Genomics
Volume 112, Issue 3, May 2020, Pages 2166-2172

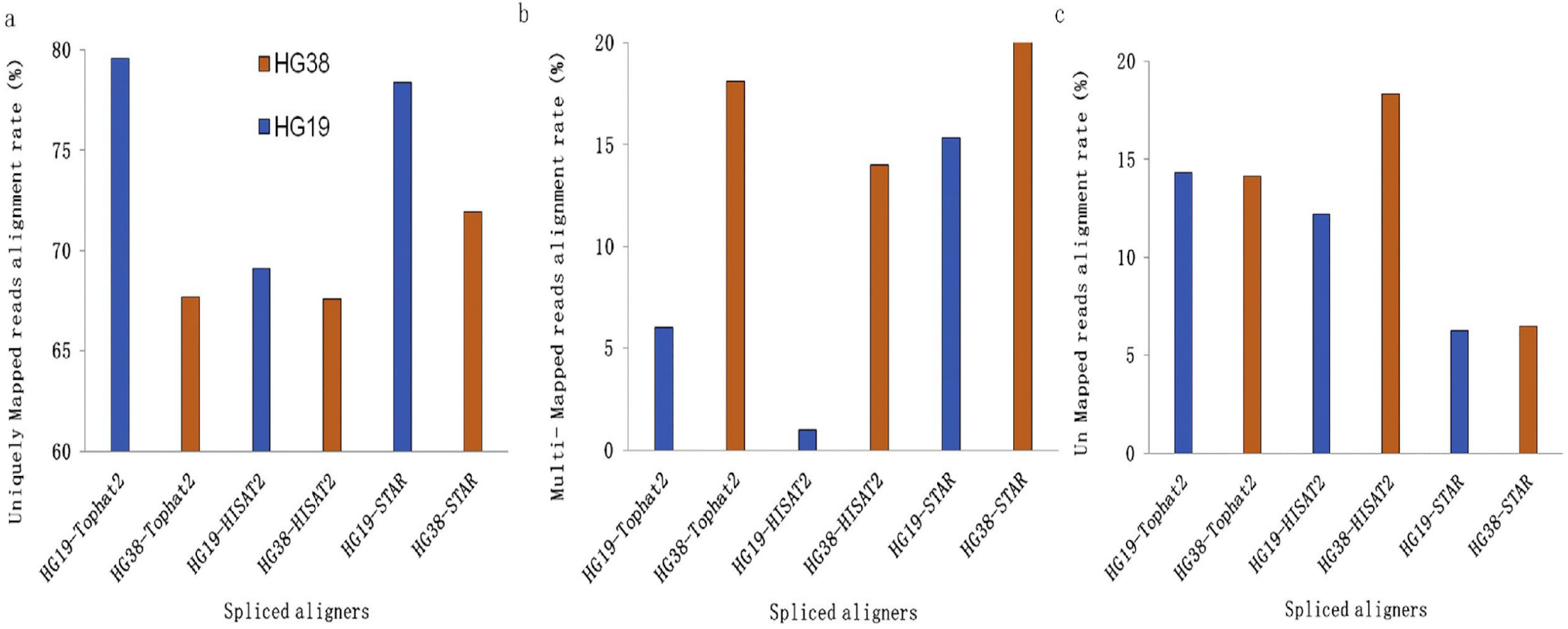
Original Article

Identifying suitable tools for variant detection and differential gene expression using RNA-seq data

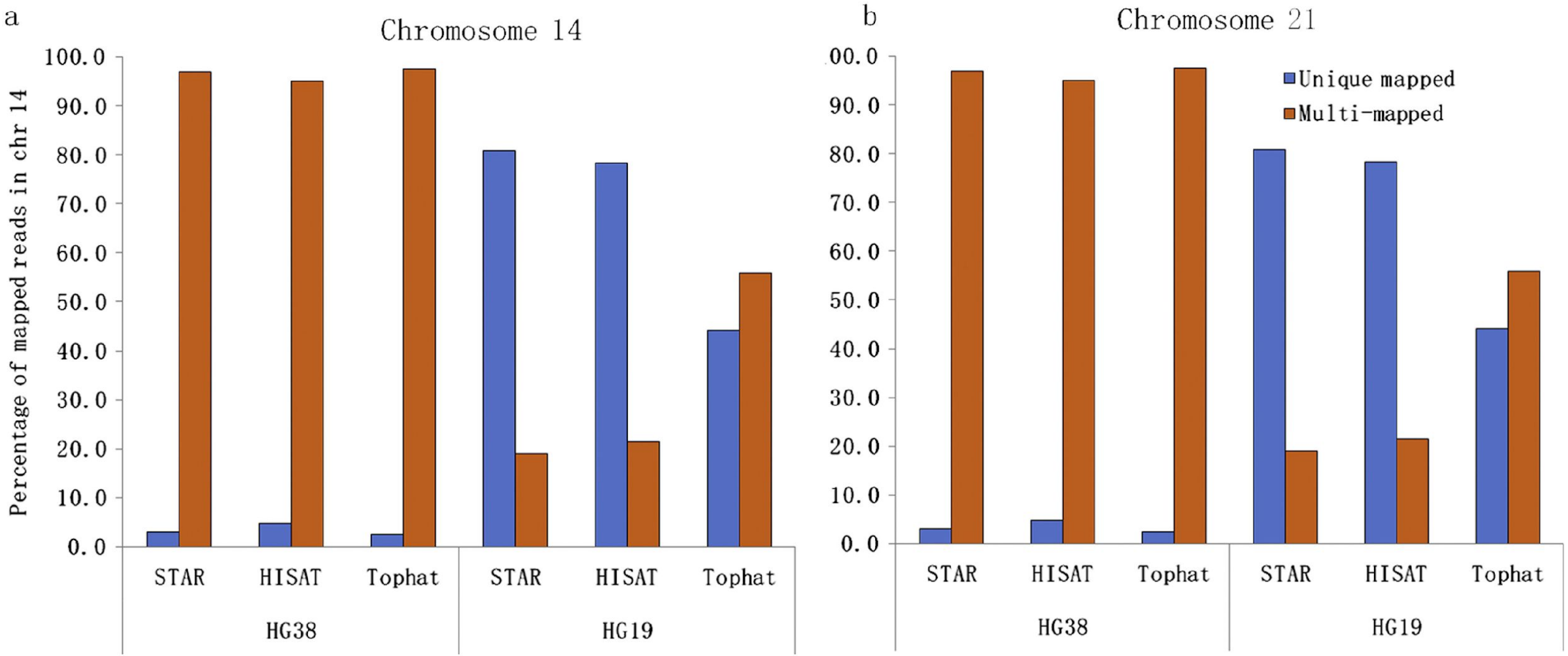
S. Akila Parvathy Dharshini [a], Y.-H. Taguchi [b], M. Michael Gromiha [a c]

https://www.sciencedirect.com/science/article/pii/S0888754319307517#f0010

The alignment rate of various spliced aligners using hg19 and hg38genomic assemblies



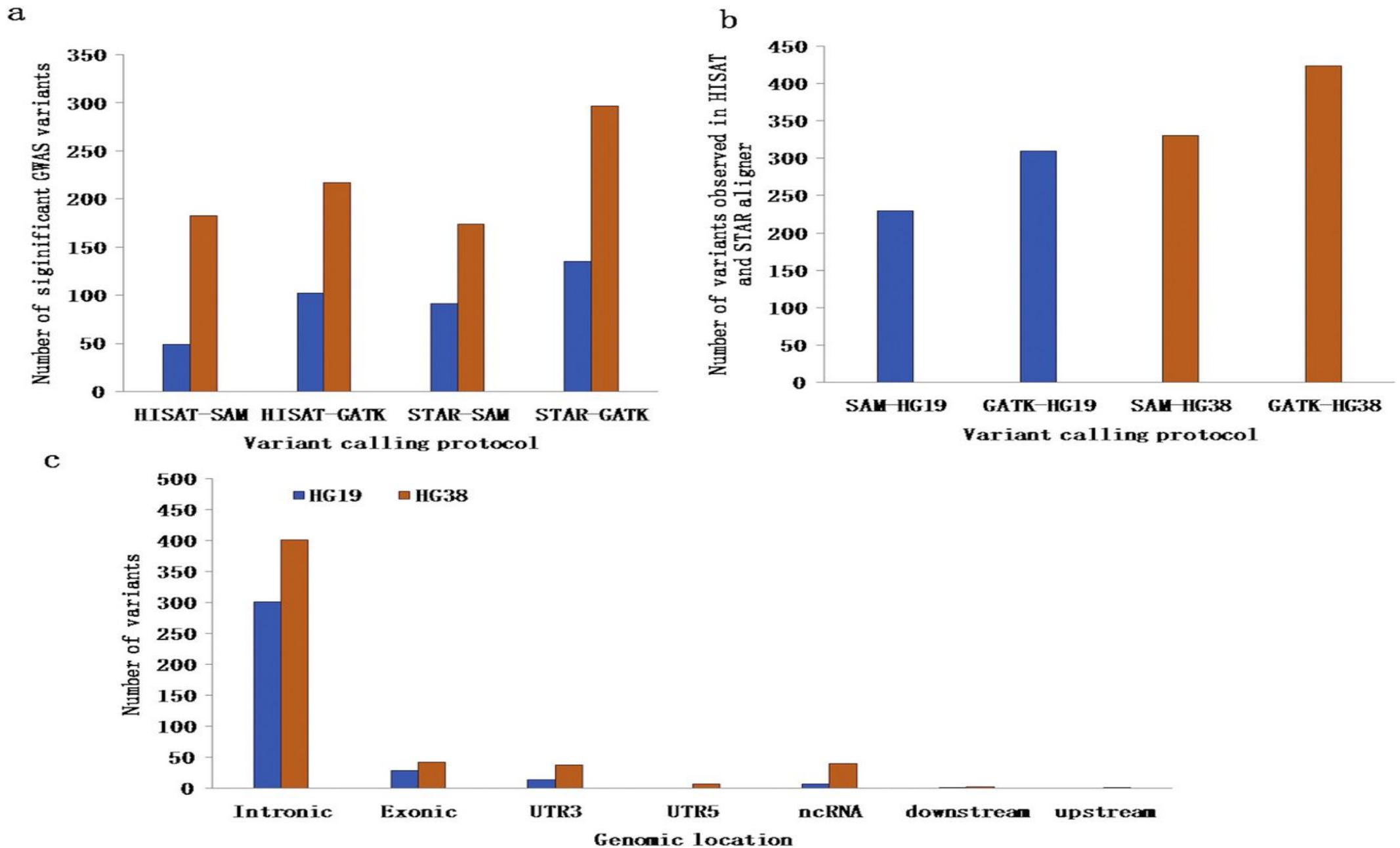
Uniquely mapped and multi-mapped reads based on chromosome location.



Number of splice junction-spanning reads using various spliced aligners (hg19/hg38)

Average spliced junction*1000		
Aligners	**hg19**	**hg38**
Tophat2	187	201
HISAT2	255	256
STAR	242	254

Number of variants predicted by various variant calling protocol (hg19/hg38)

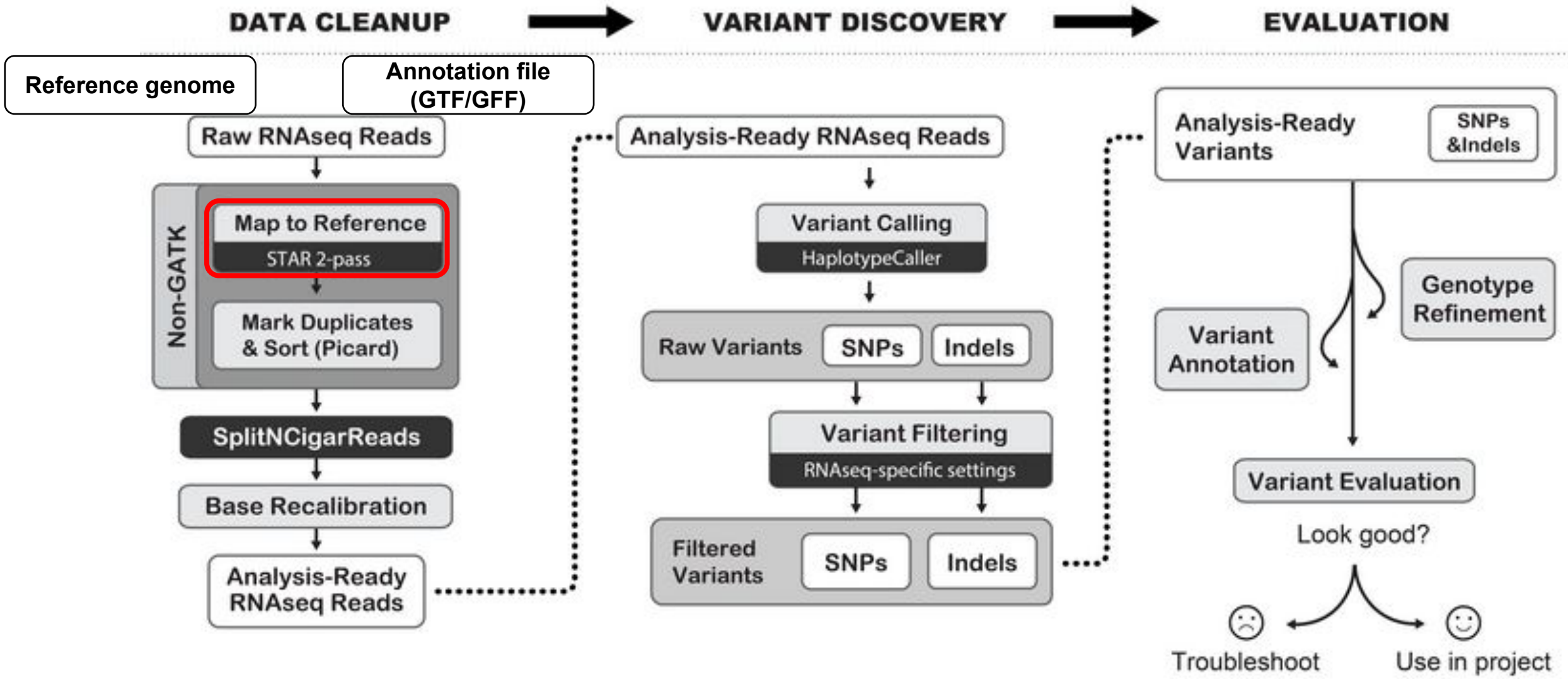


SUMMARY

Highlights

- We evaluated the RNAseq pipelines based on hg38 genomic assembly.
- In spliced alignment, the rate of multi-mapped reads are high compared to hg19.
- The aligners not able to distinguish the origin of the reads and tend to map with multiple location.
- GATK/STAR variant calling protocol yields more number of GWAS variants from RNAseq data.
- Transcriptome based quantification outperforms the genome based quantification methods.

PIPELINE


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
Reference genome
Annotation file (GTF/GFF)
Raw RNAseq Reads
Non-GATK
Map to Reference
STAR 2-pass
Mark Duplicates & Sort (Picard)
SplitNCigarReads
Base Recalibration
Analysis-Ready RNAseq Reads
Analysis-Ready RNAseq Reads
Variant Calling
HaplotypeCaller
Raw Variants
SNPs
Indels
Variant Filtering
RNAseq-specific settings
Filtered Variants
SNPs
Indels
Analysis-Ready Variants
SNPs &Indels
Genotype Refinement
Variant Annotation
Variant Evaluation
Look good?
Troubleshoot
Use in project

STAR INSTALLATION

```
wget https://github.com/alexdobin/STAR/archive/2.7.10b.tar.gz
tar -xzf 2.7.10b.tar.gz
cd STAR-2.7.10b

cd source
make STAR

alias STAR='/path/to/star-package/source/STAR'.
```

conda install

To install this package run one of the following:

```
conda install -c bioconda star
```

```
conda install -c "bioconda/label/cf201901" star
```

https://anaconda.org/bioconda/star

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

RAW DATA

European Nucleotide Archive

Enter text search terms | Search

Examples: histone, BN000065

PRJEB5348 | View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

The ENA Advanced Search API changed on 2023-05-02! Details here.

Project: PRJEB5348

High-throughput RNA sequencing (RNA-seq) is now the standard method to determine differential gene expression. Here, a 48 replicate, two condition RNA-seq experiment was designed specifically to test assumptions about RNA-seq read count variability models and the performance of methods for differential gene expression analysis by RNA-seq. Samples were run on an Illumina HiSeq for 50 cycles single-end and included ERCC RNA spike-ins. The high-replicate data allowed for strict quality control and screening of 'bad' replicates. The experiment allowed the effect of bad replicates to be assessed as well as providing guidelines for the number of replicates required for differential gene expression analysis and the most appropriate statistical tools. The mapping between technical replicates and biological replicates is provided via FigShare http://dx.doi.org/10.6084/m9.figshare.1416210 The gene read counts are also available on FigShare: https://dx.doi.org/10.6084/m9.figshare.1425503 https://dx.doi.org/10.6084/m9.figshare.1425502

Show Less

View: XML XML (STUDY)

Download: XML XML (STUDY)

Navigation: Show

Read Files: Hide

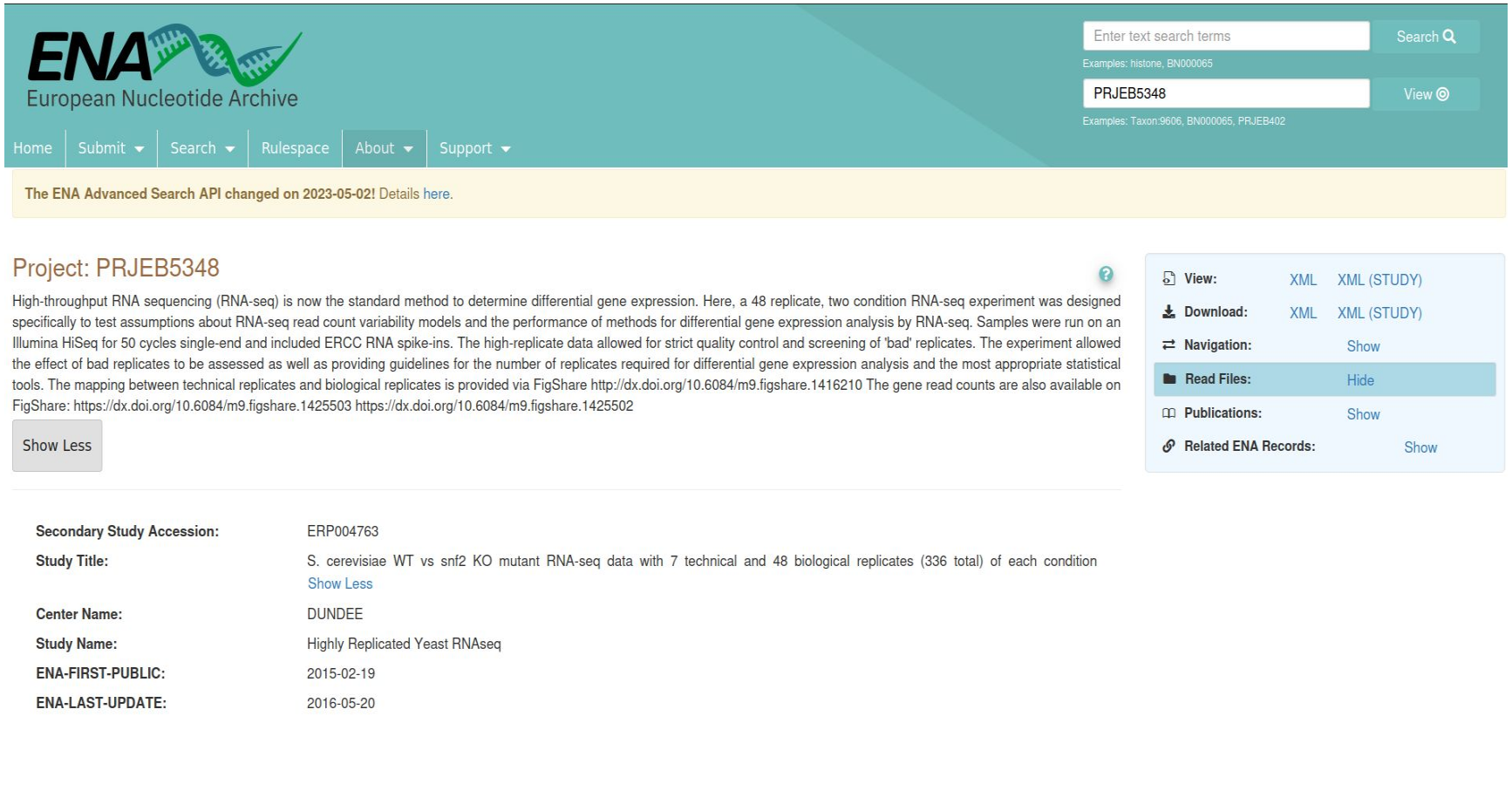
Publications: Show

Related ENA Records: Show

Secondary Study Accession:	ERP004763
Study Title:	S. cerevisiae WT vs snf2 KO mutant RNA-seq data with 7 technical and 48 biological replicates (336 total) of each condition Show Less
Center Name:	DUNDEE
Study Name:	Highly Replicated Yeast RNAseq
ENA-FIRST-PUBLIC:	2015-02-19
ENA-LAST-UPDATE:	2016-05-20

https://www.ebi.ac.uk/ena/browser/view/PRJEB5348

RAW DATA



	A	B	C	D
1	RunAccession	Lane	Sample	BiolRep
2	ERR458493	1	WT	1
3	ERR458494	2	WT	1
4	ERR458495	3	WT	1
5	ERR458496	4	WT	1
6	ERR458497	5	WT	1
7	ERR458498	6	WT	1
8	ERR458499	7	WT	1
9	ERR458500	1	SNF2	1
10	ERR458501	2	SNF2	1
11	ERR458502	3	SNF2	1
12	ERR458503	4	SNF2	1
13	ERR458504	5	SNF2	1
14	ERR458505	6	SNF2	1

2194841.tsv

ERP004763_sample_mapping.tsv (12.38 kB)

https://figshare.com/articles/dataset/Metadata_for_a_highly_replicated_two_condition_yeast_RNAseq_experiment_/1416210

What is inside GCF_000146045.2_R64_genomic.fna ?

```
cat GCF_000146045.2_R64_genomic.fna | grep '>'

>NC_001133.9 Saccharomyces cerevisiae S288C chromosome I, complete sequence
>NC_001134.8 Saccharomyces cerevisiae S288C chromosome II, complete sequence
>NC_001135.5 Saccharomyces cerevisiae S288C chromosome III, complete sequence
>NC_001136.10 Saccharomyces cerevisiae S288C chromosome IV, complete sequence
>NC_001137.3 Saccharomyces cerevisiae S288C chromosome V, complete sequence
>NC_001138.5 Saccharomyces cerevisiae S288C chromosome VI, complete sequence
>NC_001139.9 Saccharomyces cerevisiae S288C chromosome VII, complete sequence
>NC_001140.6 Saccharomyces cerevisiae S288C chromosome VIII, complete sequence
>NC_001141.2 Saccharomyces cerevisiae S288C chromosome IX, complete sequence
>NC_001142.9 Saccharomyces cerevisiae S288C chromosome X, complete sequence
>NC_001143.9 Saccharomyces cerevisiae S288C chromosome XI, complete sequence
>NC_001144.5 Saccharomyces cerevisiae S288C chromosome XII, complete sequence
>NC_001145.3 Saccharomyces cerevisiae S288C chromosome XIII, complete sequence
>NC_001146.8 Saccharomyces cerevisiae S288C chromosome XIV, complete sequence
>NC_001147.6 Saccharomyces cerevisiae S288C chromosome XV, complete sequence
>NC_001148.4 Saccharomyces cerevisiae S288C chromosome XVI, complete sequence
>NC_001224.1 Saccharomyces cerevisiae S288c mitochondrion, complete genome
```

What is inside GCF_000146045.2_R64_genomic.fna ?

```
cat GCF_000146045.2_R64_genomic.fna | head

>NC_001133.9 Saccharomyces cerevisiae S288C chromosome I, complete sequence
ccacaccacacccacacacccacacaccacacaccacacaccacacccacacacacacatCCTAACACTACCCTAAC
ACAGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCAT
TCAACCATACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATC
CAACCCACTGCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATACTGTTCTTCTACCCACCATAT
TGAAACGCTAACAAATGATCGTAAATAACACACACGTGCTTACCCTACCACTTTATACCACCACCACACATGCCATACTCAC
CCTCACTTGTATACTGATTTTACGTACGCACACGGATGCTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTC
CACTTCACTCCATGGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATGCACGGCACTTGCCTCAGCGG
TCTATACCCTGTGCCATTTACCCATAACGCCCATCATTATCCACATTTTGATATCTATATCTCATTCGGCGGTcccaaat
attgtataaCTGCCCTTAATACATACGTTATACCACTTTTGCACCATATACTTACCACTCCATTTATATACACTTATGTC
```

What is inside GCF_000146045.2_R64_genomic.gtf ?

```
cat GCF_000146045.2_R64_genomic.gtf | head | column -t

NC_001133.9  RefSeq  transcript  1807   2169   .  -  .  transcript_id  "rna-NM_001180043.1";  gene_id  "gene-YAL068C";    gene_name  "PAU8"
NC_001133.9  RefSeq  exon        1807   2169   .  -  .  transcript_id  "rna-NM_001180043.1";  gene_id  "gene-YAL068C";    gene_name  "PAU8";
NC_001133.9  RefSeq  CDS         1807   2169   .  -  0  transcript_id  "rna-NM_001180043.1";  gene_id  "gene-YAL068C";    gene_name  "PAU8";
NC_001133.9  RefSeq  transcript  2480   2707   .  +  .  transcript_id  "rna-NM_001184582.1";  gene_id  "gene-YAL067W-A"
NC_001133.9  RefSeq  exon        2480   2707   .  +  .  transcript_id  "rna-NM_001184582.1";  gene_id  "gene-YAL067W-A";
NC_001133.9  RefSeq  CDS         2480   2707   .  +  0  transcript_id  "rna-NM_001184582.1";  gene_id  "gene-YAL067W-A";
NC_001133.9  RefSeq  transcript  7235   9016   .  -  .  transcript_id  "rna-NM_001178208.1";  gene_id  "gene-YAL067C";    gene_name  "SEO1"
NC_001133.9  RefSeq  exon        7235   9016   .  -  .  transcript_id  "rna-NM_001178208.1";  gene_id  "gene-YAL067C";    gene_name  "SEO1";
NC_001133.9  RefSeq  CDS         7235   9016   .  -  0  transcript_id  "rna-NM_001178208.1";  gene_id  "gene-YAL067C";    gene_name  "SEO1";
NC_001133.9  RefSeq  transcript  11565  11951  .  -  .  transcript_id  "rna-NM_001179897.1";  gene_id  "gene-YAL065C"
```

- **NC_001133.9:** This is the reference genome sequence identifier (chromosome) where the feature is located.

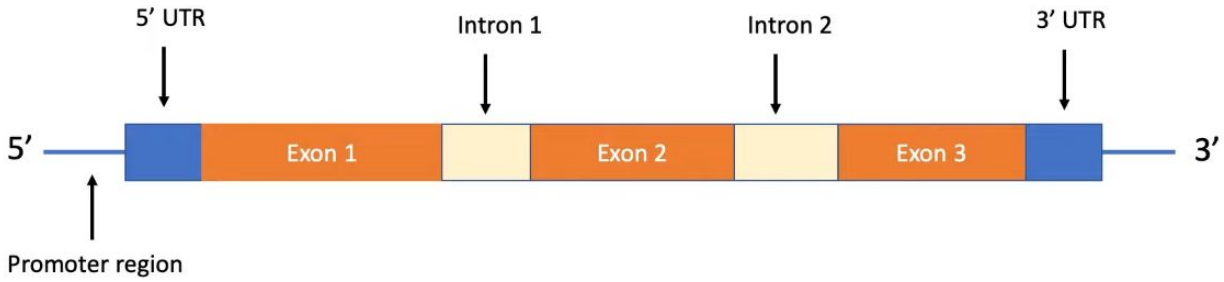
- RefSeq: The source of the annotation, in this case, it is labeled as RefSeq.

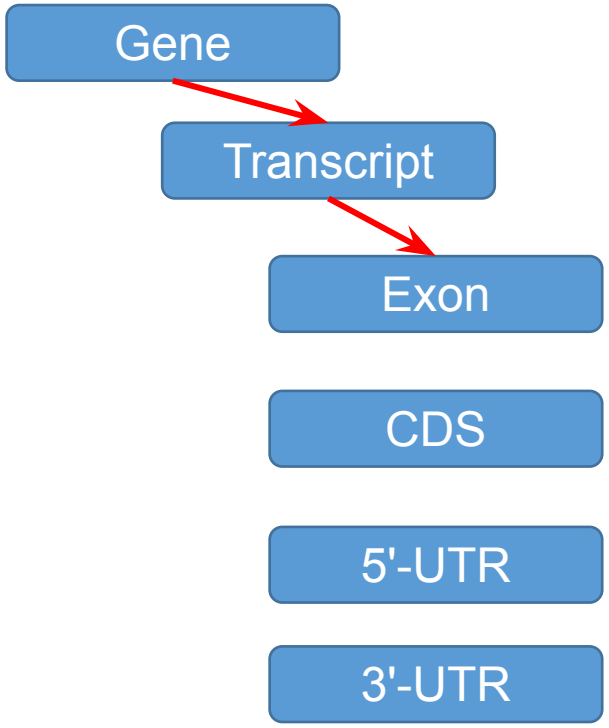
- **transcript**: This indicates that the feature is a transcript, representing a complete or partial RNA molecule transcribed from a gene.

- **1807 and 2169**: These are the start and end positions of the transcript on the reference genome (chromosome). The transcript spans from position 1807 to position 2169

Homo_sapiens.GRCh38.109.gtf.gz

```
zcat Homo_sapiens.GRCh38.109.gtf.gz | grep -v "#" | head -30 | column -t

1  ensembl_havana  gene          1471765  1497848  .  +  .  gene_id  "ENSG00000160072";  gene_version  "20";  gene_name
1  ensembl_havana  transcript    1471765  1497848  .  +  .  gene_id  "ENSG00000160072";  gene_version  "20";  transcript_id
1  ensembl_havana  exon          1471765  1472089  .  +  .  gene_id  "ENSG00000160072";  gene_version  "20";  transcript_id
1  ensembl_havana  CDS           1471885  1472089  .  +  0  gene_id  "ENSG00000160072";  gene_version  "20";  transcript_id
1  ensembl_havana  start_codon   1471885  1471887  .  +  0  gene_id  "ENSG00000160072";  gene_version  "20";  transcript_id
1  ensembl_havana  exon          1477274  1477350  .  +  .  gene_id  "ENSG00000160072";  gene_version  "20";  transcript_id
1  ensembl_havana  CDS           1477274  1477350  .  +  2  gene_id  "ENSG00000160072";  gene_version  "20";  transcript_id
1  ensembl_havana  exon          1478644  1478745  .  +  .  gene_id  "ENSG00000160072";  gene_version  "20";  transcript_id
```





https://www.youtube.com/watch?v=2Y7zxh1BiFs

GENCODE

GENCODE

Human | Mouse | How to access data | FAQ | Documentation | About us

Human

Release 44 (GRCh38.p14)

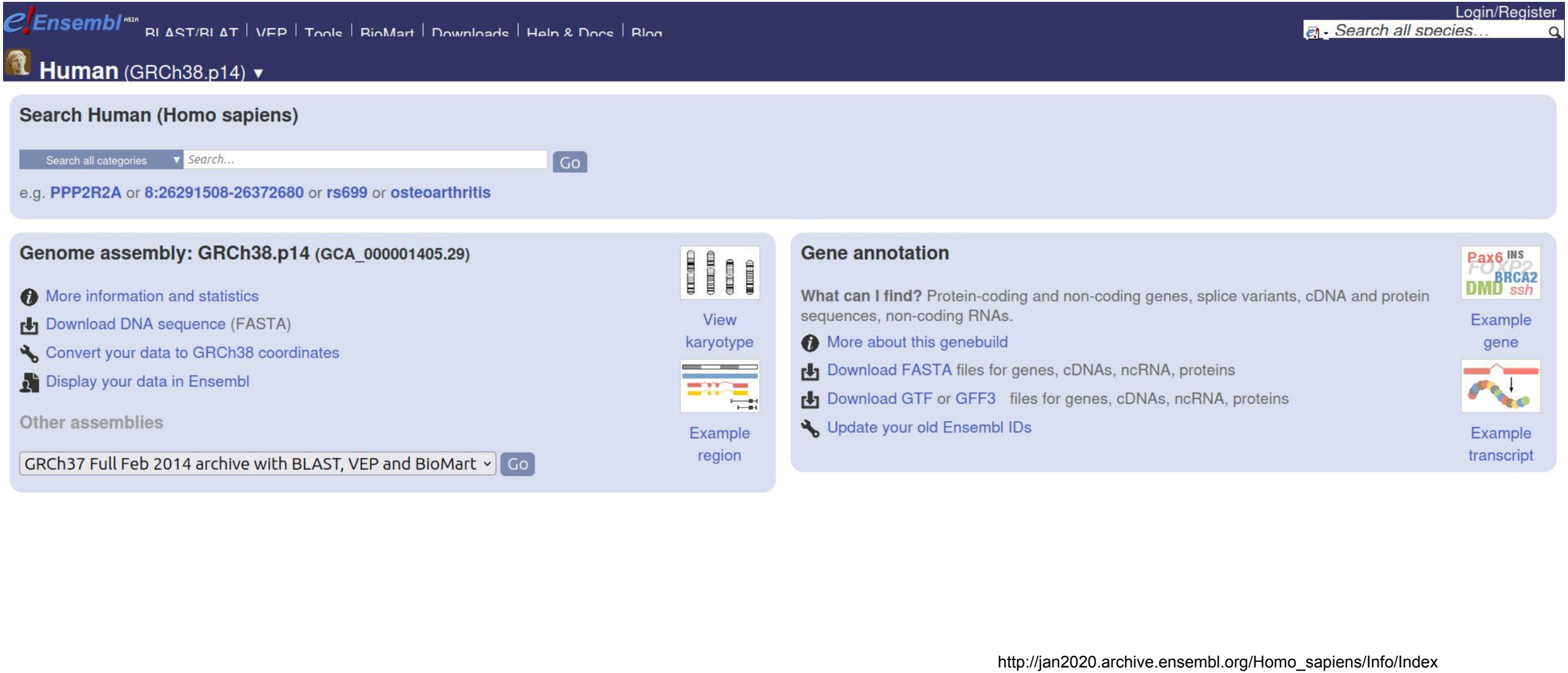
- Statistics of this release
- More information about this assembly (including patches, scaffolds and haplotypes)
- Go to GRCh37 version of this release

GTF / GFF3 files

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes)	GTF GFF3

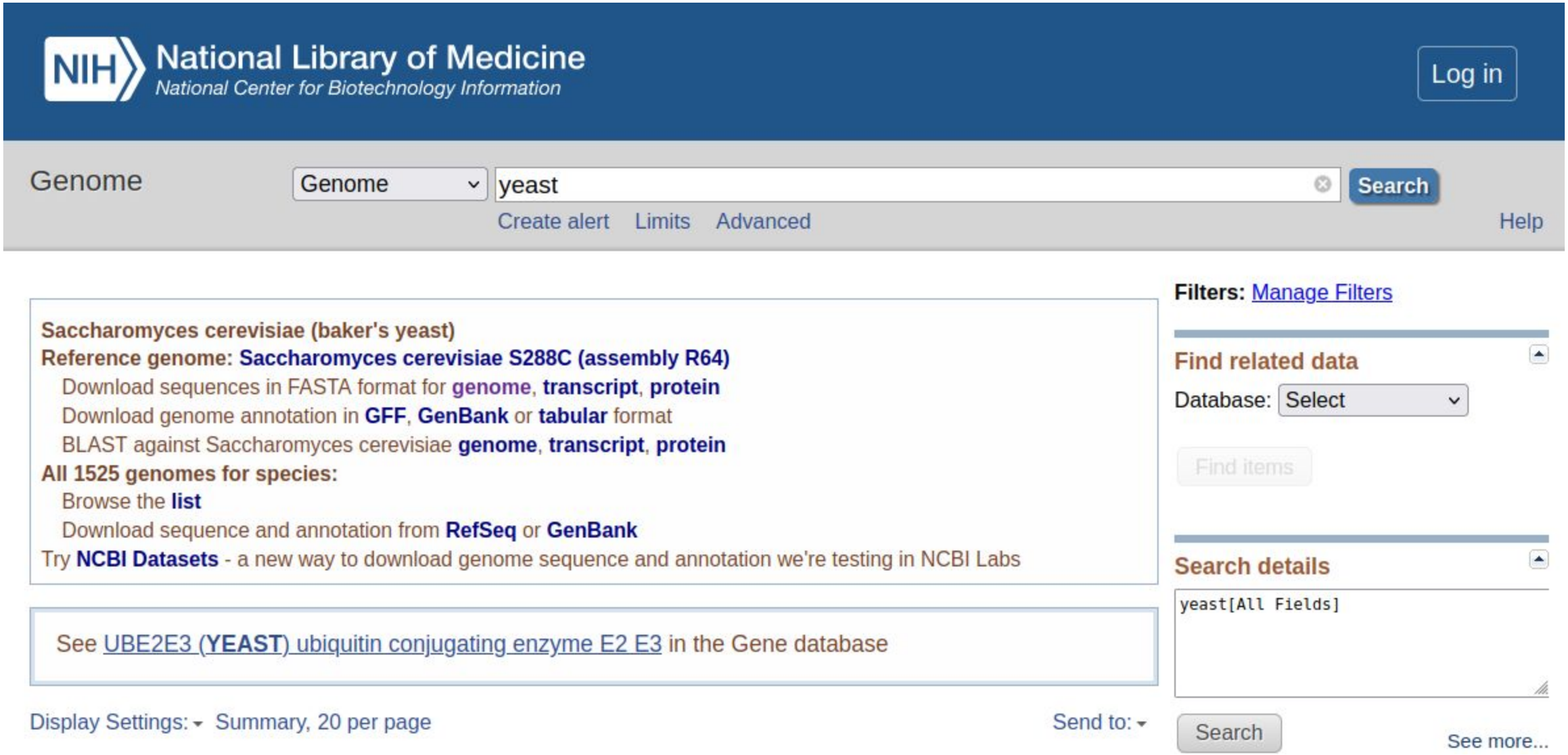
https://www.gencodegenes.org/human/

ENSEMBL



http://jan2020.archive.ensembl.org/Homo_sapiens/Info/Index

NCBI



https://www.ncbi.nlm.nih.gov/genome/?term=yeast

UCSC GENOME BROWSER

Sequence, Annotation, and Other Downloads

This page contains links to sequence and annotation downloads for the genome assemblies featured in the UCSC Genome Browser. available in the **Table Browser** or via the command-line **utilities**.

For access to the most recent assembly of each genome, see the **current genomes** directory. Previous versions of certain data are a (Genome Archive) species data can be found **here**. All data in the Genome Browser are freely usable for any purpose except as indic researchers, as listed on the Genome Browser **credits** page. Please acknowledge the contributor(s) of the data you use.

Human | **Mouse**

SARS-CoV-2 (COVID) | **Zebrafish**

Fruit fly | **Mammals**

Human genomes

Jan. 2022 (T2T-CHM13 v2.0/hs1)

- **Fileserver (bigBed, maf, fa, etc) annotations**
- **Standard genome sequence files and select annotations (2bit, GTF, GC-content, etc)**
- **LiftOver files**
- **Pairwise alignments**

Dec. 2013 (GRCh38/hg38)

- **Genome sequence files and select annotations (2bit, GTF, GC-content, etc)**
- **Sequence data by chromosome**
- **Annotations**

https://hgdownload.soe.ucsc.edu/downloads.html

1. GENERATE GENOME INDEX

```
##Create genome index
STAR --runMode genomeGenerate
     --genomeDir /media/bacdao/Data1/Rna_seq_yeast/starIndex
     --genomeFastaFiles GCF_000146045.2_R64_genomic.fna
     --sjdbGTFfile GCF_000146045.2_R64_genomic.gff
     --runThreadN 4
```

--runMode genomeGenerate: This option indicates that the command is for generating a genome index.

--genomeDir: This specifies the directory where the genome index files will be stored.

--genomeFastaFiles: This option specifies the path to the reference genome FASTA file. The reference genome is the DNA sequence of the organism being studied.

--sjdbGTFfile: This option specifies the path to the annotation file in GTF (Gene Transfer Format) or GFF (General Feature Format) format.

--runThreadN: This option specifies the number of threads (CPU cores) to use for the alignment process.

2. STAR ALIGNMENT

```
STAR --genomeDir $GENOMEDIR --readFilesIn $FILES --readFilesCommand zcat
    --outFileNamePrefix alignment_STAR/WT_1/ --outFilterMultimapNmax 1 --outReadsUnmapped Fastx
    --outSAMtype BAM SortedByCoordinate
    --runThreadN 16
    --twopassMode Basic
```

--genomeDir: Specifies the directory where the genome index files generated by STAR are located.

--readFilesIn: Specifies the paths to the fastq files.

--readFilesCommand: Specifies the command to use for reading compressed fastq files.

--outFileNamePrefix: Specifies the output directory and filename prefix for the aligned reads.

--outFilterMultimapNmax: Sets the maximum number of loci a read can map to. Here, it's set to 1, which means reads will only be aligned to one location.

--outReadsUnmapped Fastx: Requests that the unmapped reads be written to an output file in fastq format.

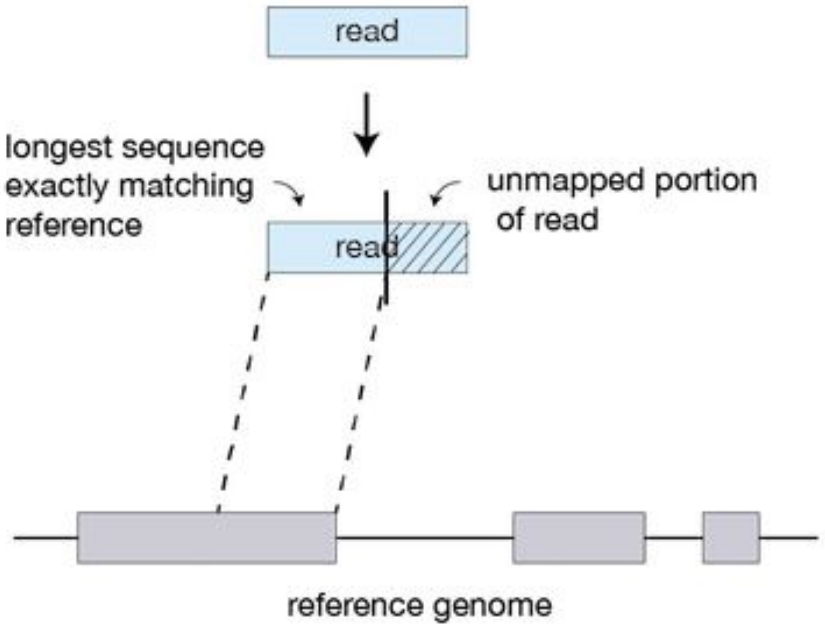
--outSAMtype: Specifies the output format for the aligned reads.

--runThreadN 16: Specifies the number of threads (CPU cores) to use for the alignment. In this case, 16 threads will be used.

--twopassMode: Specifies the mode for the two-pass alignment strategy.

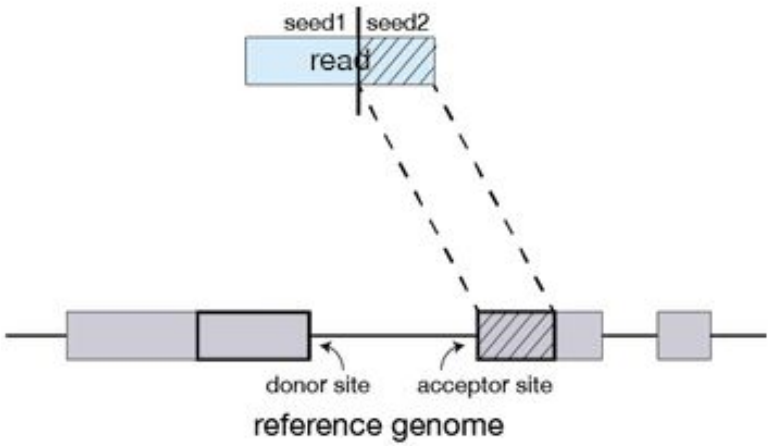
Seed searching

For every read that STAR aligns, STAR will search for the longest sequence that exactly matches one or more locations on the reference genome. These longest matching sequences are called the Maximal Mappable Prefixes (MMPs):



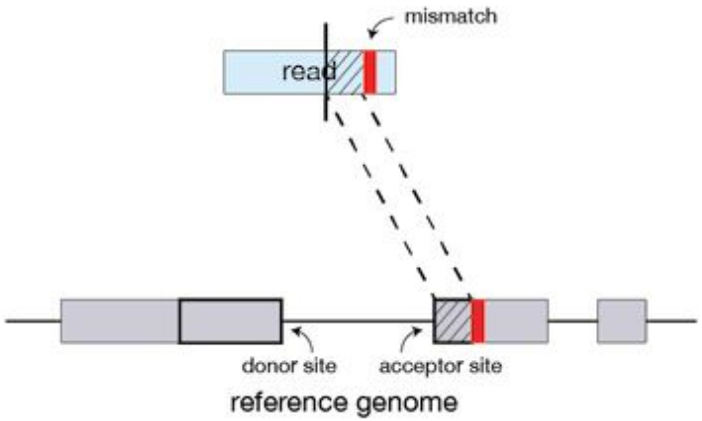
https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/STAR_alignment_strategy.html

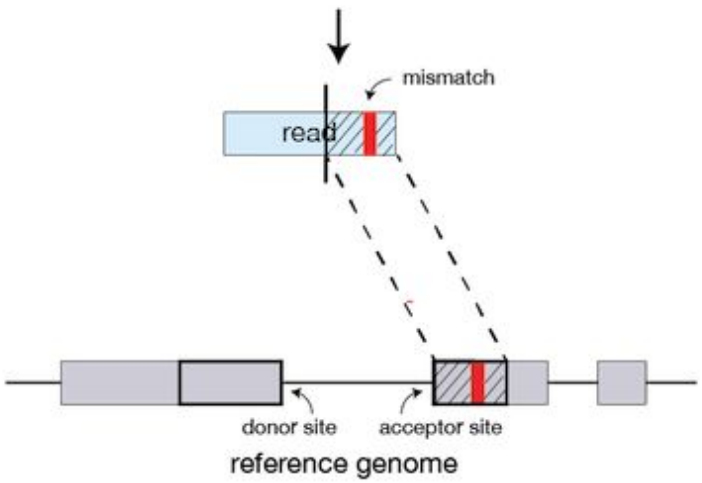
STAR will then search again for only the unmapped portion of the read to find the next longest sequence that exactly matches the reference genome, or the next MMP, which will be *seed2*.

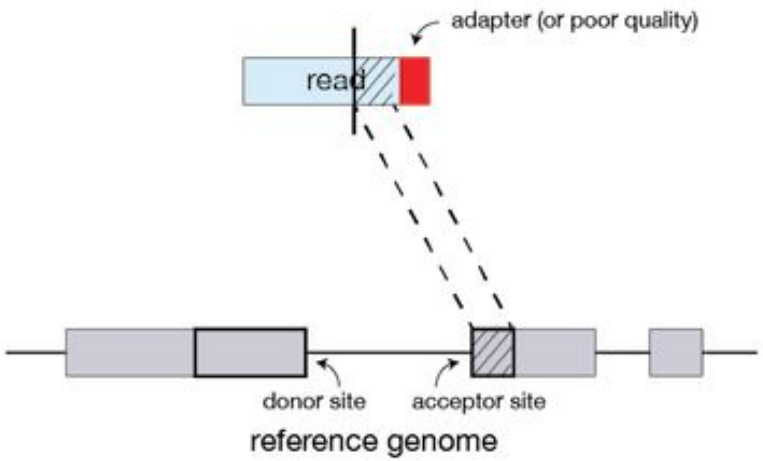


This sequential searching of only the unmapped portions of reads underlies the efficiency of the STAR algorithm. STAR uses an uncompressed suffix array (SA) to efficiently search for the MMPs, this allows for quick searching against even the largest reference genomes. Other slower aligners use algorithms that often search for the entire read sequence before splitting reads and performing iterative rounds of mapping.

If STAR does not find an exact matching sequence for each part of the read due to mismatches or indels, the previous MMPs will be extended.





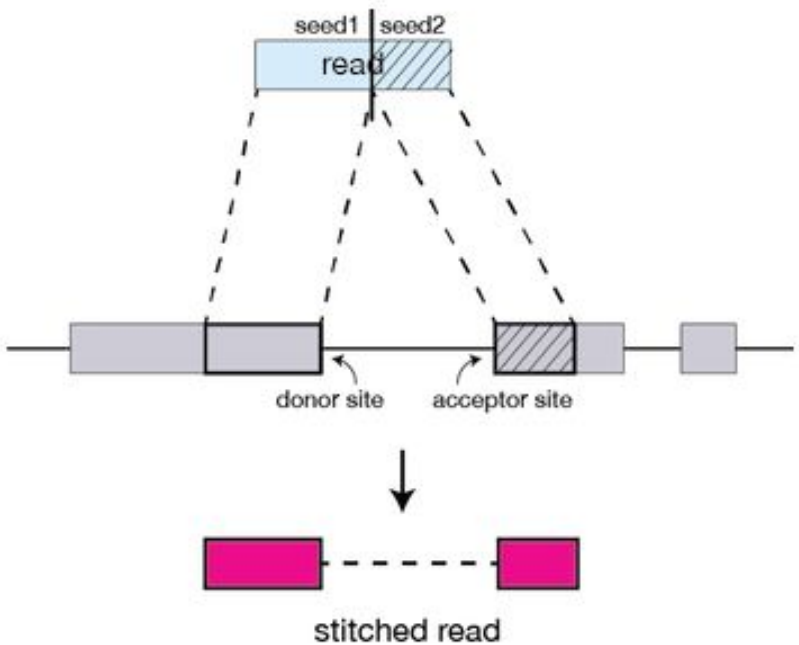


If extension does not give a good alignment, then the poor quality or adapter sequence (or other contaminating sequence) will be soft clipped.

Clustering, stitching, and scoring

The separate seeds are stitched together to create a complete read by first clustering the seeds together based on proximity to a set of ‘anchor’ seeds, or seeds that are not multi-mapping.

Then the seeds are stitched together based on the best alignment for the read (scoring based on mismatches, indels, gaps, etc.).



OUTPUT BAM FILE

```
samtools view -F 4 WT_1/Aligned.sortedByCoord.out.bam | head | column -t

ERR458493.552967  16  NC_001133.9  140   255  12M61232N37M2S  *  0  0  CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC  BB?HHJJIGHHJIGIIJJIJGIJIJJIIIGHBJJJJJJHHHHFFDDDA1+B
ERR458498.814362  0   NC_001133.9  1593  255  11M71474N40M    *  0  0  TTTCTACAAAGATTCTCTCACTTGTACAGAGGTGTCTTCCCATTCGTTTTC  CCCFFFFFHHHHHJJGIJJJIJJIHJJJIJJJHHIJJJJJJJIJJJGIJIG
ERR458496.427513  16  NC_001133.9  3782  255  51M             *  0  0  CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGAAACGCTAAG  JJJJIJJJJIJJJJJJJJIJJJJJJJGHBJJJJJJJHHHHHFFFFFCCB
ERR458493.243111  0   NC_001133.9  3873  255  51M             *  0  0  TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACCAAGGGACAA  @@<DDDDEH<<<CEFHAEIHHGGCHIGEEGG?@FF@<?ECGGGIIHIIIII
ERR458494.816646  16  NC_001133.9  3972  255  51M             *  0  0  TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAG  @BBEBG;IGFDD@C:BHC<4A?94C<CH<EEEFABBFE<CDCBBD;DD=?1
ERR458497.736781  16  NC_001133.9  3976  255  51M             *  0  0  GAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAGTTTG  IJJJJJJJJJJJJJJJJJJIIJJJJJJJJJJJJJJJJHHHHHFFFFFCCC
ERR458495.31401   0   NC_001133.9  4273  255  51M             *  0  0  CGGCAATCGCTTGCGGTTATGGCACGAAGAACAATGCAATAGCTCTTACAA  ?<+AD@BBF?F?BACGGGHI@DGHGHGEGGD>;BDG38BB*==8BGECHGE
ERR458495.334443  16  NC_001133.9  4944  255  51M             *  0  0  CCAGCCAGAAAATGCACATCATAAAGACATGTTATTCAAGAGCCACGACCG  GDBIIHAHGHF>ACIIGHIGCEHEHEEGH>G<EFC?EHDFFDDDDDD7@@@
ERR458498.150487  16  NC_001133.9  4950  255  51M             *  0  0  AGAAAATGCACATCATAAAGACATGTTATTCAAGAGCCACGACCGTCTTCA  <JIJJIJJIGJJFIIGIIGF9IIIIGIHGIJIIHEGHF@FFDGFFFFD@@@
ERR458493.531202  0   NC_001133.9  5198  255  51M             *  0  0  GTTGTGGTCACTAATGAGAACTTGAATAGTTTTCAACTGCTGGTGATAAAT  BBCFFFFFHHHHHJJJJJJJJJJEHJJJJHIJJJJJJJJJJJJJHIIJJJJJ
```

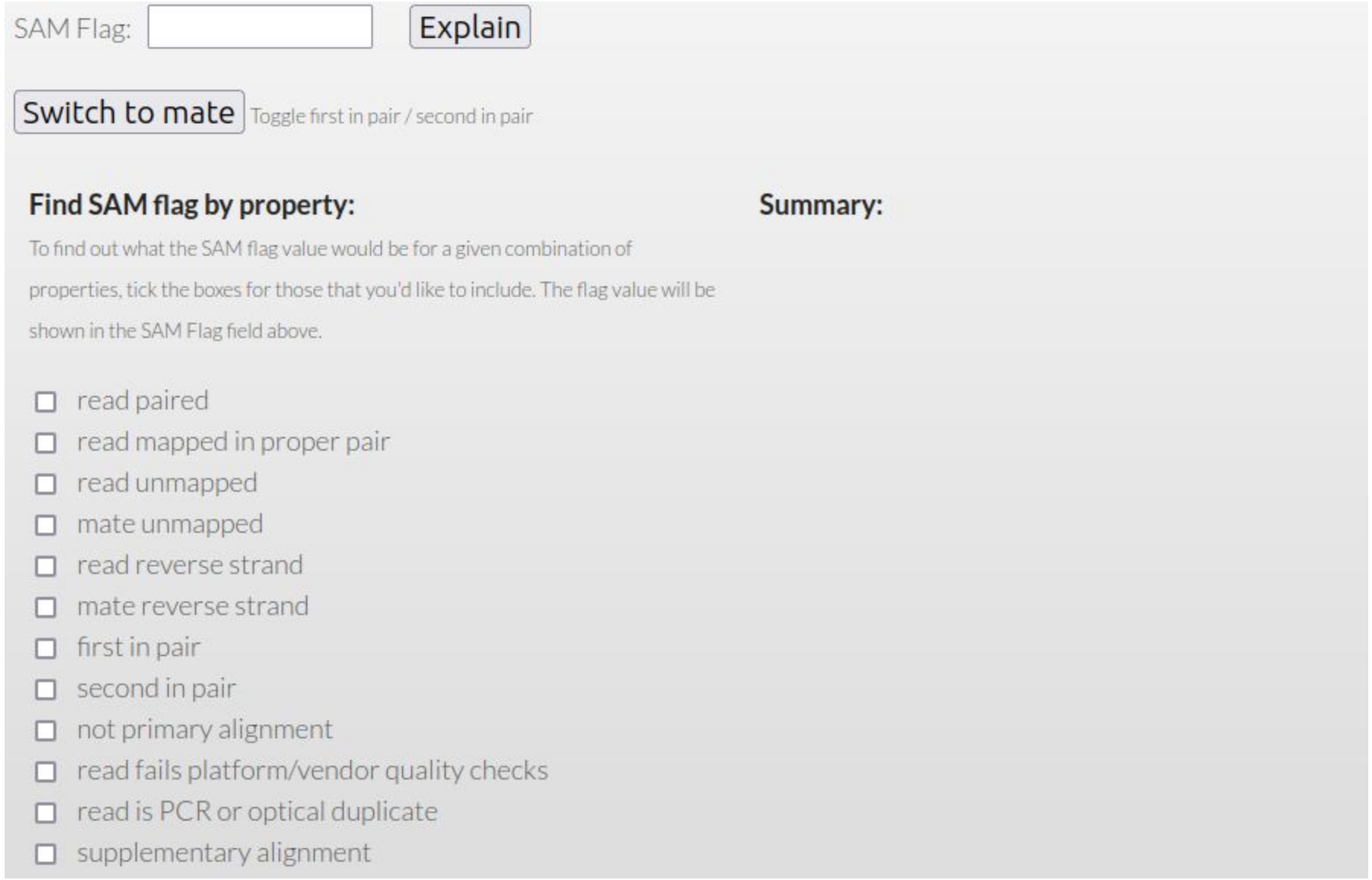
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

OUTPUT BAM FILE

```
samtools view -F 4 WT_1/Aligned.sortedByCoord.out.bam | head | column -t

ERR458493.552967  16  NC_001133.9  140   255  12M61232N37M2S  *  0  0  CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC  BB?HHJJIGHHJIGIIJJIJGIJIJJIIIGHBJJJJJJHHHHFFDDDA1+B
ERR458498.814362  0   NC_001133.9  1593  255  11M71474N40M    *  0  0  TTTCTACAAAGATTCTCTCACTTGTACAGAGGTGTCTTCCCATTCGTTTTC  CCCFFFFFHHHHHJJGIJJJIJJIHJJJIJJJHHIJJJJJJJIJJJGIJIG
ERR458496.427513  16  NC_001133.9  3782  255  51M             *  0  0  CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGAAACGCTAAG  JJJJIJJJJIJJJJJJJJIJJJJJJJGHBJJJJJJJHHHHHFFFFFCCB
ERR458493.243111  0   NC_001133.9  3873  255  51M             *  0  0  TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACCAAGGGACAA  @@<DDDDEH<<<CEFHAEIHHGGCHIGEEGG?@FF@<?ECGGGIIHIIIII
ERR458494.816646  16  NC_001133.9  3972  255  51M             *  0  0  TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAG  @BBEBG;IGFDD@C:BHC<4A?94C<CH<EEEFABBFE<CDCBBD;DD=?1
ERR458497.736781  16  NC_001133.9  3976  255  51M             *  0  0  GAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAGTTTG  IJJJJJJJJJJJJJJJJJJIIJJJJJJJJJJJJJJJJHHHHHFFFFFCCC
ERR458495.31401   0   NC_001133.9  4273  255  51M             *  0  0  CGGCAATCGCTTGCGGTTATGGCACGAAGAACAATGCAATAGCTCTTACAA  ?<+AD@BBF?F?BACGGGHI@DGHGHGEGGD>;BDG38BB*==8BGECHGE
ERR458495.334443  16  NC_001133.9  4944  255  51M             *  0  0  CCAGCCAGAAAATGCACATCATAAAGACATGTTATTCAAGAGCCACGACCG  GDBIIHAHGHF>ACIIGHIGCEHEHEEGH>G<EFC?EHDFFDDDDDD7@@@
ERR458498.150487  16  NC_001133.9  4950  255  51M             *  0  0  AGAAAATGCACATCATAAAGACATGTTATTCAAGAGCCACGACCGTCTTCA  <JIJJIJJIGJJFIIGIIGF9IIIIGIHGIJIIHEGHF@FFDGFFFFD@@@
ERR458493.531202  0   NC_001133.9  5198  255  51M             *  0  0  GTTGTGGTCACTAATGAGAACTTGAATAGTTTTCAACTGCTGGTGATAAAT  BBCFFFFFHHHHHJJJJJJJJJJEHJJJJHIJJJJJJJJJJJJJHIIJJJJJ
```

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

FLAG EXPLAINATION



https://broadinstitute.github.io/picard/explain-flags.html

OUTPUT BAM FILE

```
samtools view -F 4 WT_1/Aligned.sortedByCoord.out.bam | head | column -t

ERR458493.552967  16  NC_001133.9  140   255  12M61232N37M2S  *  0  0  CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC  BB?HHJJIGHHJIGIIJJIJGIJIJJIIIGHBJJJJJJHHHHFFDDDA1+B
ERR458498.814362  0   NC_001133.9  1593  255  11M71474N40M    *  0  0  TTTCTACAAAGATTCTCTCACTTGTACAGAGGTGTCTTCCCATTCGTTTTC  CCCFFFFFHHHHHJJGIJJJIJJIHJJJIJJJHHIJJJJJJJIJJJGIJIG
ERR458496.427513  16  NC_001133.9  3782  255  51M             *  0  0  CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGAAACGCTAAG  JJJJIJJJJIJJJJJJJJIJJJJJJJGHBJJJJJJJHHHHHFFFFFCCB
ERR458493.243111  0   NC_001133.9  3873  255  51M             *  0  0  TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACCAAGGGACAA  @@<DDDDEH<<<CEFHAEIHHGGCHIGEEGG?@FF@<?ECGGGIIHIIIII
ERR458494.816646  16  NC_001133.9  3972  255  51M             *  0  0  TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAG  @BBEBG;IGFDD@C:BHC<4A?94C<CH<EEEFABBFE<CDCBBD;DD=?1
ERR458497.736781  16  NC_001133.9  3976  255  51M             *  0  0  GAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAGTTTG  IJJJJJJJJJJJJJJJJJJIIJJJJJJJJJJJJJJJJHHHHHFFFFFCCC
ERR458495.31401   0   NC_001133.9  4273  255  51M             *  0  0  CGGCAATCGCTTGCGGTTATGGCACGAAGAACAATGCAATAGCTCTTACAA  ?<+AD@BBF?F?BACGGGHI@DGHGHGEGGD>;BDG38BB*==8BGECHGE
ERR458495.334443  16  NC_001133.9  4944  255  51M             *  0  0  CCAGCCAGAAAATGCACATCATAAAGACATGTTATTCAAGAGCCACGACCG  GDBIIHAHGHF>ACIIGHIGCEHEHEEGH>G<EFC?EHDFFDDDDDD7@@@
ERR458498.150487  16  NC_001133.9  4950  255  51M             *  0  0  AGAAAATGCACATCATAAAGACATGTTATTCAAGAGCCACGACCGTCTTCA  <JIJJIJJIGJJFIIGIIGF9IIIIGIHGIJIIHEGHF@FFDGFFFFD@@@
ERR458493.531202  0   NC_001133.9  5198  255  51M             *  0  0  GTTGTGGTCACTAATGAGAACTTGAATAGTTTTCAACTGCTGGTGATAAAT  BBCFFFFFHHHHHJJJJJJJJJJEHJJJJHIJJJJJJJJJJJJJHIIJJJJJ
```

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- 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.

IGV VIZUALIZATION

```
cat Log.final.out
                          Number of input reads |   7014609
                      Average input read length |   51
                                    UNIQUE READS:
                     Uniquely mapped reads number |   6010344
                          Uniquely mapped reads % |   85.68%
                            Average mapped length |   50.73
                         Number of splices: Total |   52174
              Number of splices: Annotated (sjdb) |   51044
                         Number of splices: GT/AG |   51671
                         Number of splices: GC/AG |   104
                         Number of splices: AT/AC |   7
                     Number of splices: Non-canonical |   392
                        Mismatch rate per base, % |   0.36%
                           Deletion rate per base |   0.00%
                          Deletion average length |   1.37
                          Insertion rate per base |   0.00%
                         Insertion average length |   1.04
                                  MULTI-MAPPING READS:
            Number of reads mapped to multiple loci |   0
                 % of reads mapped to multiple loci |   0.00%
            Number of reads mapped to too many loci |   798677
                 % of reads mapped to too many loci |   11.39%
                                  UNMAPPED READS:
       Number of reads unmapped: too many mismatches |   0
            % of reads unmapped: too many mismatches |   0.00%
                 Number of reads unmapped: too short |   203102
                      % of reads unmapped: too short |   2.90%
                     Number of reads unmapped: other |   2486
                          % of reads unmapped: other |   0.04%
                                  CHIMERIC READS:
                            Number of chimeric reads |   0
                                 % of chimeric reads |   0.00%
```

uT : for unmapped reads, reason for not mapping:

0 : no acceptable seed/windows, "Unmapped other" in the Log.final.out

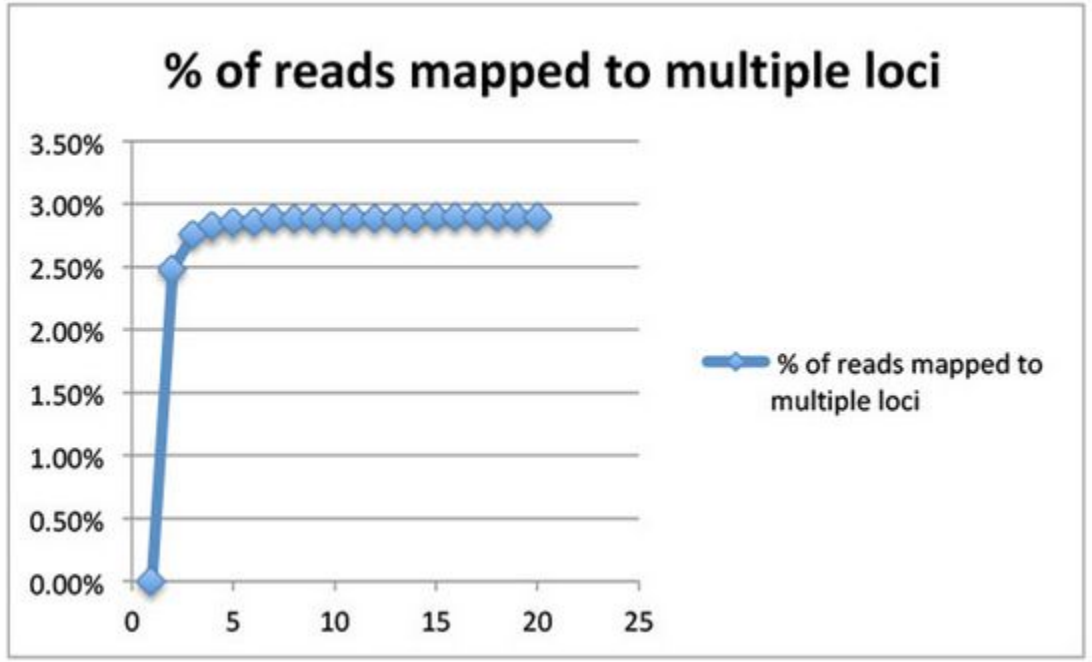
1 : best alignment shorter than min allowed mapped length, "Unmapped: too short" in the Log.final.out

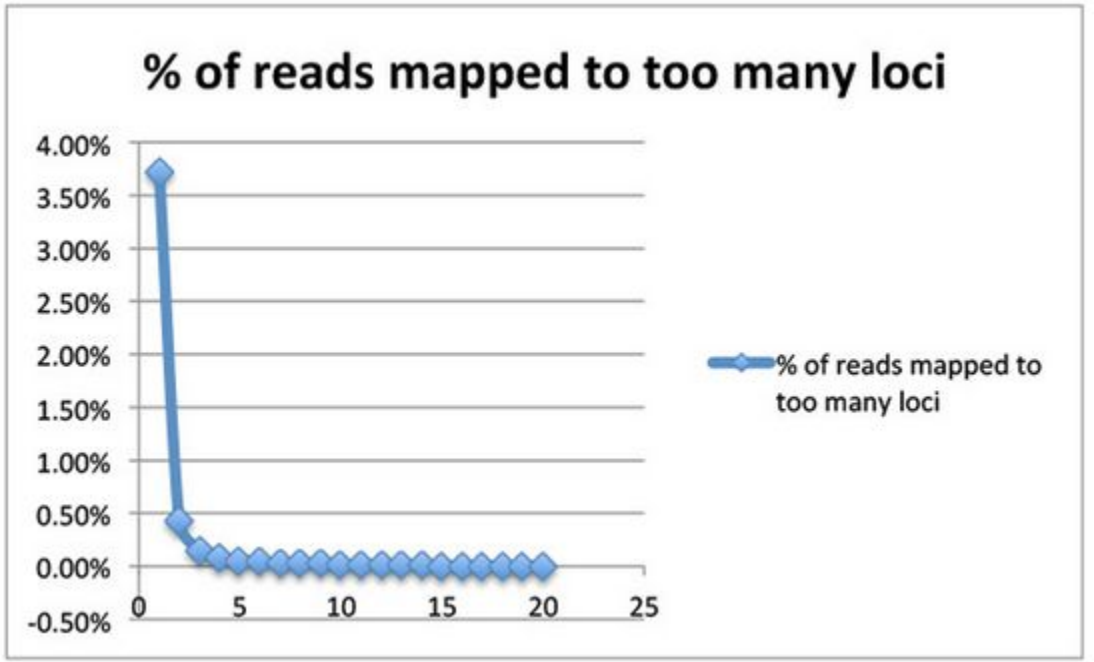
2 : best alignment has more mismatches than max allowed number of mismatches, "Unmapped: too many mismatches" in the Log.final.out

3 : read maps to more loci than the max number of multimappng loci, "Multimapping: mapped to too many loci" in the Log.final.out

4 : unmapped mate of a mapped paired-end read

MultimapNmax	Uniquely mapped reads %	% of reads mapped to multiple loci	% of reads mapped to too many loci	Multi	Too Many
1	94.50%	0.00%	3.72%	1	> 1
2	95.31%	2.48%	0.42%	1 - 2	> 2
3	95.31%	2.75%	0.15%	1 - 3	> 3
4	95.31%	2.82%	0.08%	1 - 4	> 4
5	95.31%	2.85%	0.05%	1 - 5	> 5
6	95.31%	2.86%	0.04%	1 - 6	> 6
7	95.31%	2.88%	0.02%	1 - 7	> 7
8	95.31%	2.88%	0.02%	1 - 8	> 8
9	95.31%	2.88%	0.02%	1- 9	> 9
10	95.31%	2.89%	0.01%	1 - 10	> 10
11	95.31%	2.89%	0.01%	1 - 11	> 11
12	95.31%	2.89%	0.01%	1 - 12	> 12
13	95.31%	2.89%	0.01%	1 - 13	> 13
14	95.31%	2.89%	0.01%	1 - 14	> 14
15	95.31%	2.90%	0.00%	1 - 15	> 15
16	95.31%	2.90%	0.00%	1 - 16	> 16
17	95.31%	2.90%	0.00%	1 - 17	> 17
18	95.31%	2.90%	0.00%	1 - 18	> 18
19	95.31%	2.90%	0.00%	1 - 19	> 19
20	95.31%	2.90%	0.00%	1 - 20	> 20





https://www.biostars.org/p/402546/

ALIGNMENT QC - SAMTOOLS

```
for SAMPLE in WT_1 SNF2_1;
do for i in /media/bacdao/Data1/Rna_seq_yeast/alignment_STAR/${SAMPLE}/Aligned.sortedByCoord.out.bam;
do samtools flagstat $i > ${SAMPLE}/${SAMPLE}_flagstat.out; done; done
```

```
cat WT_1_flagstat.out

6010344 + 0 in total (QC-passed reads + QC-failed reads)
6010344 + 0 primary
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
6010344 + 0 mapped (100.00% : N/A)
6010344 + 0 primary mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

RSeQC

- mRNA is an unstable molecule. It is important to check if your sample has been degraded before performing further analysis.
- RSeQC is a Python package (collection of scripts) that provides some measures of sample/transcript quality:

https://github.com/MonashBioinformaticsPlatform/RSeQC

MAPPING STATISTICS

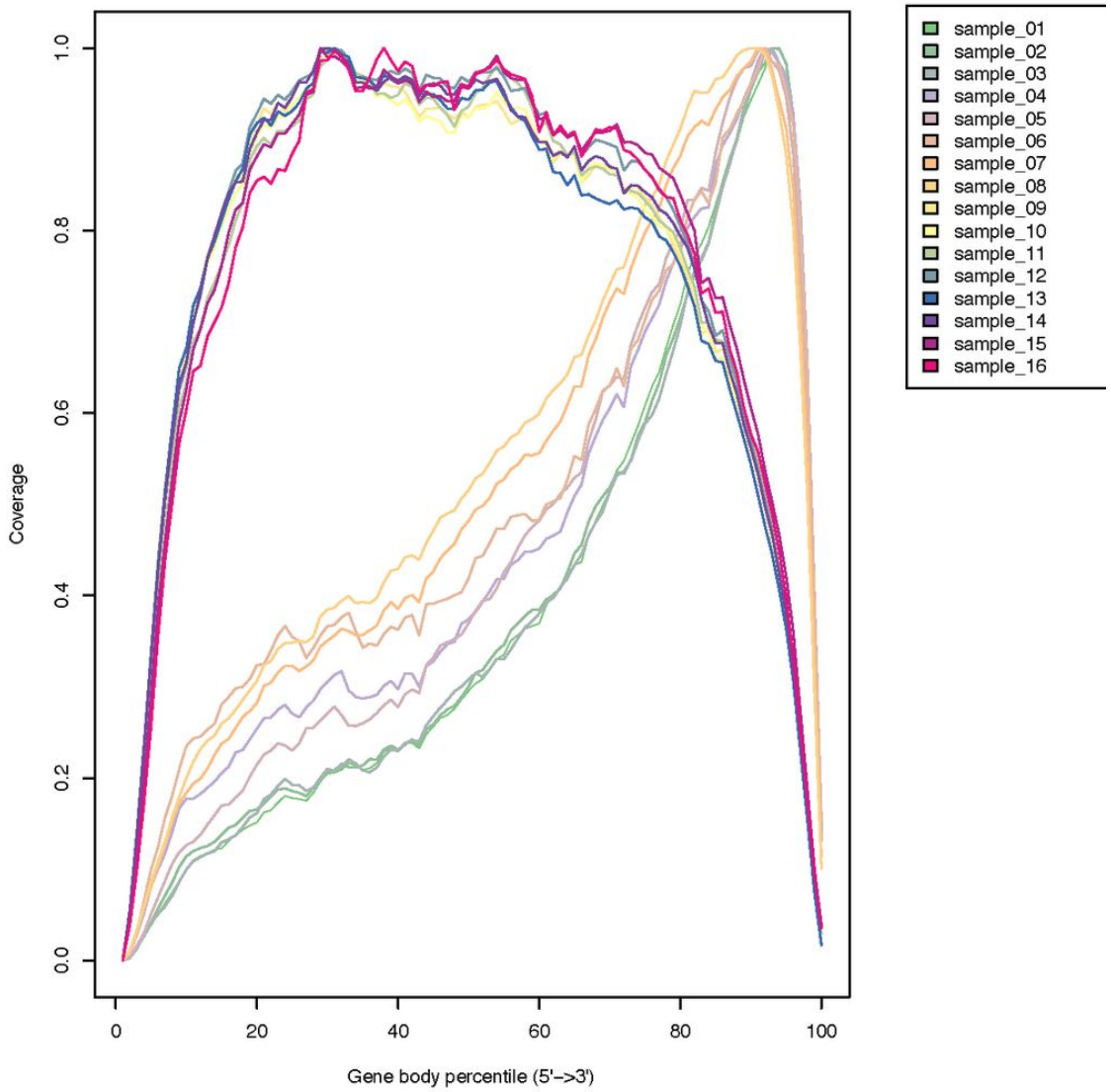
```
#Output (all numbers are read count)
#================================================
Total records:                                41465027
QC failed:                                    0
Optical/PCR duplicate:                        0
Non Primary Hits                              8720455
Unmapped reads:                               0

mapq < mapq_cut (non-unique):                 3127757
mapq >= mapq_cut (unique):                    29616815
Read-1:                                       14841738
Read-2:                                       14775077
Reads map to '+':                             14805391
Reads map to '-':                             14811424
Non-splice reads:                             25455360
Splice reads:                                 4161455
Reads mapped in proper pairs:                 21856264
Proper-paired reads map to different chrom:   7648
```

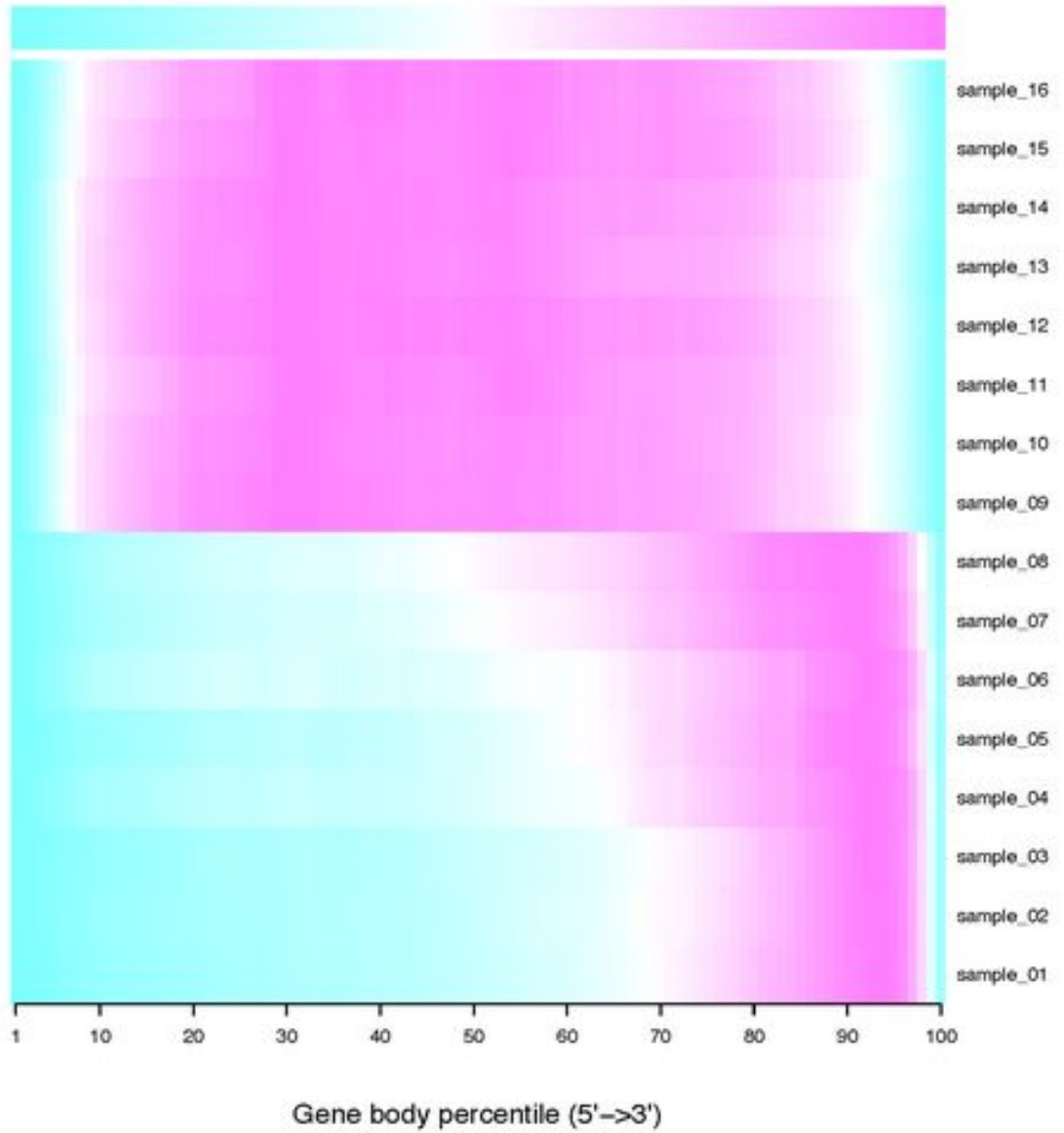
- Non Primary Hits: The number of reads that have multiple alignments to different genomic locations.

- **mapq < mapq_cut (non-unique)**: The number of reads with mapping quality (MAPQ) scores less than a specified threshold (mapq_cut). These reads are considered non-unique alignments.

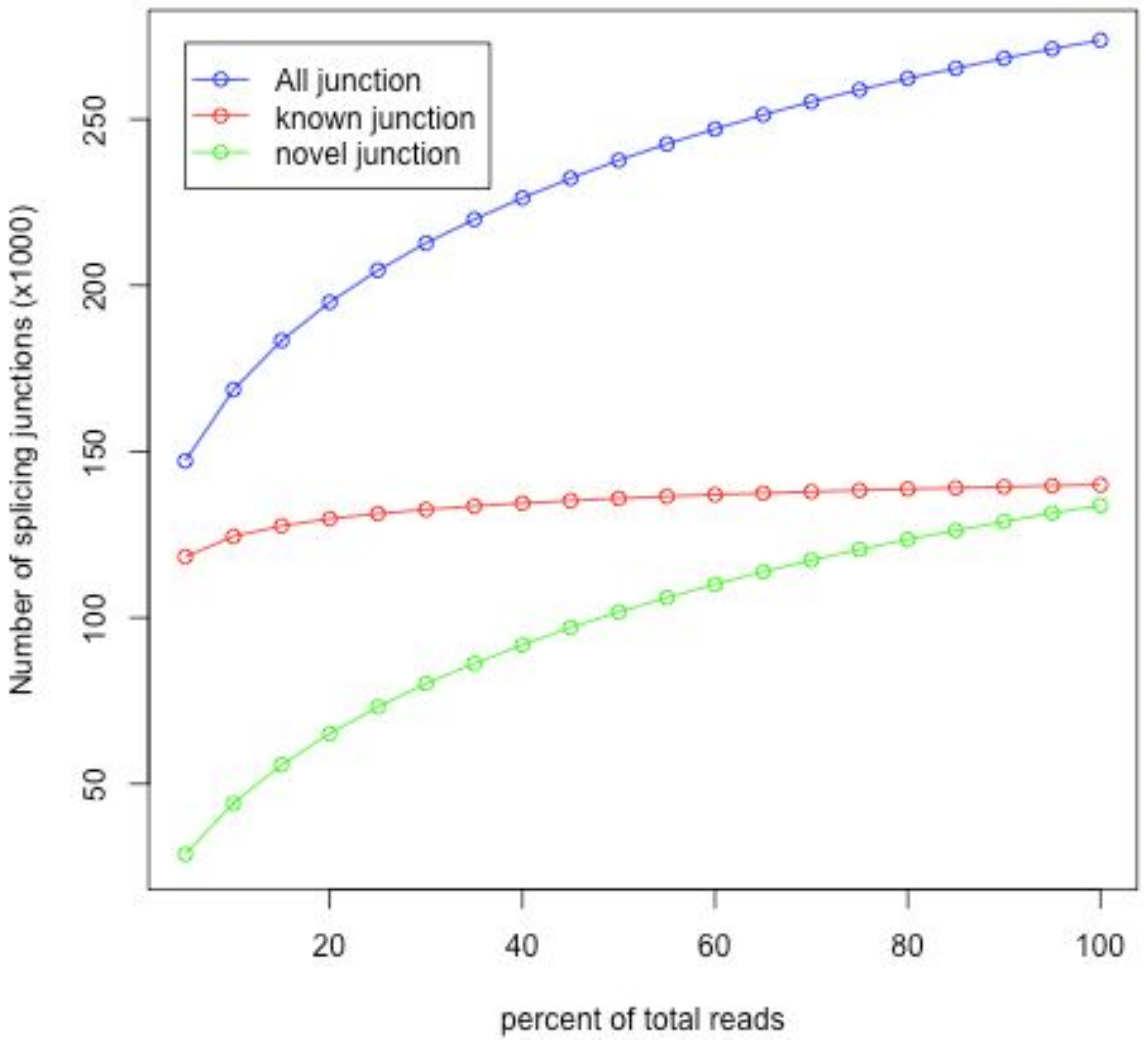
- **Proper-paired reads map to different chrom:** The number of paired-end reads that are properly paired but map to different chromosomes. This could indicate potential structural variations or misalignments.

GENE BODY COVERAGE



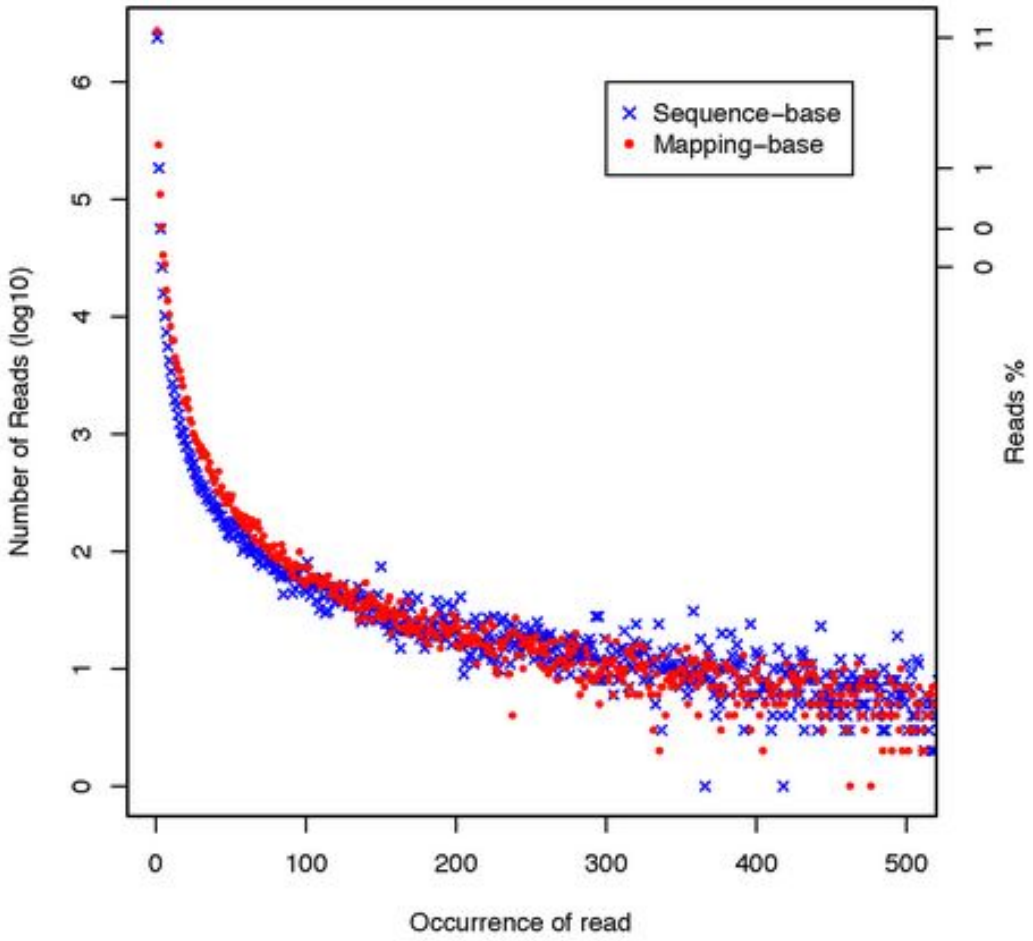


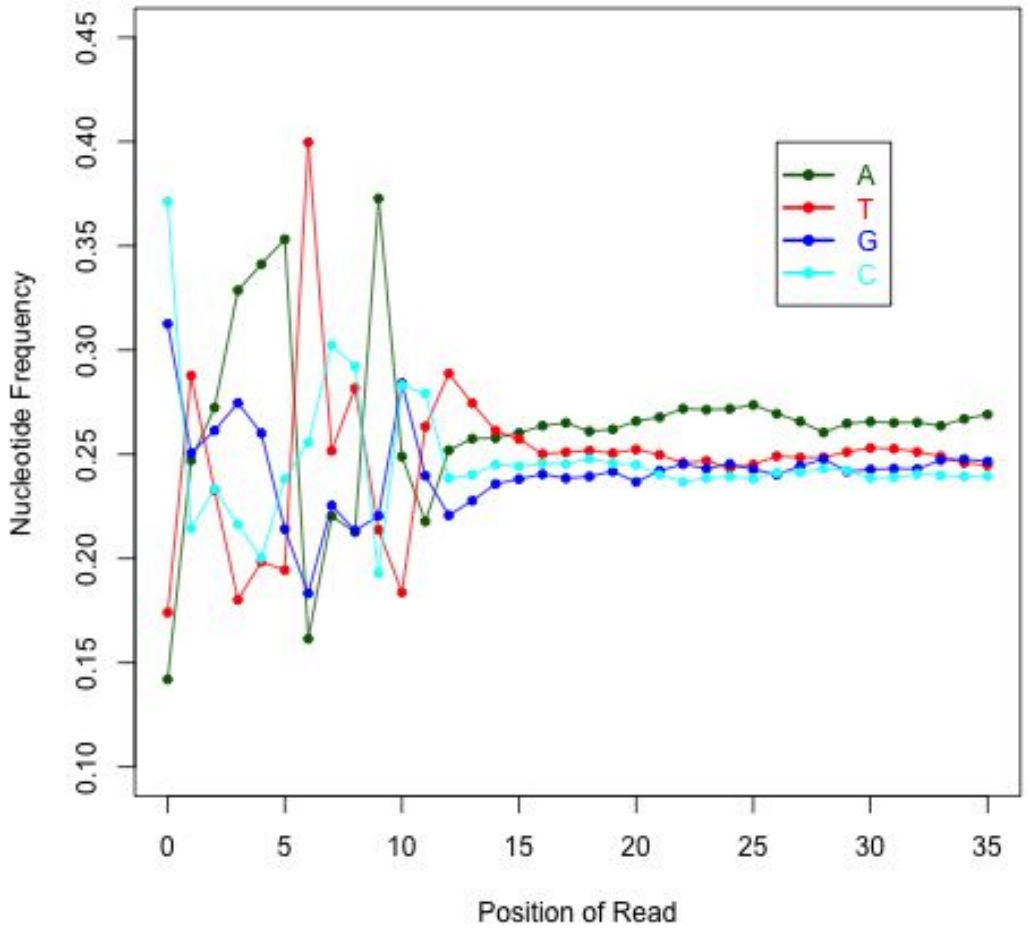
JUNCTION SATURATION



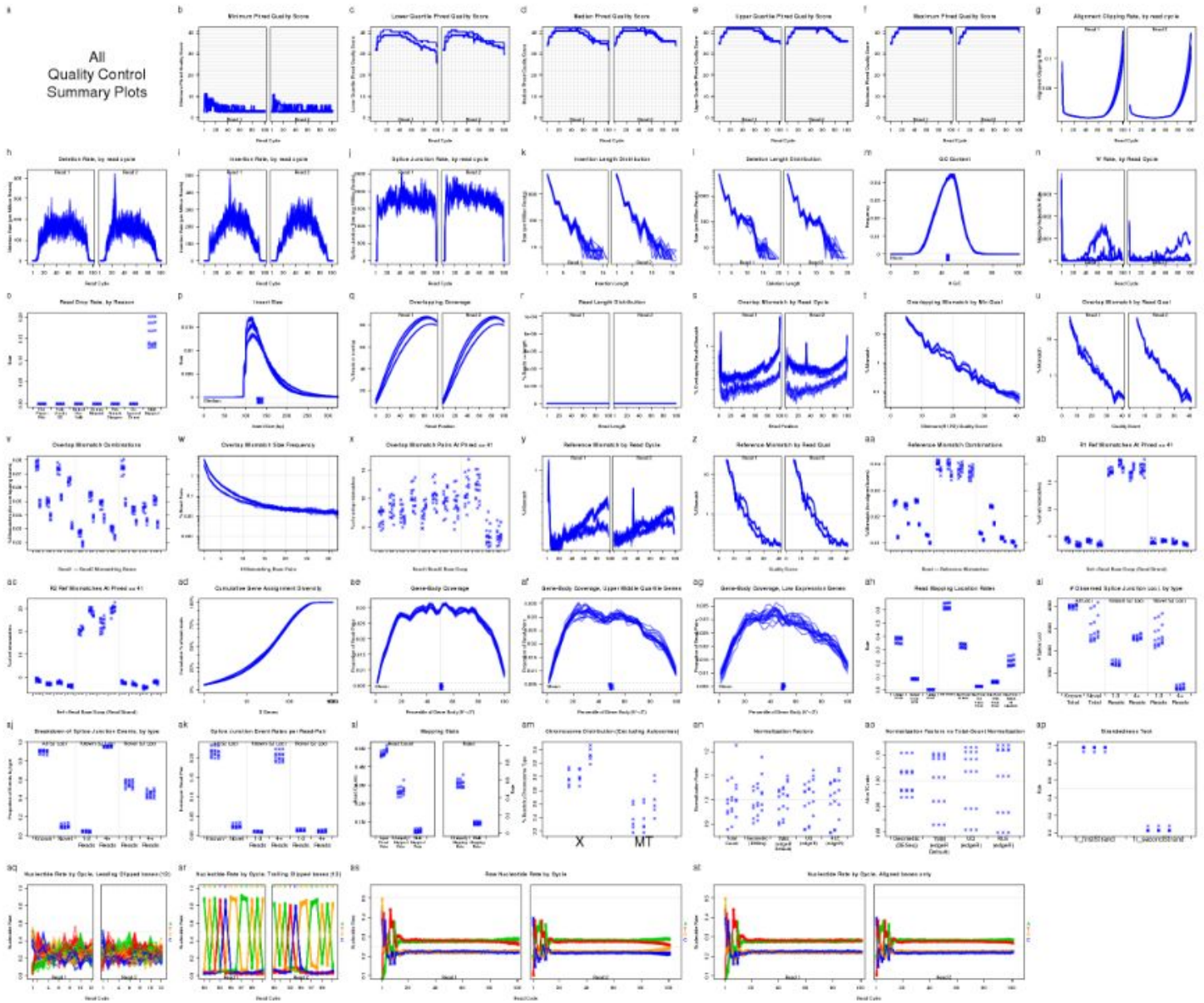
In this example, current sequencing depth is almost saturated for “known junction” (red line) detection because the number of “known junction” reaches a plateau. In other words, nearly all “known junctions” (expressed in this particular tissue) have already been detected, and deeper sequencing will not likely to detect additional “known junction” and will only increase junction coverage (i.e. junction covered by more reads).

READ DUPLICATION AND NUCLEOTIDE VERSUS CYCLE



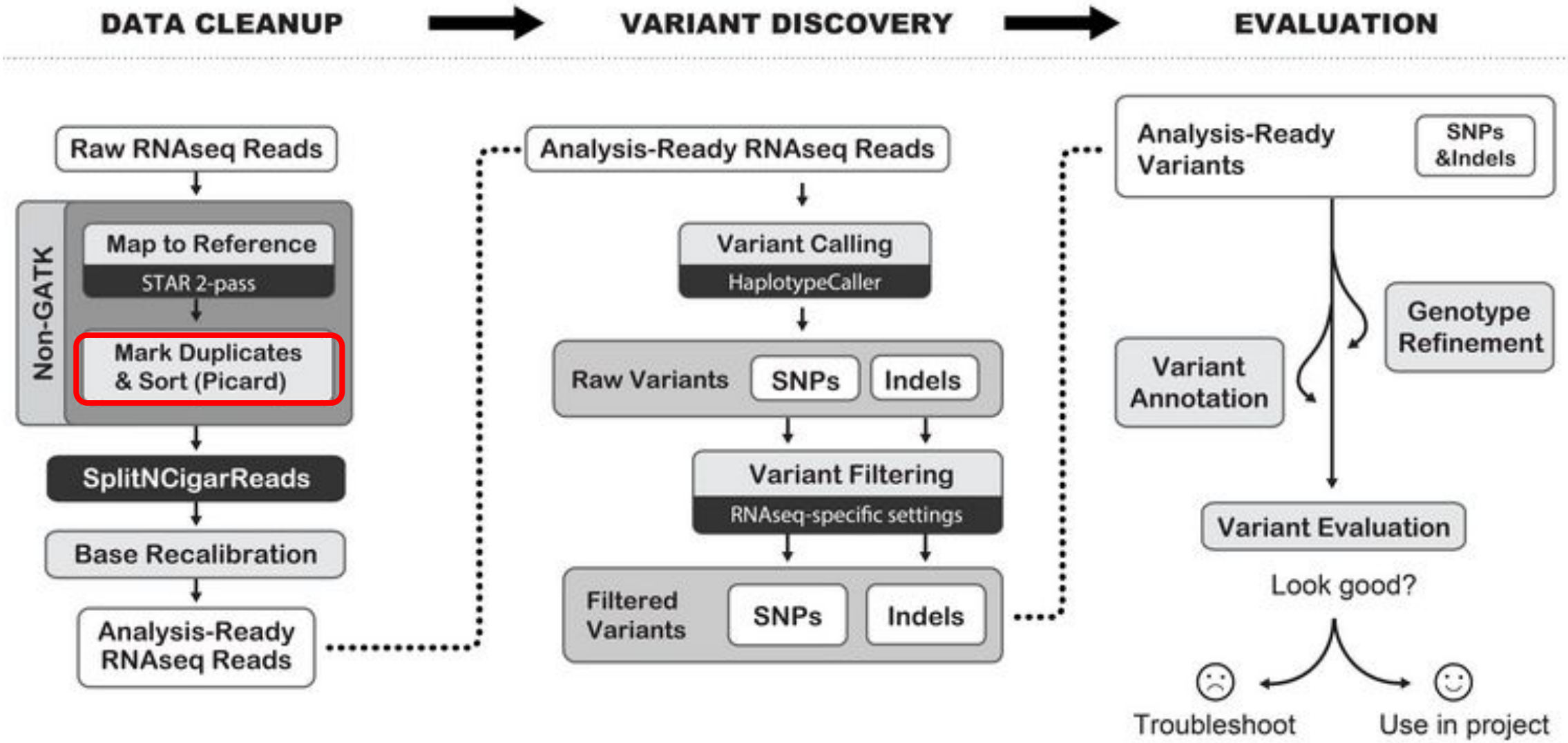


ALIGNMENT QC - QoRTs



http://hartleys.github.io/QoRTs/doc/QoRTs-vignette.pdf

PIPELINE


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
Raw RNAseq Reads
Non-GATK
Map to Reference
STAR 2-pass
Mark Duplicates & Sort (Picard)
SplitNCigarReads
Base Recalibration
Analysis-Ready RNAseq Reads
Analysis-Ready RNAseq Reads
Variant Calling
HaplotypeCaller
Raw Variants
SNPs
Indels
Variant Filtering
RNAseq-specific settings
Filtered Variants
SNPs
Indels
Analysis-Ready Variants
SNPs &Indels
Variant Annotation
Genotype Refinement
Variant Evaluation
Look good?
Troubleshoot
Use in project

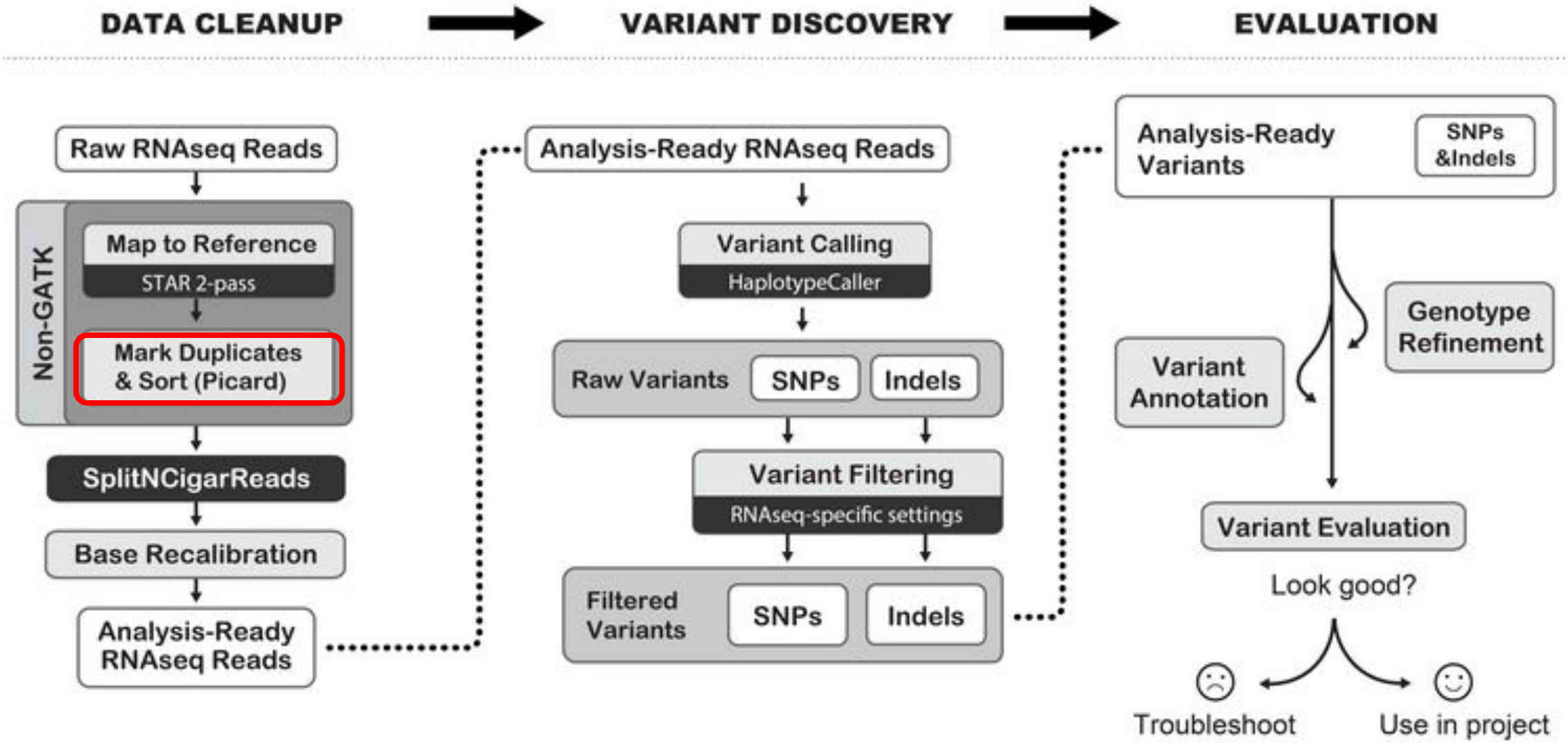
ADD READ GROUPS

```
for SAMPLE in WT_1 SNF2_1;
    do  java -jar /media/bacdao/Data1/RNA_seq/tools/picard.jar
    AddOrReplaceReadGroups
    I=${REF_DIR}/alignment_STAR/${SAMPLE}/Aligned.sortedByCoord.out.bam
    o=${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}.bam
    RGLB=lib1
    RGPL=Illumina
    RGPU=HiSeq2000
    RGSM=${SAMPLE} ; done
```

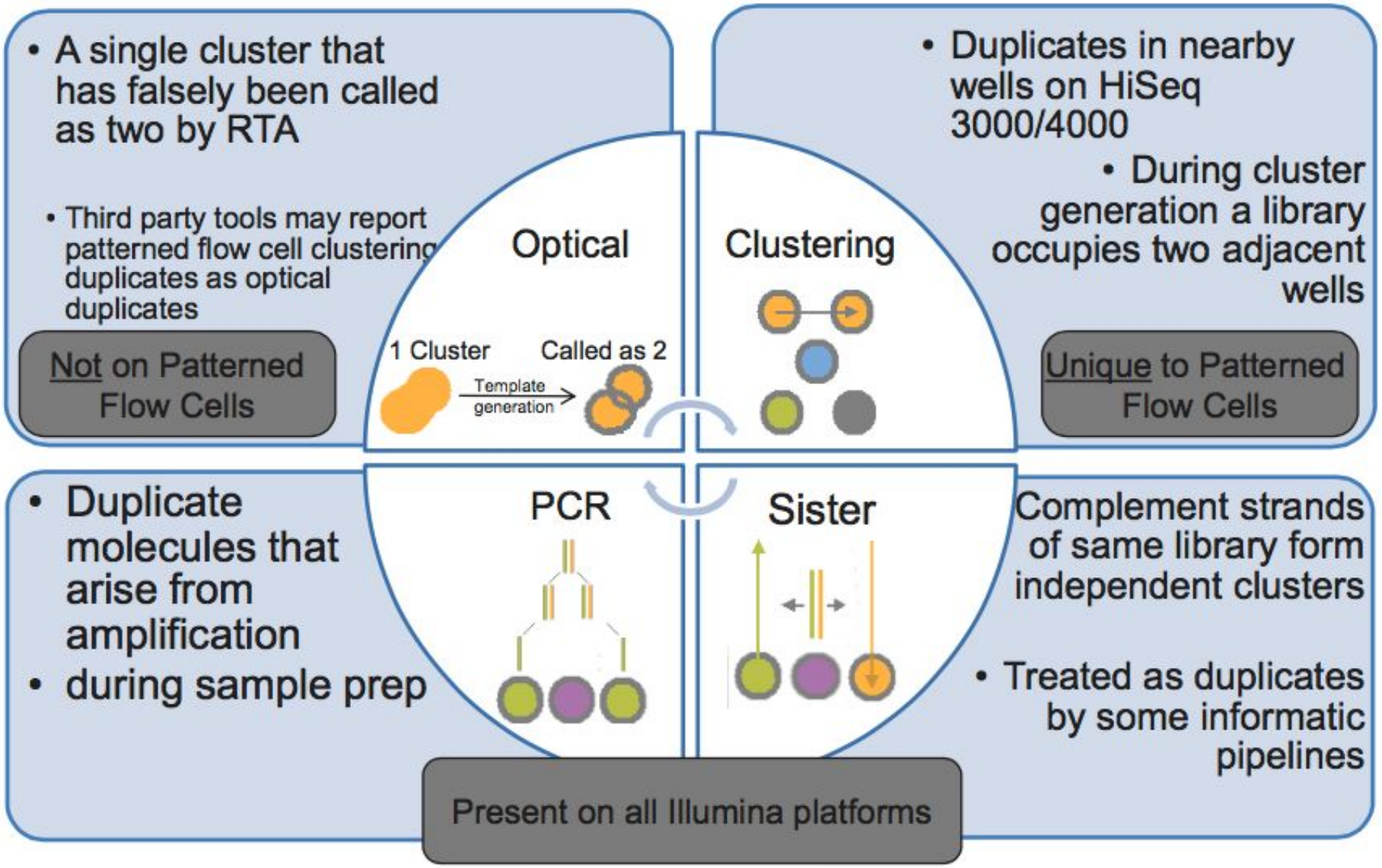
Experiment View all 672 results.	
ERX424885	Illumina HiSeq 2000 sequencing
Run View all 672 results.	
ERR458607	Illumina HiSeq 2000 sequencing
Study	
ERP004763	S. cerevisiae WT vs snf2 KO mutant RNA-seq data with 7 technical and 48 biological replicates (336 total) of each condition
Project	
PRJEB5348	S. cerevisiae WT vs snf2 KO mutant RNA-seq data with 7 technical and 48 biological replicates (336 total) of each condition

https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB5348

PIPELINE


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
Raw RNAseq Reads
Non-GATK
Map to Reference
STAR 2-pass
Mark Duplicates & Sort (Picard)
SplitNCigarReads
Base Recalibration
Analysis-Ready RNAseq Reads
Analysis-Ready RNAseq Reads
Variant Calling
HaplotypeCaller
Raw Variants
SNPs
Indels
Variant Filtering
RNAseq-specific settings
Filtered Variants
SNPs
Indels
Analysis-Ready Variants
SNPs &Indels
Variant Annotation
Genotype Refinement
Variant Evaluation
Look good?
Troubleshoot
Use in project

MARK DUPLICATES



http://enseqlopedia.com/2016/05/increased-read-duplication-on-patterned-flowcells-understanding-the-impact-of-exclusion-amplification/

MARK DUPLICATES

```
for SAMPLE in WT_1 SNF2_1;
    do  java -jar /media/bacdao/Data1/RNA_seq/tools/picard.jar
    MarkDuplicates
    I=${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}.bam
    o=${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_dedupped.bam
    CREATE_INDEX=true VALIDATION_STRINGENCY=SILENT
    M=output.metrics ; done

samtools flagstat ${REF_DIR}/var_cal/WT_1/WT_1.bam

samtools flagstat ${REF_DIR}/var_cal/WT_1/WT_1_dedupped.bam
```

```
samtools flagstat ${REF_DIR}/var_cal/WT_1/WT_1_dedupped.bam

6010344 + 0 in total (QC-passed reads + QC-failed reads)
6010344 + 0 primary
0 + 0 secondary
0 + 0 supplementary
3860084 + 0 duplicates
3860084 + 0 primary duplicates
6010344 + 0 mapped (100.00% : N/A)
6010344 + 0 primary mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

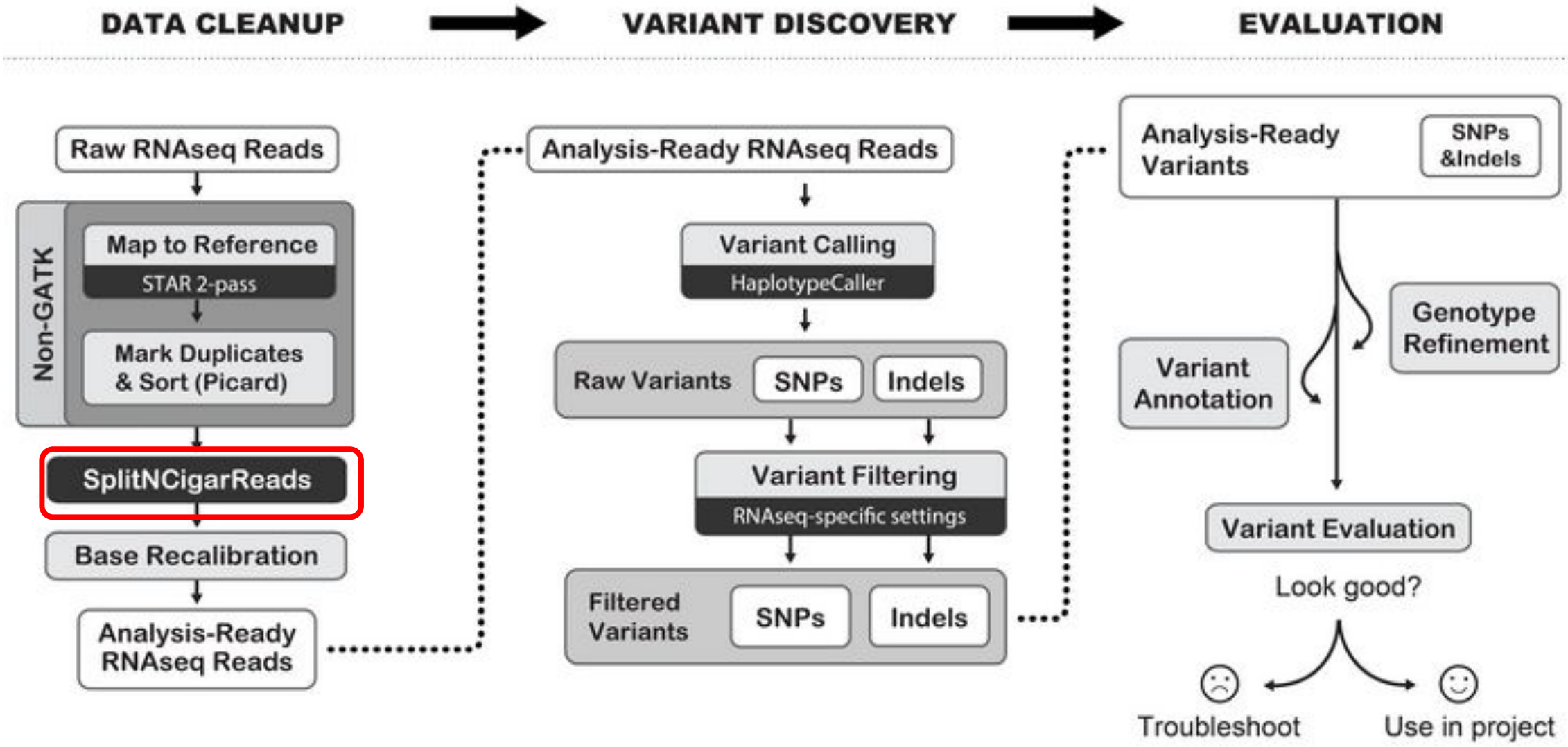
MARK DUPLICATES

```
samtools view WT_1.bam | head -5 | column -t

ERR458493.552967  16    NC_001133.9  140   255  12M61232N37M2S  *  0  0  CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC
ERR458498.814362  0     NC_001133.9  1593  255  11M71474N40M    *  0  0  TTTCTACAAAGATTCTCTCACTTGTACAGAGGTGTCTTCCCATTCGTTTTC
ERR458496.427513  16    NC_001133.9  3782  255  51M             *  0  0  CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGAAACGCTAAG
ERR458493.243111  0     NC_001133.9  3873  255  51M             *  0  0  TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACCAAGGGACAA
ERR458494.816646  16    NC_001133.9  3972  255  51M             *  0  0  TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAG
```

```
samtools view WT_1_dedupped.bam | head -5 | column -t

ERR458493.552967  1040  NC_001133.9  140   255  12M61232N37M2S  *  0  0  CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC
ERR458498.814362  0     NC_001133.9  1593  255  11M71474N40M    *  0  0  TTTCTACAAAGATTCTCTCACTTGTACAGAGGTGTCTTCCCATTCGTTTTC
ERR458496.427513  16    NC_001133.9  3782  255  51M             *  0  0  CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGAAACGCTAAG
ERR458493.243111  0     NC_001133.9  3873  255  51M             *  0  0  TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACCAAGGGACAA
ERR458494.816646  16    NC_001133.9  3972  255  51M             *  0  0  TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAG
```

https://broadinstitute.github.io/picard/explain-flags.html

PIPELINE


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
Raw RNAseq Reads
Non-GATK
Map to Reference
STAR 2-pass
Mark Duplicates & Sort (Picard)
SplitNCigarReads
Base Recalibration
Analysis-Ready RNAseq Reads
Analysis-Ready RNAseq Reads
Variant Calling
HaplotypeCaller
Raw Variants
SNPs
Indels
Variant Filtering
RNAseq-specific settings
Filtered Variants
SNPs
Indels
Analysis-Ready Variants
SNPs &Indels
Genotype Refinement
Variant Annotation
Variant Evaluation
Look good?
Troubleshoot
Use in project

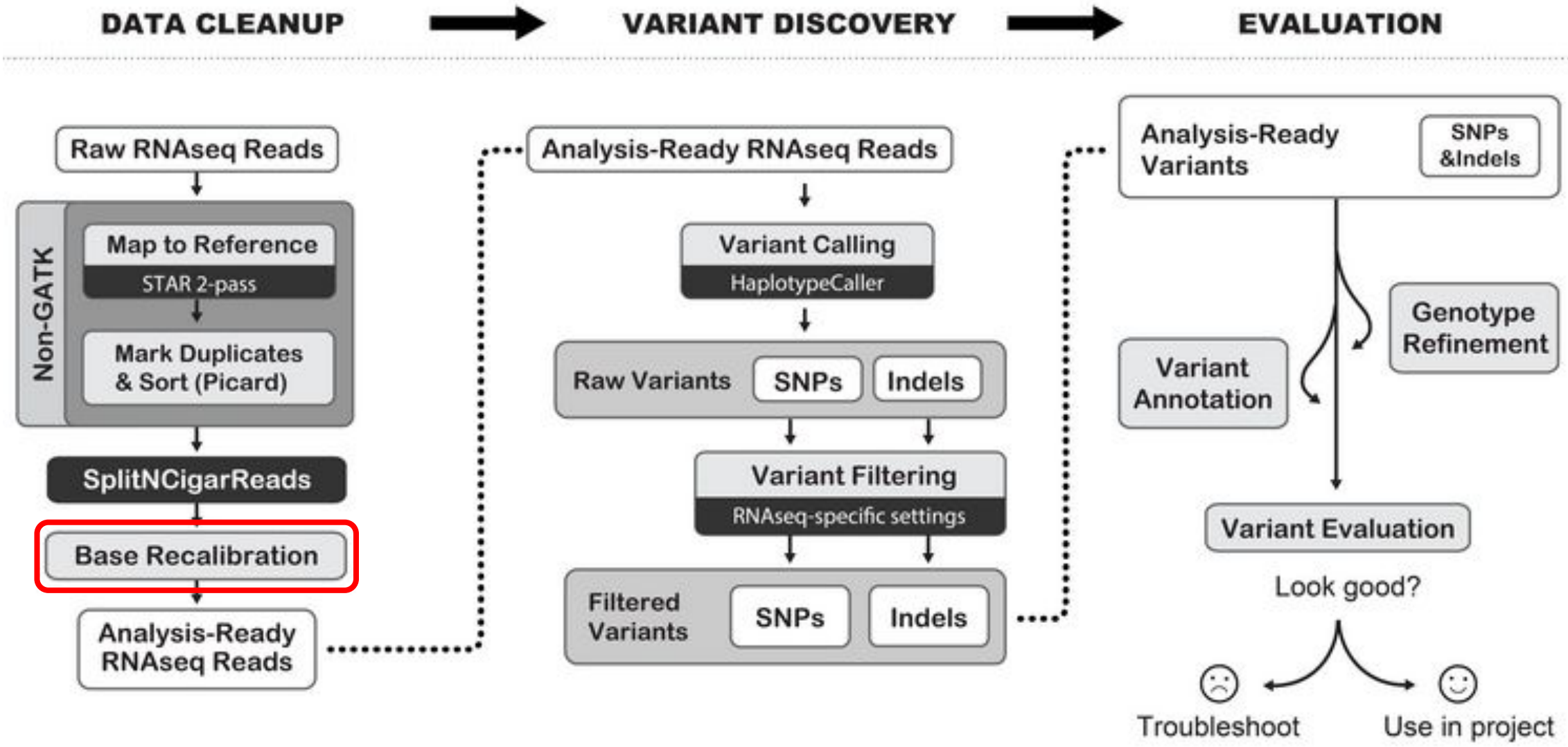
SPLIT 'N' CIGAR READ

```
for SAMPLE in WT_1 SNF2_1;
    do gatk SplitNCigarReads
    -R ${REF_DIR}/GCF_000146045.2_R64_genomic.fna
    -I ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_dedupped.bam -
    O ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_split.bam; done
```

```
samtools view WT_1_dedupped.bam | head -5 | column -t

ERR458493.552967  1040  NC_001133.9  140   255  12M61232N37M2S  *  0  0  CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC
ERR458498.814362  0     NC_001133.9  1593  255  11M71474N40M    *  0  0  TTTCTACAAAGATTCTCTCACTTGTACAGAGGTGTCTTCCCATTCGTTTTC
ERR458496.427513  16    NC_001133.9  3782  255  51M             *  0  0  CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGAAACGCTAAG
ERR458493.243111  0     NC_001133.9  3873  255  51M             *  0  0  TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACCAAGGGACAA
ERR458494.816646  16    NC_001133.9  3972  255  51M             *  0  0  TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAG
```

```
samtools view WT_1_split.bam | head -5 | column -t

ERR458493.552967  1040  NC_001133.9  140   60  12M39S  *  0  0  CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC
ERR458498.814362  0     NC_001133.9  1593  60  11M40S  *  0  0  TTTCTACAAAGATTCTCTCACTTGTACAGAGGTGTCTTCCCATTCGTTTTC
ERR458496.427513  16    NC_001133.9  3782  60  51M     *  0  0  CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGAAACGCTAAG
ERR458493.243111  0     NC_001133.9  3873  60  51M     *  0  0  TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACCAAGGGACAA
ERR458494.816646  16    NC_001133.9  3972  60  51M     *  0  0  TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCCTGAAG
```

PIPELINE


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
Raw RNAseq Reads
Non-GATK
Map to Reference
STAR 2-pass
Mark Duplicates & Sort (Picard)
SplitNCigarReads
Base Recalibration
Analysis-Ready RNAseq Reads
Analysis-Ready RNAseq Reads
Variant Calling
HaplotypeCaller
Raw Variants
SNPs
Indels
Variant Filtering
RNAseq-specific settings
Filtered Variants
SNPs
Indels
Analysis-Ready Variants
SNPs &Indels
Genotype Refinement
Variant Annotation
Variant Evaluation
Look good?
Troubleshoot
Use in project

BASE RECALIBRATION

```
for SAMPLE in WT_1 SNF2_1;
    do gatk BaseRecalibrator
    -R ${REF_DIR}/GCF_000146045.2_R64_genomic.fna
    -I ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_split.bam
    -known-sites sac_cer.vcf.gz
    -O ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_recal_data.table; done
```

```
for SAMPLE in WT_1 SNF2_1;
    do gatk ApplyBQSR
    -R ${REF_DIR}/GCF_000146045.2_R64_genomic.fna
    -I ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_split.bam
    -bqsr-recal-file ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_recal_data.table
    -O ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_recal_split.bam; done
```

BASE RECALIBRATION

IGSR: The International Genome Sample Resource

Supporting open human variation data

Home About Data Help

About variant identifiers

Assembly conversion / Data access / Identifiers / VCF

Answer:

All of the 1000 Genomes SNPs and indels have been submitted to dbSNP, and will have rsIDs in the main 1000 Genomes release files.

If you are using some of the older working files that were used during the data gathering phase of the 1000 Genomes Project, you may fi
internally by the groups that did that set of particular variant calling, and are not found anywhere other than these files, as they will have

KGP identifiers

You may also see kgp identifiers, which were created by Illumina for their genotyping platform before some variants identified during the

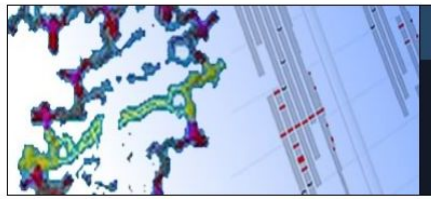
We do not possess a mapping of these identifiers to current rs numbers. As far as we are aware no such list exists.

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

dbSNP SNP Search
Advanced Help



dbSNP

dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

Getting Started	Submission	Access Data
dbSNP 20th Anniversary	How to Submit	Web Search
Overview of dbSNP	Hold Until Published (HUP) Policies	eUtils API
About Reference SNP (rs)	Submission Search	Variation Services
Factsheet		FTP Download
Entrez Updates (May 26, 2020)		Tutorials on GitHub

```
#dbsnp vcf (dbSNPs for GATK) is needed for the run: Download the vcf and the index file into this directory

wget http://ftp.ensembl.org/pub/release-100/variation/vcf/saccharomyces_cerevisiae/saccharomyces_cerevisiae.vcf.gz

wget http://ftp.ensembl.org/pub/release-100/variation/vcf/saccharomyces_cerevisiae/saccharomyces_cerevisiae.vcf.gz.csi
```

BASE RECALIBRATION

```
cat GCF_000146045.2_R64_genomic.fna | grep -E ">"

>NC_001133.9 Saccharomyces cerevisiae S288C chromosome I, complete sequence
>NC_001134.8 Saccharomyces cerevisiae S288C chromosome II, complete sequence
>NC_001135.5 Saccharomyces cerevisiae S288C chromosome III, complete sequence
>NC_001136.10 Saccharomyces cerevisiae S288C chromosome IV, complete sequence
>NC_001137.3 Saccharomyces cerevisiae S288C chromosome V, complete sequence
>NC_001138.5 Saccharomyces cerevisiae S288C chromosome VI, complete sequence
>NC_001139.9 Saccharomyces cerevisiae S288C chromosome VII, complete sequence
>NC_001140.6 Saccharomyces cerevisiae S288C chromosome VIII, complete sequence
>NC_001141.2 Saccharomyces cerevisiae S288C chromosome IX, complete sequence
>NC_001142.9 Saccharomyces cerevisiae S288C chromosome X, complete sequence
>NC_001143.9 Saccharomyces cerevisiae S288C chromosome XI, complete sequence
>NC_001144.5 Saccharomyces cerevisiae S288C chromosome XII, complete sequence
>NC_001145.3 Saccharomyces cerevisiae S288C chromosome XIII, complete sequence
>NC_001146.8 Saccharomyces cerevisiae S288C chromosome XIV, complete sequence
>NC_001147.6 Saccharomyces cerevisiae S288C chromosome XV, complete sequence
>NC_001148.4 Saccharomyces cerevisiae S288C chromosome XVI, complete sequence
>NC_001224.1 Saccharomyces cerevisiae S288c mitochondrion, complete genome
```

```
zcat saccharomyces_cerevisiae.vcf.gz | grep -v "#" | head

I   84  s01-84  G   A   .   .   SGRP;TSA=SNV
I   109 s01-109 G   C   .   .   SGRP;TSA=SNV
I   111 s01-111 C   T   .   .   SGRP;TSA=SNV
I   114 s01-114 T   C   .   .   SGRP;TSA=SNV
I   115 s01-115 C   G   .   .   SGRP;TSA=SNV
I   136 s01-136 G   A   .   .   SGRP;TSA=SNV
I   165 s01-165 C   T   .   .   SGRP;TSA=SNV
I   166 s01-166 C   T   .   .   SGRP;TSA=SNV
I   169 s01-169 A   G   .   .   SGRP;TSA=SNV
I   177 s01-177 G   C   .   .   SGRP;TSA=SNV
```

https://regexr.com/

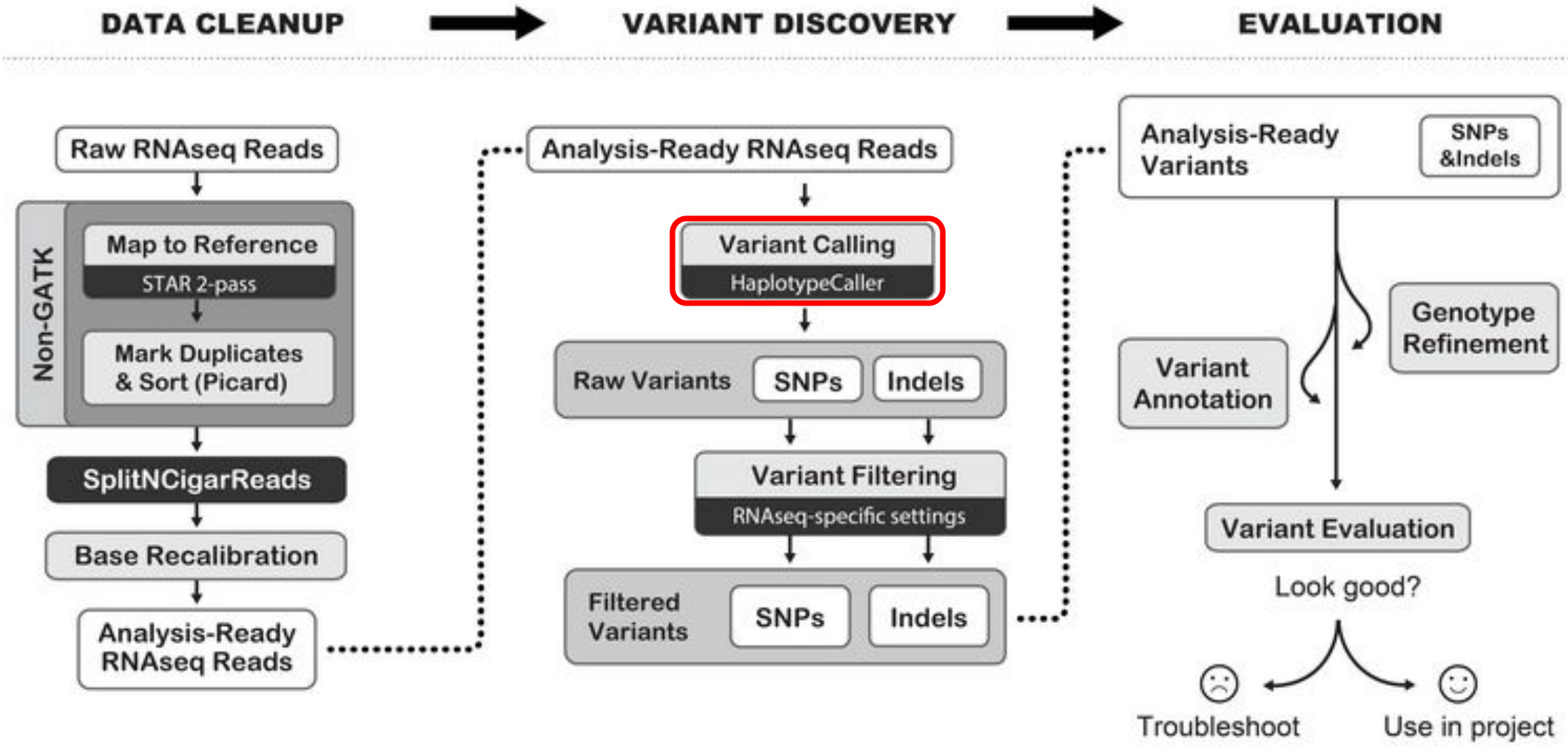
BASE RECALIBRATION

```
#Because the chromosome name in reference fasta file is "NC_001..." which is different from the one in vcf file "I","II",...; we need to use bcftools to transfer them into the similar name.
sudo apt install bcftools

cat ${REF_DIR}/GCF_000146045.2_R64_genomic.fna | grep ">" | grep -E '^>NC_[0-9]+\.[0-9]+'| awk '{ print $1, $6 }' | sed 's/^>//g' | sed -e 's/,$//' -e '$d' | awk '{print $2, $1}' > ${REF_DIR}/annotate.txt

bcftools annotate --rename-chrs ${REF_DIR}/annotate.txt ${REF_DIR}/saccharomyces_cerevisiae.vcf.gz > ${REF_DIR}/sac_cer.vcf
```

```
cat annotate.txt | column -t
I      NC_001133.9
II     NC_001134.8
III    NC_001135.5
IV     NC_001136.10
V      NC_001137.3
VI     NC_001138.5
VII    NC_001139.9
VIII   NC_001140.6
IX     NC_001141.2
X      NC_001142.9
XI     NC_001143.9
XII    NC_001144.5
XIII   NC_001145.3
XIV    NC_001146.8
XV     NC_001147.6
XVI    NC_001148.4
```

https://regexr.com/

PIPELINE


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
Raw RNAseq Reads
Non-GATK
Map to Reference
STAR 2-pass
Mark Duplicates & Sort (Picard)
SplitNCigarReads
Base Recalibration
Analysis-Ready RNAseq Reads
Analysis-Ready RNAseq Reads
Variant Calling
HaplotypeCaller
Raw Variants
SNPs
Indels
Variant Filtering
RNAseq-specific settings
Filtered Variants
SNPs
Indels
Analysis-Ready Variants
SNPs &Indels
Variant Annotation
Genotype Refinement
Variant Evaluation
Look good?
Troubleshoot
Use in project

VARIANT CALLING

```
for SAMPLE in WT_1 SNF2_1;
    do gatk HaplotypeCaller
    -R ${REF_DIR}/GCF_000146045.2_R64_genomic.fna
    -I ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_recal_split.bam
    -dont-use-soft-clipped-bases
    -stand-call-conf 20
    -O ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_called_variants.vcf; done
```

-dont-use-soft-clipped-bases: This option tells the HaplotypeCaller not to use soft-clipped bases (bases that were partially aligned to the reference genome) during variant calling.

-stand-call-conf: The minimum phred-scaled confidence threshold at which variants should be called. Variants with a quality score lower than this threshold will not be included in the final output. (Default = 30)

https://gatk.broadinstitute.org/hc/en-us/articles/13832687299739-HaplotypeCaller

VARIANT FILTERING

```
for SAMPLE in WT_1 SNF2_1;
    do gatk VariantFiltration
    -R ${REF_DIR}/GCF_000146045.2_R64_genomic.fna
    -V ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_called_variants.vcf
    -window 35
    -cluster 3
    -filter-name FS
    -filter "FS>30.0"
    -O ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_filtered_called_variants.vcf; done
```

-**window**: The window size (in bases) in which to evaluate clustered SNPs. Works together with the --cluster-size argument.

-cluster: The number of SNPs which make up a cluster. Must be at least 2

-filter-name: Names to use for the list of filters. This name is put in the FILTER field for variants that get filtered

-filter: Names to use for the list of filters. This name is put in the FILTER field for variants that get filtered.

```
for SAMPLE in WT_1 SNF2_1;
    do gatk VariantFiltration
    -R ${REF_DIR}/GCF_000146045.2_R64_genomic.fna
    -V ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_called_variants.vcf
    -window 35
    -cluster 3
    -filter-name FS
    -filter "FS>30.0"
    -O ${REF_DIR}/var_cal/${SAMPLE}/${SAMPLE}_filtered_called_variants.vcf; done
```

```
cat WT_1_filtered_called_variants.vcf | grep -v '#' | head | column -t
#CHROM        POS   ID REF ALT QUAL     FILTER                                    INFO
NC_001133.9  25452   .  T   A   73.64   SnpCluster  AC=1;AF=0.500;AN=2;BaseQRankSum=0.000;DP=4;ExcessHet=0.0000;FS=0.000;MLEAC=1;
NC_001133.9  25460   .  T   C   73.64   SnpCluster  AC=1;AF=0.500;AN=2;BaseQRankSum=-0.967;DP=3;ExcessHet=0.0000;FS=0.000;MLEAC=1
NC_001133.9  25461   .  G   A   73.64   SnpCluster  AC=1;AF=0.500;AN=2;DP=3;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.0
NC_001133.9  25464   .  G   A   73.64   SnpCluster  AC=1;AF=0.500;AN=2;DP=3;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.0
NC_001133.9  25470   .  C   T   73.64   SnpCluster  AC=1;AF=0.500;AN=2;BaseQRankSum=-0.967;DP=4;ExcessHet=0.0000;FS=0.000;MLEAC=1
NC_001133.9  25479   .  T   C   74.32   SnpCluster  AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00

NC_001133.9  51930   .  T   G   70.64   FS          AC=1;AF=0.500;AN=2;BaseQRankSum=-3.256;DP=26;ExcessHet=0.0000;FS=33.331;MLEAC

NC_001133.9  188881  .  G   C   88.65   PASS        AC=1;AF=0.500;AN=2;BaseQRankSum=0.319;DP=4;ExcessHet=0.0000;FS=0.000;MLEAC=1;
NC_001133.9  188939  .  A   G   160.64  PASS        AC=1;AF=0.500;AN=2;BaseQRankSum=1.180;DP=11;ExcessHet=0.0000;FS=0.000;MLEAC=1
NC_001133.9  192737  .  T   C   39.64   PASS        AC=1;AF=0.500;AN=2;BaseQRankSum=-0.524;DP=7;ExcessHet=0.0000;FS=0.000;MLEAC=1
```