Mapping Next Generation Sequence Reads

Bingbing Yuan

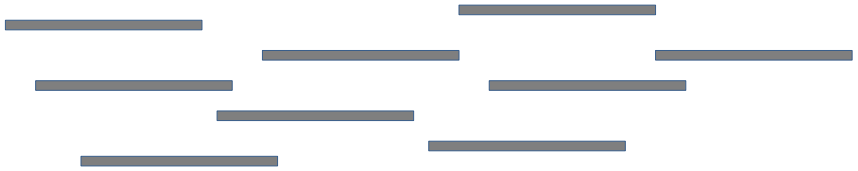
Dec. 2, 2010

What happen if reads are not mapped properly?

- Some data won't be used, thus fewer reads would be aligned.

- Reads are mapped to the wrong location, creating false positives and false negatives

Our pipeline outline



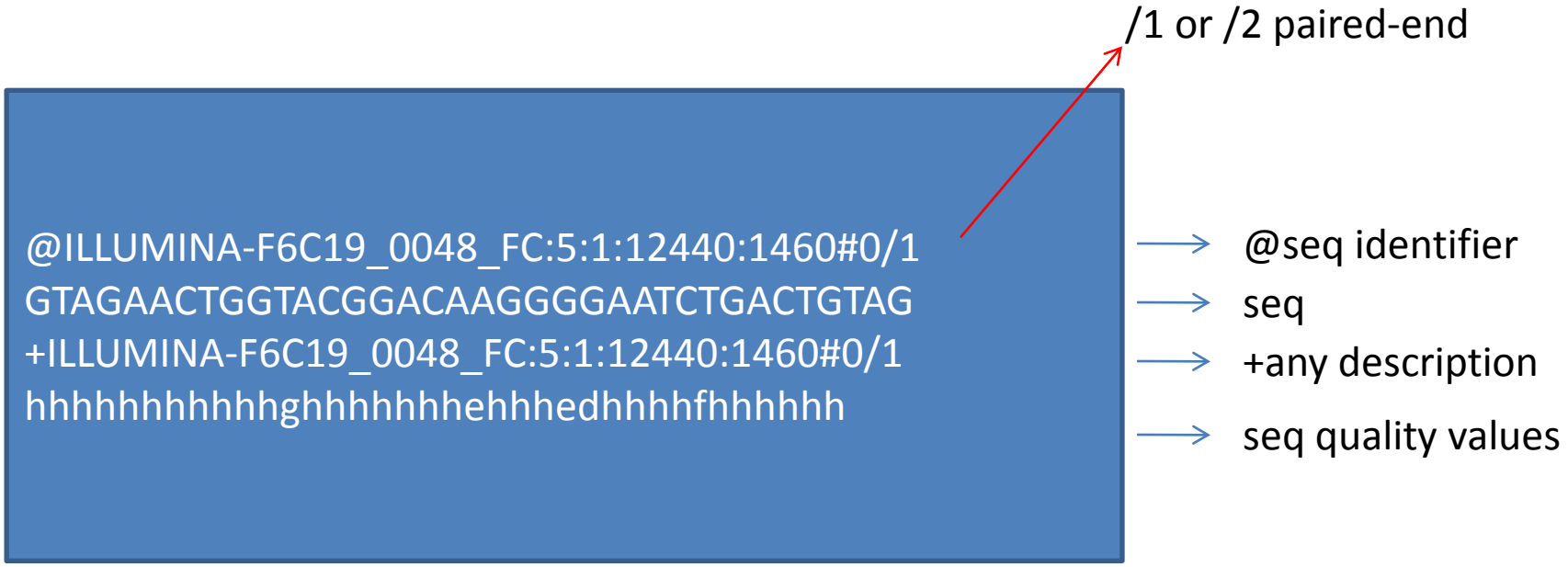
Part 1 — Quality Control

Part 2 — Mapping
(Chip-seq + RNA-seq)

Part 3 — Visualization

Illumina data format

- Fastq format: (QualityScore/s_1_sequence.txt)



Check read quality

- Overall read distribution, read quality
- Per-cycle base call, quality scores
- May need to
  - remove reads with lower quality
  - Trim the read seq
  - Remove adapter/linker seq

Freely Available Tools for QC

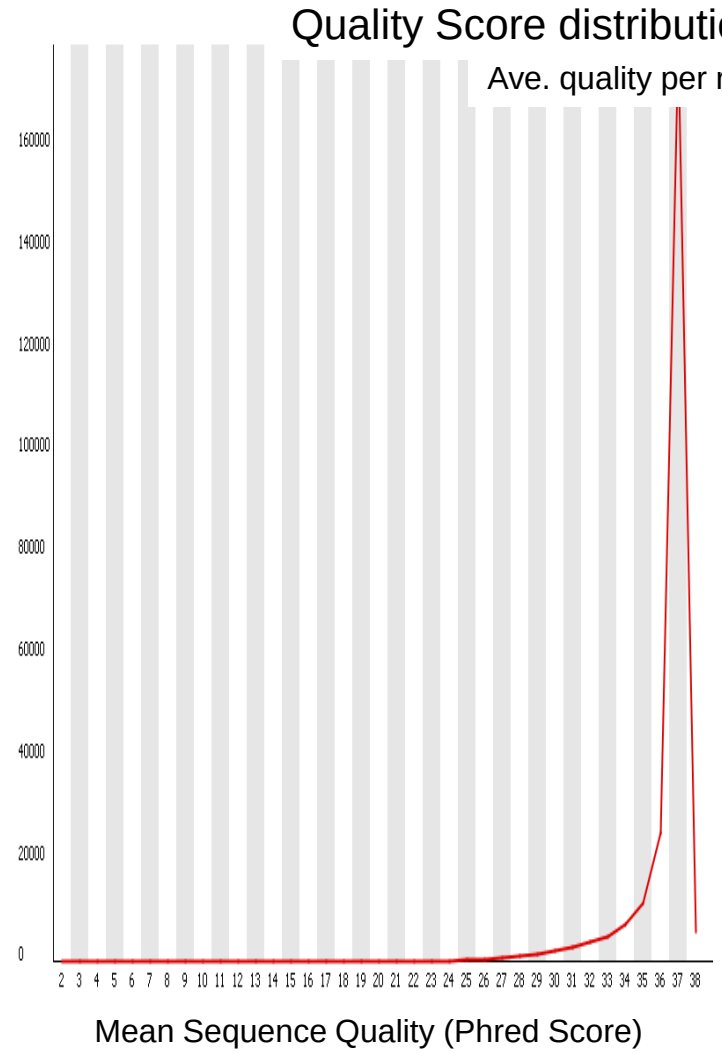
- Galaxy:
  - http://main.g2.bx.psu.edu/
  - Many functions
  - Long time for uploading files since it is on remote server
- Fastx toolkit:
  - http://hannonlab.cshl.edu/fastx_toolkit/
  - galaxy integration, Linux(Tak), MacOSX
- FastQC (picard):
  - http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc
  - Linux(Tak), Window, MacOSX
- Shortread:
  - http://www.bioconductor.org/packages/release/bioc/html/ShortRead.html
  - R package, Linux (Tak), Window, Mac

General information about reads

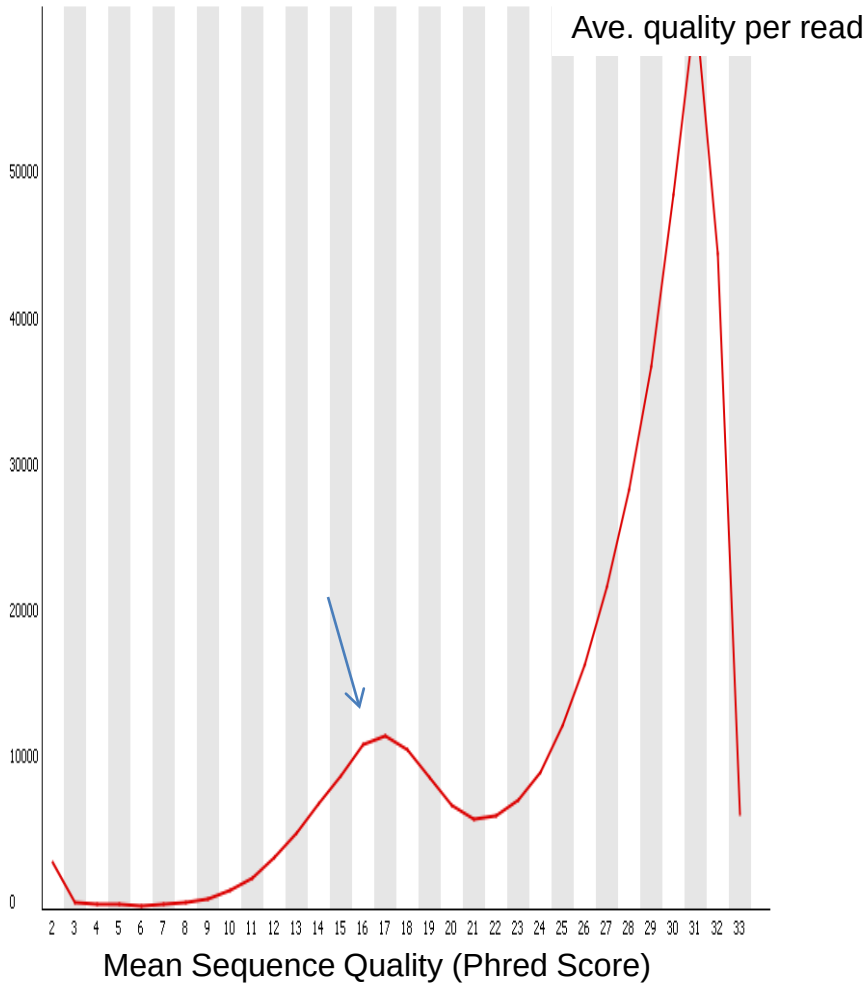
Measure	Value
Filename	SRR015149.fastq
File type	Conventional base calls
Total Sequences	8923918
Sequence length	26
%GC	43

Created with FastQC

Overall read quality



FastQC from Babraham Bioinformtics



Galaxy: Filter FASTQ, filter by quality

FASTX toolkit: *fastq_quality_filter*

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.90%
40	1 in 10000	99.99%

Most abundant Reads

	sequence	count	lane
1	GATCGGAAGAGCTCGTATGCCGTCTT	231002	character
2	GNNNNNNNNNNNNNNNNNNNNNNNN	8626	character
3	ANNNNNNNNNNNNNNNNNNNNNNNN	7405	character
4	AAAAAAAAAAAAAAAAAAAAAAAAA	5539	character
5	TNNNNNNNNNNNNNNNNNNNNNNNN	4502	character
6	CNNNNNNNNNNNNNNNNNNNNNNNN	4334	character
7	GATCGGAAGGAGCTCGTATGCCGTCT	3809	character
8	AAAATCATGGAAAATGATTTTAGATC	3171	character
9	GATCGGAAGAGCTCGGTATGCCGTCT	2988	character
10	GATCGGAAGAGCTCGTATGCAGTCTT	2143	character
11	GATNNNNNNNNNNNNNNNNNNNNN	2054	character
12	GTTTTCTCGCCATATTCCAGGTCCTT	2012	character
13	GATCGGAAGAGGCTCGTATGCCGTCT	1999	character
14	GAATATGGCAAGAAAACTGAAAATCA	1956	character
15	GATCGGAAGAGCTCGTATGCCGTATT	1905	character
16	GTTTTCCTCGCCATATTTCACGTCCT	1774	character
17	GATCGGAAGAGCTCGTATGCCGCCTT	1698	character
18	AAANNNNNNNNNNNNNNNNNNNNN	1659	character
19	GATCGGAAGAGCTCGTATGACGTCTT	1603	character
20	GAANNNNNNNNNNNNNNNNNNNNN	1502	character

provide clues to the source of over-represented sequences. Some of these reads are filtered by the alignment algorithms; other duplicate reads might point to sample preparation issues.

Created with shortread

Most abundant Reads

Sequence	Count	Percentage	Possible Source
CTGTAGGCACCATCAATTCGTATGCCGTCTTCTGCT	1175220	5.90	No Hit
AAGAGGTGCACAATCGACCGATCCTGACTGTAGGCA	359160	1.80	No Hit
TACACGGAGTCGACCCGCAACGCGACTGTAGGCACC	77161	0.38	No Hit
TGTAGGCACCATCAATTCGTATGCCGTCTTCTGCTT	70591	0.34	Illumina Single End Apapter 2 (95% over 21bp)
ACGCGAAACTCAGGTGCTGCAATCTCTGTAGGCACC	67674	0.34	No Hit
TCGAAGAGTCGAGTTGTTTGGGAATGCCTGTAGGCA	66160	0.33	No Hit

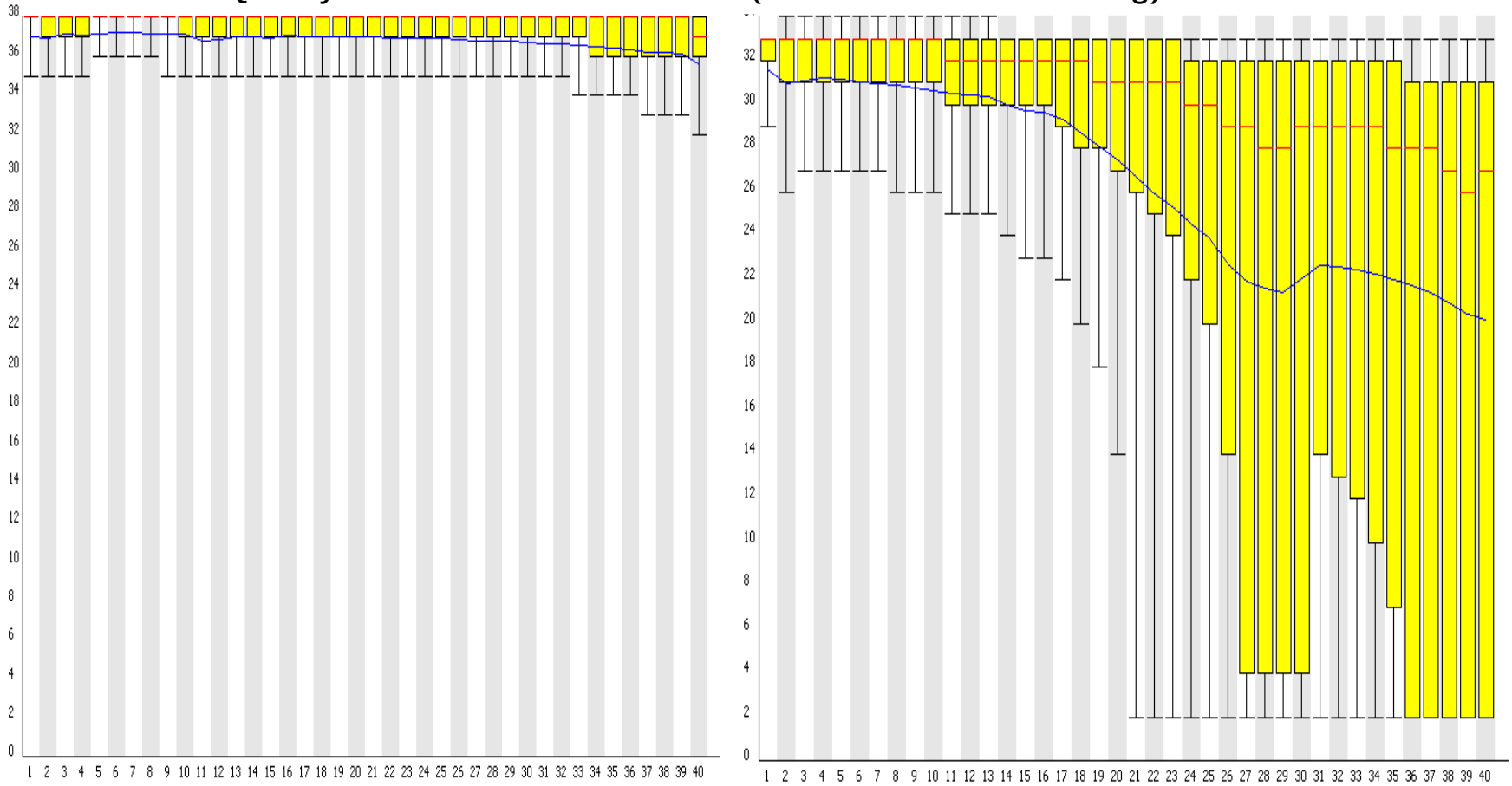
Created with FastQC

Per-cycle quality score

Good

Bad

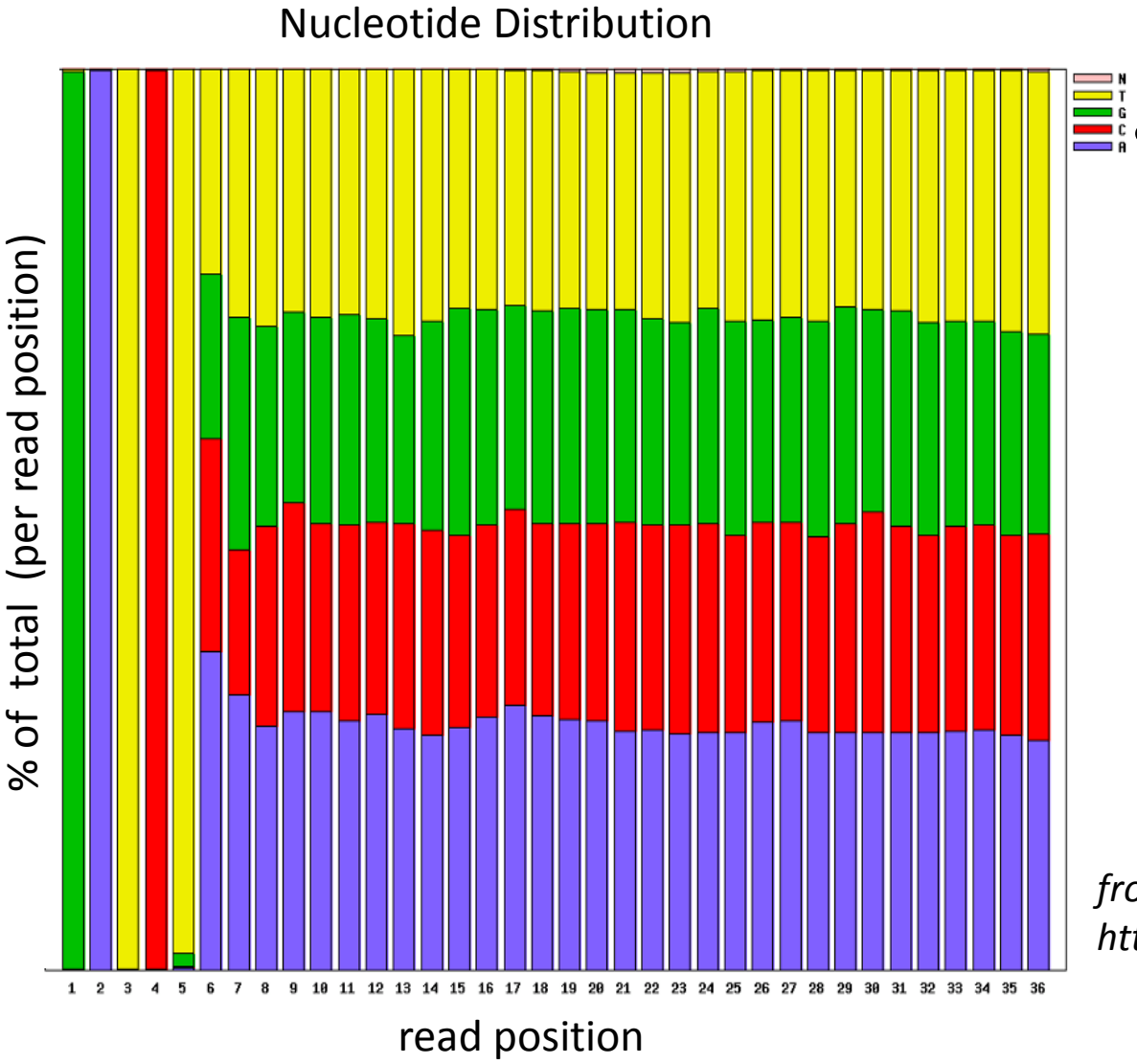
Quality Scores across all bases (Illumina >v1.3 encoding)



Position in read (bp)

From FastQC in Babraham Bioinformtics

Per-cycle base call



- Trim reads:
  - FASTX toolkit:
    - *fastx_trimmer*
  - Galaxy:
    - FASTQ Trimmer
    - Trim sequences

from Fastx in
http://hannonlab.cshl.edu/fastx_toolkit

Remove adapter/linker

What it does

This tool clips adapters from the 3'-end of the sequences in a FASTA/FASTQ file.

Clipping Illustration:

Clipping Example:

```
>1
ATGTAATGTTTATATATATCGTAAATCCAACACAAT
>2
TATTTTGGAATTCCACGACCCTGTAGGCACCATCAA
>3
ACGTTGTTCGGTGCGTCCTGAACTGTAGGCACCATC
>4
TTTCTTCTTATCTCTTCGAGTCTGTAGGCACCATCA
>5
TGGAACTTGCTGTAGGCACCATCATTATTTATATAA
>6
TTTACCGGAAGCATAACTCTTCTGTAGGCACCATCA
>7
TGTATTAGCGGTGGGGCCCGACTGTAGGCACCATCA
```

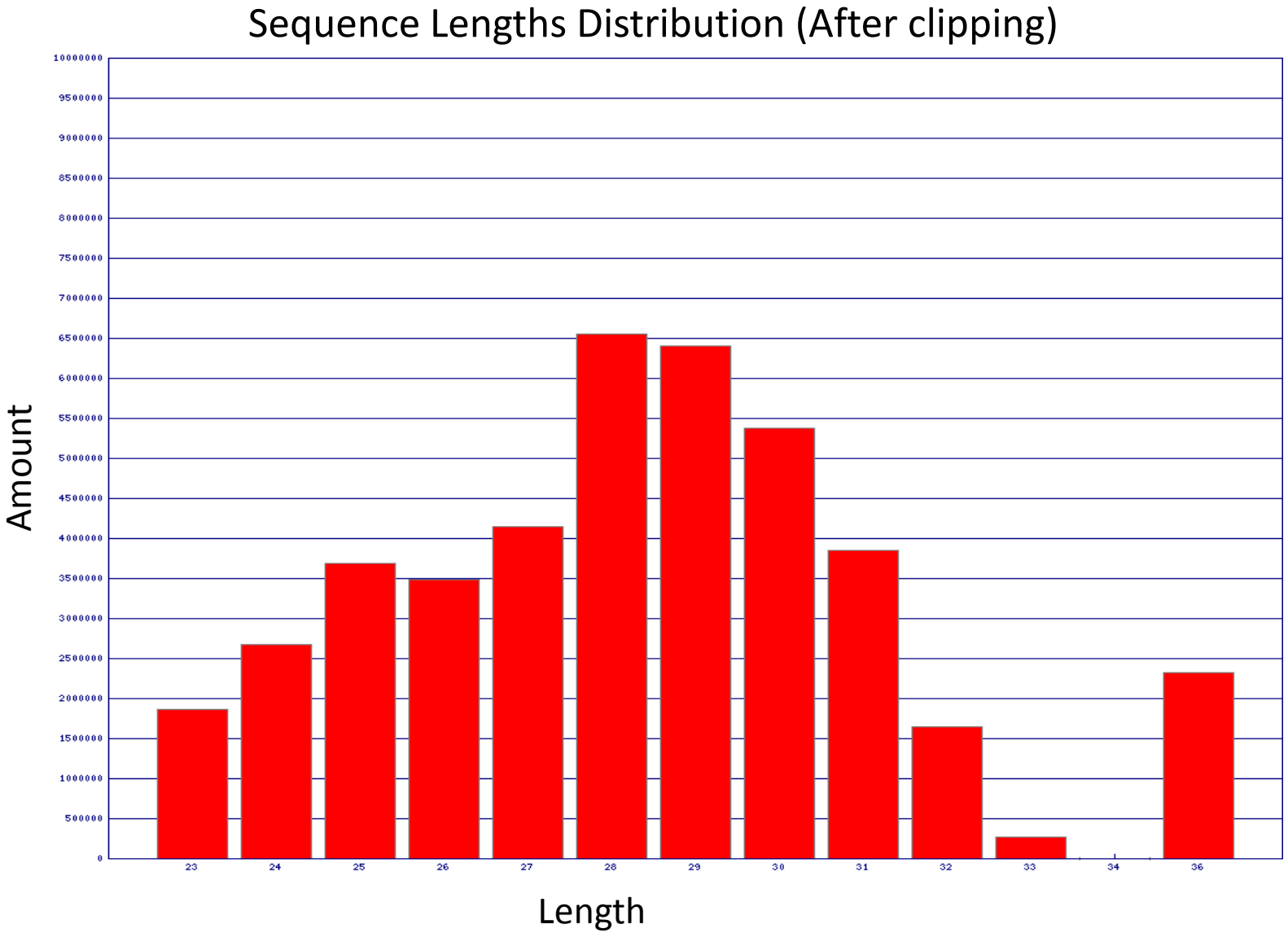
```
>2
TATTTTGGAATTCCACGACC
>3
ACGTTGTTCGGTGCGTCCTGAA
>4
TTTCTTCTTATCTCTTCGAGT

>6
TTTACCGGAAGCATAACTCTT
>7
TGTATTAGCGGTGGGGCCCGA
```

In the above example:

- Sequence no. 1 was discarded since it wasn't clipped (i.e. didn't contain the adapter sequence). (**Output** parameter).
- Sequence no. 5 was discarded --- it's length (after clipping) was shorter than 15 nt (**Minimum Sequence Length** parameter).

From fastx: ***fastq_clipper*** *in http://hannonlab.cshl.edu/fastx_toolkit*

Sequence length distribution after clipping



Created with FASTX toolkit: *fastx_quality_stats, fastx_nucleotide_distribution_graph.sh*

Challenges of mapping short reads

- Large genome
- Billions of reads
- Speed
- Repeat regions
- Sequencing errors, reference genome variations

Mapping Techniques:

- Index genome:
  - Burrows-Wheeler Transform: order the genome
  - FM index: index genome

Free Mapping software for Chip-seq

- Bowtie (tak):
  - Langmead *et al*. (2009) Genome Biology, 10:R25
  - http://bowtie-bio.sourceforge.net/index.shtml
  - One of the fastest alignment software for short reads
  - Not gapped-alignment
  - Base quality can be used for evaluating alignments
  - Mismatch: 0-3
  - Flexible reporting mode including SAM format
- BWA (Burrows-Wheeler Alignment Tool):
  - http://bio-bwa.sourceforge.net/bwa.shtml
  - Short reads up to 200bp
  - Gapped alignment
  - Base quality not used for evaluating alignments
  - Allow >3 mismatches
  - Need to run samse/sampe to get SAM format

Reporting the alignments by bowtie

- Unique alignment only
- Reporting ambiguous hits
  - Randomly report one
  - Report all alignments above cutoff parameters
  - Report all alignments with best alignment scores

Bowtie options

- Index genome once: /nfs/genome/
  - mm9.1.ebwt mm9.2.ebwt mm9.3.ebwt mm9.4.ebwt mm9.rev.1.ebwt mm9.rev.2.ebwt
- Alignment:
  - Seed: The first L bases are called the “seed” (-l)
  - Max mismatches in seed (-n)
- Reporting:
  - Report number of alignment per read (-k)
  - Suppress all alignments (-m)
  - Report best hits (--best, --strata)
- Output:
  - unalign reads (--un)
  - reads over –m cutoff (--max)
  - Sam format (-S)

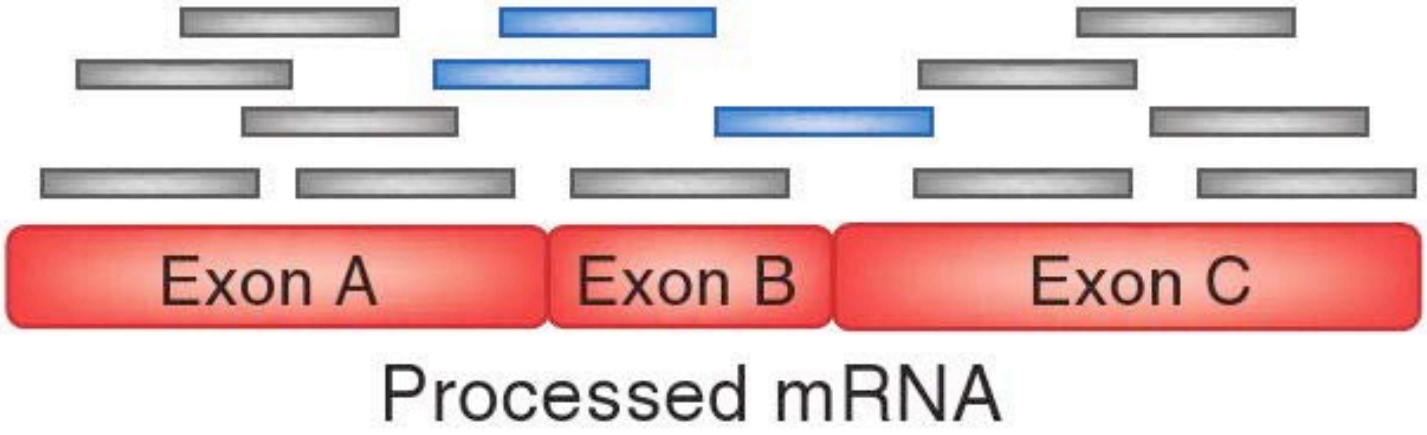
Bowtie examples

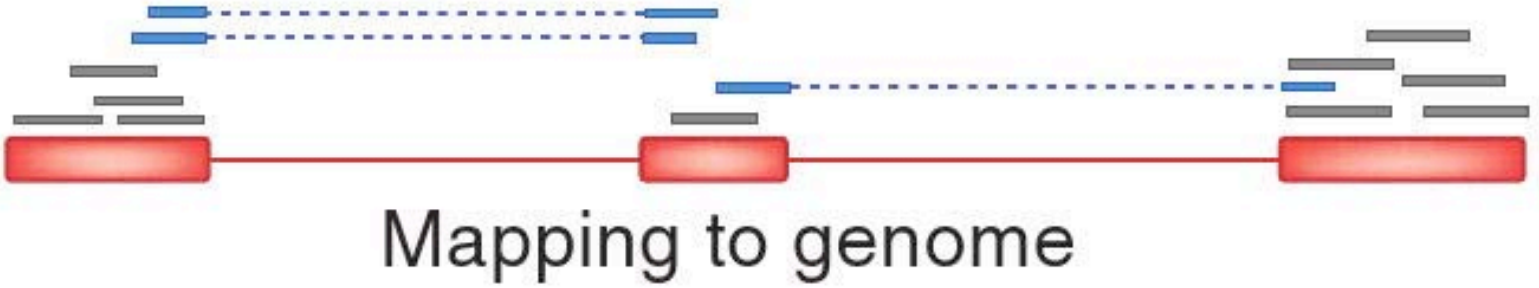
- bowtie solexa1.3-quals -n 1 -l 36 /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9 s3_sequence.txt

- bowtie solexa1.3-quals -n 1 -l 36 –m 10 –k 10 –max mapOver10.fq /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9 s3_sequence.txt

- bowtie solexa1.3-quals -n 1 -l 36 –m 10 –k 2 ---best --strata /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9 s3_sequence.txt

RNA-Seq





Trapnell C, Salzberg SL. *Nat Biotechnol.* 455-7(2009).

Mapping RNA-seq

- Map to genome:
  - Computationally expensive
  - Potential novel transcripts
  - reads across exon-exon junction will not be aligned
- Map to transcripts:
  - Computationally inexpensive
  - limited by the annotation files
- De novo assembly
  - No reference genome
  - With reference genome: allows detection of chimeric transcripts
    - Assemble into contigs and align(BLAT) against genome

Mapping RNA-seq with Tophat

- http://tophat.cbcb.umd.edu/
- Trapnell et al, 2009. PMID: 19289445
- built on the ultrafast short read mapping program Bowtie
- find splice junctions without a reference annotation.

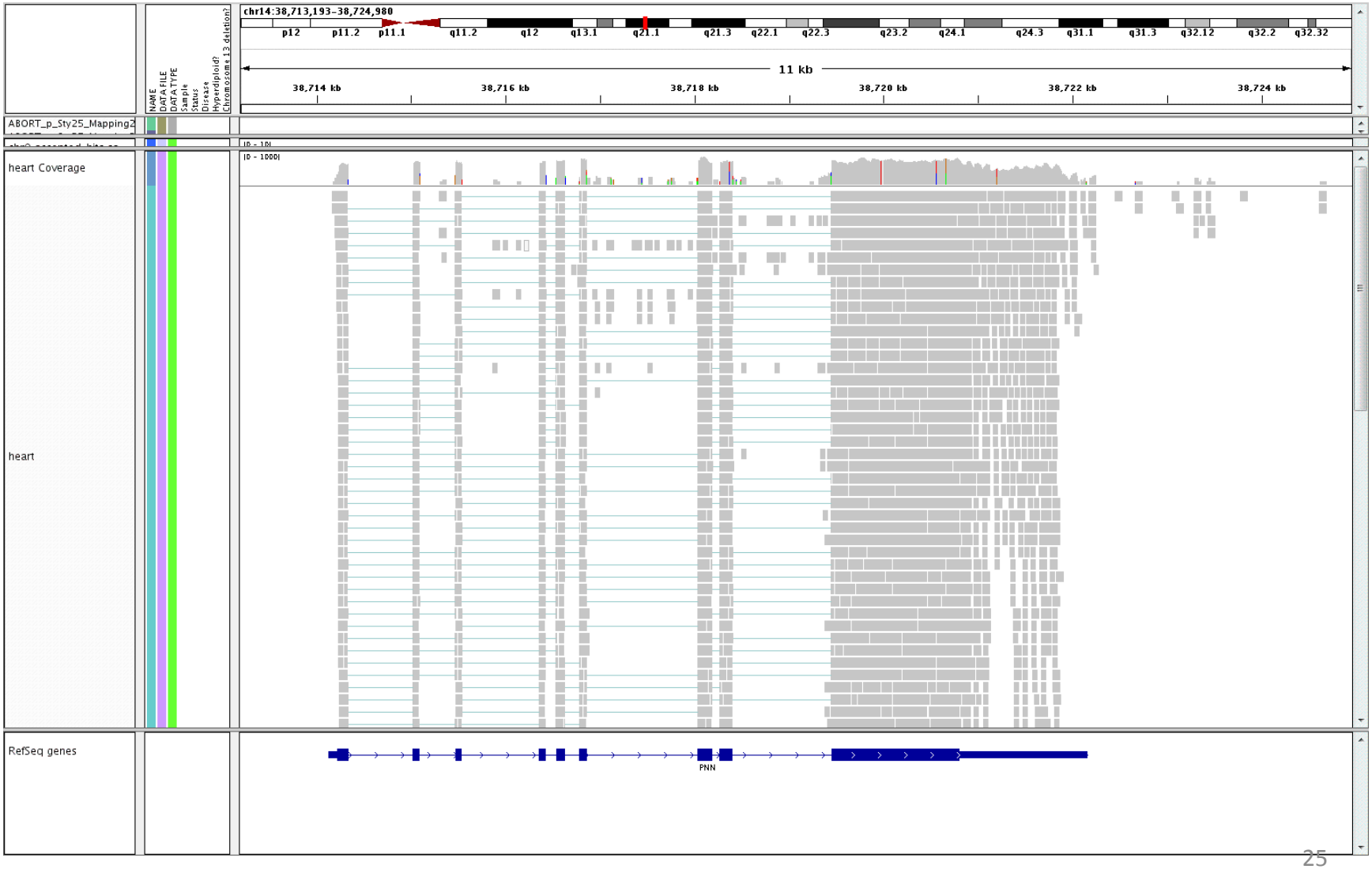
1. first mapping RNA-Seq reads to the genome
2. builds a database of possible splice junctions
    a. distinct regions of piled up reads in the initial mapping
    b. evidence for a splice junction: such as alignments across "GT-AG" introns
    c. paired end reads

- Linux (Tak)

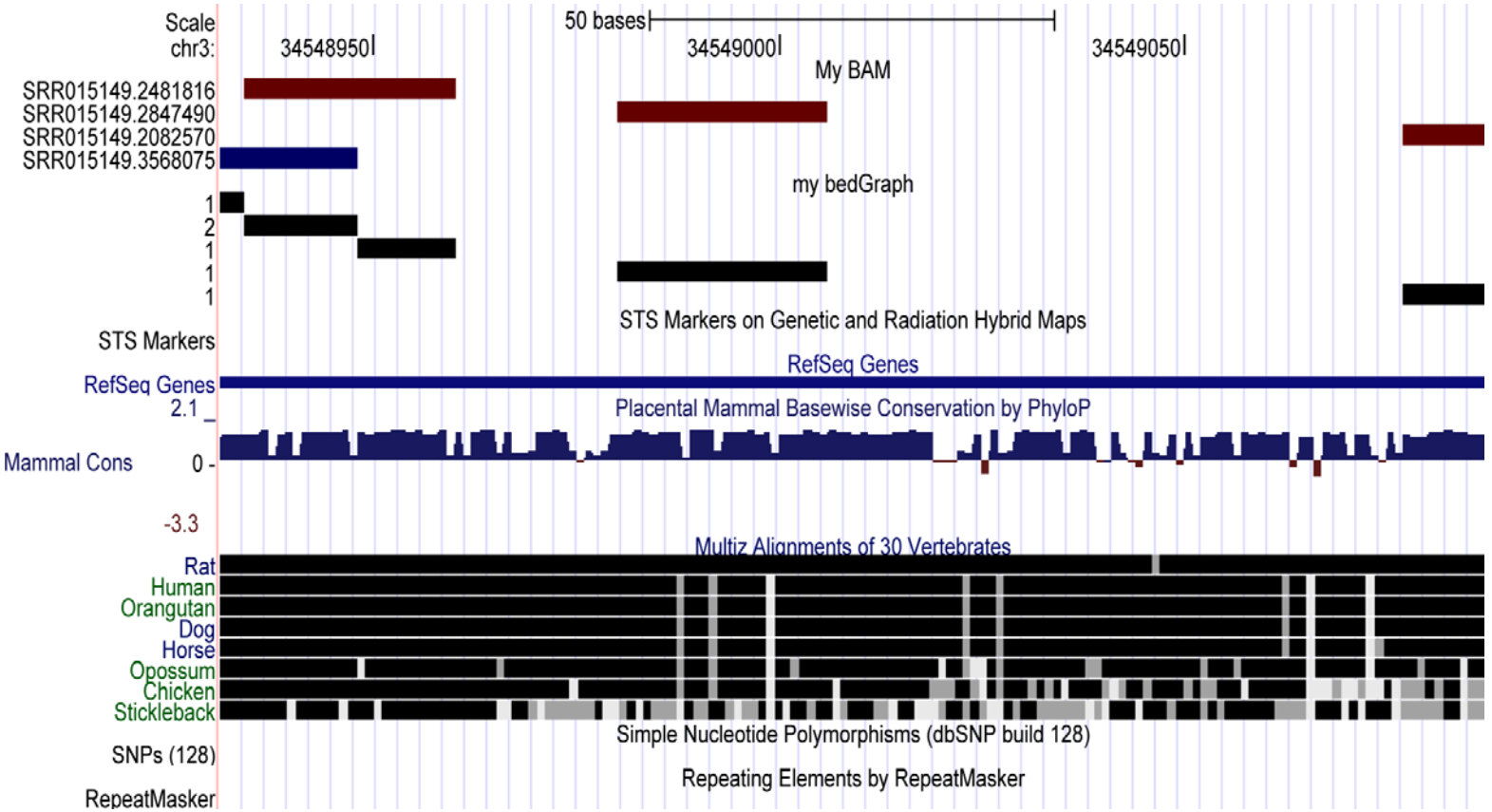
SAM/BAM

- Examples:
  - read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
    AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG
    <<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \ NM:i:1 RG:Z:L1
  - read_28701_28881_323b 147 chr20 28834 30 35M = 28701 -168 \
    ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA
    <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \ MF:i:18 RG:Z:L2
  - SRR015149.61819 16 chr3 29065583 255 26M * 0 0
    ACNAATTGCNATGCAGACACTTCACC ""!.6,"=1!I+IIIIIIBII)*CII XA:i:2
    MD:Z:2G6G16 NM:i:2
- Files for browsers:
  - Convert SAM to BAM, then sort, index bam files with samtools (tak) (http://samtools.sourceforge.net/)
  - Sort, index SAM with IGV tools (http://www.broadinstitute.org/igv/)
  - UCSC genomes browser:
    - BAM format needs to be on http/ftp server
    - convert BAM to bedGraph with genomeCoverageBed from bedtools (tak) (http://bioinformatics.oxfordjournals.org/content/26/6/841.full)

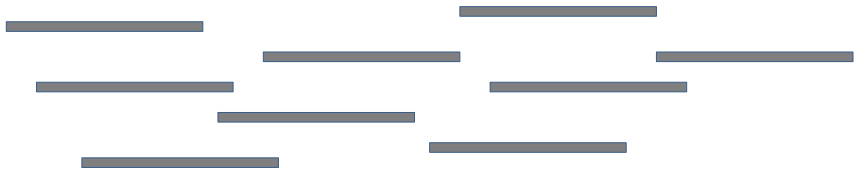
IGV


chr14:38,713,193-38,724,980
p12
p11.2
p11.1
q11.2
q12
q13.1
q21.1
q21.3
q22.1
q22.3
q23.2
q24.1
q24.3
q31.1
q31.3
q32.12
q32.2
q32.32
11 kb
38,714 kb
38,716 kb
38,718 kb
38,720 kb
38,722 kb
38,724 kb
NAME
DATA FILE
DATA TYPE
Sample
Status
Disease
Hyperdiploid?
Chromosome 13 deletion?
ABORT_p_Sty25_Mapping2
heart Coverage
heart
RefSeq genes
PNN

UCSC Genome Browser



Our pipeline



fastq

Part 1: QC

FastQC, FASTX-toolkit, shortread

Chip-Seq

RNA-Seq

Part 2: mapping

Bowtie

TopHat

Part 3: view

IGV

References

- FastQC to check the quality of high throughput sequence http://www.youtube.com/watch?v=bz93ReOv87Y
- RNA Seq (by Ryan Morin) http://www.bioinformatics.ca/files/CBW%20-%20presentations/HTSeq_2010_Module%203/HTSeq_2010_Module%203.mp4
- Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. *Nat Biotechnol.* 455-7(2009).
- Wang Z. *et al.* RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 57-63 (2009).
- Assessing Sequence Data Quality from Bioinfo-core: http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010