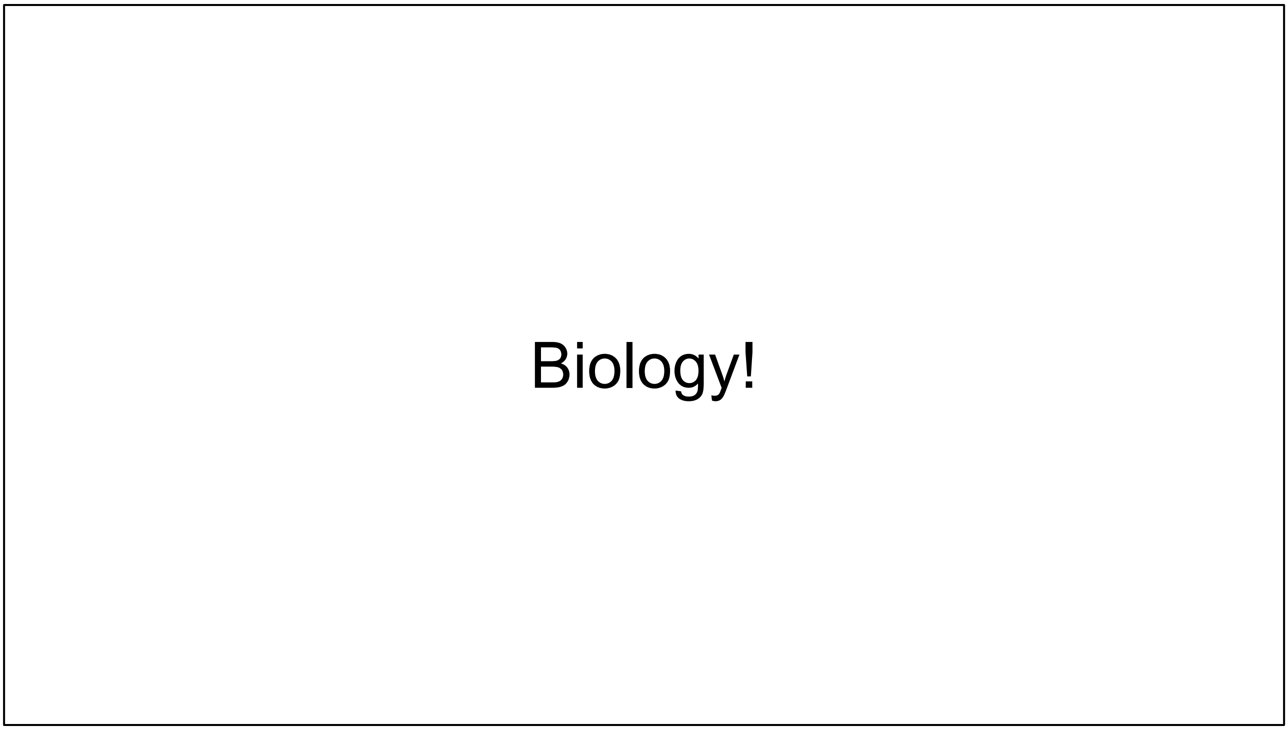
Vials - VIsualizing ALternative splicing of genes

By: Louie Dinh

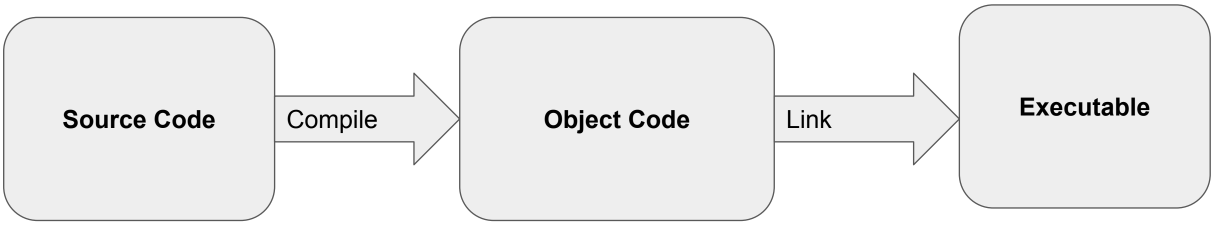


Biology!

I’m going to have to explain a bit about the cell works before the paper makes any sense.

Central Dogma [For Computer Scientists]

- REVIEW: Compiling Source Code



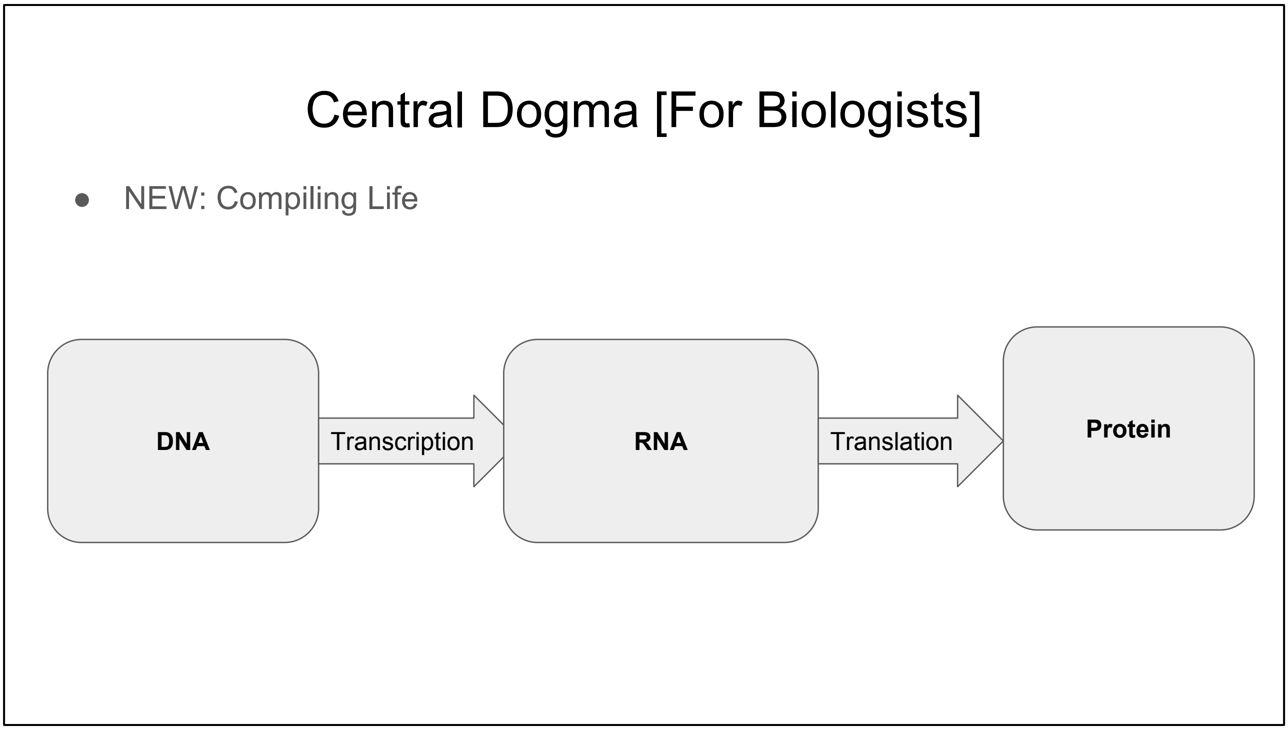
Source Code

Compile

Object Code

Link

Executable



Central Dogma [For Biologists]

- NEW: Compiling Life

DNA

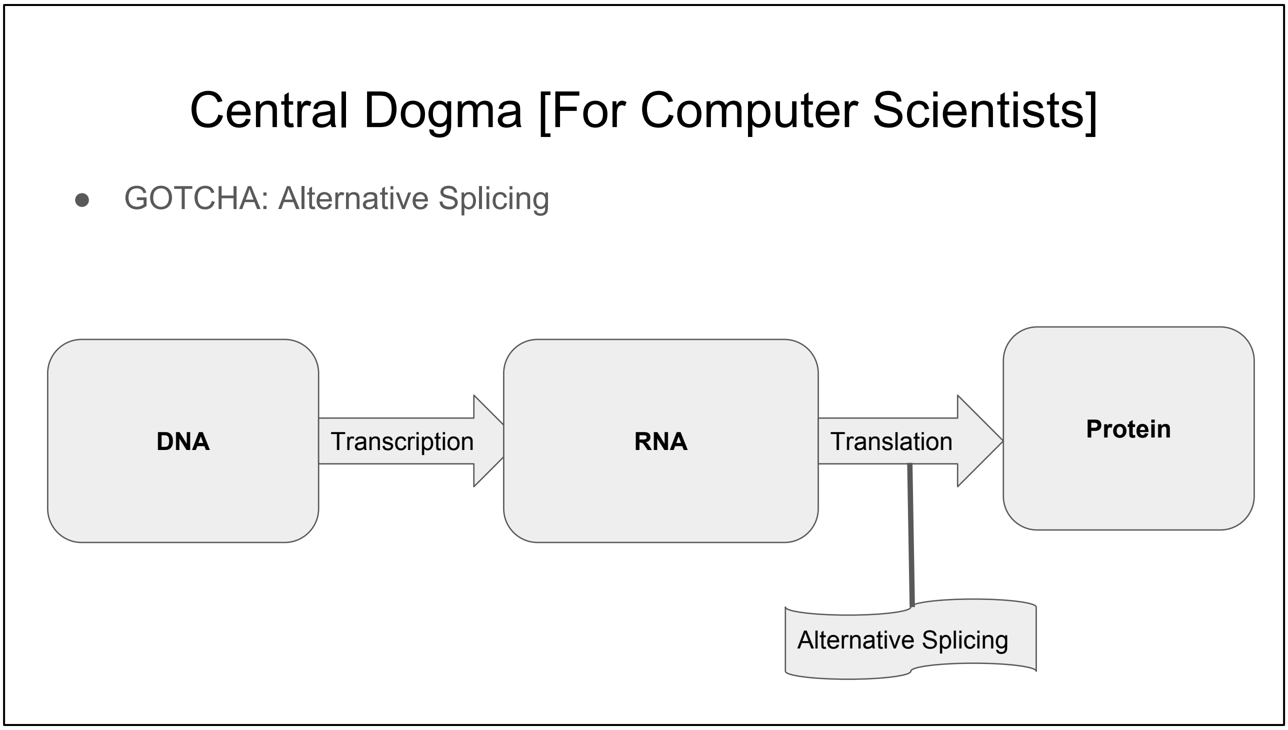
Transcription

RNA

Translation

Protein

DNA is like source code. Instructions for how to create an organism. ACTG. Highly Stable. Passed between generations.
- Cell is like a computer. Source code needs an architecture to execute on! Many different types in your body. Similar to how your source code must run in multiple environments. Multiplatform is hard!
- Protein is like an executable. It’s the thing that does stuff. It modules the chemical interactions, ties your ligaments together, digests the food you eat.



Central Dogma [For Computer Scientists]

- GOTCHA: Alternative Splicing

DNA

Transcription

RNA

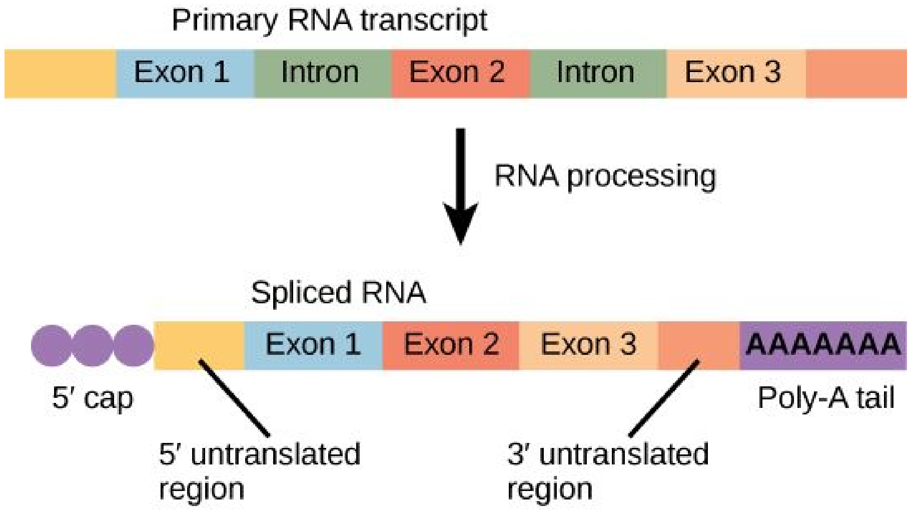
Translation

Protein

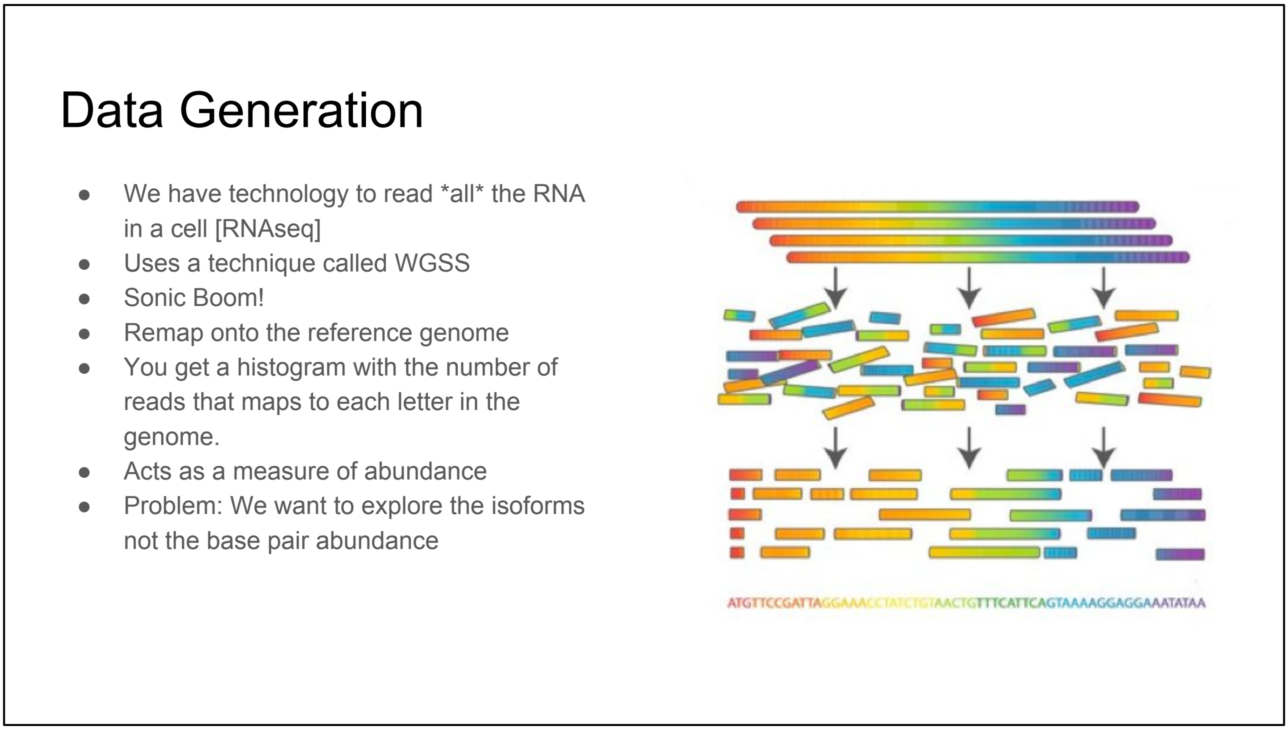
Alternative Splicing

- Actually I left something out
- Just like how you can have compiler flags to modify your programs during linking
- This is called alternative splicing. You have the same source code but you can optionally compile in different bits so that you can adapt it to different architectures and environments.

Architecture Of A Gene



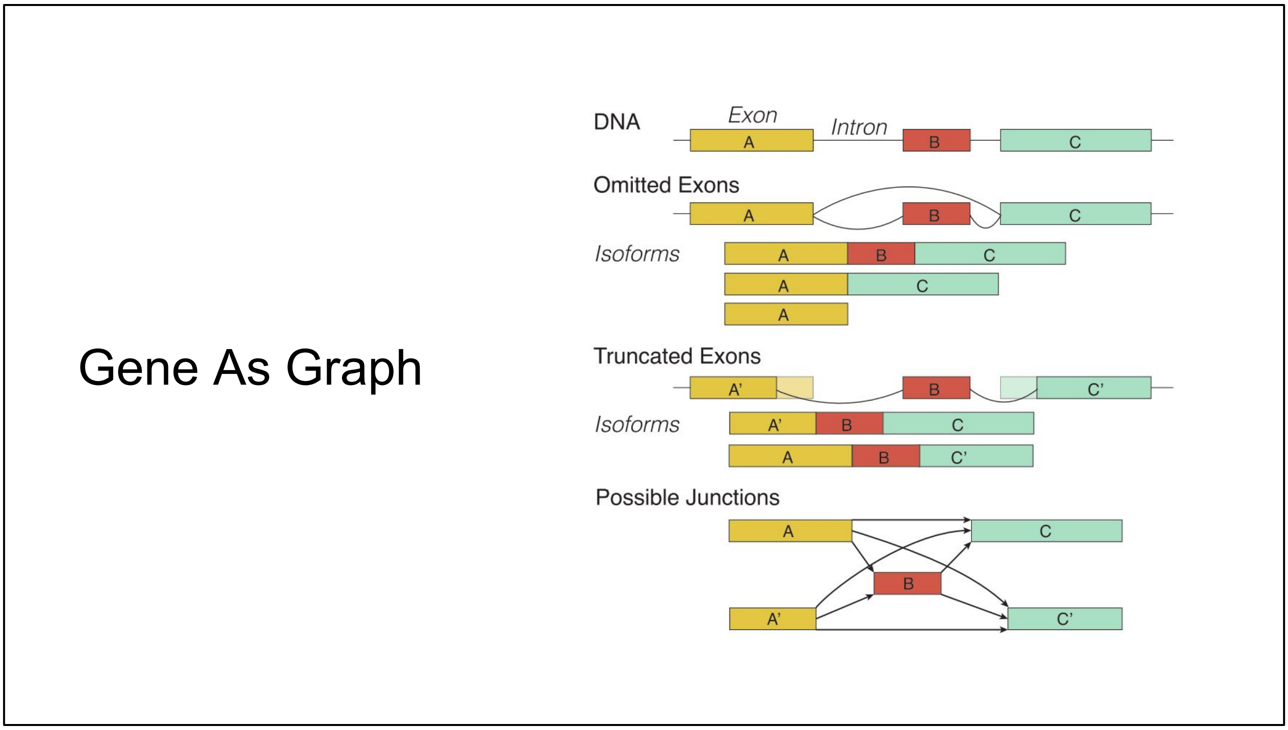
- Just one more thing.
- In a gene there are introns and exons.
- Introns are like comments. They don’t get compiled into the final transcript
- They get processed out in the cell.
- Also the start and end sites of these exons are wobbly. Early truncation happens.



Data Generation

- We have technology to read *all* the RNA in a cell [RNAseq]
- Uses a technique called WGSS
- Sonic Boom!
- Remap onto the reference genome
- You get a histogram with the number of reads that maps to each letter in the genome.
- Acts as a measure of abundance
- Problem: We want to explore the isoforms not the base pair abundance

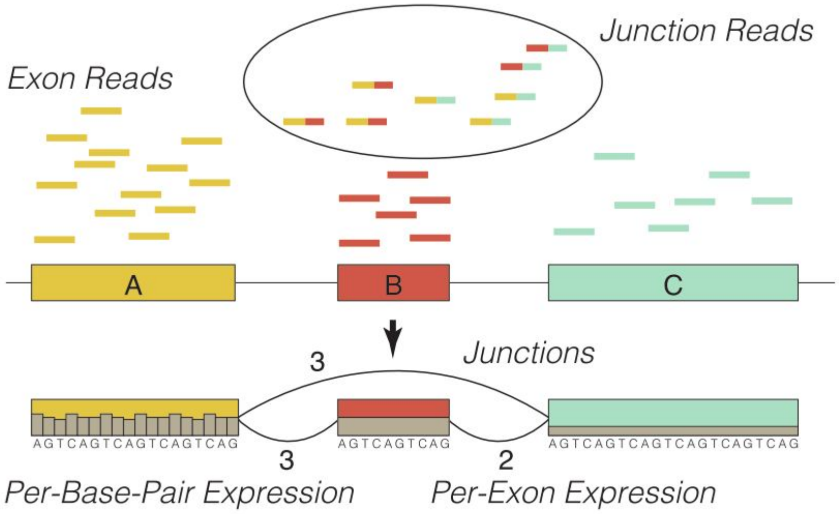
- We have technology to read *all* the RNA in a cell [RNAseq]
- Uses a technique called WGSS
- Sonic Boom!
- Remap onto the reference genome
- You get a histogram with the number of reads that maps to each letter in the genome.
- Acts as a measure of abundance
- Problem: We want to explore the isoforms not the base pair abundance



Gene As Graph

- A gene can really be thought of as a graph
- Nodes are the exon variants.

Junction Reads as Edges



- Junction reads are reads that span between two exons.
- Comes from a particular isoform

What Is The Data?

Tabular and Graph

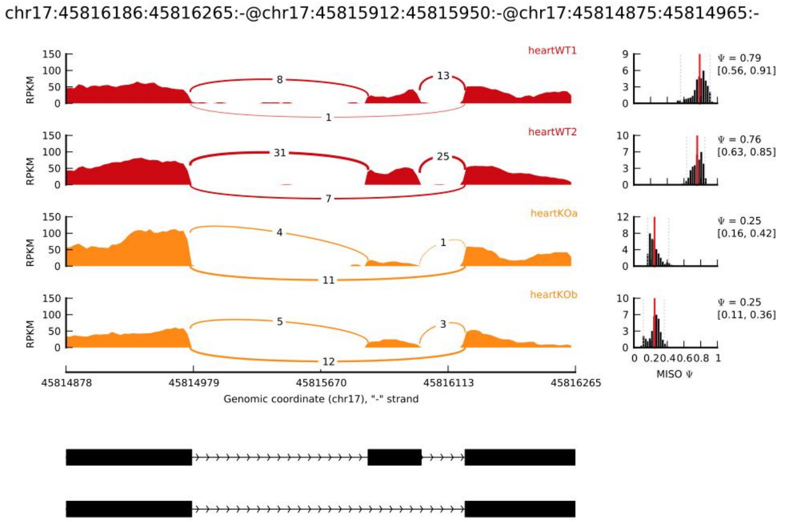
- Tabular data - base pair abundance. How many reads cover every single letter?
    - Key = (Sample ID, Genome Location), Value = Count
- Tabular data [derived] - Isoform Abundance
    - Key = (Sample ID, Genome Location, Exon Inclusion/Exclusin Mask) Value= Count
- Multivariate DAG - Junction Support
    - Nodes = Exon, Edges = Junction Reads. Isoform = Path through graph.

Why: Tasks

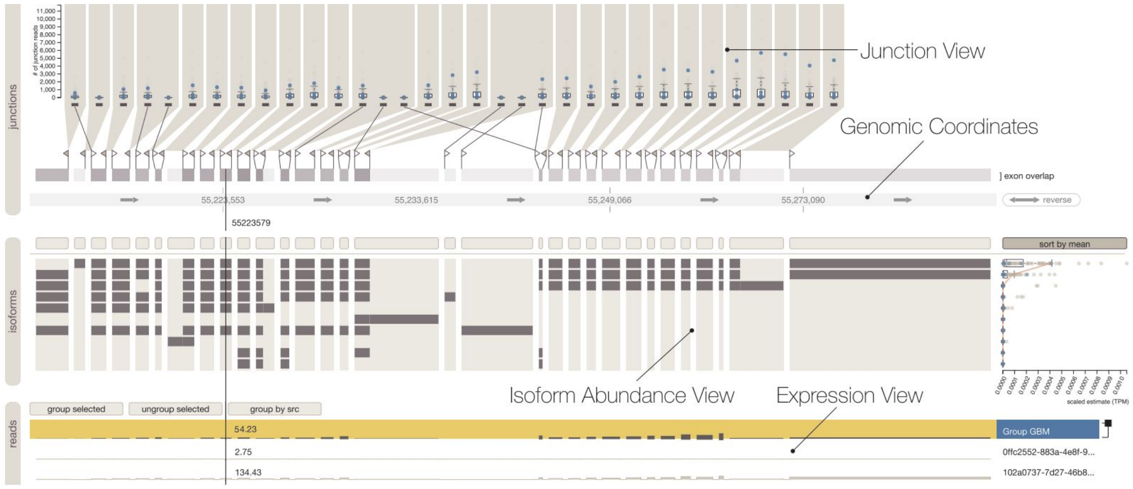
3 Main Tasks

1. Compare isoforms between samples [e.g one particular person] and between groups [e.g Glioblastoma versus Lymphoma Patients]
2. Discover new isoforms.
3. Control data quality

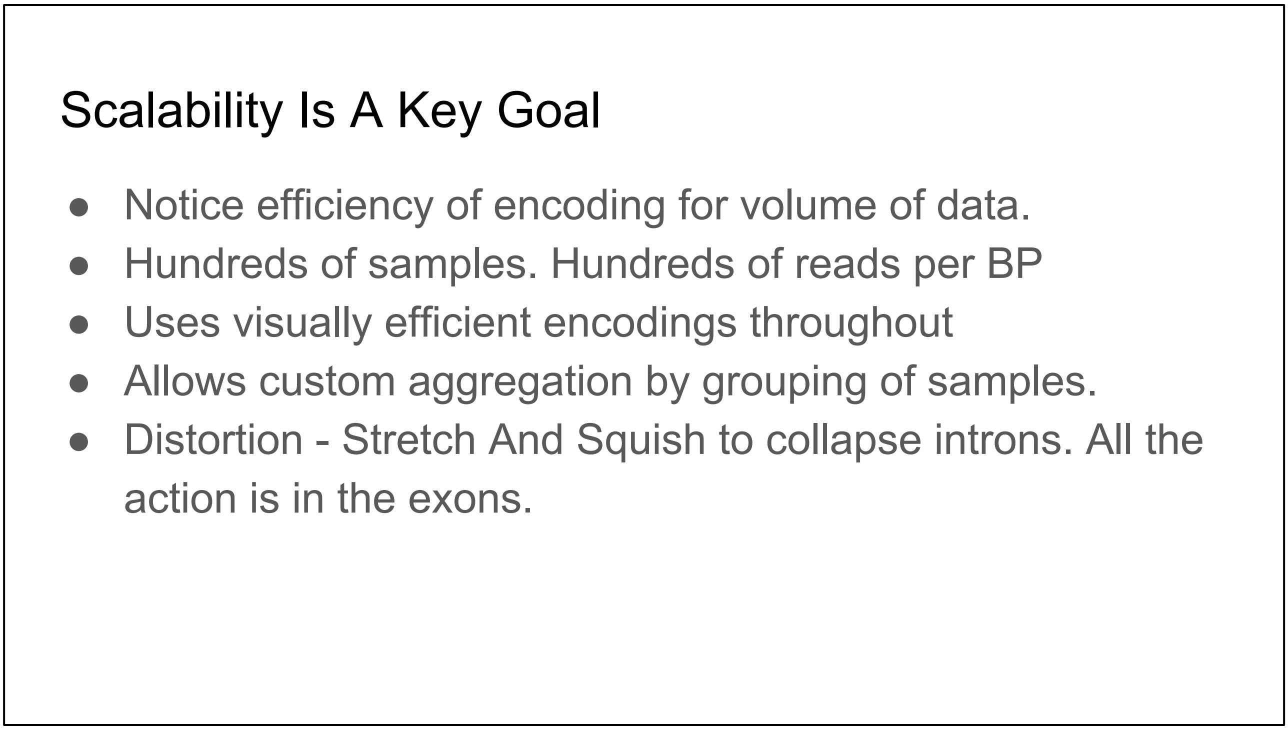
Designed For Scalability. Fixes Sashimi Graphs.



Overview



- 3 Views (Junction, Isoform Abundance, Expression Abundance)
- Abundance =raw. Junction support = raw, Isoform abundance = derived
- Heavy use of linked highlighting. Selection in any one view will affect all other views
- Small multiples - Each view shows different data but all views are anchored to the absolute position in the genome. Very clever.

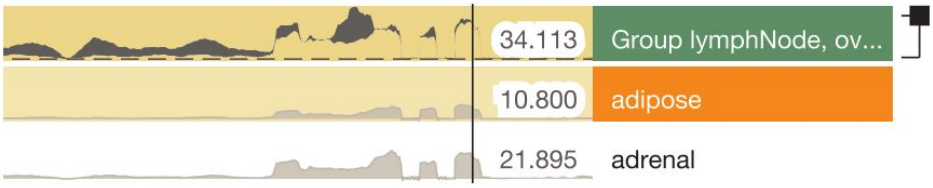


Scalability Is A Key Goal

- Notice efficiency of encoding for volume of data.
- Hundreds of samples. Hundreds of reads per BP
- Uses visually efficient encodings throughout
- Allows custom aggregation by grouping of samples.
- Distortion - Stretch And Squish to collapse introns. All the action is in the exons.

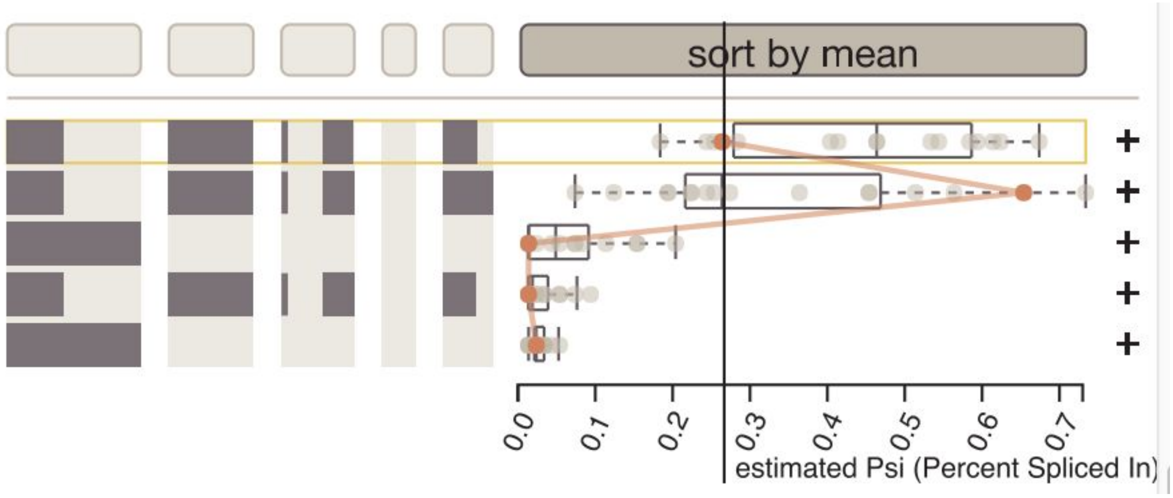
Expression View - Key = (Loc, Sample), Val =

- Abundance at the per base pair level
- Allows custom aggregation via user defined groups
- Hue is used for encoding group memberships in other views
- Focus by hover - linked highlighting across all views



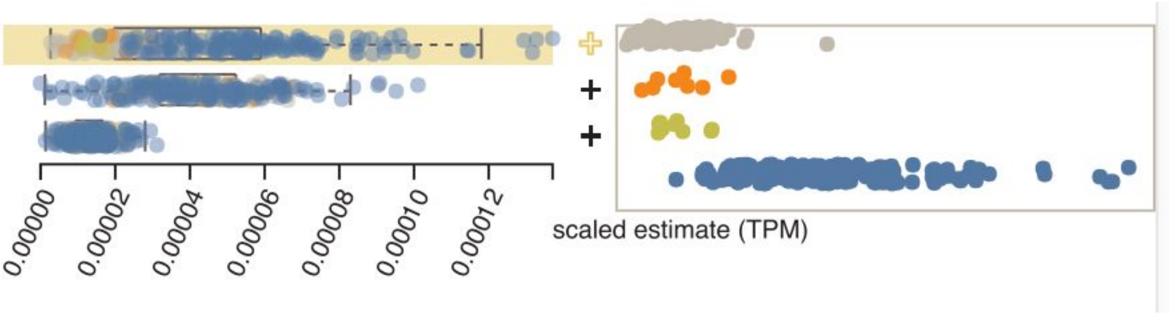
- Shows how often each base pair is read in the sample
- Aggregates samples additively into the group view at the top
- Focus by hover - will do a linked highlight the sample across all views
- Main place for user defined aggregation. Lets you group samples which is propagated to all other views and encoded with color.

Isoform Abundance View



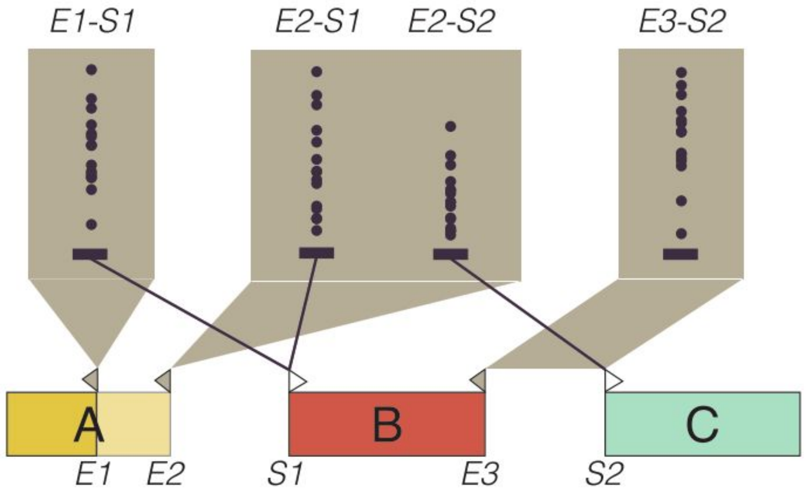
- Each row is a particular isoform
- Dark bars represent the exon that is included
- Grayed out area represents the full spectrum of that exon’s splicing
- Spatial position on an aligned scale.
- Dot plots showing abundance per sample. Embedded barplot showing distribution.
- If you click the “+”

Overview + Detail On Demand

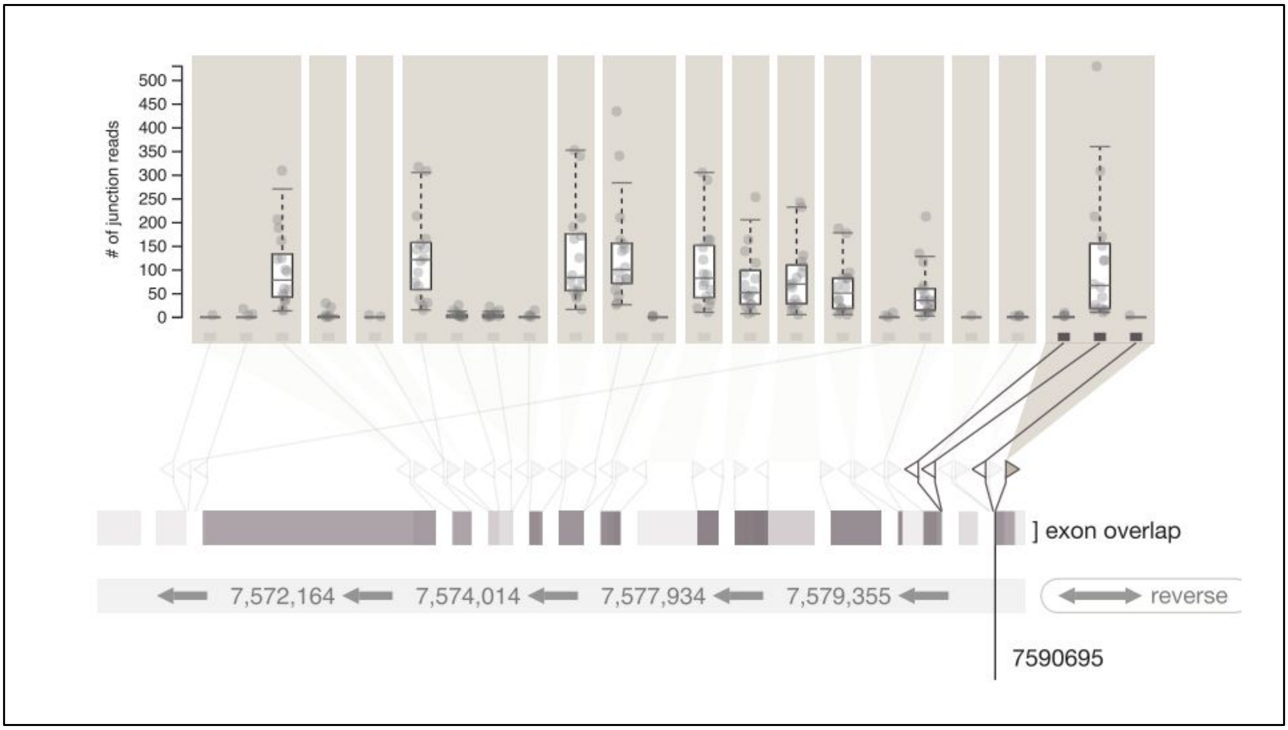


- One dot per sample. Aggregate by group.

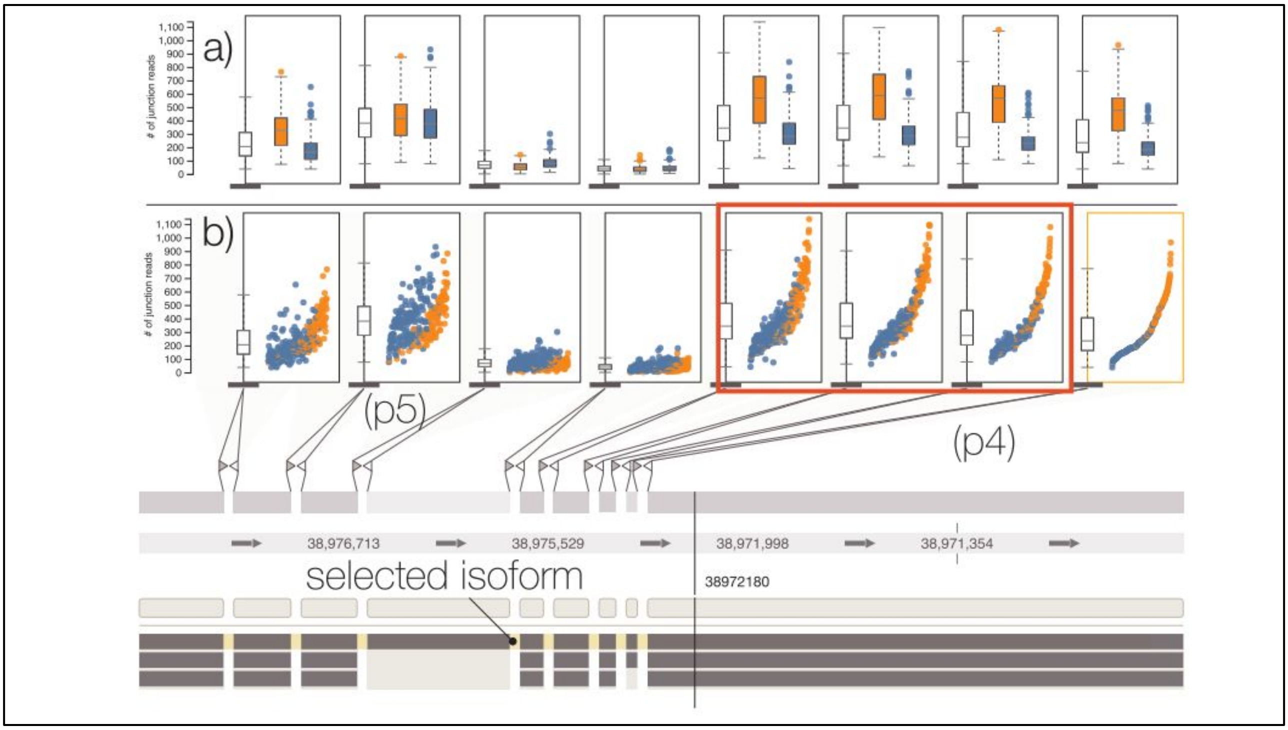
Junction View - Graph Based Data



- Shows junction reads.
- For a particular start site, shows the projection.
- Line marks into the projection show the end of the junction
- Dot plots showing abundance of junction support

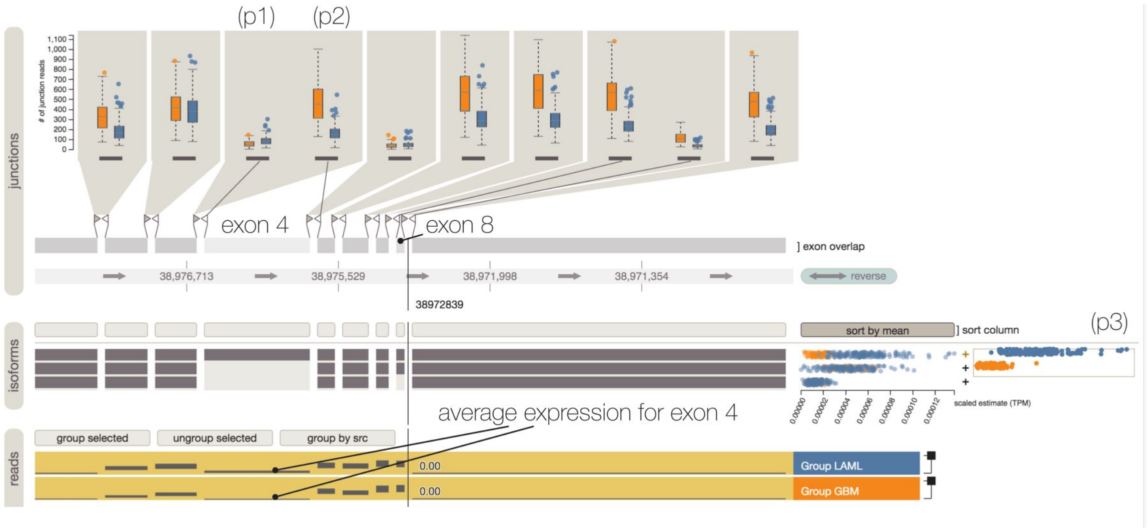


- Actual view in Vials
- Fades all other junctions on hover
- Triangle glyphs are distorted to fit adjacent exon truncation sites



- Again the junction support view.
- Multiform with data shown as both dotplots and boxplots
- Map group membership to hue
- Allows for comparison between groups and samples.
- Actually a study of the SRSF7 gene. Regulates alternative splicing.
-

GBM vs LAML



- You can see that exon 4 is differentially expressed
- Included much more in GBM than in LAML
- Sanity check on the other edge.

Synthesis + Critique

- Great job on scalability.
  - Details on Demand, Distortion, Custom Aggregation + Filtering
- Visual efficiency was prioritized (dot plots + boxplots, aligned position)
- Global coordinate system allows easy navigation and browsing. Keeps orientation.
- Excellent analysis of tasks. Co-authors are analysts.

- Didn't address mismapped reads. Some motifs can be very prevalent.
- No facilities for annotation. Hard to remember discoveries.

Questions?