Decoding Next-Gen Sequencing Data using Custom-built Computational and Quantitative Methods

Yongsheng Bai, Ph.D.

Department of Biology
Indiana State University

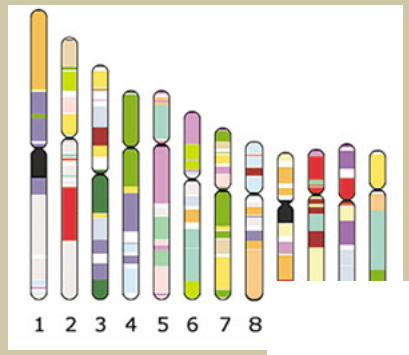
August 4, 2014

Outline

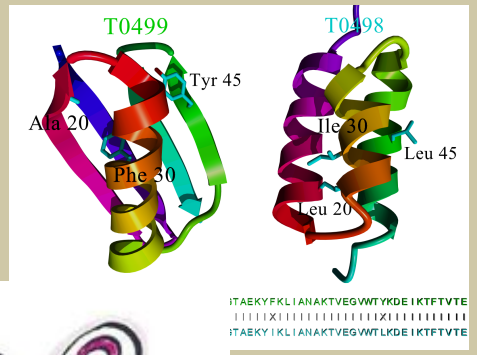
- Read-Split-Walk (RSW): Identification of genome-wide, non-canonical spliced regions using RNA-Seq data
- SNPAAMapper: Efficient genome-wide variant analysis pipeline for whole-exome and whole-genome sequencing data

Biology in the Post-Genomic Era

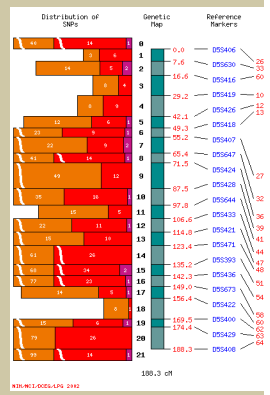
Genome



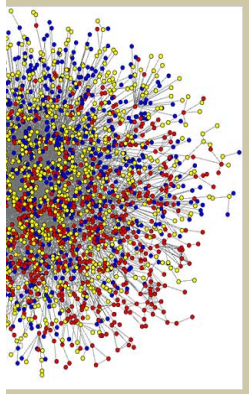
Structure



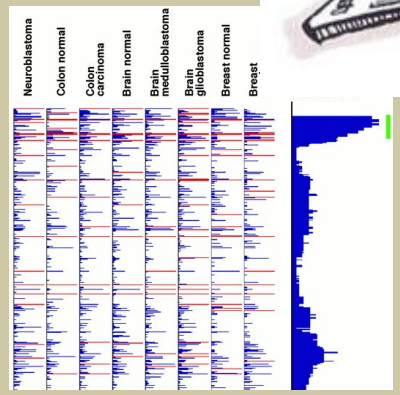
SNP



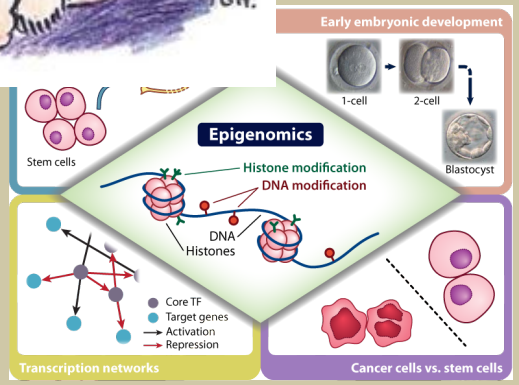
Interaction



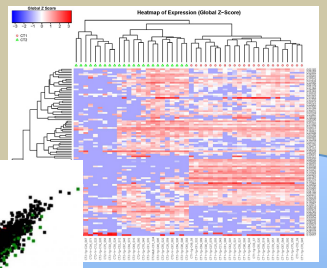
Expression

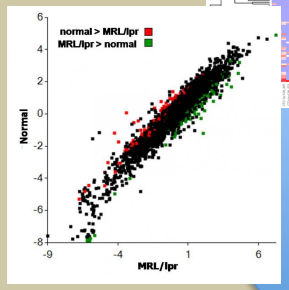


Epigenome

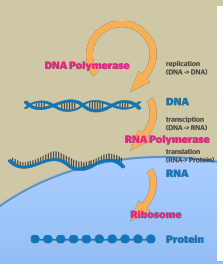


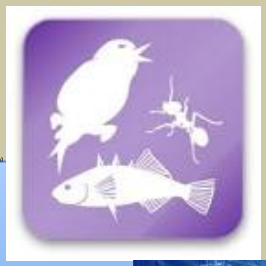
Bioinformatics: an *Interdisciplinary* Field and a *Helpful* Hand





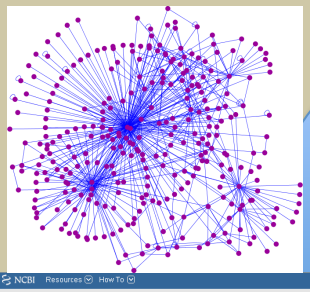
Mathematics/ Statistics



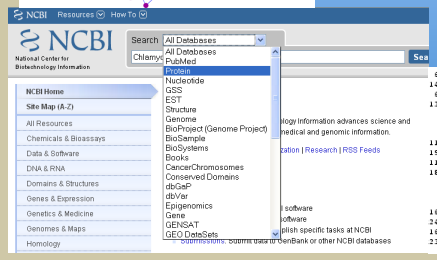


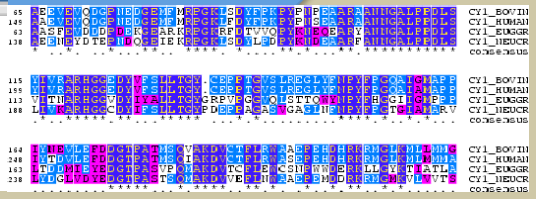
Biology/ Medical Science

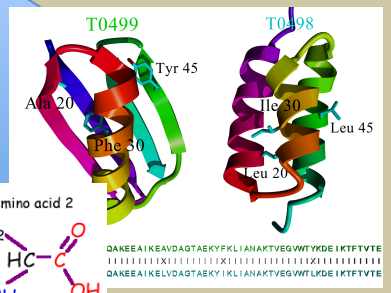
Bioinformatics



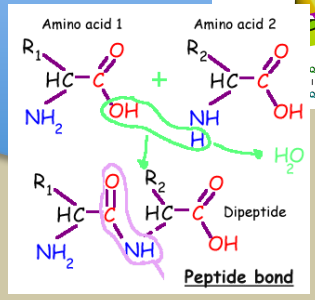
Computer Science



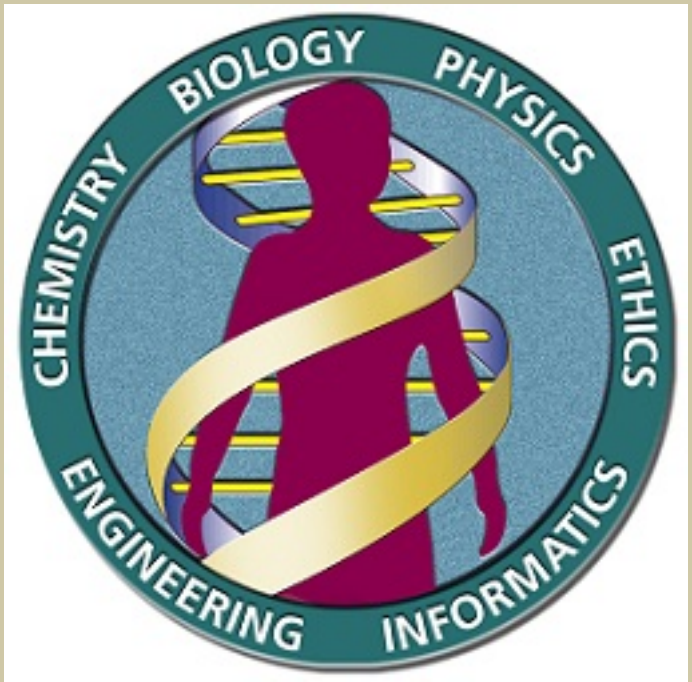


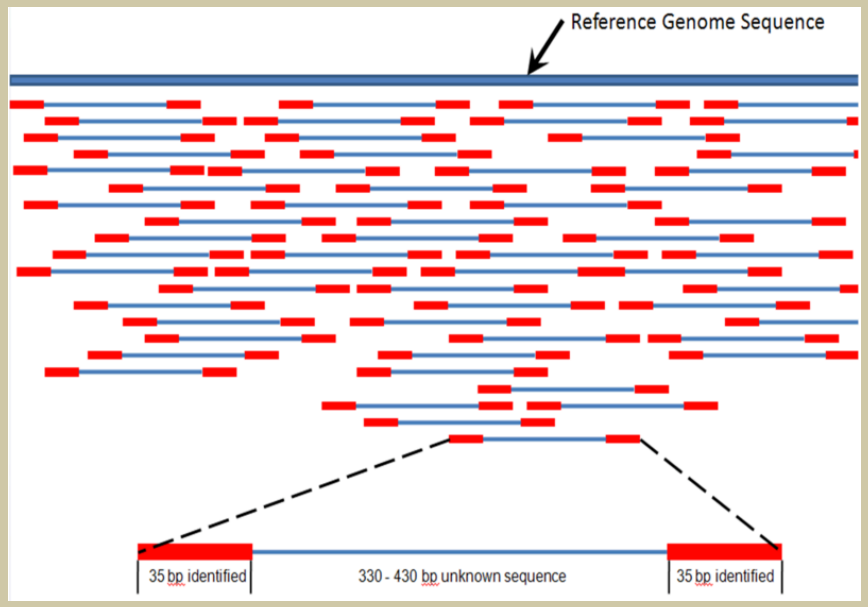


Physics/ Chemistry

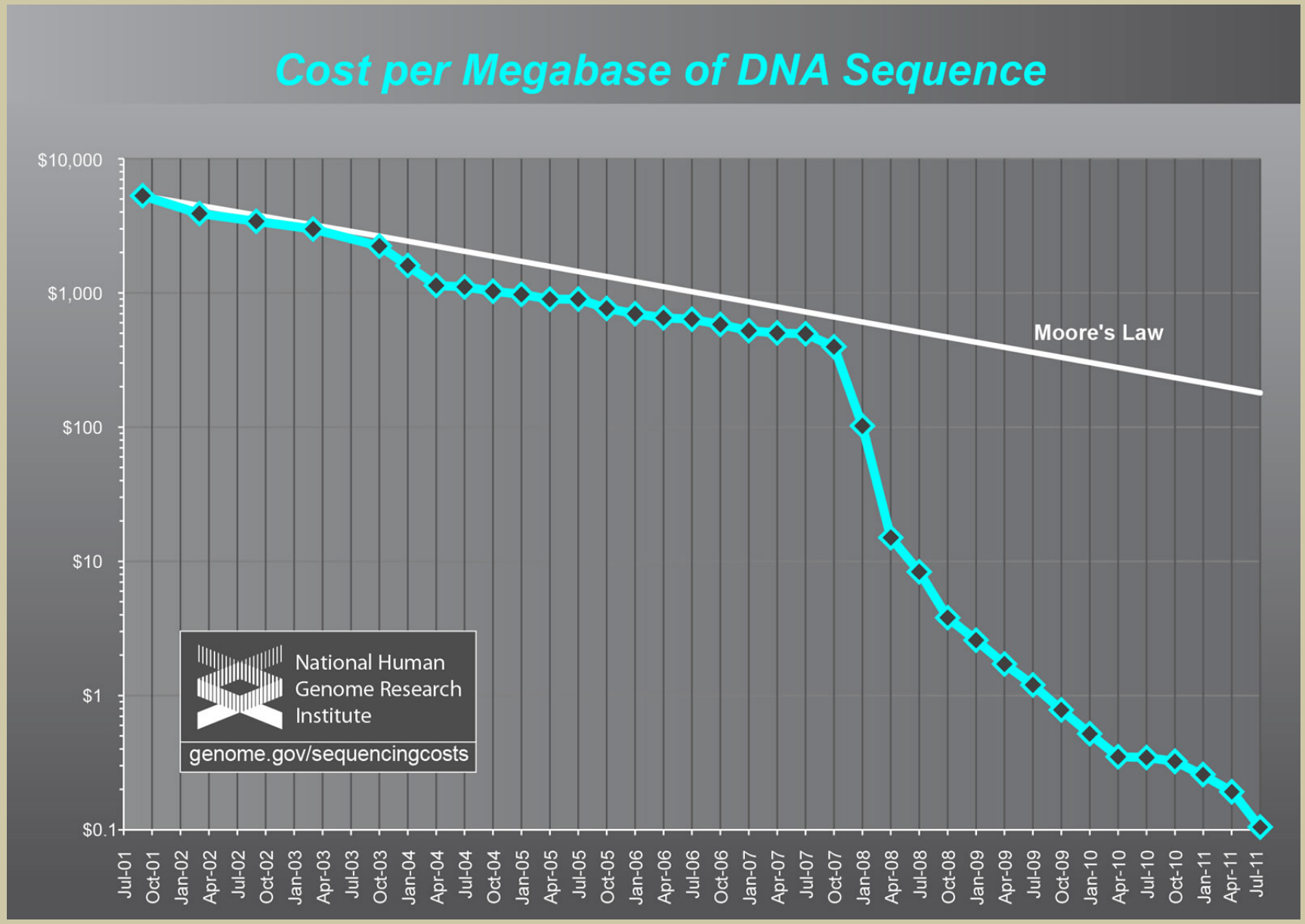


Human Genome Project and Next-Generation Sequencing





Bioinformatics analysis tools/software were also subsequently developed!


Cost per Megabase of DNA Sequence
$10,000
$1,000
$100
$10
$1
$0.1
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11
Apr-11
Jul-11

Next-Generation Sequencing

- Enhanced sequencing capabilities
  - Increased throughput (huge increase in # of reads, decrease in time to produce)
  - Decreased costs per base
  - Ability to sequence from individual samples
- Drawbacks
  - Different error profiles and rates for different platforms
  - Short reads make assembly and mapping to reference difficult
  - Level of coverage becomes important to revealing signatures because repeated regions are collapsed
  - Low complexity regions do not have high-quality assembly

Standard Downstream Uses

- **Whole genome shotgun sequencing**: Whole genome assembly and genome comparisons within and between species
- **Targeted region sequencing (Exome-Seq)**: Reference mapping and SNP calling
- **Whole transcriptome sequencing (RNA-Seq)**: Expression quantification and novel splice junction detection
- **Chromatin Immunoprecipitation Sequencing (ChIP-Seq)**: Regional DNA-protein interaction sequencing
- **Random regions sequenced across samples (RAD-seq)**: Next-generation population genetics

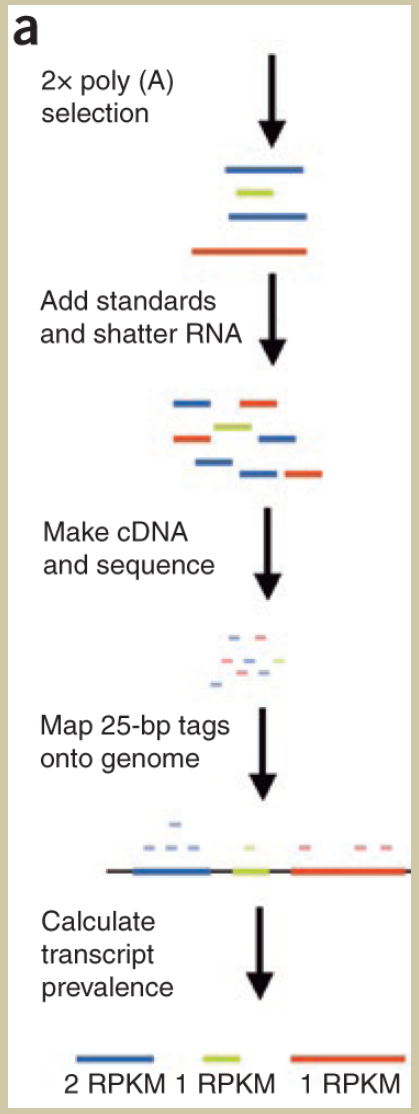
Why Do We Need Customized NGS Analysis Tools?

- Given the complexity of the genome, a program coded to one problem may not deal with a broader range of problems.
- Each project has a different goal and different types of questions to answer. It is also difficult to design a universal tool due to the complexity of biological questions.

Outline

- Read-Split-Walk (RSW): Identification of genome-wide, non-canonical spliced regions using RNA-Seq data
- SNPAAMapper: Efficient genome-wide variant analysis pipeline for whole-exome and whole-genome sequencing data

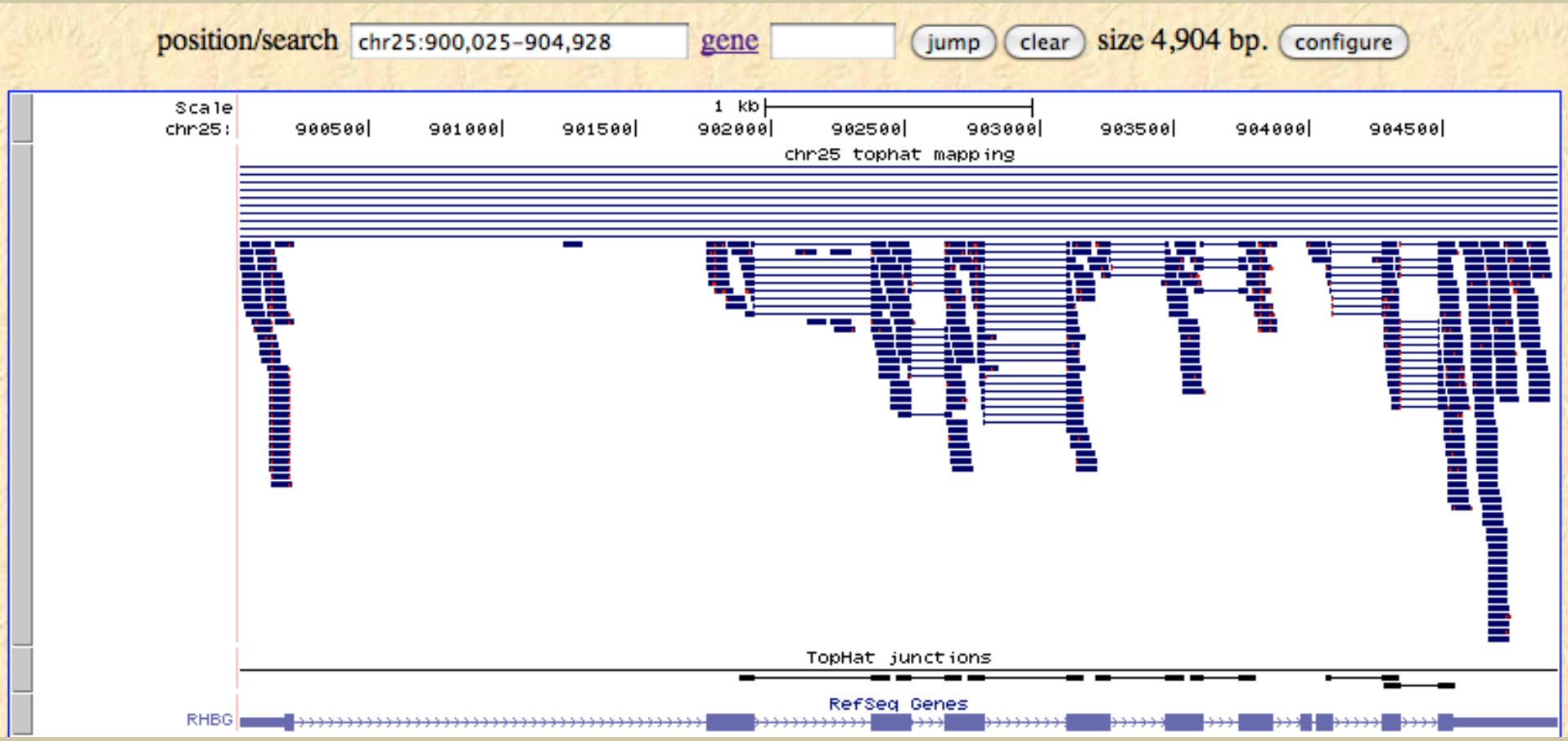
RNA-Seq Method



1. mRNA is enriched using polyA selection
2. mRNA is sheared to library insert size (200-400 nt)
3. cDNA is prepared from mRNA
4. Linkers/adapters ligated to each end
5. Individual sequences determined
6. Transcripts are mapped and quantified

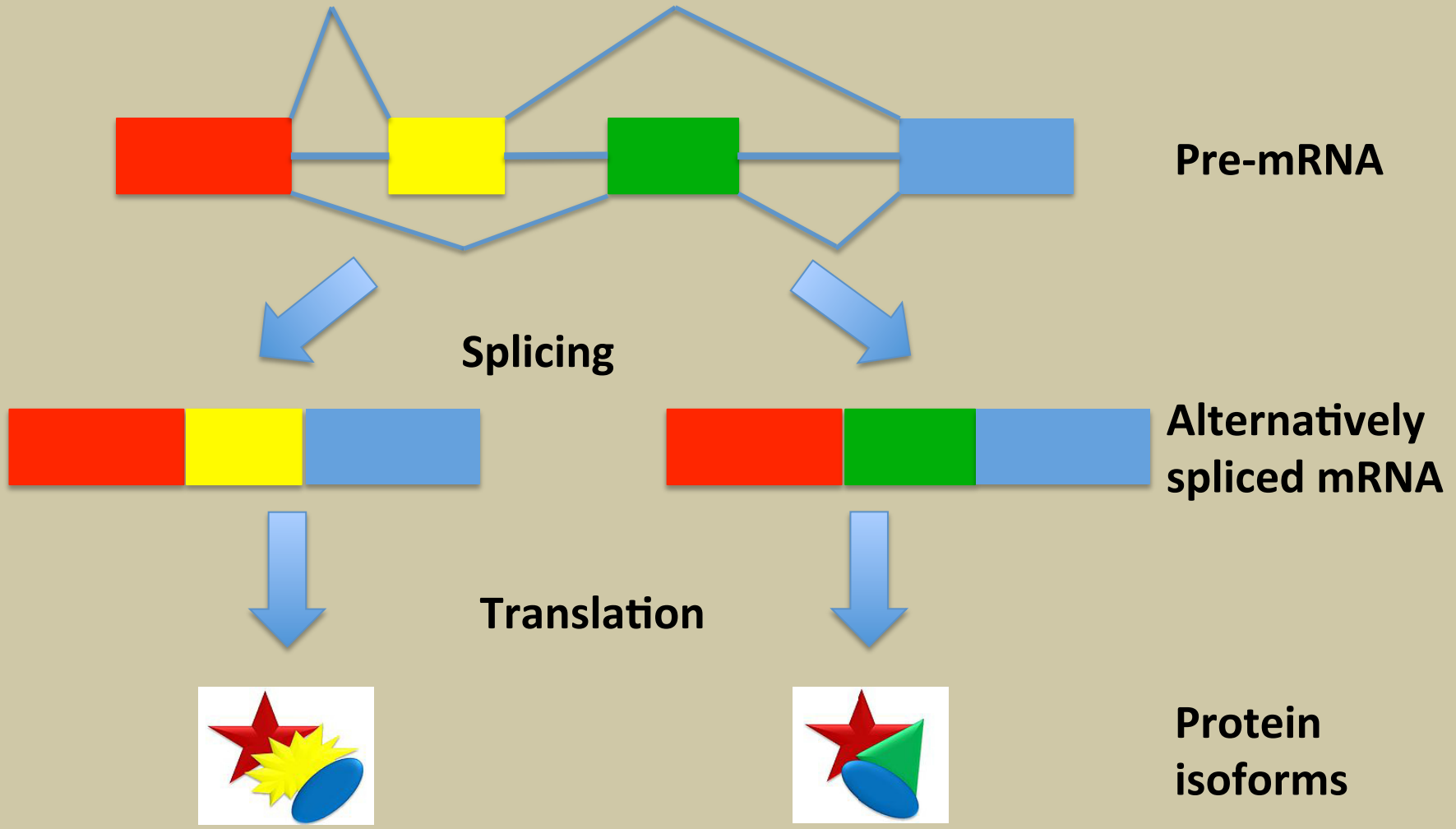
(Mortazavi *et al.*, 2008)

RNA-Seq Visualization

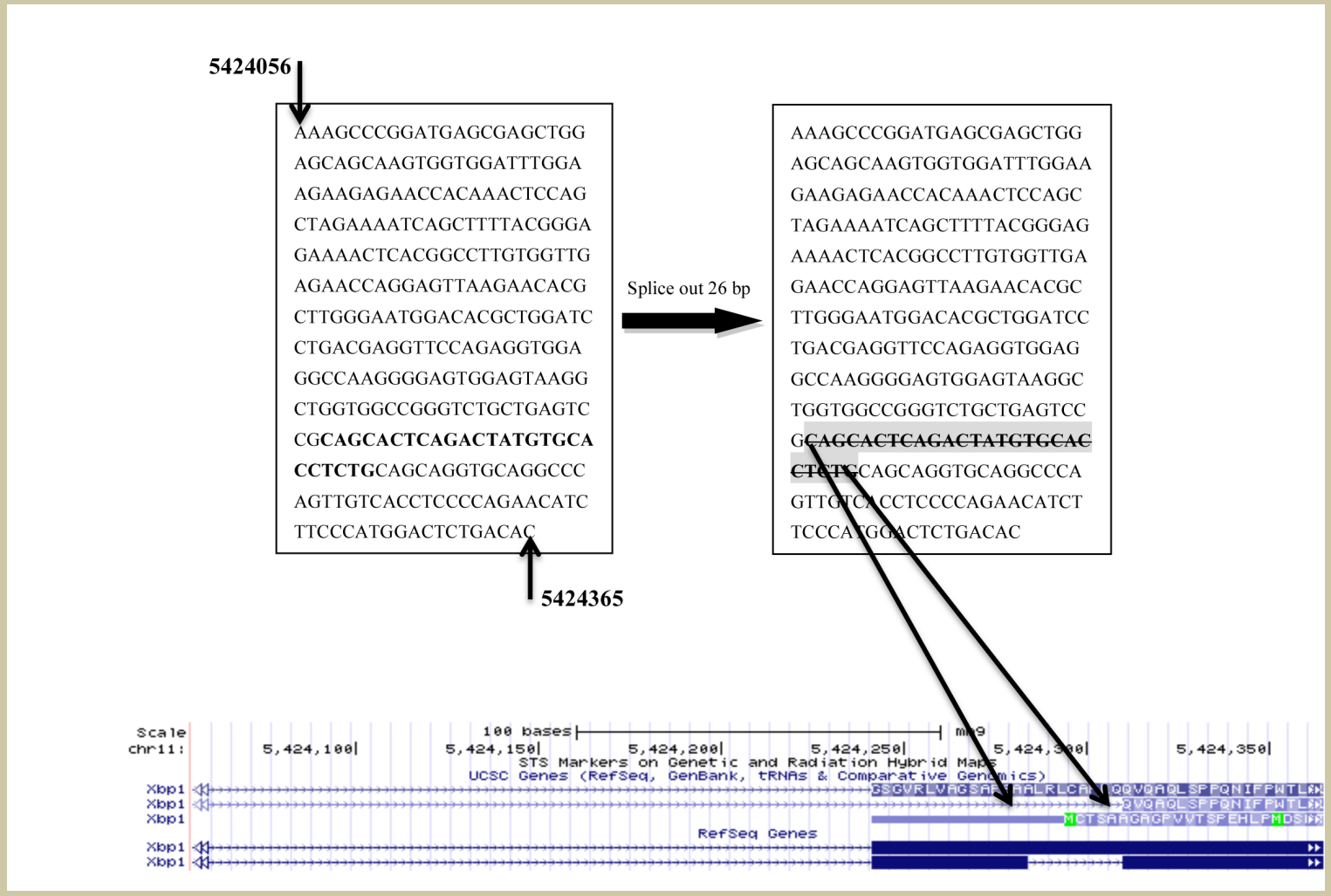


(UCSC Genome Browser: http://genome.ucsc.edu)

Alternative Splicing



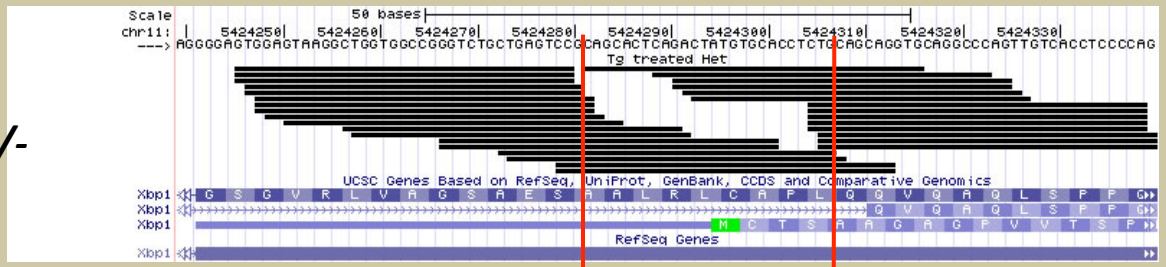
Xbp1's 26nt Splicing



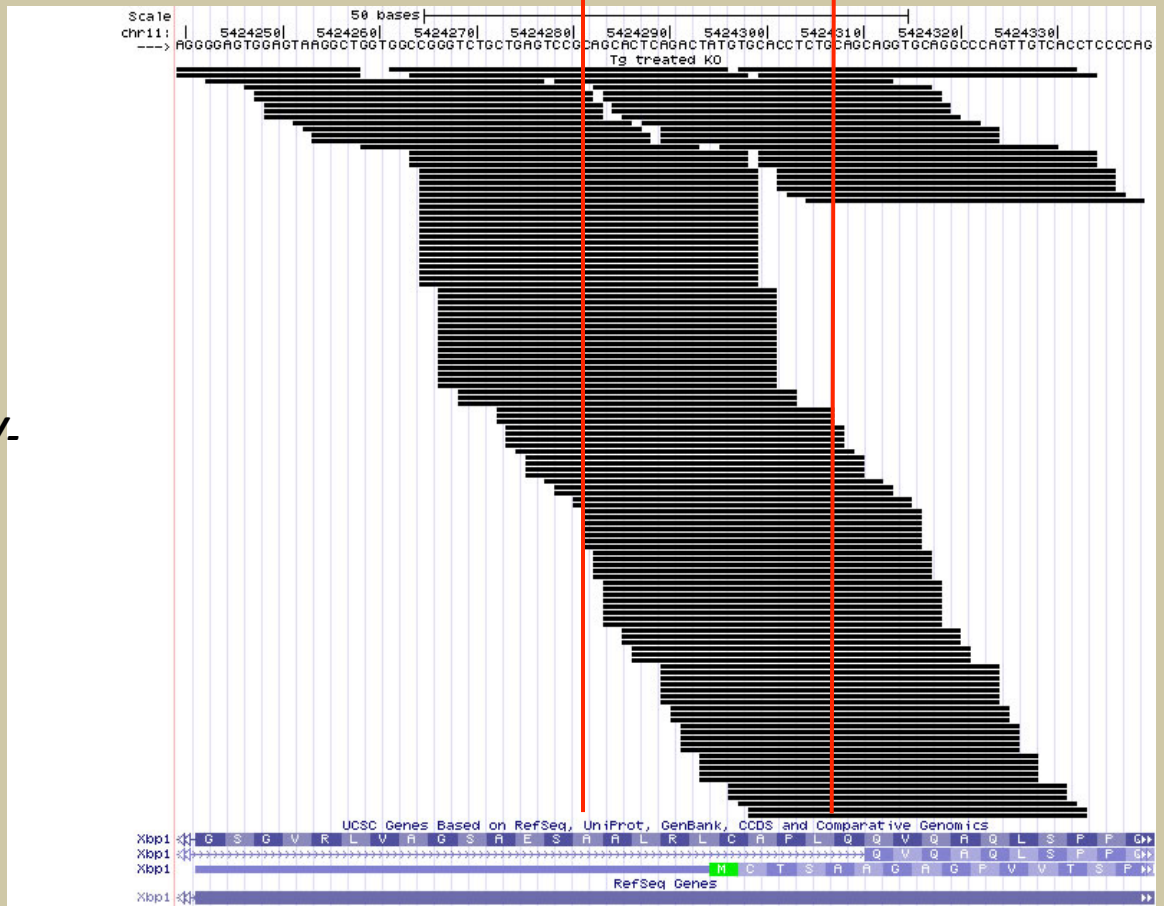
(Bai *et al.*, 2014)

Change in Coverage at *Xbp1*'s Unconventional Spliced Region (Tg treated MEFs)

***Ire1α*$^{+/-}$**



***Ire1α*$^{-/-}$**



Motivation

- Does *Ire1α* have additional targets?
- The goal is to identify *Ire1α* dependent alternative splice variants (genes) in the Heterozygous (Het) sample (*Ire1α*$^{+/-}$) but not in the Knockout (KO) sample (*Ire1α*$^{-/-}$)

Dr. Randal Kaufman
Sanford-Burnham Medical Research Institute

Two RNA-Seq Experiments

- *Ire1α*$^{+/-}$ *vs Ire1α*$^{-/-}$ in **Thapsigargin** (Tg) treated MEF (mouse embryonic fibroblast) cells
- *Ire1α*$^{+/-}$ *vs Ire1α*$^{-/-}$ in **Dithiothreitol** (DTT) treated MEF (mouse embryonic fibroblast) cells
- **Thapsigargin and/or Dithiothreitol** are expected to cause a stronger unfolded protein response (UPR)

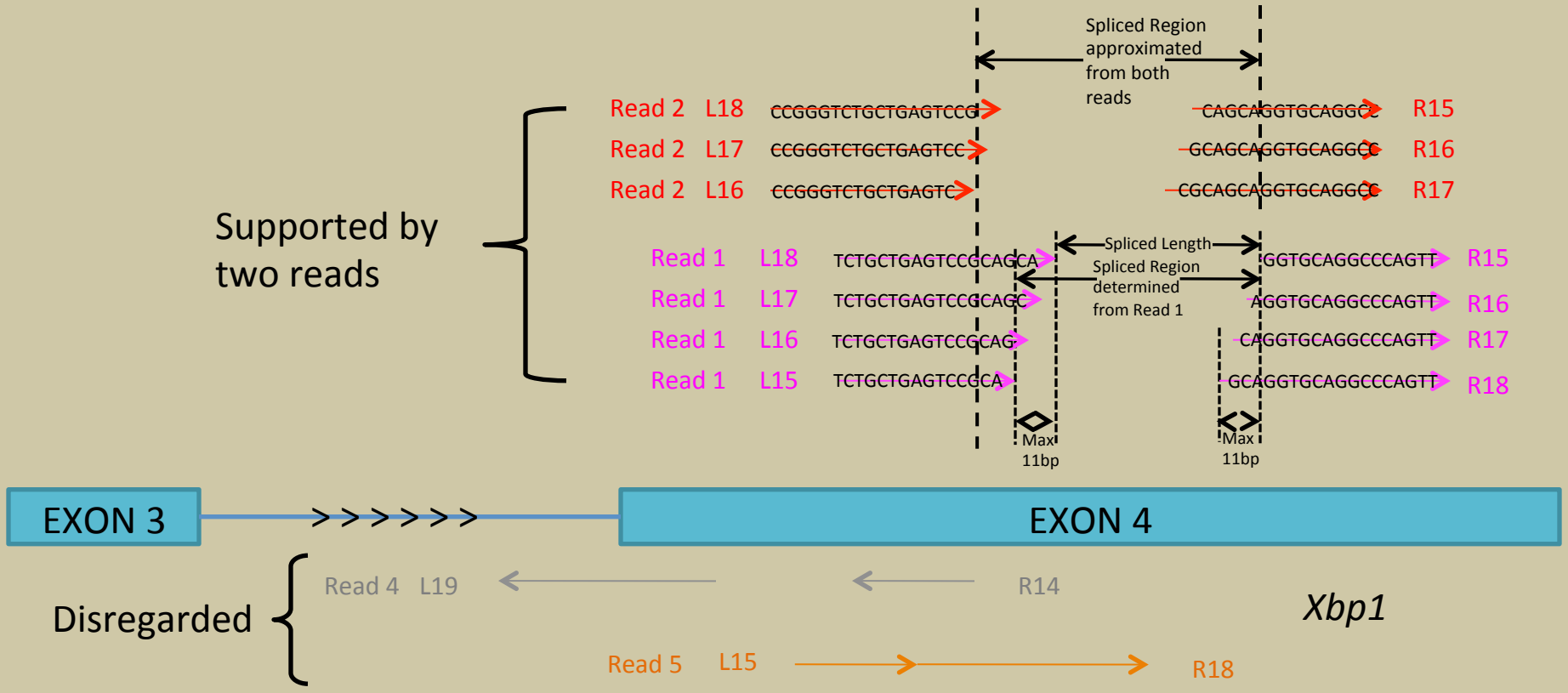
Tools to Identify Novel Splice Sites in RNA-Seq Data

- **Alt Event Finder** (Zhou *et al*., 2012)
  - Processes the mapped reads to report splice regions and does not consider unmapped reads in the analysis input
- **TrueSight** (Li *et al*., 2013)
  - Utilizes all identified junctions and builds a regression model to report the best alignment. The algorithm is not suitable for our *Xbp1*-like non-canonical spliced region detection study due to the lack of pre-known Xbp1-like spliced regions for model training
- **SplicingCompass** (Aschoff M., 2013)
  - Maps reads to the reference genome initially and does not utilize unmapped reads. It then predicts genes that undergo differential splicing based on the expression level

But NONE of these tools were designed for detecting the type of non-canonical, short splicing pattern observed in *Xbp1*!

"READ-SPLIT-WALK (RSW)" Algorithm

- Align original reads and split unmapped reads
- Identify split pairs and report spliced regions



(Bai *et al.*, 2014)

Results

Table 1. The number of reads aligned to *Xbp1*'s 26 bp splice junction by our "Read-Split-Walk" (RSW) algorithm for MEF samples

Sample type	Gene	Splice length	Spliced region	Supported by # of reads	
				Het	KO
Tg treated	*Xbp1*	26	chr11:5424280-5424312	21	0
DTT treated	*Xbp1*	26	chr11:5424280-5424312	173	0

(Bai *et al*., 2014)

Additional Top Ranked Genes Identified by RSW

Samples	Gene name	Chromosome	# of reads mapped	Splice length	Spliced region
Tg treated	Fhl1	chrX	4	1797	54043221--54045021
	Cyr61	chr3	3	2	145310282--145310291
	Tatdn2	chr6	3	121	113659910--113660036
DTT treated	Rps9	chr7	80	4	3658432--3658436
	1810013L24 Rik	chr16	30	3	8858087--8858093
	Sparc	chr11	23	2	55208920--55208938

(Bai *et al*., 2014)

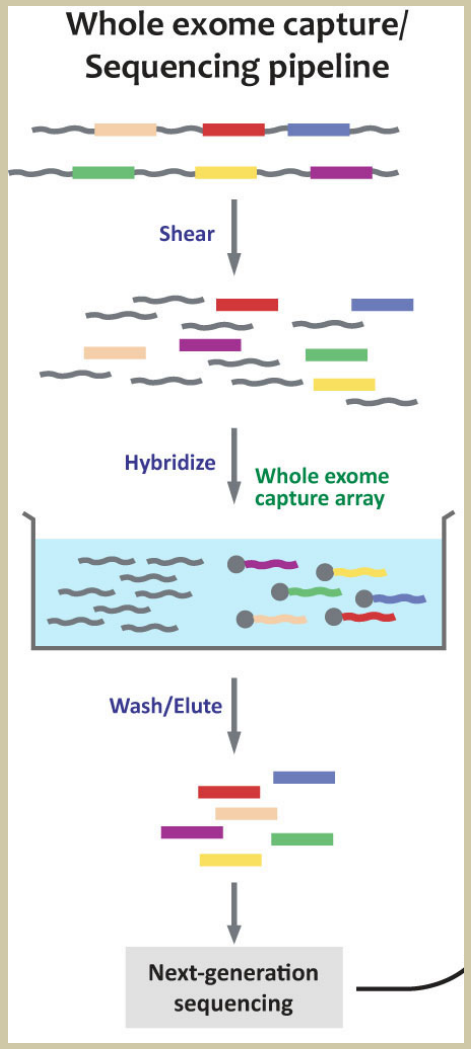
Summary of Read-Split-Walk

- We concluded that *Xbp1* mRNA is likely the only significant splicing target of *Ire1α* in the MEF transcriptome.
- Our RSW method of detecting a particular type of non-canonical splicing from RNA-Seq data is complementary to existing novel splice junction detection approaches in the field that are not able to detect these events.

Outline

- Read-Split-Walk (RSW): Identification of genome-wide, non-canonical spliced regions using RNA-Seq data
- SNPAAMapper: Efficient genome-wide variant analysis pipeline for whole-exome and whole-genome sequencing data

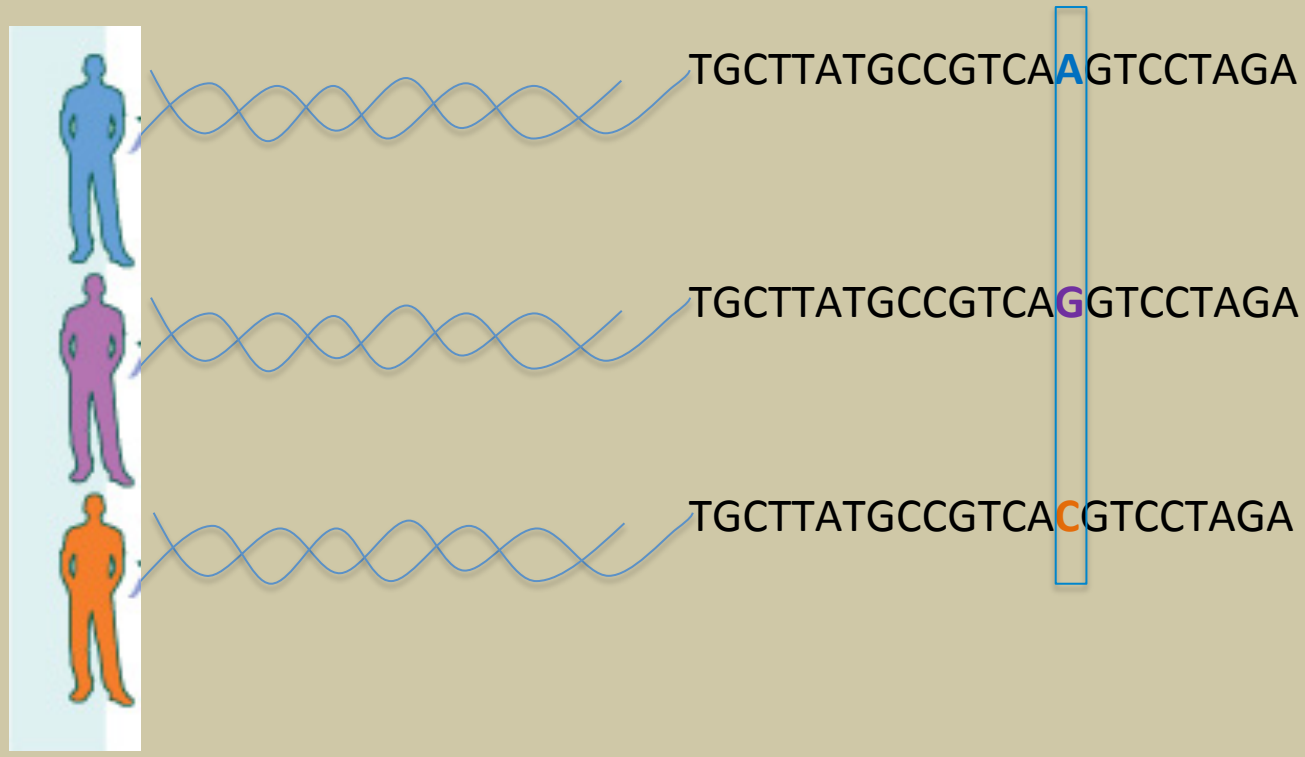
Exome Sequencing



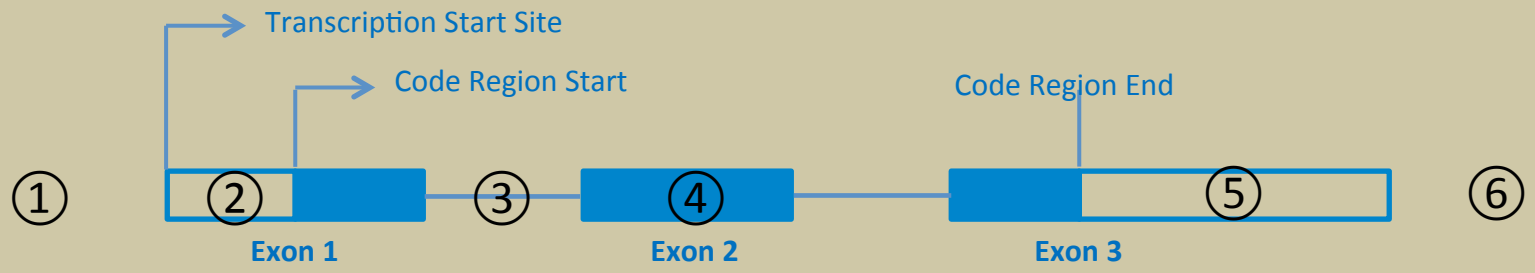
1. Sample DNA (QC)
2. Prepare fragment library
3. Hybridize exon sequences
4. Amplify enriched library
5. Library QC and quantification
6. Sequencing

(Goh and Choi, 2012)

Single Nucleotide Polymorphism (SNP)



SNP Classes



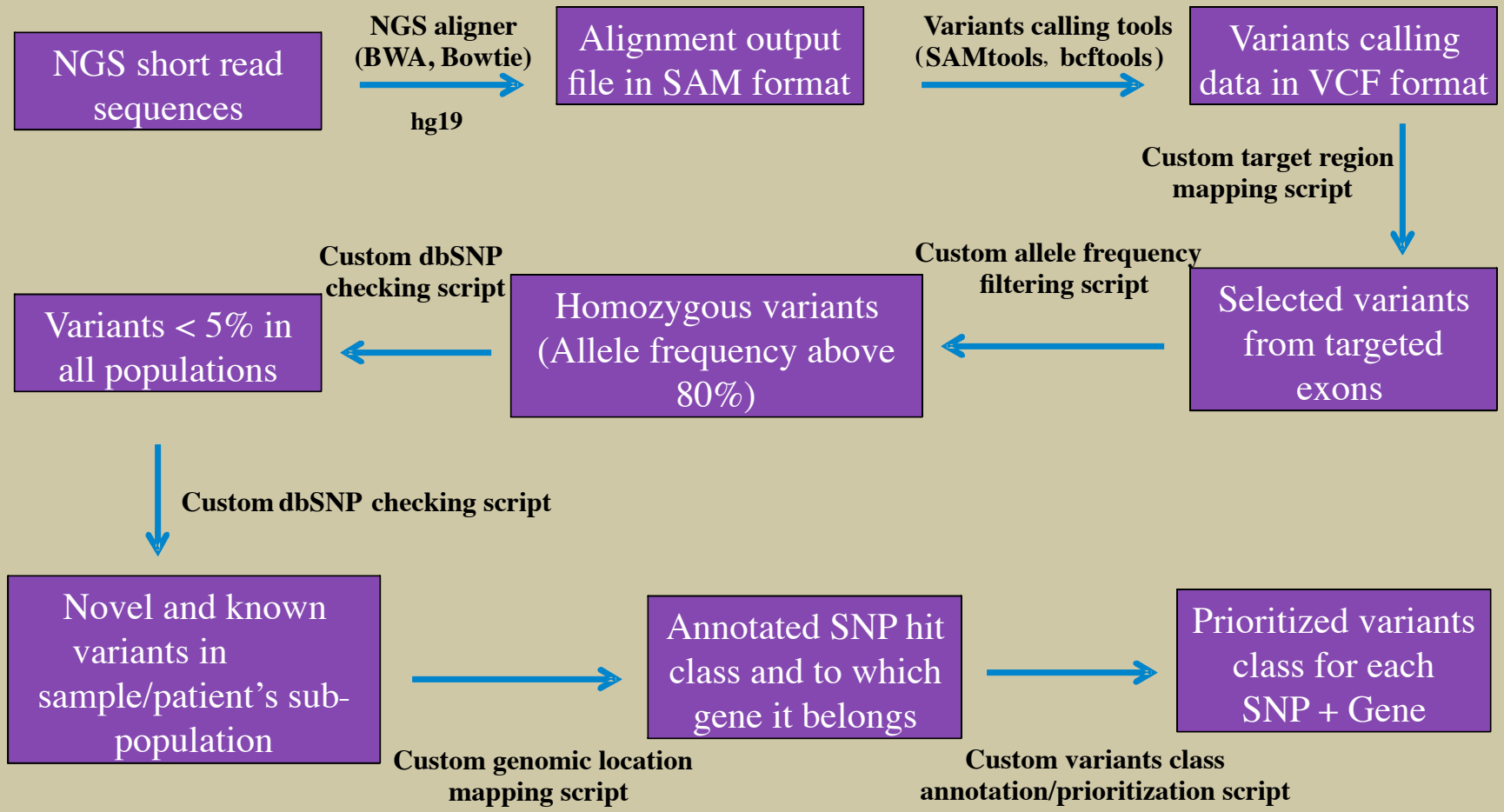
- Nongenic SNPs ①, ⑥
  - May still affect gene splicing, transcription factor binding…
  - Could be present at upstream or downstream from the gene
- Genic SNPs
  - 5' UTR SNPs ②
  - Intronic SNPs ③
  - CDS SNPs ④
    - Synonymous SNPs
    - Non-synonymous SNPs
      - Missense SNPs
      - Nonsense SNPs
  - 3' UTR SNPs ⑤

Need for Powerful Downstream Variant Analysis Tools

- Many short read aligners and variant calling tools have been developed in recent years
- **But few downstream analysis tools for annotating and analyzing detected genetic variants are available**
- **A powerful variation analysis tool that can analyze multiple samples and classify and prioritize identified variants is valuable**

(Bai* and Cavalcoli, 2013)

SNPAAMapper Pipeline



Features of SNPAAMapper

- SNPAAMapper pipeline provides a convenient tool that allows detected variants to be further elucidated and facilitates the fast downstream interpretation of identified variants.
- SNPAAMapper accepts input data generated from several NGS platforms and has the capability of annotating multiple samples simultaneously.
- SNPAAMapper has significant implications in studying genetic differences both within a population and among different populations. It can also help us understand how populations have diverged through evolution.

Acknowledgements

Department of Computational Medicine and Bioinformatics at University of Michigan

Maureen Sartor

James Cavalcoli

Sanford-Burnham Medical Research Institute

Justin Hassler

Randal Kaufman