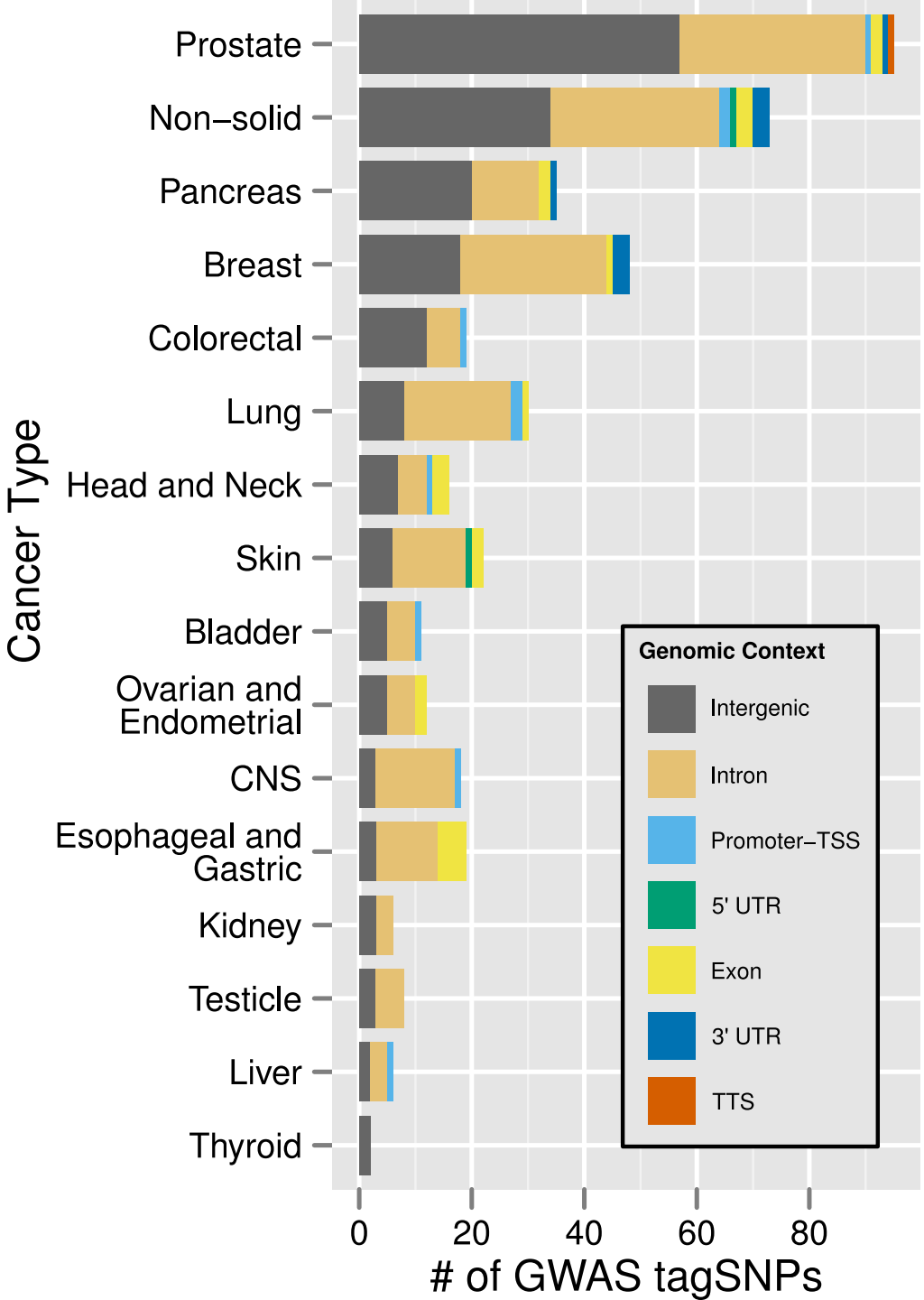
USC
UNIVERSITY
OF SOUTHERN
CALIFORNIA

NCI CCC
A Comprehensive Cancer Center Designated by the National Cancer Institute

Unlocking the Secrets of Enhancer Biology with GWAS

Dennis J. Hazelett, Cedars-Sinai, LA

1. Best SNPs
2. Best Enhancers



"Houston we have a Problem": (Tom Hanks)

NAR 2012, Downloaded 3758 times from Bioconductor

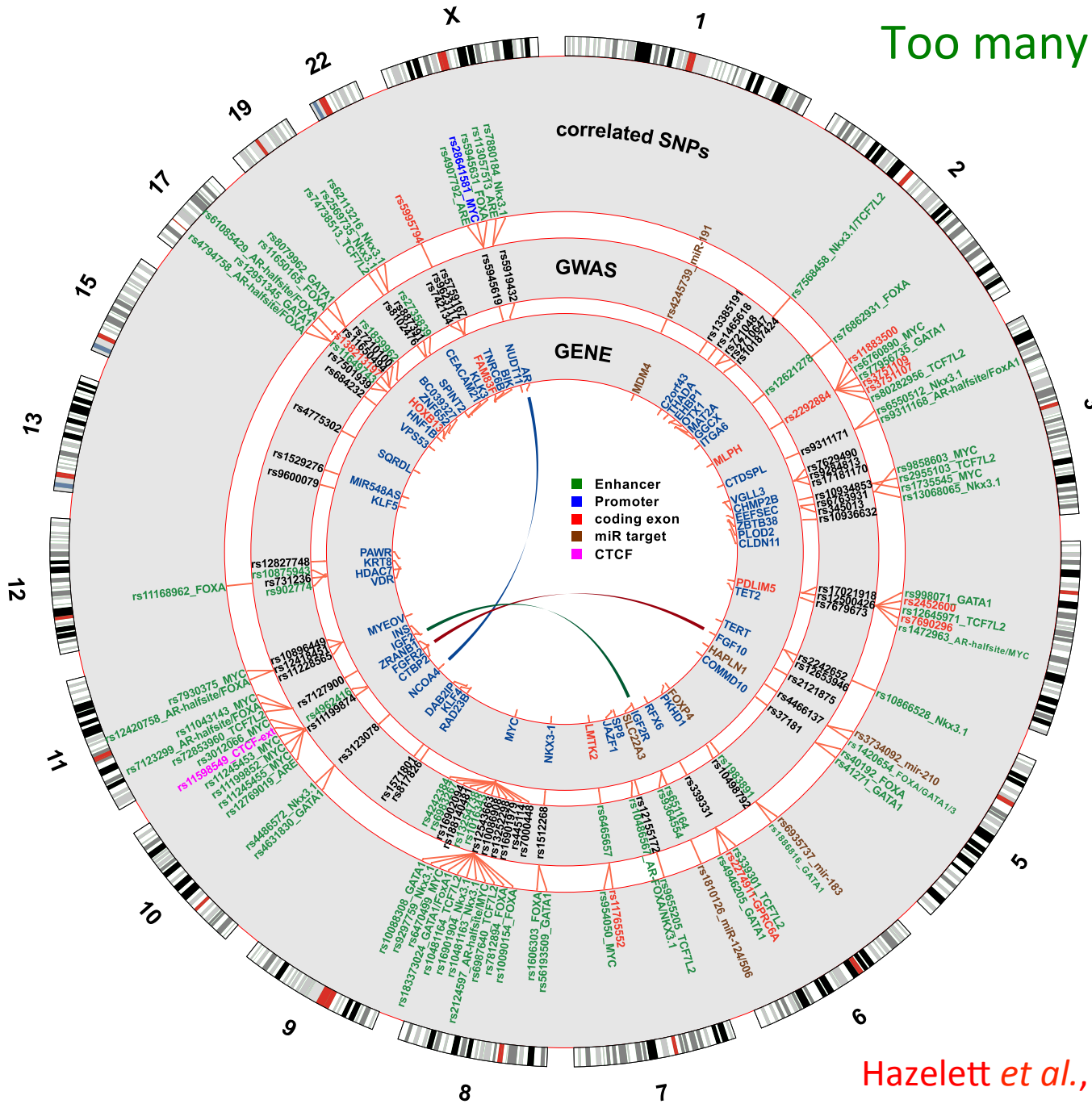
FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs

Simon G. Coetzee[1,2], Suhn K. Rhie[1,2], Benjamin P. Berman[1,2,3], Gerhard A. Coetzee[1,2,4,*] and Houtan Noushmehr[1,2,3,*]

[1]Norris Cancer Center, [2]Department of Preventive Medicine, [3]Epigenome Center and [4]Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

Received March 20, 2012; Revised April 26, 2012; Accepted May 13, 2012

This results in a significant reduction in candidate functional SNPs to be tested.
(From 10's of thousands to <1000)



Too many candidates!

Hazelett *et al.*, PLoS Gen 2014

Genome-wide two important Questions

Show me the best risk SNPs/enhancers

And

Show me the Genes
(Cuba Gooding)

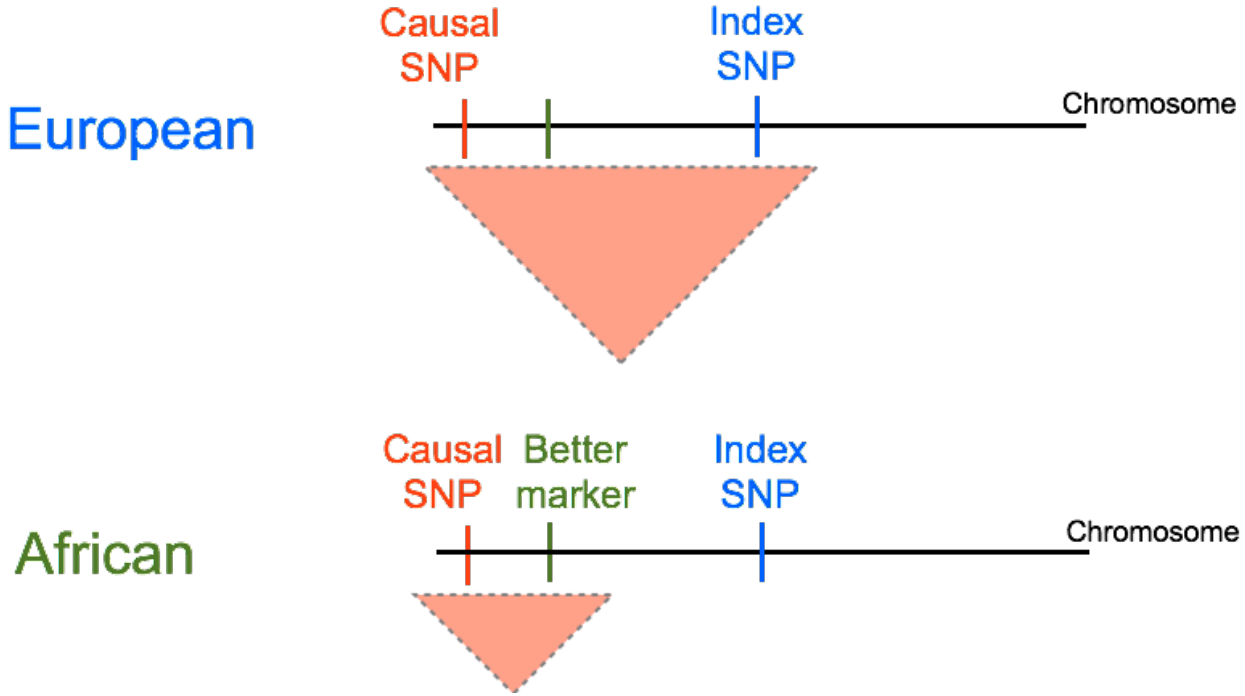
1. Best risk SNPs

Prostate Cancer fine-mapping of multi-ethnic cohorts overlaid FunciSNP

- Ying Han (Haiman)
- Dennis Hazelett (Coetzee)

Rationale

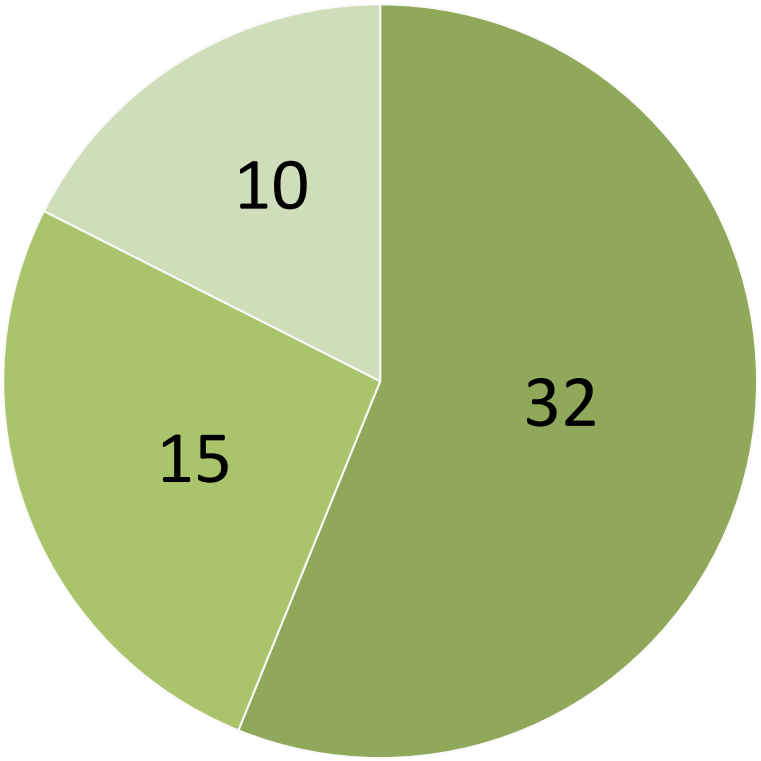
- Better coverage of genetic variation: imputed to 1000 Genomes
- Empowered by the large sample size
- Multiethnic populations help reduce the set of candidate SNPs



- Comprehensive functional annotation of all candidate SNPs

Ying Han

The most associated SNPs at 57 loci

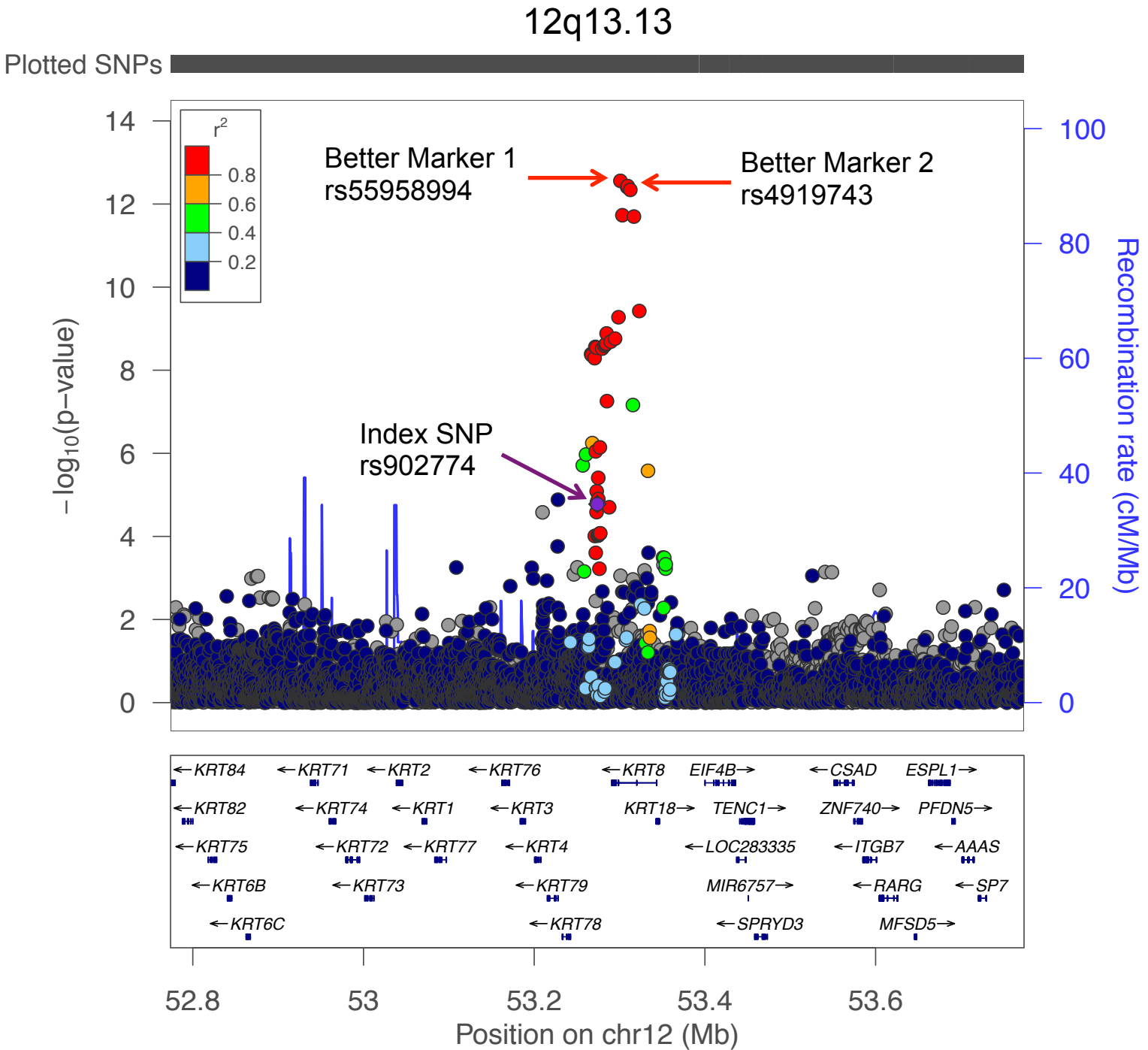


Compared to the index SNP

- > 1 order of magnitude change in the p-value
- < 1 order of magnitude change in the p-value
- Index variant remains the most significant

Better marker identified for 32 index SNPs

25 index SNPs or their proxies remain the most associated SNPs


12q13.13
Plotted SNPs
r²
0.8
0.6
0.4
0.2
Better Marker 1
rs55958994
Better Marker 2
rs4919743
Index SNP
rs902774
−log10(p−value)
Recombination rate (cM/Mb)
14
12
10
8
6
4
2
0
100
80
60
40
20
0
←KRT84 ←KRT71 ←KRT2 ←KRT76 ←KRT8 EIF4B→ ←CSAD ESPL1→
←KRT82 ←KRT74 ←KRT1 ←KRT3 KRT18→ TENC1→ ZNF740→ PFDN5→
←KRT75 ←KRT72 ←KRT77 ←KRT4 ←LOC283335 ←ITGB7 ←AAAS
←KRT6B ←KRT73 ←KRT79 MIR6757→ ←RARG ←SP7
←KRT6C ←KRT78 ←SPRYD3 MFSD5→
52.8
53
53.2
53.4
53.6
Position on chr12 (Mb)

SNPs per region

Selection Criteria	SNPs/region
All SNPs (1K) at 1MB/region	~3,000 (MAF ≥ 1%)
SNPs r^2 ≥ 0.1 with index SNP	~40
SNPs r^2 ≥ 0.5 in biofeatures	9.8
SNPs significantly fine mapped	12 - 22
SNPs significantly fine mapped AND in biofeatures	3.5 - 5.4 (2.4 common in two independent studies)

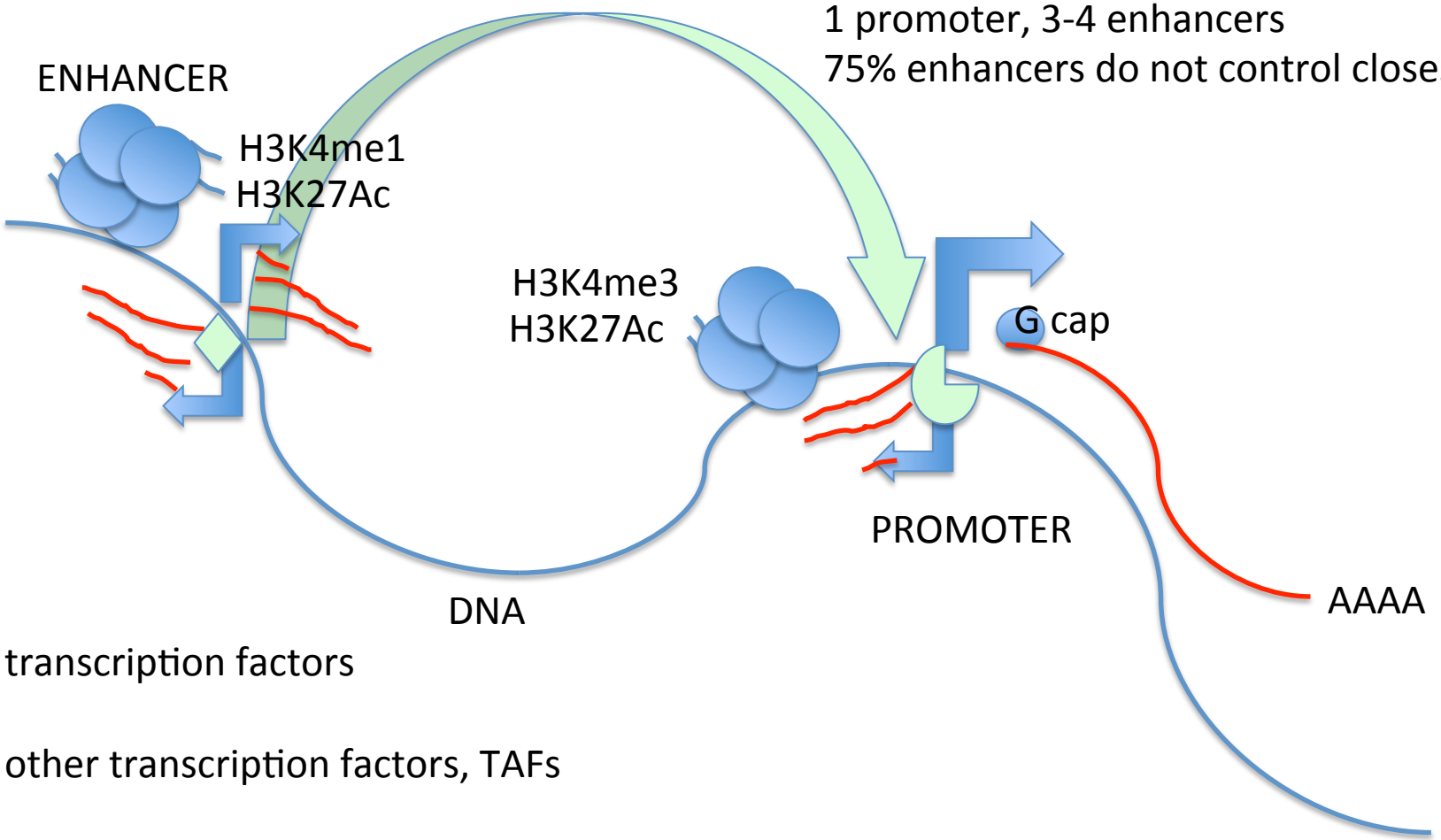
Summary

Among 69 known prostate cancer risk loci

- 57 loci were statistically significant in our study
- We identified better markers at 32 loci
- 3-5 putatively functional, finemapped SNPs for followup studies

2. Best Enhancers

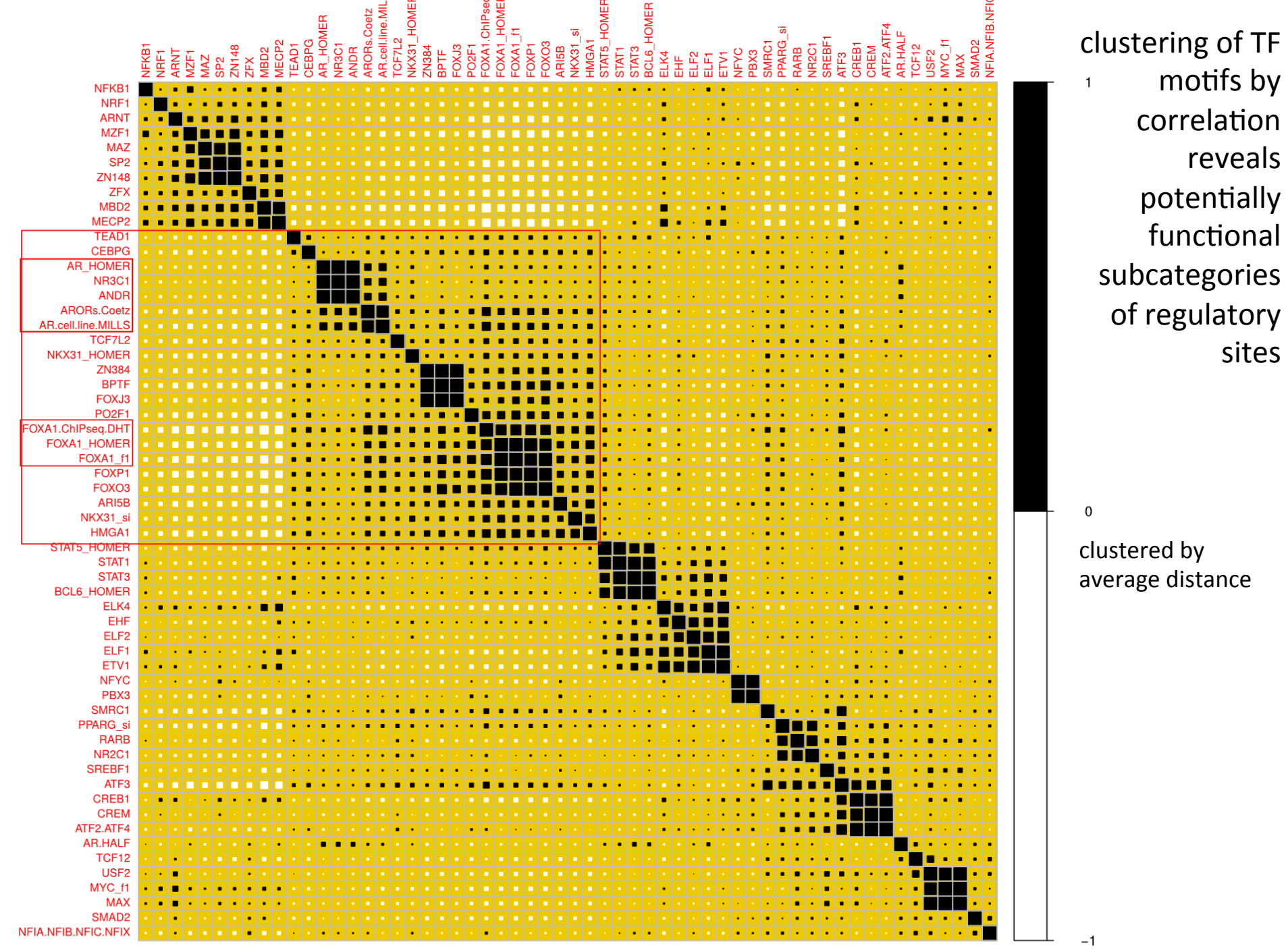
Average:
1 enhancer, 3-4 promoters
1 promoter, 3-4 enhancers
75% enhancers do not control closest gene



Andersson *et al.*, 2014 (Nature)

How we chose to define and classify putative regulatory sites in LNCaP

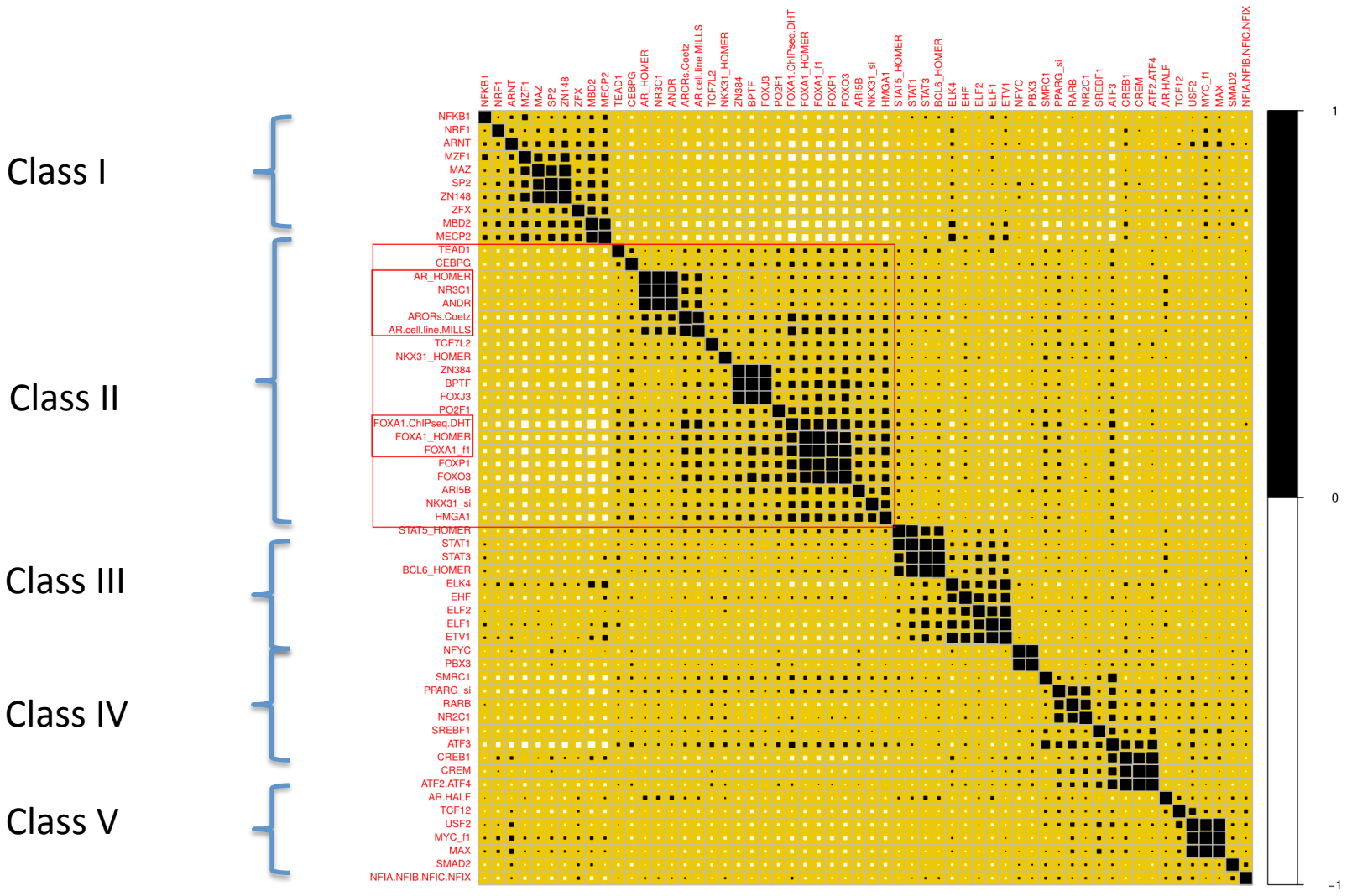
- DNaseI hypersensitivity sites within H3K27Ac regions (Duke set)
- response elements for all motifs in Factorbook and HOCOMOCO, plus a few from Homer
- throw out bottom 5% of response elements
- make a big data grid; rows of DHS sites, columns of response element counts. Pearson correlation of REs, followed by unsupervised clustering (avg link)
- regulatory sites = includes “Enhancer” >1kb distance from TSS, “Promoter” <1 kb TSS



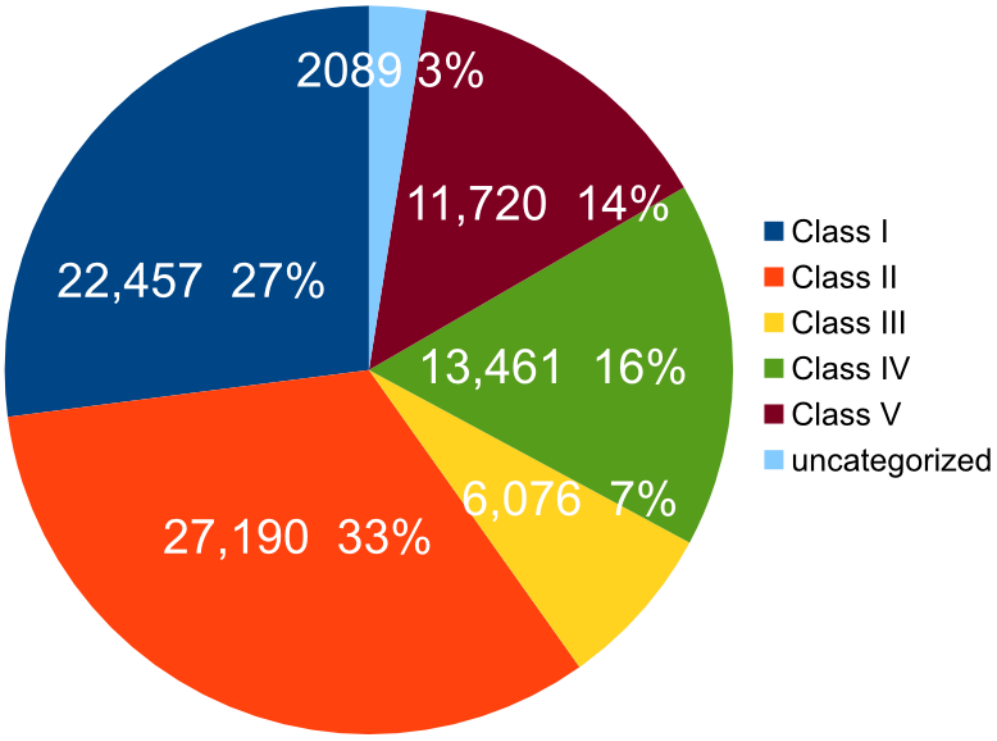
clustering of TF motifs by correlation reveals potentially functional subcategories of regulatory sites

clustered by average distance

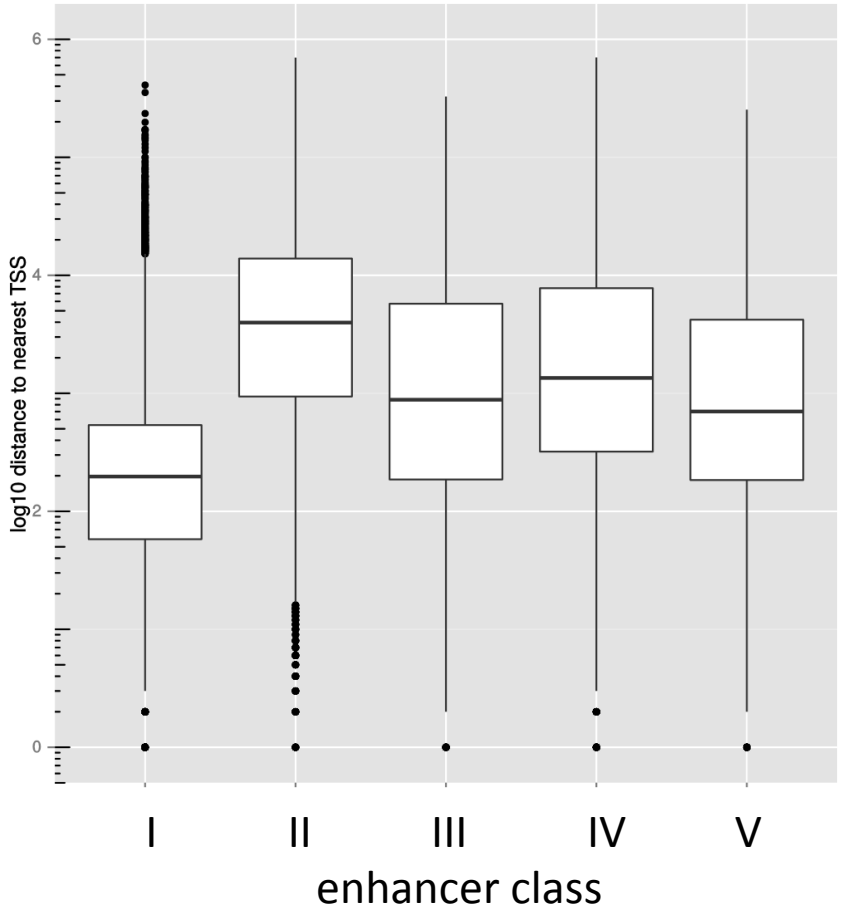
Create classification scheme based on TF motif clustering



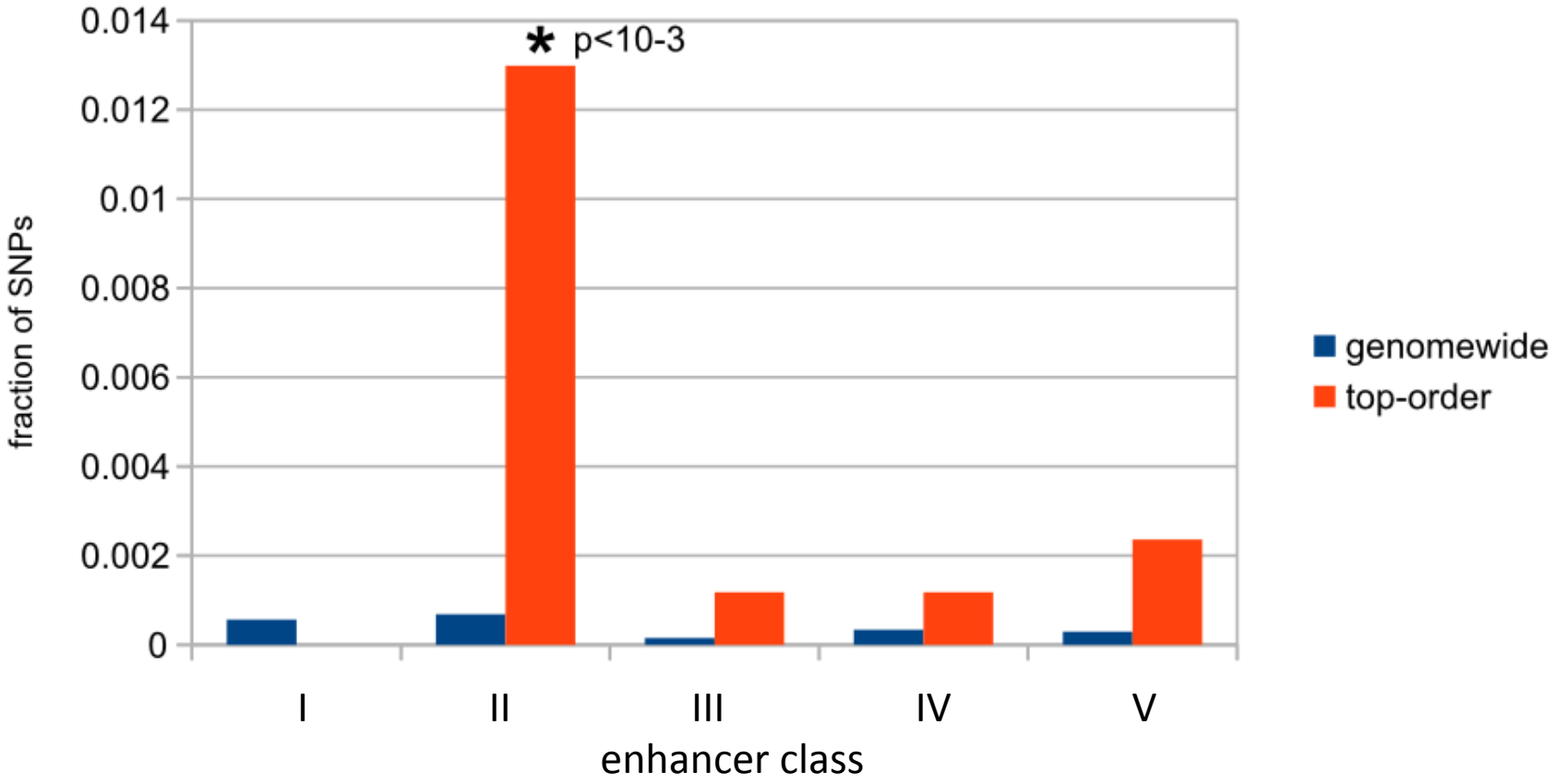
Number of DNaseI HS sites by class in LNCaP

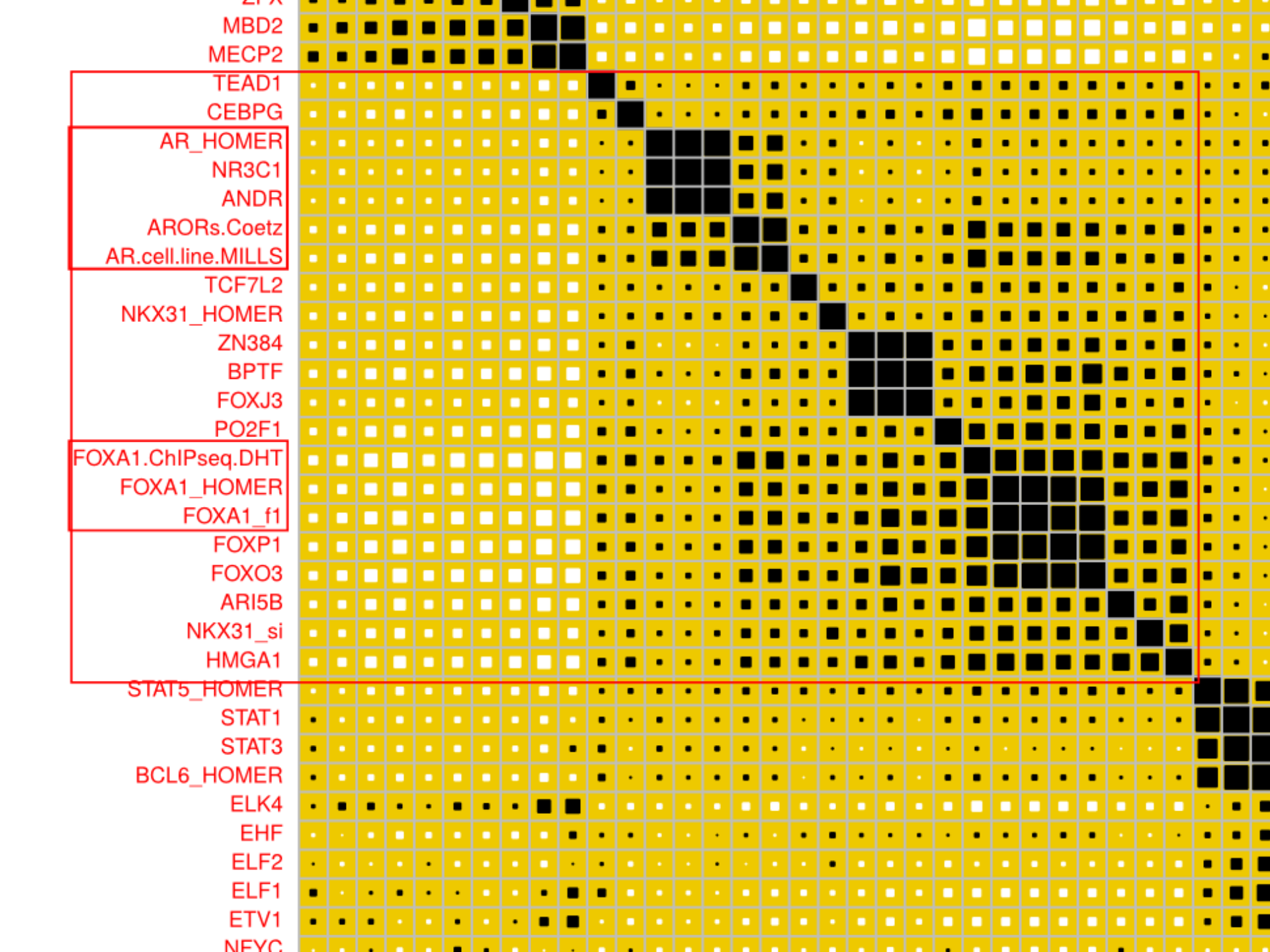


Class I – Class V regulatory regions are spatially differentiated relative to transcription start sites



Prostate Cancer risk SNPs are enriched for Group II enhancers




MBD2
MECP2
TEAD1
CEBPG
AR_HOMER
NR3C1
ANDR
ARORs.Coetz
AR.cell.line.MILLS
TCF7L2
NKX31_HOMER
ZN384
BPTF
FOXJ3
PO2F1
FOXA1.ChIPseq.DHT
FOXA1_HOMER
FOXA1_f1
FOXP1
FOXO3
ARI5B
NKX31_si
HMGA1
STAT5_HOMER
STAT1
STAT3
BCL6_HOMER
ELK4
EHF
ELF2
ELF1
ETV1

Summary

- LNCaP enhancers can be classified according to TF binding motifs
- Risk SNPs are highly enriched in a specific class of enhancers
- A short-list of best risk enhancers/ SNPs was compiled

So, back to two important questions

1. Show me the best risk SNPs/enhancers ✔

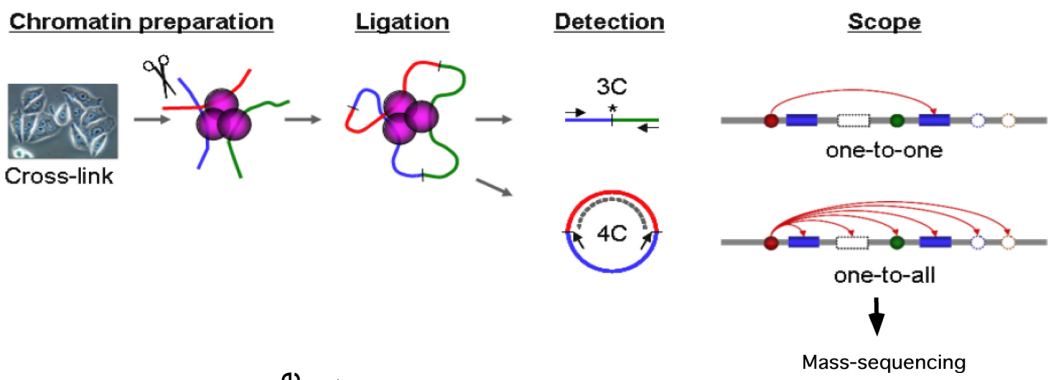
<u>But what remains:</u>

2. Show me the Genes as Cuba Gooding would ask

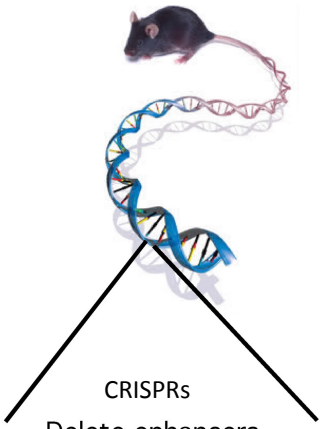
- eQTL
- 4C
- CRISPR/Cas9

Target Gene Identification

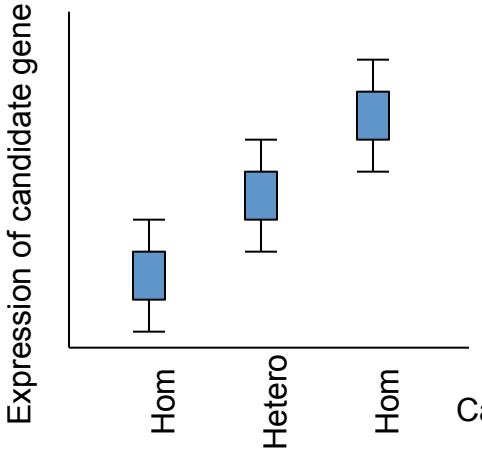
3C, 4C, 5C and Hi-C



Transgenic Mouse Models

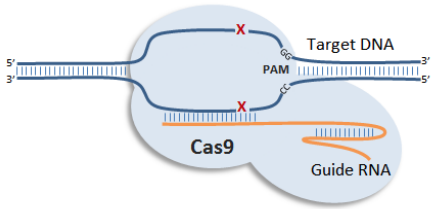


eQTL



CRISPR/Cas9

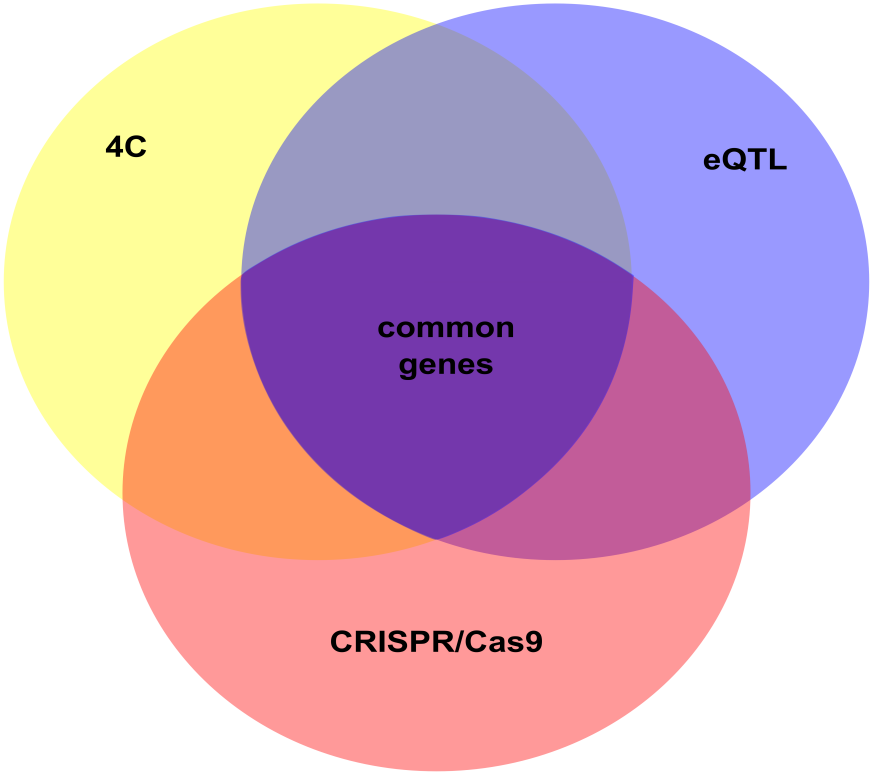
Delete enhancer in cultured cells



RNA

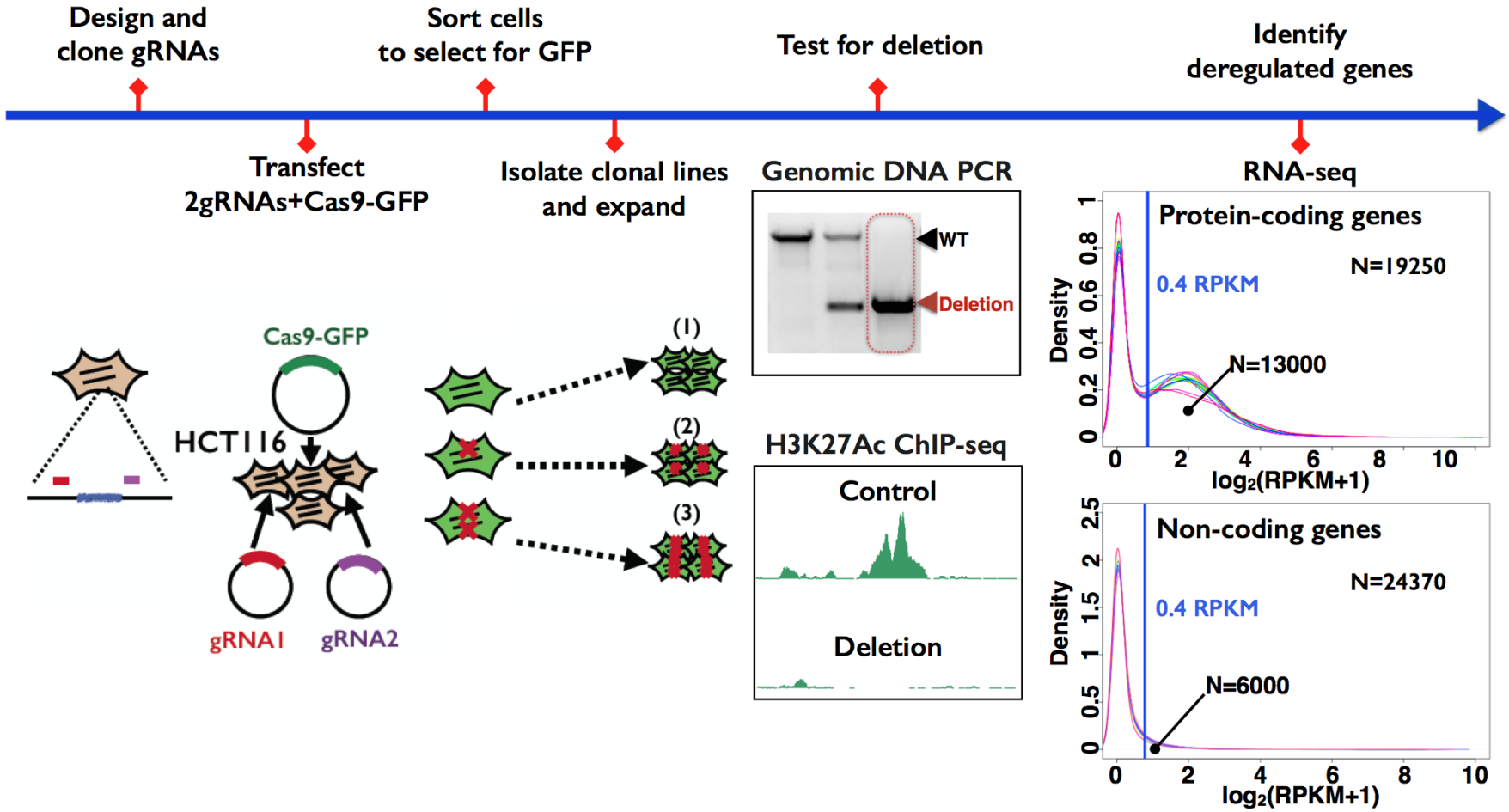
Hi-seq mass sequencing

Integration



Venn diagram of idealized target genes identified using our 3 independent approaches, resulting in common genes most likely to functionally be involved in PCa risk at each enhancer.

CRISPR Experimental Pipeline



Acknowledgements and Thanks

- Dennis Hazelett (Coetzee lab)
- Ying Han (Haiman group)

- Gerry Coetzee
  - Chunli Yan
  - Yu Guo (Phoebe)
  - Hang Yang
  - Sonya Liu
- Dennis Hazelett
  - Simon Coetzee
- Peggy Farnham
  - Malaina Gaddis
  - Esther Tak
  - Zhifei Luo
- Charlie Nicolet & NGS
- Chris Haiman
- Wange Lu
- Josh Millstein

Funding: R01CA136924, R01CA190182