Introduction to RNA-Seq

Part I: Mapping Reads

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology

2 July 2013

Example RNA-Seq Experiment

Generate Idea

Collect Samples

Isolate RNA

Create Libraries

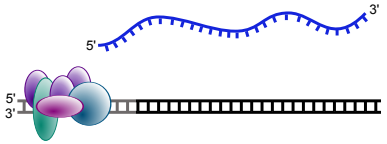
Sequence Samples

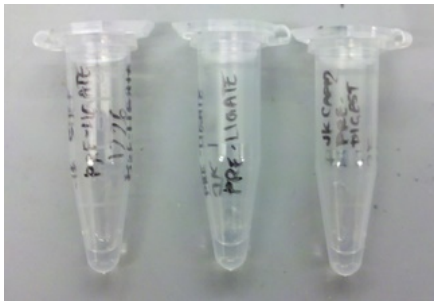
Map Reads

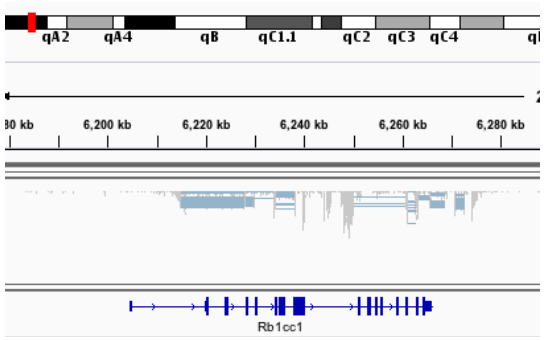
Quantitate Abundance

Test Significance

Rinse, Repeat







	A	B	C	D	E	F
1	Locus	WT (fpkm)	Mut (fpkm)	log2(change)	q_value	Significant
2	XLOC_000001	19.6	16.0	-0.3	0.999	no
3	XLOC_000002	8.6	2.6	-1.7	0.816	no
4	XLOC_000003	34.5	23.6	-0.5	0.999	no
5	XLOC_000004	3.3	20.8	2.6	0.027	yes
6	XLOC_000005	9.2	10.0	0.1	0.999	no
7	XLOC_000006	31.1	36.2	0.2	0.999	no

A Three Part Tour

Part I: Mapping Reads

- FastQC
- Bowtie
- SAMtools
- IGV
- TopHat

Part II: Quantitating Abundance

- Annotations
- RNA-SeQC
- Cufflinks
- HTSeq

Part III: Comparing Genes

- SeqMonk
- Cuffdiff
- CummeRbund
- DESeq

What to Expect

WILL COVER	High-level overview Programs you CAN use Some commonly used options Quality control checks
WON'T COVER	Nitty-gritty theory Programs you SHOULD use All the options How to install programs, use Unix, etc.

Keep Track Along the Way

sample source

extraction method

amount used

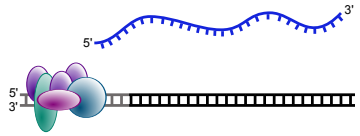
insert size

library kit

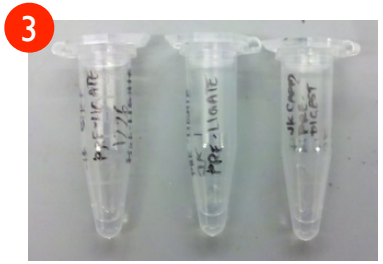
processed batch

1

2



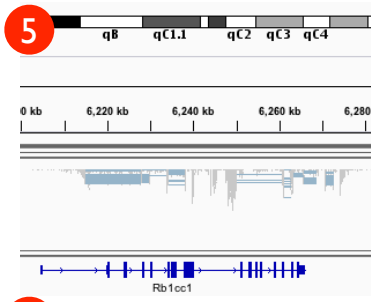
3



4

platform, # reads, cycles, quality, base composition, duplicates

5



genome, program, rRNA, unmapped, exon/intron/intergenic, junctions, isoform, multimap, coverage

6

A	B	C	D
Locus	WT (fpkm)	Mut (fpkm)	log2(change)
OC_000001	19.6	16.0	-0.3
OC_000002	8.6	2.6	-1.7
OC_000003	34.5	23.6	-0.5
OC_000004	3.3	20.8	2.6
OC_000005	9.2	10.0	0.1
OC_000006	31.1	36.2	0.2

annotation, normalization, value +/- variance, replicates, error model, multiple-correction, top100

Characterizing Sequencing Reads

Part I: Mapping Reads	*Part II: Quantitating Abundance*	*Part III: Comparing Genes*
FastQC	Annotations	SeqMonk
Bowtie	RNA-SeQC	Cuffdiff
SAMtools	Cufflinks	CummeRbund
IGV	HTSeq	DESeq
TopHat		

FastQ Stores Both Sequence and Quality

```
$ cd bodymap
$ ls data
$ zcat data/ER-chr22-001_1.fastq.gz | head
```

FASTQ

```
@HS2:192:C0JW3AC:5:1101:1231:2149 1:N:0:CGATGT   ← name
NGCGAAGTATATGGCACAATTGAGCAGTAAAAGTGTTAAAATATAG   ← sequence
+
#1=DDDFDDHFHHIJJJJGHIIHGICGFGIIIHCGIIIEHIJJFHG   ← quality
```

worse

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
0.........................26...31........41
```

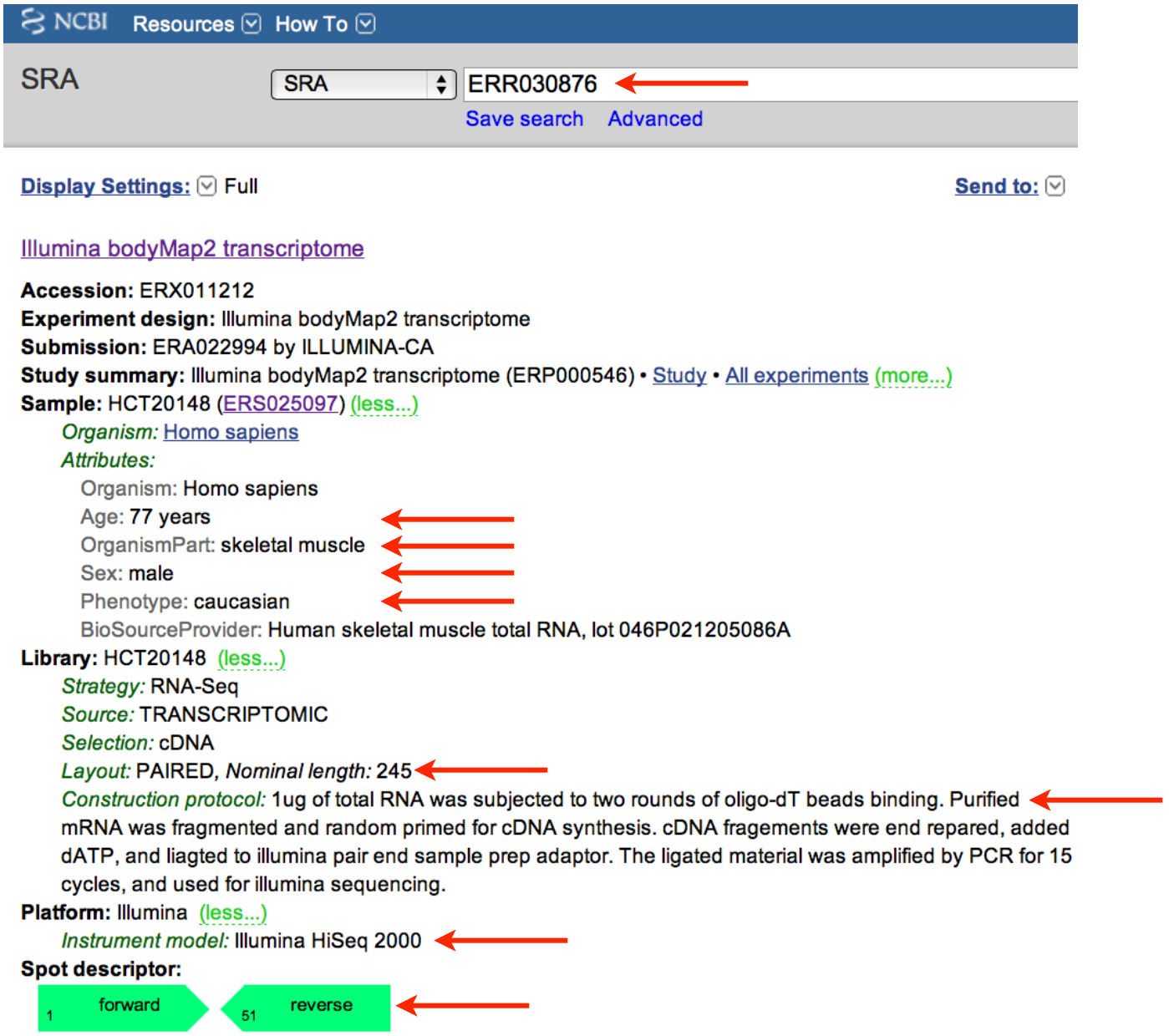
better

FASTA

```
>SequenceA
NGCGAAGTATATGGCACAATTGAGCAGTAAAAGTGTTAAAATATAG
```

wikipedia.org/wiki/FASTQ_format

Sequence Data Is Available from SRA



Assess Read Quality with FastQC

```
$ zcat data/*_1.fastq.gz > reads.fastq
$ wc -l reads.fastq
$ fastqc reads.fastq
```

FastQC Report

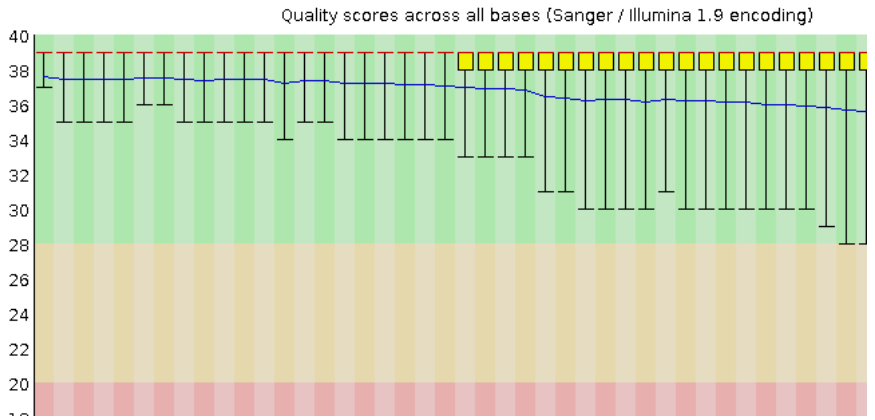
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	reads.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2265661
Filtered Sequences	0
Sequence length	50
%GC	49

Per base sequence quality



bioinformatics.babraham.ac.uk/projects/fastqc/

Mapping Reads to a Genome

Part I: Mapping Reads

FastQC

Bowtie

SAMtools

IGV

TopHat

Part II: Quantitating Abundance

Annotations

RNA-SeQC

Cufflinks

HTSeq

Part III: Comparing Genes

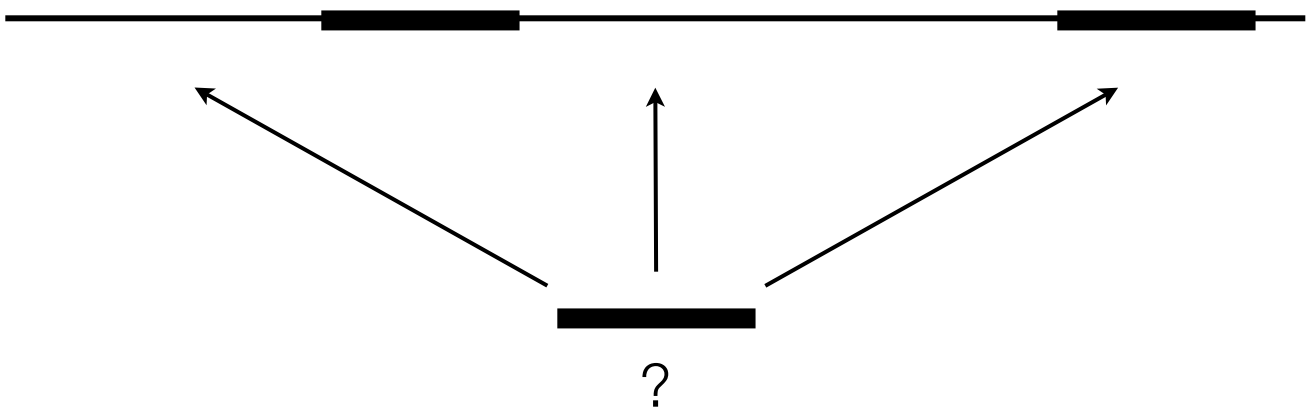
SeqMonk

Cuffdiff

CummeRbund

DESeq

The Alignment Problem



GOALS

Find match quickly

Find true match

Avoid missing a match

COMPLICATIONS

Large genome, many reads

Paralogs, pseudo-genes, repeats

Sequencing errors, polymorphisms

Map Reads to Genome with Bowtie

```
$ cd ~/genomes/human
$ wc hg19-chr22.fa
$ bowtie2-build hg19-chr22.fa hg19-chr22
```

```
$ cd ~/bodymap
$ bowtie ~/genomes/human/hg19-chr22 reads.fastq -S reads.sam
```

```
# reads processed: 2265661
# reads with at least one reported alignment: 1436115 (63.39%)
# reads that failed to align: 829546 (36.61%)
Reported 1436115 alignments to 1 output stream(s)
```

SAM Stores Alignments and Reads

```
$ wc -l reads.sam
$ less -S reads.sam
```

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr22        LN:51304566
@PG     ID:Bowtie       VN:0.12.9       CL:"bowtie /home/workshop/genomes/human/hg19-chr22 re
ER.10/1  16 chr22 44414087 255 50M * 0 0 GCCTT... #####... XA:i:0  MD:Z:4G3A0A5T3A30   NM:i:5
ER.18/1  16 chr22 44567995 255 50M * 0 0 CATCA... A=AA@... XA:i:1  MD:Z:6T26G16        NM:i:2
ER.28/1  4  *     0        0   *   * 0 0 CAAGC... GGGGE... XM:i:0
ER.133/1 16 chr22 50657513 255 50M * 0 0 TCGCG... #####... XA:i:1  MD:Z:4T0G3G9A1T1G26 NM:i:6
ER.145/1 16 chr22 29182951 255 50M * 0 0 GCCGG... 55444... XA:i:0  MD:Z:3T0T45         NM:i:2
ER.185/1 4  *     0        0   *   * 0 0 GAGAA... #####... XM:i:0
ER.236/1 16 chr22 16123595 255 50M * 0 0 CTGGA... #####... XA:i:2  MD:Z:6C11A7C9T13    NM:i:4
```

The SAM Format Specification (v1.4-r985)

The SAM Format Specification Working Group

September 7, 2011

1 The SAM Format Specification

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to

samtools.sourceforge.net/SAM1.pdf

Manipulating Sequence Alignments

Part I: Mapping Reads	*Part II: Quantitating Abundance*	*Part III: Comparing Genes*
FastQC	Annotations	SeqMonk
Bowtie	RNA-SeQC	Cuffdiff
SAMtools	Cufflinks	CummeRbund
IGV	HTSeq	DESeq
TopHat		

Manipulate SAM/BAM Files with SAMtools

```
$ samtools
```

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.19-44428cd

Usage:   samtools <command> [options]

Command: view         SAM<->BAM conversion
         sort         sort alignment file
         mpileup      multi-way pileup
         depth        compute the depth
         faidx        index/extract FASTA
         tview        text alignment viewer
         index        index alignment
```

Li, *et al.*, 2009 Bioinformatics

Determine Why Reads "Unmapped"

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

```
$ samtools view -Sf 0x4 reads.sam | head
```

```
ER.28/1   4 * 0 0 * * 0 0 CAAGCACAAGATCCCCGTGAAGTACCTGGAGTTCATCTCGGAATGCATCA GGGGEFFFFC4,5.444
ER.185/1  4 * 0 0 * * 0 0 GAGAACATCCTGGACCTGAAGATAGAGGATTTCTTAGAGAGACGCCTGCA ################
ER.334/1  4 * 0 0 * * 0 0 NCAAGATCGCACCACTGCACTCCAGCCTGGCGACAGAGCGAGGCTCCGTT #+++*+')++*77666FF
ER.373/1  4 * 0 0 * * 0 0 CCCACGCATCTCCCTGTGCTTGTGGACTGGTTTGGTGATCCACTGGGTGT HHHIHHHIHHHGFHHHH
ER.391/1  4 * 0 0 * * 0 0 ATTATTAGGGATCCAGAGGATCACTACAAAGTAGTAAAACTTTAATTCAT HHHHHAA@>@45555HH
```

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast

Quickly Extract Sorted, Indexed BAMs

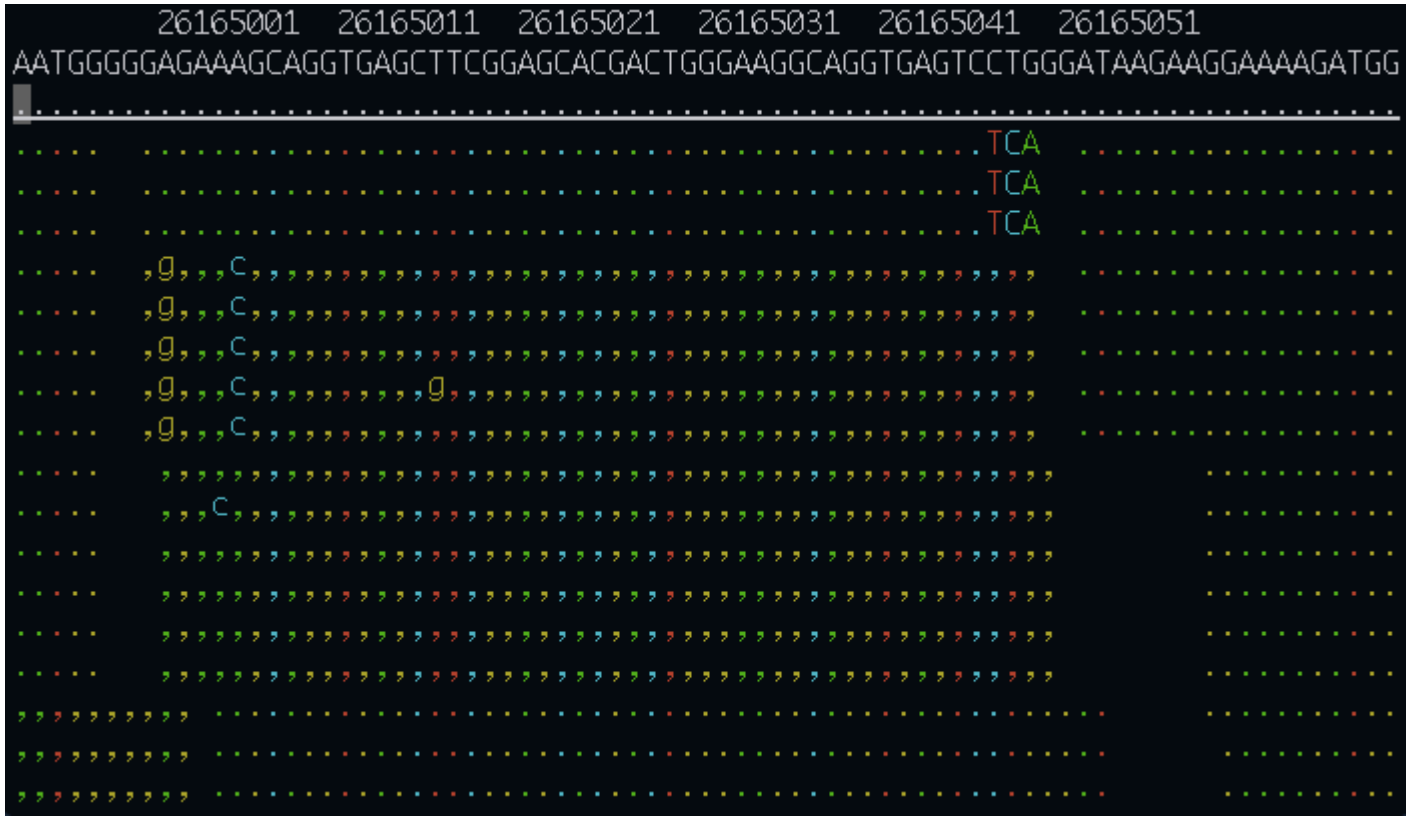
```
$ samtools view -bS reads.sam > reads.bam
$ samtools sort reads.bam reads-sorted
$ samtools index reads-sorted.bam
```

```
$ samtools view reads-sorted.bam chr22:26,164,000-26,164,085
```

```
ER.26848640/1 16 chr22 26164062 255 50M * 0 0 TCGCCCCTCCAGAACGAGAGACCCCAGAAATTTCCATCAGCCA
ER.40045895/1 16 chr22 26164078 255 50M * 0 0 AGAGATCCCAGAAATTTCCATCAGCCAACCCAACAGCAAGTCC
ER.8036632/1  16 chr22 26164081 255 50M * 0 0 GATCCCAGAAATTTCCATCAGCCAACCCAACAGCAAGTCCAGC
ER.16689872/1 16 chr22 26164081 255 50M * 0 0 GATCCCAGAAATTTCCATCAGCCAACCCAACAGCAAGTCCAGC
ER.40220564/1 16 chr22 26164081 255 50M * 0 0 GATCCCAGAAATTTCCATCAGCCAACCCAACAGCAAGTCCAGC
ER.42881111/1 16 chr22 26164081 255 50M * 0 0 GATCCCAGAAATTTCCATCAGCCAACCCAACAGCAAGTCCAGC
ER.51649994/1 16 chr22 26164081 255 50M * 0 0 GATCCCAGAAATTTCCATCAGCCAACCCAACAGCAAGTCCAGC
ER.56732719/1 16 chr22 26164081 255 50M * 0 0 GATCCCAGAAATTTCCATCAGCCAACCCAACAGCAAGTCCAGC
```

View Sorted, Indexed BAM Files

```
$ samtools tview reads-sorted.bam ~/genomes/human/hg19-chr22.fa
```



(?)help
(q)uit
(g)o

chr22:26,164,000

(.) toggle
(n) color
(r) reads

<SPACE>
<BACKSPACE>

Viewing Your Alignments

Part I: Mapping Reads

- FastQC
- Bowtie
- SAMtools
- **IGV**
- TopHat

Part II: Quantitating Abundance

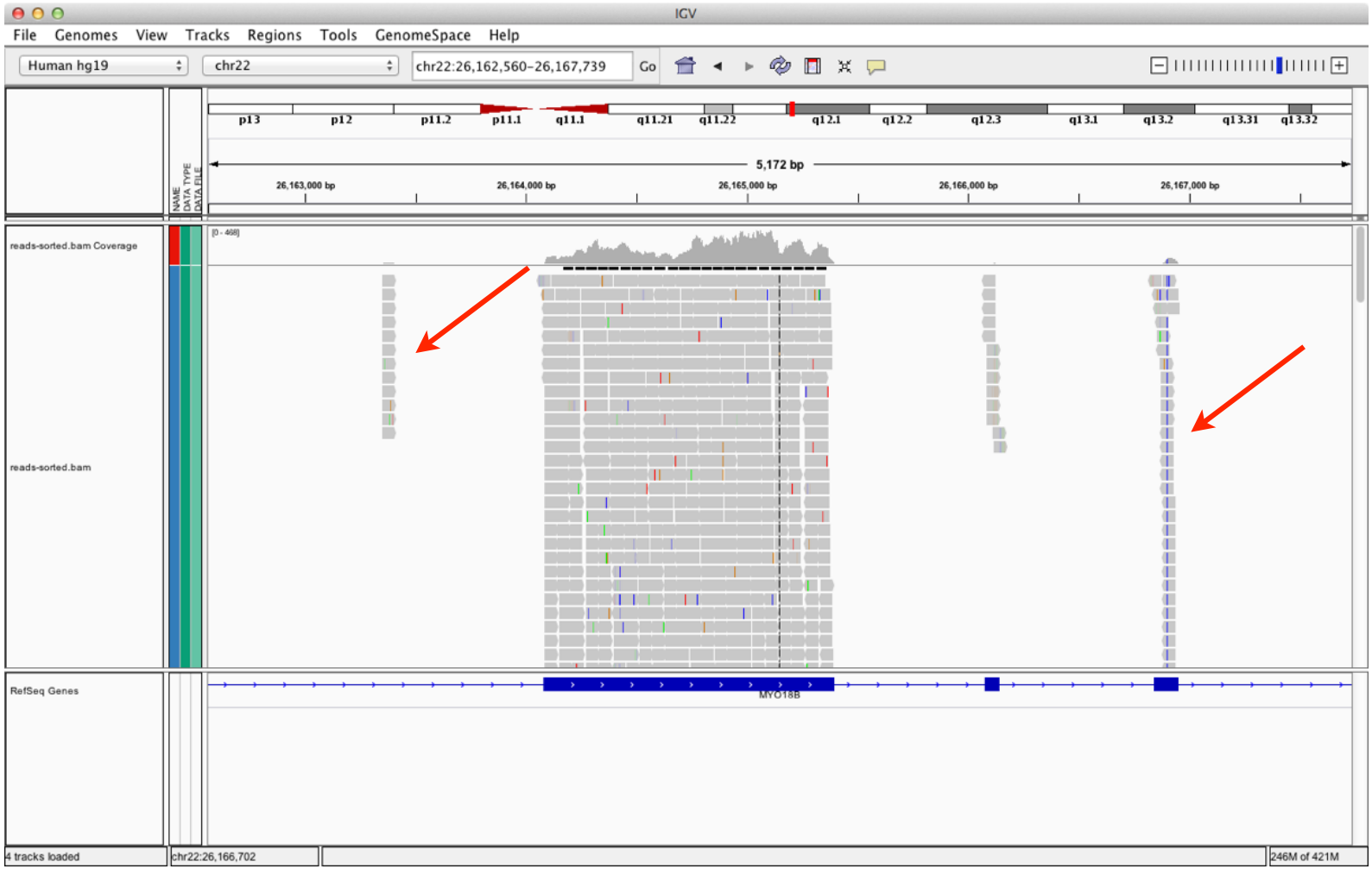
- Annotations
- RNA-SeQC
- Cufflinks
- HTSeq

Part III: Comparing Genes

- SeqMonk
- Cuffdiff
- CummeRbund
- DESeq

“High-Performance Genomics Data Visualization and Exploration”

```
$ scp workshop@192.168.56.101:bodymap/reads-sorted.* .
```

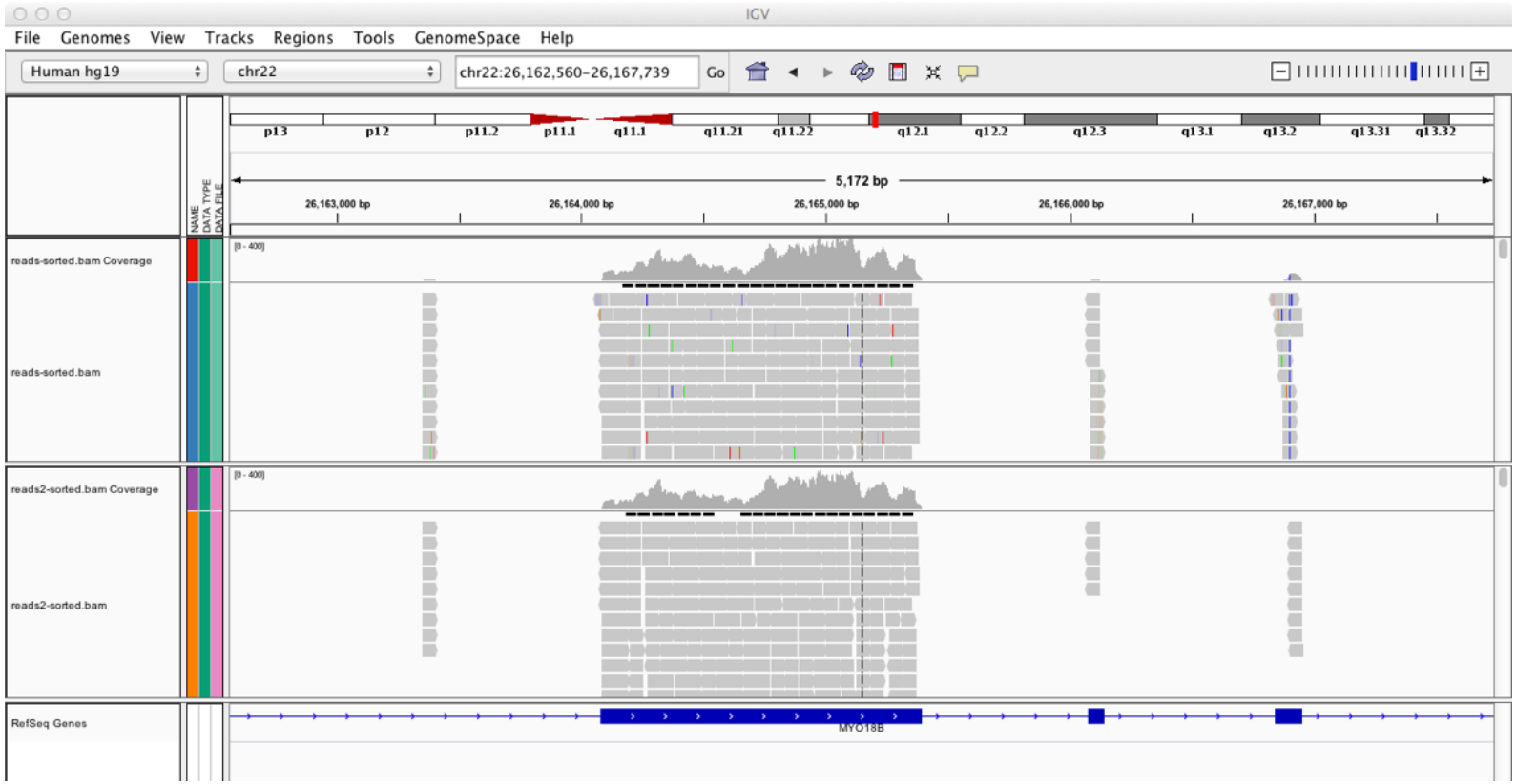


Robinson, *et al.*, 2011 Nature Biotechnology

Rinse, Revise, Repeat

```
$ bowtie -v0 -m1 ~/genomes/human/hg19-chr22 reads.fastq -S reads2.sam
```

```
# reads processed: 2265661
# reads with at least one reported alignment: 693092 (30.59%)
# reads that failed to align: 1517326 (66.97%)
# reads with alignments suppressed due to -m: 55243 (2.44%)
Reported 693092 alignments to 1 output stream(s)
```



Mapping Junction Reads

Part I: Mapping Reads

- FastQC
- Bowtie
- SAMtools
- IGV
- **TopHat**

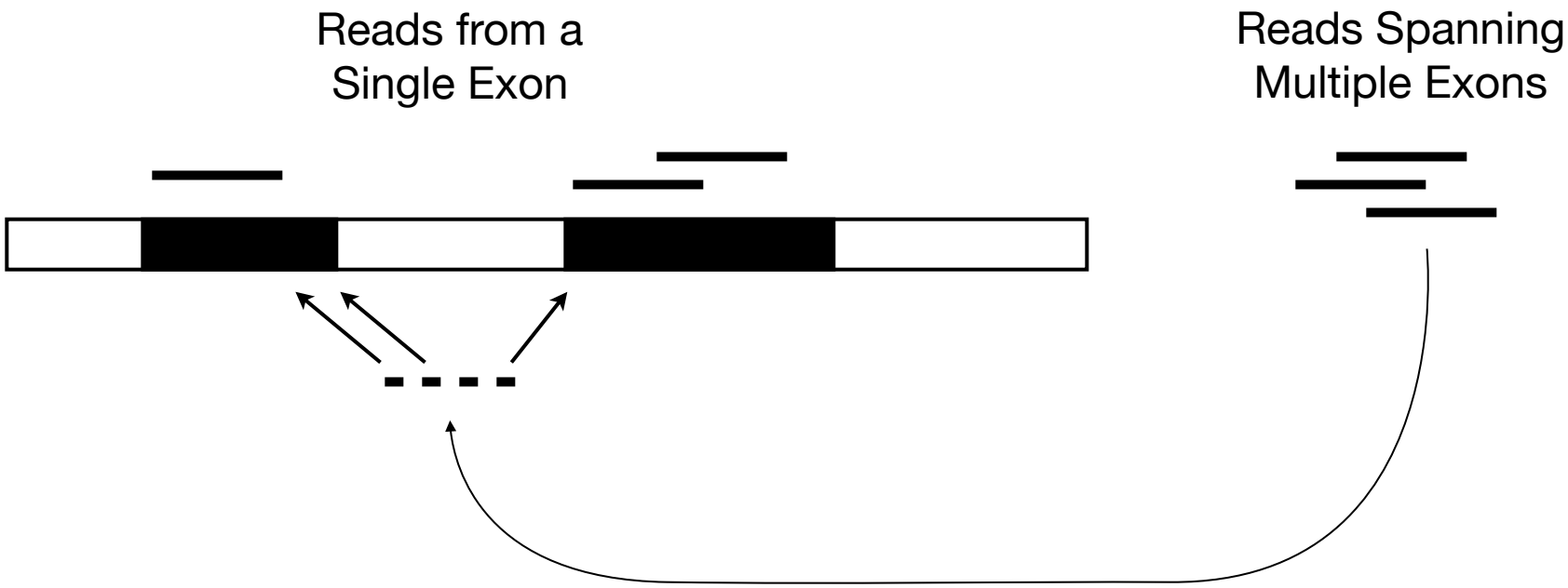
Part II: Quantitating Abundance

- Annotations
- RNA-SeQC
- Cufflinks
- HTSeq

Part III: Comparing Genes

- SeqMonk
- Cuffdiff
- CummeRbund
- DESeq

The Junctions Read Problem



Map Non-Junction and Junction Reads Using TopHat

```
$ tophat -T --no-novel-juncs -G ~/genomes/human/hg19-chr22-iGenomes.gtf
~/genomes/human/hg19-chr22 reads.fastq
```

```
[2013-07-02 13:25:43] Beginning TopHat run (v2.0.3)
-----------------------------------------------
[2013-07-02 13:25:43] Checking for Bowtie
        Bowtie version:   2.0.0.6
[2013-07-02 13:25:43] Checking for Samtools
      Samtools version:   0.1.19.0
[2013-07-02 13:25:43] Checking for Bowtie index files
[2013-07-02 13:25:43] Checking for reference FASTA file
[2013-07-02 13:25:43] Generating SAM header for /home/workshop/
genomes/human/hg19-chr22
   format:       fastq
   quality scale:   phred33 (default)
[2013-07-02 13:25:43] Reading known junctions from GTF file
[2013-07-02 13:25:43] Preparing reads
                [...]
-----------------------------------------------
[2013-07-02 13:34:20] Run complete: 00:08:37 elapsed
```

Trapnell, Pachter, and Salzberg, 2009 Bioinformatics; Kim, et al., 2013 Genome Biology

View Junction Reads in IGV

```
$ cd tophat_out
$ ls
$ samtools index accepted_hits.bam
```

```
$ scp workshop@192.168.56.101:bodymap/tophat_out/accepted_hits.* .
```

A Three Part Tour

Part I: Mapping Reads

- FastQC
- Bowtie
- SAMtools
- IGV
- TopHat

Part II: Quantitating Abundance

- Annotations
- RNA-SeQC
- Cufflinks
- HTSeq

Part III: Comparing Genes

- SeqMonk
- Cuffdiff
- CummeRbund
- DESeq

Workshop Slides and More

Carnegie-Style Science | Labs | Research Programs | Faculty | Resources | Seminars | Outreach | Contact

Home

Tan Lab

COURSE MATERIALS

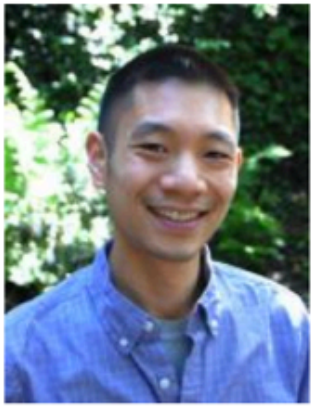
Workshops

- Intro to RNA-Seq - Summer 2013 - Day 1 / Day 2 / Day 3
- Intro to Unix - Spring 2013 - CCG Short Course - Day 1 slides / Day 2 slides

Perl Thursdays

- July 18, 2013 - refining session
- June 27, 2013 - create average coverage map for a set of genes (top200.cov groupCoverage.pl)
- May 30, 2013 - refining session (gt1kbexon-v2.pl)
- May 2, 2013 - "Hello, world!", filter for exons on + strand >1 kb (mouse.gtf gt1kbexon.pl)

Nitty Gritty Workflows



Frederick Tan
Bioinformatics
Office Telephone: (410) 246-3084
Lab Telephone: (410) 246-3083
Department Fax: (410) 243-6311
Email

Lab Members »

Rosa Alcazar, Visiting Scientist
- Research Summary
- Course Material
- Sequence Analysis