BMEG3105 | Data Analytics for Personalized Genomics and Precision Medicine | Lecture 17
Lecturer: Yu Li (李煜) | 02-NOV-22

## Genomics Data Analysis

Variant Calling Pipeline

- Reasons for the steps
- File interpretation
- Factors affect variant calling

Gene Fusion

- Definition
- RNA-seq can detect it

GWAS

- P-value correction

Epigenetics

- Gene expression regulation: structure and environment
- Data analytics pipeline

## Why do we care about variants?

- 3.2 billion sites in the human genome
    - Any two humans share 99.5% DNA
    - We can efficiently describe a genome with relation to a reference
- 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: genetic variants at multiple levels

## Sequence Mapping Recap

- **TAATGCCATGGATD | TAA, CCA, GAT, GCC, CCA, ATG**

Slide each read along the genome, calculate the difference

- Each time, we may use dynamic programming to calculate the difference
- For simplicity, we would not use it for now

**T A A T G C C A T G G A T G**

**C C A**

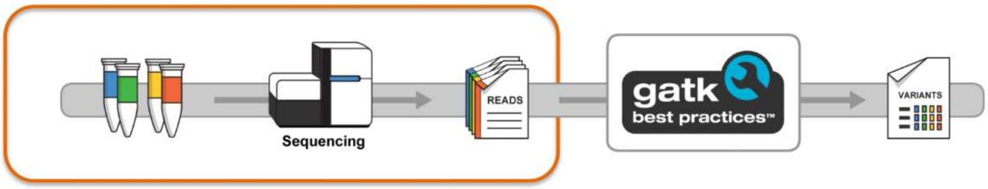
**2 3 3 3 2 0 2 3 3 2 3 3**

**T A A T G C G A T G G A T G**

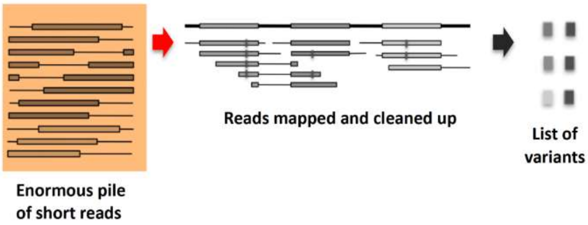
**C C A**

**2 3 3 3 2 1 3 3 3 2 3 3**

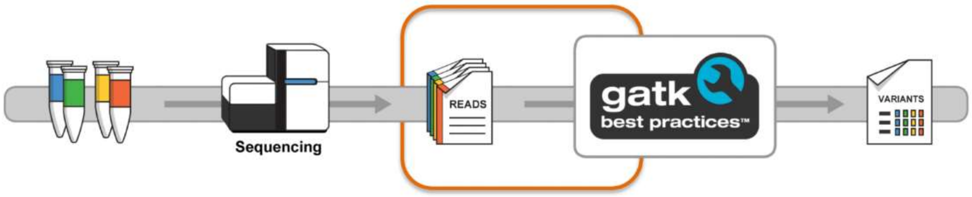
## How to Discover the Genetic Variants?



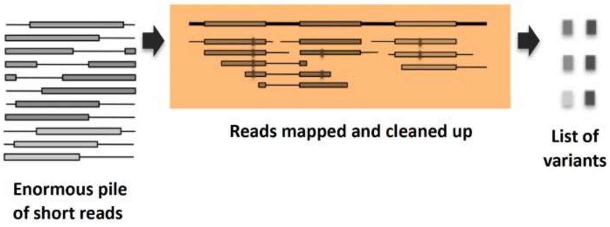
1. Library preparation
2. Sequencing



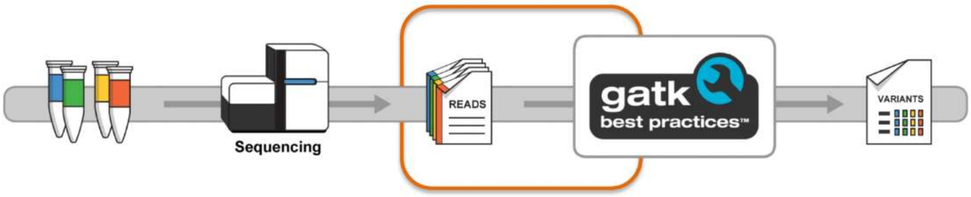
## Data Pre-Processing Step



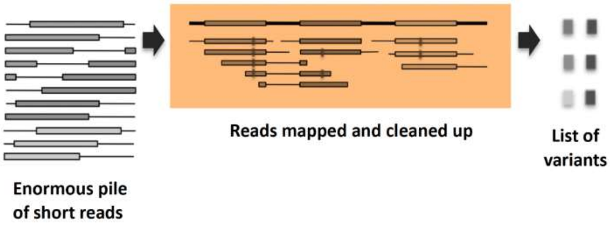
1. Mapping
2. Marking duplicates
3. Base recalibration



## Variant Calling



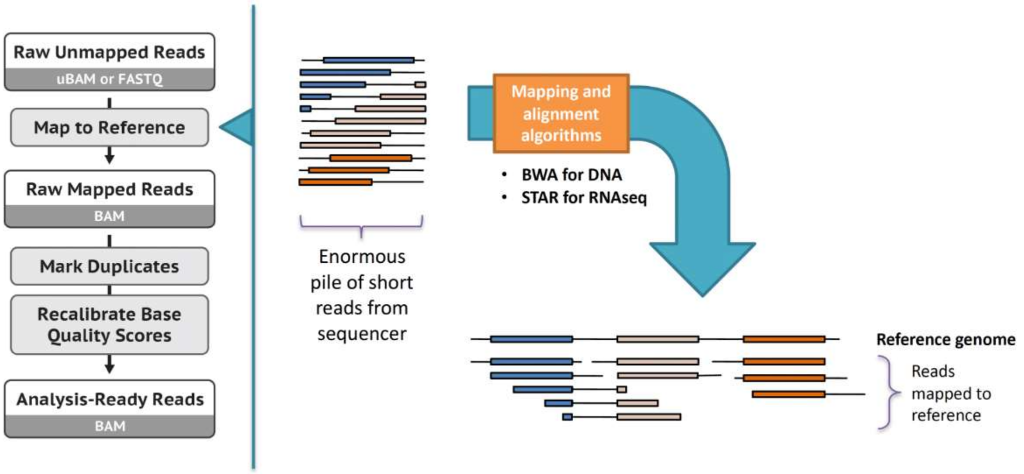
1. Mapping
2. Marking duplicates
3. Base recalibration



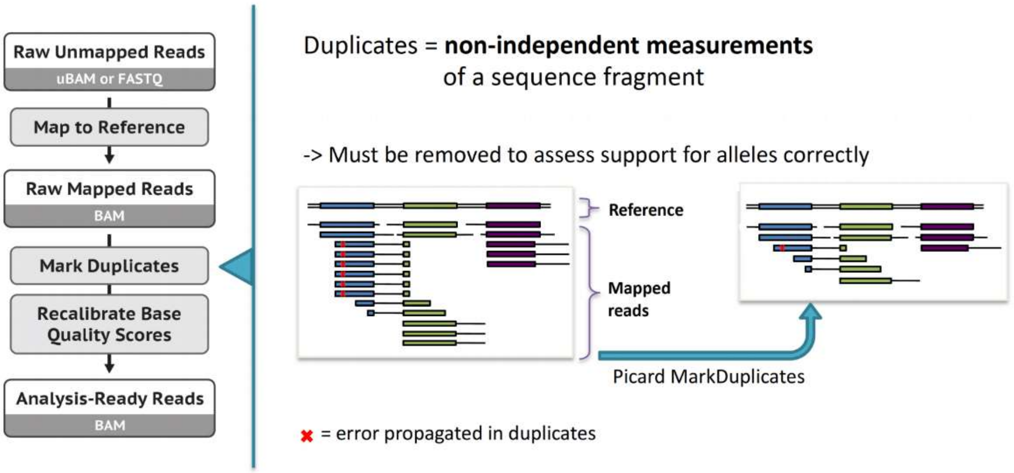
## Variants vs Errors

- Must distinguish between actual variation (real change) and errors (artifacts) introduced into the analysis
- Errors can creep in on various levels:
  - PCR artifacts (amplification of errors)
  - Sequencing (errors in base calling)
  - Alignment (misalignment, mis-gapped alignment)
  - Variant calling (low depth of coverage, few samples)
  - Genotyping (poor annotation)

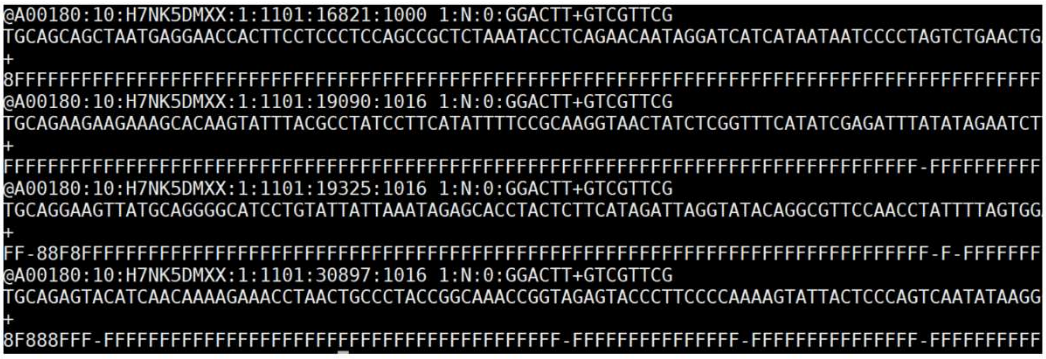
## Step 1: Map the Reads Produced by the Sequencer to the Reference



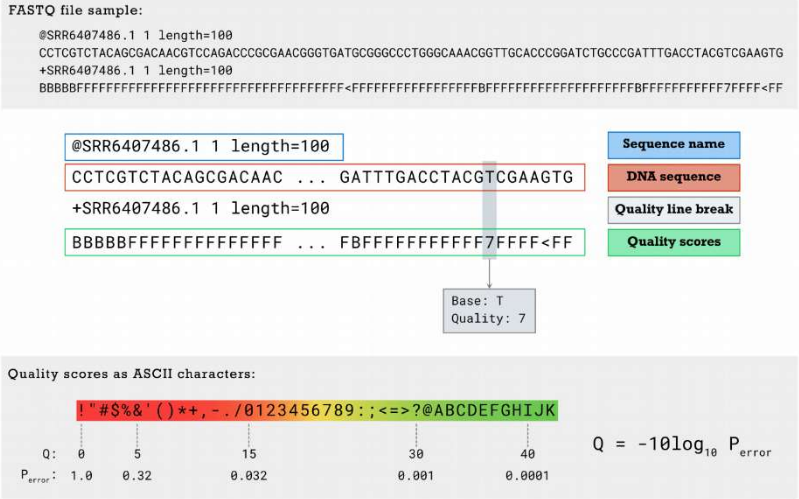
## Step 2: Mark Duplicates to Mitigate Duplication Artifacts


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
Duplicates = non-independent measurements of a sequence fragment
-> Must be removed to assess support for alleles correctly
Reference
Mapped reads
Picard MarkDuplicates
✖ = error propagated in duplicates

## Input Format: FASTQ



```
@A00180:10:H7NK5DMXX:1:1101:16821:1000 1:N:0:GGACTT+GTCGTTCG
TGCAGCAGCTAATGAGGAACCACTTCCTCCCTCCAGCCGCTCTAAATACCTCAGAACAATAGGATCATCATAATAATCCCCTAGTCTGAACTG
+
8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19090:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAAGAAGAAAGCACAAGTATTTACGCCTATCCTTCATATTTTCCGCAAGGTAACTATCTCGGTTTCATATCGAGATTTATATAGAATCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19325:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGGAAGTTATGCAGGGGCATCCTGTATTATTAAATAGAGCACCTACTCTTCATAGATTAGGTATACAGGCGTTCCAACCTATTTTAGTGG
+
FF-88F8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-F-FFFFFFF
@A00180:10:H7NK5DMXX:1:1101:30897:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAGTACATCAACAAAAGAAACCTAACTGCCCTACCGGCAAACCGGTAGAGTACCCTTCCCCAAAAGTATTACTCCCAGTCAATATAAGG
+
8F888FFF-FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFFFFFFFF-FFFFFFFFFFFFFFFF-FFFFFFFFFFF
```



FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFF7FFFF<FF
```

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAAC ... GATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFF ... FBFFFFFFFFFFF7FFFF<FF
```

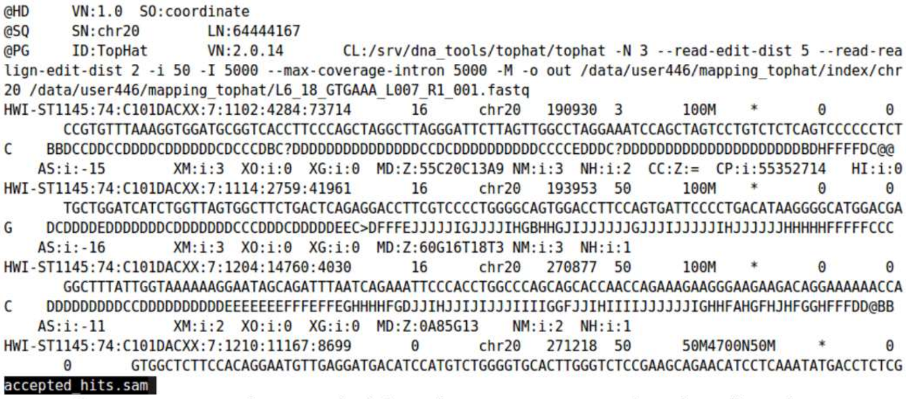
Quality scores as ASCII characters:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJK
```

| Q: | 0 | 5 | 15 | 30 | 40 |
|---|---|---|---|---|---|
| $P_{error}$: | 1.0 | 0.32 | 0.032 | 0.001 | 0.0001 |

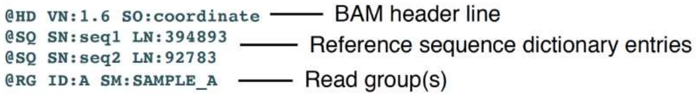
$$Q = -10\log_{10} P_{error}$$

## Output Format: Sequence/Binary Alignment Map (SAM/BAM)



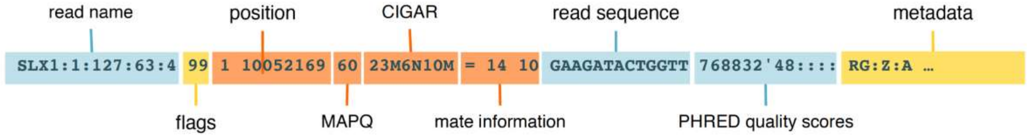
```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C       BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?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>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    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** lines starting with @ symbol describing various metadata for *all* reads



```
@HD VN:1.6 SO:coordinate
@SQ SN:seq1 LN:394893
@SQ SN:seq2 LN:92783
@RG ID:A SM:SAMPLE_A
```

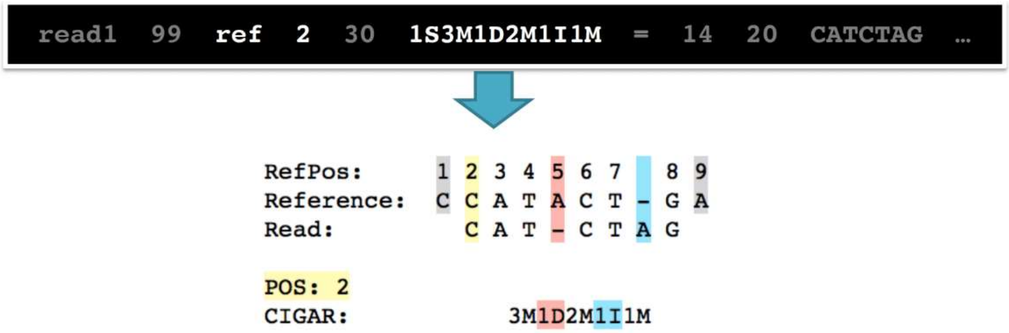
**RECORDS** containing structured read information (1 line per read/record)



```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: RG:Z:A …
```

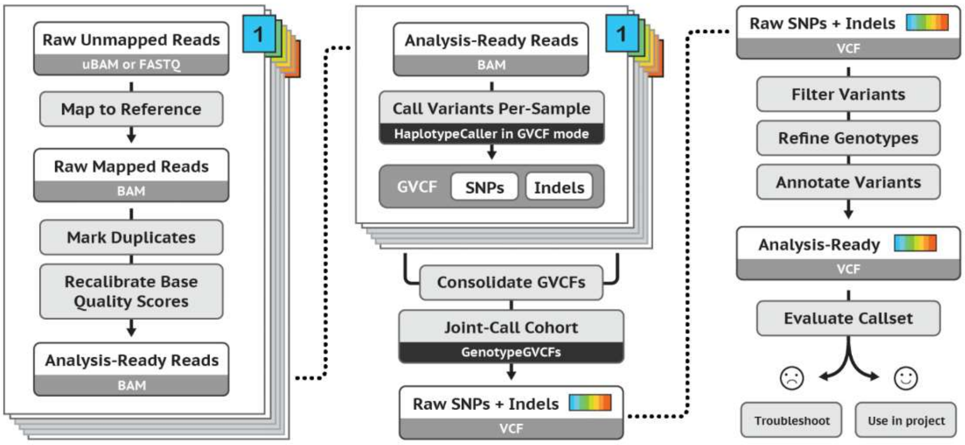
- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**
- Mate information points to the read from the other end of the molecule (other in a pair)

## CIGAR (Concise Idiosyncratic Gapped Alignment Report) Summarizes Alignment Structure

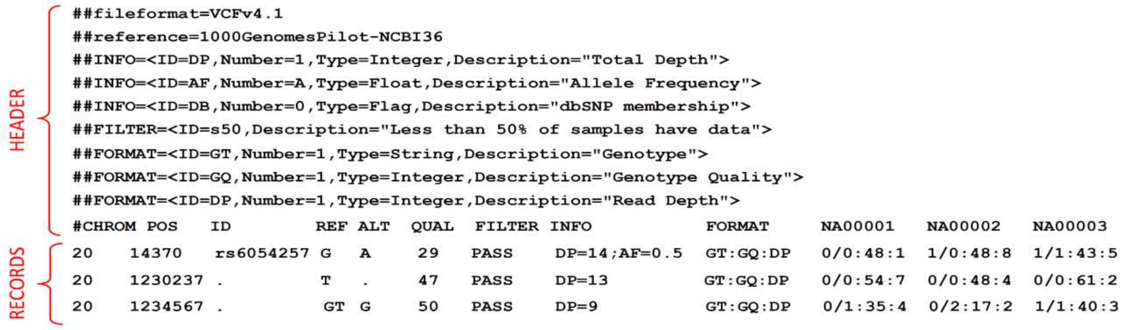


```
read1  99  ref  2  30  1S3M1D2M1I1M  =  14  20  CATCTAG  …

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
```

## Variant Calling in More Detail

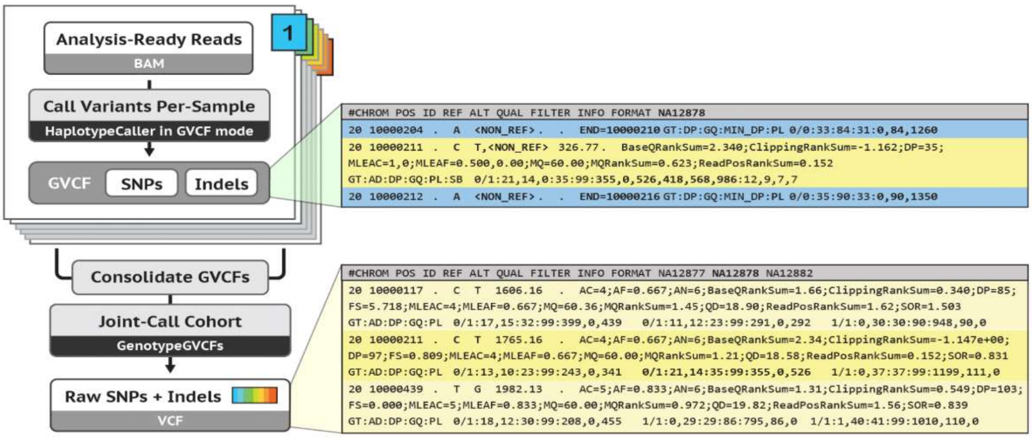


## Variant Call Format (VCF)


HEADER
RECORDS

```
##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">
#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
```

## From Per-Sample GVCFs to Final Multi-Sample VCF



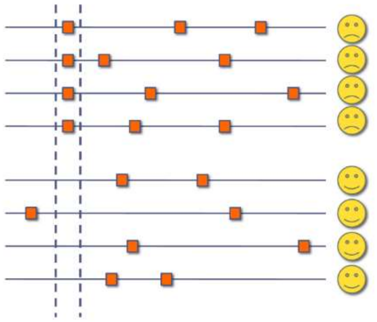
```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 .  A  <NON_REF> .  .  END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 .  C  T,<NON_REF> 326.77 .  BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB  0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 .  A  <NON_REF> .  .  END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12877 NA12878 NA12882
20 10000117 .  C  T  1606.16  .  AC=4;AF=0.667;AN=6;BaseQRankSum=1.66;ClippingRankSum=0.340;DP=85;
FS=5.718;MLEAC=4;MLEAF=0.667;MQ=60.36;MQRankSum=1.45;QD=18.90;ReadPosRankSum=1.62;SOR=1.503
GT:AD:DP:GQ:PL  0/1:17,15:32:99:399,0,439   0/1:11,12:23:99:291,0,292   1/1:0,30:30:90:948,90,0
20 10000211 .  C  T  1765.16  .  AC=4;AF=0.667;AN=6;BaseQRankSum=2.34;ClippingRankSum=-1.147e+00;
DP=97;FS=0.809;MLEAC=4;MLEAF=0.667;MQ=60.00;MQRankSum=1.21;QD=18.58;ReadPosRankSum=0.152;SOR=0.831
GT:AD:DP:GQ:PL  0/1:13,10:23:99:243,0,341   0/1:21,14:35:99:355,0,526   1/1:0,37:37:99:1199,111,0
20 10000439 .  T  G  1982.13  .  AC=5;AF=0.833;AN=6;BaseQRankSum=1.31;ClippingRankSum=0.549;DP=103;
FS=0.000;MLEAC=5;MLEAF=0.833;MQ=60.00;MQRankSum=0.972;QD=19.82;ReadPosRankSum=1.56;SOR=0.839
GT:AD:DP:GQ:PL  0/1:18,12:30:99:208,0,455   1/1:0,29:29:86:795,86,0  1/1:1,40:41:99:1010,110,0
```

## Further Downstream Analysis

### Genome-Wide Association Studies (GWAS)

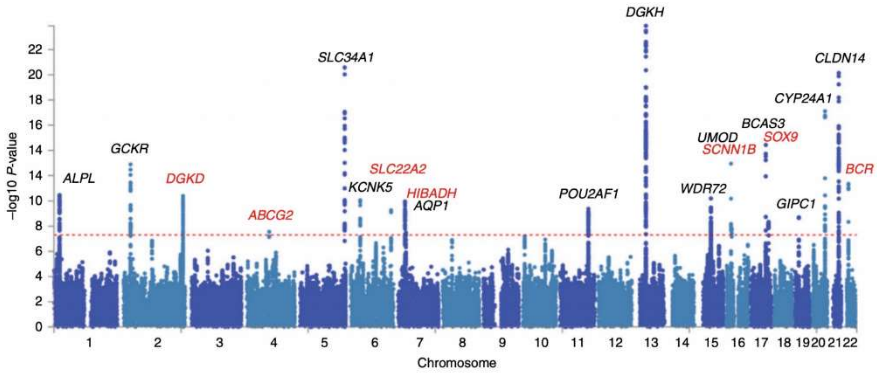
- Trying to determine whether specific variant(s) in many individuals can be associated with a trait (disease)



Spot the variant that is common amongst all affected but absent in all unaffected

The ideal case (for some rare Mendelian diseases)

In reality – 3.5 million SNPs



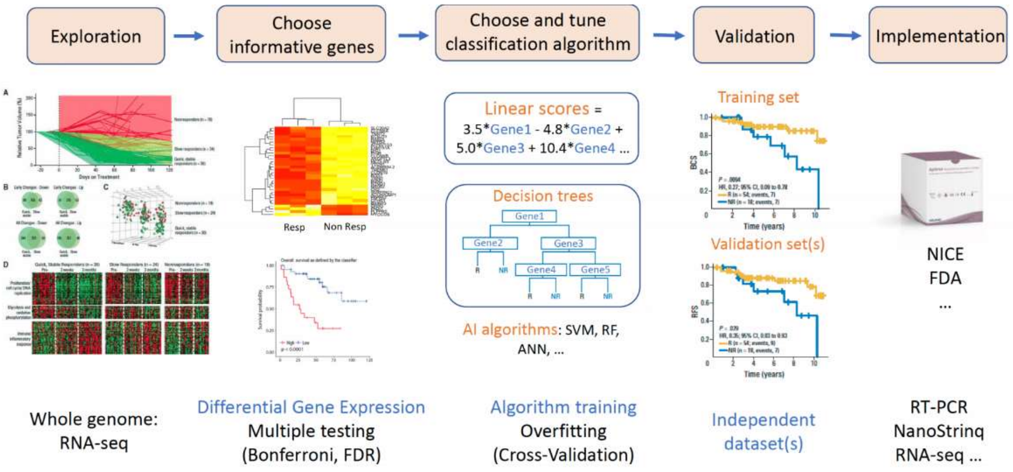
## Bonferroni Correction

Adjusted p-value = p-value / number of tests

Suppose we have 1 million SNPs to test

- Adjusted p-value = 0.05/1,000,000
  = $5*10^{-8}$

## RNA-seq Data Analysis



Exploration → Choose informative genes → Choose and tune classification algorithm → Validation → Implementation

Linear scores = 3.5*Gene1 - 4.8*Gene2 + 5.0*Gene3 + 10.4*Gene4 ...

Decision trees

AI algorithms: SVM, RF, ANN, ...

Training set

Validation set(s)

NICE
FDA
...

Whole genome: RNA-seq

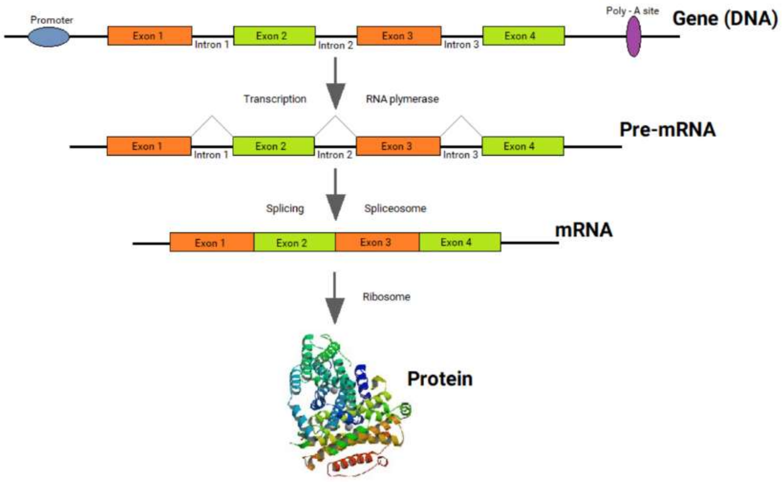
Differential Gene Expression
Multiple testing
(Bonferroni, FDR)

Algorithm training
Overfitting
(Cross-Validation)

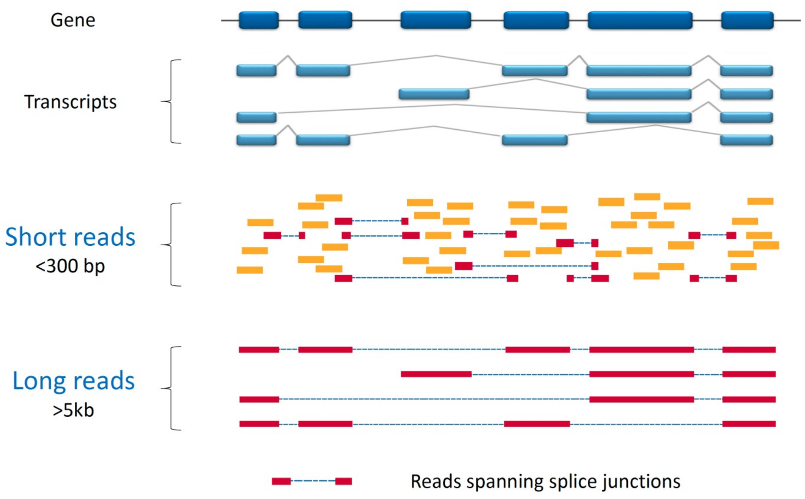
Independent dataset(s)

RT-PCR
NanoStrinq
RNA-seq ...

## Transcription, Splicing and Translation of a Eukaryotic Gene



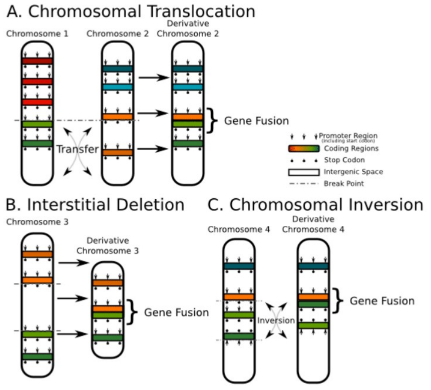
## Mapping Spanning Splice Junctions



The mapping algorithm should be modified slightly. But it’s helpful for identifying gene fusion.

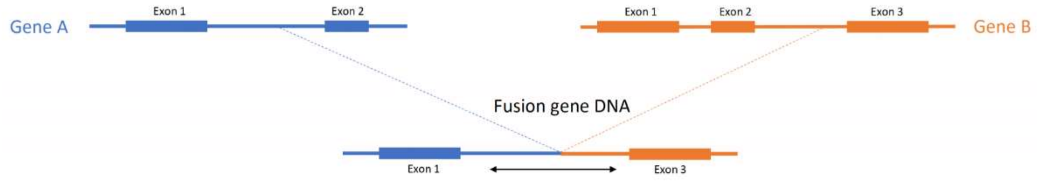
## What is Gene Fusion?

- The first fusion gene was described in cancer cells in the early 1980s
- Novel gene formed by fusion of two distinct wild type genes
- In cancer: produced by somatic genome rearrangements

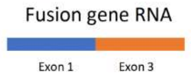


Gene fusion is a specific kind of structural variant related to cancer

## RNA-seq for Gene Fusion Detection

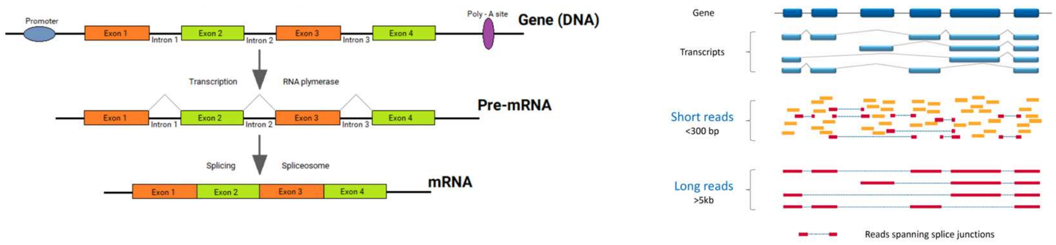


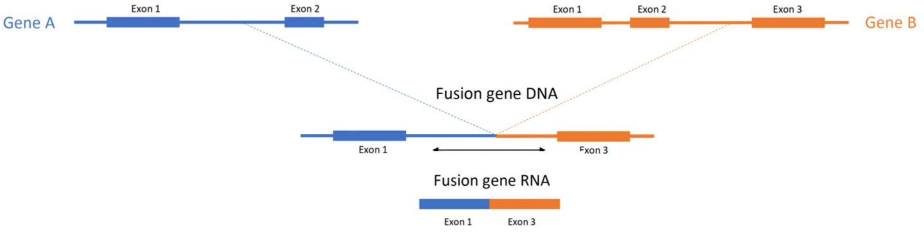
**Break-points are in introns**
**We need whole genome sequencing**
**Whole exome sequencing is not enough**



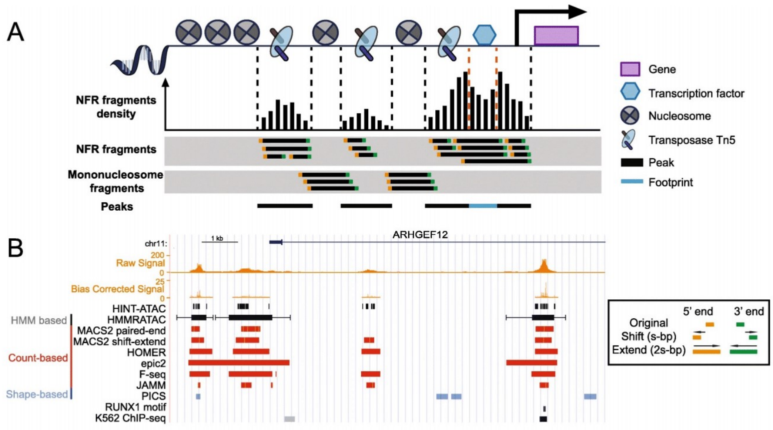
**Detecting fusion in RNA-seq requires much less sequencing than WGS, especially with long reads**

## Why can it be Detected by RNA-seq?





## Epigenetics – Peak Calling

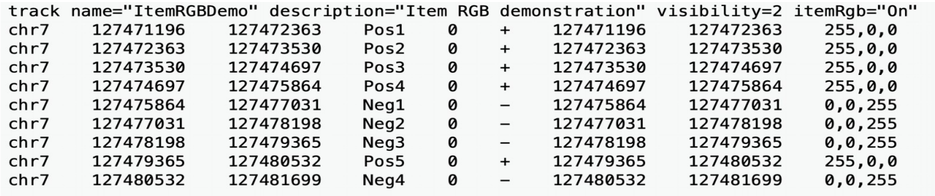


**Statistical testing: Peak shape VS random background**

## Peak Calling Output – BED file

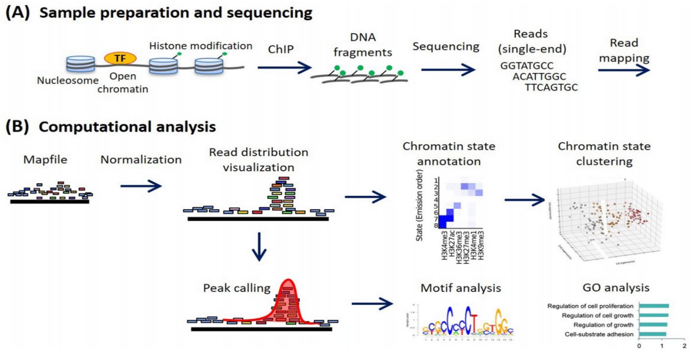
Browser Extensible Data (BED) format

- Chromosome
- Start
- End
- Label
- …

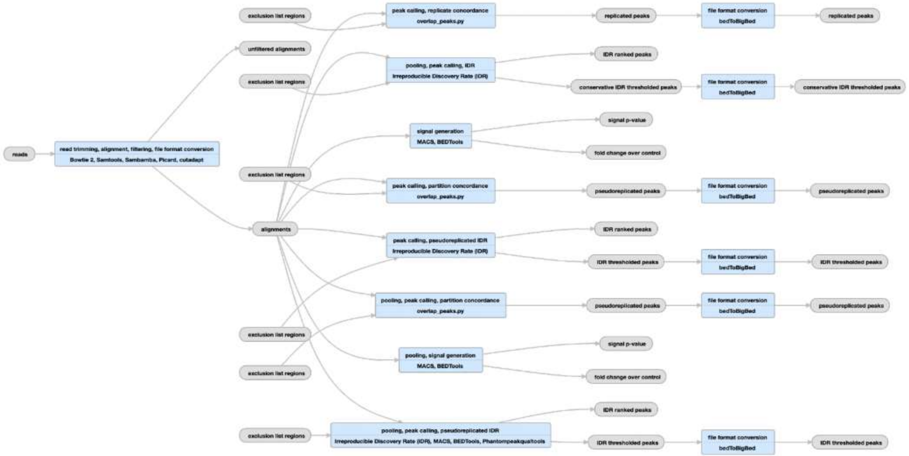


```
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    –    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    –    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    –    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    –    127480532    127481699    0,0,255
```

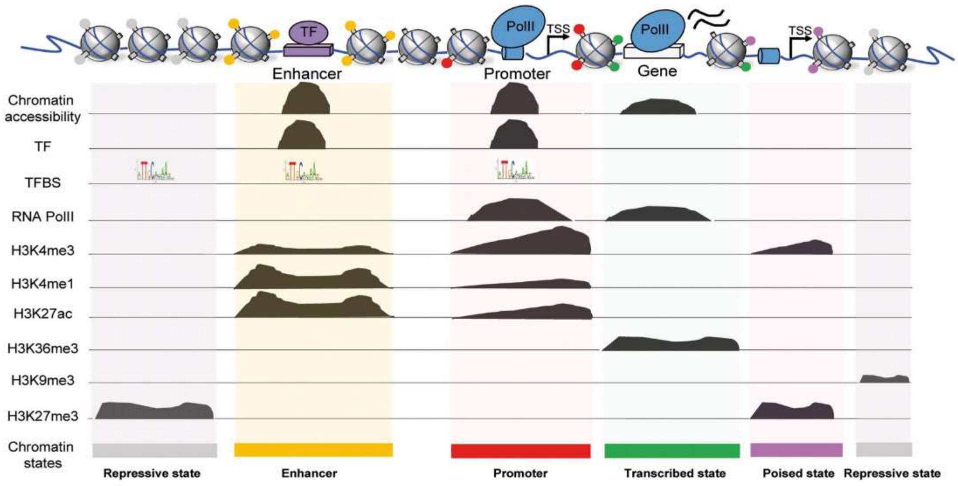
## The Overall Data Analytics Pipeline for Epigenetics



## The Entire Detailed Pipeline (ATAC-seq as an example)


reads
read trimming, alignment, filtering, file format conversion
Bowtie 2, Samtools, Sambamba, Picard, cutadapt
exclusion list regions
unfiltered alignments
alignments
peak calling, replicate concordance
overlap_peaks.py
pooling, peak calling, IDR
Irreproducible Discovery Rate (IDR)
signal generation
MACS, BEDTools
peak calling, partition concordance
overlap_peaks.py
peak calling, pseudoreplicated IDR
Irreproducible Discovery Rate (IDR)
pooling, peak calling, partition concordance
overlap_peaks.py
pooling, signal generation
MACS, BEDTools
pooling, peak calling, pseudoreplicated IDR
Irreproducible Discovery Rate (IDR), MACS, BEDTools, Phantompeakqualtools
replicated peaks
IDR ranked peaks
conservative IDR thresholded peaks
signal p-value
fold change over control
pseudoreplicated peaks
IDR thresholded peaks
file format conversion
bedToBigBed

## Histone Marks and Chromatin Accessibility


TF
PolII
TSS
PolII
TSS
Enhancer
Promoter
Gene
Chromatin accessibility
TF
TFBS
RNA PolII
H3K4me3
H3K4me1
H3K27ac
H3K36me3
H3K9me3
H3K27me3
Chromatin states
Repressive state
Enhancer
Promoter
Transcribed state
Poised state
Repressive state