Digitalization of biology, a history in statistics

History [edit]

The method of Sequential analysis is first attributed to Abraham Wald[1] with Jacob Wolfowitz, W. Allen Wallis, and Milton Friedman[2] while at Columbia University's Statistical Research Group as a tool for more efficient industrial quality control during World War II. Its value to the war effort was immediately recognised, and led to its receiving a "restricted" classification.[3] At the same time, George Barnard led a group working on optional stopping in Great Britain. Another early contribution to the method was made by K.J. Arrow with D. Blackwell and M.A. Girshick.[4]

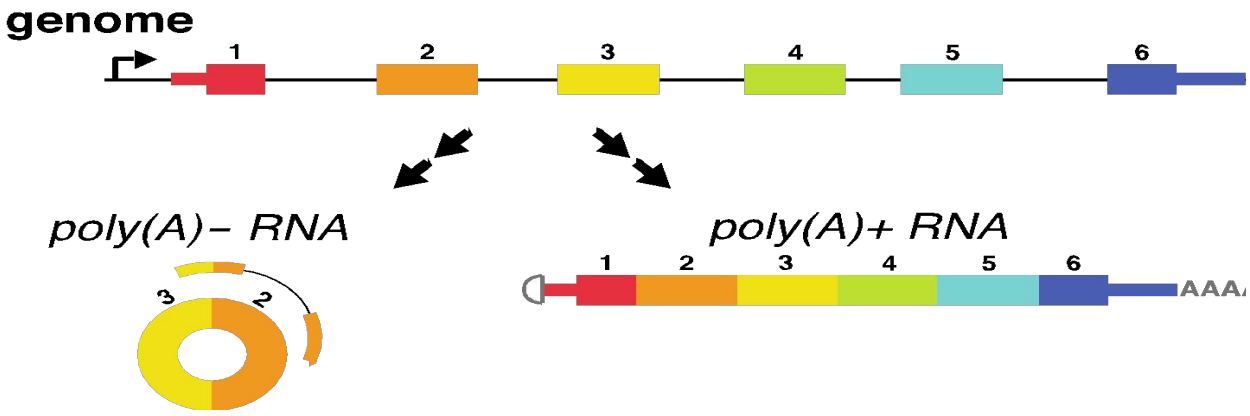
A similar approach was independently developed from first principles at about the same time by Alan Turing, as part of the Banburismus technique used at Bletchley Park, to test hypotheses about whether different messages coded by German Enigma machines should be connected and analysed together. This work remained secret until the early 1980s.[5]

Peter Armitage introduced the use of sequential analysis in medical research, especially in the area of clinical trials. Sequential methods became increasingly popular in medicine following Stuart Pocock's work that provided clear recommendations on how to control Type 1 error rates in sequential designs.[6]

Example from wikipedia, explore

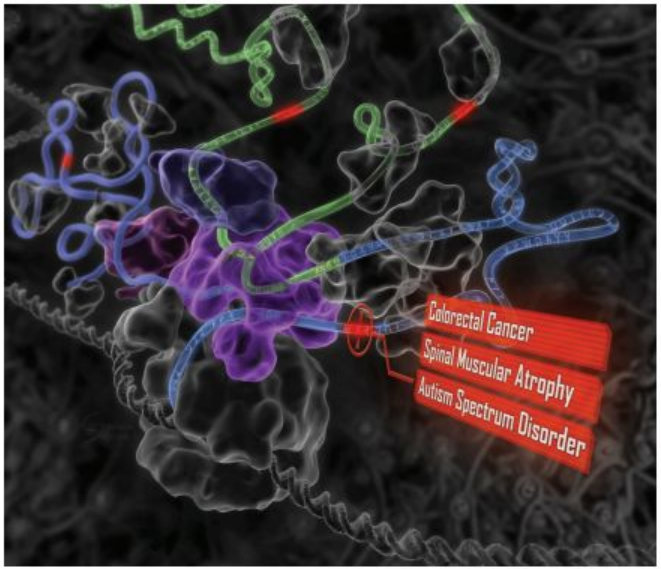
1. What is missing from CRAN and Wikipedia?
2. **Data is new, theoretical framework for analyzing them best is usually old**
3. New applied problems in biology, new biology and new theoretical problems

What is RNA splicing?



More on the board… definition of exon, junction, isoform

Splicing is a cellular code yet to be broken



Conclusions from deep learning on DNA variants, lacks answers to:

- What is the “cause”
- What is the consequence?

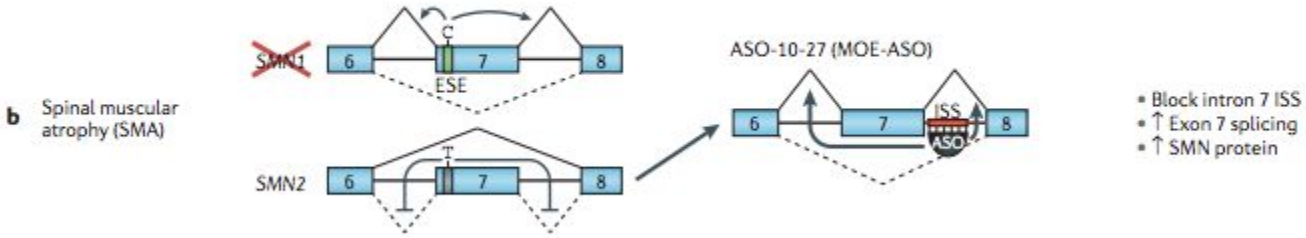
→ how can the disease be treated?

http://science.sciencemag.org/content/sci/347/6218/1254806.full.pdf

Splicing is essential in development and mis-regulation implicated in many disease

The genomic age, more than coding SNPs

1. Cancer genomes: recurrent non-coding variants
2. Neurological diseases: SMA



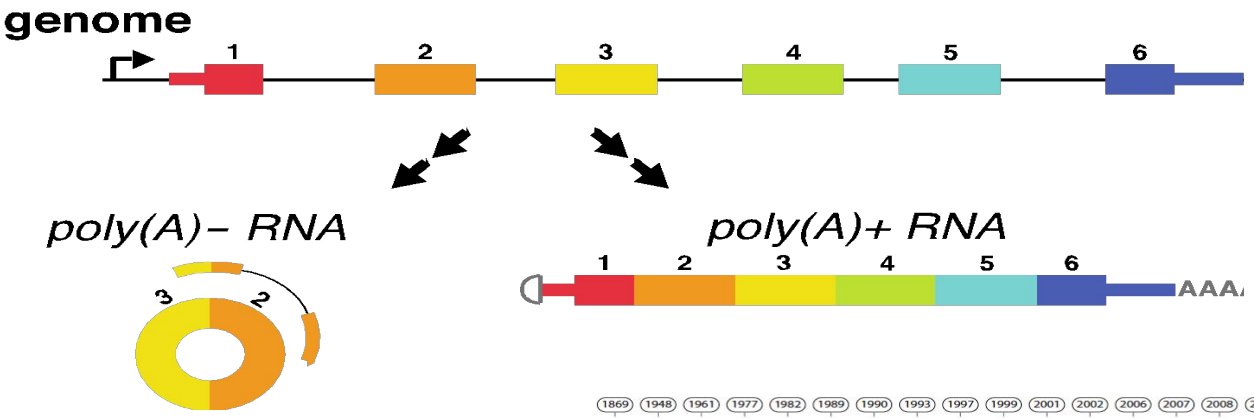
http://www.learnaboutsma.org/antisense/

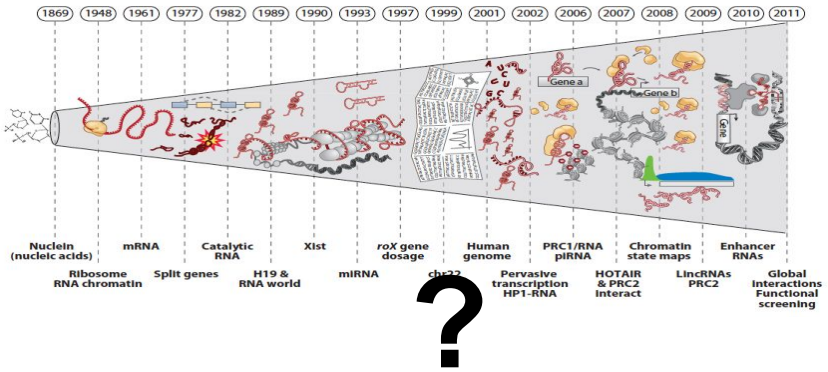
http://www.nature.com/nrg/journal/v17/n1/pdf/nrg.2015.3.pdf

Table 1 | **Disease-associated splicing alterations**

Disease	Gene (mutation)	M
Cis		
Limb girdle muscular dystrophy type 1B (LGMD1B)	*LMNA*[24] (c.1608+5G>C)	5′
Familial partial lipodystrophy type 2 (FPLD2)	*LMNA*[25] (c.1488+5G>C)	5′
Hutchinson–Gilford progeria syndrome (HGPS)	*LMNA*[26] (c.1824C>T)	Al
Dilated cardiomyopathy (DCM)	*LMNA*[28] (c.640-10A>G)	Al
Familial dysautonomia (FD)	*IKBKAP*[128] (c.2204+6T>C)	D
Duchenne muscular dystrophy (DMD)	*DMD*[129] Exon 45–55 deletions are common	Ex
Becker muscular dystrophy (BMD)	*DMD*[130] (c.4250T>A)	Ex
Early-onset Parkinson disease (PD)	*PINK1* (REF. 131) (c.1488+1G>A)	U
Frontotemporal dementia with parkinsonism chromosome 17 (FTDP-17)	*MAPT*[132] (c.892A>G)	Ex
X-linked parkinsonism with spasticity (XPDS)	*ATP6AP2* (REF. 133) (c.345C>T)	N
Spliceosome		
Retinitis pigmentosa (adRP)	*PRPF6* (REF. 134) (c.2185C>T)	Al lo
	SNRNP200 (REF. 135) (c.3260C>T), (c.3269G>T)	• •
Myelodysplastic syndromes (MDS)	*U2AF1* (REF. 46) (c.101G>A)	Al
Microcephalic osteodysplastic primordial dwarfism type 1 (MOPD I)	*RNU4ATAC*[54–56] (g.30G>A), (g.50G>A), (g.50G>C), (g.51G>A), (g.53C>G), (g.55G>A), (g.111G>A)	5′ & di
Trans		
Spinal muscular atrophy (SMA)	*SMN1* (REFS 136,137) (c.922+6T/G), deletion	Lo pr

More diseases like SMA?
Detecting quantitative RNA expression





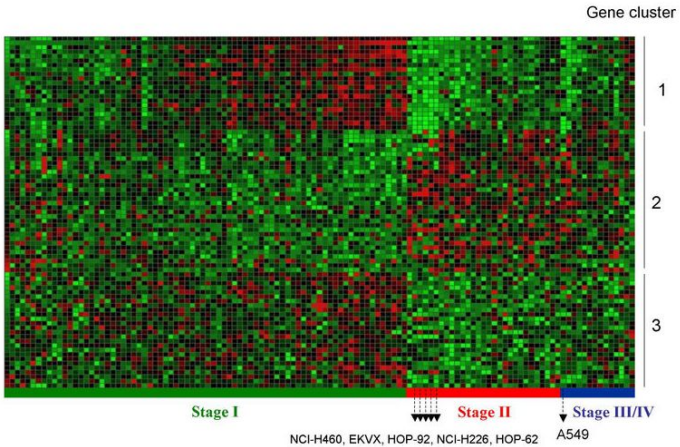
For therapies, quantitative precision and mechanism is needed→ foundational statistics

.3* chromatin mark X + .6 * SNP #1 doesn’t make a SMA therapy

What are the needed statistical algorithms?

1. Quantifying exon expression, junction expression
2. Deconvolving isoform expression
3. Some are trying to discover new RNA

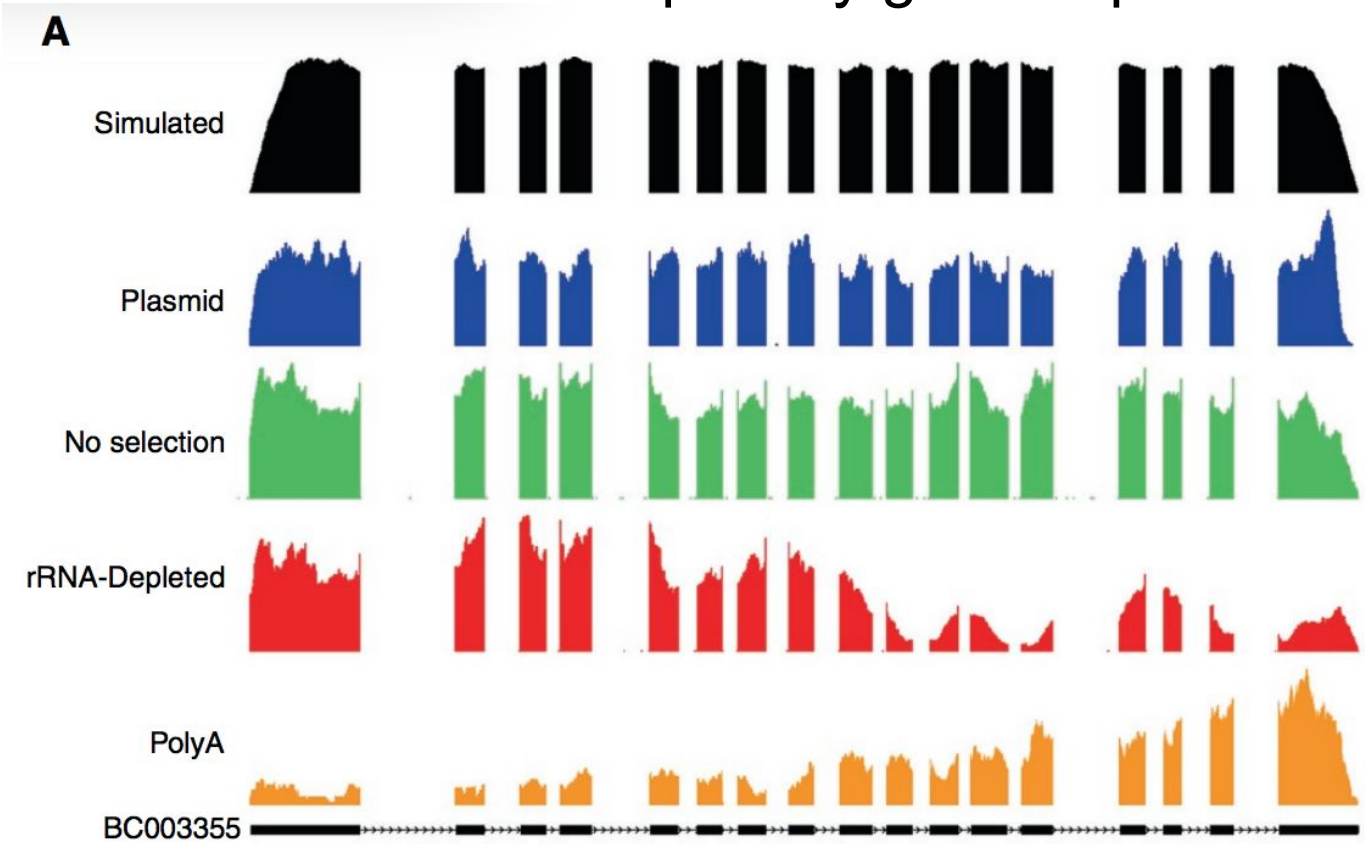
We want to know the copies of RNA per cell



From:
http://media.springernature.com/lw785/springer-stati
mage/art%3A10.1186%2F1471-2164-7-166/MediaC
ects/12864_2006_Article_549_Fig4_HTML.jpg

General problem: alignment as a black box, read densities

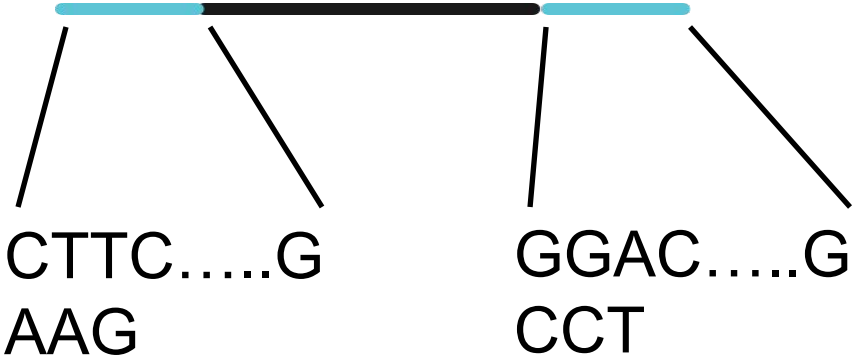
Use read densities to quantify gene expression



Lahens *et al. Genome Biology* 2014, **15**:R86
http://genomebiology.com/2014/15/6/R86

The data: paired-end RNA-seq

Matched sequences are obtained for each library molecule



First, use statistical modeling: why?

- Statistics underlies all of the algorithms used to quantify gene expression from RNA-Seq

- Most simple is the Poisson model

- Named for Poisson, who used it to model rare events:
  - # horse kickings in the Prussian army per year
- Po(λ), the larger λ, the more likely the rare event

  - Defined as $Po(X=k)=e^{-\lambda} \lambda^k/k!$
  - $k>0$

The statistical modeling

- Po(λ), the larger λ, the larger the rate of the rare event
  - Defined as $Po(X=k)=e^{-\lambda} \lambda^k/k!$
  - k>0

  - In RNA-Seq, each transcript (compared to all others) will be rare, so each transcript abundance modeled as λ_i
- In statistics, we take observed data and use it to estimate parameters, in this case, λ_i

- This is formally accomplished by, for example the MLE
- In RNA seq, “RPKM” is conceptually like λ_i

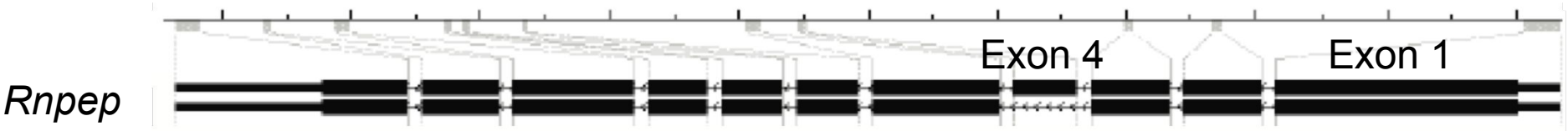
More on the model

- Po(λ), the larger λ, the larger the rate of the rare event
  - Defined as $Po(X=k)=e^{-\lambda} \lambda^k/k!$
  - $k>0$
- For the Poisson distribution, the abundance of each transcript is proportional to λ, so estimation seems easy.

- Caveat: we have to control for sequencing depth.. Why?

- In reality, as we will see, alternative splicing (in 99% of human genes and “all multiexon LINCS”) makes the situation “much more complicated”

Intuition for the statistical problem



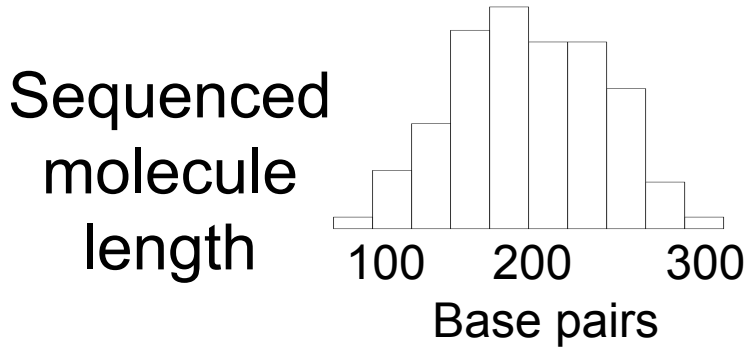
Estimate the expression of each isoform?

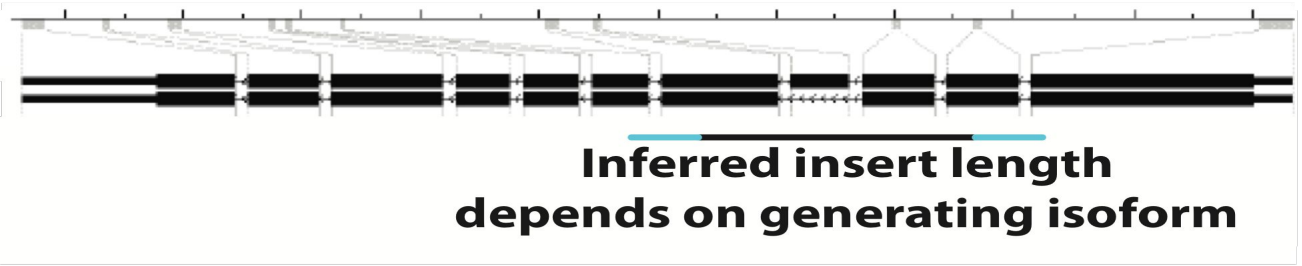
Nontrivial : we only observe fragments of sequences

- Since the size distribution of library molecules is known, inferred insert lengths can be used to increase statistical power and inference

Intuition for the most powerful modeling

- Compute genome-wide insert length distribution





•Mapped to Isoform 1
→ length 150
•Mapped to Isoform 2
→ length 90

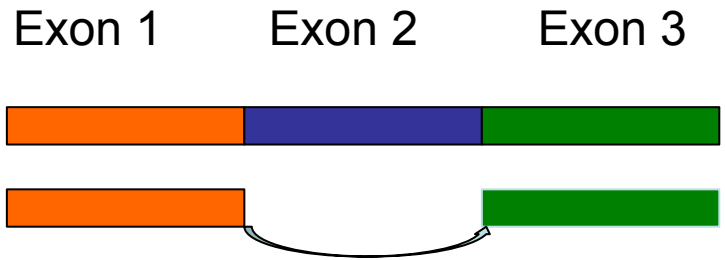
- Statistical improvement over naïve models
- Optimal information reduction
- Quantifies information gain using PE Sequencing

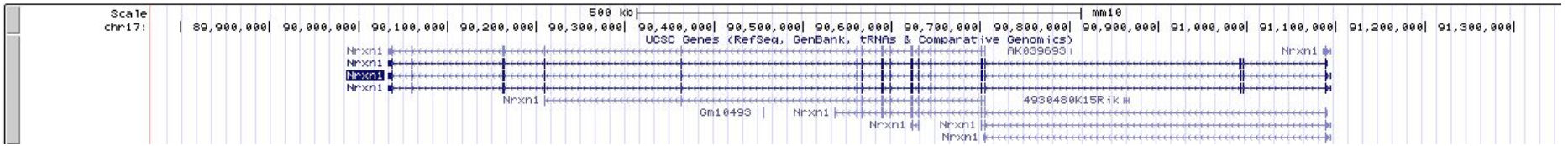
Why do we care: just fun math?

- Not knowing the isoforms means we don’t know the gene level expression
- Off the shelf tools are “mostly right” but many times wrong
- Most labs don’t use their latest published software
- Current tools only provide approximate answers

Intuition for statistically quantifying isoforms

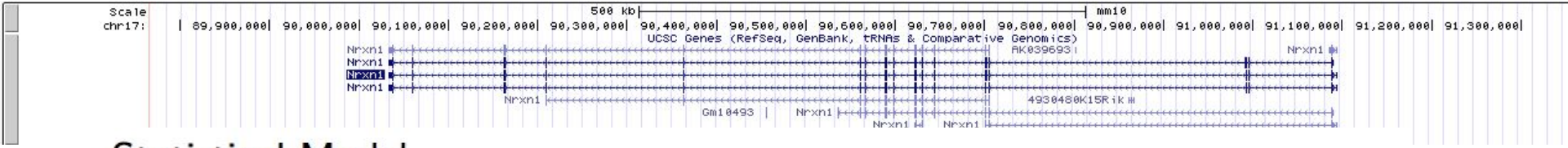
1. Exon-level and junctional reads are observed
2. There is a deconvolution problem
    a. Quantifying exon expression, junction expression
    b. Deconvolving isoform expression





Sufficient statistics, statistical problem, Poisson models

Formalizing the problem and model



Statistical Model

- The relative abundance for the I isoforms are the parameters of interest and denoted $\{\theta_i\}_{i=1}^{I}$.

Solving the problem with statistics

Data: observe $\{n_{\cdot,j}\}_{j=1}^{J}$; n_{ij} are unobservable.

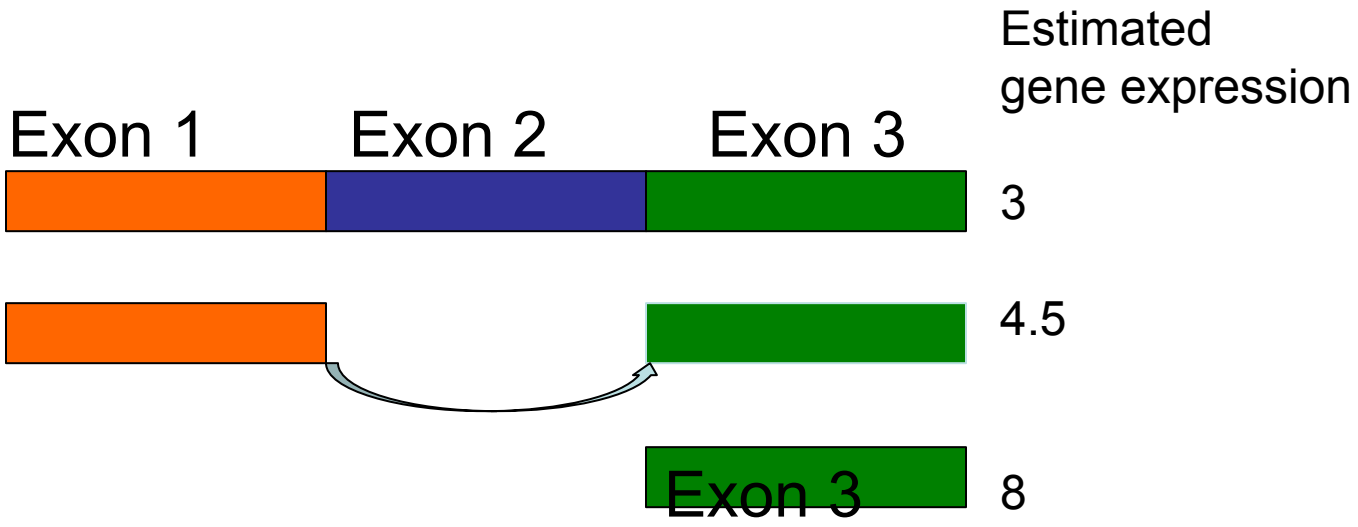
Likelihood function for statistics $\{n_i\}_{i=1}^{I}$: $n_j = n_{\cdot,j}$ follows a Poisson distribution with parameter $\sum_{i=1}^{I} \theta_i a_{i,j} = \theta \cdot a_j$, where

Each isoform expression is independent:

The importance of statistics

Exon	1	2	3
Count	1	0	8

Remember, counts =“expression” in RNA-Seq



Without taking isoforms into account, gene expression estimates (and differential gene expression will be wrong)!

Gene and isoform expression are inextricably linked

Quantify alternative splicing is needed to reliably measure gene expression

Sailfish and other recently developed algorithms compute coverage
At per nucleotide resolution, improving (but not eliminating) some problems

Also, significant implications for differential gene expression

1 rpkm

3 rpkm

1 rpkm

Even more “problems”: count data is noisy

Example, idea: clean it up w/ robust statistics

Bayesian analysis