Distal regulatory modules (first draft)

Jing and Kevin

METHODS

Method overview

- Identify distal regulatory modules (DRMs)
  - Distal to avoid influence from enclosing gene
- Determine expression levels of genes
- Perform filtering
- Find DRM *r* and gene *g* where the signal of a histone mark at *r* correlates (positively or negatively) with the expression of *g* across multiple cell lines
- Find TFs that bind to *r* in cell lines with strong signal of the histone mark as potential regulators of *g*

Identifying DRM

- Use TF binding data to train a model for binding active regions (BARs). Use it to find BARs in the whole genome.
- Filter out regions within 10kb of annotated genes.

Basic filtering

- Not to consider a (DRM, histone mark) pair if:
  - The histone mark signal is too low (<5) or changes too little in the related cell lines (<2 fold)
- Not to consider a gene if:
  - Its expression level is too low (<5) or changes to little in the related cell lines (<2 fold)

Additional filtering

- Filter out a (DRM, target gene) pair if

  (Applied, otherwise too many candidates:)

  - They are on different chromosomes
  - They are too far apart (100kb)

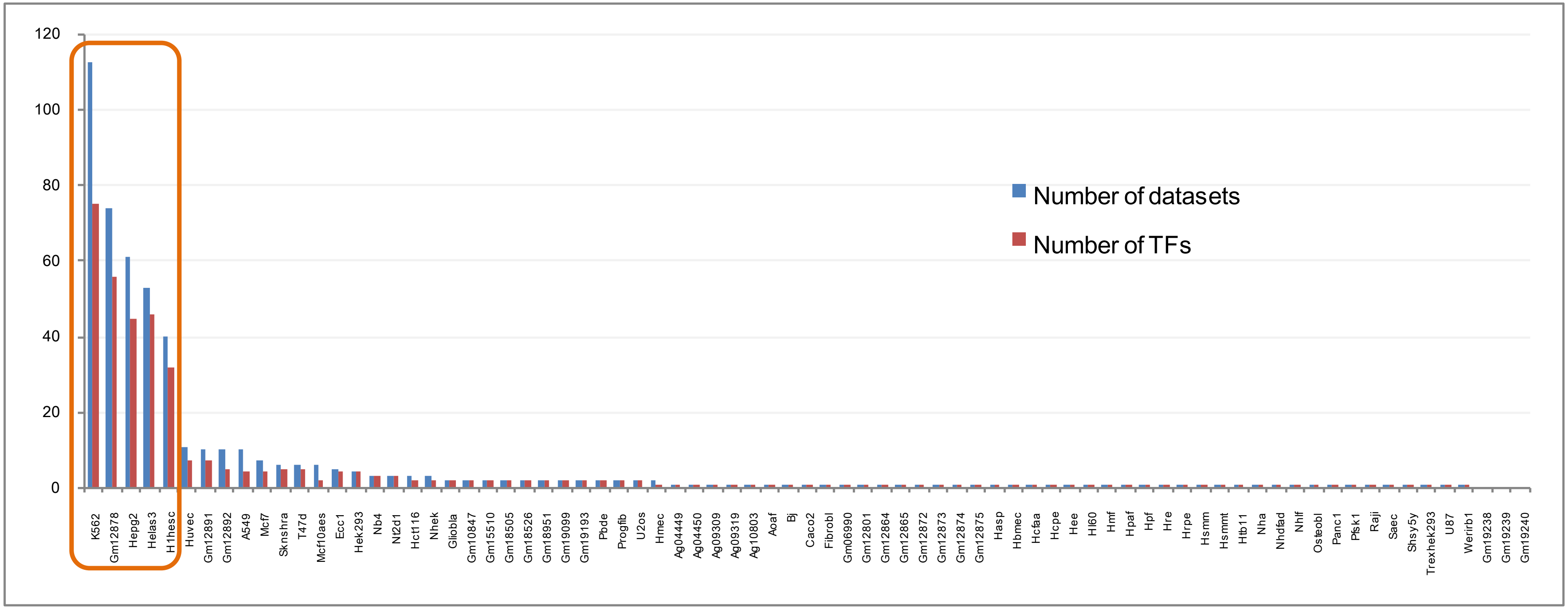
  (Not yet applied:)

  - There is CTCF binding between them in the cell lines that the DRM is supposed to be active
  - There is no long-range interaction data that supports the connection
  - There is no expression (eRNA) at the DRM in the cell lines that the DRM is supposed to be active
  - Absence of conserved motifs in DMRs

DATA

*See louise:/home/yy222/chromod/conf/human_grch37_jan2011.config for list of datasets and locations

TF datasets



- Decision: Use GM12878, H1-hESC, HeLa-S3, Hep-G2 and K562 to find BARs

Histone mark datasets

	Ag04449	Ag04450	Ag09309	Ag09319	Ag10803	Aoaf	Bj	Caco2	Gm06990	Gm12878	H1hesc	H7es	Hasp	Hbmec	Hcf	Hcfaa	Hcm	Hcpe	Hct116	Hee	Hek293	Helas3	Hepg2	Hl60	Hmec	Hmf	Hpaf	Hpf	Hre	Hrpe	Hsmm	Hsmmt	Huvec	Hvmf	Jurkat	K562	Mcf7	Nb4	Nha	Nhdfad	Nhdfneo	Nhek	Nhlf	Nt2d1	Osteobl	Saec	Sknshra	U2os	Dataset count	Cell line count
H2az										1													1								1	1				1									1				6	6
H3k27ac										1	1											1	1		1						1	1	1			1			1	1		1	1		1				14	14
H3k27me3							1	1	1	2	1	1										2	2		2				1		1		2			3			1	1		2	1	1		1	1		28	20
H3k36me3							1	1	1	2	1	1										2	2		1				1		1	1	2			2			1	1		2	1	1	1	1	1		28	22
H3k4me1										1	1											1	1		1						1	1	1			2			1			1	1	1	1				15	14
H3k4me2										1	1											1	1		1						1	1	1			1				1		1	1		1				13	13
H3k4me3	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	2	1	2	1	1	1	1	1	1	1	2	1	1	3	1	1	1	1	1	2	1	1		1	1		54	46
H3k79me2										1												1	1								1	1				1													6	6
H3k9ac										1	1											1	1		1						1	1	1			2				1		1	1	1					14	13
H3k9me1																																	1			1						1							3	3
H3k9me3										1																					1					1								1	1			1	6	6
H4k20me1										1	1											1	1		1						1	1	1			1						1	1						11	11

Correlation and sample size

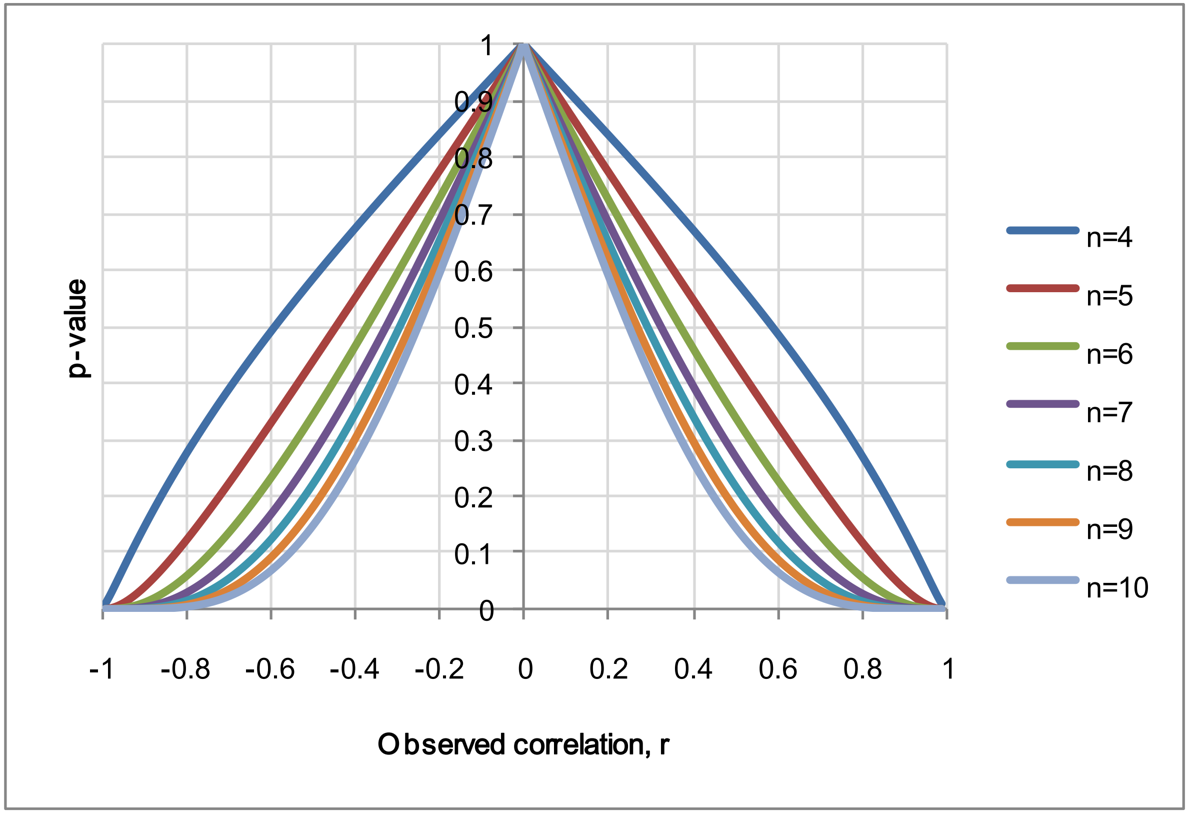
- Fisher transformation
  - Suppose the correlation between two random variables is r_0. For a sample size of n, let r be the observed correlation. The following function

    $$\mathrm{F(r)} = \frac{1}{2}\ln\frac{1+\mathrm{r}}{1-\mathrm{r}}$$

    approximately follows a Gaussian distribution with mean = $F(r_0)$ and standard deviation = $\frac{1}{\sqrt{\mathrm{n}-3}}$
  - Therefore to test the null hypothesis that the correlation is 0, the two-sided p-value can be computed as

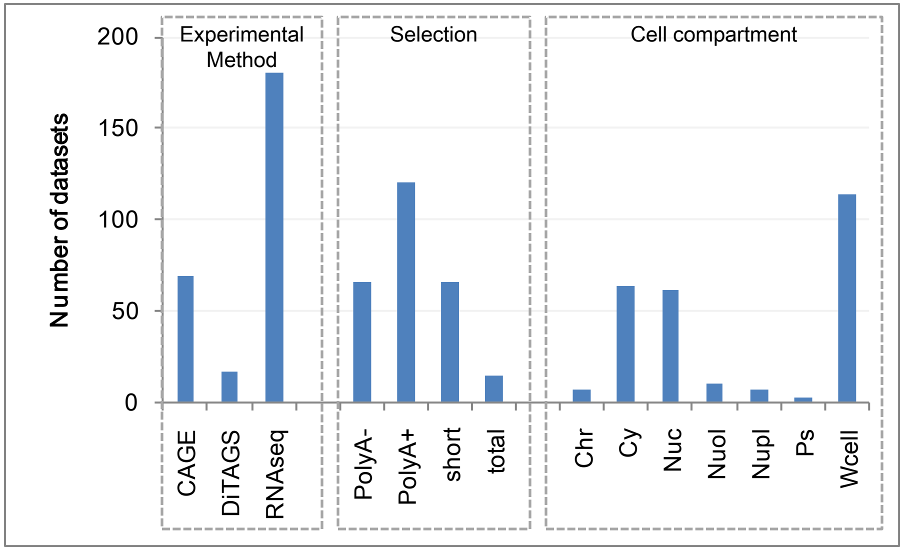
    $$2\Phi\left(-|\mathrm{F(r)}-\mathrm{F(0)}|\sqrt{\mathrm{n}-3}\right) = 2\Phi\left(-|\mathrm{F(r)}|\sqrt{\mathrm{n}-3}\right)$$

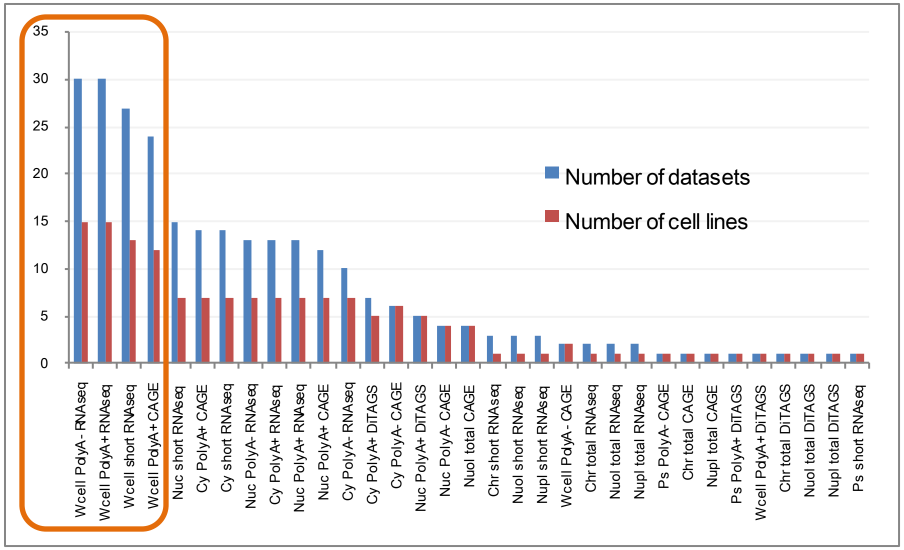
Correlation and sample size



- If multiple hypothesis testing correction is applied, need very extreme r to get a significant p-value
- Decision: Focus on cases with 7 or more cell lines

RNA datasets





- Decision: Focus on 4 combinations:
  - (Wcell, PolyA-, RNAseq)
  - (Wcell, PolyA+, RNAseq)
  - (Wcell, short, RNAseq)
  - (Wcell, PolyA+, CAGE)

Long-range interaction datasets

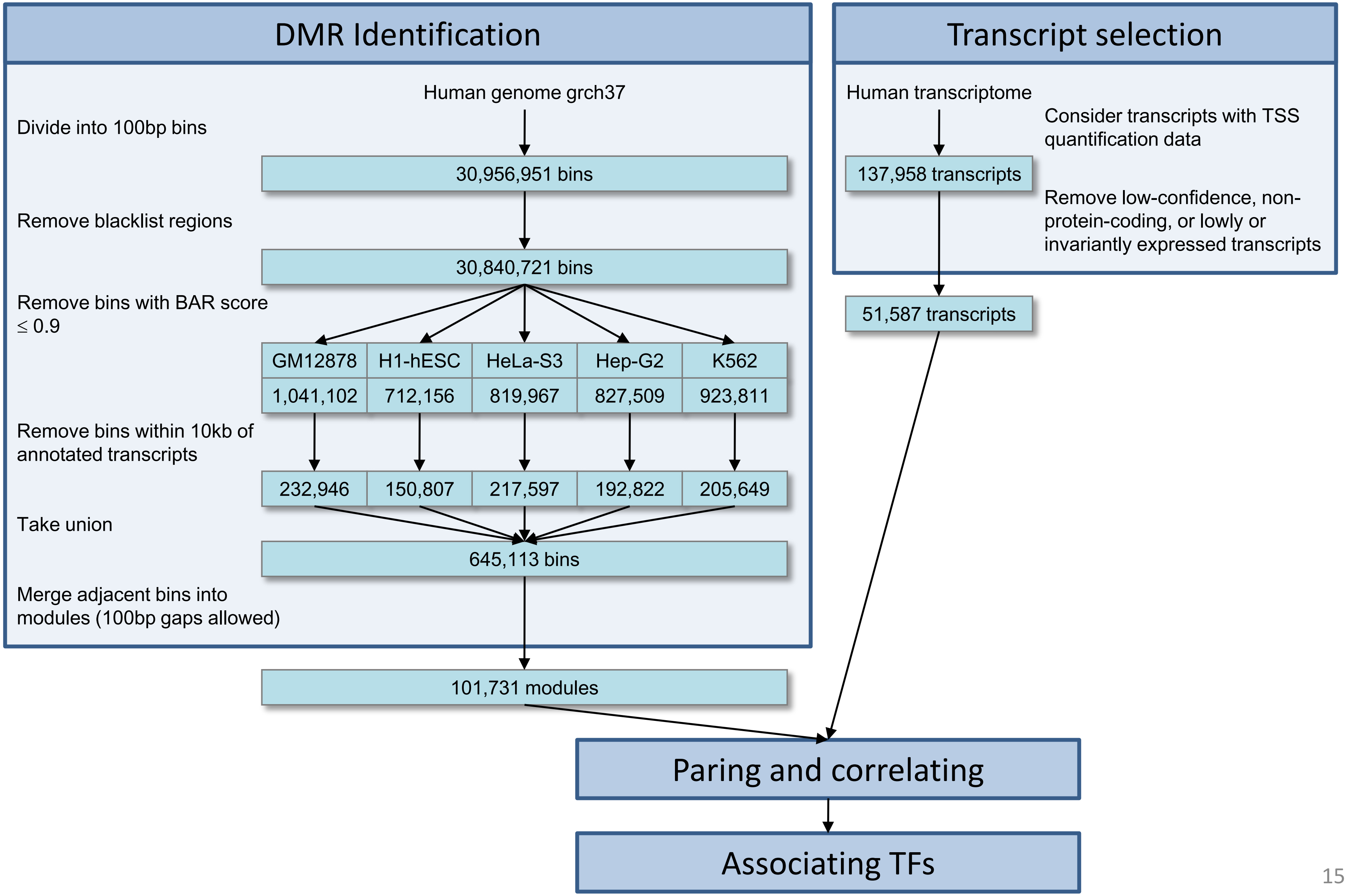
	Bj	Caco2	Gm06990	Gm12878	H1hesc	Hct116	Helas3	Hepg2	K562	Lncap	Mcf7	Nb4	Sknshra
GIS ChIA-PET (Pol2)						T	T		T		T	T	
UMass 5C				T	T		T		T				
Uw 5C	T	T	T	T				T	T	T	T		T

- Data quality not certain
- Decision: Use whenever possible

RESULTS

*Result files are stored in louise:/home/yy222/chromod/results/drm/

Pipeline (1)



DMR Identification

Human genome grch37

Divide into 100bp bins

30,956,951 bins

Remove blacklist regions

30,840,721 bins

Remove bins with BAR score ≤ 0.9

GM12878	H1-hESC	HeLa-S3	Hep-G2	K562
1,041,102	712,156	819,967	827,509	923,811

Remove bins within 10kb of annotated transcripts

232,946	150,807	217,597	192,822	205,649

Take union

645,113 bins

Merge adjacent bins into modules (100bp gaps allowed)

101,731 modules

Transcript selection

Human transcriptome

Consider transcripts with TSS quantification data

137,958 transcripts

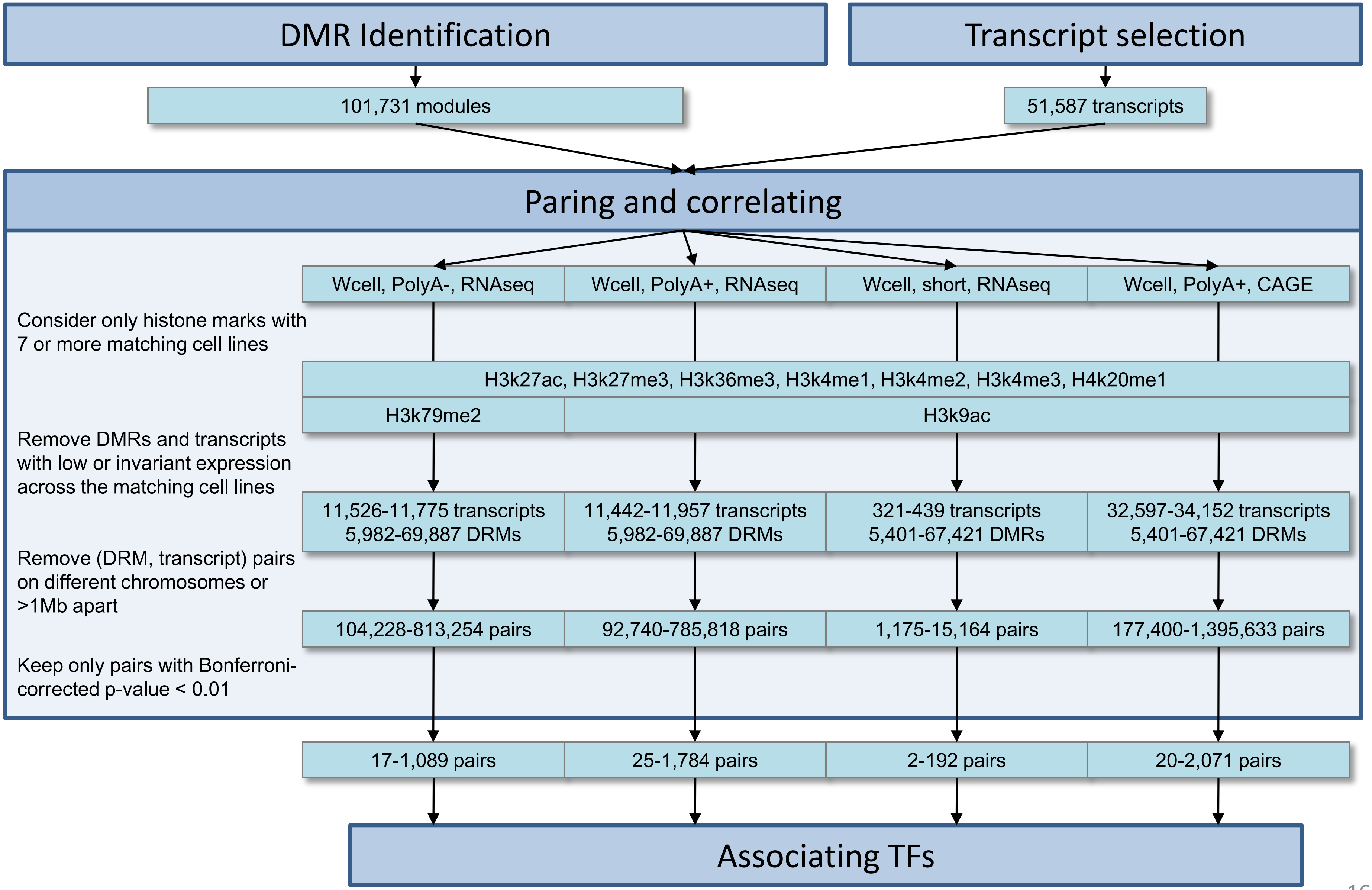
Remove low-confidence, non-protein-coding, or lowly or invariantly expressed transcripts

51,587 transcripts

Paring and correlating

Associating TFs

Pipeline (2)



DMR Identification

101,731 modules

Transcript selection

51,587 transcripts

Paring and correlating

Consider only histone marks with 7 or more matching cell lines

Remove DMRs and transcripts with low or invariant expression across the matching cell lines

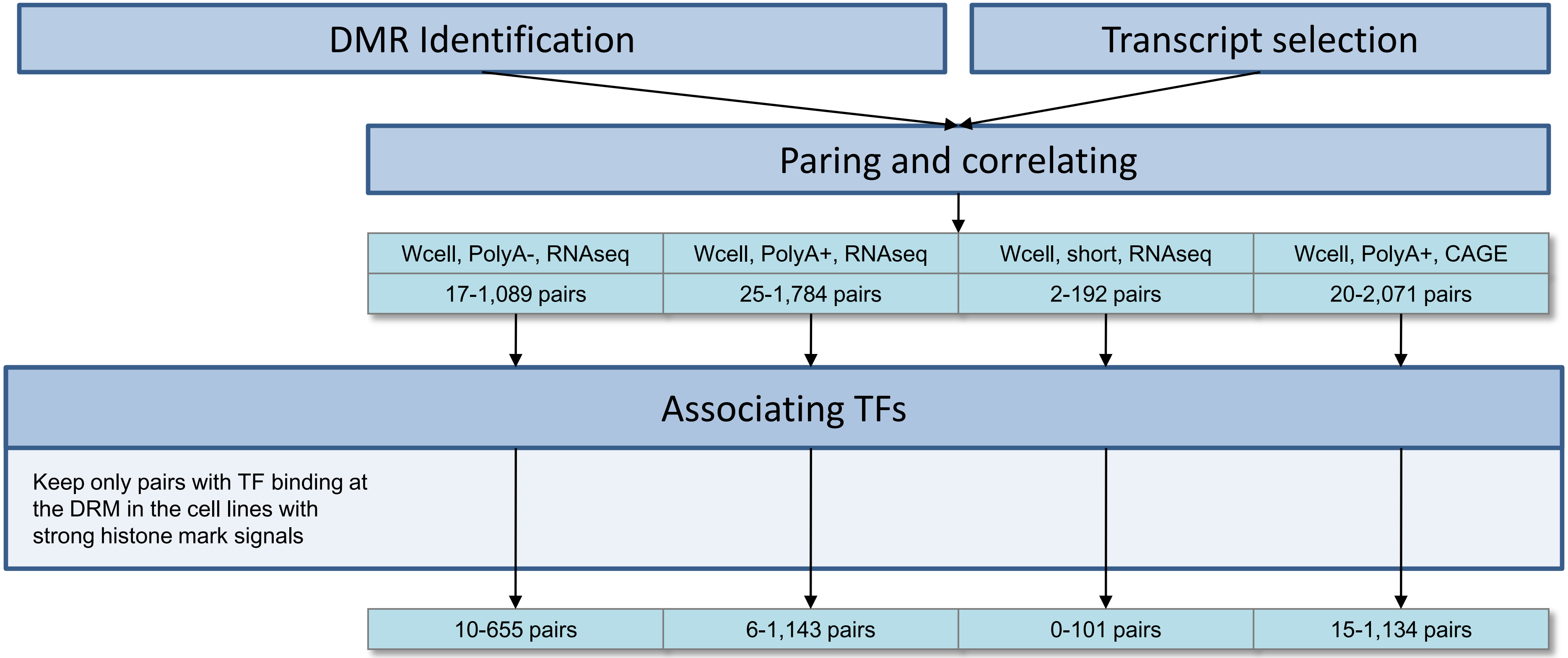
Remove (DRM, transcript) pairs on different chromosomes or >1Mb apart

Keep only pairs with Bonferroni-corrected p-value < 0.01

Wcell, PolyA-, RNAseq	Wcell, PolyA+, RNAseq	Wcell, short, RNAseq	Wcell, PolyA+, CAGE
H3k27ac, H3k27me3, H3k36me3, H3k4me1, H3k4me2, H3k4me3, H4k20me1			
H3k79me2	H3k9ac		
11,526-11,775 transcripts 5,982-69,887 DRMs	11,442-11,957 transcripts 5,982-69,887 DRMs	321-439 transcripts 5,401-67,421 DMRs	32,597-34,152 transcripts 5,401-67,421 DRMs
104,228-813,254 pairs	92,740-785,818 pairs	1,175-15,164 pairs	177,400-1,395,633 pairs
17-1,089 pairs	25-1,784 pairs	2-192 pairs	20-2,071 pairs

Associating TFs

Pipeline (3)



DMR Identification

Transcript selection

Paring and correlating

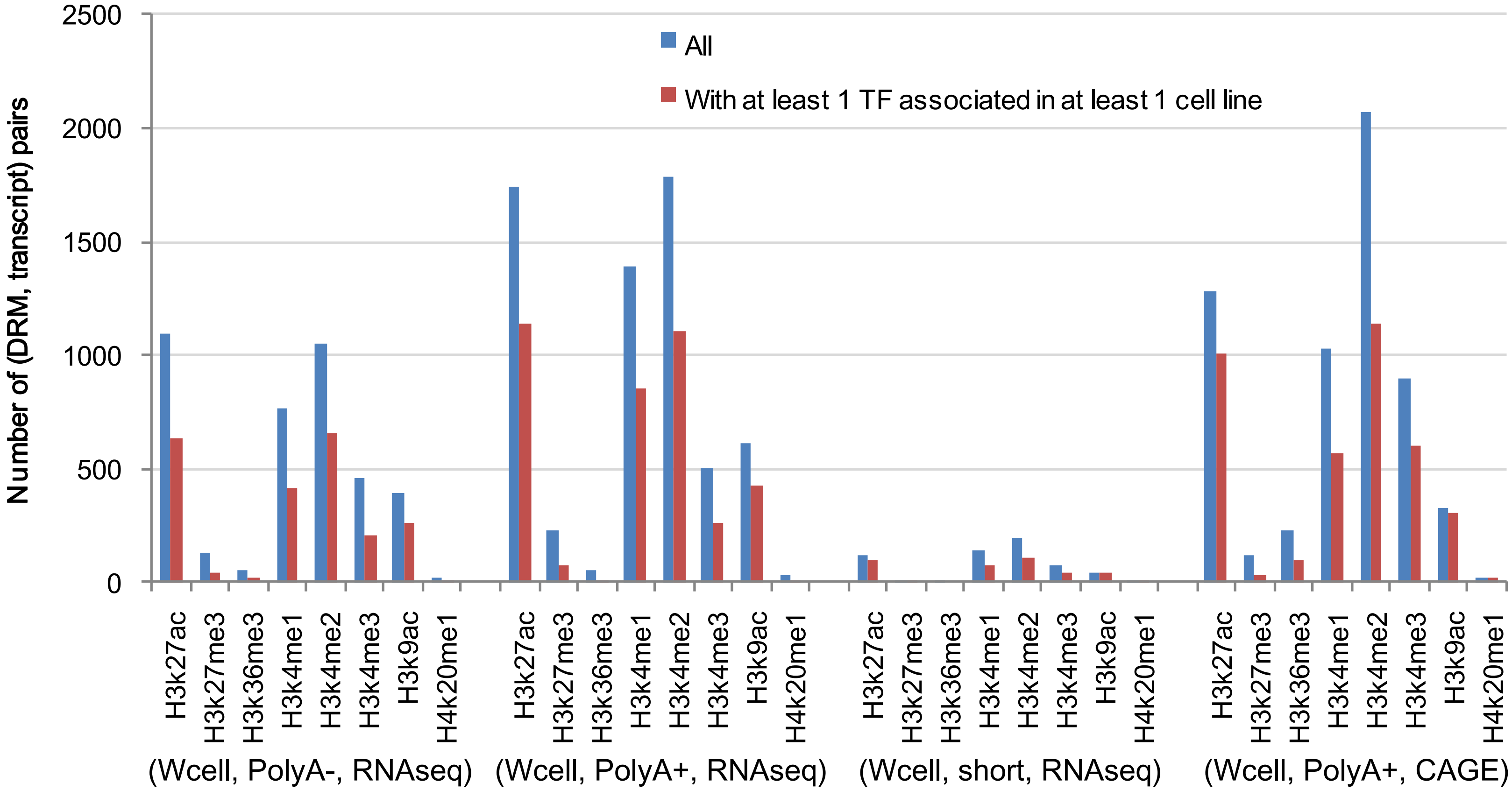
Wcell, PolyA-, RNAseq	Wcell, PolyA+, RNAseq	Wcell, short, RNAseq	Wcell, PolyA+, CAGE
17-1,089 pairs	25-1,784 pairs	2-192 pairs	20-2,071 pairs

Associating TFs

Keep only pairs with TF binding at the DRM in the cell lines with strong histone mark signals

10-655 pairs	6-1,143 pairs	0-101 pairs	15-1,134 pairs

Number of called pairs



Current limitations

- Assume (unrealistically) that gene expression has a simple correlation with histone mark at DRMs
  - A better model needs to consider at least histone mark signals and TF binding at promoters
- Use of distance threshold and overly stringent p-value cutoff to reduce the number of (DRM, transcript) pairs
  - Would be good if long-range interaction, eRNA and/or motif data can be used instead
- Ad-hoc thresholds
- Low DRM resolution (100bp units)
- Small number of matched cell lines