Indiana State University

ENCODE

Evaluation of U12-type non-canonical splicing in human ENCODE RNA-Seq datasets and analysis of biological functions for spliced sequences by *Read-Split-Fly* algorithm

Yongsheng Bai and Jeff Kinne

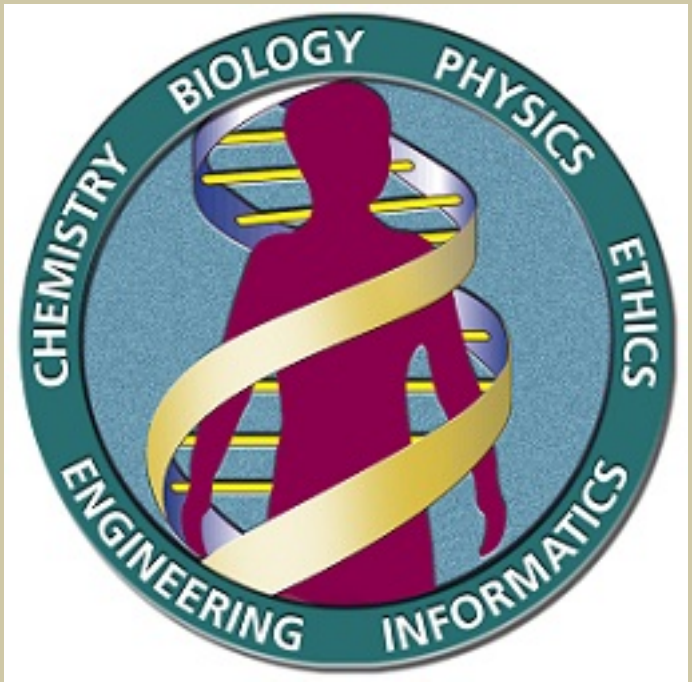
Department of Biology
Department of Mathematics & Computer Science

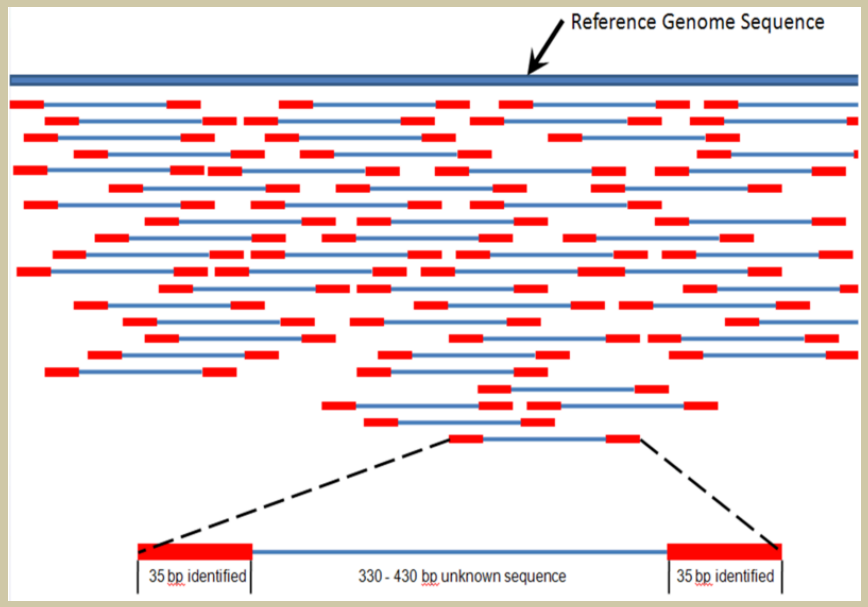
June 8, 2016

THE CENTER FOR GENOMIC ADVOCACY AT INDIANA STATE UNIVERSITY

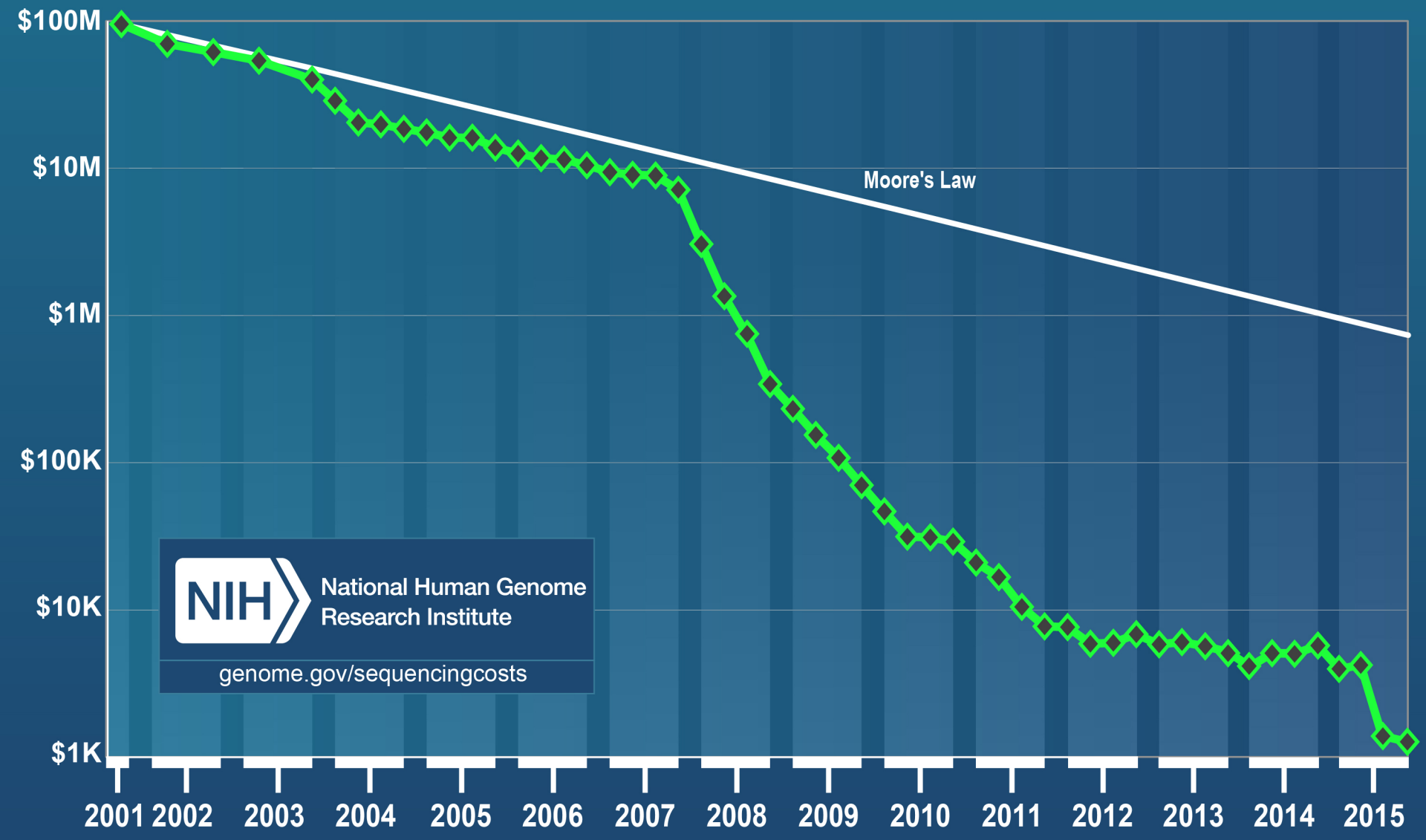
UP UNBOUNDED POSSIBILITIES Innovate. Strengthen. Unify.

Human Genome Project and Next-Generation Sequencing



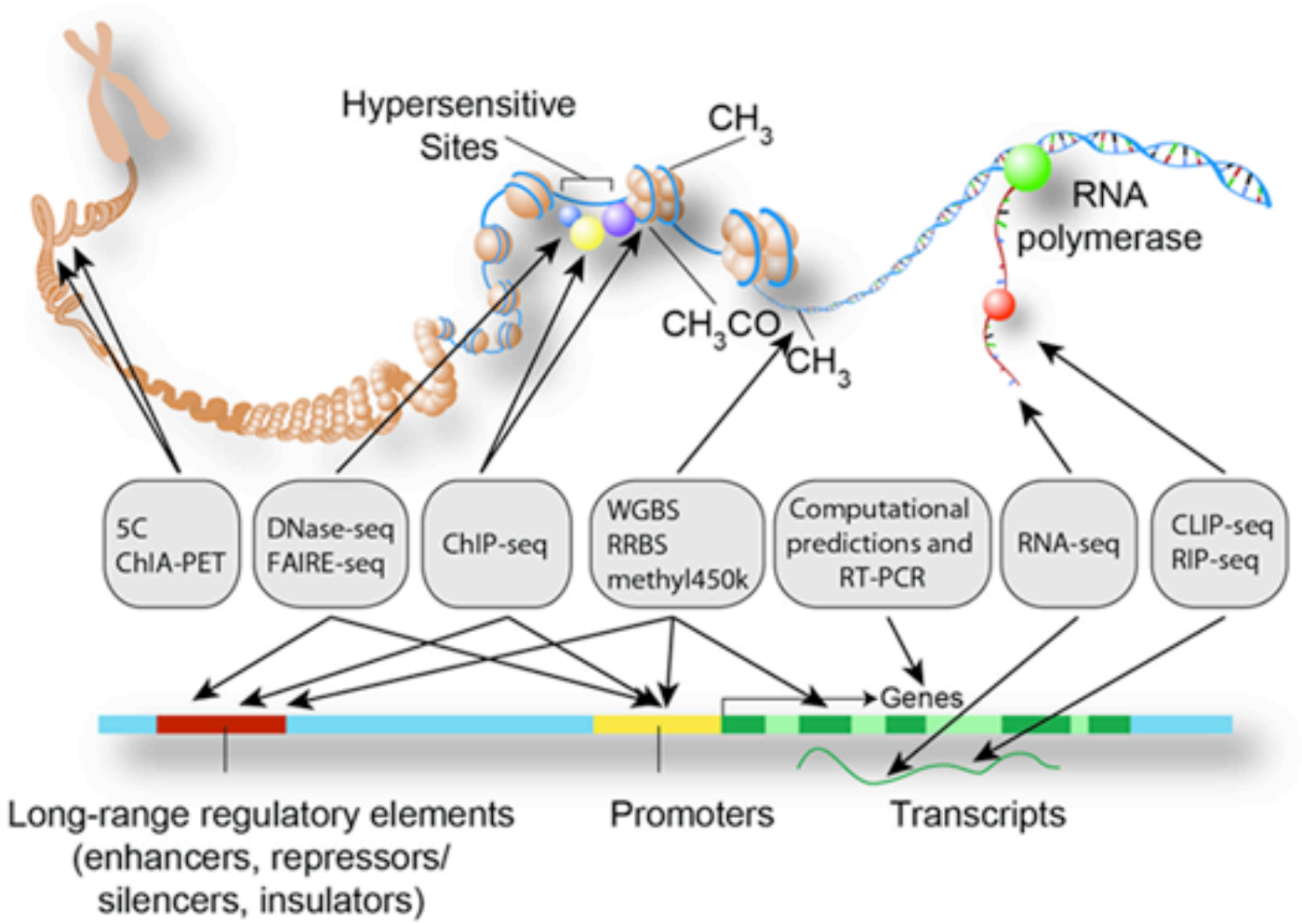


Cost per Genome



ENCODE Project

(https://www.encodeproject.org/)

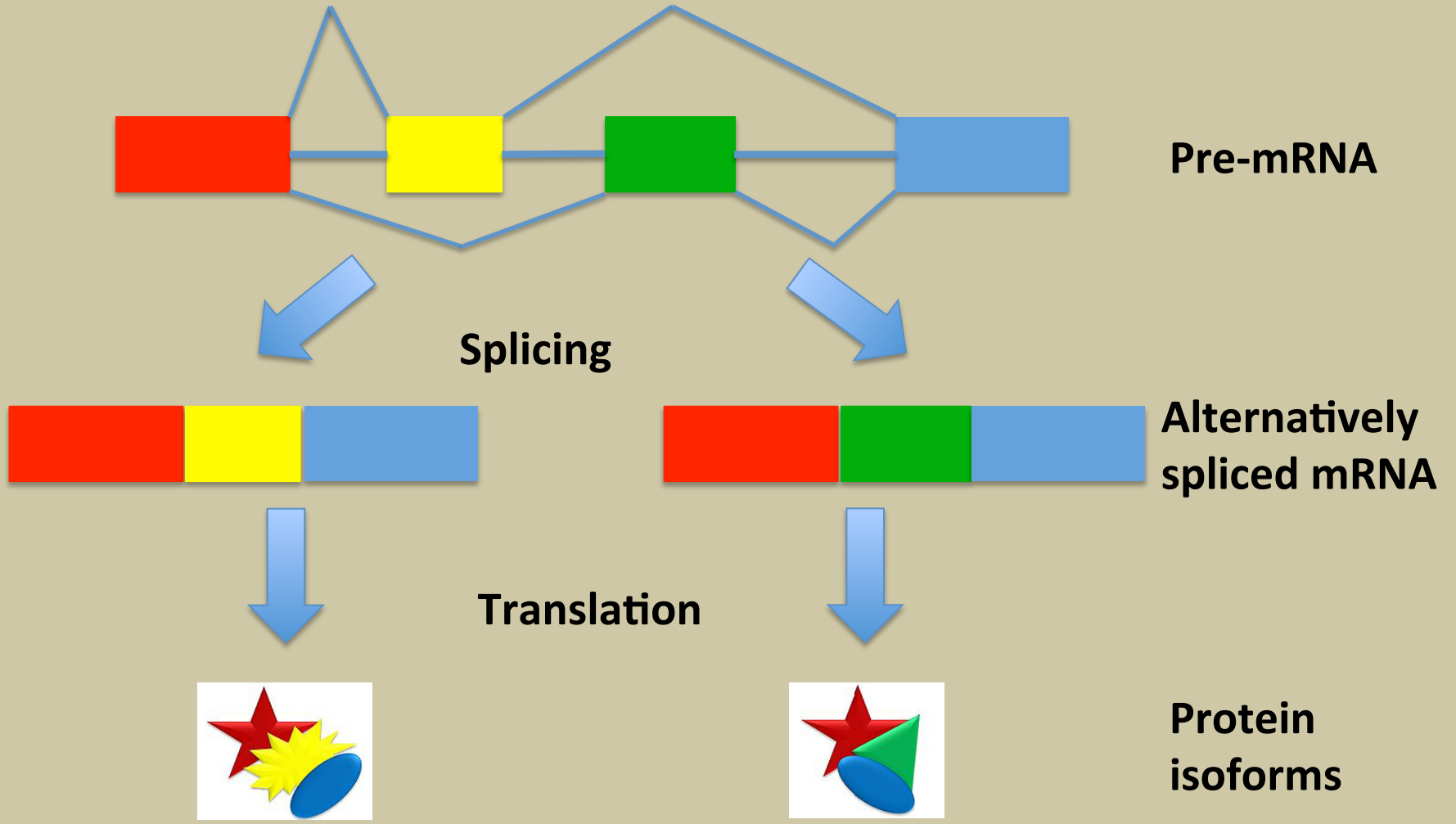


Datasets – ENCODE

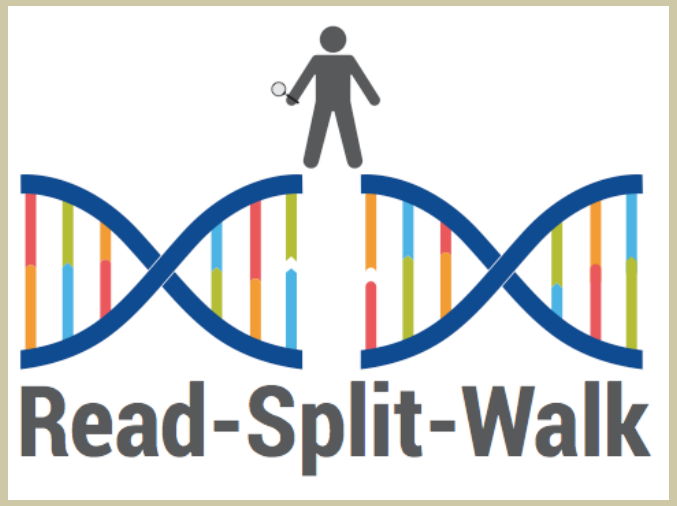
(https://www.encodeproject.org/)

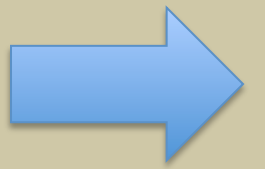
Dataset Classification	Experiment	File Name	Run type	Forward Or Reverse Mate	Biological Replicate	Technical Replicate
Paired-Ended with Biological Replicate	ENCSR468ION	ENCFF002DJH	paired-ended	1	1	1
	ENCSR468ION	ENCFF002DJU	paired-ended	2	1	1
	ENCSR468ION	ENCFF002DJV	paired-ended	1	2	1
	ENCSR468ION	ENCFF002DJX	paired-ended	2	2	1
Paired-Ended without Biological Replicate	ENCSR001UXR	ENCFF576OBS	paired-ended	1	1	1
	ENCSR001UXR	ENCFF063OPQ	paired-ended	2	1	1
Single-Ended with Biological Replicate	ENCSR000BYG	ENCFF000MVL	single-ended	1	1	1
	ENCSR000BYG	ENCFF000MVN	single-ended	1	2	1

Transcription and Translation



RSR Pipeline





The original code for the RSW algorithm worked by considering all pairs of parts of unmapped reads at once. The software was written in **Perl**.

Results are computed much faster in RSR by considering each unmapped read separately. The algorithm was implemented in **C++**.

Comparison of number of reads for supporting *Xbp1* 26 nt spliced regions reported by RSR and other tools

	500 nM Thapsigargin (Tg)		1 mM Dithiothreitol (Dtt)	
Software	Het (Ire1α +/-)	KO (Ire1α -/-)	Het (Ire1α +/-)	KO (Ire1α -/-)
Read-Split-Run (RSR)	**21**	**0**	**173**	**0**
TopHat	23	86	59	289
BWA	0	0	67	0
Bowtie2	0	0	171	0
STAR	0	0	0	0
Alt Event Finder	0	0	0	0

(Bai *et al.*, 2016. To appear in BMC Genomics)

RSF Pipeline – Improved algorithm

- Rescue read halves from bowtie max file

RSF Pipeline – Improved algorithm

- Increased sensitivity
  - Rescue read halves from bowtie max file

***Xbp1* splice Supporting reads**	**Read-Split-Run**	**Read-Split-Fly**
Tg treated sample	21	**27**
Dtt treated sample	173	**209**

RSF Pipeline – Improved algorithm

- Increased performance (Memory and Time Usage)

Memory (GB) / Time (Hours) Usage	Read-Split-Run	Read-Split-Fly	Change (%)
ENCSR558LHB	123.2 / 15.8	18.9 / 11.5	-84.7 / -27.2
ENCSR905FLM	110.2 / 13.5	20.7 / 12.6	-81.2 / -6.7
ENCSR000CQY	55.2 / 0.55	46.5 /0.42	-15.8 / -24

Extracting Spliced Sequences

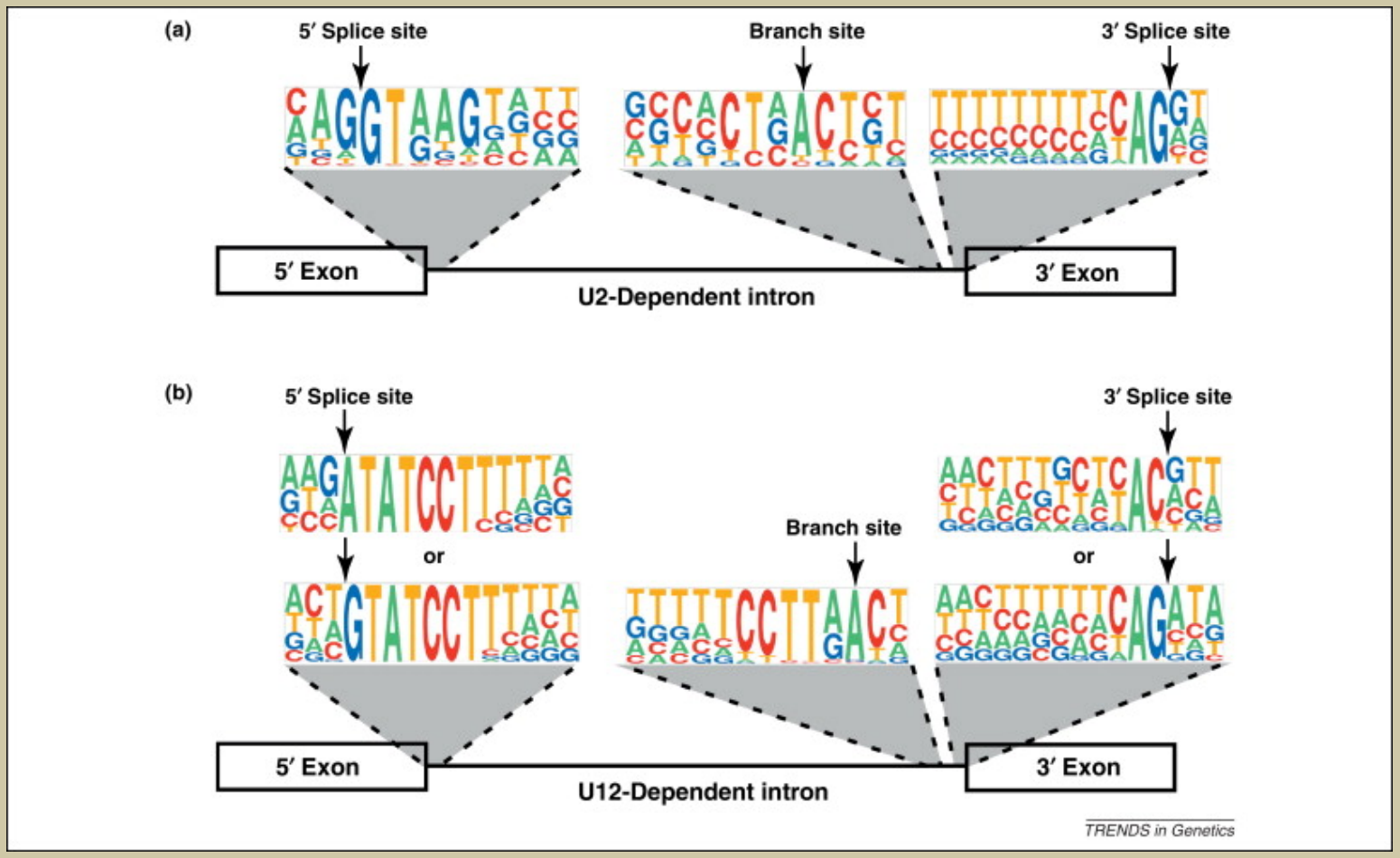
- Sample data before and after bowtie-inspect-RSR file processing:

GeneName	Chromosome	# supporting reads	splice length	range of supporting reads	Novel or not (*)	Bracketed sequence
Xbp1	chr11 22 26	5424280--5424312	Novel	GGGTCTGCTGAGTCCGCA--GCAGGTGCAGGCCCAGT		
Ddx41	chr13 2 74	55635449--55635524	*	CCGAGCCAGCTCTCGC--GAGGGGCAGATGATGAGCC		
Nol7	chr13 4 76	43494024--43494104	*	GCGAGTGCTCGCA--GGGATAAAACGCTGTTGAAGGA		
Gtpbp4	chr13 2 80	8991097--8991177	*	GCACATAATCTTTAGCAACA--TTGTCCACCAAGTTT		
Pdlim7	chr13 5 81	55610171--55610255	*	TGCTCTGAACAGGCTGGGCT--CTGCTAAGACCCAGG		

GeneName	Chromosome	# supporting reads	splice length	range of supporting reads	Novel or not (*)	Bracketed sequence	spliced sequence
Xbp1	chr11 22 26	5424279]--[5424312	Novel	GGGTCTGCTGAGTCC]--[GGTGCAGGCCCAGTT	GCAGCACTCAGACTATGTGCACCTCTGCAGCA		
Ddx41	chr13 2 74	55635448]--[55635524	*	CCGAGCCAGCTCTCG]--[GAGGGGCAGATGATG	CTACAGGCAGAGACAAGGCTGGTACCGGAAGCCAGCACAGTCCTTCTCCCGACCCCAGGATTAGCCTATCTTACC		
Nol7	chr13 4 76	43494023]--[43494104	*	GCGCGCGAGTGCTCG]--[GATAAAACGCTGTTG	CAGGTGCAGCGCCCGGGAATGTTGGCGCCTCGGGACTCCCCAGGCTCGGCCGGCGCTGACCCGACTCCTTTCCTTCCAGG		
Gtpbp4	chr13 2 80	8991096]--[8991177	*	TAATCTTTAGCAACA]--[TTGTCCACCAAGTTT	CTTAGAAAAAGAGAAAGTAATCTAAGGGTTTGTATTTTTTGGTAATAACTTATAAGATATATAGTAAGAAAATGCCTTAC		
Pdlim7	chr13 5 81	55610170]--[55610255	*	TGAACAGGCTGGGCT]--[CTAAGACCCAGGCTG	CTGGAGAAGAAGAGAGACGCAGTCACAGCAGGTAAGAGACCAGAGGCTATCCACACCCCTTCTGCCTGTCAGCACCTGTACCTG		

- Consequently, the spliced sequence fits snugly within the “range of supporting reads.”
  - For example, the ***XBP1*** candidate spliced sequence begins at index 5424280 and ends at index 5424311

Consensus sequences of human U12- and U2-type intron



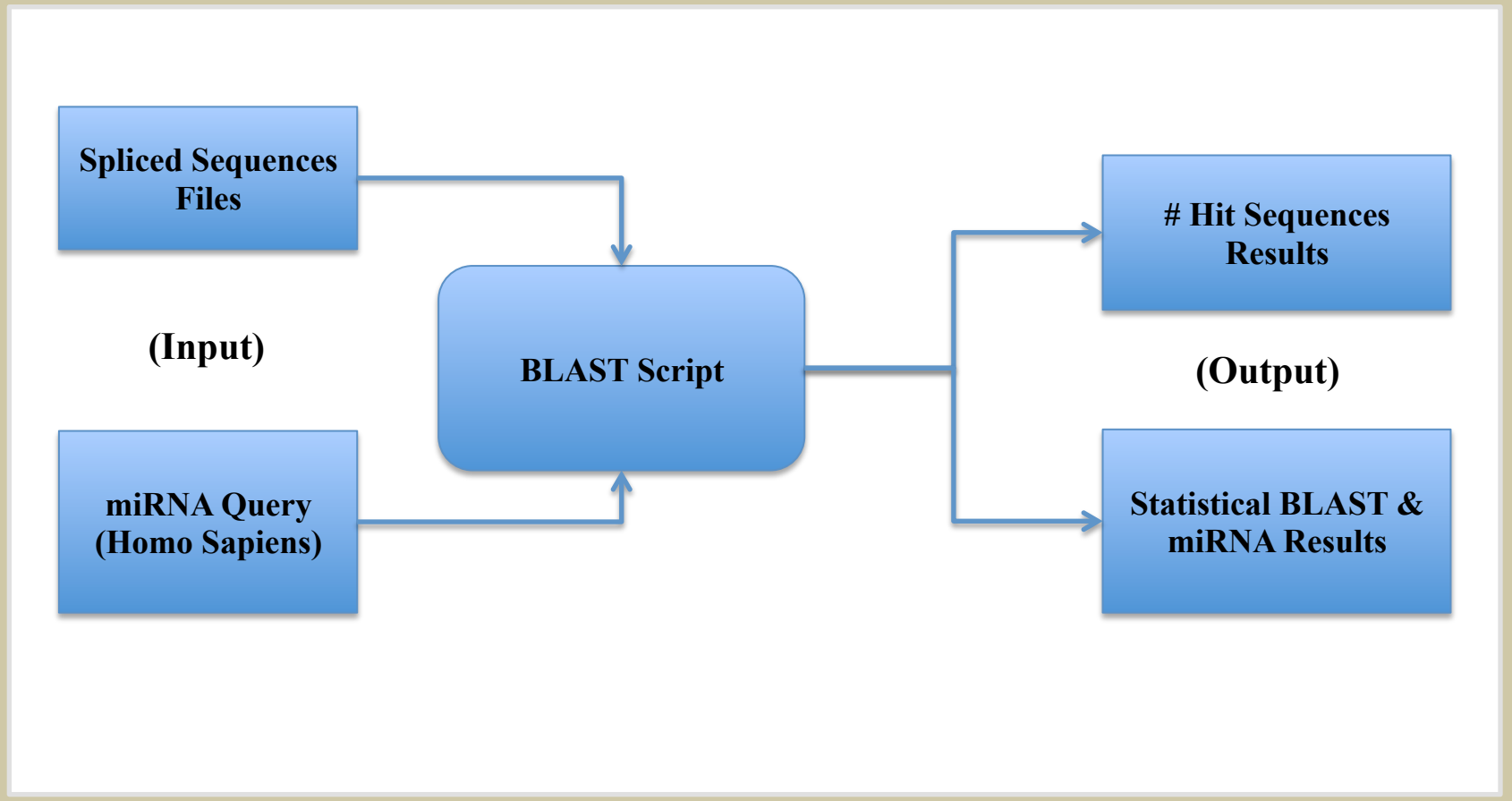
(Padgett, R. 2012)

U12-type non-canonical splicing in 21 human ENCODE RNA-Seq datasets

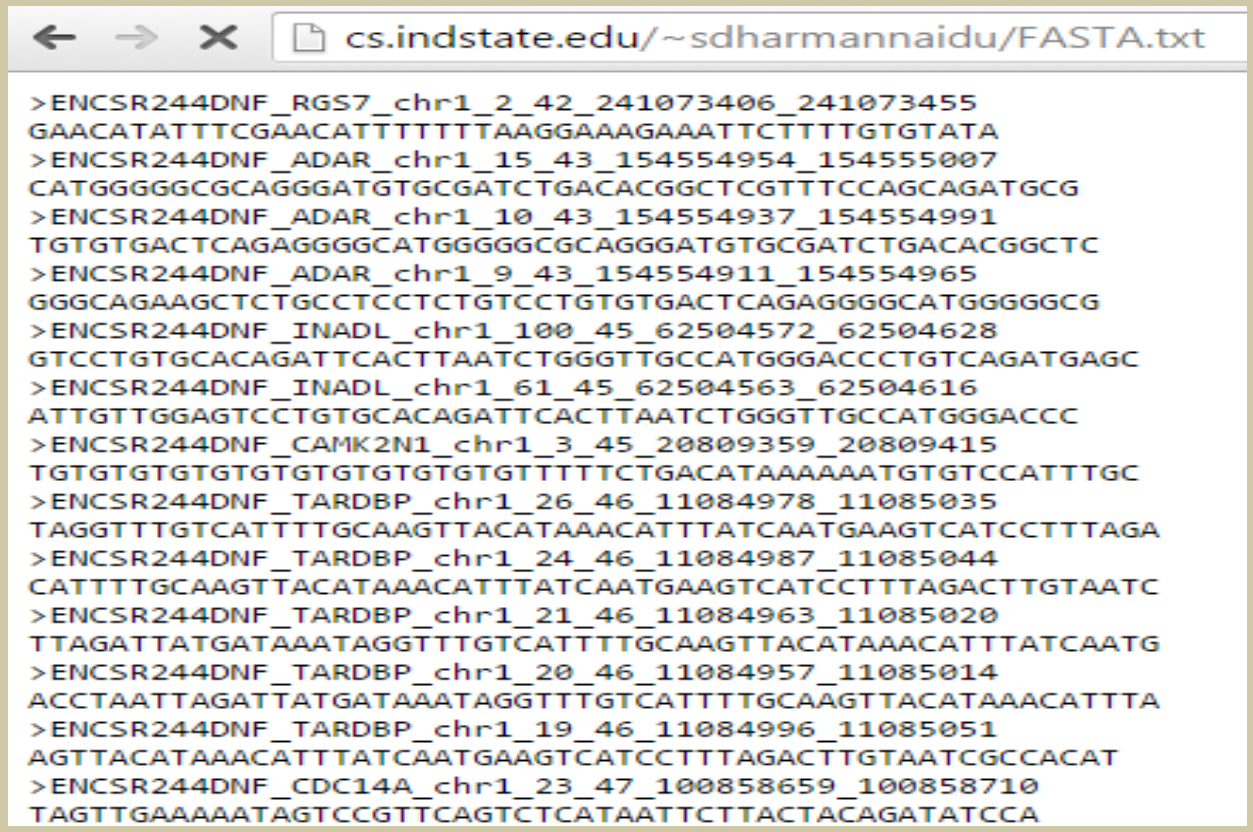
E-Value	1

	u12		u12 Total	u2		u2 Total	Grand Total
Row Labels	known	novel		known	novel		
3pFull	52977	36760	89737	13094	8295	21389	111126
5pFull	8840	11618	20458	3466	2140	5606	26064
branch	222	397	619	NA	NA	NA	619
branch#extend	12968	13342	26310	NA	NA	NA	26310
Grand Total	75007	62117	137124	16560	10435	26995	164119

RSF -BLAST-Workflow



Sample DB File



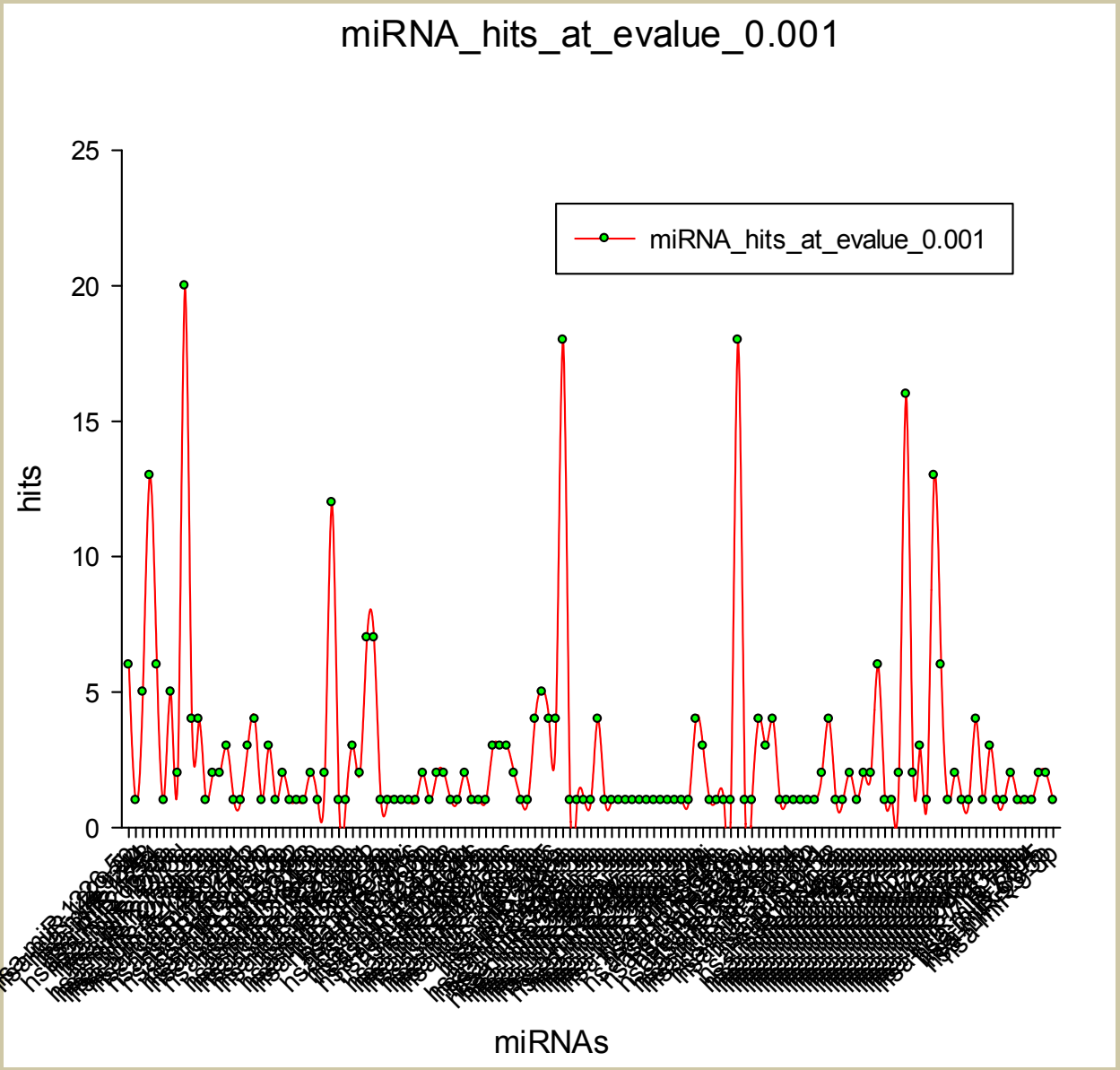
Query – Sample Data

>hsa-let-7a-5p MIMAT0000062 Homo sapiens let-7a-5p
TGAGGTAGTAGGTTGTATAGTT

>hsa-let-7a-3p MIMAT0004481 Homo sapiens let-7a-3p
CTATACAATCTACTGTCTTTC

>hsa-let-7a-2-3p MIMAT0010195 Homo sapiens let-7a-2-3p
CTGTACAGCCTCCTAGCTTTCC

miRNA BLAST hits for 21 ENCODE Samples



Summary

- Preliminary results from 21 samples of ENCOCDE datasets show that there are several miRNAs are prevalent (> 50%) in studied ENCODE samples. Two of them (*hsa-miR-1273d* and *hsa-miR-548*) are associated with many diseases as suggested in the literature.
- U12-type non-canonical splicing could be noteworthy in ENCODE datasets.
- RSF for identifying U12-type splicing events using ENCODE datasets is applicable to study a range of diseases across biological systems under different experimental conditions.

http://bioinf1.indstate.edu/RSR

Read-Split-Run

A pipeline for detecting non-canonical spliced-regions in RNA-Seq data.

Experiment Settings

Mode: Non-comparative

Reads Type: Single

Experiment Replicates: 1

Data

Select Genome Reference: mm9

Select RNA-Seq File (FASTQ format): Tg_Het.txt — Select file...

Quality Encoding: Solexa — Check files

Length of Reads (in bp): 35

Pre-processing

Minimum Split Size (2+) bp: 11

Maximum Good Alignments Allowed Per Read: 11

Candidate Selection

Minimum Distance Between Candidate Pairs (2 to 12) bp: 2

Maximum Distance Between Candidate Pairs (bp): 30k

Read Mapping Region Boundary Buffer Size (0 to 11) bp: 3

Minimum Number of Supporting Reads: 10

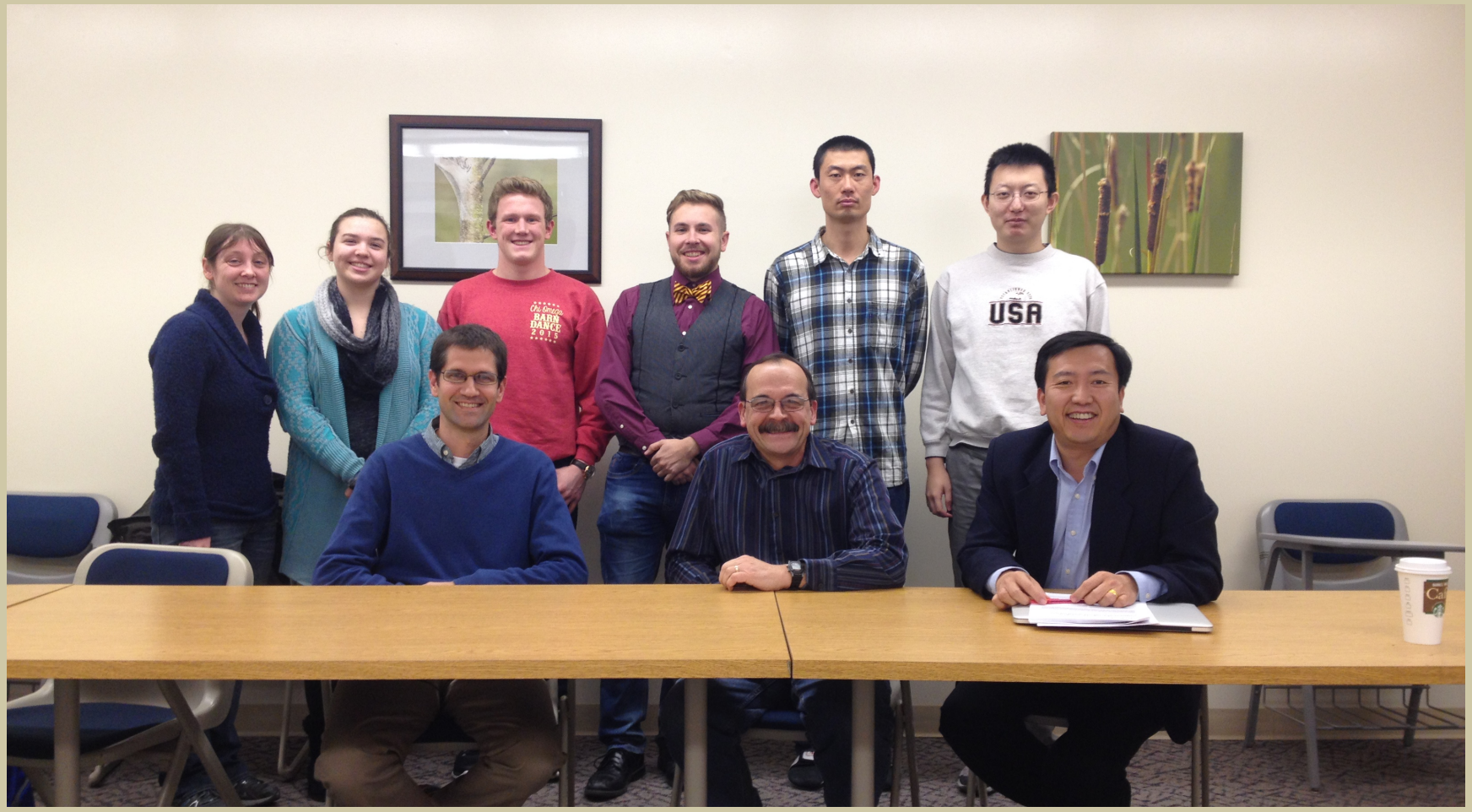
Output

Email: Yongsheng.Bai@indstate.edu

Submit Reset

Bai Lab at Indiana State University © Copyright 2015 All Rights Reserved

Bai Lab (Spring 2016)




bits
4
3
2
1
0
N
1
2
3
4
5
6
7
8
C
weblogo.berkeley.edu