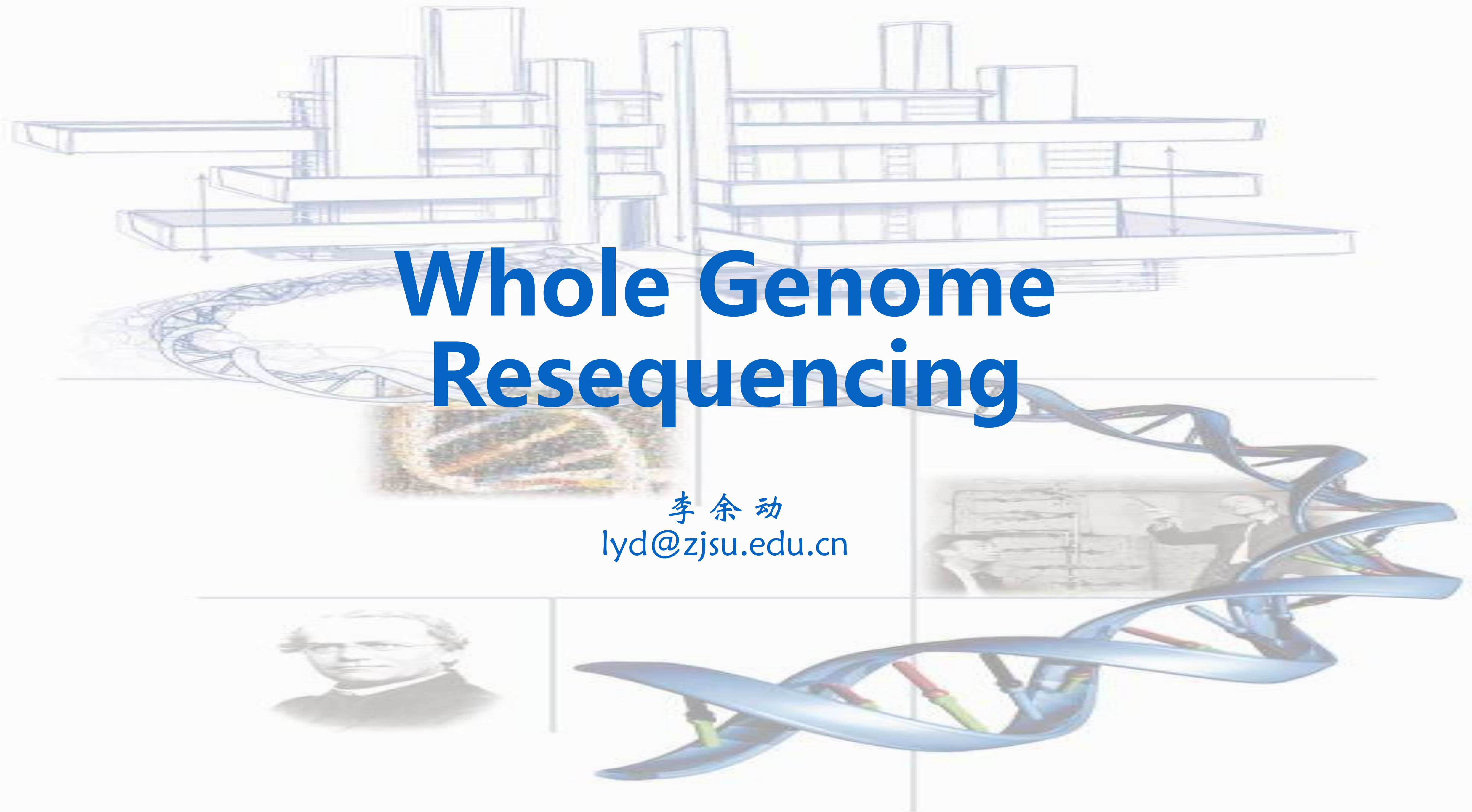


Whole Genome Resequencing

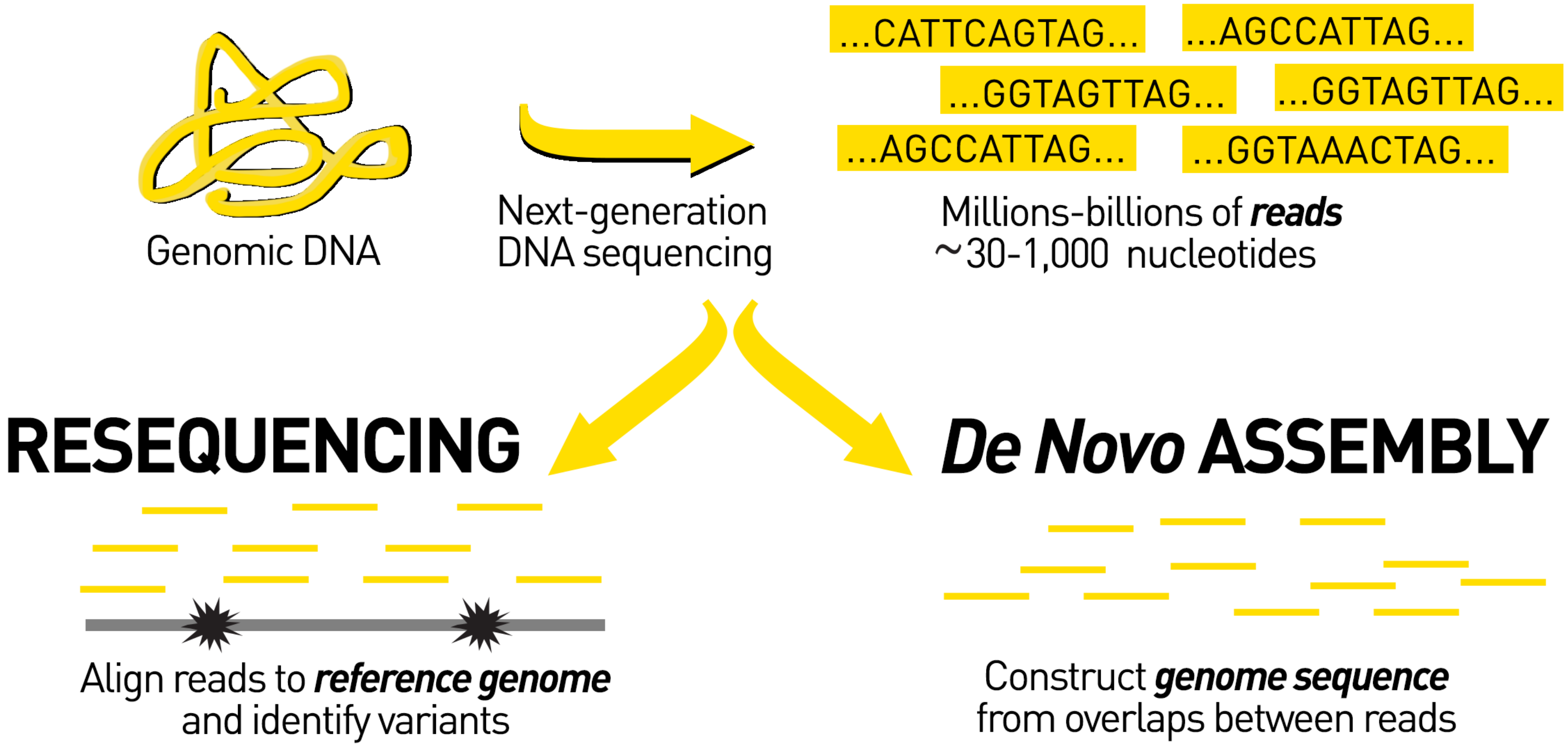
李余动
lyd@zjsu.edu.cn

Topics

- **Genetic Variations**
- **Read Mapping**
    - ✦BWA, Bowtie2…
    - ✦SAM/Samtools
- **Variant calling**
    - ✦Bcftools, GATK…
    - ✦VCF format
- **SARS-CoV-2 resequencing**
    - ✦Data analysis pipeline

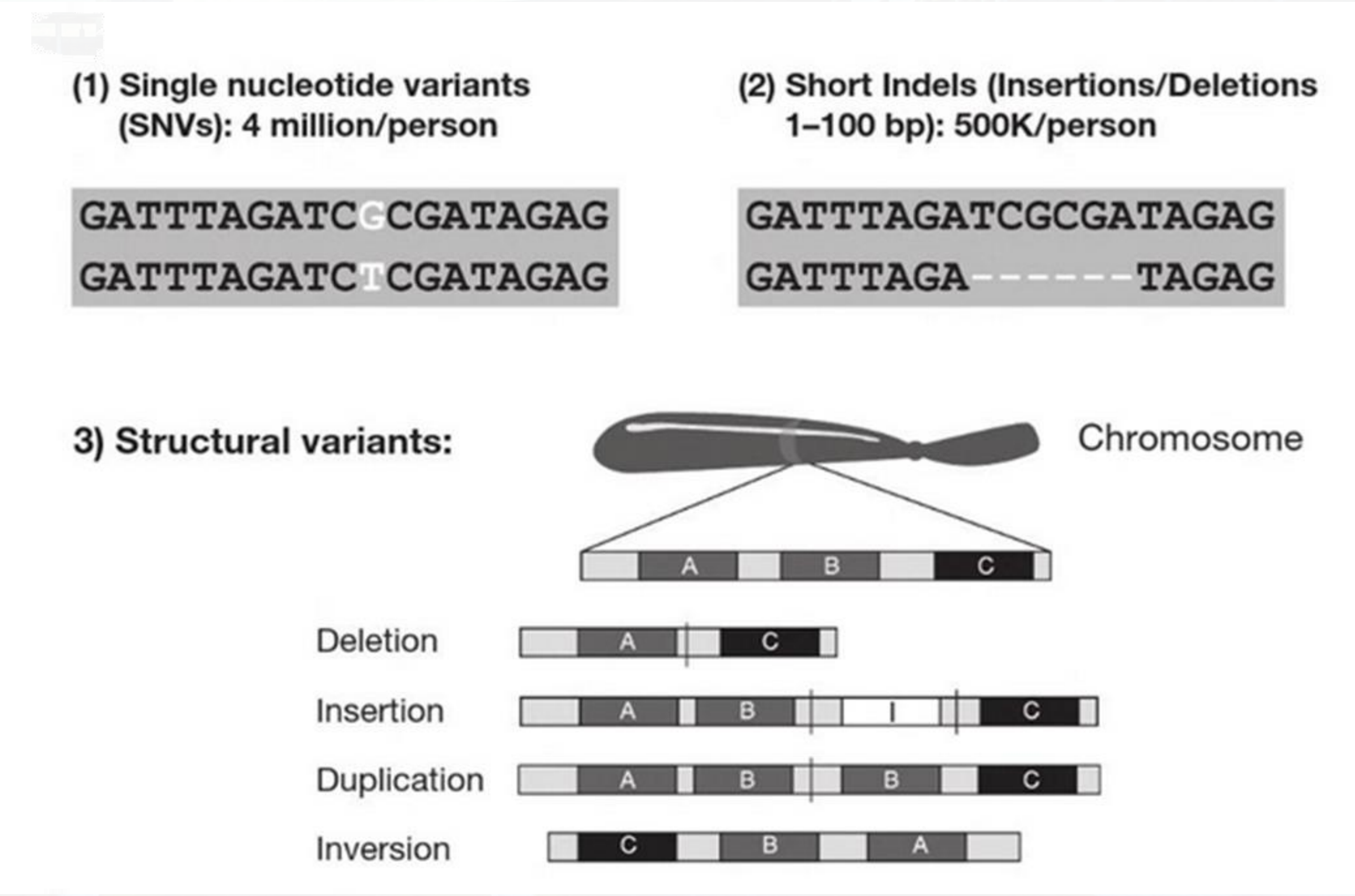
全基因组测序(Whole Genome Sequencing)

- 当前基因组测序主要是针对已有基因组物种的重测序。
- 重测序主要比较个体基因组与参考基因组的差异


...CATTCAGTAG...
...AGCCATTAG...
...GGTAGTTAG...
...GGTAGTTAG...
...AGCCATTAG...
...GGTAAACTAG...
Genomic DNA
Next-generation DNA sequencing
Millions-billions of ***reads***
~30-1,000 nucleotides
RESEQUENCING
De Novo ASSEMBLY
Align reads to ***reference genome*** and identify variants
Construct ***genome sequence*** from overlaps between reads

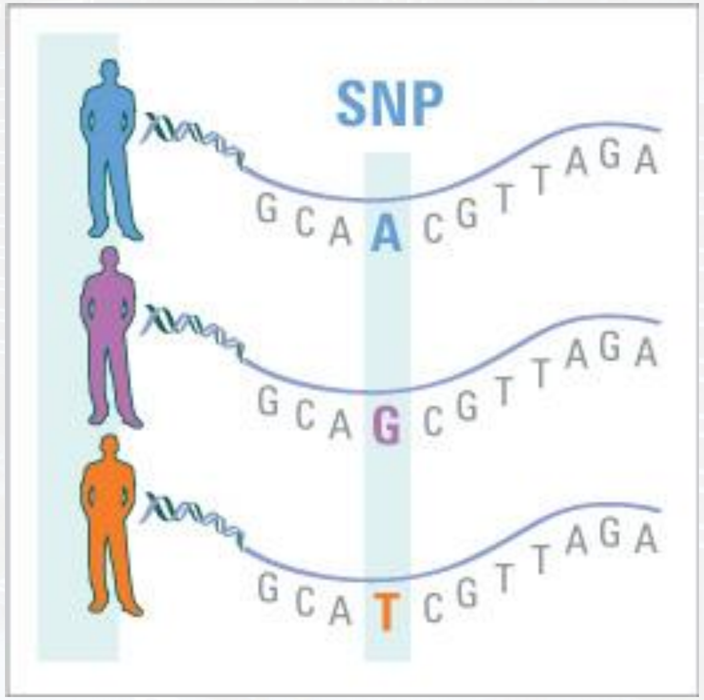
Genetic Variations(遗传变异)

- **SNV**: Single Nucleotide Variant
- **Indel**: Insertion/Deletion
- **SV**: Structural Variants
  - **CNV**: Copy Number Variation



Single Nucleotide Polymorphisms (SNPs)

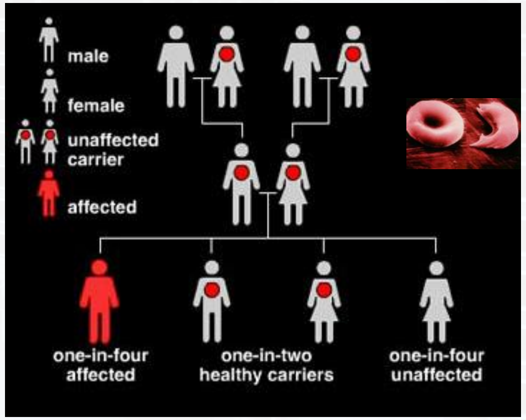
- 从种群概念上讲，一般把那些在种群中发生频率大于1%的SNVs称为SNPs。
- ~ 97 % of the genome between any two individuals is identical
  - ~ 1% of the differences are single nucleotide variations (SNPs)
  - ~2% Other changes – copy number variations, deletions
- 人类基因组每隔500至1000个碱基就会存在一处SNP位点。Between 11-12 million SNPs have been identified (dbSNP)



SNP: 单核苷酸多态性

SNPs can lead to altered protein sequence and function

- Inherited (Germline) Mendelian genetic disorders are caused by variations in DNA (SNPs)
- Most of these deleterious variations affect the function of the encoded protein
  - e.g., Sickle cell anemia Val to Glu codon 6.

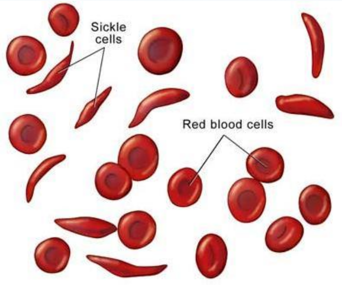


http://www.orgsites.com/va/pasca/

```
Normal HbA     ATGGTGCACCTGACTCCTGTGGAGAAGTC
Disease HbS    ATGGTGCACCTGACTCCTGAGGAGAAGTC
```

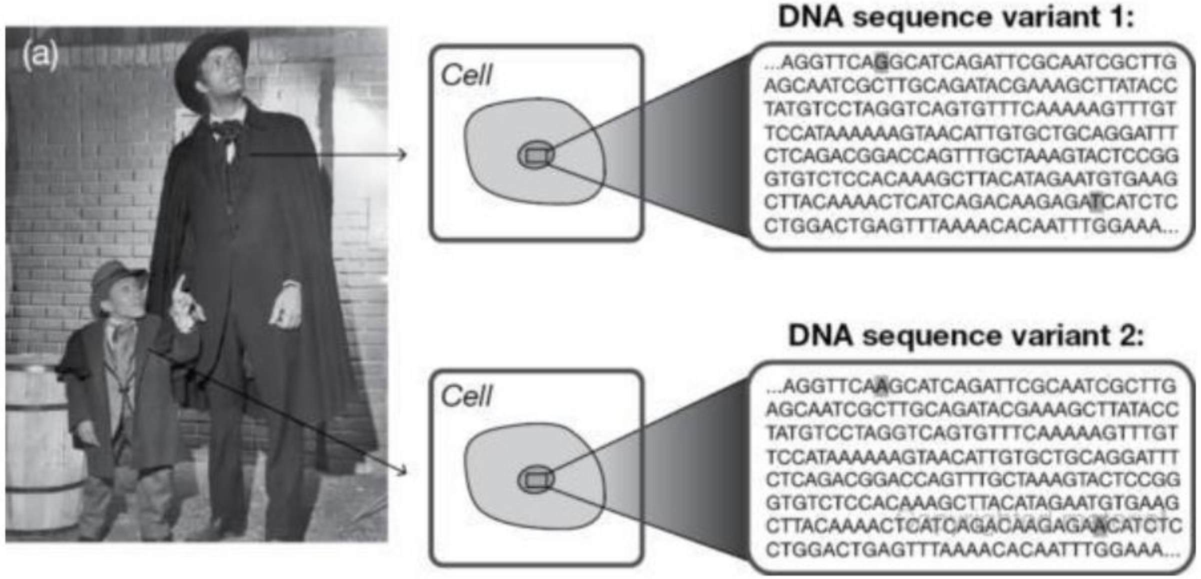
↓

```
Normal HbA   MVHLTPVEKSAVTA
Disease HbS  MVHLTPEEKSAVTA
```



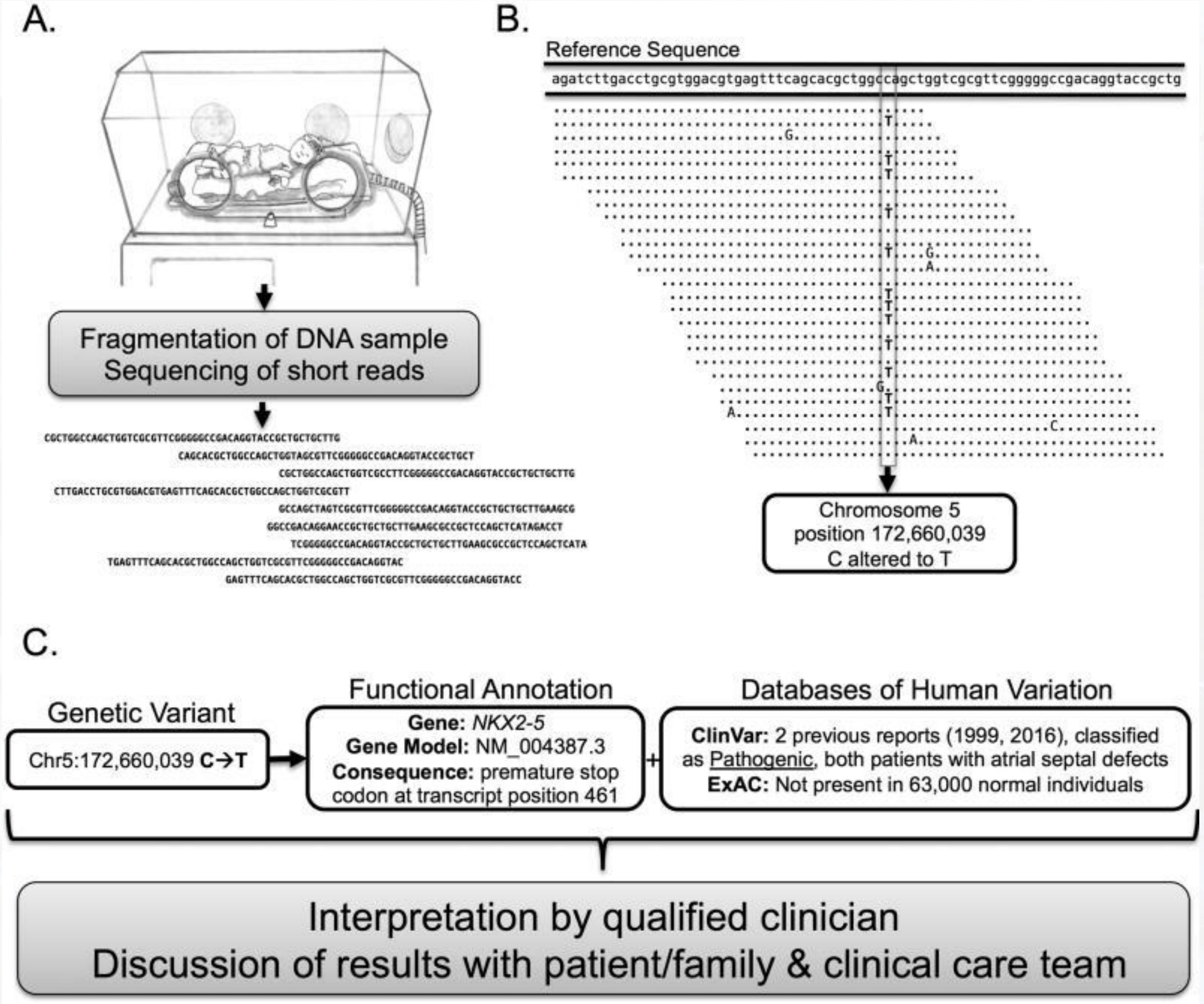
biologycorner.com

基因型(genotype)决定表型(phenotype)

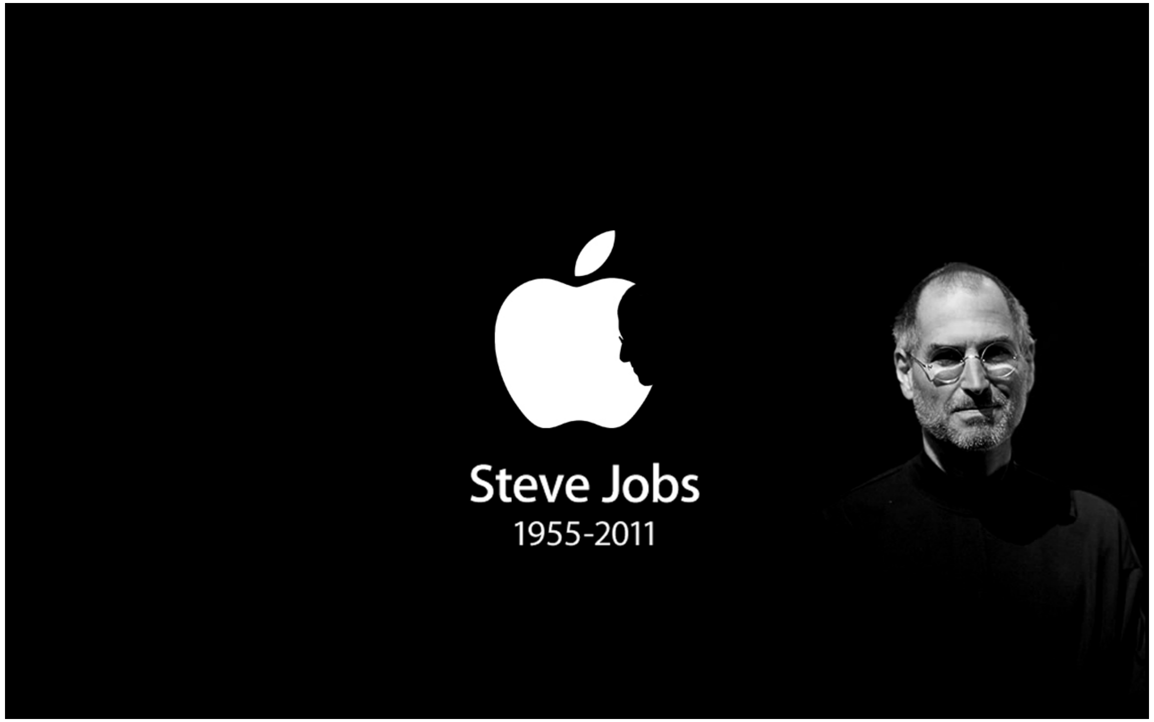

(a)
Cell
DNA sequence variant 1:
...AGGTTCAGGCATCAGATTCGCAATCGCTTG
AGCAATCGCTTGCAGATACGAAAGCTTATACC
TATGTCCTAGGTCAGTGTTTCAAAAAGTTTGT
TCCATAAAAAAGTAACATTGTGCTGCAGGATTT
CTCAGACGGACCAGTTTGCTAAAGTACTCCGG
GTGTCTCCACAAAGCTTACATAGAATGTGAAG
CTTACAAAACTCATCAGACAAGAGATCATCTC
CTGGACTGAGTTTAAAACACAATTTGGAAA...
Cell
DNA sequence variant 2:
...AGGTTCAAGCATCAGATTCGCAATCGCTTG
AGCAATCGCTTGCAGATACGAAAGCTTATACC
TATGTCCTAGGTCAGTGTTTCAAAAAGTTTGT
TCCATAAAAAAGTAACATTGTGCTGCAGGATTT
CTCAGACGGACCAGTTTGCTAAAGTACTCCGG
GTGTCTCCACAAAGCTTACATAGAATGTGAAG
CTTACAAAACTCATCAGACAAGAGAACATCTC
CTGGACTGAGTTTAAAACACAATTTGGAAA...

通过不同表型差异与基因组的关联分析(Association Study)，研究表型差异的遗传学机制。

- While **single gene** and well-known Mendelian genetic disorders, such as sickle-cell anemia, Tay–Sachs disease and cystic fibrosis, can be identified with simple diagnostic techniques, **Whole genome sequencing (WGS)** can be used to identify the cause of a large range of genetic diseases.


A.
Fragmentation of DNA sample
Sequencing of short reads
B.
Reference Sequence
agatcttgacctgcgtggacgtgagtttcagcacgctggccagctggtcgcgttcgggggccgacaggtaccgctg
Chromosome 5
position 172,660,039
C altered to T
C.
Genetic Variant
Chr5:172,660,039 C→T
Functional Annotation
Gene: NKX2-5
Gene Model: NM_004387.3
Consequence: premature stop codon at transcript position 461
Databases of Human Variation
ClinVar: 2 previous reports (1999, 2016), classified as Pathogenic, both patients with atrial septal defects
ExAC: Not present in 63,000 normal individuals
Interpretation by qualified clinician
Discussion of results with patient/family & clinical care team

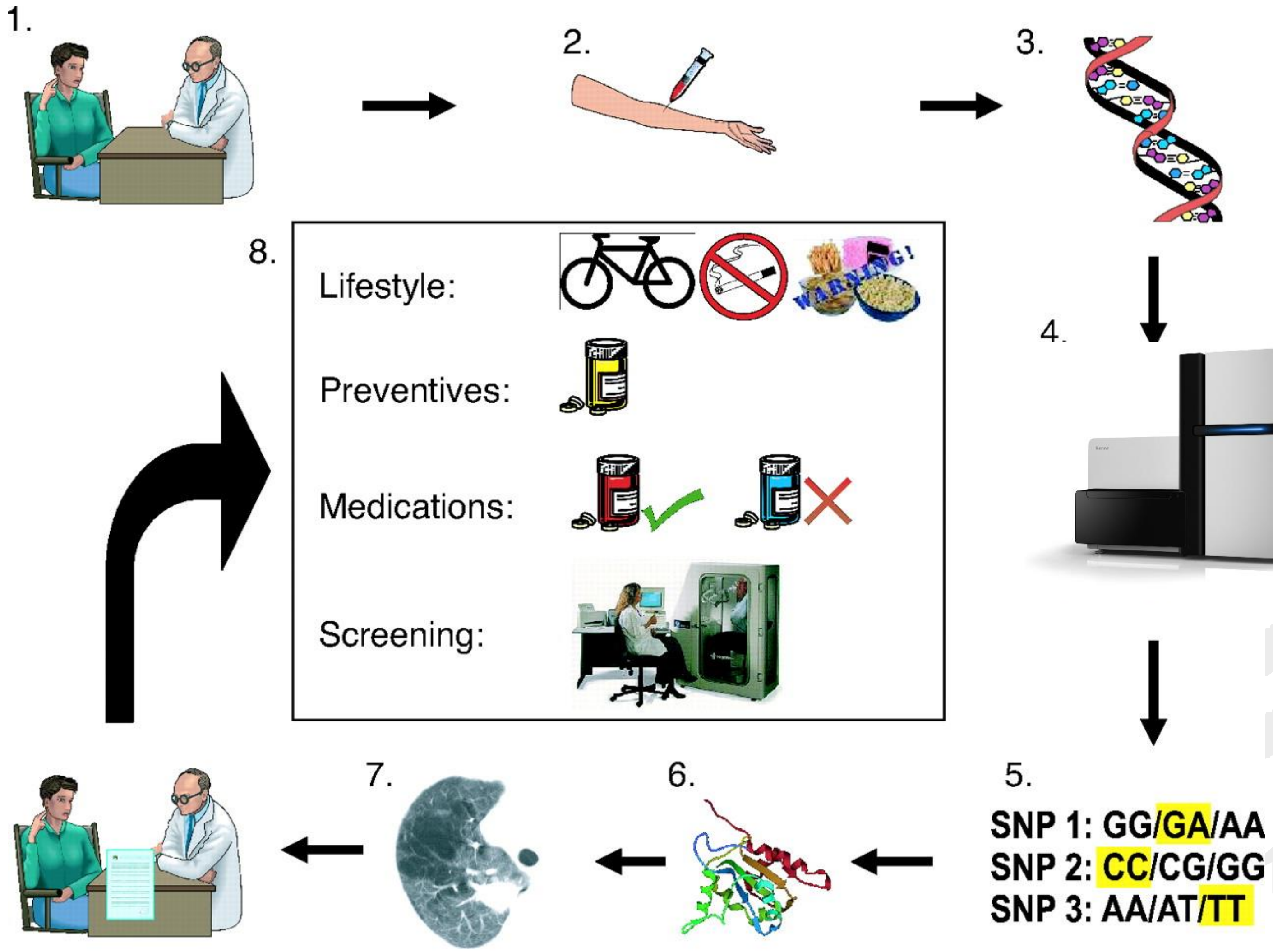
Fighting cancer through next-generation sequencing



为了赢得与癌症的斗争，史蒂夫·乔布斯曾花费10万美元巨资为自己DNA测序，寄希望于找出治疗肿瘤的基因。

“I’m either going to be one of the first to be able to outrun a cancer like this, or I’m going to be one of the last to die from it.” -- Steve Jobs

The Future of Genomics in Medicine

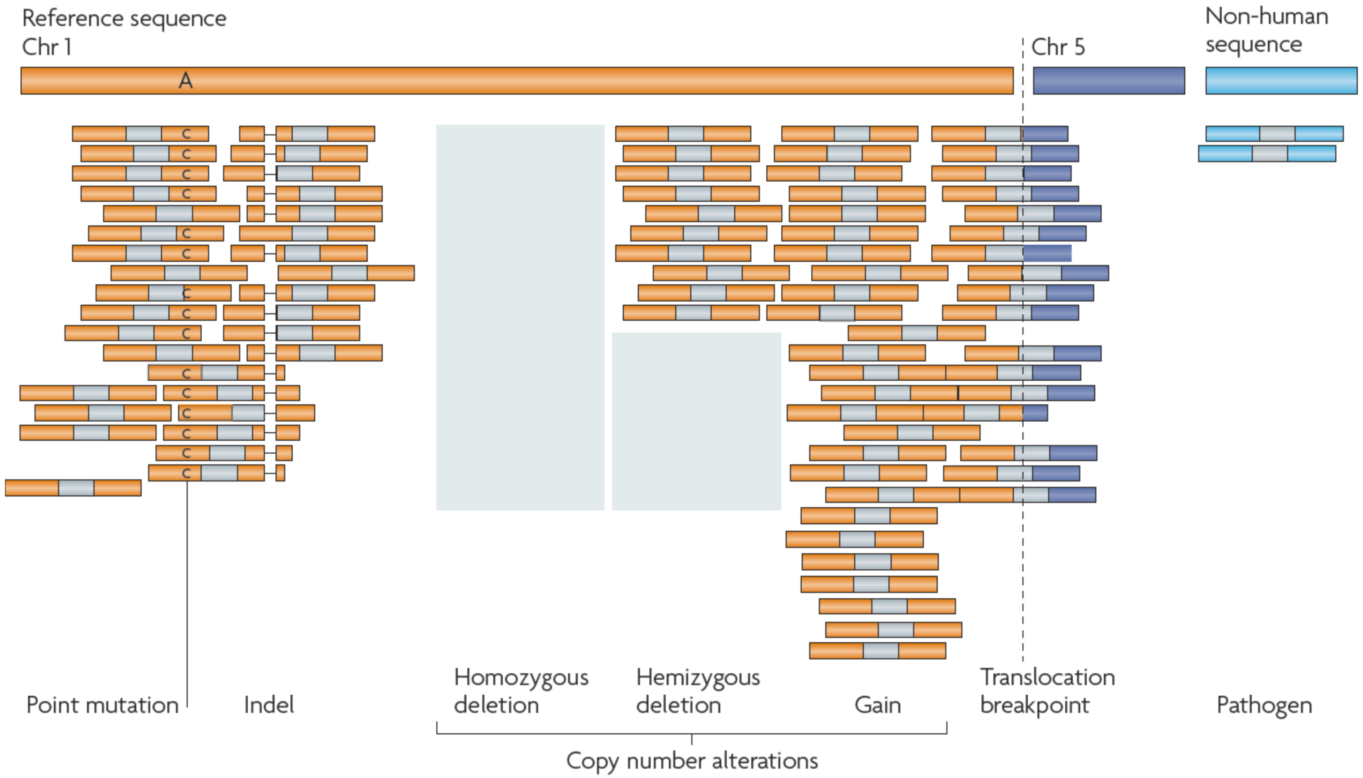


©2007 by American College of Chest Physicians

Tebbutt S J et al. Chest 2007;131:1216-1223

重测序数据分析的主要工作

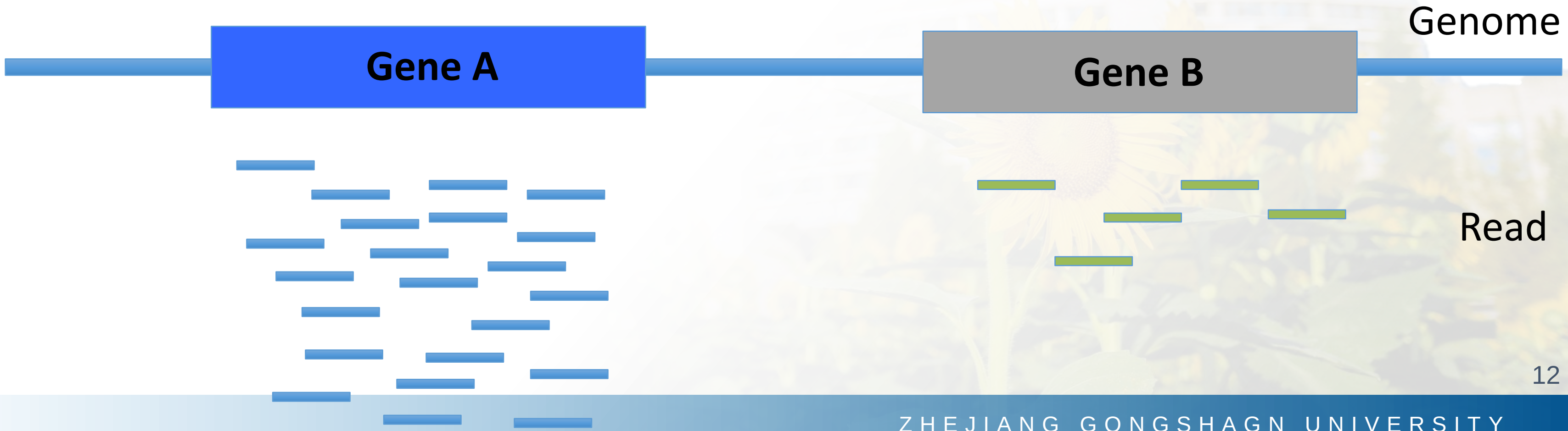
- 快速地将数百万个短读段(reads) 回帖到参考基因组上，
- 准确地鉴定SNPs和Indels等突变。



Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11, 685–696 (2010).

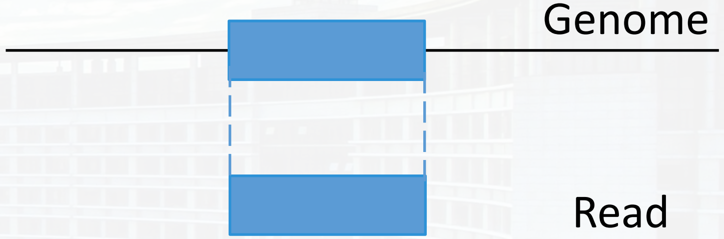
Read Mapping : 序列比对（alignment)

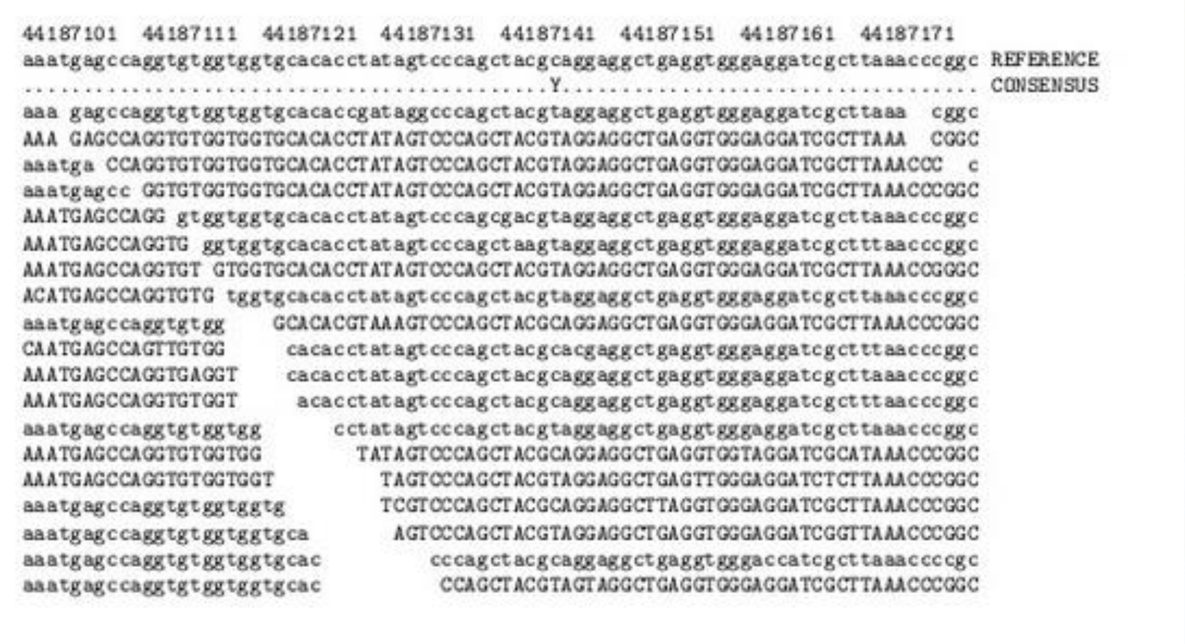
- Aligning Millions of Short Sequence Reads
- One sequence is “embedded” in the other sequence (NGS reads, PCR primer, etc.)
  - ✦“Local alignment” for long sequence
  - ✦“Global alignment” for short sequence
- Aligners:
  - ✦BWA, Bowtie2, STAR, HISAT2, …
  - ✦minimap2, STARlong… (Nanopore, PacBio)



Mapping: Input data

- Reference Genome
  - ✦Nucleotide sequence (FastA)
  - ✦Length: Hundreds of Mb per chromosome
  - ✦~3 Gb in total (for human genome)
- Reads
  - ✦Nucleotide sequence with various qualities (FastQ)
    - error rate ranges from a few tenths of a percent to several percent
  - ✦Length: ~100 bp per read
  - ✦Hundreds of Gbs per run

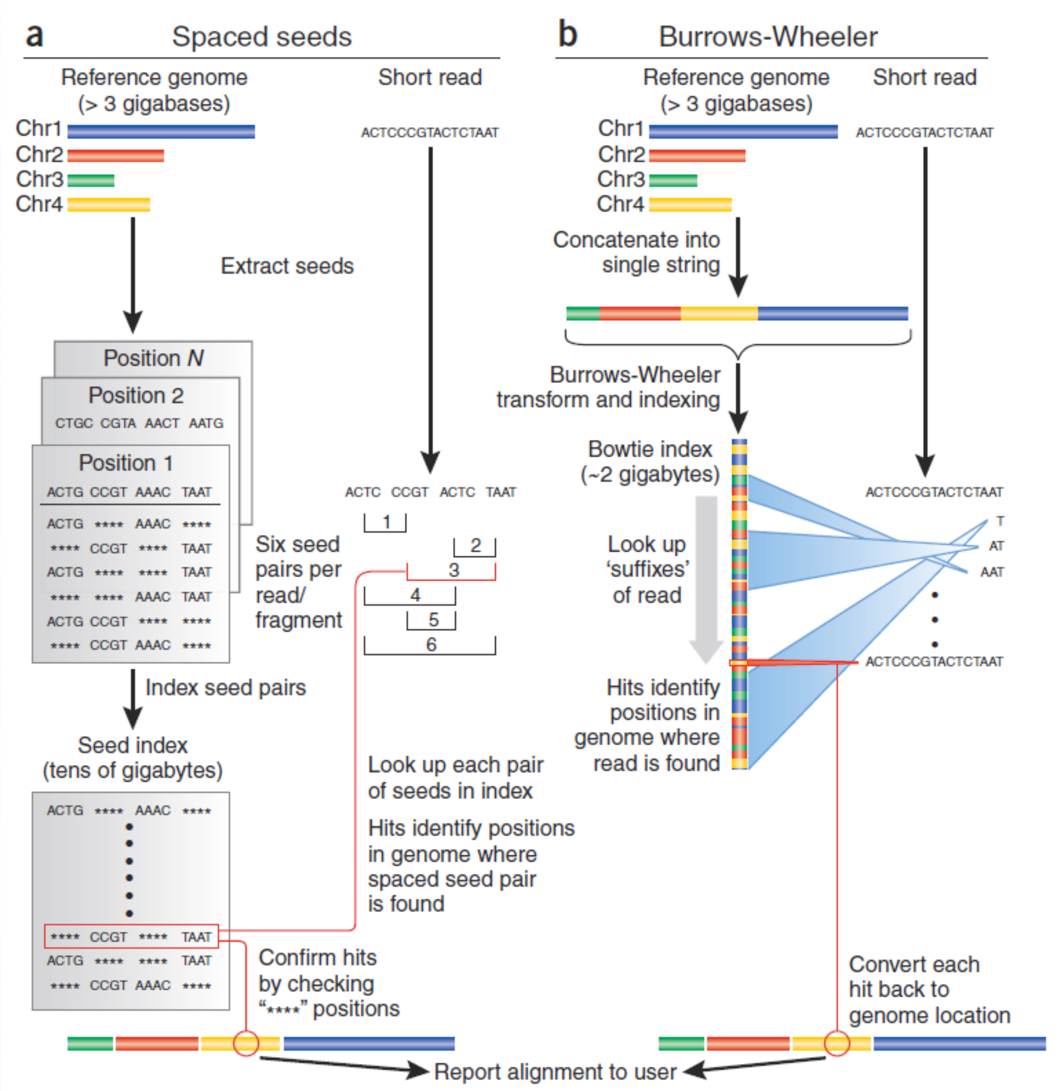




```
44187101  44187111  44187121  44187131  44187141  44187151  44187161  44187171
aaatgagccaggtgtggtggtgcacacctatagtcccagctacgcaggaggctgaggtgggaggatcgcttaaacccggc REFERENCE
............................................Y................................... CONSENSUS
aaa gagccaggtgtggtggtgcacaccgataggcccagctacgtaggaggctgaggtgggaggatcgcttaaa  cggc
AAA GAGCCAGGTGTGGTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAA  CGGC
aaatga CCAGGTGTGGTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAACCC  c
aaatgagcc GGTGTGGTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAACCCGGC
AAATGAGCCAGG gtggtggtgcacacctatagtcccagcgacgtaggaggctgaggtgggaggatcgcttaaacccggc
AAATGAGCCAGGTG ggtggtgcacacctatagtcccagctaagtaggaggctgaggtgggaggatcgctttaacccggc
AAATGAGCCAGGTGT GTGGTGCACACCTATAGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGCTTAAACCGGGC
ACATGAGCCAGGTGTG tggtgcacacctatagtcccagctacgtaggaggctgaggtgggaggatcgcttaaacccggc
aaatgagccaggtgtgg    GCACACGTAAAGTCCCAGCTACGCAGGAGGCTGAGGTGGGAGGATCGCTTAAACCCGGC
CAATGAGCCAGTTGTGG     cacacctatagtcccagctacgcacgaggctgaggtgggaggatcgctttaacccggc
AAATGAGCCAGGTGAGGT    cacacctatagtcccagctacgcaggaggctgaggtgggaggatcgcttaaacccggc
AAATGAGCCAGGTGTGGT     acacctatagtcccagctacgcaggaggctgaggtgggaggatcgctttaacccggc
aaatgagccaggtgtggtgg      cctatagtcccagctacgtaggaggctgaggtgggaggatcgcttaaacccggc
AAATGAGCCAGGTGTGGTGG        TATAGTCCCAGCTACGCAGGAGGCTGAGGTGGTAGGATCGCATAAACCCGGC
AAATGAGCCAGGTGTGGTGGT         TAGTCCCAGCTACGTAGGAGGCTGAGTTGGGAGGATCTCTTAAACCCGGC
aaatgagccaggtgtggtggtg        TCGTCCCAGCTACGCAGGAGGCTTAGGTGGGAGGATCGCTTAAACCCGGC
aaatgagccaggtgtggtggtgca       AGTCCCAGCTACGTAGGAGGCTGAGGTGGGAGGATCGGTTAAACCCGGC
aaatgagccaggtgtggtggtgcac         cccagctacgcaggaggctgaggtgggaccatcgcttaaaccccgc
aaatgagccaggtgtggtggtgcac          CCAGCTACGTAGTAGGCTGAGGTGGGAGGATCGCTTAAACCCGGC
```

Mapping algorithms

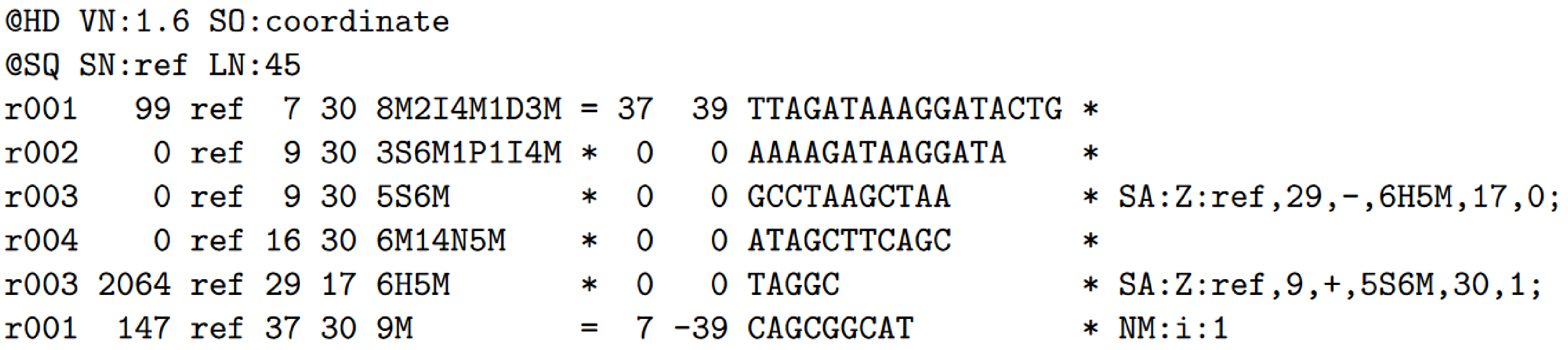
- Burrows-Wheeler transform (BWT)
  - ✦数据压缩算法(bzip2)
  - ✦BWT-based tools: BWA, Bowtie2, SOAP2
  - ✦Fast, memory-efficient, Less sensitive
- Hashing (哈希)
  - ✦Hash-based tools: MAQ, Novoalign, Stampy
  - ✦most accurate overall results



Trapnell, C. & Salzberg, S. L. (2009) How to map billions of short reads onto genomes. Nat Biotech 27, 455–457.

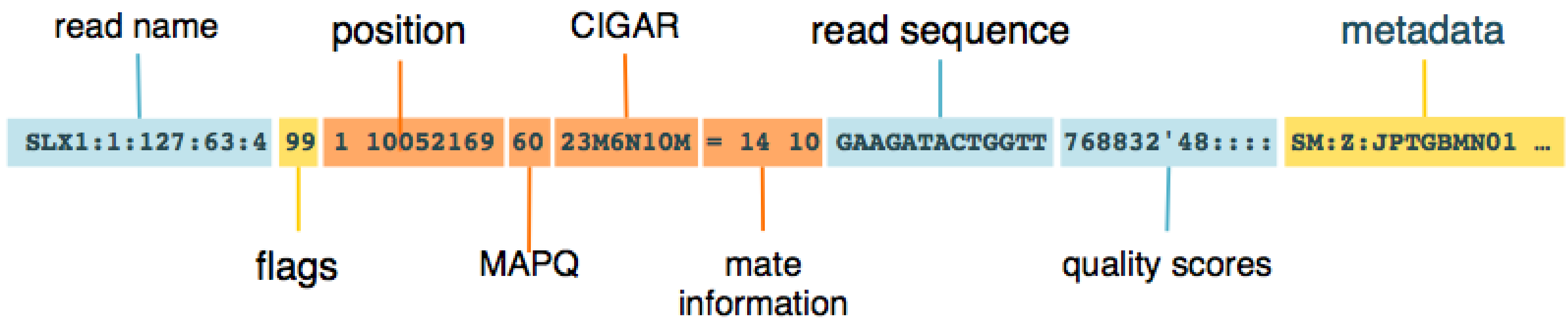
比对结果文件 — SAM

- SAM(Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments
  - ✦在SAM格式中，每一行表示 一个read的比对结果



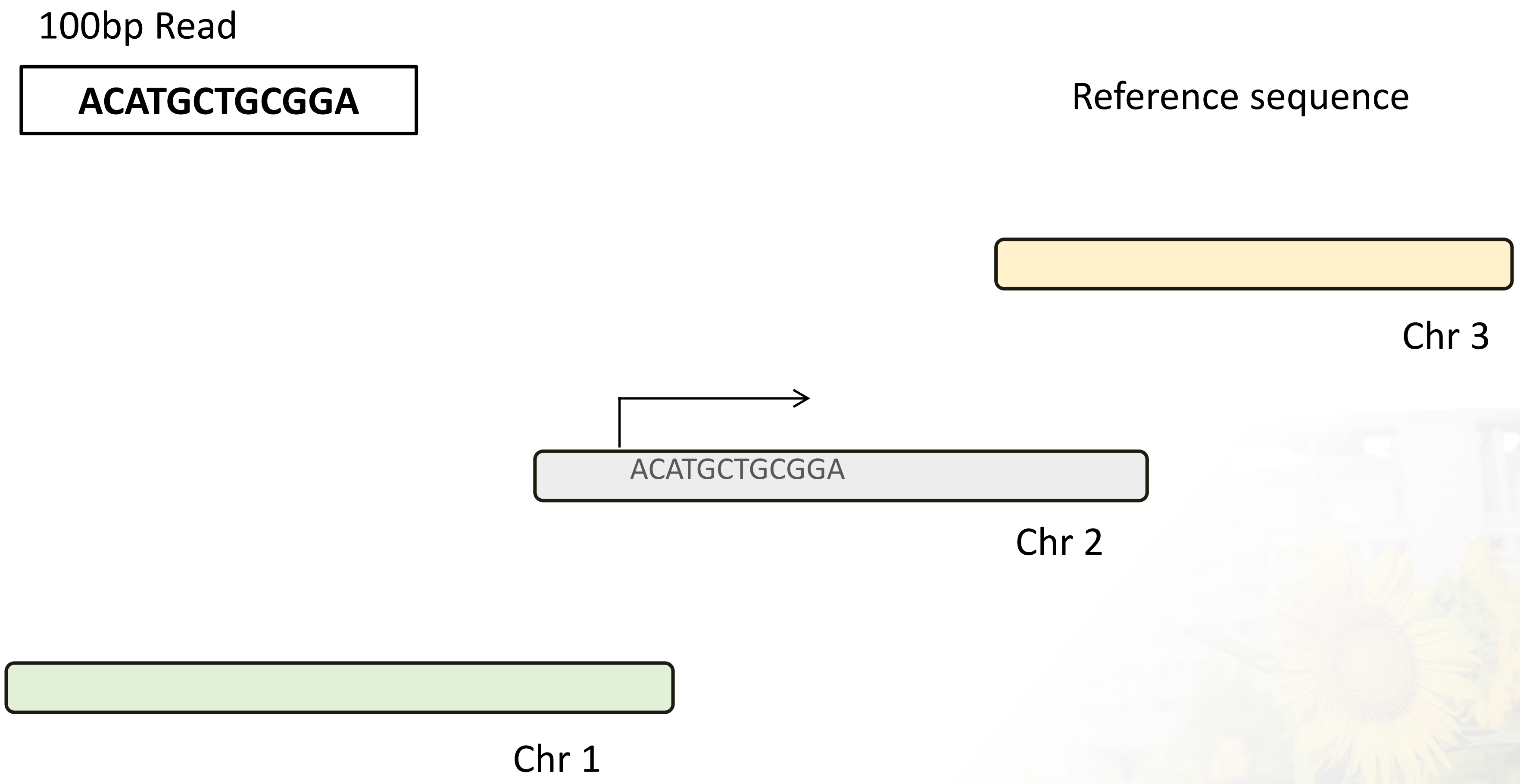
```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



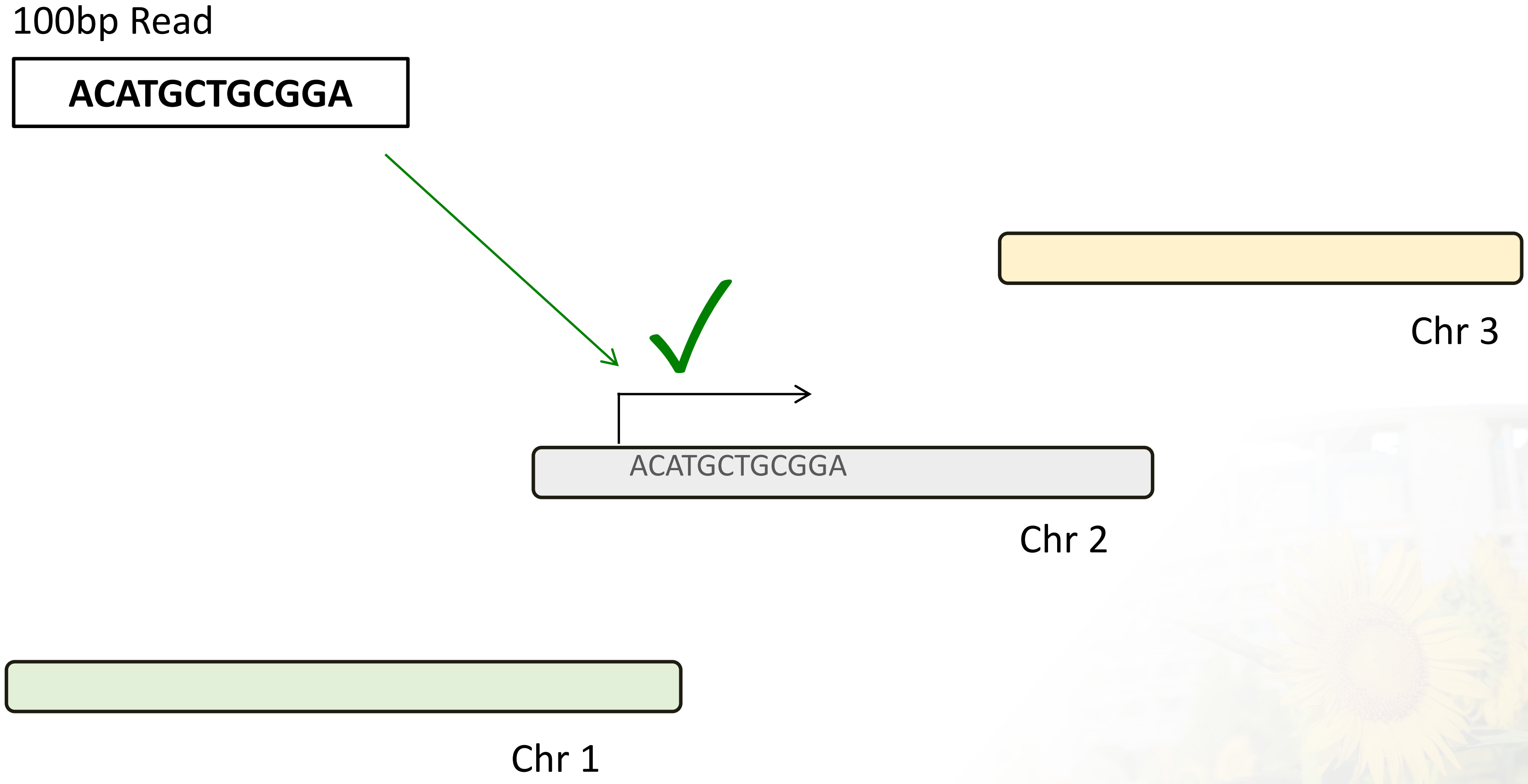
Mapping quality (MAPQ) score is the probability that the read is incorrectly mapped, or more importantly, the probability that the read maps uniquely.

Read Alignment



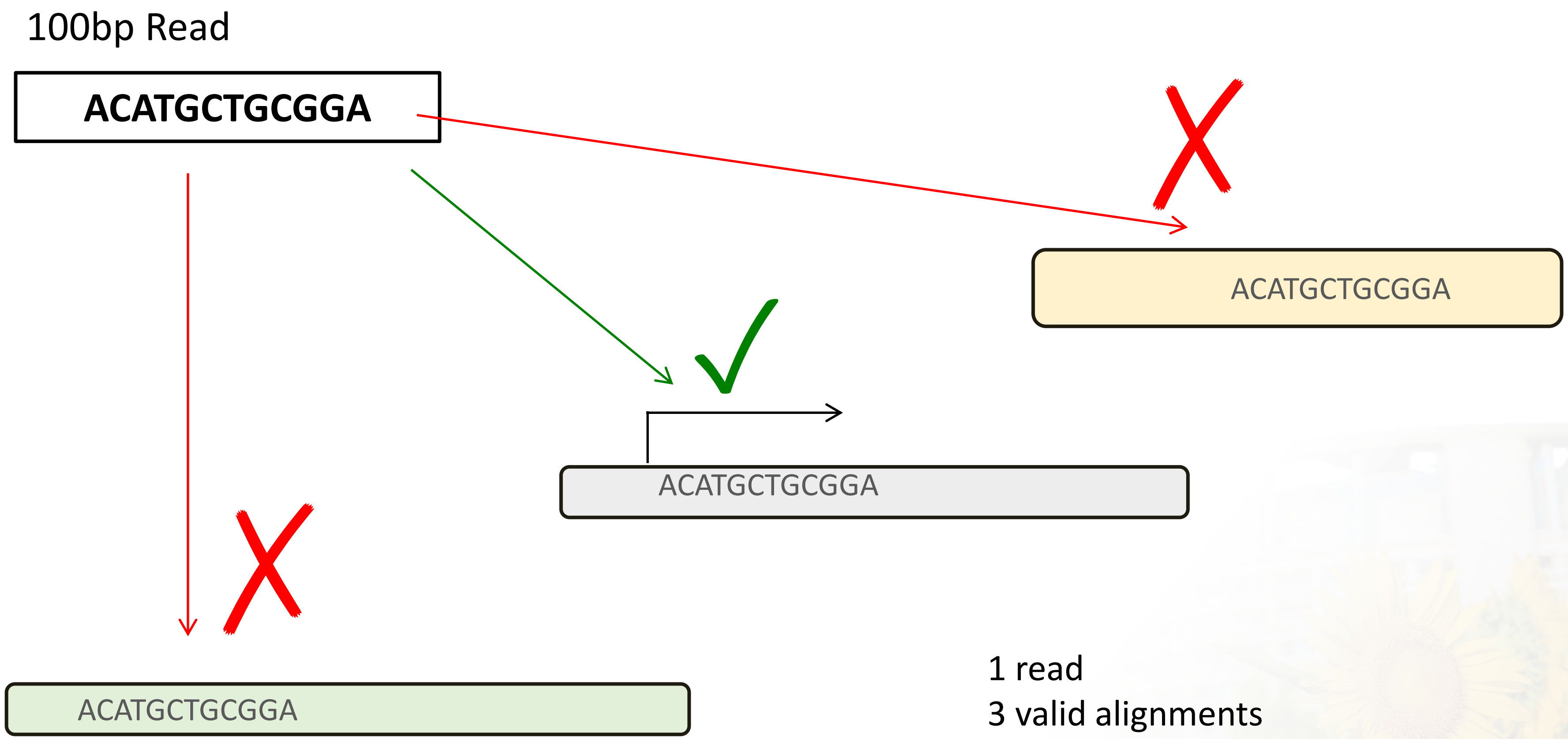
Steve Munger, 2017

The perfect read: 1 read = 1 unique alignment.



Steve Munger, 2017

Some reads will align equally well to multiple locations. “Multi-mapped reads”



1 read
3 valid alignments
Only 1 alignment is correct

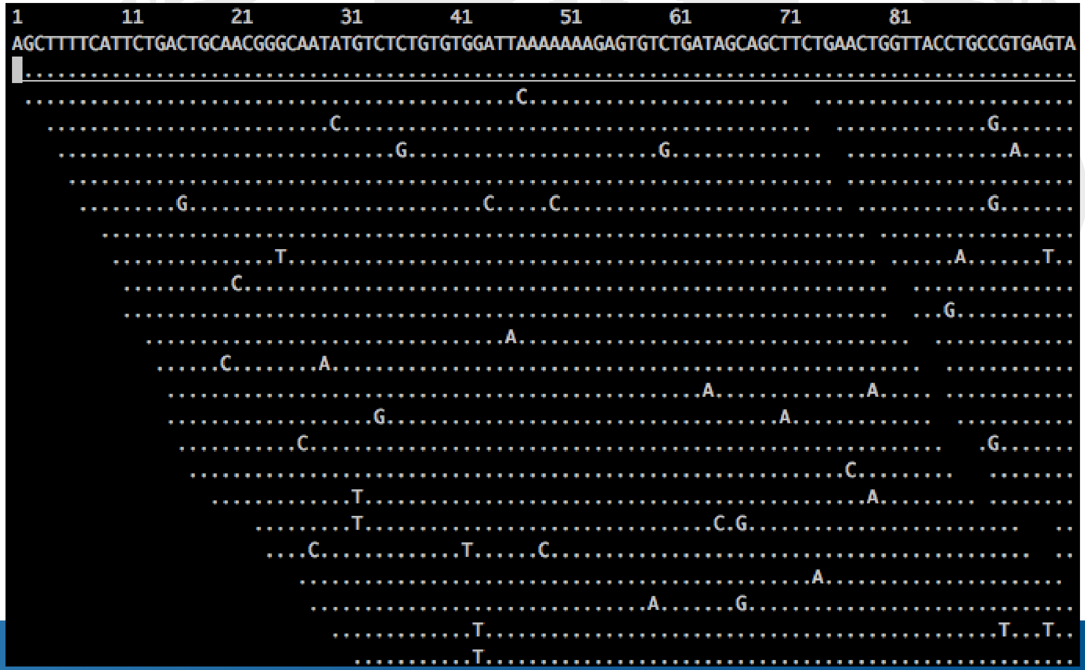
- Ignore them?
- Weight them?

Steve Munger, 2017

SAM文件工具 — SAMtools

- Tools to handle Bam/Sam files: SAMtools
  - ✦$samtools view test.bam
  - ✦$samtools view –h test.bam | less #show headers
  - ✦$samtools flagstat ./data/SRR3096662_Aligned.sort.bam
- View alignment with samtools
  - ✦$samtools index ./SRR3096662_Aligned.sort.bam
  - ✦$samtools tview ./SRR3096662_Aligned.sort.bam --reference ./GRCh37.genome.fa

  #Need to make the index for the bam file



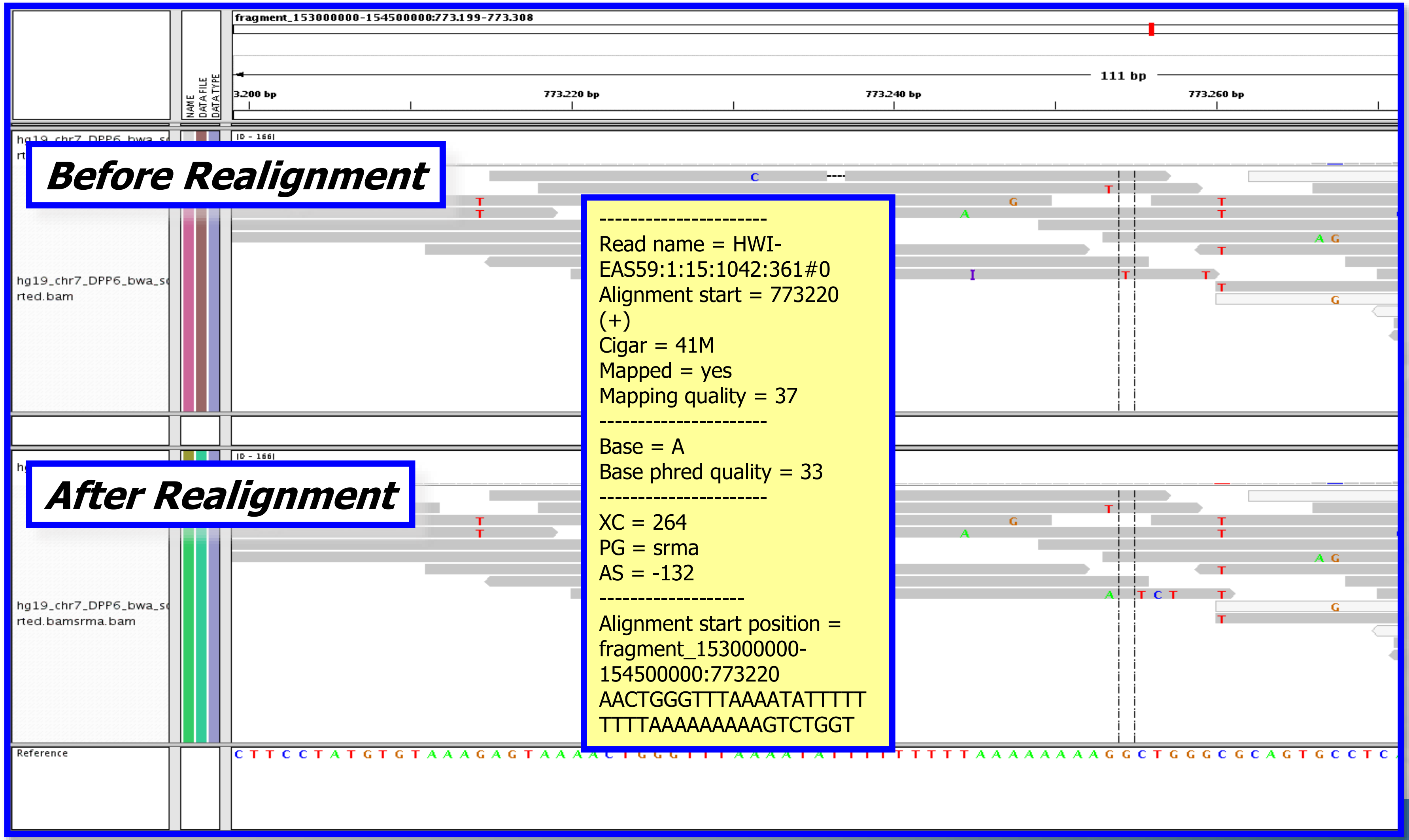
```
1         11        21        31        41        51        61        71        81
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTA
 ................................................................................................
 .............................................C........................ .........................
   ..........................C........................................... ...............G.......
    ...............................G.......................G.............. ................A.....
     ..................................................................... ......................
      .........G...........................C....C.......................... .............G.......
        ..................................................................... ...................
         ...............T.....................................................  ......A.......T..
          ..........C..........................................................  ................
          .....................................................................   ...G...........
            .................................A...................................  ..............
             ......C........A.....................................................  .............
              .................................................A..............A.... .............
              ...................G....................................A............  ............
               ...........C.........................................................   .G........
                ............................................................C........    ........
                  .............T..............................................A........  ........
                      .........T................................C.G.........................   ..
                       ....C.............T.....C.............................................  ..
                          ...............................................A......................
                           ...............................A......G...............................
                             .............T...............................................T...T..
                               ...........T......................................................
```

Aligned reads

比对结果可视化(Visualization):IGV

Integrative Genome Viewer (IGV)：

http://software.broadinstitute.org/software/igv/download



fragment_153000000-154500000:773,199-773,308
111 bp
3,200 bp
773,220 bp
773,240 bp
773,260 bp
NAME
DATA FILE
DATA TYPE
hg19_chr7_DPP6_bwa_sorted.bam
hg19_chr7_DPP6_bwa_sorted.bamsrma.bam
Reference

Before Realignment

After Realignment

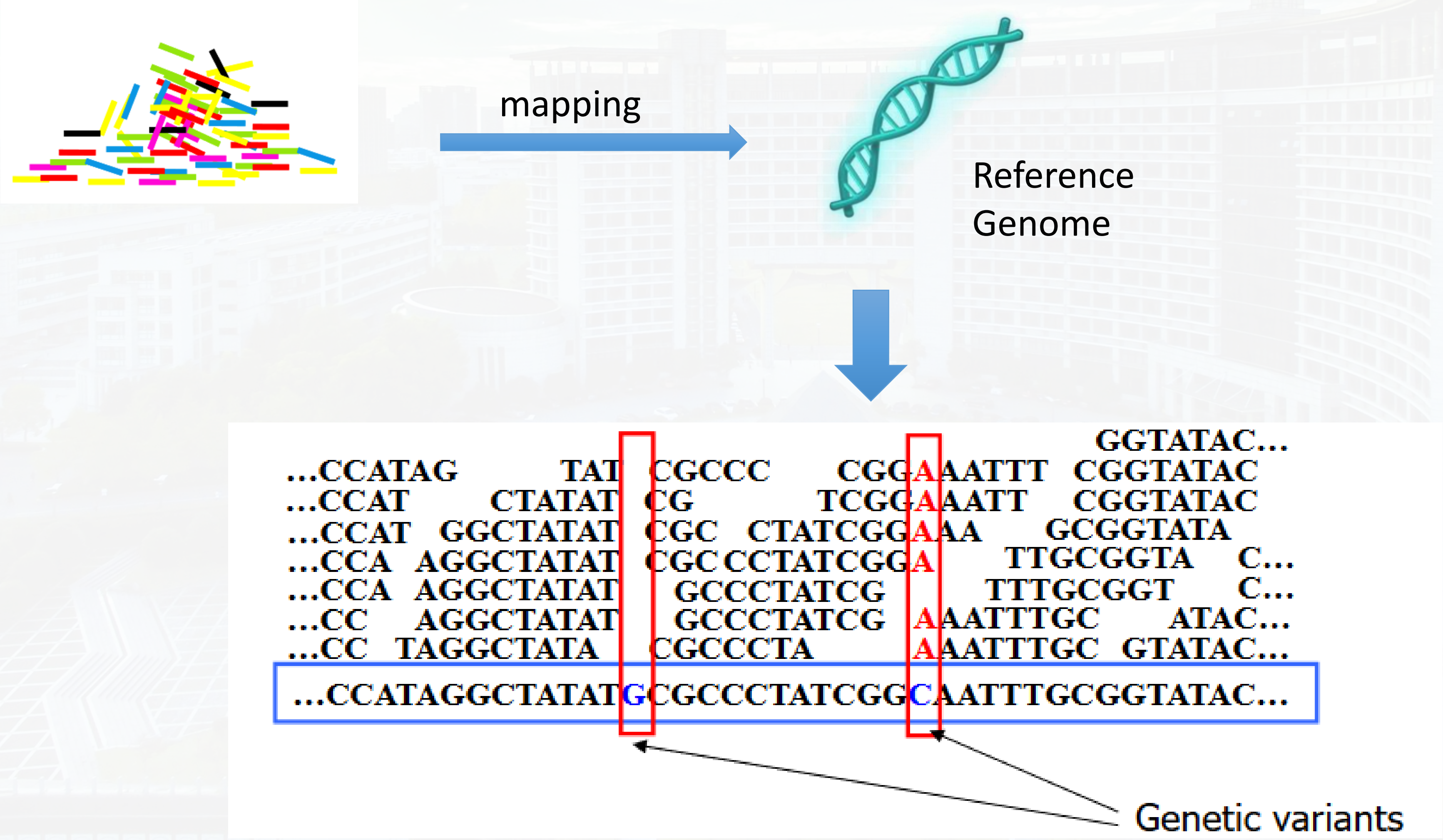
Read name = HWI-EAS59:1:15:1042:361#0
Alignment start = 773220 (+)
Cigar = 41M
Mapped = yes
Mapping quality = 37

Base = A
Base phred quality = 33

XC = 264
PG = srma
AS = -132

Alignment start position = fragment_153000000-154500000:773220
AACTGGGTTTAAAATATTTTTTTTTAAAAAAAAAGTCTGGT

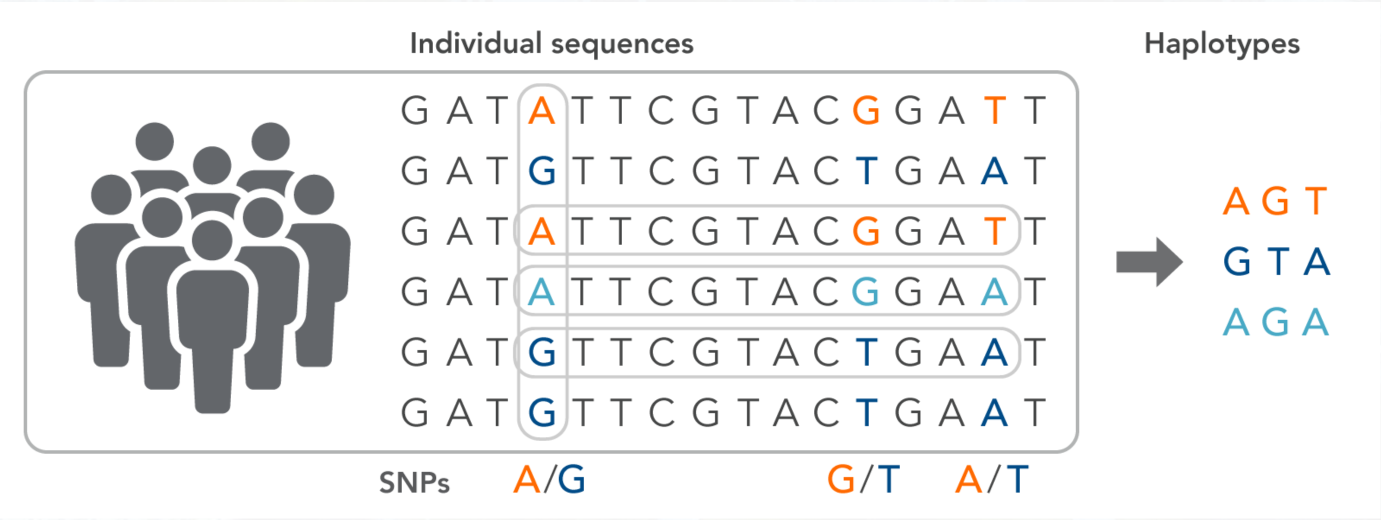
将读段比对到基因组鉴定遗传变异



```
                                                   GGTATAC...
...CCATAG       TAT  CGCCC     CGGAAATTT  CGGTATAC
...CCAT      CTATAT  CG       TCGGAAATT   CGGTATAC
...CCAT  GGCTATAT  CGC  CTATCGGAAA    GCGGTATA
...CCA  AGGCTATAT  CGC CCTATCGGA    TTGCGGTA    C...
...CCA  AGGCTATAT    GCCCTATCG     TTTGCGGT    C...
...CC   AGGCTATAT    GCCCTATCG  AAATTTGC     ATAC...
...CC  TAGGCTATA    CGCCCTA     AAATTTGC  GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

SNP calling VS. Genotyping

- SNP calling: identifies variable sites (variants).
- Genotyping: determines the genotype for each individual at each site.
  - ✦The number of alleles (等位基因) or ploidy (染色体倍性) is decided and fixed.

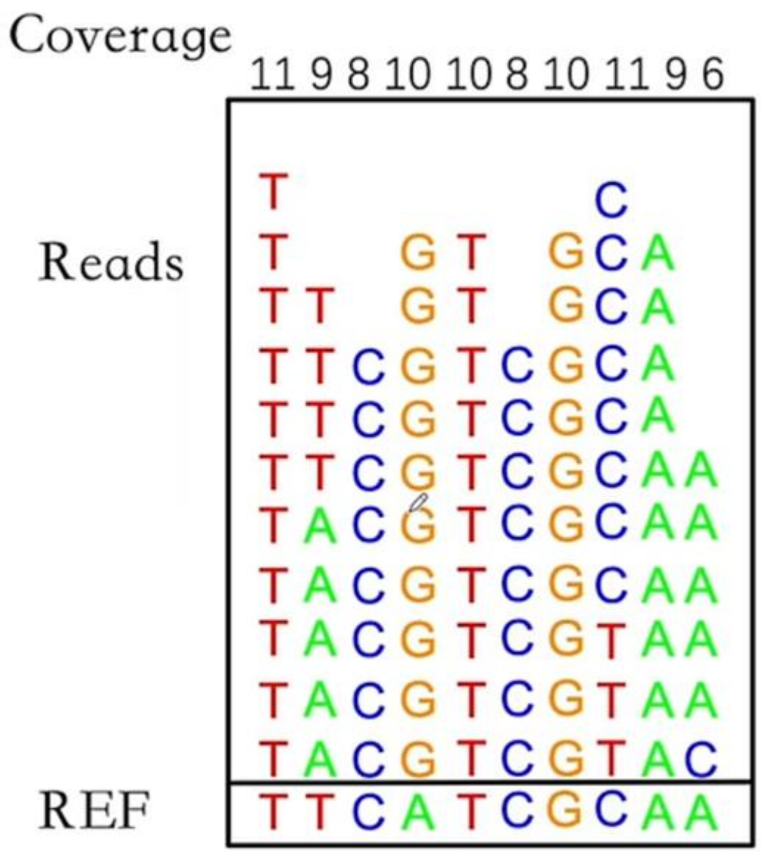


```
G A T A T T C G T A C G G A T T
G A T G T T C G T A C T G A A T
G A T A T T C G T A C G G A T T
G A T A T T C G T A C G G A A T
G A T G T T C G T A C T G A A T
G A T G T T C G T A C T G A A T
```

```
A G T
G T A
A G A
```

Algorithms for genotype and SNP calling

- The sequenced DNA fragments ("reads") are aligned to the reference sequence and the base at each position is determined, counted and then run through a number of statistical tests to determine whether the site is different from the reference sequence.



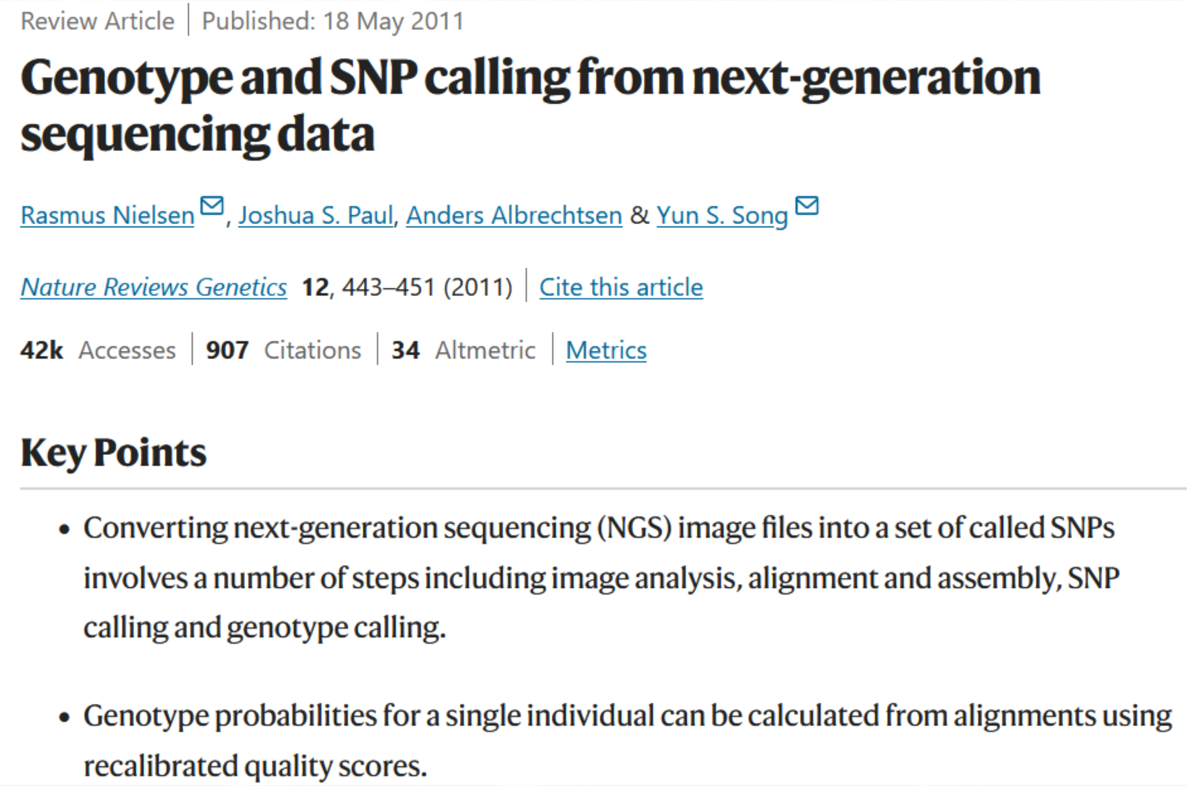
Heterozygous
Homozygous

POS	REF	ALT	Allele Fraction (AF)
2	T	A	(5+1)/(9+2) = 54.5%
4	A	G	100%
8	C	T	(3+1)/(11+2) = 30.8%

Confidence?

Algorithms for genotype and SNP calling

- SNP calling can be done using likelihood ratio tests or Bayesian procedures, …
- 遗传变异的统计学方法可以参考文献:



Review Article | Published: 18 May 2011

Genotype and SNP calling from next-generation sequencing data

Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen & Yun S. Song

Nature Reviews Genetics **12**, 443–451 (2011) | Cite this article

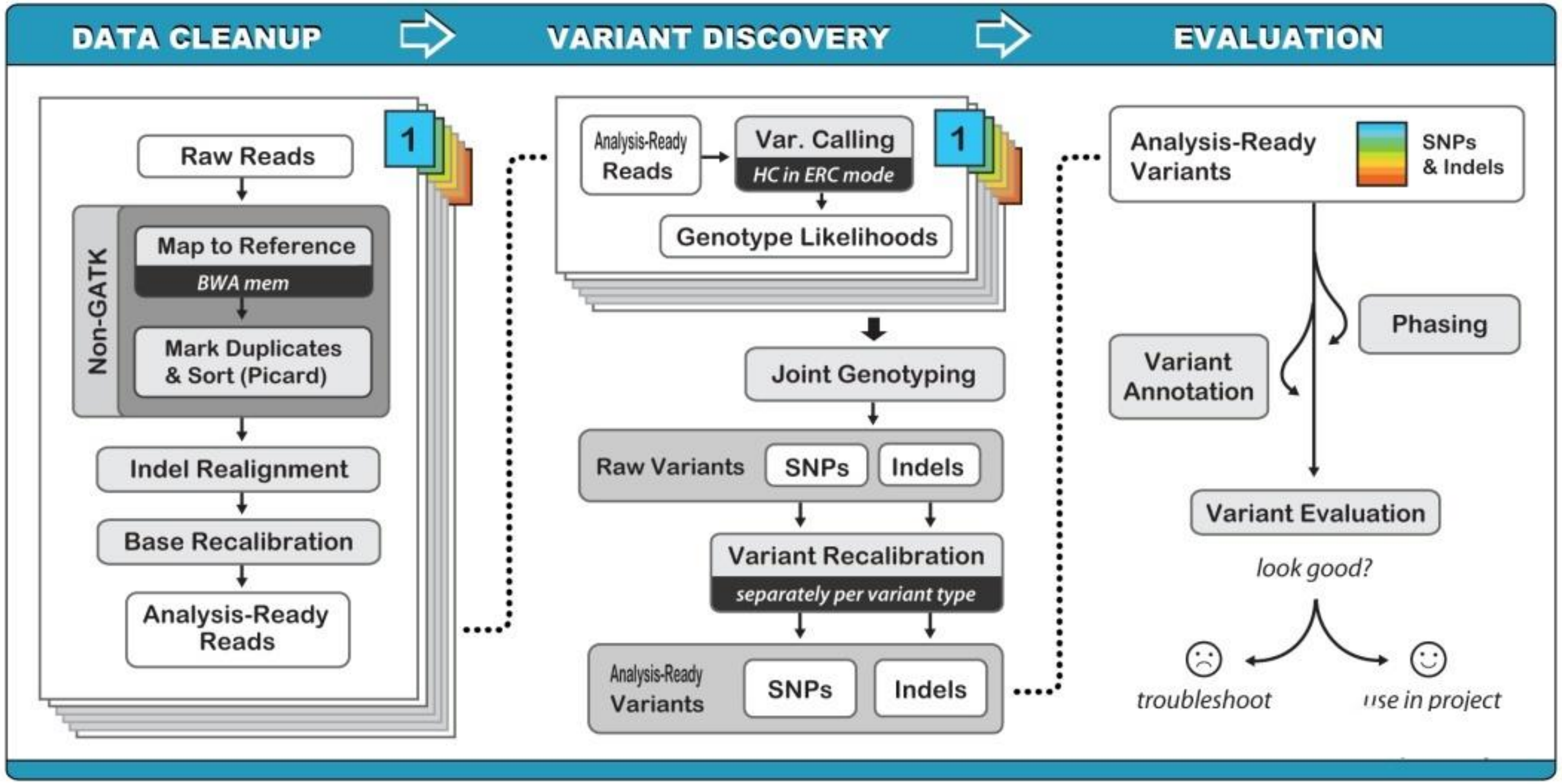
42k Accesses | **907** Citations | **34** Altmetric | Metrics

Key Points

- Converting next-generation sequencing (NGS) image files into a set of called SNPs involves a number of steps including image analysis, alignment and assembly, SNP calling and genotype calling.
- Genotype probabilities for a single individual can be calculated from alignments using recalibrated quality scores.

Nielsen, R., Paul, J., Albrechtsen, A. et al. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12, 443–451 (2011).

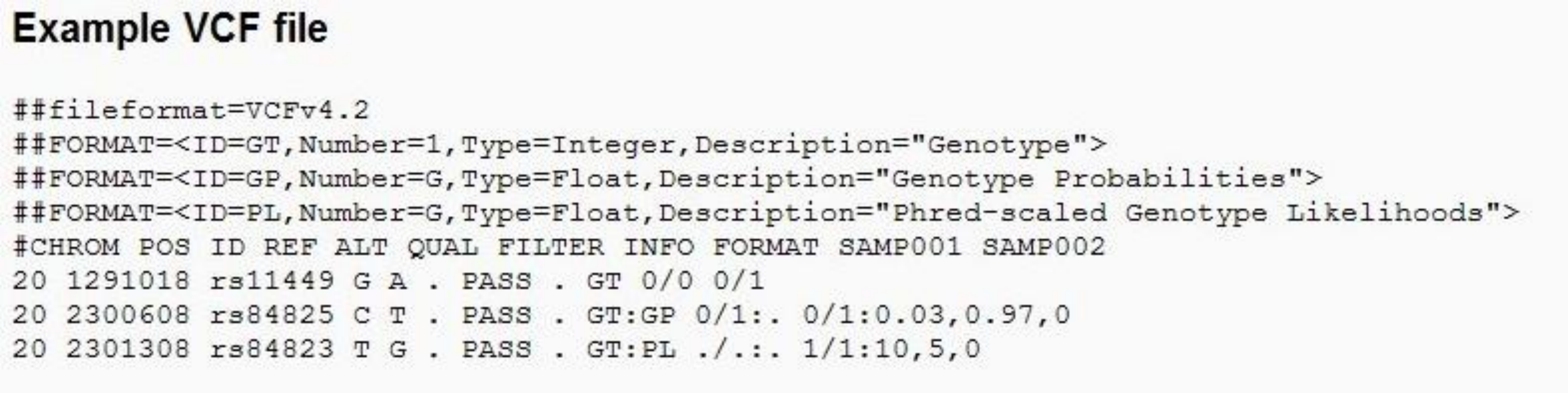
常用的变异鉴定工具：
GATK (Genome Analysis ToolKit)



best practices:
https://gatk.broadinstitute.org/hc/en-us/articles/360036194592-Getting-started-with-GATK4

变异识别结果文件 — VCF

- VCF(Variant Call Format) 是存储遗传变异类型，如SNPs, Indels和SVs的标准文件格式
  - ✦在VCF格式中，每一行表示 一个遗传变异的结果

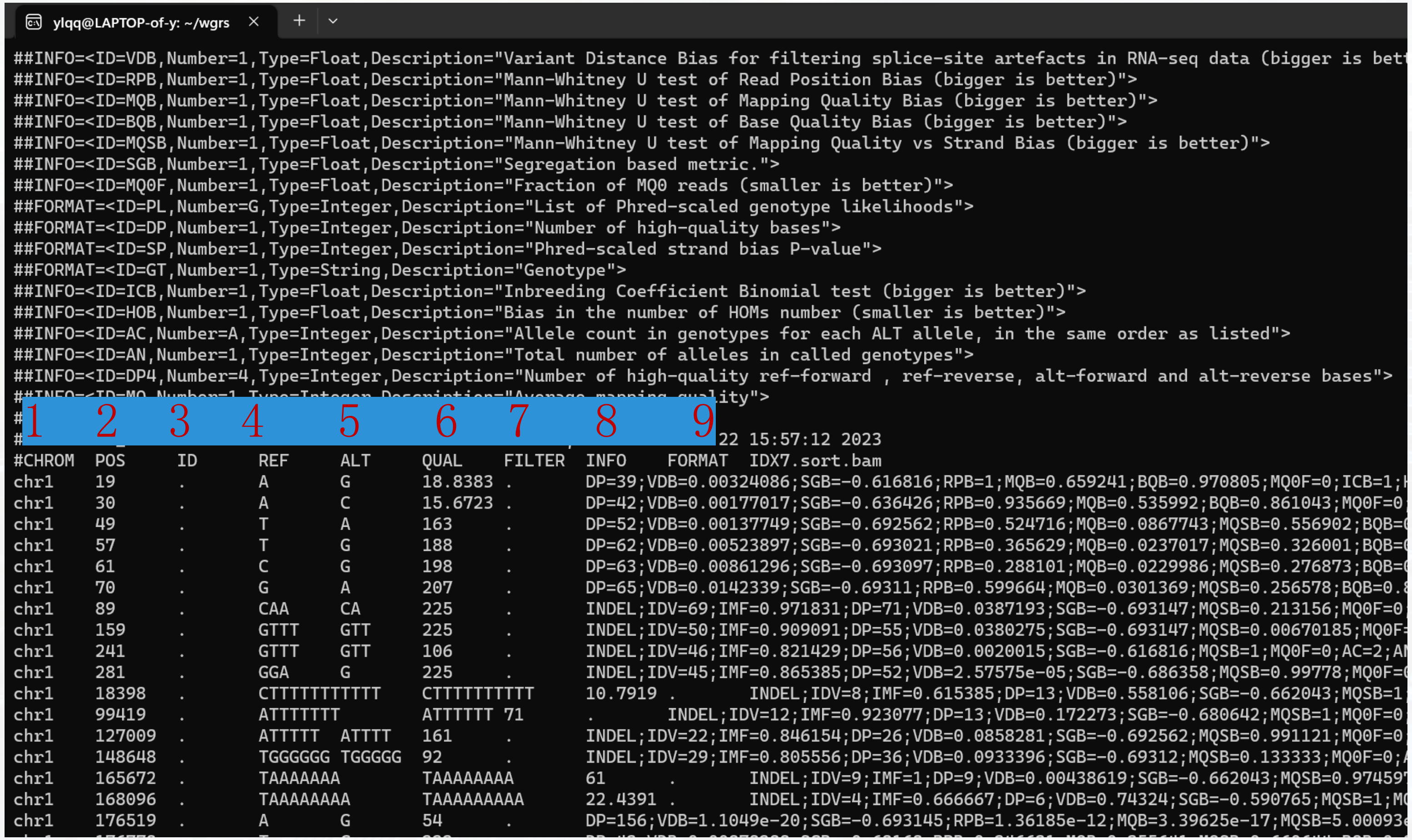


Example VCF file

```
##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype Probabilities">
##FORMAT=<ID=PL,Number=G,Type=Float,Description="Phred-scaled Genotype Likelihoods">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMP001 SAMP002
20 1291018 rs11449 G A . PASS . GT 0/0 0/1
20 2300608 rs84825 C T . PASS . GT:GP 0/1:. 0/1:0.03,0.97,0
20 2301308 rs84823 T G . PASS . GT:PL ./.:. 1/1:10,5,0
```

http://samtools.github.io/hts-specs/

基因组变异的检测结果-VCF



```
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is bett
##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Number of high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">
##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
#
#                                                                           22 15:57:12 2023
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  IDX7.sort.bam
chr1    19      .       A       G       18.8383 .       DP=39;VDB=0.00324086;SGB=-0.616816;RPB=1;MQB=0.659241;BQB=0.970805;MQ0F=0;ICB=1;H
chr1    30      .       A       C       15.6723 .       DP=42;VDB=0.00177017;SGB=-0.636426;RPB=0.935669;MQB=0.535992;BQB=0.861043;MQ0F=0;
chr1    49      .       T       A       163     .       DP=52;VDB=0.00137749;SGB=-0.692562;RPB=0.524716;MQB=0.0867743;MQSB=0.556902;BQB=0
chr1    57      .       T       G       188     .       DP=62;VDB=0.00523897;SGB=-0.693021;RPB=0.365629;MQB=0.0237017;MQSB=0.326001;BQB=0
chr1    61      .       C       G       198     .       DP=63;VDB=0.00861296;SGB=-0.693097;RPB=0.288101;MQB=0.0229986;MQSB=0.276873;BQB=0
chr1    70      .       G       A       207     .       DP=65;VDB=0.0142339;SGB=-0.69311;RPB=0.599664;MQB=0.0301369;MQSB=0.256578;BQB=0.8
chr1    89      .       CAA     CA      225     .       INDEL;IDV=69;IMF=0.971831;DP=71;VDB=0.0387193;SGB=-0.693147;MQSB=0.213156;MQ0F=0;
chr1    159     .       GTTT    GTT     225     .       INDEL;IDV=50;IMF=0.909091;DP=55;VDB=0.0380275;SGB=-0.693147;MQSB=0.00670185;MQ0F=
chr1    241     .       GTTT    GTT     106     .       INDEL;IDV=46;IMF=0.821429;DP=56;VDB=0.0020015;SGB=-0.616816;MQSB=1;MQ0F=0;AC=2;AN
chr1    281     .       GGA     G       225     .       INDEL;IDV=45;IMF=0.865385;DP=52;VDB=2.57575e-05;SGB=-0.686358;MQSB=0.99778;MQ0F=0
chr1    18398   .       CTTTTTTTTTTT    CTTTTTTTTTT     10.7919 .       INDEL;IDV=8;IMF=0.615385;DP=13;VDB=0.558106;SGB=-0.662043;MQSB=1;
chr1    99419   .       ATTTTTTT        ATTTTTT 71      .       INDEL;IDV=12;IMF=0.923077;DP=13;VDB=0.172273;SGB=-0.680642;MQSB=1;MQ0F=0;
chr1    127009  .       ATTTTT  ATTTT   161     .       INDEL;IDV=22;IMF=0.846154;DP=26;VDB=0.0858281;SGB=-0.692562;MQSB=0.991121;MQ0F=0;
chr1    148648  .       TGGGGGG TGGGGG  92      .       INDEL;IDV=29;IMF=0.805556;DP=36;VDB=0.0933396;SGB=-0.69312;MQSB=0.133333;MQ0F=0;A
chr1    165672  .       TAAAAAAA        TAAAAAAAA       61      .       INDEL;IDV=9;IMF=1;DP=9;VDB=0.00438619;SGB=-0.662043;MQSB=0.974597
chr1    168096  .       TAAAAAAAA       TAAAAAAAAA      22.4391 .       INDEL;IDV=4;IMF=0.666667;DP=6;VDB=0.74324;SGB=-0.590765;MQSB=1;MQ
chr1    176519  .       A       G       54      .       DP=156;VDB=1.1049e-20;SGB=-0.693145;RPB=1.36185e-12;MQB=3.39625e-17;MQSB=5.00093e
```

以#开头的注释部分

主体部分

1：参考序列名称　　4：参考序列碱基　　7：位点是否需要过滤

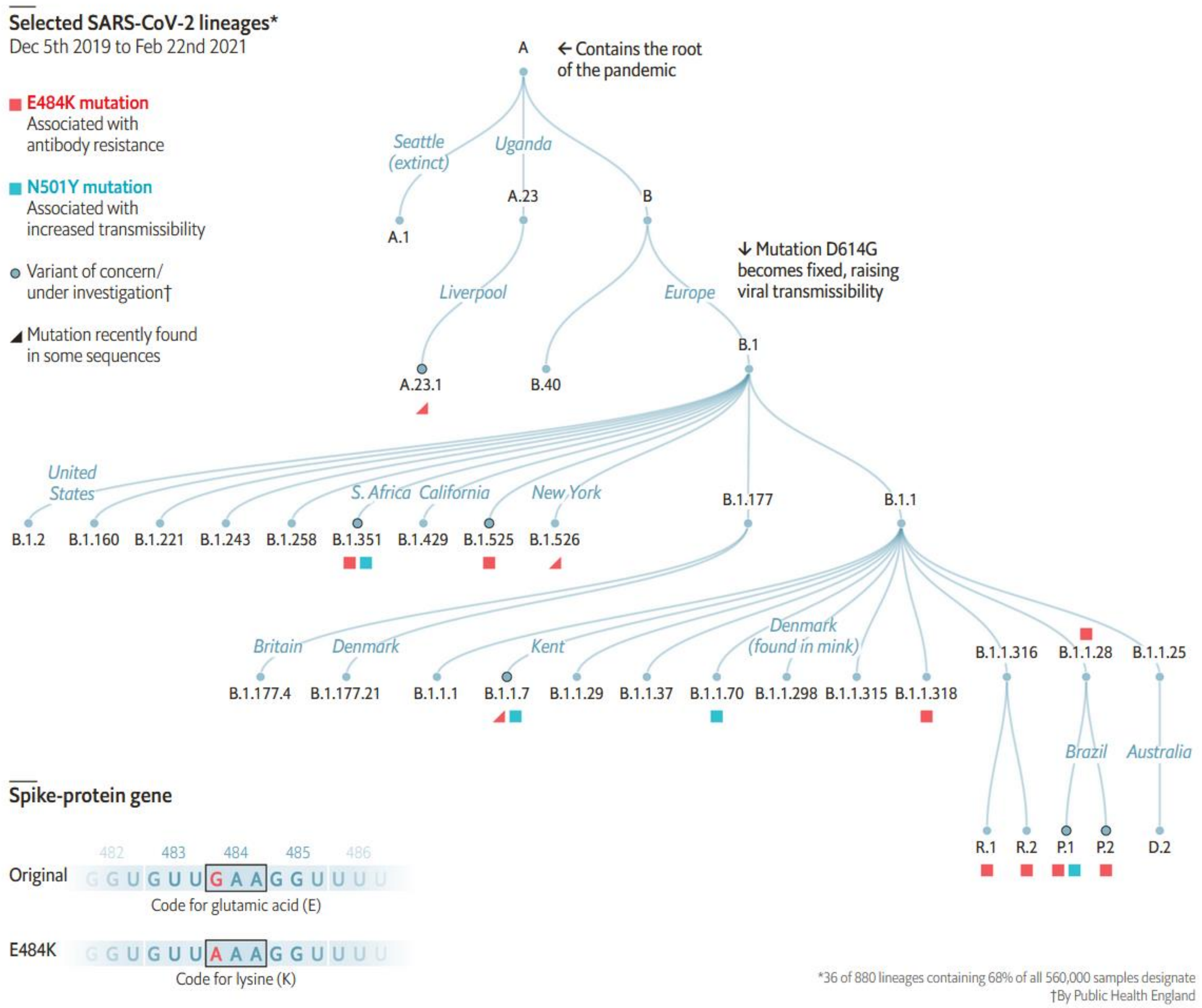
2：变异所在位置　　5：目的碱基　　8：变异相关信息

3：变异点ID　　6：变异质量值　　9：变异格式

SARS-CoV-2基因组重测序分析

Lineages of SARS-CoV-2

- WuHan-Hu-1
- UK, South African, Brazilian, and India variants
- The World Health Organization (WHO) has designated SARS-CoV-2 variants Alpha(B.1.1.7), Beta(B.1.351), Gamma(P.1), and Delta(B.1.617.2) as Variants of Concern (VOC), among which Delta variant with remarkable transmission and immune escape ability has attracted great attentions.



Selected SARS-CoV-2 lineages*
Dec 5th 2019 to Feb 22nd 2021

🟥 **E484K mutation** Associated with antibody resistance

🟦 **N501Y mutation** Associated with increased transmissibility

○ Variant of concern/ under investigation†

◢ Mutation recently found in some sequences

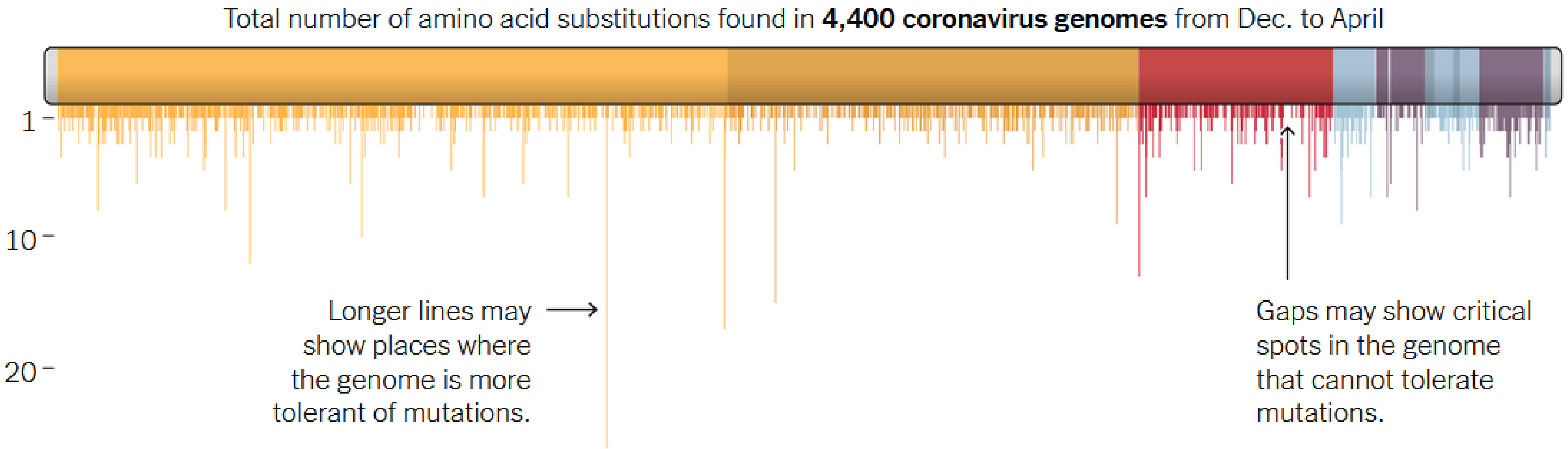
Spike-protein gene

482
483
484
485
486
Original
G G U G U U G A A G G U U U U
Code for glutamic acid (E)
E484K
G G U G U U A A A G G U U U U
Code for lysine (K)

*36 of 880 lineages containing 68% of all 560,000 samples designate
†By Public Health England

Mutations of SARS-CoV-2

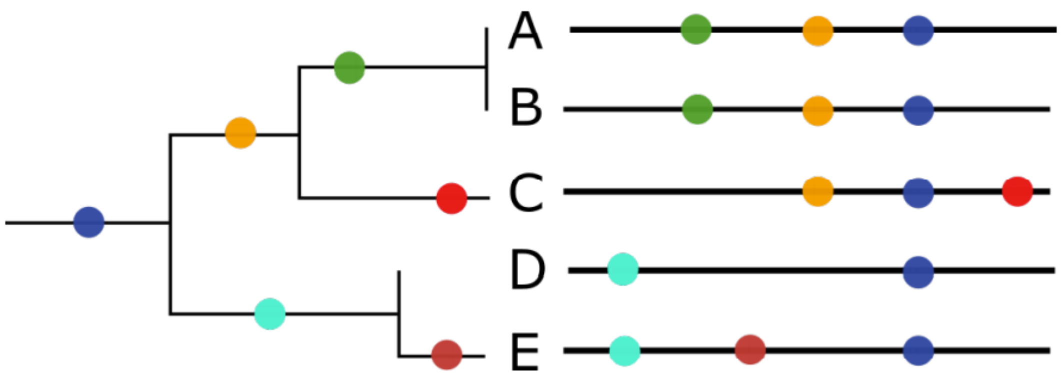
基因组的保守区与易变区

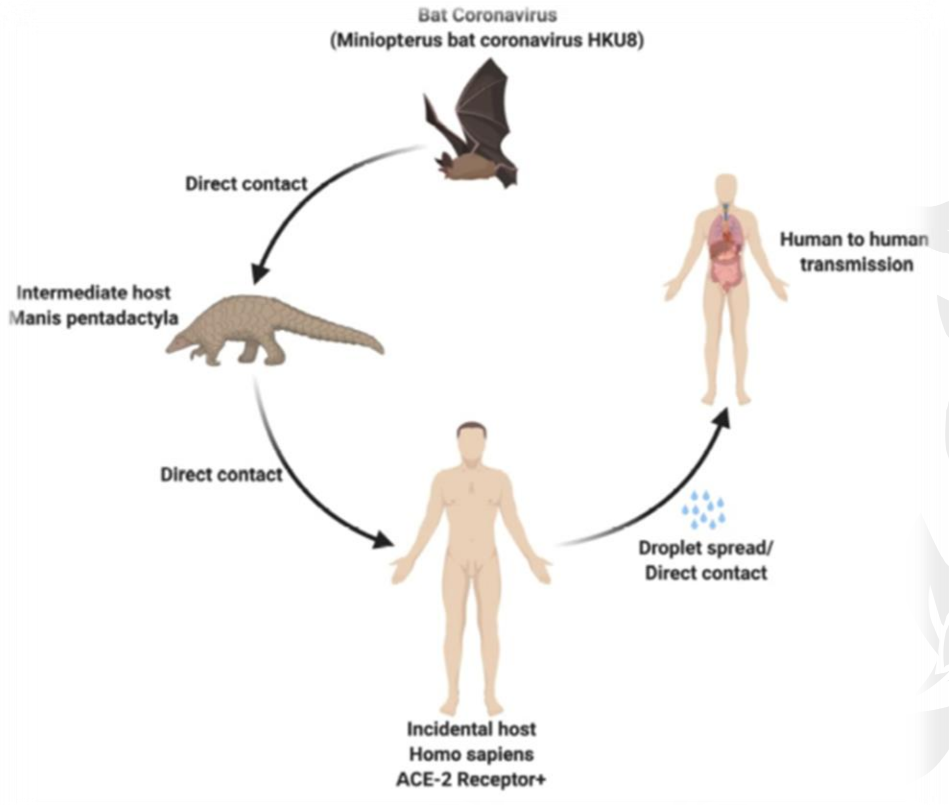


哪个区适合适作为抗病毒药物或疫苗的靶标？

Origin of SARS-CoV-2 (病毒溯源)

- Mutations in the genome produce a fingerprint that can be used to infer ancestral relationships (phylogeny).
- Zoonosis, jumped from animals to humans





ansmission Cycle of SARS CoV 2. Contributed by Rohan Bir Singh, MD; Made with Biorender.com

目录
CONTENT

01
软件与数据准备

02
原始数据质控

03
短序列回帖

04
比对可视化

1. 软件与数据准备

1.1 软件工具的准备

(1)操作系统: Linux/macOS

(2)测序序列质量控制软件: FastQC、Trimmomatic

(3)序列比对工具: BWA，包括BWA index、BWA mem、BWA aln、BWA sample

(4)序列分析工具: Samtools，包括Samtools view、Samtools sort、Samtools mpileup、 bcftools call、bcftools consensus

1. 软件与数据准备

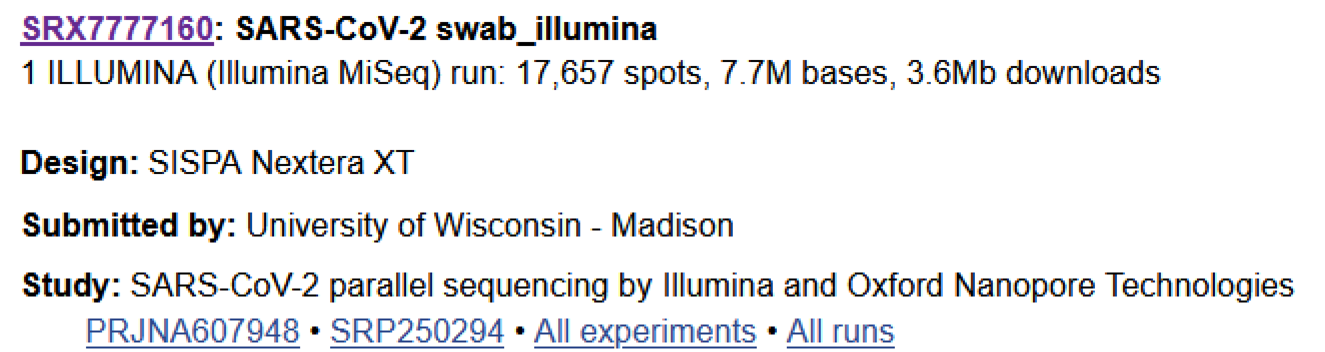
1.2 数据的准备

(1)参考序列

- SARS-CoV-2 Genome (NC_045512.2)

(2)测序数据(Illumina PE)

- NCBI accession number: SRR11140750
- A clinical swab obtained from a patient in Madison, WI, USA
- Data released on 21st Feb 2020.



SRX7777160: SARS-CoV-2 swab_illumina

1 ILLUMINA (Illumina MiSeq) run: 17,657 spots, 7.7M bases, 3.6Mb downloads

Design: SISPA Nextera XT

Submitted by: University of Wisconsin - Madison

Study: SARS-CoV-2 parallel sequencing by Illumina and Oxford Nanopore Technologies

PRJNA607948 • SRP250294 • All experiments • All runs

2. 原始数据质控

2.1 FastQC 用于测序数据概况的分析

```
$mkdir fastqc_out
$fastqc SRR11140750.fastq -o ./fastqc_out/
```

-o --outdir FastQC生成的报告文件的储存路径；

--extract 默认情况下生成的报告打包成1个压缩文件，设置该参数是无需打包；

-t –threads 选择程序运行的线程数，每个线程占用250MB内存，越多程序运行越快

2. 原始数据质控

-c --contaminants 污染物选项，输入的是一个文件，文件格式是 Name [Tab] Sequence，里面是可能的污染序列，如果有这个选项，FastQC会在计算时评估污染的情况，并在统计的时候进行分析；

-a –adapters 也是输入一个文件，文件格式是 Name [Tab] Sequence，储存的是测序的adpater序列信息；如果不输入，当前版本的 FastQC 会按照通用引物来评估序列是否有 adapter的残留。

2. 原始数据质控

2.2 除去接头和低质量序列: Trimmomatic

```
$java -jar Trimmomatic-0.39/trimmomatic-0.39.jar SE -phred33
SRR11140750.fastq  SRR11140750.out.fq
ILLUMINACLIP:Trimmomatic-0.39/adapters/TruSeq3-
SE.fa:2:30:10  SLIDINGWINDOW:5:20 LEADING:5
TRAILING:5 MINLEN:50
```

2. 原始数据质控

-ILLUMINACLIP: 设置切除接头序列的参数

-SLIDINGWINDOW: 设置滑动窗口长度的参数，

-LEADING: 设置是否切除read开头碱基的质量阈值；

-TRAILING: 设置是否切除read末尾碱基的质量阈值；

-MINLEN: 设置read被切除后至少需要保留的长度；如果低于该长度，序列将被丢弃。

3.短序列回帖

```
#参考序列NC_045512.2文件重命名为covid19.fasta
$cp NC_045512.2.fasta covid19.fasta
#Index the reference genome for use with BWA
$bwa index covid19.fasta
#Align the Illumina reads
$bwa mem covid19.fasta SRR11140750.fastq.gz > SRR11140750.sam
#Coordinate sort SAM file, and output to BAM
$samtools sort -o SRR11140750.bam SRR11140750.sam
```

Can you figure out how to do the bwa mem and samtools sort commands in a pipeline so as to avoid writing the large intermediary SAM file?

4. BAM File Visualisation

```
#Generate index of the genome file
$samtools faidx covid19.fasta
#Index the BAM file
$samtools index SRR11140750.bam
```

最终生成四个文件：

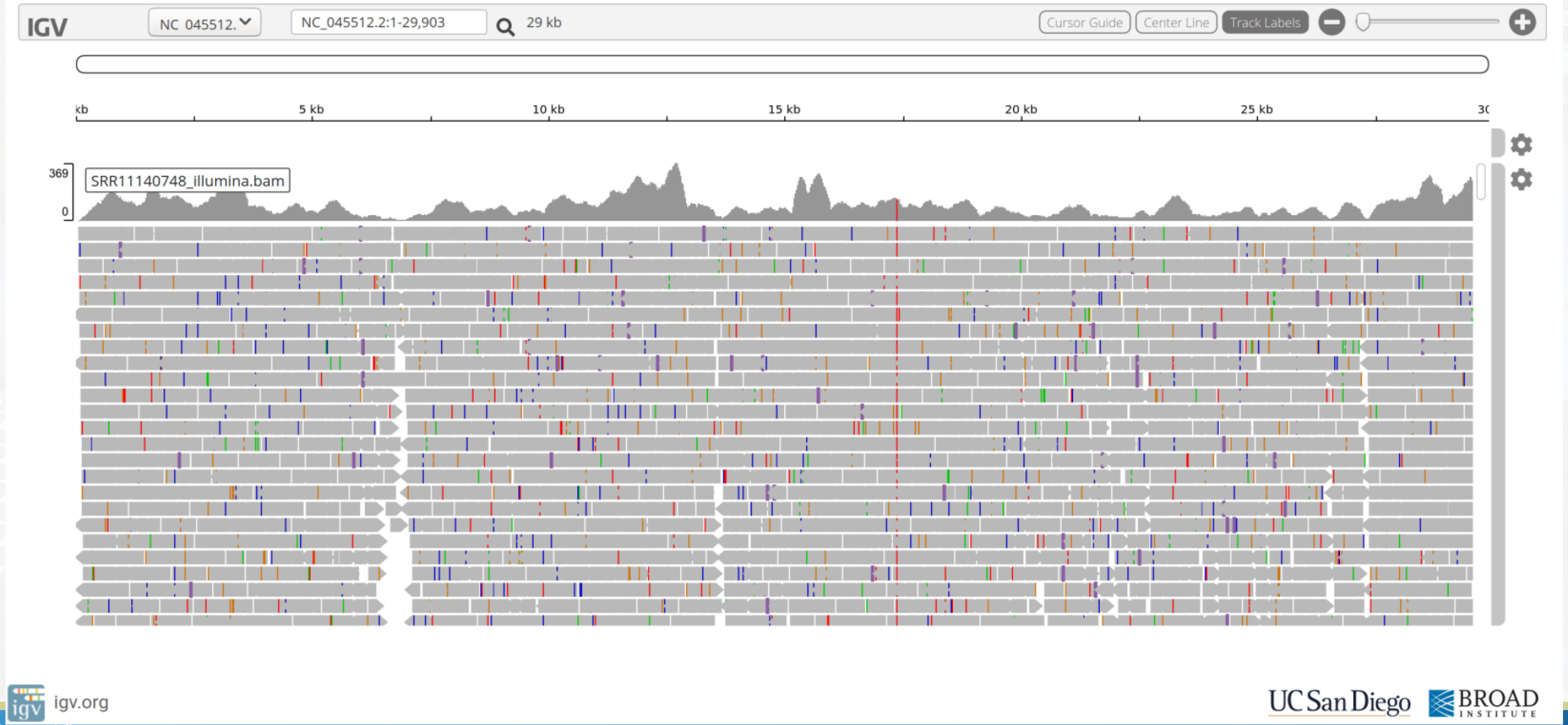
- COVID19.fasta
- COVID19.fasta.fai
- SRR11140750.bam
- SRR11140750.bam.bai

```
# View alignment with samtools
$samtools tview ./SRR11140750.bam --reference ./covid19.fasta
```

#Press “q” to quit view

4.BAM file visualization with IGV

- Run local IGV, or visit IGV-web (https://igv.org/app/).
- Load the genome from a “Local File ...” by selecting both the COVID-19.fasta and COVID-19.fasta.fai files.
- Load a “Track” from a “Local File ...” by selecting both the SRR11140750.bam and SRR11140750.bam.bai files.



The stacked grey arrows represent the reads aligned to the SARS-CoV-2 reference genome. What do the coloured vertical bars within the reads indicate?

作 业

- 使用本课提供的SARS-CoV-2测序数据(SRR11140750.fastq)，利用BWA、Samtools、IGV等工具进行reads mapping与可视化等。

ZHEJIANG GONGSHANG UNIVERSITY
1911
浙江工商大学

谢谢观看