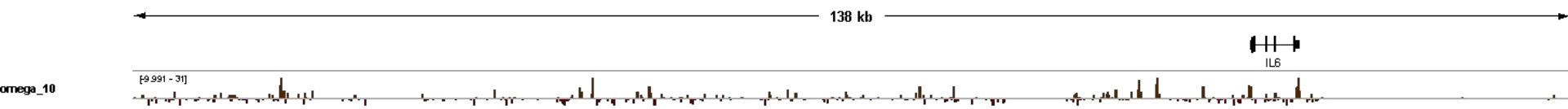
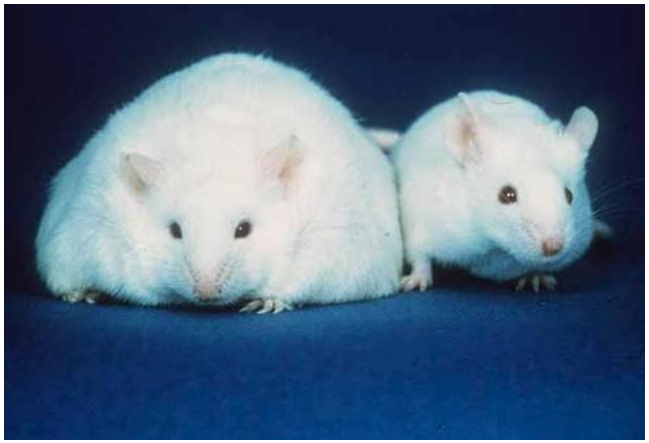
Introduction to read mapping and RNA-Seq


138 kb
IL6
[-9.991 - 31]
ornega_10

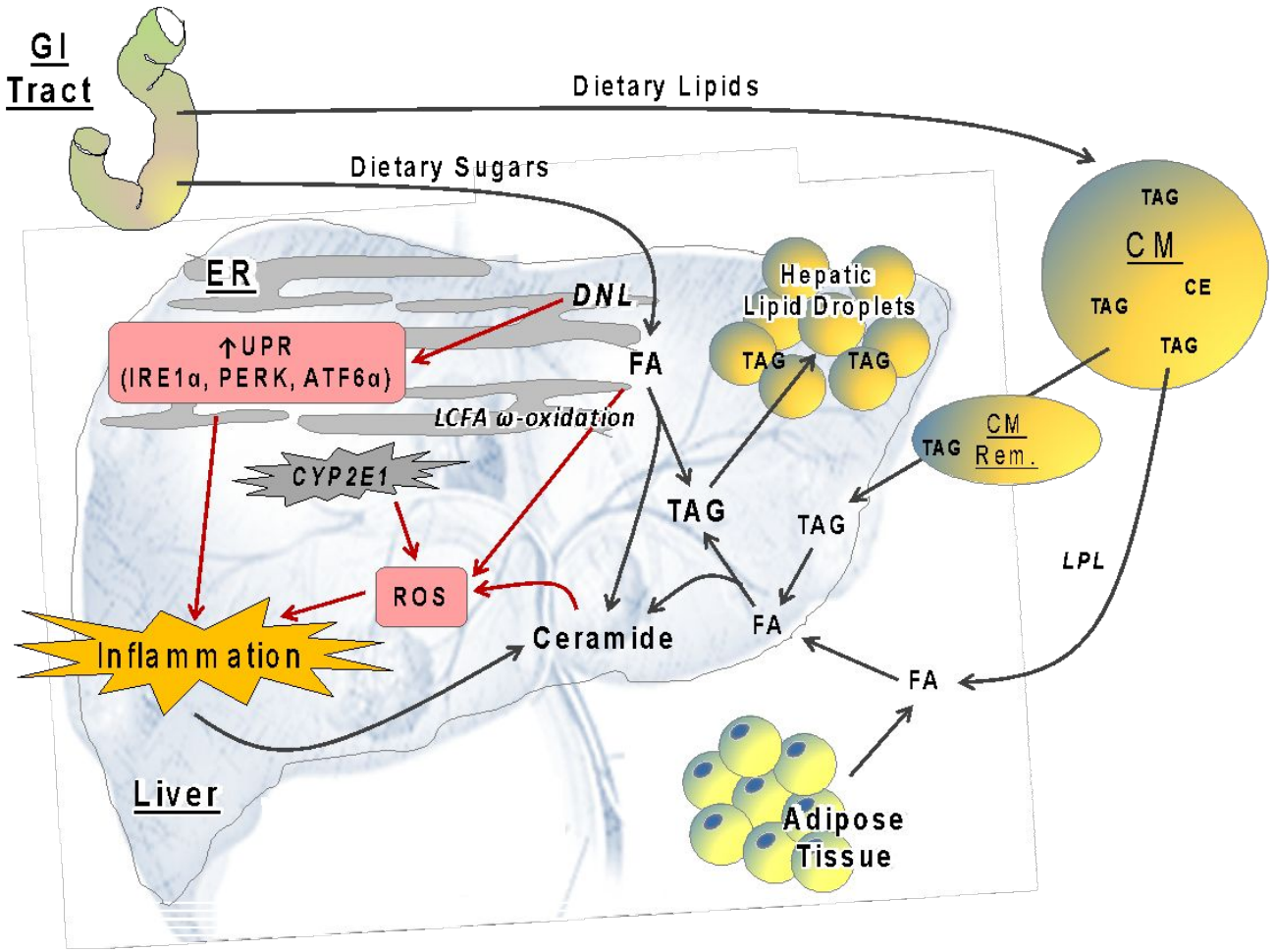
Gene expression as a readout of cell state

- Cells use signaling cascades to process information from their environment.
  - A typical cascade involves a extracellular receptor that triggers a transcriptional response
- What is the effect of the loss of function of a cellular component?
  - Map the function of a gene by measuring its impact on transcriptional output.

Sample dataset: Fat diet makes you fat…

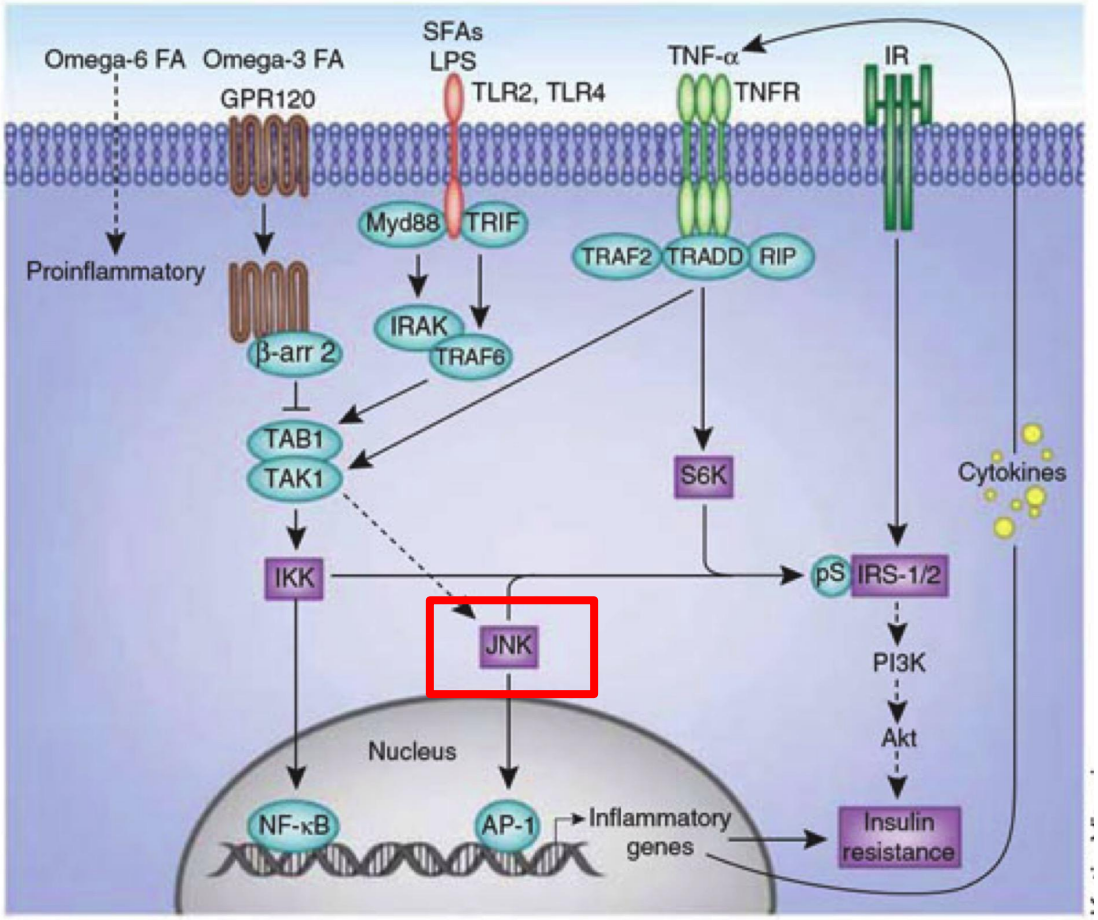


Why? Why not to everyone?



Ip, B and Wang X, Nutrients 2014

In particular JNK signaling is key



Osborne O. and Olefsky J, Nat. Medicine 2012

Experimental approach

$L^{\Delta 1}$ = Jnk1 KO

$L^{\Delta 2}$ = Jnk2 KO

$L^{\Delta 1,2}$ = Double KO

L^{WT} = Wildtype

Genotypes

Diets

$L^{\Delta 1}$		$L^{\Delta 2}$		$L^{\Delta 1,2}$		L^{WT}	
HF	CW	HF	CW	HF	CW	HF	CW
Rep1 Rep2 Rep3	Rep1 Rep2 Rep3	Rep1 Rep2 Rep3	Rep1 Rep2 Rep3	Rep1 Rep2 Rep3	Rep1 Rep2 Rep3	Rep1 Rep2 Rep3	Rep1 Rep2 Rep3

24 samples: 4 Genotypes x 2 Diets x 3 replicates

Vernia et al, Cell Matabolism 2014

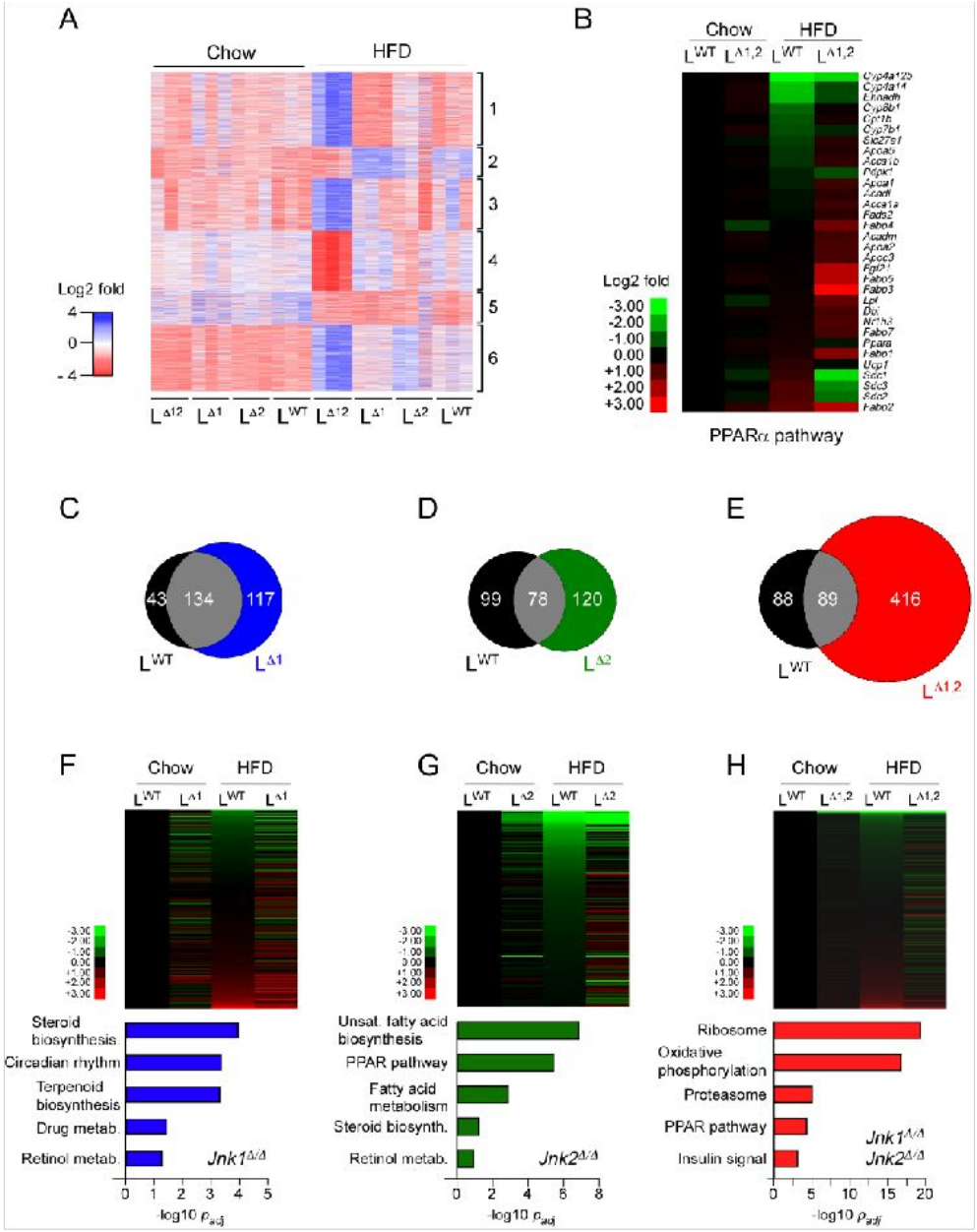
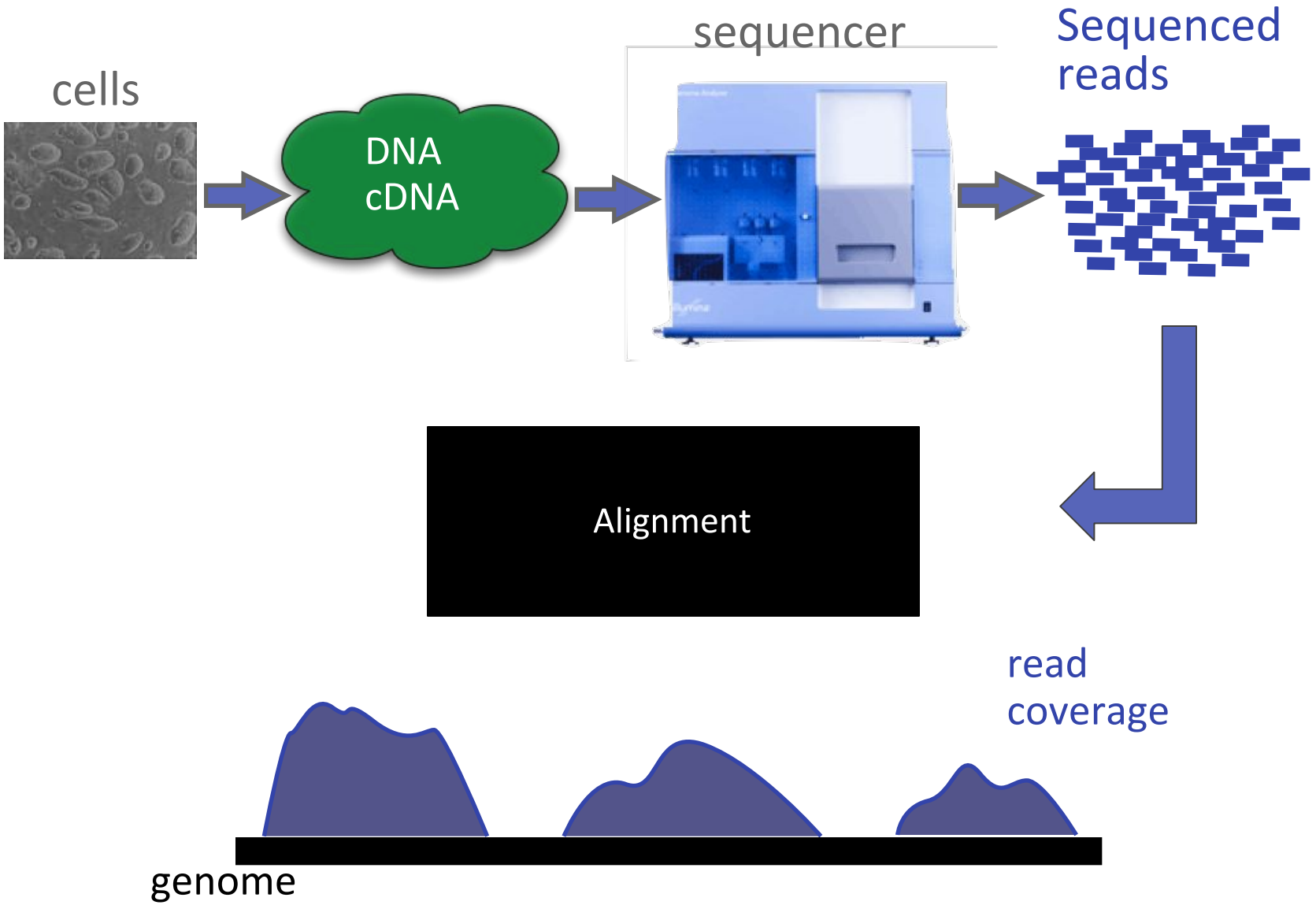


Figure S3. Analysis of hepatic genes differentially regulated by high fat diet in control and liver-specific JNK-deficient mice, Related to Figure 3.

How do we do this?

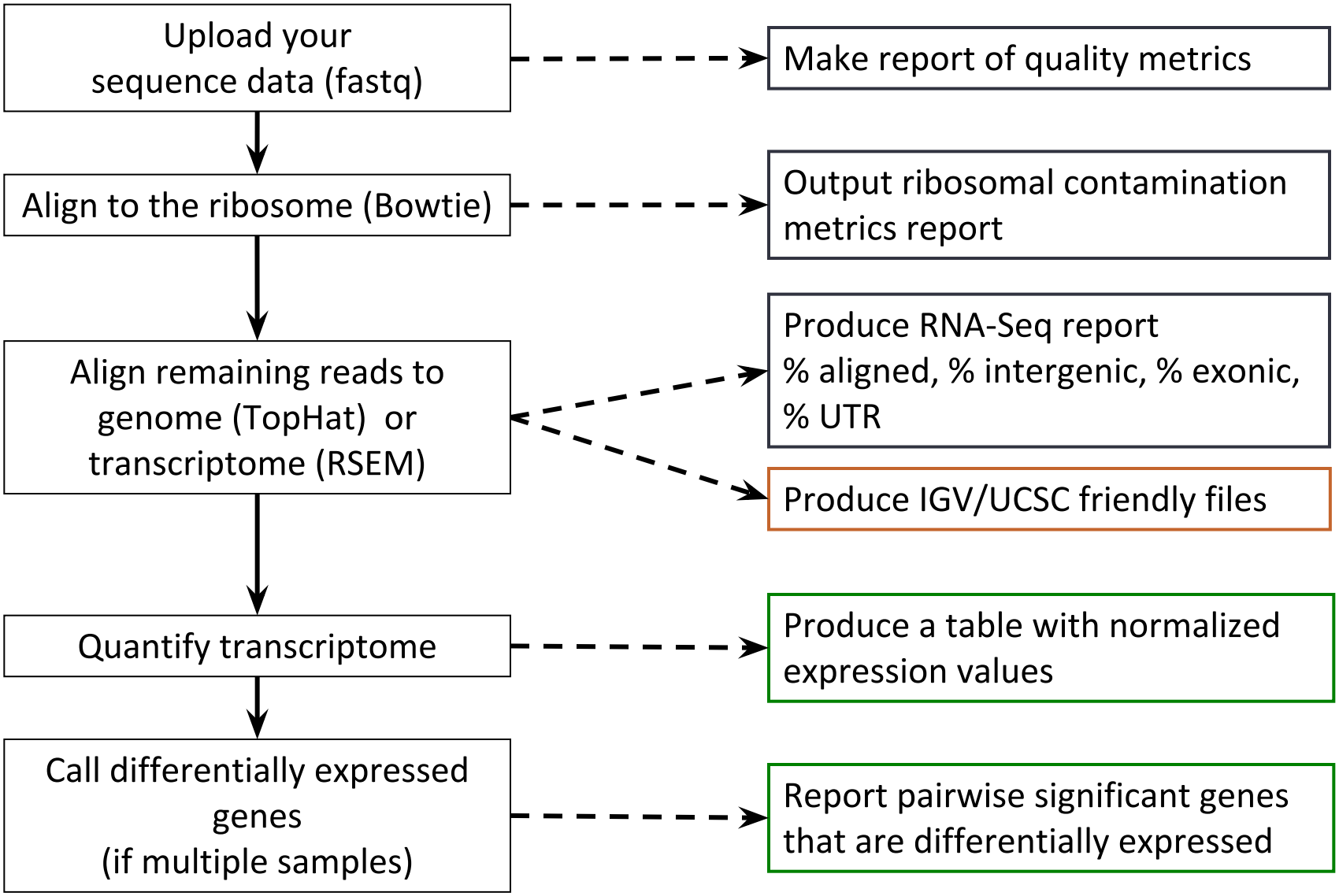
Measuring gene expression by sequencing



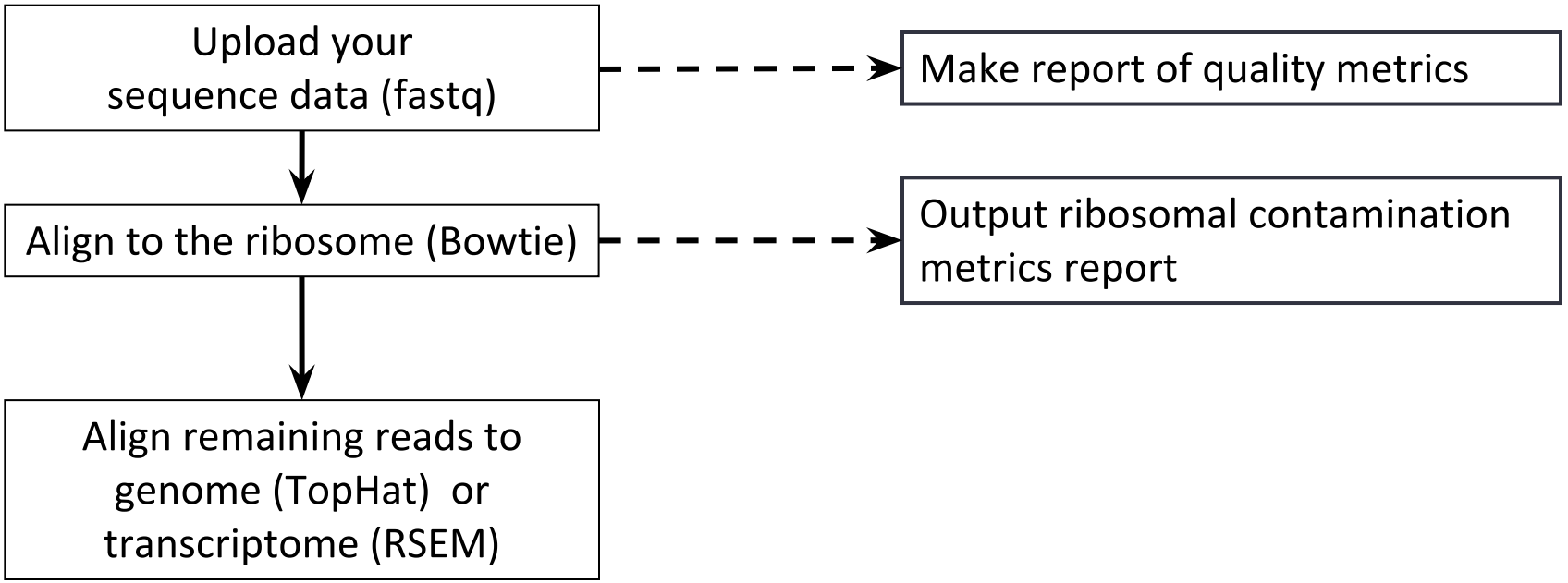
Let's pause for a moment and look at what we are doing

- Goal: quantifying nucleic acid species in a sample
- Collected nucleic acids – all mixed up
- Sheared it into roughly equally large pieces – all mixed up
- Added adapters so that we can read the pieces in the sequencer
- We sequence a small fraction (70 bases) of the pieces
- **Objective**: Given the 70 bases read from each piece, can we determine:
  - **what nucleic acids sequenced were there to begin with?**
  - **in what proportion?**
- Strategy: Try to find the place were the 70 base read originated by finding the best place were it matches the genome
- Difficulties:
  - The genome is large!
  - The genome is complicated – lots of *repetitive* sequence

Our typical RNA quantification pipeline



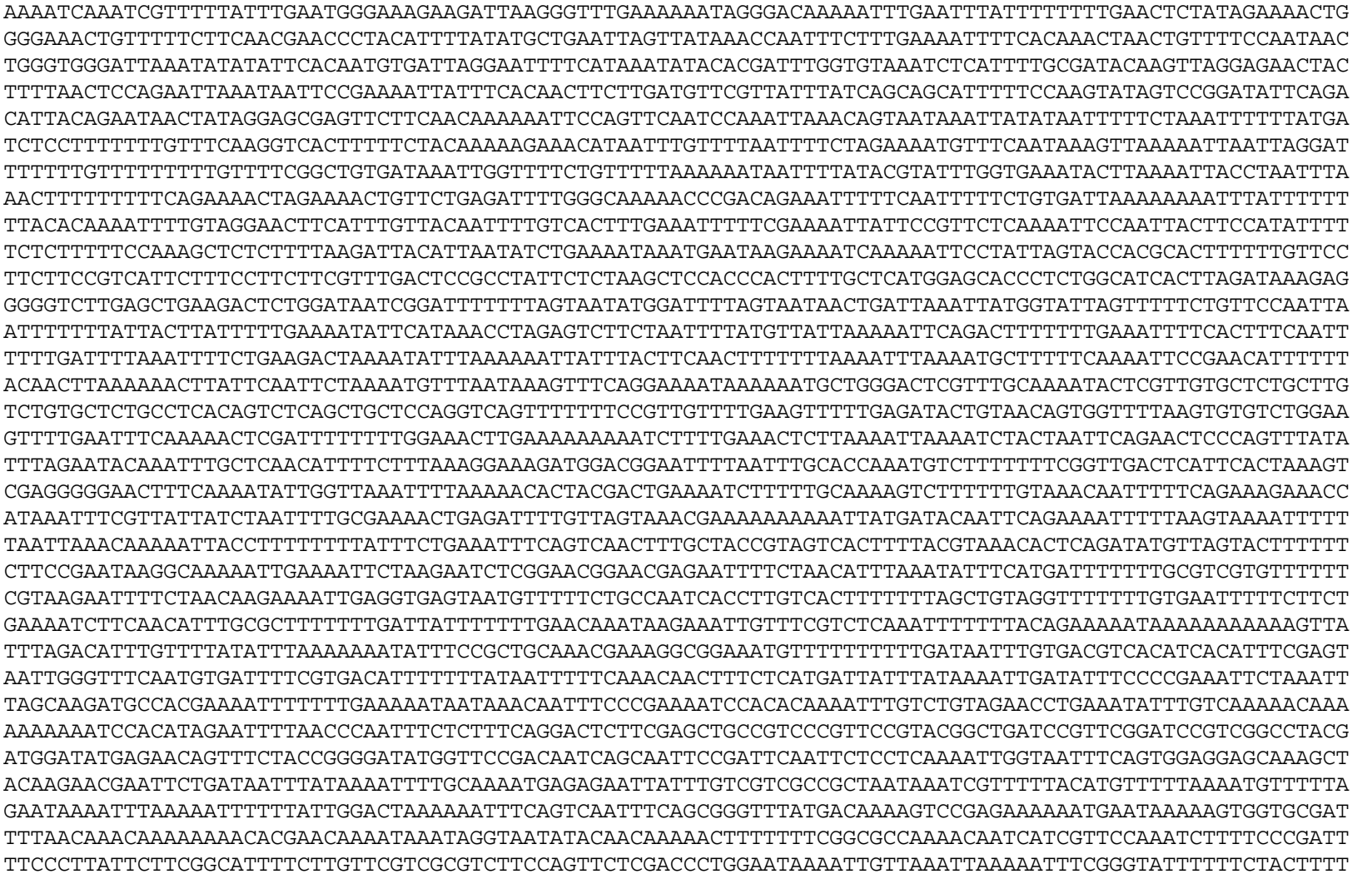
Alignment requires pre-processing



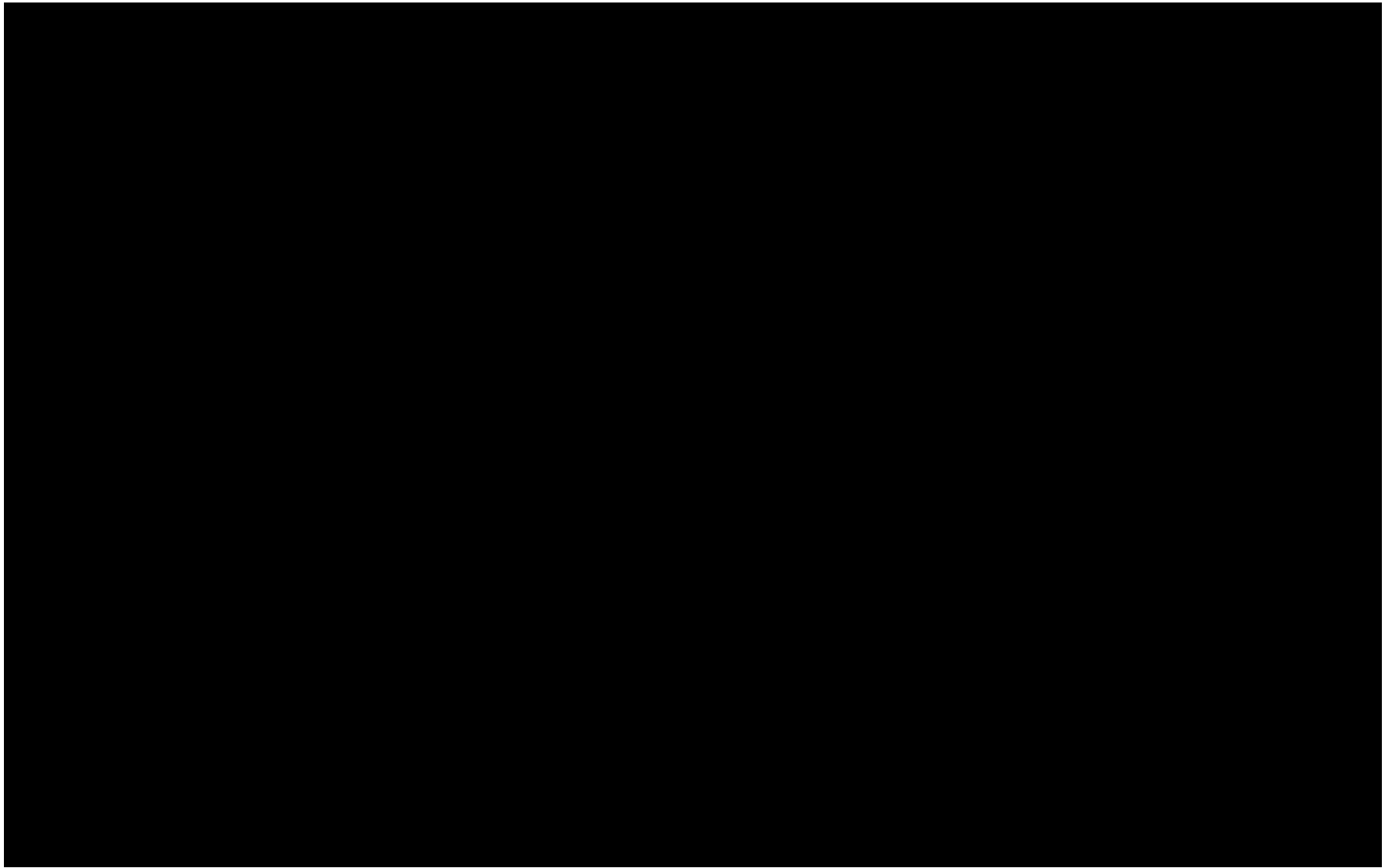

```
bowtie2-build -f mm10.fa mm10

rsem-prepare-reference \
--gtf ucsc.gtf --transcript-to-gene-map ucsc_into_genesymbol.rsem \
mm10.fa mm10.rsem
```

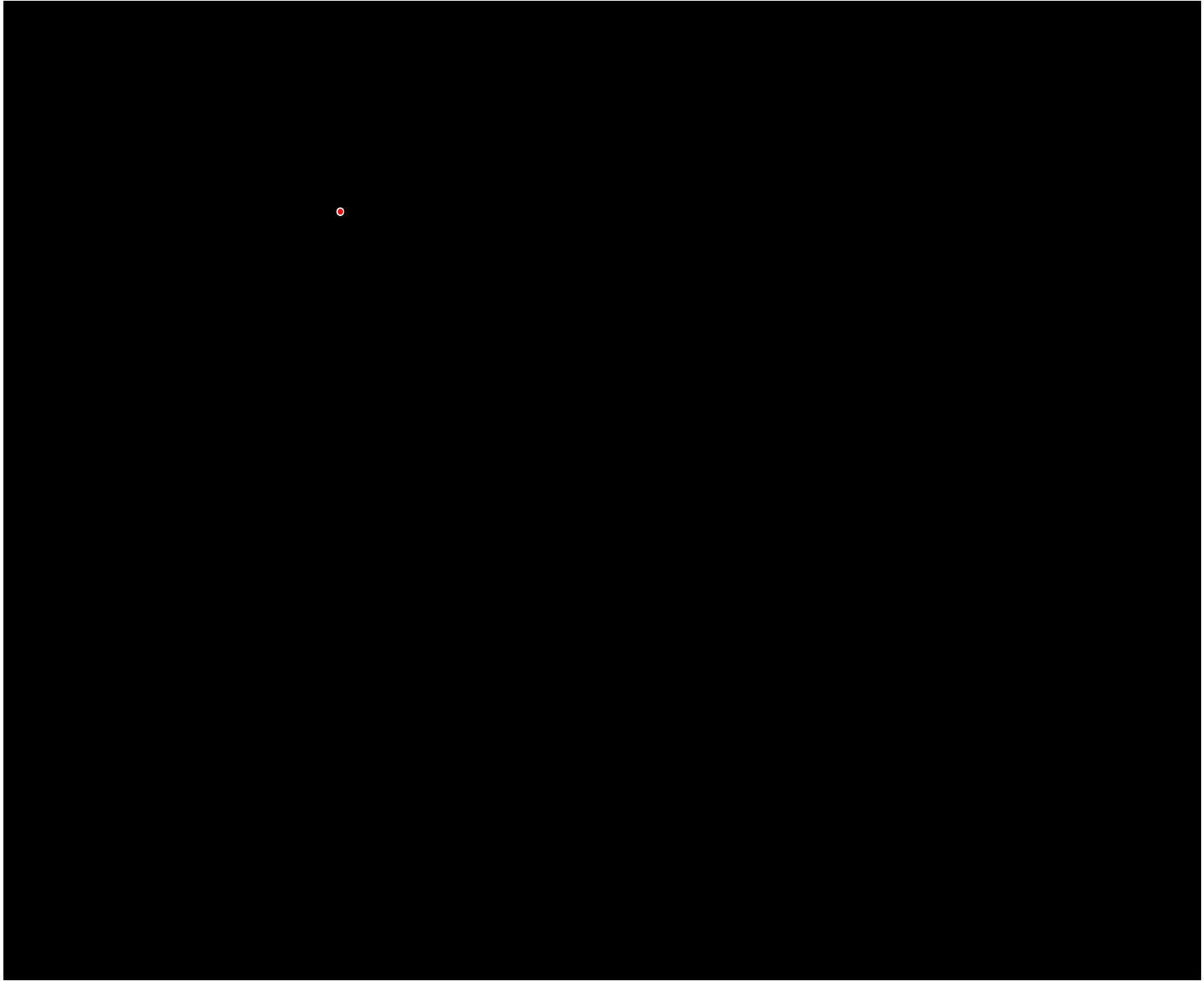
Why do we create an “index” of the genome?

ACTTGACCTACT**N**GGACCCT



Why do we create an “index” of the genome?





Hash-based mapping:

Step1: hash/index the genome

Toy genome
(16 bp)

CATGGTCATTGGTTCC

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4
GTC	5

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4
GTC	5
TCA	6

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3
GGT	4
GTC	5
TCA	6

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Complete hash/kmer index of our toy genome (forward strand only)

What is a good choice of hash size?

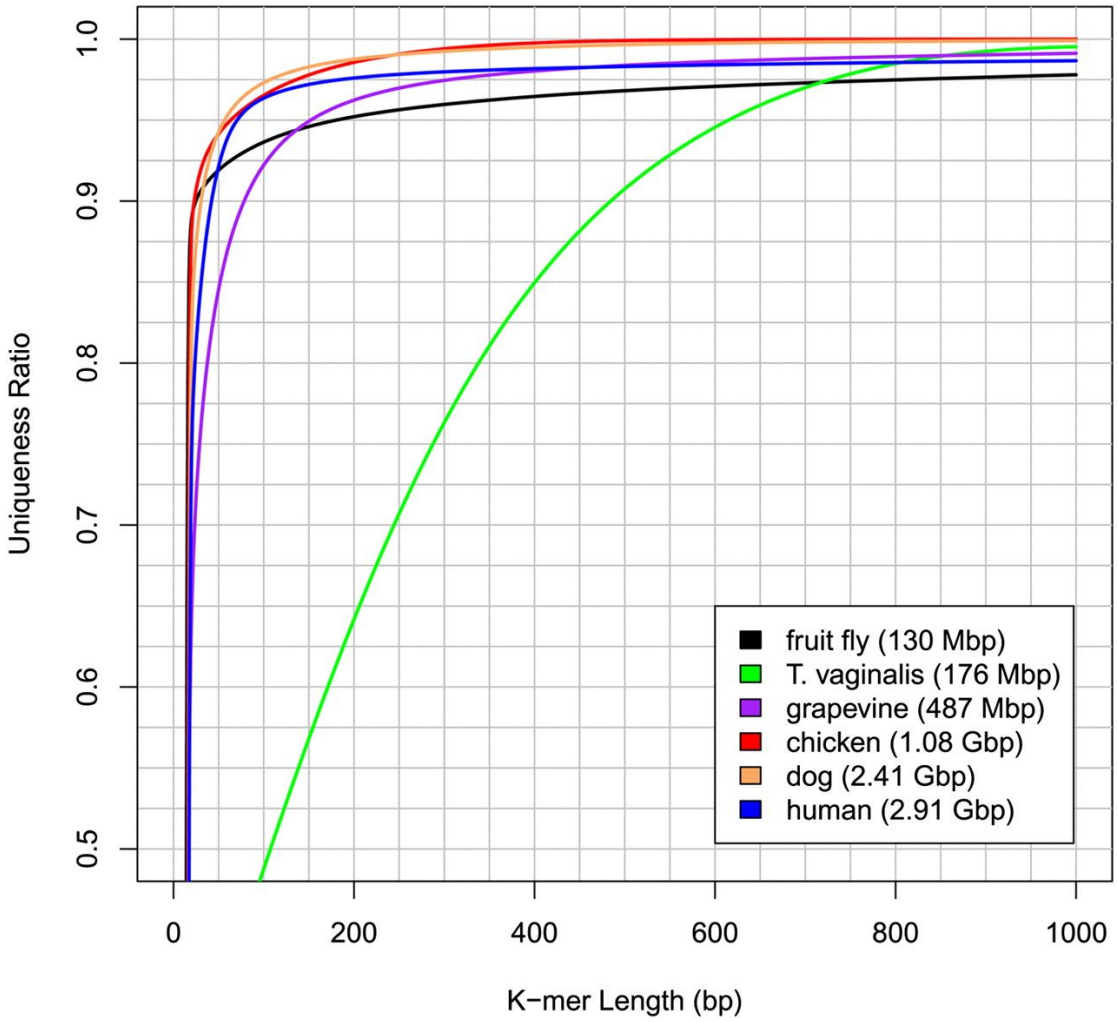
k=10?

4^{10} (1,048,576)

Every one of these is present in the human genome at least once

http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-167

It takes a very long k-mer to be unique in most genomes!



http://genome.cshlp.org/content/20/9/1165

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

→ Read TGGTCA

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

kmer index is used to quickly find candidate alignment locations in genome.

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

→ Read TGGTCA

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Read TGGTCA

Hash matches 3,10

Hash match

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

→

Read TGGTCA

Hash matches 3,10,6

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

→

Read TGGTCA

Hash matches 3,10,6

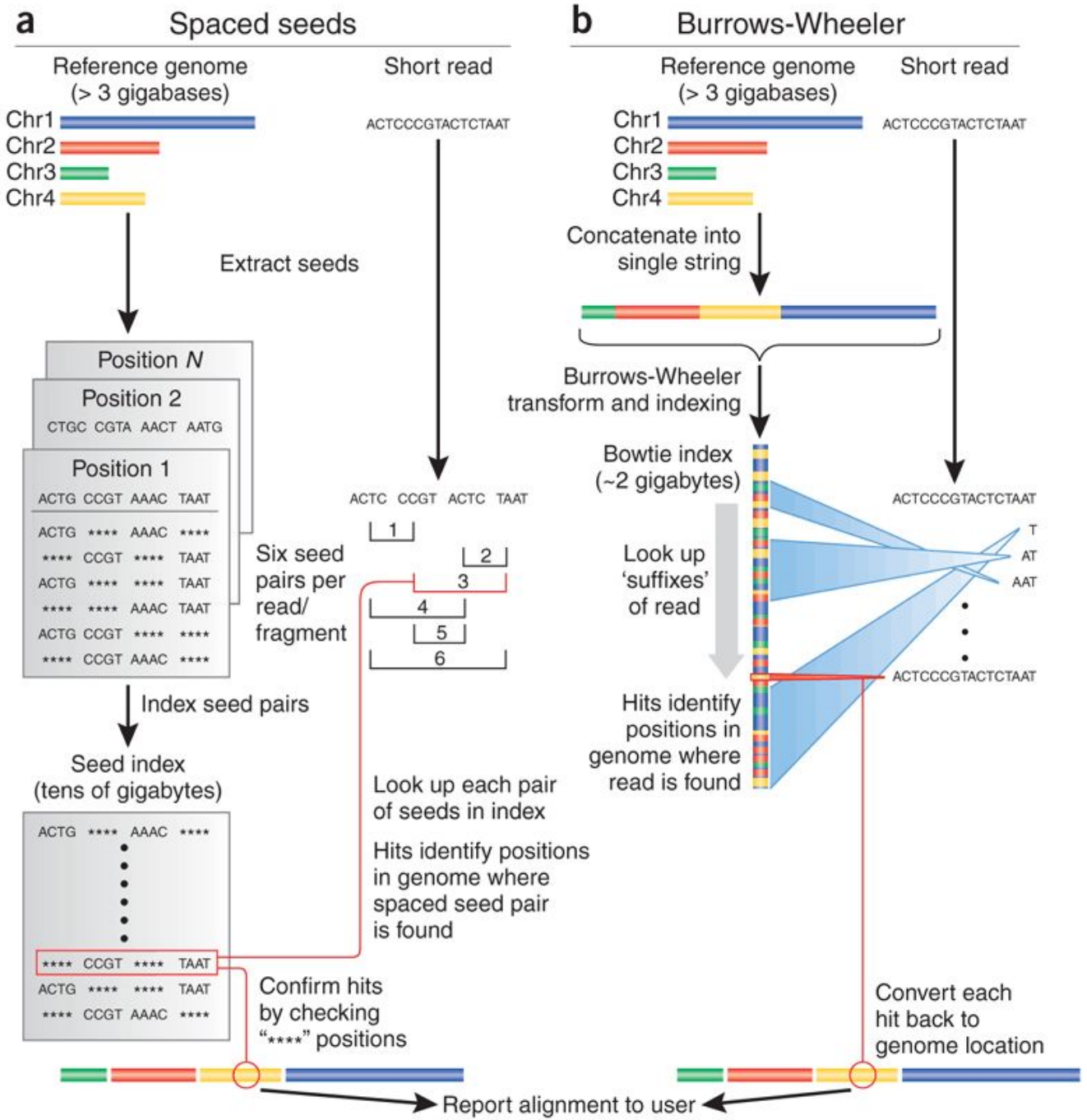
Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Mapping quality (MAPQ)

What is the probability that the sequence should be mapped here and only here?

MAPQ uses the Phred (log) scale: $MAPQ = -10*\log_{10}(P_{map\_loc\_wrong})$

$(P_{map\_loc\_wrong})$	$\log_{10}(P_{map\_loc\_wrong})$	MAPQ
1	0	0
0.1	-1	10
0.01	-2	20
0.001	-3	30
0.0001	-4	40



Trapnell, Salzberg, Nature Biotechnology 2009

Read alignment relies on a dictionary

- Once loaded onto memory, they can process thousands of reads/second
- Different aligners offer different advantages:
  - Sensitivity to variability (aligning to highly variable organisms)
  - Speed
  - Memory
  - Specificity (at some cost of sensitivity)
- It is NOT a solved problem
- You really do not know **which region was sequenced** you only know what is **the best match** to the reference.
  - There could be sequencing errors
  - Variability
  - Non unique alignments

Short read aligners

Table 2. Alignment algorithms and software tools.

Name	Website	Reference	Remark
SOAP *	soap.genomics.org.cn	[32–35]	k-mer inexact match seed; support at most 3 mismatches; GPU calculation supported
CUSHAW $	cushaw3.sourceforge.net/homepage.htm#downloads	[36–39]	k-mer inexact match, maximal exact match and hybrid seeds; GPU supported
Bowtie &	bowtie-bio.sourceforge.net	[40,41]	k-mer inexact match seed; high speed; double-index; up to 3 mismatches
BWA	bio-bwa.sourceforge.net	[42,43]	k-mer inexact match and maximal exact match seed
GASSST	www.irisa.fr/symbiose/projects/gassst/	[44]	k-mer exact match seed; it currently has been tested for reads up to 500 bp
GNUMAP	dna.cs.byu.edu/gnumap/	[45]	k-mer exact match seed; probabilistically mapping reads to repeat regions
MOSAIK	gkno.me/pipelines.html#mosaik	[46]	k-mer exact match seed
NextGenMap	cibiv.github.io/NextGenMap/	[47]	q-gramq-gram filter; GPU calculation supported
QPALMA	www.raetschlab.org/suppl/qpalma	[48]	k-mer inexact match; incorporate read quality score and splice site
RMAP	rulai.cshl.edu/rmap/	[49,50]	k-mer inexact match seed; 10 mismatches allowed; incorporate read quality score
Segemehl	www.bioinf.uni-leipzig.de/Software/segemehl/	[51]	k-mer inexact match seed; enhanced suffix arrays
SeqMap	www-personal.umich.edu/~jianghui/seqmap/	[52]	k-mer inexact match; support windows, linux, Mac OS

Pharmaceutics 2015, 7, 523-541; doi:10.3390/pharmaceutics7040523

Table 2. *Cont.*

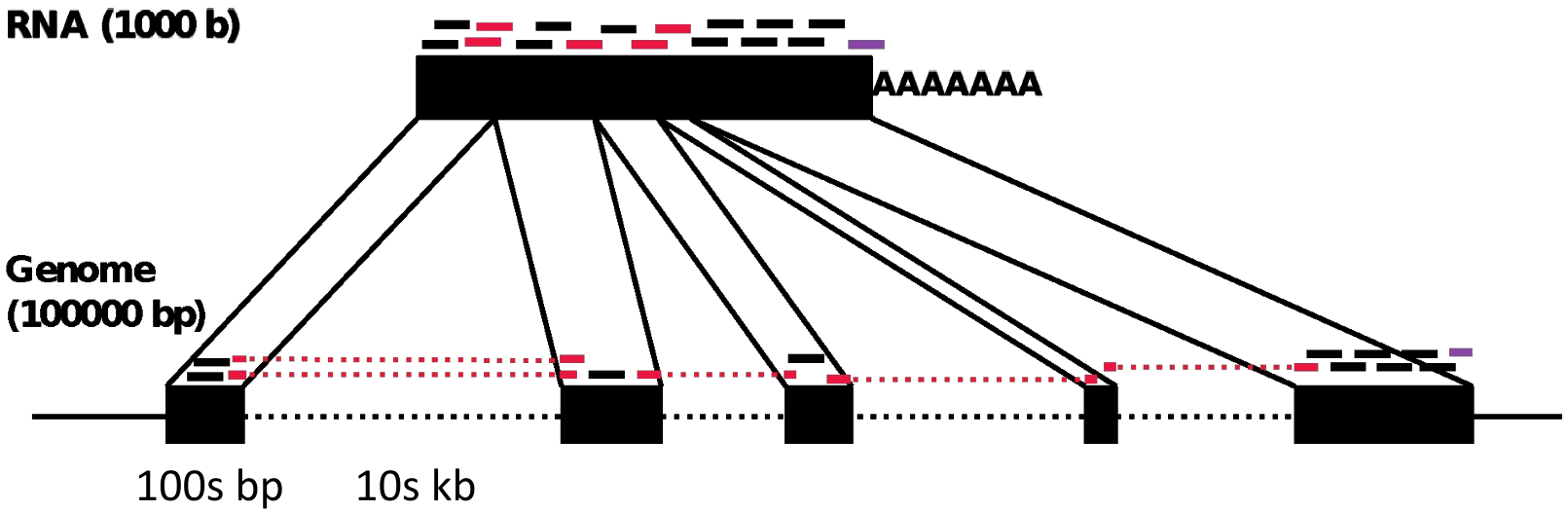
Name	Website	Reference	Remark
Stampy	www.well.ox.ac.uk/project-stampy	[53]	k-mer inexact match; support up to 30 bp indels in paired-end reads alignment
Cloudburst	sourceforge.net/projects/cloudburst-bio/	[54]	Highly sensitive read mapping with MapReduce.
drFAST	drfast.sourceforge.net/	[55]	k-mer inexact match; specially designed for better delineation of structural variants
BFAST	sourceforge.net/projects/bfast/	[56]	k-mer spaced seeds
MAQ	maq.sourceforge.net	[57]	k-mer spaced seeds; incorporate quality scores of reads in alignment
MOM	go.vcu.edu/mom	[58]	k-mer spaced seeds; unlimited mismatches in the 3′ and 5′ flanking regions.
PASS	pass.cribi.unipd.it	[59]	k-mer spaced seeds; implemented in C++ and supported on Linux and Windows
PerM	code.google.com/p/perm/	[60]	k-mer spaced seeds; 9 mismatches are allowed
SHRiMP2	compbio.cs.toronto.edu/shrimp/	[61,62]	combined k-mer spaced seeds and q-gram filter
ZOOM	www.bioinfor.com/zoom/general/overview.html	[63]	k-mer spaced seeds; tolerate 2 mismatches by default
BarraCUDA	seqbarracuda.sourceforge.net/	[64]	Incorporate GPU to speed up BWA
GEM	gemlibrary.sourceforge.net/	[65]	q-gram filter
MPSCAN	www.atgc-montpellier.fr/mpscan/	[66]	q-gram filter; support Windows, linux, Mac OS
ERNE	iga-rna.sourceforge.net/	[67]	long gap support; Works on Windows, Mac OS X, linux
SARUMAN	www.cebitec.uni-bielefeld.de/brf/saruman/saruman.html	[68]	k-mer inexact matched seed; support GPU calculation
LAST	last.cbrc.jp/	[69]	adaptive seed
Genalice MAP	www.genalice.com/product/genalice-map/	NA	cloud calculation; High sensitivity for SNPs and long INDELS
Novoalign	www.novocraft.com/	NA	support up to 7 and 16 mismatches in single-end and pair-end reads.
PRIMEX	bioinformatics.cribi.unipd.it/primex	[70]	k-mer inexact match seed; written in C++; lookup table and server functionality
SOCS	solidsoftwaretools.com/gf/project/socs/	[71]	good at align CpG methylation-enriched reads
SToRM	bioinfo.lifl.fr/yass/iedera_solid/storm/	[72]	doesn't support pair-end reads
iSAAC	https://github.com/sequencing/isaac_aligner	[73]	k-mer inexact match seed; high speed
RazerS	www.seqan.de/projects/razers/	[74]	q-gram filter; support Windows, linux, Mac OS X
SSAHA2	www.sanger.ac.uk/resources/software/ssaha2/	[75]	k-mer inexact match seed; support various output formats
UGENE	ugene.unipro.ru/	[76]	works on Windows, linux and Mac OS X

* Include SOAP, SOAP2, SOAP3 and SOAP3-dp; $ Include CUSHAW (k-mer inexact match seed), CUSHAW2 (maximal exact match seed) and CUSHAW3 (hybrid seeds); & Include Bowtie and Bowtie2. NA: commercial software, no reference available.

Sequence alignment software

Aligner	Approach	Applications	Availability
BWA-mem	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Bowtie2	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Novoalign	hash-based	DNA, SE, PE	free for academic use
STAR	hash-based (reads)	RNA-seq	open-source
GSNAP	hash-based (reads)	RNA-seq	open-source
TopHat2	Burrows-Wheeler	RNA-seq	open-source
HISAT2	Burrows-Wheeler	RNA-seq	open-source

The RNA-Seq alignment problem



Challenges:

- Genes exist at many different expression levels, spanning several orders of magnitude.
- Reads originate from both mature mRNA (exons) and immature mRNA (introns) and it can be problematic to distinguish between them.
- Reads are short and genes can have many isoforms making it challenging to determine which isoform produced each read.

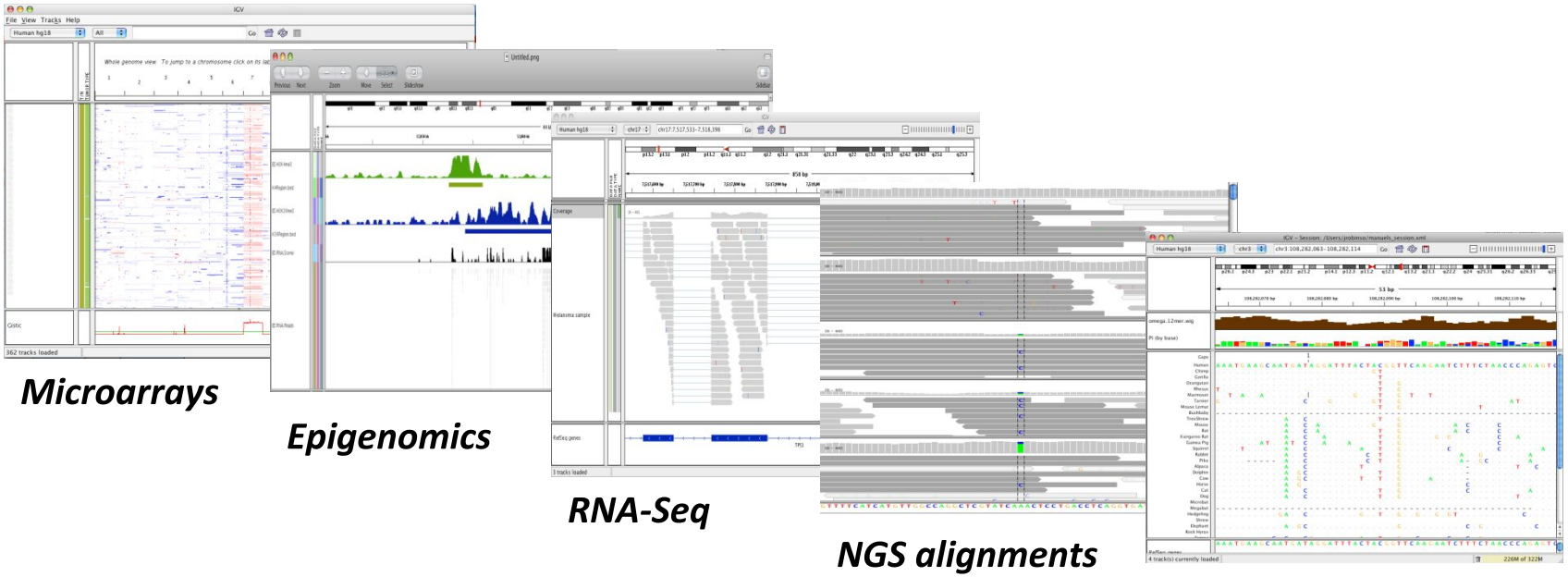
Mapping RNA-Seq reads: Exon-first spliced alignment (e.g. TopHat)

Integrative Genomics Viewer

A desktop application

for the visualization and interactive exploration

of genomic data



Microarrays

Epigenomics

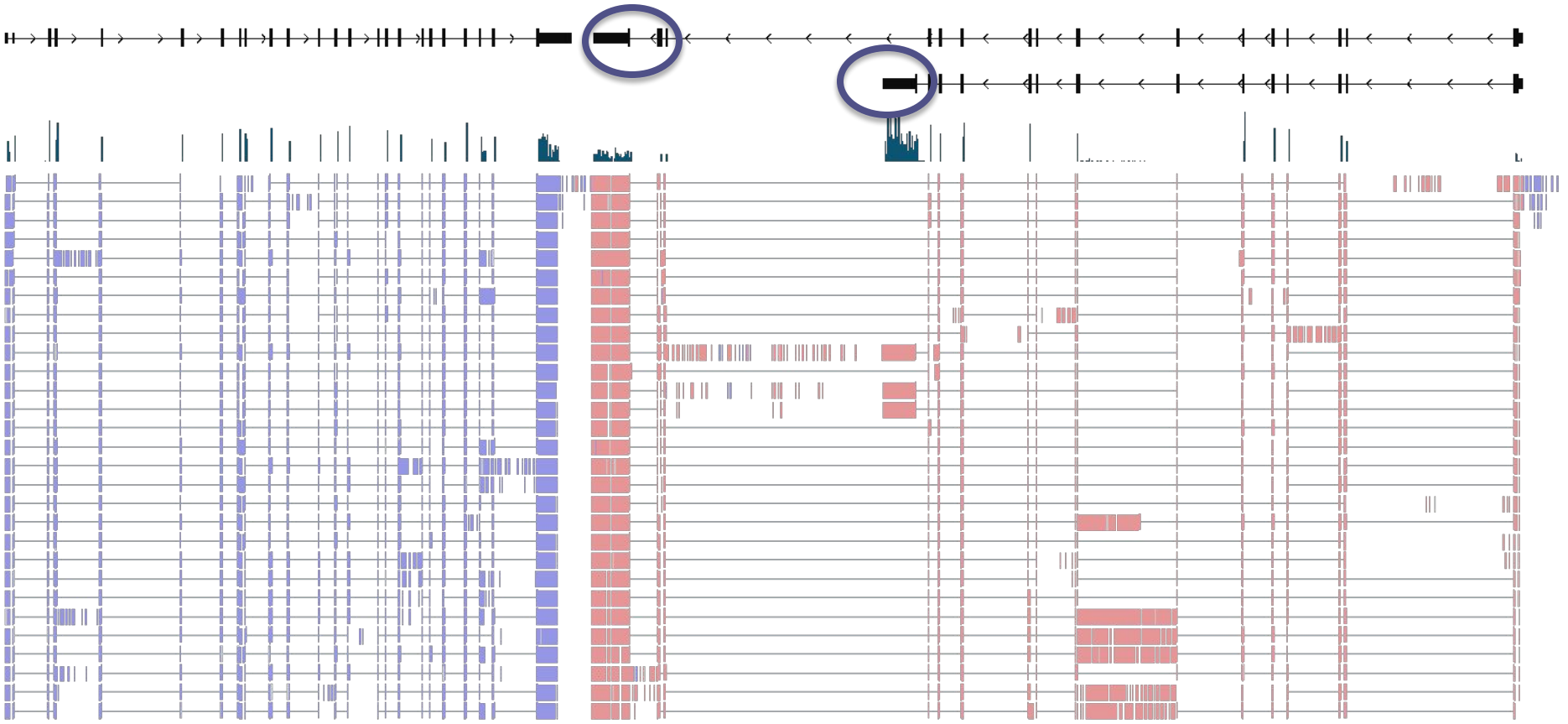
RNA-Seq

NGS alignments

Comparative genomics

BROAD INSTITUTE

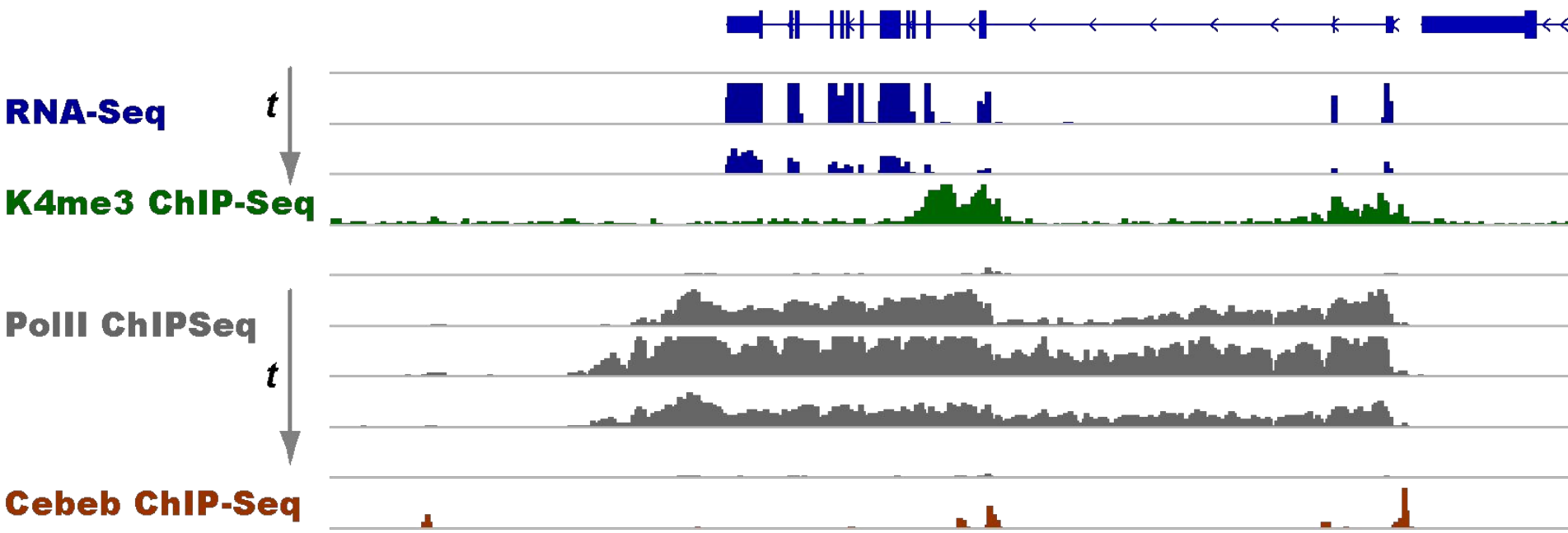
Visualizing read alignments with IGV — RNASeq



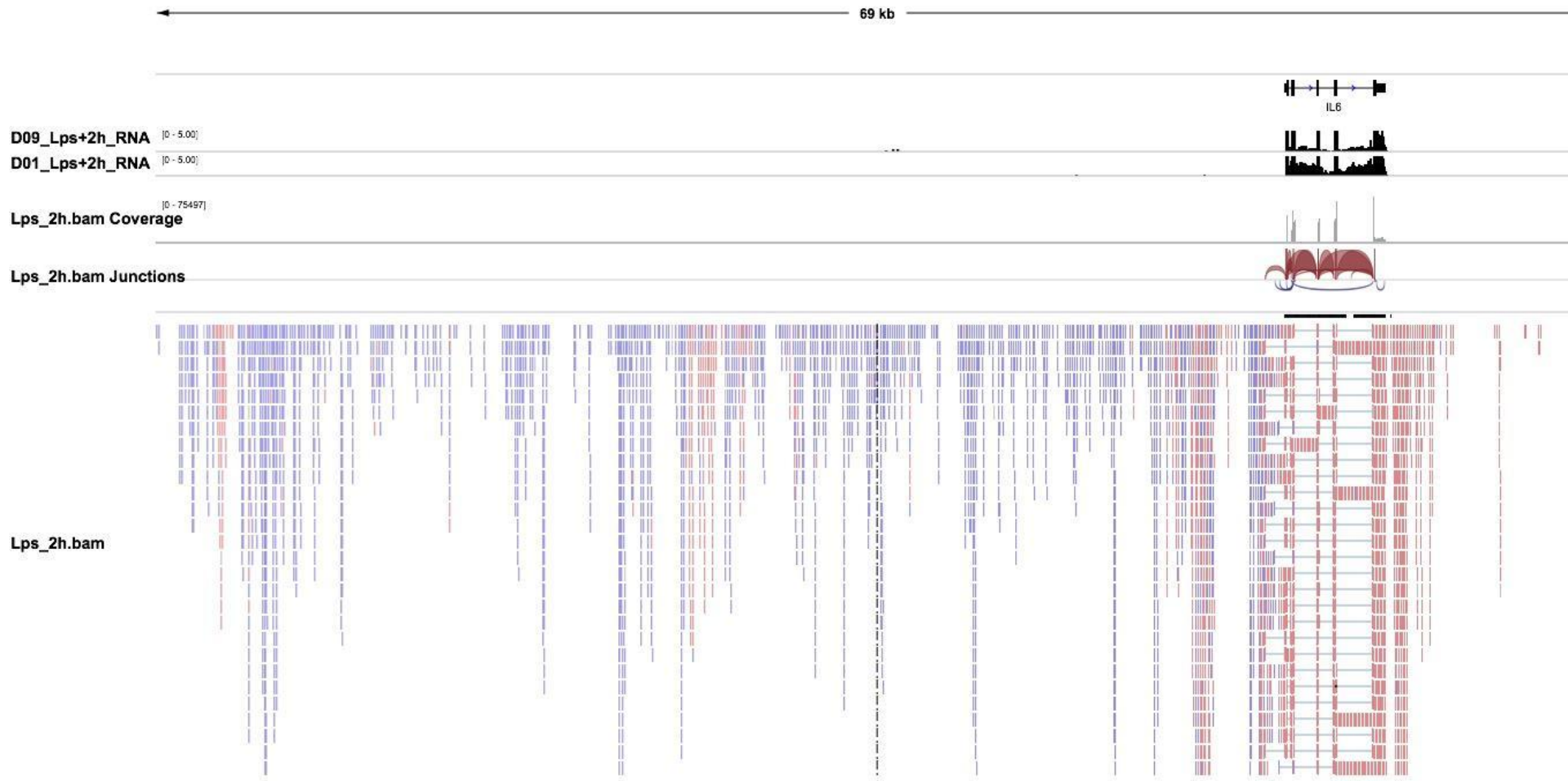
Strand specific library!

Gap between reads spanning exons

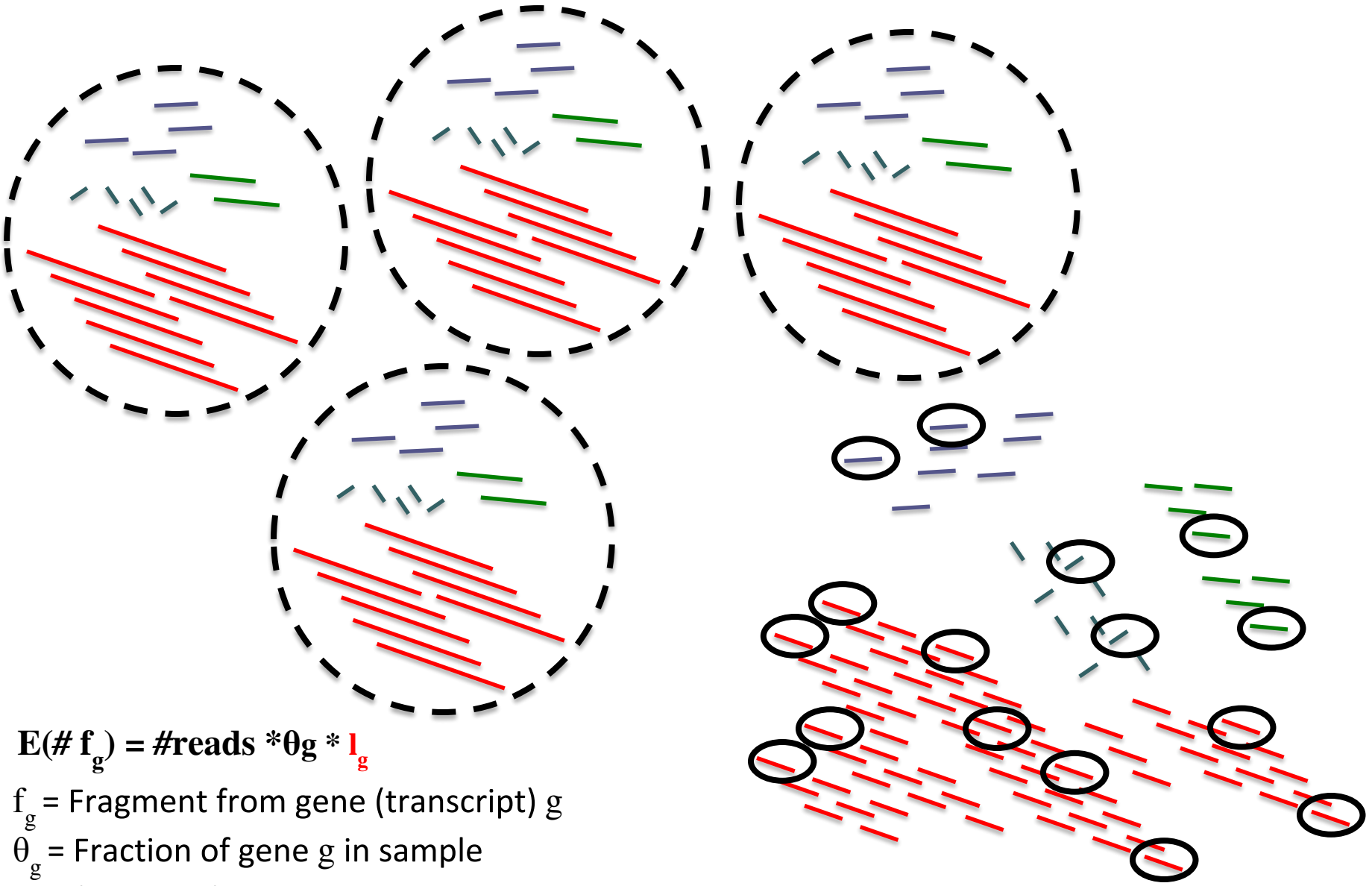
Visualizing read alignments with IGV — zooming out



Or to troubleshoot



Quantification is the next problem



$$E(\#\, f_g) = \#reads * \theta_g * l_g$$

f_g = Fragment from gene (transcript) g

θ_g = Fraction of gene g in sample

l_g = (effective) length of gene g