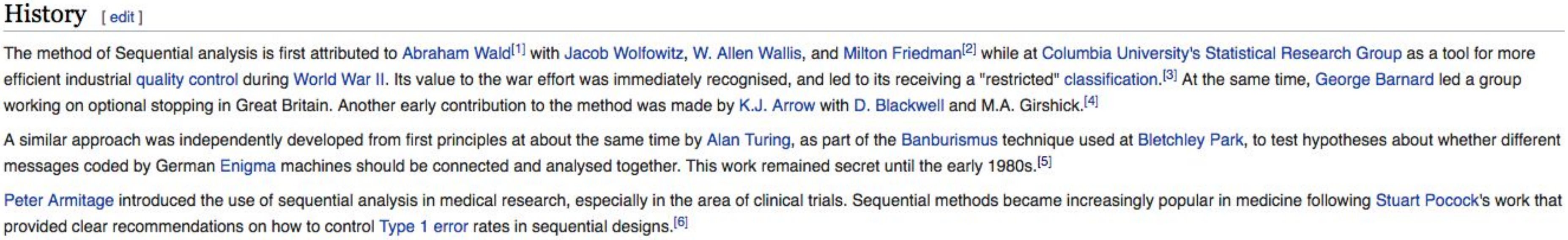
Digitalization of biology, a history in statistics



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

Data is new, theoretical framework for analyzing them best is usually old

What is missing from CRAN and Wikipedia?

Summary

- Biomedical background: DNA, RNA and its role in disease
    - RNA: the new medicine, and the promise for biomedical data science
    - What data is available?
    - What can be discovered

- Statistical concepts and modeling motivated by detecting RNA splicing in disease
    - Parametric statistical models for RNA-seq
    - Non-parametric statistical modeling

Outline of Lecture 1

1. **Biomedical background: DNA, RNA and its role in disease**
    a. **What data is available?**

2. RNA: the new medicine, and the promise for biomedical data science
3. Foundations for modeling RNA-seq
    a. Rank tests
        i. Properties of rank tests
            1. Robustness
            2. Speed
            3. Theoretical tractability

Counter-examples!

Biological motivation

DNA coding variants: the classical phenotypes

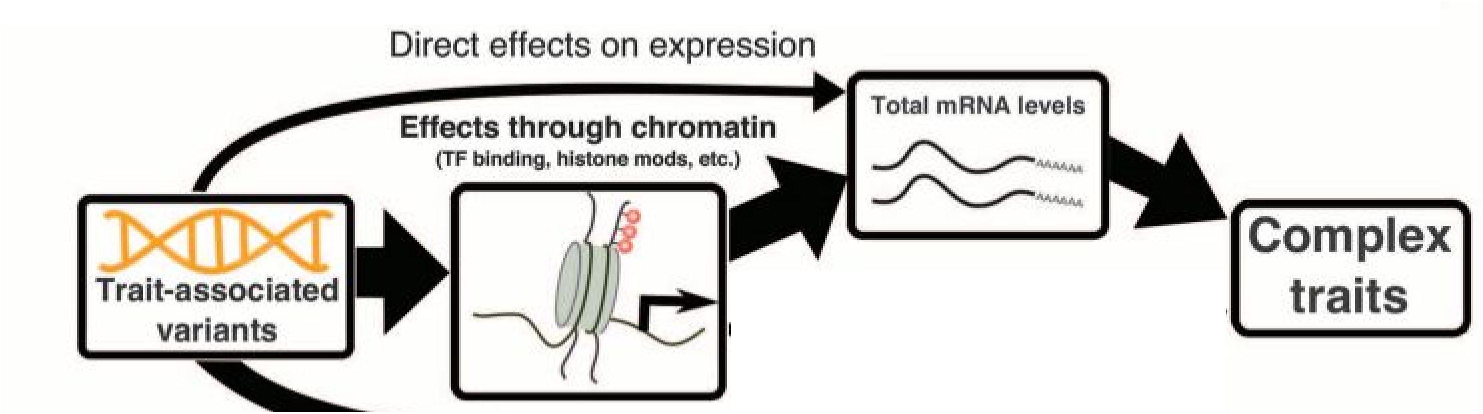

Direct effects on expression
Effects through chromatin
(TF binding, histone mods, etc.)
Total mRNA levels
AAAAAA
AAAAAA
Trait-associated variants
Complex traits

Figure from .. Pritchard, Science, 2016

- The GWAS hope: simple DNA variants will explain disease
- The reality: SNPs leave much to explain: ~50% mendelian disorders cannot be explained by whole exome sequencing (http://biorxiv.org/content/early/2016/07/29/066738)

The expanding role of RNA and regulation

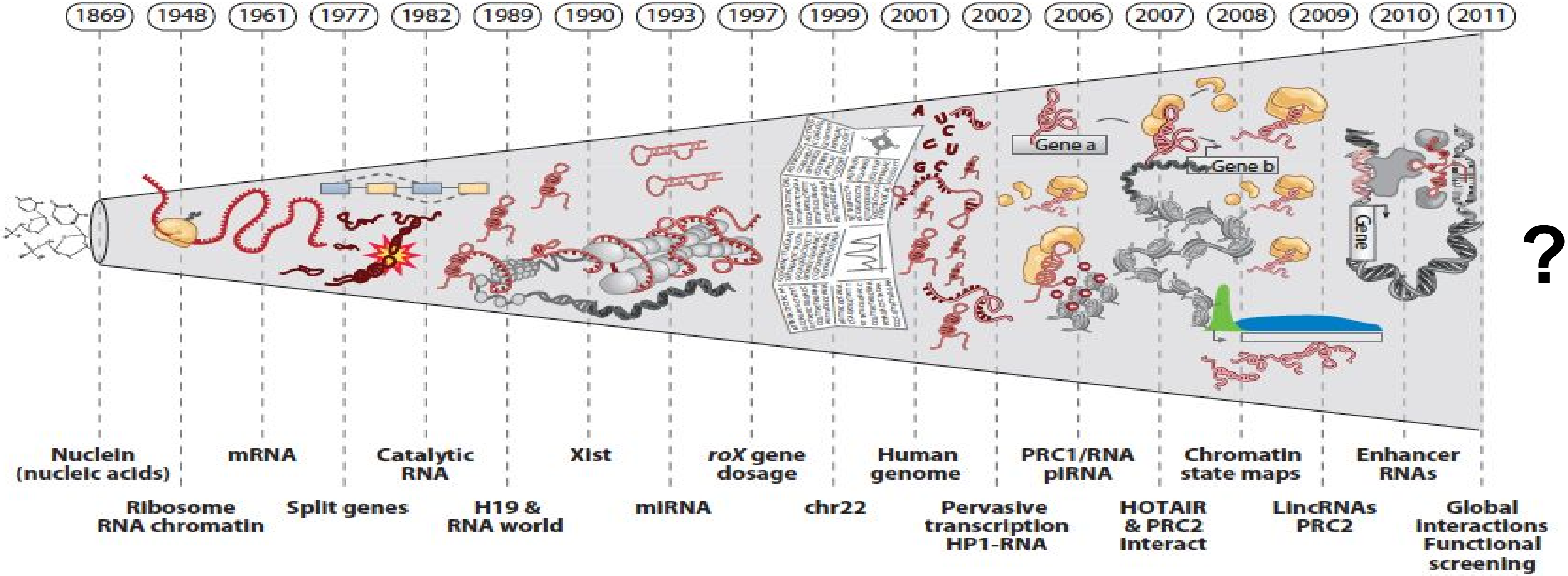
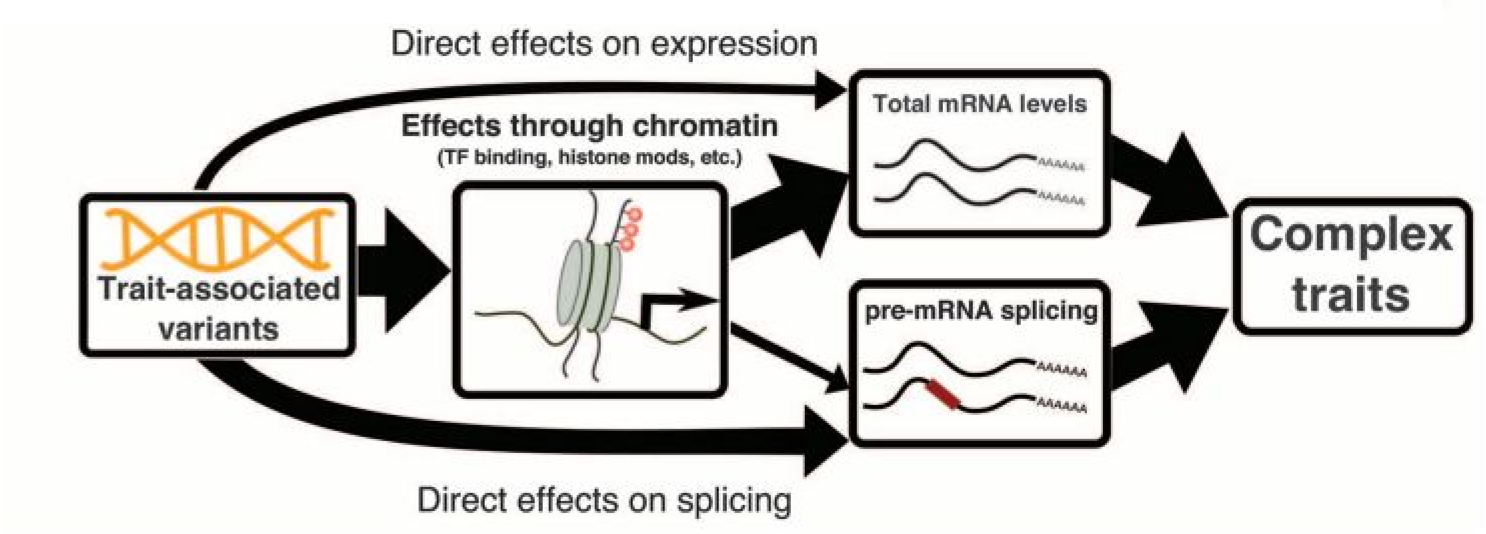


Figure from Rinn et al, 2012

The genomic age: more than just the exome

1. The DNA
    a. Modifications
    b. Epigenetic marks
    c. Hidden variants
        i. the unassembled genome
        ii. the unassembled personal genome
2. The RNA
    a. Non-coding RNA
    b. RNA processing defects, defective RNA-- the most quantitative, direct observable in a diseased tissue
3. The protein

Splicing, a biological and medical mystery

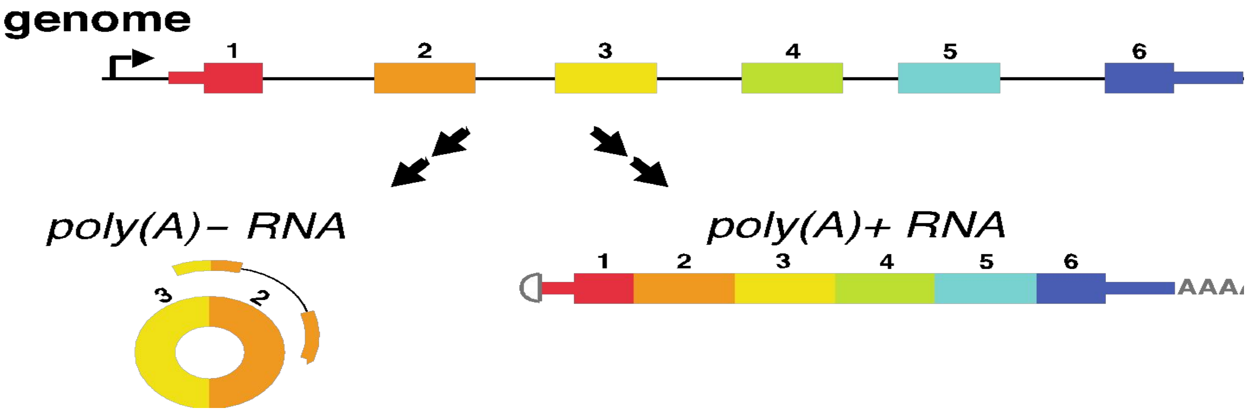

Direct effects on expression
Effects through chromatin
(TF binding, histone mods, etc.)
Total mRNA levels
AAAAAA
AAAAAA
Trait-associated variants
Complex traits
pre-mRNA splicing
AAAAAA
AAAAAA
Direct effects on splicing

Splicing variants explain some Mendelian disorders

Li et al, 2016; http://biorxiv.org/content/early/2016/07/29/066738

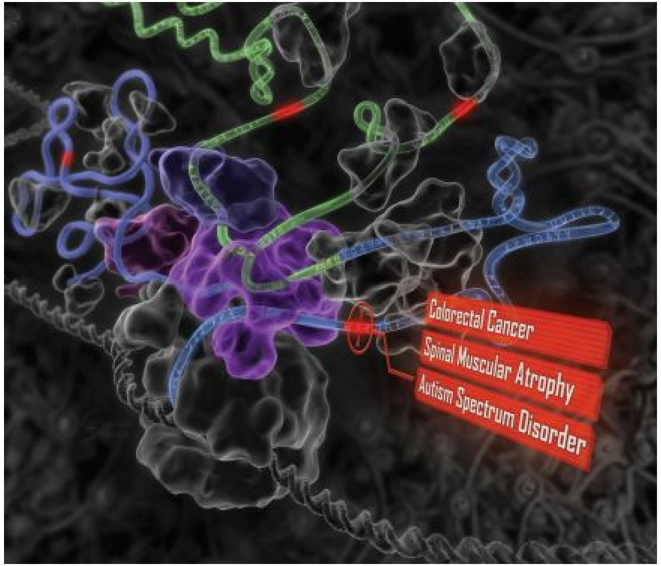
Li .. Pritchard, Science, 2016

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?

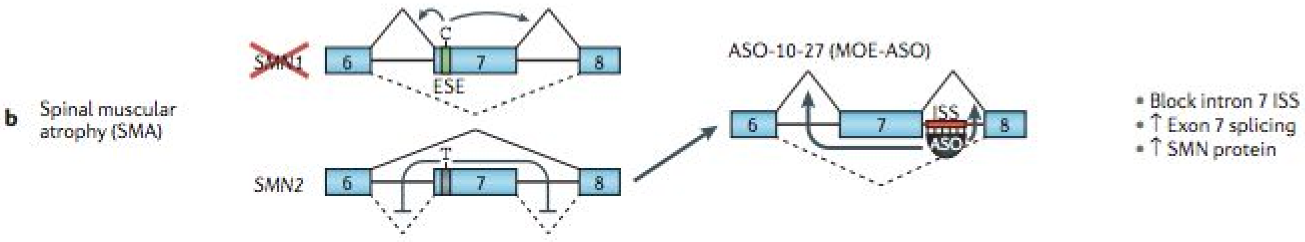
→ how can the disease be treated?

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

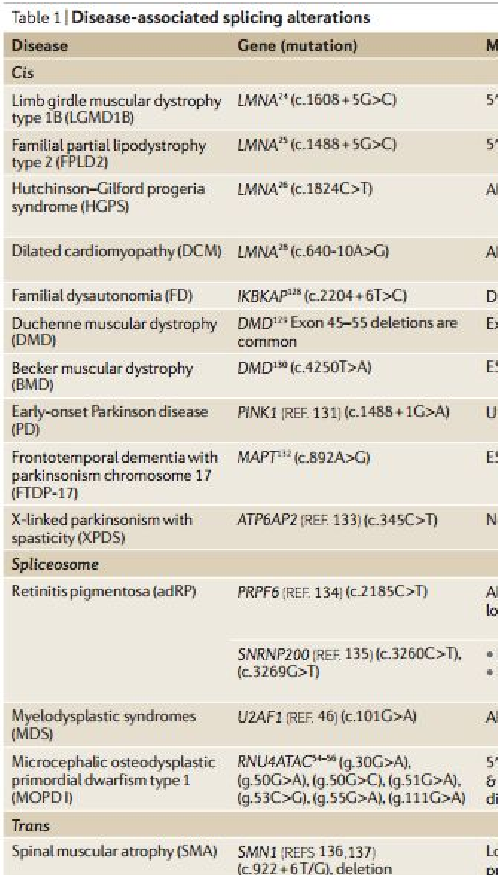
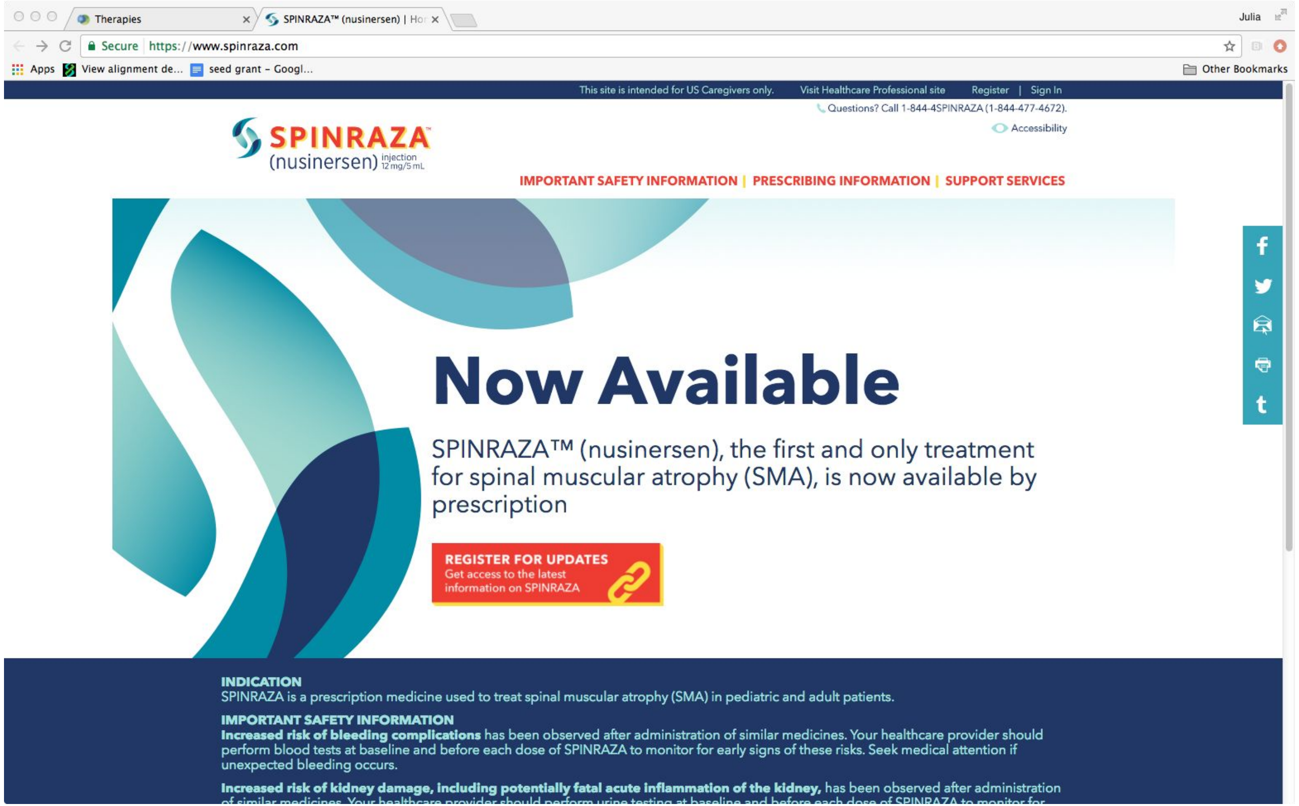


Table 1 | **Disease-associated splicing alterations**

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)	A
Dilated cardiomyopathy (DCM)	*LMNA*[28] (c.640-10A>G)	A
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)	E
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)	E
X-linked parkinsonism with spasticity (XPDS)	*ATP6AP2* [REF. 133] (c.345C>T)	N
Spliceosome		
Retinitis pigmentosa (adRP)	*PRPF6* [REF. 134] (c.2185C>T)	A lo
	SNRNP200 [REF. 135] (c.3260C>T), (c.3269G>T)	• •
Myelodysplastic syndromes (MDS)	*U2AF1* [REF. 46] (c.101G>A)	A
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 p

SMA: the first drug, an RNA


Therapies
SPINRAZA™ (nusinersen) | Ho
Julia
Secure https://www.spinraza.com
Apps
View alignment de...
seed grant - Googl...
Other Bookmarks
SPINRAZA™
(nusinersen) injection 12 mg/5 mL

This site is intended for US Caregivers only. Visit Healthcare Professional site Register | Sign In

Questions? Call 1-844-4SPINRAZA (1-844-477-4672).

Accessibility

IMPORTANT SAFETY INFORMATION | PRESCRIBING INFORMATION | SUPPORT SERVICES

Now Available

SPINRAZA™ (nusinersen), the first and only treatment for spinal muscular atrophy (SMA), is now available by prescription

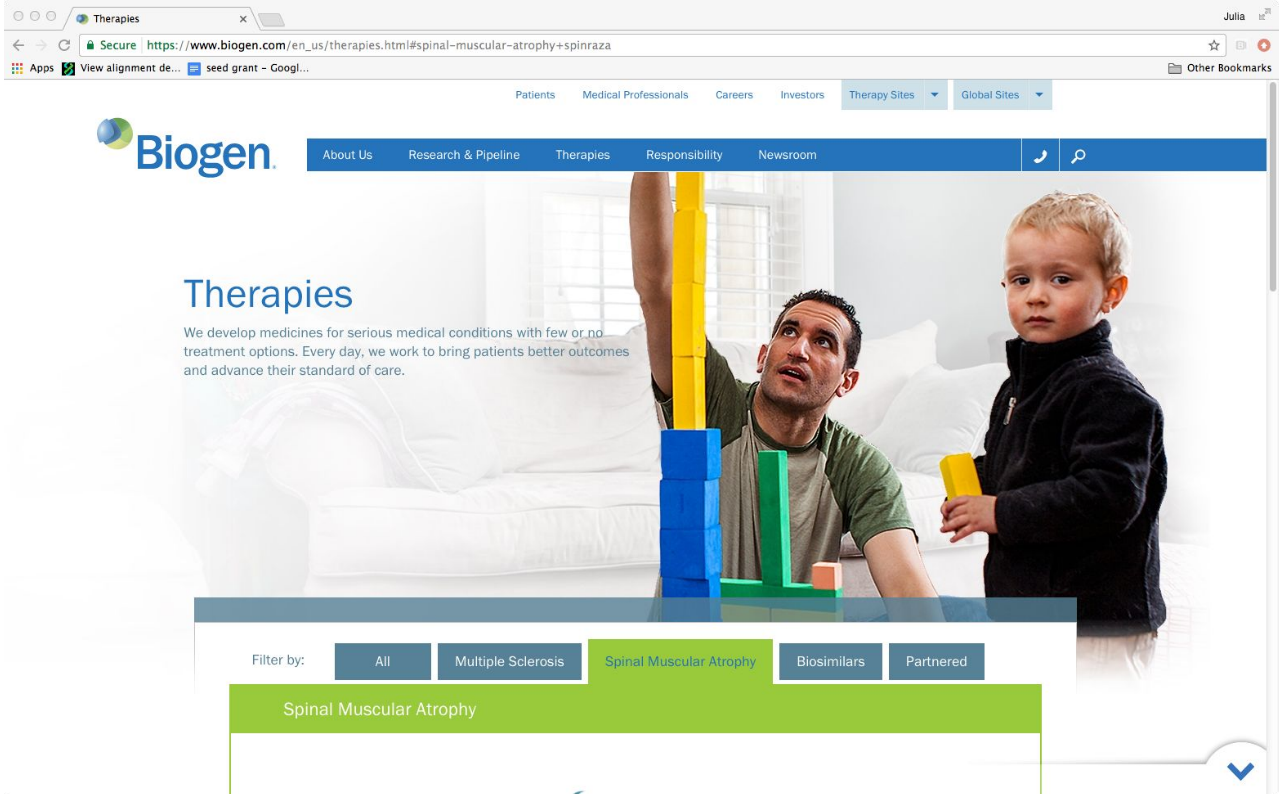
REGISTER FOR UPDATES
Get access to the latest information on SPINRAZA

INDICATION
SPINRAZA is a prescription medicine used to treat spinal muscular atrophy (SMA) in pediatric and adult patients.

IMPORTANT SAFETY INFORMATION
Increased risk of bleeding complications has been observed after administration of similar medicines. Your healthcare provider should perform blood tests at baseline and before each dose of SPINRAZA to monitor for early signs of these risks. Seek medical attention if unexpected bleeding occurs.

Increased risk of kidney damage, including potentially fatal acute inflammation of the kidney, has been observed after administration of similar medicines. Your healthcare provider should perform urine testing at baseline and before each dose of SPINRAZA to monitor for

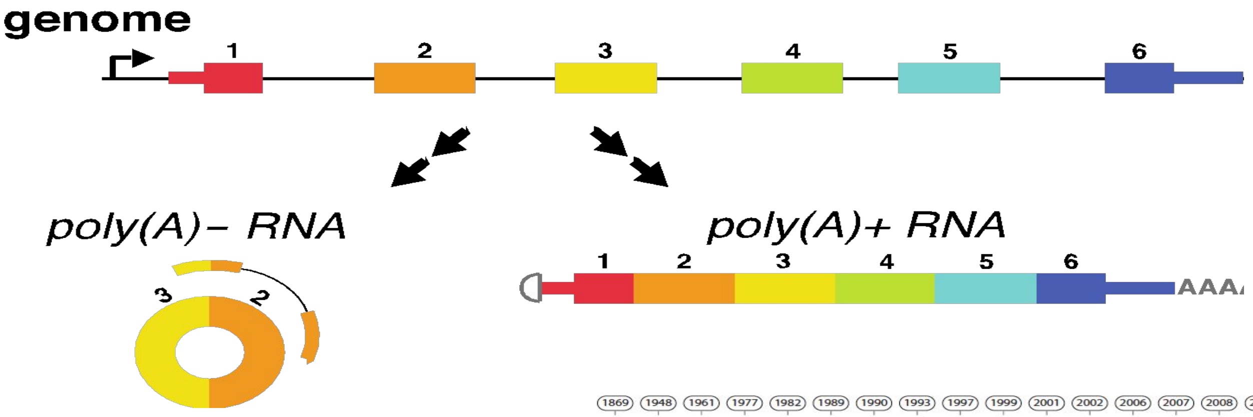
Biogen, a company founded on RNA therapeutics

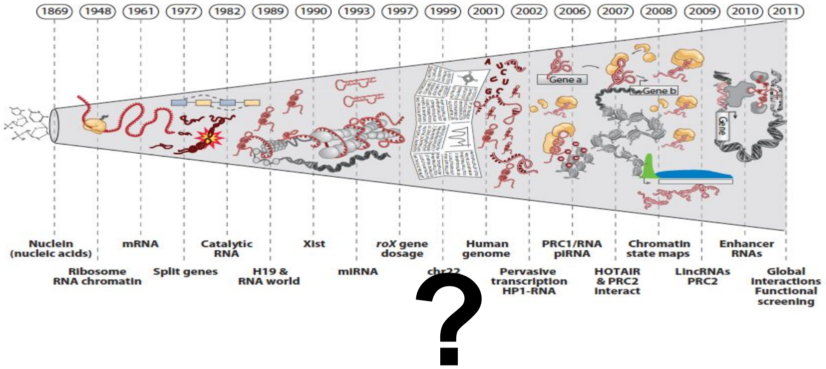


Therapies

We develop medicines for serious medical conditions with few or no treatment options. Every day, we work to bring patients better outcomes and advance their standard of care.

More diseases like SMA?
Detecting quantitative RNA expression





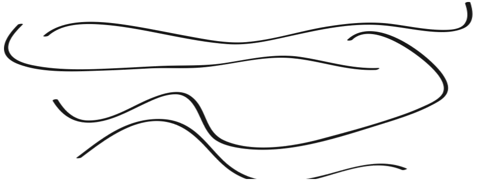
For therapies, quantitative precision and mechanism is needed→ foundational statistics

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

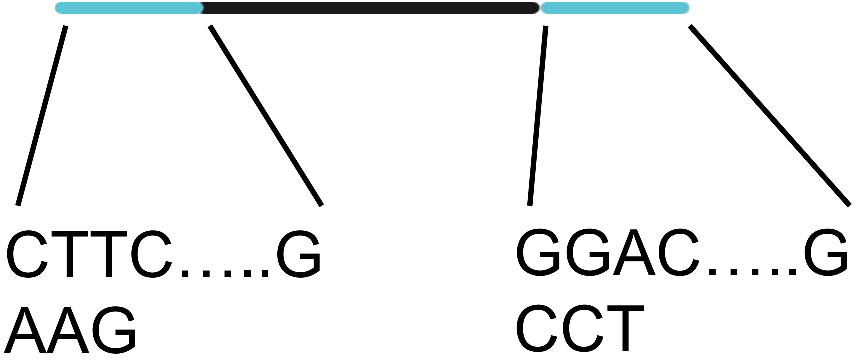
What are the needed statistical algorithms?

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

The data: paired-end RNA-seq



Matched sequences are obtained for each library molecule



CTTC…..G
AAG

GGAC…..G
CCT

The statistical model intuition

- 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 gets a λ value
- In statistics, we take observed data and use it to estimate parameters, in this case, λ

- This is formally accomplished by, for example the MLE
- In RNA seq, "RPKM" is conceptually like λ

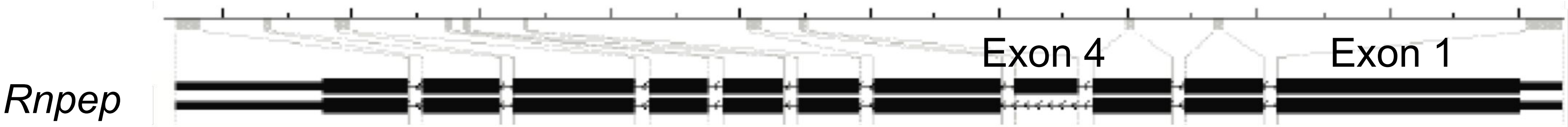
More on the model

- 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 makes the situation “much more complicated”

Intuition for the statistical problem



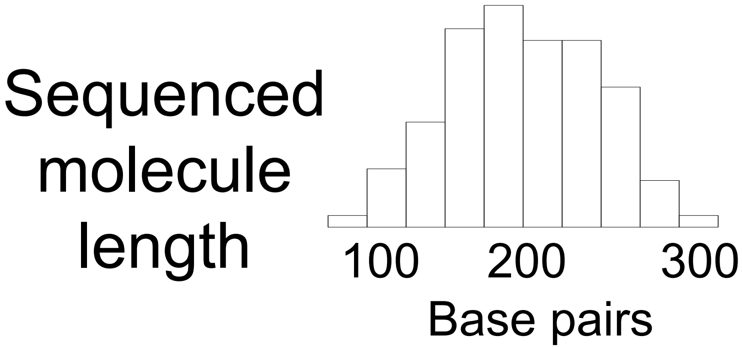
Estimate the expression of each isoform?

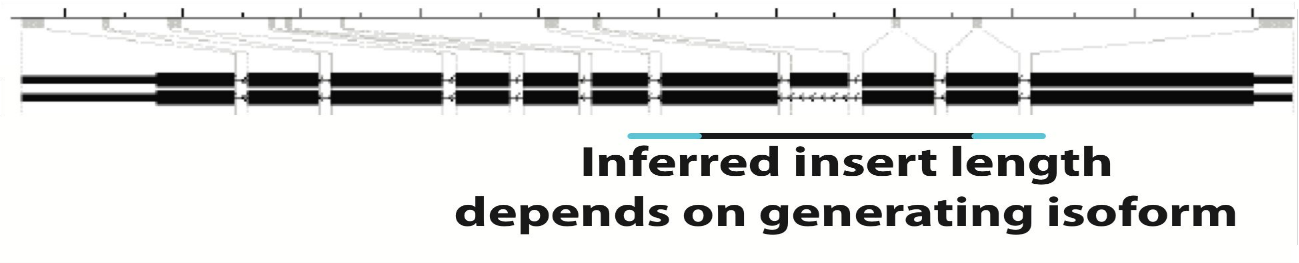
Nontrivial : we only observe fragments of sequences

- 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




Inferred insert length depends on generating isoform

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

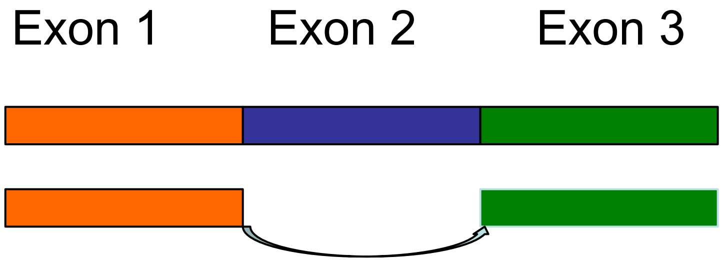
- Statistical improvement over naïve models
- Optimal information reduction
- Quantifies information gain using PE Sequencing

Why do we care: just fun math?

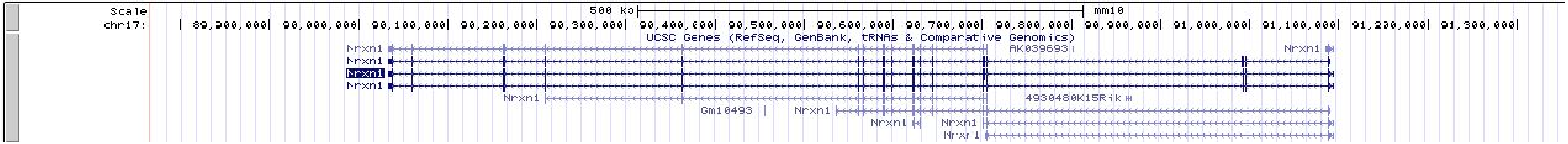
- Not knowing the isoforms means we don’t know the gene level expression
- Off the shelf tools are “mostly right” but many times wrong
- Most labs don’t use their latest published software
- Current tools only provide approximate answers

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

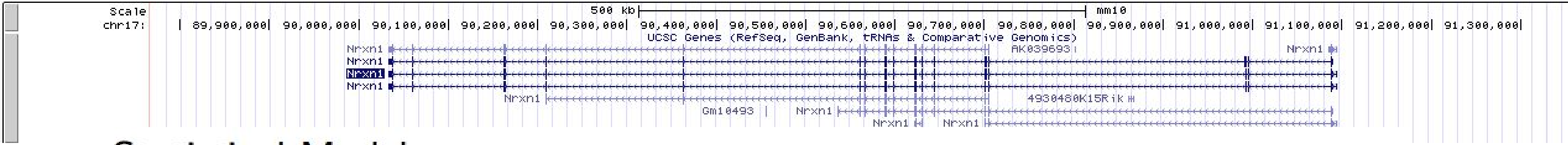


Exon 1 Exon 2 Exon 3



Sufficient statistics, statistical problem, Poisson models

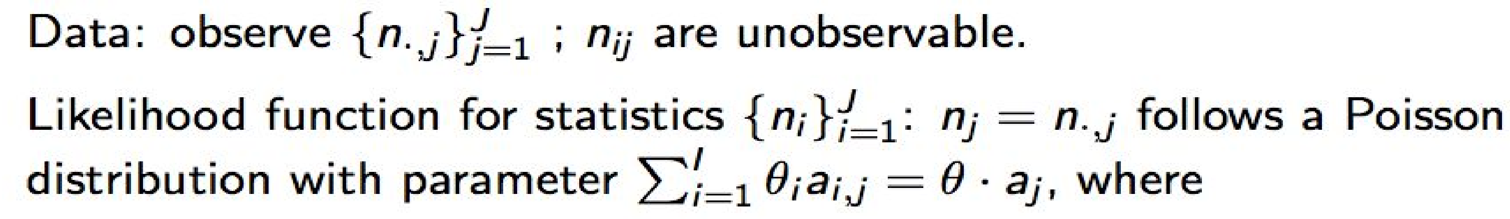
Formalizing the problem and model



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

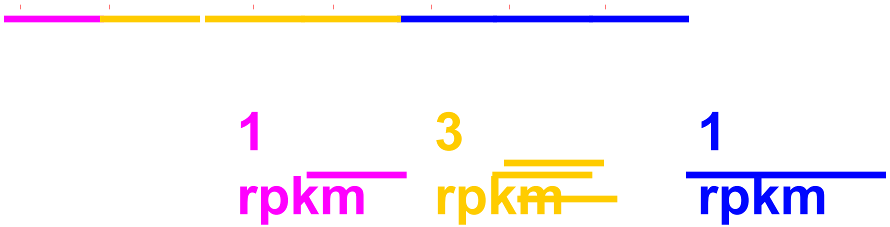


Data: observe $\{n_{\cdot,j}\}_{j=1}^{J}$; n_{ij} are unobservable.

Likelihood function for statistics $\{n_i\}_{i=1}^{J}$: $n_j = n_{\cdot,j}$ 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 application (biology) is impacted!



1 rpkm

3 rpkm

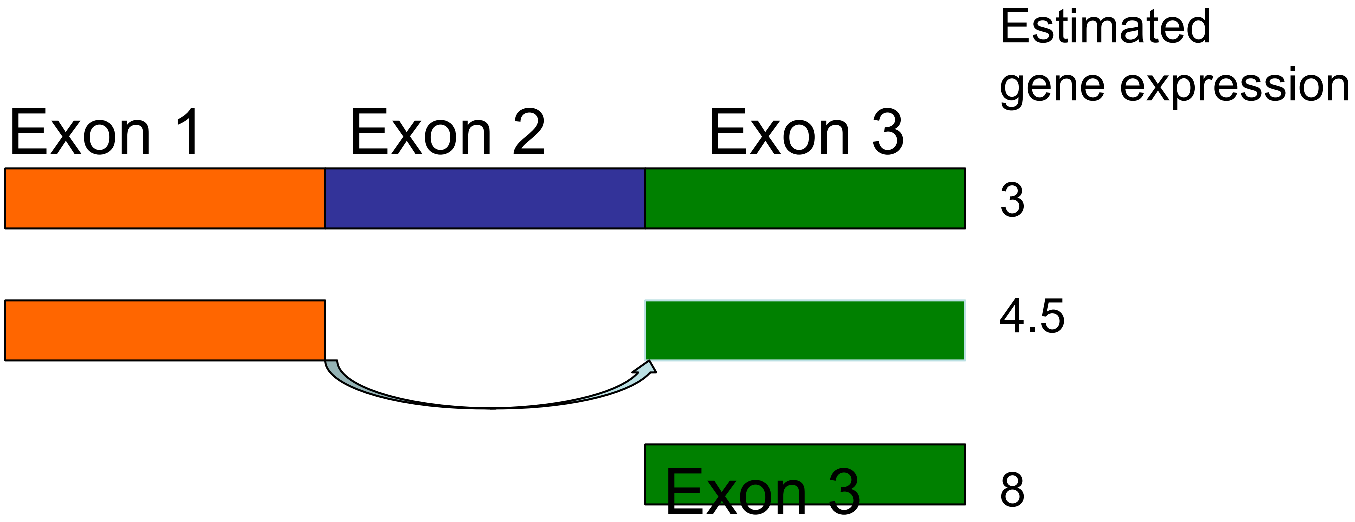
1 rpkm

Remember, RPKM is like lambda

The importance of statistics

Exon	**1**	**2**	**3**
Count	1	0	8

Remember, counts =“expression” in RNA-Seq

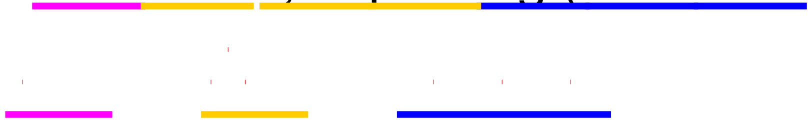


Without taking isoforms into account, gene expression estimates (and differential gene expression will be wrong)!

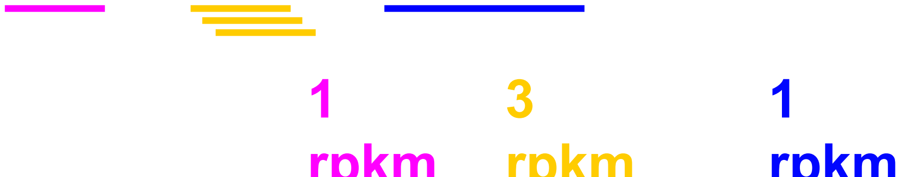
Gene and isoform expression are inextricably linked

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



Even more “problems”: count data is noisy

Example, idea: clean it up w/ robust statistics

Properties of statistical inference

1. Theoretically best
   a. Under the given null and alternative, test is best
   b. Fisher’s efficient estimator
   c. Uniformly Most Powerful test (illustration)
2. Fast
   a. Inexpensive to store data
      i. Reduction to sufficient or minimal sufficient statistics
   b. Computationally inexpensive
      i. Computing test statistics is simple
3. Mechanistic
   a. Tests and scientific/medical interventions easy to perform
   b. Few predictors, LASSO and NMF move in this direction

The first modern, efficient, theoretically tractable tests: Rank tests

1. Theoretically ~~best~~ tractable
2. Fast
   a. Computationally inexpensive
3. Inexpensive to store data

   Downside? Lose power

4. Next lectures will move onto more powerful tests

Rank tests

General idea:

1. Replace data by ranks
2. Perform a test on the ranked data to test if deviation from expectation

Advantage: requires simply sorting the data and a single computation

1. Sort time: O(n log n) (worst case, O(n^2): data storage benefits

Disadvantage: power (brainstorm example)

On board: derivation of Mann-Whitney test and introduction to random permutations

Mann-Whitney test

- Derivation, useful
- Conceptual example of how to apply approach in general
- Kruskall-Wallis

Theoretical analysis is interesting, but not required

Computing the null by simulation : more safeguards

How would we do this?

Lecture 2: bootstrap for significance testing

What is missing from rank test?

1. Power
2. Effect size calculations

Motivation by GTEx and IVT-Seq

Exon level data-- discovering relationships and isoforms?

Opportunities for discovery

Introduction to the GTEX data

Opportunities for discovery

GTEx -- statistical detection of splicing variants

Efficient approaches to statistical testing w/o knowlege of the null

-- motivating example, but important to learn history

Motivation by Gtex

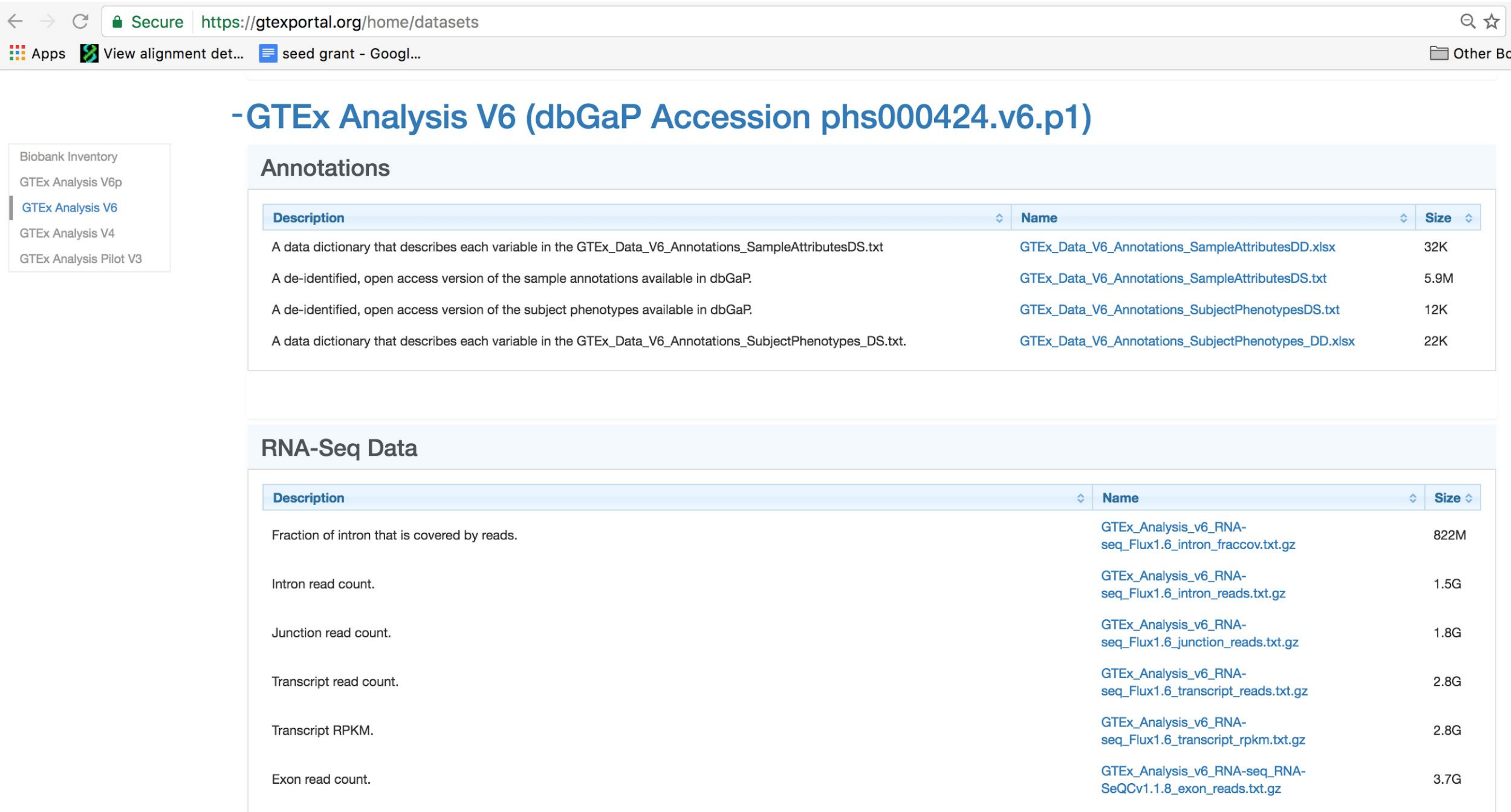
Describe data: clinical data https://gtexportal.org/home/datasets

And a great deal of information on genotype/RNA expression

https://gtexportal.org/home/tissueSummaryPage#cause

https://gtexportal.org/home/gene/SMN2

But, statistics are not interpretable



-GTEx Analysis V6 (dbGaP Accession phs000424.v6.p1)

Annotations

Description	Name	Size
A data dictionary that describes each variable in the GTEx_Data_V6_Annotations_SampleAttributesDS.txt	GTEx_Data_V6_Annotations_SampleAttributesDD.xlsx	32K
A de-identified, open access version of the sample annotations available in dbGaP.	GTEx_Data_V6_Annotations_SampleAttributesDS.txt	5.9M
A de-identified, open access version of the subject phenotypes available in dbGaP.	GTEx_Data_V6_Annotations_SubjectPhenotypesDS.txt	12K
A data dictionary that describes each variable in the GTEx_Data_V6_Annotations_SubjectPhenotypes_DS.txt.	GTEx_Data_V6_Annotations_SubjectPhenotypes_DD.xlsx	22K

RNA-Seq Data

Description	Name	Size
Fraction of intron that is covered by reads.	GTEx_Analysis_v6_RNA-seq_Flux1.6_intron_fraccov.txt.gz	822M
Intron read count.	GTEx_Analysis_v6_RNA-seq_Flux1.6_intron_reads.txt.gz	1.5G
Junction read count.	GTEx_Analysis_v6_RNA-seq_Flux1.6_junction_reads.txt.gz	1.8G
Transcript read count.	GTEx_Analysis_v6_RNA-seq_Flux1.6_transcript_reads.txt.gz	2.8G
Transcript RPKM.	GTEx_Analysis_v6_RNA-seq_Flux1.6_transcript_rpkm.txt.gz	2.8G
Exon read count.	GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_exon_reads.txt.gz	3.7G

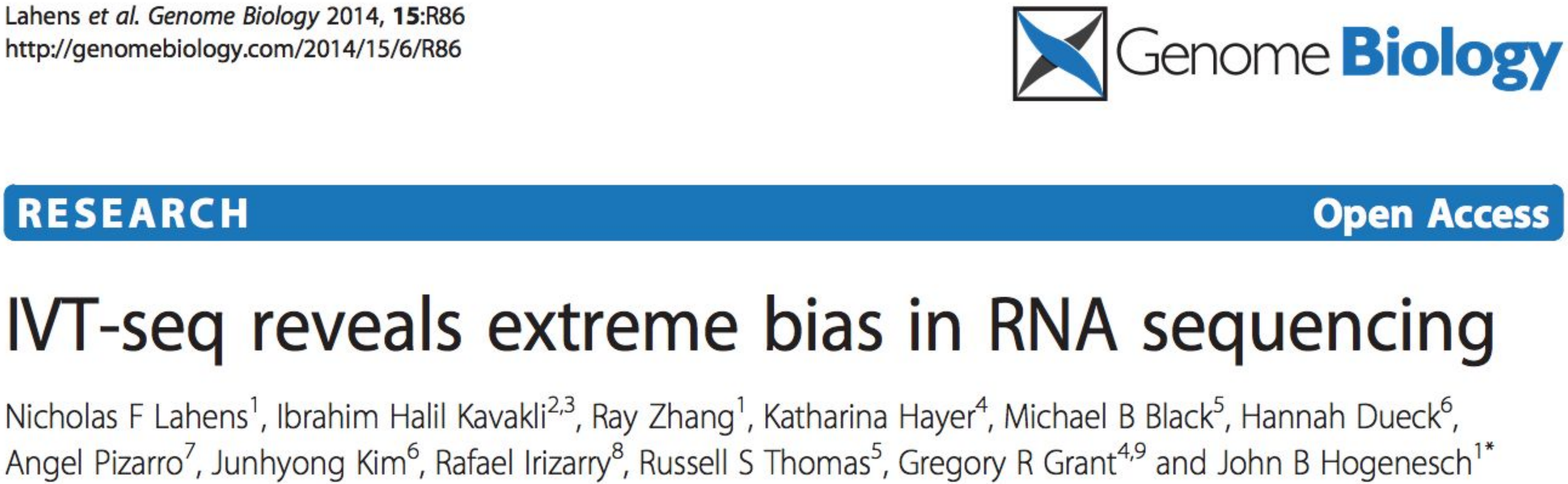
Motivation by Gtex

Describe question: differential isoform expression

Motivation by Gtex

Describe question: differential isoform expression

Extreme biases in RNA-seq: no theoretical null



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

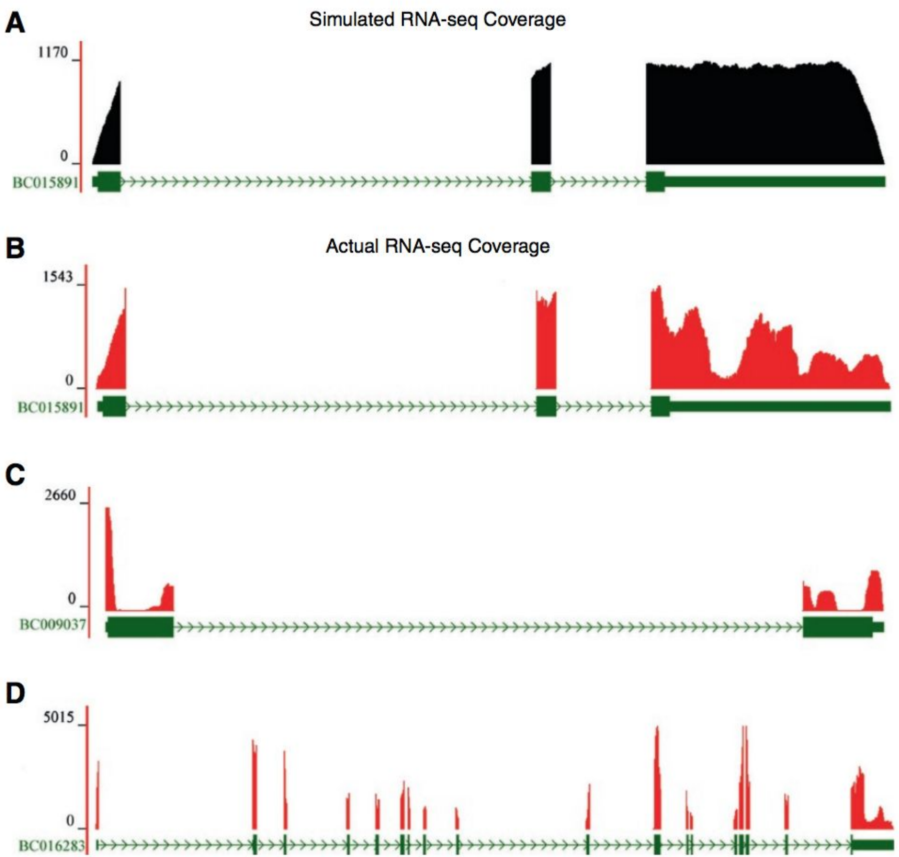
Genome **Biology**

RESEARCH **Open Access**

IVT-seq reveals extreme bias in RNA sequencing

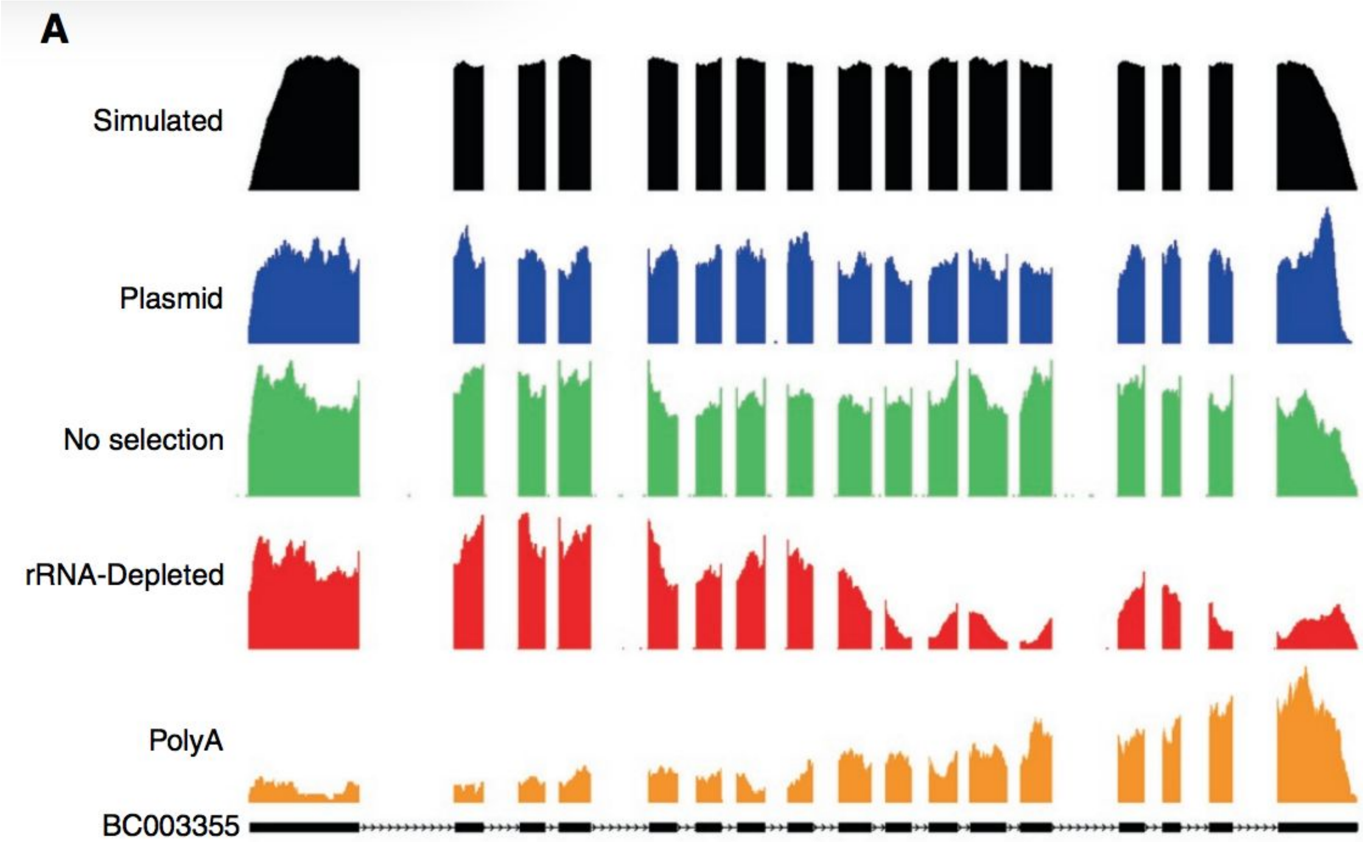
Nicholas F Lahens[1], Ibrahim Halil Kavakli[2,3], Ray Zhang[1], Katharina Hayer[4], Michael B Black[5], Hannah Dueck[6], Angel Pizarro[7], Junhyong Kim[6], Rafael Irizarry[8], Russell S Thomas[5], Gregory R Grant[4,9] and John B Hogenesch[1*]

Simulations and intuition don’t match real data



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

Selection and efficiency confound naive estimation



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

How do we overcome these problems?

- Learn statistical theory and methods
- Designing our own custom test that captures intuition, then analyze its properties

BREAK and brainstorm

Designing our own custom test that captures intuition, then analyze its properties

Go through procedure with real data

Give an example

Define bootstrap theory

Go through why this is true

Define bootstrap

Did we need to do the computation?

Classes of problems

Reduction to combinatorial CLT?

Use of permutation testing to control FDR

Example of permutation testing and FDR estimation

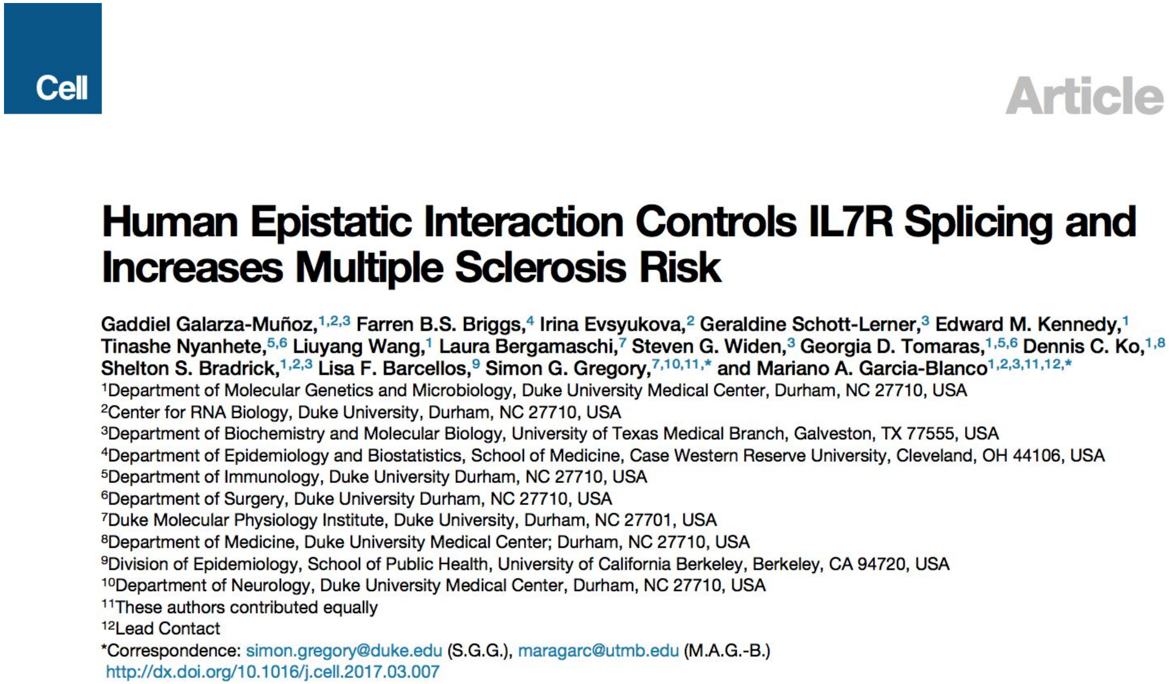
When the bootstrap breaks down?

Candes’ example

Lecture 3: speeding up testing

Biological motivation: many diseases are caused by dysregulated splicing

MS, a recent discovery



Cell

Article

Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk

Gaddiel Galarza-Muñoz,[1,2,3] Farren B.S. Briggs,[4] Irina Evsyukova,[2] Geraldine Schott-Lerner,[3] Edward M. Kennedy,[1] Tinashe Nyanhete,[5,6] Liuyang Wang,[1] Laura Bergamaschi,[7] Steven G. Widen,[3] Georgia D. Tomaras,[1,5,6] Dennis C. Ko,[1,8] Shelton S. Bradrick,[1,2,3] Lisa F. Barcellos,[9] Simon G. Gregory,[7,10,11,*] and Mariano A. Garcia-Blanco[1,2,3,11,12,*]

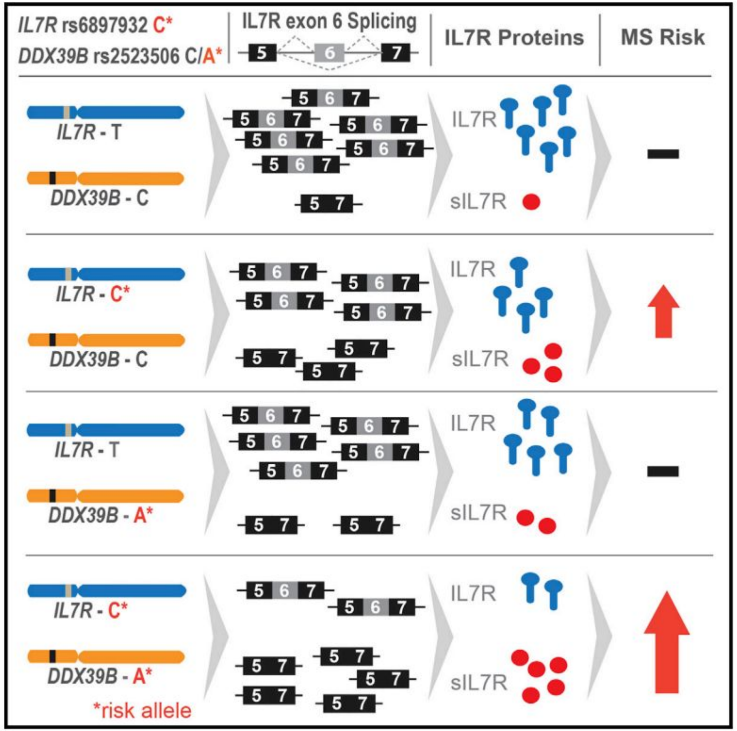
[1]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, USA
[2]Center for RNA Biology, Duke University, Durham, NC 27710, USA
[3]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555, USA
[4]Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA
[5]Department of Immunology, Duke University Durham, NC 27710, USA
[6]Department of Surgery, Duke University Durham, NC 27710, USA
[7]Duke Molecular Physiology Institute, Duke University, Durham, NC 27701, USA
[8]Department of Medicine, Duke University Medical Center; Durham, NC 27710, USA
[9]Division of Epidemiology, School of Public Health, University of California Berkeley, Berkeley, CA 94720, USA
[10]Department of Neurology, Duke University Medical Center, Durham, NC 27710, USA
[11]These authors contributed equally
[12]Lead Contact
*Correspondence: simon.gregory@duke.edu (S.G.G.), maragarc@utmb.edu (M.A.G.-B.)

http://dx.doi.org/10.1016/j.cell.2017.03.007

Important and interesting, suggests a bigger opportunity with massive data

Splicing pinpointed as ‘causal factor’ in MS

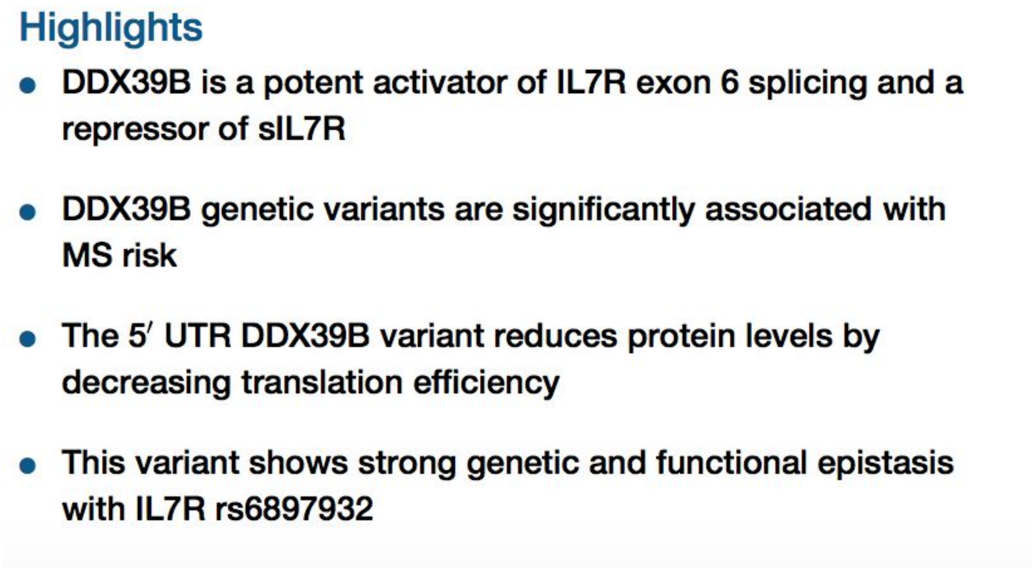
Graphical Abstract



IL7R splicing changes its interaction with the immune system

The splicing factor has a mutant with epistatic control over this variant

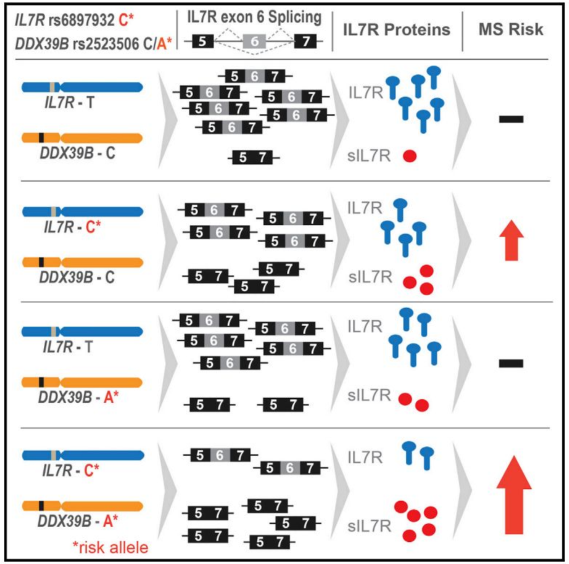
From genetics to mechanism



Highlights

- DDX39B is a potent activator of IL7R exon 6 splicing and a repressor of sIL7R
- DDX39B genetic variants are significantly associated with MS risk
- The 5′ UTR DDX39B variant reduces protein levels by decreasing translation efficiency
- This variant shows strong genetic and functional epistasis with IL7R rs6897932

Graphical Abstract

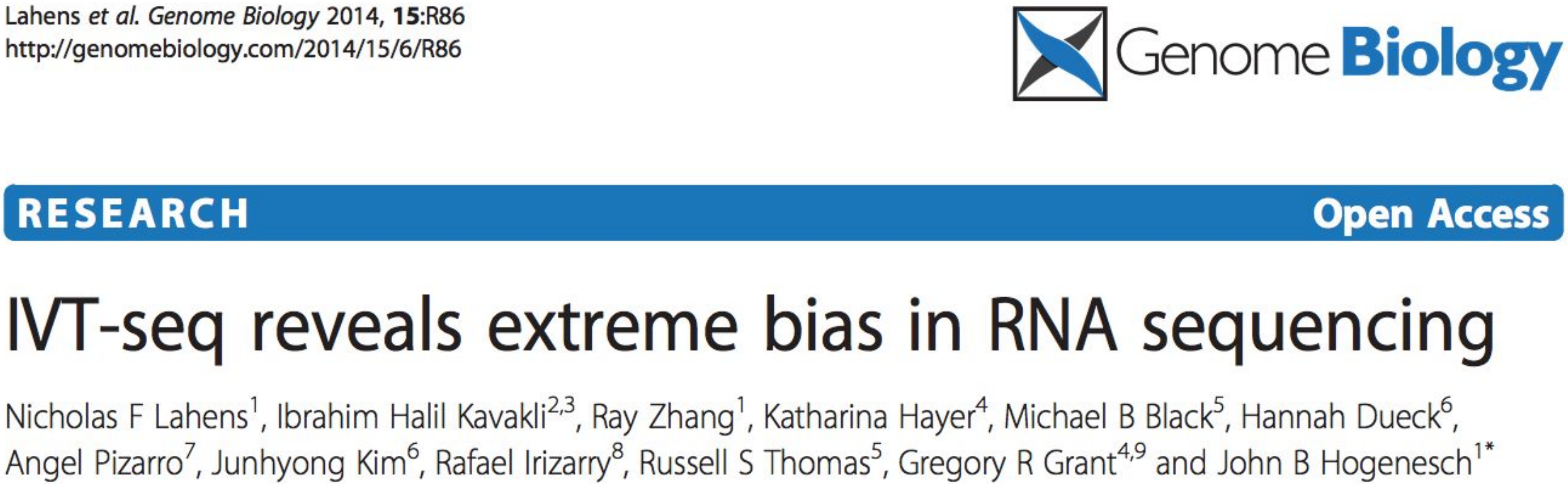


The fantasy of RNA-seq: perfect statistical modeling

1. Many models assume each RNA isoform is sampled at Poisson(a) where a is a constant proportional to the abundance of the transcript
2. Modified models use the negative binomial
3. These assumptions doesn’t hold, as we will see
4. (similar problems with DNA)

Testing for differential expression of RNA requires non-parametric approaches

Extreme biases in RNA-seq: no theoretical null



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

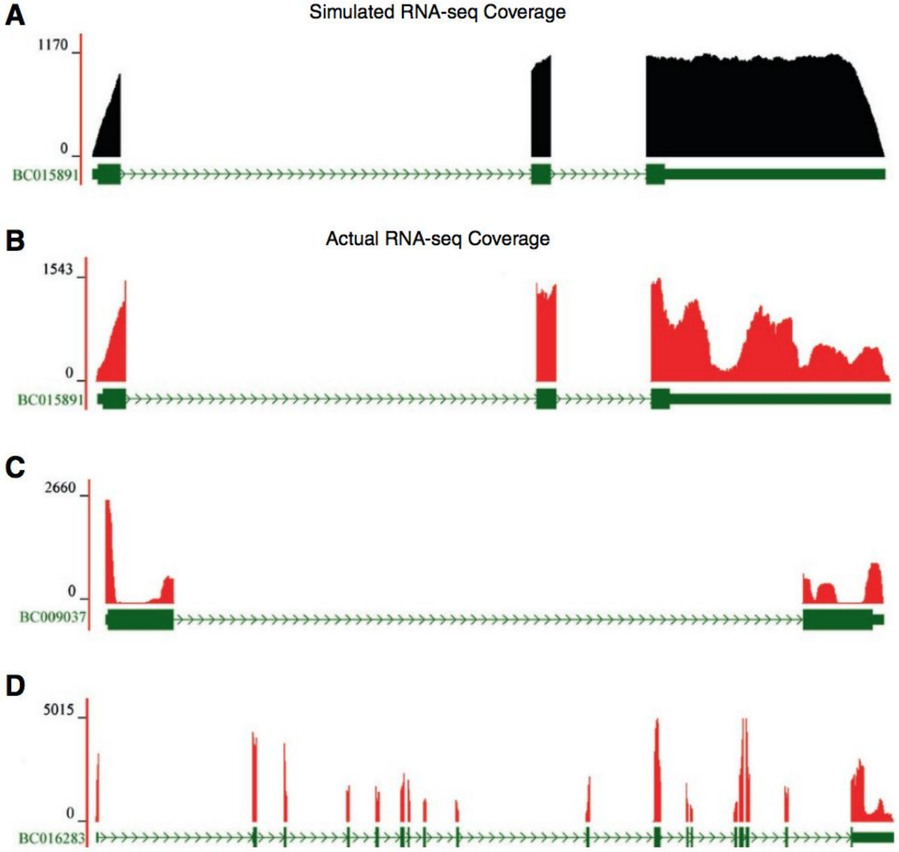
Genome **Biology**

RESEARCH **Open Access**

IVT-seq reveals extreme bias in RNA sequencing

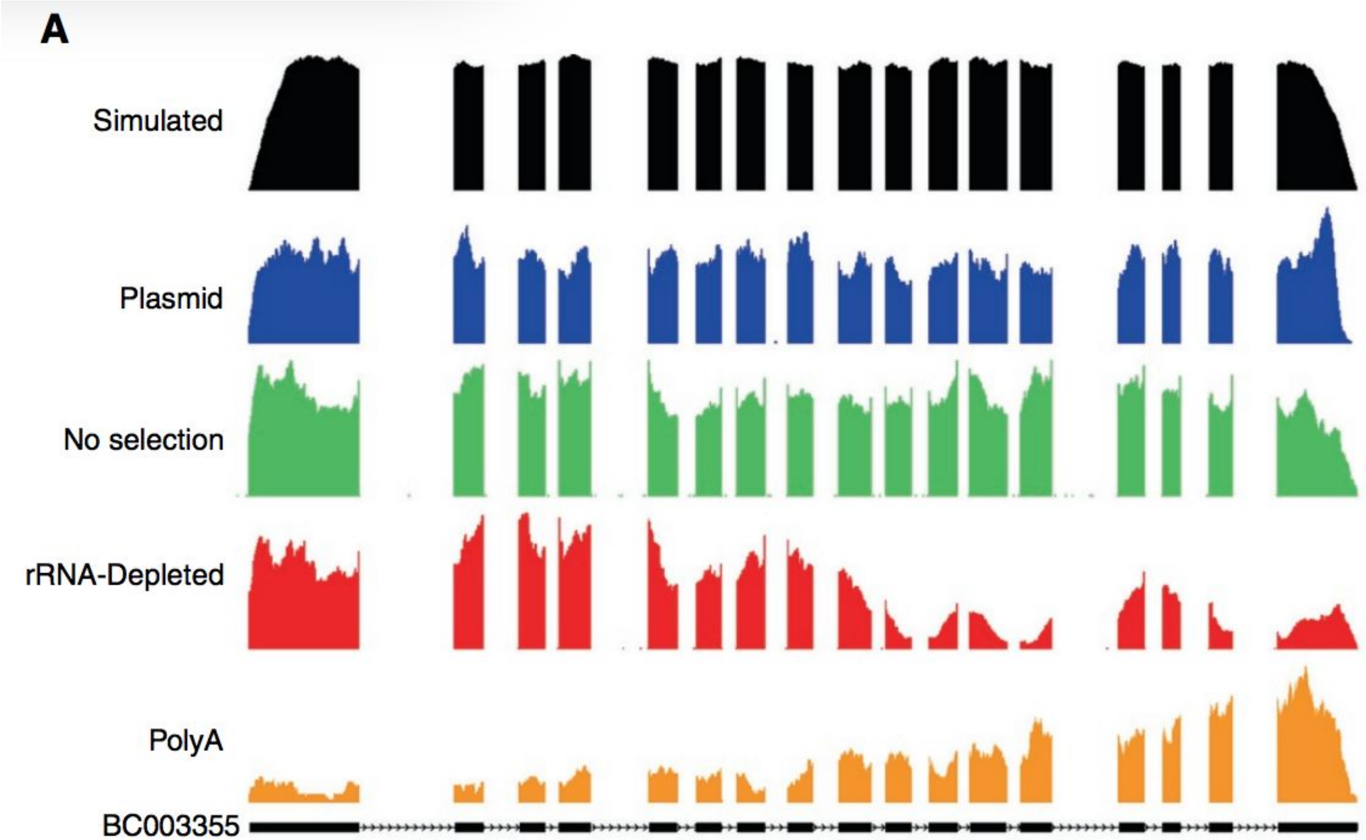
Nicholas F Lahens[1], Ibrahim Halil Kavakli[2,3], Ray Zhang[1], Katharina Hayer[4], Michael B Black[5], Hannah Dueck[6], Angel Pizarro[7], Junhyong Kim[6], Rafael Irizarry[8], Russell S Thomas[5], Gregory R Grant[4,9] and John B Hogenesch[1*]

Simulations and intuition don’t match real data



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

Selection and efficiency confound naive estimation



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

Modeling differential isoform expression

- Bias means that we can’t rely on closed form theoretical distribution
- Have to model the exon-level bias empirically
- Some approaches exist, but what if you want a robust new algorithm?

ILR7 example: genetic interactions with splicing

- Some approaches exist, but what if you want a robust new algorithm?
- Needs to be fast
- Every simulation “counts”
-