Comparison of Massively Parallel Assays for Enhancer Activity

Anurag Sethi
P2-TECH
March 2015

A number of massively parallel assays have been developed in the last 5 years for testing enhancer activity

Technique	Plasmid/Chromatin	Length of element tested	Elements
In-vitro transcription (Shendure, Nat. Biotech, 2009)	In-vitro (100K)	200 bp / 3-4 promoters	Effect of variants
MPRA (Tarjei, Nature Biotech, 2012)	Plasmid/ human cells (40K) - RNAseq	87 bp / 2 enhancers	Effect of variants (indels/subs)
MPFD (Shendure, Nat. Methods, 2012)	Plasmid/ mouse	1kb /3-4 enhancers	Effect of variants
eFS (Bulyk, Nat. Methods, 2013)	Genome/fly 1 clone per cell	1 kb/ ChIP-seq of TF	Finding enhancers
STARR-Seq (Stark, Nature, 2013)	Plasmid/fly	600 bp/whole genome	Finding enhancers
CRE-Seq (Cohen, Genome Res, 2014)	Plasmid/human	132 bp/chromHMM and Segway	Accuracy of predictions
FIREWACh (Dailey, Nature Methods, 2014)	Genome/mouse 1 clone per cell	100-300bp/DNase	Finding enhancers
SIF-Seq (Pennacchio, Nature Methods, 2014)	Genome/mouse 1 clone per cell2	1-2 kb/specific regions of genome	Finding enhancers

Pros and cons of method

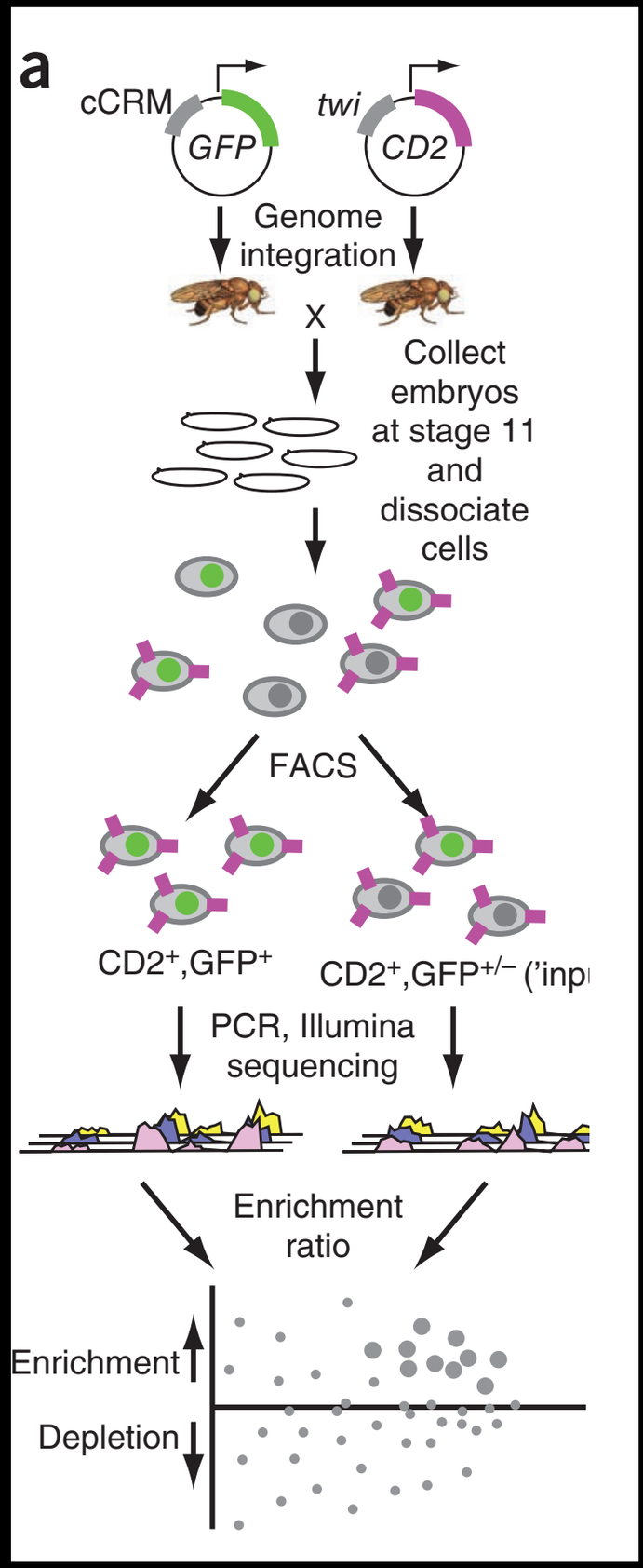
These assays are the only tests for high-throughput functional enhancer validation or for measuring the effect of sequence on these enhancers.
All these methods are supposed to be specific (so a positive is presumably a functional enhancer).

All methods do not contain 3D information of native chromosome.
All methods do not work for cooperative enhancers.
These methods are not sensitive (negatives may be positives).

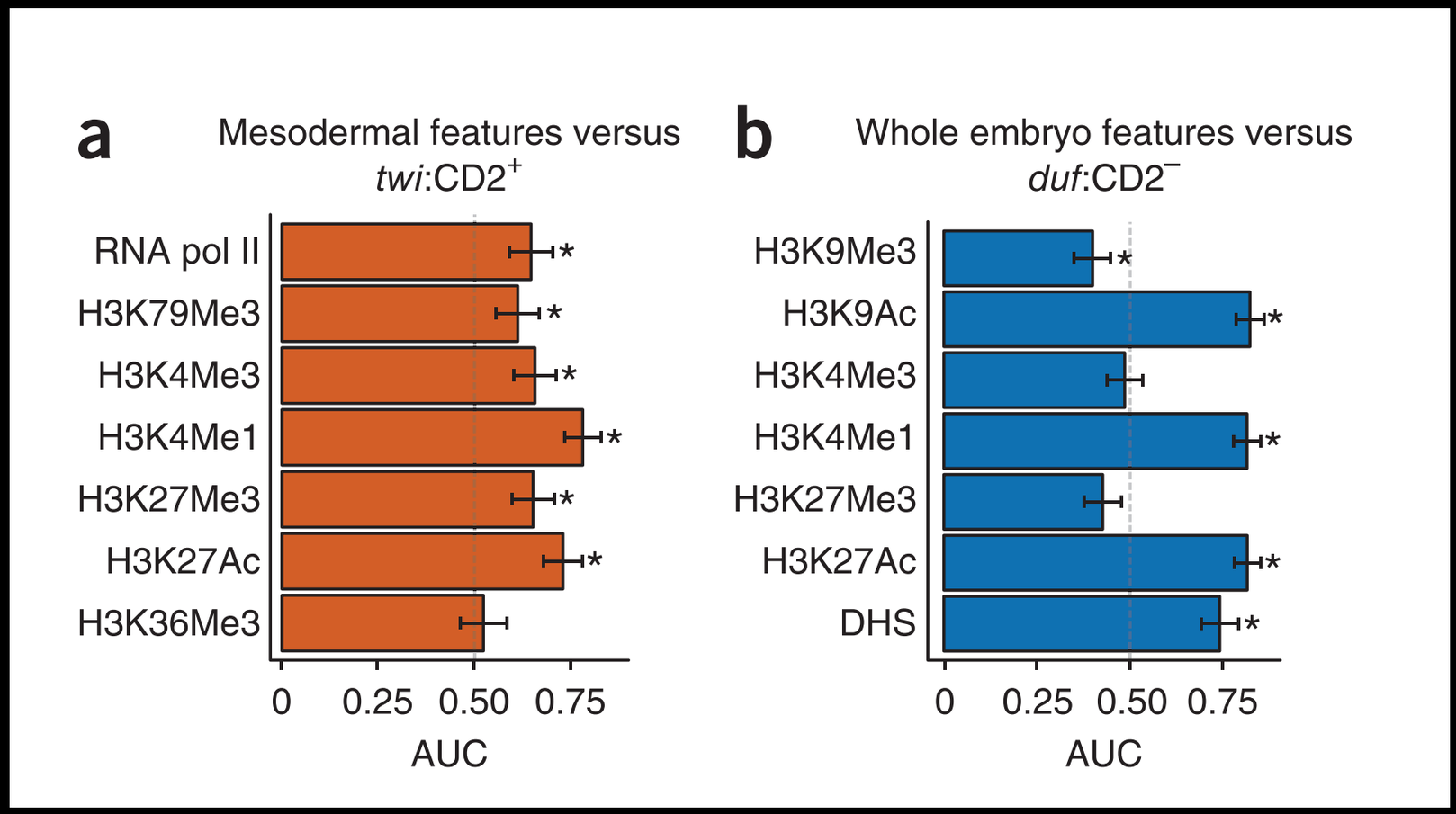
Methods that work for plasmids do not contain epigenetic information.
Methods that work with integration into genome can contain noise due to random insertion into functional regions (repetition is key).
Methods that reduce the tested region to < 500 bp are seriously limiting for mammalian enhancers.

eFACS-seq (eFS) study design

- Enhancer candidates chosen by ChIP-Seq of mesoderm-specific TF-binding sites.
- Each candidate is approximately **1kb** long.
- Each candidate cloned upstream of Hsp70 promoter and EGFP gene. A second gene added in tissue specific manner to ensure that they can identify tissue-specific enhancers. Both genes **integrated** into host genome.
- 100s of candidates in each tissue.
- Only **one** clone gets integrated per cell.
- 61 traditional enhancer assays with majority reporting enhancer activity.



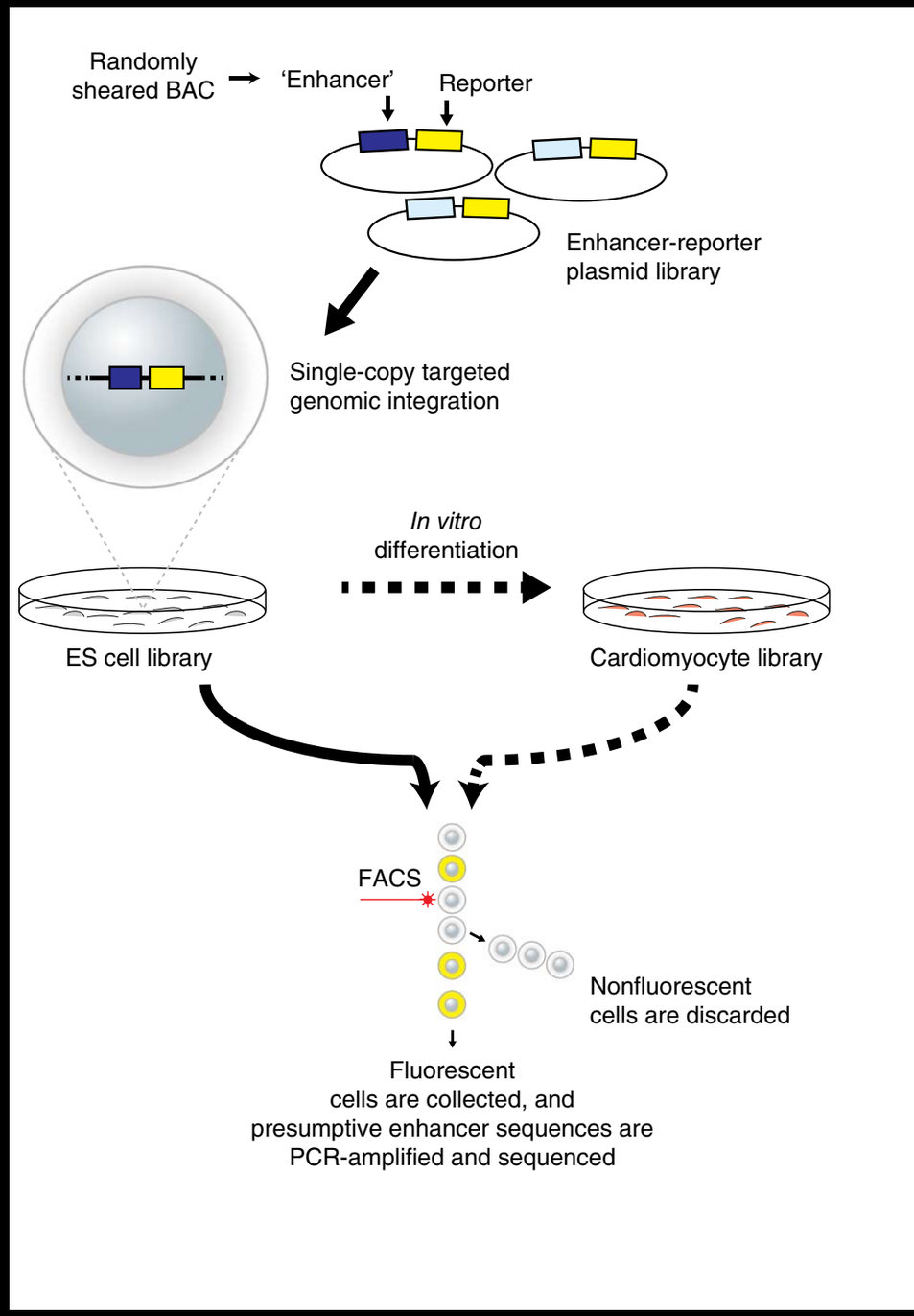
eFACS-seq (eFS) results



In addition to activating marks, repressive marks can also be enriched in positives.

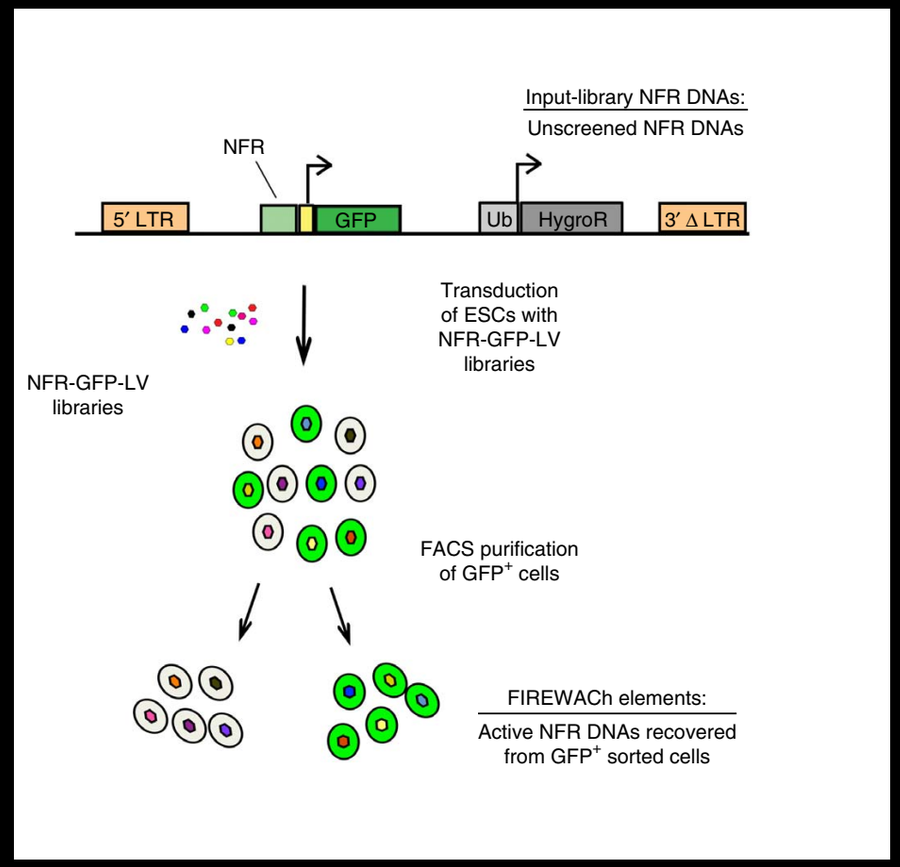
SIF-Seq study design

- Enhancer candidates chosen based on locii of interest in mouse ES cell (50kb) (and ChIP-Seq) and human heart/neurons (160 kb).
- Create BAC with mouse DNA and then randomly shear into approx. **1 kb length**.
- Integrated into plasmid close to a minimal promoter and YFP.
- **Integrated into X chromosome.**
- **One potential enhancer** per cell.
- FACS to sort cells expressing YFP.
- Small population of cells show positive enhancer activity.
- Amplified positive enhancer sequences with PCR using primers recognizing the flanking sequences.
- Unsorted sample used as input.
- Tested enhancer activity using traditional assays.



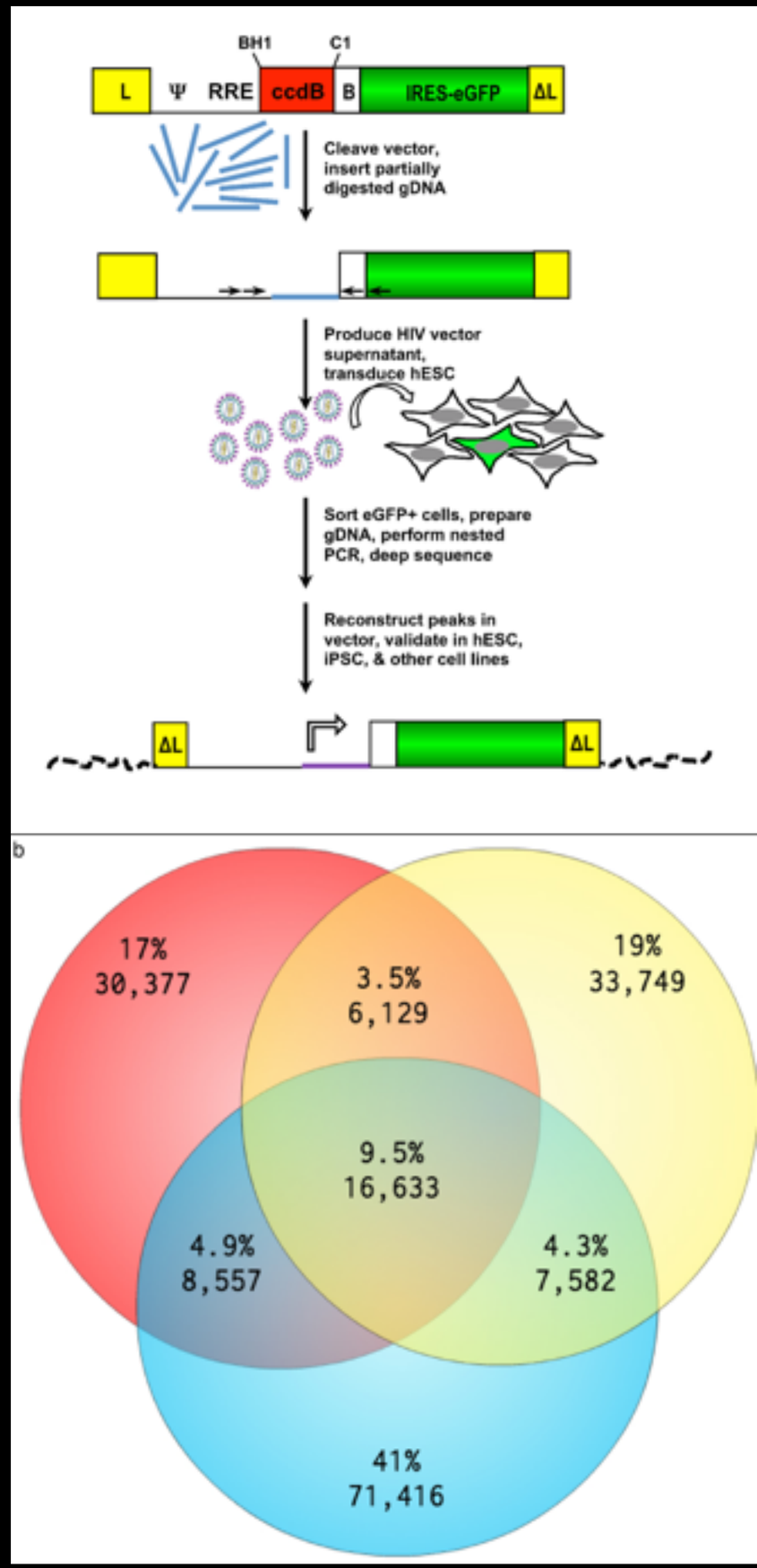
FIREWACh study design

- Enhancer candidates chosen based on open DNA in cell-line (murine ESC).
- Integrated into virus particles close to a minimal promoter and GFP.
- **Integrated into genome randomly with 1 clone per cell (H1-hESC).**
- **One potential enhancer of length 100-300 bp** per cell.
- FACS to sort cells expressing GFP.
- Small population of cells show positive enhancer activity.
- Amplified positive enhancer sequences with PCR using primers recognizing the flanking sequences.
- Tested enhancer activity using traditional assays.



Sutton-Seq study design

- Enhancer candidates chosen by shredding whole genome - **2.5 kb length**.
- Integrated into virus particles close to a minimal promoter and GFP.
- **Integrated into genome randomly with 1-5 clones per cell (H1-hESC). This leads to carrier plasmids in a single cell but the hope is that different replicates will not have the same DNA regions as carrier enhancers.**
- FACS to sort cells expressing GFP.
- ?? population of cells show positive enhancer activity. about 70K peaks per replica but only 17K peaks common to all three replicas (treated as single experiment).
- Amplified positive enhancer sequences with PCR using primers recognizing the flanking sequences.
- Tested enhancer activity using similar assays.



What Sutton has asked of us?

Figure out how to improve analysis of data?
Can we do further bioinformatic analysis of data?

Pros and Cons of this Method

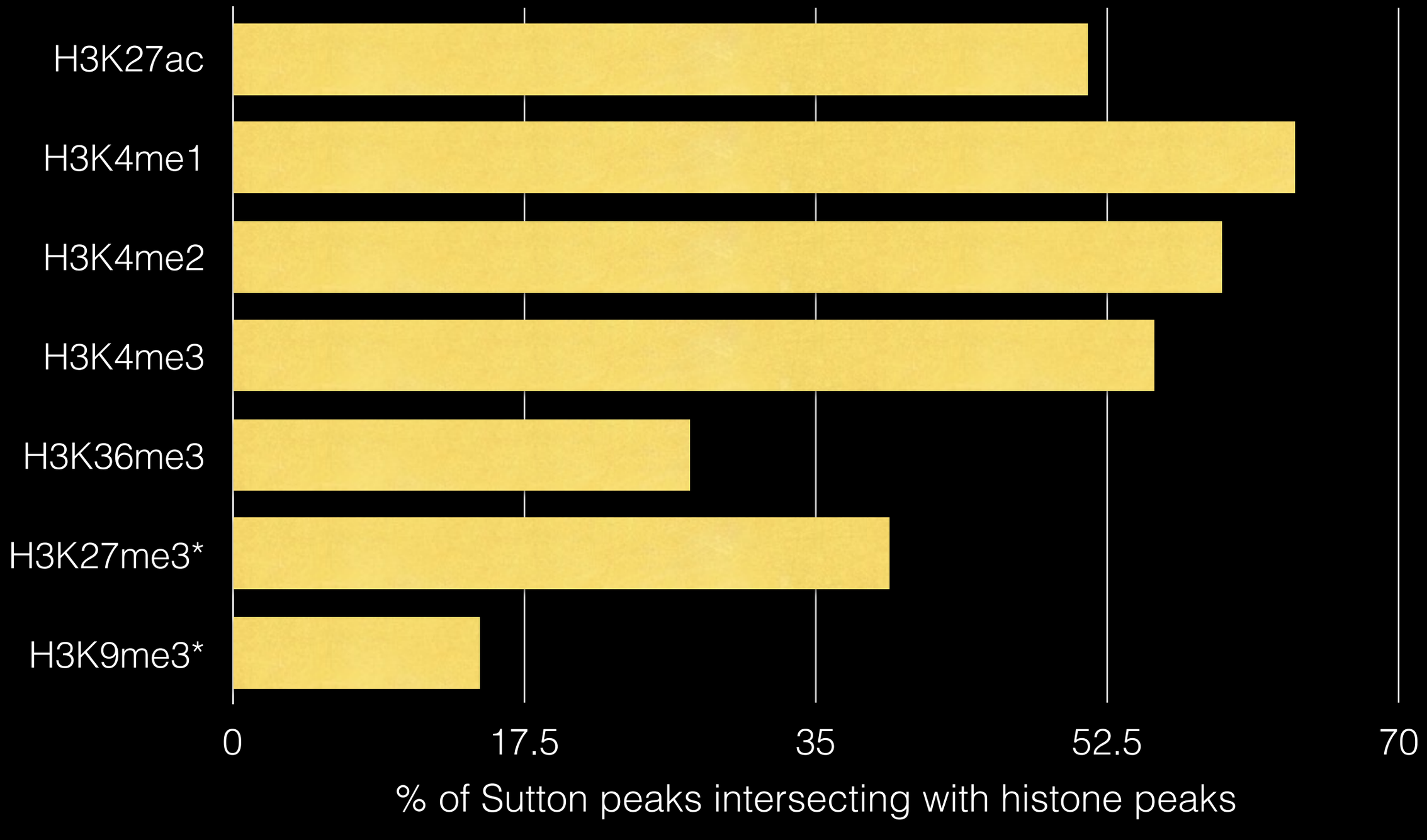
It is the first method to test enhancer activity on a genome wide scale for humans.

Can the problem of carrier enhancers be solved using 3 replicates in an experiment? - Don’t know at the moment!

Improving analysis of paper

Multimapping.
PCR duplicates.
Peak calling?
Simulations to find FDR when looking at intersection of peaks from 3 different replicates.
Post processing analysis - histone marks, TF binding, GRN

Any promise from current paper



<25% of peaks do not intersect with these histone peaks.
1914 peaks in DRM and 7734 peaks in PRM - ENCODE metatrack (55-60% of peaks).
49 peaks intersect with VISTA positives.

Extra Slides for Other Methods

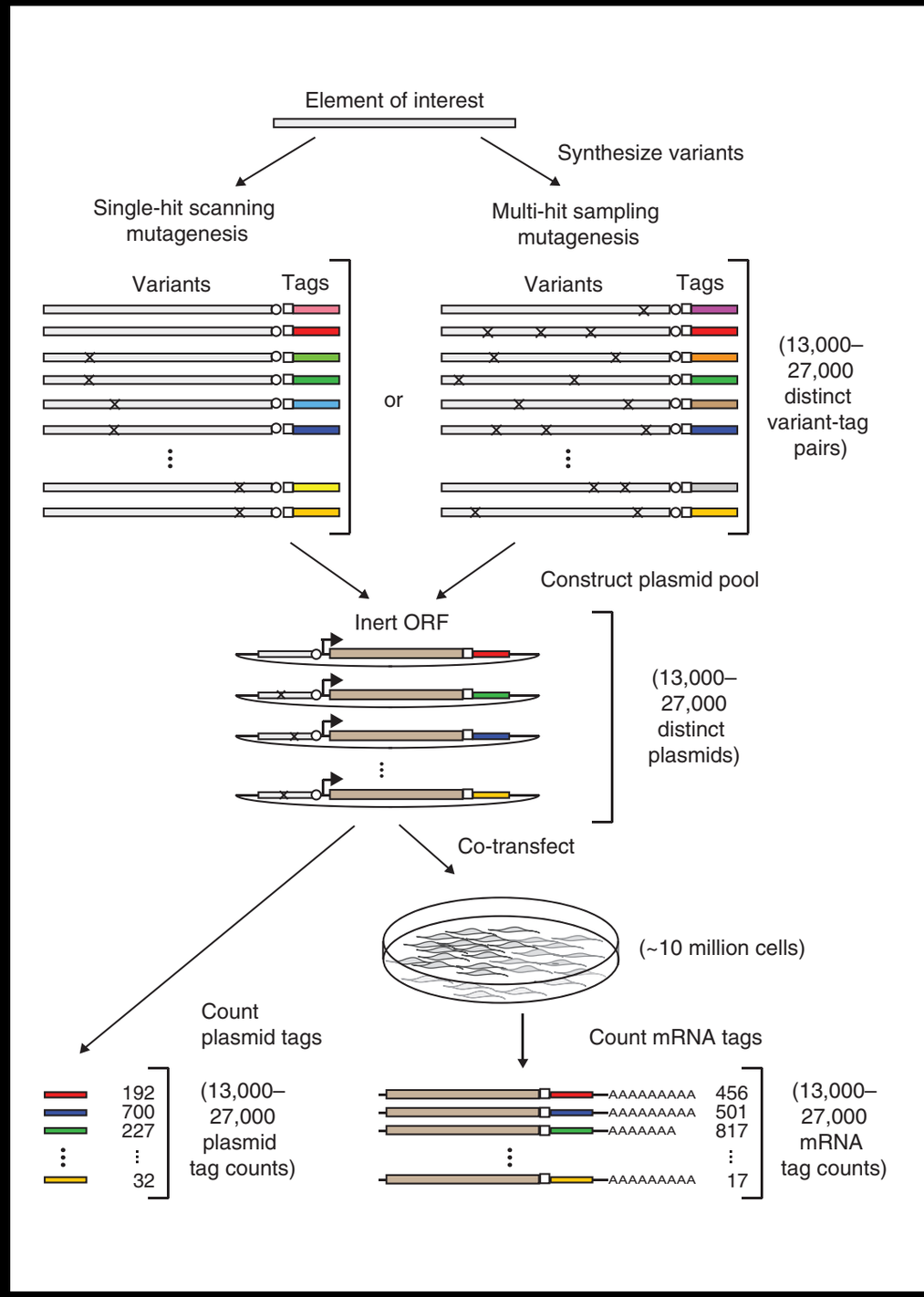
CRE-Seq study design

- 600 strong enhancers, 600 weak enhancers, and 300 repressed enhancers in **K562**. Also 600 predicted enhancers in **H1** that were not enhancers in **K562**.
- **Compared to random controls** (scrambled sequences with same di-nucleotide frequency). Total number of elements = 3237.
- Predictions inserted into construct had a hard **130 bp limit.** 300 of the 600 strong and 600 weak enhancers were full-length predictions while the rest were central region of the prediction. All 300 repressed region were central regions of the prediction.
- Cloned upstream of Hsp68 minimal promoter with a unique sequence barcode on 3' UTR.
- RNA-seq to figure out which enhancers were active.
- Replicates show correlation of 0.95-0.97!!
- Luciferase assays to measure expression driven by 12 CREs with a minP basal promoter.
- 26% of enhancers are active and repressors do not repress activity.
- Weak enhancers stronger than strong enhancers in assay.
- DHS best indicator of enhancer activity. Adding chromatin features and TF motifs improves the model.

Barak Cohen (Genome Research 2014)

MPRA study design

- Well studied enhancers chosen for variant study (NFKB and CRE) - single mutation, deletion, insertion, and multihit mutations.
- Create plasmid with variants added randomly. **87 bp length**. (28000 variants).
- Integrated into plasmid close to a minimal promoter and luciferase gene in human embryonic kidney cell-line.
- RNA-Seq of PCR-amplified RNA to measure the effect of mutations.
- Developed linear models to measure the effects of mutations on enhancer activity.
- The largest effect observed in activity is only 2-fold in most drastic case but inducibility changed by a higher amount.



MPFD study design

- 3 well studied enhancers chosen for variant study - single mutation, and average of 2 mutations.
- Create plasmid with variants added randomly. **< 1kb length**. (>100K variants).
- Integrated into plasmid close to a minimal promoter and luciferase gene in mouse tail vein cells.
- DNA-Seq (cDNA) of PCR-amplified RNA to measure the effect of mutations.
- Substitutions tend to have additive effects.