Analysis of RNA-Seq Data

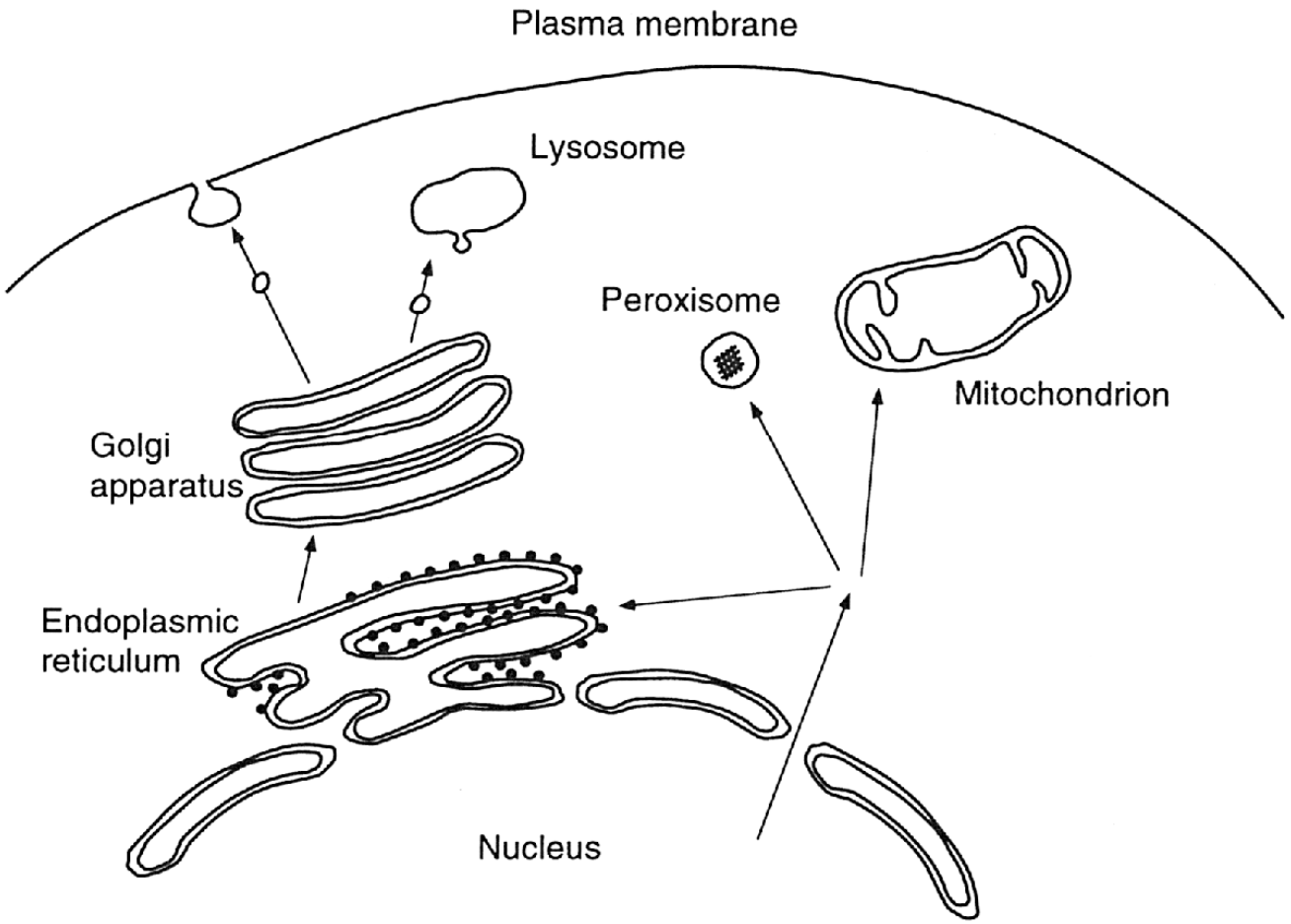
Wing Hung Wong
Stanford University

JSM-2010

Outline

- **Scientific background**
- Mapping of reads
- Read rates modeling
- Quantification of expression
- Splice junction discovery
- Isoform discovery
- Future outlook

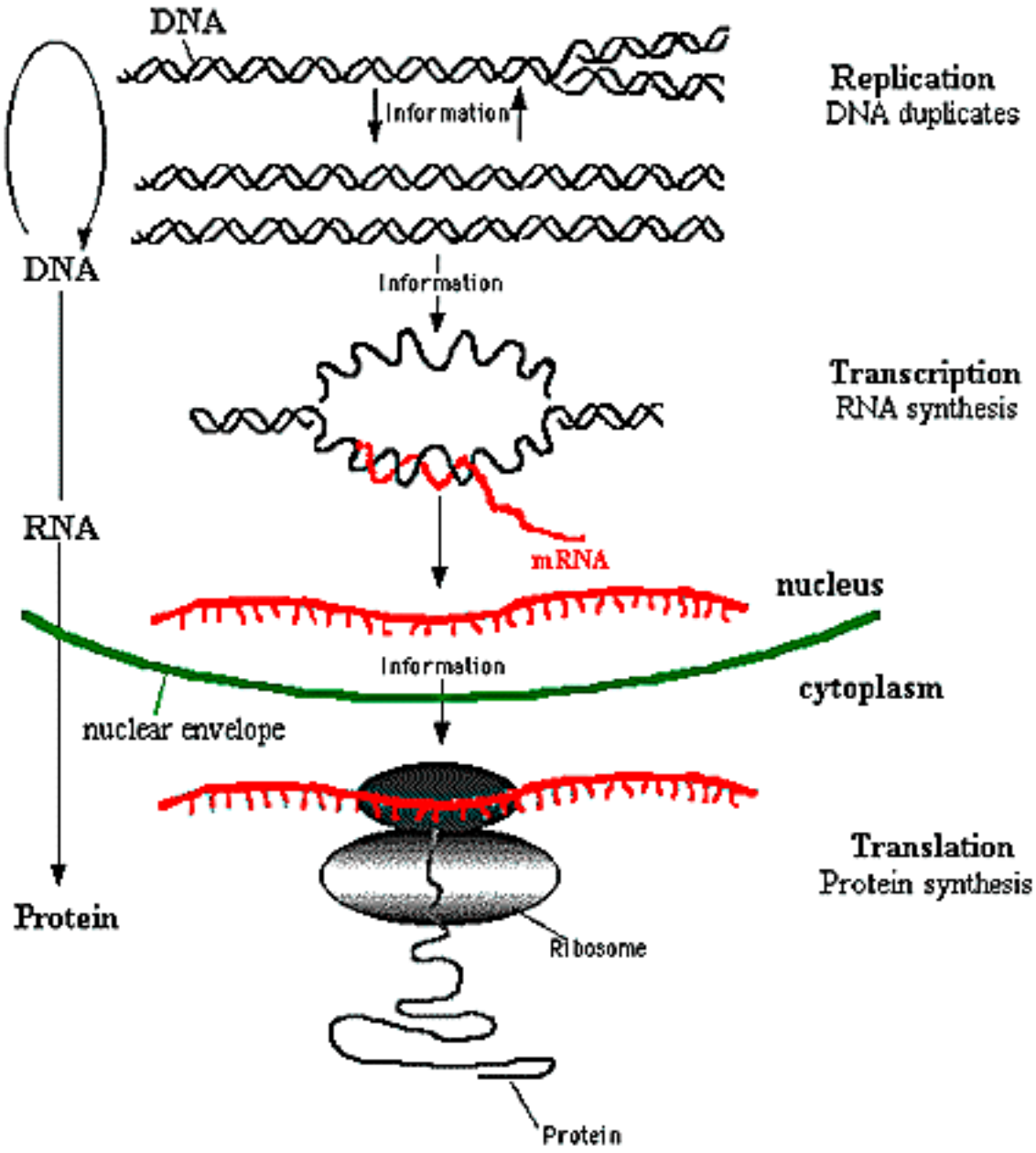
Cells are basic units of life



Schematic illustration of a eukaryotic cell

Basic working of a cell

- DNA contains genetic information
- Proteins (with RNA, lipids,..) self-assemble into the functional components of the cell.
- DNA replicates during cell division, so genetic information is passed to daughter cells
- Central dogma dictates how genetic information is utilized



The Central Dogma of Molecular Biology

Different proteins may be made in different cell types

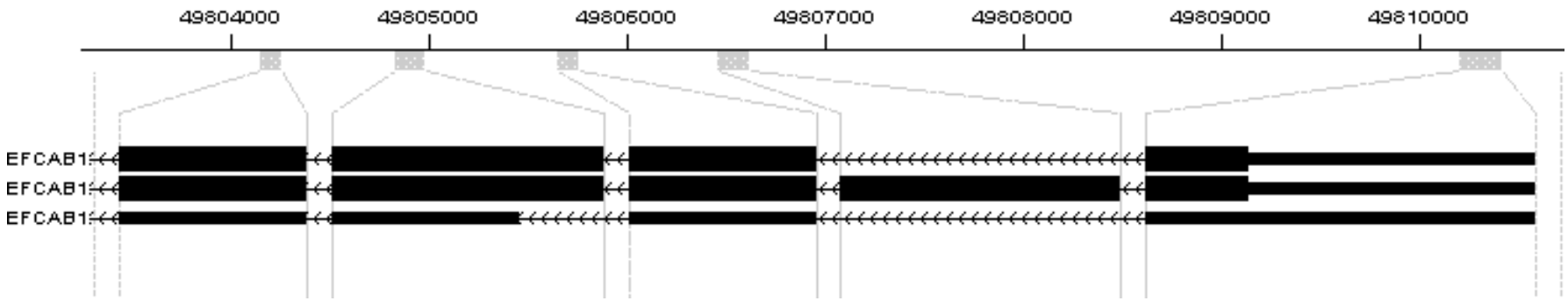
- Hemoglobin in red blood cells
- Myosins in muscle cells
- Albumin in liver cells

Cell types are different mainly because of differential gene expression.

In particular, gene not needed are not transcribed

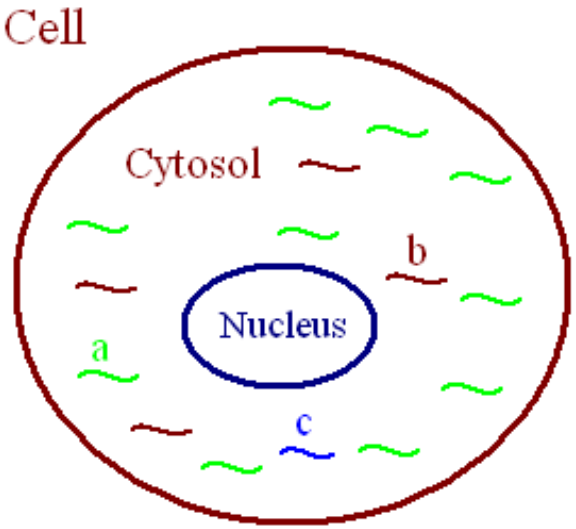
Alternative splicing

Multiple mRNA “isoforms” may be produced from the same genetic locus



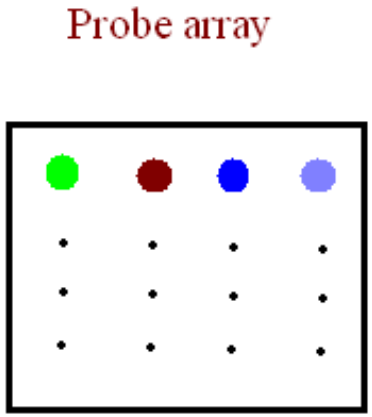
This allows a single gene to create multiple proteins

Measurement of gene expression by microarray

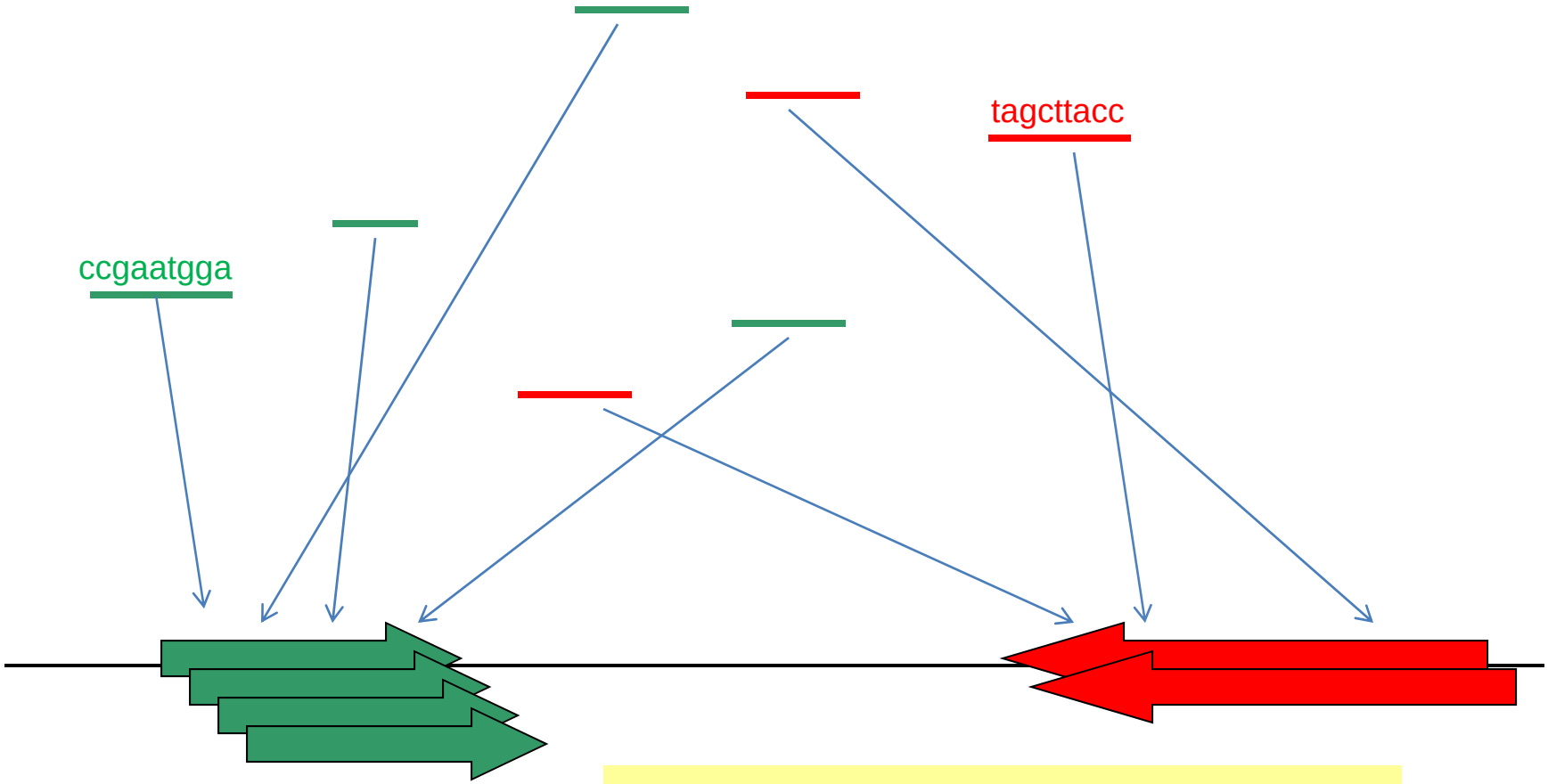


Expressoin pattern

Gene	Level
a	high
b	Medium
c	low
⋮	⋮



Measurement of gene expression by RNA sequencing



Need large sequencing capacity!

Revolution in sequencing

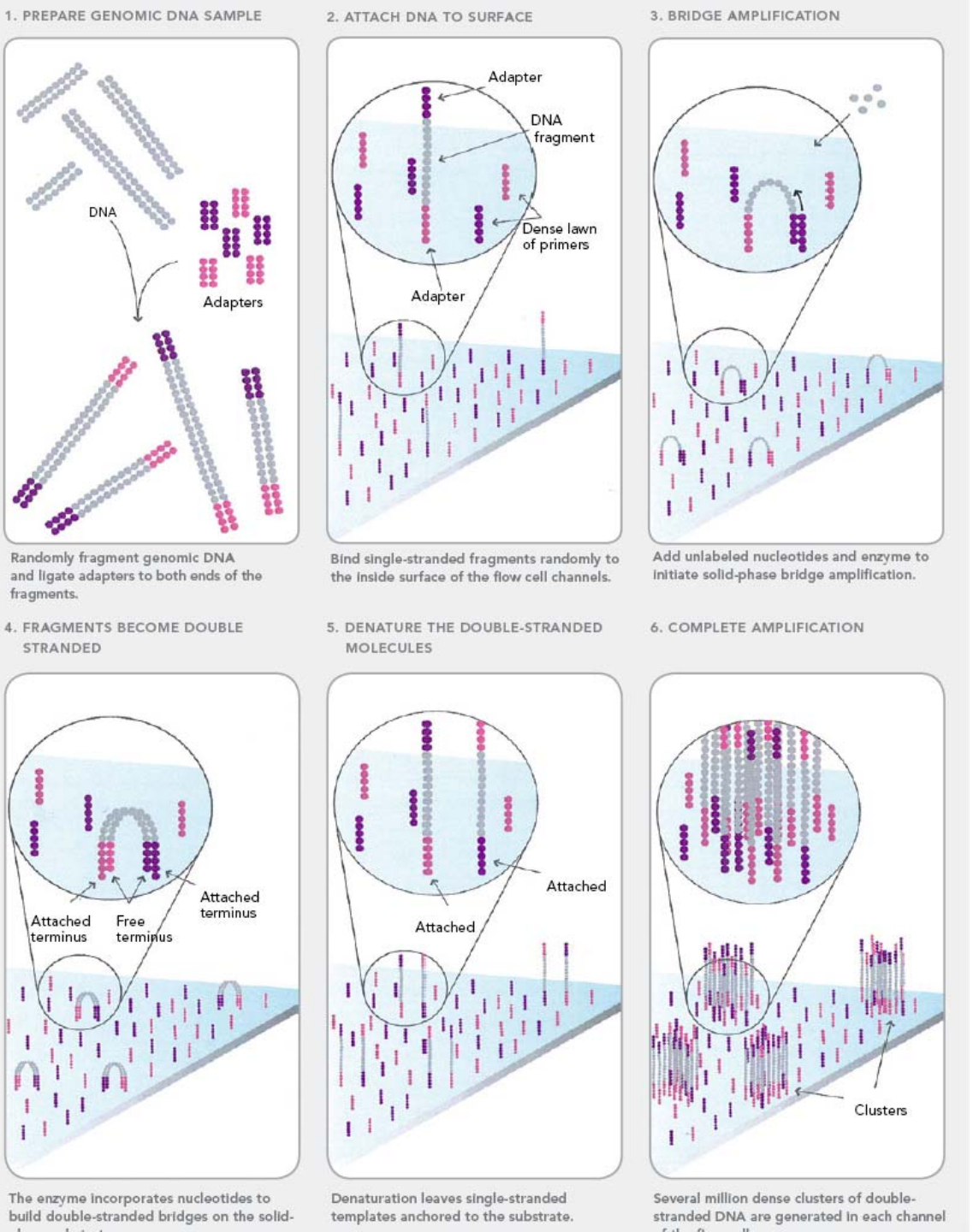
- Starting around 2004, new technologies have increased sequencing capacity at a rate faster than Moore's law.
- In 2008, a Solexa run could produce about 48 million x 32 bp . Just two years later, it is 480 million x 200 bp.
- RNA-Seq allows us to leverage this capacity for transcriptome analysis.

Outline

- Scientific background
- **Mapping of reads**
- Read rates modeling
- Quantification of expression
- Splice junction discovery
- Isoform discovery
- Future outlook

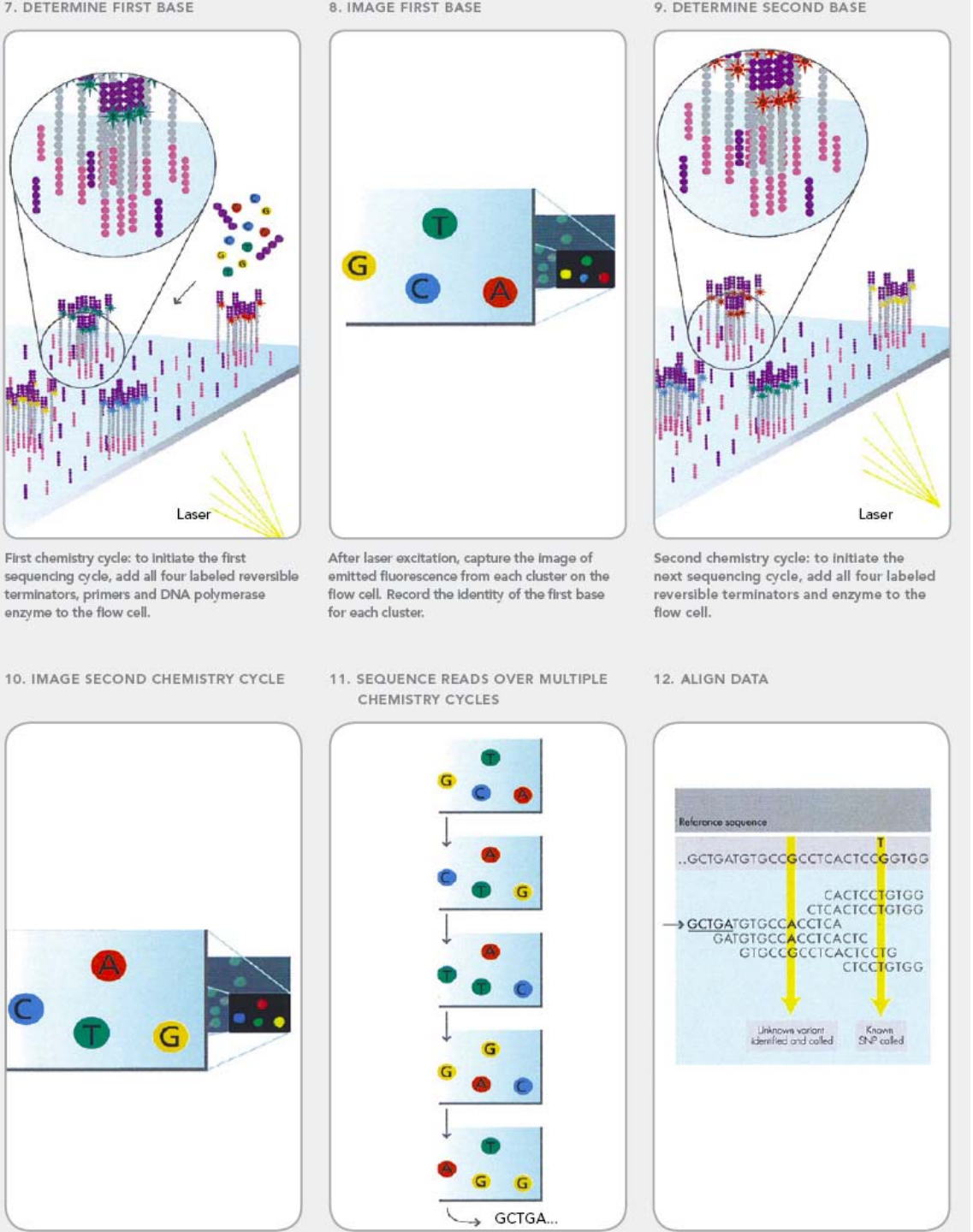
Solexa sequencing

Monitor single base extension by imaging



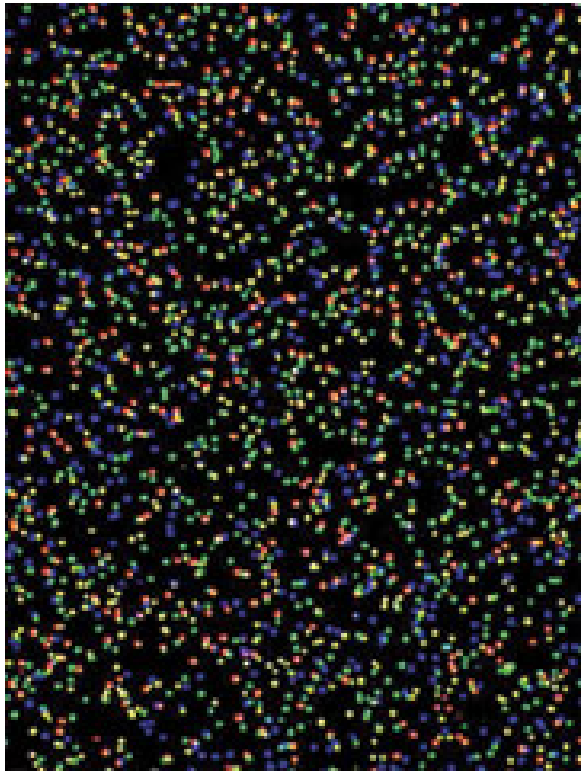
From: www.illumina.com

Sample preparation to imaging cycle takes about 3-4 days



From: www.illumina.com

What does the data look like?



Base calling

AAAAATCTCTTCCTGAACCATTCAGAAAATGC
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AACAGACCTAAAATCGCTCATTGCATATTCTT
AACCAGGCGACCTGCGACTCCTTGACGTTGAC
ATGTTAGGGTTGTACGGTAGAACTCCTATTAT
ATTGCCCAGAAAGTACCTGAGCTATCAGTGAT
ATCCCGATCCCGGTTACAGAGTCCATTGTAGA
ACCACCCAACAATGACTAATCAAACTAACCTC
ATGGGGGAAATATTGCAATTATGTAAAGGTAA
ATGTTTAAAAGTCCACTTTTAAAACTATATTT
ATATAACTCTCTTCCCTCTCACTCTTTCTCTC
AGGGAACTACTCCCACCCTGGAGCCTCCGTAG
AAAAGATATATATATATATATTCATAATTA
AGTCGACCCTGCACCTGGTCCTGCGTCTGAGA
ATTTGGTGAGTAATTAAAGAGAGTAGTAGCAT
GGTCTGTTTGTCGTATGCCGTCTTCTTCTTTT
ATTGAAAGAAGTCTTTCTAGAAATGTTAAATA
AGGGACTGAAGCTGCTGGGGCCATGTTTTTAG
AGAAAATATTAAAATCTTTGAAGAAGAAGAAG
AAGGGGATTTAGAGGGTTCTGCGGGCAAATTT
AGAACCCTCCATAAACCTGGAGTGACTATATG
AATAAGTCGGTTCAGGAGATCCAAGGAACCTT
ATTGGGTTTGGCTGTATCCCACCCCGTTACAA
CGGGGATAAGTGTGGTTTCGAAGAAGATATAA

Stages of data analysis

- Stage 1:
  - Base calling (Illumina, ABI, Phil Green)
- Stage 2:
  - Sequencing mapping (for known genome)
  - SNP calling, variation detection (for known genome)
  - De novo assembly (for unknown genome)
- Stage 3:
  - Gene transcription analysis (for RNA-Seq)
  - Discovery of novel splices & isoforms
  - Comparative analysis, etc

Sequence mapping

Find all the matches for a read in the genome

A DNA Sequence: ACATAGGATCATGAAGTACCCATATCTAGTGGG

reads: AGGA, CATC, ATAT, TTTG, GTGT

Matched Results: ACATAGGATCATGAAGTACCCATATCTAGTGGG

Perfect Match: AGGA ATAT

1bp Mismatch: CATC CATC CATC GTGT

Efficiency is crucial: >200 millions reads per run

Basic approaches to mapping

- Index the genome (or the reads)
- Mapping algorithms
  - Seed-based algorithms (BLAT)
  - Pigeonhole principle
  - Suffix tree/array, BWT (Burrows-Wheeler transform)

Simple example

An example of mismatch = 1

Reads:

ATTCCG

CGTATG

TTCCTT

GATATT

AAATGC

GGACTA

TACTGT

Split into two parts

ATT-CCG

CGT-ATG

TTC-CTT

GAT-ATT

AAA-TGC

GGA-CTA

TAC-TGT

Simple example

Reads:

ATT-CCG

CGT-ATG

TTC-CTT

GAT-ATT

AAA-TGC

GGA-CTA

TAC-TGT

Sorting by each part

List 1

AAA-TGC

ATT-CCG

CGT-ATG

GAT-ATT

GGA-CTA

TAC-TGT

TTC-CTT

&

List 2

CGT-ATG

GAT-ATT

ATT-CCG

GGA-CTA

TTC-CTT

AAA-TGC

TAC-TGT

Simple example

Query sequence

AATTGC

split into two parts

AAT-TGC

look up in both lists, find match

List 1

AAA-TGC
ATT-CCG
CGT-ATG
GAT-ATT
GGA-CTA
TAC-TGT
TTC-CTT

&

List 2

CGT-ATG
GAT-ATT
ATT-CCG
GGA-CTA
TTC-CTT
AAA-TGC
TAC-TGT

Some short-read mapping tools

Software	Max. mismatches	Gap	Max. read length	Report All matches	Reference
ELAND	2	No	32	N	A. Cox (Illumina)
SOAP	2	1	60	N	R. Li (2008)
RMAP	>4	No	64	N	A. D. Smith (2008)
SeqMap	>4	>4	>200	Y	H. Jiang (2008)
ZOOM!	>4	1	64	Y	H. Lin (2008)
MAQ	3	No	64	N	H. Li (2008)
Bowtie	**3**	**No**	**>200**	**Y**	**B. Langmead (2009)**

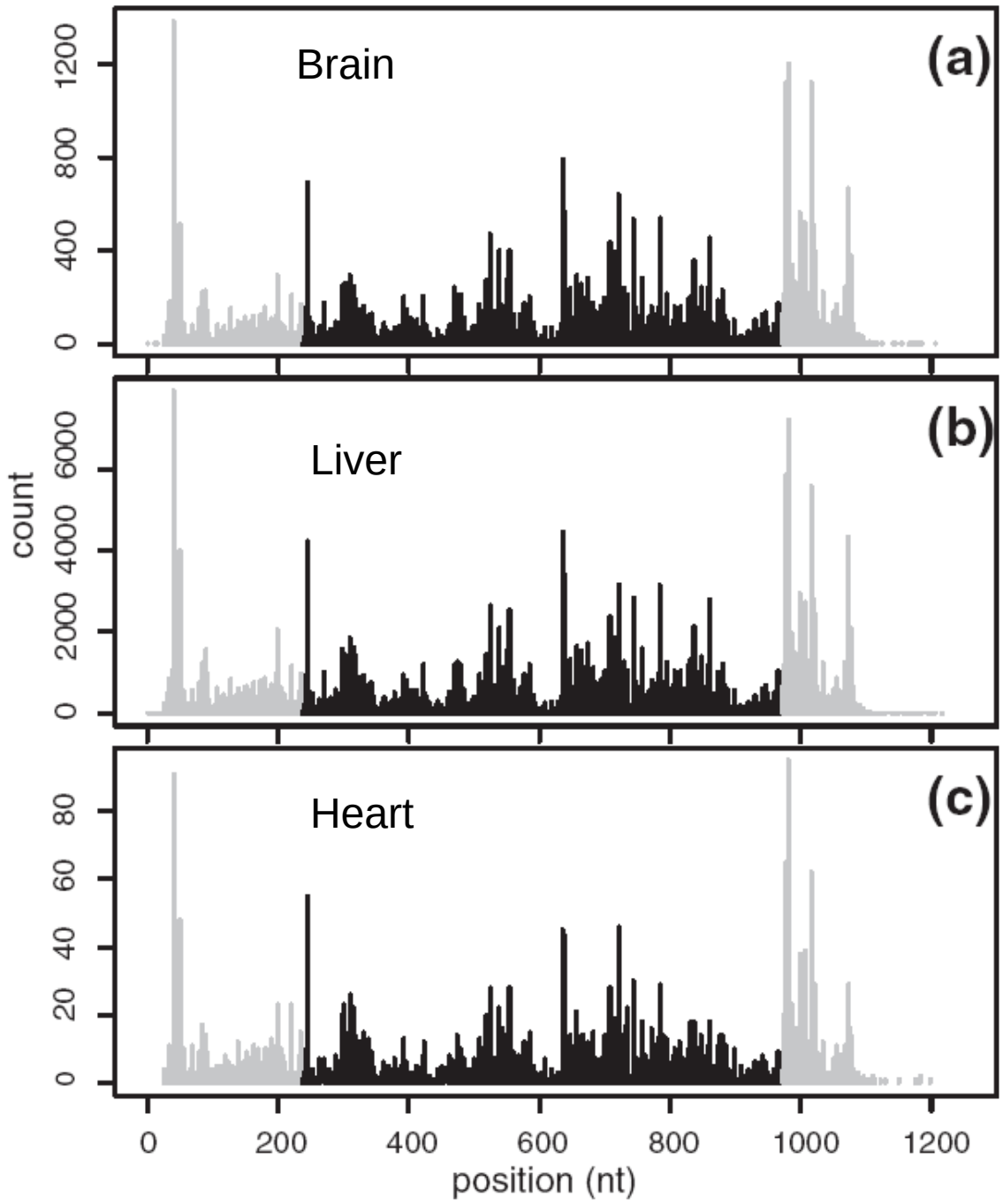
For 100 million reads of 50 bp each, mapping to human genome:
BOWTIE takes ~ 40 minutes on a 24-core server with large memory.
BLAT (with short seed length to ensure sensitivity) takes ~400 minutes

Challenge: storage and processing power

Outline

- Scientific background
- Mapping of reads
- **Read rates modeling**
- Quantification of expression
- Splice junction discovery
- Isoform discovery
- Future outlook

Read are non-uniformly distributed, but same pattern across tissues with large differences in expression levels.



Example: read counts along the transcript of the Apoe gene in mouse.

Modeling read rates:

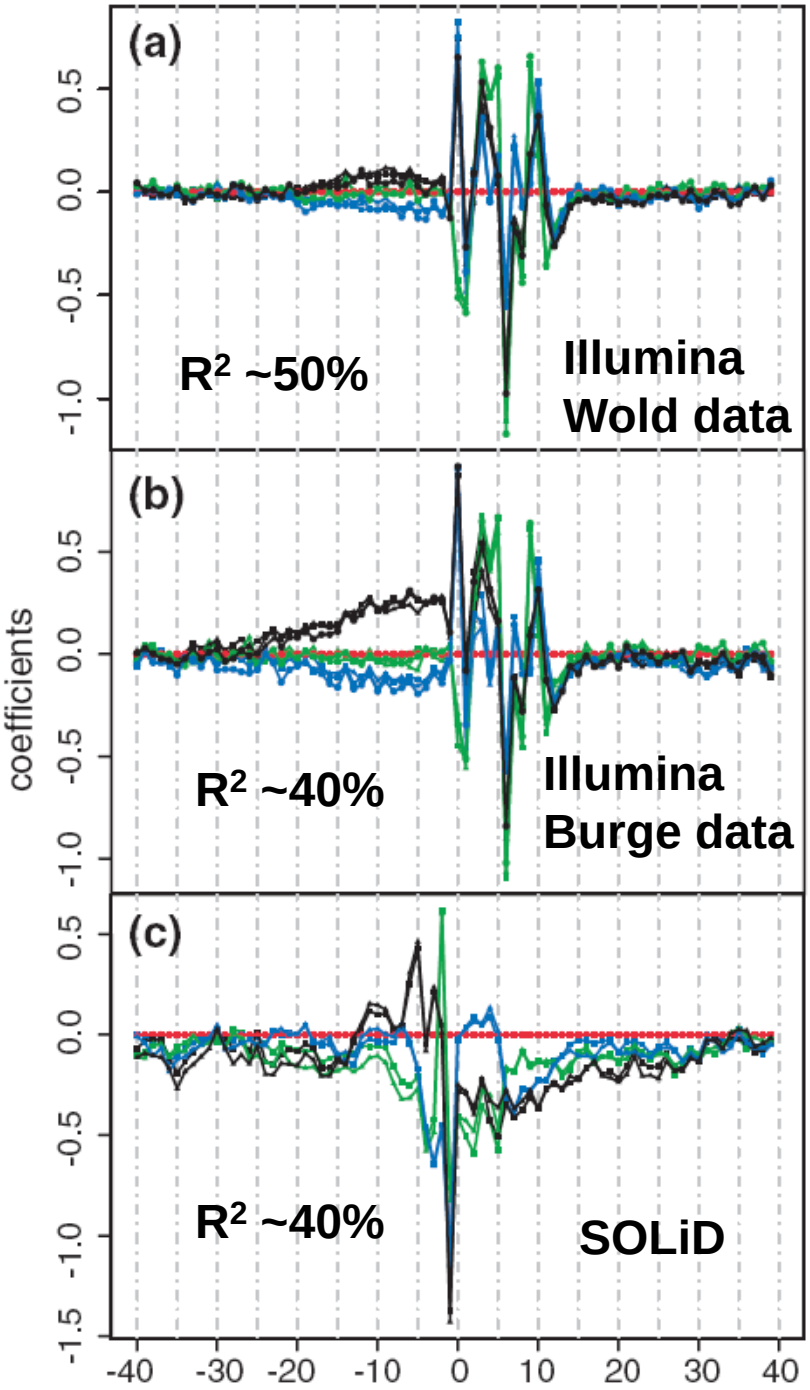
n_{ij}= count at nucleotide *j* of gene *i*

Assume $n_{ij} \sim Poisson\ (\mu_{ij})$

$$\log(\mu_{ij}) = \nu_i + \alpha + \sum_{k=1}^{K} \sum_{h \in \{A,C,G\}} \beta_{kh} I(b_{ijk} = h)$$

Results (Jun Li et al 2010):

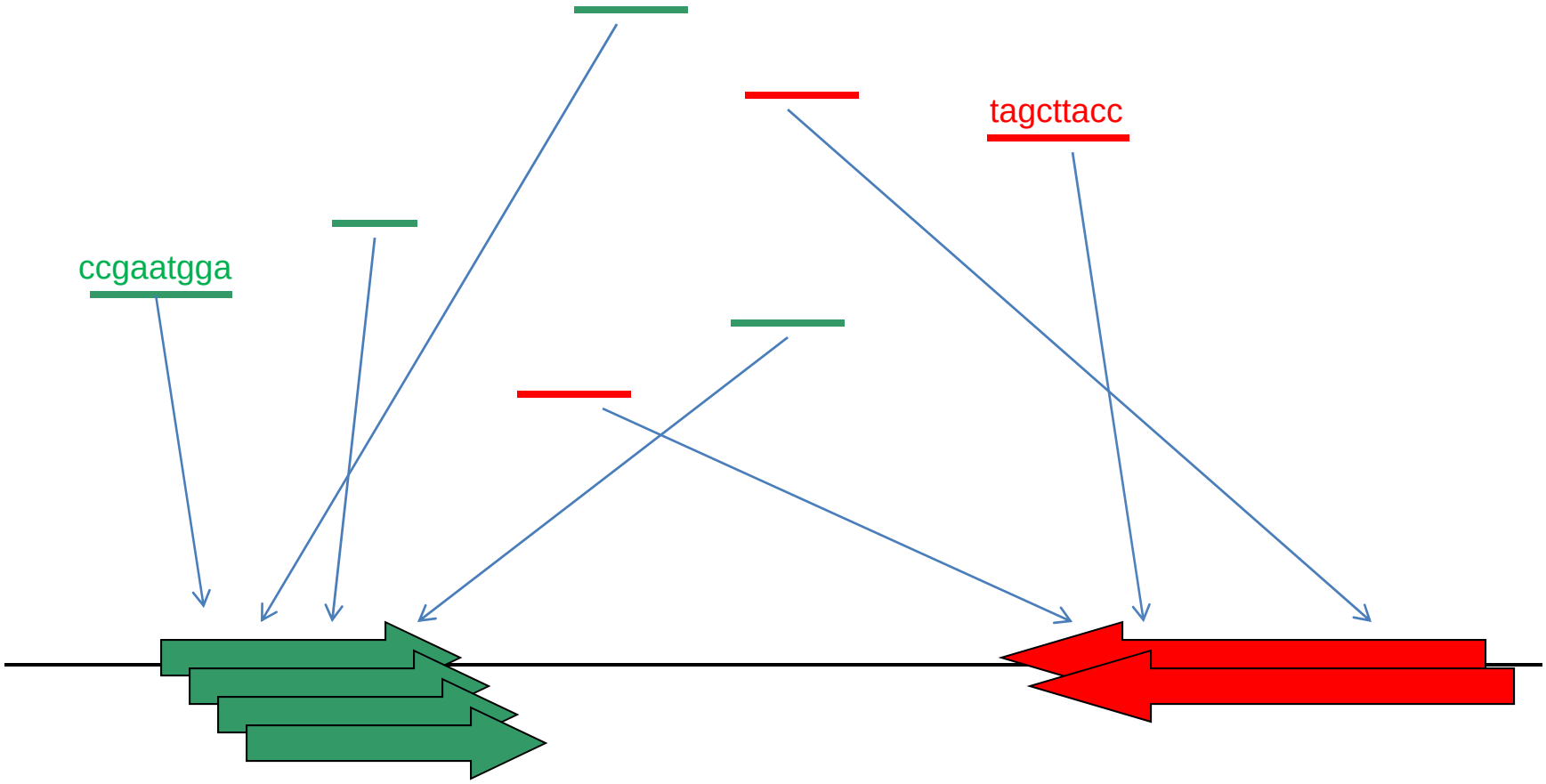
- Local sequence predicts read rate variation. (see also Hansen et al 2010)
- Model is platform dependent
- Nonlinear models (e.g. MART) predict even better (R^2 ~ 50%-70%)



Outline

- Scientific background
- Mapping of reads
- Read rates modeling
- **Quantification of expression**
- Splice junction discovery
- Isoform discovery
- Future outlook

Expression level is revealed by the counts of reads mapped to the gene



RPKM as gene-level expression index

- More reads mapped to gene if transcript is long
- More reads mapped to gene if sequencing is deep
- Expression index (Mortazavi et al 2008, Wold Lab)

  Let l = size of transcript in kb

  N= total # of mappable reads

  then the gene expression index is

  RPKM = (# reads mapped to gene)/ (l * N)

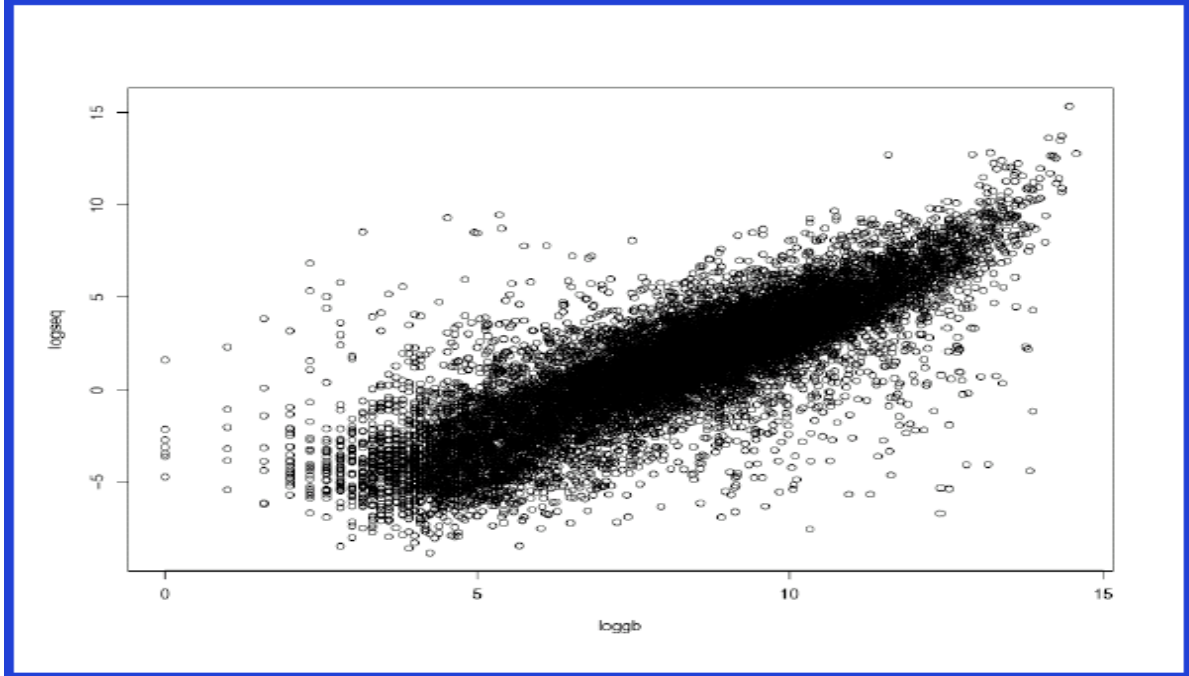
  "reads per kb per million reads"

  1 RKPM ~ 0.3 to 1 transcript per cell

Consistency with microarray

(Wold data, exon array indexes by Karen Kapur)

Log-Log Correlation Sequencing/Exon Array

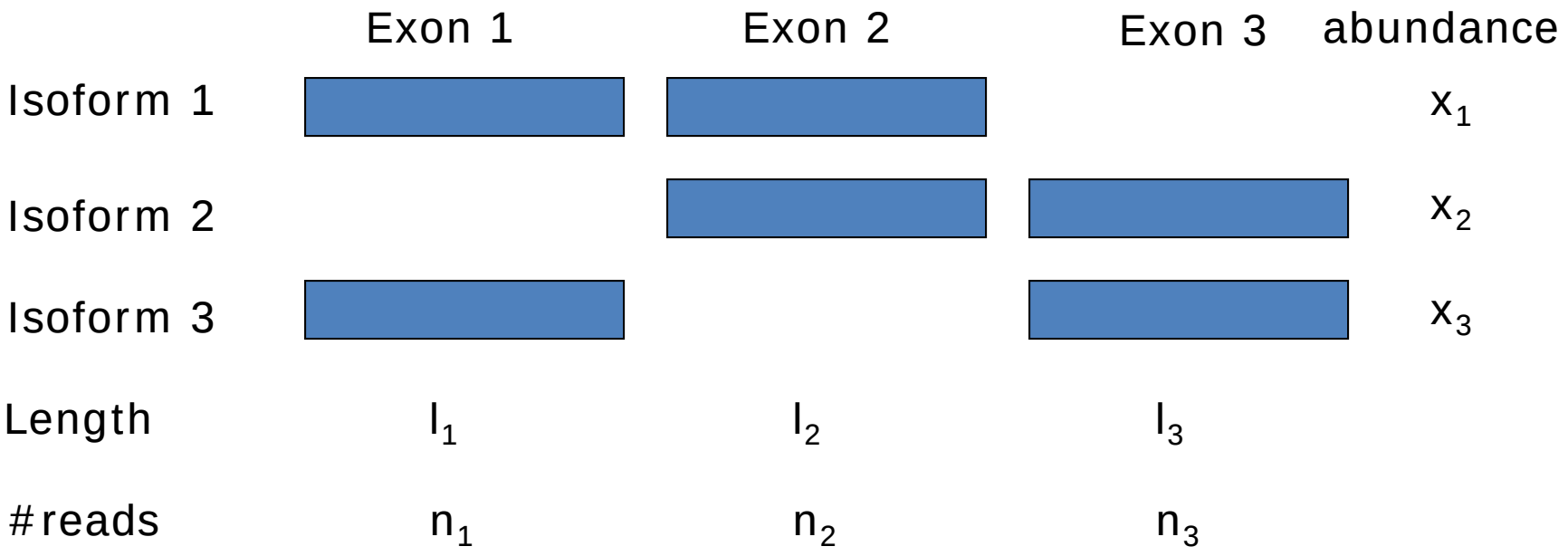


Tissue	Log-Log Correlation
Liver	0.8350
Muscle	0.8247
Brain	0.7566

Isoform expression estimation

- In the future, estimation experiments may be done separately from discovery experiments
- Assuming the set of isoforms is given, how to estimate the RPKM for each of the isoform?

Simple RPKM computation may fail in many cases



Model-based approach

(Jiang & Wong, 2008)

- Assume each read is sampled uniformly along the length of each transcript in the sample, and that longer transcripts are proportionally more likely to be sampled.

- Under this model, n_1, n_2, .. are independent Poisson variables.

- Draw inference on x_1, x_2,… from the likelihood or the posterior distribution

Concavity of log-likelihood

- Let A_{ij} = Indicator {isoform j contains exon i }

$$f = \log lik = \sum_i \left(n_i \log \sum_j A_{ij} x_j \right) - n \sum_i \left(l_i \sum_j A_{ij} x_j \right)$$

- Gradient

$$\frac{\partial f}{\partial x_k} = \sum_i A_{ik} \left(\frac{n_i}{\sum_{j \in A_i} x_j} - n l_i \right)$$

- Hessian

$$\frac{\partial^2 f}{\partial x_k \partial x_l} = -\sum_i A_{ik} A_{il} \left(\frac{n_i}{\left(\sum_{j \in A_i} x_j \right)^2} \right)$$

Concavity

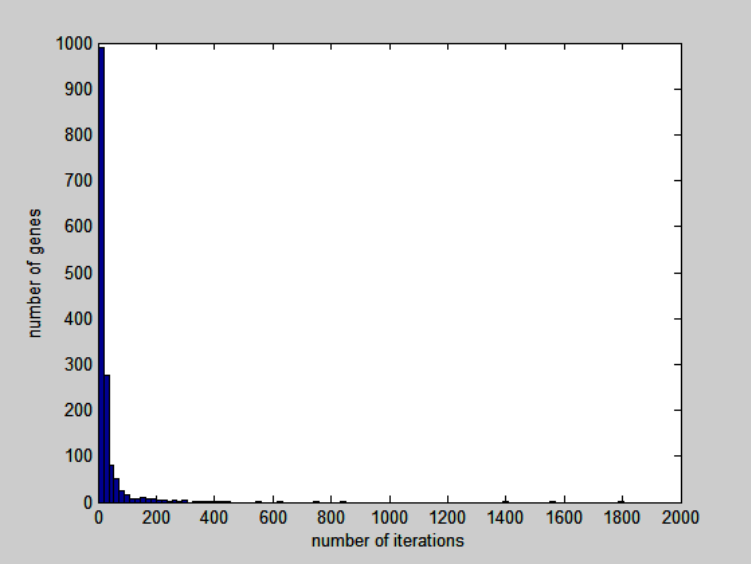
- Hessian in matrix form $$Hf = \frac{\partial^2 f}{\partial X^2} = -A' DA$$

- Where $A = \{A_{ij}\}$, D is a diagonal matrix, with
$$D_{ii} = n_i / \left(\sum_{j \in A_i} x_j \right)^2 \geq 0$$

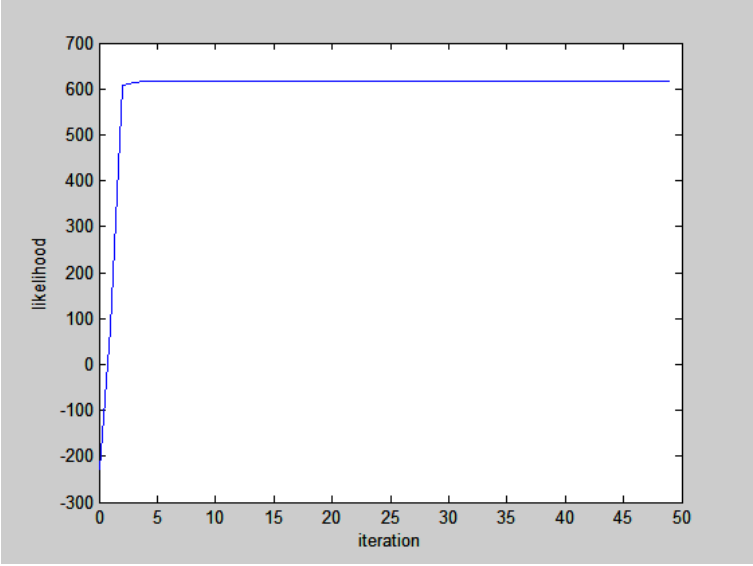
- Thus Hessian is negative semidefinite, and *f* is concave. It suffices to find local maximum

Numerical optimization

- Iterative method (hill climbing)
  - For the 1510 genes that have multiple isoforms, max(num_it) = 1805, mean(num_it) = 32.87, median(num_it) = 15

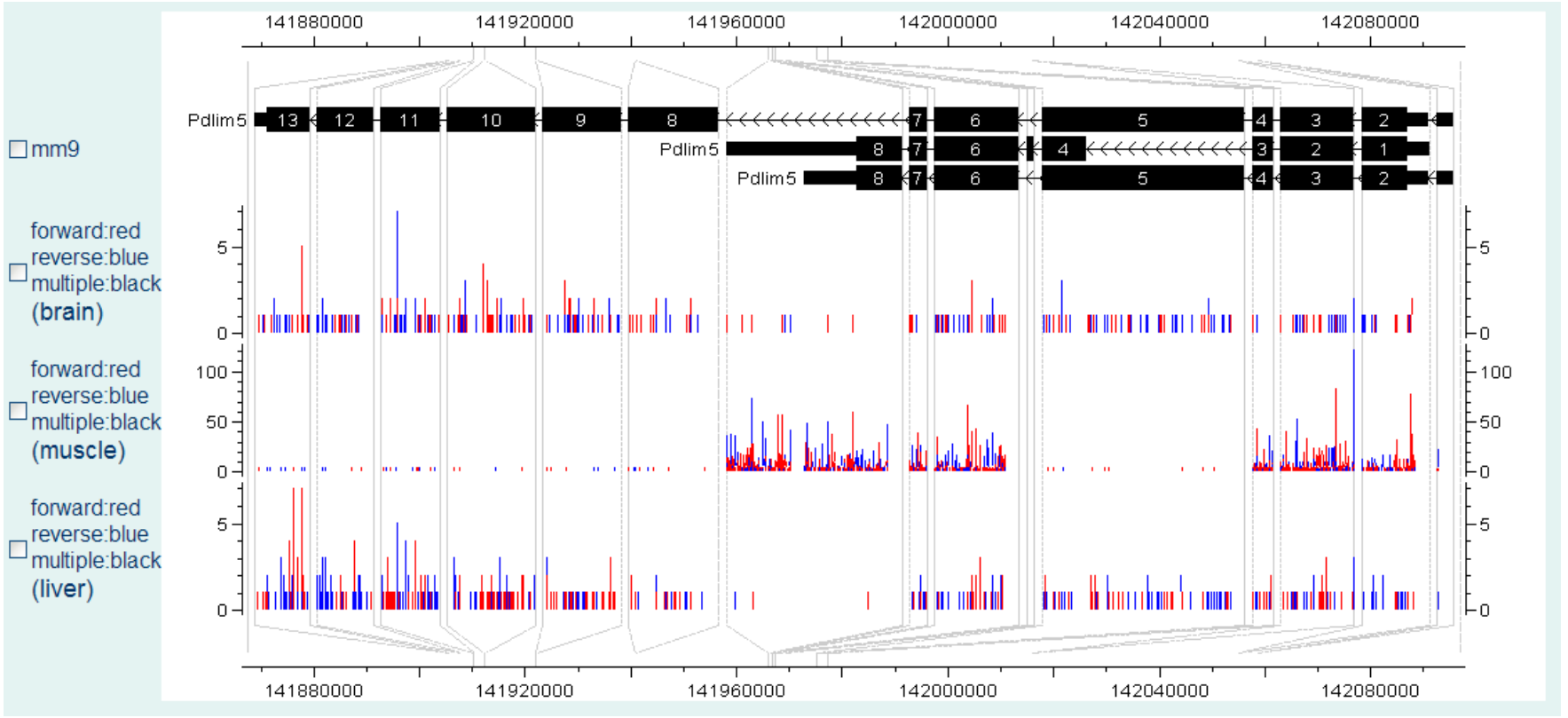


Histogram of iteration counts



Convergence profile for gene Rtn4

Example



Tissue	Isoform 1	Isoform 2	Isoform 3
Brain	5.05	0.42	0
Muscle	1.91	238.67	14.89
Liver	7.96	0.12	0

Statistical inference

- Multiple isoforms
  - Correlated expression
  - Asymptotics of the MLE

$$\hat{\theta} \sim N(\theta, I(\theta)^{-1})$$

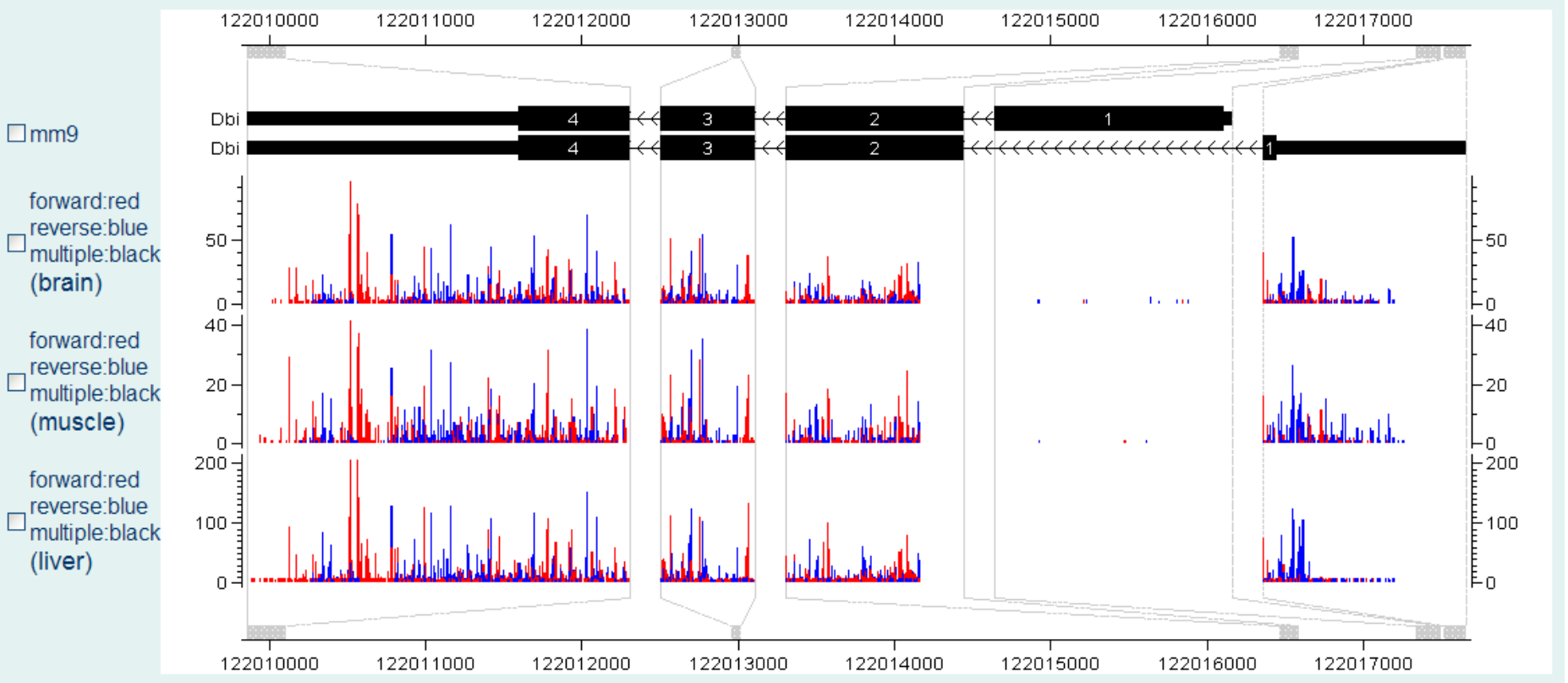
  - Fisher information matrix

$$I_{jk} = Cov\left(\frac{\partial}{\partial \theta_j} \log f(X;\theta), \frac{\partial}{\partial \theta_k} \log f(X;\theta) \right) = -E_\theta \left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X;\theta) \right]$$

Statistical inference

- Difficulty: when some isoform(s) are not expressed, Fisher Information becomes singular
- Our approach: Use importance sampling to draw from the posterior, starting with a proprosal density related to the asymptotic distribution
- Summarize marginal inferences for single or pairs of isoform expressions

Example – 95% probability interval



Tissue	Isoform 1	95% Interval	Isoform 2	95% Interval
Brain	3.87	(2.22, 6.76)	580.68	(559.52, 601.86)
Muscle	1.04	(0.39, 3.04)	330.64	(314.02, 347.24)
Liver	0.32	(0.08, 1.82)	1376.04	(1343.51, 1408.42)

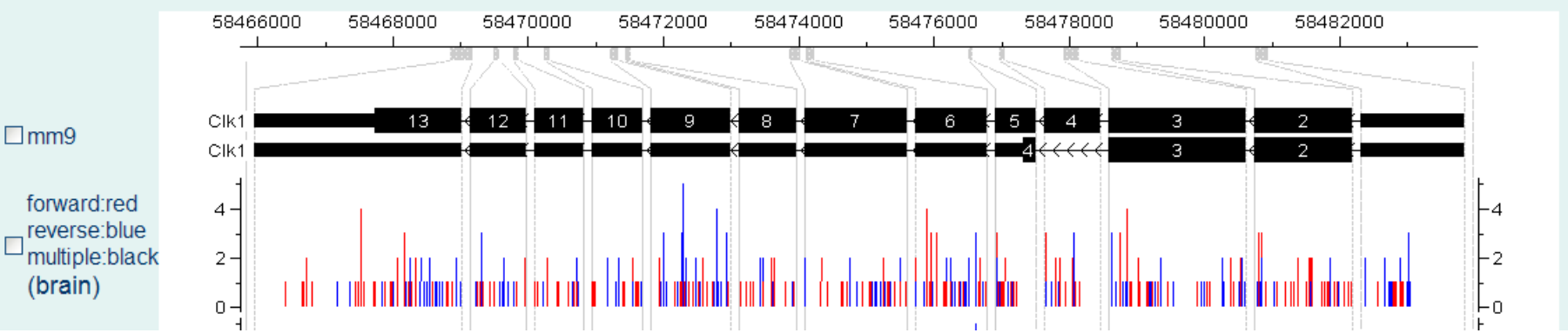
Gene level expression

- Gene expression is obtained by summing isoform expressions

$$g = \sum_i t_i, \text{ where } t = \hat{\theta}$$

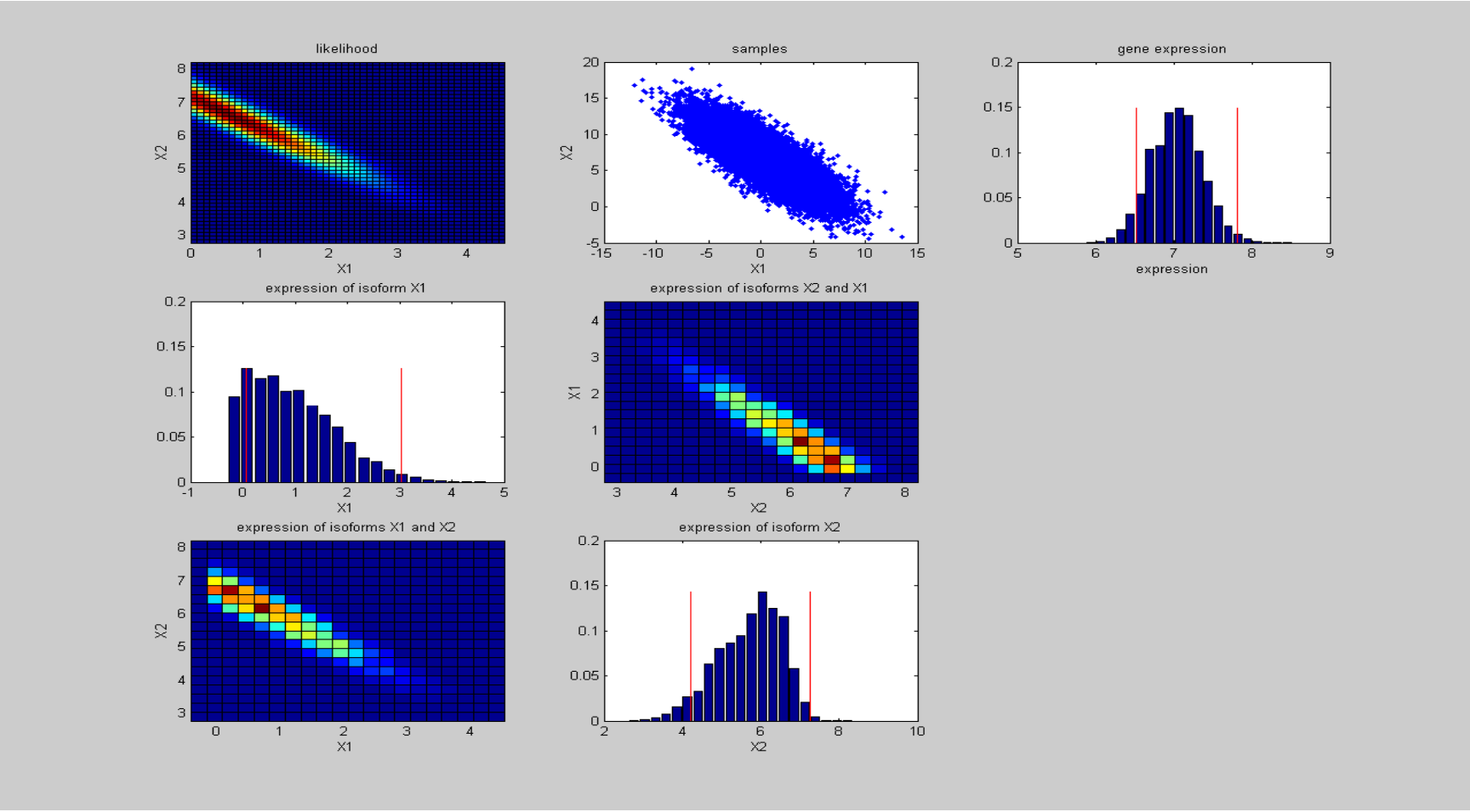
- Marginal posterior for g can be obtained from that of θ
- **In many cases we may have tight inference for the gene level expression but yet have great uncertainty about the expression for individual isoforms**

Example – gene level expression

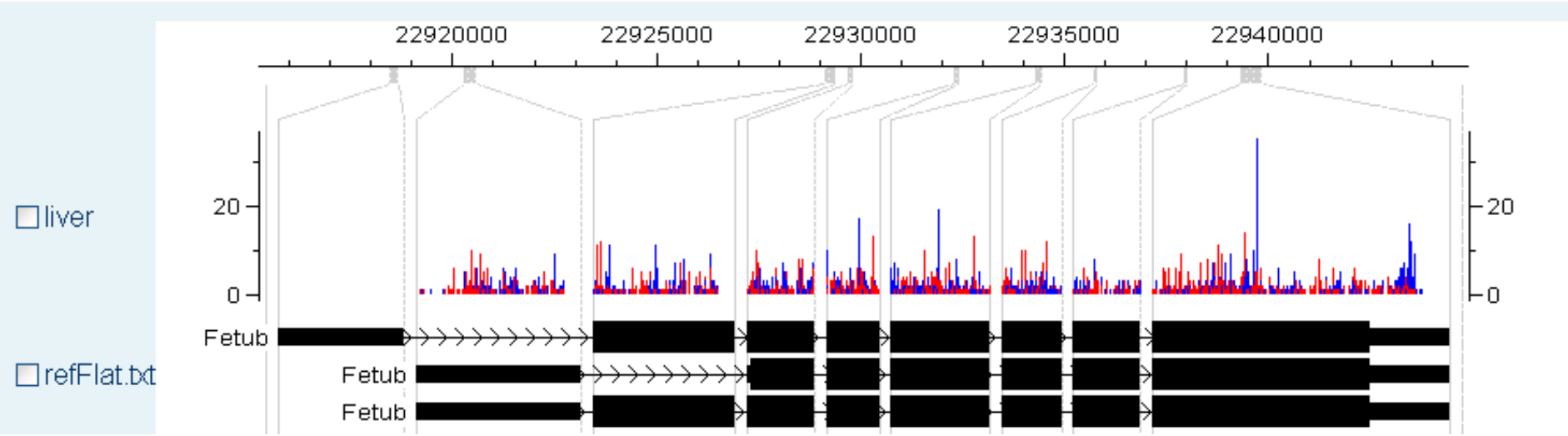


	Expression	95% interval
Isoform 2 (upper)	6.60	(4.20, 7.28)
Isoform 1 (lower)	0.48	(0.05, 3.01)
Gene level (Isoform 1 + Isoform 2)	7.09	(6.52, 7.84)

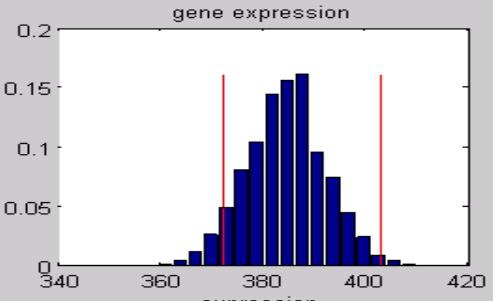
Marginal inference

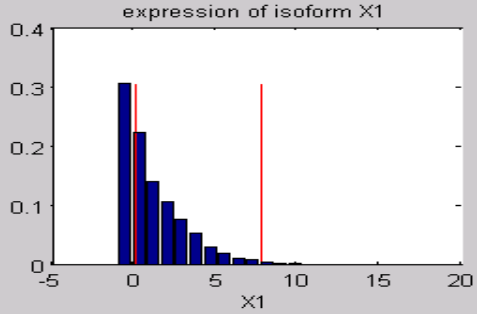


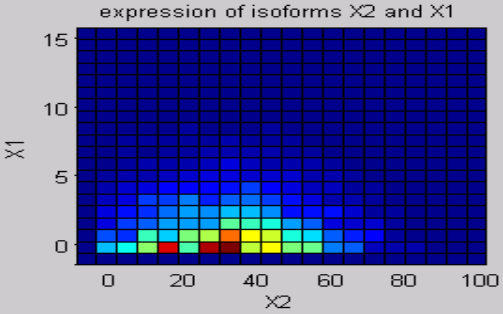
Another example

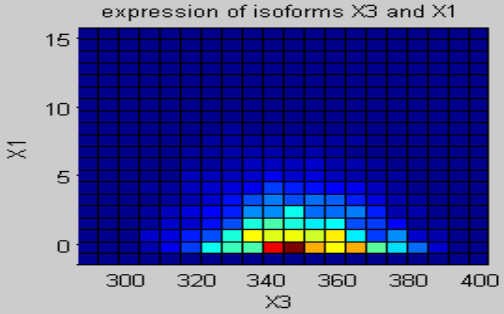


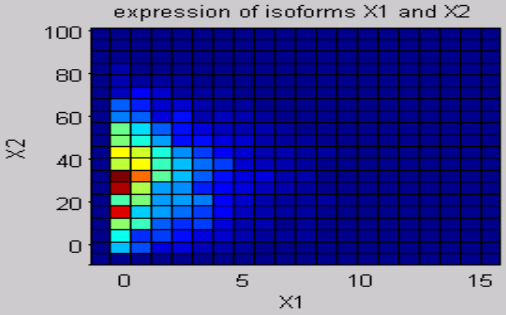
	Expression	95% interval
Isoform 1 (upper)	2.1	(0.05, 7.76)
Isoform 2 (middle)	35	(4, 71)
Isoform 3 (lower)	350	(316, 379)
Gene level	387	(371, 402)

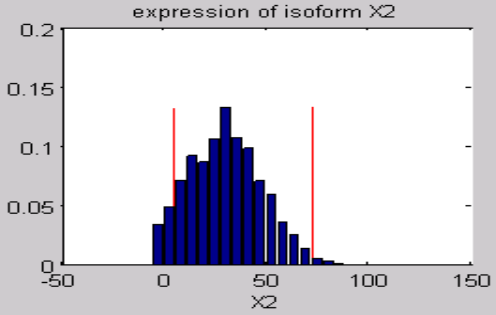

gene expression
expression

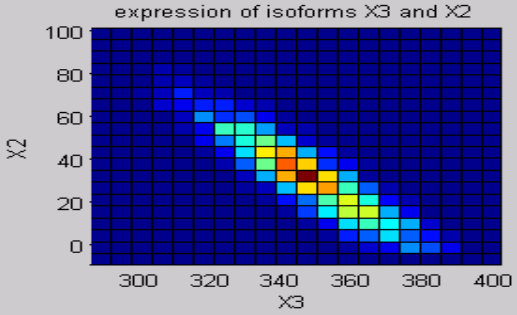

expression of isoform X1
X1

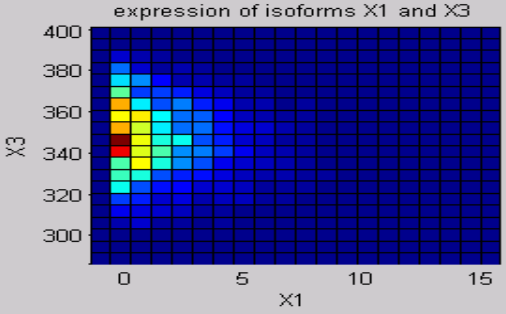

expression of isoforms X2 and X1
X1
X2

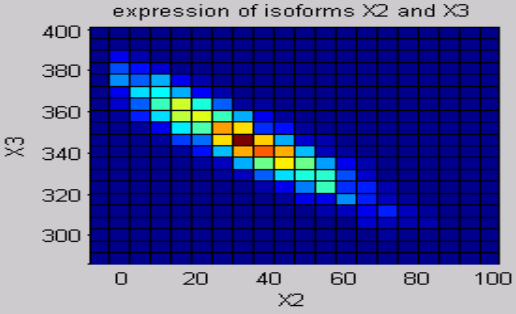

expression of isoforms X3 and X1
X1
X3

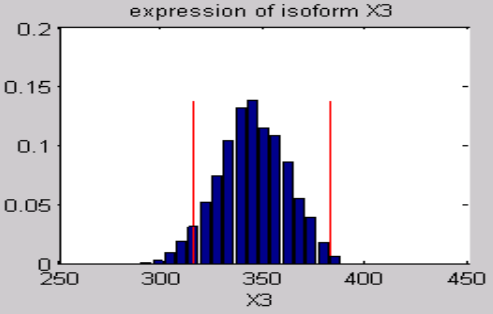

expression of isoforms X1 and X2
X2
X1


expression of isoform X2
X2


expression of isoforms X3 and X2
X2
X3


expression of isoforms X1 and X3
X3
X1


expression of isoforms X2 and X3
X3
X2


expression of isoform X3
X3

Outline

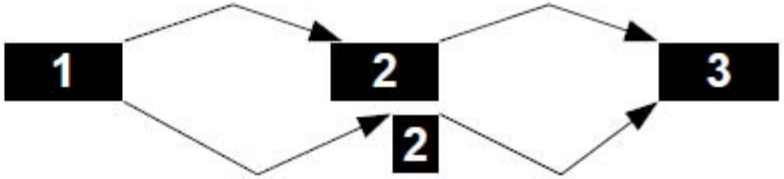
- Scientific background
- Mapping of reads
- Read rates modeling
- Quantification of expression
- **Splice junction discovery**
- Isoform discovery
- Future outlook

Discovery of novel isoforms

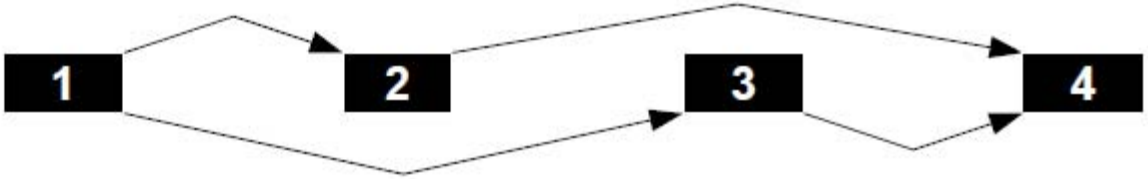
First step: detect splice junctions

Exon Skipping

Alternate Exon Length

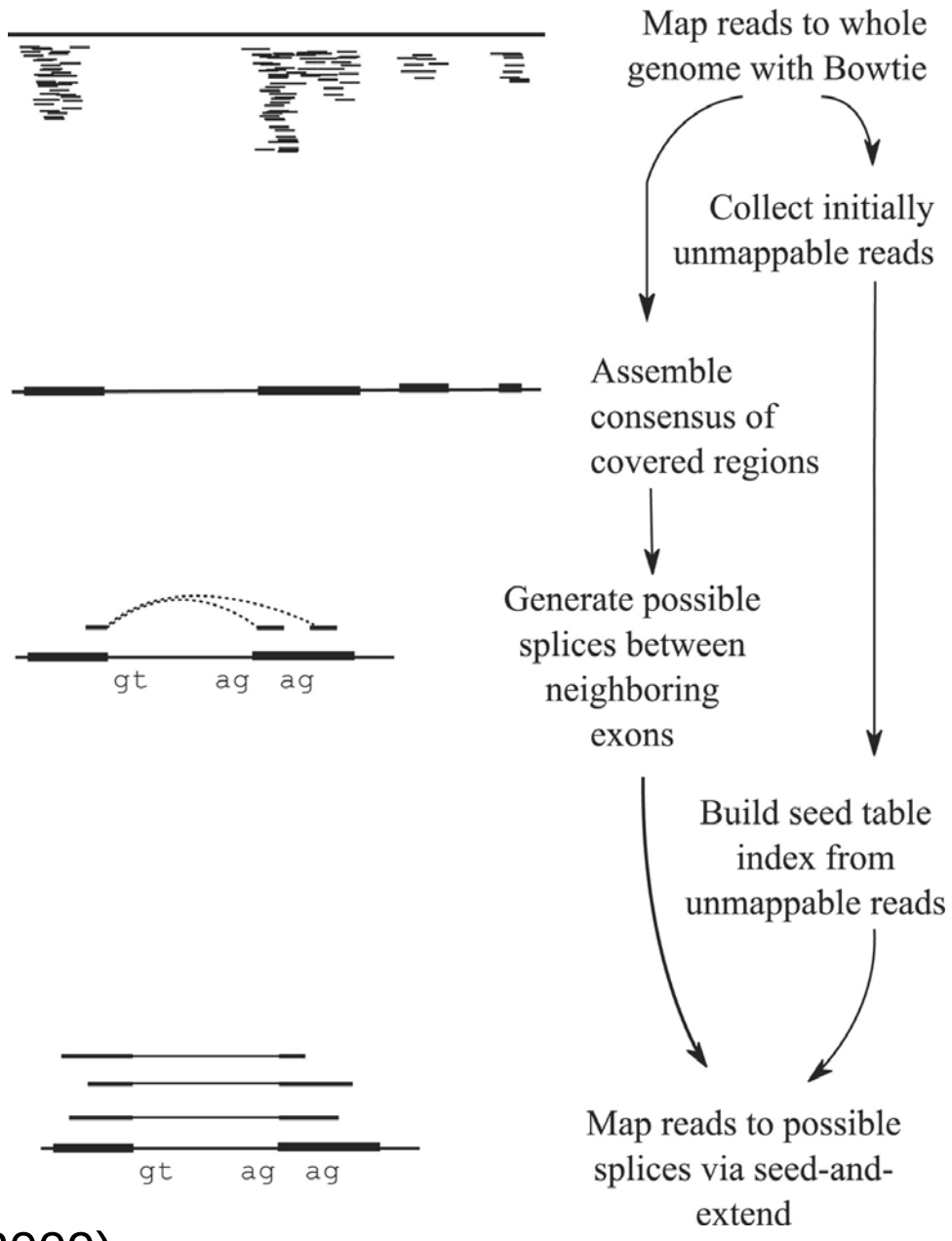


Mutually Exclusive Exons



TopHat Software

- Do not assume known annotation
- Putative exon definition by clustering mappable reads



Trapnell, et al. Bioinformatics (2009)

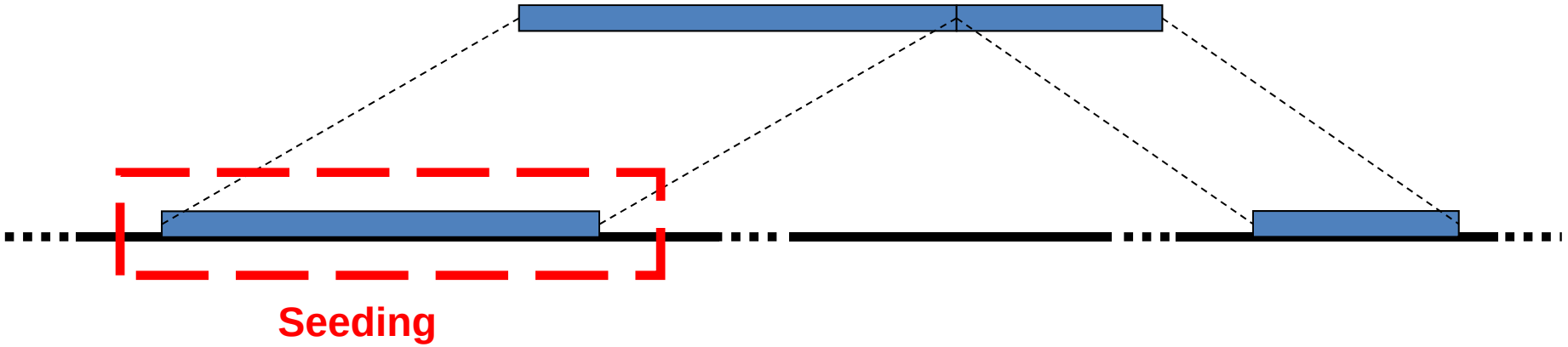
SpliceMap Software (Au et. el. NAR 2010)

- do not assume known annotations
- directly find split map of reads
- customizable to balance sensitivity/specificity
- fast

http://www.stanford.edu/group/wonglab/SpliceMap/index.html

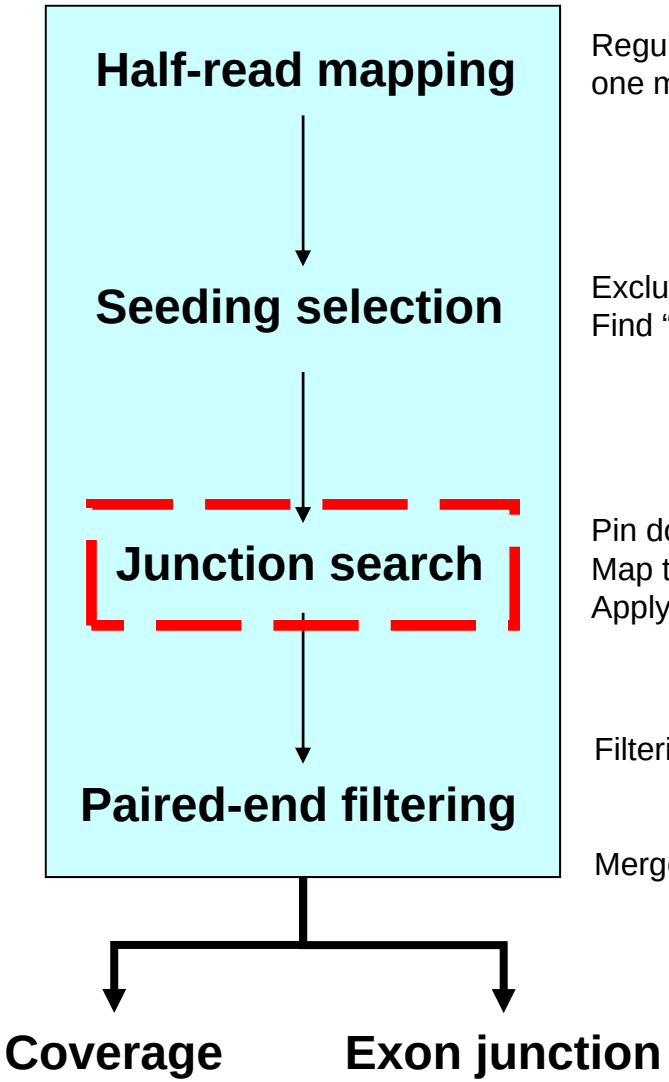
Basic concept

- **Split map**



- **if read length >=50 bp**
- **at least one of the halves will have non-split map**

Algorithm:



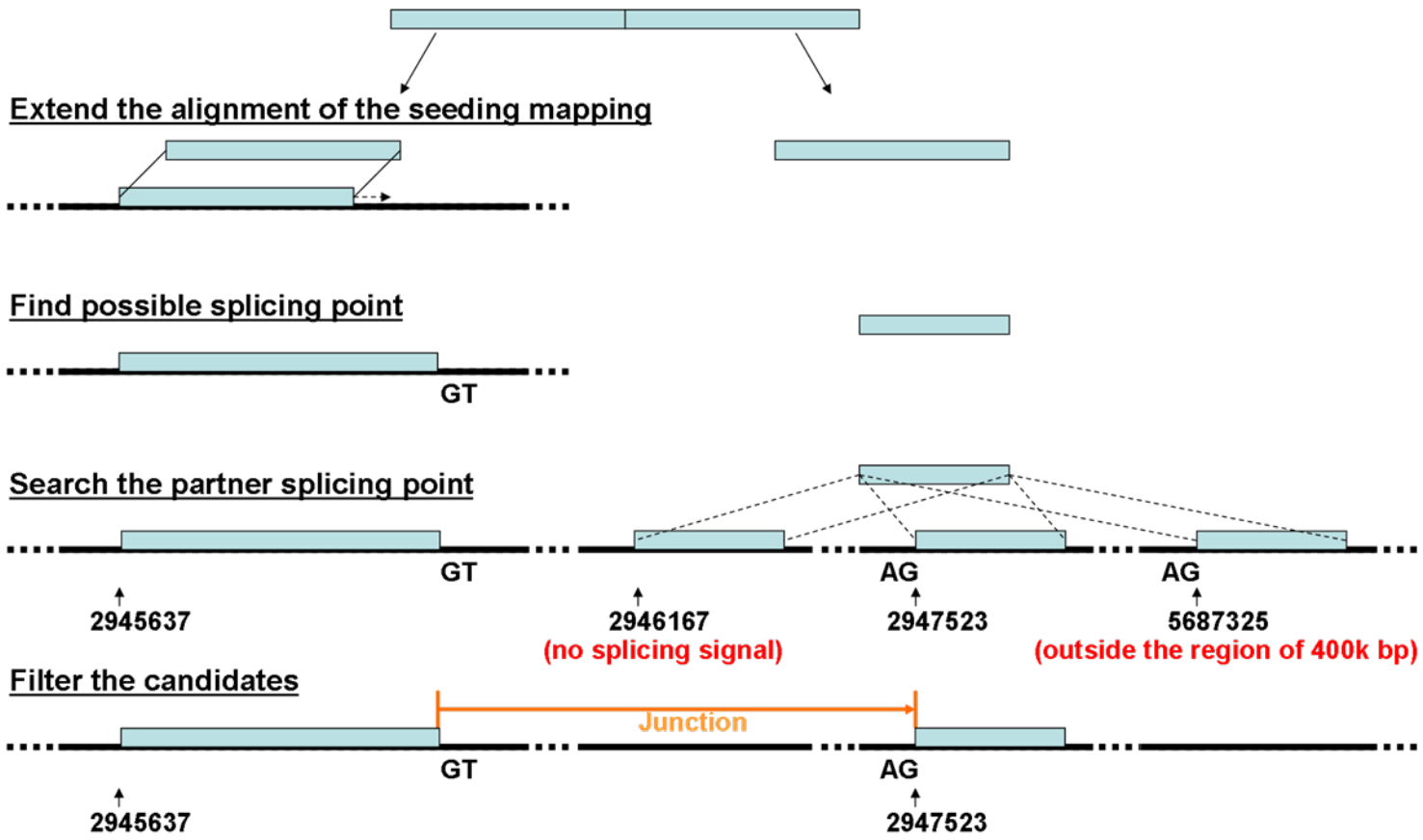
Regular mapping of both halves of read on genome
one mismatch allowed

Exclude local duplicated mapped hits
Find “exonic hits”

Pin down and extend a seeding alignment
Map the residual sequence by hash table look up
Apply splicing signal and distance restriction

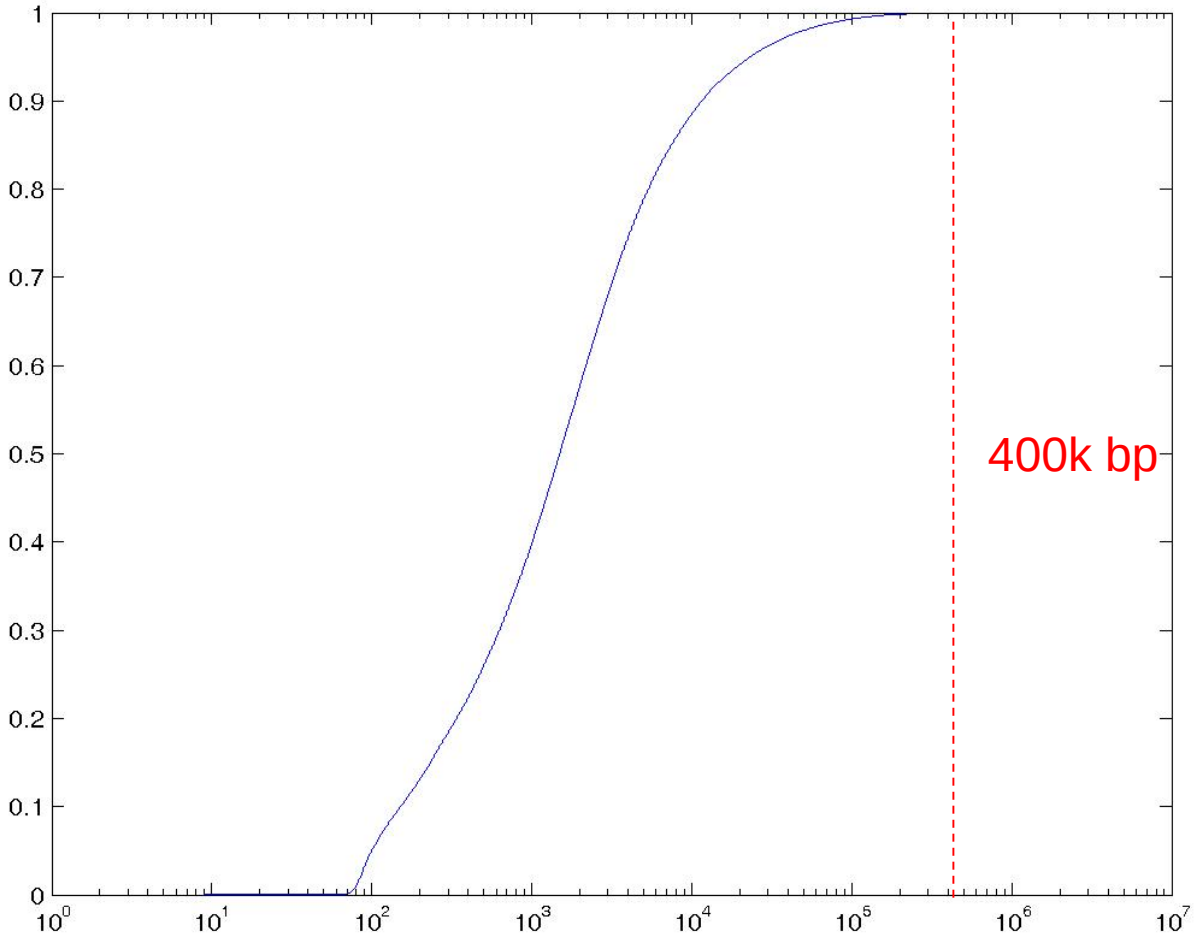
Filtering: 1) good hits
2) direction and positional order
3) distance
Merge results and calculate coverage

Junction search:



- Residual length >= 10 bp
- Canonical splicing signal GT-AG (appears in 98% splices)
- Distance < 400k bp (existence of intron between two splicing exons)

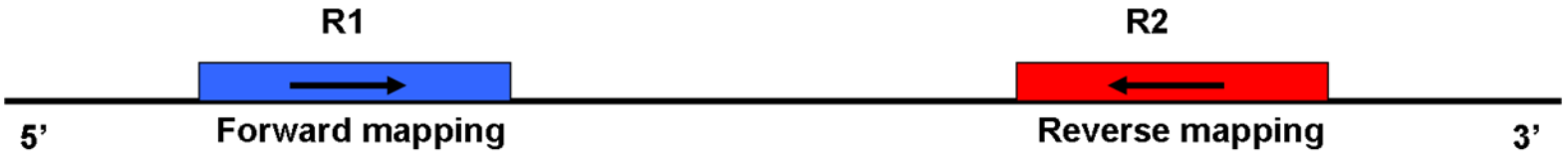
The cumulative distribution function of the intron sizes



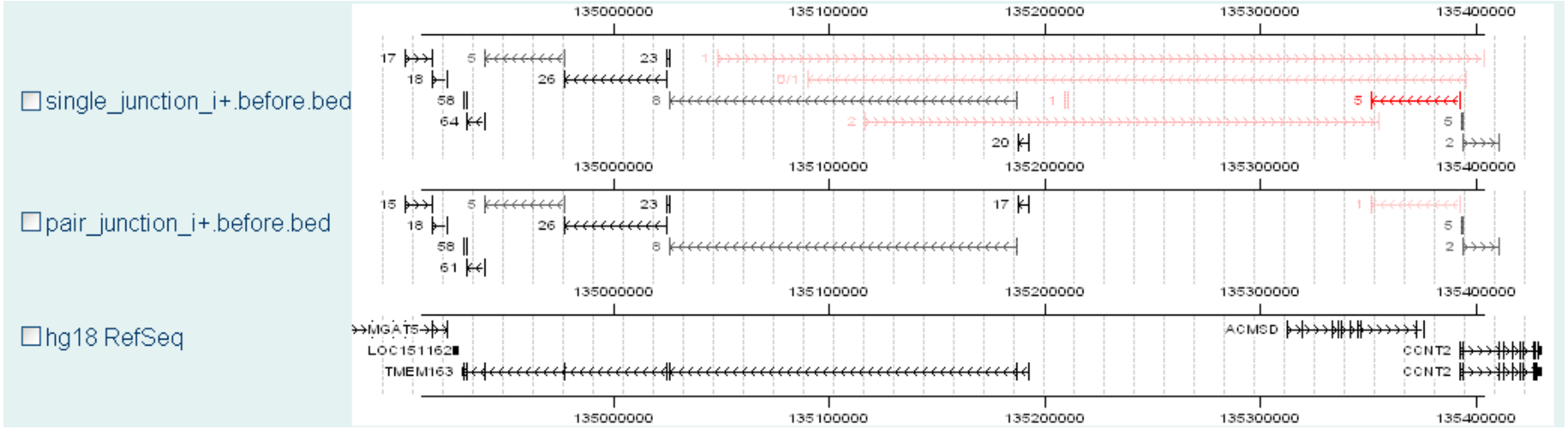
* Based on hg19 human Refseq annotation.

Paired-end filtering: (Illumina data)

- Both are “good hits” (exonic, extension or junction)
- Opposite sequencing direction (“bridge sequencing”)
- Distance < 400k bp

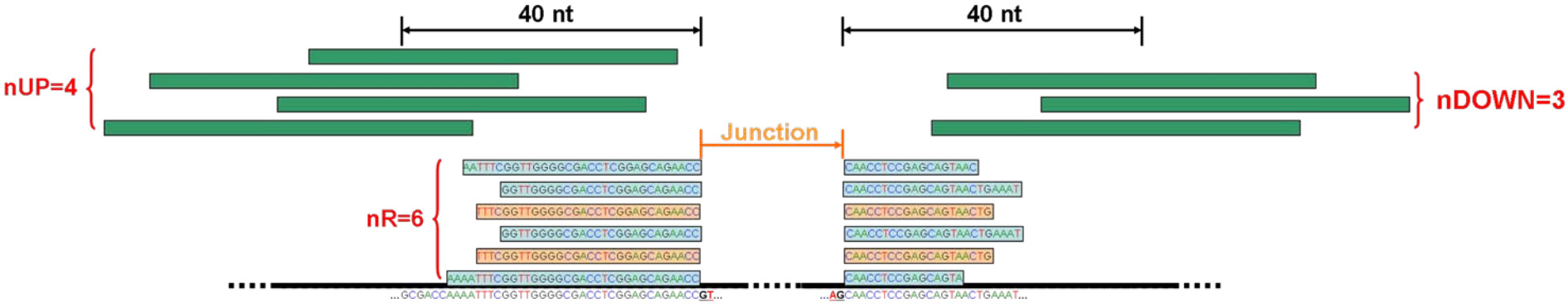


- Example



Parameters for junction quality

- **nR: number of reads supporting this junction**
- **nNR: number of non-redundant supporting reads**
- **nUM: number of uniquely mapped supporting reads**
- **nUP: number of mapped reads in upstream adjacent regions (within K bp)**
- **nDOWN: number of mapped reads in downstream adjacent regions (within K bp)**



nR of this junction is 6
The deep green reads are uniquely mapped supporting reads (nUM=4). The wheat reads are multiply mapped supporting reads. Two supporting reads are redundant, so nNR=4.
There are 4 and 3 uniquely mapped reads (grey green) in upstream and downstream adjacent regions of 40 bp respectively, so nUP=4 and nDOWN=3.

nR distribution

- Novel junctions tends to have low expression

nR	All junctions	Novel junctions*
1	47,995 (27.73%)	24,887 (73.27%)
2~5	49,118 (28.38%)	7,746 (22.81%)
6~20	44,503 (25.72%)	1,170 (3.44%)
21~50	19,055 (11.01%)	131 (0.39%)
51~200	10,464 (6.05%)	30 (0.09%)
201~1000	1,791 (1.03%)	2 (0.01%)
1000+	133 (0.08%)	0 (0%)

***Novel ⇔ not in RefSeq, KnownGene or Ensembl**

Improve specificity by filters

	SpliceMap				
Optional filters*	**---**	**nUM**	**nUP/nDOWN**	**nNR**	**nUM + nUP/nDOWN**
Total junctions	**173,059**	**171,407**	**151,169**	**122,925**	**150,287**
Novel junctions	**33,966**	**32,999**	**26,574**	**9,160**	**25,939**
Junctions with EST validation	**145,232**	**144,380**	**130,059**	**114,768**	**129,690**
Novel junctions with EST	**11,964**	**11,723**	**9,956**	**4,454**	**9,809**
EST validation rate	**83.92%**	**84.23%**	**86.04%**	**93.36%**	**86.29%**
EST validation rate (novel)	**35.22%**	**35.53%**	**37.47%**	**48.62%**	**37.82%**

- **"---" presents no application of any parametric filters;**
- **"nUM" filter requires nUM>0; "nUP/nDOWN" filter requires nUP+nDOWN>0; and "nNR" filter requires nNR>1. For all "nUP/nDOWN" filters, we set K=40.**

Specificity Comparison

23,412,226 paired 50-bp reads from human brain

	SpliceMap	TopHat
Total junctions	**150,287**	**147,712**
Novel junctions	**25,937**	**31,432**
Junctions with EST validation	**129,690 (86.29%)**	**119,835 (81.13%)**
Novel junctions with EST	**9,809 (37.82%)**	**7,967 (25.34%)**

Sensitivity Comparison

Junction detection sensitivity, stratified by gene expression

	SpliceMap	TopHat
0<RPKM<=1 (2993)	5.17%	6.20%
1<RPKM<=2 (1199)	33.73%	29.31%
2<RPKM<=5 (2049)	61.52%	53.80%
5<RPKM<=20 (3245)	88.89%	81.81%
20<RPKM<=50 (1340)	96.12%	91.38%
50<RPKM<=100 (522)	97.84%	94.15%
RPKM>100 (408)	96.39%	90.08%

Sensitivity:

how complete is the junction discovery

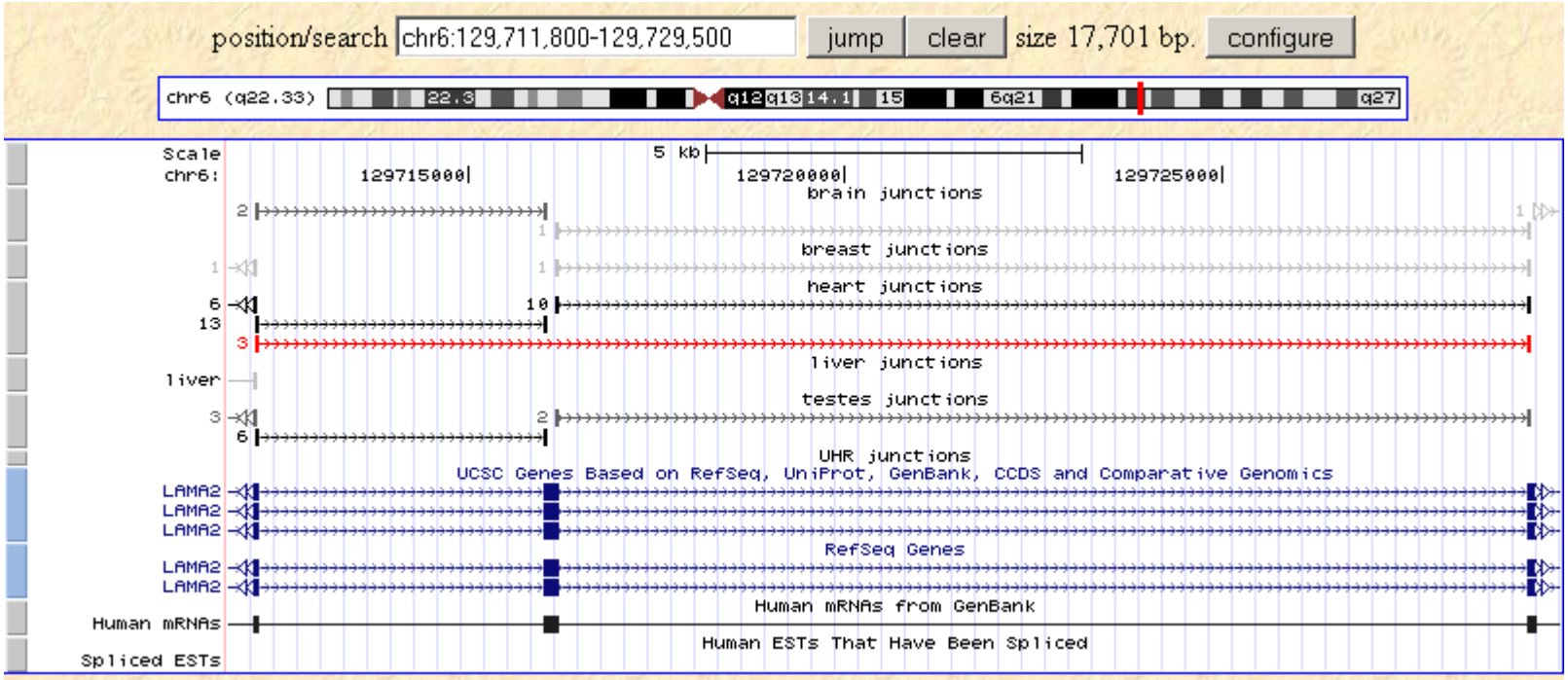
	SpliceMap	TopHat
Number of genes detected	**8774**	**8886**
1<=p<50	**1433**	**2072**
51<=p<80	**1599**	**1983**
81<=p<100	**1496**	**1388**
p=100	**4,246**	**3443**

A gene is detected if at least one junction of the gene is detected.

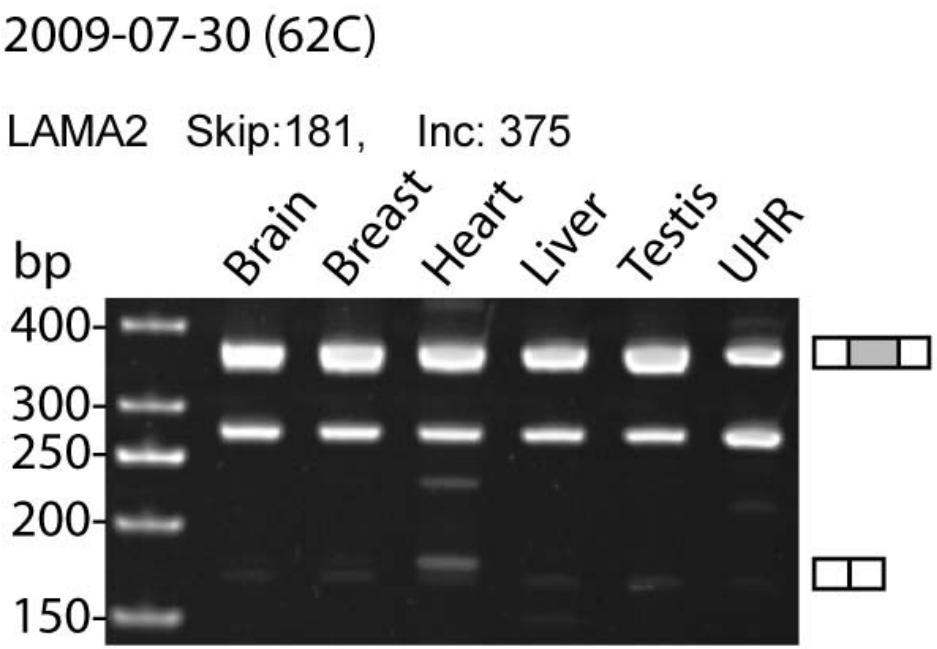
p is the percentage of junctions (in the gene) detected

PCR validation

Gene name		Cassette exon	Length	Result
nNR=1	PTH1R	chr3:46912228-46912363	135	+
	ARL13A	chrX:100128440-100128546	106	+
	PIAS4	chr19:3979928-3980034	106	-
	SPAG17	chr1:118443751-118443946	195	+
nNR=2	NDUFA13	chr19:19499089-19499161	72	-
	RTKN	chr2:74508058-74508227	169	+
	BAT2L	chr9:133311608-133311850	242	+
	PSAP	chr10:73248793-73248874	81	-
nNR:3~5	OSBP2	chr22:29615476-29615623	147	+
	ARHGEF12	chr11:119805630-119805750	120	+
	GTPBP1	chr22:37453163-37453299	136	+
nNR:6~10	MYH7	chr14:22973242-22973298	56	+
	TTN	chr2:179197436-179197703	267	+
	ITGB1BP3	chr19:3892068-3892175	107	+
nNR>10	GIPR	chr19:50872420-50872481	61	+
	LAMA2	chr6:129716001-129716195	194	+
	ATG4D	chr19:10516458-10516542	84	+
	DAB2IP	chr9:123576359-123576512	153	+
FHOD3		chr18:32593078-32593102	24	+
C6orf145		chr6:3682849-3682963	114	+

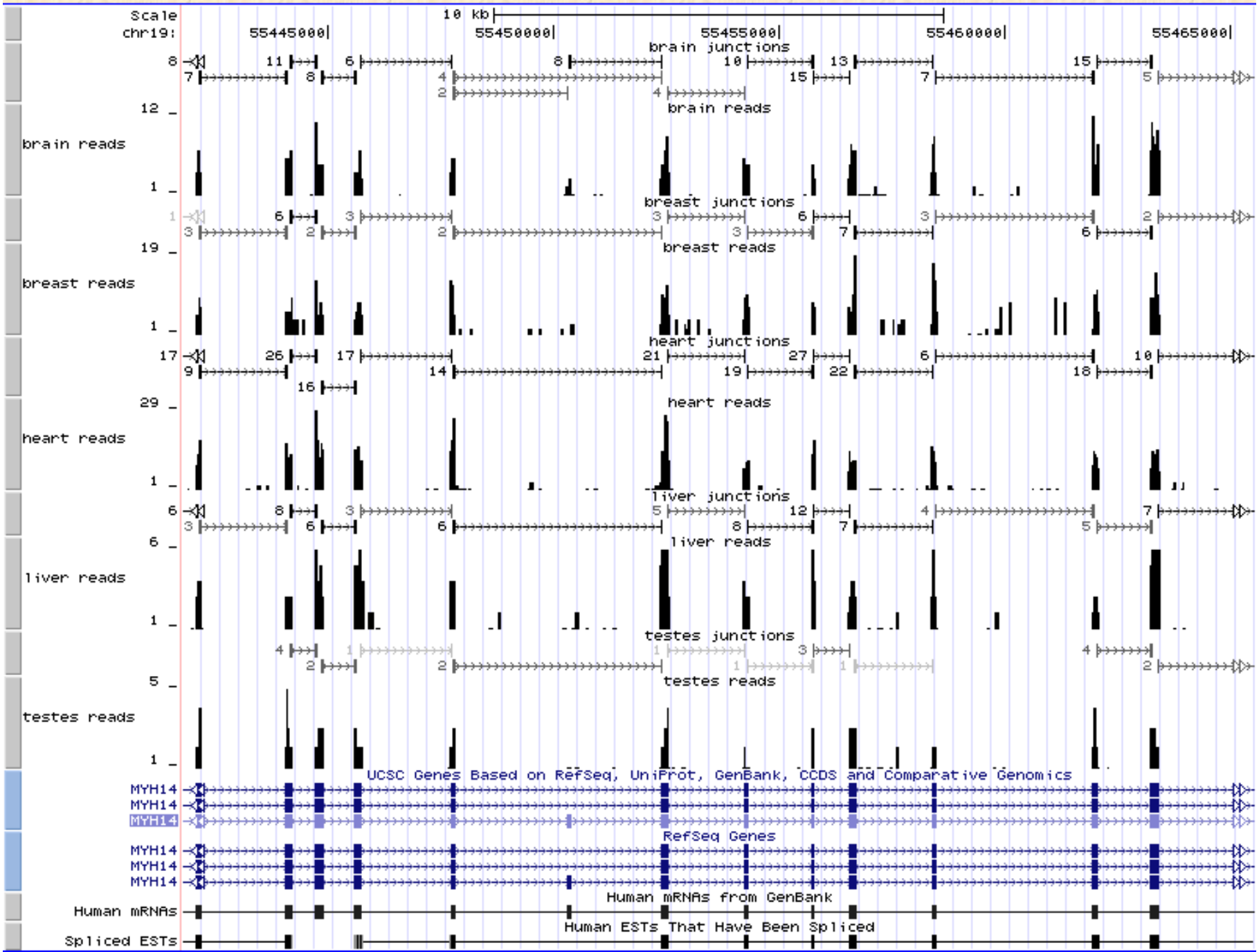


Gene name	Upstream exon	Cassette exon	Downstream exon	Length
LAMA2	chr6:129712135-129712222	chr6:129716001-129716195	chr6:129729056-129729199	194

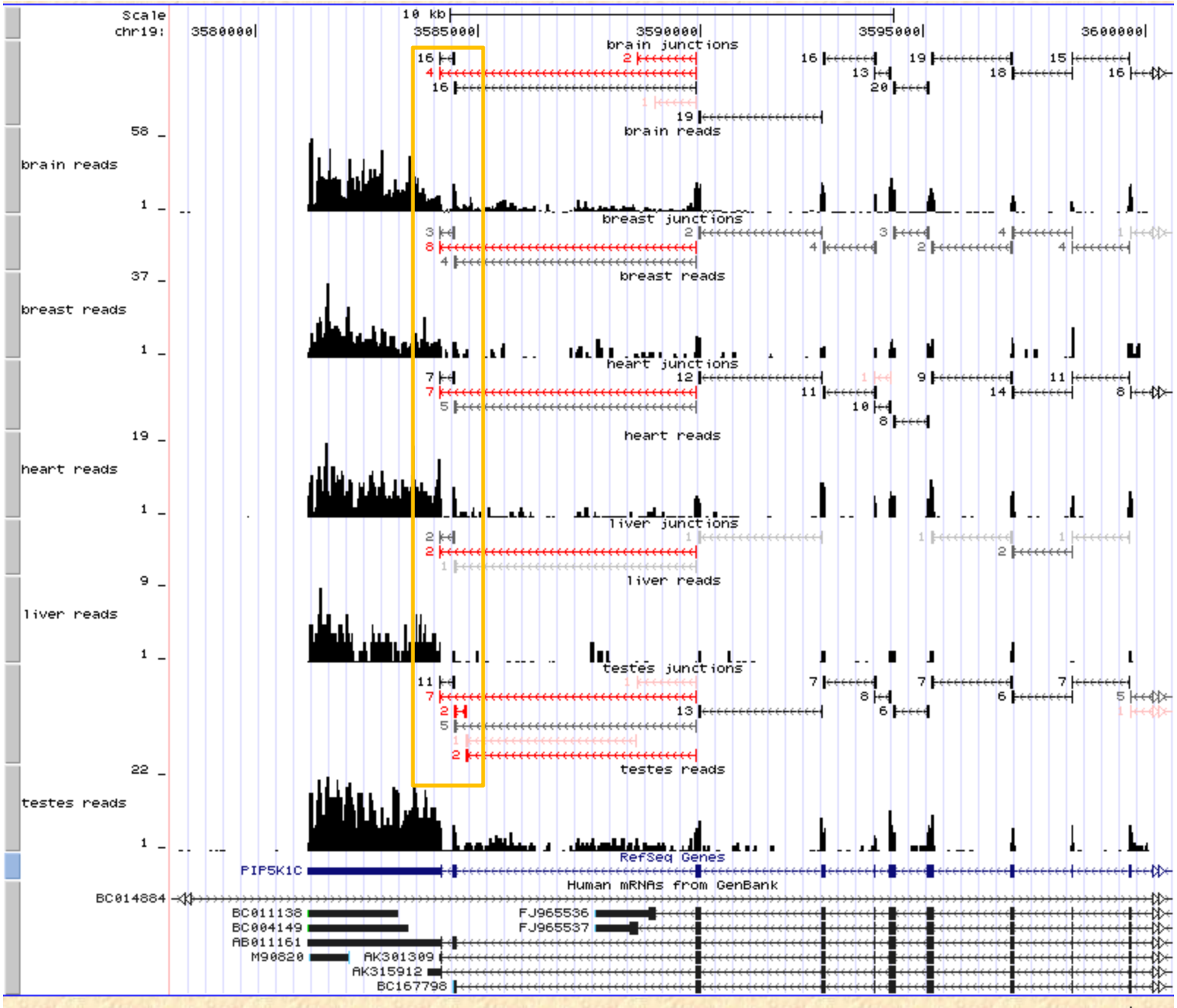


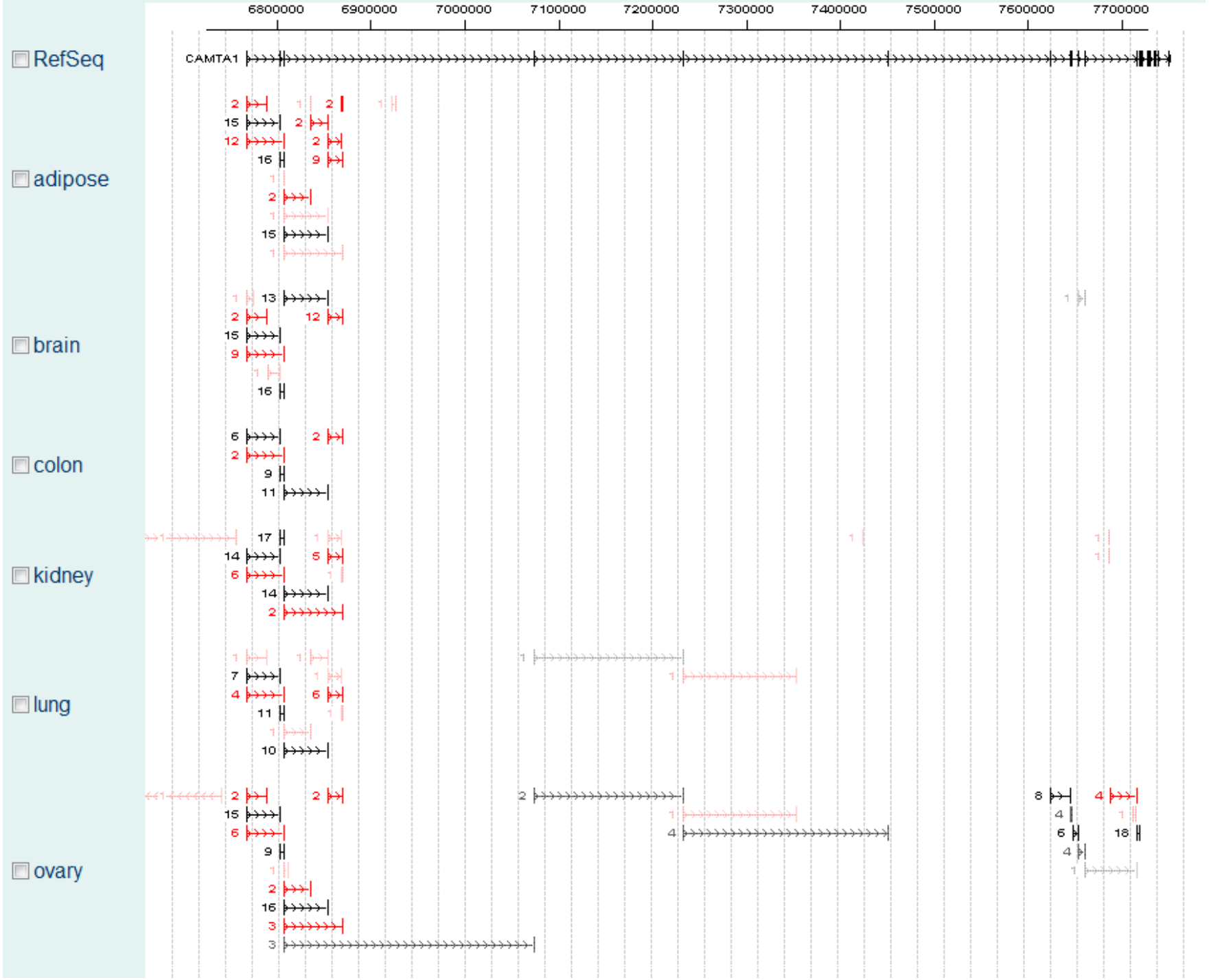
Examples

Chr19:55441784-55465542



Chr19: 3578107 - 3600612




6800000
6900000
7000000
7100000
7200000
7300000
7400000
7500000
7600000
7700000
RefSeq
CAMTA1
adipose
brain
colon
kidney
lung
ovary

Outline

- Scientific background
- Mapping of reads
- Read rates modeling
- Quantification of expression
- Splice junction discovery
- **Isoform discovery**
- Future outlook

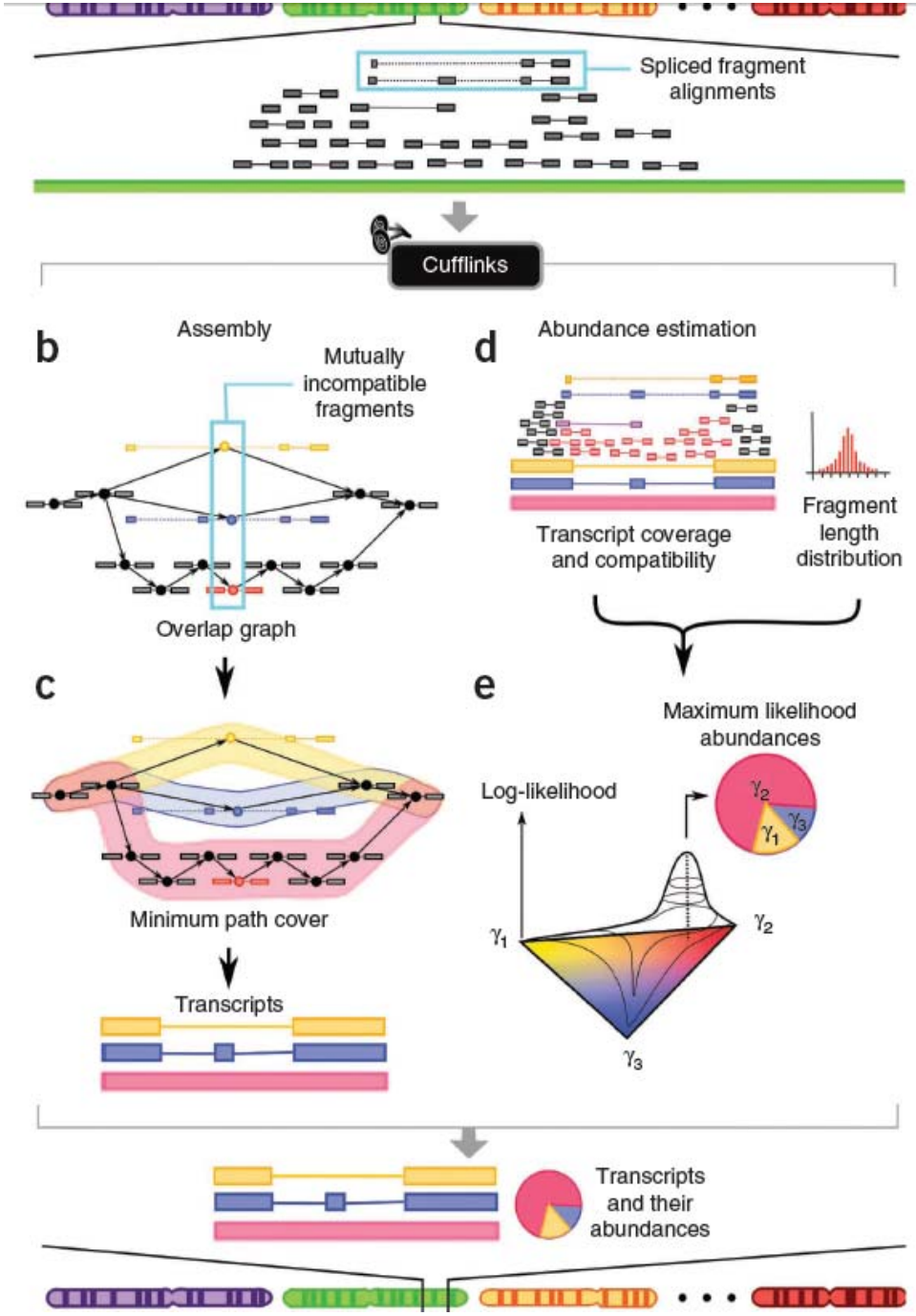
Cufflinks Software

Trapnell, *Nature Biotech*, 2010

Identify all compatible pairs of reads, connect them with an edge

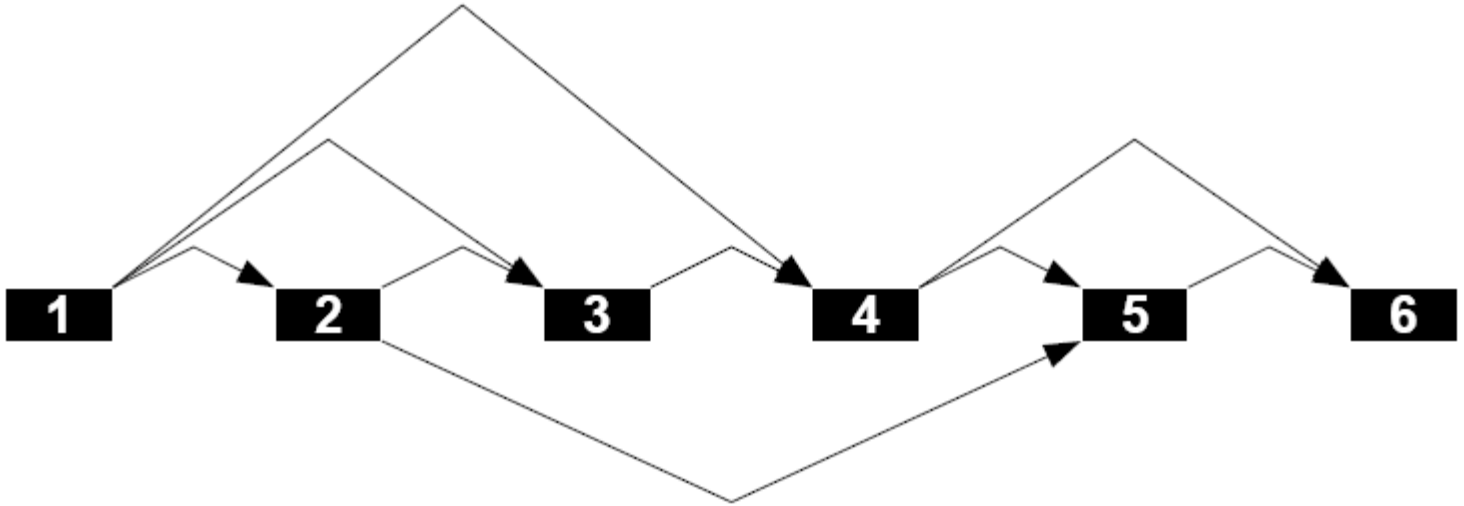
Find a minimal set of paths that cover all the fragments in the overlap graph.

Regard the the paths as isoforms



While Cufflinks gives very useful results, the isoform discovery problem is not yet completely resolved.

With current single and pair-end protocols, isoforms are non-identifiable from the reads (David Hiller 2009). This raises great conceptual and practical difficulties.



Future outlook

- Data rate doubles every few months
- Computing infrastructure needs to scale
- Downstream analyses: comparing samples, allele specific expression, regulation of splicing, etc
- Beware! 3rd generation technology may change the statistical issues

Credits

Kinfai Au

John Mu

SpliceMap

Hui Jiang

SeqMap,
CisG Browser,
Isoform expr.

Identifiability

David Hiller

Jun
Li

Non-uniformity