Statistical algorithms towards precision genomics

Julia Salzman
Department of Biochemistry and Biomedical Data Science

BMI-215 2018

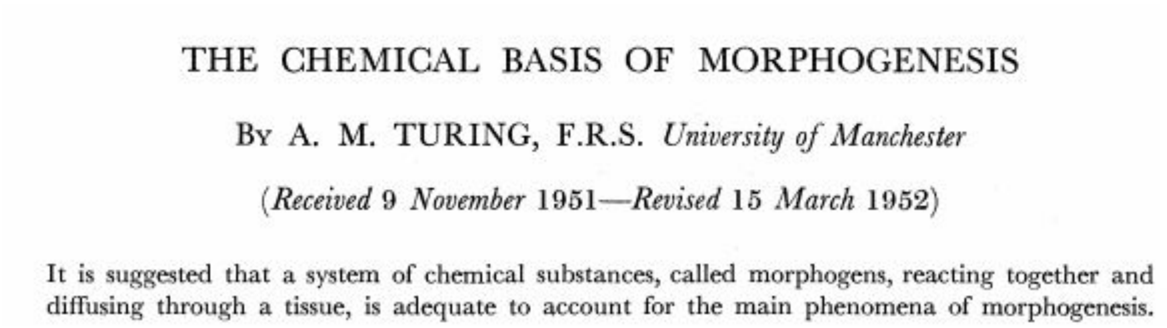
Introduction to RNA/DNA sequence genomics

Digitalization of biology, a history in statistics

Key biological principles were discovered by the founders of computer science and statistics using mathematical modeling

Images of Alan Turing and R. A. Fisher from Wikipedia

Mathematical theory in biology

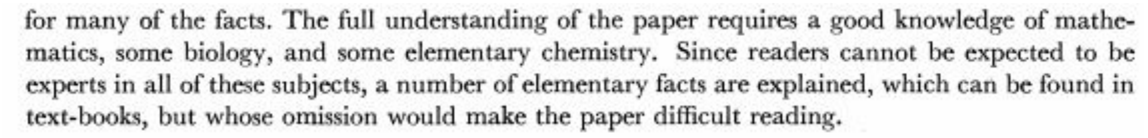


THE CHEMICAL BASIS OF MORPHOGENESIS

BY A. M. TURING, F.R.S. *University of Manchester*

(*Received* 9 *November* 1951—*Revised* 15 *March* 1952)

It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis.



for many of the facts. The full understanding of the paper requires a good knowledge of mathematics, some biology, and some elementary chemistry. Since readers cannot be expected to be experts in all of these subjects, a number of elementary facts are explained, which can be found in text-books, but whose omission would make the paper difficult reading.

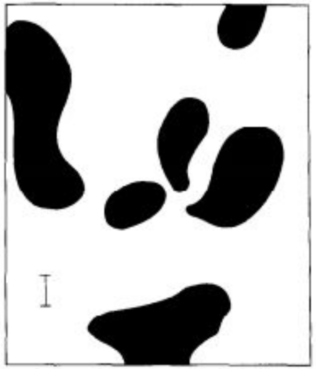


Figure 2. An example of a "dappled" pattern as resulting from a type (*a*) morphogen system. A marker of unit length is shown (see text, Sections 9 and 11).

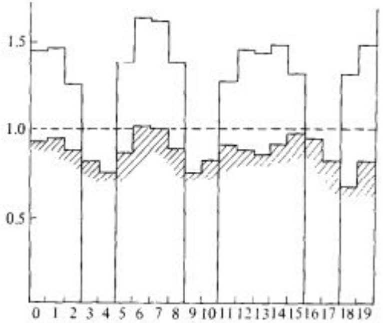


Fig. 3. Concentrations of *Y* in the development of the first specimen (taken from Table 1): (---------) original homogeneous equilibrium; (/////////) incipient pattern; (———) final equilibrium.

Overview

- Biomedical background: DNA, RNA and its role in disease
  - RNA: the new medicine, and the promise for biomedical data science
  - What biomedical problems can be studied?
  - What analytic tools are needed?
    - Biomedical background
    - Foundations for RNA-seq algorithms and analysis

Opportunities with statistics

From genome to phenome: the expanding role of RNA

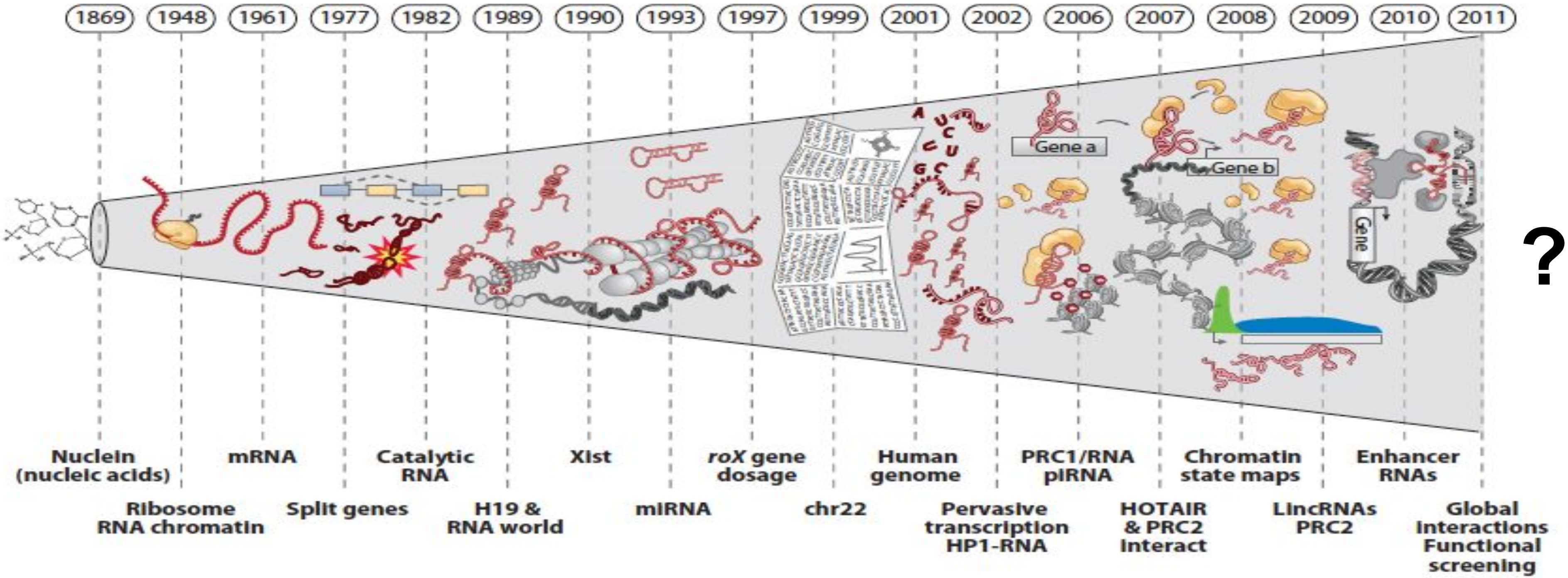


Figure from Rinn et al, 2012

Human variation and dysregulation in disease: **more than the DNA**

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

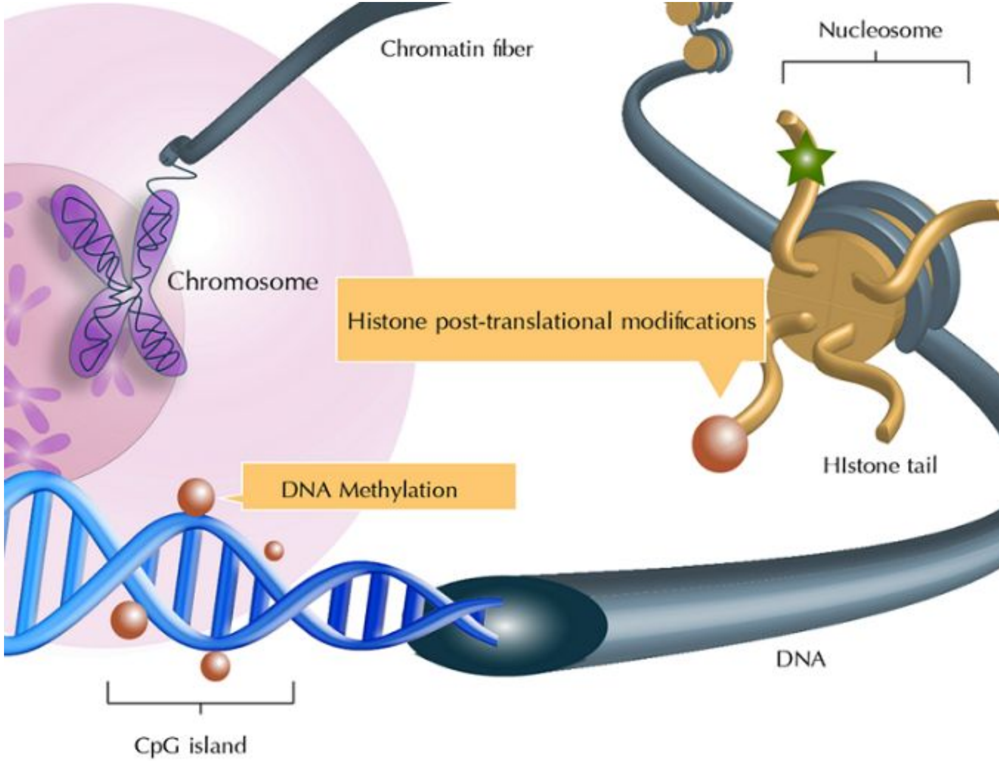
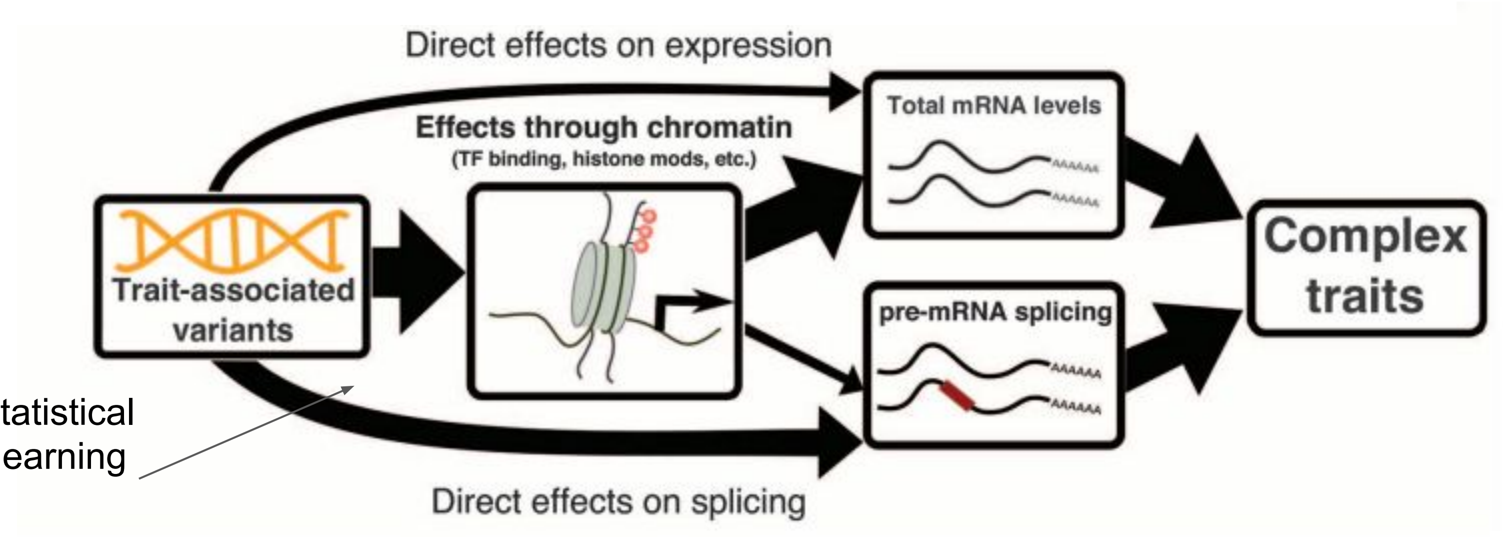


Image from http://www.visembryo.com/images/DNA%20methylation%20Imperial%20UK.png

RNA processing, splicing, a biomedical mystery


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

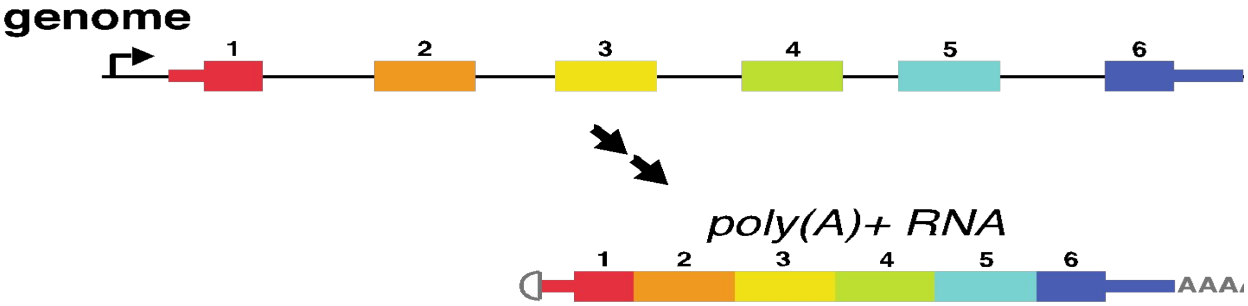
Find using statistical or machine learning approaches

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?

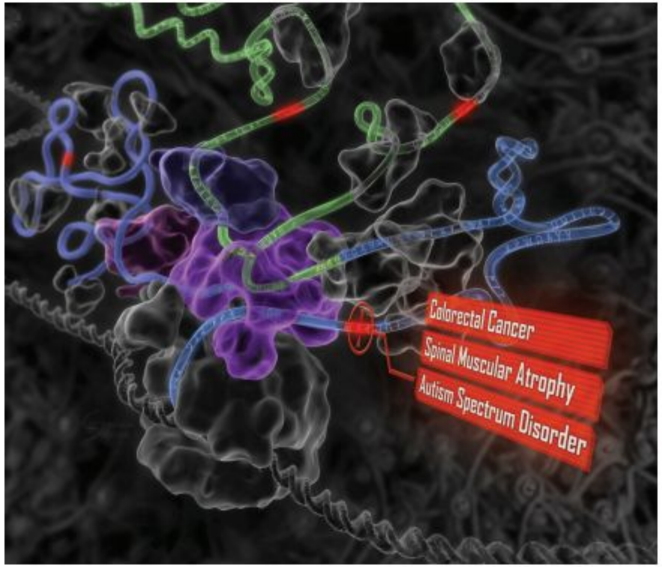


- A/C/G/T code for how to process RNA is only partially understood, even in 2018!
- Circular RNA, new biomarkers, new functions?

Splicing is a cellular code yet to be broken

Mechanistic insights from massive data?

Analysis of Discrete data



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

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

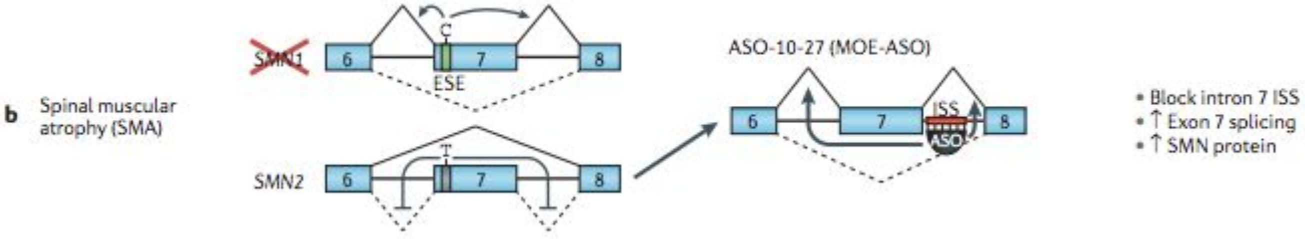
- What is the “cause”
- What is the consequence?

→ how can the disease be treated?

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

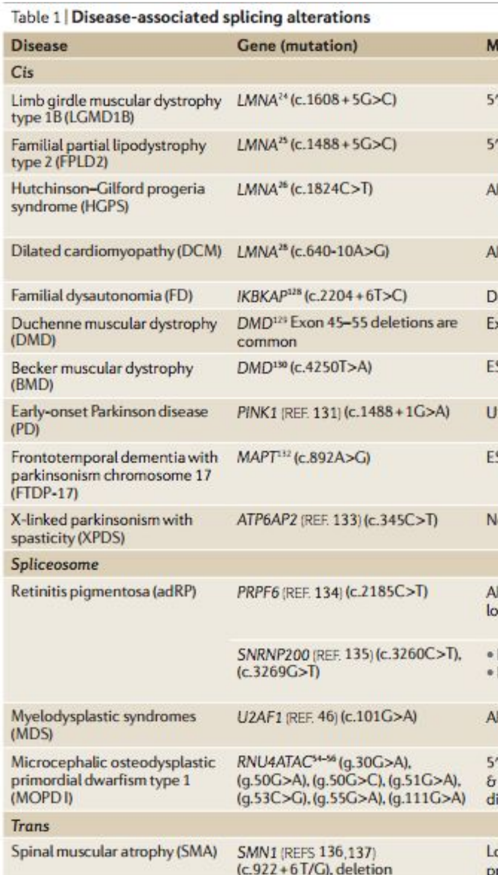
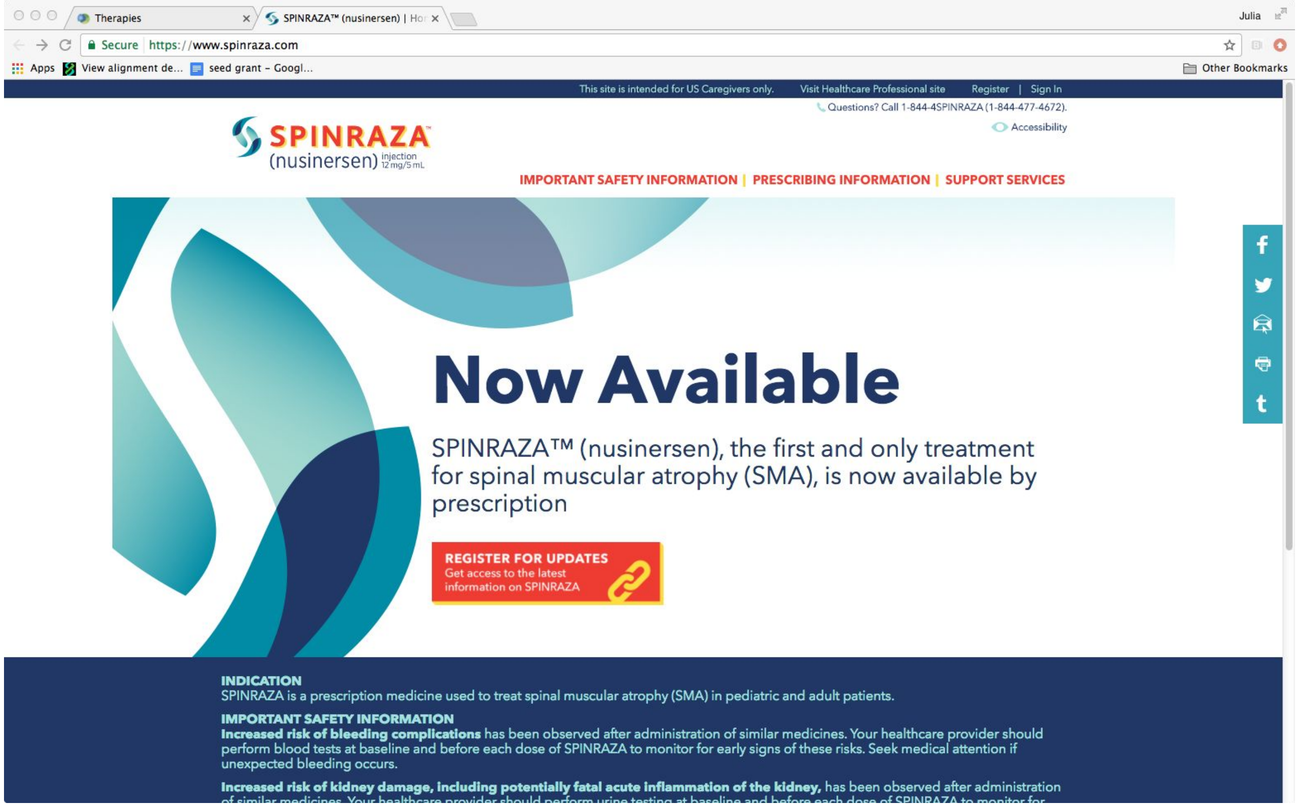


Table 1 | **Disease-associated splicing alterations**

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*[125] Exon 45–55 deletions are common	Ex
Becker muscular dystrophy (BMD)	*DMD*[130] (c.4250T>A)	ES
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)	ES
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

SMA: the first drug, an RNA



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

SPINRAZA™
(nusinersen) injection 12 mg/5 mL

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

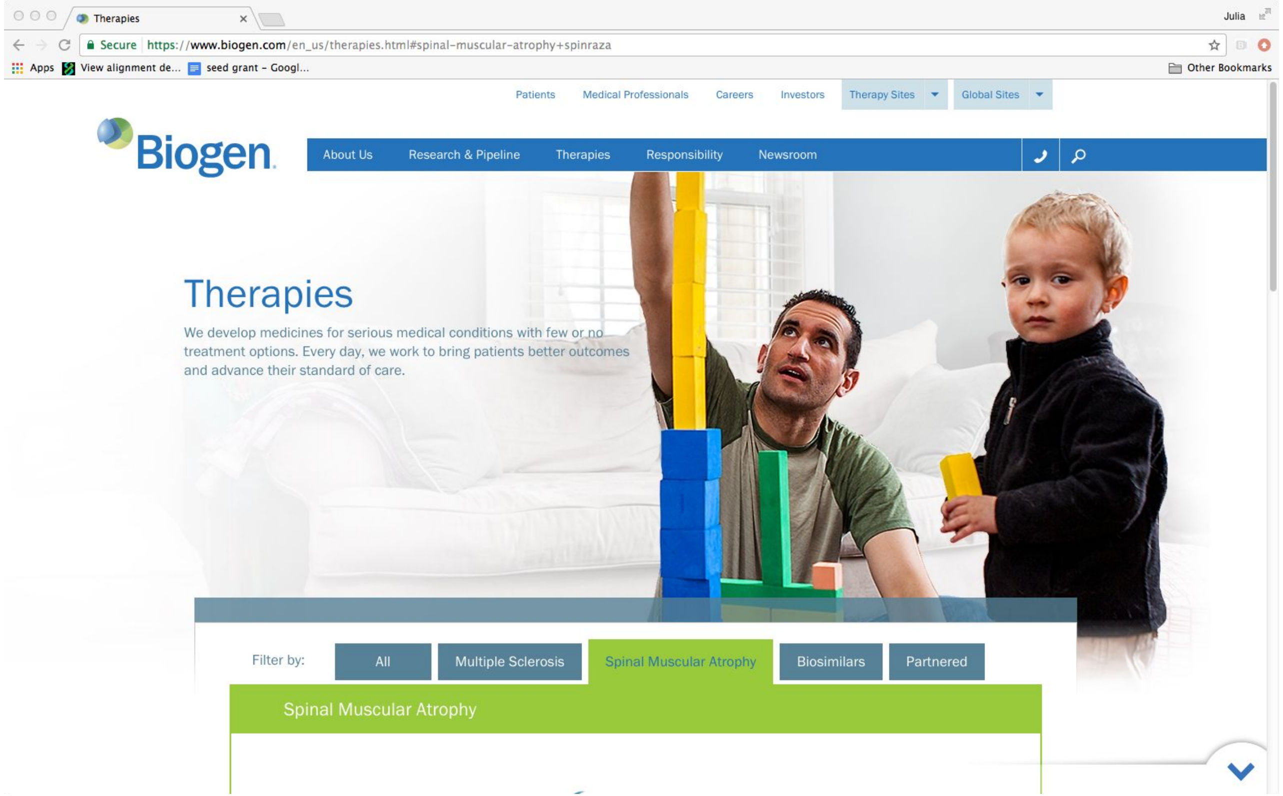
REGISTER FOR UPDATES
Get access to the latest information on SPINRAZA

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

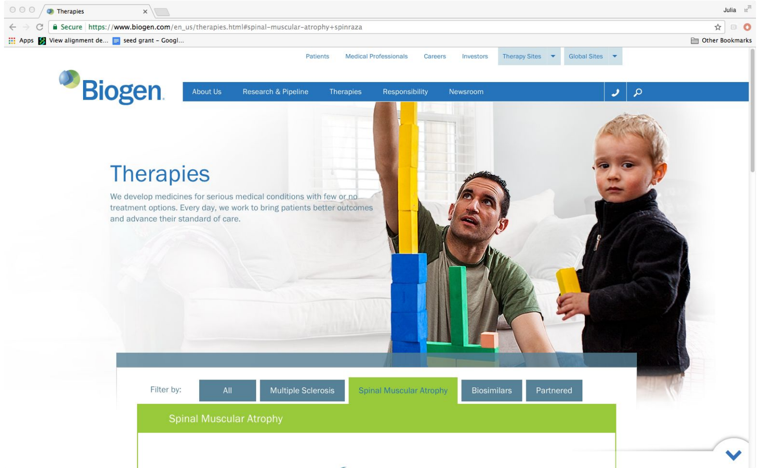
We develop medicines for serious medical conditions with few or no treatment options. Every day, we work to bring patients better outcomes and advance their standard of care.

Disease-causing RNA variants?

I. Prerequisite: statistical algorithms for splice detection

1. circular and linear RNA
2. **gene fusion detection -- cancer**
3. **de novo sequence detection -- many diseases**

Early disease detection



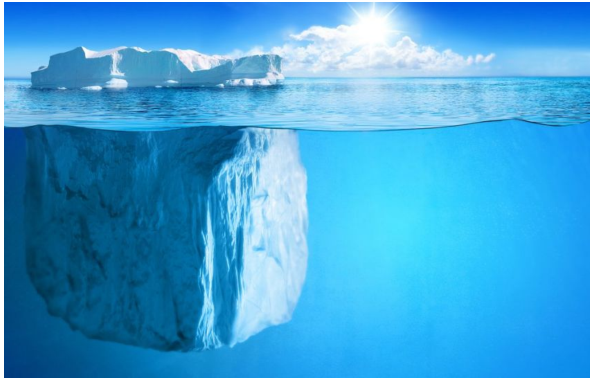
https://www.biogen.com/en_us/therapies.html#spinal-muscular-atrophy+spinraza

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.

Spinal Muscular Atrophy

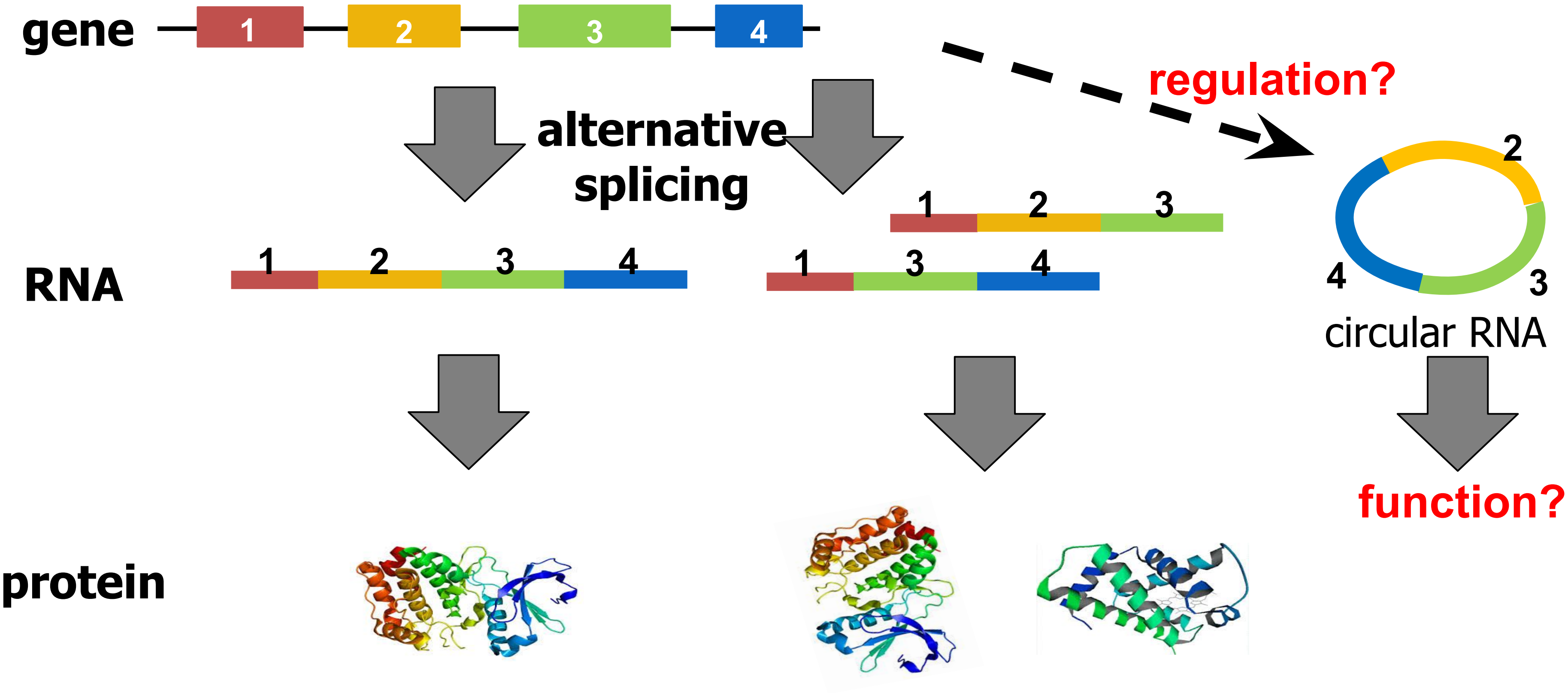
We know important information is missing!



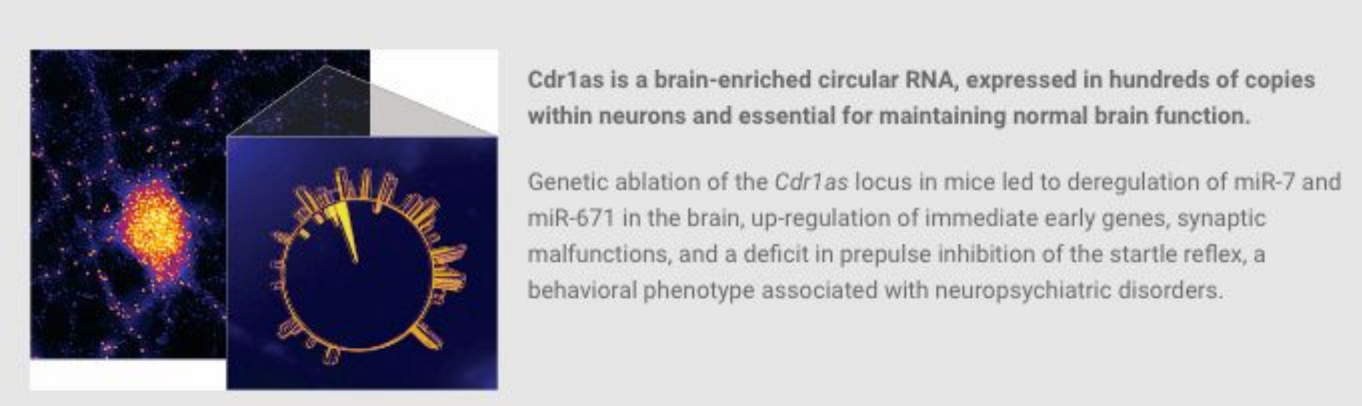
Current annotations/knowledge

Disease causing variants

Circular RNA:
expanding the role of protein-coding genes

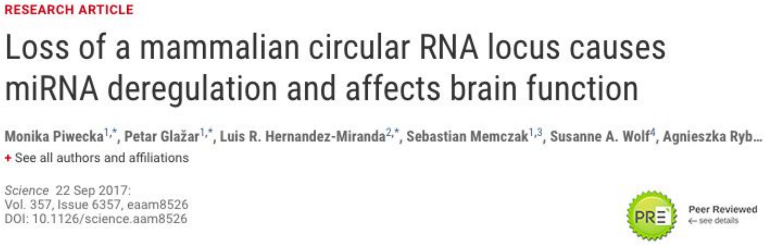


Circular RNAs likely have function in the nervous system



Cdr1as is a brain-enriched circular RNA, expressed in hundreds of copies within neurons and essential for maintaining normal brain function.

Genetic ablation of the *Cdr1as* locus in mice led to deregulation of miR-7 and miR-671 in the brain, up-regulation of immediate early genes, synaptic malfunctions, and a deficit in prepulse inhibition of the startle reflex, a behavioral phenotype associated with neuropsychiatric disorders.



RESEARCH ARTICLE

Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function

Monika Piwecka[1,*], Petar Glažar[1,*], Luis R. Hernandez-Miranda[2,*], Sebastian Memczak[1,3], Susanne A. Wolf[4], Agnieszka Ryb...

+ See all authors and affiliations

Science 22 Sep 2017:
Vol. 357, Issue 6357, eaam8526
DOI: 10.1126/science.aam8526

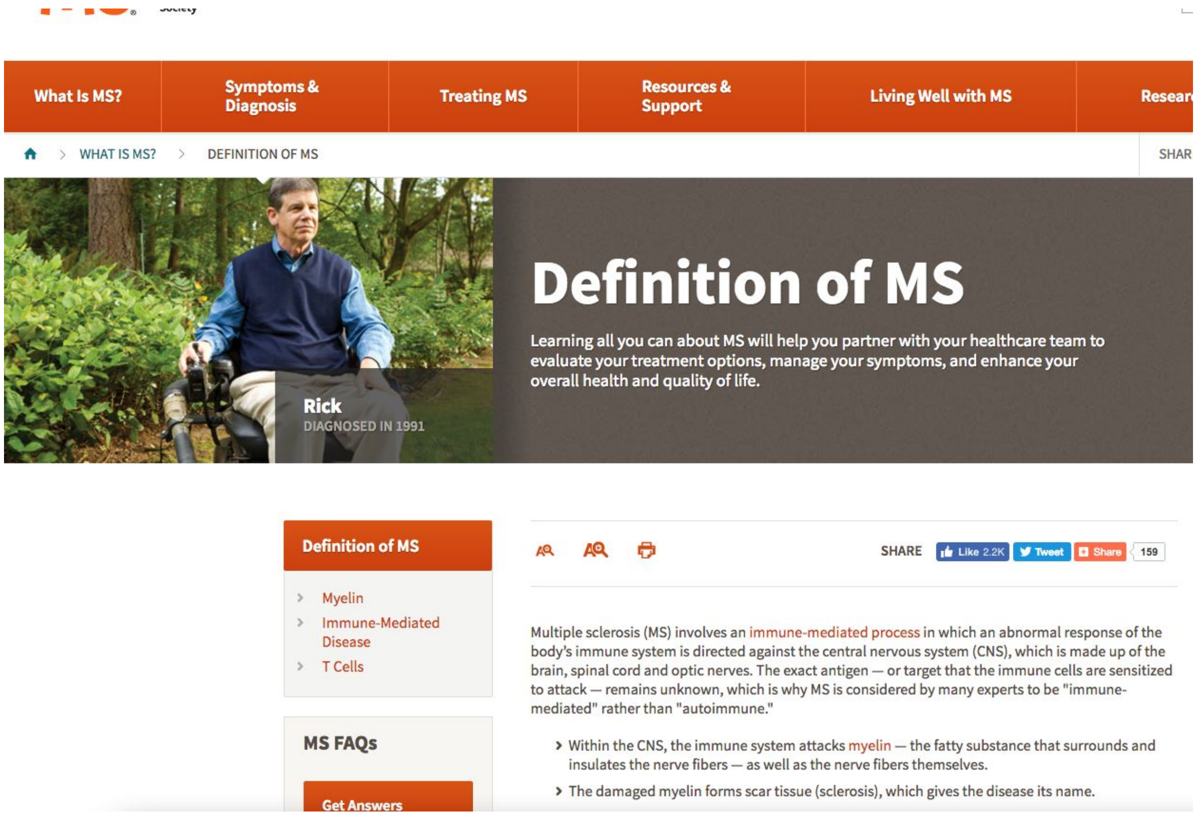
Why better statistical algorithms are needed

- Many algorithms are now ‘patched’ to detect circRNA, but still inaccurate for detecting RNA ‘normal’ cells

- Those patches don’t fix blind spots for other biomedically critical RNA splicing events

- Missing circular RNA means more biomedically relevant RNA are missed in biomedical context, especially cancer and neurodegeneration

Applications for high dimensional statistical inference?



Definition of MS

Learning all you can about MS will help you partner with your healthcare team to evaluate your treatment options, manage your symptoms, and enhance your overall health and quality of life.

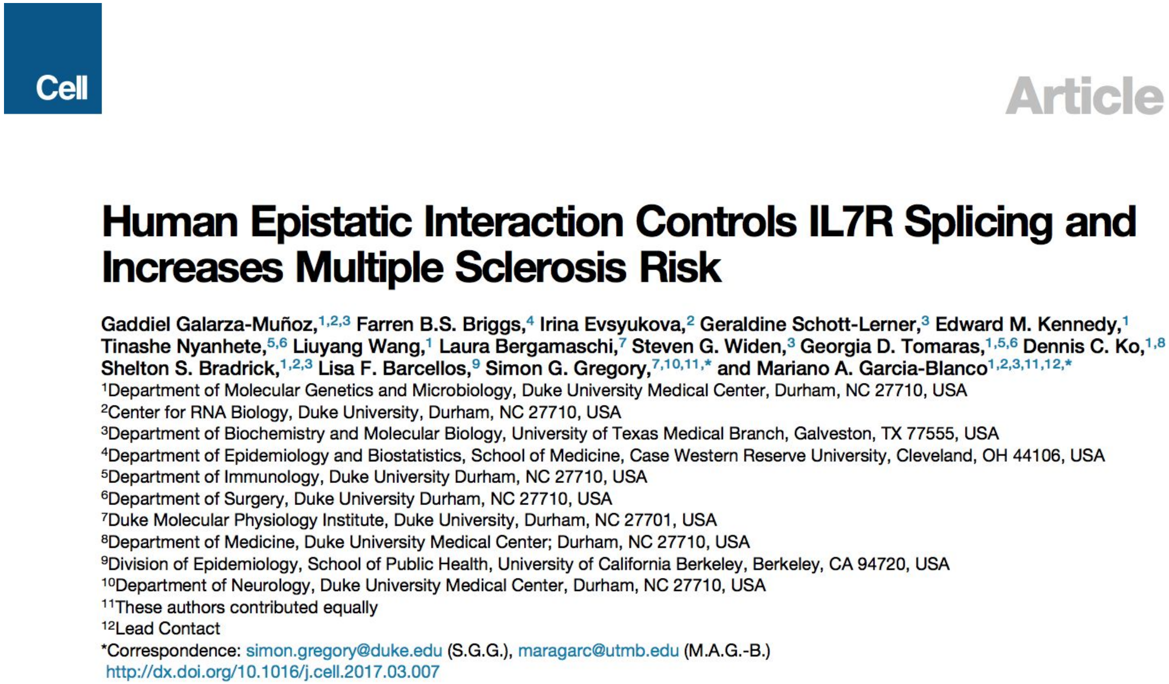
Multiple sclerosis (MS) involves an immune-mediated process in which an abnormal response of the body's immune system is directed against the central nervous system (CNS), which is made up of the brain, spinal cord and optic nerves. The exact antigen — or target that the immune cells are sensitized to attack — remains unknown, which is why MS is considered by many experts to be "immune-mediated" rather than "autoimmune."

- Within the CNS, the immune system attacks myelin — the fatty substance that surrounds and insulates the nerve fibers — as well as the nerve fibers themselves.
- The damaged myelin forms scar tissue (sclerosis), which gives the disease its name.

http://www.nationalmssociety.org/What-is-MS/Definition-of-MS

Biological motivation: many diseases are caused by dysregulated splicing

A recent discovery in MS



Cell

Article

Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk

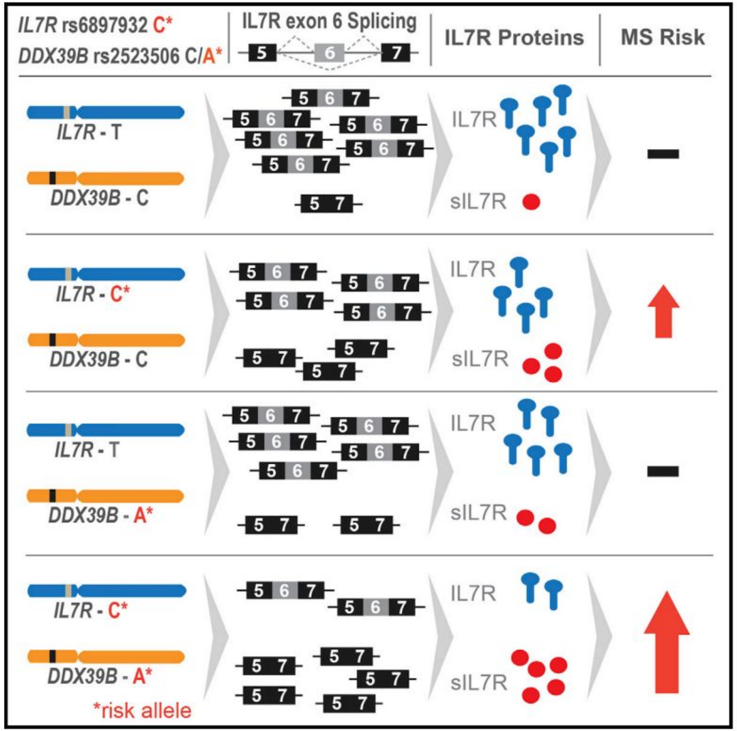
Gaddiel Galarza-Muñoz,[1,2,3] Farren B.S. Briggs,[4] Irina Evsyukova,[2] Geraldine Schott-Lerner,[3] Edward M. Kennedy,[1] Tinashe Nyanhete,[5,6] Liuyang Wang,[1] Laura Bergamaschi,[7] Steven G. Widen,[3] Georgia D. Tomaras,[1,5,6] Dennis C. Ko,[1,8] Shelton S. Bradrick,[1,2,3] Lisa F. Barcellos,[9]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



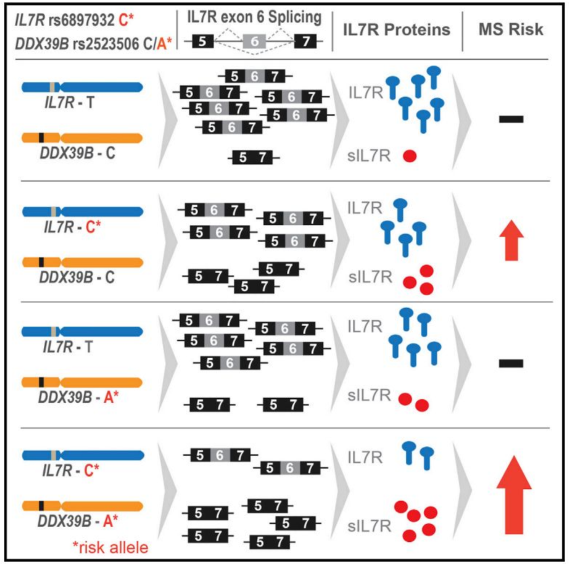
IL7R splicing changes its interaction with the immune system

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

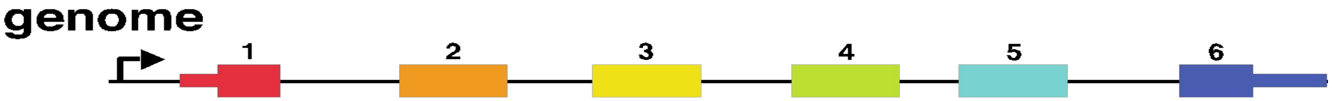
From genetics to mechanism

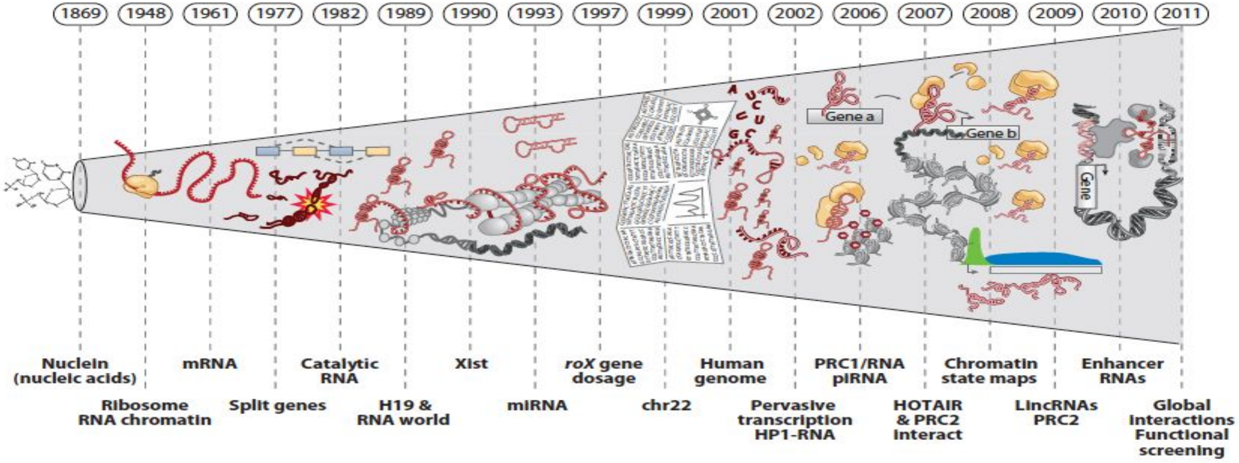
- This discovery was made w/ GWAS and time-intensive experiments
- Risk allele is known
- Pull down proteins associated w/ gene
- Experimentally test for synergy in risk allele
- Unexplained mysteries in some ethnic groups (for example)
- **Better 'systems biology' of splicing and gene expression**

Graphical Abstract

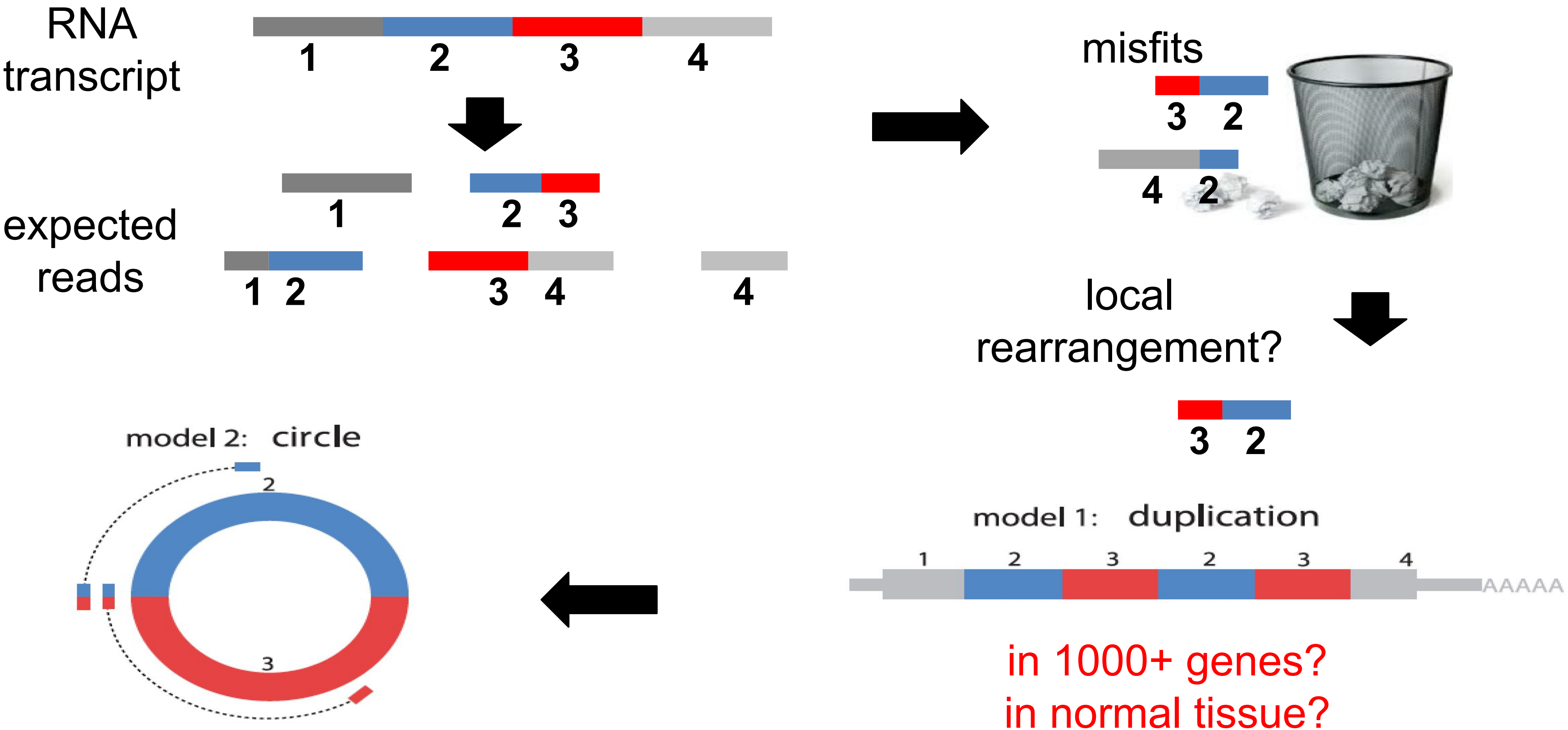


Need for statistical quantification of RNA variants





Circular RNA



RNA transcript

expected reads

misfits

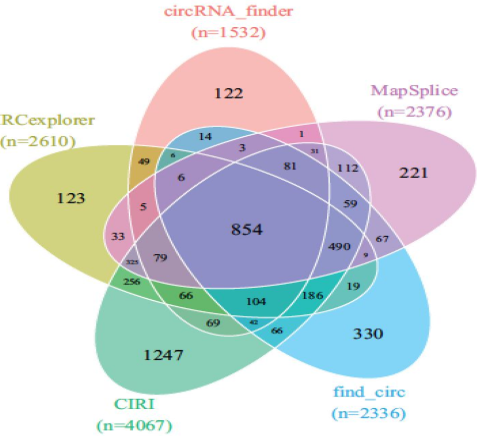
local rearrangement?

model 1: duplication

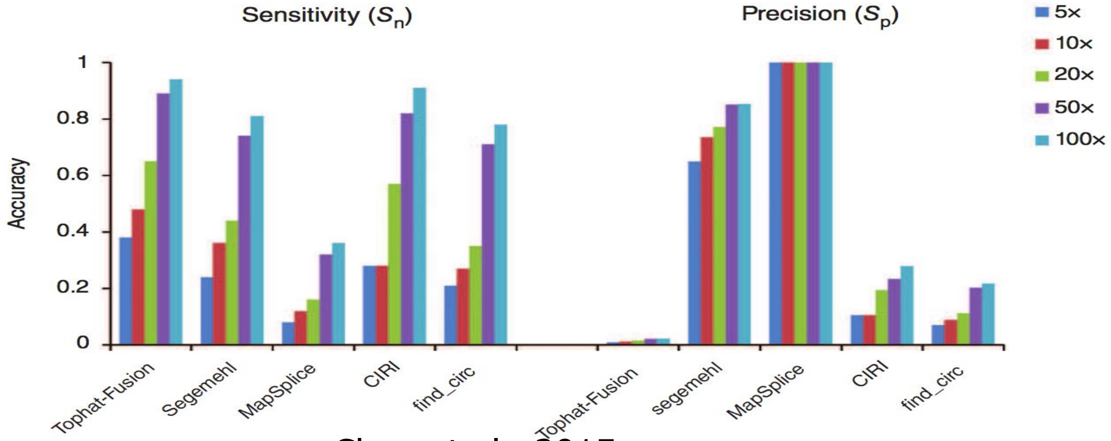
model 2: circle

in 1000+ genes?
in normal tissue?

Linear algorithms patched for circRNA



Hansen et al., 2015



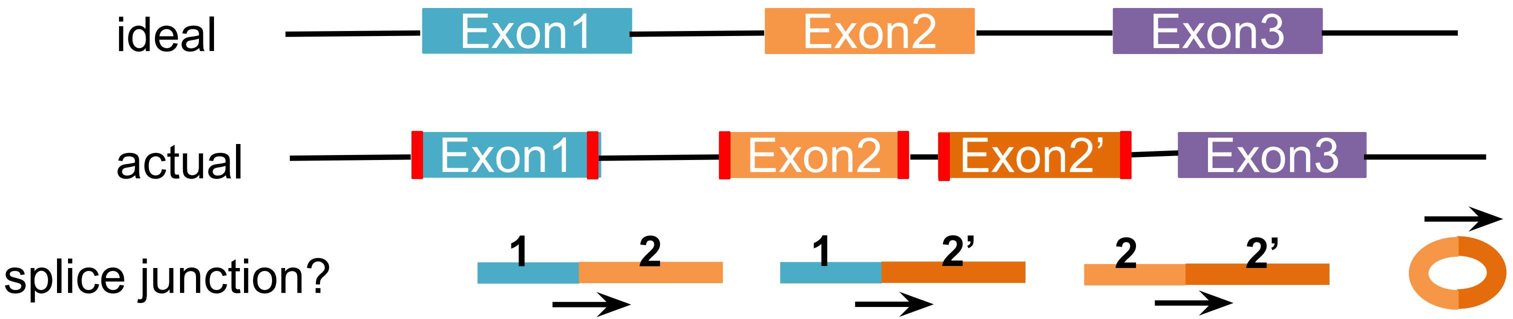
Chen et al., 2015

- Low overlap in genome-wide predictions of circRNA
- Trade-off between sensitivity and specificity
- **Not measuring how well algorithms discover disease variants**
- **Statistical problem: how to discovery and quantify splicing precisely? Need to learn applied statistical methodology**

Splice detection in RNA-Seq is an important unsolved problem

- circular RNA was overlooked in 30+ years of studying splicing

- RNA-Seq: complete and precise transcriptome reconstruction?

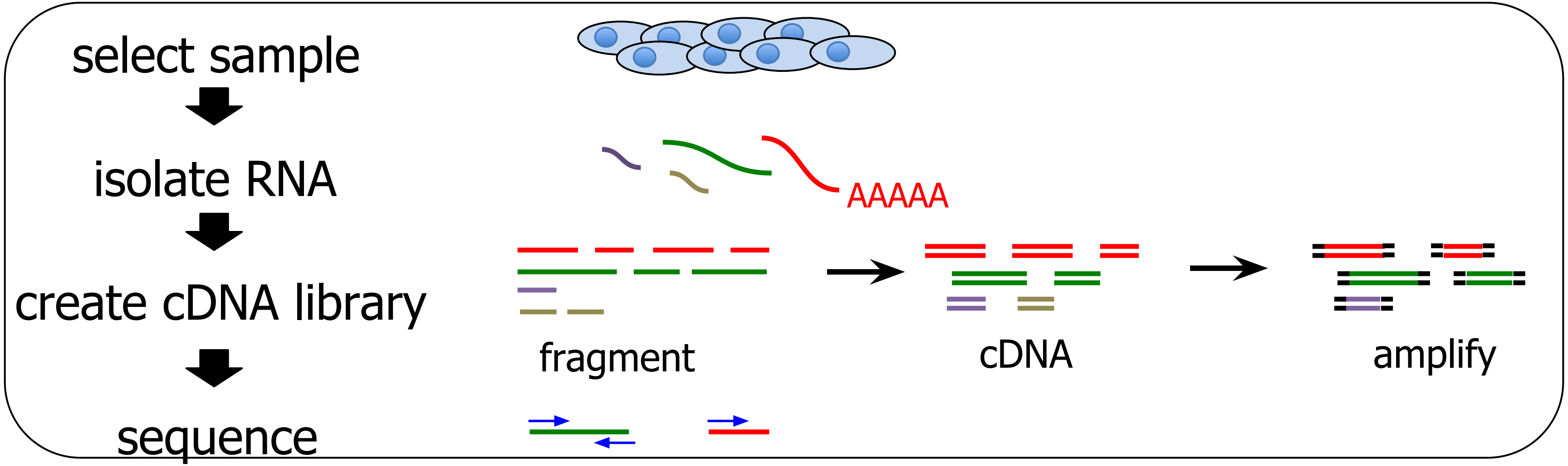


- > 95% of human genes produce alternative isoforms

- splicing errors and dysregulation known to cause many diseases

Biases and errors complicate RNA-Seq analysis

RNA-Seq Experiment



select sample

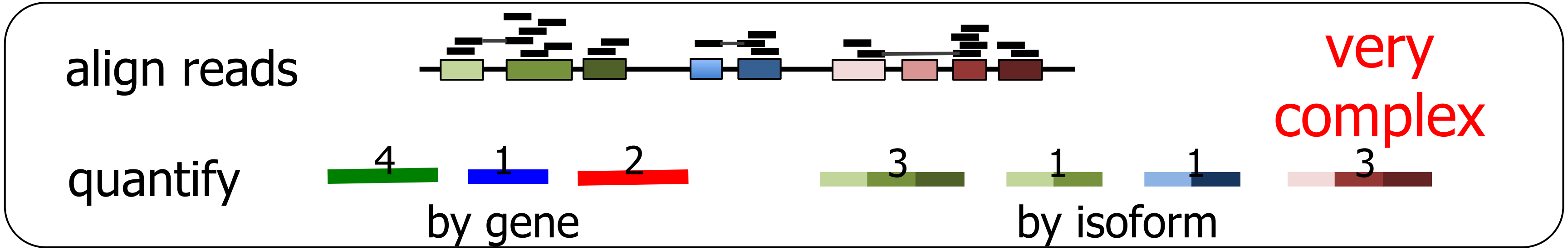
isolate RNA

create cDNA library

sequence

> 100,000,000 reads

Data Analysis



align reads

quantify

very complex

Many spliced aligners, but no clear winner

Numerous independent benchmarks of dozens of linear spliced aligners

(Engstrom 2013; Florea and Salzberg 2013; Hatem 2013; Liu 2014; Carrara 2015)

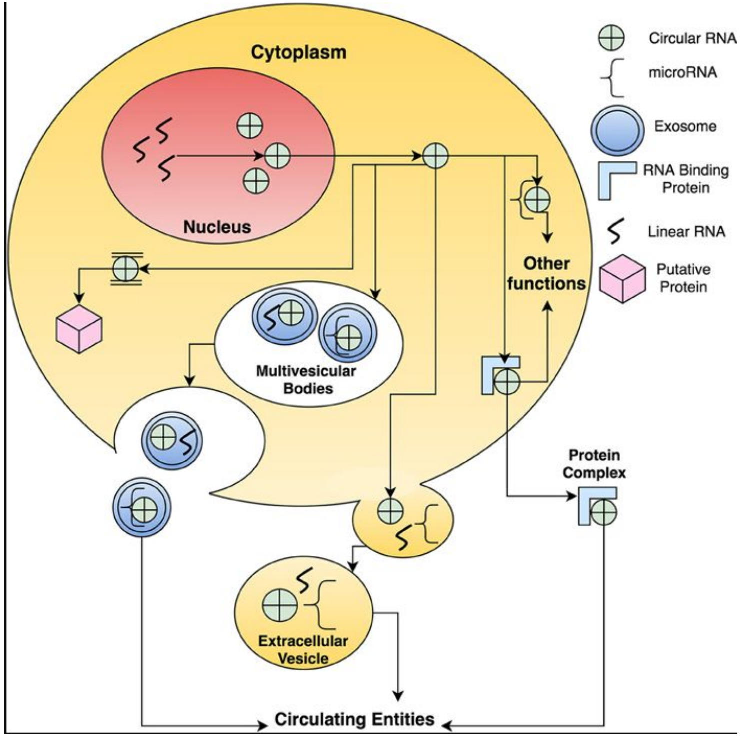
conclude that existing algorithms suffer from:

- high false positive rate
- low sensitivity for < 5 reads
- different biases against some isoforms

Patched linear splice detection algorithms to detect circular RNA:

- little overlap in genome-wide predictions of circRNA
- poor accuracy

RNA as a sensitive and specific biomarker

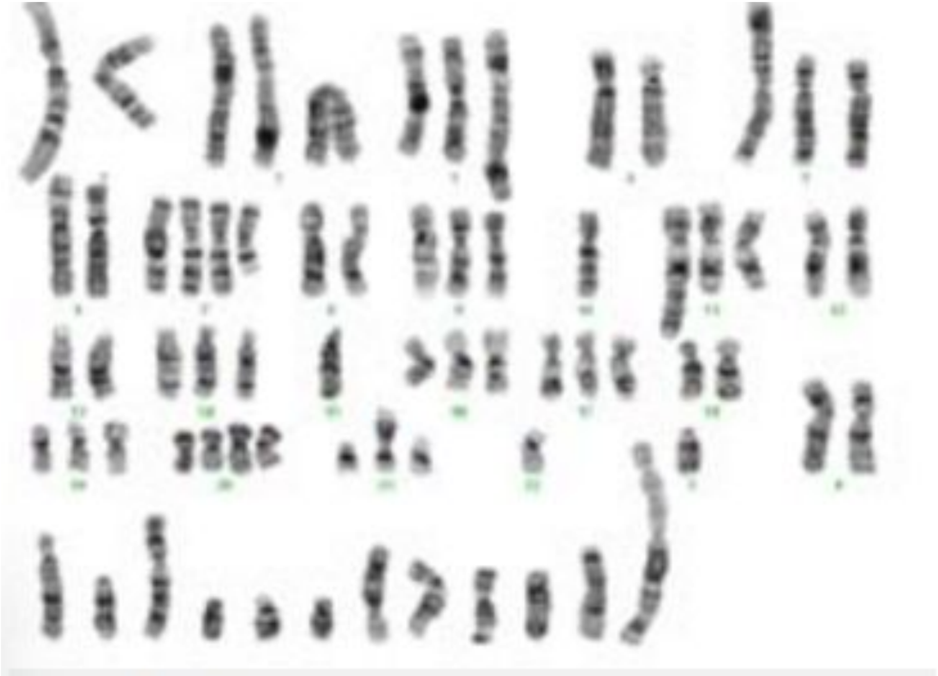


RNAs in diseased cells escape into the blood stream and are ‘digital’ markers

https://www.frontiersin.org/files/Articles/213360/fphys-07-00355-HTML/image_m/fphys-07-00355-g001.jpg

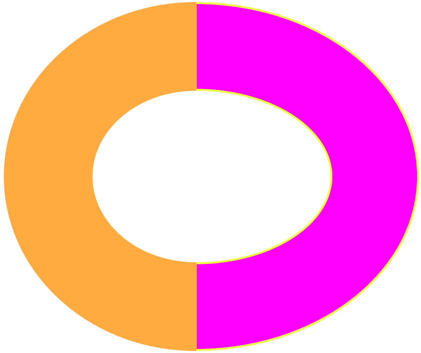
Gene fusion detection

The function of genome instability in cancer?

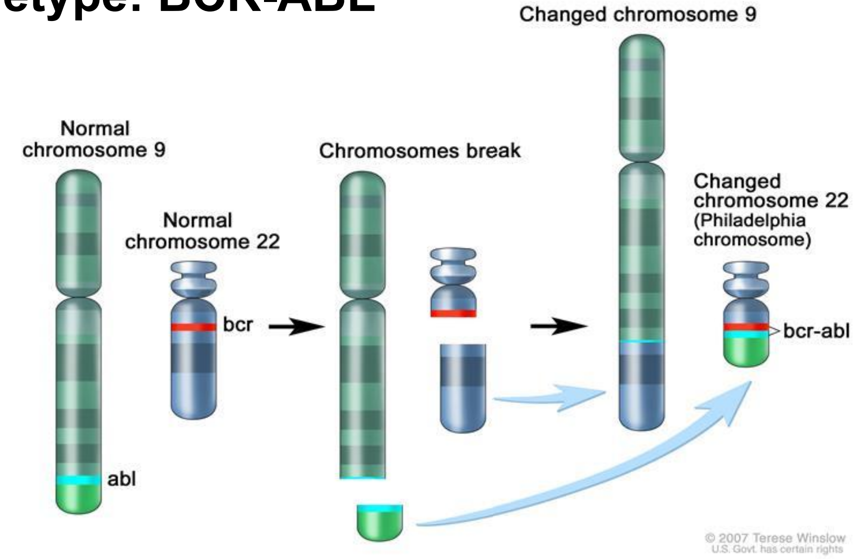


Gene fusions, cancer specific drivers

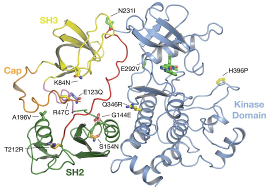
Fusion Circular RNA?
Circular RNA as biomarkers?



The archetype: BCR-ABL

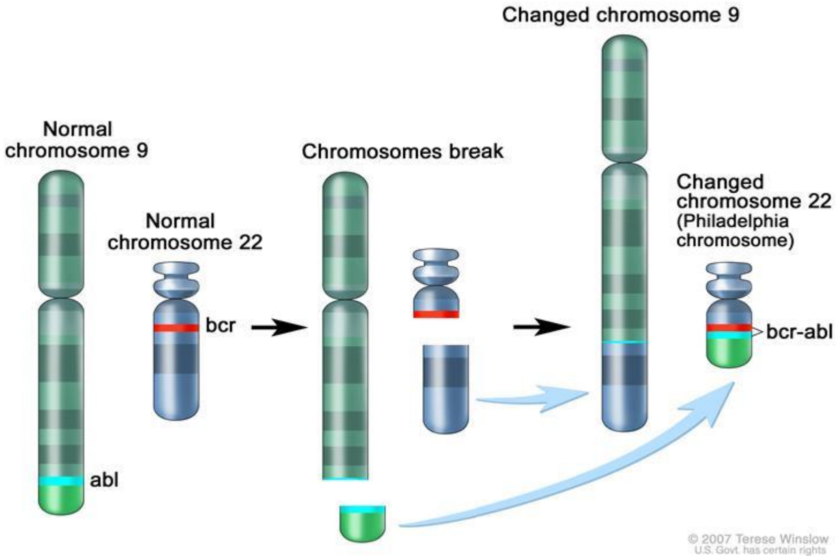


© 2007 Terese Winslow
U.S. Govt. has certain rights

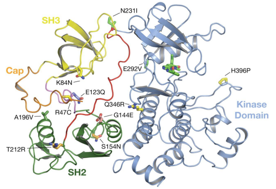


To find fusions: a step “back”

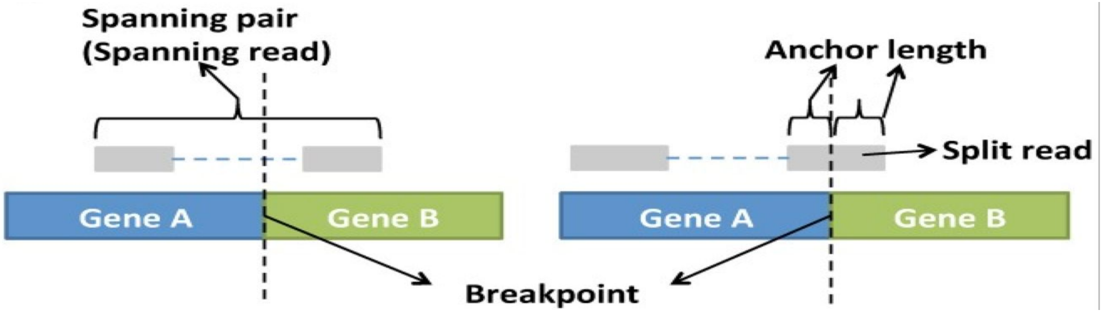
- State-of-the art for unbiased fusion detection -- needed improvements
  - Similar biases leading to missing circular RNA
  - **Unbiased fusion discovery, new cancer biology?**
  - **New druggable fusions (with existing or to-be-created drugs)?**
    - **Bioinformatics can already tell us the list of genes that might have a drug target**



© 2007 Terese Winslow
U.S. Govt. has certain rights



Poor accuracy of gene fusion algorithms

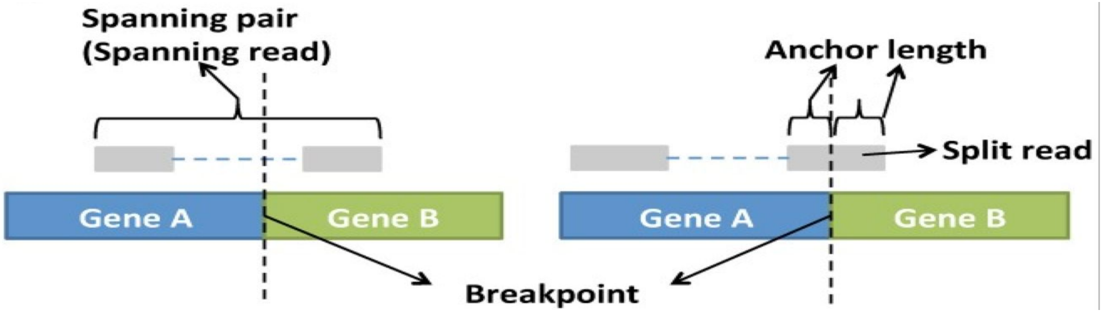


- 23+ “state of the art” algorithms, but none “trusted”
  - identify many gene fusions that are clear false positives
    - detect 100s of fusions in normal tissues
    - real fusions deep in list of false positives
  - small changes to parameters have big impact on results

Liu et al., 2016

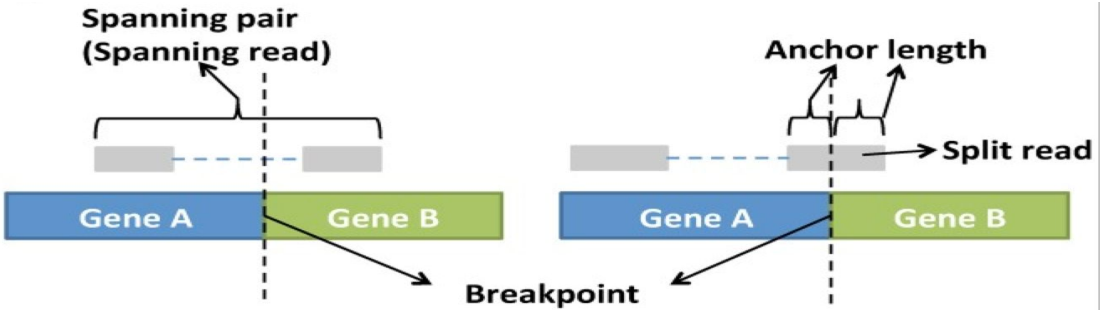
- human-guided filtering and targeted design of tumor sequencing

Why is gene fusion detection difficult?



- Mutations: surrounding sequence defines putative position
  - Number is N where N nucleotides of coding sequence
- Fusion: could include cryptic exon
  - Number is quadratic in number of exons in the human genome
- Challenge: assume exons are 100nt, and each gene has 10 exons. Compare m^2 to N as a function of g genes.

Consequence: basic and clinical biology



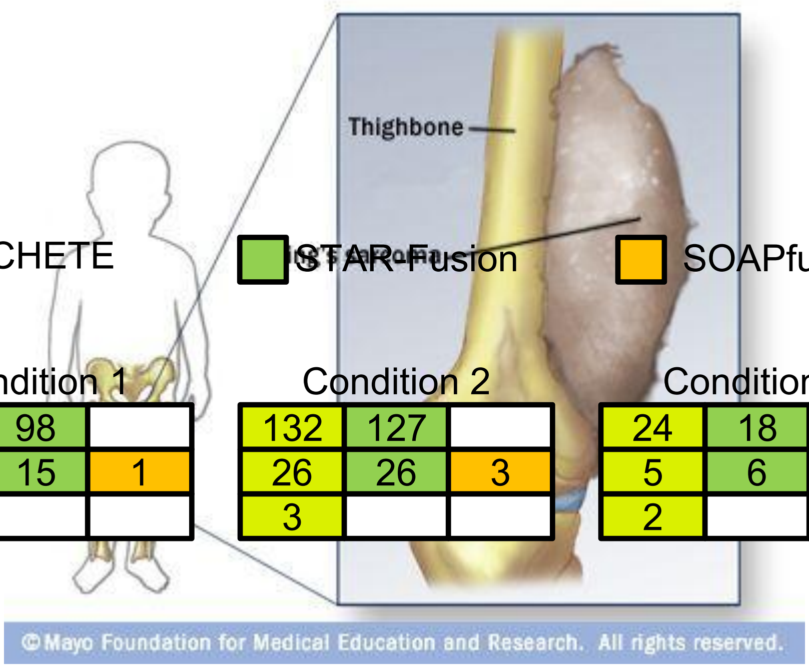
Because discovering gene fusions is so difficult, there are

1. Large numbers of FP
2. Large numbers of FN
3. No statistical estimate of the FP or FN

Clinicians focus on common gene fusions-- not discovered with RNA-seq.

Statistical challenges prevent personalized cancer genomics!

Some cancers are driven by gene fusions EWSR1-FLI1



© Mayo Foundation for Medical Education and Research. All rights reserved.

🟨 MACHETE 🟩 STAR-Fusion 🟧 SOAPfuse

	Condition 1			Condition 2			Condition 3			Condition 4		
Isoform 1	104	98		132	127		24	18	4	12	11	
Isoform 2	13	15	1	26	26	3	5	6		4	5	1
Isoform 3				3			2					

Clinical discovery in *Big Data*

The Cancer Genome Atlas

With the right algorithms, known fusions identifie

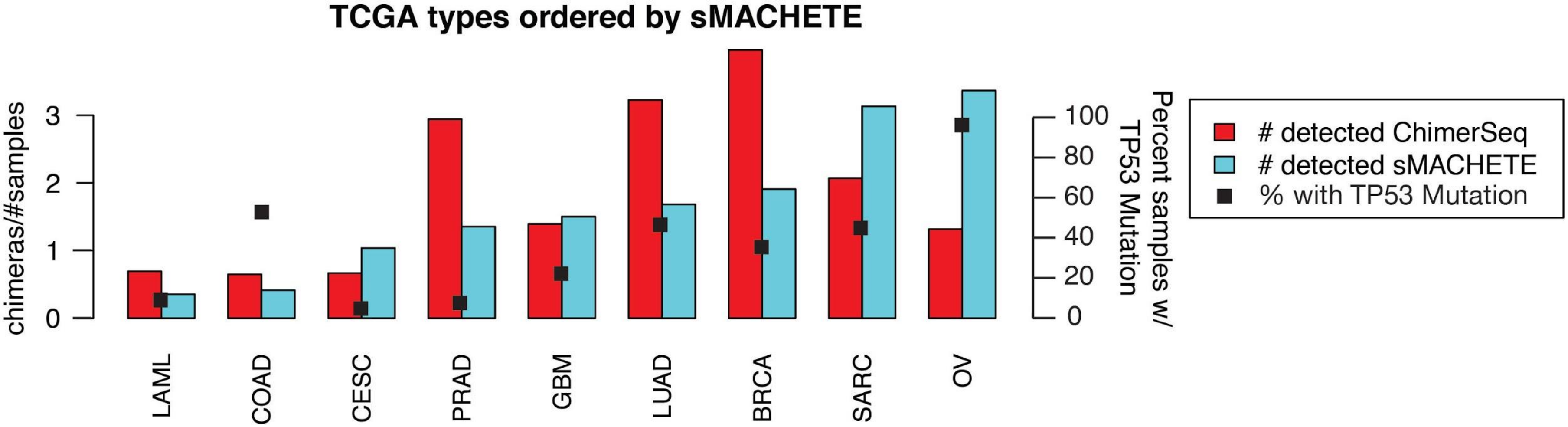
- Prostate cancer: TMPRSS2-ERG
- Acute Myeloid Leukemia: BCR-ABL, PML-RARA, and other fusions
- Glioblastoma: EGFR fusions

New predicted recurrent fusions

Rare/private potential drivers

Potentially druggable-- but statistics are needed to bring cancer genomics to your PC

A taste of big data cancer genomics:
Statistical analysis (rather than classical cancer biology):
fusions could drive some cancers



Freeman et al

Methodology covered in lectures

1. Know how to eyeball statistical significance w/o a computer (quiz)
    a. Review of z scores, hypothesis testing, duality with confidence intervals
2. How to pose and formulate statistical models
3. How to use computer-intensive methods for statistical inference
    a. Permutations testing
        i. What is it?
        ii. Does it always work?
        iii. Can it be done in closed form?
    b. Bootstrapping, early stopping -- the theory
        i. Suppose you decide to do resampling: how much do you need?
        ii. How much is necessary?
        iii. How much is sufficient (Martingales)?
4. How to think about deep learning from a statistical point of view
    a. Machine learning: intrinsic limitations
        i. Theoretical and empirical examples
    b. Moving forward to maximize discovery

Summary of applications

1. Biological problems
   a. Disease genomics
2. Approaching their solution with parametric models, significance testing
3. Using statistical models to discover biological mechanisms and dysregulation in disease

Example Project proposals

Design a statistical algorithm (or implement comparisons) with associated statistical confidence to precisely identify any of the following variants linked to or causing disease:

1. Gene fusions in cancer
2. Mutations in cancer statistically associated with splicing or gene fusions
3. Splicing events present only in diseased genomes (such as neurodegenerative diseases)
4. Novel promoters/epigenetic marks corresponding to transcription initiation sites
5. Pick a disease and discover biomarkers with RNA-seq, use statistical tests to quantify sensitivity and specificity