BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

Lecture 15: Genomics data

24 October, 2025| Lecturer: Yu LI | Covered pages: 1-43, 71

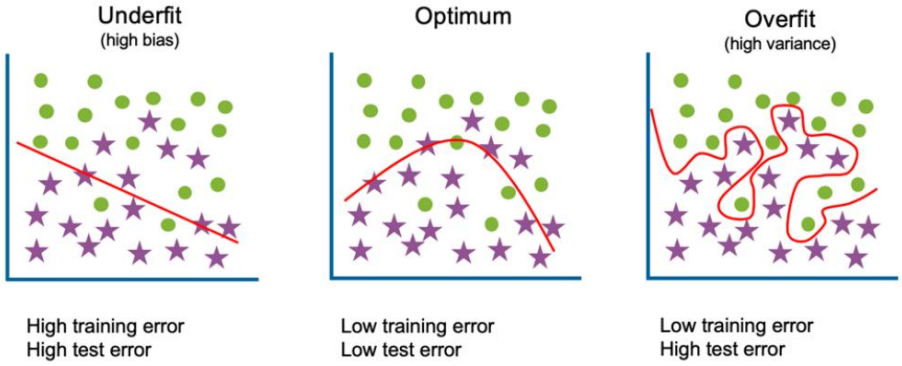
Scriber: Chan Wai

Recap:

Linear regression problems:

Underfitting: when facing complicated problems, the model capacity of simple linear regression is insufficient to capture complex relationships among variables. In practice, we need to first overfit the data.

Overfitting: Statistically, it is the production of an analysis that corresponds too closely to a particular dataset and may fail to fit additional data or predict future observations reliably. In machine learning terms, it occurs when a method is more complex than the problem, performing well on training data but poorly on testing data.



Multi-Omics

The multi-omics approach includes data from the genome, epigenome, transcriptome, proteome, metabolome, and phenome. The core multi-omics data analysis techniques are the same as other: sequence alignment and comparison, dimension reduction and visualization, and clustering & classification.

Statistical analysis for differential gene expression analysis

To discover quantitative changes in expression levels between experimental groups.

- ✓ **T-test**: Its purpose is to find a significant difference between 2 sets of data.
  - ➢ Calculate a test statistic based on the mean and variance of the data
  - ➢ The test statistic follows a Student's t-distribution
  - ➢ Generate a **p-value**: the probability that the result occurred by chance

- ➢ The smaller the p-value, the more confident we are in the result
- ➢ Standard threshold: p-value < 0.05 indicates a significant difference

✓ **Fisher's Exact Test**: used for analyzing contingency tables.

The p-value can be calculated exactly from the table, unlike the t-test where we calculate a t-value and then derive the p-value from a distribution.

Application: Used for gene enrichment analysis and testing associations between pathways and diseases.

1. Genes involved in KEGG biological pathway
2. Genes not involved in KEGG biological pathway
3. Genes related to type-2 diabetes
4. Genes not related to type-2 diabetes

Contingency Table:

	In gene set	Not in gene set	Total
In pathway	100 (a)	9000 (b)	9100
Not in pathway	113 (c)	11000 (d)	11113
Total	213	20000	20213

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

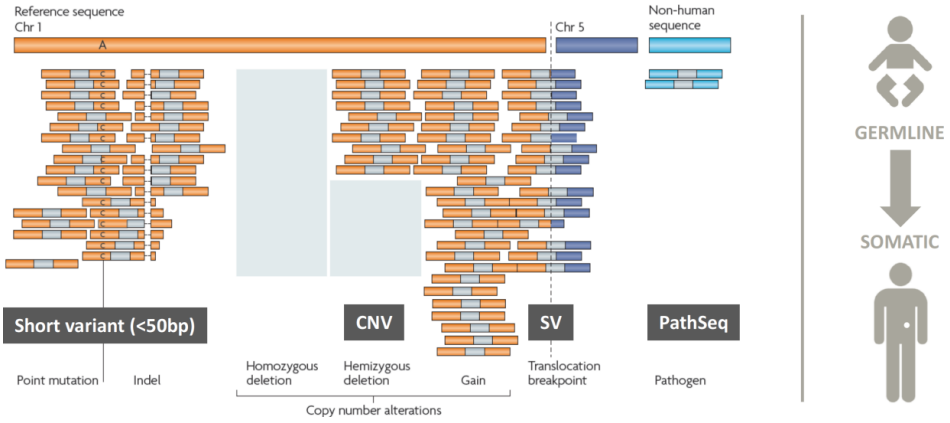
$p = 0.5802 > 0.05$. This pathway is not related to type-II diabetes

Lec15: Genome data analysis

Variant Calling

Variant calling is the computational process of identifying genetic differences (variants) between sequenced DNA and a reference genome. These variants can be:

- **Point mutations** (single nucleotide changes)
- **Indels** (insertions/deletions, <50bp)
- **Copy Number Variations (CNV**, > 1000bp**)**
- **Structural Variants (SV)** including translocations (alter up to millions bp)
- **Germline** (inherited) and **somatic** (cancer-specific) variants



Biological significance of variants:

- Any two humans share 99.5% of DNA; we can efficiently describe a genome relative to a reference using variants
- Genetic differences among people lead to differences in disease risk and response to treatment
- Genetic variation is used to find genes and variants that contribute to disease
- Cancer involves genetic variants at multiple levels

Pipeline for discovering genetic variants:

Step 1: Library Preparation & Sequencing

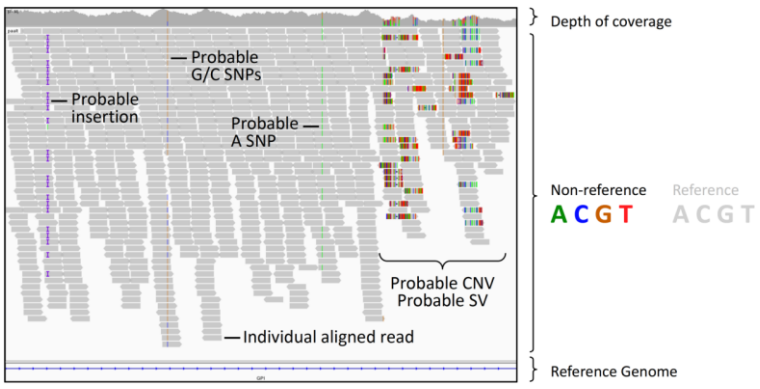
I. Slide each read along the reference genome, calculate the difference by dynamic programming.

```
T A A T G C C A T G G A T G     T A A T G C G A T G G A T G
        C C A                             C C A
2 3 3 3 2 0 2 3 3 2 3 3   VS   2 3 3 3 2 1 3 3 3 2 3 3
```

No variant (exact match) VS variant (point mutation)

II. Distinguish between actual variation (real change) & errors (artifacts):
   i. PCR artifacts (amplification of errors)
   ii. Sequencing errors (base calling)
   iii. Alignment errors (misalignment, mis-gapped alignments)
   iv. Variant calling errors (low depth of coverage, few samples)
   v. Genotyping errors (poor annotation)

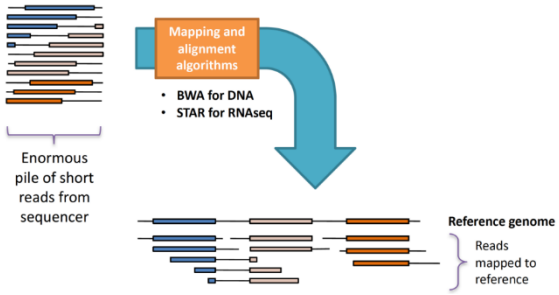
Variant visualization in genome browser:



Step 2: Data pre-processing

I. Map the enormous pile of short reads produced by the sequencer to the reference genome by mapping and alignment algorithms:

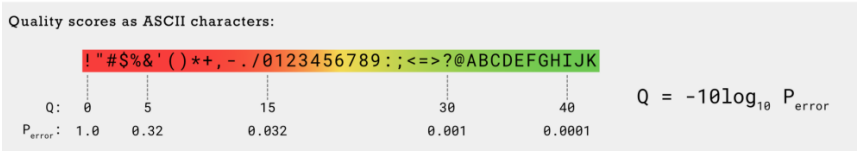
- BWA for DNA
- STAR for RNAseq



Input format: FASTQ

FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFBFFFFFFFFFF7FFFF<FF
```



Output format: Sequence/Binary Alignment Map (SAM/BAM)

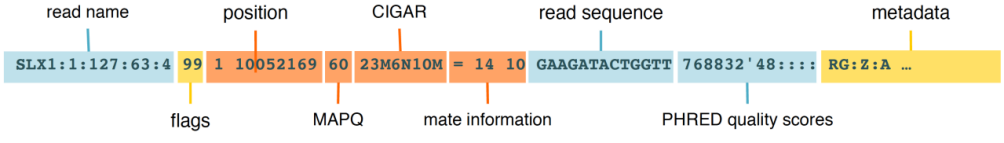
```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C       BBDCCDDCCDDDDCDDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDDCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
        AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G       DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
        AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C       DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
        AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

- Header information: MAPQ is quality

HEADER lines starting with @ symbol describing various metadata for *all* reads

```
@HD VN:1.6 SO:coordinate ——— BAM header line
@SQ SN:seq1 LN:394893 ——— Reference sequence dictionary entries
@SQ SN:seq2 LN:92783
@RG ID:A SM:SAMPLE_A ——— Read group(s)
```

RECORDS containing structured read information (1 line per read/record)



- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**
- Mate information points to the read from the other end of the molecule (other in a pair)

CIGAR summarizes alignment structure

CIGAR = Concise Idiosyncratic Gapped Alignment Report

```
read1  99  ref  2  30  1S3M1D2M1I1M  =  14  20  CATCTAG  …
```

```
RefPos:     1 2 3 4 5 6 7   8 9
Reference:  C C A T A C T - G A
Read:         C A T - C T A G

POS: 2
CIGAR:          3M1D2M1I1M
```

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

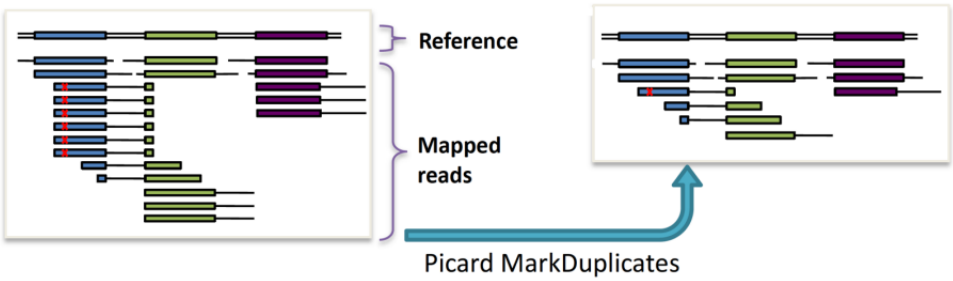
Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

- "Consumes query" and "consumes reference" indicate whether the CIGAR operation causes the alignment to step along the query sequence and the reference sequence respectively.
- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

II. Mark duplicates to mitigate duplication artifacts.

Duplicates = **non-independent measurements** of a sequence fragment

-> Must be removed to assess support for alleles correctly

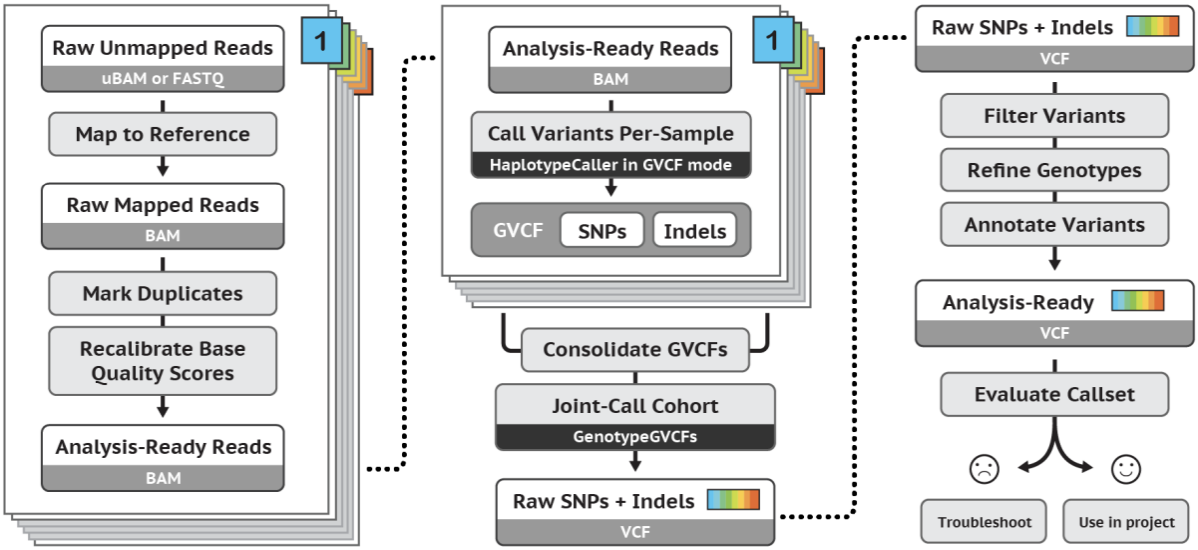


✖ = error propagated in duplicates

Cause of duplication:

i. Library Duplicates, caused by PCR
ii. Optical Duplicates, occur during sequencing

Step 3: Variant Calling



Variant Call Format (VCF):

HEADER

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```

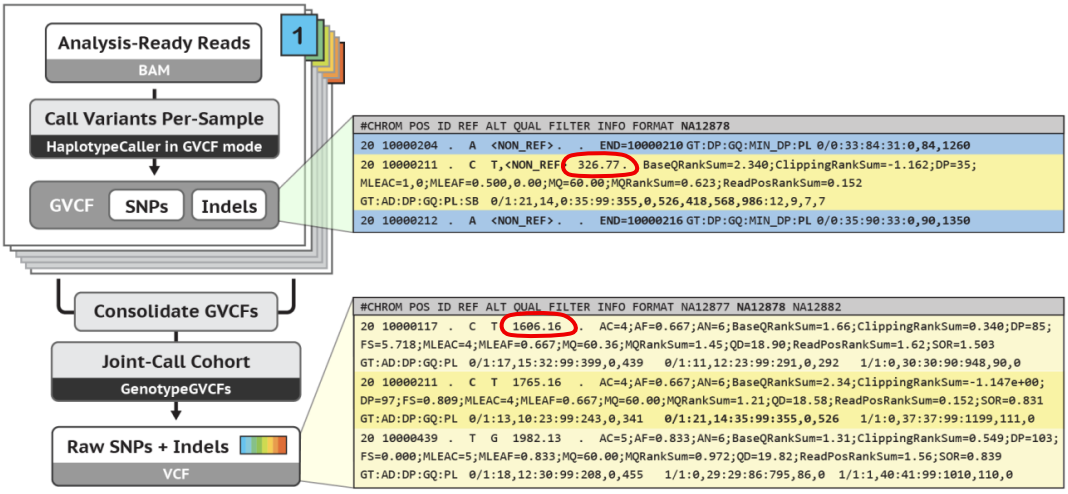
RECORDS

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	DP=14;AF=0.5	GT:GQ:DP	0/0:48:1	1/0:48:8	1/1:43:5
20	1230237	.	T	.	47	PASS	DP=13	GT:GQ:DP	0/0:54:7	0/0:48:4	0/0:61:2
20	1234567	.	GT	G	50	PASS	DP=9	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Joint analysis increases variants & empowers discovery, as family or population data add valuable information:

- Rarity of variants
- *de novo* mutation
- Ethnic background

From per-sample GVCFs to final multi-sample VCF: compare quality across samples



Summarization of Variant Calling:

- The pipeline
  - ✓ A concrete tool you can use in the future
  - ✓ You know what you are expecting from each step. And which file you are looking for

- The file format
  - ✓ We talked about reads a lot of time. What are they in the real analysis?
  - ✓ It's for practice. We want to avoid the case that you learn a lot but you still cannot resolve reallife problems
  - ✓ You know what to input to a specific step. If you get an error, you know what to change

- Trouble-shooting
  - ✓ For example, in real-life, you have a nice BAM/SAM file, but your VCF file is empty. Is it because of programming bugs, file formats, or no variants?
  - ✓ Hopefully, our introduction to the pipeline will be useful
  - ✓ Usefulness is more important than exams

Potential Projects-4,5,6

4. Genetic variant calling pipeline

5. Epigenetic data processing pipeline

6. Gene fusion detection pipeline