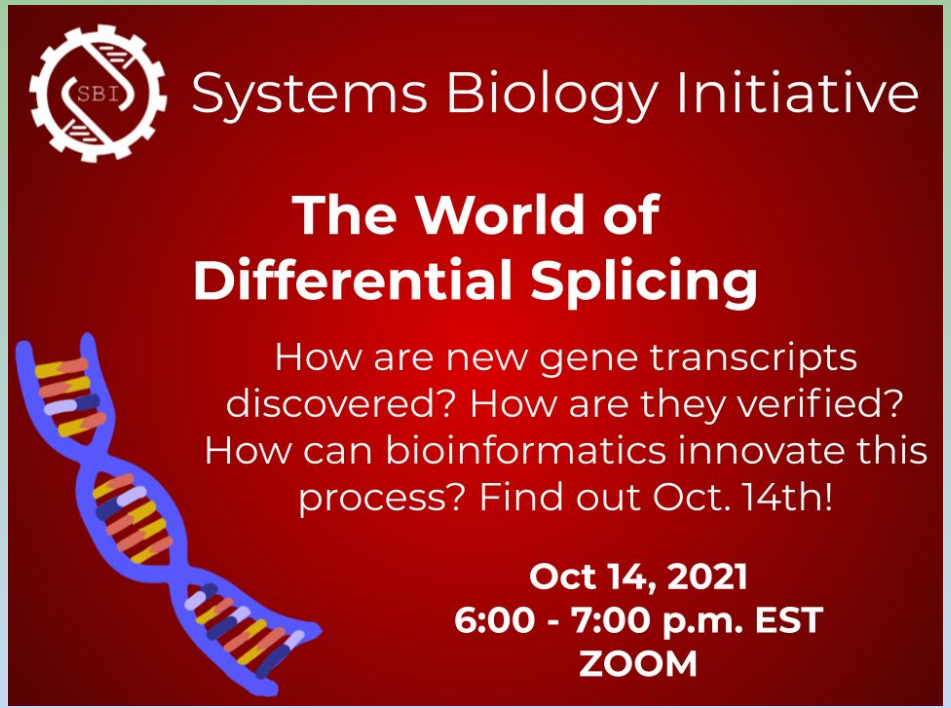
Systems Biology Initiative

The World of Differential Splicing

Presented by Theo Nelson



Introductions

Theo Nelson

tmn2126@columbia.edu

CC ‘24

Computer Science / PreMed

Dry and Wet Lab Research Experience

Systems Biology Initiative

sbi.columbia@gmail.com

Student Organization

Systems Biology - Computational Biology

https://sbicolumbia.wixsite.com/cusbi

Send a wave to a fellow listener!

Core Concepts

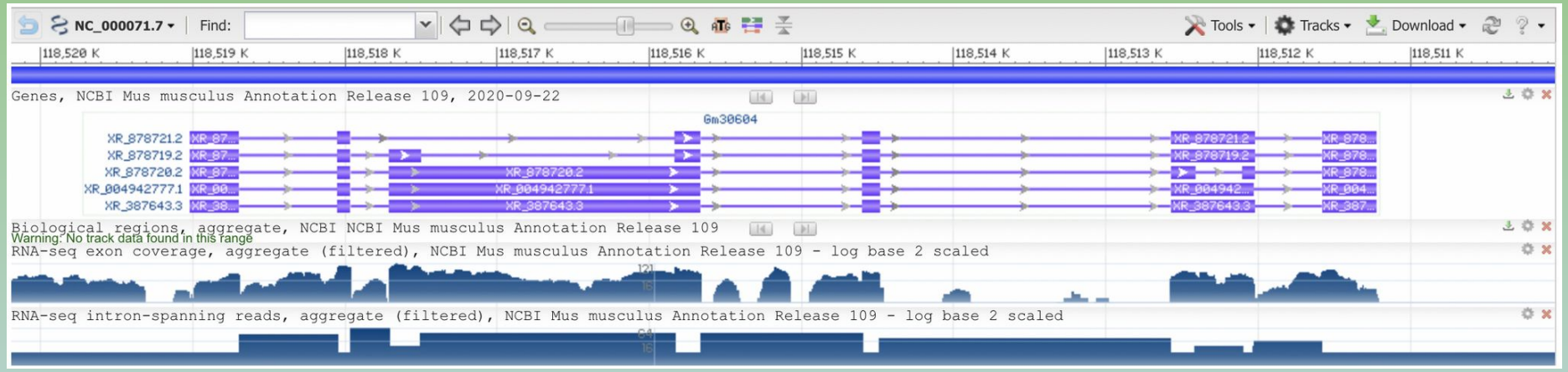
Variability

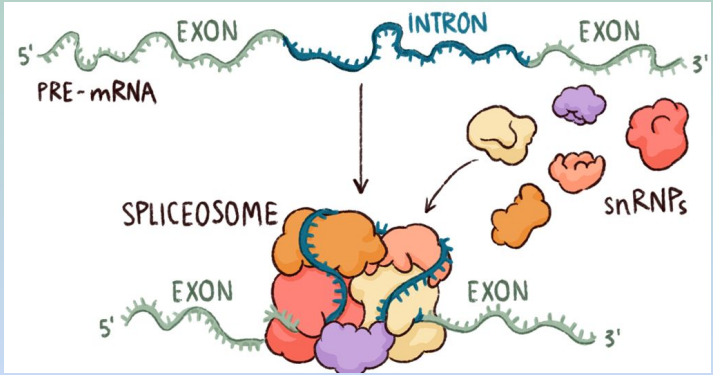
DNA

RNA

Protein

Core Concepts





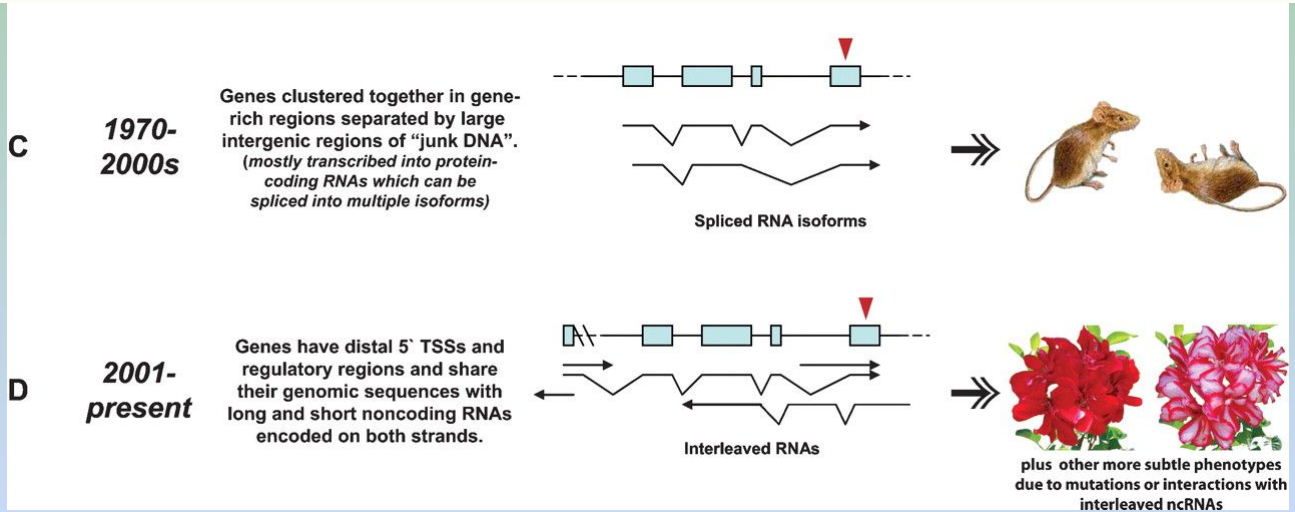
Genes ➜ Transcripts

The Evolution of “Genes”

Time Period	Key Facts	Gene Model	Phenotype

gene (n.)

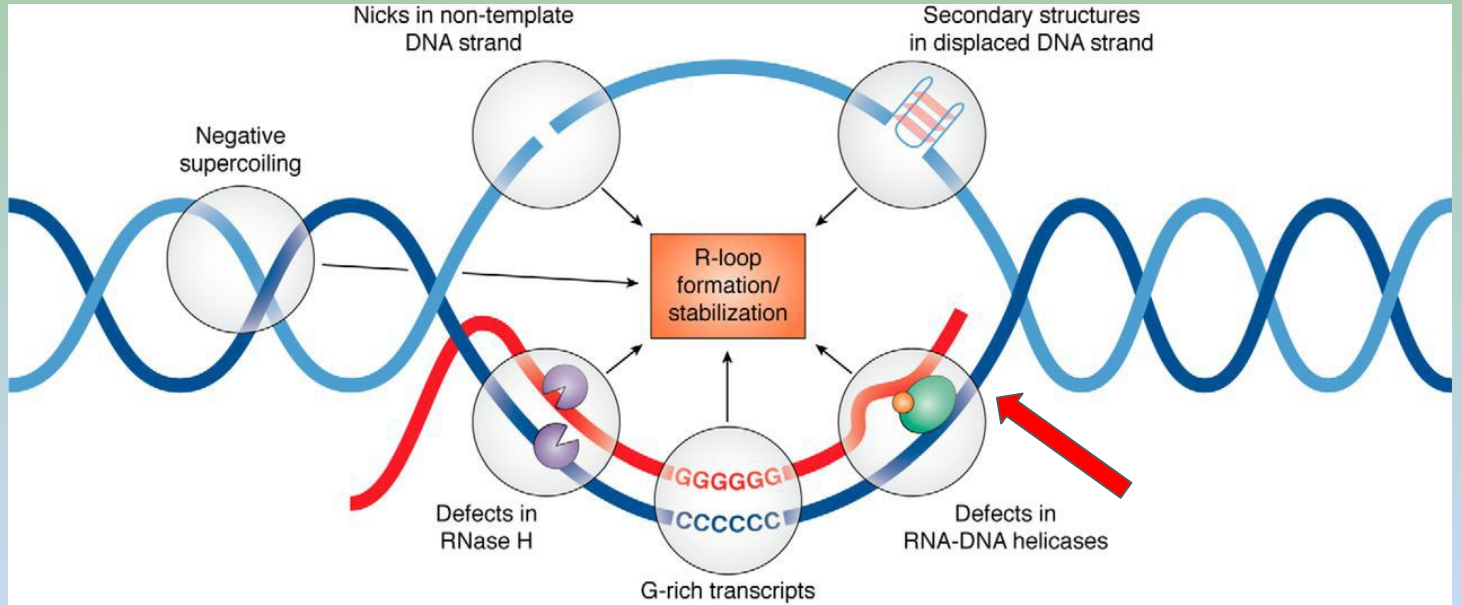
1911, from German *Gen*, coined 1905 by Danish scientist Wilhelm Ludvig Johannsen (1857-1927), from Greek *genea* "generation, race" (from PIE root ***gene-** "give birth, beget"). De Vries had earlier called them *pangenes*. ***Gene pool*** is attested from 1946.



10.1101/gr.6525007

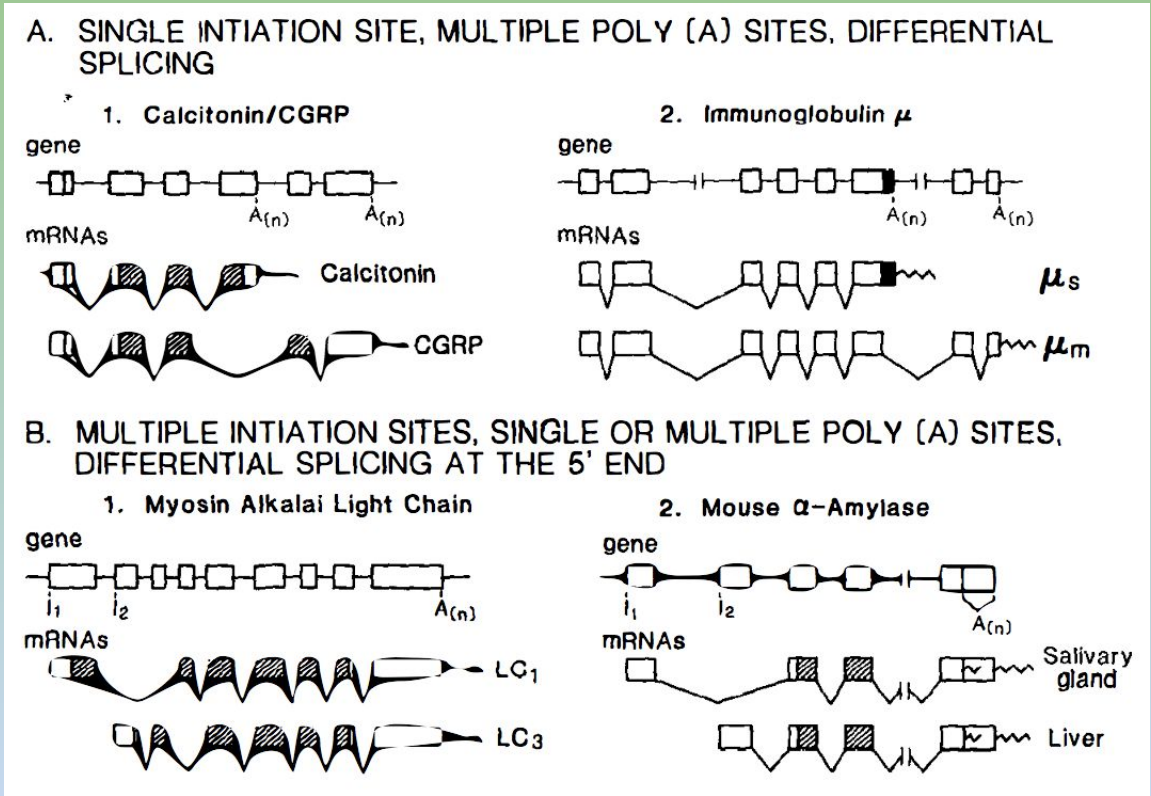
Alternative Splicing - 1977

An Amazing Sequence Arrangement at the 5′ Ends of Adenovirus 2 Messenger RNA



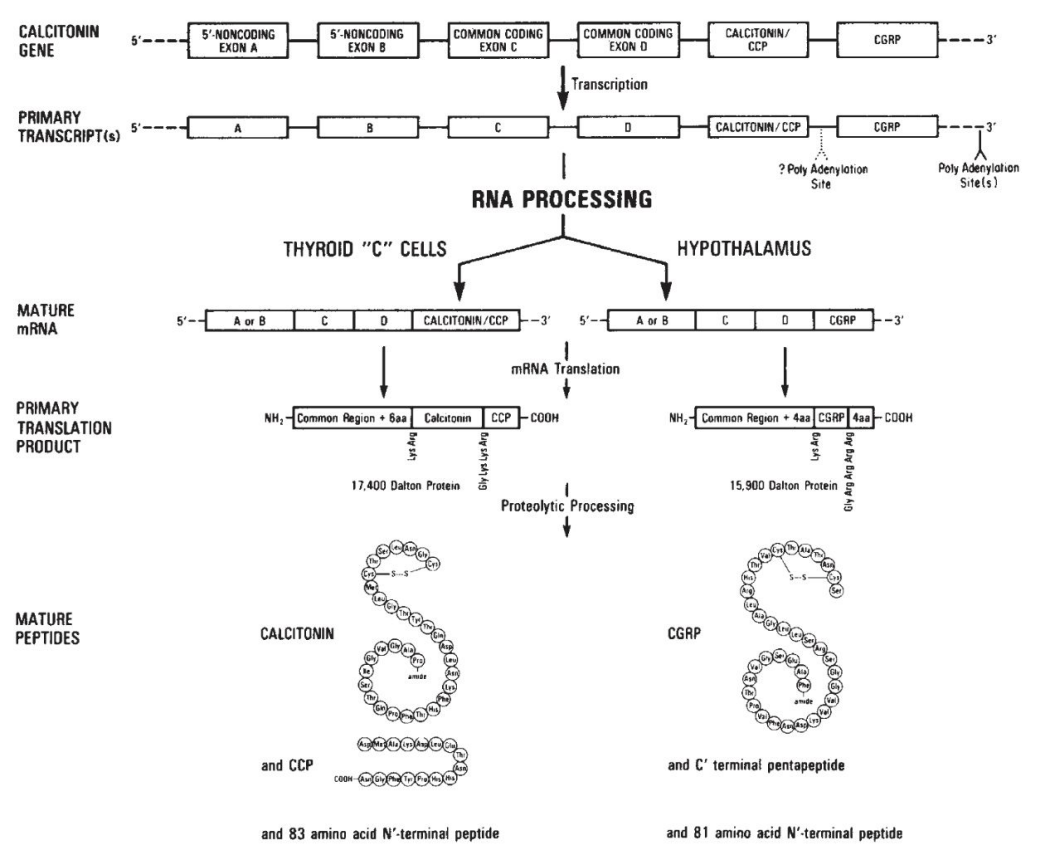
10.1016/0092-8674(77)90180-5

Alternative Splicing - 1981



10.1146/annurev.bi.55.070186.005303

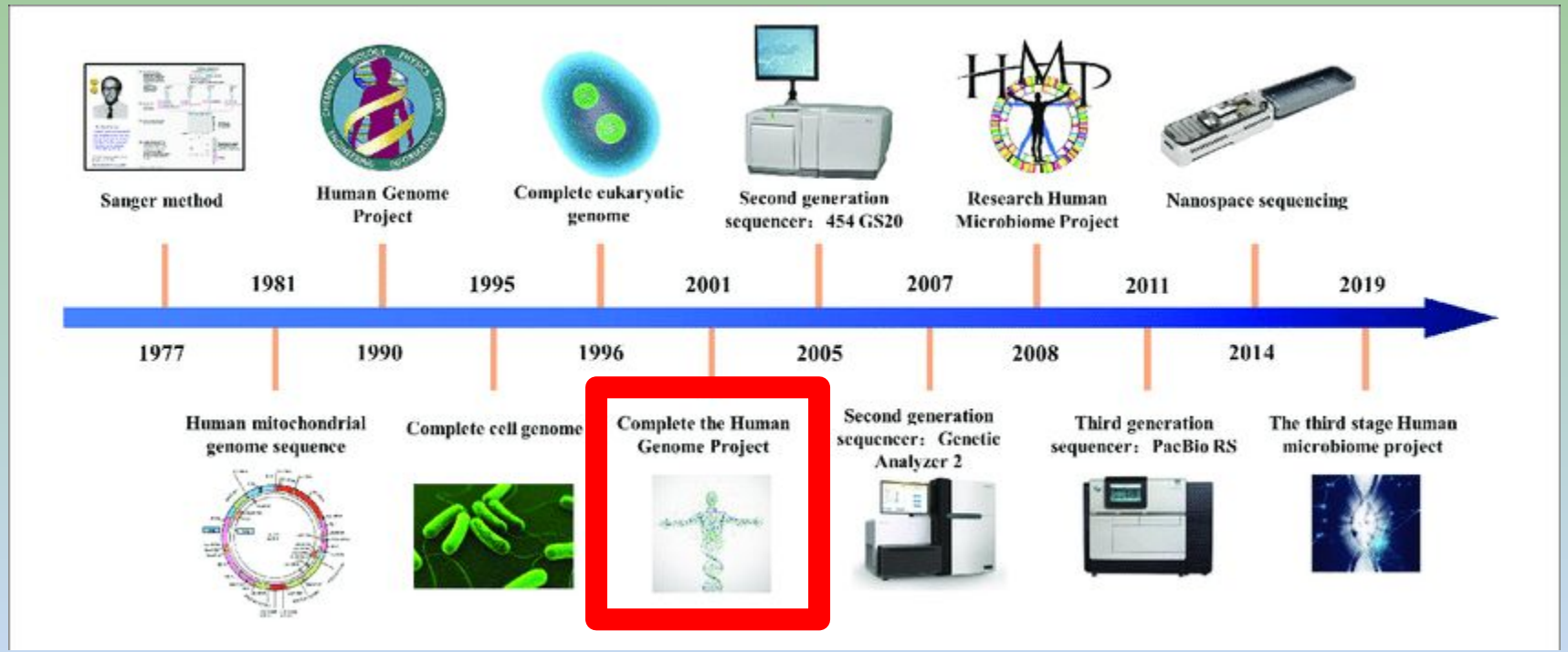
Alternative Splicing - 1981



10.1038/298240a0

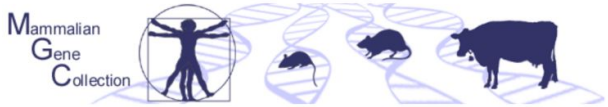
Sequencing

Changing the Game



10.3389/fbioe.2020.01032

Mammalian Gene Collection



MGC Home

Clone Info

- **Where to Buy**
- Vectors & Method Overviews

Sequencing Info

- MGC ESTs

MGC Info

- Project Summary
- Project Teams
- NIH Institutes
- References

Other Species Collections

- Danio (ZGC)
- Xenopus (XGC)

Useful Links

- UCSC
- NCBI Home
- NCBI MGC Retrieval
- Full-length cDNA Projects
- CGAP
- OCG

MAMMALIAN GENE COLLECTION

17-Dec-12	Human	Mouse	Rat	Bovine
Total MGC full ORF clones	29,818	27,285	6,763	9,104
Non-redundant genes	17,592	17,701	6,486	8,724

About the MGC

The goal of the Mammalian Gene Collection (MGC), a trans-NIH initiative, is to provide researchers with unrestricted access to sequence-validated full-length protein-coding (FL-CDS) cDNA clones for human, mouse, and rat genes. In 2005, the project added the cow cDNAs generated by Genome Canada.

MGC cDNA clones were obtained by screening of cDNA libraries, by transcript-specific RT-PCR cloning, and by DNA synthesis of cDNA inserts. (See References 1, 2, 3)

All MGC sequences are deposited in GenBank and the clones can be purchased from distributors of the IMAGE consortium. You can use "A Guide to Finding Mammalian Gene Collection (MGC) Clones and Evaluating Their Sequence" to assist in determining whether MGC cDNA clones for human, mouse, or rat genes and transcripts of interest are available for purchase or sequence investigation.

ORFeome Collaboration (OC) was formed to provide the research community with sequence-validated, full-ORF human cDNA clones in the Gateway® vector format. The Project Summary provides background information and additional details about the MGC and the ORFeome Collaboration.

With the conclusion of the MGC project in March 2009, the GenBank records of MGC sequences will be frozen, without further updates. (See Reference 4) Since the definition of what constitutes a full-length coding region for some of the genes and transcripts for which we have MGC clones will likely change in the future, users planning to order MGC clones will need to monitor for these changes. Users can make use of genome browsers and gene-specific databases, such as the UCSC Genome browser, NCBIs Map Viewer, and Entrez Gene, to view the relevant regions of the genome (browsers) or gene-related information (Entrez Gene).

Note: Please check the GenBank record of each MGC full-length clone for detailed sequence annotation. Some MGC sequences have nucleotide differences that are not supported by other experimental data.

Search for Full-length MGC Clones by Gene Symbol or Keyword

Select **Human** ◉ or **Mouse** ○ or **Rat** ○ or **Bovine** ○			
Enter Gene Symbol		Search	Help
Enter Gene Keyword		Search	Help

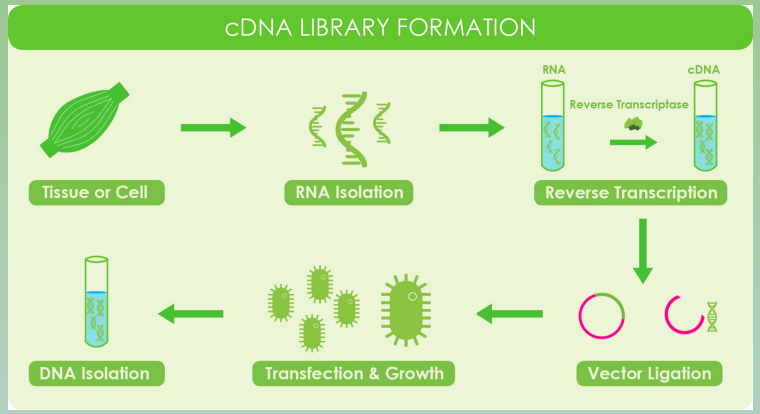
Nucleotide BLAST against MGC Clones

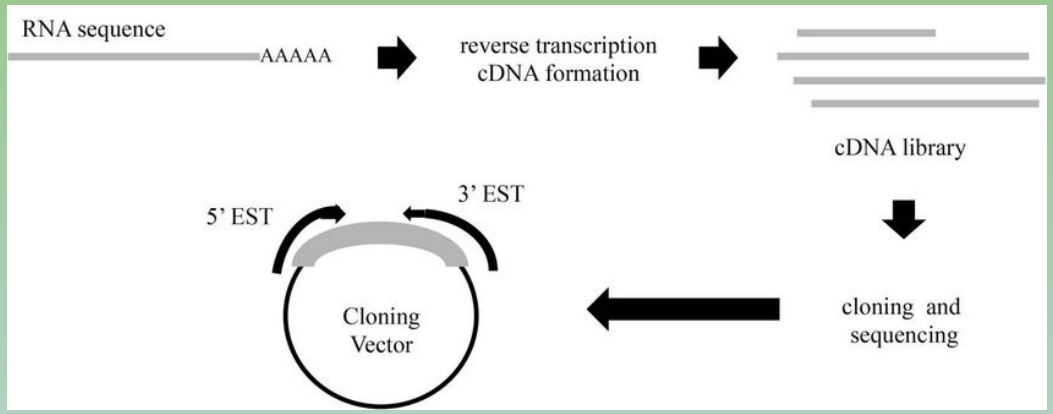
Nucleotide BLAST a sequence against MGC **full-length** sequences.

MGC Full-length Clone Information

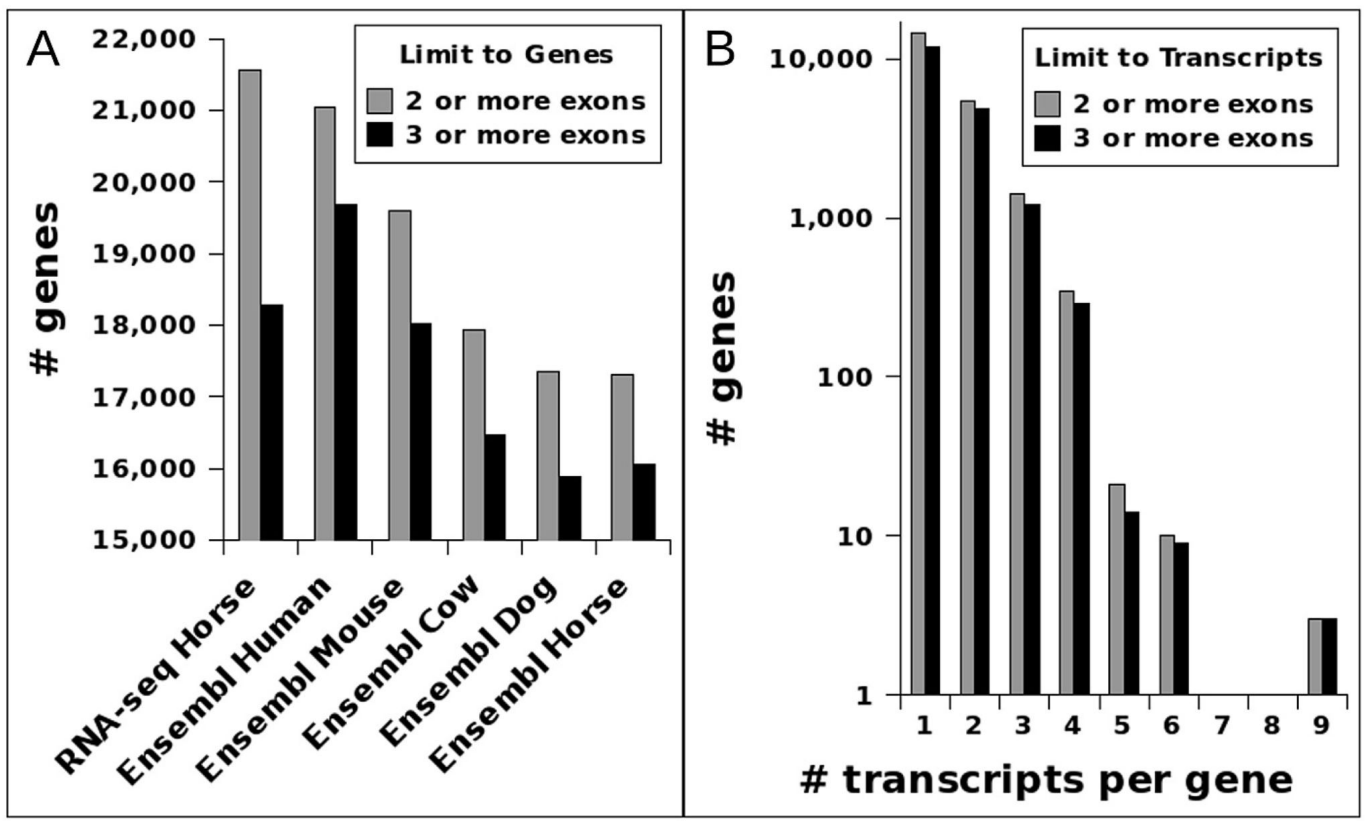
Select **Human** ◉ or **Mouse** ○ or **Rat** ○ or **Bovine** ○

Clone Collection



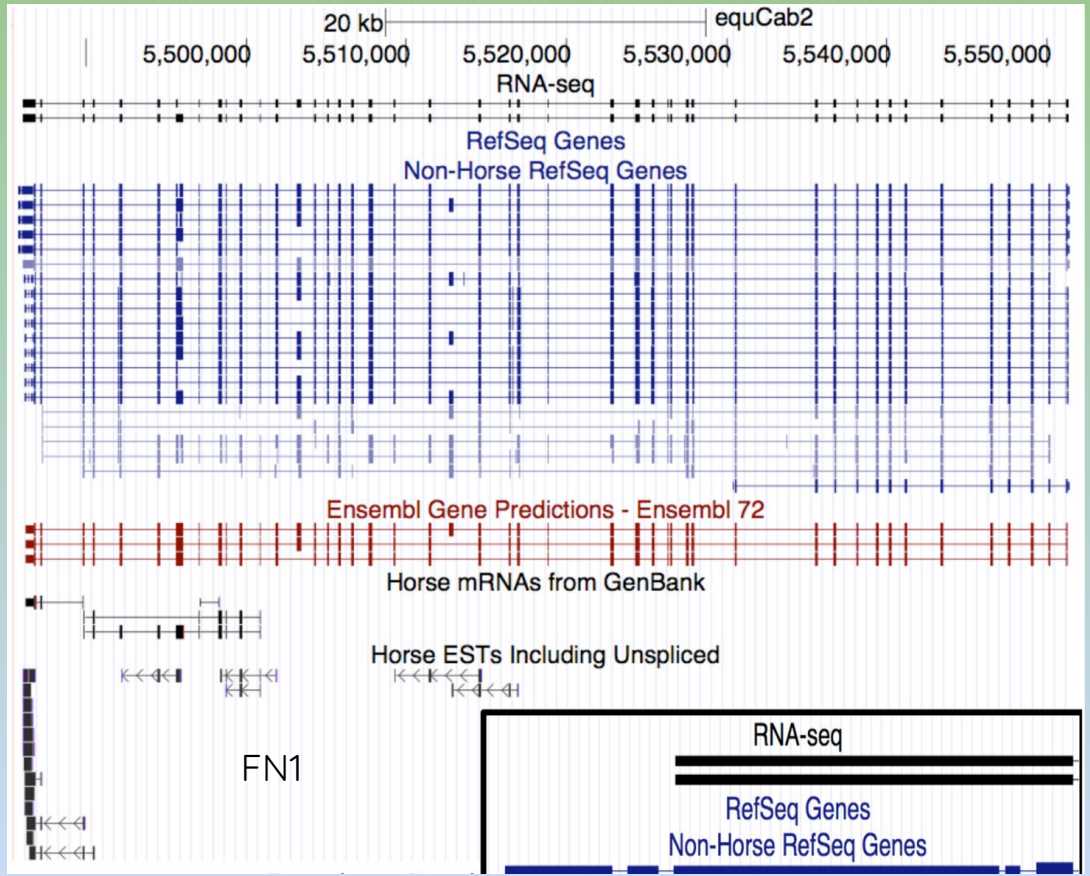


RNA Sequencing



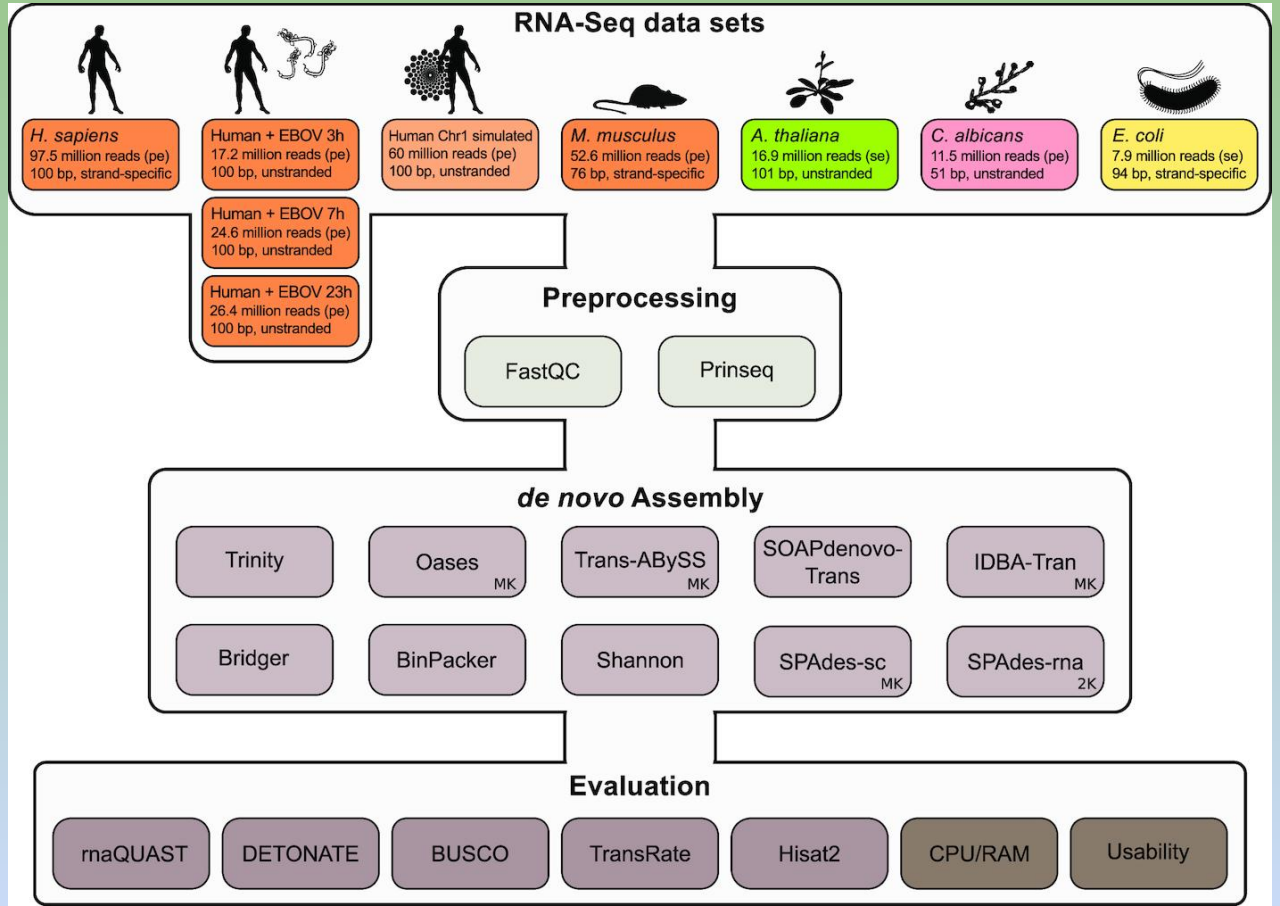
10.1371/journal.pone.0124375.g002

RNA Sequencing



10.1371/journal.pone.0124375.g002

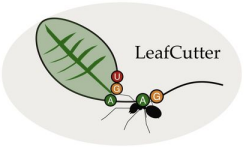
De Novo Assembly



10.1093/gigascience/giz039

LeafCutter

LeafCutter: Annotation-free quantification of RNA splicing



Yang I. Li[1], David A. Knowles[1], Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, Jonathan K. Pritchard

[1]*Equal contribution*

Leafcutter quantifies RNA splicing variation using short-read RNA-seq data. The core idea is to leverage spliced reads (reads that span an intron) to quantify (differential) intron usage across samples. The advantages of this approach include

- easy detection of novel introns
- modeling of more complex splicing events than exonic PSI
- avoiding the challenge of isoform abundance estimation
- simple, computationally efficient algorithms scaling to 100s or even 1000s of samples

For details please see our bioRxiv preprint and corresponding Nature Genetics publication.

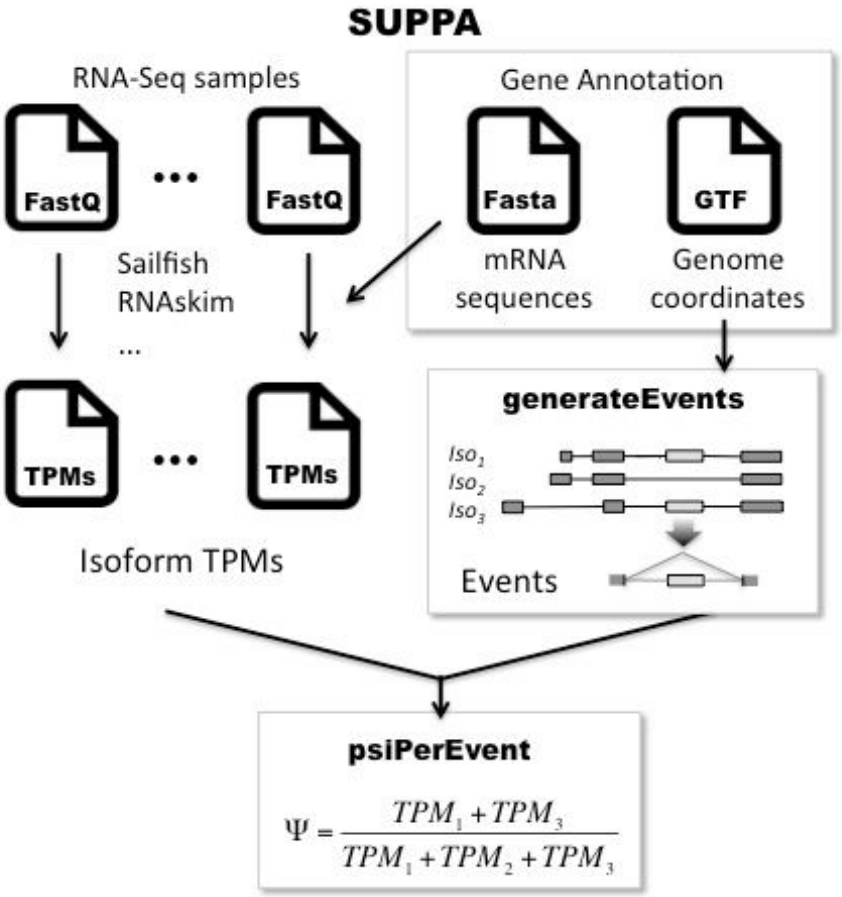
Additionally, for full details on the leafcutter for Mendelian Diseases (leafcutterMD) method that performs outlier splicing detection, see our Bioinformatics publication.

- Installation
- Differential splicing
- Outlier splicing
- Visualization
- SplicingQTL

Check out a demo leafcutter shiny app here: 10 brain vs. 10 heart samples from GTEx.

We have a Google group for user questions at https://groups.google.com/forum/#!forum/leafcutter-users

SUPPA2


SUPPA
RNA-Seq samples
FastQ
FastQ
Sailfish
RNAskim
...
TPMs
TPMs
Isoform TPMs
Gene Annotation
Fasta
GTF
mRNA sequences
Genome coordinates
generateEvents
Iso1
Iso2
Iso3
Events
psiPerEvent
$\Psi = \frac{TPM_1 + TPM_3}{TPM_1 + TPM_2 + TPM_3}$

Shark

Shark: fishing relevant reads in an RNA-Seq sample

Luca Denti, Yuri Pirola ✉, Marco Previtali, Tamara Ceccato, Gianluca Della Vedova, Raffaella Rizzi, Paola Bonizzoni Author Notes

Bioinformatics, Volume 37, Issue 4, 15 February 2021, Pages 464–472, https://doi.org/10.1093/bioinformatics/btaa779

Published: 14 September 2020 **Article history** ▾

Abstract

Motivation

Recent advances in high-throughput RNA-Seq technologies allow to produce massive datasets. When a study focuses only on a handful of genes, most reads are not relevant and degrade the performance of the tools used to analyze the data. Removing irrelevant reads from the input dataset leads to improved efficiency without compromising the results of the study.

Bioinformatics Approaches

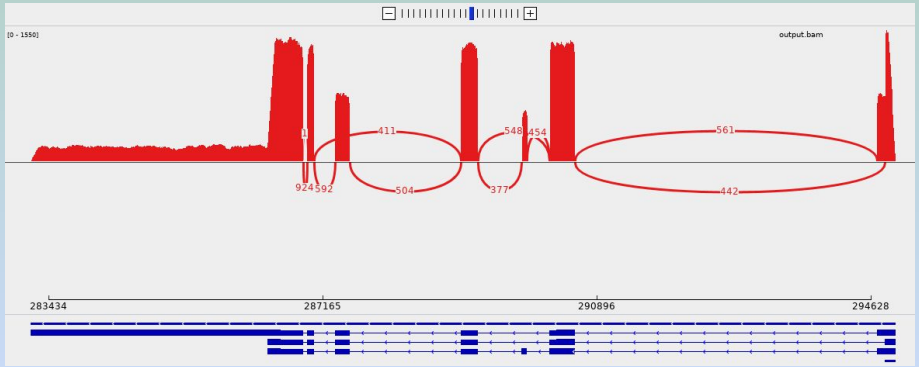
ASGAL https://asgal.algolab.eu/documentation

ASGAL (**A**lternative **S**plicing **G**raph **AL**igner) is a tool for detecting the alternative splicing events expressed in a RNA-Seq sample with respect to a gene annotation. The **main idea** behind *ASGAL* is the following one: the alternative splicing events can be detected by aligning the RNA-Seq reads against the splicing graph of the gene.

The instructions to install and use *ASGAL* are at http://asgal.algolab.eu.

- SUPPA2
- LeafCutter
- IsoformSwitchAnalyzeR
- DiffSplice
- Shark
- EventPointer
- ASGAL
- FRASER

Novel splicing event – Fruit Fly Example



Example ASGAL Run

Series GSE99479	Query DataSets for GSE99479
Status	Public on Oct 16, 2017
Title	A class of GATA3 mutation reprograms the breast cancer transcriptional network through gain and loss of function
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing Genome binding/occupancy profiling by high throughput sequencing
Summary	A pioneer transcription factor, GATA3, is one of the most frequently mutated genes in breast cancer, yet the impact of these mutations is largely unknown. We generated a GATA3 mutant cell line (T47D wt/R330fs) by CRISPR. Mutation of one allele of GATA3 led to loss of binding and decreased expression at a subset of genes, including Progesterone Receptor. At other loci, associated with epithelial to mesenchymal transition, gain of binding at a novel sequence motif correlated with increased gene expression. Our results illuminate tumor-promoting functions of specific GATA3 mutations in breast cancer.
Overall design	Genome-wide mapping of chromatin localization of luminal transcription factors in GATA3 mutant cells
Contributor(s)	Takaku M, Grimm SA, Paul WA
Citation(s)	Takaku M, Grimm SA, De Kumar B, Bennett BD et al. Cancer-specific mutation of GATA3 disrupts the transcriptional regulatory network governed by Estrogen Receptor alpha, FOXA1 and GATA3. *Nucleic Acids Res* 2020 May 21;48(9):4756-4768. PMID: 32232341

T47D wt/R330fsl tumors
10 weeks
SRR5929949

Implementation
+ Processes

Results