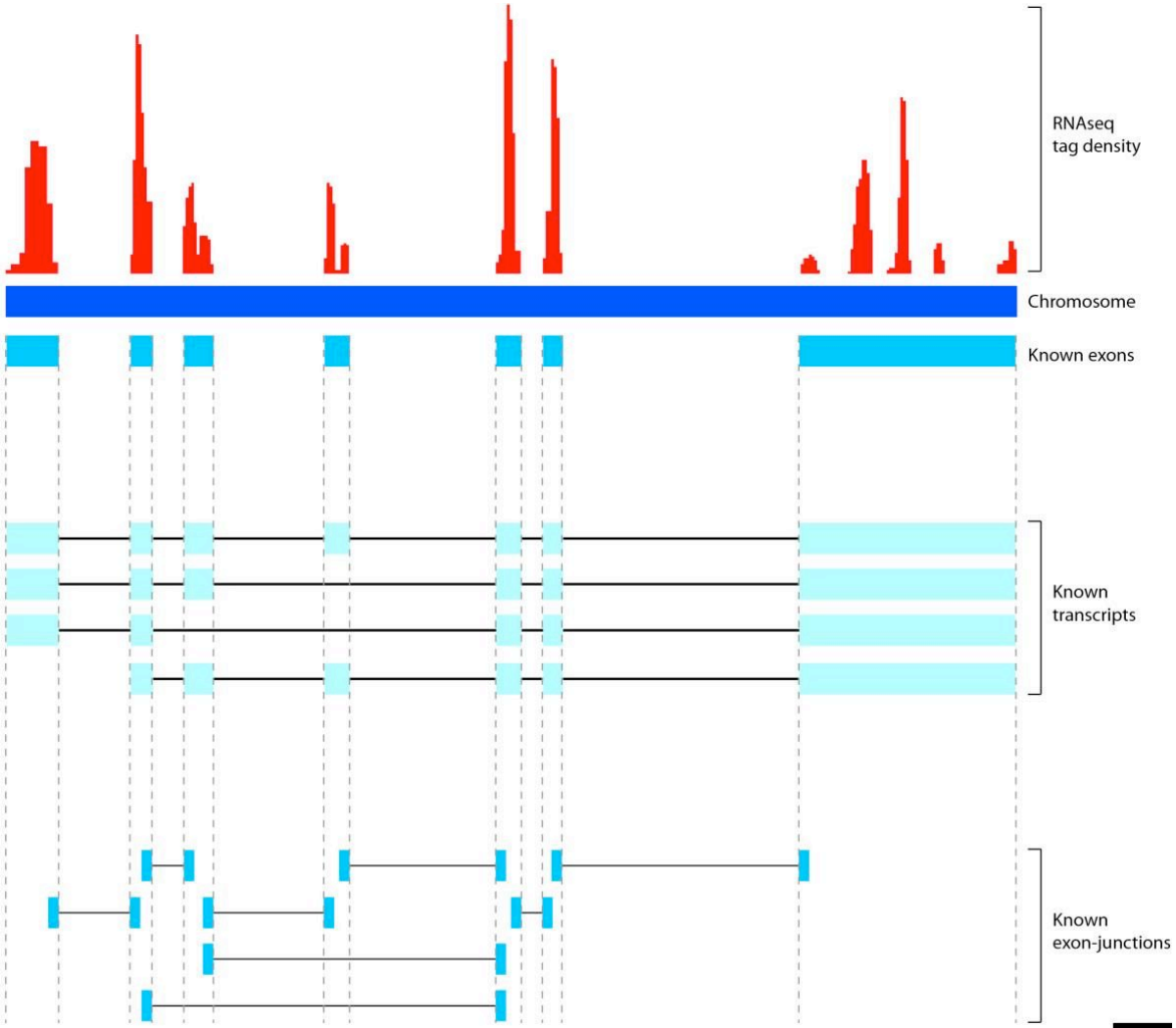
Next-Generation Sequencing (NGS) Technologies and Data Analysis

Christopher E. Mason
TA: Paul Zumbo
Spring 2010

Class #3:
Algorithms for RNA-Seq Analysis

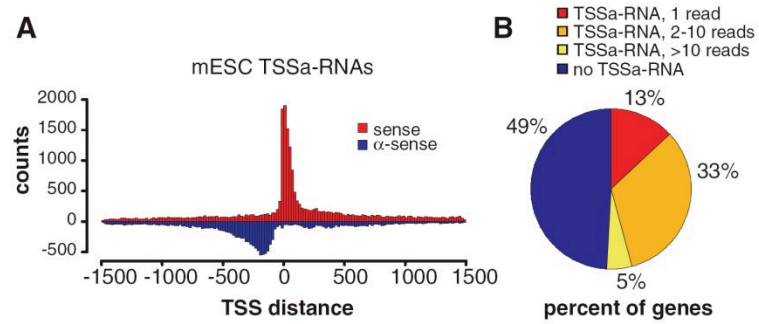


Many types of RNA-Seq

1. TSSa-RNAs (short RNAs)

Divergent Transcription from Active Promoters

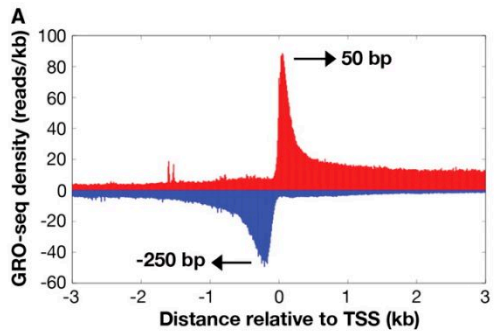
Amy C. Seila,[1*] J. Mauro Calabrese,[1,2*†] Stuart S. Levine,[3] Gene W. Yeo,[4‡] Peter B. Rahl,[3] Ryan A. Flynn,[1] Richard A. Young,[2,3] Phillip A. Sharp[1,2§]



2. GRO-Seq (global run-on) from engaged RNA-Pol II

Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters

Leighton J. Core,* Joshua J. Waterfall,* John T. Lis†



3. Antisense-Transcription

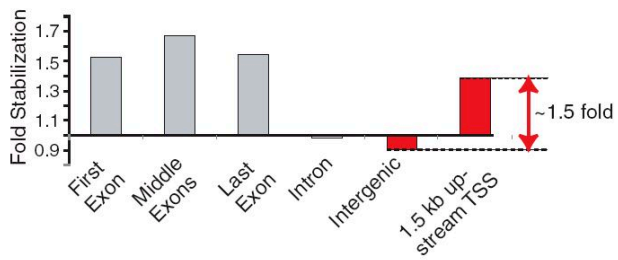
The Antisense Transcriptomes of Human Cells

Yiping He, Bert Vogelstein, Victor E. Velculescu, Nickolas Papadopoulos,* Kenneth W. Kinzler

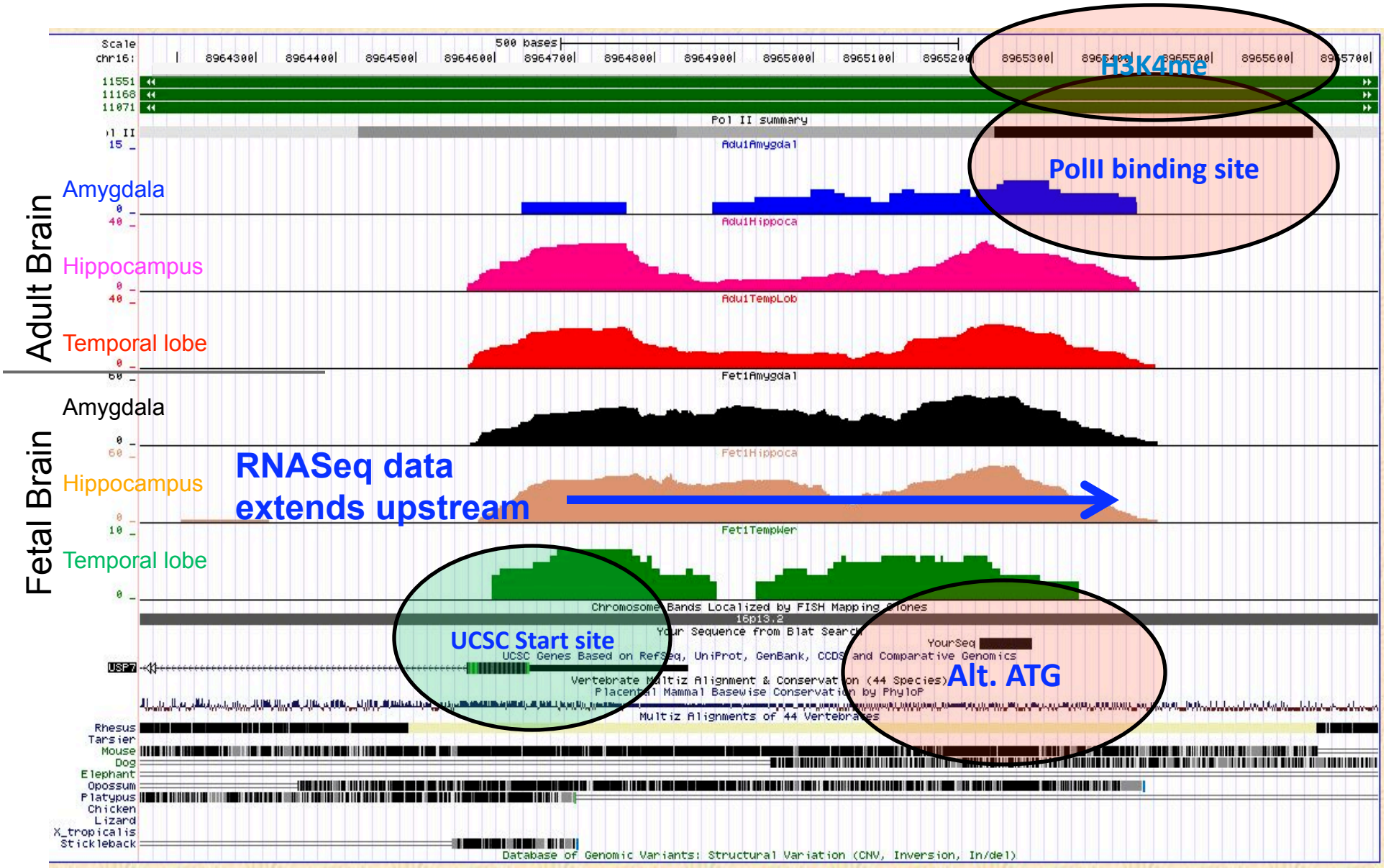
4. PROMPTs (promoter upstream transcripts)

RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters

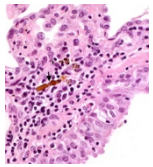
Pascal Preker,[1] Jesper Nielsen,[2] Susanne Kammler,[1*] Søren Lykke-Andersen,[1] Marianne S. Christensen,[1] Christophe K. Mapendano,[1] Mikkel H. Schierup,[2] Torben Heick Jensen[1†]



Ideally, we should Integrate RNA-Seq, Chip-Seq, and methylation data to find/refine genesa and loci

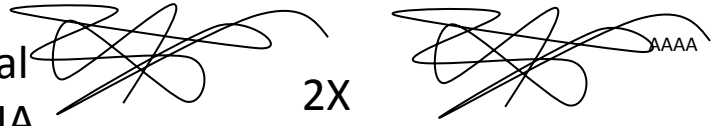


Illumina sample prep



Dissect & Homogenize sample

Extract total RNA & gDNA eliminator column & DNAase



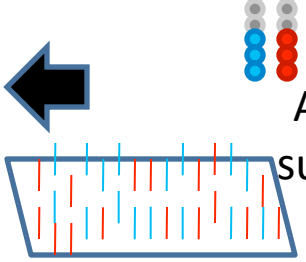
2X Oligo-dT Isolate polyA RNA

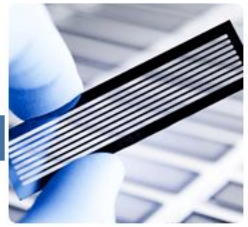
Fragment RNA

1st and 2nd strand cDNA synthesis

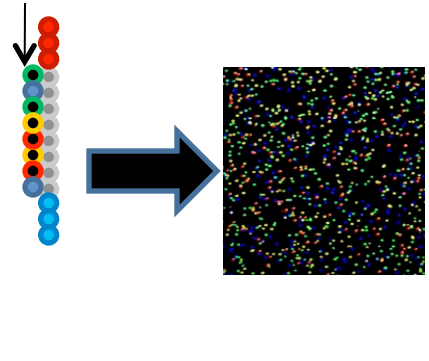
Ligate adaptors to DNA

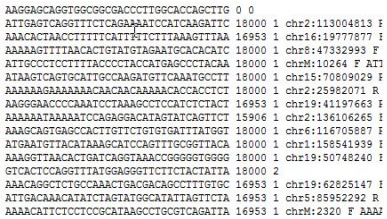
Attach DNA to surface, amplify in place





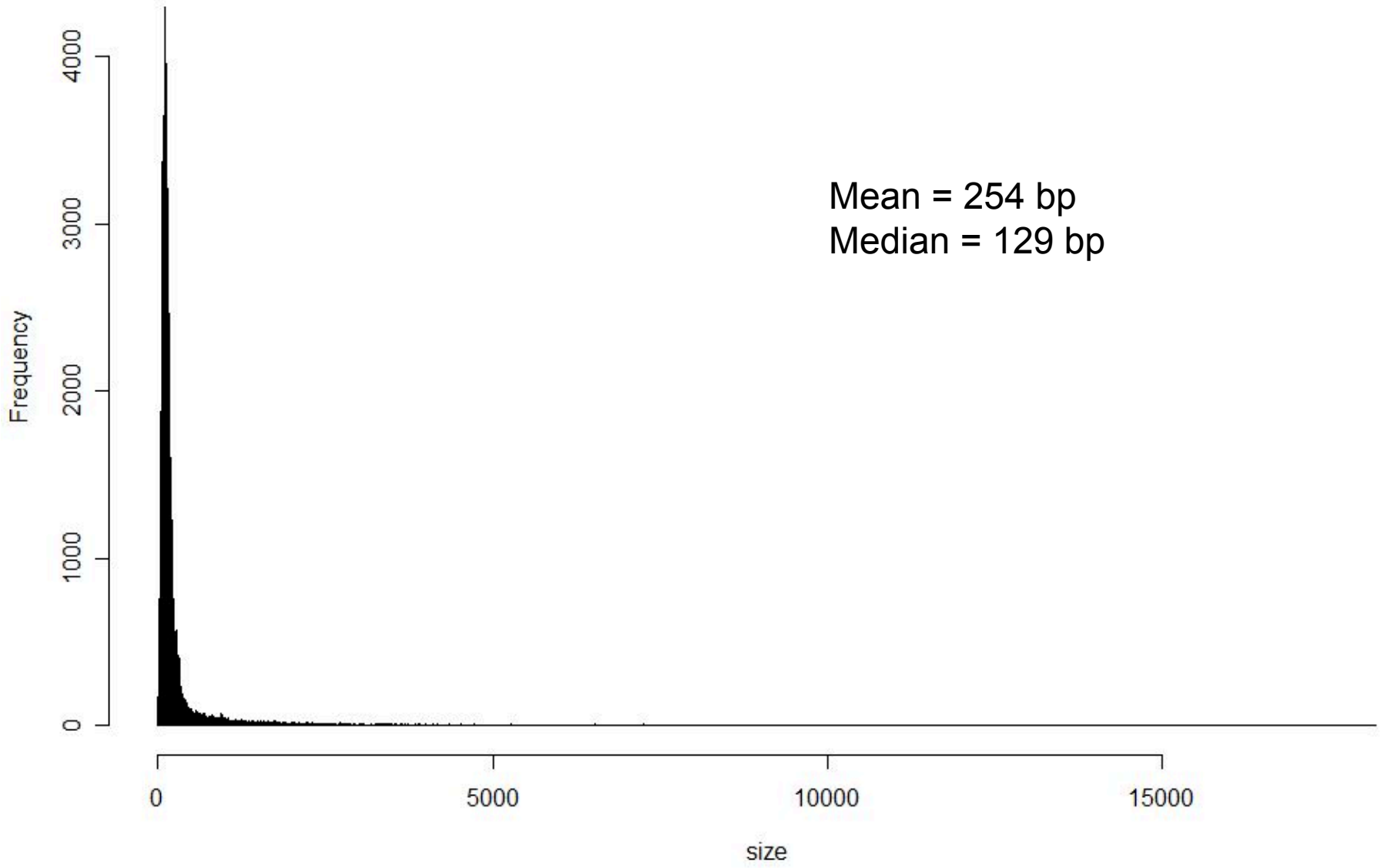
1. Add terminators, primers, polymerase.
2. Take picture.
3. Remove 3' terminus.
4. Repeat.



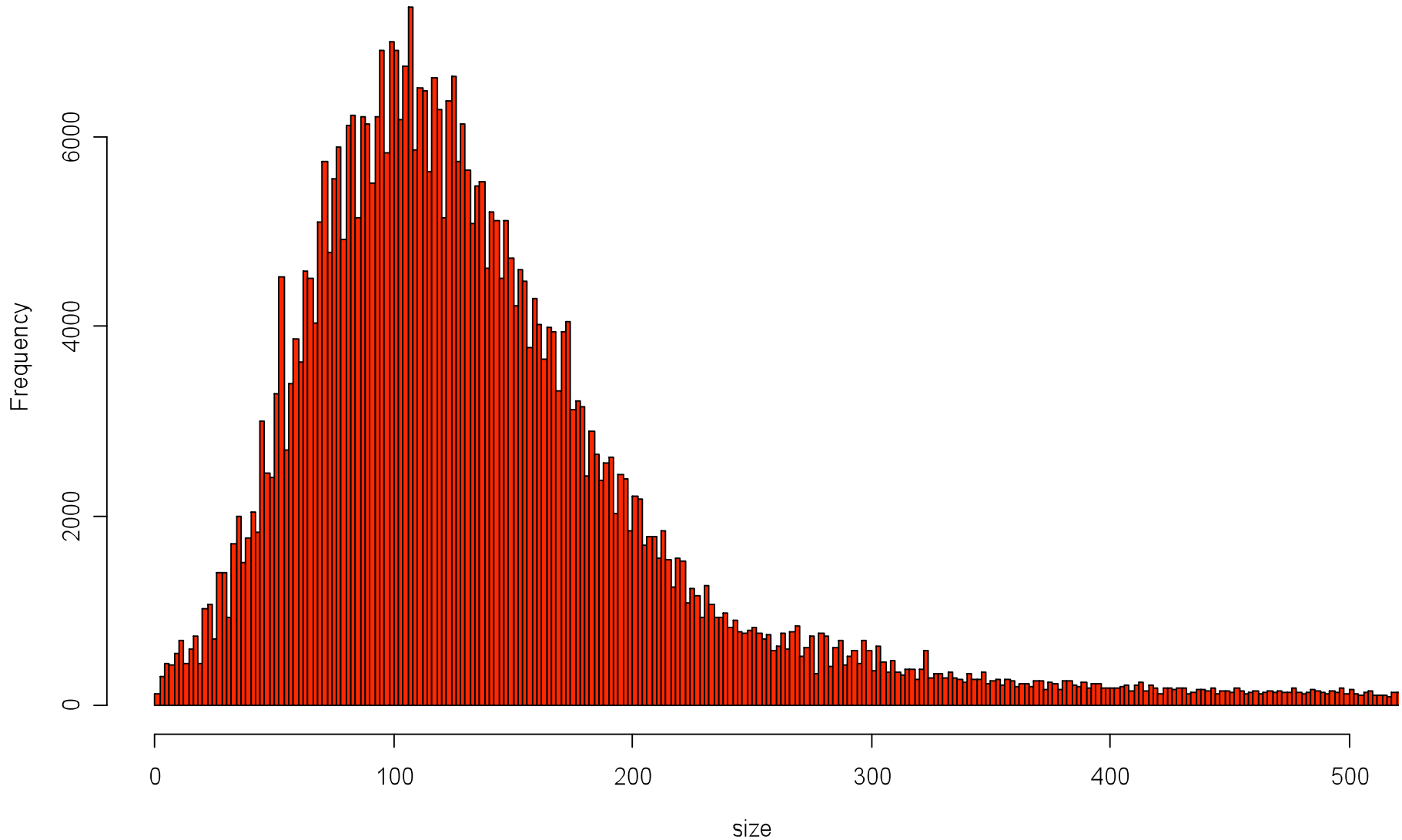


Millions of reads…
Where do we put them?

Exons of the Human Genome (n=500k)

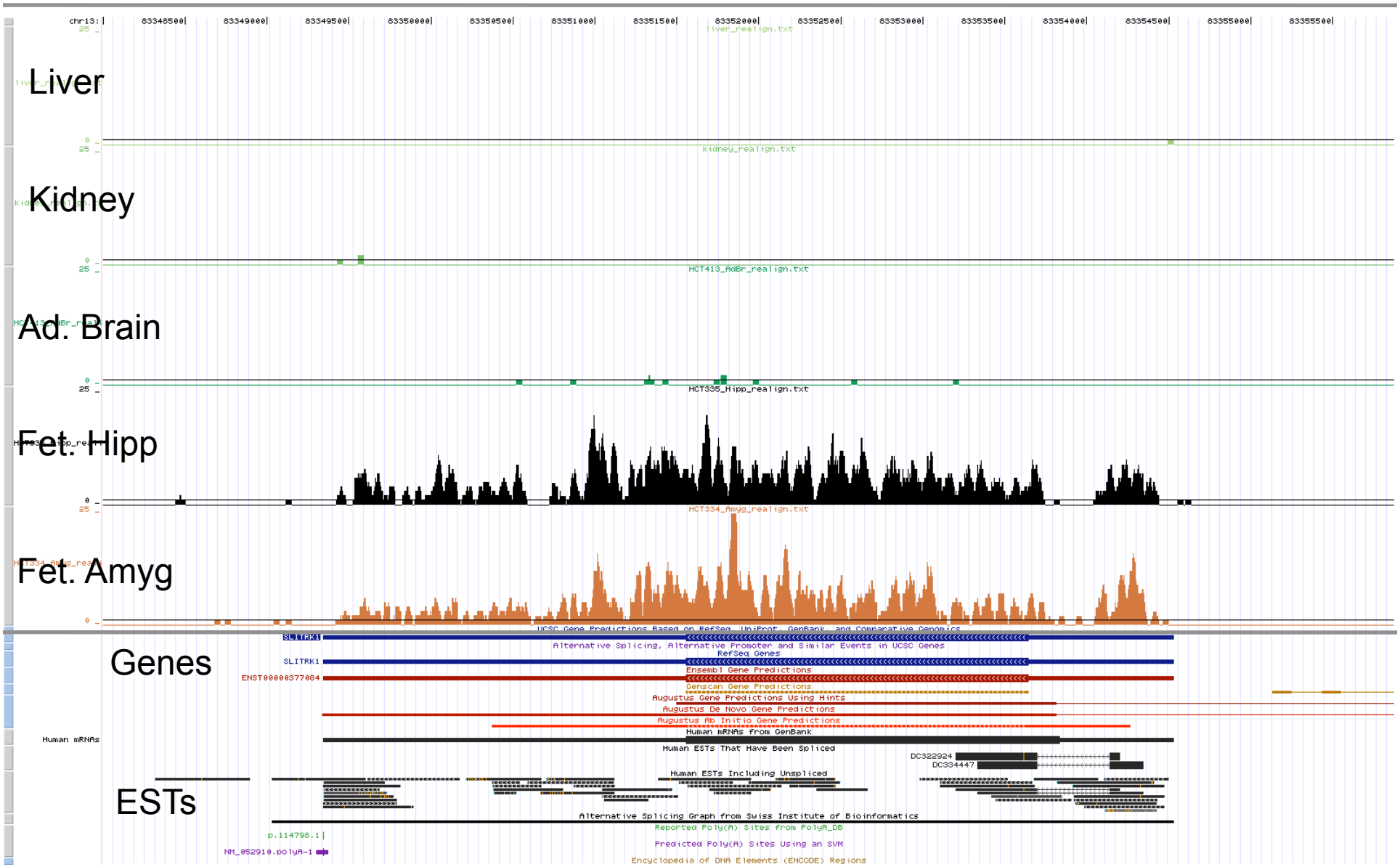


Exons of the Human Genome (ENSEMBL v54)

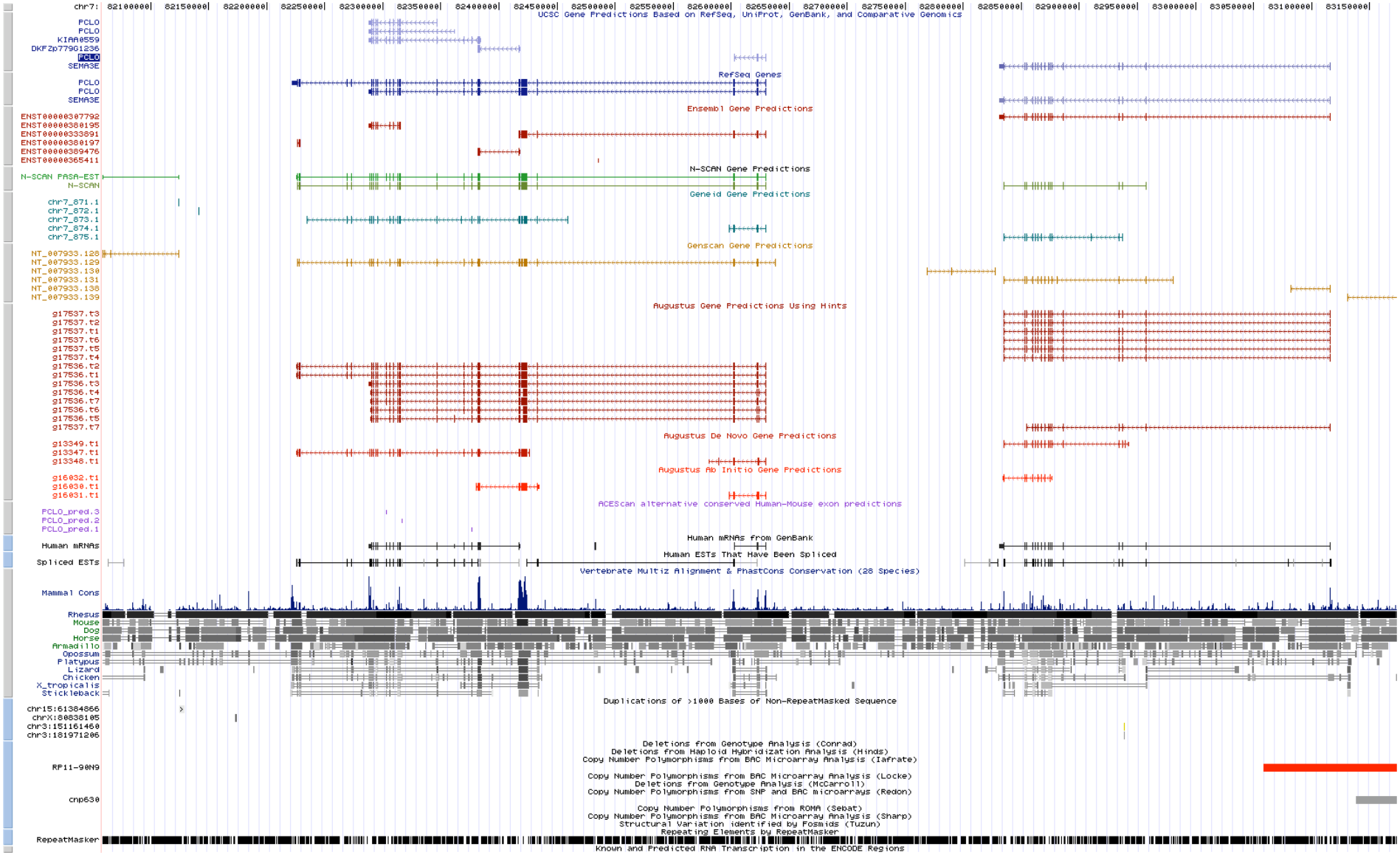


Average Intron=5,450

What is the source of the wiggles?



Gene Annotation is One Source



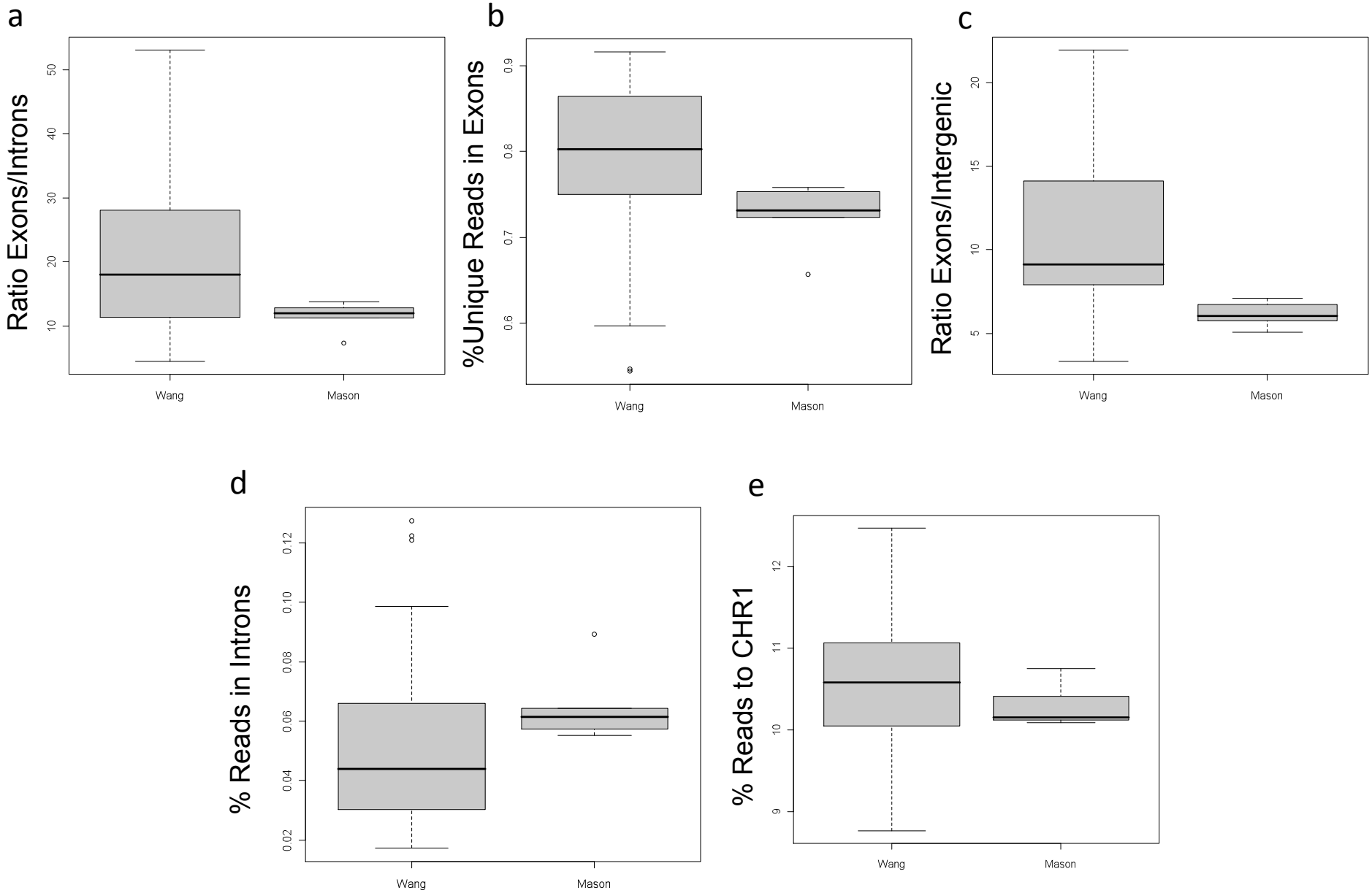
There are a three-fold range of options on gene annotation:

How many exonic Mb of the 3Gb genome?

	Aceview	UCSC	Vega	ENSEMBL	Refseq
Aceview	178				
UCSC	76	81			
Vega	51	42	58		
ENSEMBL	64	60	43	70	
RefSeq	60	61	37	57	62

Mb of exonic regions, discarding strand information.
Have we reached the limit?

QC Parameters have a wide range

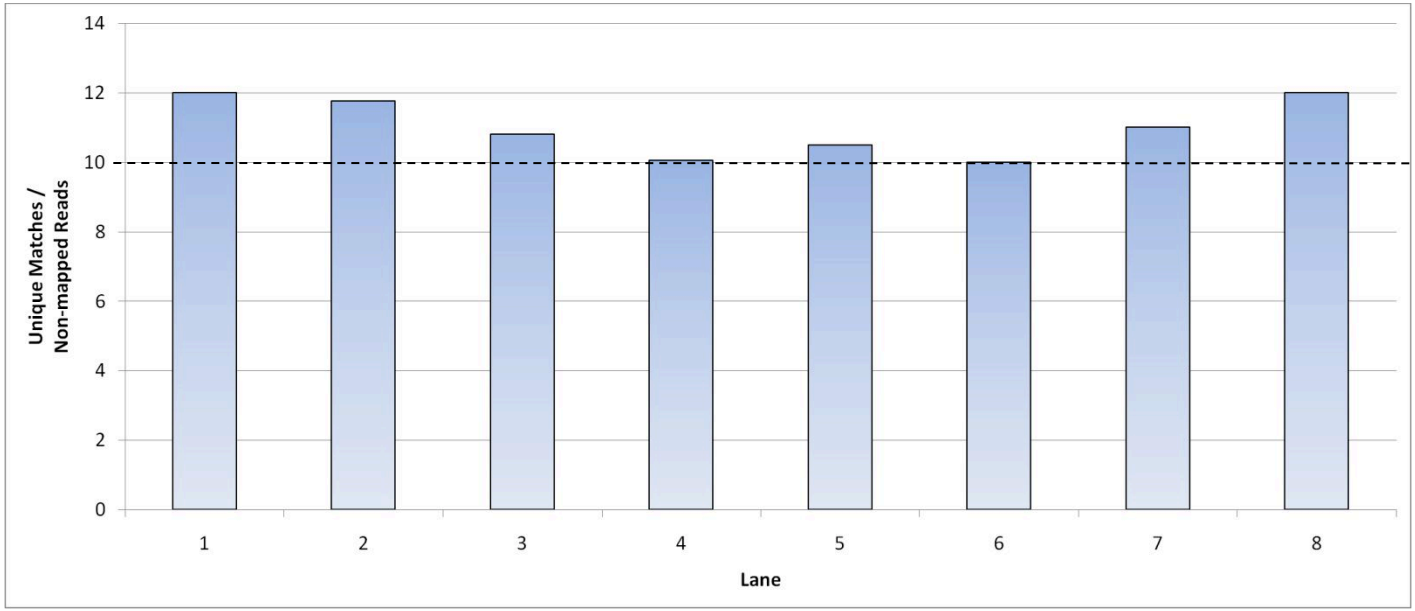


Each lane is not the same

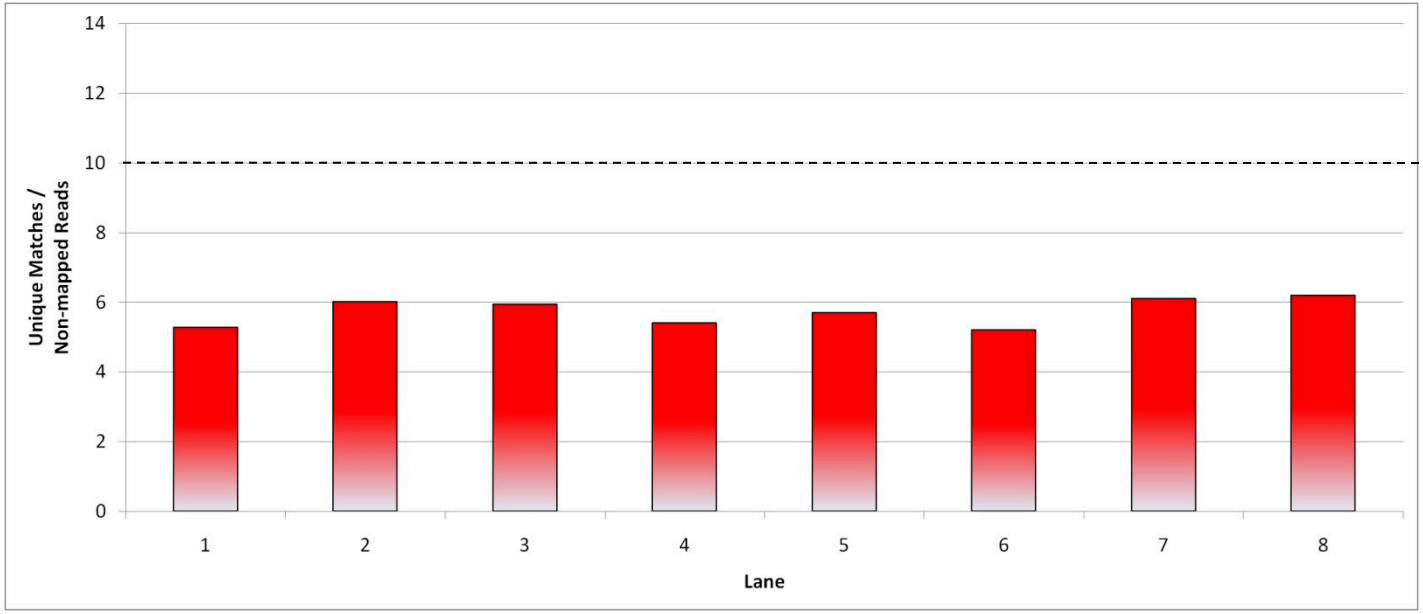
Lane Info		Phasing Info		Raw Data (tile mean)		Filtered Data (tile mean)						
Lane	Clusters (tile mean) (raw)	% Phasing	% Prephasing	% Error Rate (raw)	Equiv Perfect Clusters (raw)	% retained	Cycle 2-4 Av Int (PF)	Cycle 2-10 Av % Loss (PF)	Cycle 10-20 Av % Loss (PF)	% Align (PF)	% Error Rate (PF)	Equiv Perfect Clusters (PF)
1	27703	0.9100	0.4900	3.09	11530	45.18	920 +/- 172	-0.60 +/- 0.30	0.28 +/- 0.24	69.05	2.24	8085
2	36843	0.9100	0.4900	2.73	14834	42.22	1015 +/- 180	-0.65 +/- 0.23	0.27 +/- 0.29	69.91	1.77	10236
3	28518	0.9100	0.4900	3.74	9500	32.85	788 +/- 176	-1.13 +/- 0.38	0.41 +/- 0.31	64.83	2.72	5781
4	15138	0.9100	0.4900	3.17	7504	41.60	775 +/- 185	-1.88 +/- 2.07	0.07 +/- 0.28	85.71	2.46	5318
5	29043	0.9100	0.4900	3.77	9437	30.28	781 +/- 150	-1.32 +/- 0.39	0.43 +/- 0.32	65.23	2.67	5506
6	35408	0.9100	0.4900	2.24	15242	45.72	1029 +/- 210	-0.98 +/- 0.22	0.31 +/- 0.33	71.73	1.40	11006
7	23658	0.9100	0.4900	1.93	12087	58.22	1116 +/- 196	-0.91 +/- 0.27	0.42 +/- 0.35	73.82	1.36	9673
8	14177	0.9100	0.4900	7.78	207	62.43	1068 +/- 213	-0.67 +/- 0.20	0.41 +/- 0.34	2.05	7.79	141

RNA-Seq QC Can Be Seen in Read Distributions

Good Flowcell:
Error <0.5%
Phasing < 1



Bad Flowcell:
Error > 2%
Phasing 1.5



Metric for RNA-Seq Expression

RPKM:

Reads per Kilobase per Million Reads

Normalizes for (1) gene size and (2) sequencing depth

(~0.1-1 transcript/cell)

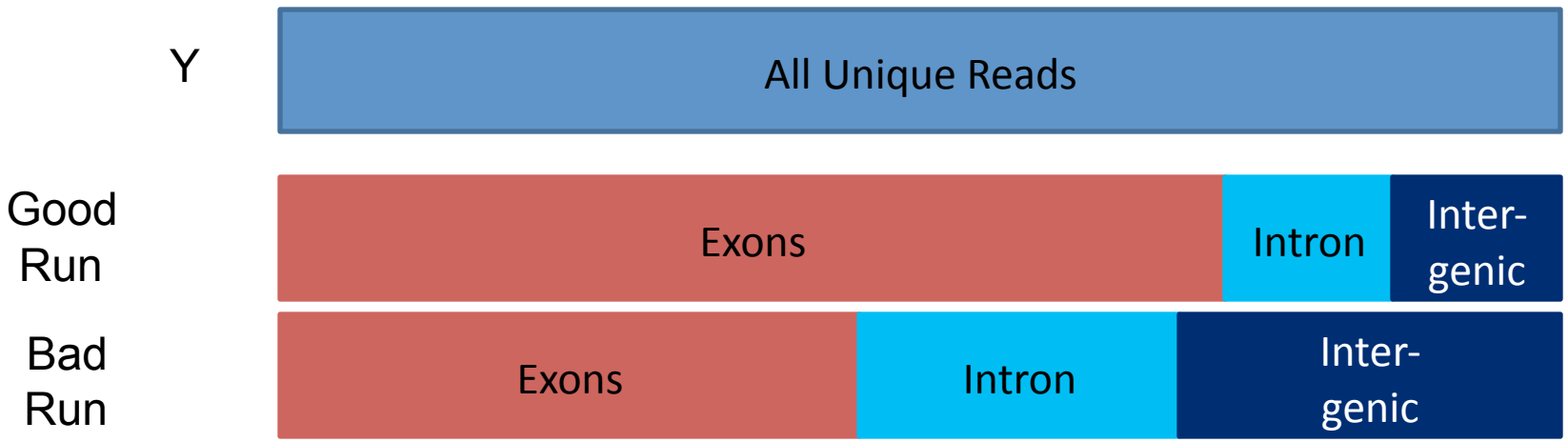
$$\text{RPKM} = \frac{N\ reads}{1\ gene} \times \frac{1\ \text{gene}}{B\ bp} \times \frac{1000\ \text{bp}}{1Kb} \times \frac{1\ Million\ reads}{Y\ total\ reads}$$

Y = (exons, introns, intergenic reads)

FPKM=fragments-PKM
is for paired-end data

Mortazavi, Williams, *et al.*
Nature Methods, 2008

Bias is introduced if these ratios are not kept:



Good Run RPKM = 128.4

Bad Run RPKM = 72.7

Mortazavi, Williams, *et al.* *Nature Methods*, 2008

Coverage Requirements: How many lanes/plates/wells?

<u>Depends on:</u>

1. Read Length
2. Size of Transcriptome
3. Complexity of Transcriptome
4. Complexity of Tissue
5. Biological Variance
6. Errors (random and systematic)

Plateau of Information Starts @ ~500Mb

Number of lanes compared	Differentially expressed genes	Overlap with genes called from the array	Correlation of fold changes between Solexa and the array
One vs One	5670	4208	0.67
Two vs Two	7994	5340	0.70
Three vs Three	9482	5909	0.71
Four vs Four	10580	6278	0.72
Five vs Five	11493	6534	0.73

Liver

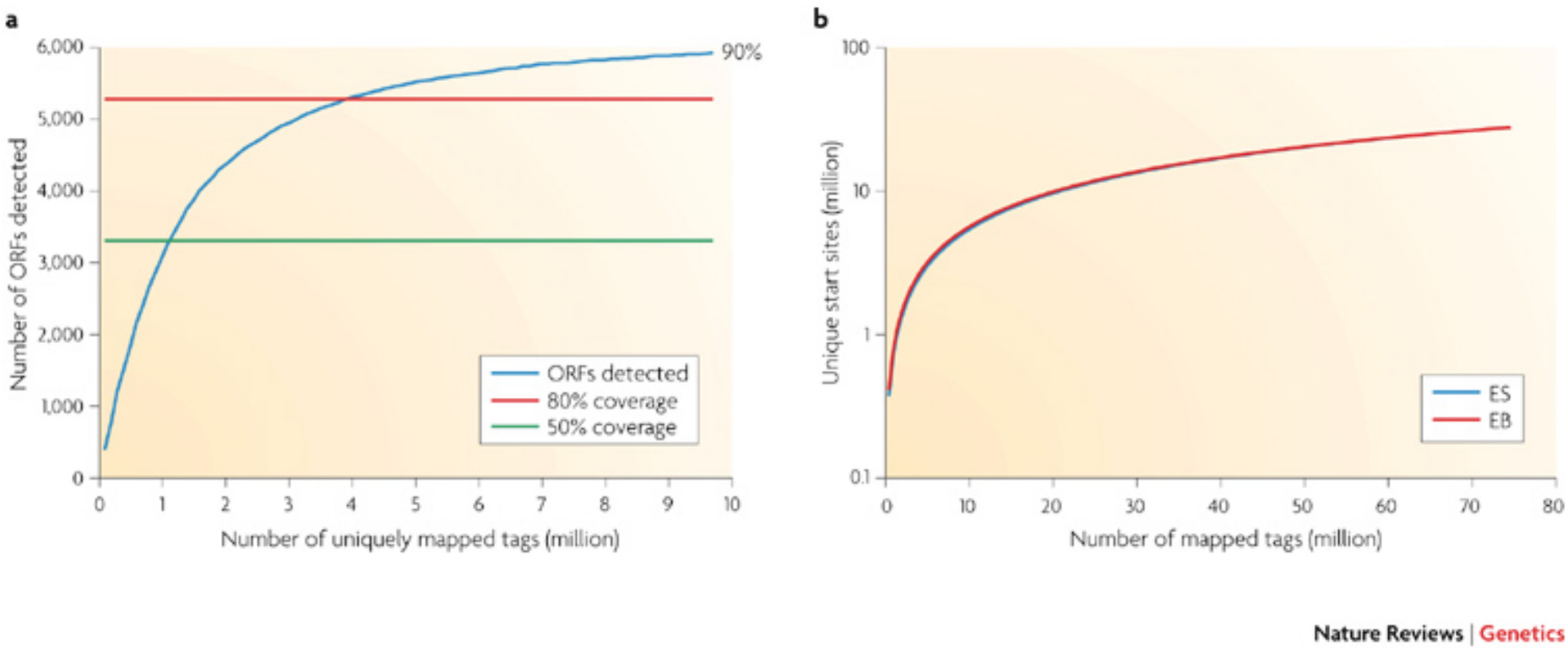
	No genes	Percentage
Five Lanes	20080	100
Four Lanes	19695	97.9
Three Lanes	19170	95.5
Two Lanes	18390	91.6
One Lane	16973	84.5

Kidney

	No genes	Percentage
Five Lanes	20921	100
Four Lanes	20552	98.2
Three Lanes	20064	96.0
Two Lanes	19355	92.5
One Lane	18080	86.4

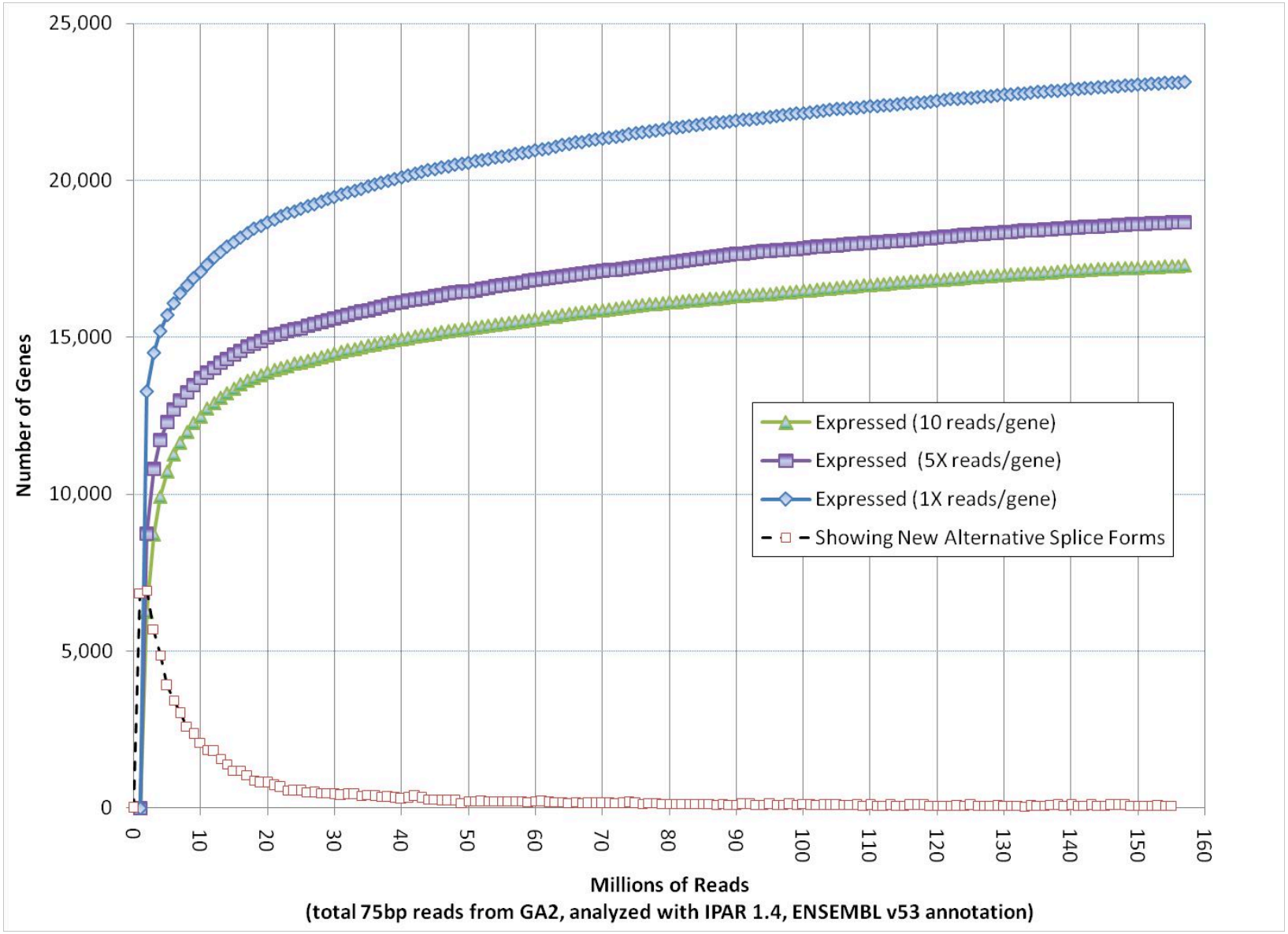
Marioni,Mason et al, Genome Research, 2008

Coverage Requirements

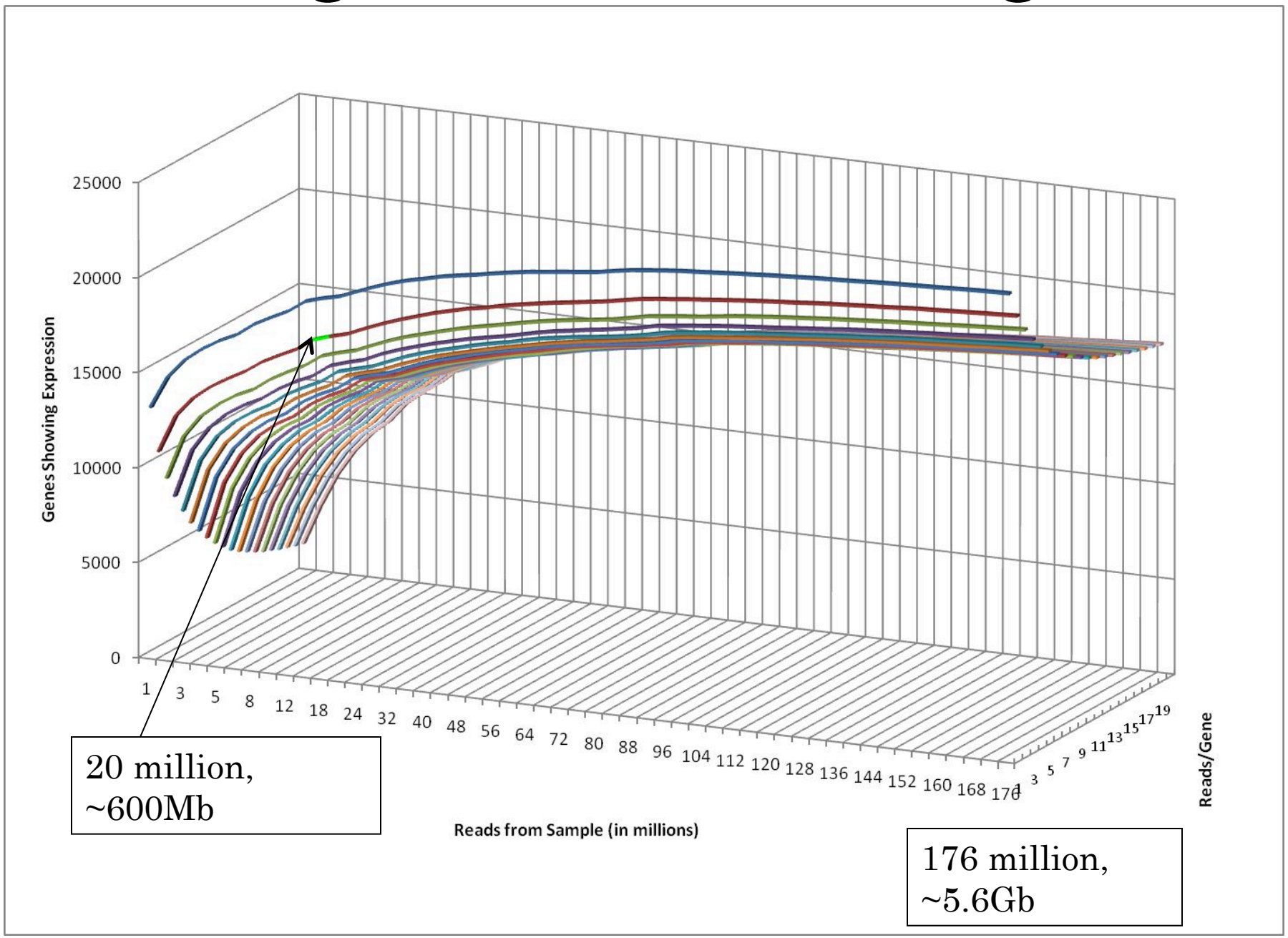


Wang, Gerstein, and Snyder, 2008

No current visible end of gene discovery



Regardless of coverage



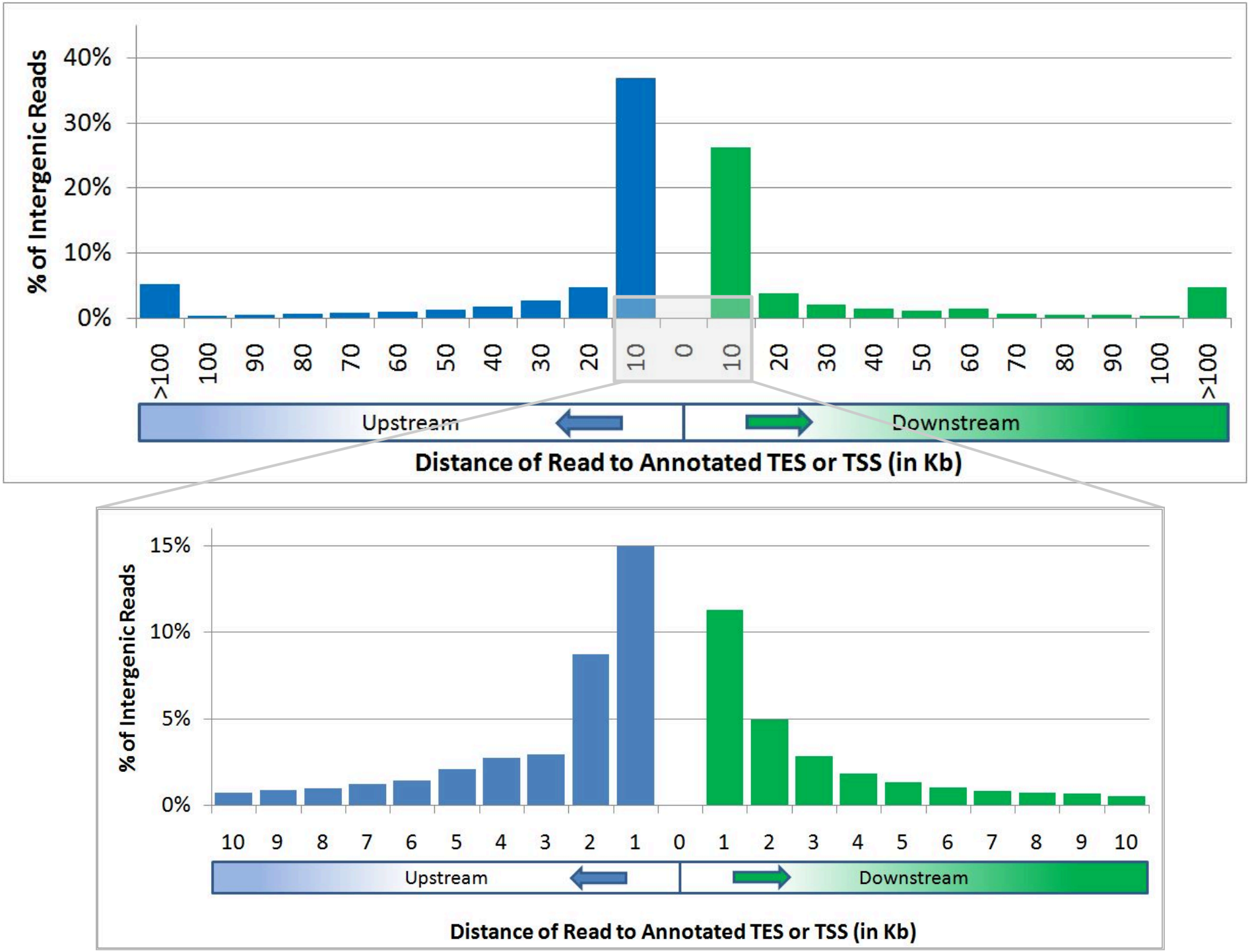
Summary of Example Illumina RNA-Seq Data

- 40% of reads mapped uniquely to a genomic location
- Of these, 65% mapped to autosomal or sex chromosomes

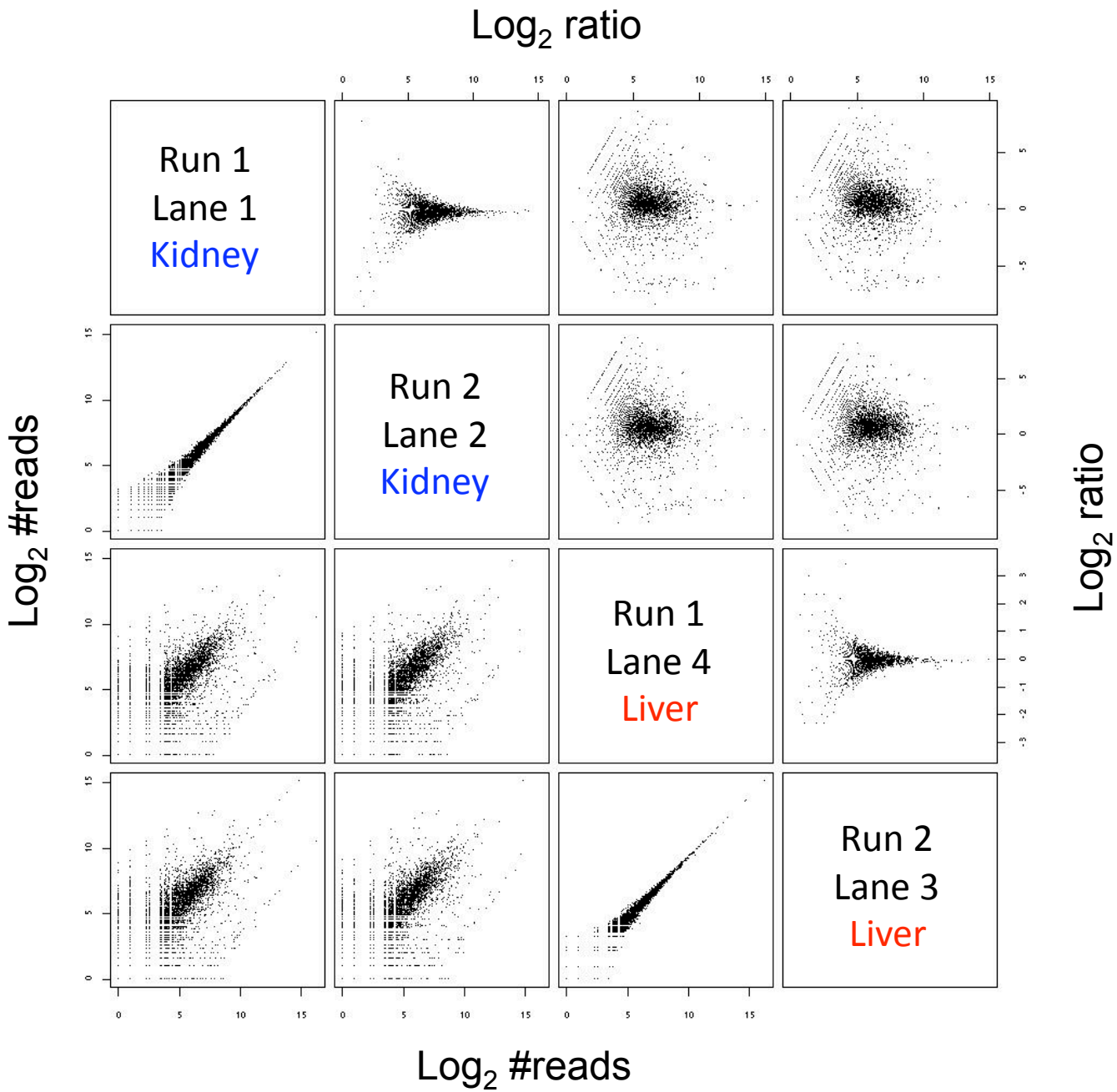
	Lane 1	Lane 2	Lane 3	Lane 4	Lane 6	Lane 7	Lane 8
	Solexa Run 1						
	kidney	liver	kidney	liver	liver	kidney	liver
Concentration (pM)	3	3	3	3	3	3	3
# Reads	13,017,169	14,003,322	13,401,343	14,230,879	13,525,355	12,848,201	13,096,715
Total Sequence (Mb)	417	448	429	455	433	411	419
# Mapped Reads	5,025,044	5,142,214	5,199,295	5,167,290	4,997,324	4,901,266	4,822,319
Mapped to chr1-22,X,Y	3,261,380	3,460,175	3,369,521	3,480,325	3,363,455	3,179,248	3,249,417
Mapped in Genes	2,706,150	2,847,704	2,792,026	2,861,877	2,761,468	2,630,987	2,668,148
Mapped in Exons	1,926,217	1,815,816	1,981,182	1,821,860	1,752,042	1,861,126	1,692,041
	Solexa Run 2						
	liver	kidney	liver	kidney	kidney	liver	kidney
Concentration (pM)	1.5	3	3	1.5	3	1.5	1.5
# Reads	9,096,595	13,687,929	14,761,931	8,843,158	13,449,864	9,341,101	8,449,276
Total Sequence (Mb)	291	438	472	283	430	299	270
# Mapped Reads	4,138,533	5,293,547	5,320,141	4,394,988	5,422,895	4,437,111	4,266,893
Mapped to chr1-22,X,Y	2,794,909	3,456,114	3,591,760	2,885,222	3,533,100	2,989,819	2,799,046
Mapped in Genes	2,328,896	2,875,214	2,959,436	2,416,834	2,938,079	2,488,832	2,345,160
Mapped in Exons	1,532,142	2,055,876	1,896,001	1,751,854	2,096,458	1,634,684	1,701,056

- Of these, 83% were located in genic regions
- Of those outside…

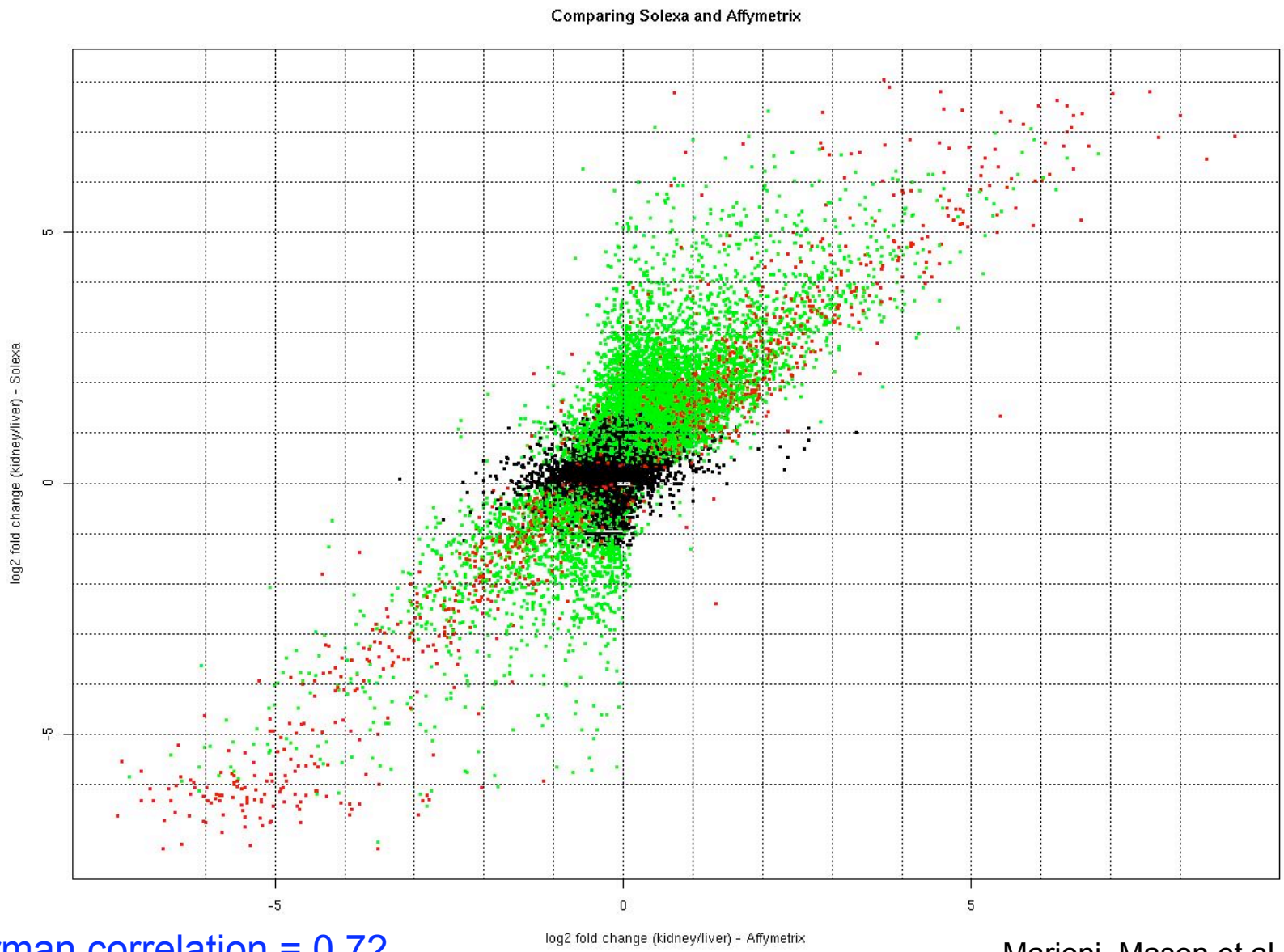
Marioni and Mason et al, 2008


% of Intergenic Reads
40%
30%
20%
10%
0%
>100
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
70
80
90
100
>100
Upstream
Downstream
Distance of Read to Annotated TES or TSS (in Kb)
% of Intergenic Reads
15%
10%
5%
0%
10 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 10
Upstream
Downstream
Distance of Read to Annotated TES or TSS (in Kb)

Variance is highest at low coverage genes



Good correlation between arrays, but some “dead zones” exist



Spearman correlation = 0.72

Marioni, Mason et al, 2008

RNA-Seq Error Sources

<u>Depends on:</u>

1. Read Length
2. Aligner (seed-based, gapped)
3. Sample Prep
4. Quality Scores
5. All Errors from sequencing
6. Platform-specific errors
7. Number of Genes Expressed in Tissue

RNA-Seq Analysis Options

- **ERANGE: RPKM and known junction mapper**
  http://woldlab.caltech.edu/rnaseq
- **TopHat: Aligns your RNA-Seq reads, and creates view of the junctions, or aligns to a known set of junctions**
  http://tophat.cbcb.umd.edu
- **Cufflinks: Assembles transcripts, estimates DE, and differential regulation of genes**
  http://cufflinks.cbcb.umd.edu/downloads/cufflinks-0.8.1.tar.gz
- **G-Mo.R-SE: Constructs gene models from spliced alignments**
  http://www.genoscope.cns.fr/externe/gmorse
- **BEGAT**: Per base, Exon, Gene Allele, Transcript, PolyA, and Fusion
  http://physiology.med.cornell.edu/faculty/mason/lab/BEGAT.html
- **GSNAP**
  http://research-pub.gene.com/gmap/

RNA-Seq Analysis with TopHat/Cufflinks

First, download the GTF file from ENSEMBL:

] ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh37.57.gtf.gz

Unzip the file

] gunzip Homo_sapiens.GRCh37.57.gtf.gz

Get a copy of Bowtie, TopHat, Cufflinks, and Boost libraries:

] http://sourceforge.net/projects/bowtie-bio/files/bowtie/0.12.3/bowtie-0.12.3-linux-x86_64.zip/download

] http://tophat.cbcb.umd.edu/downloads/tophat-1.0.13.tar.gz

] http://cufflinks.cbcb.umd.edu/downloads/cufflinks-0.8.1.tar.gz

] http://sourceforge.net/projects/boost/files/boost/1.42.0/boost_1_42_0.tar.gz/download

RNA-Seq Alignment with TopHat

Download the BWT human genome for Bowtie

```
]ftp://ftp.cbcb.umd.edu/pub/data/bowtie_indexes/hg19.ebwt.zip
```

Unzip Bowtie, then assign it to your PATH

```
]export PATH=$PATH:/Users/chm2042/Desktop/bowtie-0.12.3/
```

Unzip, Untar, and install TopHat

```
]./configure
]make
]make install
```

Run Tophat

```
]./tophat-1.0.13/bin/tophat genomes/hg19 burge_liver.fq
```

Rename Output Folder

```
]mv tophat_out X_tophat_out
output:
    accepted_hits.sam – read alignments in SAM format
    junctions.bed – junctions for viewing in UCSC browser
    coverage.wig – a BEdGraph track for the UCSC browser
```

TopHat Is Expecting a Mammalian Genome

```
-v/--version
    -o/--output-dir                <string>    [ default: ./tophat_out ]
    -a/--min-anchor                <int>       [ default: 8            ]
    -m/--splice-mismatches         <0-2>       [ default: 0            ]
    -i/--min-intron                <int>       [ default: 50           ]
    -I/--max-intron                <int>       [ default: 500000       ]
    -g/--max-multihits             <int>       [ default: 40           ]
    -F/--min-isoform-fraction      <float>     [ default: 0.15         ]
    --solexa-quals
    --solexa1.3-quals                          (same as phred64-quals)
    --phred64-quals                            (same as solexa1.3-quals)
    -p/--num-threads               <int>       [ default: 1            ]
    -G/--GFF                       <filename>
    -j/--raw-juncs                 <filename>
    -r/--mate-inner-dist           <int>
    --mate-std-dev                 <int>       [ default: 20           ]
    --no-novel-juncs
    --no-gff-juncs
    --no-coverage-search
    --coverage-search
    --no-closure-search
    --closure-search
    --fill-gaps
    --microexon-search
    --butterfly-search
    --no-butterfly-search
    --keep-tmp
```

TopHat Options

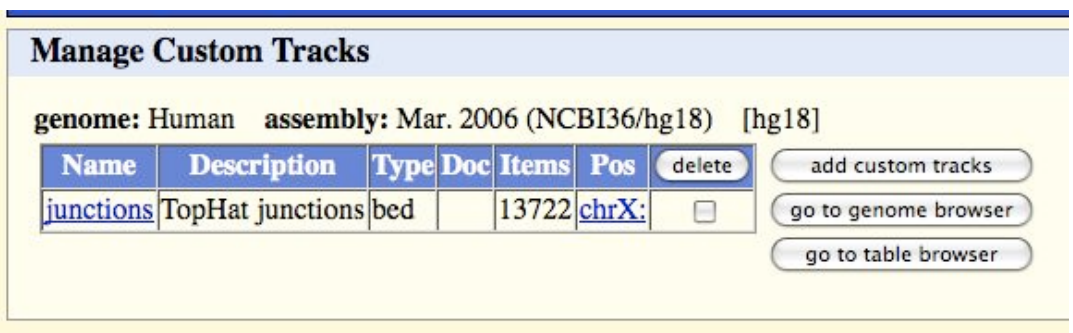
```
Advanced Options:

    --segment-mismatches           <int>       [ default: 2            ]
    --segment-length               <int>       [ default: 25           ]
    --min-closure-exon             <int>       [ default: 100          ]
    --min-closure-intron           <int>       [ default: 50           ]
    --max-closure-intron           <int>       [ default: 5000         ]
    --min-coverage-intron          <int>       [ default: 50           ]
    --max-coverage-intron          <int>       [ default: 20000        ]
    --min-segment-intron           <int>       [ default: 50           ]
    --max-segment-intron           <int>       [ default: 500000       ]

SAM Header Options (for embedding sequencing run metadata in output):
    --rg-id                        <string>    (read group ID)
    --rg-sample                    <string>    (sample ID)
    --rg-library                   <string>    (library ID)
    --rg-description               <string>    (descriptive string, no tabs allowed)
    --rg-platform-unit             <string>    (e.g Illumina lane ID)
    --rg-center                    <string>    (sequencing center name)
    --rg-date                      <string>    (ISO 8601 date of the sequencing run)
    --rg-platform                  <string>    (Sequencing platform descriptor)
```

Visualizing the Junctions

Go to the UCSC genome browser:

http://genome.ucsc.edu/cgi-bin/hgGateway

Upload the junctions.bed file as a custom track



Go to: chrX:43,625,858-43,741,721

RNA-Seq Processing with Cufflinks

If you cannot use the pre-compiled binaries, you will need to:
Unzip, Untar, and install Cufflinks with Boost Libraries

```
] ./bootstrap.sh
] ./bjam install --prefix=/Users/chm2032/Desktop/boost_1_42_0/boost

./configure --prefix=/home/ngsst19/cufflinks-0.8.1/ --with-boost=/home/
ngsst19/boost_1_42_0/booster/

] make

] make install
```

GFF file - General Feature Format

GFF is a format used to hold information about annotation of a genome.
It has nine REQUIRED fields that must be TAB-DELIMITED

1. Seqname: Chromosome or scaffold
2. Source: The program that generated this feature
3. Feature: The feature (exon, CDS, start_codon, stop_codon, etc.
4. Start: The 1-based starting position of the feature
5. End: The 1-based ending position of the feature
6. Score: Between 0-1000. If no score, stored as “.”
7. Strand: The Watson/Crick or +/- strand of the feature
8. Frame: If a coding exon, the reading frame needs to be 0-2 that is the frame of the first base. If feature is not coding, stored as “.”.
9. Group: All lines with the same group are linked as a single item

GTF file
Gene Transfer Format

GTF is a format used to hold information about gene structure, and the first eight fields are the same as GFF.

The ninth field (group) is expanded into attributes of at least one item. Each attribute is stored as a type/value pair.

Attributes must end in a semi-colon.

Attributes must be separated by one space.

There are two required attributes: gene_id and transcript_id.

Example: gene_id “BRCA1”; transcript_id “BRCA1.1”;

Cufflinks RNA-Seq Toolbox

Cufflinks -- assembles transcripts and estimates abundances

Sort the sam file by reference position (this step is NOT necessary if reads were aligned with tophat)

```
]sort -k 3,3 -k 4,4n hits.sam > hits.sam.sorted
```

Perfrom an *ab initio* construction

```
]./cufflinks --num-thread 4 hits.sam.sorted
```

Rename the output files

```
]mv genes.expr brain.expr
]mv transcripts.expr brain.expr
]mv transcripts.gtf brain.gtf
```

Constuction based on a known annotation

```
]./cufflinks --num-thread 4 --GTF aceview_acembly.gtf hits.sam.sorted 2>&1 cufflinks.log
```

Cuff-Compare and other functions

cuffcompare -- compares cufflinks ab initio assembly to a reference annotation or compares cufflinks transcript abundance estimates across multiple samples

Compare cufflink ab initio to a reference annotation:

```
]./bin/cuffcompare -o summary.stats -r Homo_sapiens.GRCh37.57.gtf -V transcripts.gtf
```

(output = summary.loci, summary.stats, transcripts.refmap, transcripts.tmap)

Compare multiple cufflinks transcript abundance estimates for N samples

```
]./cuffcompare -o stats.txt -V transcripts_liver.gtf transcripts_brain.gtf
```

(output = brain.refmap, brain.tmap, liver.refmap, liver.tmap, stats.loci, stats.tracking, stats.txt, stats.combined.gtf)

Compare multiple ab initio cufflinks transcript abundance estimates from different experimental conditions and to a reference annotation

```
]./bin/cuffcompare -o summary.stats -r Homo_sapiens.GRCh37.57.gtf -V transcripts.gtf brain.gtf liver.gtf
```

(output = brain.refmap, brain.tmap, liver.refmap, liver.tmap, summary.loci, summary.tracking, summary.stats, summary.combined.gtf)

Cuff-Diff

Cuffdiff -- finds 'significant' differences in transcript expression, splicing, and promoter use between two sets or a time-series

Compare two samples

```
]./cuffdiff summary.combined.gtf brain.sam liver.sam
```

Cufflinks Caveats

1. Does NOT output the no-hits
2. Space intensive
3. No support for PolyA sites or fusions (yet)