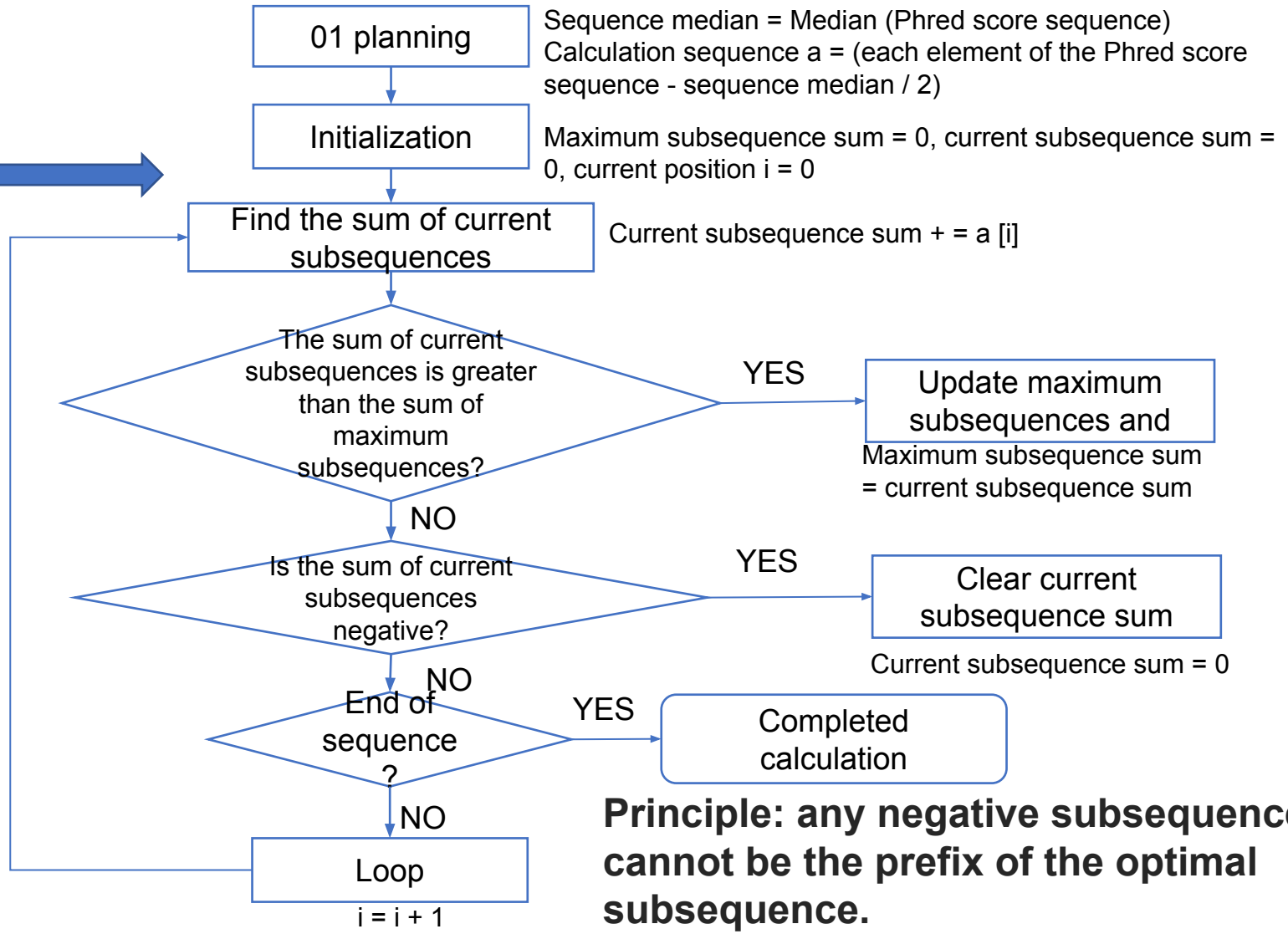
Higher Phred Score Average >=30

Remove Head and Tail Noise: Modified Mott Trimming Algorithms

1. Overall process of noise removal

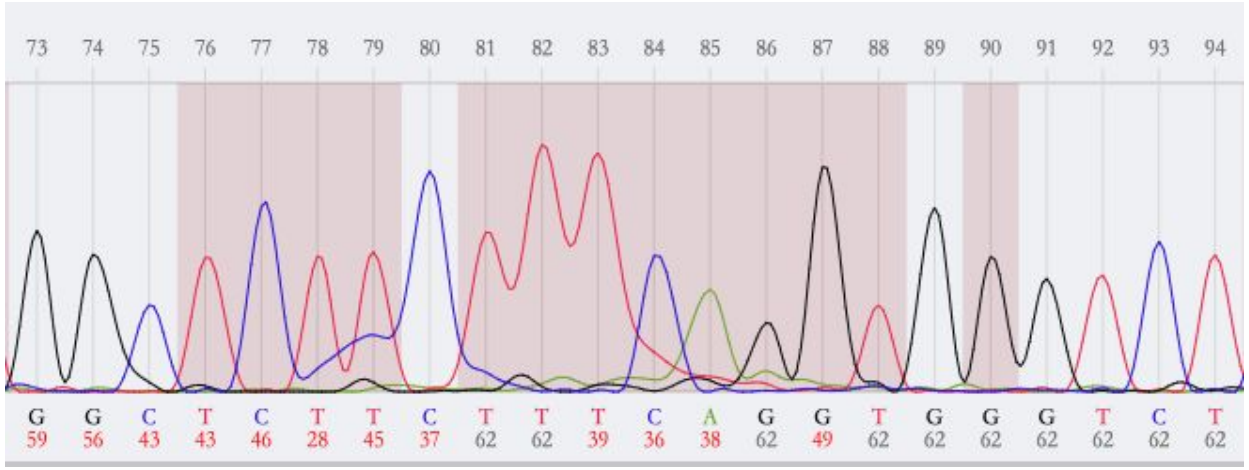


2. Maximum subsequence algorithm with O (n) time complexity



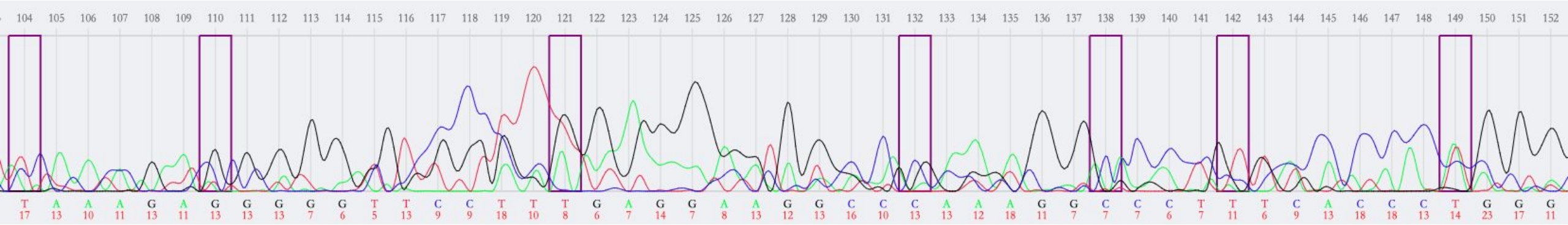
Principle: any negative subsequence cannot be the prefix of the optimal subsequence.

Middle Noise and Sequencing Failure



The noise points in the non head tail region are not removed as head tail noise. Thus, it will affect the subsequent detection steps of heterozygous mutation and homozygous mutation.

If there is noise in all areas, the sequencing fails. At this time, the whole data should be discarded. Otherwise, a large number of pseudopositive heterozygous mutations will be identified.

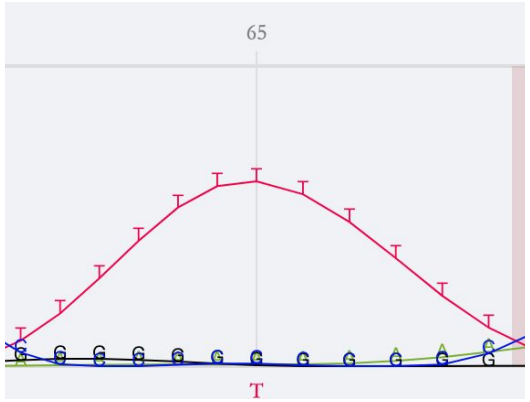


Noise Region Detection based on Convolutional Neural Network

1d version of Le-Net5

The distribution density of data points is not uniform

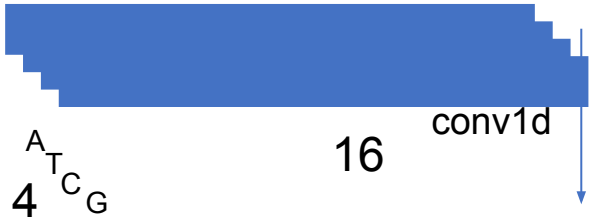
Signal curve of fluorescence intensity of each base



16 points/bp

sampling

Input Shape 4x16

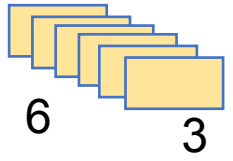


6 filters
kernel size 2x1
strides: 1

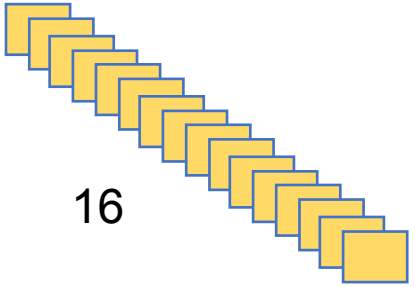
Feature Map 6x3

16 filters
kernel size 5x1
strides: 1

conv1d

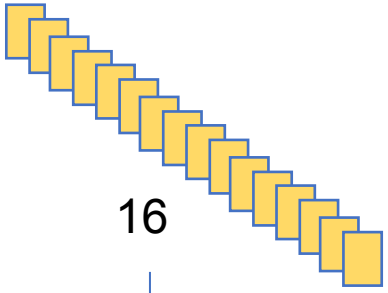


Feature Map 2x16



pool size 2x1
strides 2x1

maxPooling1d

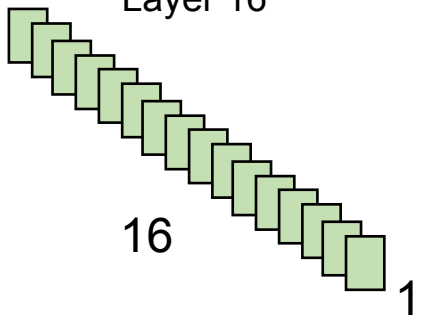


Feature Map 1x16

full connection

dense

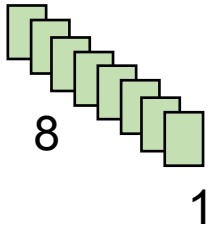
Layer 16



full connection

dense

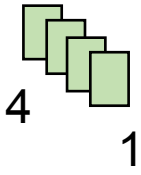
Layer 8



full connection

dense

Layer 4



full connection

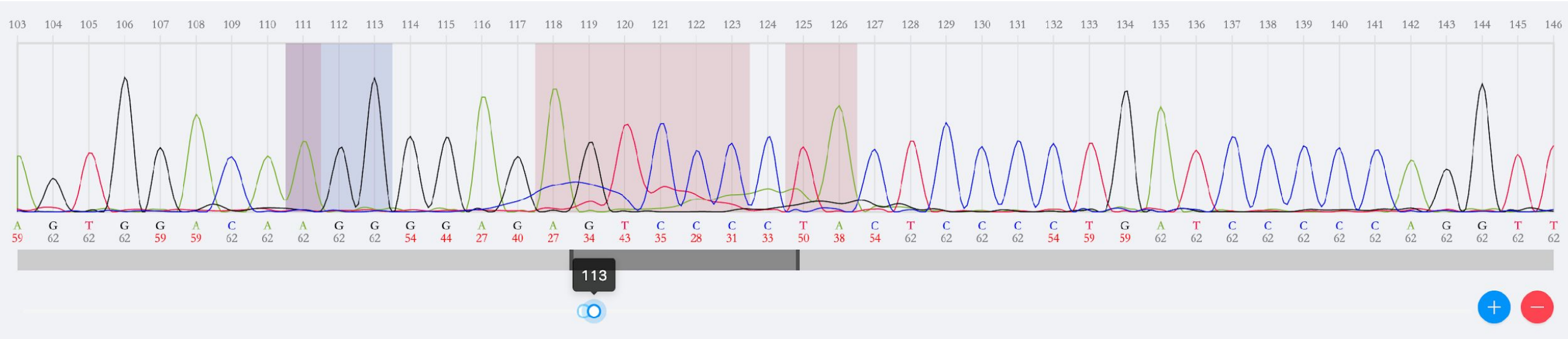
dense

Output 1

Noise Area Detection Training

Operation area

Noise area



1 selected operation range

2. Mark + noise-data

Detect the head and tail noise regions by maximum subsequence algorithm

Use trained model to identify noise area

Manually add/delete noise regions

Export training data

Train the Neural Network

```
Epoch 2798 / 20000
27604ms 1087us/step - acc=0.958 loss=0.0347
Epoch 2799 / 20000
27641ms 1089us/step - acc=0.957 loss=0.0352
Epoch 2800 / 20000
25308ms 997us/step - acc=0.957 loss=0.0353
Epoch 2801 / 20000
25912ms 1021us/step - acc=0.959 loss=0.0344
Epoch 2802 / 20000
```

98.3% accuracy can be achieved after 2 days of training with one GTX2070

Protein mutation detection

Blast with subsequence which has removed head and tail noise area

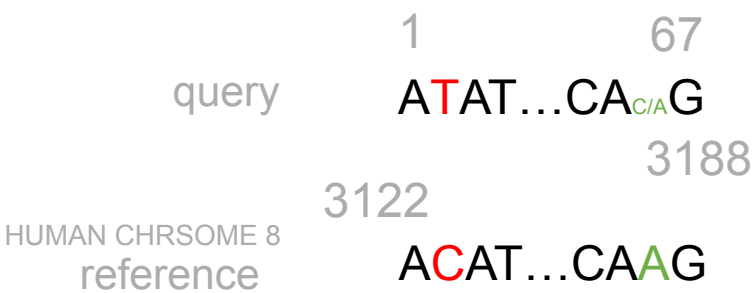
Sequence fragment retrieval and comparison

Feature retrieval

Detection of protein mutation

Blast Query

Blast Response



Mutations

TYPE	Mutation
Homozygous	CHR -8 3123 C->T
Heterozygous	CHR-8 3287 A->C

CDS Database

CHR	Gene	From	To	Exon Location
8	LPL	2000	4000	Join(2000…3126, 3200…3286)

Exon matched mutations

In Exon	Mutation
CHR-8 LPL-1	CHR -8 3123 C->T
Not in Exon Will no effect	CHR-8 3287 A->C

Standard Genetic Code

碱基1	碱基2								碱基3
	T		C		A		G		
T	TTT	(Phe/F) 苯丙氨酸	TCT	(Ser/S) 丝氨酸	TAT	(Tyr/Y) 酪氨酸	TGT	(Cys/C) 半胱氨酸	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) 亮氨酸	TCA		TAA[B]	终止（赭石）	TGA[B]	终止（蛋白石）	A
	TTG		TCG		TAG[B]	终止（琥珀）	TGG	(Trp/W) 色氨酸	G
C	CTT		CCT	(Pro/P) 脯氨酸	CAT	(His/H) 组氨酸	CGT	(Arg/R) 精氨酸	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) 谷氨酰胺	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) 异亮氨酸	ACT	(Thr/T) 苏氨酸	AAT	(Asn/N) 天冬酰胺	AGT	(Ser/S) 丝氨酸	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) 赖氨酸	AGA	(Arg/R) 精氨酸	A
	ATG[A]	(Met/M) 甲硫氨酸	ACG		AAG		AGG		G
G	GTT	(Val/V) 缬氨酸	GCT	(Ala/A) 丙氨酸	GAT	(Asp/D) 天冬氨酸	GGT	(Gly/G) 甘氨酸	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) 谷氨酸	GGA		A
	GTG		GCG		GAG		GGG		G

TAT->Y
TAC->Y

TAT->Y
CAT->H

Display of Detected Mutations

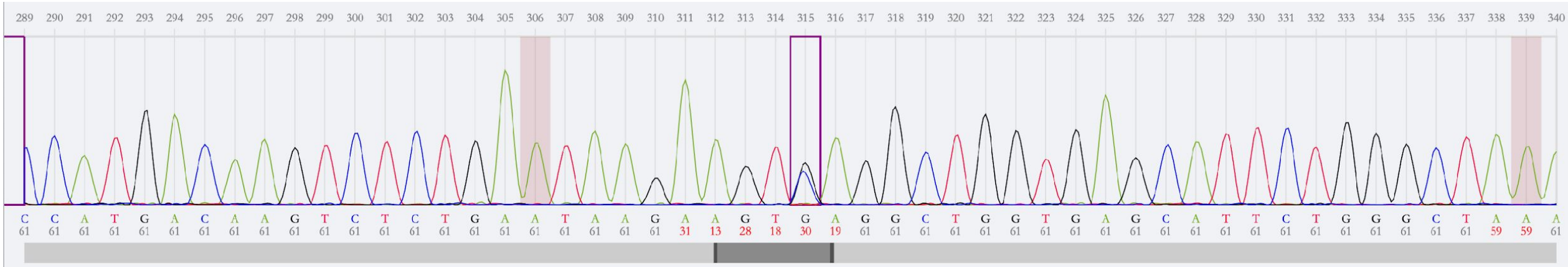
Mutation not on exon

No mutation of protein

TAT->Y
TAC->Y

Protein mutation

TAT->Y
CAT->H



	序号	序列文件	波形质量分	突变点位 核苷酸突变 外显子突变 蛋白质突变
○	1	LPL9-1.ab1	66	
◉	2	LPL9-2.ab1	67	纯合 30(A->G) 纯合 127(T->G) 杂合 CDS位置1421(C->G) LPL 蛋白质突变 474(S->X)

Mutation not on exon

No mutation of protein

Protein mutation

Comparison

Comparison	South China Agricultural University's method	Our method
Noise removal method	Filtering based on wavelet transformation	Modified Mott trimming algorithm Convolutional Neural Network
Heterozygous mutation detection method	Back Propagation Neural Network Parameters: peak distance, height ratio and fluctuation ratio of two peaks	Computational geometry
Test data set	Eucalyptus urophylla 26 sequencing files	Homosapiens HTG-AP 3500 sequencing files
Accuracy rate Accurate number / (accurate number + missed number) * 100%	96.5%	94.59%
Missed judgement rate Number of missed judgments / (accurate number + number of missed judgments) * 100%	3.5%	5.01%
False positive rate Number of misjudgments / (accurate number + number of misjudgments) * 100%	24.6%	16%

SNP基因分析系统

gene.i51cy.com/snp

基因SNP
分析系统

突变识别　MegaBLAST　ABIF详解　智能去噪训练

数据库：人类(9606)

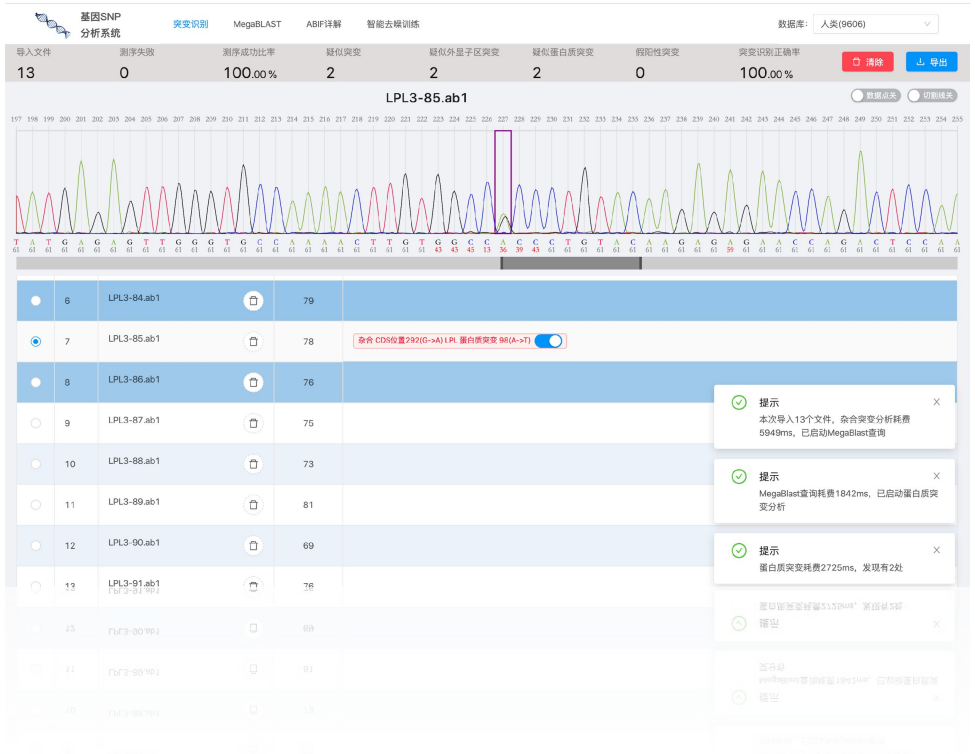
数据导入

点击选择文件 或 将文件拖放至此

支持ab1格式，支持多选

感谢南京外国语学校 王禹听 李滕昊 金川杨 三位同学提供算法支持

Advantages



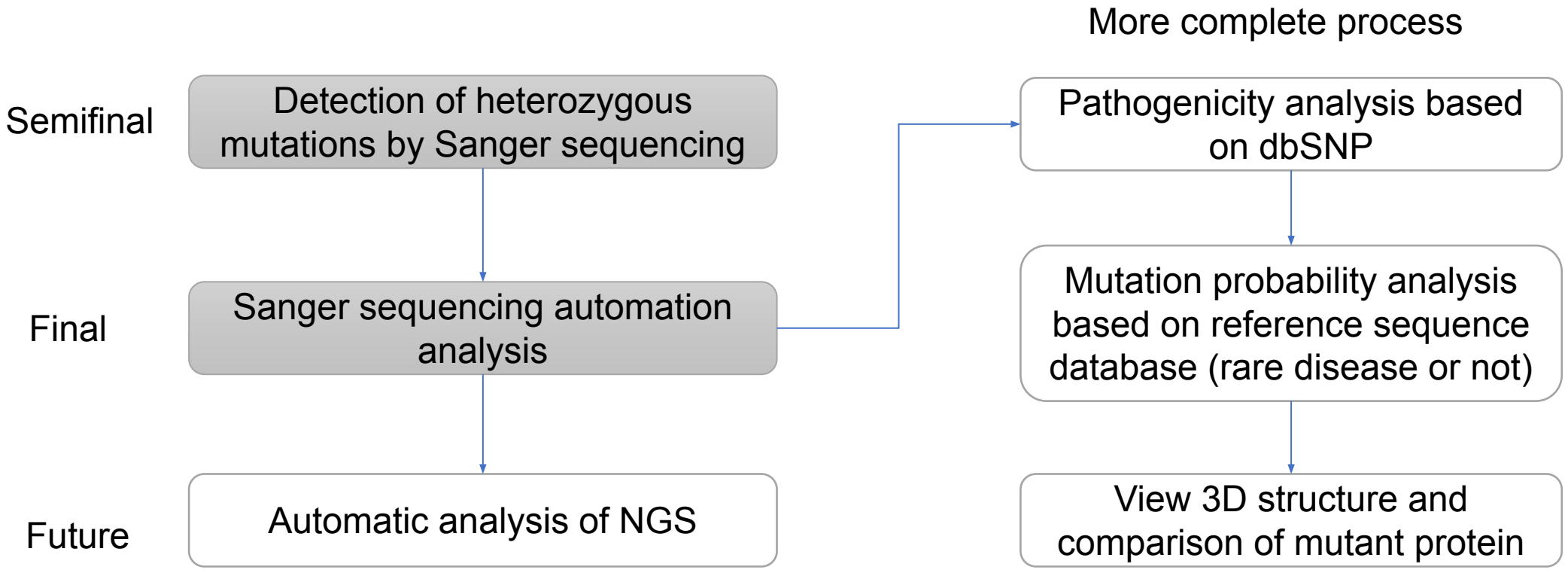
1. 900x faster than human;

2. No setup required;

3.Complete Process of gene mutation analysis;

4.Good user interaction experience.

Expectation

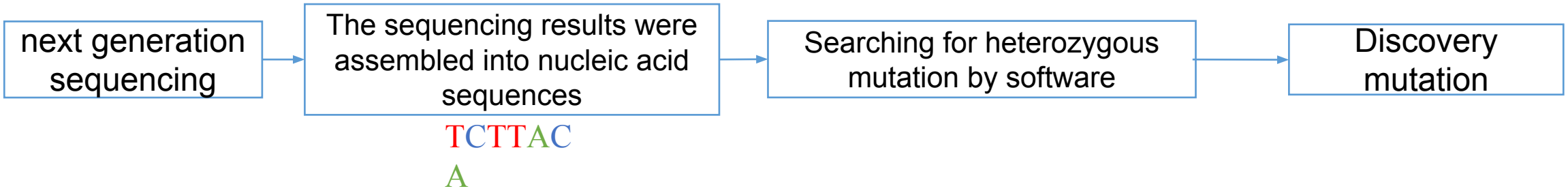


Thank you

Chuanyang Jin
Yuting Wang
Tenghao Li

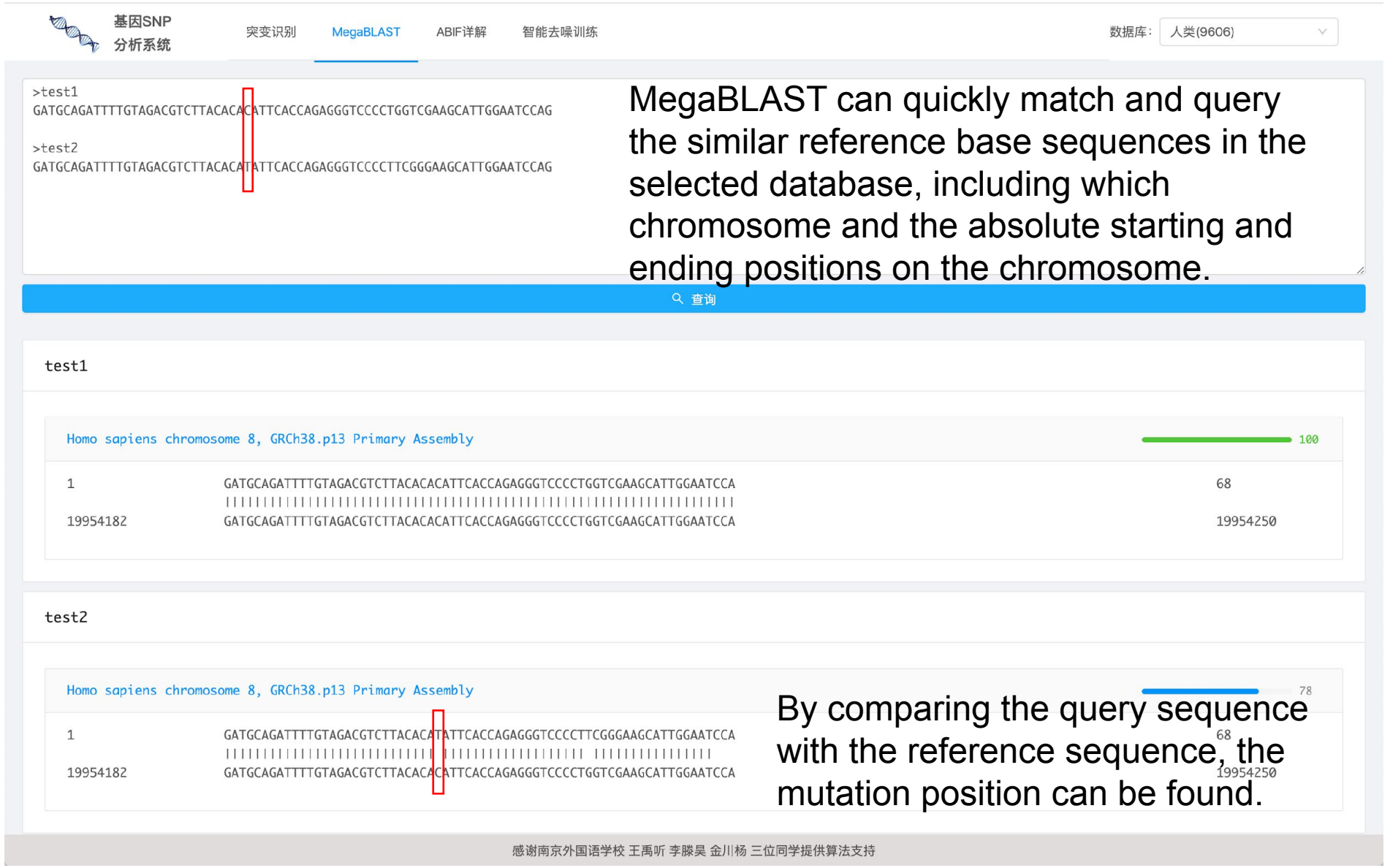
Backup Slide: Next-Generation Sequencing (NGS)

Sequencing technology category	Maximum flux of single sequencing	Output format	Sequencing accuracy
New generation sequencing technology of NGS	All sequences on all chromosomes	Base sequence	False positive



Only 50% (diploid, if n-ploidy is 1 / N * 100%) of heterozygous mutations were detected successfully from nucleotide sequence

Backup Slide: MegaBLAST based alignment



MegaBLAST can quickly match and query the similar reference base sequences in the selected database, including which chromosome and the absolute starting and ending positions on the chromosome.

By comparing the query sequence with the reference sequence, the mutation position can be found.

Backup Slide: Platform application construction

Under the guidance of Zhang Weibo, chief engineer of Nanjing YOUPU IT Co., Ltd