Yale

Analysis of Genomic Variation in Non-Coding Elements Using Population-Scale Sequencing Data from the 1000 Genomes Project

Xinmeng Jasmine Mu
Ph.D. Candidate
Program in Computational Biology & Bioinformatics
Mark Gerstein group, Yale University
X-Gen Congress & Expo
San Diego, March 6, 2012

Introduction & Background

Non-coding regions in the genome:

- have more sequences under natural selection than coding DNA in humans (1.5% of genome is coding, 5% under selection).
- are involved in biological functions and disease associations (conserved non-coding elements and GWAS studies).
- biochemically active within the genome, such as interacting with transcription factors (ENCODE Project).

Much less effort has been invested on non-coding elements, compared to coding regions!

Introduction & Background

Population-based approaches to evaluate the functional relevance of non-coding elements:

- the level of naturally occurring genomic variations –polymorphism. A reduction of polymorphism, compared to sequences under neutral evolution, suggests natural selection or lower mutation rates.
- Polymorphism naturally co-varies with divergence between species regardless of the mutation rate. Nucleotide diversity vs. divergence (McDonald–Kreitman test).
- Selective constraints result in a skew of derived allele frequency (DAF) spectrum towards the low-frequency alleles.

Introduction & Background

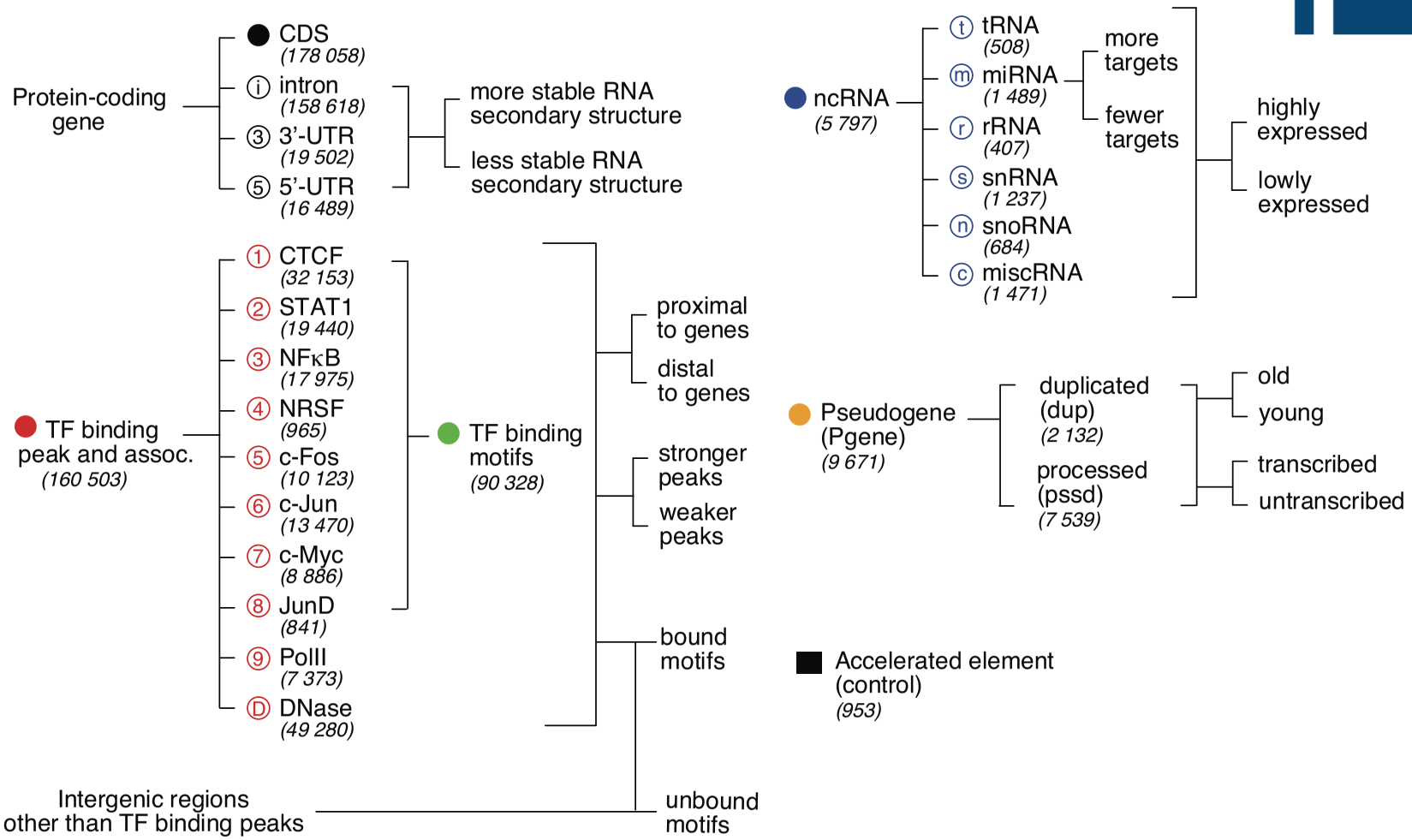
The 1000 Genomes Project low coverage pilot, 2010

- A comprehensive catalog of genomic variations (genome wide, many individuals in multiple populations: 60 CEU; 59 YRI; 60 CHBJPT).
- A full spectrum of genomic variations, including single nucleotide polymorphisms (SNPs), short insertions and deletions (indels), and structural variations (SVs).

  - The latter two types have not previously been well studied of functional significance.

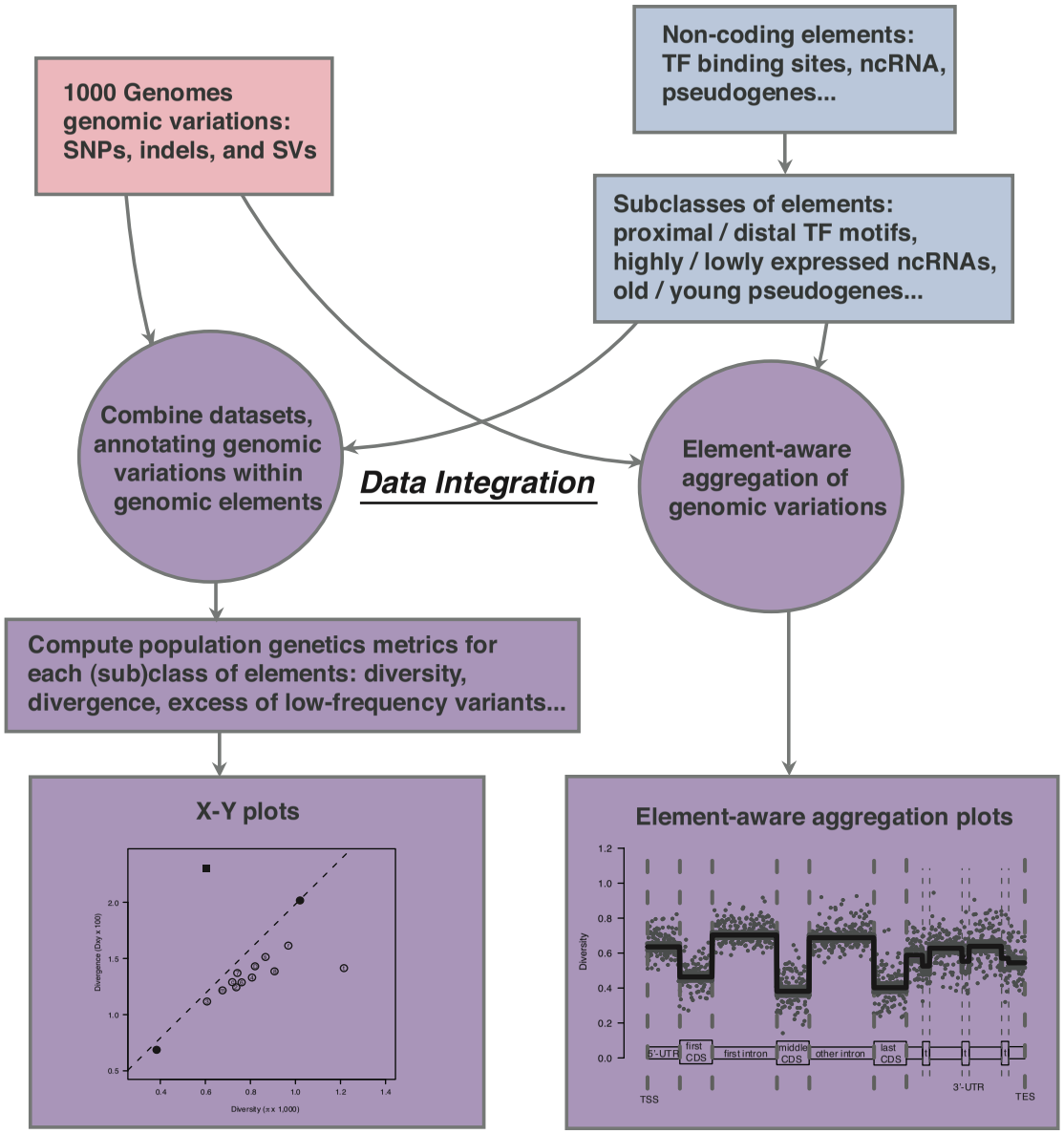
  - Improved SV detection in terms of number, size-range and breakpoint-precision.

Overview of the genomic elements surveyed



Mu et al., *NAR* (2011)

ncVAR - framework for an integrative analysis of genomic variations in non-coding elements



Mu et al., *NAR* (2011)

Results

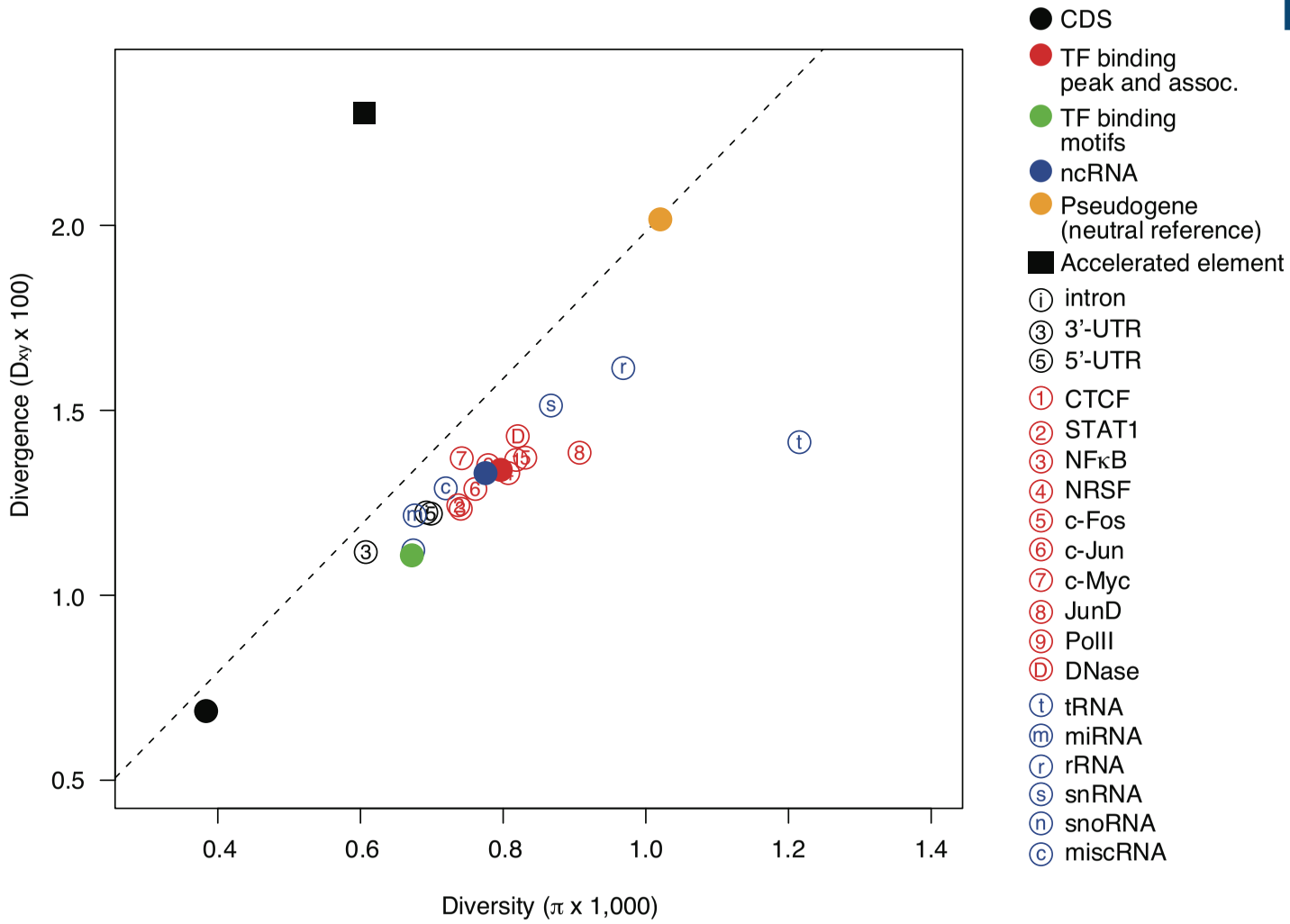
- Comparing classes of elements
- Comparing subclasses within an element class
- Intra-element differences of a given element
- More binding reactions introduce more selective constraints

Results

- **Comparing classes of elements**
- Comparing subclasses within an element class
- Intra-element differences of a given element
- More binding reactions introduce more selective constraints

Diversity vs. divergence

- Non-coding elements are under selective constraints for SNPs



Mu et al., *NAR* (2011)

McDonald–Kreitman test

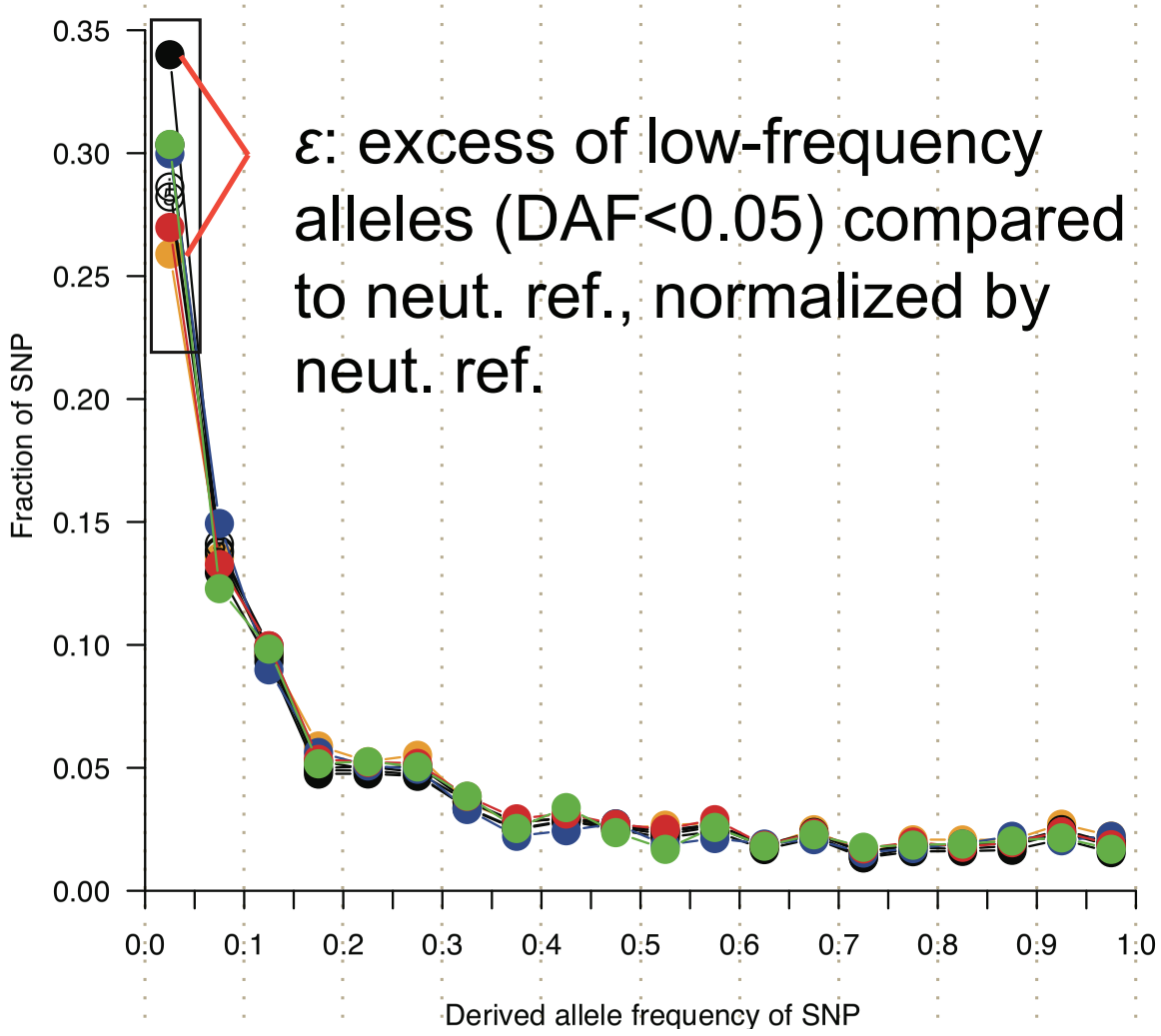
Element	SNP diversity ($\pi \times 1000$)	Divergence ($D_{xy} \times 100$)	Polymorphism (P)	Number of fixed differences (D)	Neutrality index (NI)	McDonald–Kreitman test P-value
Pseudogene	1.02	2.02	46 122	206 922	1.00	–
CDS	0.38	0.69	49 636	181 193	1.23	2.38E-179
Intron	0.69	1.22	2 244 675	8 610702	1.17	3.03E-205
3′UTR	0.61	1.12	60 129	232 581	1.16	3.53E-103
5′UTR	0.70	1.22	293 916	1 116 579	1.18	3.78E-202
TF peak	0.80	1.34	111 140	417 405	1.19	5.30E-186
TF motif	0.67	1.11	2409	8545	1.26	2.13E-22
ncRNA	0.78	1.33	2254	8023	1.26	1.42E-20
Accelerated element	0.60	2.30	701	5656	0.56	5.07E-55

NI = $\frac{P_i/D_i}{P_n/D_n}$, where *i* is the region to study, and *n* is the neutral reference.

NI > 1: negative selection
NI < 1: positive selection
NI = 1: neutral evolution

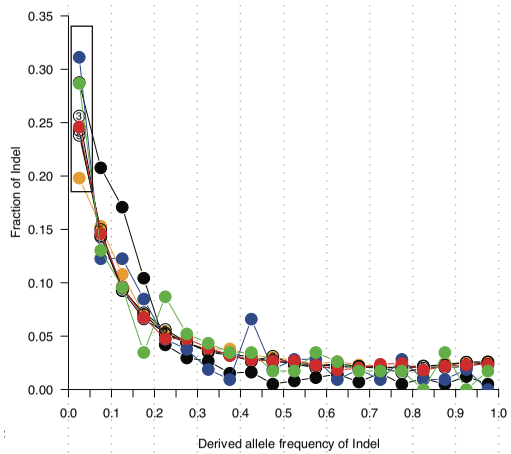
Mu et al., *NAR* (2011)

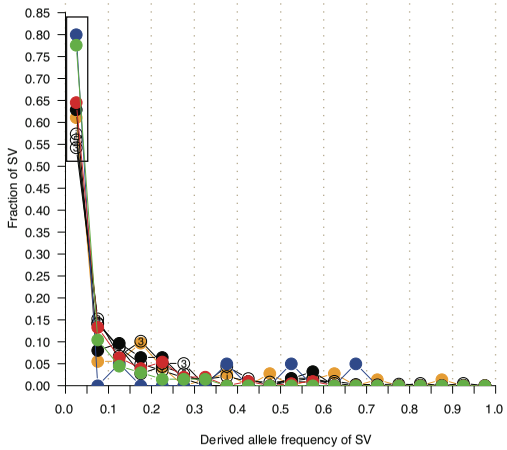
Variant allele frequency spectra



ε: excess of low-frequency alleles (DAF<0.05) compared to neut. ref., normalized by neut. ref.

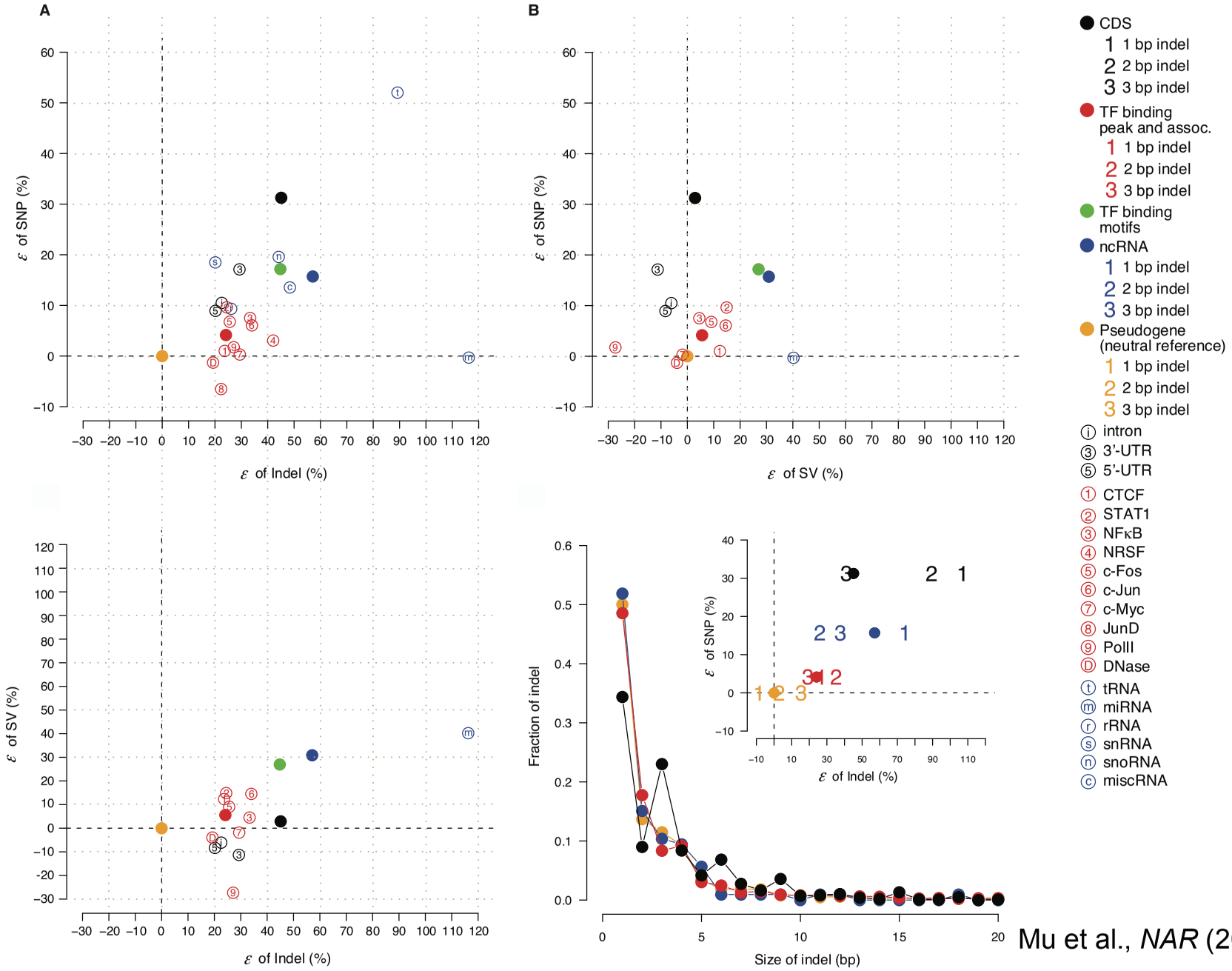
CDS
TF-binding peak and assoc.
TF-binding motifs
ncRNA
Pseudogene (neutral reference)
ⓘ intron
③ 3'UTR
⑤ 5'UTR





Mu et al., *NAR* (2011)

ε: excess of low frequency alleles (DAF<0.05) rel. to neut. ref.

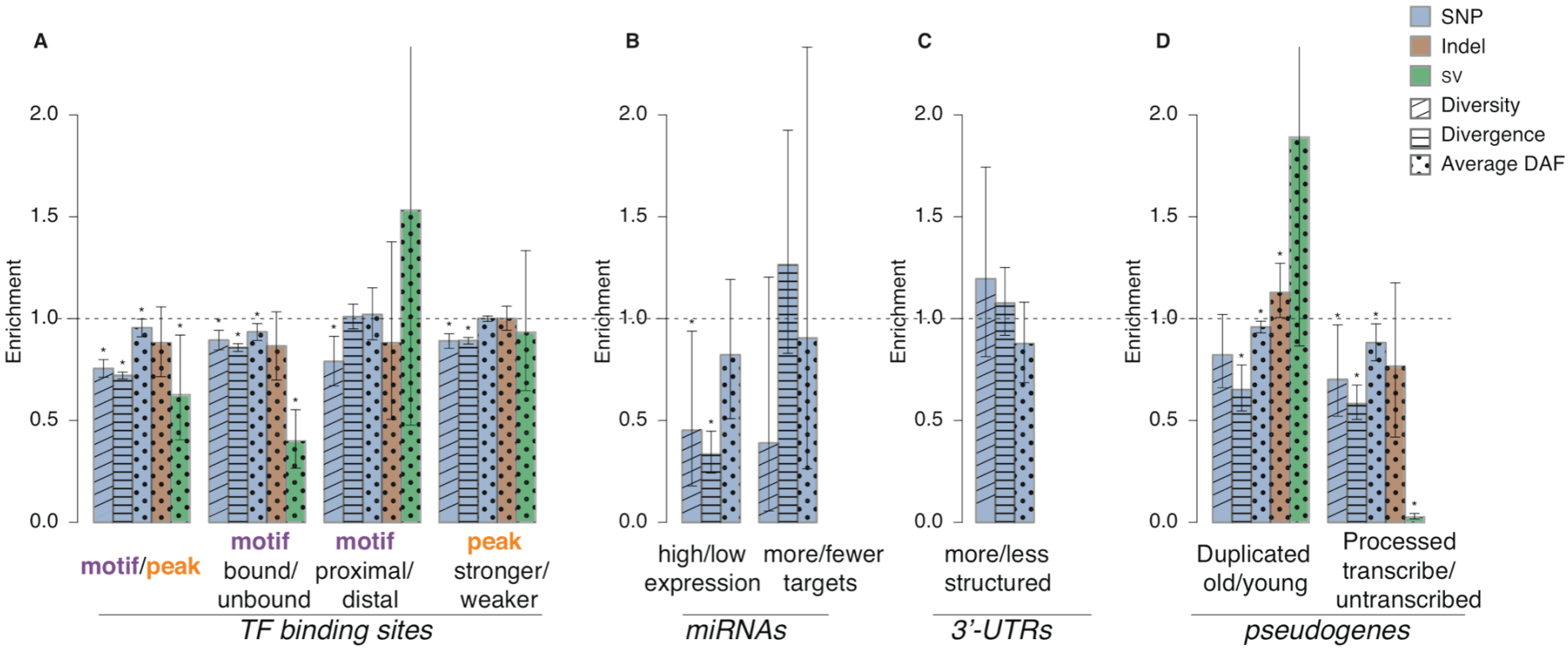


Mu et al., *NAR* (2011)

Results

- Comparing classes of elements
- **Comparing subclasses within an element class**
- Intra-element differences of a given element
- More binding reactions introduce more selective constraints

Comparison between subgroups of elements



Mu et al., *NAR* (2011)

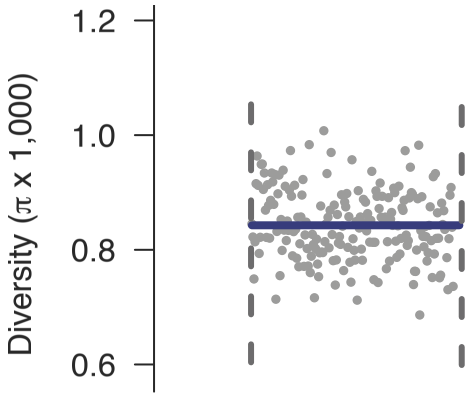
Results

- Comparing classes of elements
- Comparing subclasses within an element class
- **Intra-element differences of a given element**
- More binding reactions introduce more selective constraints

Aggregation plot for SNP and indel diversity

Aggregation plot for an element with *M* sequences (e.g. *M* ~20,000 for genes).

- Divide each sequence of the element into *N* bins (e.g. *N* = 200)
- Calculate nucleotide diversity within each bin of each sequence
- Average diversity from each bin over all the sequences => a data point (grey) on the aggregation plot
- Average all the data points => aggregation mean (blue line)



Mu et al., *NAR* (2011)

Assess the confidence interval of aggregation means

Simple bootstrapping

- Assume the sequences of the element are independent.
- Randomly resample the same number of sequences for n = 10,000 times, with replacement, from the original set of sequences.
- Calculate average from all the resampled aggregation means

=> bootstrapping mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, and

=> bootstrapping standard deviation (SD). $S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$,

where x_1, x_2,…, x_n are the aggregation mean from resampled datasets 1, 2, ..., n.

- 95% confidence interval is calculated from $\bar{x} \pm 1.96S$.

Mu et al., *NAR* (2011)

Assess the confidence interval of aggregation means

Block bootstrapping

- Linkage disequilibrium (LD) => nucleotide diversity for genomic sequences that are sufficiently close to each other are dependent.
- Simple bootstrapping underestimates the SD.
- Randomly resample *n* = ~1M blocks from the genome.

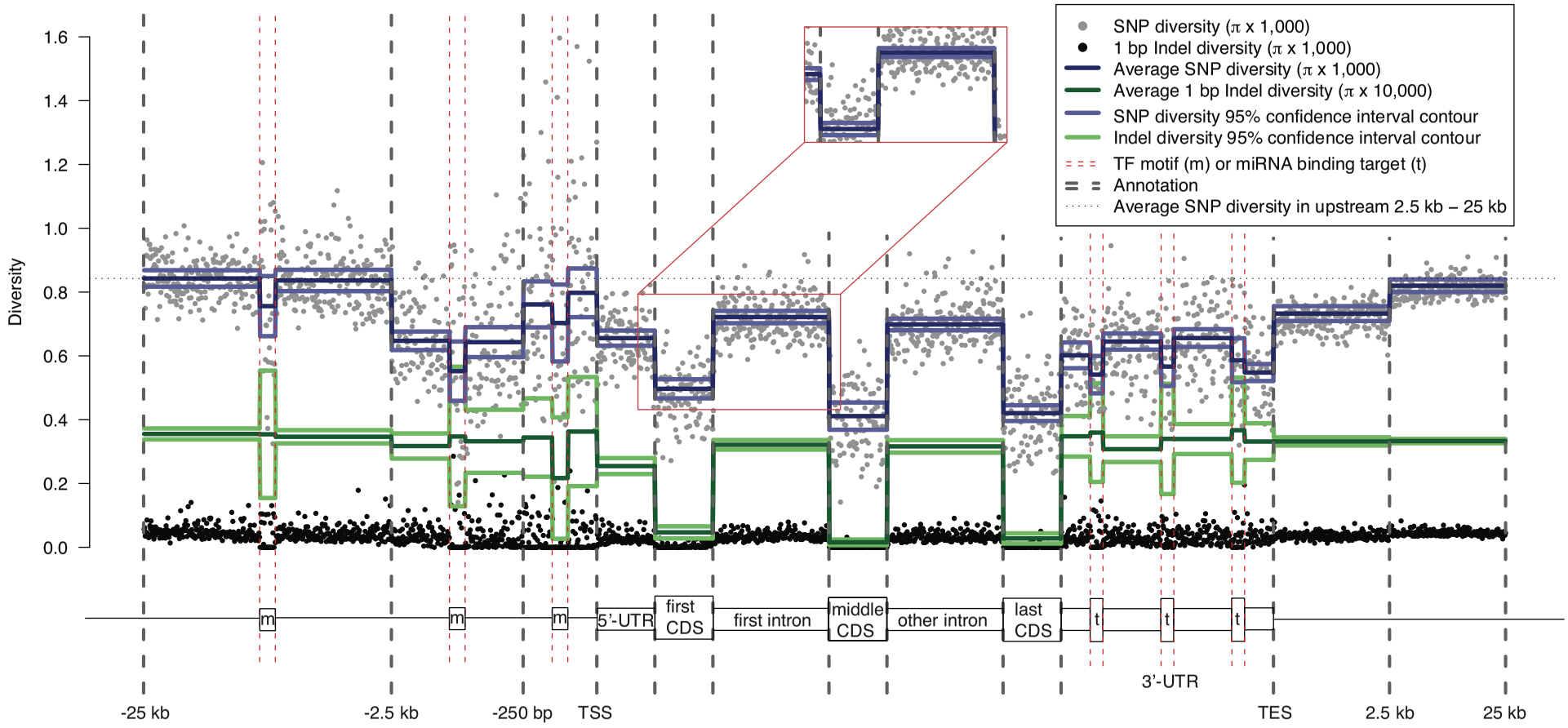
  => bootstrapping mean $\bar{x} = \frac{\sum_{i=1}^{n} w_i x}{\sum_{i=1}^{n} w_i}$

  => unbiased estimator of SD for weighted samples of blocks $S' = \sqrt{\frac{\sum_{i=1}^{n} w_i}{\left(\sum_{i=1}^{n} w_i\right)^2 - \sum_{i=1}^{n} w_i^2} \times \sum_{i=1}^{n} w_i (x_i - \bar{x})^2},$

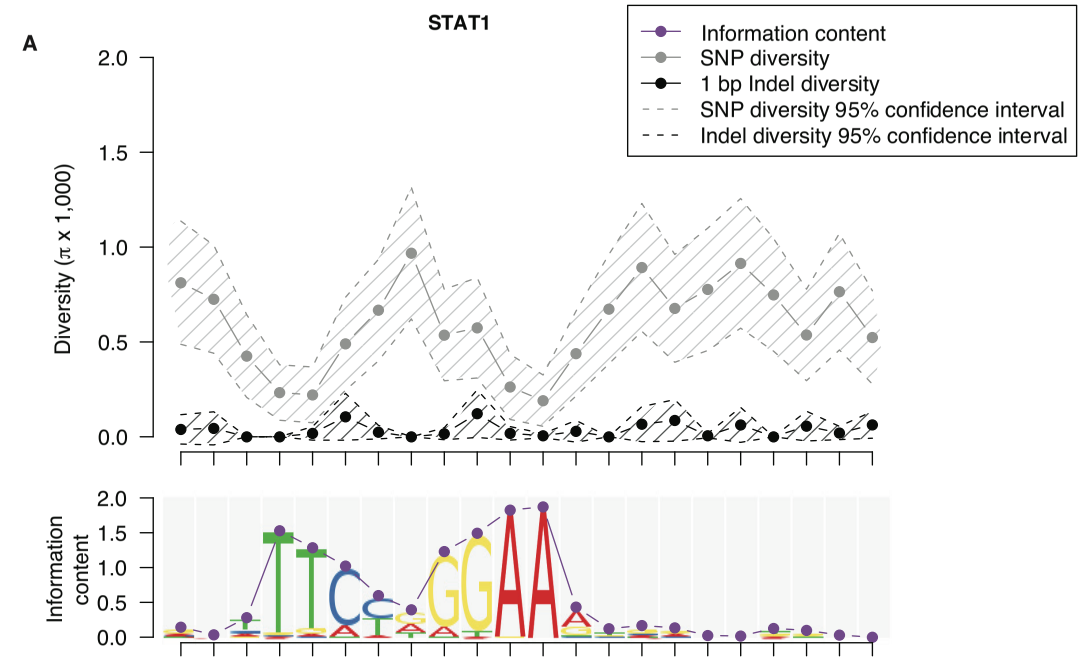
  where $x_1, x_2, \ldots, x_n$ are the aggregation mean from resampled blocks 1, 2, ..., *n*, and $w_1, w_2, \ldots, w_n$ are the number of sequences of the element within in blocks 1, 2, ..., *n*.
- *S'* is then renormalized according to the effective genome size *G* and the block size *L* to obtain the bootstrapping SD for the whole genome (*S*): $S = \frac{S'}{\sqrt{G/L}}$
- 95% confidence interval is calculated from $\bar{x} \pm 1.96S$.

Mu et al., *NAR* (2011)

Aggregation plot for SNP and indel diversity in coding genes and surrounding regions

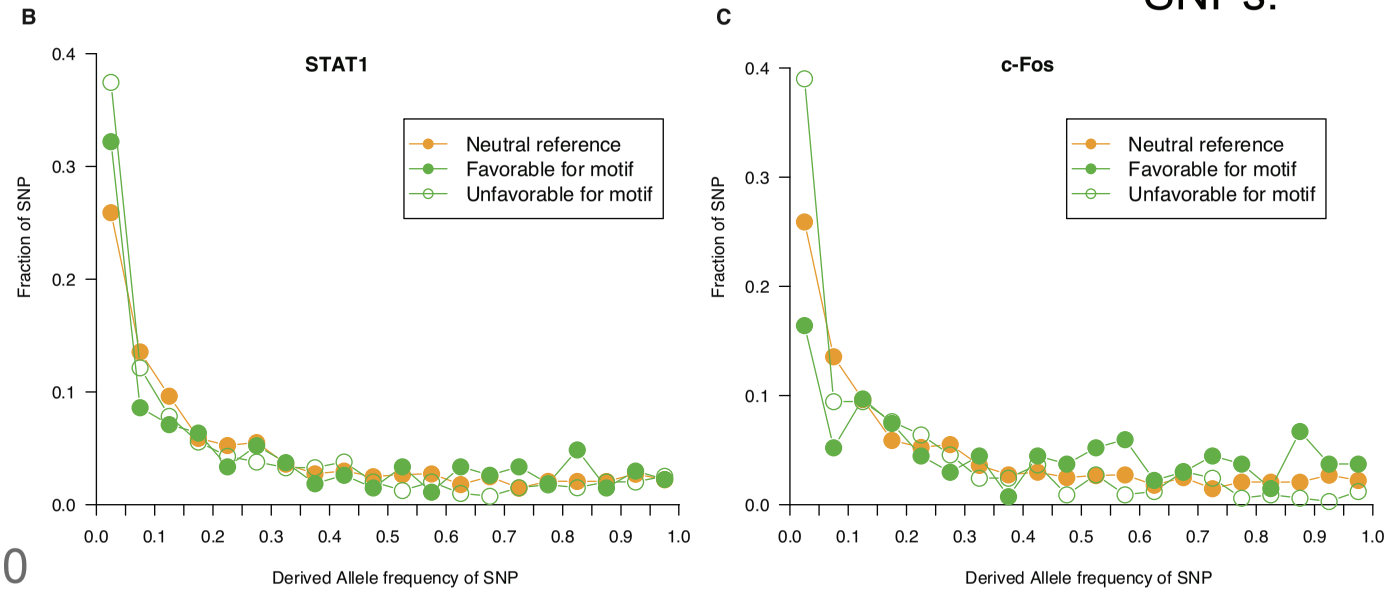


Mu et al., *NAR* (2011)

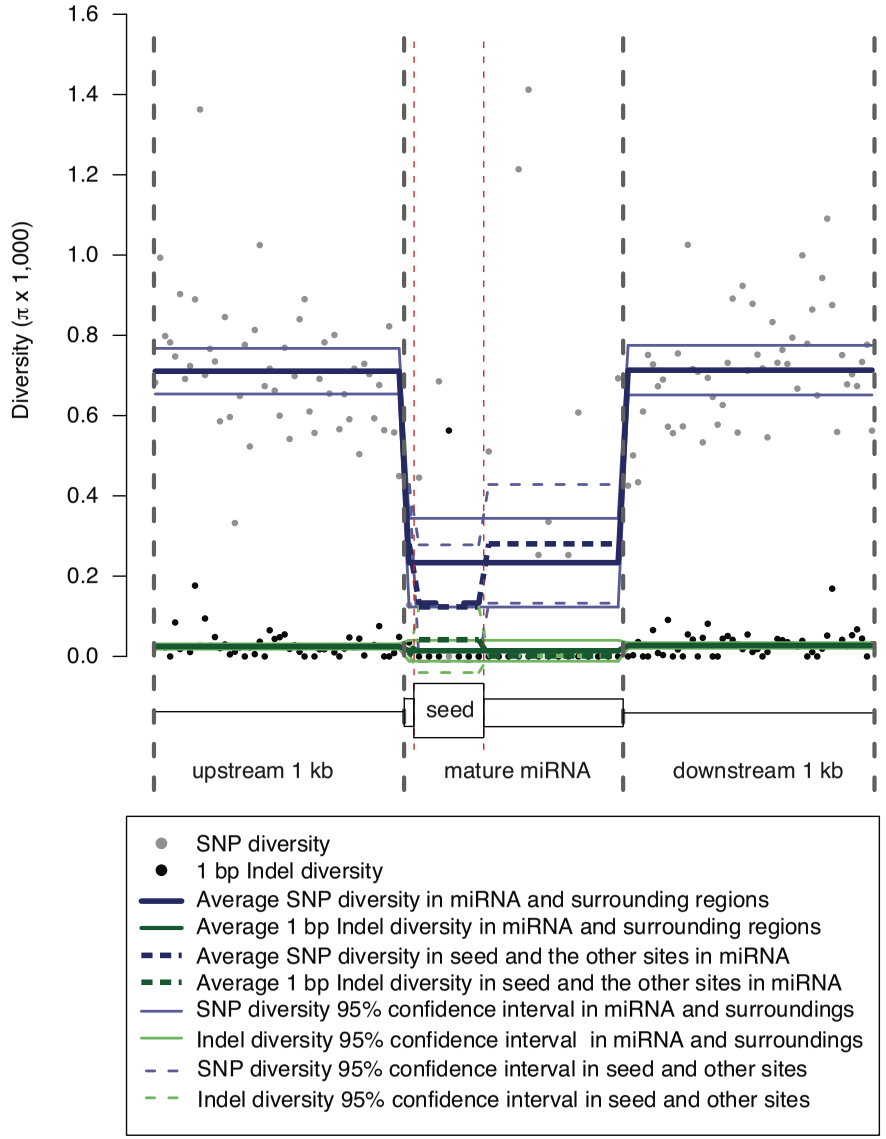


TF-binding motifs:

- SNP diversity reversely correlates with information content.
- SNPs unfavorable for motifs are under more selective constraints than favorable SNPs.



Mu et al., *NAR* (2011)



miRNAs:

- miRNAs (particularly the seed regions) are under stronger selective pressure than their surroundings.

Mu et al., *NAR* (2011)

SVs interacting with genomic elements (enrichment w.r.t. randomized control)

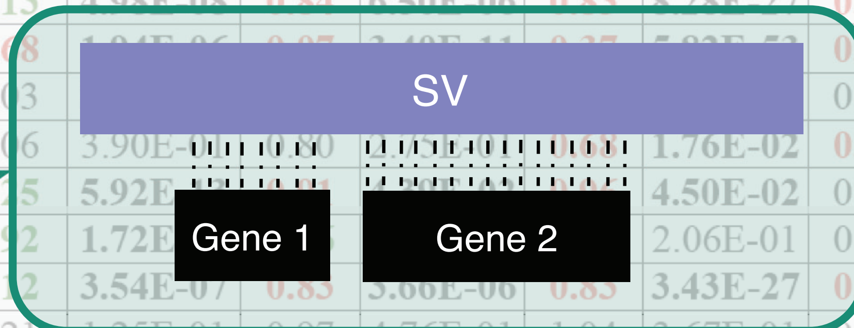
Element	All SVs		NAHR		VNTR		NHR		TEI	
	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value
Gene	0.90	**8.68E-20**	1.13	**4.98E-08**	0.84	**6.50E-06**	0.83	**8.28E-27**	0.87	**6.96E-09**
CDS	0.37	**8.72E-85**	0.68	**1.94E-06**	0.07	**3.40E-11**	0.37	**5.82E-53**	0.04	**3.47E-24**
5'UTR Only	0.96	2.17E-01	1.03	3.86E-01	0.83	1.44E-01	0.97	3.45E-01	0.95	3.11E-01
3'UTR Only	0.72	**3.47E-03**	1.06	3.90E-01	0.80	2.75E-01	0.68	**1.76E-02**	0.46	**6.16E-03**
Intron Only	1.02	7.60E-02	1.25	**5.92E-13**	0.91	**4.39E-02**	0.96	**4.50E-02**	0.99	3.57E-01
Whole Gene	1.41	**8.96E-03**	1.92	**1.72E-03**	2.76	**2.89E-02**	1.18	2.06E-01	0.00	2.34E-01
Partial Gene	0.90	**1.06E-20**	1.12	**3.54E-07**	0.83	**3.66E-06**	0.83	**3.43E-27**	0.87	**7.75E-09**
ncRNA	1.08	2.06E-01	1.21	1.25E-01	0.97	4.76E-01	1.04	3.67E-01	0.76	3.13E-01
Whole ncRNA	1.03	3.94E-01	1.18	1.64E-01	0.76	3.41E-01	0.99	4.83E-01	0.37	1.51E-01
Partial ncRNA	1.83	**2.58E-02**	1.73	2.17E-01	2.10	2.26E-01	1.96	6.28E-02	1.59	2.54E-01
Motif	0.73	**3.74E-13**	0.87	**3.86E-2**	1.44	**5.70E-03**	0.71	**5.91E-10**	0.13	**8.45E-10**
Whole Motif	0.73	**5.58E-13**	0.90	7.35E-02	1.39	**1.48E-02**	0.71	**2.52E-10**	0.14	**4.11E-09**
Partial Motif	0.75	1.74E-01	0.00	**4.66E-02**	2.48	5.03E-02	1.11	3.93E-01	0.00	**4.10E-02**
Pseudogene	1.24	**1.11E-05**	1.56	**3.37E-07**	1.54	**1.73E-02**	1.24	**6.94E-04**	0.50	**3.58E-03**
Whole Pseudogene	1.51	**1.15E-12**	1.95	**3.98E-13**	2.50	**1.22E-04**	1.33	**1.44E-04**	0.51	1.63E-01
Partial Pseudogene	0.93	2.39E-01	0.97	4.40E-01	1.05	4.37E-01	1.10	2.16E-01	0.50	**6.26E-03**

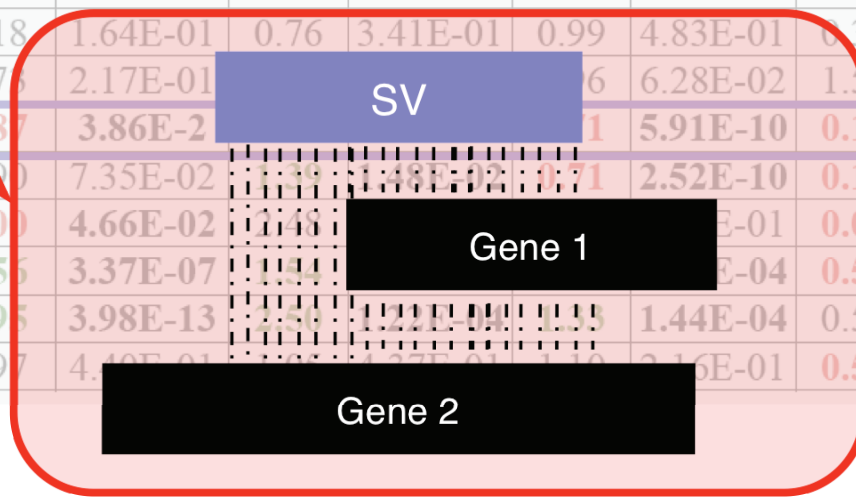
- SVs are shuffled in the whole genome.
- Significant *P*-values (<0.05) in black and bold
- Significant enrichments in green; Significant depletions in red.

Mu et al., *NAR* (2011)

SVs interacting with genomic elements (enrichment w.r.t. randomized control)

Element	All SVs		NAHR		VNTR		NHR		TEI	
	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value
Gene	0.90	8.68E-20	1.13	4.98E-08	0.84	6.50E-06	0.83	8.28E-27	0.87	6.96E-09
CDS	0.37	8.72E-85	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3.47E-24
5'UTR Only	0.96	2.17E-01	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3.11E-01
3'UTR Only	0.72	3.47E-03	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.76E-02	0.46	6.16E-03
Intron Only	1.02	7.60E-02	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.50E-02	0.99	3.57E-01
Whole Gene	1.41	8.96E-03	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.06E-01	[illegible]	2.34E-01
Partial Gene	0.90	1.06E-20	[illegible]	3.54E-07	0.83	3.66E-06	0.83	3.43E-27	[illegible]	7.75E-09
ncRNA	1.08	[illegible]	1.21	1.25E-01	0.97	4.76E-01	1.04	3.67E-01	0.76	3.13E-01
Whole ncRNA	1.03	[illegible]	1.18	1.64E-01	0.76	3.41E-01	0.99	4.83E-01	[illegible]	1.51E-01
Partial ncRNA	1.83	[illegible]	[illegible]	2.17E-01	[illegible]	[illegible]	[illegible]	6.28E-02	[illegible]	2.54E-01
Motif	0.73	3.74E-13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.91E-10	[illegible]	8.45E-10
Whole Motif	0.73	5.58E-13	[illegible]	7.35E-02	[illegible]	[illegible]	[illegible]	2.52E-10	[illegible]	4.11E-09
Partial Motif	0.75	1.74E-01	[illegible]	4.66E-02	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.10E-02
Pseudogene	1.24	1.11E-05	[illegible]	3.37E-07	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3.58E-03
Whole Pseudogene	1.51	1.15E-12	[illegible]	3.98E-13	[illegible]	[illegible]	[illegible]	1.44E-04	[illegible]	1.63E-01
Partial Pseudogene	0.93	2.39E-01	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6.26E-03



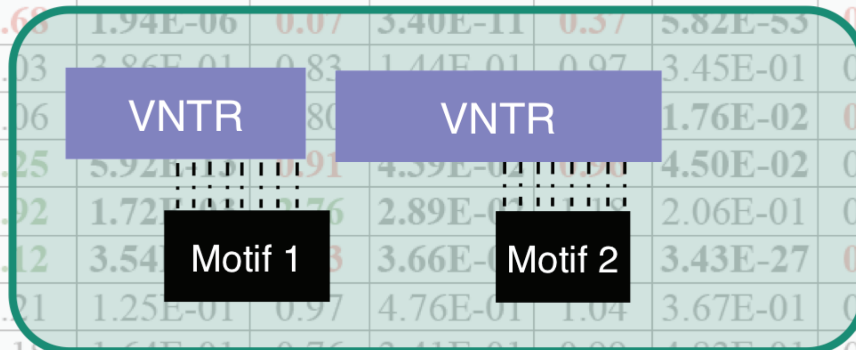


- SVs are shuffled in the whole genome.
- Significant *P*-values (<0.05) in black and bold
- Significant enrichments in green; Significant depletions in red.

Mu et al., *NAR* (2011)

SVs interacting with genomic elements (enrichment w.r.t. randomized control)

Element	All SVs		NAHR		VNTR		NHR		TEI	
	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value
Gene	0.90	**8.68E-20**	1.13	**4.98E-08**	0.84	**6.50E-06**	0.83	**8.28E-27**	0.87	**6.96E-09**
CDS	0.37	**8.72E-85**	0.68	**1.94E-06**	0.07	**3.40E-11**	0.37	**5.82E-53**	0.04	**3.47E-24**
5'UTR Only	0.96	2.17E-01	1.03	3.86E-01	0.83	1.44E-01	0.97	3.45E-01	0.95	3.11E-01
3'UTR Only	0.72	**3.47E-03**	1.06	[illegible]	[illegible]	[illegible]	[illegible]	**1.76E-02**	0.46	**6.16E-03**
Intron Only	1.02	7.60E-02	1.25	[illegible]	[illegible]	[illegible]	[illegible]	**4.50E-02**	0.99	3.57E-01
Whole Gene	1.41	**8.96E-03**	1.92	[illegible]	[illegible]	[illegible]	[illegible]	2.06E-01	0.00	2.34E-01
Partial Gene	0.90	**1.06E-20**	1.12	[illegible]	[illegible]	[illegible]	[illegible]	**3.43E-27**	0.87	**7.75E-09**
ncRNA	1.08	2.06E-01	1.21	1.25E-01	0.97	4.76E-01	1.04	3.67E-01	0.76	3.13E-01
Whole ncRNA	1.03	3.94E-01	1.18	1.64E-01	0.76	3.41E-01	0.99	4.83E-01	0.37	1.51E-01
Partial ncRNA	1.83	**2.58E-02**	1.73	2.17E-01	[illegible]	2.26E-01	1.96	6.28E-02	1.59	2.54E-01
Motif	0.73	**3.74E-13**	0.87	**3.86E-2**	1.44	**5.70E-03**	0.71	**5.91E-10**	0.13	**8.45E-10**
Whole Motif	0.73	**5.58E-13**	0.90	7.35E-02	1.39	**1.48E-02**	0.71	**2.52E-10**	0.14	**4.11E-09**
Partial Motif	0.75	1.74E-01	0.00	**4.66E-02**	2.48	5.03E-02	1.11	3.93E-01	0.00	**4.10E-02**
Pseudogene	1.24	**1.11E-05**	1.56	**3.37E-07**	1.54	**1.73E-02**	1.24	**6.94E-04**	0.50	**3.58E-03**
Whole Pseudogene	1.51	**1.15E-12**	1.95	**3.98E-13**	2.50	**1.22E-04**	1.33	**1.44E-04**	0.51	1.63E-01
Partial Pseudogene	0.93	2.39E-01	0.97	4.40E-01	1.05	4.37E-01	1.10	2.16E-01	0.50	**6.26E-03**

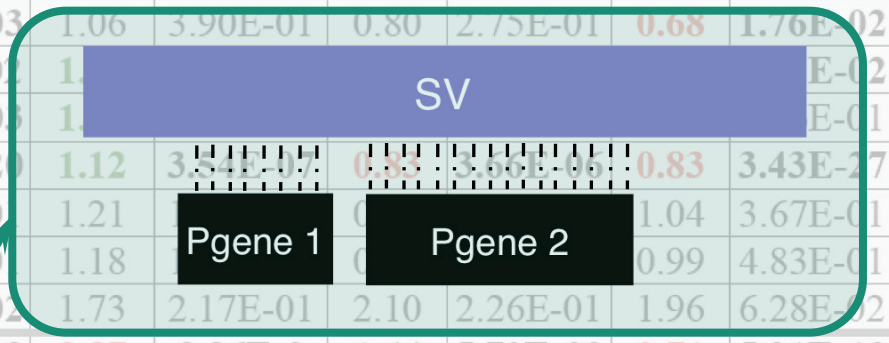


- SVs are shuffled in the whole genome.
- Significant *P*-values (<0.05) in black and bold
- Significant enrichments in green; Significant depletions in red.

Mu et al., *NAR* (2011)

SVs interacting with genomic elements (enrichment w.r.t. randomized control)

Element	All SVs		NAHR		VNTR		NHR		TEI	
	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value
Gene	0.90	8.68E-20	1.13	4.98E-08	0.84	6.50E-06	0.83	8.28E-27	0.87	6.96E-09
CDS	0.37	8.72E-85	0.68	1.94E-06	0.07	3.40E-11	0.37	5.82E-53	0.04	3.47E-24
5'UTR Only	0.96	2.17E-01	1.03	3.86E-01	0.83	1.44E-01	0.97	3.45E-01	0.95	3.11E-01
3'UTR Only	0.72	3.47E-03	1.06	3.90E-01	0.80	2.75E-01	0.68	1.76E-02	0.46	6.16E-03
Intron Only	1.02	7.60E-02	1.				E-02		0.99	3.57E-01
Whole Gene	1.41	8.96E-03	1.				E-01		0.00	2.34E-01
Partial Gene	0.90	1.06E-20	1.12	3.54E-07	0.83	3.66E-06	0.83	3.43E-27	0.87	7.75E-09
ncRNA	1.08	2.06E-0	1.21				1.04	3.67E-01	0.76	3.13E-01
Whole ncRNA	1.03	3.94E-0	1.18				0.99	4.83E-01	0.37	1.51E-01
Partial ncRNA	1.83	2.58E-02	1.73	2.17E-01	2.10	2.26E-01	1.96	6.28E-02	1.59	2.54E-01
Motif	0.73	3.7 E-13	0.87	3.86E-2	1.44	5.70E-03	0.71	5.91E-10	0.13	8.45E-10
Whole Motif	0.73	5.68E-13	0.90	7.35E-02	1.39	1.48E-02	0.71	2.52E-10	0.14	4.11E-09
Partial Motif	0.75	.74E-01	0.00	4.66E-02	2.48	5.03E-02	1.11	3.93E-01	0.00	4.10E-02
Pseudogene	1.24	1.11E-05	1.56	3.37E-07	1.54	1.73E-02	1.24	6.94E-04	0.50	3.58E-03
Whole Pseudogene	1.51	1.15E-12	1.95	3.98E-13	2.50	1.22E-04	1.33	1.44E-04	0.51	1.63E-01
Partial Pseudogene	0.93	2.39E-01	0.97	4.40E-01	1.05	4.37E-01	1.10	2.16E-01	0.50	6.26E-03



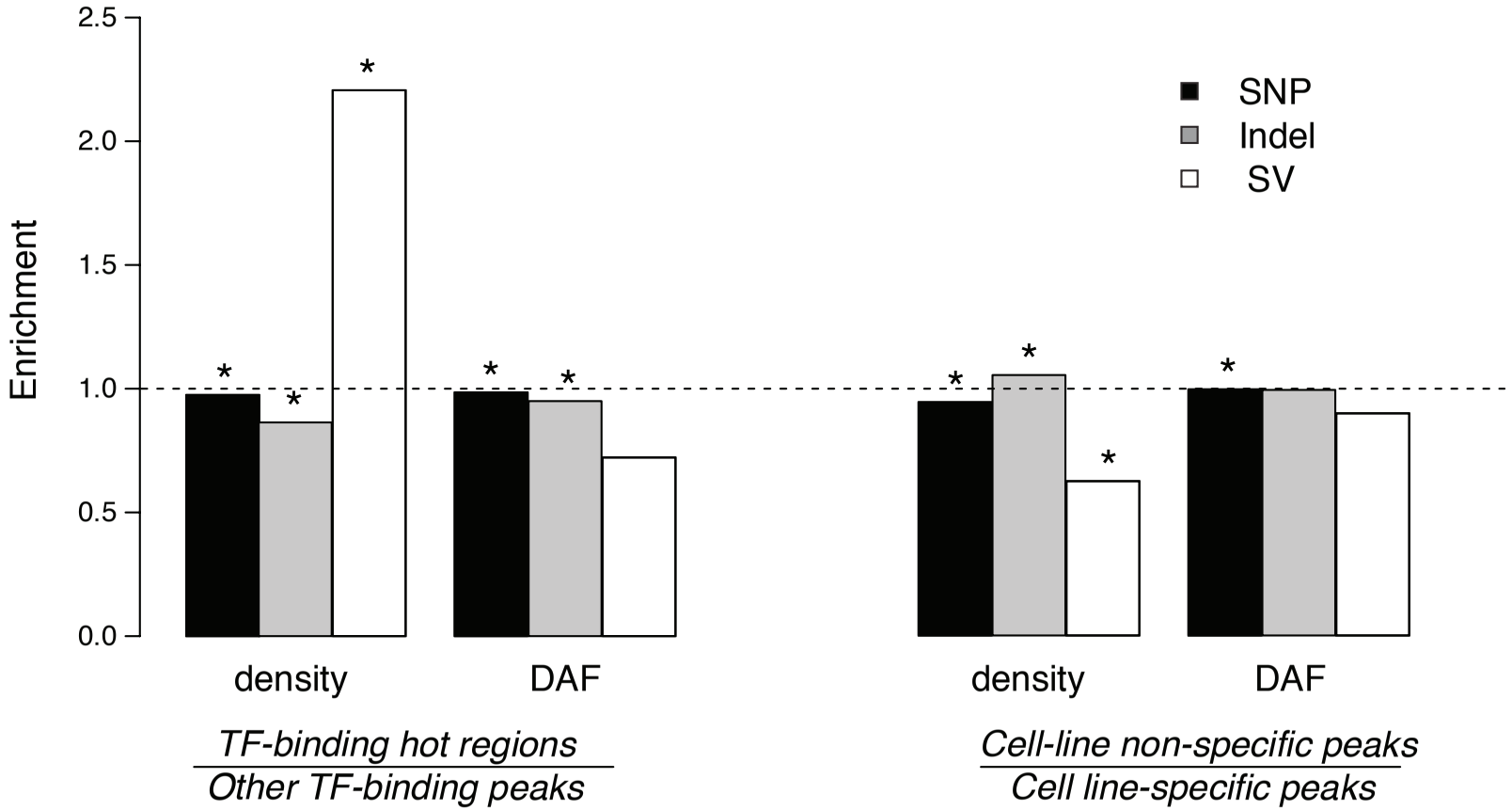
- SVs are shuffled in the whole genome.
- Significant *P*-values (<0.05) in black and bold
- Significant enrichments in green; Significant depletions in red.

Mu et al., *NAR* (2011)

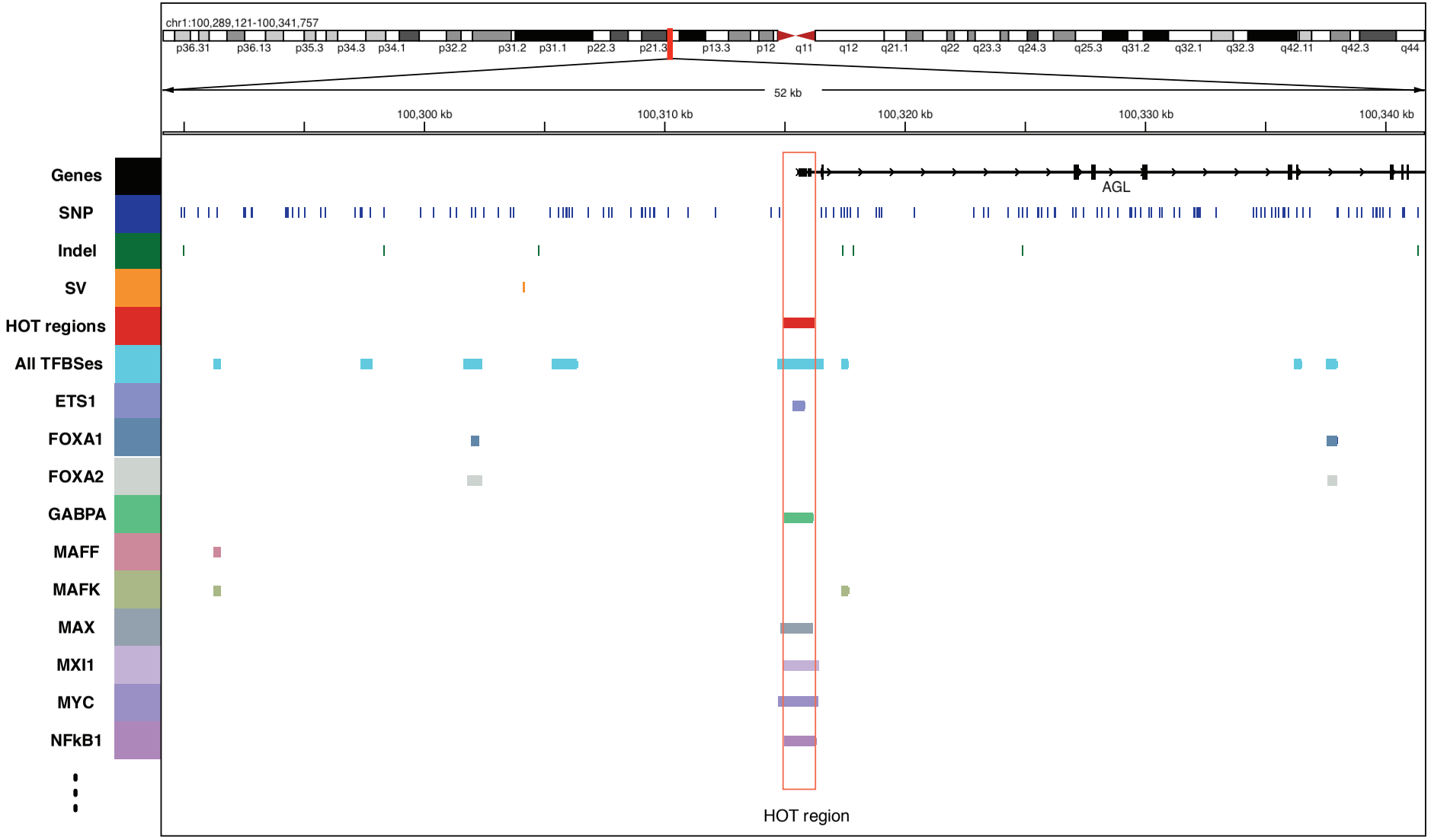
Results

- Comparing classes of elements
- Comparing subclasses within an element class
- Intra-element differences of a given element
- **More binding reactions introduce more selective constraints**

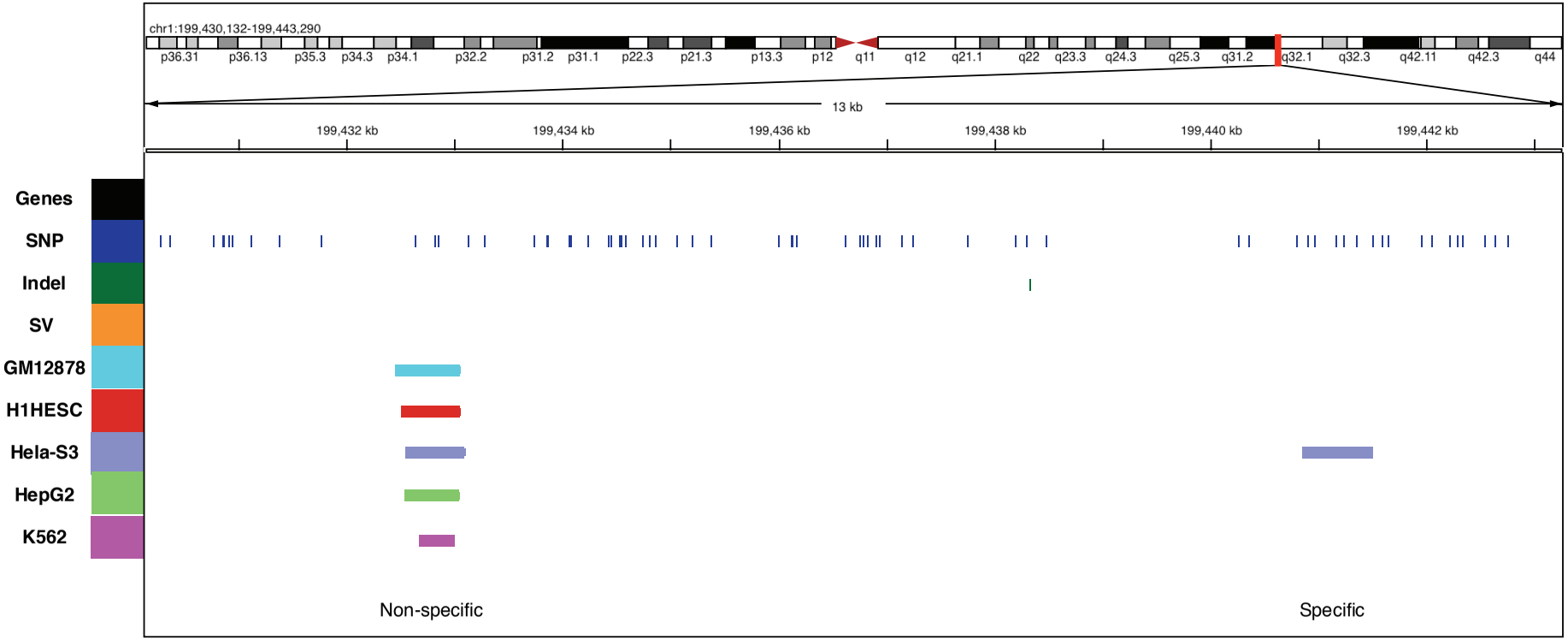
More binding reactions tend to introduce more selective constraints



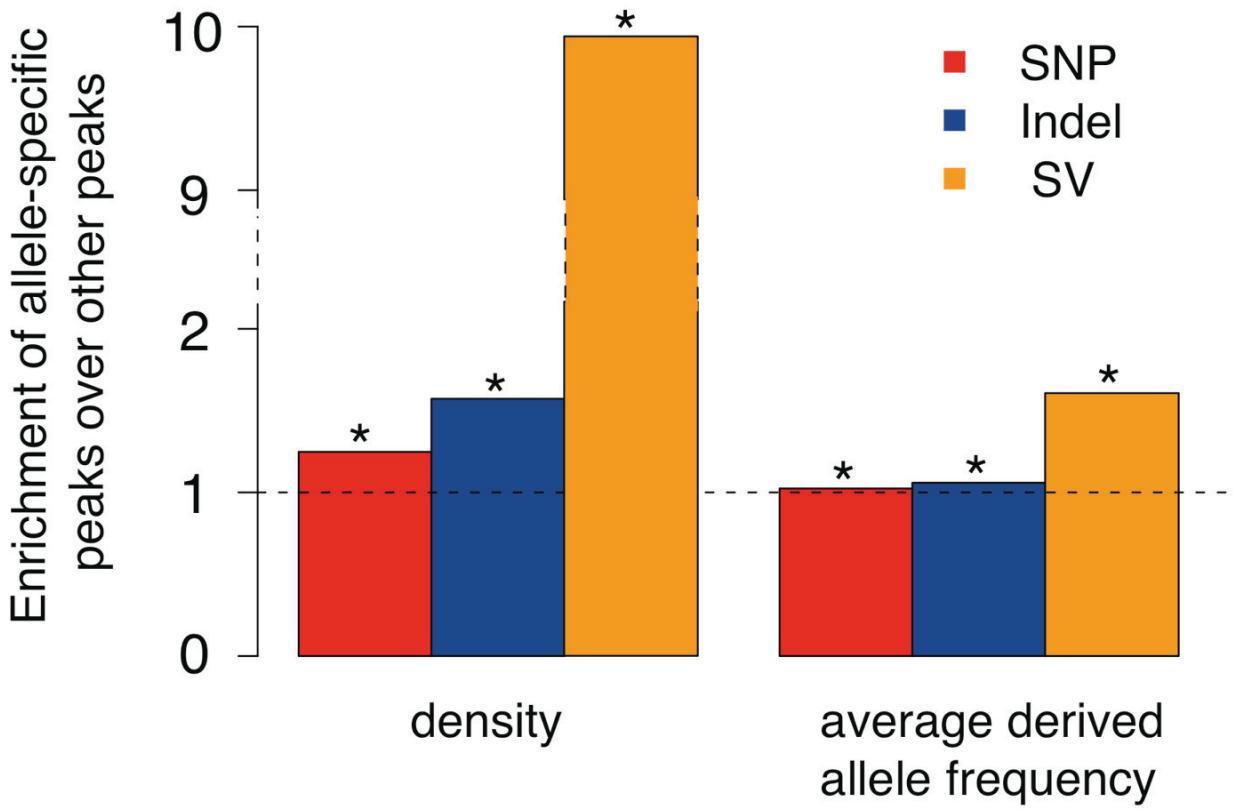
TF-binding HOT regions



Cell-line specific TF-binding sites



Allele-specific binding sites are under less selective constraints



Allele-specific binding sites are under less selective constraints

ASB	Low DAF (<0.05)	Medium DAF (0.05-0.5)	High DAF (>0.5)
Allelic score (median ± MAD)	0.99947±0.00053	0.99973±0.00027	0.99982±0.00018
Number of SNPs	143	782	398

ASE	Low DAF (<0.05)	Medium DAF (0.05-0.5)	High DAF (>0.5)
Allelic score (median ± MAD)	0.62±0.34	0.70±0.34	0.74±0.31
Number of SNPs	56	393	177

TFs with more target genes tend to have higher degree of “allelicity”

- *TF Allelicity* is used to quantify *the degree of allele-specific behavior of each TF*.
- We define the allelicity of a TF as the fraction of SNPs that exhibit ASB out of all the SNPs that may potentially exhibit ASB (for heterozygous SNPs within binding regions of a given TF that pass a read depth filter).
- We find that TFs with higher degrees of allelicity tend to have more regulatory target genes (spearman correlation ρ = 0.28, P-valule = 0.044).

Conclusion

- Developed a framework, ncVAR, for the integrative analysis of three types of genomic variations (SNPs, indels, and SVs) in a number of different types of non-coding elements (TF-binding sites, ncRNAs, pseudogenes, etc.).
- Developed an element-aware aggregation method using block bootstrapping procedures.
- Characterized and compared selection pressure between different classes of elements, subclasses within an element class, and intra-element differences of a given element.
- More binding reactions tend to introduce more selective constraints.

Future Directions

- Higher resolution analyses using the production phase data of the 1000 Genomes Project.
- Analyses using the ENCODE full production phase data: binding sites for >120 TFs and chromatin modifiers in >40 cell types, ncRNAs, DNase I hypersensitive sites in >80 tissue types, chromatin modification marks, DNA methylation modifications, etc.

Acknowledgement

Yale University

Arif Harmanci
Jieming Chen
Joel Rozowsky
Robert Bjornson
Zhi Lu
Yong Kong
Hugo Lam
Mark Gerstein

Other Gerstein lab members

UC Berkeley

Ben Brown
Peter Bickel
Haiyan Huang

The ENCODE Consortium
The 1000 Genomes Project