Introduction to RNA-Seq

Part II: Quantitating Abundance

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

16 July 2013

TopHat Aligns RNA-Seq Data Using Bowtie As an Alignment Engine


```
tophat -o skeletal_thout -G hg19-chr22-iGenomes.gtf hg19-chr22
ERR030876_1.fastq ERR030876_2.fastq

tophat -o heart_thout -G hg19-chr22-iGenomes.gtf hg19-chr22
ERR030886_1.fastq ERR030886_2.fastq

tophat -o wbc_thout -G hg19-chr22-iGenomes.gtf hg19-chr22
ERR030875_1.fastq ERR030875_2.fastq
```

TopHat Reports Additional Alignment Information

```
$ cd bodymap
$ ls heart_thout
```

```
accepted_hits.bam   insertions.bed   logs                unmapped.bam
deletions.bed       junctions.bed    prep_reads.info
```

```
$ samtools view heart_thout/accepted_hits.bam | less -S
```

```
ER.18184843   73 chr22 16051721 50 50M *        0        0 ... AS:i:-10 ... NH:i:1
ER.41951719   73 chr22 16051721 50 50M *        0        0 ... AS:i:-10 ... NH:i:1
ER.53800159   73 chr22 16051721 50 50M *        0        0 ... AS:i:-10 ... NH:i:1
ER.75785528  353 chr22 16051885  0 50M = 36294994 20243159 ... AS:i:-5  ... NH:i:6 ..
ER.75785528  321 chr22 16051885  0 50M = 36927193 20875358 ... AS:i:-5  ... NH:i:6 ..
ER.75785528  353 chr22 16051885  0 50M = 41267865 25216030 ... AS:i:-5  ... NH:i:6 ..
```

```
$ samtools view heart_thout/accepted_hits.bam | wc -l
```

```
5,862,208 + 162,238,981 > 165,837,568
```

unmapped FastQ

Comparing Genome Annotations

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

There Are Many Sources for Genome Annotations

UCSC Genome Bioinformatics

e!Ensembl

good first choice
(gene_id, gene_name)

Illumina iGenomes

WormBase

genome.ucsc.edu/cgi-bin/hgTables
ensembl.org
support.illumina.com/sequencing/sequencing_software/igenome.ilmn

GTF Stores Position and Other Attributes

GFF
(General Feature Format)

GFF2 **GTF2** GFF3

```
$ cd ~/genomes/human
$ less -S Homo_sapiens.GRCh37.72-chr22.gtf
```

seq	source	feature	start	end	score	strand	frame	attributes
		exon						gene_id
		CDS						transcript_id
		start_codon						gene_name
		stop_codon						tss_id
								p_id

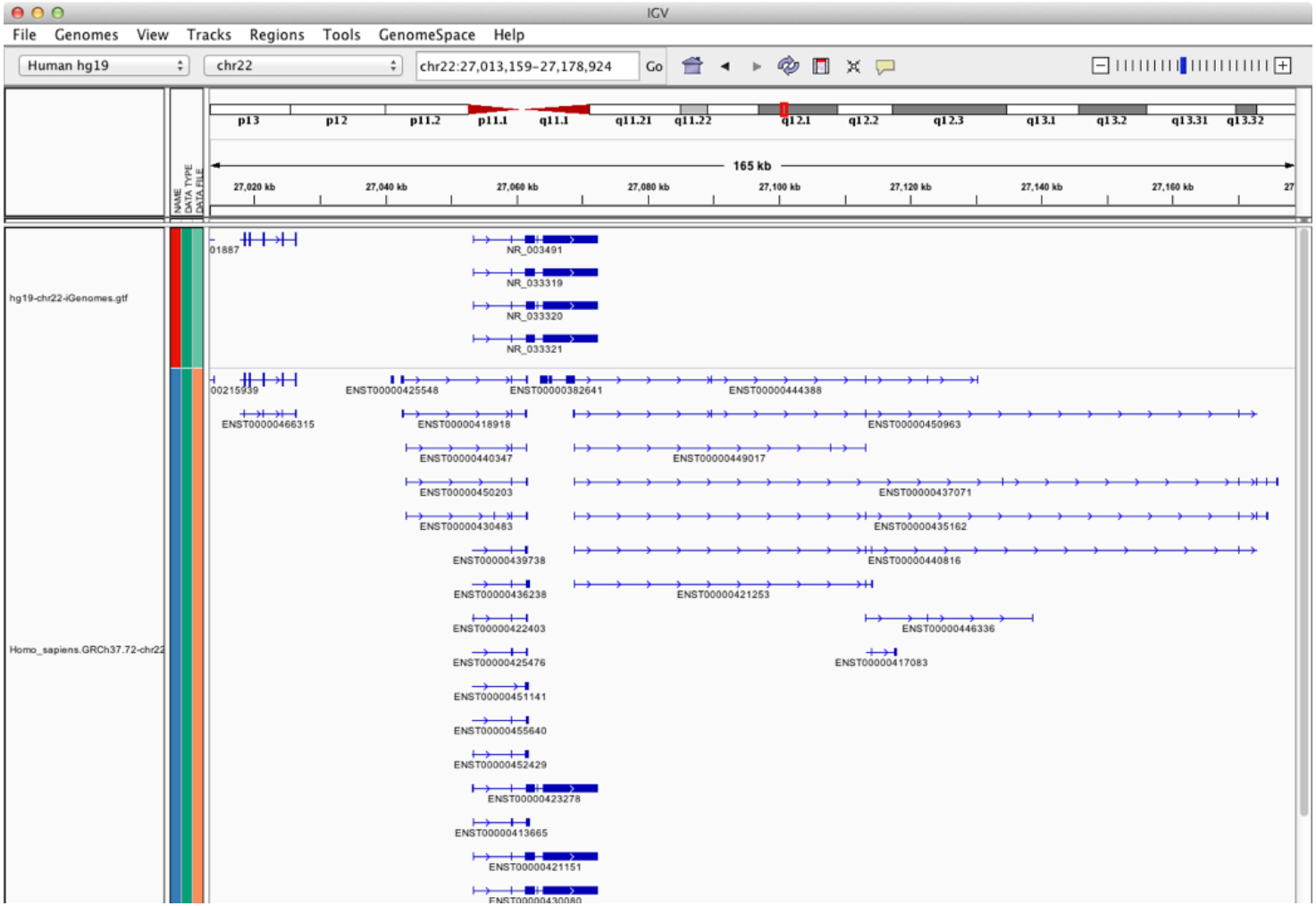
```
$ cut -f2 Homo_sapiens.GRCh37.72-chr22.gtf | sort | uniq -c
```

sanger.ac.uk/resources/software/gff/spec.html
mblab.wustl.edu/GTF22.html
sequenceontology.org/gff3.shtml

Visually Compare Annotations in IGV

```
$ wc -l hg19-chr22-iGenomes.gtf
$ wc -l Homo_sapiens.GRCh37.72-chr22.gtf
```

```
$ scp workshop@192.168.56.101:genomes/human/*.gtf .
```



Track Versions for Annotations and Genomes

Homo_sapiens.GRCh37.72.gtf

species

genome assembly

Ensembl release

```
grep "^>" GRCh37.72.fa
                  84 total
>10
>11
>12
>13
   [...]
>GL000248.1
>GL000249.1
>MT
>X
>Y
```

```
grep "^>" hg19.fa
                  93 total
>chr10
>chr11
>chr11_gl000202_random
      [...]
>chr17_ctg5_hap1
      [...]
>chrUn_gl000248
>chrUn_gl000249
>chrX
>chrY
```

ensembl.org/info/docs/genebuild/assembly.html
ensembl.org/Homo_sapiens/Info/Annotation
genome.ucsc.edu/FAQ/FAQreleases.html

Characterizing Sequence Alignments

Part I: Mapping Reads

- FastQC
- Bowtie
- SAMtools
- IGV
- TopHat

Part II: Quantitating Abundance

- Annotations
- **RNA-SeQC**
- HTSeq
- Cufflinks

Part III: Comparing Genes

- SeqMonk
- Cuffdiff
- CummeRbund
- DESeq

Understand Where Your Reads Map in a Genome

Mapped/Unmapped?
Unique/Multimapped?
Intra/Intergenic?
Exon/Intron?
Ribosomal RNA?

RNA-SeQC

Sorted/indexed .BAM
Sorted/indexed .FA
Read Groups

```
$ cd ~/bodymap
$ AddOrReplaceReadGroups I=skeletal_thout/accepted_hits.bam
  O=skeletalRG.bam LB=a PL=illumina PU=a SM=a
$ samtools index skeletalRG.bam
$ samtools faidx ~/genomes/human/hg19-chr22.fa
$ bwa index ~/genomes/human/hg19-rDNA.fa
```

```
$ RNA-SeQC -BWArRNA ~/genomes/human/hg19-rDNA.fa
  -o skeletal-Ensembl-rDNA -r ~/genomes/human/hg19-chr22.fa
  -s "a|skeletalRG.bam|a"
  -t ~/genomes/human/Homo_sapiens.GRCh37.72-chr22.gtf -n 50
```

gatkforums.broadinstitute.org/discussion/1317/collected-faqs-about-bam-files
picard.sourceforge.net
Li and Durbin, 2009 Bioinformatics; Deluca, *et al.*, 2012 Bioinformatics

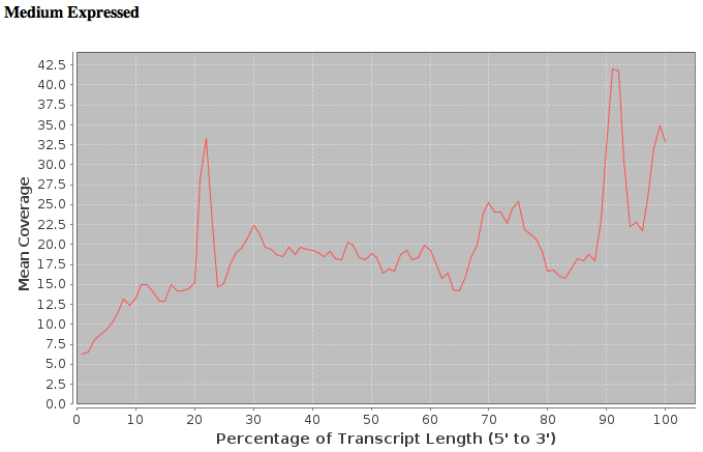
Working through an RNA-SeQC Report

Total Reads

Sample	Note	Total Purity Filtered Reads Sequenced	Alternative Aligments	
a	a	3,480,685	2,012,914	

	rRNA	rRNA rate
.3	46,508	0.013

	Fragment Length Mean	Fragment Length StdDev	
4	102	171	



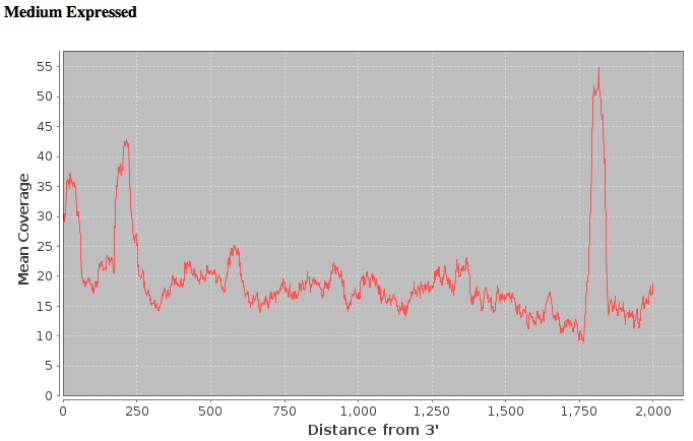
Transcript-associated Reads

Sample	Note	Intragenic Rate	Exonic Rate	Intronic Rate	Intergenic Rate	Expression Profiling Efficiency	Transcripts Detected	Genes Detected
a	a	0.910	0.737	0.173	0.086	0.737	3,209	643

Coverage Metrics for Middle 50 Expressed Transcripts

The metrics in this table are calculated across the transcripts that were determined to have th

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'200Base Norm	No. C
a	a	19.07	1.22	37	0.34	



broadinstitute.org/webservices/gpModuleRepository/download/prod/module/?file=/RNASeQC/broad.mit.edu:cancer.software.genepattern.module.analysis/00263/2/RNA-SeQC.pdf

Compare Replicates with RNA-SeQC

```
$ RNA-SeQC -BWArRNA ~/genomes/human/hg19-rDNA.fa
  -o skeletal-Ensembl-rDNA -r ~/genomes/human/hg19-chr22.fa
  -s samples.txt
  -t ~/genomes/human/Homo_sapiens.GRCh37.72-chr22.gtf
```

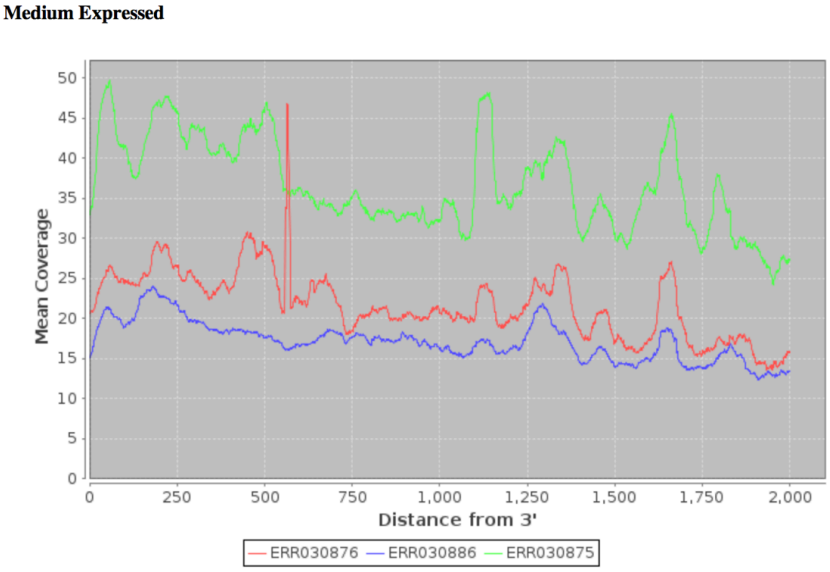
```
Sample ID     Bam File          Notes
ERR030876     skeletal.bam      76
ERR030886     heart.bam         86
ERR030875     wbc.bam           75
```

Spearman Correlation Matrix

Sample	Note	ERR030876	ERR030886	ERR030875
ERR030876	76	1.000	0.900	0.714
ERR030886	86	0.900	1.000	0.765
ERR030875	75	0.714	0.765	1.000

Pearson Correlation Matrix

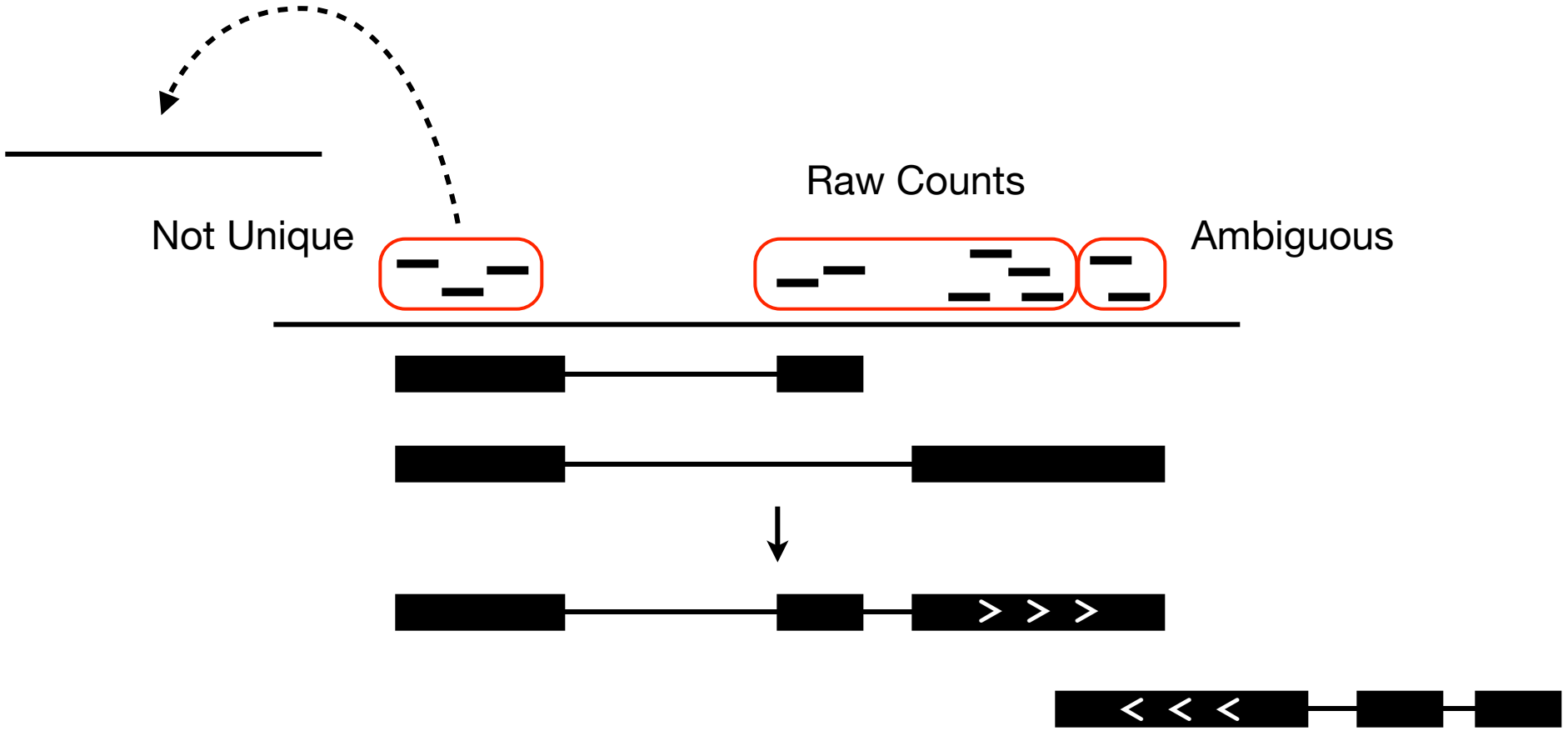
Sample	Note	ERR030876	ERR030886	ERR030875
ERR030876	76	1.000	0.871	0.714
ERR030886	86	0.871	1.000	0.769
ERR030875	75	0.714	0.769	1.000



Quantitating Abundance - Strategy A

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

Strategy A: Only Count Reads that Map Unambiguously



Create Counts Table Using HTSeq

```
$ samtools sort -n skeletalRG.bam skeletal-nameSort
$ samtools view skeletal-nameSort.bam |
  python -m HTSeq.scripts.count -s no -
  ~/genomes/human/Homo_sapiens.GRCh37.72-chr22.gtf > skeletal.counts
$ less skeletal.counts
```

```
ENSG00000008735         24
ENSG00000015475         1251
ENSG00000025708         100
ENSG00000025770         1248
      [...]
no_feature              386808
ambiguous               111714
too_low_aQual           0
not_aligned             0
alignment_not_unique    1851790
```

Sanity Check Top Hits Multiple Ways

```
$ sort -rnk2 skeletal.counts | head -n100
```

Gene: MB ENSG00000198125

Description	myoglobin [Source:HGNC Symbol;Acc:6915]
Location	Chromosome 22: 36,002,811-36,033,998 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000684.1:36002811:36033998:1
Transcripts	

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
76 DAVID IDs

Options **Classification Stringency** Medium

Rerun using options | Create Sublist

23 Cluster(s) Download File

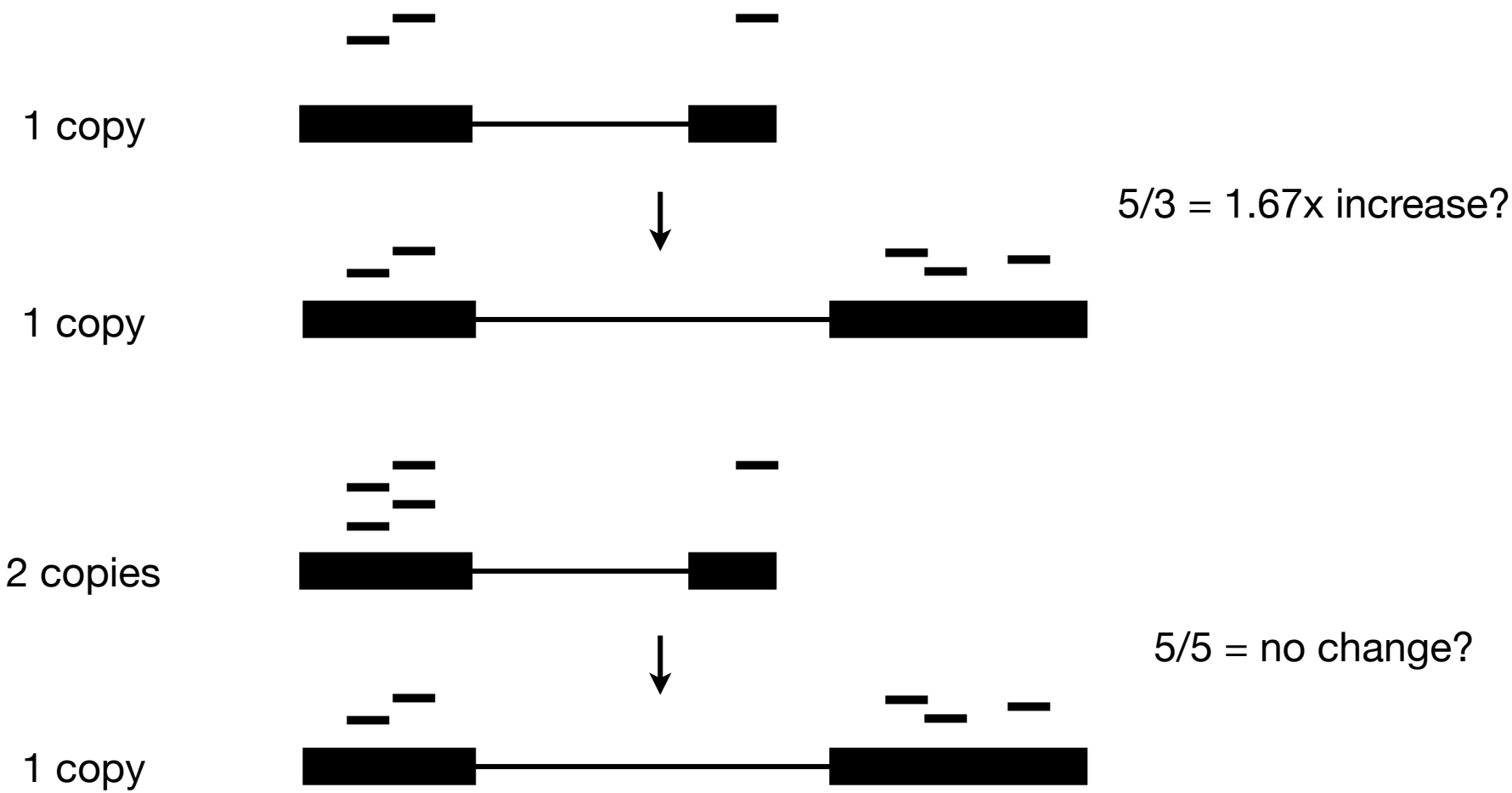
	Annotation Cluster 1	Enrichment Score: 1.68			Count	P_Value	Benjamini
	GOTERM_CC_FAT	nuclear lumen	RT		12	7.1E-3	7.6E-1
	GOTERM_CC_FAT	intracellular organelle lumen	RT		13	1.2E-2	5.5E-1
	GOTERM_CC_FAT	organelle lumen	RT		13	1.4E-2	5.1E-1
	GOTERM_CC_FAT	membrane-enclosed lumen	RT		13	1.7E-2	4.9E-1
	GOTERM_CC_FAT	nucleolus	RT		7	2.8E-2	5.1E-1
	GOTERM_CC_FAT	nucleoplasm part	RT		6	3.7E-2	5.7E-1
	GOTERM_CC_FAT	nucleoplasm	RT		7	7.3E-2	6.9E-1
	Annotation Cluster 2	**Enrichment Score: 1.45**			**Count**	**P_Value**	**Benjamini**
	GOTERM_MF_FAT	RNA polymerase II transcription factor activity	RT		6	3.6E-3	4.8E-1

Dennis G Jr, *et al.*, 2003 Genome Biology; Huang da, Sherman, and Lempicki 2009 Nature Protocols

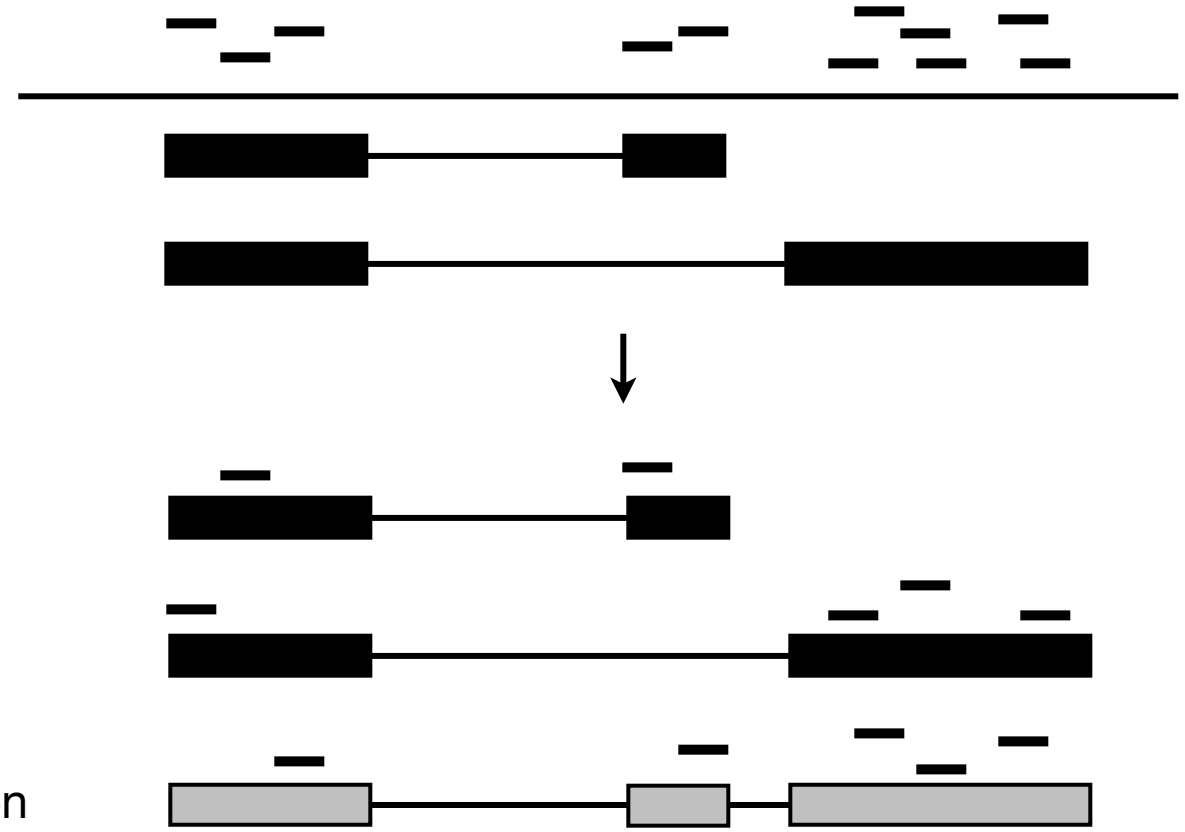
Quantitating Abundance - Strategy B

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

How Does One Account for Changes in Isoform Abundance?



Strategy B: Account for Uncertainty to Recover Ambiguous Reads



One alternative: DEXSeq (Anders, Reyes, and Huber, 2012 Genome Research)

Quantitate Abundance Using Cufflinks

```
$ cufflinks
```

```
-G/--GTF          quantitate against reference transcript annotations
-g/--GTF-guide    use reference transcript annotation to guide assembly
```

```
$ cufflinks -o skeletal_clout
  -G ~/genomes/human/Homo_sapiens.GRCh37.72-chr22.gtf skeletalRG.bam
```

```
[09:05:04] Loading reference annotation.
[09:05:04] Inspecting reads and determining fragment length distribution.
> Processed 745 loci.                          [*************************] 100%
> Map Properties:
>   Normalized Map Mass: 2276157.23
>   Raw Map Mass: 2276157.23
>   Fragment Length Distribution: Empirical (learned)
>                 Estimated Mean: 195.77
>               Estimated Std Dev: 78.52
[09:05:34] Estimating transcript abundances.
> Processed 745 loci.                          [*************************] 100%
```

```
$ cufflinks -o skeletal_clout2 -u -b hg19-chr22.fa
  -G ~/genomes/human/Homo_sapiens.GRCh37.72-chr22.gtf skeletalRG.bam
```

Trapnell, *et al.*, 2010 Nature Biotech; Roberts, *et al.*, 2011 Genome Biology; Trapnell, *et al.*, 2012 Nature Protocols; Trapnell, *et al.*, 2013 Nature Biotech

Cufflinks Reports Additional Quantitation Information

```
$ ls skeletal_clout
```

```
genes.fpkm_tracking isoforms.fpkm_tracking skipped.gtf transcripts.gtf
```

FPKM = **F**ragments **P**er **K**ilobase of transcript per **M**illion fragments mapped

```
$ sort -rnk10,10 skeletal_clout/genes.fpkm_tracking | less -S
```

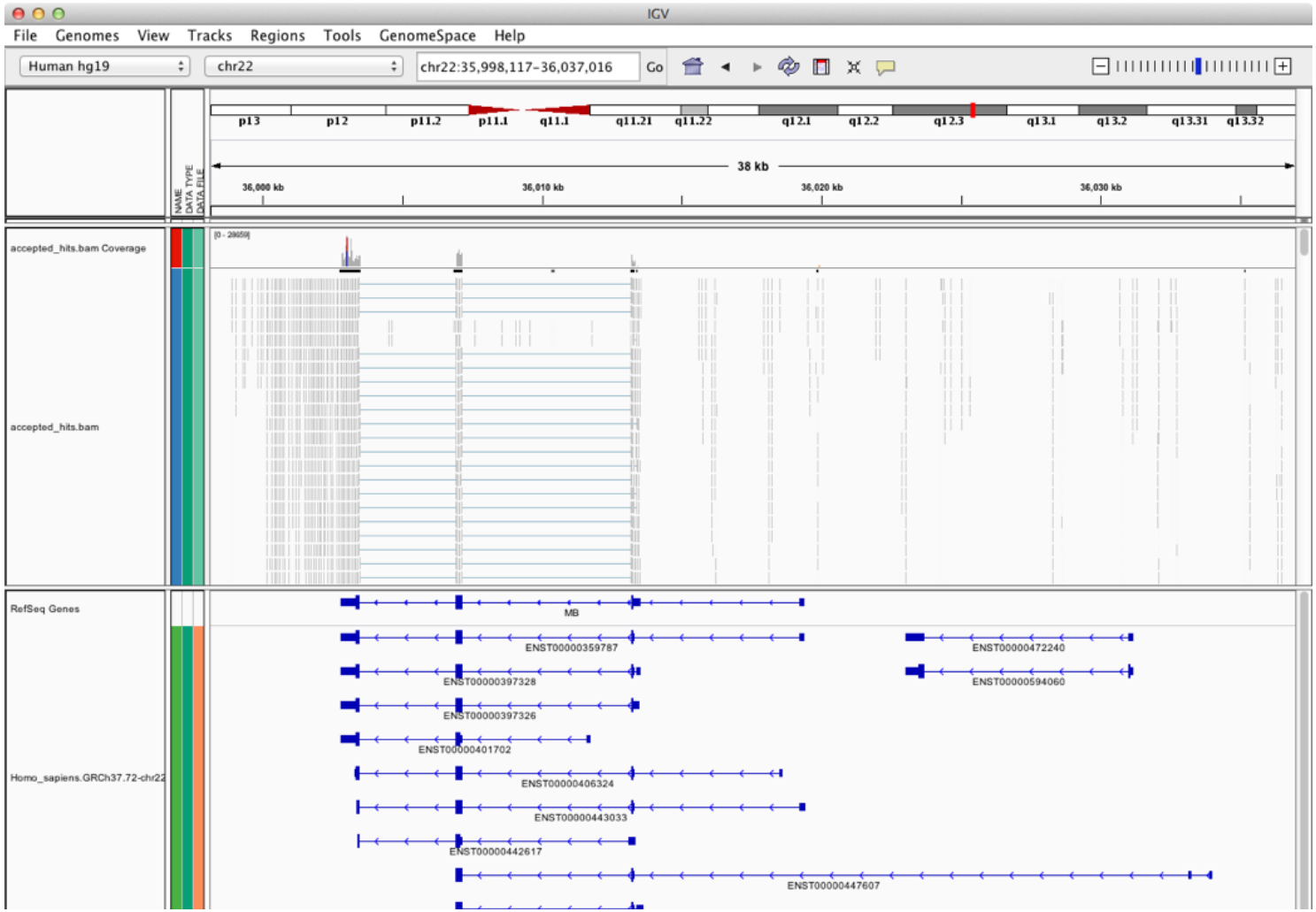
#387 by HTSeq →

tracking_id	gene_short_name	FPKM	FPKM_lo	FPKM_hi
ENSG00000199248	RNU6-28	123500	100666	146334
ENSG00000226160	LA16c-17H1.3	89451.6	59634.4	119269
ENSG00000266837	Z97055.1	72812.7	44078.7	101547
ENSG00000206615	U6	55872.3	39593.4	72151.3
ENSG00000264757	MIR3198-1	51306.4	22846.7	79766.1
ENSG00000266175	AL080243.1	44860.9	3433.17	86288.6
ENSG00000214093	RP11-247I13.3	44397.6	43853.9	44941.4
ENSG00000198125	MB	44090.6	43796.5	44384.8

Sanity Check Visually Using IGV

```
$ samtools index skeletal_thout/accepted_hits.bam
```

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



Wrap-Up

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

Workshop Slides and More

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

Home

Tan Lab

COURSE MATERIALS

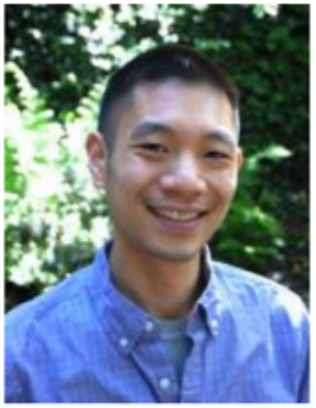
Workshops

- 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