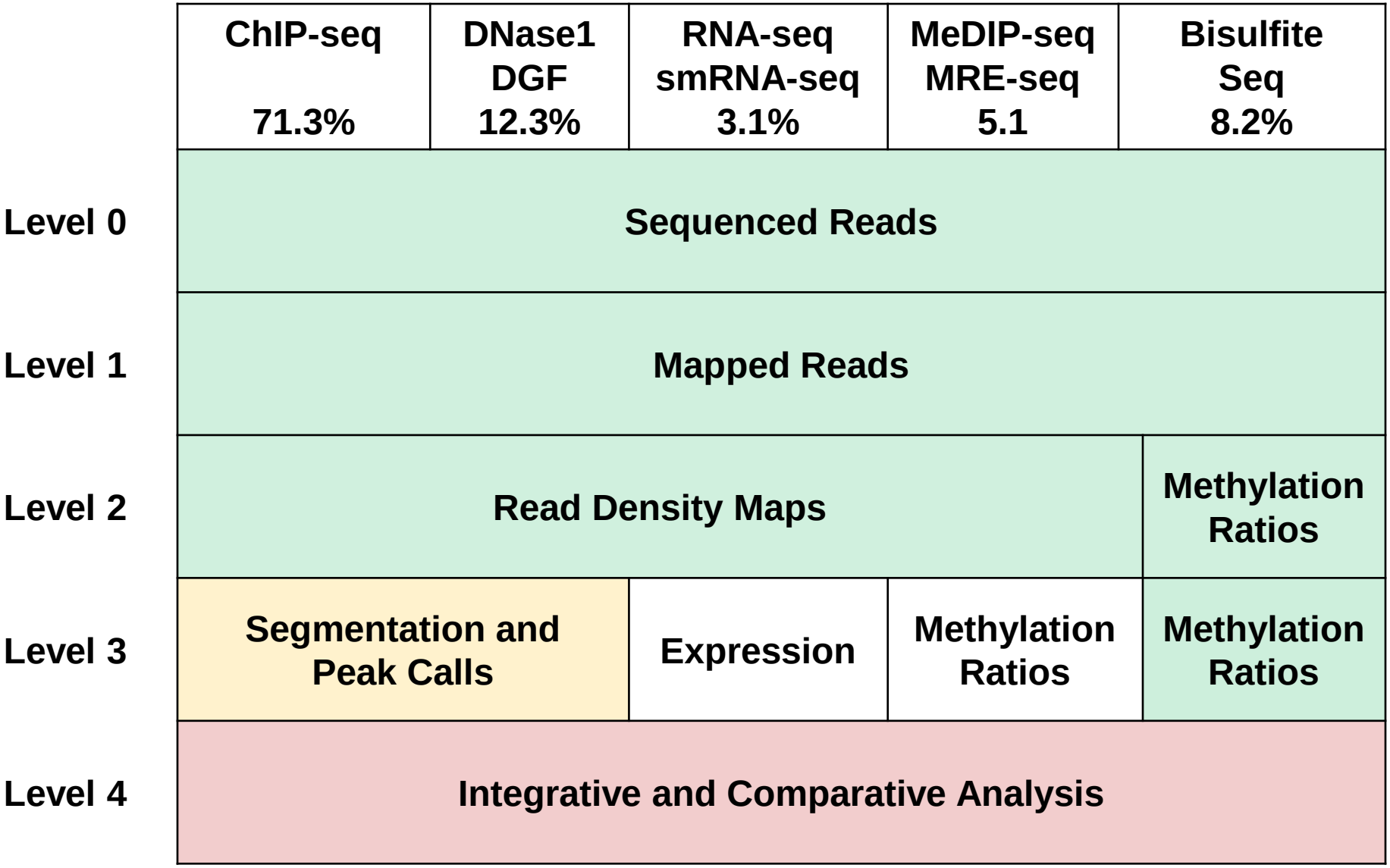
Strategies for Computing and Exposing Peaks for the 50 Epigenomes NIH Epigenomic Roadmap Data

Cristian Coarfa, Anshul Kundaje
Alan R Harris, Aleksandar Milosavljevic

Integrative and Comparative Analysis

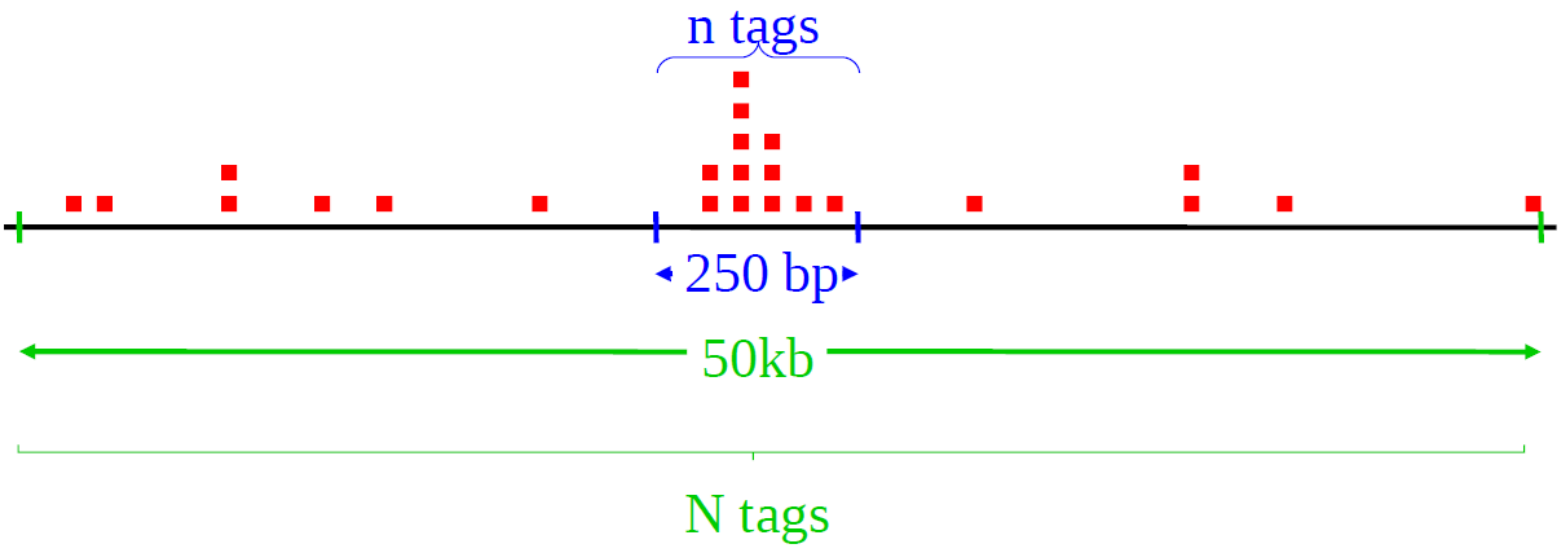


Basic Processing of Consortium Data

- Currently expose
  - Read mappings (BED)
  - Read density signals (WIG)

- Call discrete peaks
  - HotSpot
    - DNAse hypersensitivity
    - Digital Genomic Footprinting
  - IDR2 methodology
    - Punctate histone modifications
    - H3K4me3, H3K9ac, H3K27ac, H3K4me1, H3K4me2

HotSpot



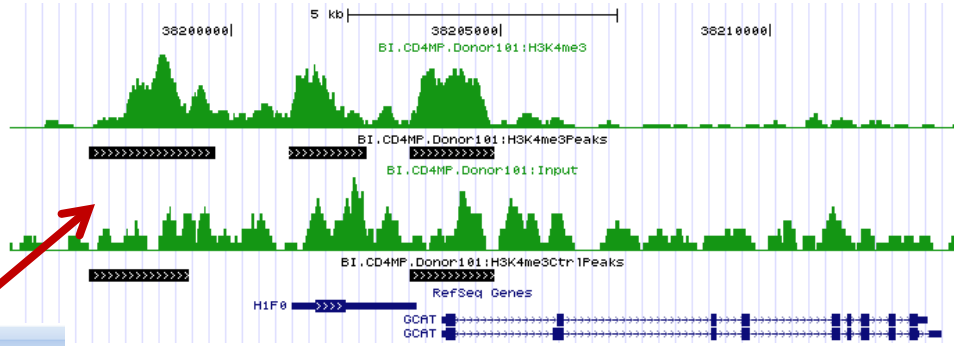
- Scan statistic gauging enrichment
  - z-score based on the binomial distribution.
- Binomial distribution
  - probability of n tags in small window given N tags large window.
  - adjust for local background fluctuations

IDR

- **I**rreproducible **D**iscovery **R**ate
- **Measuring reproducibility of high-throughput experiments,** Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Annals of Applied Statistics. Volume 5, Number 3 (2011), 1752-1779.
- IDR2
  - Manuscript in preparation
- Key changes
  - Optional use of technical replicates
    - Present in UCSD ChIP-Seq data only
  - Use of pseudo-replicates

Exposing Peak Calls

UCSC Visualization



Human Epigenome Atlas Release 5 (hg19)

View Selections In | Clear Selections | Save Selections

Assay Type

Sample Type
Filter: CD3 (e.g. "cell line")

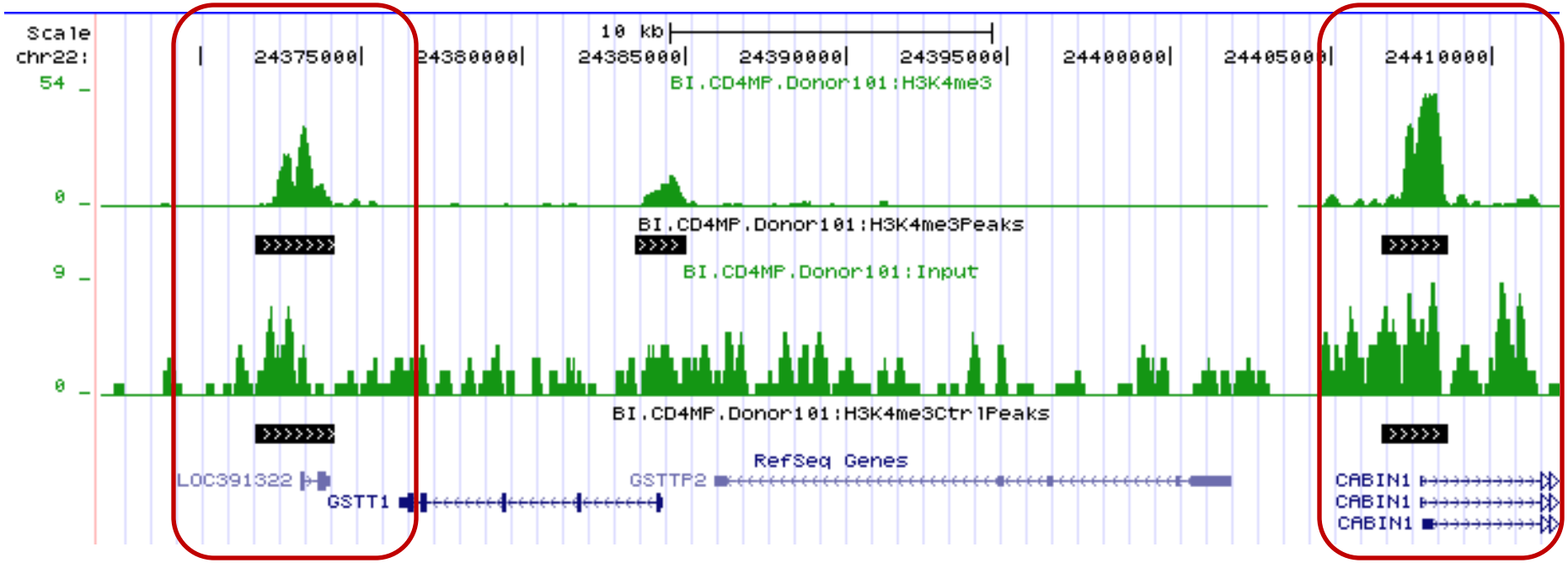
	Bisulfite-Seq	MeDIP-Seq	MRE-Seq	RRBS	DNase Hypersensitivity	Digital Genomic Footprinting	mRNA-Seq	smRNA-Seq	ChIP-Seq Input	Histone H3K27me3	Histone H3K36me3	Histone H3K4me1	Histone H3K4me3	Histone H3K9ac	Histone H3K9me3	Histone H2AK5ac	Histone H2BK5ac	Histone H2BK120ac
CD34 Cultured Cells									1	1	2	2	2		2			
CD34 Primary Cells				2					1	2	2	3	2		2			
CD3 Primary Cells				1	4				1	2	2	2	2		2			
Mobilized CD34 Primary Cells				7	15	1			11	8	6	5	7		6			
Mobilized CD3 Primary Cells					2													

HTTP/FTP Download

Index of /EdaccData/Release-5/study-sample-experiment /BI/CD4_Memory_Primary_Cells/Histone_H3K4me3/

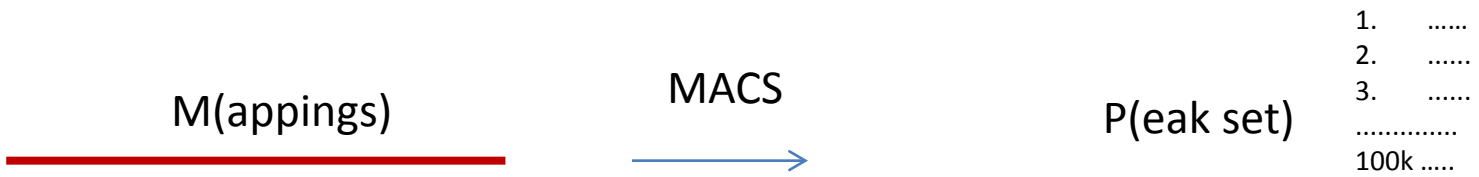
../		
BI.CD4_Memory_Primary_Cells.H3K4me3.20.bed.gz	28-Dec-2011 18:27	84M
BI.CD4_Memory_Primary_Cells.H3K4me3.20.wig.gz	21-Dec-2011 21:14	18M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_100_7..>	28-Dec-2011 18:27	418M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_100_7..>	21-Dec-2011 21:14	45M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_101_8..>	28-Dec-2011 18:27	312M
BI.CD4_Memory_Primary_Cells.H3K4me3.Donor_101_8..>	21-Dec-2011 21:14	37M

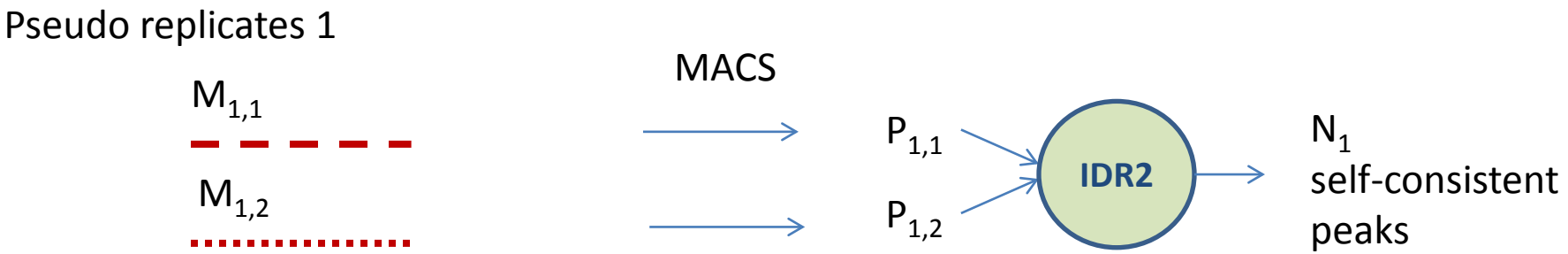
UCSC Visualization



Only peaks above background are reported

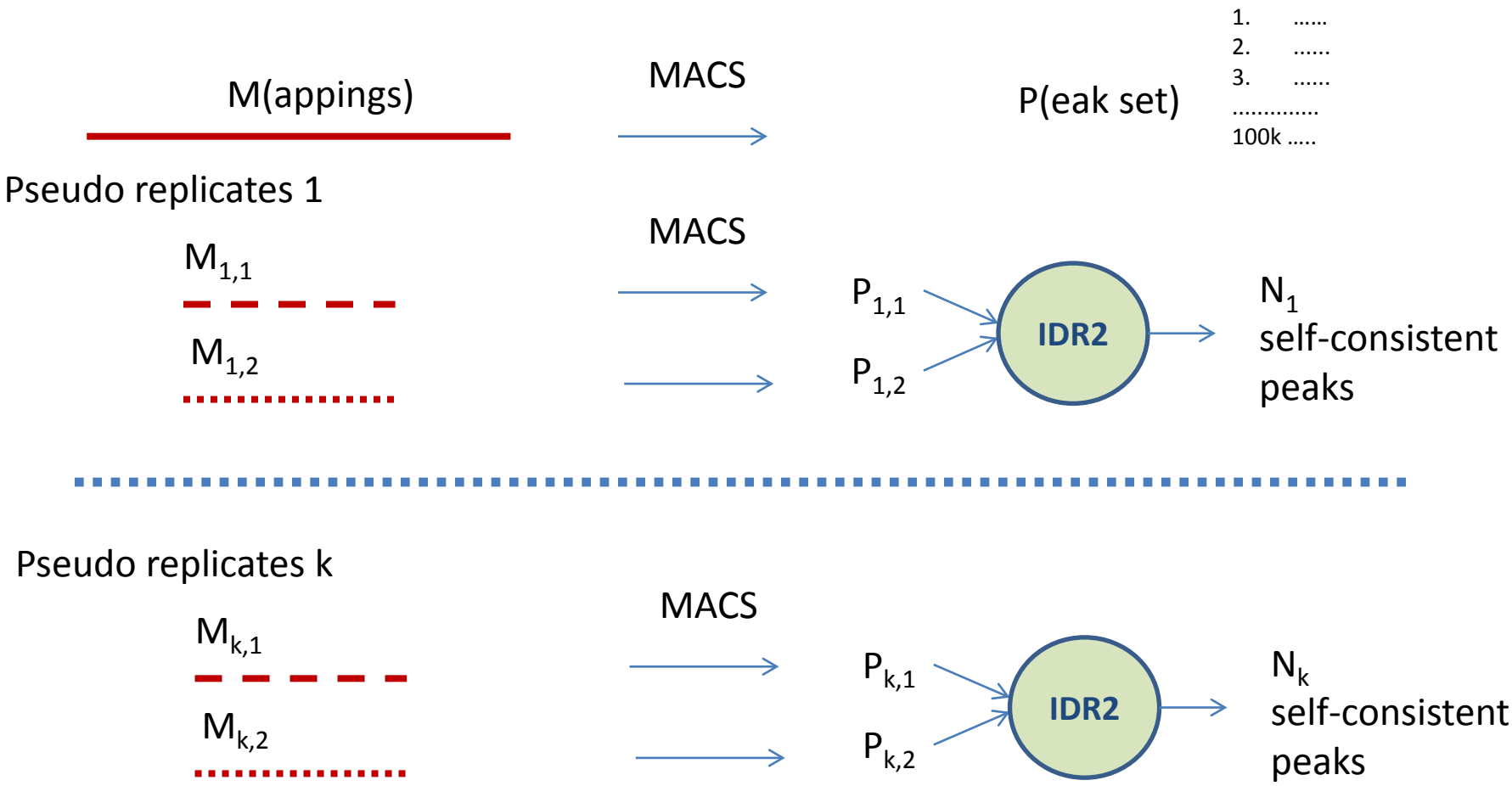
IDR2 Strategy
No Technical Replicates





Select top N_1 peaks from P

IDR2 Strategy
No Technical Replicates, Multiple IDR2 rounds



Select max($N_1, N_2, \ldots N_k$) peaks from P

Evaluation

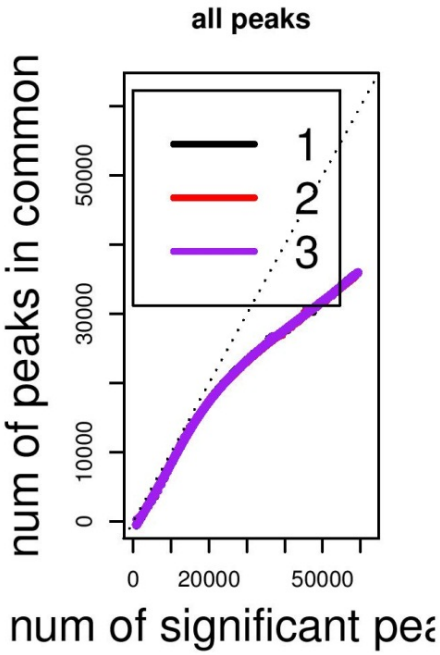
- UCSF-UBC Penis Foreskin Fibroblast H3K4me3 skin01, skin02, skin03
- UCSF-UBC H1 H3K4me3

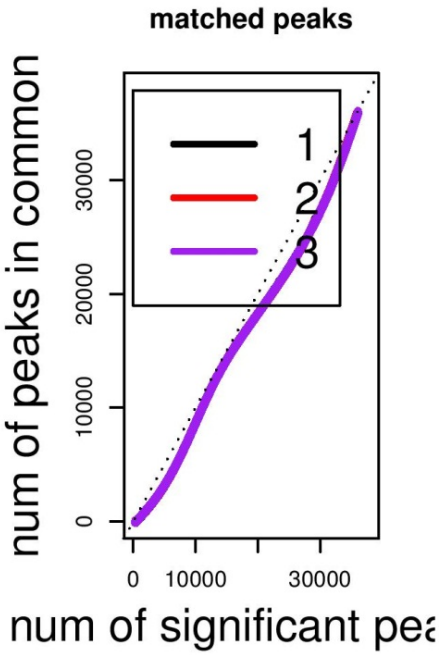
- Punctate mark
- 9-41 million reads
- Mapping
  - Pash
  - Uniquely mapping reads kept
  - Duplicates removed
  - Reads extended to 200bp in mapping direction
- IDR2
  - 3 rounds of splitting, peak calling, IDR
  - 0.01 IDR score threshold
    - based on number of peaks in pseudoreplicates
  - Infer fragment size using spp
    - Strand cross-corelation
  - break ties by adding small random number to the -10log10(pvalue) in MACS output
  - 80-80 split
    - datasets under 20 million reads

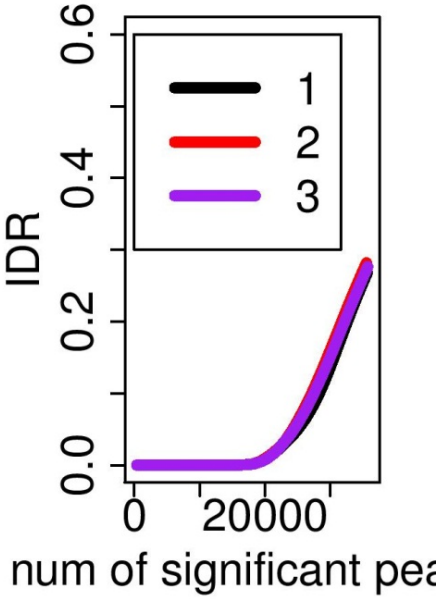
Stats

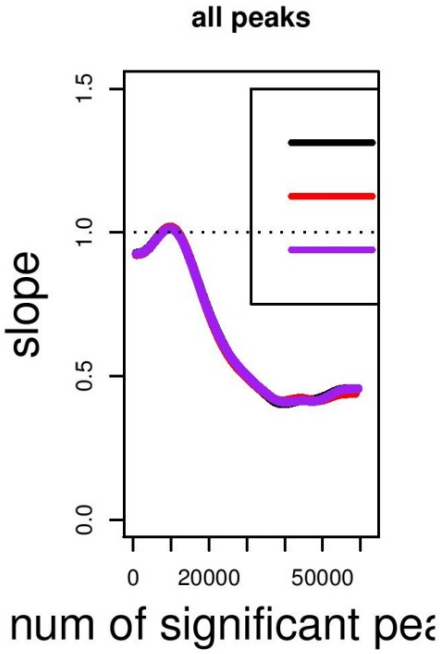
Sample	**Mapped Reads (duplicates removed)**	**Fragment size**	**Peaks at IDR threshold 0.01**	**Peaks at IDR threshold 0.02**
skin01	18.3 million	230	17,703	19,995
skin02	25.9 million	175	39,517	44,732
skin03	41.3 million	235	22,538	27,296
H1	8 million	120-150	40,908	63,247

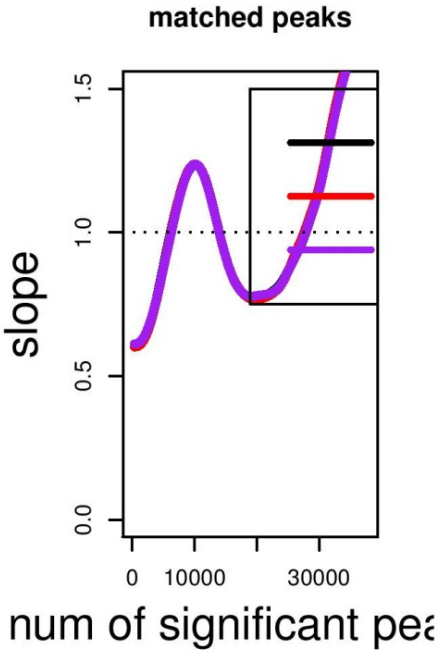
IDR consistency plots

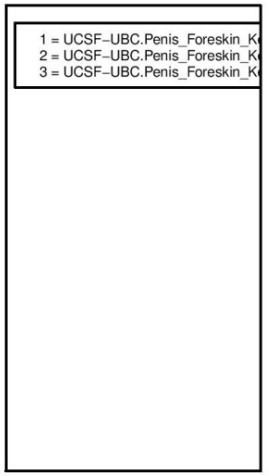






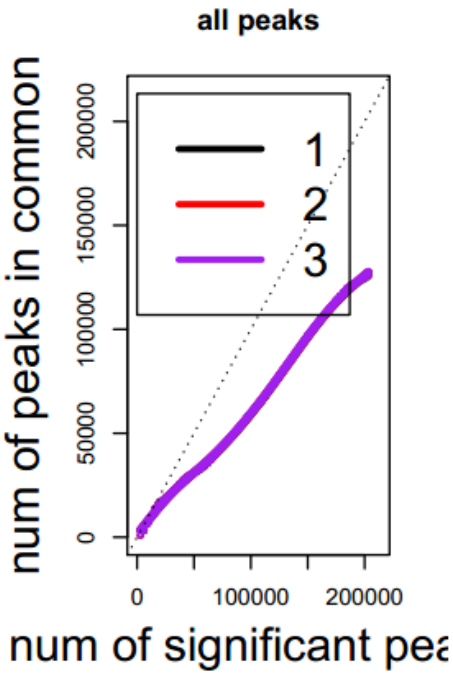


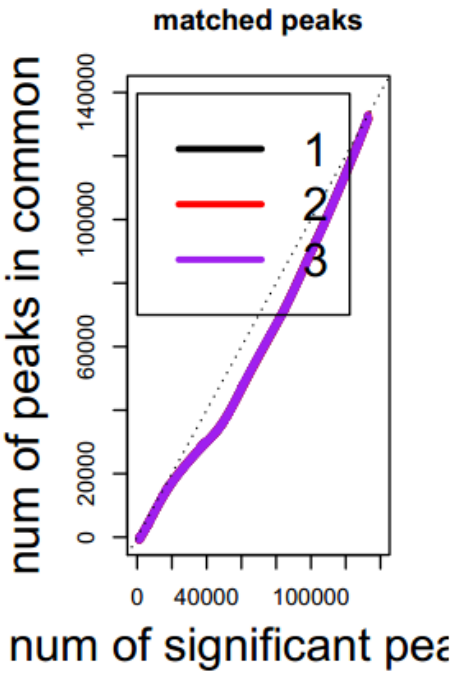


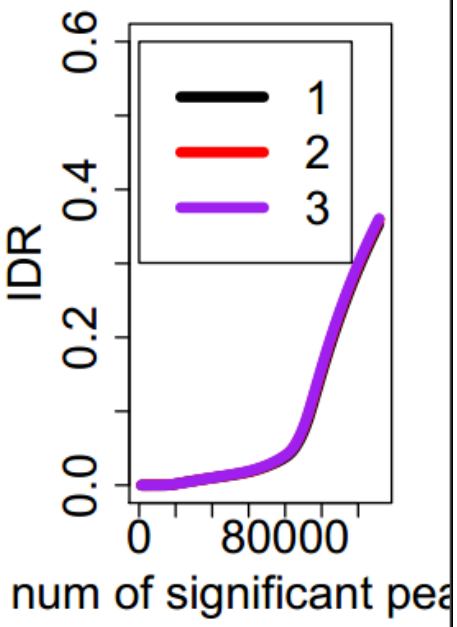


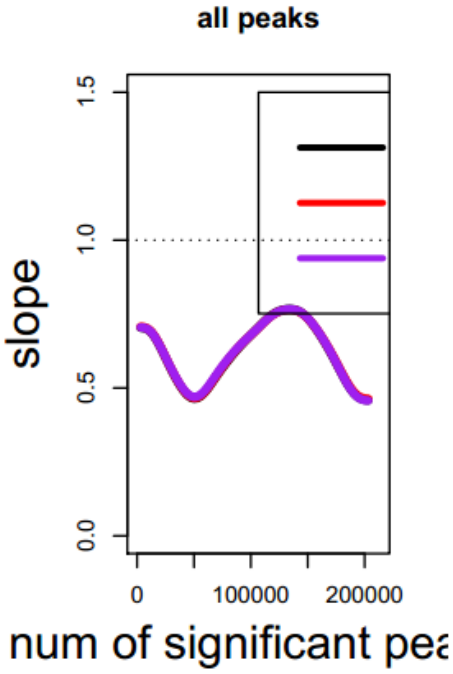
skin02

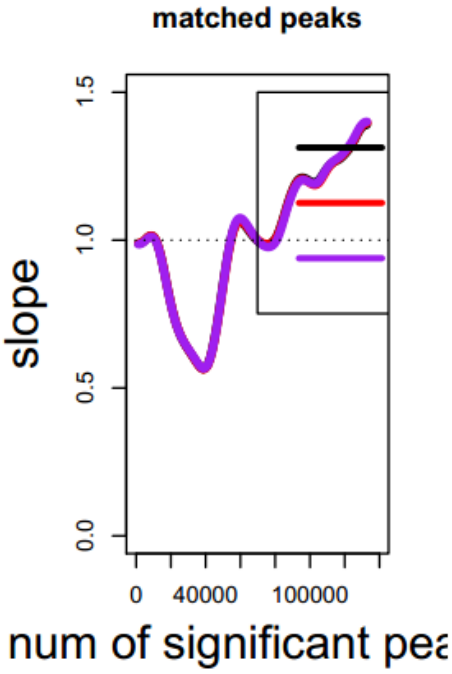
IDR consistency plots

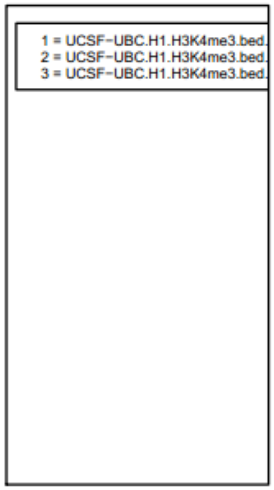












H1

Peak Call Evaluation

- Good concordance among replicates
- Robust set of peak calls available
  - Visualization: UCSC and mirrors
  - Download: BED format

Acknowledgments

- Bob Thurman (UW)
- Noam Shoresh (BI)
- Martin Hirst (BCGSC)
- Lee McDaniel (NCBI)
- Sriram Raghuram, Alan Harris, Andrew Jackson (EDACC)
- Consortium