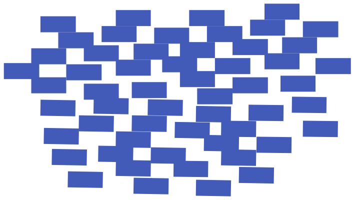
UMASS University of
Massachusetts
Medical School

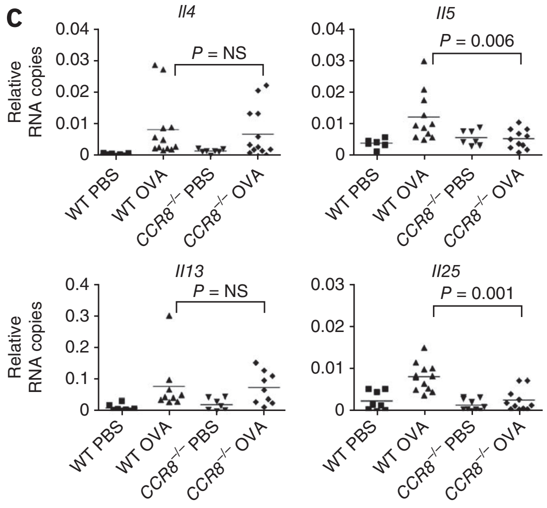
Week 4
FDR
Alignment

Biology slowly becoming a “big data” science

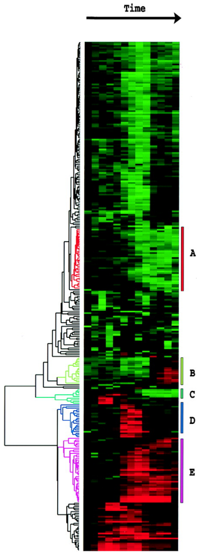
2010s

Sequenced reads





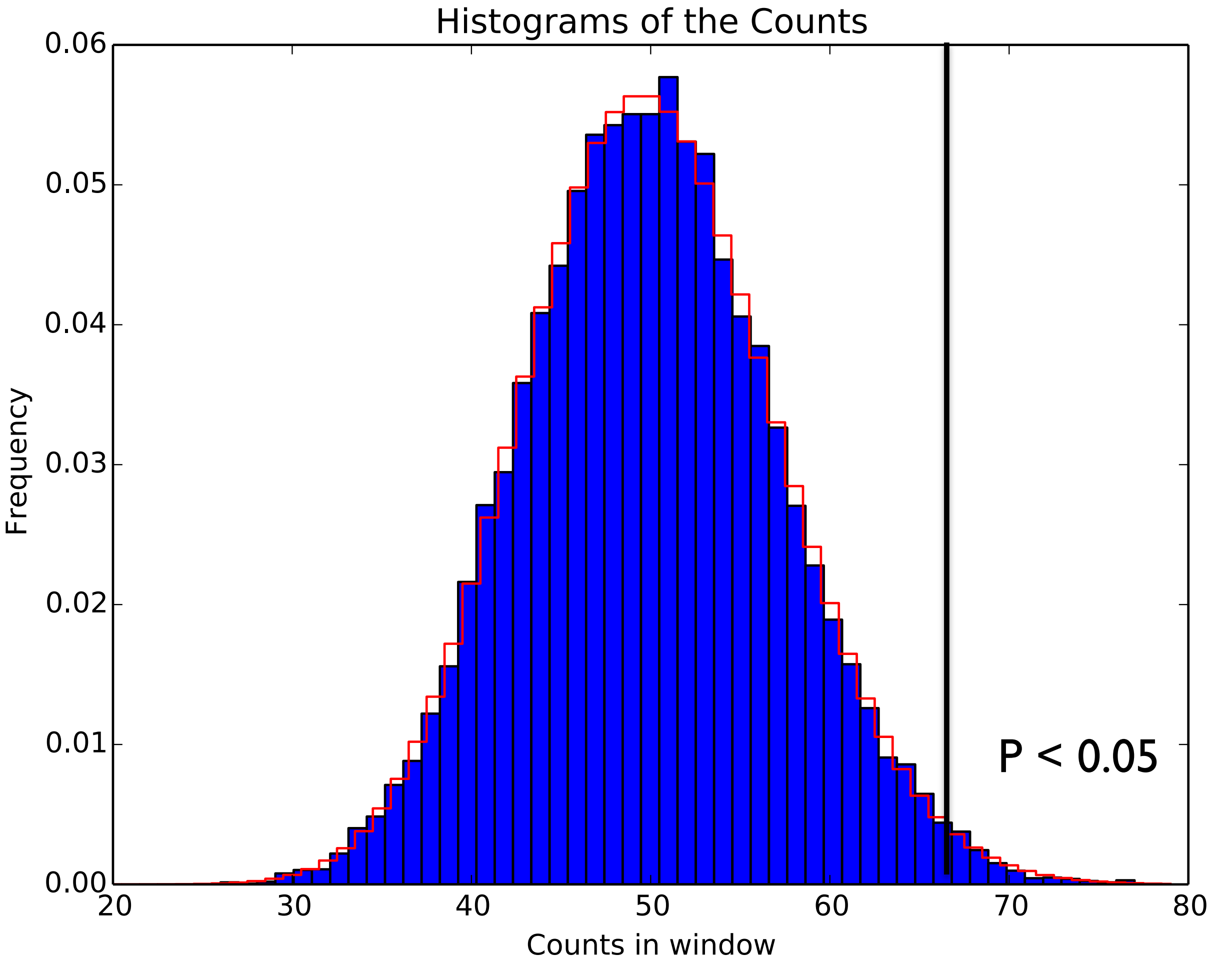
Millions-billions



Statistical methods are deeply embedded – two concepts

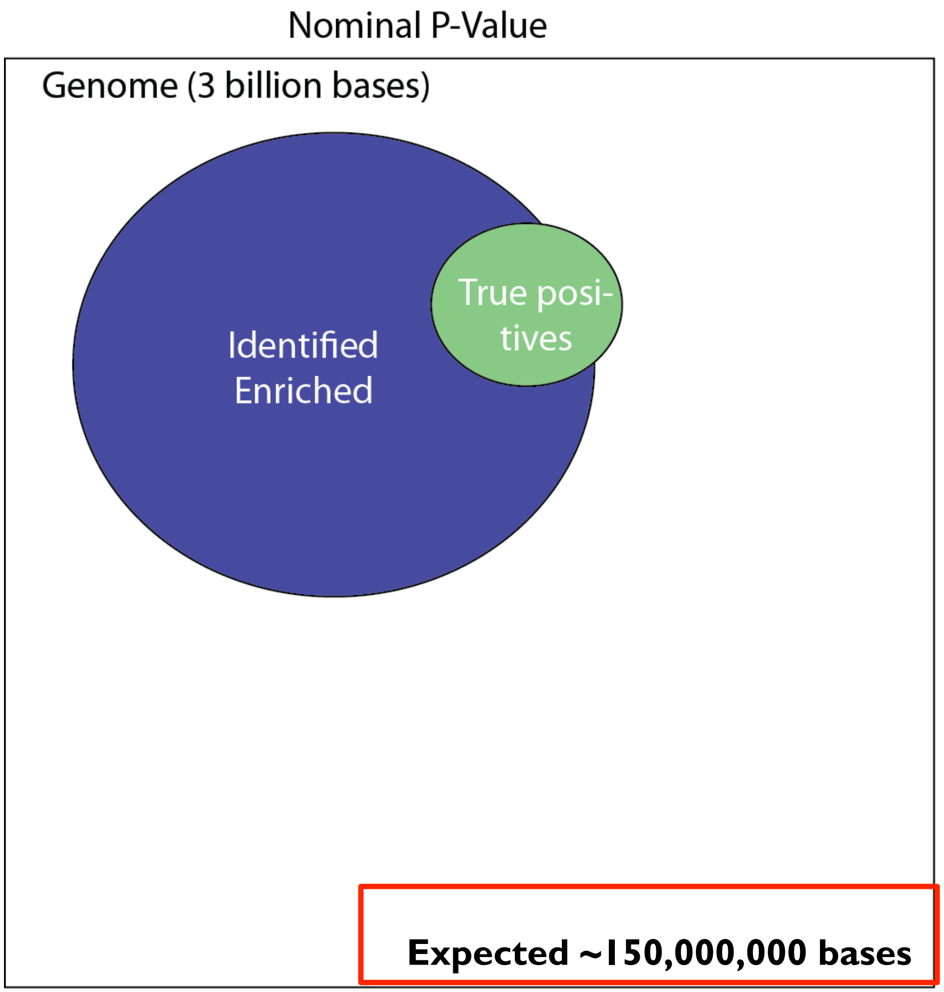
Multiple testing problems
Modeling count data

We can’t use a nominal p-value any longer



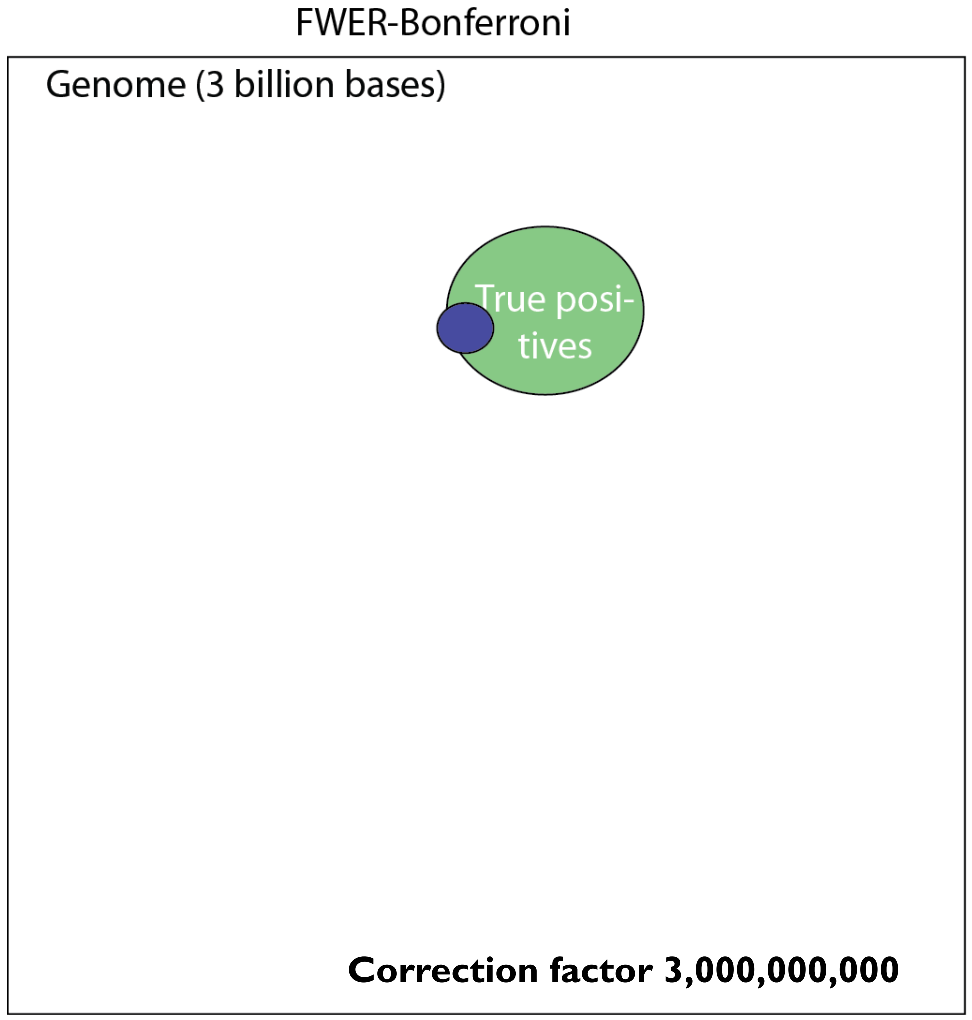
All will be noise!

The genome is large, many things happen by chance



We need to correct for multiple hypothesis testing

Bonferroni correction is way to conservative



FWER-Bonferroni

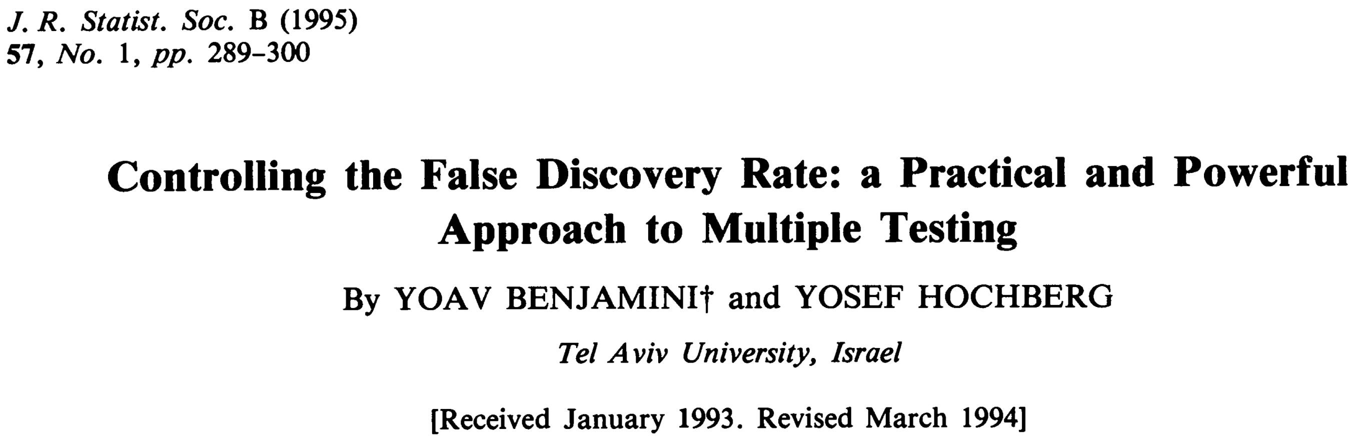
Genome (3 billion bases)

True positives

Correction factor 3,000,000,000

Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

How do we compute significance when we have this much data?



J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

Downloadable from: http://garberlab.umassmed.edu/bootcamp.2015/BH.pdf

A refresher in math notation

The product operator $\prod$

$$\prod_{i=2}^{5} i^2 = 2^2 \times 3^2 \times 4^2 \times 5^2$$

The summation operator $\sum$

$$\sum_{i=2}^{5} i^2 = 2^2 + 3^2 + 4^2 + 5^2$$

Expected Value

The expected value of a random variable is the sum of its values weighted by their probability. So, if X is a random variable (e.g. lottery ticket pay-off, Gene expression value) and x_i are its discrete values then:

$$E(X) = \sum_i x_i p(x_i)$$

The mean of observed values is an “unbiased” estimator of the expected value of the underlying distribution

Problem formulation

We test m hypothesis (e.g. gene i is differentially expressed m =20,000).
We wish to detect (m-m_0) genes that say change between conditions

	Do not pass significance	pass significance	Total
Random noise (Null hypothesis is true)	$\mathbf{U}$	$\mathbf{V}$	$\mathbf{m_0}$
True signal (Null hypothesis is false)	$\mathbf{T}$	$\mathbf{S}$	$\mathbf{m\text{-}m_0}$
Total	$\mathbf{m\text{-}R}$	$\mathbf{R}$	$\mathbf{m}$

V = **# Type I errors** (False Positives)
T = **# Type II errors** (False Negatives)

We want to infer $E(V \mid \text{significance})$, but we only observe R and m
Statistical Power is the ability to reduce **Type II** errors, a topic for another time!

Definitions

- Absent any signal in the data at a nominal p-value α. We have $E(V \mid m,\alpha) = m\alpha$.
- If we have signal then $E(V \mid m,\alpha) \leq m\alpha$
- The probability of making at least one error: $P(V>0)$. Also called Family Wise Error Rate (FWER).
- For a given dataset, we would like to compute the fraction of type I errors: (oFDR) = $\mathbf{V}/\mathbf{R}$.
- $\mathbf{FDR := E(V/R)}$

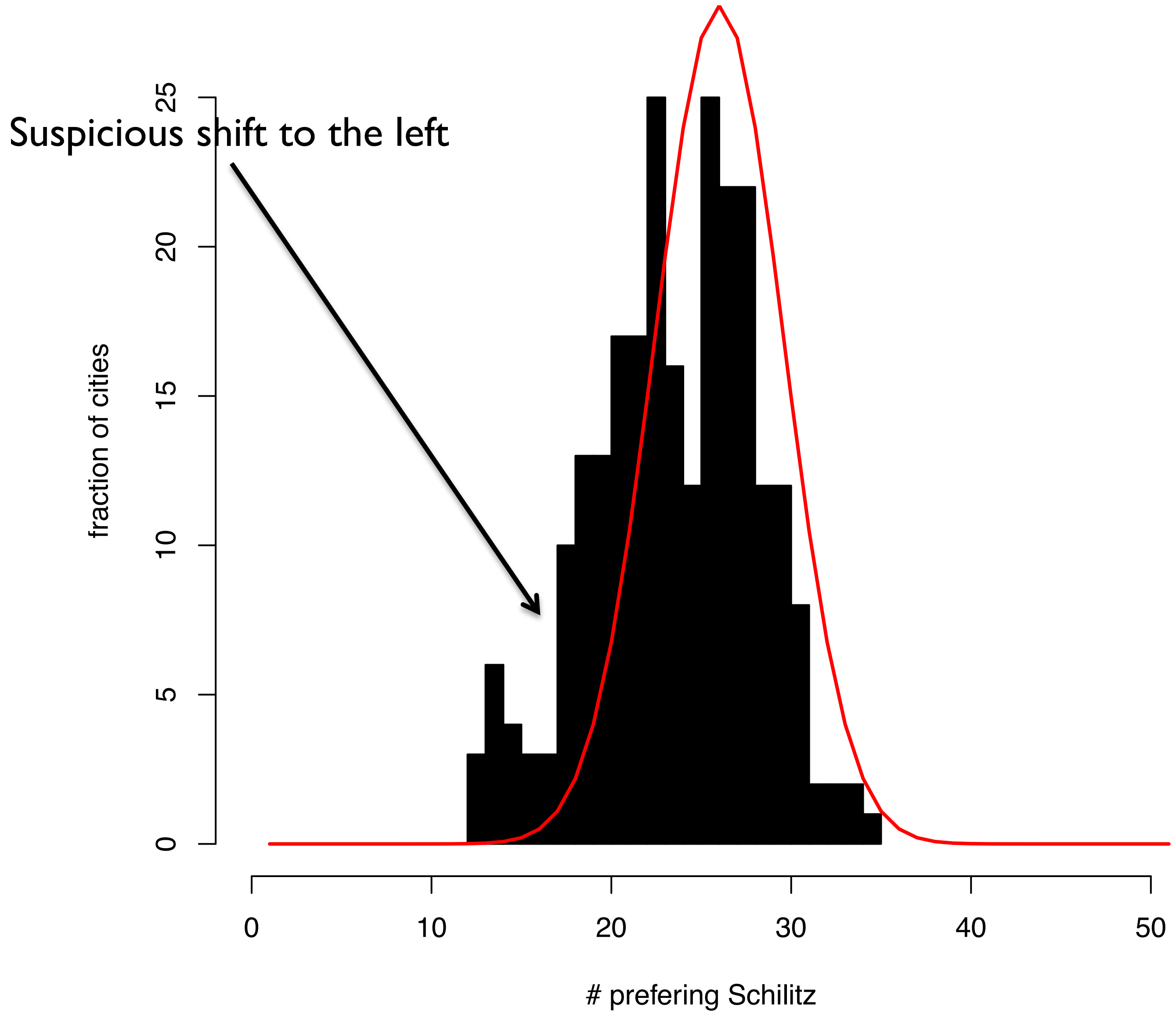
Example: Michelob strikes back

- Schlitz goes on tour, in every town they conduct a blind test of 50 Michelob faithful
- In a few places, Michelob manages to find out who are the blind testers and attempts to train them

Lets assume:

1. Training improves Michelob detection to 65%
2. The tour goes to 250 cities

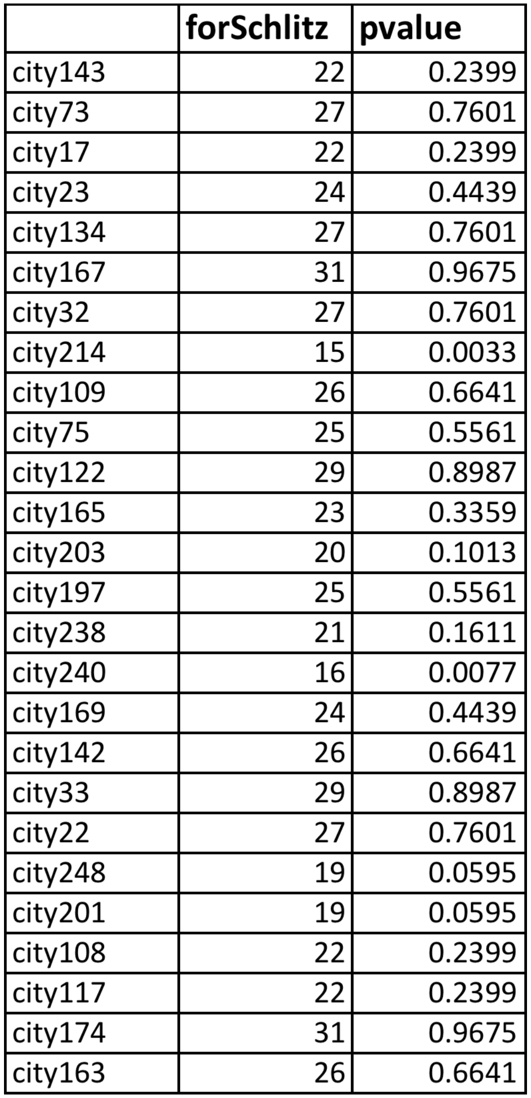
Something is funny



Suspicious shift to the left

Can Schlitz detect which cities had undergone training?

At 10% we only would talk to less than 10% of cities that did not underwent training. Not too bad

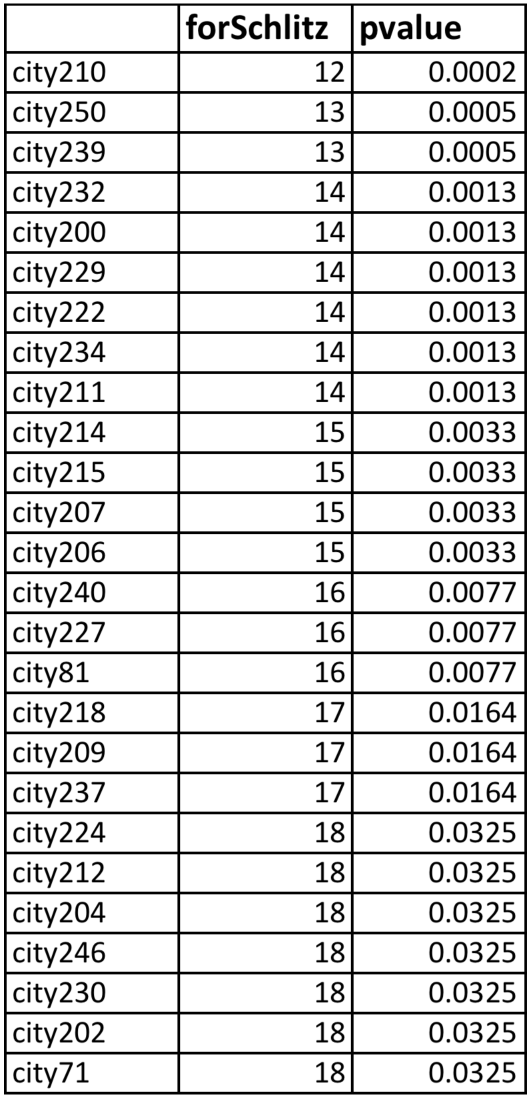


	forSchlitz	pvalue
city143	22	0.2399
city73	27	0.7601
city17	22	0.2399
city23	24	0.4439
city134	27	0.7601
city167	31	0.9675
city32	27	0.7601
city214	15	0.0033
city109	26	0.6641
city75	25	0.5561
city122	29	0.8987
city165	23	0.3359
city203	20	0.1013
city197	25	0.5561
city238	21	0.1611
city240	16	0.0077
city169	24	0.4439
city142	26	0.6641
city33	29	0.8987
city22	27	0.7601
city248	19	0.0595
city201	19	0.0595
city108	22	0.2399
city117	22	0.2399
city174	31	0.9675
city163	26	0.6641

Can Schlitz detect which cities had undergone training?

At 10% we only would talk to less than 10% of cities that did not underwent training. Not too bad

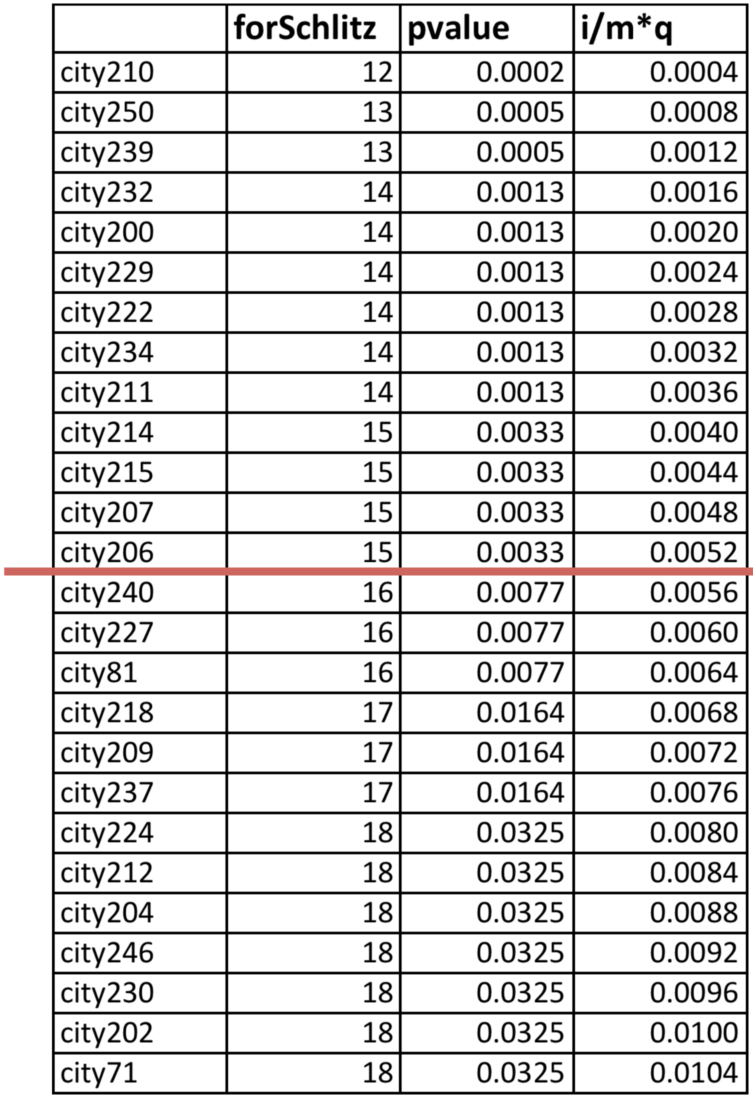
If we were to use Bonferroni, and a 0.1 significant value, we would need a corrected p-value of 0.1/250 = 0.0004. Which would give us **ONLY 1 city**



	forSchlitz	pvalue
city210	12	0.0002
city250	13	0.0005
city239	13	0.0005
city232	14	0.0013
city200	14	0.0013
city229	14	0.0013
city222	14	0.0013
city234	14	0.0013
city211	14	0.0013
city214	15	0.0033
city215	15	0.0033
city207	15	0.0033
city206	15	0.0033
city240	16	0.0077
city227	16	0.0077
city81	16	0.0077
city218	17	0.0164
city209	17	0.0164
city237	17	0.0164
city224	18	0.0325
city212	18	0.0325
city204	18	0.0325
city246	18	0.0325
city230	18	0.0325
city202	18	0.0325
city71	18	0.0325

Can Schlitz detect which cities had undergone training?

At 10% we only would talk to less than 10% of cities that did not underwent training. Not too bad



	forSchlitz	pvalue	i/m*q
city210	12	0.0002	0.0004
city250	13	0.0005	0.0008
city239	13	0.0005	0.0012
city232	14	0.0013	0.0016
city200	14	0.0013	0.0020
city229	14	0.0013	0.0024
city222	14	0.0013	0.0028
city234	14	0.0013	0.0032
city211	14	0.0013	0.0036
city214	15	0.0033	0.0040
city215	15	0.0033	0.0044
city207	15	0.0033	0.0048
city206	15	0.0033	0.0052
city240	16	0.0077	0.0056
city227	16	0.0077	0.0060
city81	16	0.0077	0.0064
city218	17	0.0164	0.0068
city209	17	0.0164	0.0072
city237	17	0.0164	0.0076
city224	18	0.0325	0.0080
city212	18	0.0325	0.0084
city204	18	0.0325	0.0088
city246	18	0.0325	0.0092
city230	18	0.0325	0.0096
city202	18	0.0325	0.0100
city71	18	0.0325	0.0104

B.H. Procedure:
Find max i, s.t. $P_i < i/m*q$

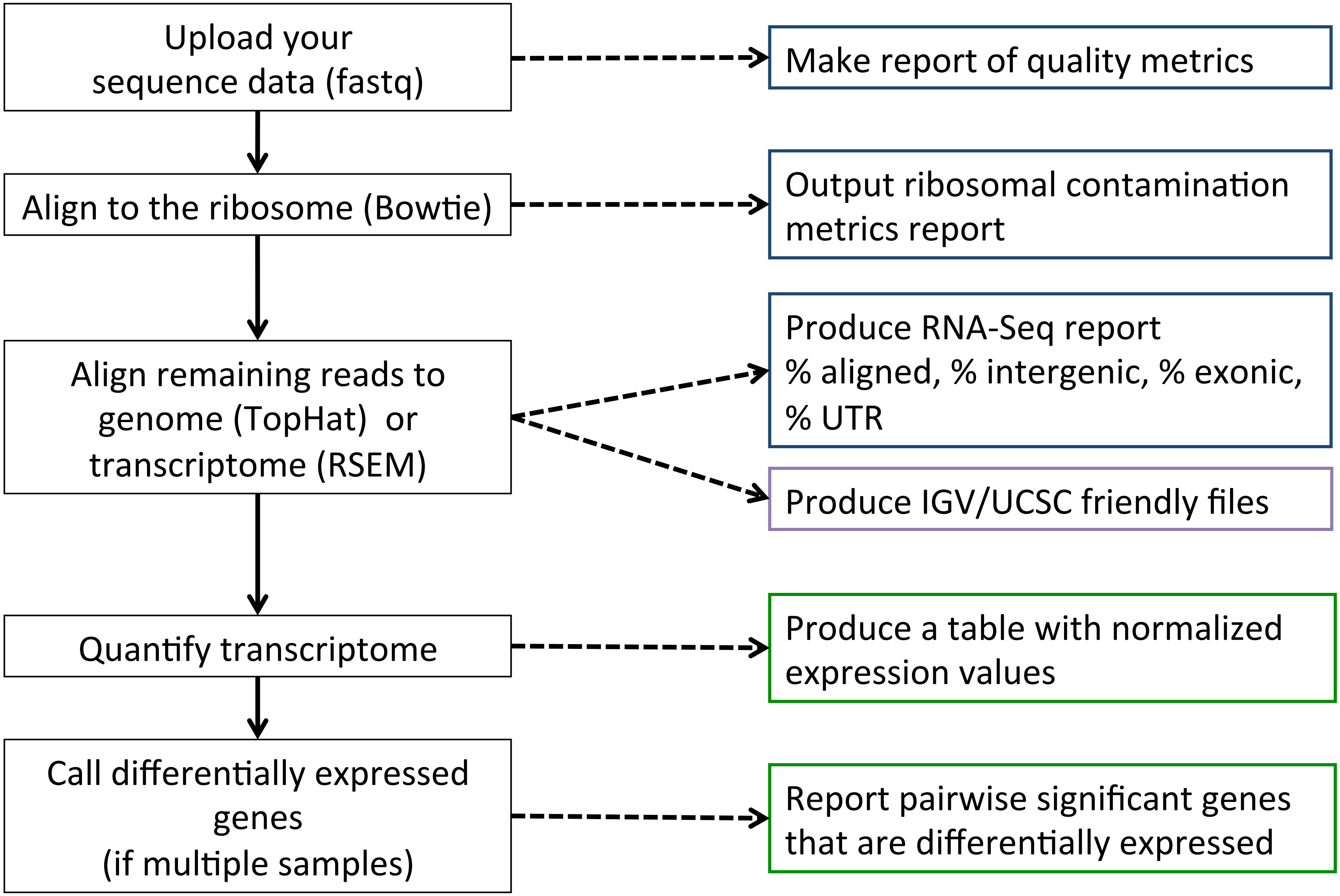
Which gives **13 cities!**

If we were to use Bonferroni, and a 0.1 significant value, we would need a corrected p-value of 0.1/250 = 0.0004. Which would give us **ONLY 1 city**

Script that generated this data available at class site

Lets get back to sequencing

Our typical RNA quantification pipeline



Upload your sequence data (fastq)

Make report of quality metrics

Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat) or transcriptome (RSEM)

Produce RNA-Seq report
% aligned, % intergenic, % exonic, % UTR

Produce IGV/UCSC friendly files

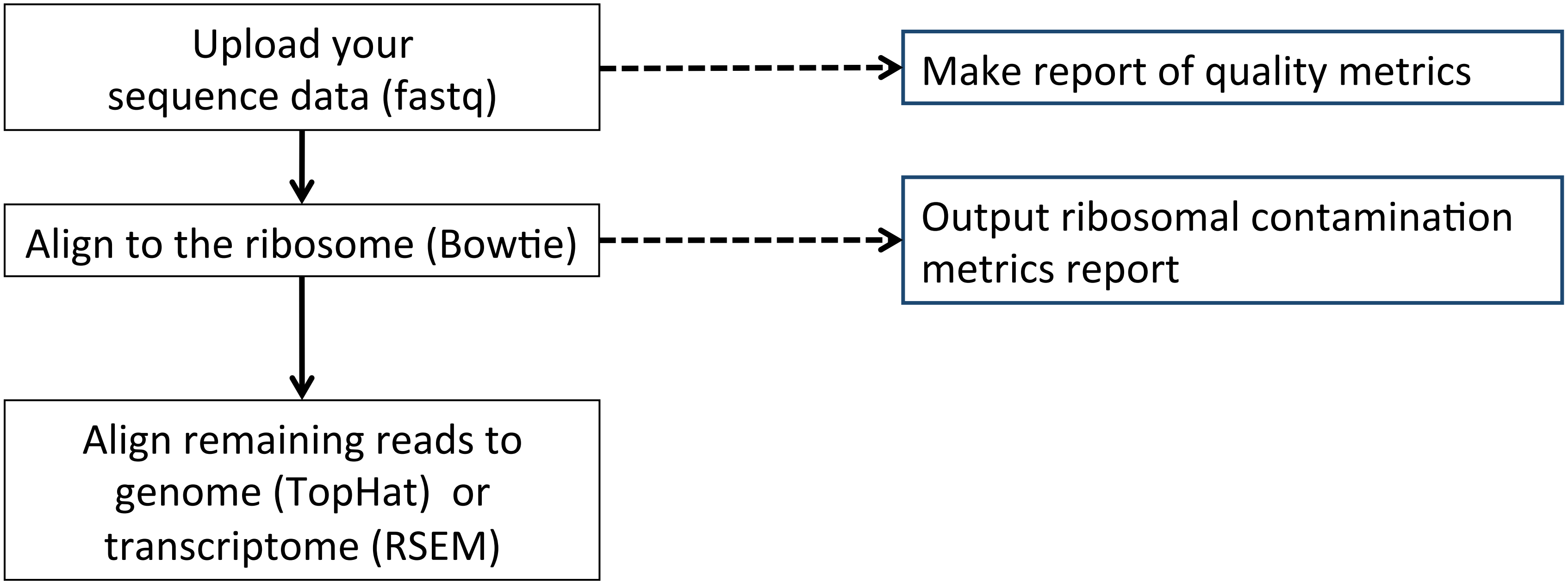
Quantify transcriptome

Produce a table with normalized expression values

Call differentially expressed genes
(if multiple samples)

Report pairwise significant genes that are differentially expressed

Alignment requires pre-processing



Upload your sequence data (fastq)

Make report of quality metrics

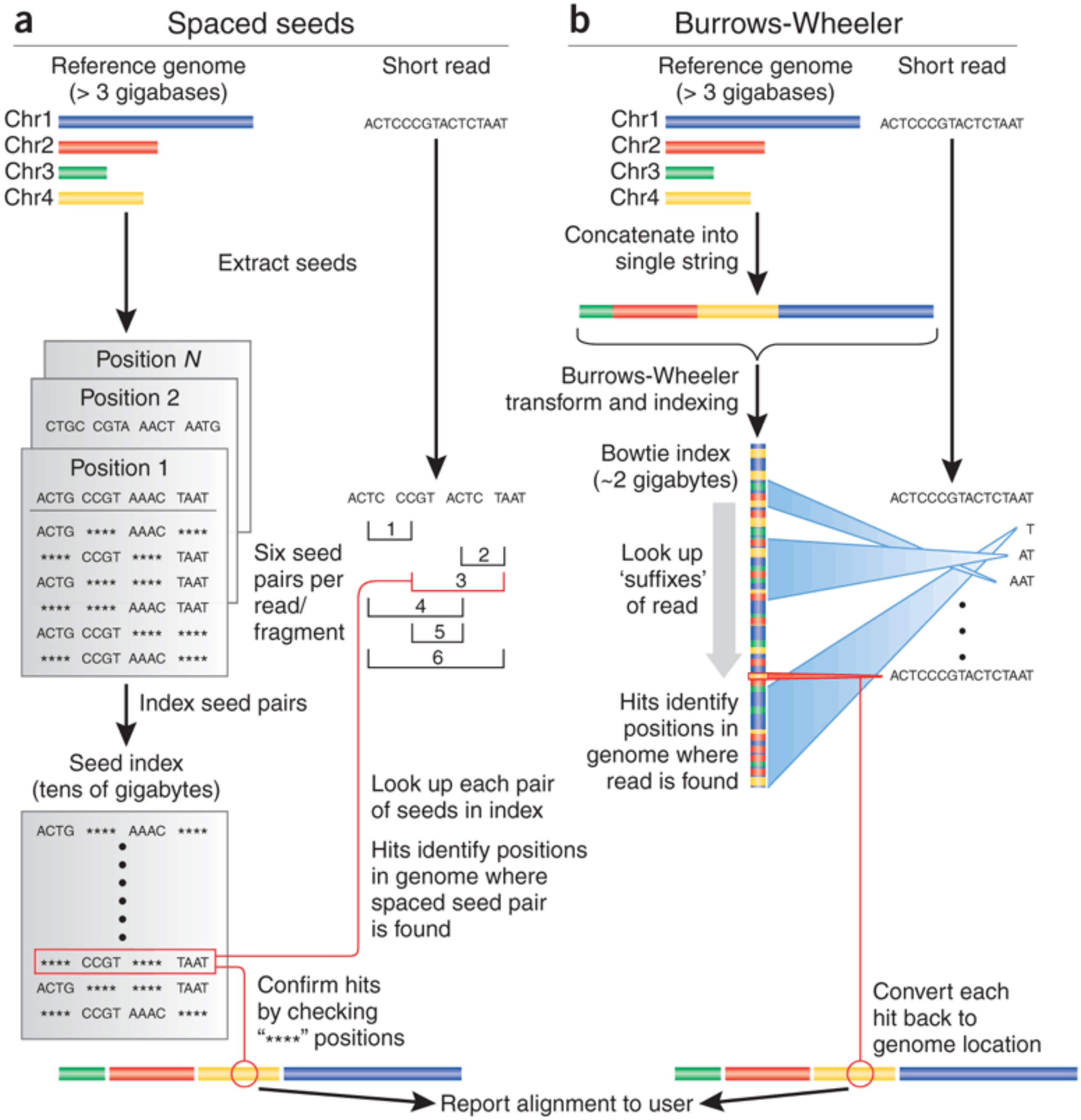
Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat) or transcriptome (RSEM)

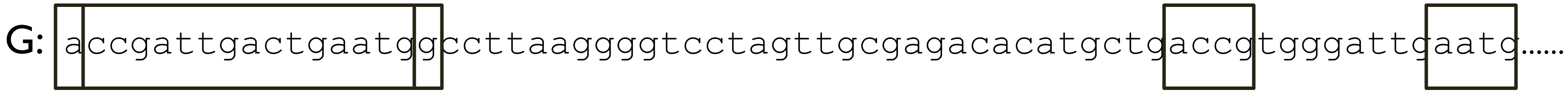
```
bowtie2-build -f mm10.fa mm10

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



Trapnell, Salzberg, Nature Biotechnology 2009

Spaced seed alignment – Hashing the genome



G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg......

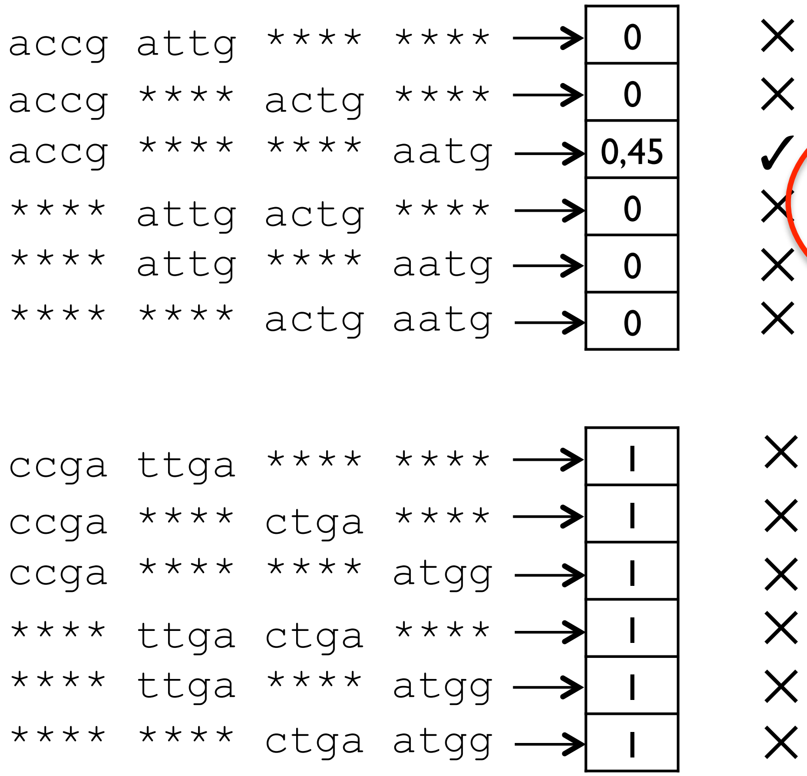
Store spaced seed positions

accg attg **** **** →	0
accg **** actg **** →	0
accg **** **** aatg →	0,45
**** attg actg **** →	0
**** attg **** aatg →	0
**** **** actg aatg →	0

ccga ttga **** **** →	1
ccga **** ctga **** →	1
ccga **** **** atgg →	1
**** ttga ctga **** →	1
**** ttga **** atgg →	1
**** **** ctga atgg →	1

Spaced seed alignment – Mapping reads

G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg......



```
accg attg **** **** → 0     ✗
accg **** actg **** → 0     ✗
accg **** **** aatg → 0,45  ✓
**** attg actg **** → 0     ✗
**** attg **** aatg → 0     ✗
**** **** actg aatg → 0     ✗

ccga ttga **** **** → 1     ✗
ccga **** ctga **** → 1     ✗
ccga **** **** atgg → 1     ✗
**** ttga ctga **** → 1     ✗
**** ttga **** atgg → 1     ✗
**** **** ctga atgg → 1     ✗
```

q: accg atag accg aatg

accgattgactgaatg accgtgggattgaatg

2 missmatches 5 missmatches

Report position 0

But, how confident are we in the placement?

$$q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$$

Mapping quality

What does $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$ mean?

Lets compute the probability the read originated at genome position i

q: accg atag accg aatg

q_s: 30 40 25 30 30 20 10 20 40 30 20 30 40 40 30 25

$q_s[k] = -10\log_{10} P(\text{sequencing error at base k})$, the PHRED score. Equivalently:

$$P(\text{sequencing error at base k}) = 10^{-\frac{q_S[k]}{10}}$$

So the probability that a read originates from a given genome position i is:

$$P(q \mid G,i) = \prod_{j \text{ match}} P(q_j \text{good call}) \prod_{j \text{ missmatch}} P(q_j \text{bad call}) \approx \prod_{j \text{ missmatch}} P(q_j \text{bad call})$$

In our example

$$P(q \mid G,0) = \left[(1-10^{-3})^6(1-10^{-4})^4(1-10^{-2.5})^2(1-10^{-2})^2\right]\left[10^{-1}10^{-2}\right] = [0.97]*[0.001] \approx 0.001$$

Mapping quality

What we want to estimate is $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$

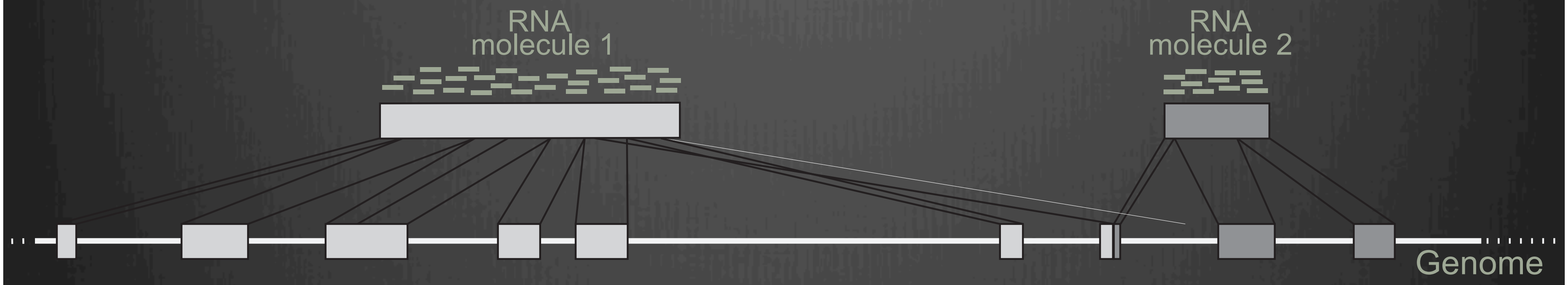
That is, the posterior probability, the probability that the region starting at i was sequenced *given* that we observed the read *q*:

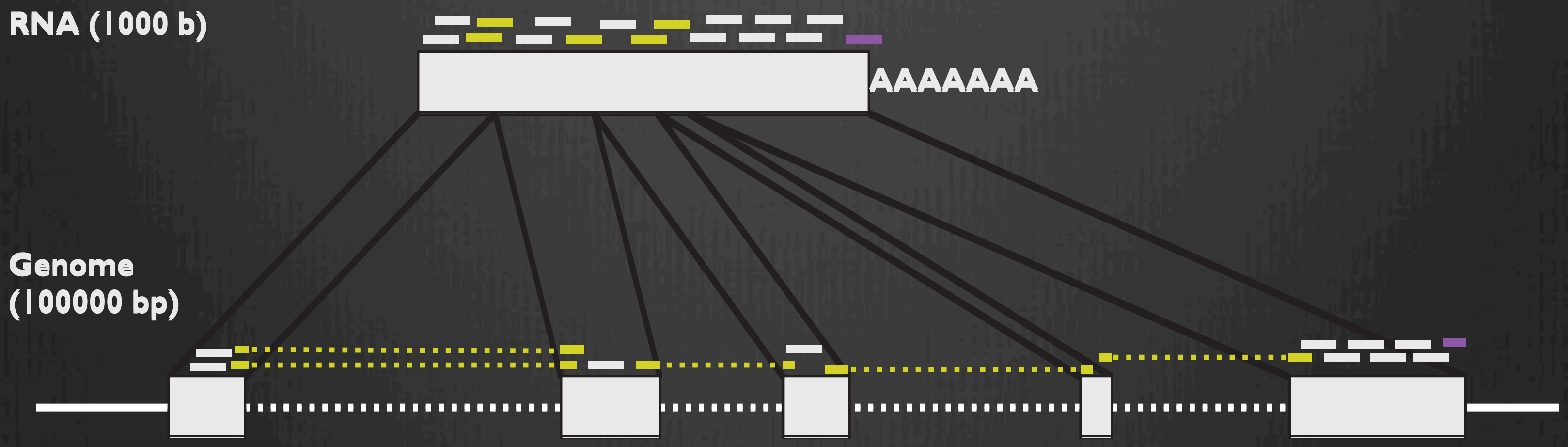
$$P(i \mid G,q) = \frac{P(q \mid G,i)P(i \mid G)}{P(q \mid G)} = \frac{P(q \mid G,i)P(i \mid G)}{\sum_j P(q \mid G, j)}$$

Fortunately, there are efficient ways to approximate this probability (see Li, H *genome Research* 2008, for example)

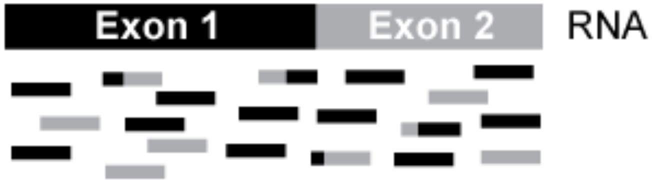
$$q_{MS} = -10\log_{10}(1 - P(i \mid G,q))$$

RNA-Seq Read mapping

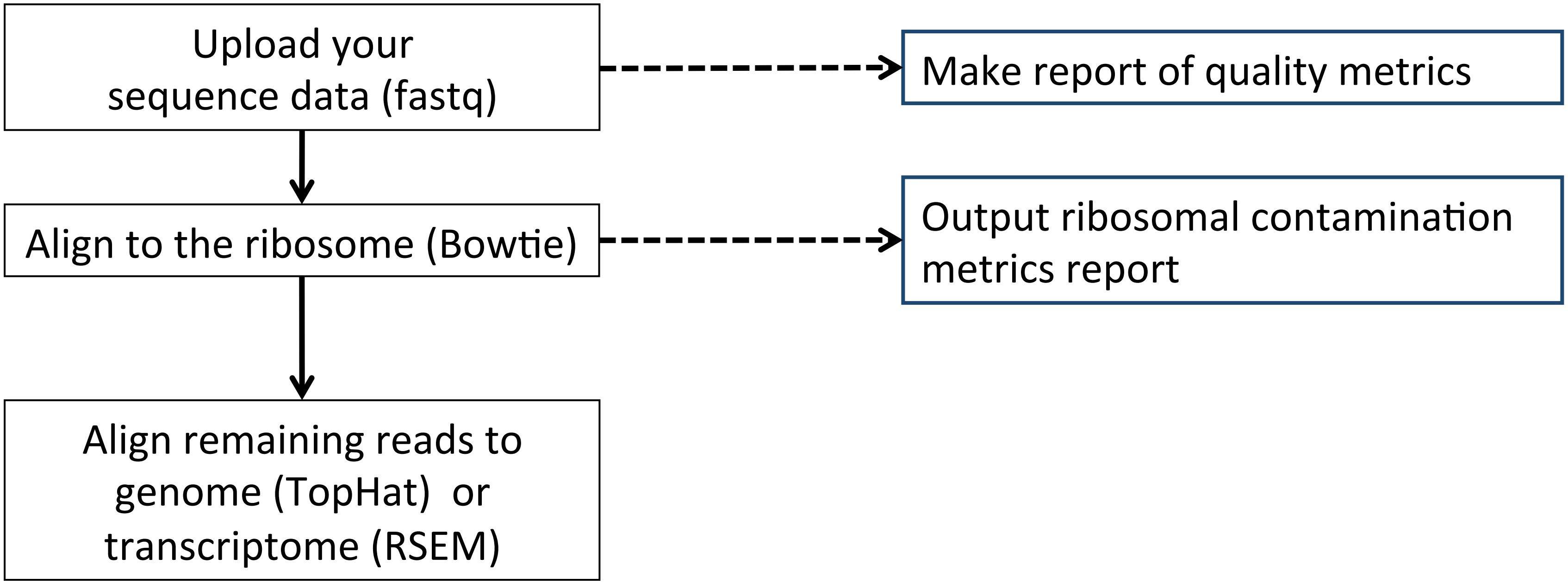




Mapping RNA-Seq reads: Exon-first spliced alignment (e.g. TopHat)



Short read alignment



Upload your sequence data (fastq)

Make report of quality metrics

Align to the ribosome (Bowtie)

Output ribosomal contamination metrics report

Align remaining reads to genome (TopHat) or transcriptome (RSEM)

```
tophat2 --library-type fr-firststrand --segment-length 20 \
-G  genome.quantification/ucsc.gtf -o  tophat/th.quant.ctrl1 \
genome.quantification/mm10 fastq.quantification/control_rep1.1.fq \
fastq.quantification/control_rep1.2.fq
```

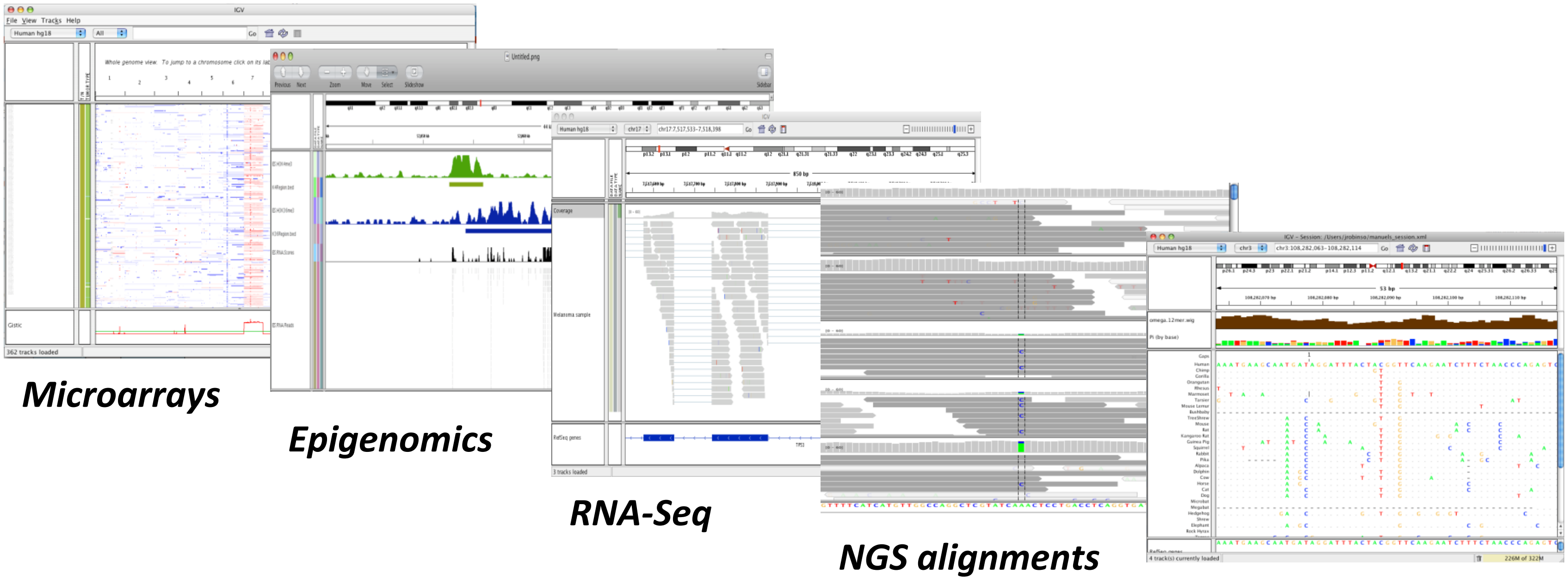
```
/project/umw_biocore/bin/igvtools.sh count -w 5 tophat/th.quant.ctrl1.bam \
tophat/th.quant.ctrl1.bam.tdf genome.quantification/mm10.fa
```

IGV: Integrative Genomics Viewer

A desktop application

for the visualization and interactive exploration

of genomic data



Microarrays

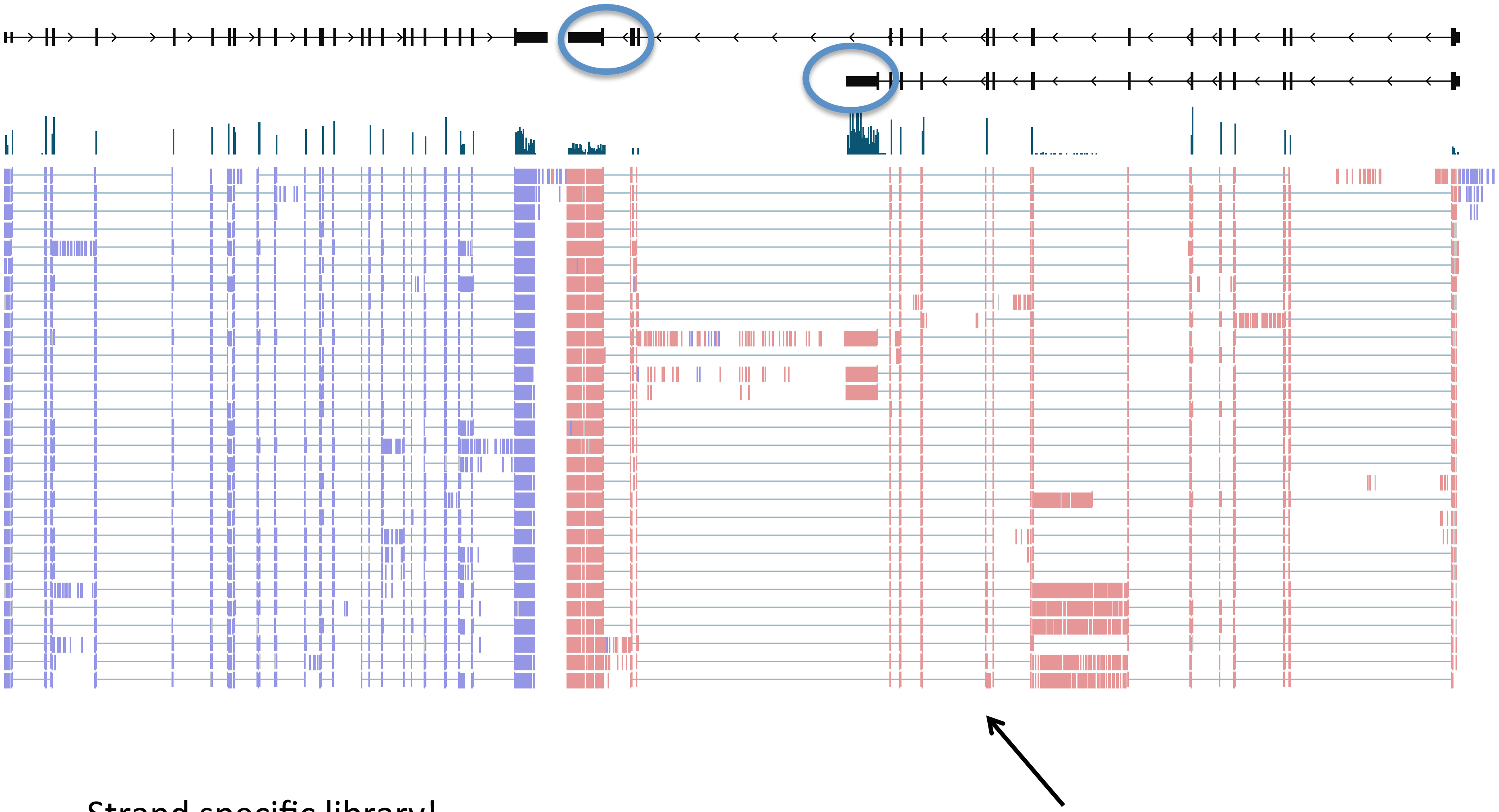
Epigenomics

RNA-Seq

NGS alignments

Comparative genomics

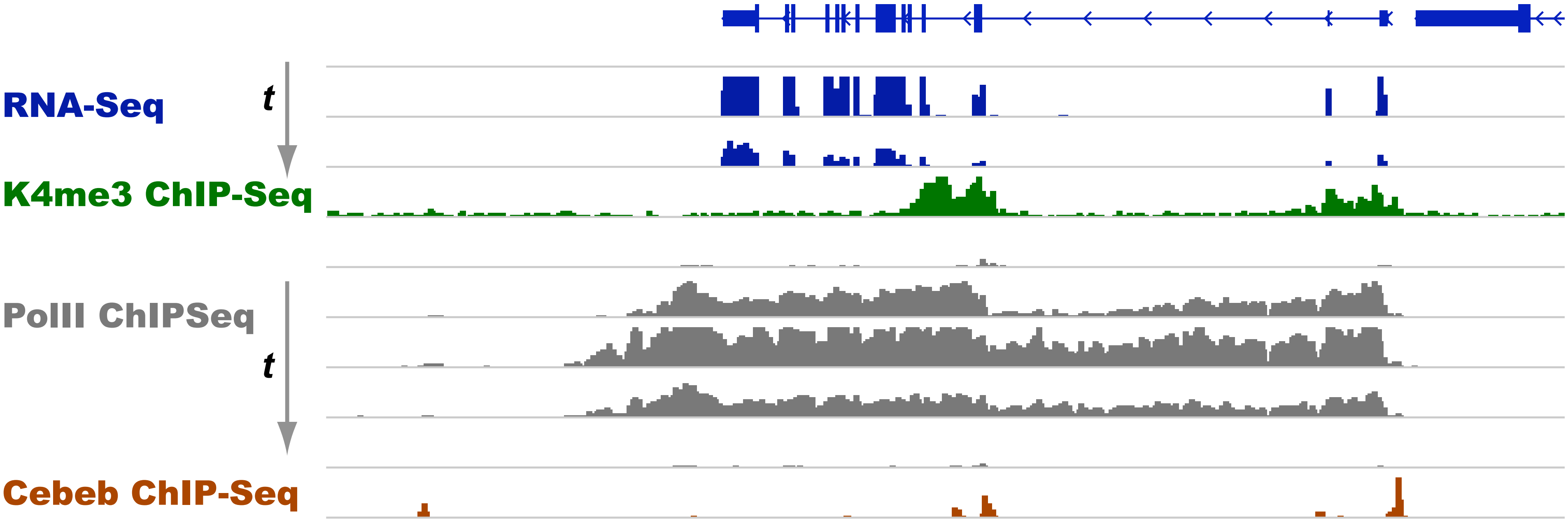
Visualizing read alignments with IGV — RNASeq



Strand specific library!

Gap between reads spanning exons

Visualizing read alignments with IGV — zooming out


RNA-Seq
t
K4me3 ChIP-Seq
PolII ChIPSeq
t
Cebeb ChIP-Seq