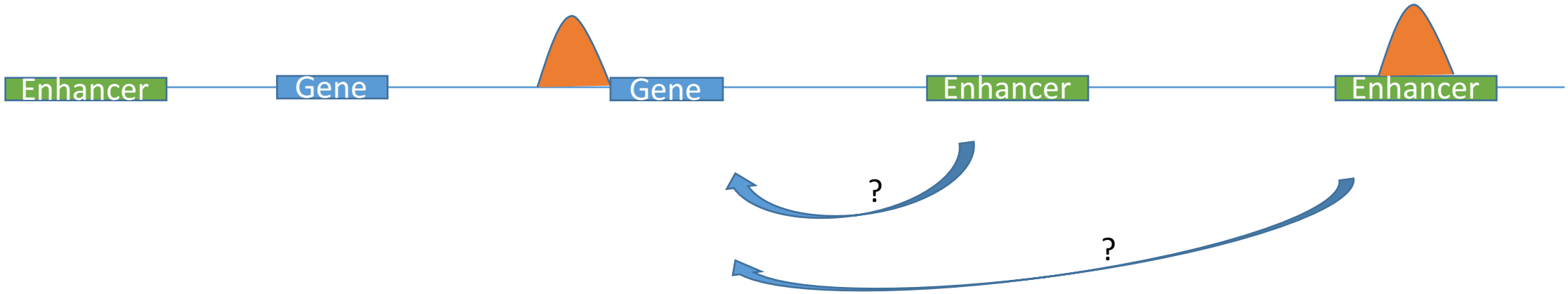
Rotation summary

Spring 2016 rotation

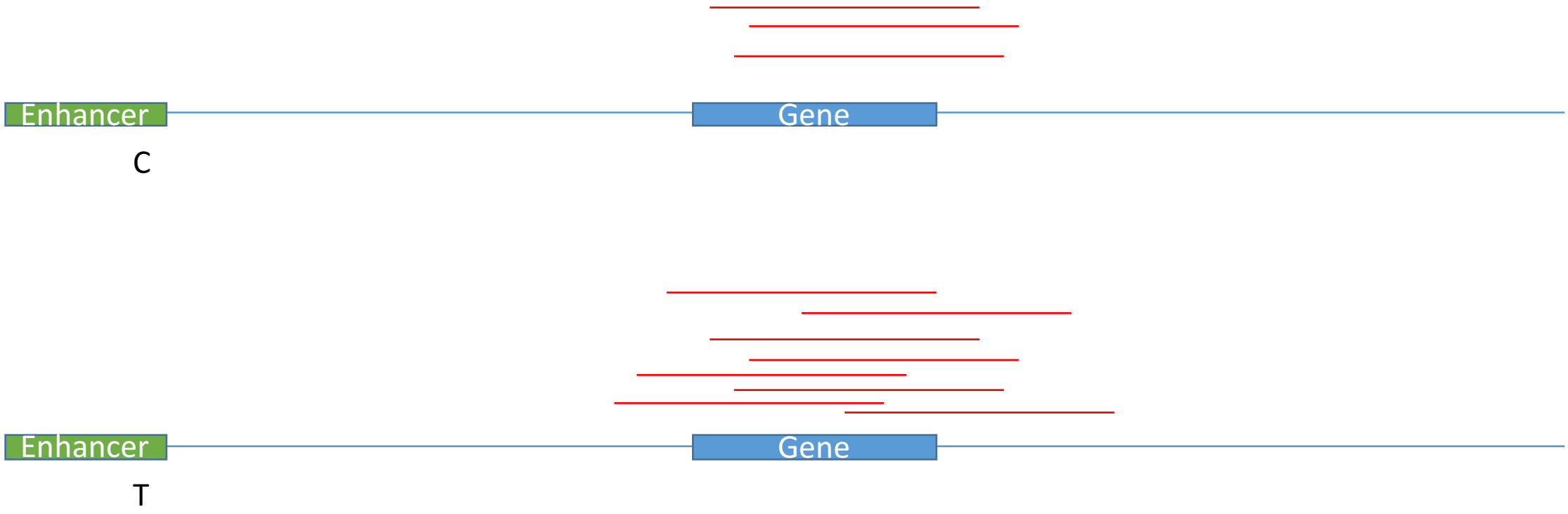


→What is the most “effective” manner for assigning distal peaks to genes?

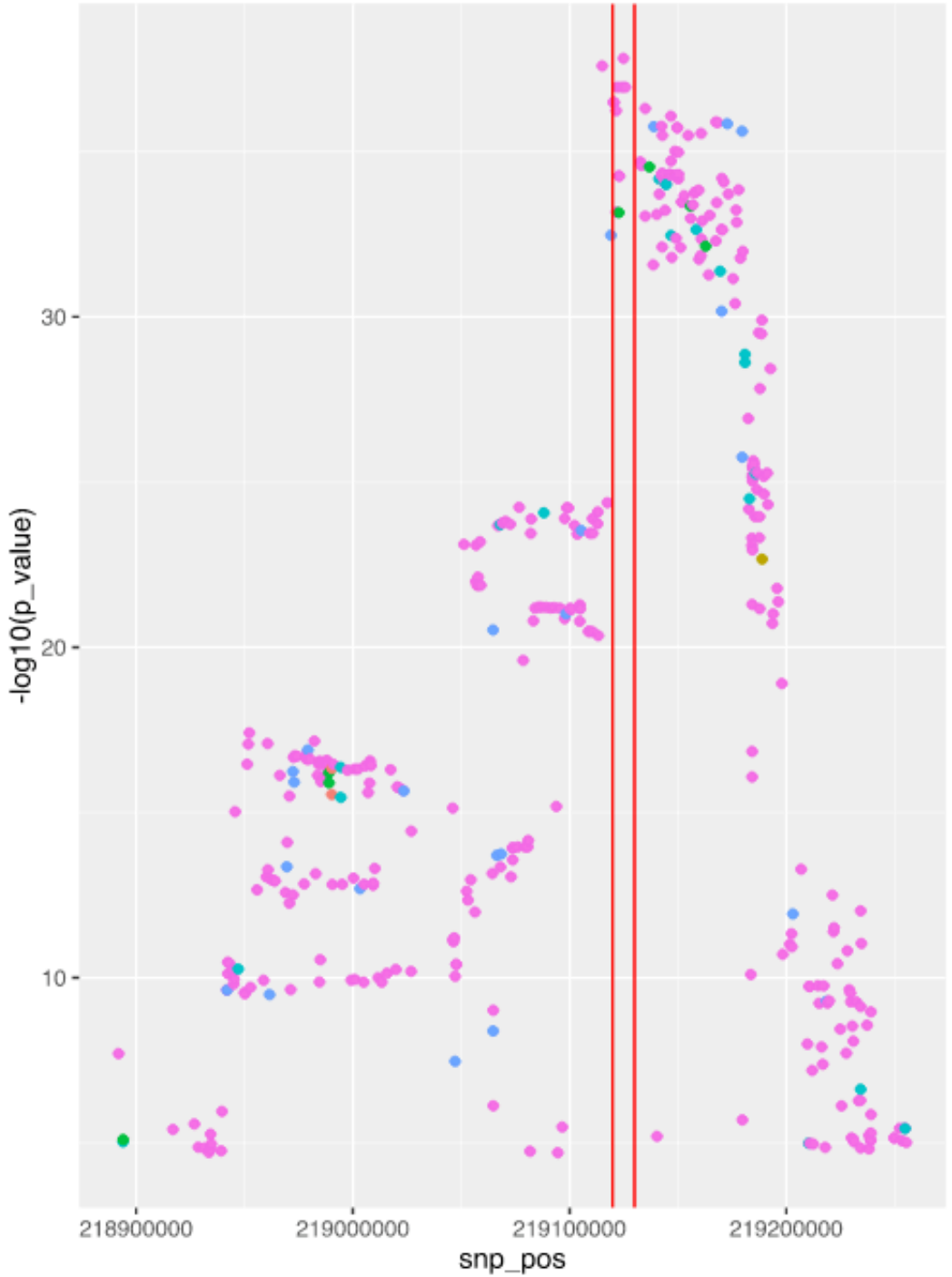
Potential methods

- eQTLs
- Chromosome capture (ChIA-PET / HiC / 5C / etc)

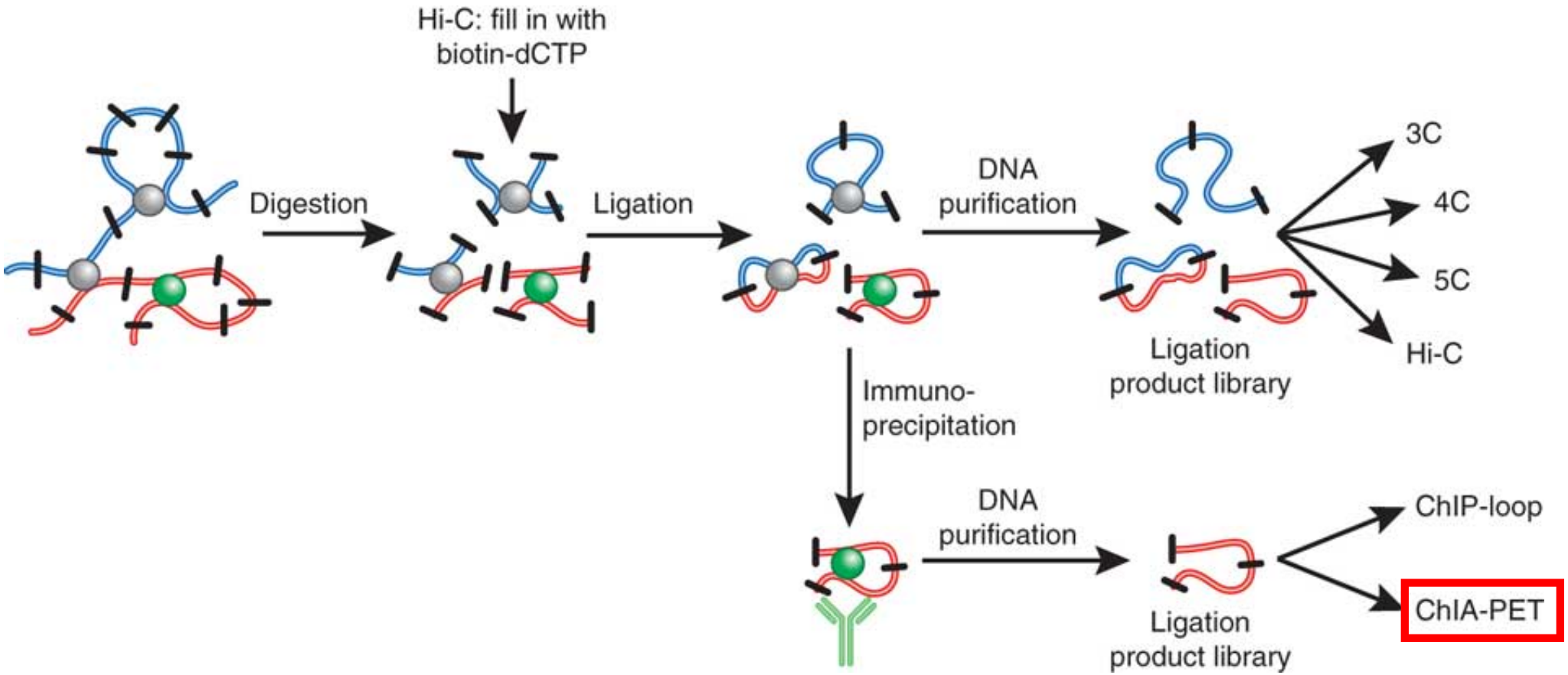
eQTLs



GTEx


30
20
10
-log10(p_value)
218900000
219000000
219100000
219200000
snp_pos

ChIA-PET



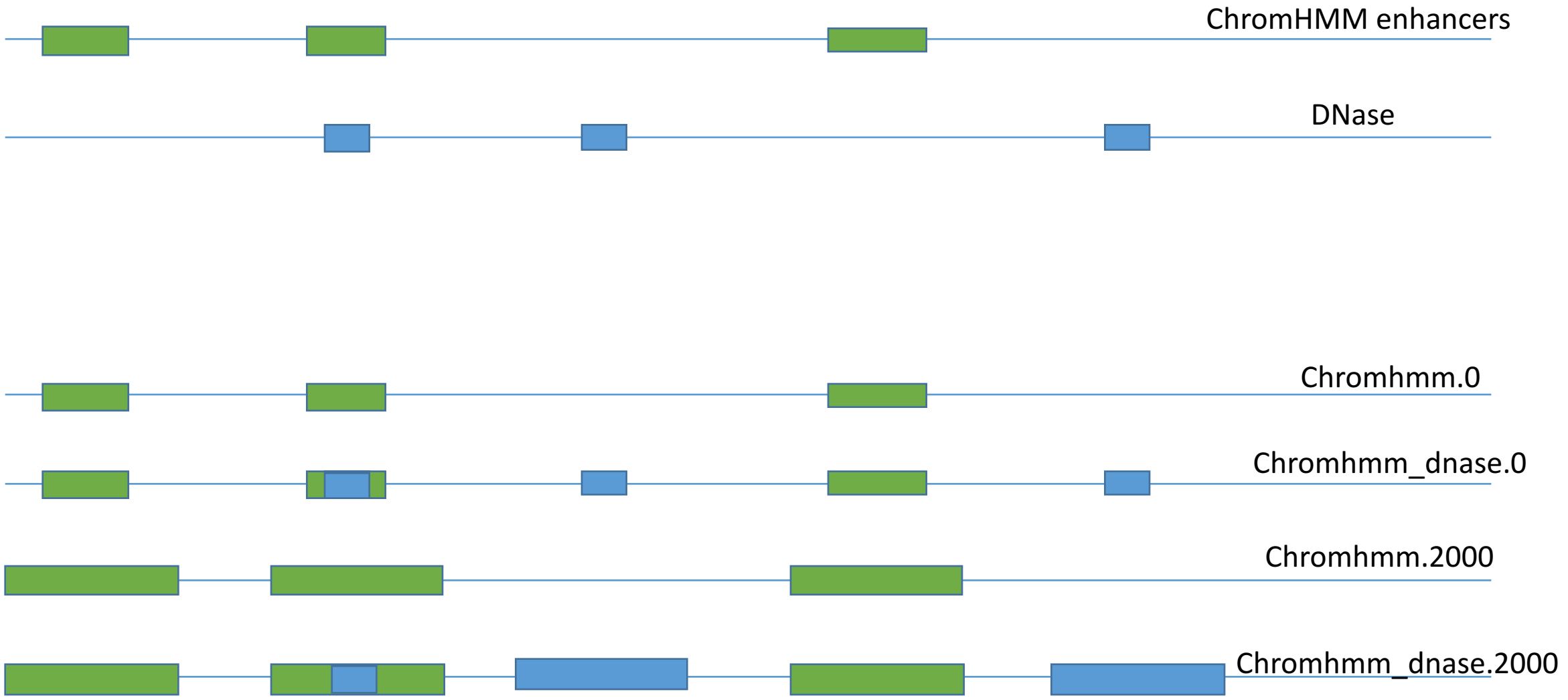
Steensel and Dekker 2010. Genomic tools for unraveling chromosome architecture (Nature)

ChIA-PET data sources

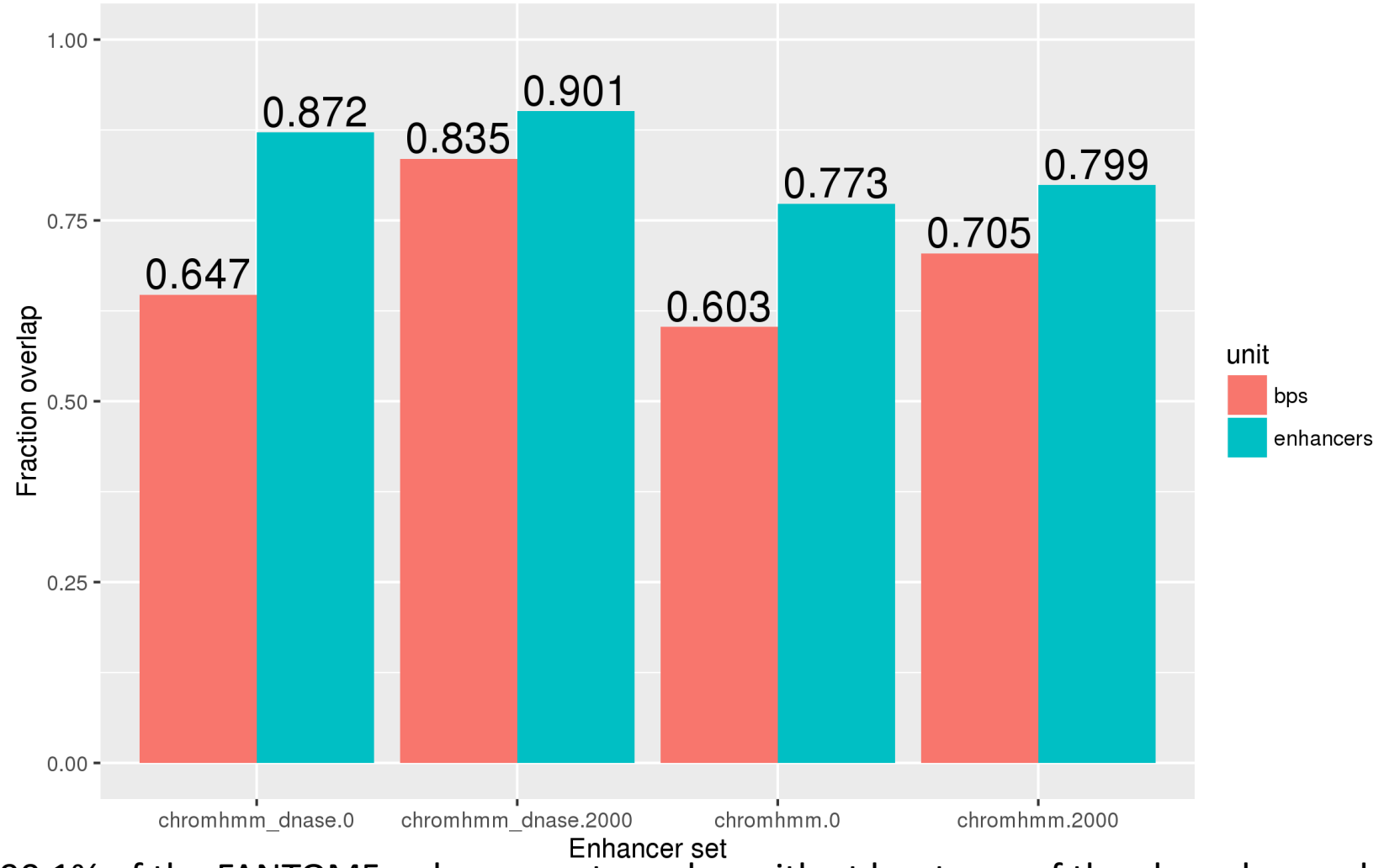
- ENCODE
  - 3x Pol2
  - 2x CTCF
  - 3x histone marks -- H3K4me1, H3K4me2, H3K4me3
  - 1x Rad21
- Young lab
  - 2x Cohesin (hESCs)
- Ruan lab (recent publication)
  - 2x CTCF (GM12878, HeLa)
  - 1x Pol2 (HeLa)

Expanding enhancer lists

- Currently FANTOM5
  - Bi-directional transcription, CAGE
  - (Extended to 2kb)
- ChromHMM (ENCODE)
  - 9 cell types
- DNase-seq (ENCODE)
  - 125 cell types

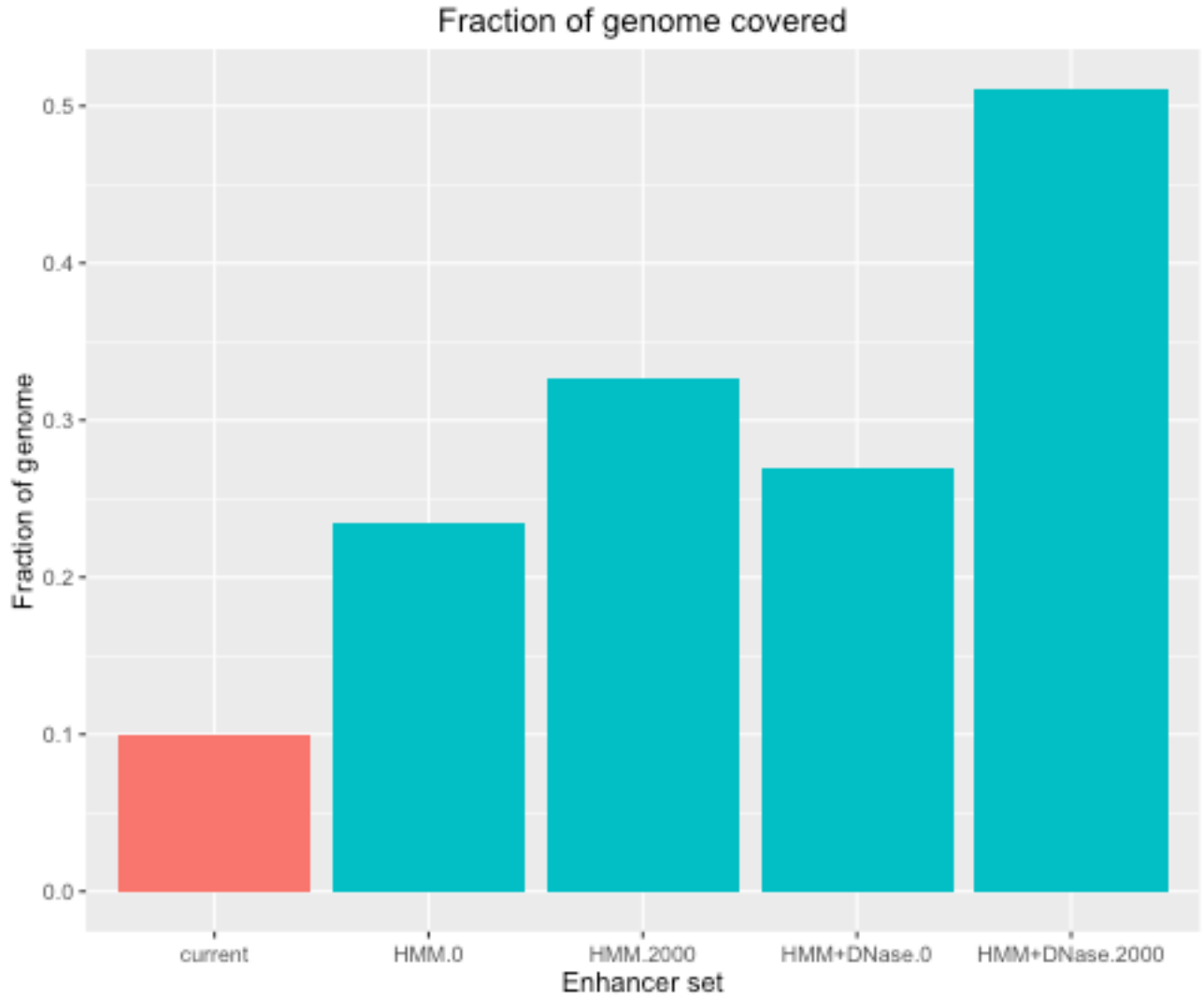

ChromHMM enhancers
DNase
Chromhmm.0
Chromhmm_dnase.0
Chromhmm.2000
Chromhmm_dnase.2000

Overlap between old and new

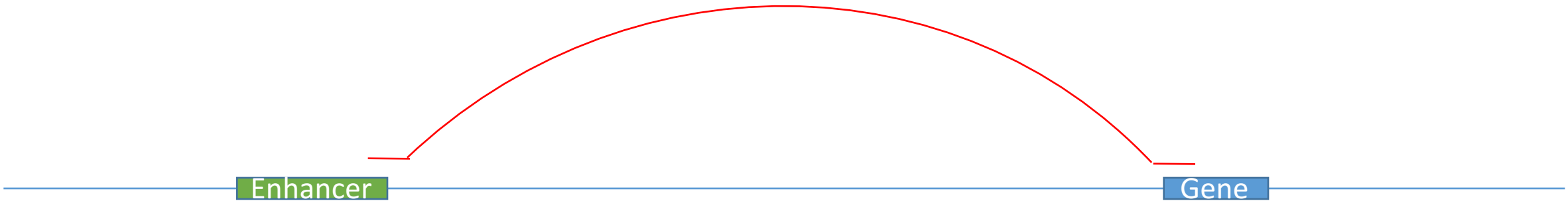


→ e.g., 90.1% of the FANTOM5 enhancer set overlap with at least one of the chromhmm_dnase.2000 set

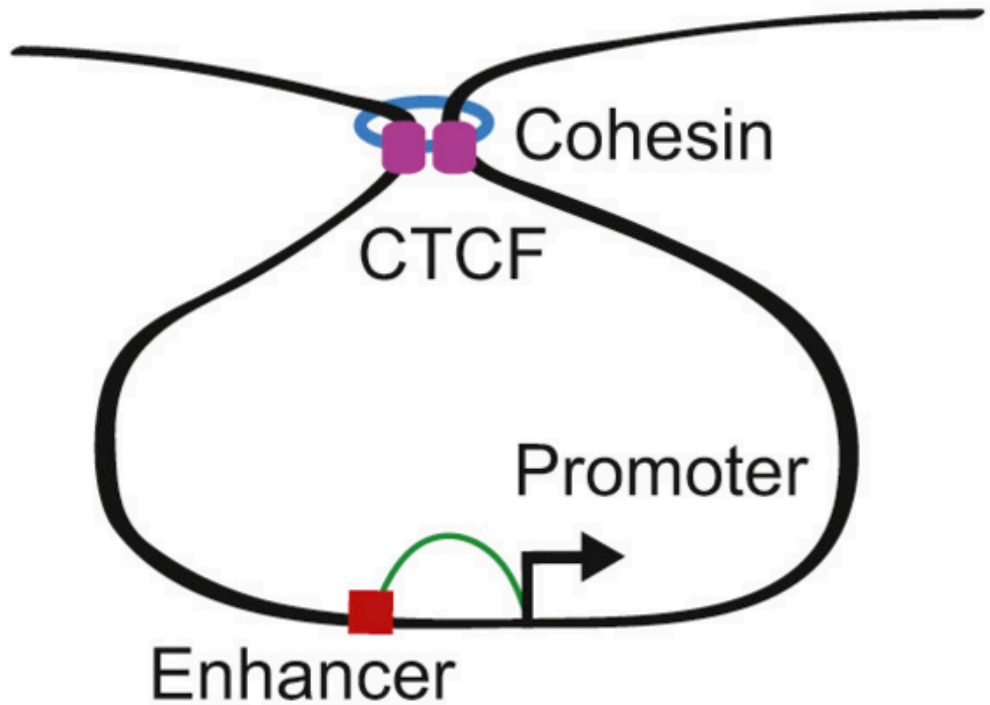
Total genome coverage



Linking genes to enhancers

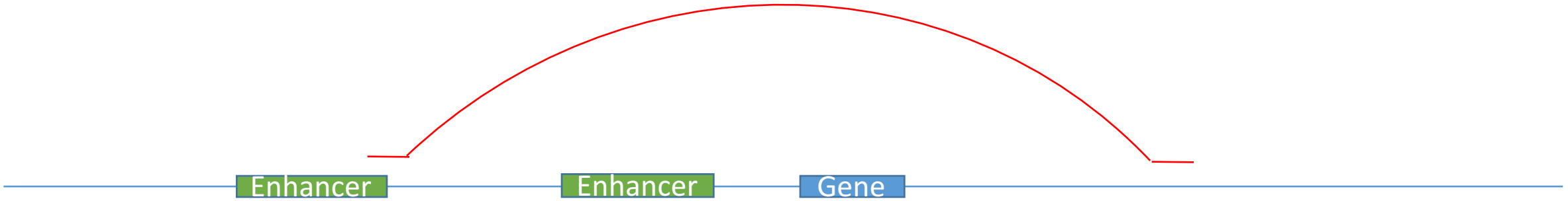


“Point-to-point” (P2P)



Ji et al 2016 – 3D chromosome regulatory landscape of human pluripotent stem cells

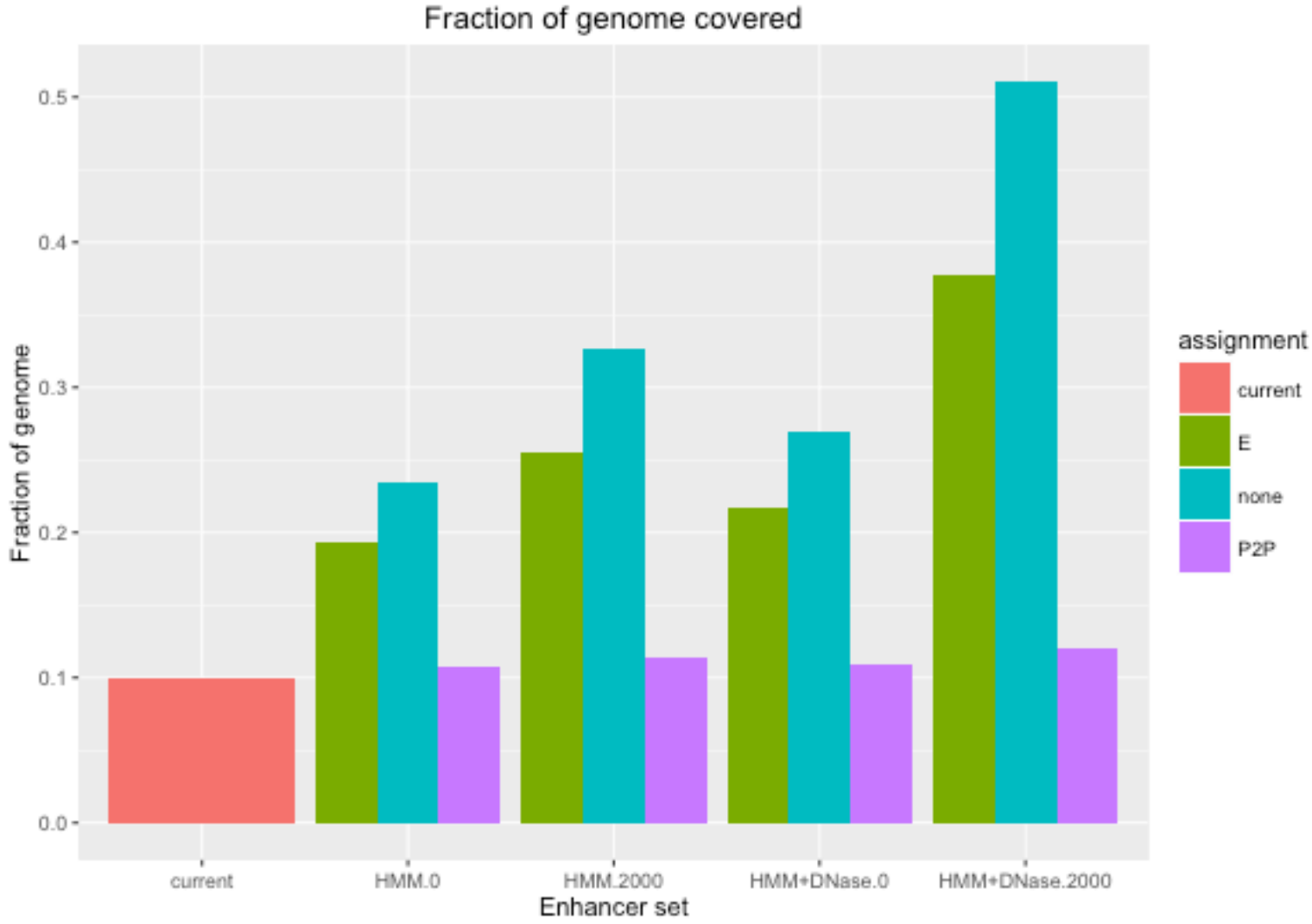
Linking genes to enhancers

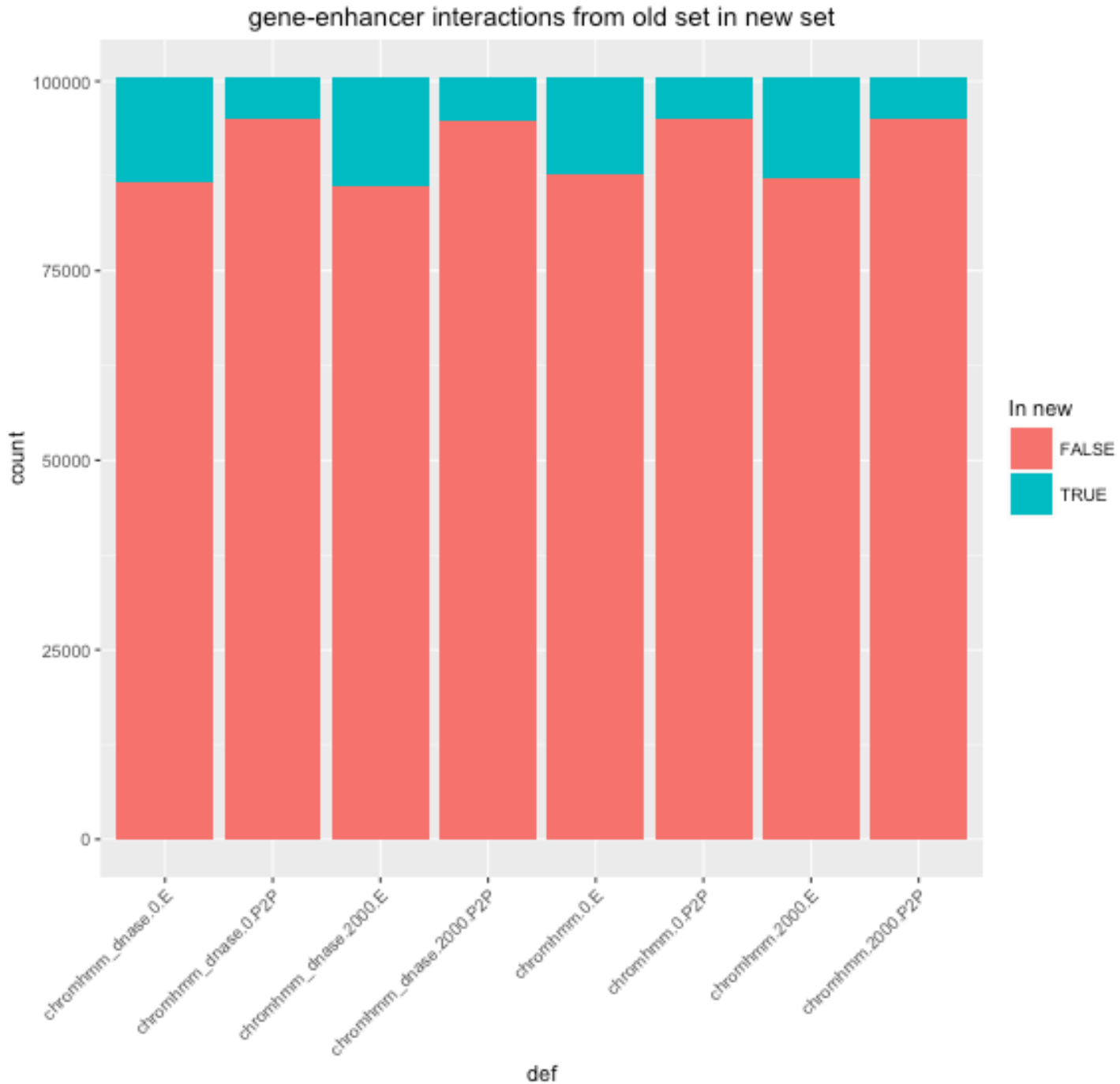


“Encompassing” (E)

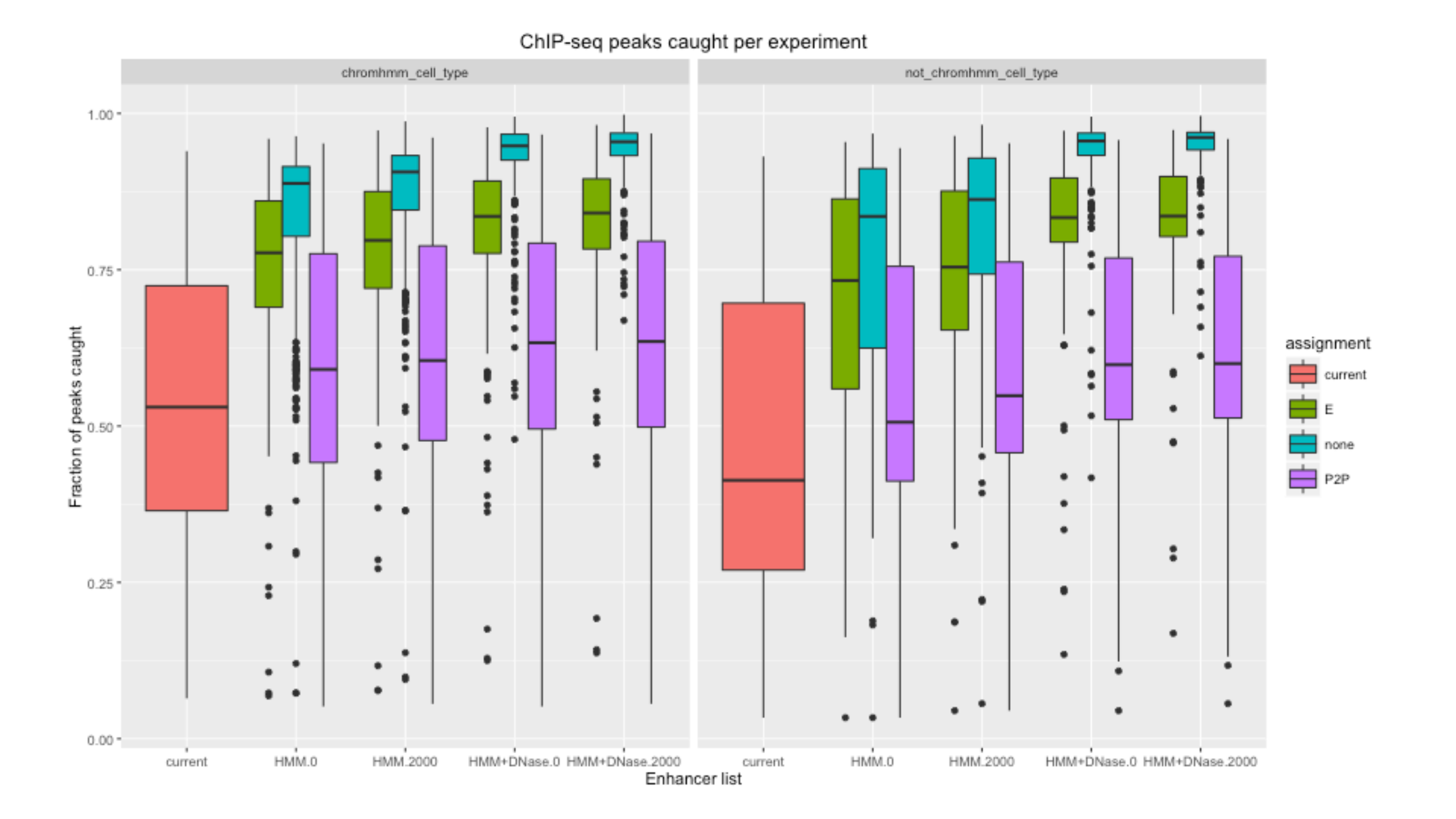
(CTCF/Cohesin ChIA-PET only)

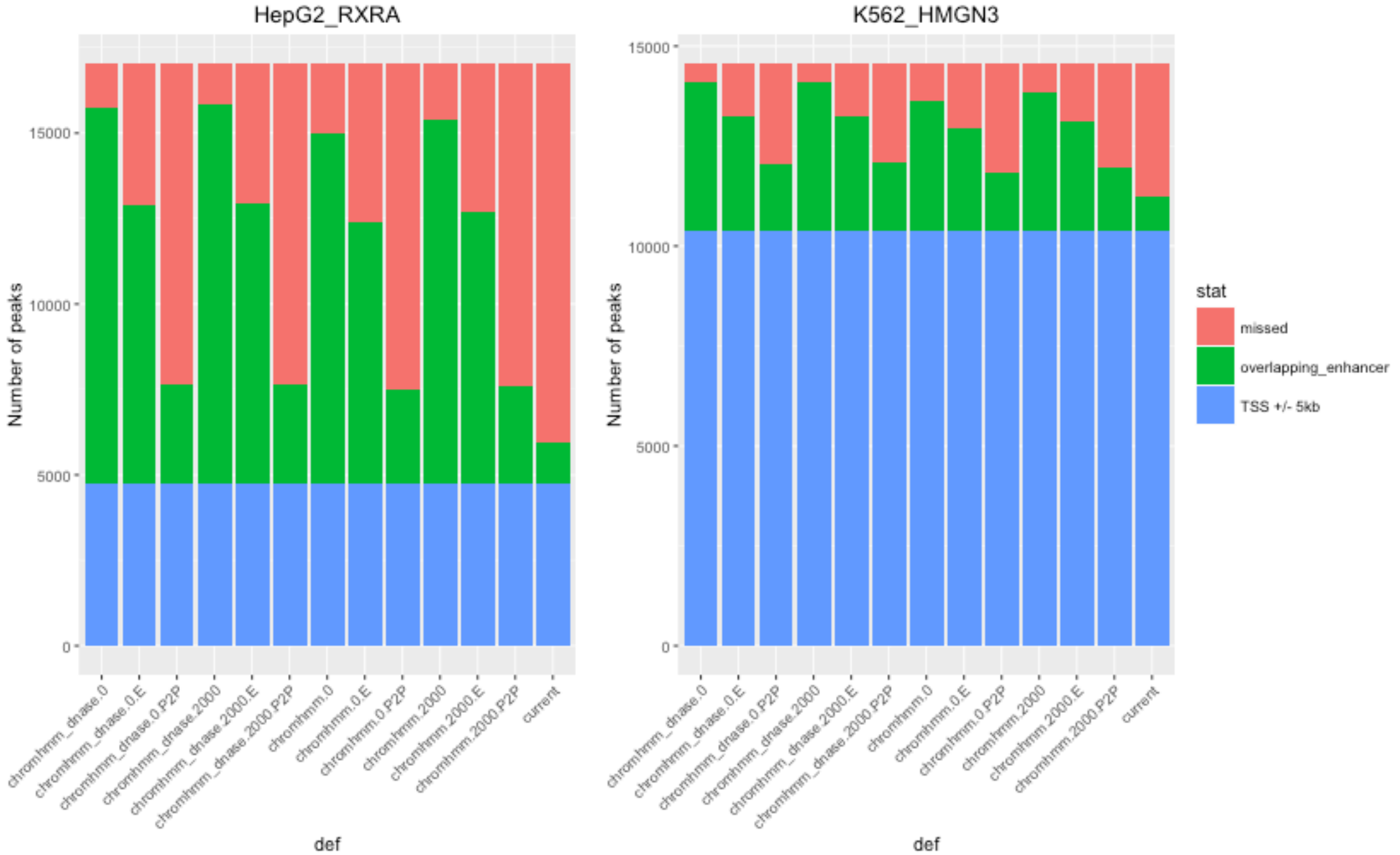
Total genome coverage



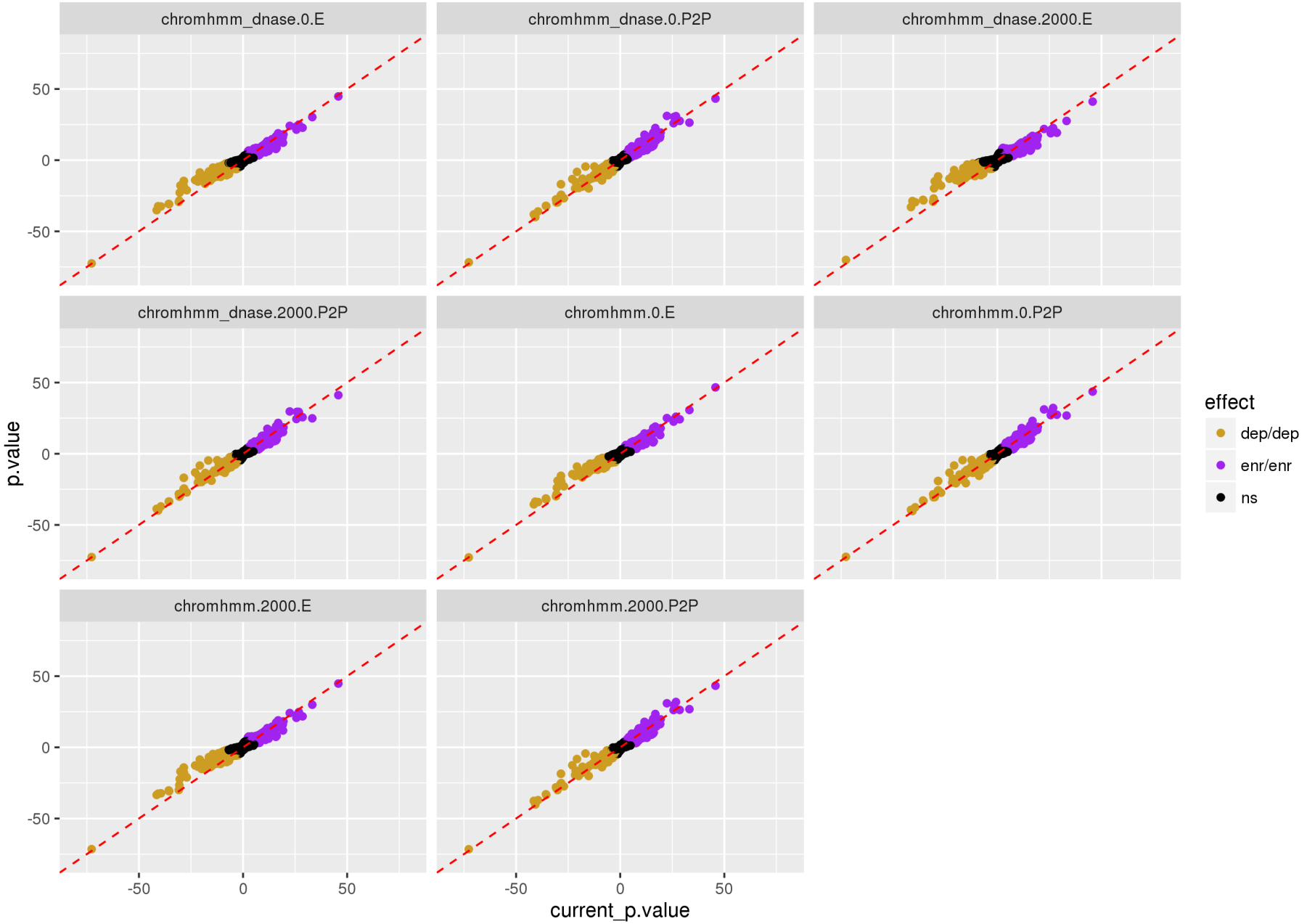

gene-enhancer interactions from old set in new set
100000
75000
50000
25000
0
count
chromhmm_dnase.0.E
chromhmm_dnase.0.P2P
chromhmm_dnase.2000.E
chromhmm_dnase.2000.P2P
chromhmm.0.E
chromhmm.0.P2P
chromhmm.2000.E
chromhmm.2000.P2P
def
In new
FALSE
TRUE

How well do these definitions capture peaks?


ChIP-seq peaks caught per experiment
chromhmm_cell_type
not_chromhmm_cell_type
Fraction of peaks caught
1.00
0.75
0.50
0.25
0.00
current
HMM.0
HMM.2000
HMM+DNase.0
HMM+DNase.2000
Enhancer list
assignment
current
E
none
P2P


HepG2_RXRA
K562_HMGN3
Number of peaks
15000
10000
5000
0
def
chromhmm_dnase.0
chromhmm_dnase.0.E
chromhmm_dnase.0.P2P
chromhmm_dnase.2000
chromhmm_dnase.2000.E
chromhmm_dnase.2000.P2P
chromhmm.0
chromhmm.0.E
chromhmm.0.P2P
chromhmm.2000
chromhmm.2000.E
chromhmm.2000.P2P
current
stat
missed
overlapping_enhancer
TSS +/- 5kb

Chipenrich


chromhmm_dnase.0.E
chromhmm_dnase.0.P2P
chromhmm_dnase.2000.E
chromhmm_dnase.2000.P2P
chromhmm.0.E
chromhmm.0.P2P
chromhmm.2000.E
chromhmm.2000.P2P
p.value
current_p.value
50
0
-50
effect
dep/dep
enr/enr
ns

Overview

This repository allows one to repeat much of the analysis that I did in my rotation in the Sartor lab in May/June 2016.

It includes the following:

- Create new enhancers lists using ENCODE ChromHMM/DNase-seq data
- Process ENCODE ChIA-PET data using Mango / download other processed ChIA-PET data from other sources
- Use processed ChIA-PET data to link enhancers to genes
- Evaluate how well different enhancer lists capture ENCODE ChIP-seq peaks
- Run chipenrich using different enhancer lists and compare the results

I've set it up so that everything can be run using a series of `make` commands.

Instructions

Download public data

You'll need to start by downloading some public datasets (ENCODE ChIP-seq, ChromHMM tracks, and master DNase). You can do this from this directory by running:

```
make chipseq # To fetch ENCODE chip-seq data
make dnase # To fetch ENCODE master DNase data
make chromhmm # To fetch the ~9 cell types with ChromHMM results on ENCODE
```

Generate enhancer lists

Once these have been downloaded, you can generate enhancer lists:

```
cd new_enhancer_lists
make enhancer_lists
cd ..
```

As it is now, this will generate four different enhancer lists:

- One list based on the ChromHMM enhancers alone
- One list based on the ChromHMM enhancers alone, extending enhancers less than 2kb in length to 2kb
- One list based on the ChromHMM enhancers unioned with the DNase regions (requiring that a DNase region shows up in at least 2 samples)
- One list based on the ChromHMM enhancers unioned with the DNase regions (requiring that a DNase region shows up in at least 2 samples), and extending enhancers less than 2kb in length to 2kb

Gather ChIA-PET interaction data

Next you need to make sure that you have ChIA-PET interaction lists (simple text files listing the significantly interacting regions based on ChIA-PET data). I've included all of the ones I've used in the `interaction_lists` directory (all files ending with *interactions), so if you don't want to change anything then you can skip this step. If you'd like to add new interactions files, just add them to this

Peak catching