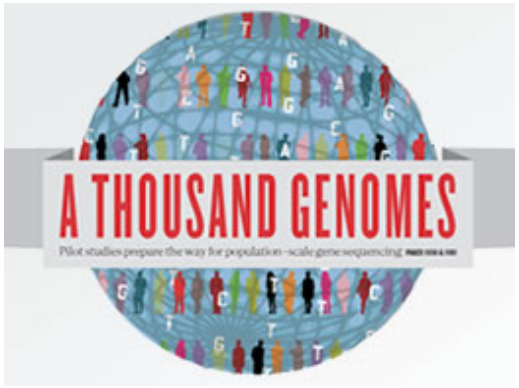
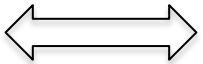
1000 Genomes Data Tutorial: Functional analysis

Functional Interpretation Group, 1000 Genomes Consortium

Ekta Khurana, PhD
Yale University
New Haven, CT

Functional annotation of sequence variants from 1,092 genomes

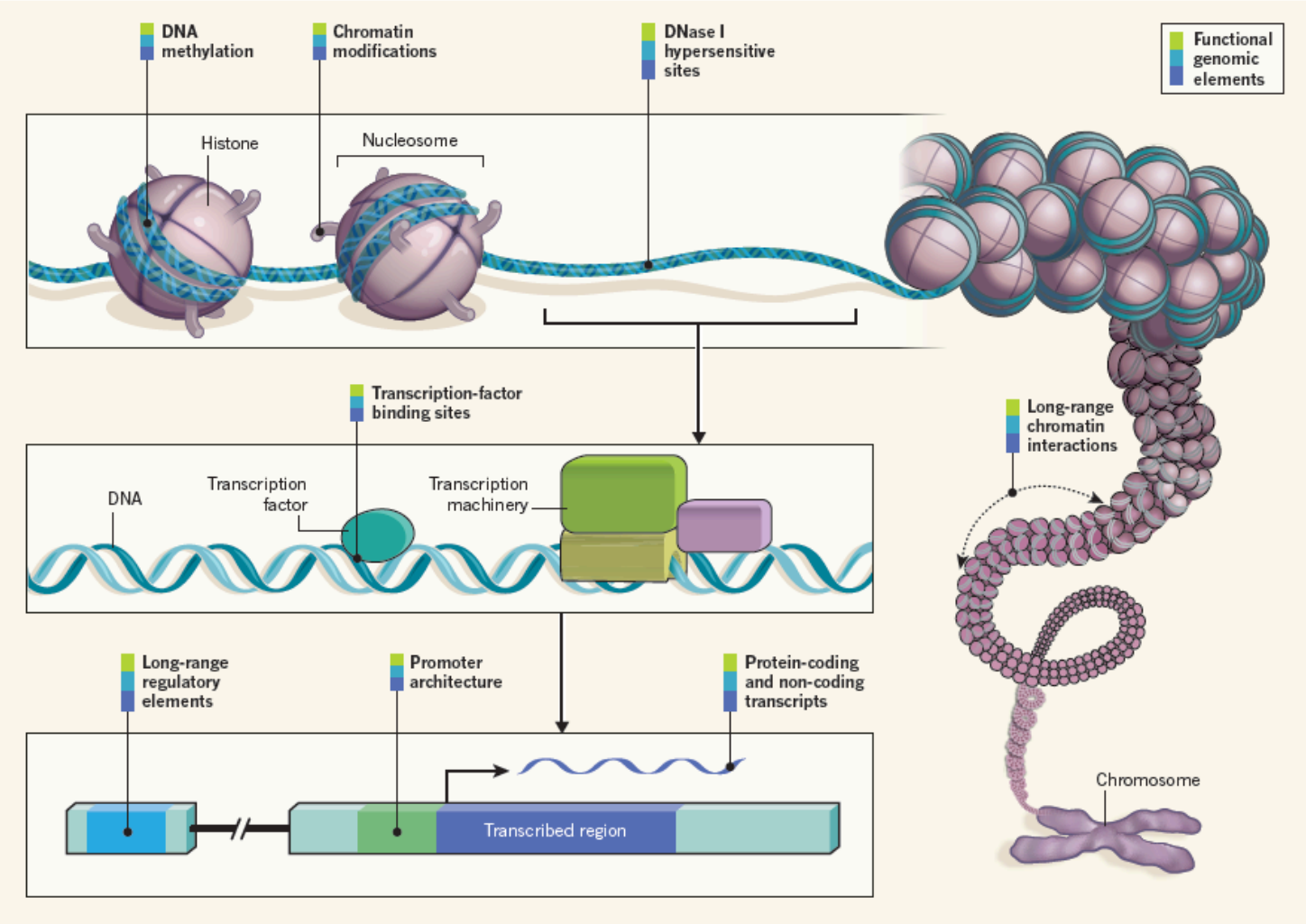




Many studies have now linked disease SNPs to regulatory regions

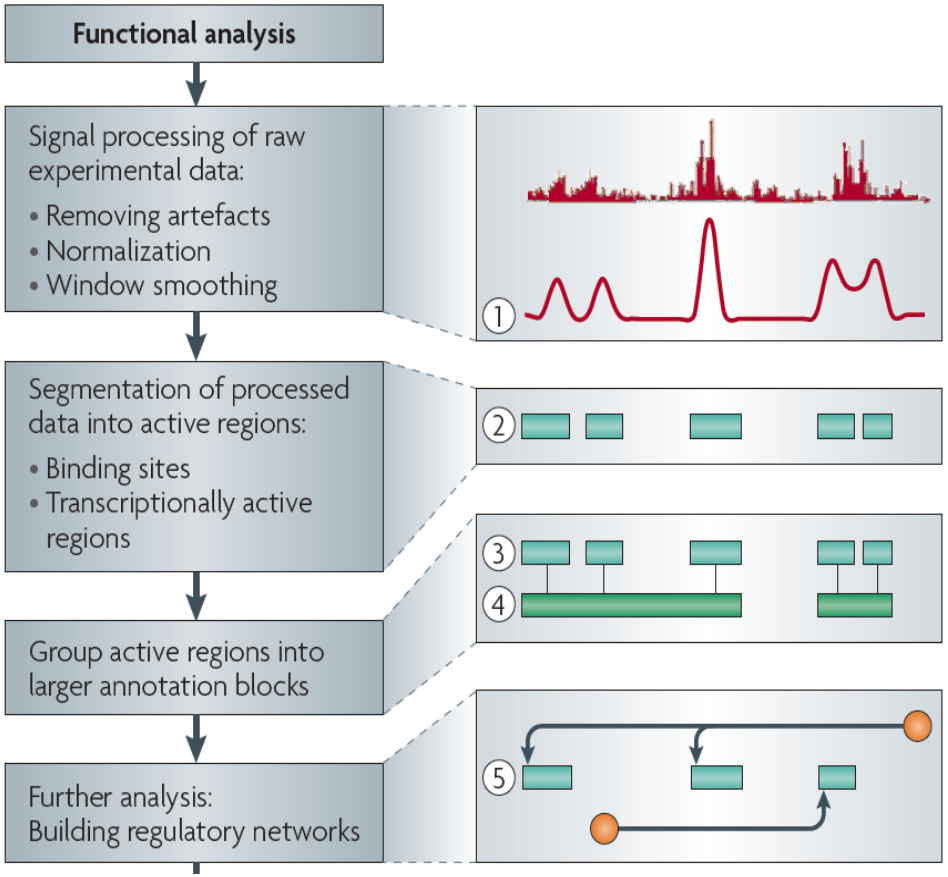
Visel et al, Nature, 2009; Maurano et al, Science, 2012

Functional genomic elements



Nature, News and Views, 2012

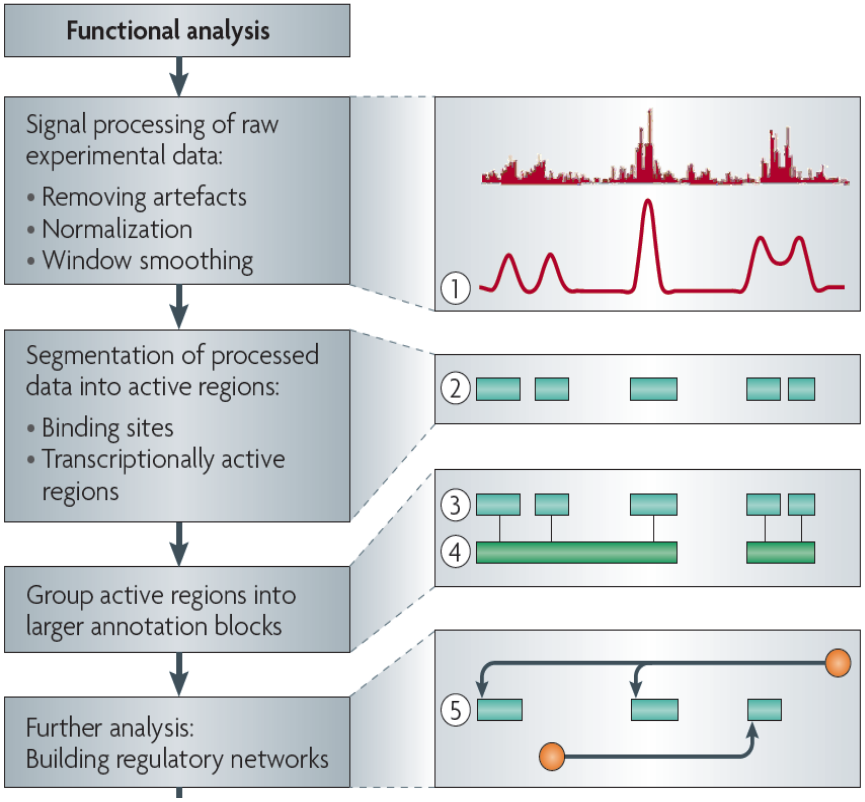
Steps in annotation of non-coding regions



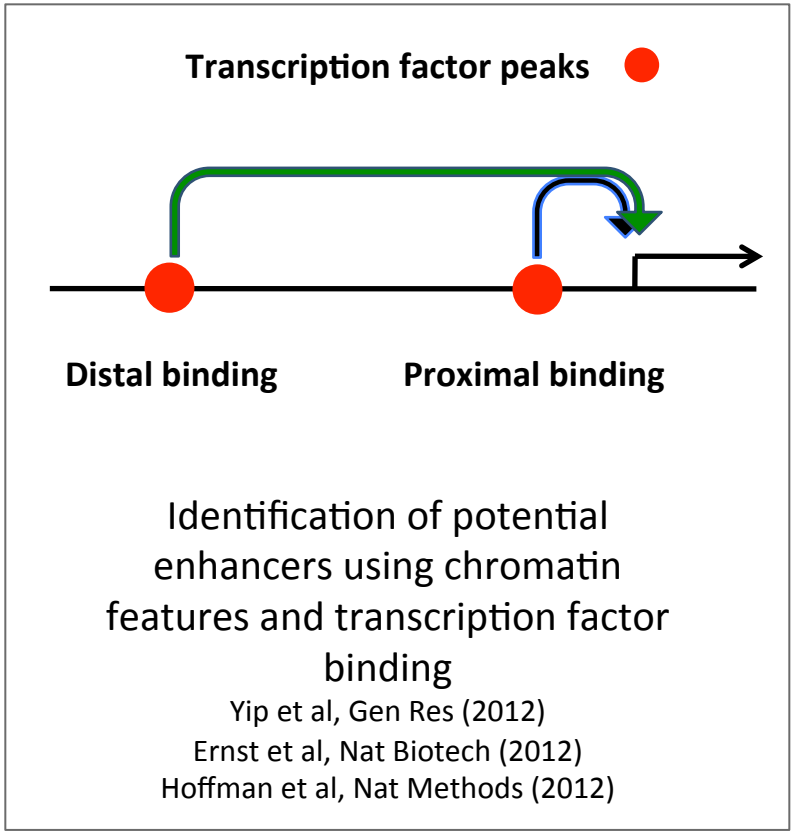
Transcription factor peaks, complex non-coding RNA transcripts

Alexander et al, Nat Rev Gen, 2010

Steps in annotation of non-coding regions

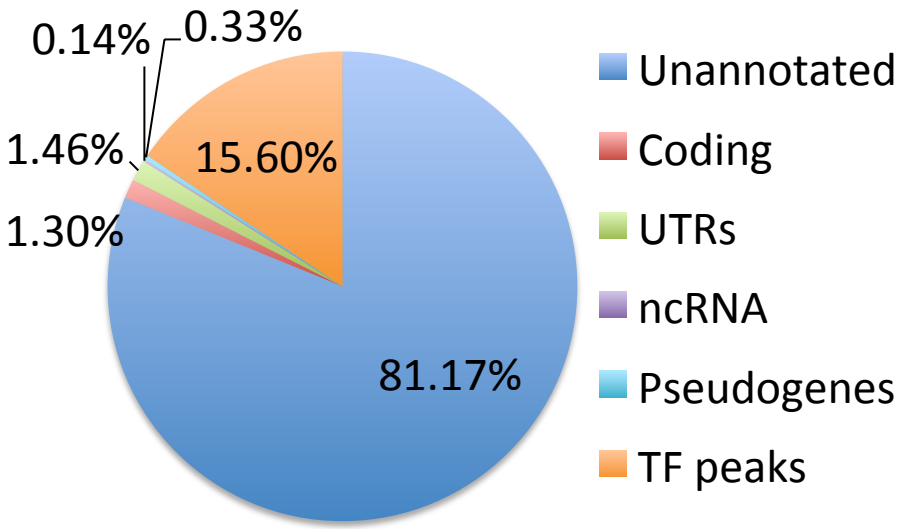


Alexander et al, Nat Rev Gen, 2010



Annotation choices for 1000 Genomes variants

- Protein-coding genes
- Non-coding RNAs, UTRs and pseudogenes from GencodeV7
- Transcription factor (TF) peaks of 119 factors from 5 ENCODE cell lines and higher resolution motifs
- Enhancer annotations
- Resource for researchers to pinpoint potential functional roles of variants



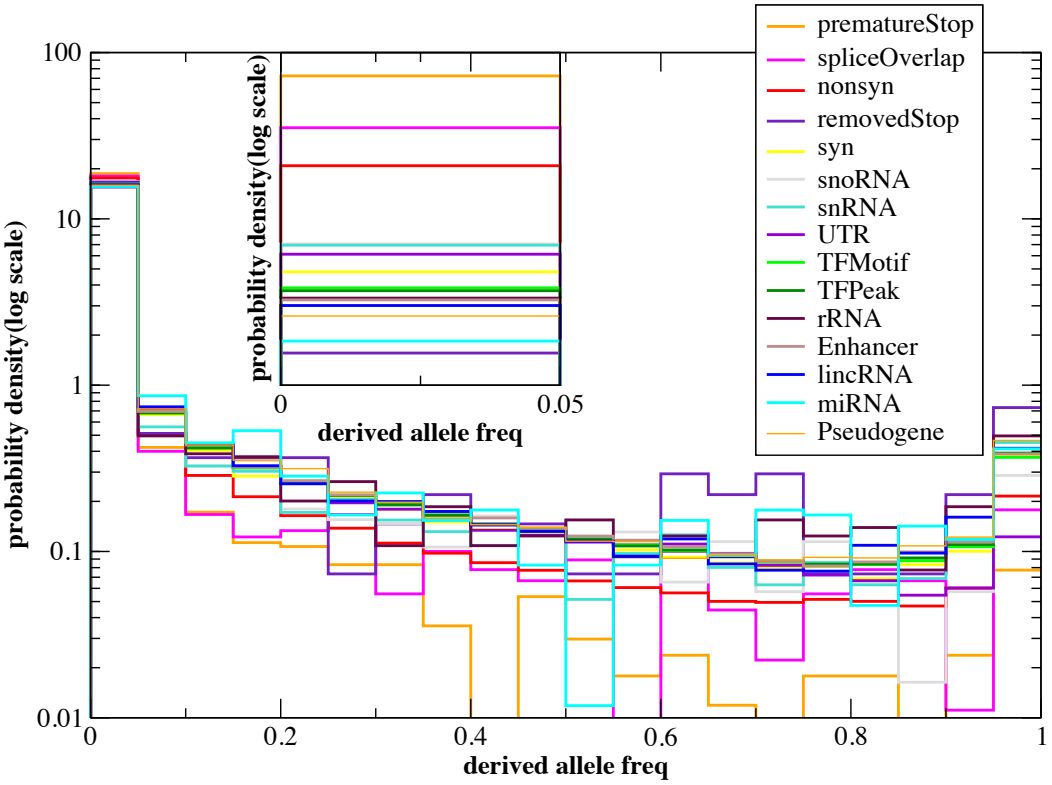
Data details

- Variant annotations for integrated Phase I release (SNPs, Indels and SVs) provided in VCF files
- Tags such as TFpeak, TFMotif, miRNA, Enhancer etc provided in VCF files
- Coordinates of functional elements provided in separate BED files
- Link:

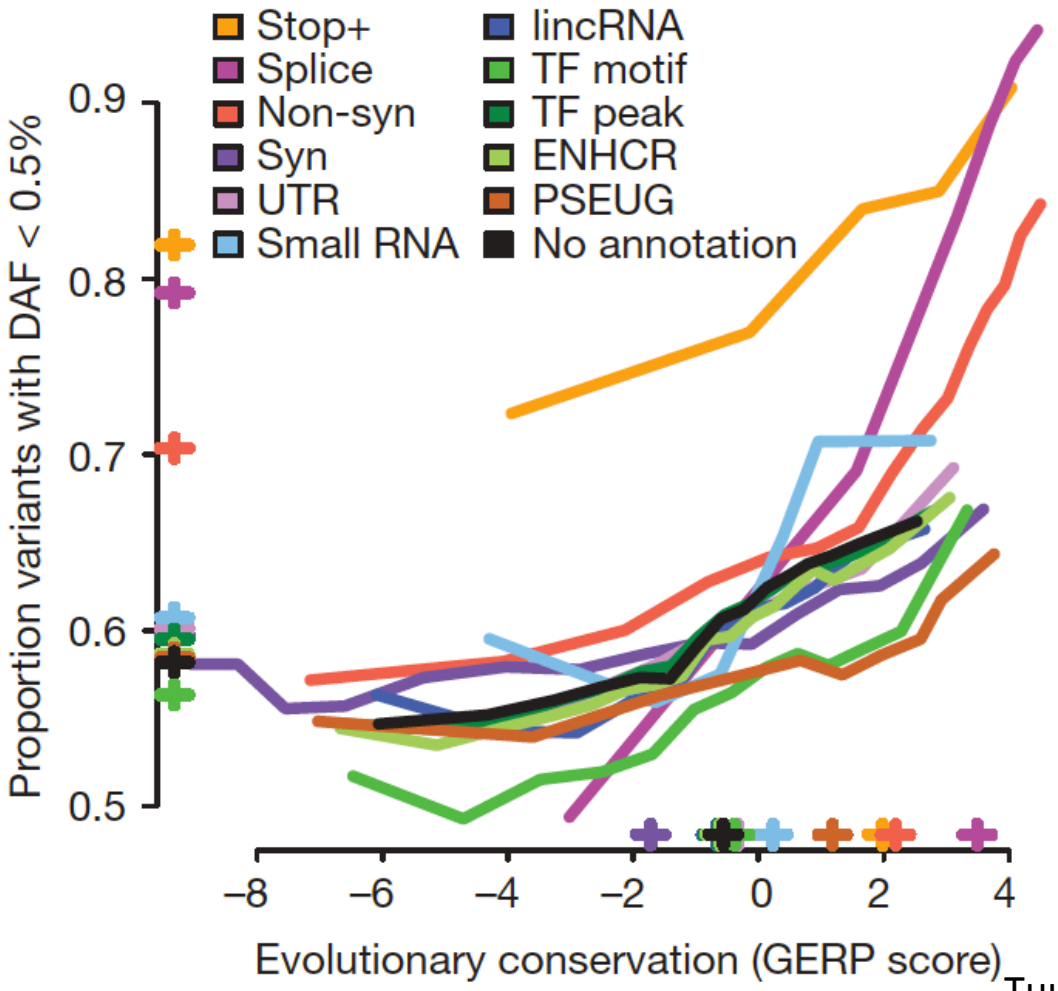
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/functional_annotation/

Purifying selection in various functional categories

- Enrichment of rare SNPs shows most functional categories are under selection relative to pseudogenes
- Coding exons under stronger selection than non-coding regulatory regions
- Transcription factor motifs within peaks under stronger selection than the entire peak region

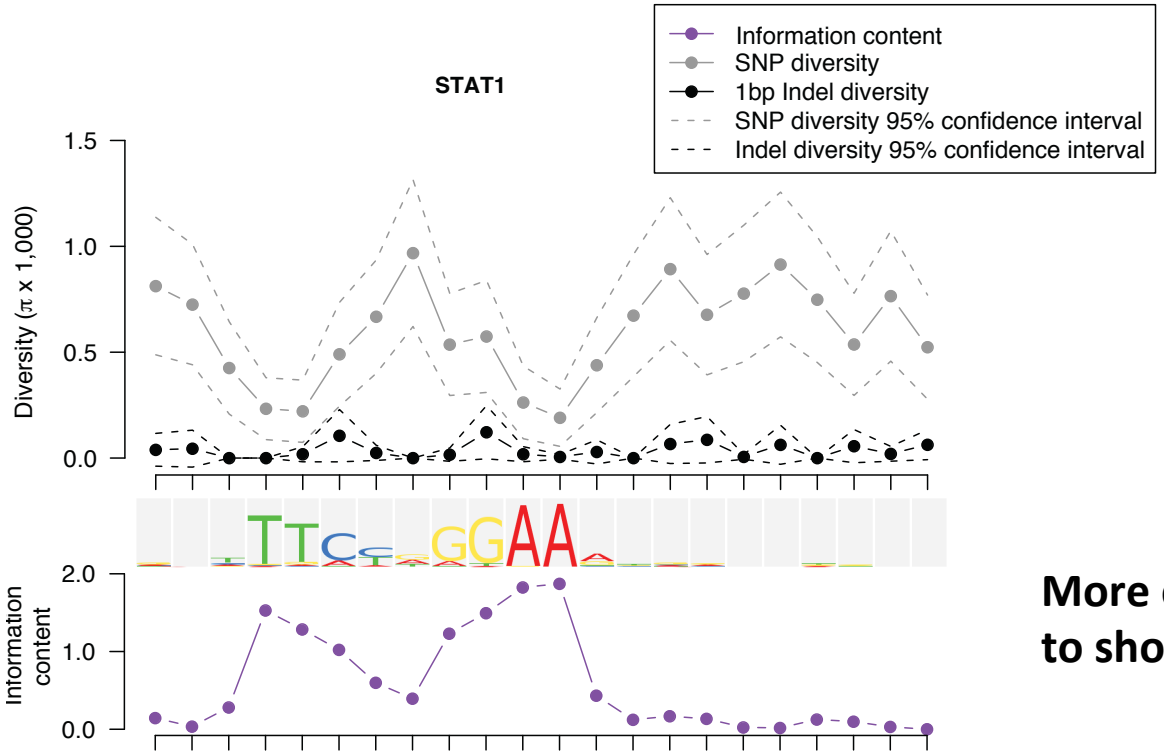


Purifying selection amongst humans vs evolutionary conservation

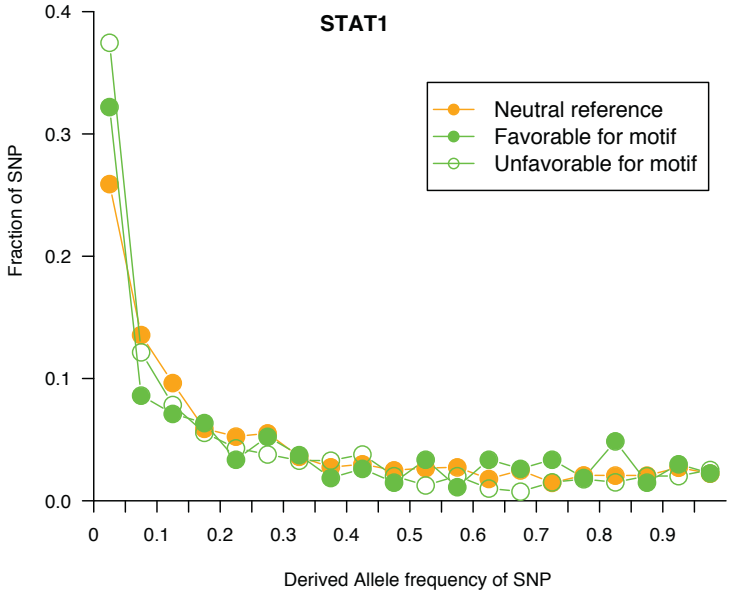


Tuuli Lappalainen

Selection constraints in TF-binding motifs



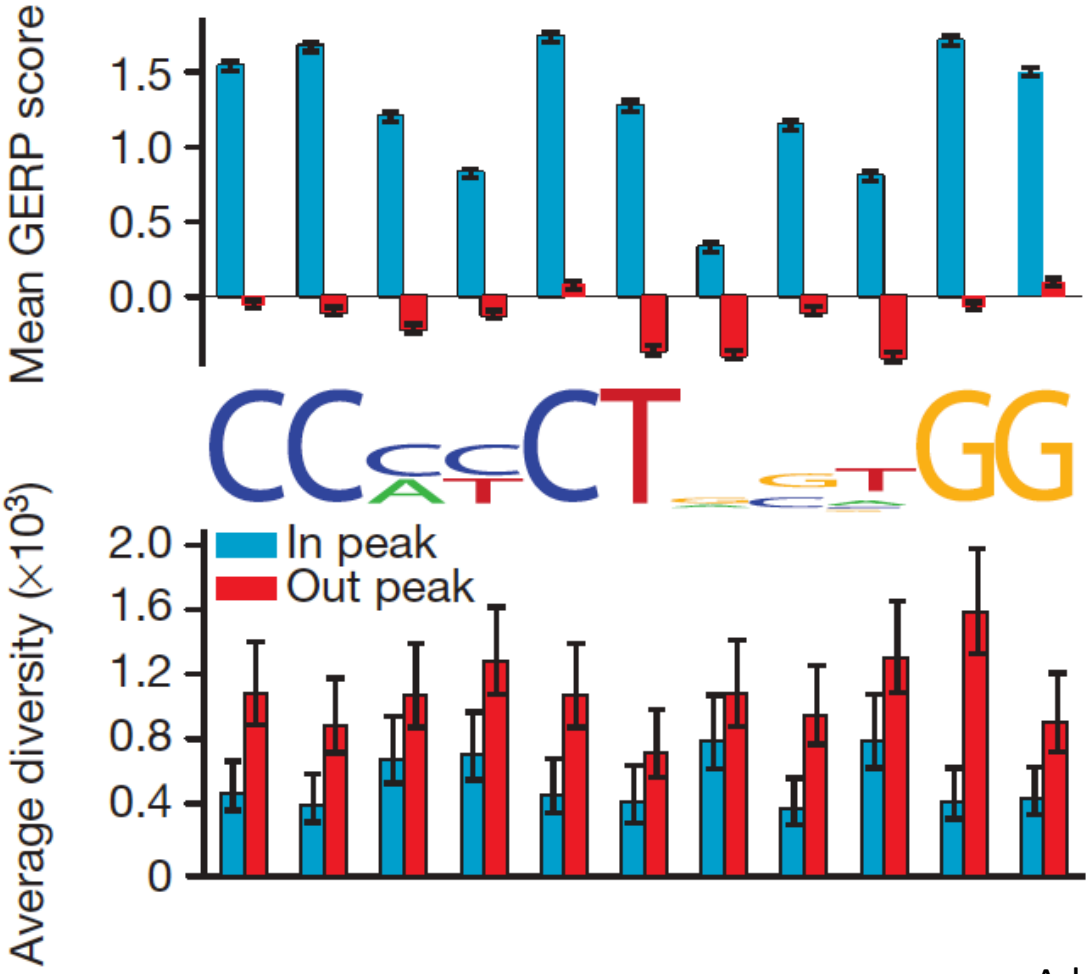
More conserved motif sites tend to show lower SNP diversity



SNPs in STAT1 motif show enrichment of rare alleles compared to neutral reference

Mu et al, Nucleic Acids Res, 2011

Selection constraints in CTCF binding motif



Adam Auton

Average number of potentially functional variants per individual

Table 2 | Per-individual variant load at conserved sites

Variant type	Number of derived variant sites per individual		
	Derived allele frequency across sample		
	<0.5%	0.5–5%	>5%
All sites	30–150 K	120–680 K	3.6–3.9 M
Synonymous*	29–120	82–420	1.3–1.4 K
Non-synonymous*	130–400	240–910	2.3–2.7 K
Stop-gain*	3.9–10	5.3–19	24–28
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8
HGMD-DM*	2.5–5.1	4.8–17	11–18
COSMIC*	1.3–2.0	1.8–5.1	5.2–10
Indel frameshift	1.0–1.3	11–24	60–66
Indel non-frameshift	2.1–2.3	9.5–24	67–71
Splice site donor	1.7–3.6	2.4–7.2	2.6–5.2
Splice site acceptor	1.5–2.9	1.5–4.0	2.1–4.6
UTR*	120–430	300–1,400	3.5–4.0 K
Non-coding RNA*	3.9–17	14–70	180–200
Motif gain in TF peak*	4.7–14	23–59	170–180
Motif loss in TF peak*	18–69	71–300	580–650
Other conserved*	2.0–9.9 K	7.1–39 K	120–130 K
Total conserved	2.3–11 K	7.7–42 K	130–150 K

* Sites with GERP>2

1000 Genomes Consortium, Nature, 2012

Future

- More non-coding annotations, Ensembl Regulatory build
- Functional impact of structural variants
- Relationship of eQTLs with functional annotation and purifying selection

Acknowledgements

Functional Interpretation Group (FIG)

~40 participants

Yale University: M Gerstein (co-chair), Y Fu, X Mu, S Balasubramanian, A Harmanci, C Sisu, J Chen, D Clarke
Wellcome Trust Sanger Institute: C Tyler-Smith (co-chair), Y Xue, V Colonna, Y Chen
University of Geneva: T Lappalainen
University of Michigan: HM Kang
Cornell University: J Das, H Yu
Massachusetts General Hospital: D MacArthur
University of Montreal: P Awadalla, A Hodgkinson
University of Medicine and Dentistry of New Jersey: J Rosenfeld
EBI: L Clarke, J Herrero, F Cunningham, P Flicek