



3000788 Intro to Comp Molec Biol

Lecture 4: Processing of DNA sequencing data

August 25, 2022



Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)



Quality control

FASTQ format

```
1 @ERR000589.41 EAS139_45:5:1:2:111/1
2 CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCAGGGAACATCTTGTCAT
3 +
4 3IIIIIIIIIIII>1IIIF9BG08E00I%IG+&?(4)%00646.C1#&(
5 @ERR000589.42 EAS139_45:5:1:2:1293/1
6 AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
7 +
8 IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

- Header: Location of cluster on Illumina's flow cell
- Sequence
- Quality score

FASTQ for paired-end sequencing

```

1 @ERR000589.41 EAS139_45:5:1:2:111/1
2 CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
3 +
4 3IIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(<
5 @ERR000589.42 EAS139_45:5:1:2:1293/1
6 AGTTGTTAAATCCAAGCCAATTAAGATAGTCTTATCTTTTAAAGAAAT
7 +
8 IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
    
```

Same order

```

1 @ERR000589.41 EAS139_45:5:1:2:111/1
2 CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
3 +
4 3IIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(<
5 @ERR000589.42 EAS139_45:5:1:2:1293/1
6 AGTTGTTAAATCCAAGCCAATTAAGATAGTCTTATCTTTTAAAGAAAT
7 +
8 IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
    
```

- Two FASTQ files
 - SAMPLE1_R1.fastq / SAMPLE1_R2.fastq
- Merged into a single FASTQ files
 - Forward & reverse reads must overlap
 - 300bp paired-end 16S rRNA sequencing

<table><tr><td>C</td><td>A</td><td>T</td><td>T</td><td>G</td><td>A</td><td>C</td><td>A</td></tr><tr><td>32</td><td>34</td><td>20</td><td>20</td><td>28</td><td>16</td><td>14</td><td>10</td></tr></table>								C	A	T	T	G	A	C	A	32	34	20	20	28	16	14	10	Forward read							
C	A	T	T	G	A	C	A																								
32	34	20	20	28	16	14	10																								
Reverse read		<table><tr><td>T</td><td>A</td><td>G</td><td>A</td><td>C</td><td>A</td><td>T</td><td>T</td></tr><tr><td>2</td><td>5</td><td>4</td><td>8</td><td>12</td><td>20</td><td>38</td><td>40</td></tr></table>								T	A	G	A	C	A	T	T	2	5	4	8	12	20	38	40	Base calls					
		T	A	G	A	C	A	T	T																						
2	5	4	8	12	20	38	40																								
										Q scores																					
<table><tr><td>C</td><td>A</td><td>T</td><td>T</td><td>G</td><td>A</td><td>C</td><td>A</td><td>T</td><td>T</td></tr><tr><td>32</td><td>34</td><td>22</td><td>16</td><td>35</td><td>28</td><td>30</td><td>34</td><td>38</td><td>40</td></tr></table>										C	A	T	T	G	A	C	A	T	T	32	34	22	16	35	28	30	34	38	40	Consensus	
C	A	T	T	G	A	C	A	T	T																						
32	34	22	16	35	28	30	34	38	40																						
								Posterior Qs																							
								Mismatch		Merged read																					

https://drive5.com/usearch/manual8.1/merge_pair.html

Phred score

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

https://www.drive5.com/usearch/manual/quality_score.html

- Q score = $-10 \times \log_{10}$ (base call error rate)
- Base call error of 10% \rightarrow Q score = ?
- Base call error of 0.0001 \rightarrow Q score = ?

Expected error at the ends of read

@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIIIII>1IIIF9BG08E00I%IG+&?(4)%00646.C1#&(

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			












FastQC tool

Basic Statistics

Measure	Value
Filename	small_rna.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	100
%GC	45

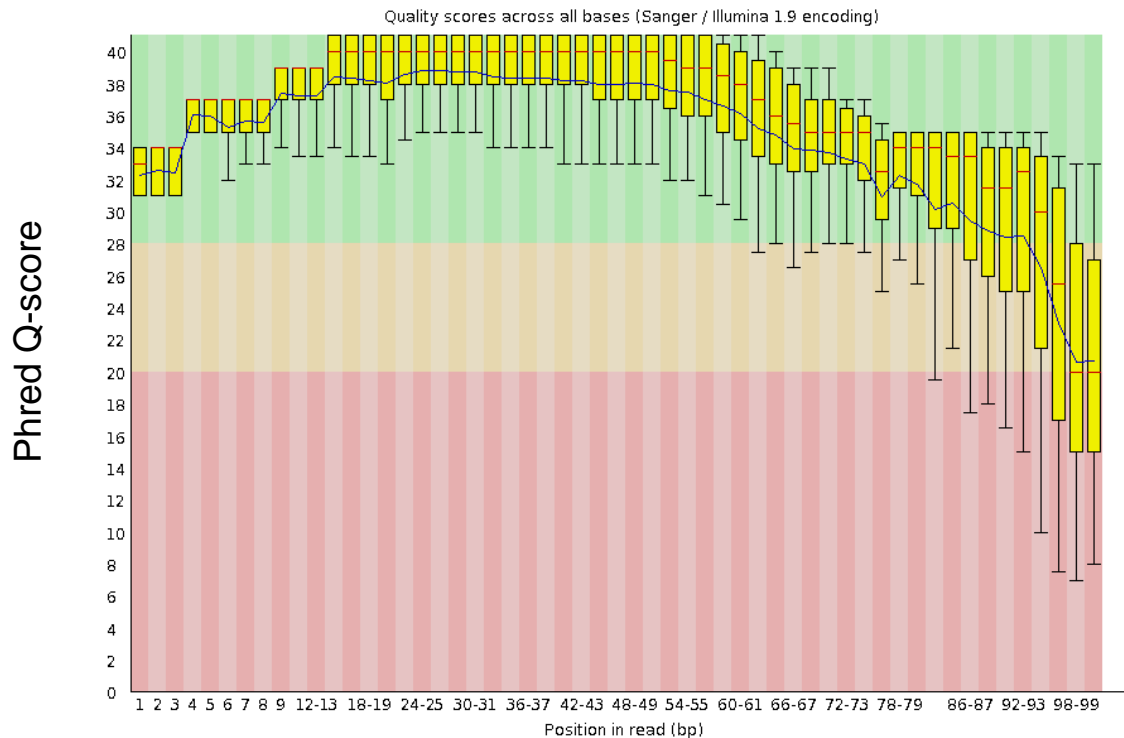
FastQC Report

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)

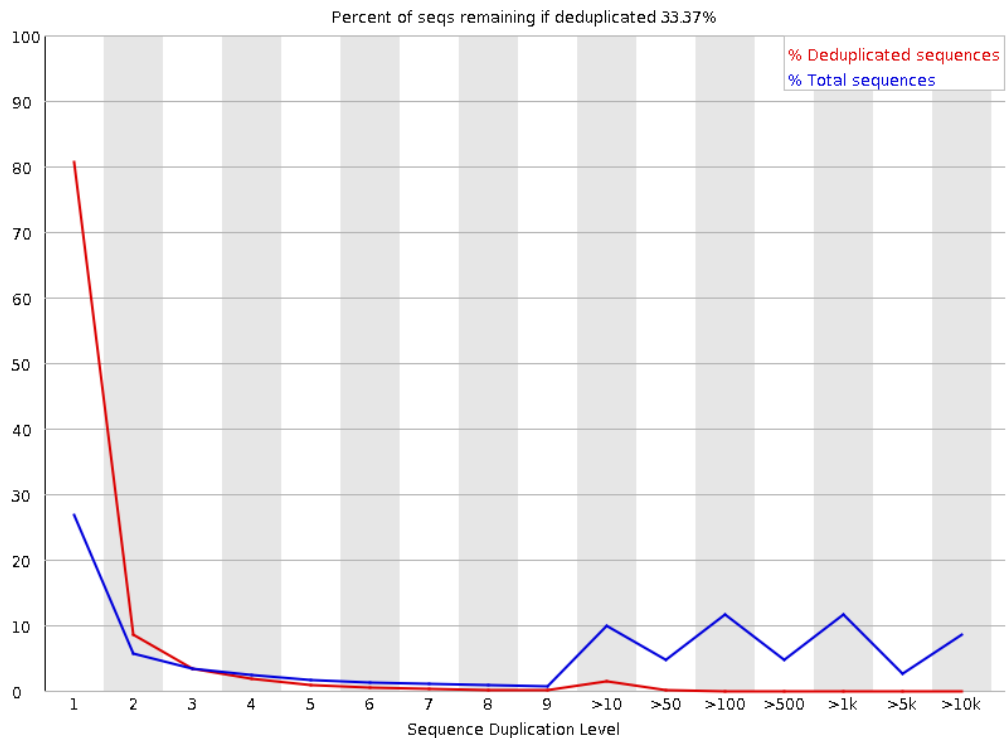
Base calling quality

! Per base sequence quality



Duplicated reads

Sequence Duplication Levels



Possible adapter read-through

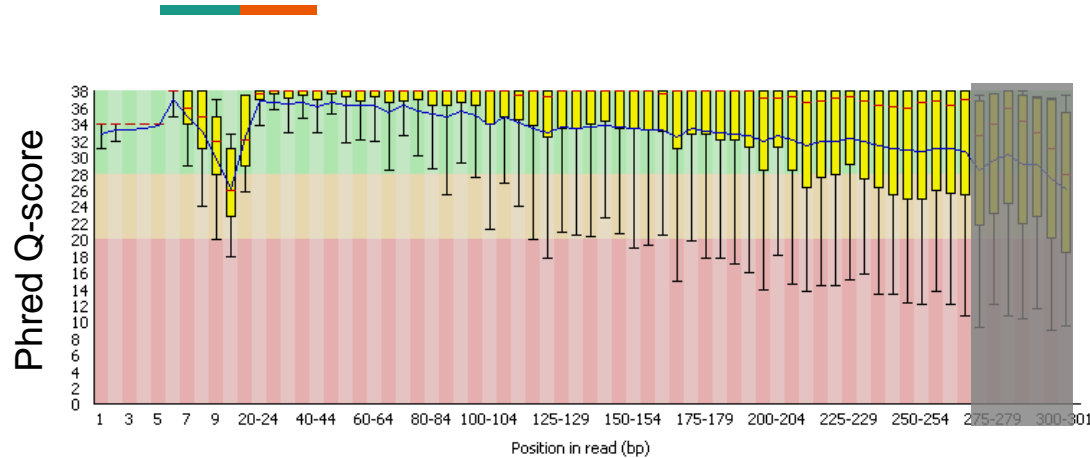
✖ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TGAGGTAGTAGATTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	10865	4.346	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TAGCTTATCAGACTGATGTTGACAGATCGGAAGAGCACACGTCTGAACTC	10845	4.338	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TCTTTGGTTATCTAGCTGTATGAGATCGGAAGAGCACACGTCTGAACTCC	7062	2.824799999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATGAAGATCGGAAGAGCACACGTCTGAACTC	4056	1.622399999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TGAGGTAGTAGTTTGTGCTGTTAGATCGGAAGAGCACACGTCTGAACTCC	3737	1.4948	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGTTTGTACAGTTAGATCGGAAGAGCACACGTCTGAACTCC	3549	1.4196	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGTTGTATGGTTAGATCGGAAGAGCACACGTCTGAACTCC	2931	1.1724	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
AACCCGTAGATCCGATCTTGTAGATCGGAAGAGCACACGTCTGAACTCCA	1910	0.764	Illumina Multiplexing PCR Primer 2.01 (100% over 29bp)
CGCGACCTCAGATCAGACGTAGATCGGAAGAGCACACGTCTGAACTCCAG	1749	0.6996	Illumina Multiplexing PCR Primer 2.01 (100% over 30bp)
TGAGGTAGTAGTTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	1647	0.6588	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATAGATCGGAAGAGCACACGTCTGAACTCCA	1622	0.648799999999999	Illumina Multiplexing PCR Primer 2.01 (100% over 29bp)
TAGCTTATCAGACTGATGTTGATAGATCGGAAGAGCACACGTCTGAACTC	1328	0.5312	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TTCAAGTAATCCAGGATAGGCTAGATCGGAAGAGCACACGTCTGAACTCC	1248	0.4992	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
AGCAGCATTGTACAGGGCTATGAAGATCGGAAGAGCACACGTCTGAACTC	1248	0.4992	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)

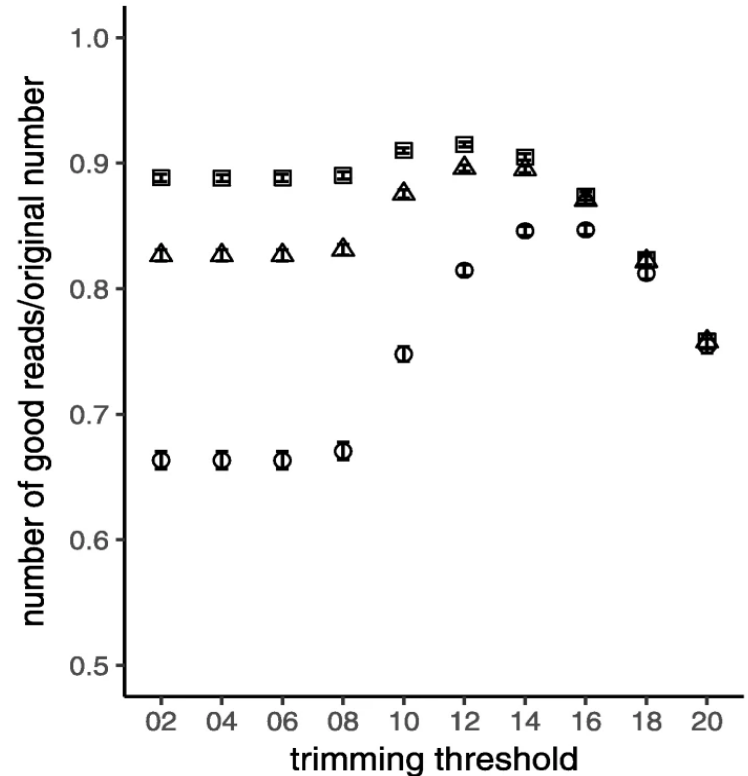


Trimming

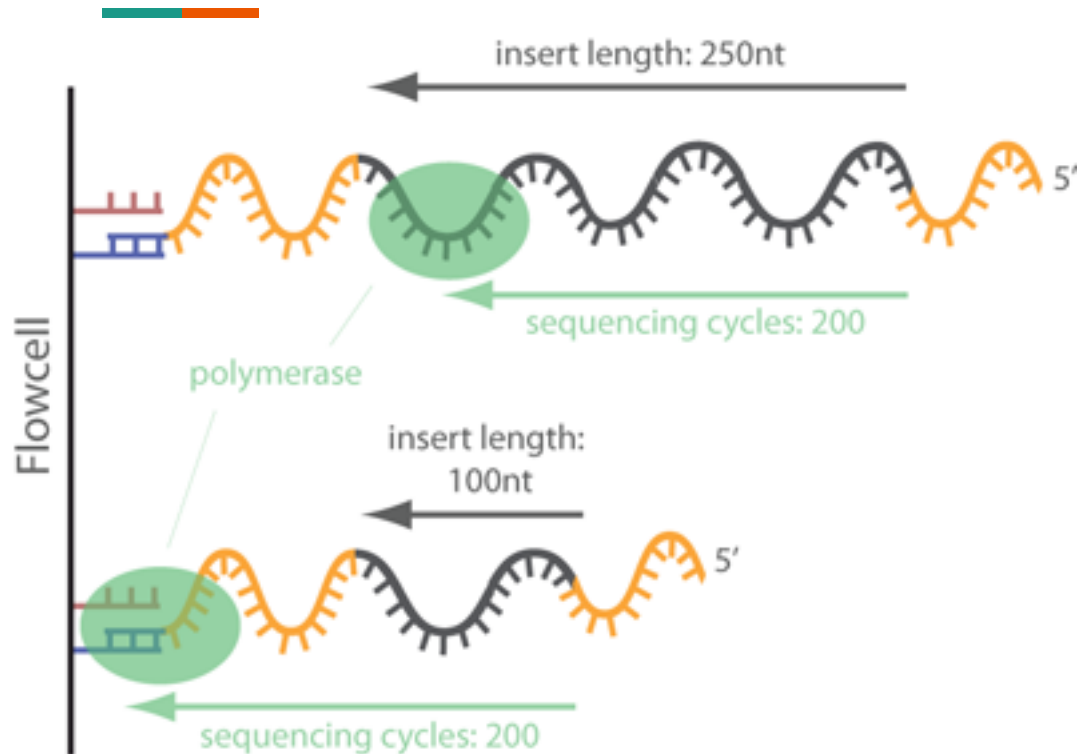
Quality trimming



- Remove bases from the end until a minimum quality is reached
- May lose reads but lead to better results in downstream analysis



Adapter trimming



main ▾

[Trimmomatic](#) / adapters /



TonyBolger Parallel Compression

..



NexteraPE-PE.fa



TruSeq2-PE.fa



TruSeq2-SE.fa



TruSeq3-PE-2.fa



TruSeq3-PE.fa



TruSeq3-SE.fa

Trimmomatic code

```
trimmomatic PE -threads 4 SRR_1056_1.fastq SRR_1056_2.fastq \
    SRR_1056_1.trimmed.fastq SRR_1056_1un.trimmed.fastq \
    SRR_1056_2.trimmed.fastq SRR_1056_2un.trimmed.fastq \
    ILLUMINACLIP:SRR_adapters.fa SLIDINGWINDOW:4:20
```

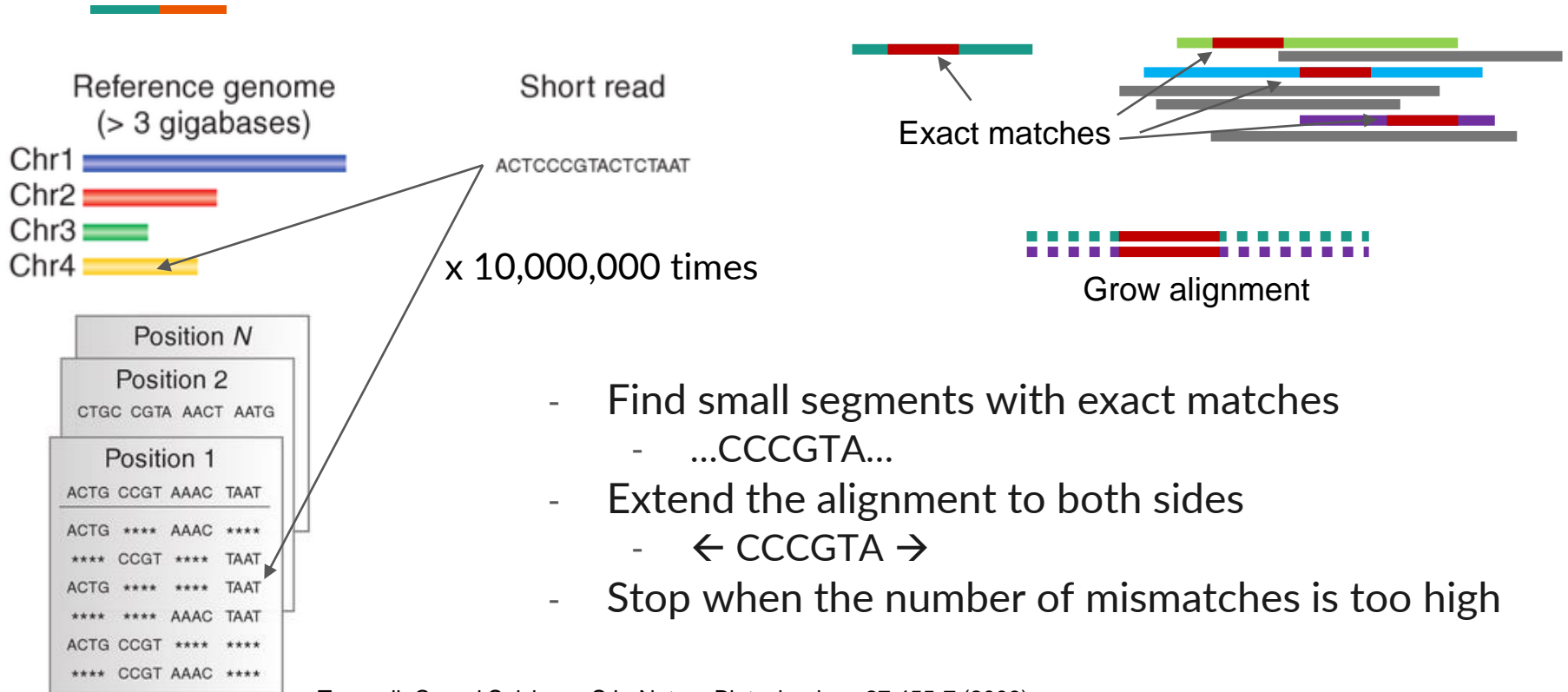
- Using 4 CPU threads
- Process 3 sets of paired end FASTQ
- Remove adapters listed in SRR_adapters.fa
- Check quality in a sliding window

header	1	>PrefixNX/1	
sequence	2	AGATGTGTATAAGAGACAG	FASTA format
header	3	>PrefixNX/2	
sequence	4	AGATGTGTATAAGAGACAG	
	5	>Trans1	
	6	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	
	7	>Trans1_rc	
	8	CTGTCTCTTATACACATCTGACGCTGCCGACGA	
	9	>Trans2	
	10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	
	11	>Trans2_rc	
	12	CTGTCTCTTATACACATCTCCGAGCCCACGAGAC	



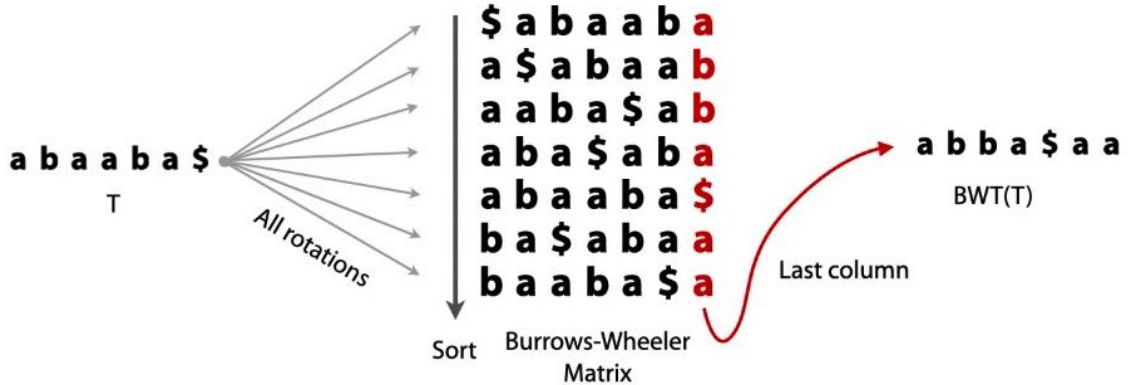
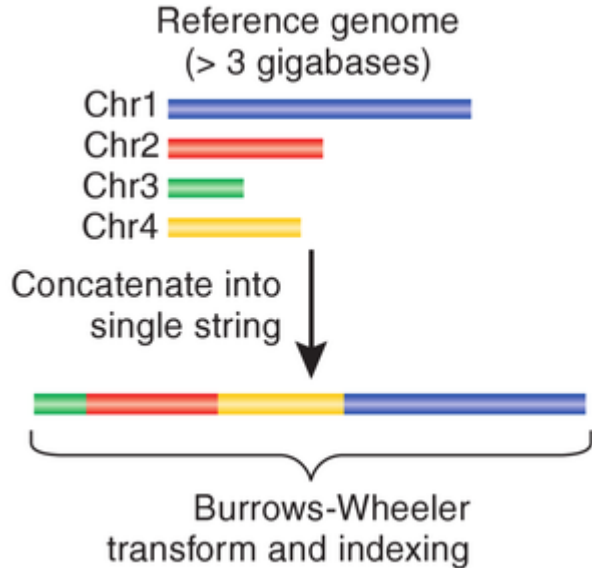
Alignment

Sequence alignment is a form of search



- Find small segments with exact matches
 - ...CCCGTA...
- Extend the alignment to both sides
 - ← CCCGTA →
- Stop when the number of mismatches is too high

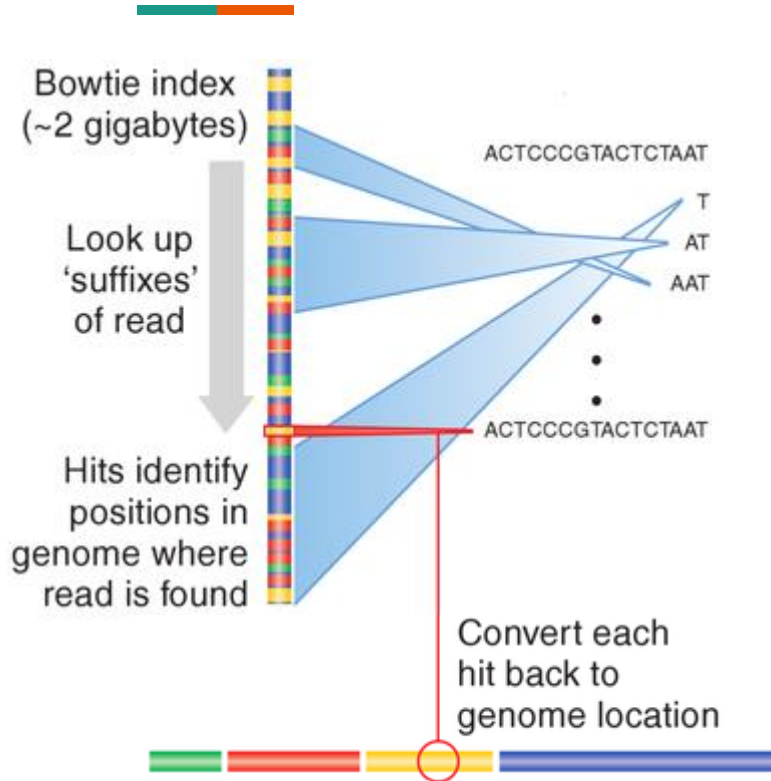
Indexing with Burrows-Wheeler transform



Burrows, M. and Wheeler, D.J. A block sorting lossless data compression algorithm. 1994

BWT is reversible!

Efficient search on BWT index



- BWT index provides information on all short segments of the genome
- 20x smaller memory than straightforward indexing of human genome
- 30x faster search speed

Dynamic programming



No. of rocks	1	2	3	4	5	6	7	8	9	10
Winner										
No. of rocks	11	12	13	14	15	16	17	18	19	20
Winner										?

- Dynamic programming build the solution of complex problem using on the solutions of simpler ones
- There is a pile of 20 rocks. Two players take turns by removing 1 or 2 rocks from the pile. Whoever removes the last rock(s) win. Who is the winner?

Dynamic programming for sequence alignment

Dynamic programming matrix:

		j → (sequence y)									
		0	1	2	3	4	5	6	7	8 = N	
			T	G	C	T	C	G	T	A	
i ↓ (sequence x)	0	0	-6	-12	-18	-24	-30	-36	-42	-48	
	1 T	-6	5	-1	-7	-13	-19	-25	-31	-37	
	2 T	-12	-1	3	-3	-2	-8	-14	-20	-26	
	3 C	-18	-7	-3	8	2	3	-3	-9	-15	
	4 A	-24	-13	-9	2	6	0	1	-5	-4	
	5 T	-30	-19	-15	-4	7	4	-2	6	0	
M = 6	A	-36	-25	-21	-10	1	5	2	0	11	

Optimum alignment scores 11:

T	-	-	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	-6	-6	+5	+5	-2	+5	+5

- The best alignments for long sequences depend on the best alignments of shorter sequences
- The best alignment for **TTCATA** vs **TGCTCGTA** is either
 - **T**/**T** + best alignment for TCATA vs GCTCGTA
 - **T**/- + best alignment for TCATA vs **T**GCTCGTA
 - -/**T** + best alignment for **T**TTCATA vs GCTCGTA

Bowtie2 command



Usage:

```
bowtie-build [options]* <reference_in> <ebwt_base>
```

Usage:

```
bowtie [options]* -x <ebwt> {-1 <m1> -2 <m2> | --12 <r> | --interleaved <i> | <s>} [<hit>]
```

- Use bowtie-build to index a genome database (FASTA file)
 - `bowtie-build GRCh38_v1.fasta GRCh38_v1`
- Use bowtie to perform alignment
 - `bowtie -x GRCh38_v1 -1 sample1_R1.fastq -2 sample1_R2.fastq -sam --threads 8 sample1.sam`



SAM/BAM manipulation

Sequence Alignment Map (SAM)

Sort Order = by genomic coordinate

```
@HD VN:1.6 SO:coordinate SN = reference sequence's name (FASTA header)
@SQ SN:ref LN:45 LN = reference sequence's length
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
```

- **r001** = read name (from sequencing FASTQ)
- **ref** = reference sequence name (from genomic FASTA)
- **7** = first position on the reference sequence
- **30** = Mapping quality score = $-10 \times \log_{10}(\text{error})$
- **8M2I4M1D3M** = CIGAR string = matches, insertion, deletion information

SAM manipulation with samtools



Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data

- By default, SAM output from aligner is **unsorted**
- Also, SAM is large
- BAM is a zipped version of SAM
- Use samtools to convert SAM→BAM and sort

Example of SAM manipulation



- `bowtie -x GRCh38_v1 -1 sample1_R1.fastq -2 sample1_R2.fastq -sam --threads 8 | samtools view -Sb - | samtools sort > sample1_sorted.bam`
- `samtools view -Sb` converts SAM to BAM
- `samtools sort` sorts the BAM file
- `|` is the `pipe` command which passes output from one software to another
- `>` is the `redirect` command which writes output of a software to a file

Pileup format

Sequence	Position	Reference Base	Read Count	Read Results	Quality
seq1	272	T	24	,.\$. , , ^+.	<<<+;<<<<<<<<<=<;<;7<&
seq1	273	T	23	,. , , A	<<<;<<<<<<<<3<=<<<;<<+
seq1	274	T	23	,.\$. , ,	7<7;<;<<<<<<<<=<;<;<<6
seq1	275	A	23	,\$. , , ^1.	<+;9*<<<<<<<<=<<:;<<<<
seq1	276	G	22	. . . T, . . . , . . . , . . . , . . .	33;+<<7=7<<7<&<<1;<<6<
seq1	277	T	22	. . . , . . . , . . . C. , . . . , . . . G.	+7<;<<<<<<<&<=<<:;<<&<
seq1	278	G	23	. . . , . . . , . . . , . . . , . . . ^k.	%38*<<;<7<<7<=<<<;<<<<<
seq1	279	C	23	A..T, . . . , . . . , . . . , . . .	75&<<<<<<<<=<<<9<<:<<<

Image from wikipedia

- Focus on each reference base pair position
 - Whether each read matches the reference or not



Sequence assembly

De novo assembly

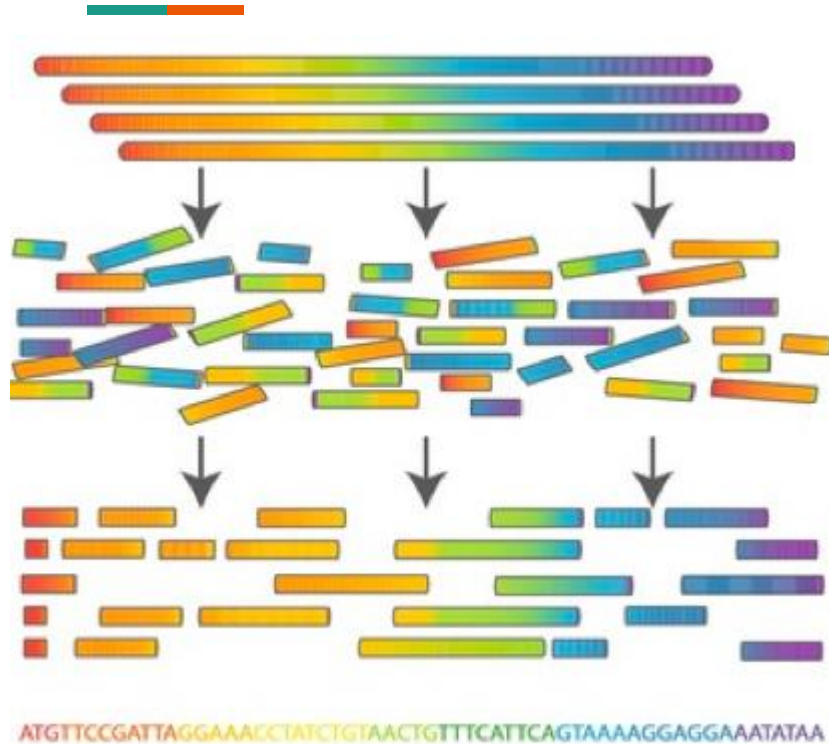


Image from wikipedia

```
CTGTGTGTT GACGTCACT
GTGTCCTGA CTG...
...ACTGT TGTCTGAC CACTG...
ACTGTGTGT CTGGCGTCA
GTGTGTCCT ACGTCACTG
```

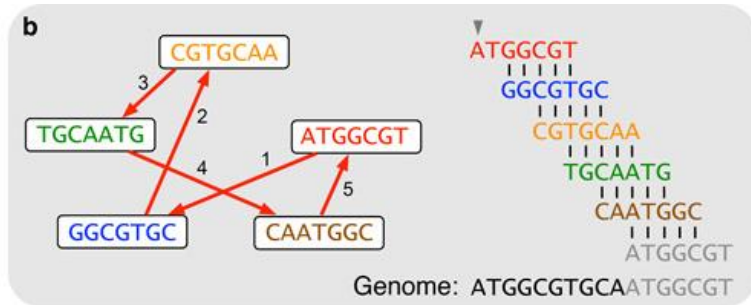


```
...ACTGTGTGTCCTGACGTCACTG...
```

Chandra Varma Bogaraju, S. Int J Embed Syst 9:74 (2017)

Assembly via de Bruijn graph

- Each directed path in a de Bruijn graph represents a possible contiguous segment of the genome

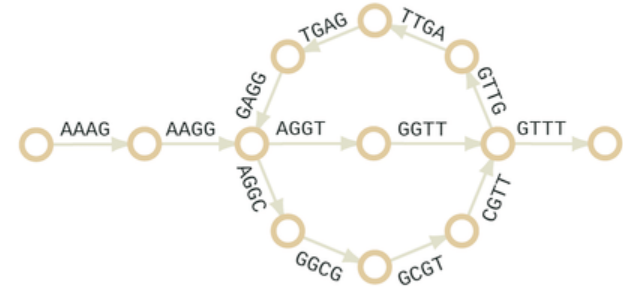


A. Short read to k -mers ($k=4$)

AAAGGCGTTGAGGTT

AAAG
AAGG
AGGC
GGCG
GCGT
CGTT
GTTG
TTGA
TGAG
GAGG
AGGT
GGTT

B. Eulerian de Bruijn graph



C. Hamiltonian de Bruijn graph



Contig and scaffold



Genome



Reads



Contigs



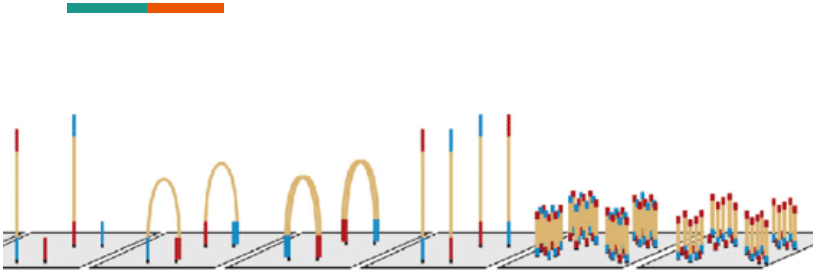
Scaffolds



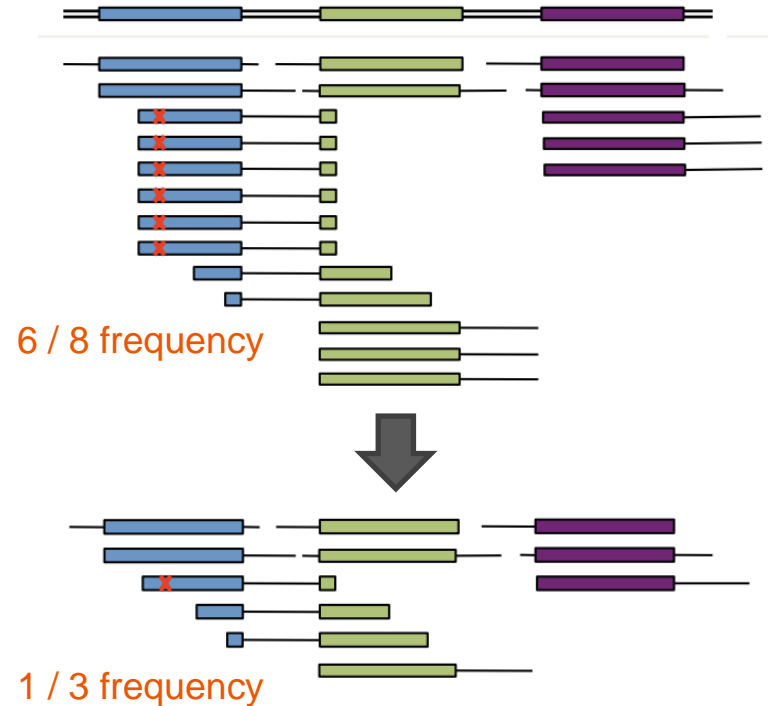


Deduplication

Duplicated = derived from the same molecule



- Similar sequences coming from nearby coordinates in Illumina flow cells
- Reads with the same start and end
 - Highly unlikely to generate the exact same DNA molecules by chance
- Lead to incorrect frequency estimates



Mark duplicate command examples

```
java -jar picard.jar MarkDuplicates \  
  I=input.bam \  
  O=marked_duplicates.bam \  
  M=marked_dup_metrics.txt
```

```
samtools markdup positionsort.bam markdup.bam
```

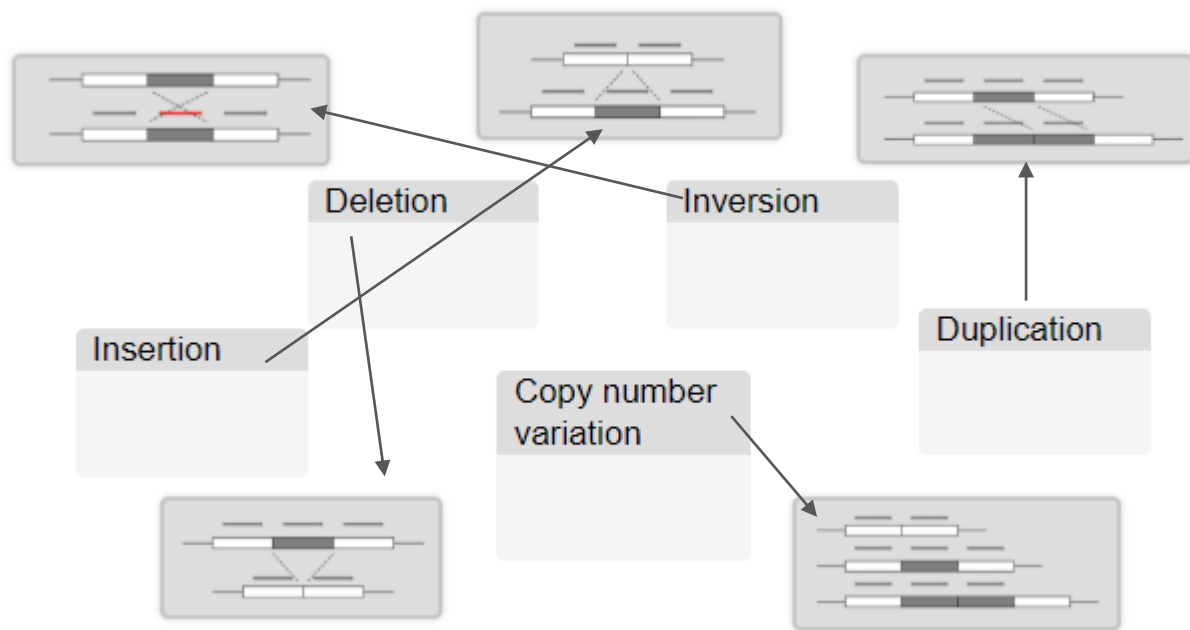
```
@SIM:1:FCX:1:15:6329:1045:GATTACT+GTCTTAAC 1:N:0:ATCCGA  
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC  
+  
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=#
```

- Require just the **sorted** alignment output
 - Illumina flow cell (x, y) coordinate is in the read name
 - Mapped chromosome position is in the SAM/BAM



Variant calling

Type of variants



Translocation

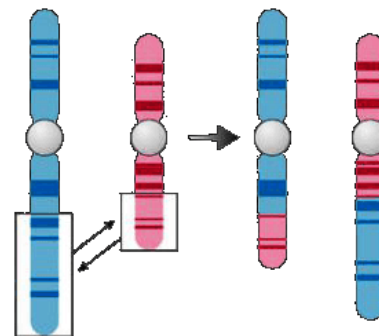



Image from wikipedia

Not all differences are true variants



Step	Sample → DNA extraction → Fragmentation → PCR amplification → Flow cell hybridization → Cluster generation → Sequencing by Synthesis						
Error source	Mutagenesis	Oxidative damage	Oxidative damage	Polymerase mistakes		Cluster PCR errors	Phasing Fluorophore crosstalk
Error category	Biological variants	DNA damage	DNA damage	PCR errors		Sequencing errors	Sequencing errors
Detection	← Ignored Called →						

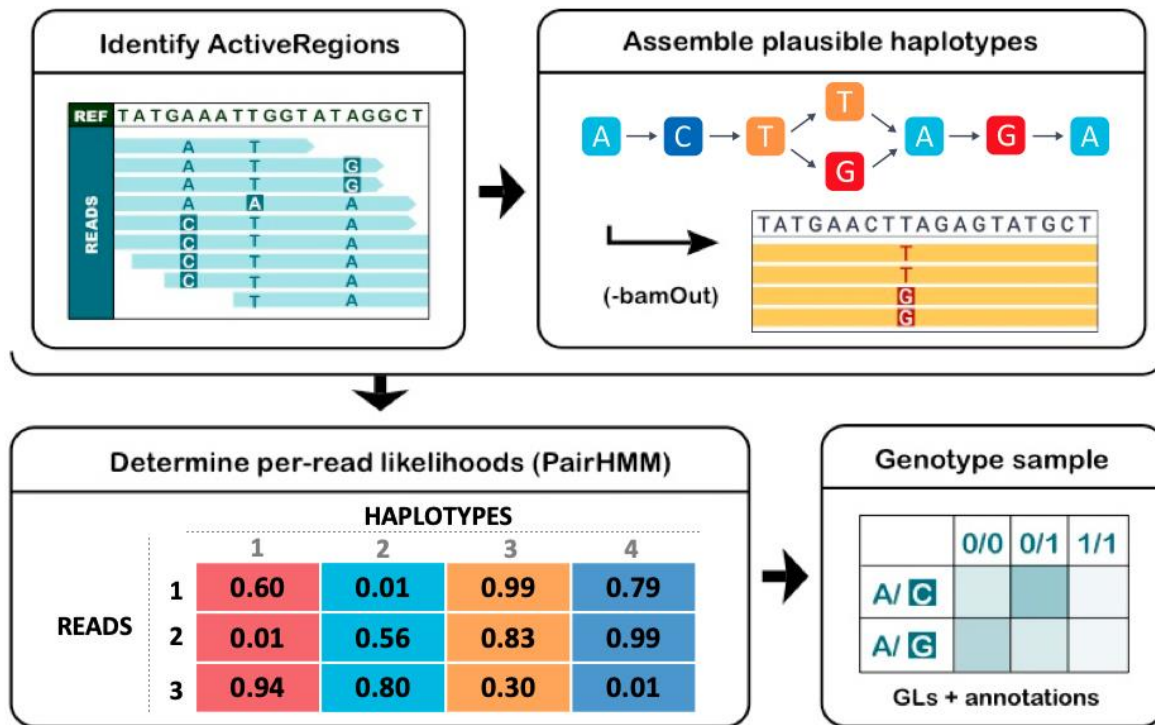
- Errors are random
 - High depth of sequencing can provide the correct consensus
 - Filter using **read depth** + **allele frequency**

Variant calling strategy



- Small-scale variants, such as SNV and short indel
 - Consider only aligned reads
 - Identify differences and assess significance
- Copy number variations
 - Consider only aligned reads
 - Look for loci with high or low frequencies compared to others
- Chromosomal translocation and inversion
 - Consider reads whose forward and reverse sequences map to different regions
 - *De novo* assembly

Small-scaled variant calling



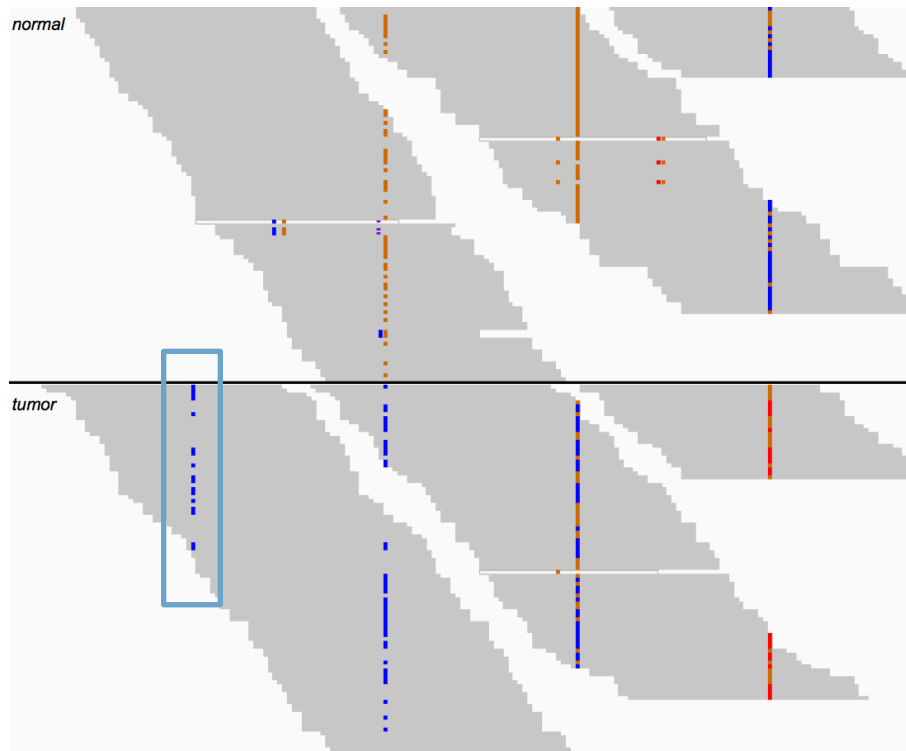
$P(\text{haplotype} \mid \text{read})$

Germline vs somatic variants

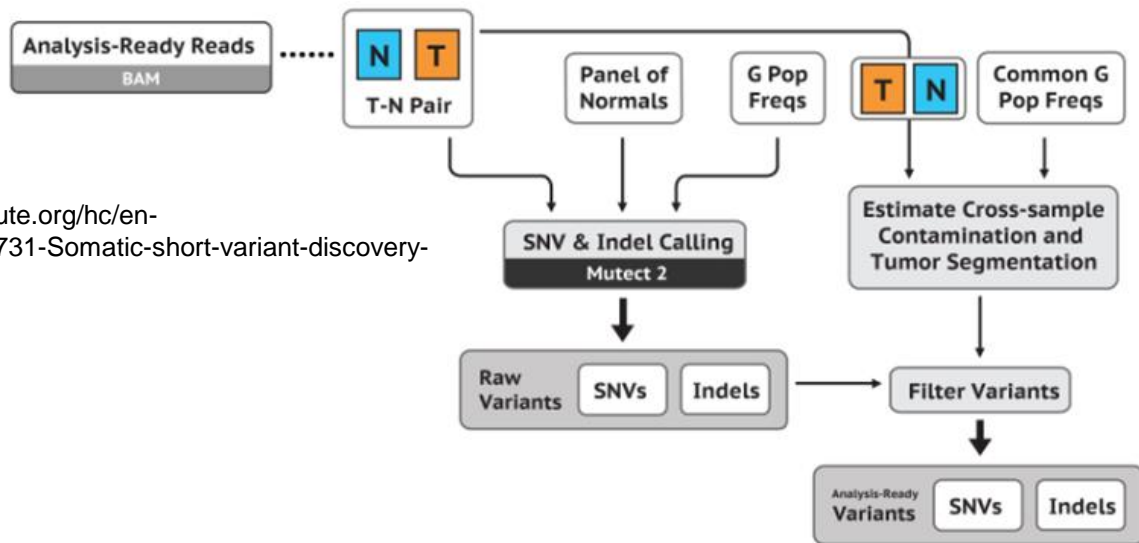


- Germline mutations = inherited and appear in every cell of the offspring
 - Compare DNA from **normal tissues** to reference genomes
 - Fixed ploidy
- Somatic mutations = occur during lifetime
 - Compare DNA from **disease tissues** to normal tissues
 - **Also compare to DNA from other healthy individuals**
 - **Allow variation in ploidy** (different disease cells can have different mutations)

Germline vs somatic variants



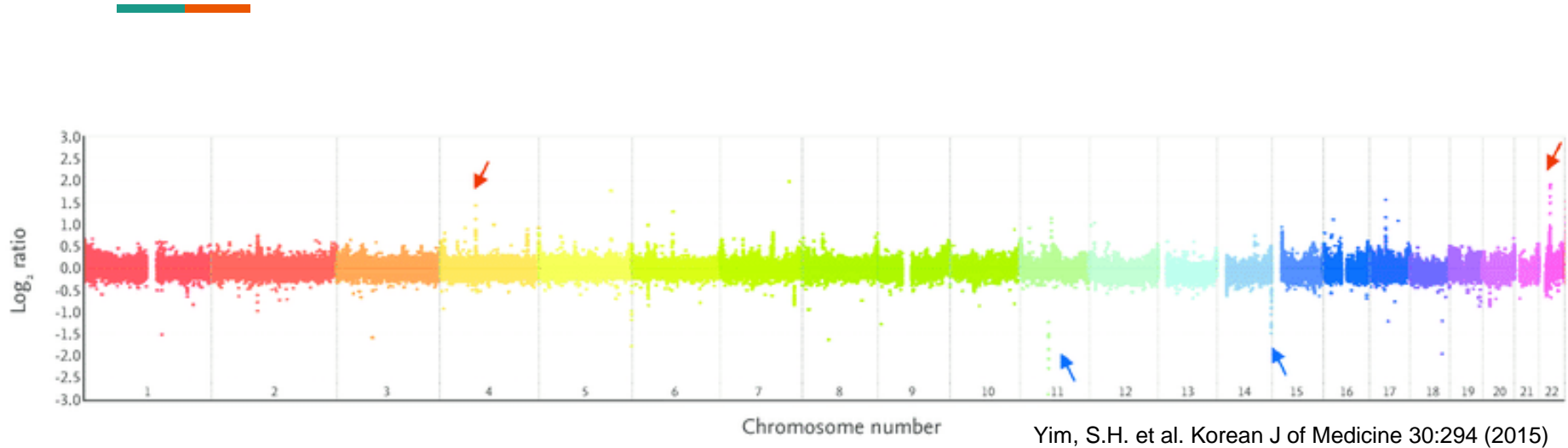
Genome Analysis Toolkit (GATK) somatic workflow



<https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels->

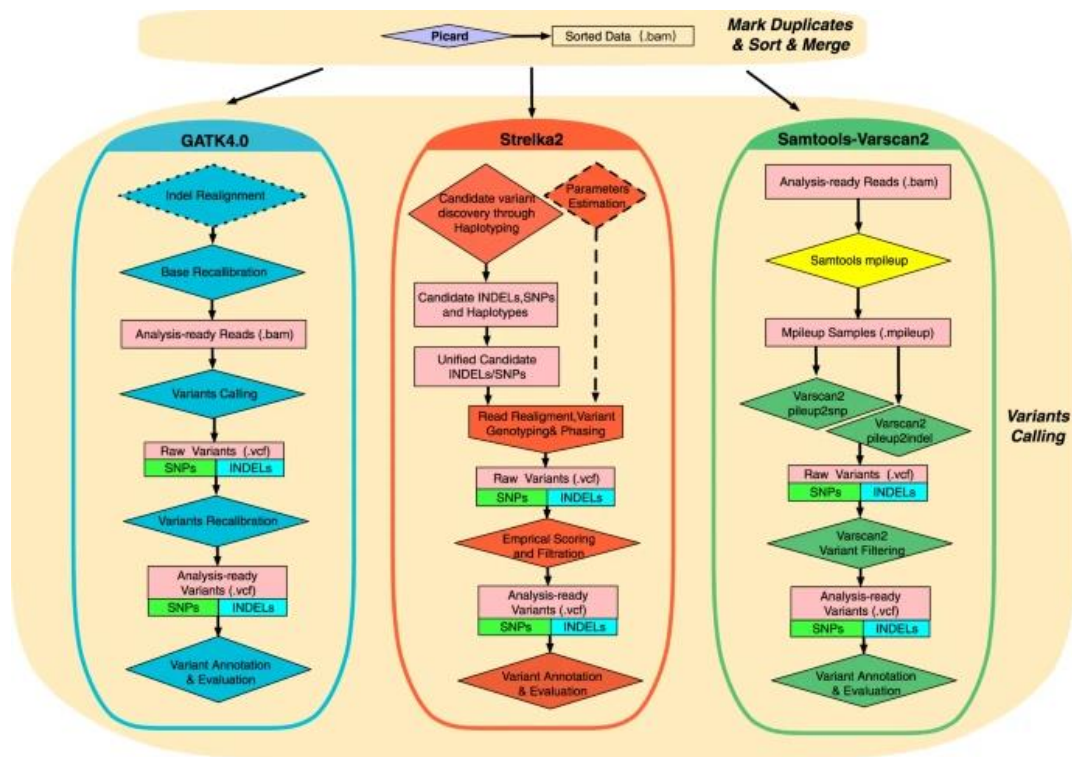
- Inclusion of matched normal (N), panels of healthy individual (Panels of Normals), and allele frequency in the general population (G Pop Freqs)
- Also estimate **contamination** = **normal cells in disease sample**

Copy number variations



- Look for loci with high or low frequencies compared to others

Utilizing multiple callers



Variant Call Format (VCF)

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA000001 NA000002
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2
```

Variant filtering



```
gatk VariantFiltration \  
  -V snps.vcf.gz \  
    -filter "QD < 2.0" --filter-name "QD2" \  
    -filter "QUAL < 30.0" --filter-name "QUAL30" \  
    -filter "SOR > 3.0" --filter-name "SOR3" \  
    -filter "FS > 60.0" --filter-name "FS60" \  
    -filter "MQ < 40.0" --filter-name "MQ40" \  
    -filter "MQRankSum < -12.5" --filter-name "MQRankSum-12.5" \  
    -filter "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum-8" \  
  -O snps_filtered.vcf.gz
```

Quality score normalized by read depth

Quality score

Strand bias scores

Mapping quality scores

- Mostly follow guideline from software developer or publications
 - But take note of the thresholds just in case of unusual outputs



Variant annotation

Understanding the importance of a variant



- Impact on sequence
 - Non-synonymous, splice site, frameshift, regulatory element
- Is it known: allele frequency
 - Genome Aggregation Database: gnomAD
 - dbSNP
- Clinical implication: observed in patients, treatment response, drug target
 - ClinVar, COSMIC, PharmGKB
- Variant effect predictor (VEP)
- Funcotator/Oncotator

ClinVar

NM_007294.3(BRCA1):c.*6207C>T

Interpretation: Benign

Review status: ★★☆☆ reviewed by expert panel

Submissions: 1

First in ClinVar: Sep 29, 2015

Most recent Submission: Sep 29, 2015

Last evaluated: Jan 12, 2015

Accession: VCV000209219.3

Variation ID: 209219

Description: single nucleotide variant

NM_007294.3(BRCA1):c.*6207C>T

Allele ID: 206177

Variant type: single nucleotide variant

Variant length: 1 bp

Cytogenetic location: 17q21.31

Genomic location: 17: 43039471 (GRCh38) [GRCh38](#) [UCSC](#)

17: 41191488 (GRCh37) [GRCh37](#) [UCSC](#)

HGVS:

Nucleotide	Protein	Molecular consequence
NC_000017.11:g.43039471G>A		
NC_000017.10:g.41191488G>A		
NG_005905.2:g.178513C>T		

[... more HGVS](#)

Protein change:

-

Other names: 11918 C>T

Canonical SPDI: [?](#) NC_000017.11:43039470:G:A

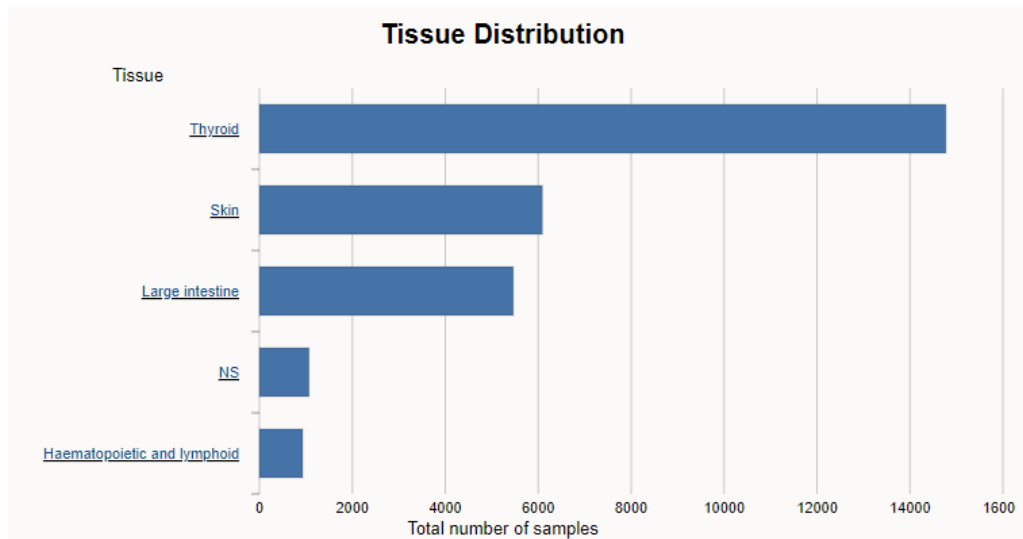
Functional consequence: -

Global minor allele frequency (GMAF): 0.00679 (A)

Allele frequency: Trans-Omics for Precision Medicine (TOPMed) 0.00211

Catalog of Somatic Mutations in Cancer (COSMIC)

Mutation
COSV56056643



Sample name ▲	Gene name ▼	Transcript ▼	Primary Tissue ▼	Tissue Subtype 1 ▼	Primary Histology ▼	Histology Subtype 1 ▼	Pubmed ID ▼
1011-mel	BRAF	ENST00000646891.1 🔗	NS	NS	Malignant melanoma	NS	15467732
1022043	BRAF	ENST00000646891.1 🔗	NS	NS	Malignant melanoma	NS	16007203

Pharmacogenomics knowledgebase



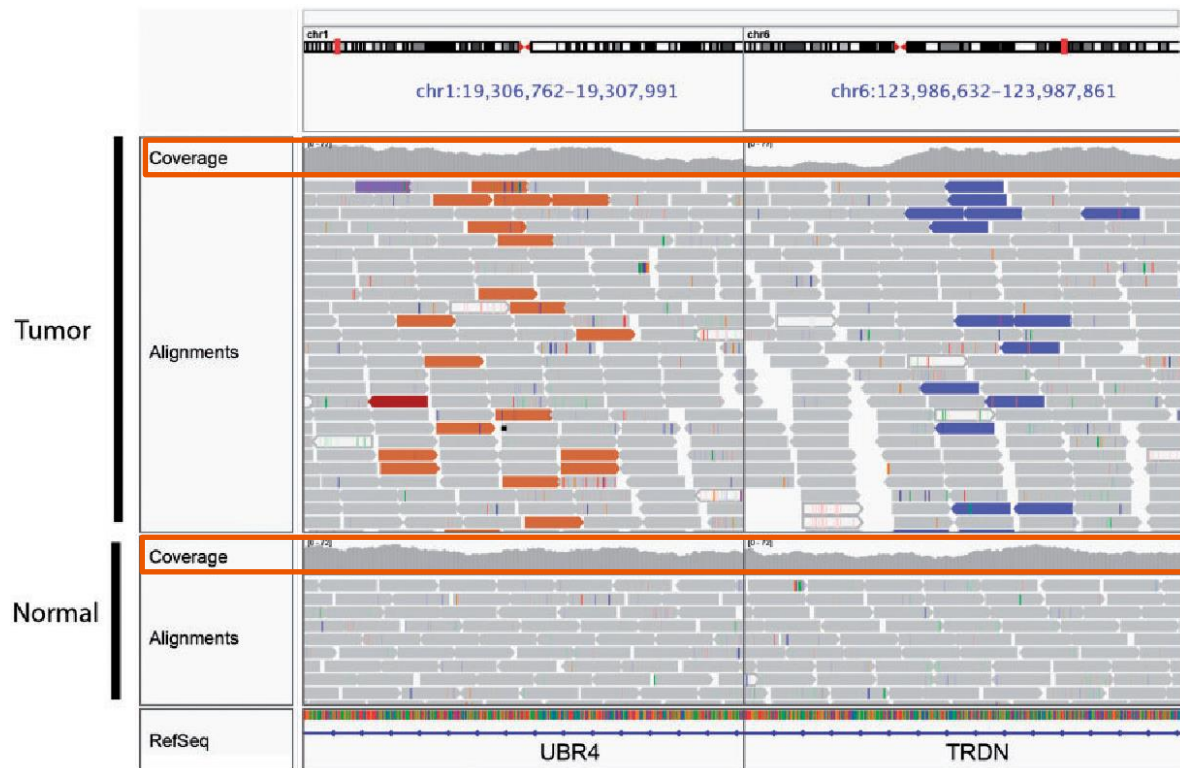
VARIANT	LITERATURE	DRUGS	GENES	ASSOCIATION
rs2069502	PMCID: PMC3959225	somatropin recombinant	CDK4	Genotype CC is associated with decreased response to somatropin recombinant in children with Turner Syndrome as compared to genotypes CT + TT.

VARIANT	SIGNIFICANCE	P-VALUE	# OF CASES	# OF CONTROLS	BIOGEOGRAPHICAL GROUPS	PHENOTYPE CATEGORIES
rs2069502	yes	< 0.05	147	0	Unknown	<ul style="list-style-type: none">Efficacy



Visualization

Integrated Genomics Viewer (IGV)



Genomics data processing workflow summary



- Check quality of FASTQ files
- Trimming
- Alignment and/or assembly
- Deduplicate
- Variant calling
 - Germline or Somatic
 - Copy number variation
 - Translocation
- Variant filtering
- Variant annotation

Any question?



- See you next week on August 30rd 9-10:30am