NGS Data Analysis on the HPRC Ada Cluster

Michael Dickens
Texas A&M University
High Performance Research Computing

June 7, 2016

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

- Available Bioinformatics Tools
- Sequence Quality Control
- *de novo* sequence assembly
- Variant Calling (SNPs and indels)
- RNA-seq Analysis
- ChIP-seq Analysis

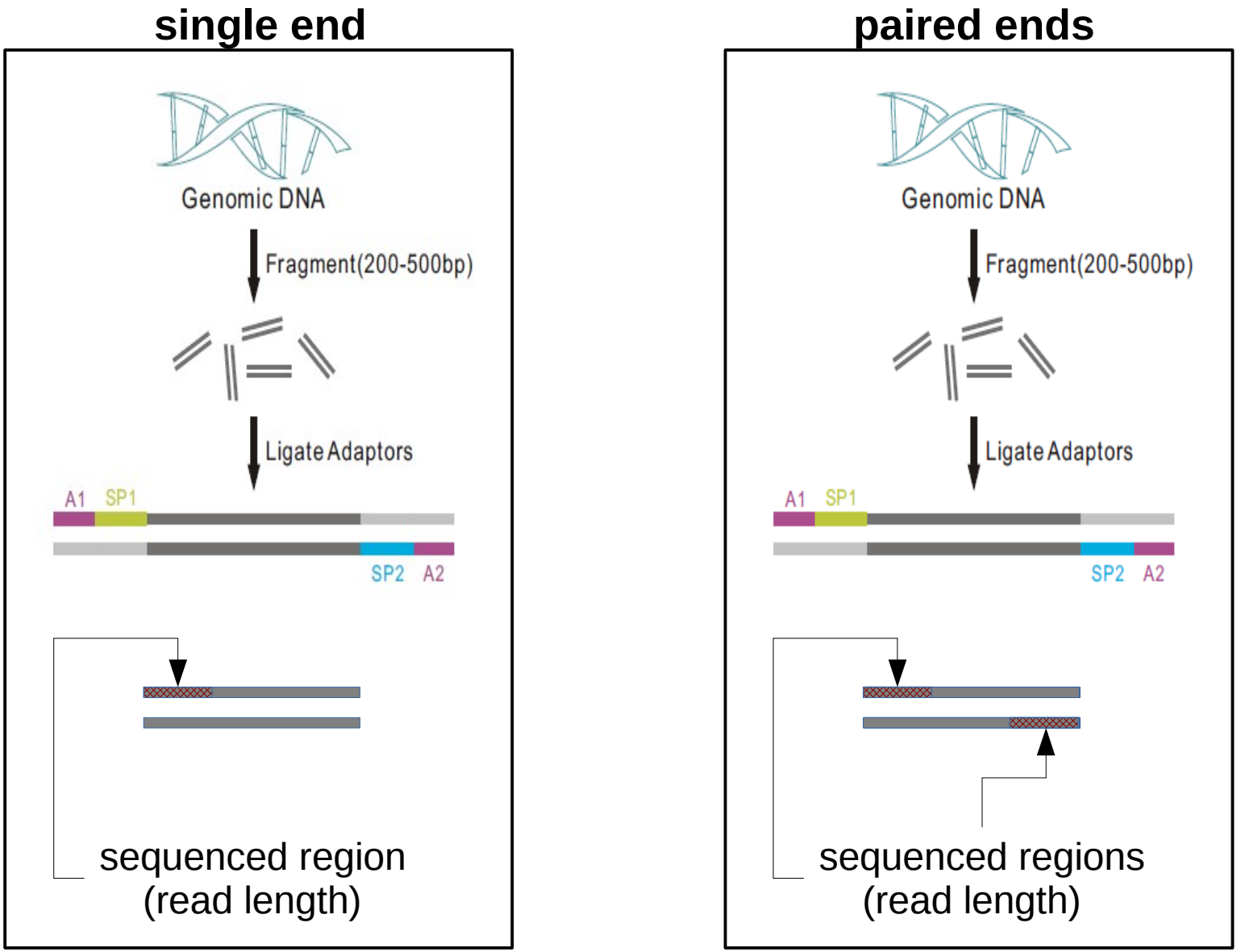
Where to Find NGS Tools

- TAMU HPRC Wiki
  - https://hprc.tamu.edu/wiki/index.php/Ada:Bioinformatics
- Tools that are already installed on Ada
  - `module avail`
  - `module spider toolname`
  - `module key assembly`

(searches tool description, some modules may be missed)

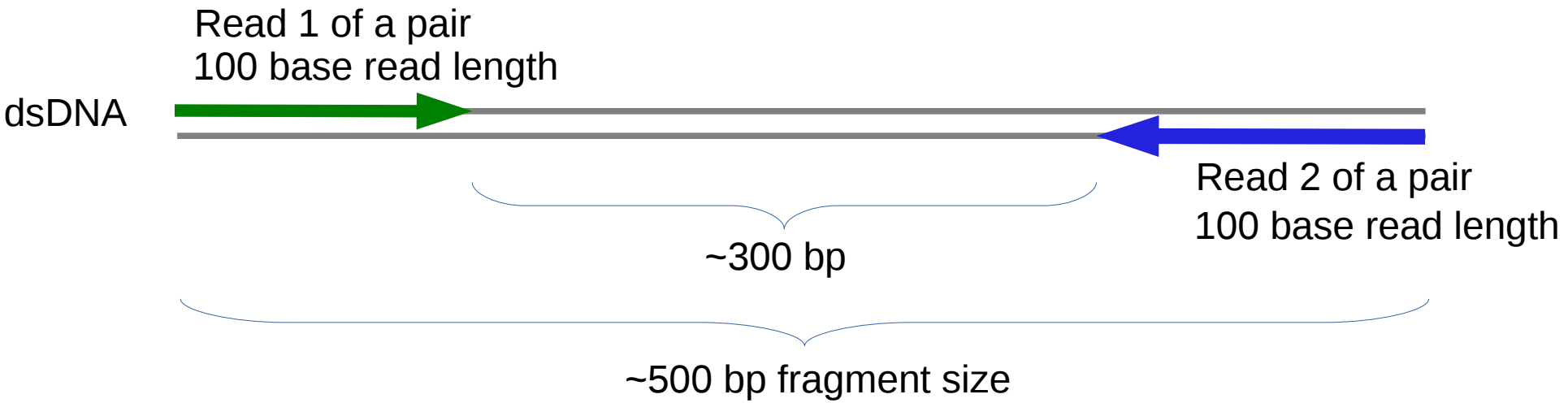
- If you find a tool that you want installed on Ada, send a request to **help@hprc.tamu.edu**
  - SeqAnswers http://seqanswers.com/wiki/Software/list
  - omictools.com
  - slideshare.net – find many NGS presentations

Illumina Sequencing Libraries



illumina.com

Paired End (PE) Reads



FASTQ format

Read 1 paired end fastq file

```
@M00861:1:000000000-A36BE:1:1101:14650:1529 1:N:0:8
TTCTTAAAAATACCATAAAAGGCTTAAACTTGCCATTTACGACGGATTAATTCCAACTCTTTTCGGCTATCTTCATCTTTTAAGGTTAAATGACTCATAA
+
FFFHBFFHHIIIIIIHFHHCGEFGHHIHHHIHD/?DGGHHH@DEB,5EGHGHHIIHIF?FGGHHCCBFDGHFHDGHGFFFGDFHH?DFHDFFHHFHFHFF
```

Read 2 paired end fastq file

```
@M00861:1:000000000-A36BE:1:1101:14650:1529 2:N:0:8
ACTAAAAATCAATTTTATCAATTTCAAGCTCTACCTTATTTACTCATTATTTTAGTGATGGCCACTTTAATAAAAATATTGGTAGCATATTTTGCAATAG
+
BFFHIHHHFHHDGHIHHIHHHGHHHHHFHHDFFHIHIIIHIHDFHHHIHIIIH=AAFHIIIHFGFHHHHHGGHHIHHFGFFFEGGHHHDGHHH/CGHIFF
```

Quality Control (QC)

QC Evaluation

- Use FastQC to visualize quality of fastq files

```
module load FastQC
```

- Displays quality score distribution of reads
  - Input is a fastq file or files
  - Can run using command line or as a GUI (using X11)
- Will alert you of poor read characteristics
- For command line processing:
  - Can view html results using lynx UNIX tool

```
lynx results.html
```

  - Can view images using eog UNIX command (X11)

```
eog per_base_sequence_content.png
```

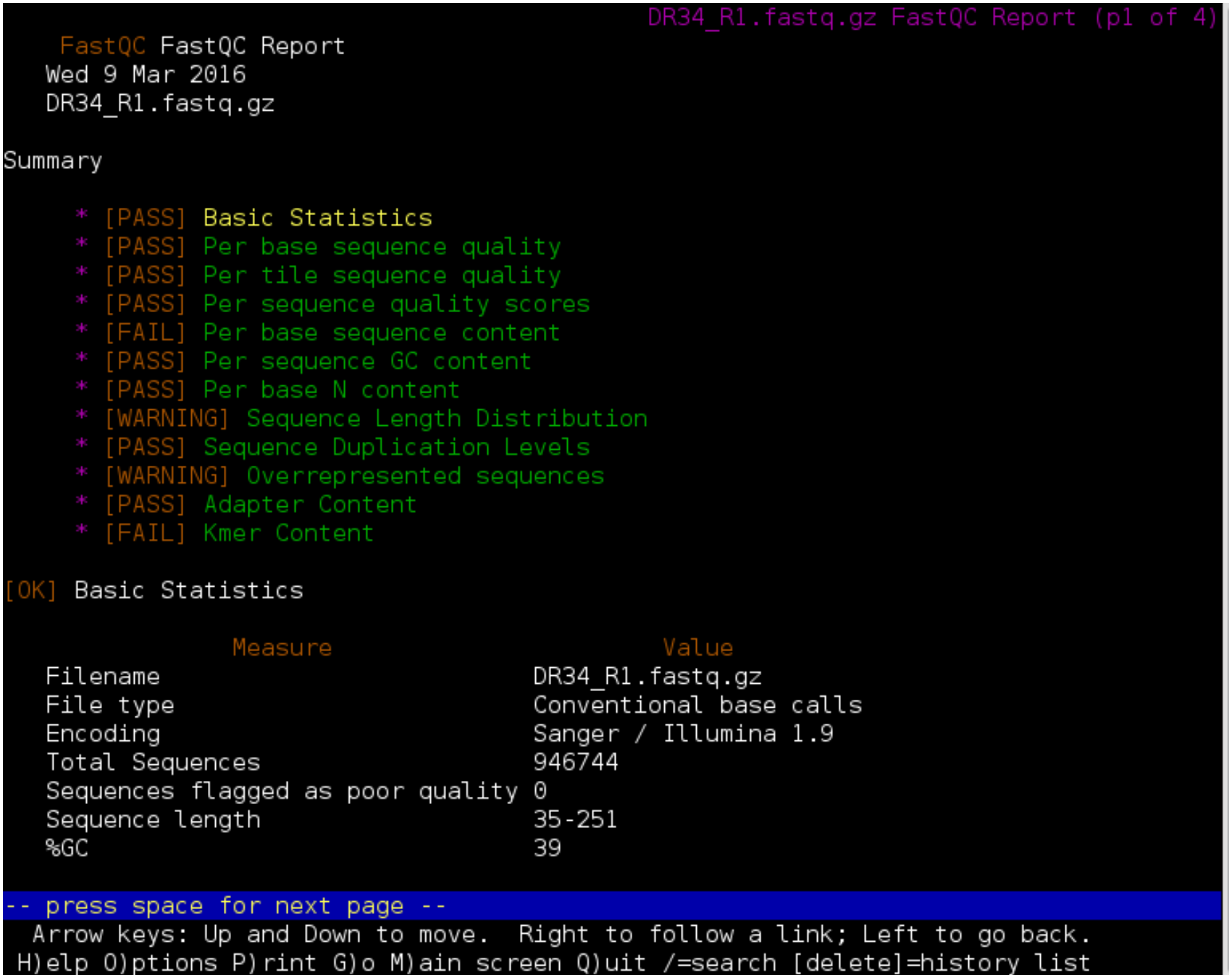
Visualize FastQC Report using GUI

FastQC Job Script Example

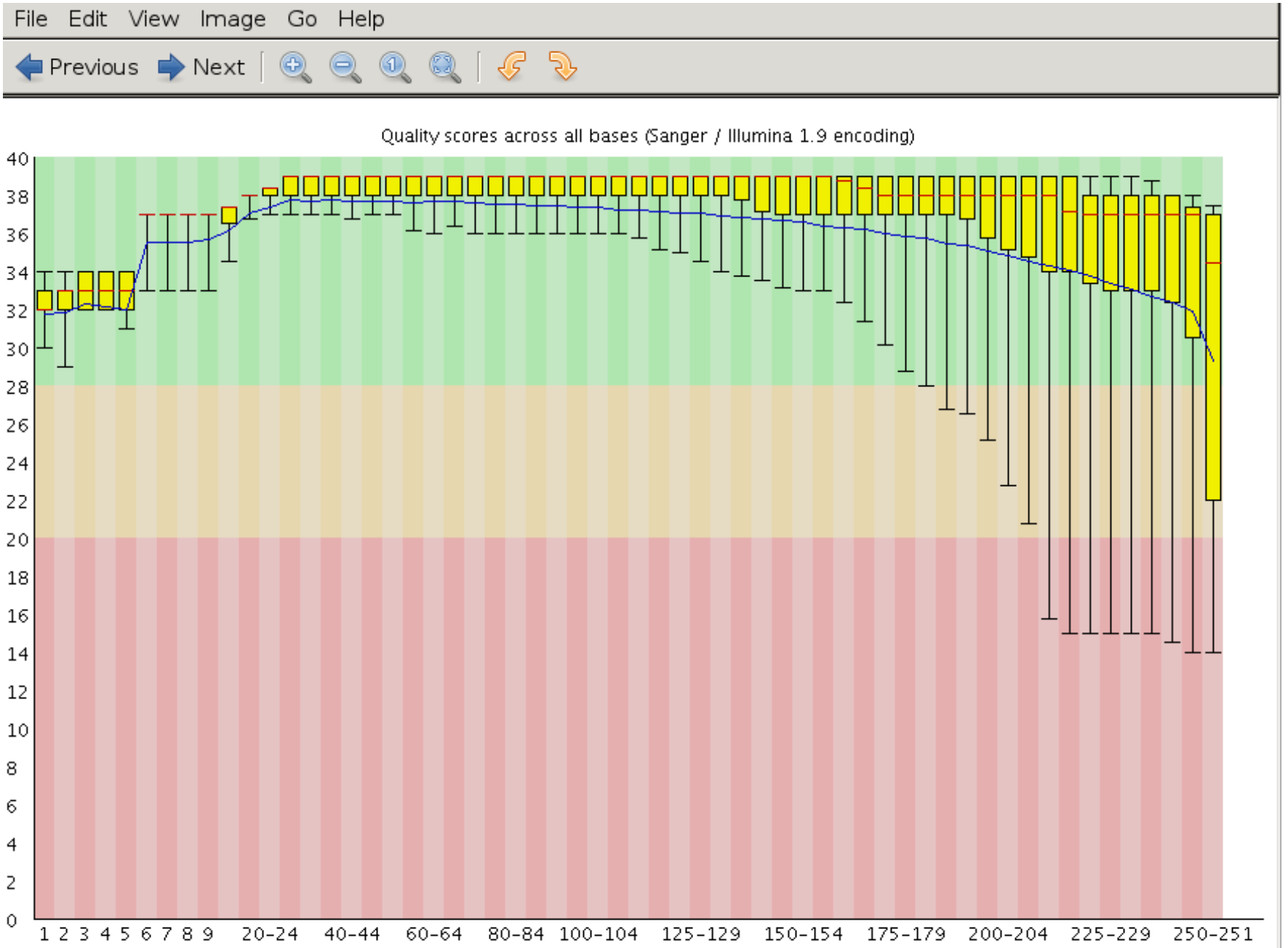
```
#BSUB -L /bin/bash                  # uses the bash login shell
#BSUB -J fastqc                     # job name
#BSUB -n 2                          # assigns 2 cores for execution
#BSUB -R "span[ptile=2]"            # assigns 2 cores per node
#BSUB -R "rusage[mem=2000]"         # reserves 2000MB memory per core
#BSUB -M 2000                       # sets to 2000MB process enforceable memory limit. (M * n)
#BSUB -W 1:00                       # sets to 1 hour the job's runtime wall-clock limit.
#BSUB -o stdout.%J                  # directs the job's standard output to stdout.jobid
#BSUB -e stderr.%J                  # directs the job's standard error to stderr.jobid

module load FastQC

fastqc --threads 2 --outdir ./ DR34_R1.fastq.gz DR34_R2.fastq.gz
```

- FastQC will process using one CPU core per file
  - If there are 10 fastq files to process and 4 cores used
    - 4 files will start processing and 6 will wait in the fastqc queue
  - If there is only one fastq file to process
    - using 10 cores does not speed up the process

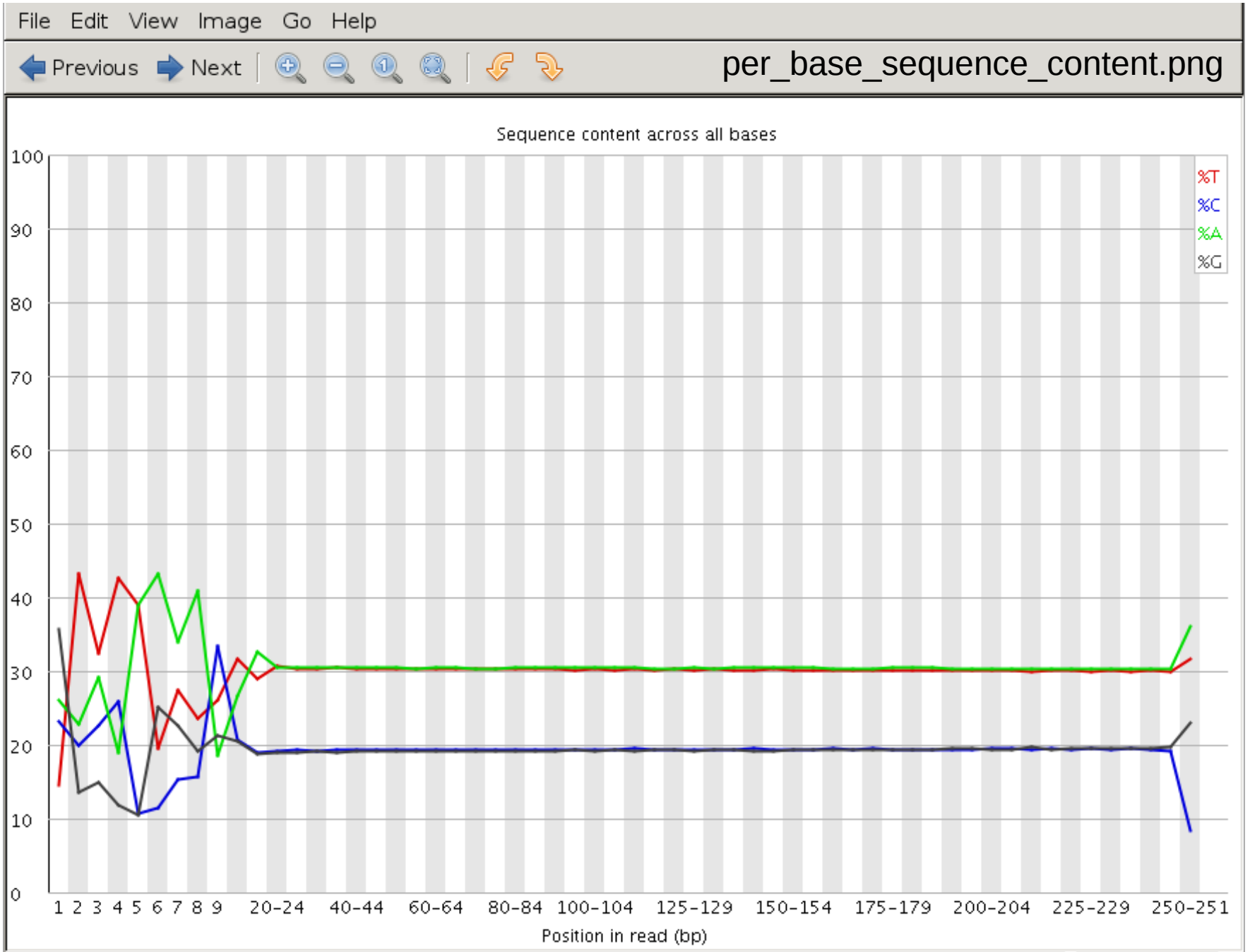
Visualize FastQC Report using lynx



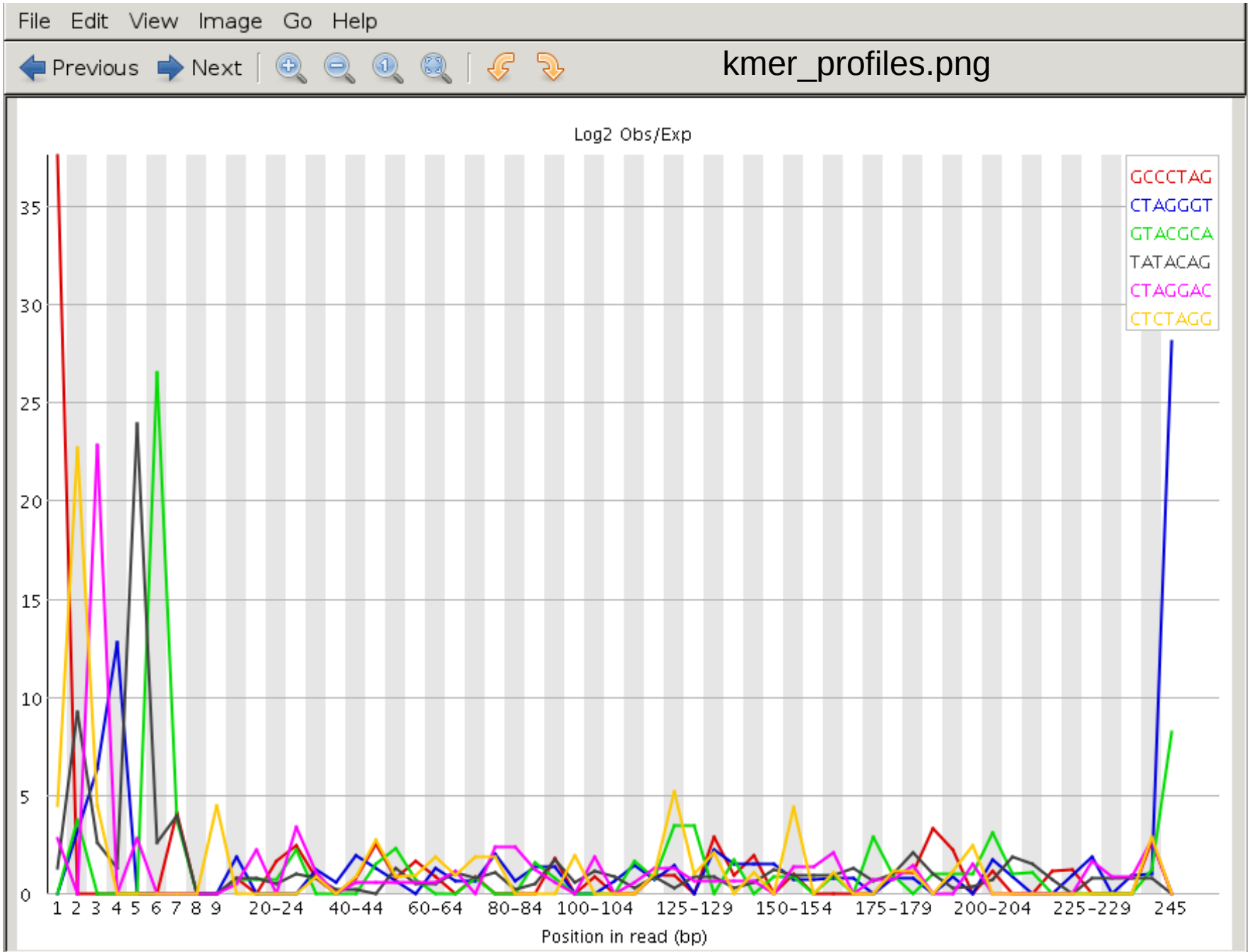
Visualize FastQC Output Images With eog



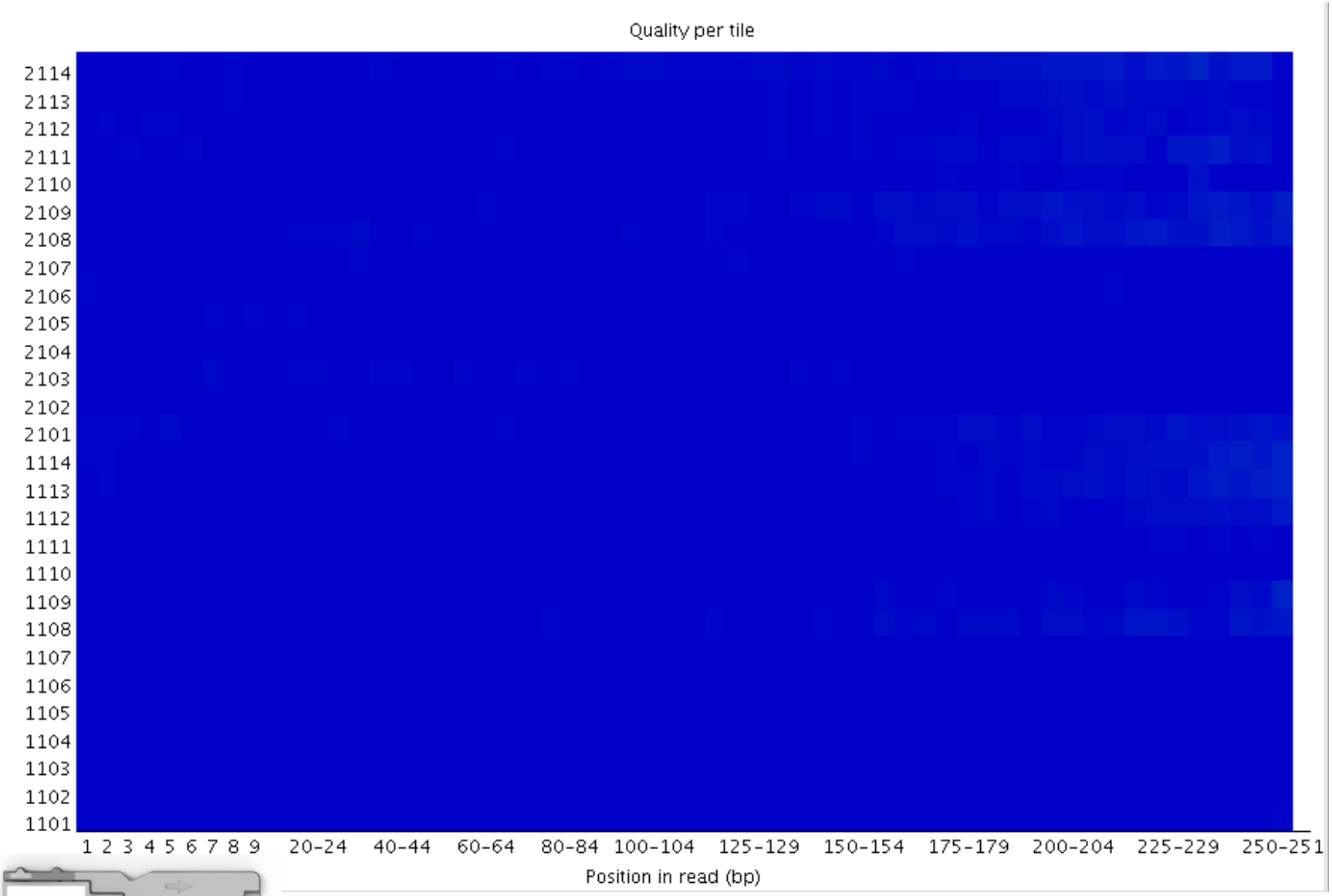
Illumina Transposon Insertion Site

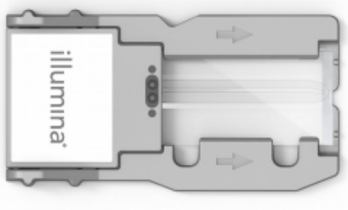


Illumina Transposon Insertion Site



FastQC Output Image
Flowcell: good per_tile quality





MiSeq flowcell

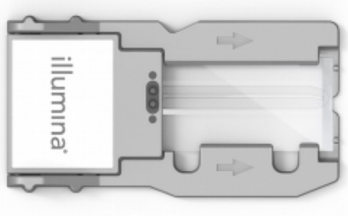
good quality

poor quality

FastQC Output Image
Flowcell: good per_tile quality

bottom of flowcell

top of flowcell



MiSeq flowcell

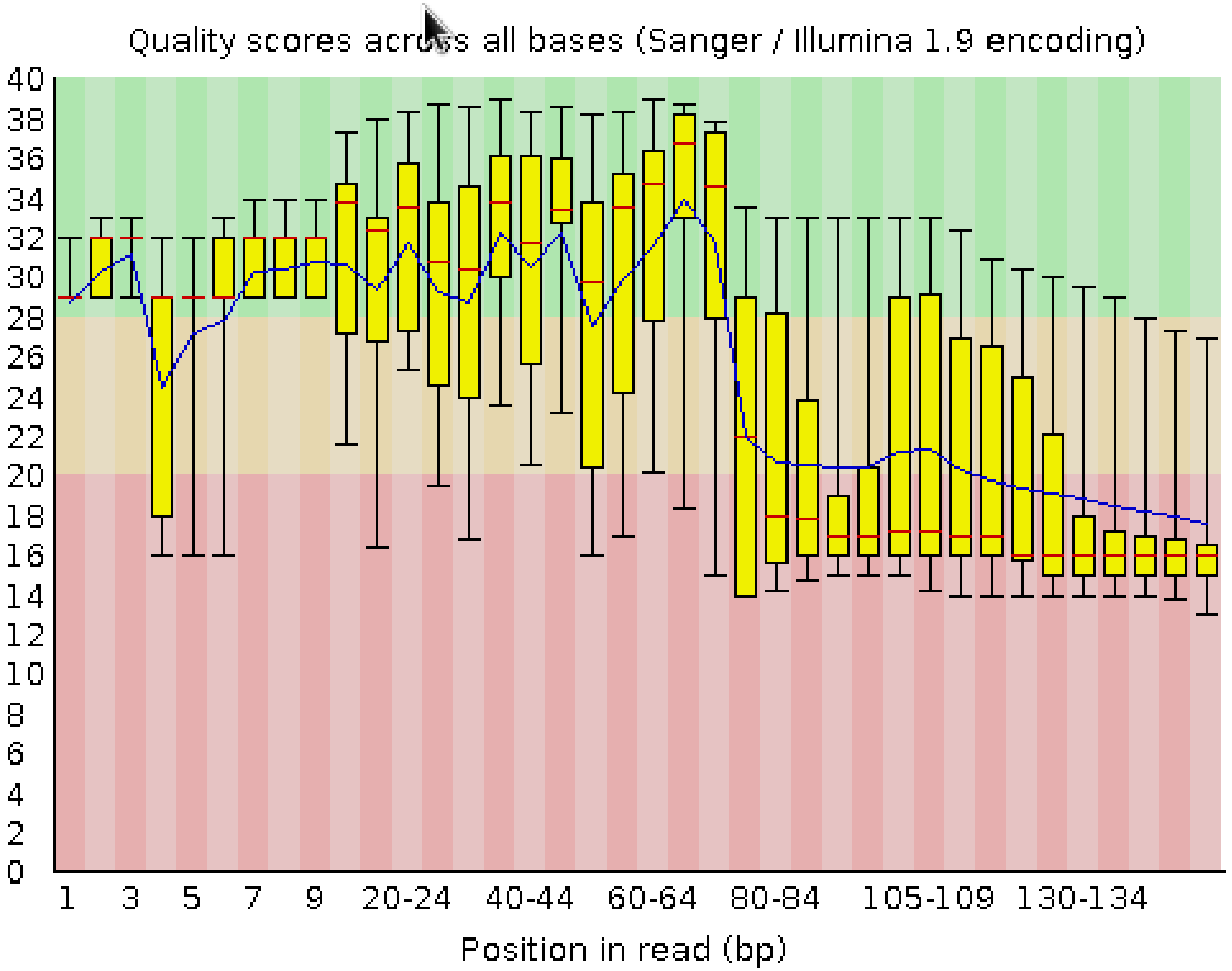
good quality

poor quality

Failed QC Examples

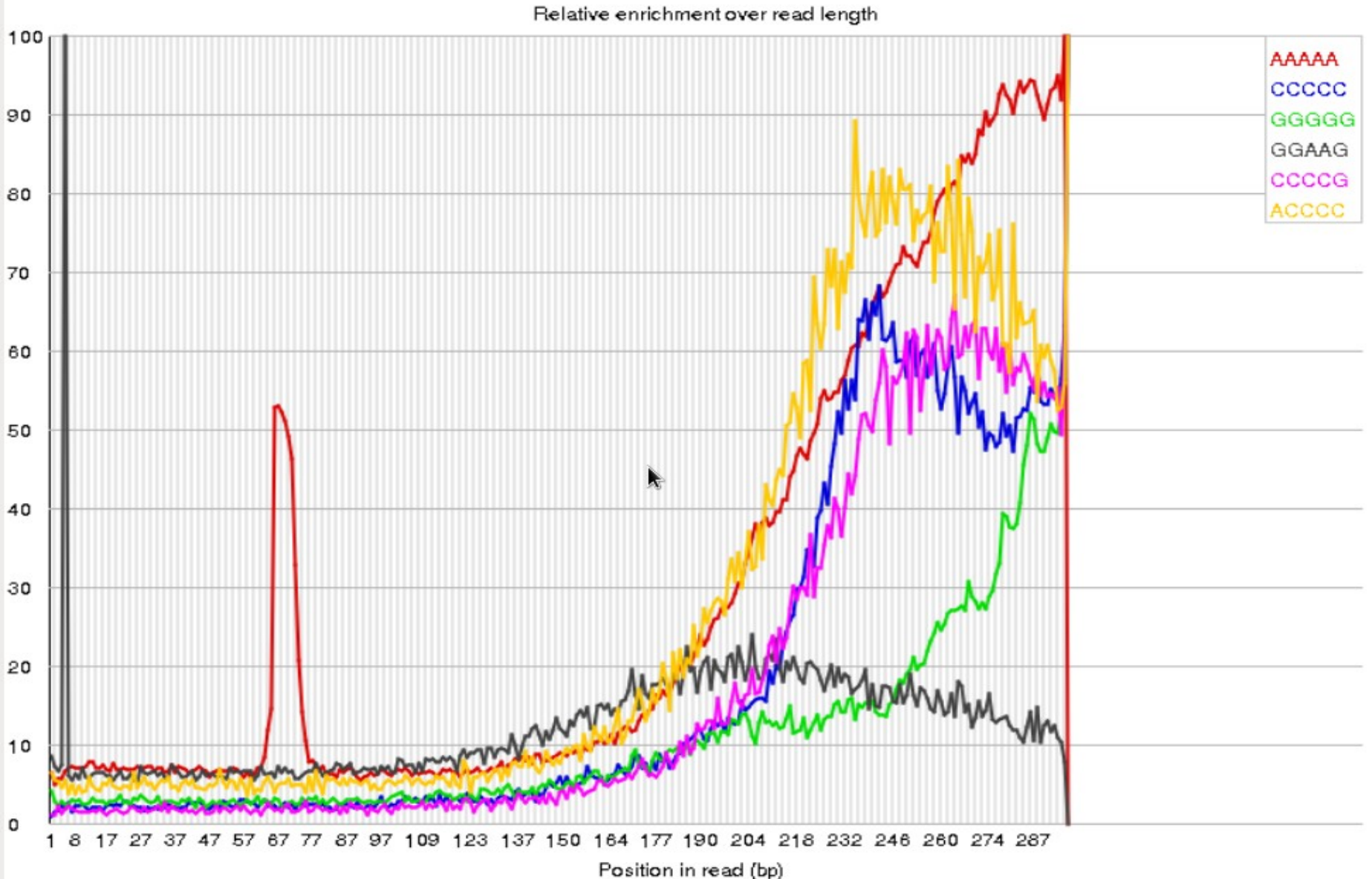
FastQC Output Image
Failed Per base sequence quality

Example 1. Expired MiSeq mate pair kit (9 months)



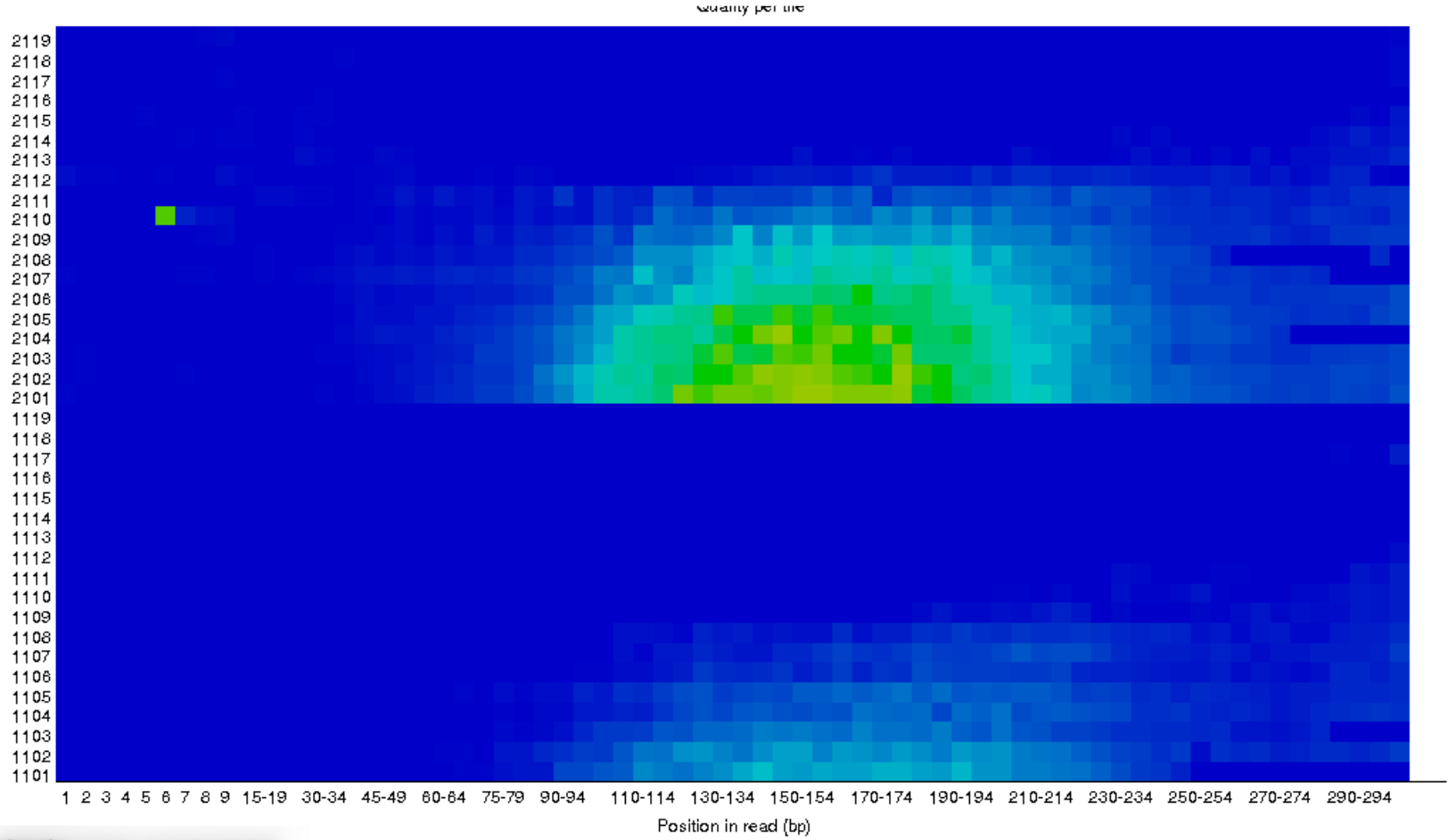
FastQC Output Image
Failed Kmer Content

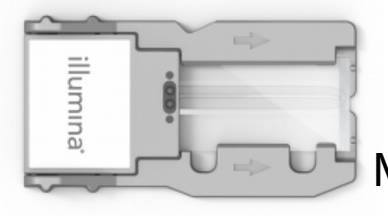
Example 2. Sequence prep adapters still on library fragments



FastQC Output Image
Flowcell: not so good per_tile quality

Example 3. Faulty flowcell





MiSeq flowcell

good quality  poor quality

QC Quality Trimming

- Trim reads based on quality scores
  - Trim the same number of bases from each read
  - Use a sliding window to calculate average quality at ends of sequences
  - Decide if you want to discard reads with Ns
    - some assemblers replace Ns with As or a random base G, C, A or T
- Trim adapter sequences
  - Trimmomatic has a file of Illumina adapter sequences
- Sequence quality trimming tools

```
module load Trimmomatic
```

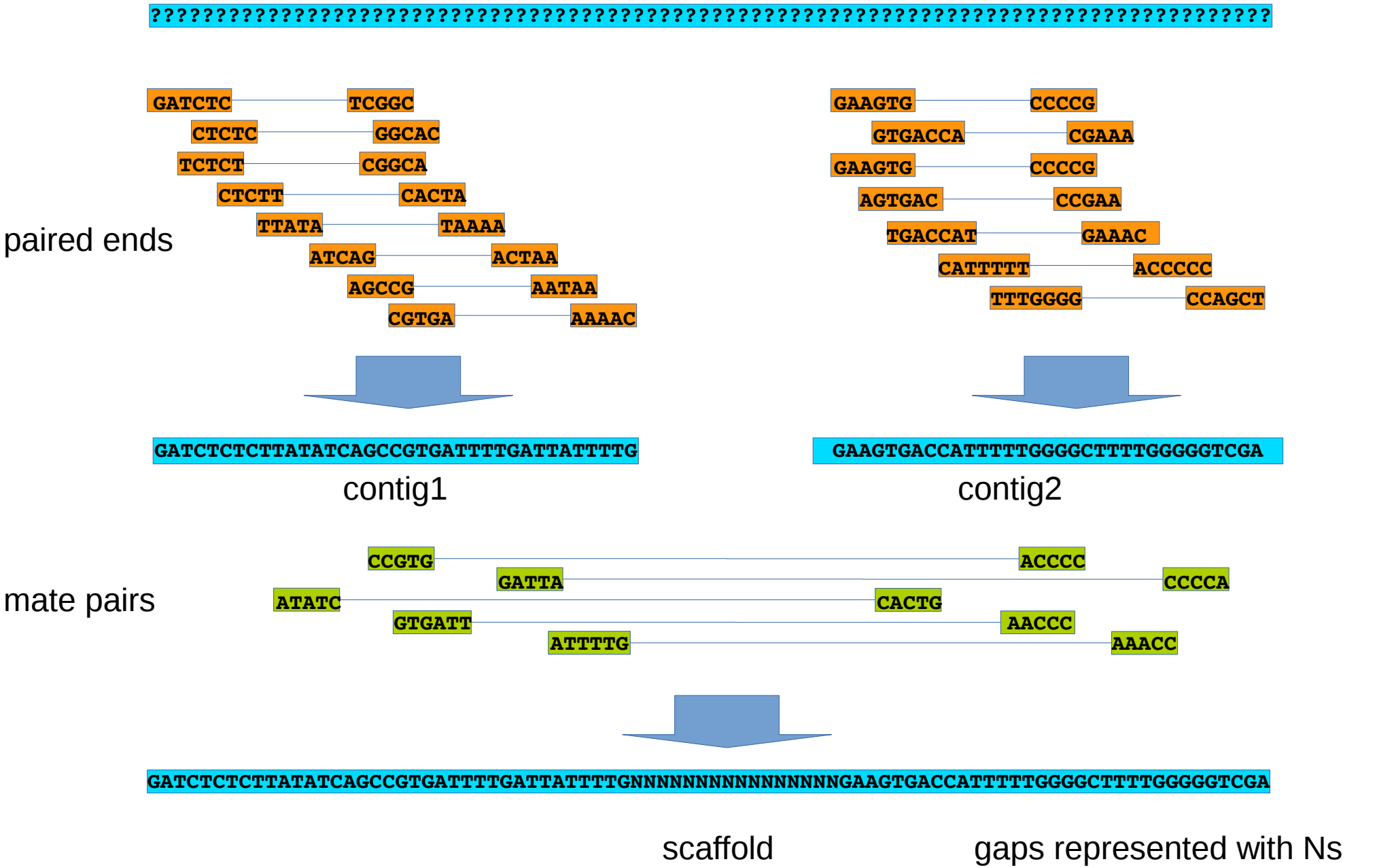
```
module load Cutadapt
```

Is QC Trimming Always Necessary?

- GATK variant calling pipeline starts with raw sequence data.
- MaSuRCA *de novo* genome assembler requires untrimmed and non error corrected reads
- "... no or modest trimming results in the most biologically accurate gene expression estimates" (RNA-seq)
  - Williams, et al. 2016 BMC Bioinformatics. 186/s12859-016-0956-2
- Variant Calling
  - "read preprocessing step did not improve the accuracy of variant calling"
    - Liu, *et al*. 2012. BMC Genomics 10.1186/1471-2164-13-S8-S8
  - "Trimming is shown to increase the quality and reliability of the analysis, with concurrent gains in terms of execution time and computational resources needed."
    - Del Fabbro, *et al*. 2013. PloS ONE. 10.1371/journal.pone.0085024
  - "Masking as an effective quality control method for next-generation sequencing data analysis
    - Yun & Yun. 2014. BMC Bioinformatics. 10.1186/s12859-014-0382-2

de novo Genome Assembly

Assembly Results in Contigs and Scaffolds



de novo Assemblers Installed on Ada

- `module load SPAdes` small genomes
- `module load Velvet` small genomes
- `module load MaSuRCA` any size genome
- `module load ABySS` any size genome
- `module load discovar` any size genome (DISCOVAR de novo)
- `module load WGS` any size genome
- `module load ALLPATHS-LG` any size genome
- `module load SGA` any size genome

Assemblathon

- Assemblathon 2 results demonstrate that there is not one assembler that performs best at assembling different genomes

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

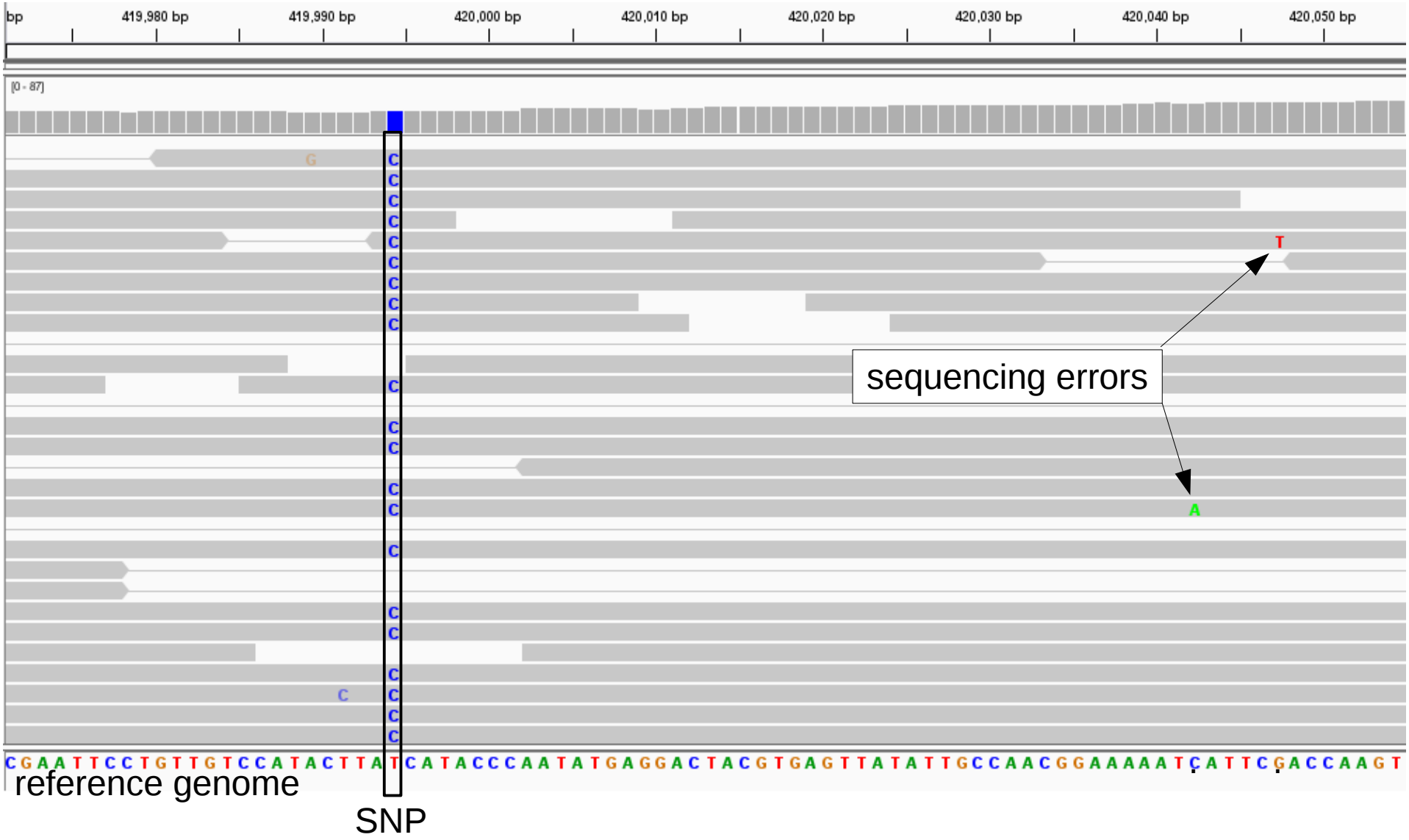
RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam[1*†], Joseph N Fass[1†], Anton Alexandrov[36], Paul Baranay[2], Michael Bechner[39], Inanç Birol[33],

Sequence Error Correction

Sequencing Errors



Tools for correcting sequencing errors

```
module load Lighter
```

Digital Normalization

Digital Normalization

- Reduce memory requirements by reducing the number of redundant sequence reads if you have a very high sequencing coverage (> 200x)

```
module load BBMap
```

Use the bbnorm.sh script in the BBMap module

A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data

C. Titus Brown[1,2,*], Adina Howe[2], Qingpeng Zhang[1], Alexis B. Pyrkosz[3], Timothy H. Brom[1]
1 Computer Science and Engineering, Michigan State University, East Lansing, MI, USA
2 Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA
3 USDA Avian Disease and Oncology Laboratory, East Lansing, MI, USA
* E-mail: ctb@msu.edu

Resolving Assembly Gaps

Gap Filling Results

134 Ns
before filling gap

TACACCTTCAGGAATATAGTTGGCTACTTTTGTTGGCAAACCTATGCCTAAGTTGACCAG
TGTATTGTGTTCTAATTCTTGTGCTACACGCTTTGCAATACGTGTTTGAATCTCTTCTTT
TGAAAGAACAGTAGCTGATGTTCCCATGTTAGGCTCCTTTCACTAAGTAGTTAACAAAAA
TGCCAGGAGTATGGACAAAATTAGGATCCATTTGCCCCACGTCAACAATTTCTCTTGCTT
CAACAATGGTGGTTTTCGCATTAGCAGCCATCACATGATTAAAGTTATTTTCAGAACCTG
CGTATTGCAAATTGCCATTTTTATCTGCCTTGTTAGCAAAAATAAGTGCGACATCAGCTT
TCAAAGGTTTTTCAAGAAGGTAGTCTTTGCCGTCAATAGTAATGACTTCTTTTCCTTTG**N**
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNTTTGTCGGCCCGCTTCACGGTTGAGACCGATATGTGATGCAATAATAG
TTGAAAACTGTTTATTGGCAACCATCTTACCAACGCCTTTGTCAGGAAAACCGGCATCGT
TACAGATTAAAGTGAGGTCTTTGACACCTTTTTCCACAAGTGCATCAATTAATTTTTCTG
GTGTGCCATTTGTCATGAAGCCACCAACCATAATGGTATCACCATCTTTAACGTGAGCTA
CCGCTTCTTTTATTGTGATTTCTTTACACATTTTTTCTACCAACTTTCACGAGACTAAGT
ATTGTTTTTCACGATAACAGCAGTTCCTTGTCCTCCACCAATGCAAAGTGTTGCTAGTCC

contig

contig

scaffold

121 Ns filled
after filling gap

TACACCTTCAGGAATATAGTTGGCTACTTTTGTTGGCAAACCTATGCCTAAGTTGACCAG
TGTATTGTGTTCTAATTCTTGTGCTACACGCTTTGCAATACGTGTTTGAATCTCTTCTTT
TGAAAGAACAGTAGCTGATGTTCCCATGTTAGGCTCCTTTCACTAAGTAGTTAACAAAAA
TGCCAGGAGTATGGACAAAATTAGGATCCATTTGCCCCACGTCAACAATTTCTCTTGCTT
CAACAATGGTGGTTTTCGCATTAGCAGCCATCACATGATTAAAGTTATTTTCAGAACCTG
CGTATTGCAAATTGCCATTTTTATCTGCCTTGTTAGCAAAAATAAGTGCGACATCAGCTT
TCAAAGGTTTTTCAAGAAGGTAGTCTTTGCCGTCAATAGTAATGACTTCTTTTCCTTTG**G**
CAACTTCGGTTCCTATTCCAGTAGGTGTTAAGAAACCGCCTAGCCCAAAACCGCCACTGC
GGATGCGCTCAGCTAATGTTCCCTGAGGTACAAGATCAATTACAGTCTCGCCCTCAGTCA
TTTGTCGGCCCGCTTCACGGTTGAGACCGATATGTGATGCAATAATAGTTGAAAACTGTT
TATTGGCAACCATCTTACCAACGCCTTTGTCAGGAAAACCGGCATCGTTACAGATTAAAG
TGAGGTCTTTGACACCTTTTTCCACAAGTGCATCAATTAATTTTTCTGGTGTGCCATTTG
TCATGAAGCCACCAACCATAATGGTATCACCATCTTTAACGTGAGCTACCGCTTCTTTTA
TTGTGATTTCTTTACACATTTTTTCTACCAACTTTCACGAGACTAAGTATTGTTTTTCAC
GATAACAGCAGTTCCTTGTCCTCCACCAATGCAAAGTGTTGCTAGTCCTCTGGTTACTTG

contig

module load GapFiller

Example of Gap in a Region of Low Coverage Coverage (< 15x)

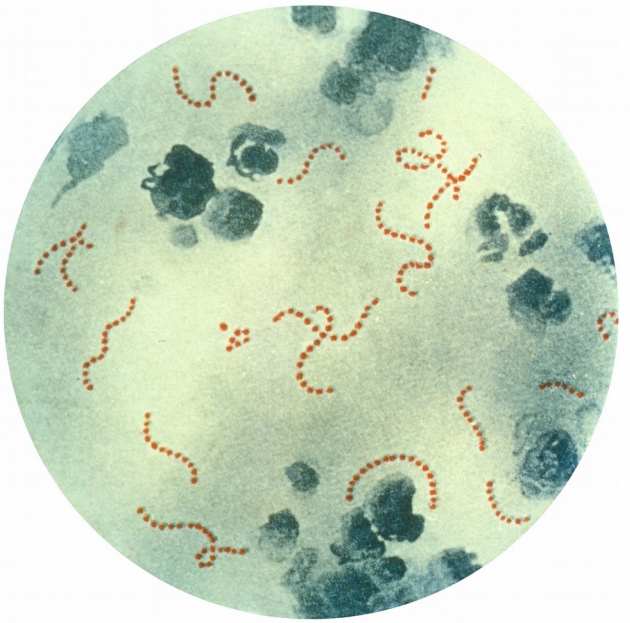
Evaluating Genome Assemblies

Genome Assembly Completeness

- The completeness of a genome can be estimated by using a set of highly conserved genes that are common to specific taxonomic groups
  - BUSCO – uses single-copy genes to access genome assembly and annotation completeness
    - `module load BUSCO`
    - evaluates % 'genes' complete, % fragmented, % missing
  - CEGMA – Core Eukaryotic Gene Mapping Approach: accesses a genome assembly for a small set of highly conserved genes.
    - Currently no further development planned
    - `module load CEGMA`

Genome Assembly Example

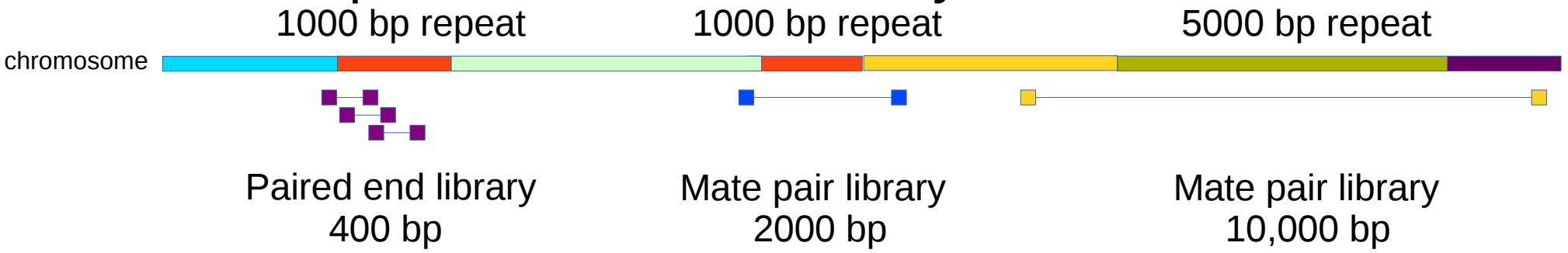
- *Streptococcus pyogenes*
- Bacteria that infect millions worldwide
- Circular DNA
  - ~1,850,000 bp long
  - ~1,800 protein coding genes
  - 20+ reference builds available



Centers for Disease Control and Prevention
http://phil.cdc.gov

Repeat Regions in *S. pyogenes*

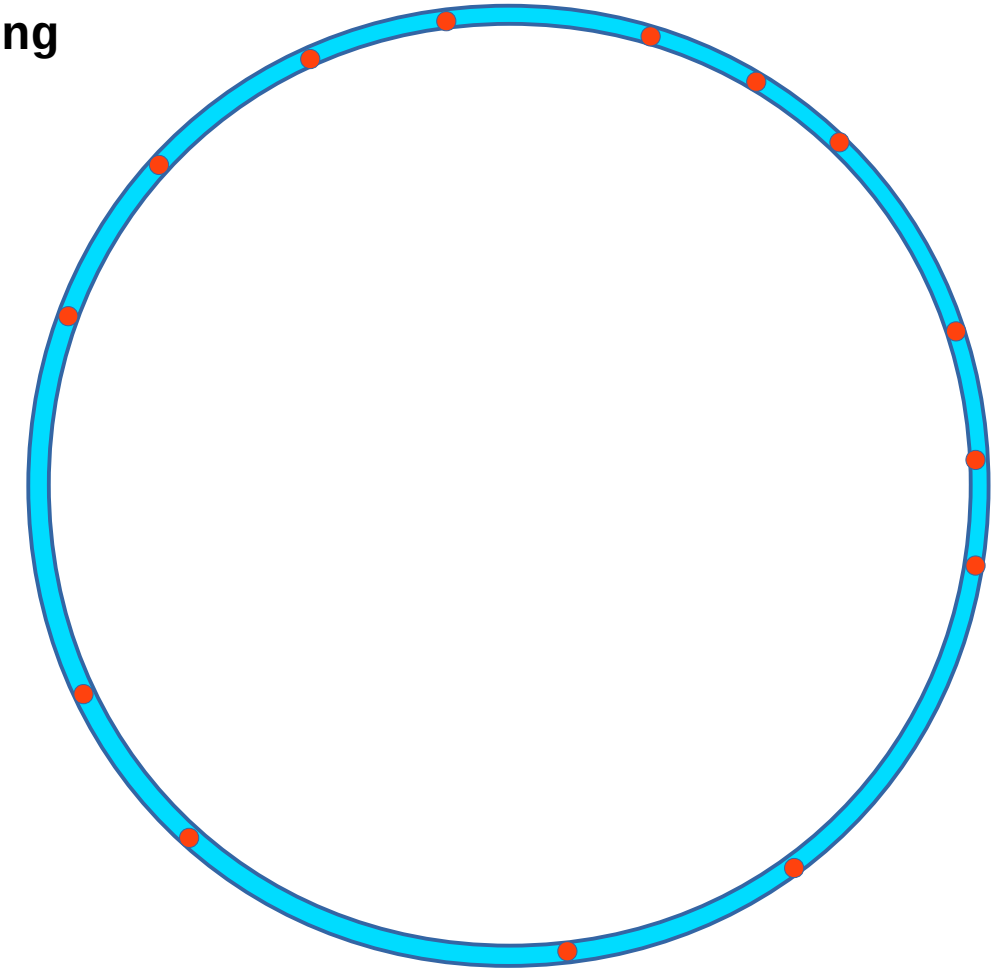
- Ribosomal Subunits
  - 16S 23S 5S tRNAs
  - 6 copies ~5000 bp each in *S. pyogenes* strain MGAS8232.
- Transposase Genes
  - 14 copies of a ~1000 bp transposase
  - 10 copies have 100% identity.



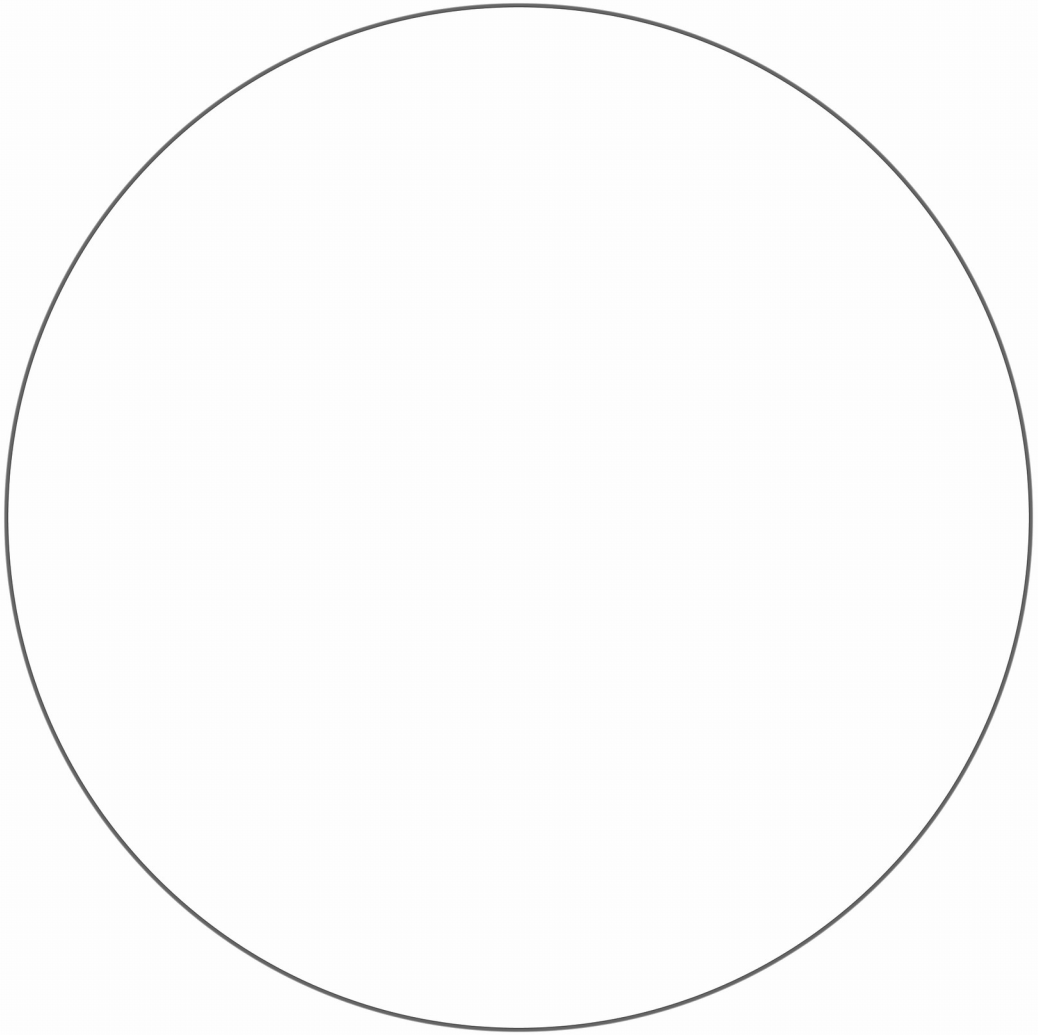
Transposase BLAST
Alignment to *S. pyogenes* strain MGAS8232

Gene symbol	spyM18_0426
Gene description	transposase
Locus tag	spyM18_0426
Gene type	protein coding

	Identity
1.	1026/1026(100%)
2.	974/974(100%)
3.	977/979(99%)
4.	973/973(100%)
5.	973/973(100%)
6.	973/973(100%)
7.	972/972(100%)
8.	972/972(100%)
9.	972/972(100%)
10.	972/972(100%)
11.	972/972(100%)
12.	743/745(99%)
13.	743/745(99%)
14.	743/745(99%)

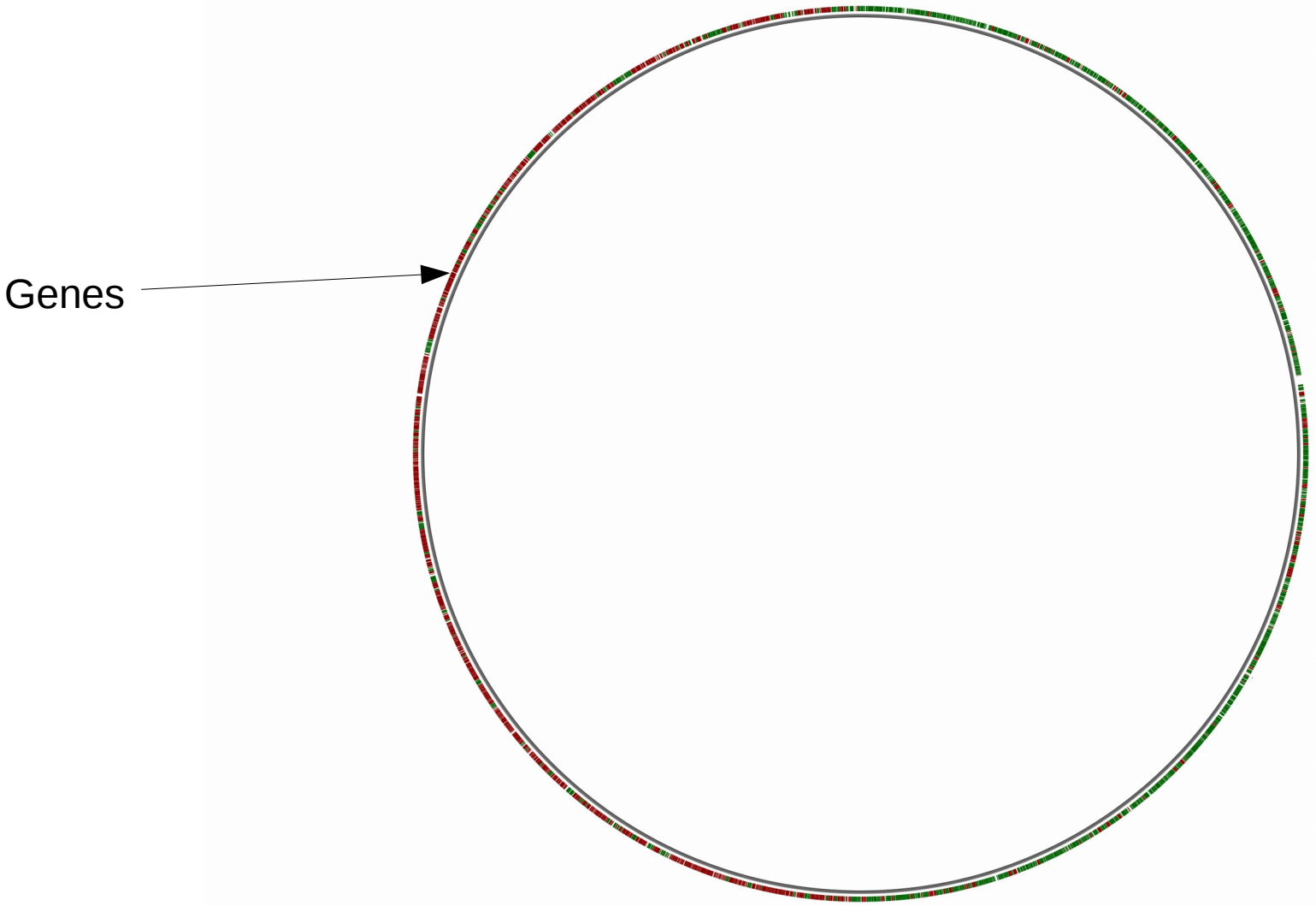


Visualizing a *de novo* Assembly with a Reference Genome



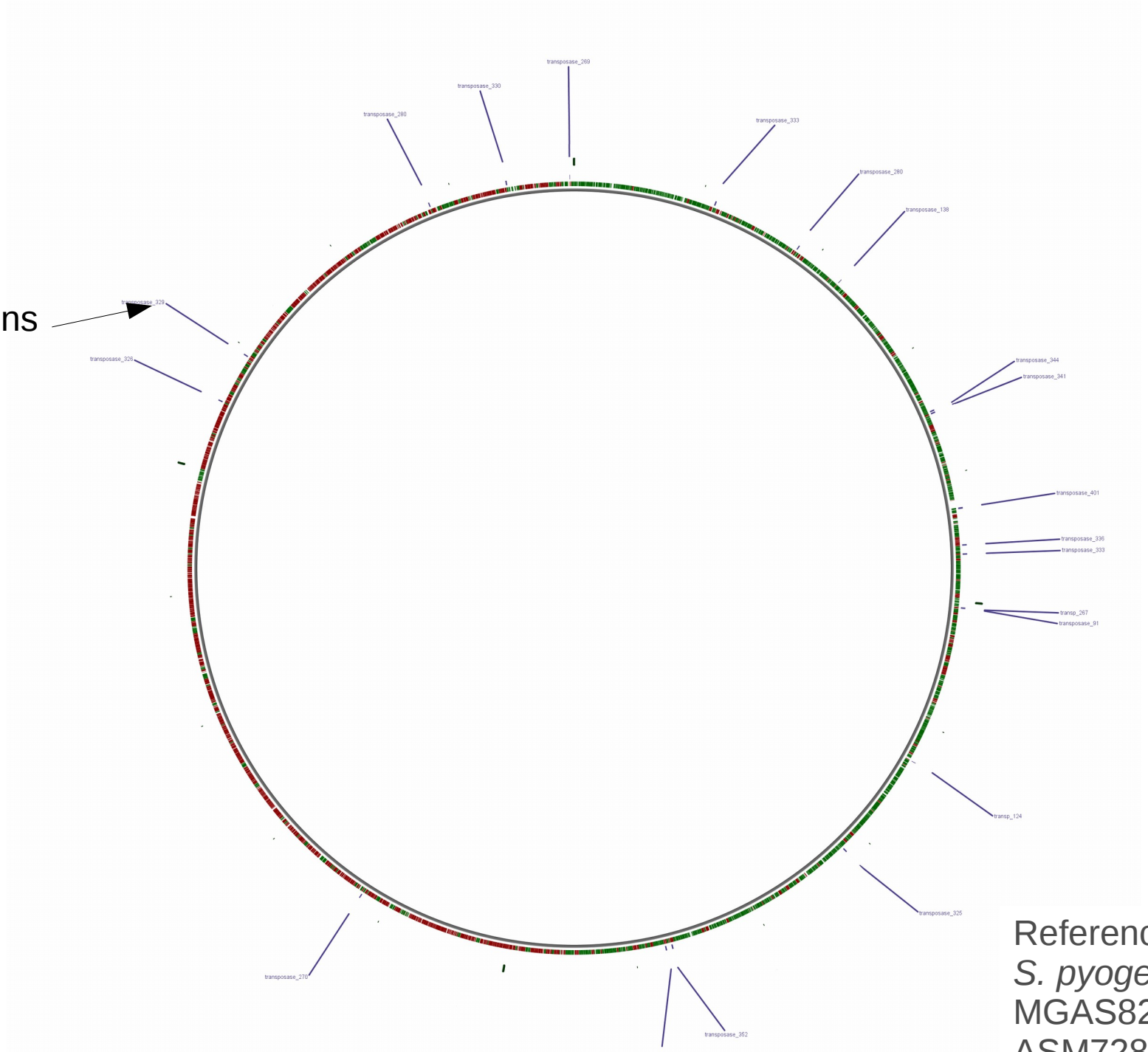
Reference genome:
S. pyogenes
MGAS8232
ASM728v1 (2002)
1,895,017 bp

Visualizing a *de novo* Assembly with a Reference Genome

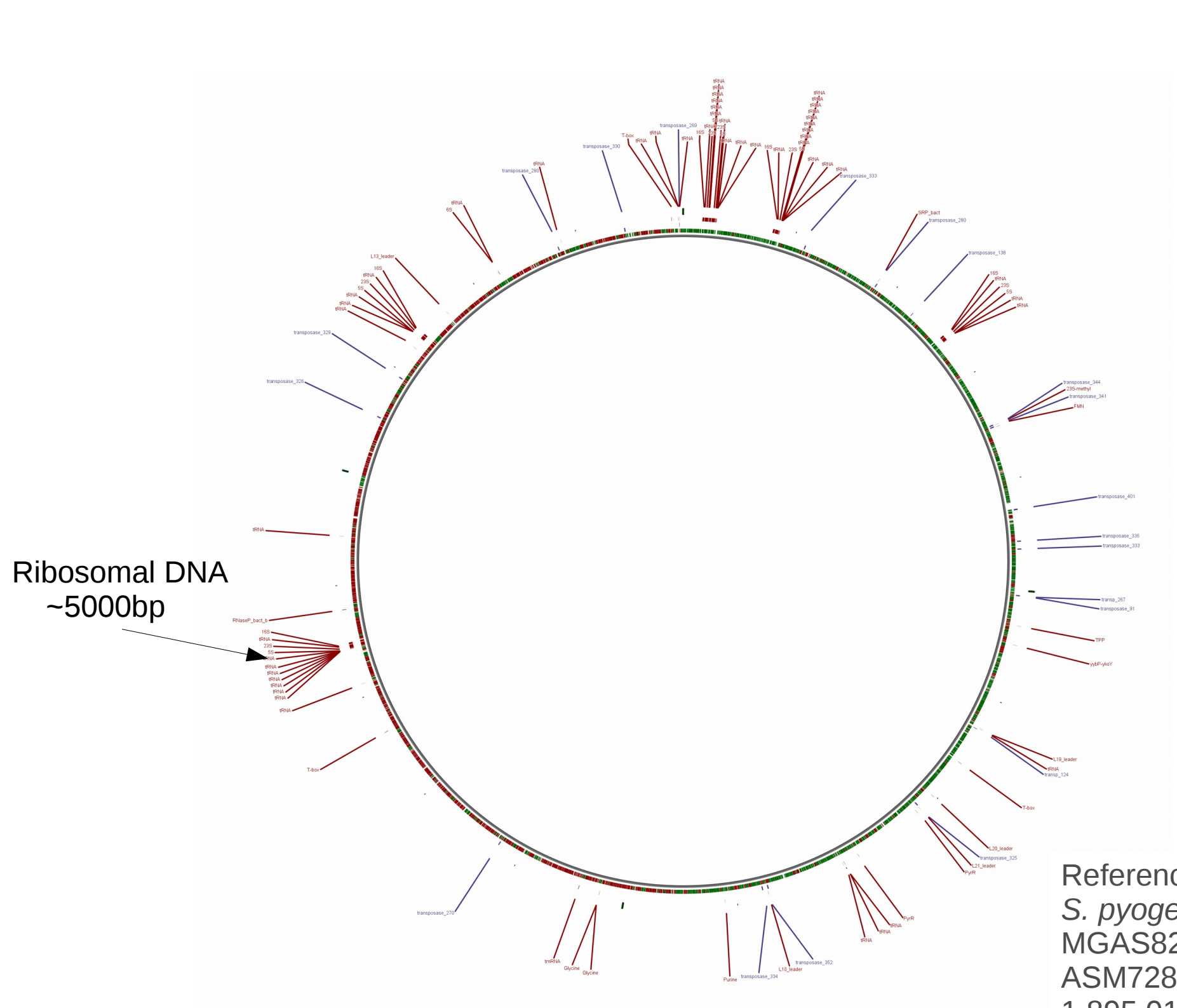


Reference genome:
S. pyogenes
MGAS8232
ASM728v1 (2002)
1,895,017 bp

Transposons
~1000bp

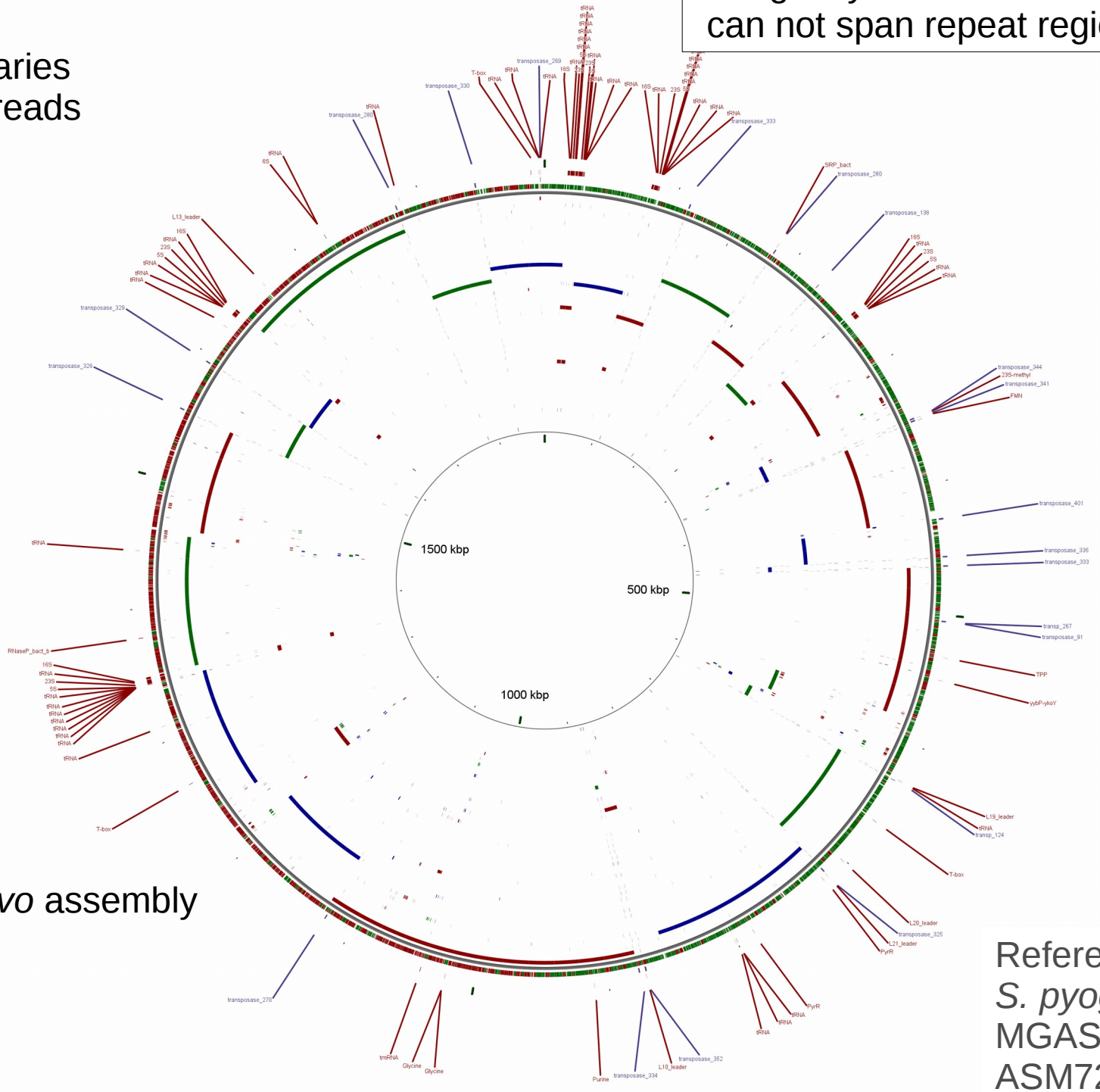


Reference genome:
S. pyogenes
MGAS8232
ASM728v1 (2002)
1,895,017 bp


Ribosomal DNA
~5000bp
Reference genome:
S. pyogenes
MGAS8232
ASM728v1 (2002)
1,895,017 bp

Using only Paired end library (400 bp) can not span repeat regions

Assembly libraries
- paired end reads



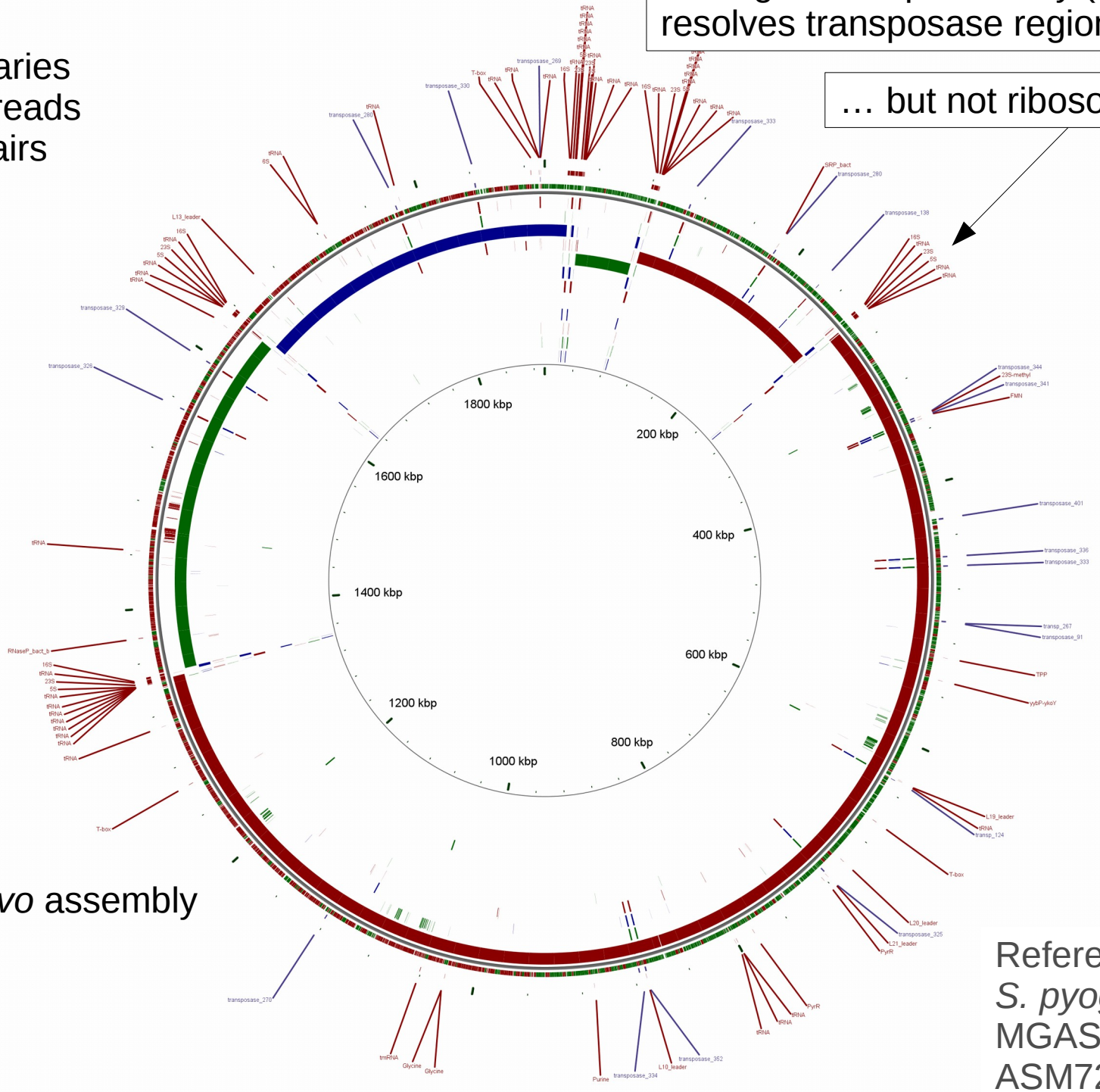
Aligned *de novo* assembly
39 contigs

Reference genome:
S. pyogenes
MGAS8232
ASM728v1 (2002)
1,895,017 bp

Assembly libraries
- paired end reads
- 2kb mate pairs

Adding a mate pair library (2000 bp) resolves transposase regions (1,000 bp)

… but not ribosomal subunits

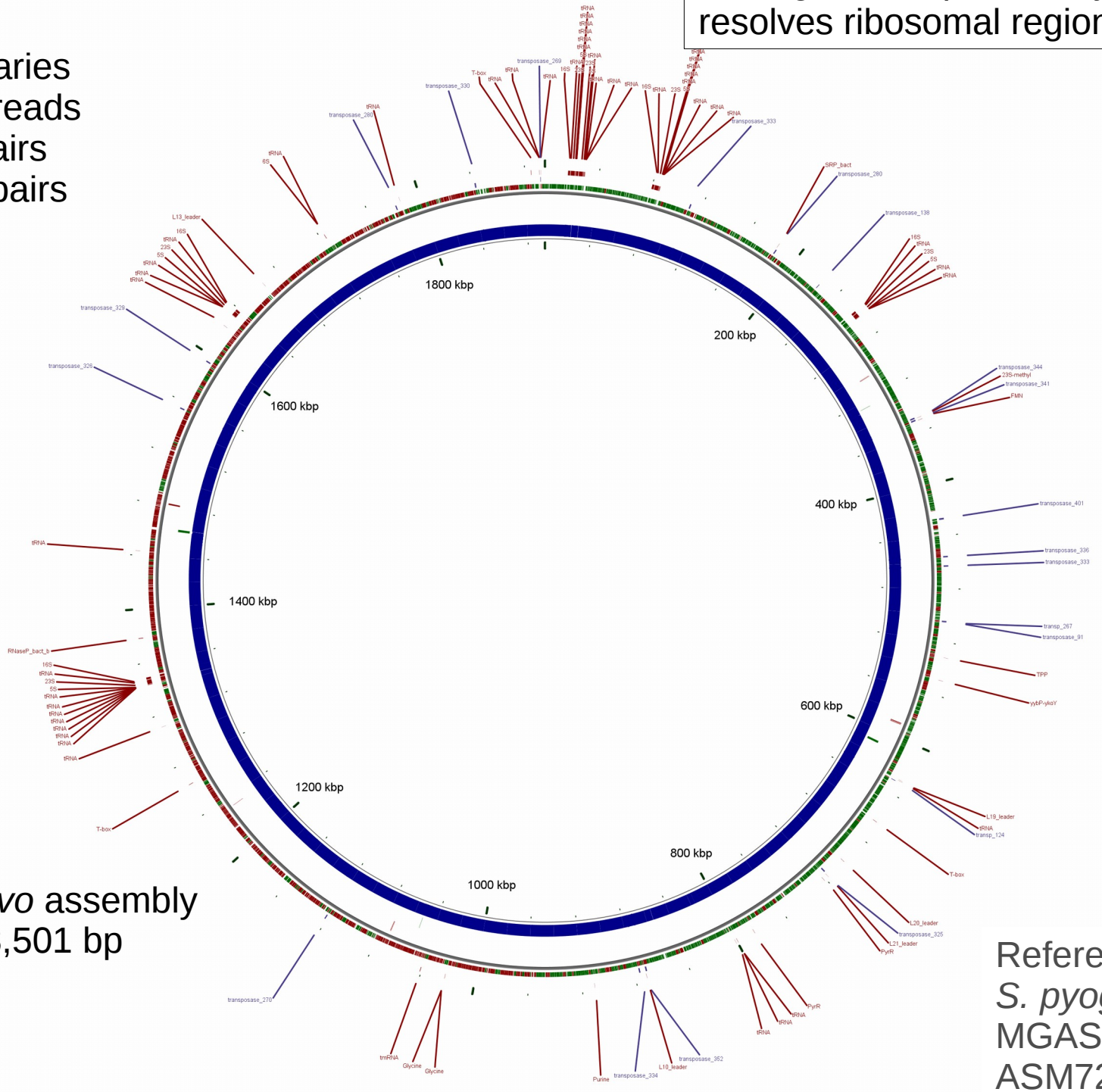


Aligned *de novo* assembly
12 contigs

Reference genome:
S. pyogenes
MGAS8232
ASM728v1 (2002)
1,895,017 bp

Adding a mate pair library (10,000 bp) resolves ribosomal regions (5,000 bp)

Assembly libraries
- paired end reads
- 2kb mate pairs
- 10kb mate pairs



Aligned *de novo* assembly
1 contig 1,878,501 bp

Reference genome:
S. pyogenes
MGAS8232
ASM728v1 (2002)
1,895,017 bp

Mapping Reads to a Reference Assembly

Sequence Read Alignment Tools

- `module load bowtie` RNA-seq
- `module load bowtie2` RNA-seq
- `module load STAR-STAR` RNA-seq
- `module load bwa` DNA-seq
- `module load SNAP` DNA-seq
- `module load BFAST` DNA-seq
- `module load GMAP-GSNAP` DNA-seq, RNA-seq
- `module load Subread` DNA-seq, RNA-seq
- `module load HISAT2` DNA-seq, RNA-seq
- `module load Stampy` DNA-seq, RNA-seq, ChIP-seq

Pre-indexed Genomes

```
/scratch/datasets/genome_indexes/ucsc/
```

(link to the indexes directly instead of copying to your directory)

- bowtie

```
/scratch/datasets/genome_indexes/ucsc/mm10/bowtie_index/
```

- bowtie2

```
/scratch/datasets/genome_indexes/ucsc/mm10/bowtie2_index/
```

- HISAT2

```
/scratch/datasets/genome_indexes/ucsc/mm10/hisat2_index/
```

- bwa

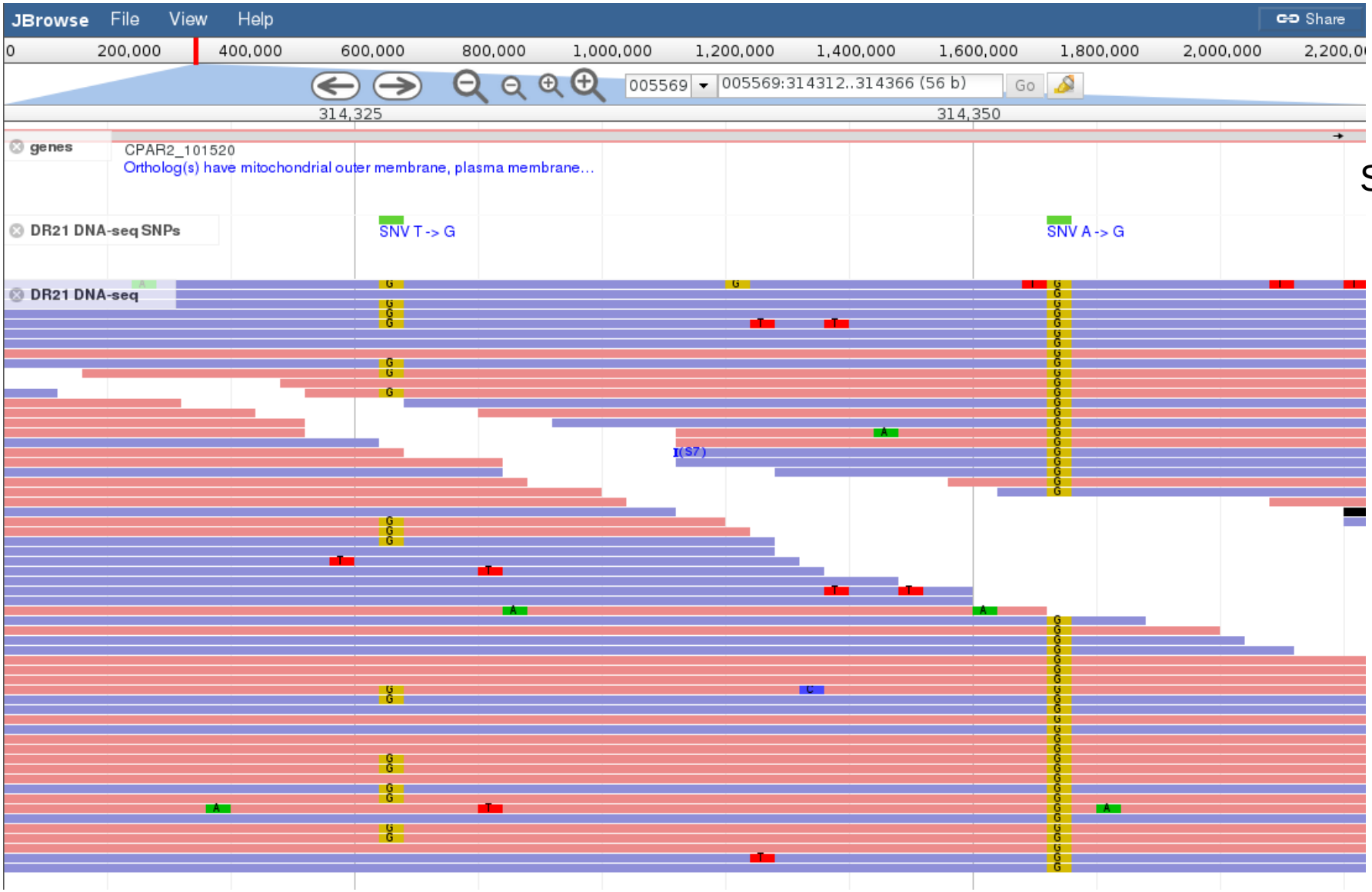
```
/scratch/datasets/genome_indexes/ucsc/mm10/bwa-0.7.12_index/
```

sam and bam format

- sam: Sequence Alignment/Map format
  - can view using the UNIX command: `more file.sam`
  - use samtools to sort sam file based on coordinate and convert to bam format
- bam: Binary Alignment/Map format
  - compressed sam file
  - Create an index of the bam file using samtools
    - The samtools index is needed by genome browsers to view bam files
    - `samtools index file.bam`
  - Viewing bam files using samtools
    - `samtools view file.bam | more` view only alignments
    - `samtools view -H file.bam` view only header
    - `samtools view -h file.bam | more` view header + alignments

Variant Calling and Annotation

Viewing SNPs in a Diploid Organism



Strand:
Blue = +
Red = -

heterozygous SNP

homozygous SNP

Variant Calling General Workflow

- Align reads to a reference
  - bwa
  - SAMtools: Sort alignments based on reference coordinates & create bam file
  - picard: Mark PCR duplicates (depends on variant calling tool)
    - PCR duplicates are artifacts resulting from a PCR amplification step during NGS library preparations.
    - PCR duplicates should be removed/marked as to not bias the frequency of variants or gene expression levels
- Call variants (SNPs and indels) distinguishing between sequencing errors and variants
  - SAMtools, BCFtools
  - Freebayes
  - VarScan
  - GATK
  - Platypus
- Annotate amino acid change using snpEff
  - missense, stop_gained, frameshift, synonymous, ...

Variant Calling Tools

- GATK

```
module load GATK BWA picard SAMtools
```

  - No need to QC trim reads, the GATK best practices pipeline will perform the necessary steps including marking PCR duplicates
  - You need a set of known variants for your species (dbSNP) or perform bootstrapping of your SNP population
  - Used in conjunction with other tools
    - bwa
    - samtools
    - picard
- SAMtools and BCFtools

```
module load SAMtools BCFtools
```

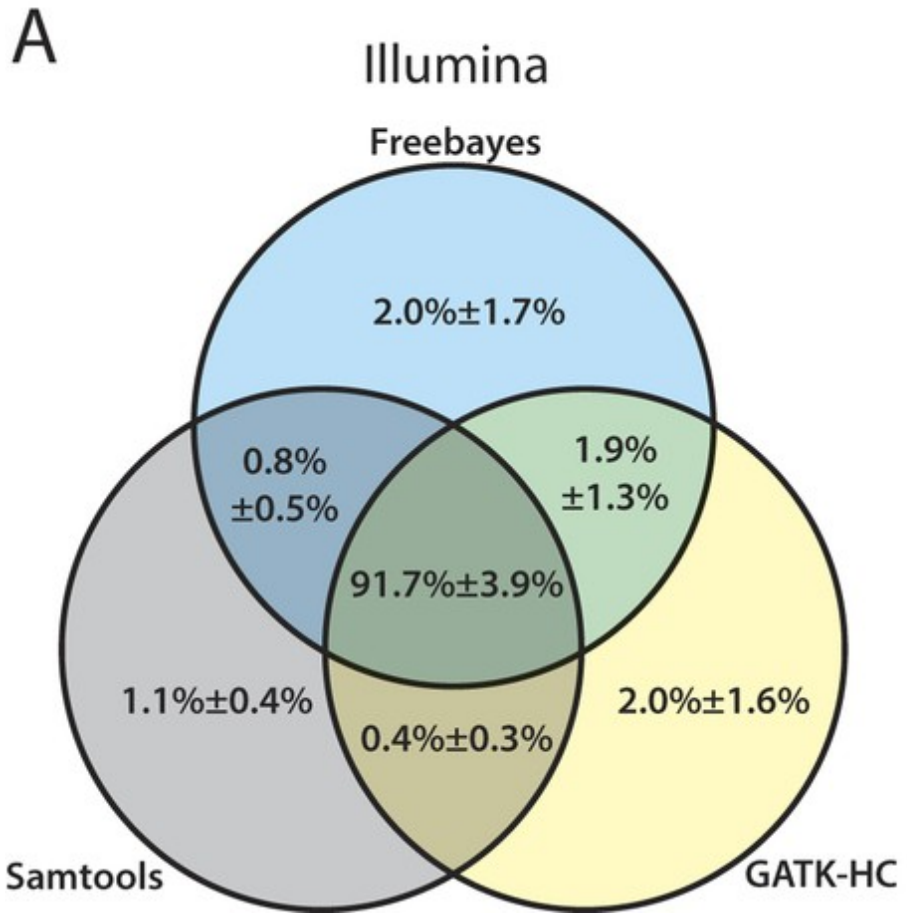
- VarScan

```
module load VarScan
```

- FreeBayes

```
module load FreeBayes
```

Summarizing Variant Calls from Different Tools



The mean percentage with standard deviation of confidence variant calls with equal to or higher than the quality score threshold of 20 are represented for (A) Illumina data sets

Huang et al. 2015 doi:10.1038/srep17875

Annotate Variants

- A file of variant calls in vcf format is needed
- A reference sequence with gene annotations is needed
- snpEff annotates a vcf file

```
module load snpEff
```

  - There are > 2,500 pre-built databases available
  - Annotates MNP (multiple nucleotide polymorphism)
    - Codon change due to two SNPs: ACA → GGA

```
5       325795  .       AC      GG      23.8901 .
AB=0.428571;ABP=3.32051;AC=1;AF=0.5;AN=2;AO=3;CIGAR=2X;DP=7;DPB=7;DPRA=0;EPP=3.73412;
EPPR=3.0103;GTI=0;LEN=2;MEANALT=1;MQM=33;MQMR=48.5;NS=1;NUMALT=1;ODDS=5.49681;PAIRED=0;
PAIREDR=0.5;PAO=0;PQA=0;PQR=0;PRO=0;QA=114;QR=150;RO=4;RPL=3;RPP=9.52472;RPPR=3.0103;
RPR=0;RUN=1;SAF=2;SAP=3.73412;SAR=1;SRF=2;SRP=3.0103;SRR=2;TYPE=mnp;technology.ILLUMINA=1;
ANN=GG|missense_variant|MODERATE|CD36_51230|CD36_51230|transcript|CAX41505.1|
protein_coding|1/1|c.1657_1658delACinsGG|p.Thr553Gly|1657/1851|1657/1851|553/616||
GT:DP:RO:QR:AO:QA:GL    0/1:7:4:150:3:114:-6.7054,0,-11.1847
```

Variant Calling Considerations

- Start with aligning reads to a reference
  - GATK does not require QC trimming
- Differentiate between sequencing errors and SNPs
  - Variants can be filtered based on alignment depth
  - Calling SNPs may require 1/3 of reads to contain SNP
- BLAST reads with SNPs to identify variant calls due to misalignment especially within duplicated genes
- Annotate SNPs (snpEff)
- Identify multiple-nucleotide polymorphism (MNP)
  - Two SNPs within a single codon

	codon	
Reference:	**TTT**	Phe
SNP 1:	**TTA**	Leu
SNP 2:	**TAT**	Tyr
SNP 1 + 2:	**TAA**	**STOP**

Consequence of Amino Acid Change

- Assess consequence of amino acid change based on sequence conservation across species using the PROVEAN tool (available as a module on Ada cluster)

PROVEAN Result (Download)

PROVEAN Prediction - Job ID: 170651083923296

- **Query sequence** (fasta)
- **Supporting sequence set used for prediction**
  Number of sequences: **422** (fasta, E-values)
  Number of clusters: **30**
- **Score thresholds for prediction**
  (1) Default threshold is -2.5, that is:
  -Variants with a score equal to or below -2.5 are considered "deleterious,"
  -Variants with a score above -2.5 are considered "neutral."
  (2) How to use a more stringent threshold.

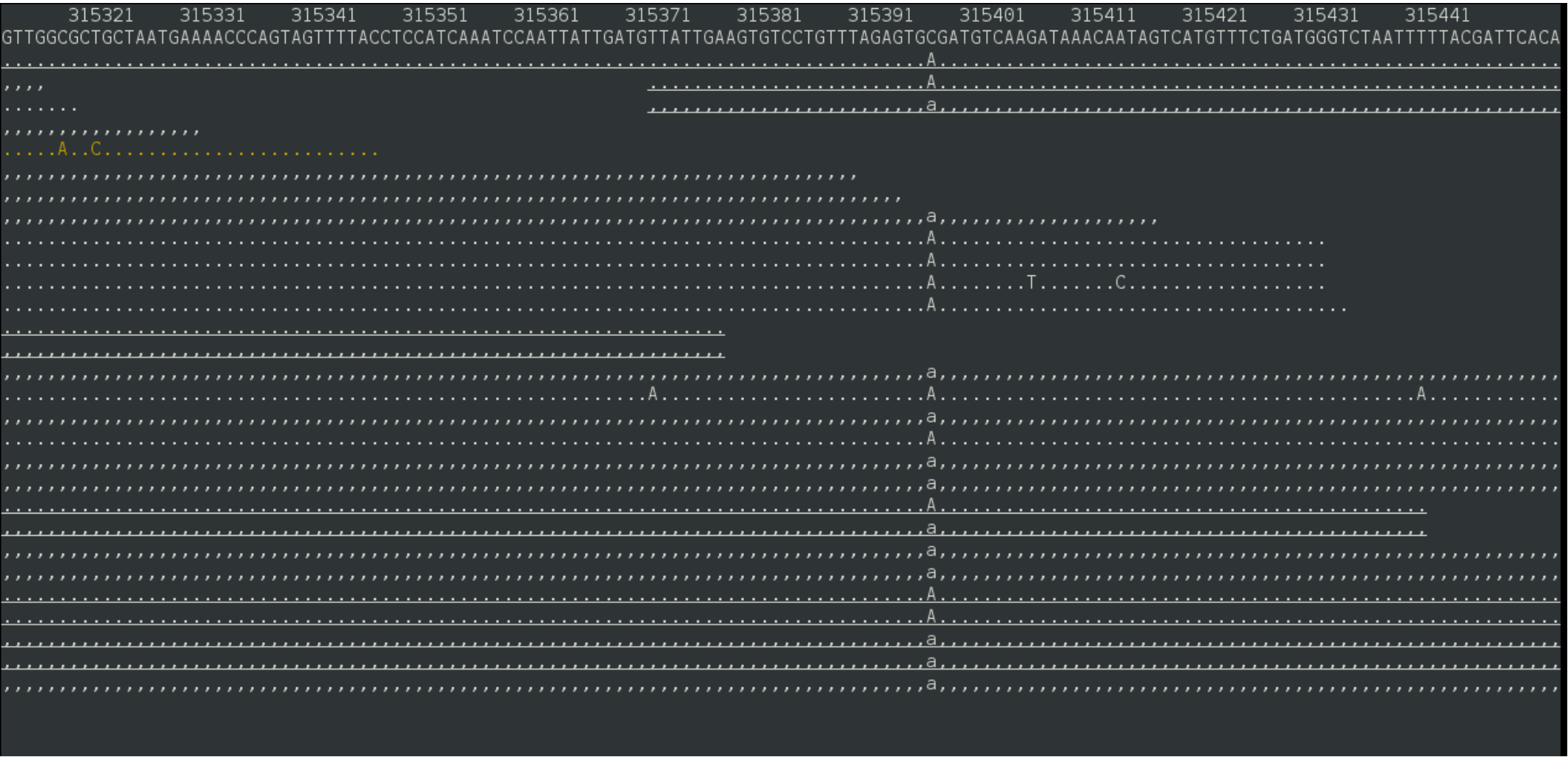
Variant	PROVEAN score	Prediction (cutoff= -2.5)
P72R	-0.230	Neutral
G105C	-8.276	Deleterious
K370del	-1.544	Neutral
H178_H179insPHP	-10.692	Deleterious
L22_W23delinsQS	-10.695	Deleterious

http://provean.jcvi.org

Alignment Viewing Tools

View Alignments with samtools

```
samtools tview sample_1_sorted.bam reference_genome.fa -p 1:315315
```



IGV (Integrative Genomics Viewer)

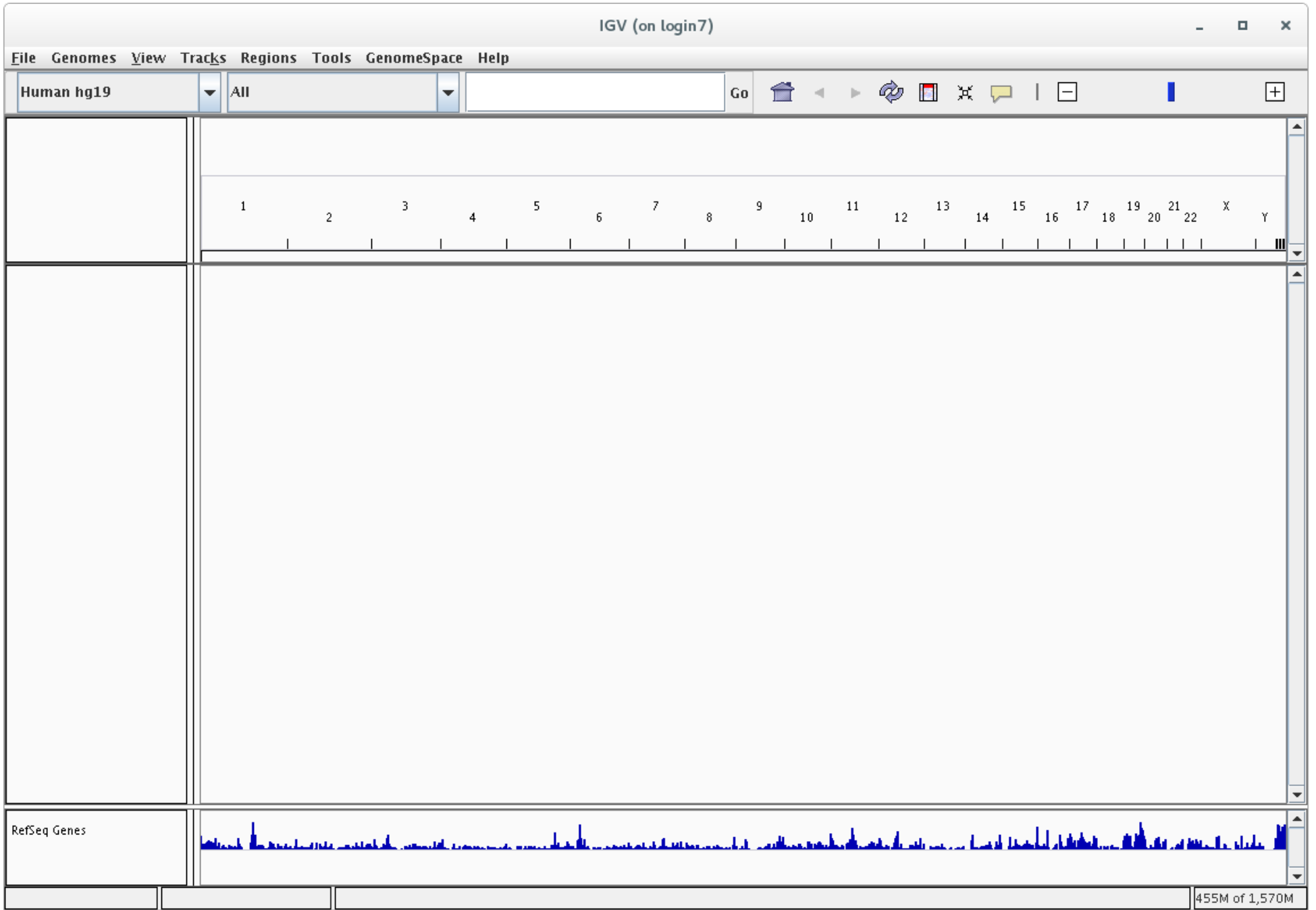
IGV is a genome browser with pre-loaded genomes available in which you can use to view .bed, .sam, .vcf files

```
module load IGV
```

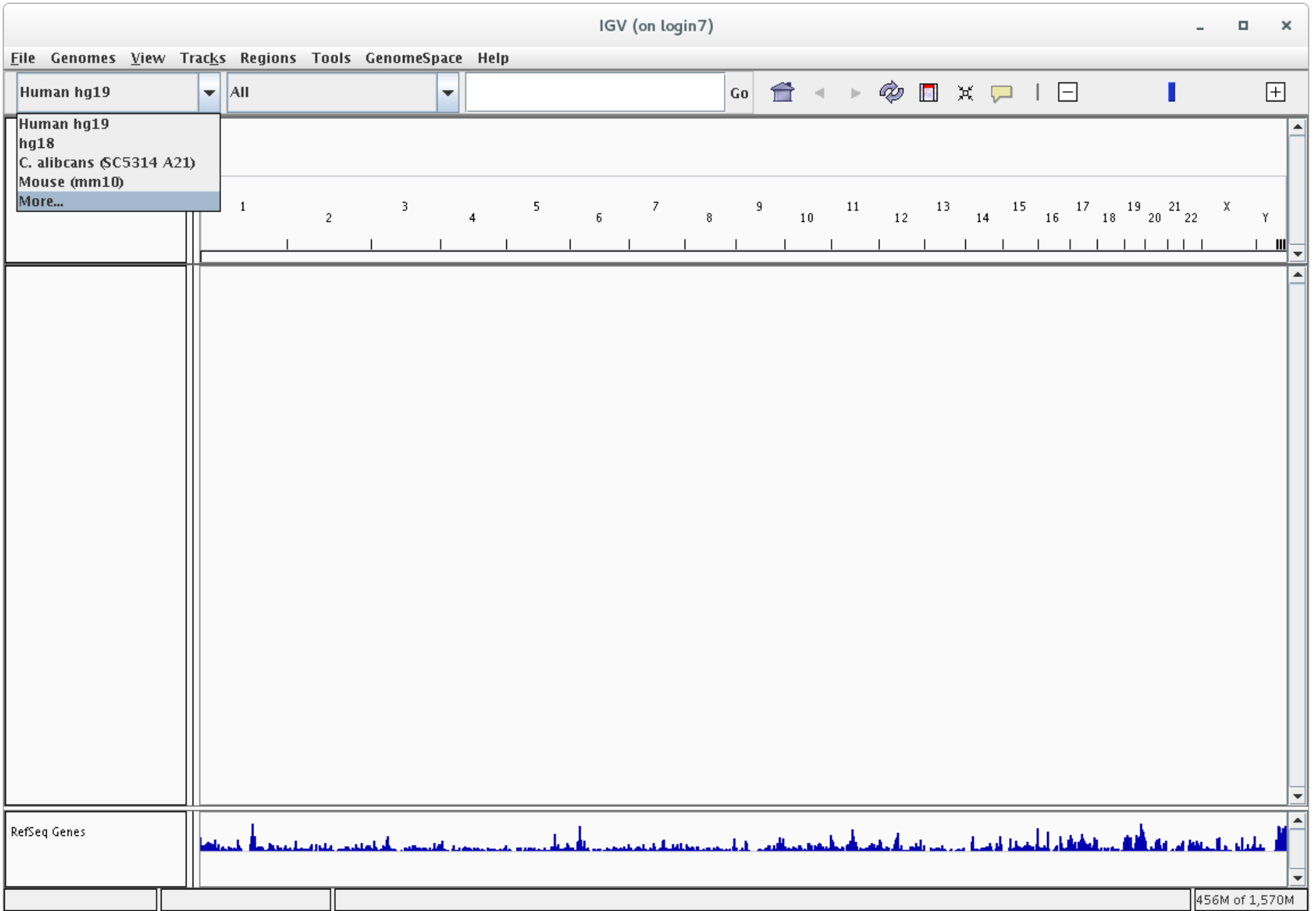
Launch IGV using the igv.sh script (X11 login required)

```
igv.sh
```

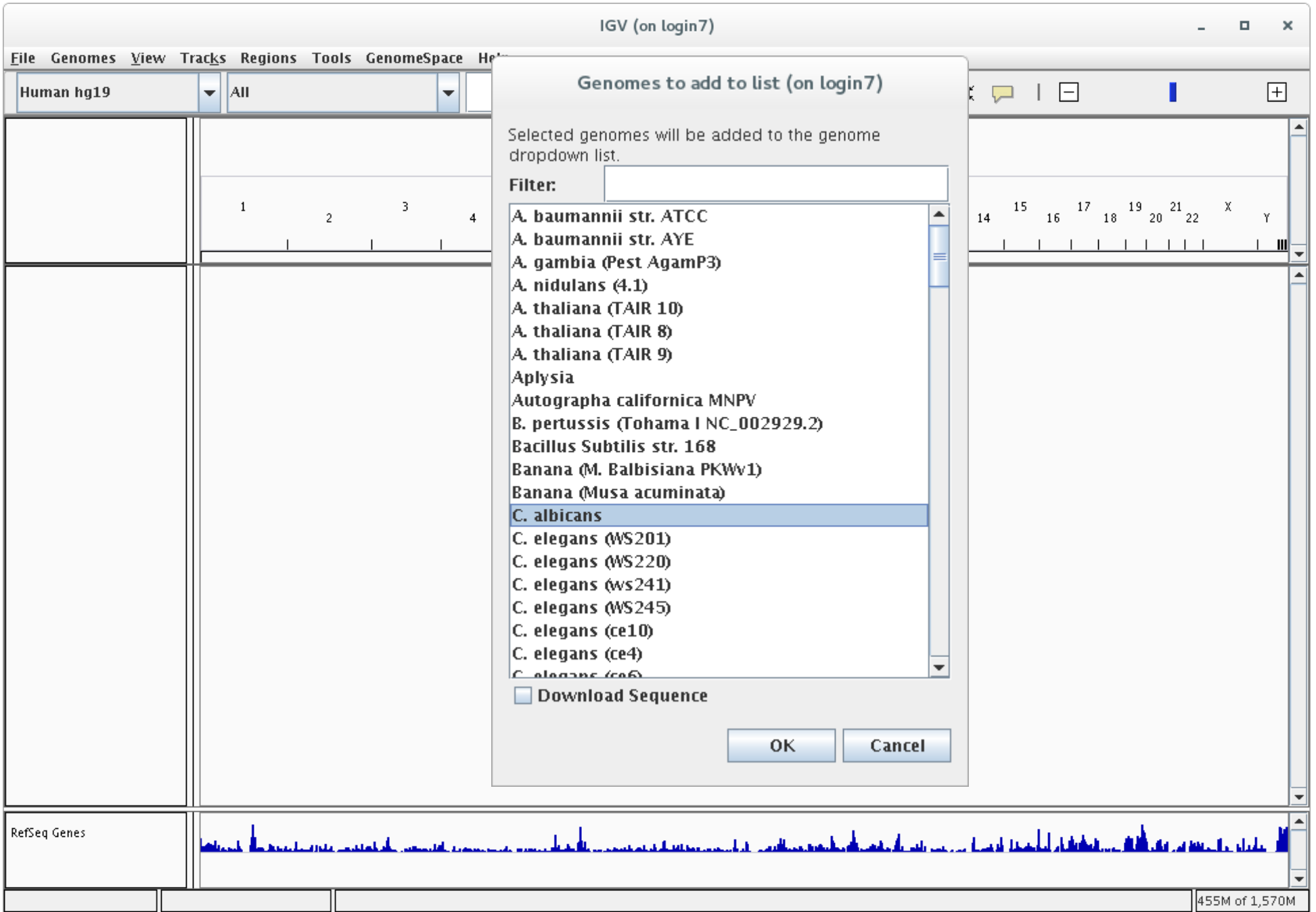
hg19 is default Reference Genome



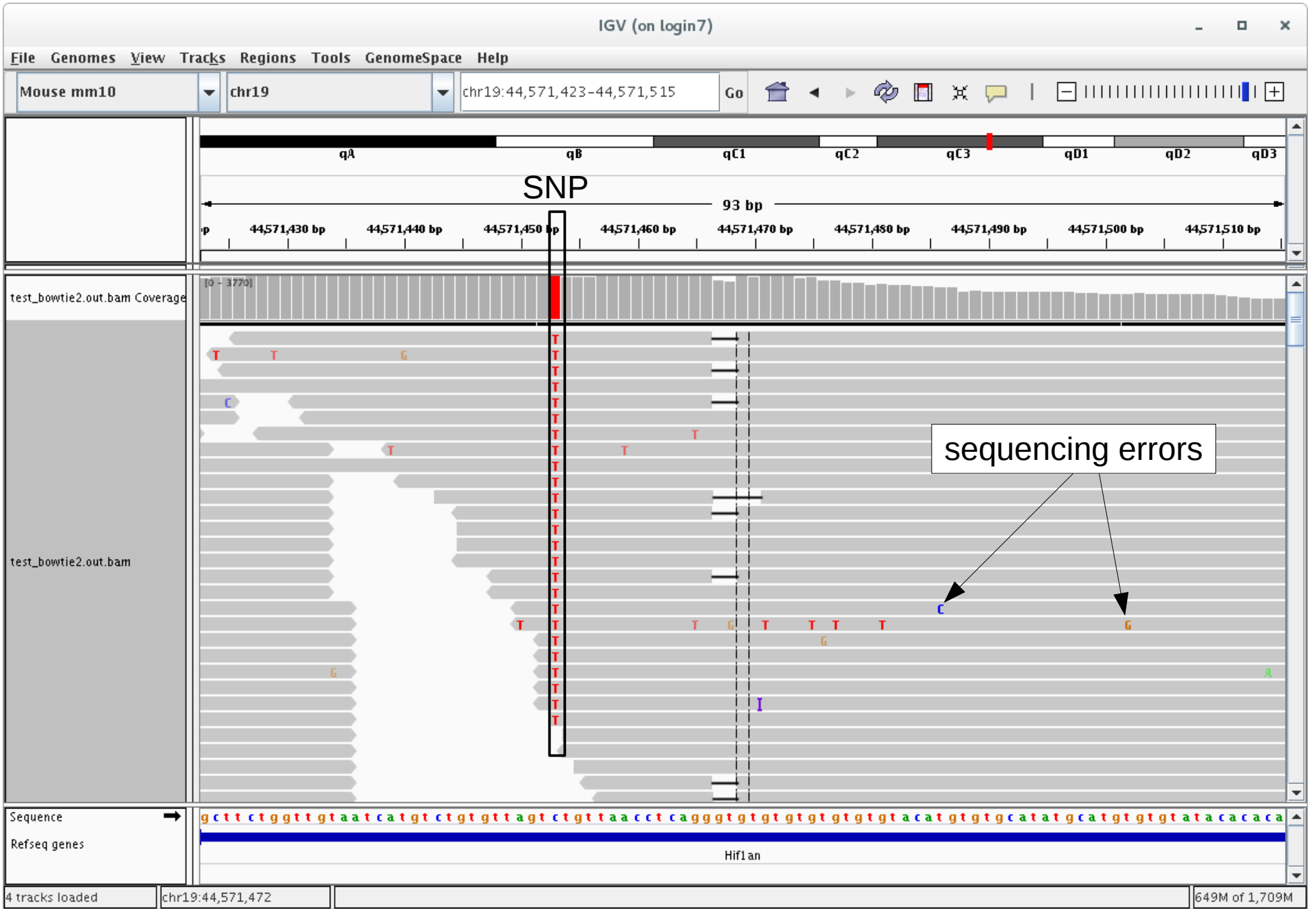
Change the Reference Genome



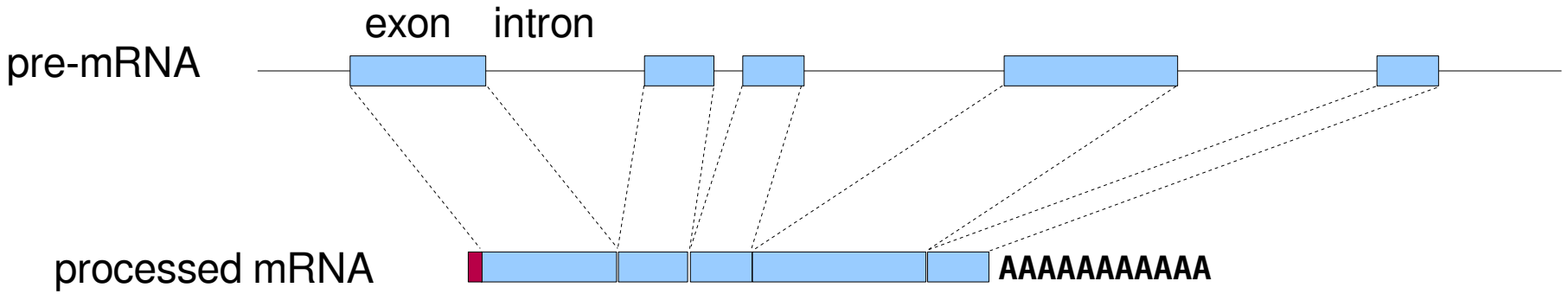
Change the Reference Genome



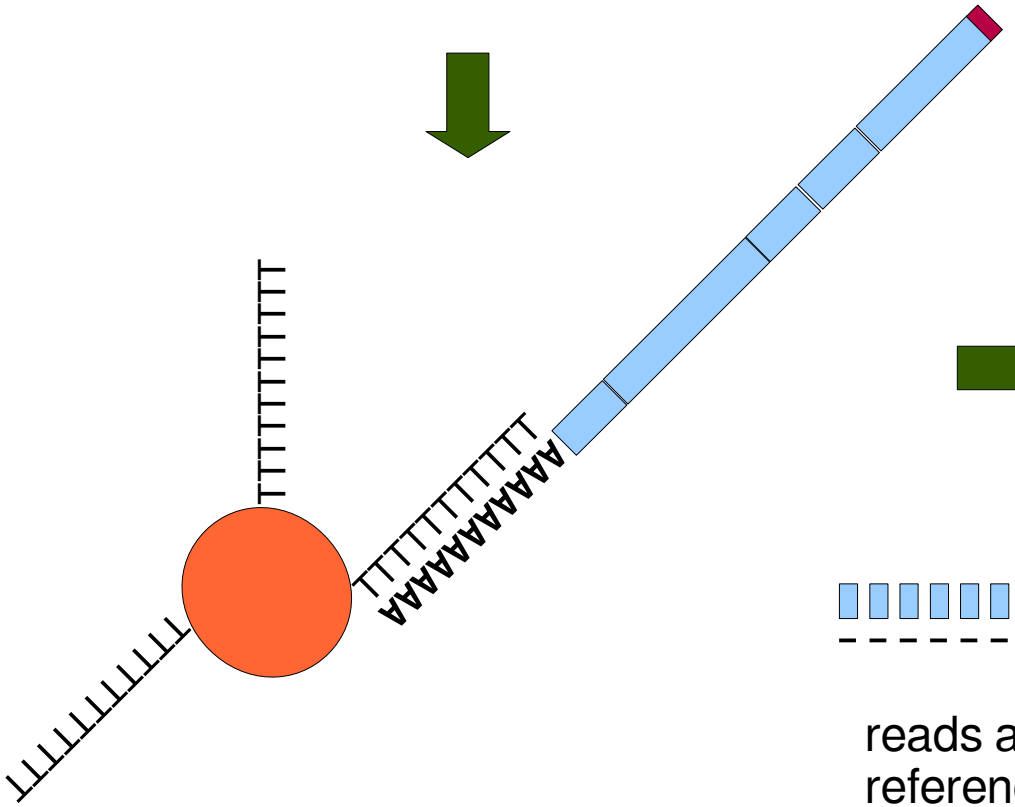
IGV viewing indexed bam file



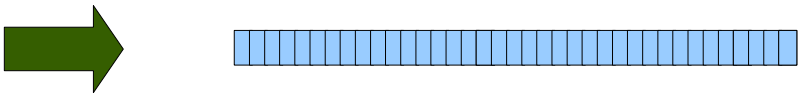
RNA-seq Analysis

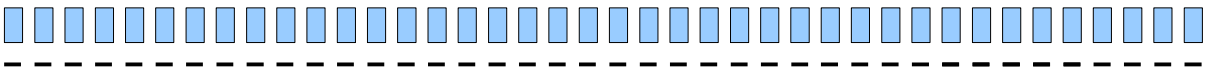


mRNA strands are captured by their Poly(A) tail using Poly(T) coated magnetic beads



exon region is fragmented and sequenced



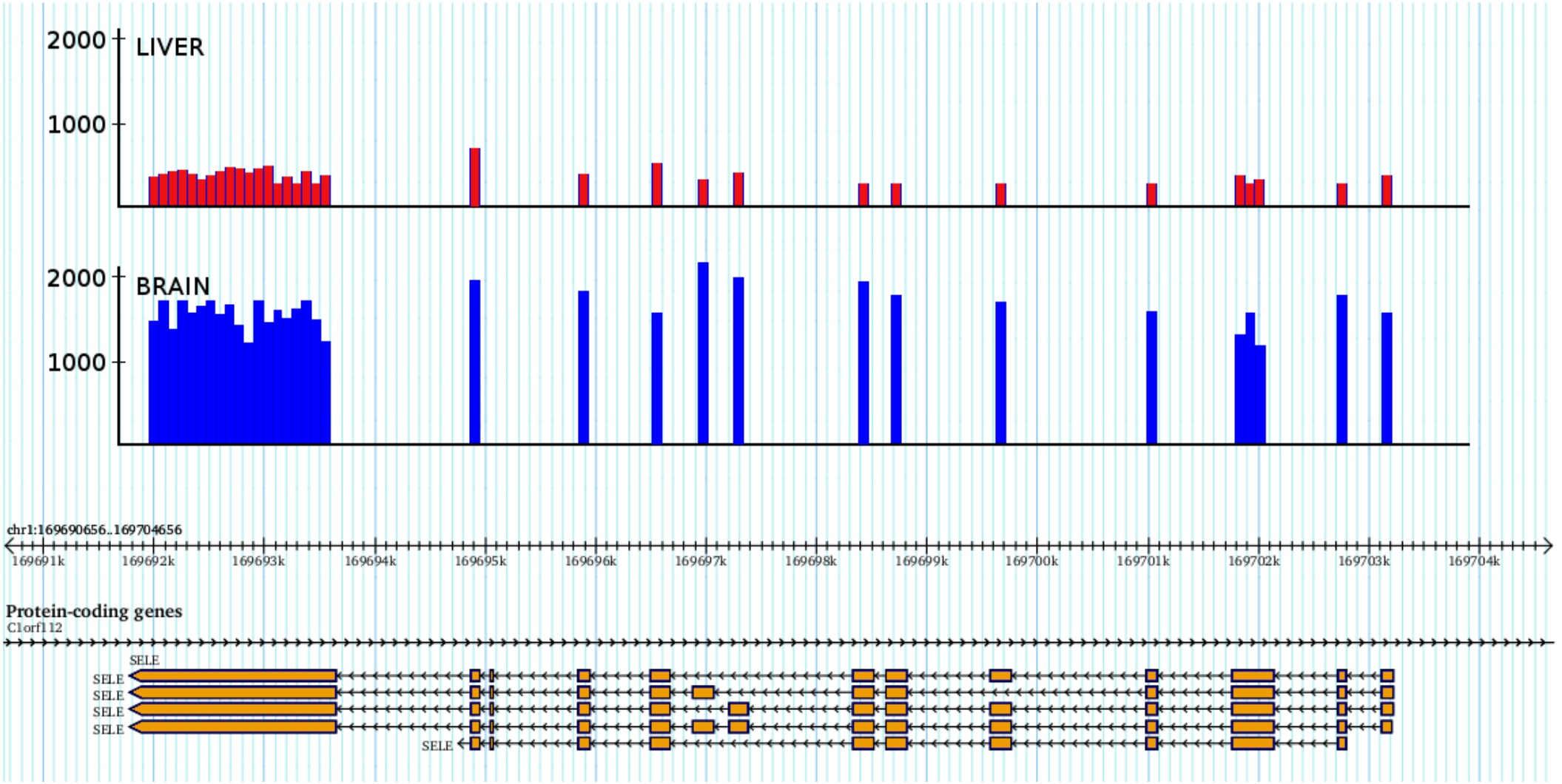


reads are then mapped to a reference genome or reference sequences

RNA-seq Applications

- Transcriptome assembly (find isoforms and rare transcripts)
  - *de novo* (Trinity, SOAPdenovo-Trans)
  - reference based (StringTie)
- Differential Expression (DE)
  - Bowtie, TopHat, Cufflinks, Cuffmerge, Cuffdiff
  - DESeq, DESeq2
  - edgeR
- Variant Calling
  - STAR, Picard, GATK (Haplotype Caller in RNA-seq mode)
- Genome Annotation
  - Align to assembly for validation of gene models
- *de novo* genome assembly scaffolding
  - L_RNA_scaffolder

RNA-seq Differential Expression (DE)



http://www.illumina.com/technology/mrna_seq.ilmn

RPKM vs FPKM vs TPM

- The number of **R**eads **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for single end reads
- The number of **F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for paired-end reads
    - If both paired reads align to a transcript then they are counted as one alignment
- **T**ranscripts **P**er kilobase **M**illion
  - Normalize for gene length first
  - Normalize for sequence depth second

http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

RNA-seq Transcript Abundance Estimation

- Alignment based
  - Express (outputs FPKM)
  - RSEM (isoform/gene level estimates without RPKM or FPKM)
- Non-Alignment based
  - Kallisto (pseudoalignment)
  - Salmon (lightweight alignment)
  - Sailfish (k-mer)
- Normalized abundance counts are used as input for DE analysis software
  - DESeq, DESeq2
  - edgeR

Tuxedo Suite

- TopHat – splice aware alignment of RNA-seq reads using Bowtie2
  - TopHat is superseded by HISAT2
- Cufflinks – assembles aligned reads into transcripts and estimates their abundances
- Cuffdiff – compares RNA-seq abundance (expression) levels of two samples or groups

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
CAWT_00001	CAWG_00001	-	chr_1.1:8373-9093	q1	q2	OK	111.944	163.869	0.549763	0.768107	0.58795	0.996768	no
CAWT_00002	CAWG_00002	-	chr_1.1:11447-12425	q1	q2	OK	14.5992	30.9037	1.08189	1.3841	0.2921	0.98312	no
CAWT_00003	CAWG_00003	-	chr_1.1:14130-14451	q1	q2	OK	248.323	259.152	0.0615814	0.172186	0.94685	0.996768	no
CAWT_00004	CAWG_00004	-	chr_1.1:14890-16045	q1	q2	OK	60.9546	86.0009	0.496617	0.604904	0.6204	0.996768	no
CAWT_00005	CAWG_00005	-	chr1.1:17512-18175	q1	q2	OK	366.694	646.846	0.818848	1.0705	0.4738	0.996768	no
…													
…													
CAWT_01628	CAWG_01628	-	chr1.2:664522-665344	q1	q2	OK	3.56447	157.849	5.46871	6.64693	0.00015	0.0482417	yes

p_value = The uncorrected p-value of the test statistic.
q_value = The FDR-adjusted p-value of the test statistic

R Bioconductor

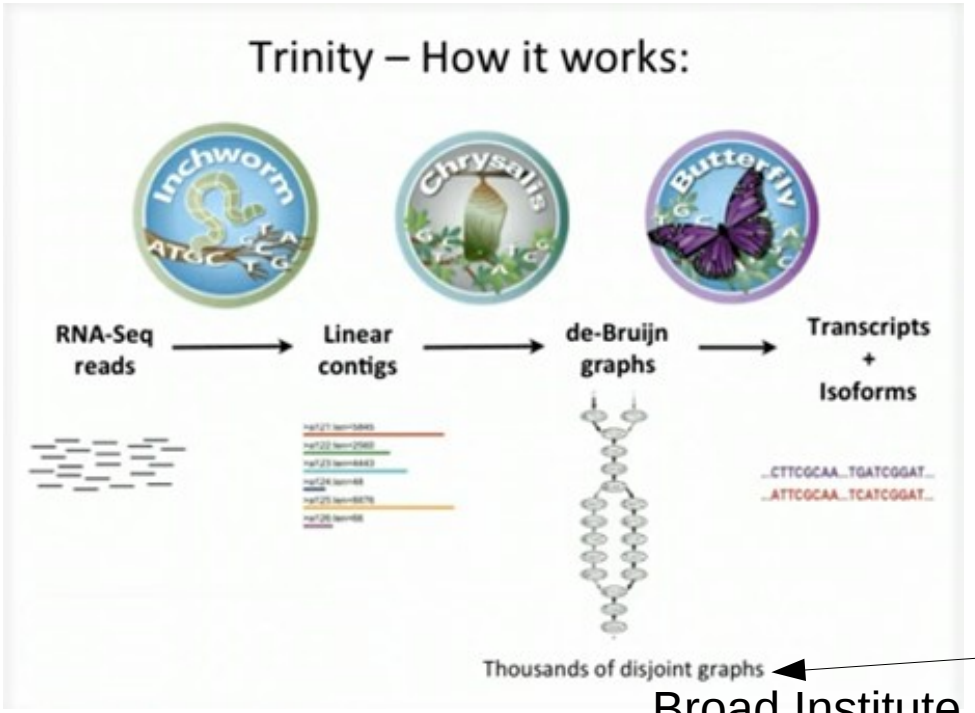
- Bioconductor packages can be found in this R version

```
module load R_tamu/3.2.5-intel-2015B
```

- Popular R bioconductor packages for RNA-seq
  - CQN – Normalization of RNA-seq data
  - edgeR – Differential gene expression
  - DESeq, DESeq2 – Differential gene expression
  - cummeRbund – analysis/visualization of cufflinks data

RNA-seq Transcriptome Assembly

- with a reference
  - `module load Scripture`
  - `module load Bowtie2 Tophat Cufflinks`
  - `module load Trinity/2.1.1-intel-2015B`
- without a reference
  - `module load Trinity/2.1.1-intel-2015B`



ideally one graph per gene

Broad Institute

RNA-Seq Blog

Transcriptome Sequencing Research & Industry News

Tag Archives: TMM

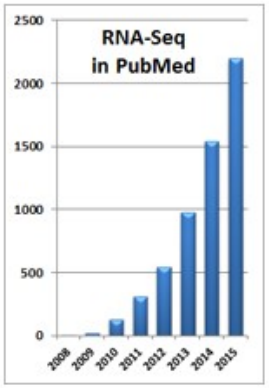
Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data

October 30, 2015 0 1,501 Views

Recently, rapid improvements in technology and decrease in sequencing costs have made RNA-Seq a widely used technique to quantify gene expression levels. Various normalization approaches have been proposed, owing to the importance of normalization in the analysis of RNA-Seq data. …

Read More »

An iteration normalization and test method for differential expression analysis of RNA-seq data



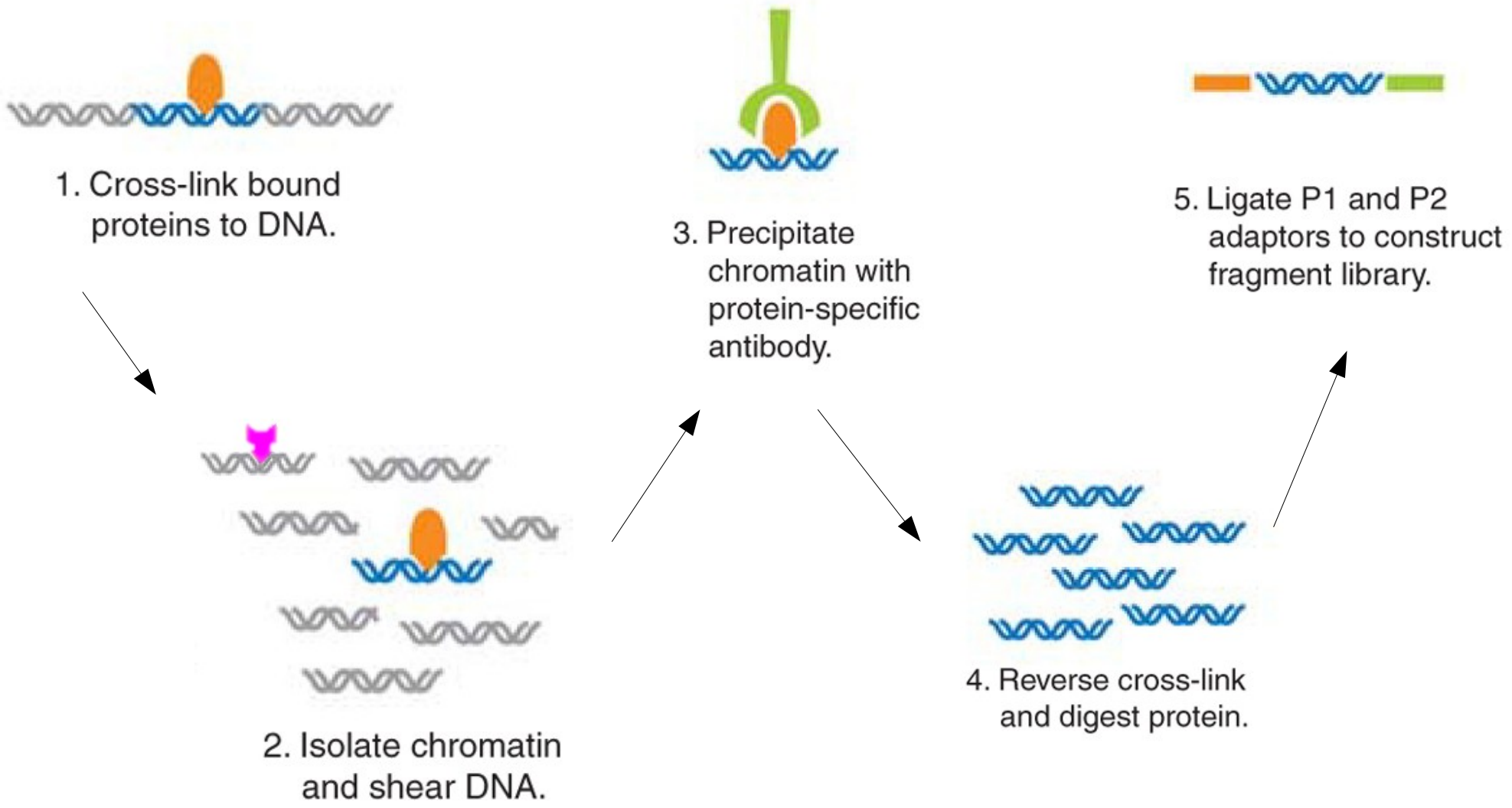
RED: A Java-MySQL Software for Identifying

email address

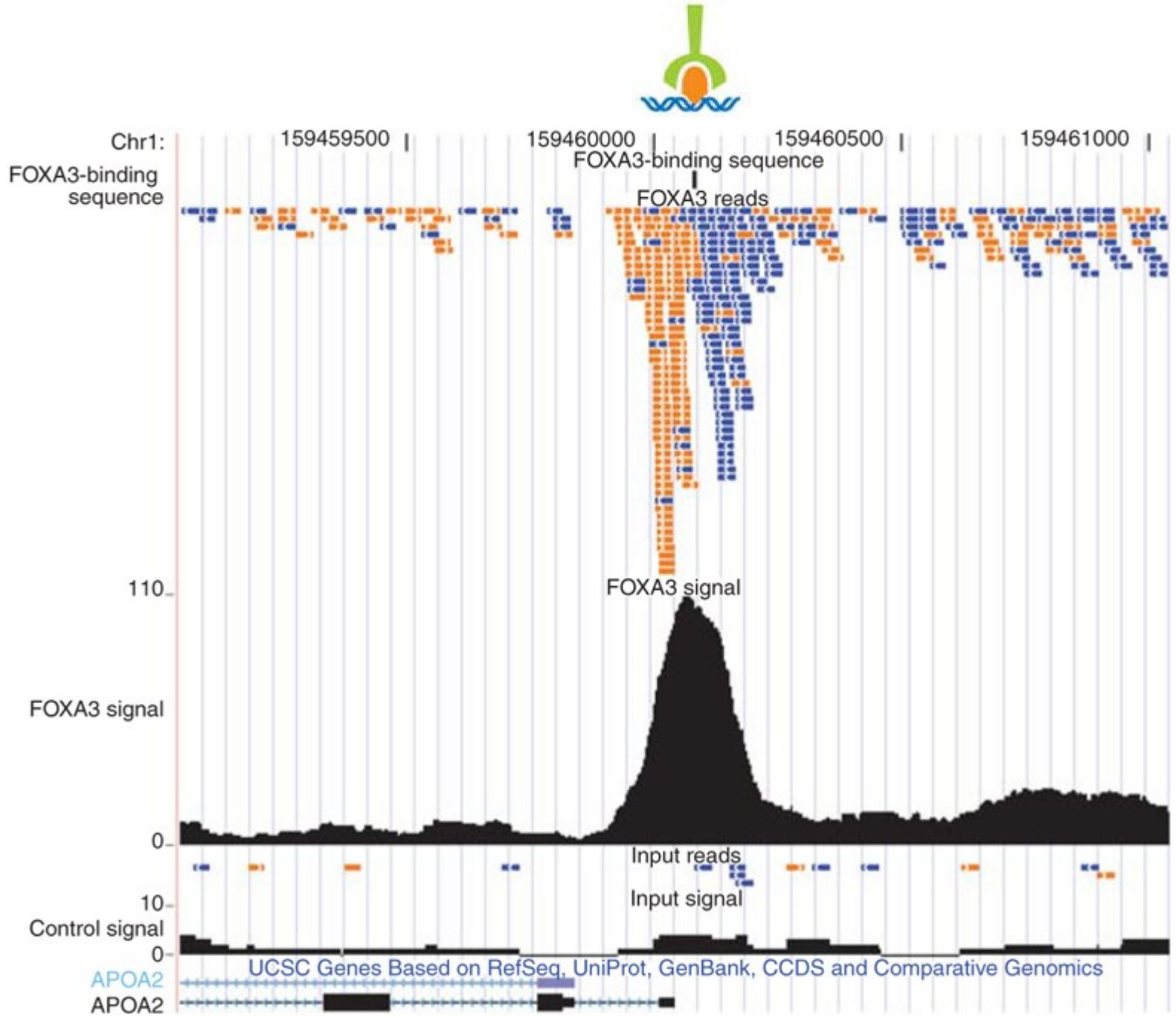
Subscribe

ChIP-seq Analysis

Chromatin immunoprecipitation (ChIP) is a technique for identifying and characterizing elements in protein-DNA interactions involved in gene regulation or chromatin organization.

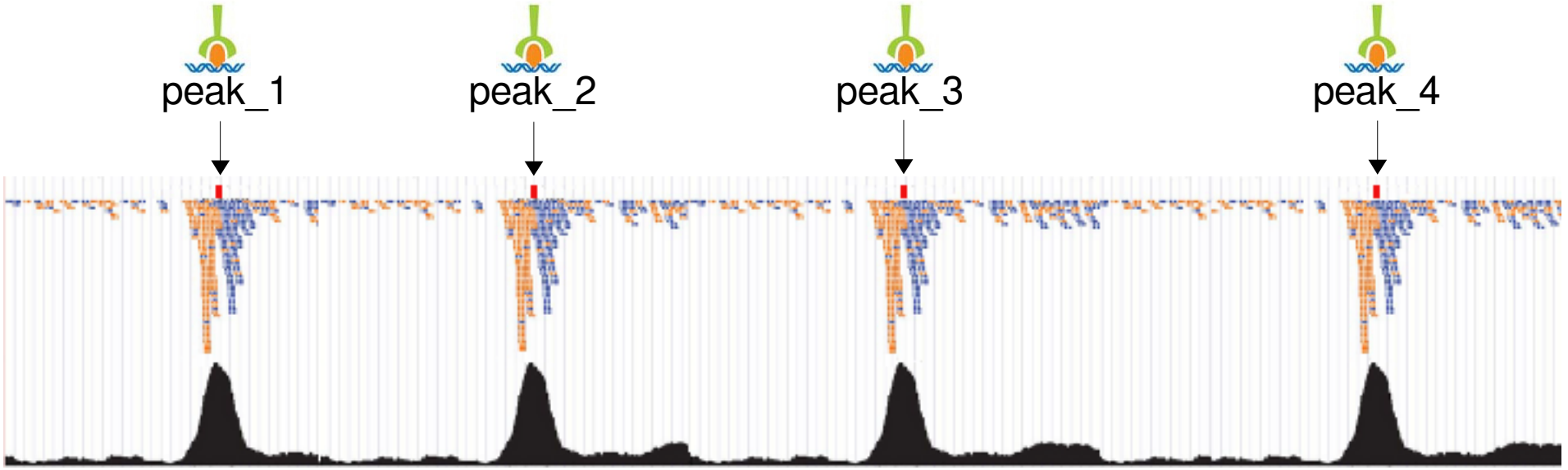


Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system
Nature Methods 6, (2009)

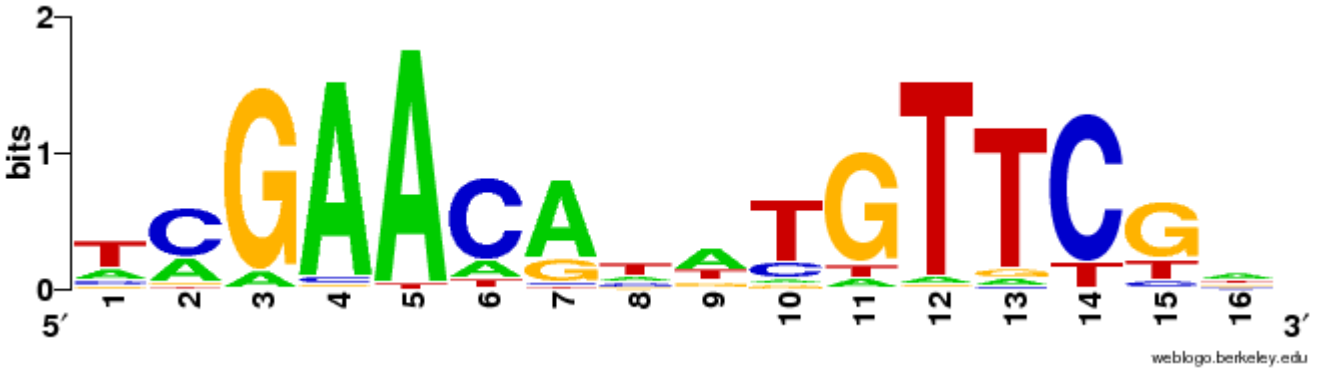


Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system
Nature Methods 6, (2009)

Select ~200 base region based on your interpolated peaks to select the sequence surrounding the protein binding site on DNA across the genome



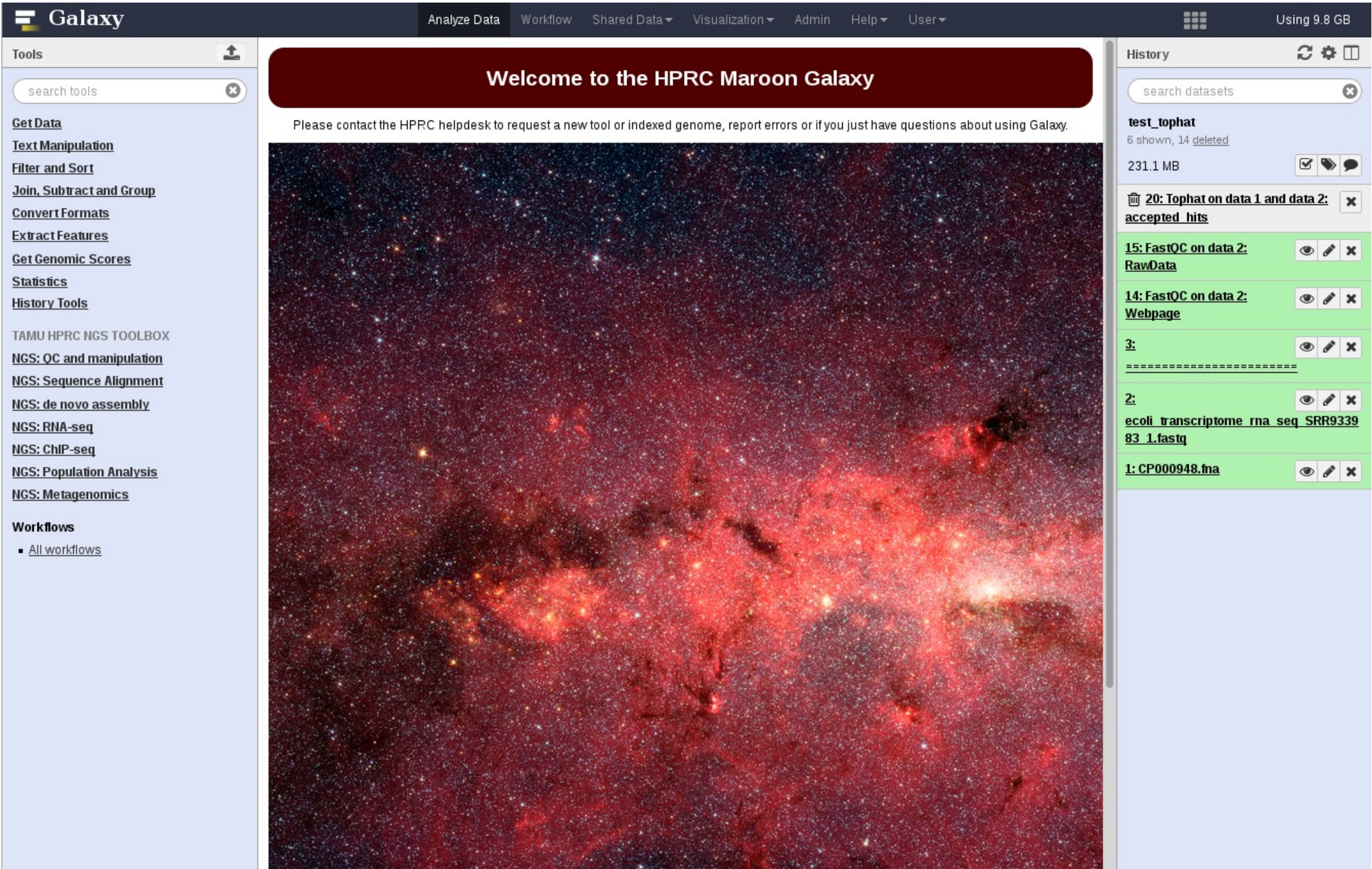
The goal is to find a DNA motif or motifs sequence among the sequences at each peak which describes the DNA sequence that a protein recognizes and binds



ChIP-seq

- Protein-DNA interactions
  - `module load MACS`
  - `module load MACS2`
- Subdivision of ChIP-seq regions into discrete signal peaks
  - `module load PeakSplitter`
- Peak caller
  - `module load PeakRanger`
  - `module load BroadPeak`
- Identify enriched domains from histone modification ChIP-seq data
  - `module load SICER`

Galaxy on Ada



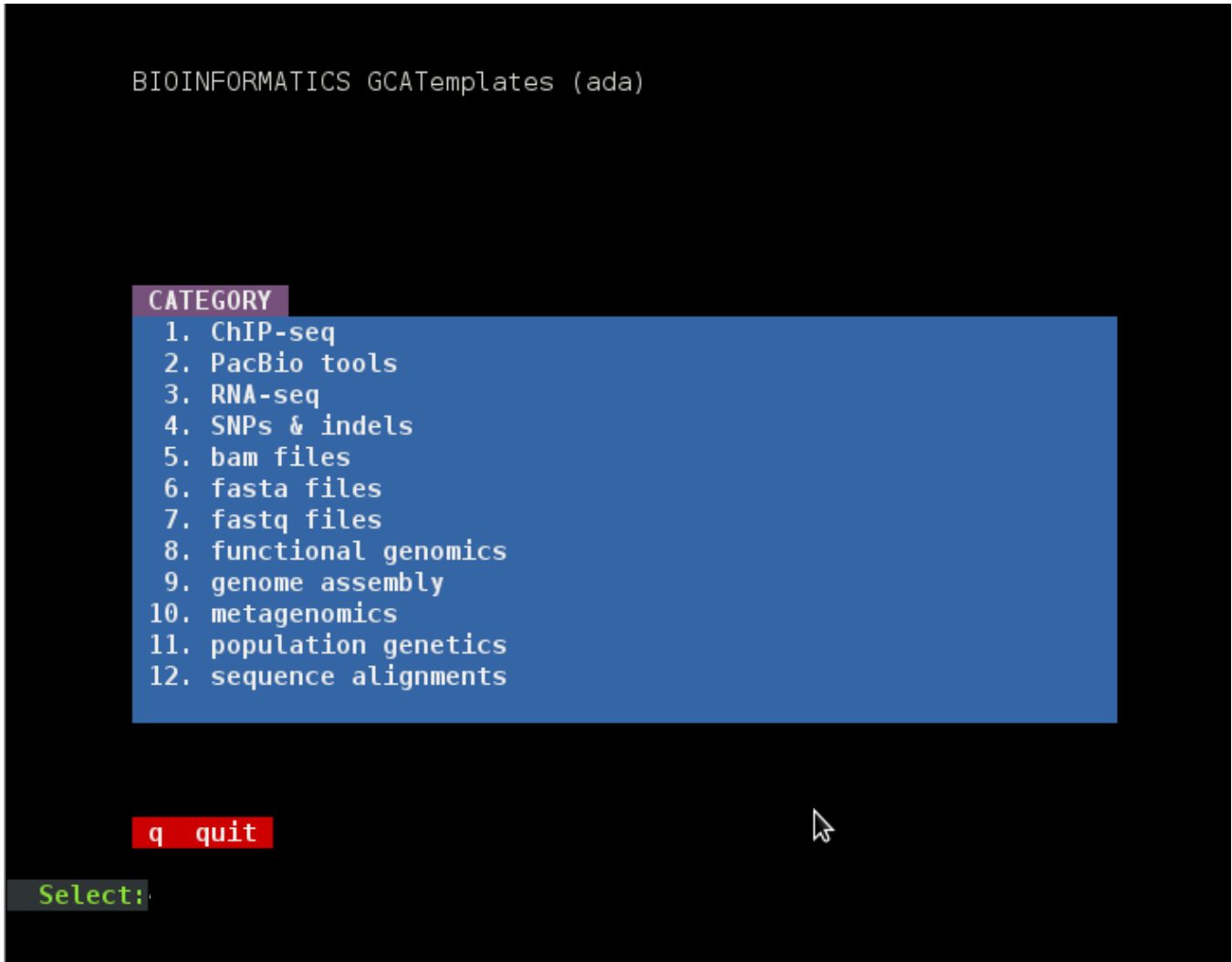
Sequence Trimming Exercise

Quality Trim Sequence Reads to be used in a *de novo* genome assembly

```
module load GCATemplates
```

```
gcatemplates
```

Select '#7 fastq files'
then find the template under 'fastq quality trim' that runs trimmomatic



Perform QC Trimming on Paired End reads

- Edit the GCATemplates template script
  - run_trimmomatic_0.35_pe_ada.sh
- Use the following paired end sequence files (same as FastQC example)

```
/scratch/helpdesk/ngs/reads/DR34_R1.fastq.gz
/scratch/helpdesk/ngs/reads/DR34_R2.fastq.gz
```

Use the full paths to the files instead of copying to your working directory