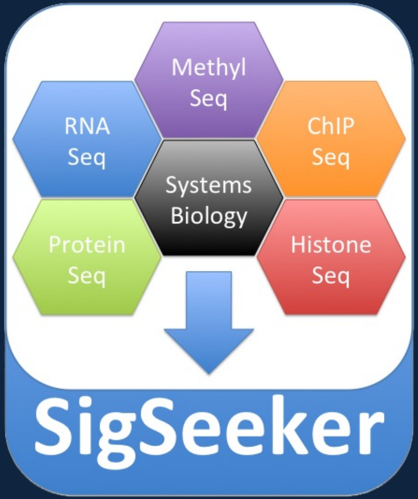
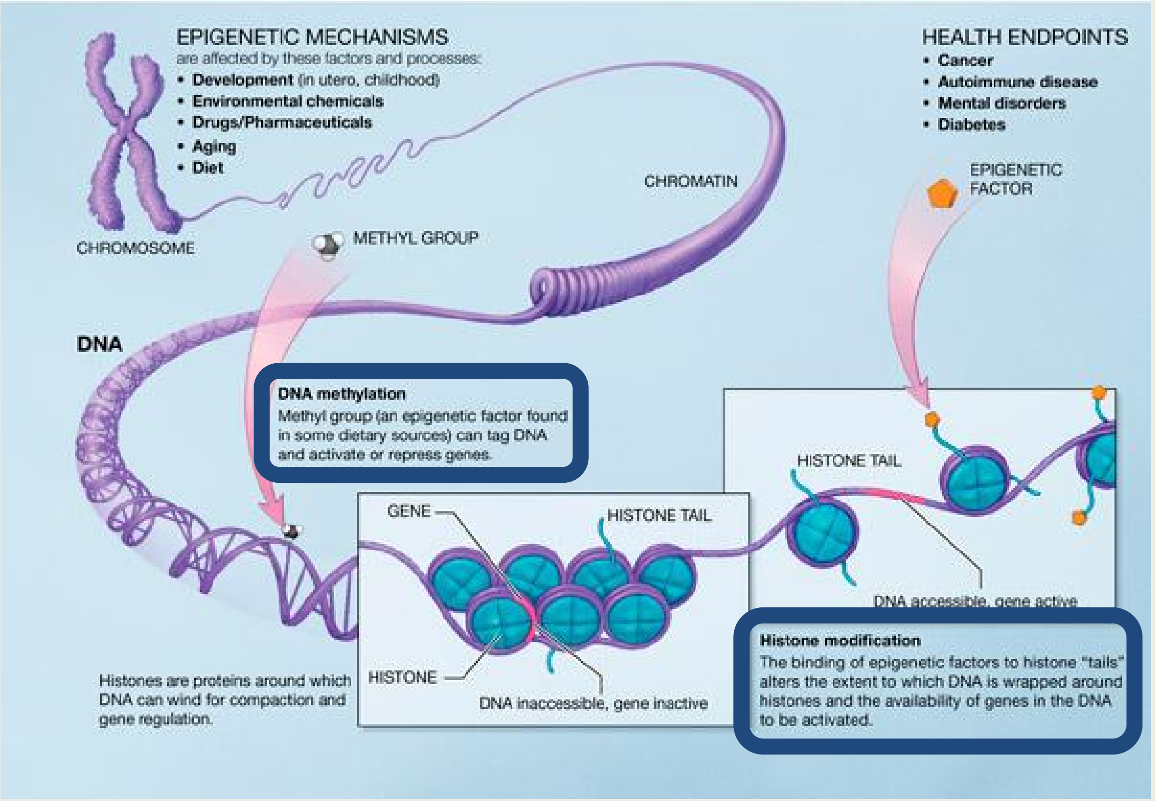


SigSeeker: An Ensemble for Analysis of Epigenetic Data

Jens Lichtenberg
Elisabeth F. Heuston
David M. Bodine

Background - Epigenetics



EPIGENETIC MECHANISMS

are affected by these factors and processes:

- **Development** (in utero, childhood)
- **Environmental chemicals**
- **Drugs/Pharmaceuticals**
- **Aging**
- **Diet**

HEALTH ENDPOINTS

- **Cancer**
- **Autoimmune disease**
- **Mental disorders**
- **Diabetes**

DNA methylation

Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

Histones are proteins around which DNA can wind for compaction and gene regulation.

Histone modification

The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

- Heritable changes in gene activity not attributed to changes in DNA
- DNA Methylation
  - Punctuated modification with often direct impact on gene expression
- Histone modification
  - Long range modifications that impacts the accessibility of DNA for transcription

Background - Motivation

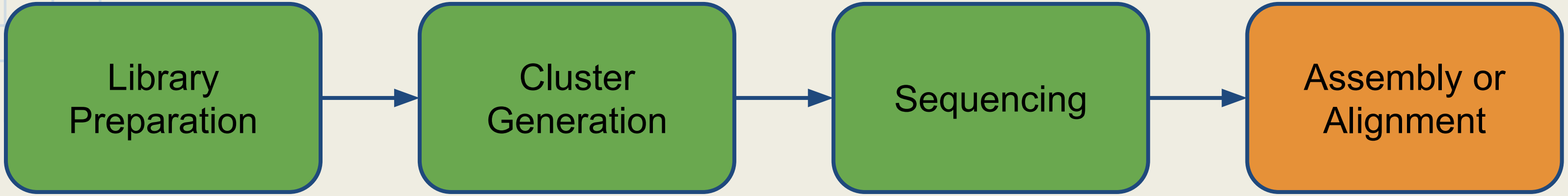
- Large variety of epigenetic approaches that produce sequencing data

- Few approaches exist that attempt the large scale comprehensive analysis of these data

- Common threads in the evaluation process

- Large number of tools for each process

Background - NGS



Library Preparation

Cluster Generation

Sequencing

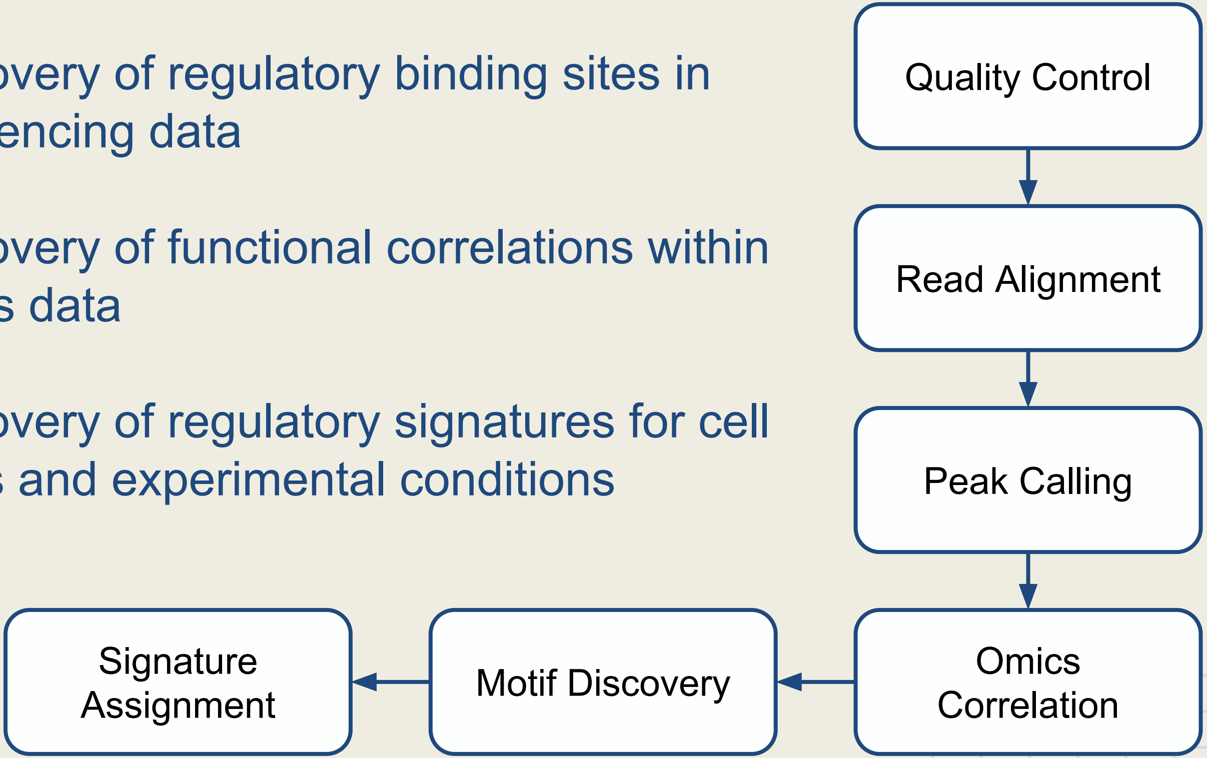
Assembly or Alignment

Next Generation Sequencing

- De Novo Sequencing and Resequencing
  - Single nucleotide polymorphisms, insertions deletions, copy number variations
- Transcript Profiling and Discovery
  - High specificity, sensitivity and read coverage
- Bisulfite and ChIP Sequencing
  - Mapping accuracy, high sensitivity, low sample input

Analysis - Objectives

- Discovery of regulatory binding sites in sequencing data
- Discovery of functional correlations within omics data
- Discovery of regulatory signatures for cell types and experimental conditions



Analysis - Algorithms

- Epigenetic analysis approaches discover regions in which mapped reads are over represented in a sample compared to a control

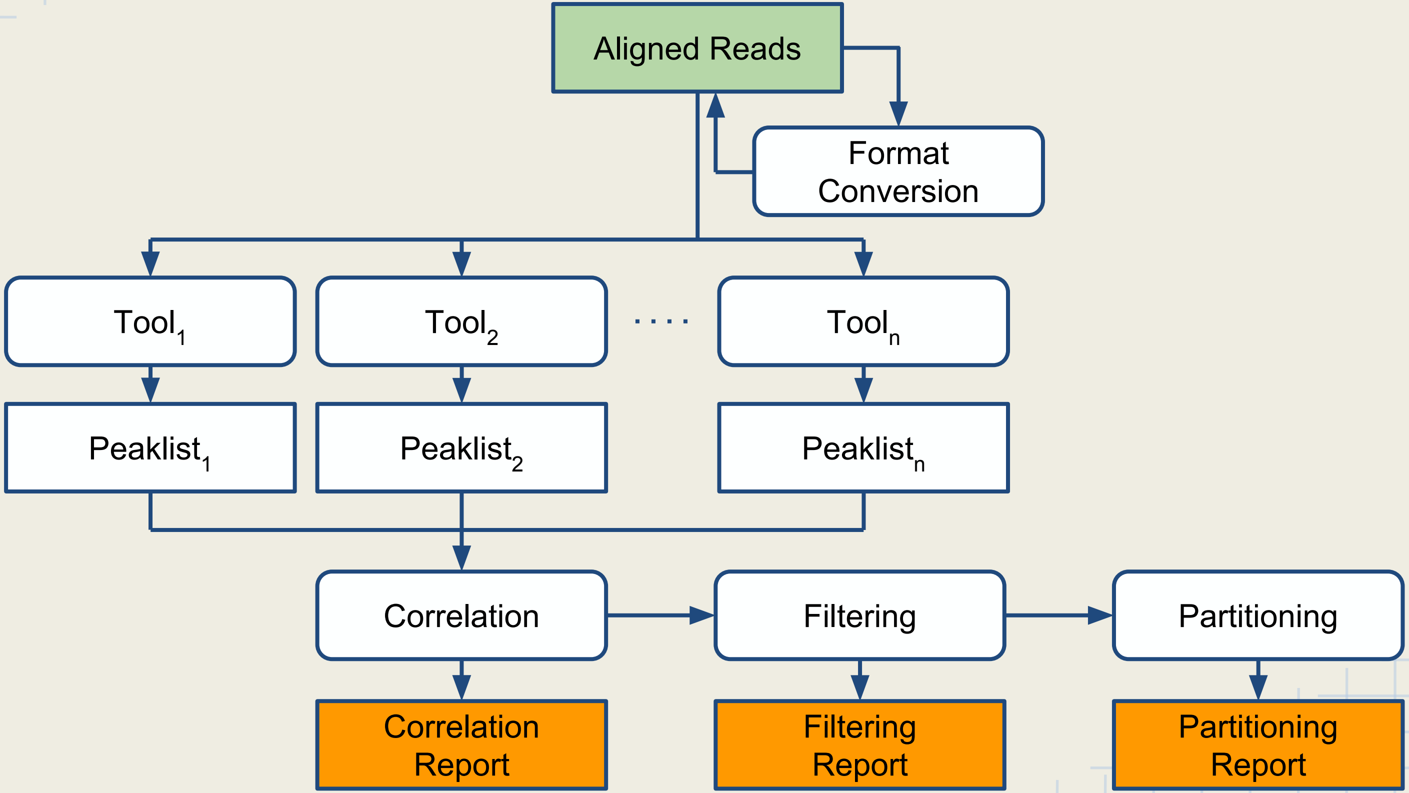
- Excess of 75 different epigenetic analysis approaches exist geared towards different research questions
  - Transcription factor binding site discovery
    - Slender regions containing specific sequence motifs
  - Histone modification localization
    - Broad regions
  - Methylation site discovery
    - Regions defined by contained methylated CpG sites

Analysis - Foundations

- Questions
  - Which approach is the best for the analysis?
  - Is there a correlation between different approaches?
  - Is there a correlation with existing data?

- Hypotheses
  - Different approaches have different advantages and disadvantages
  - Integration of tools will provide better results for the evaluation of sequencing data
  - Correlation of with established information provides deeper understanding

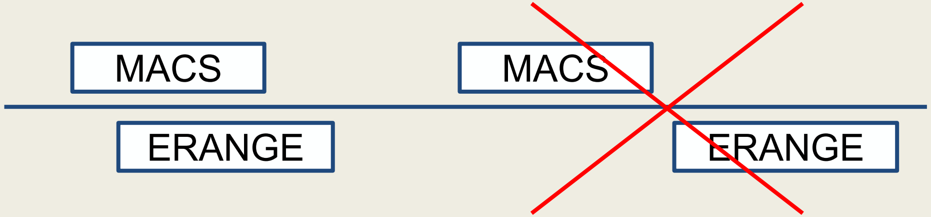
Framework - Overview

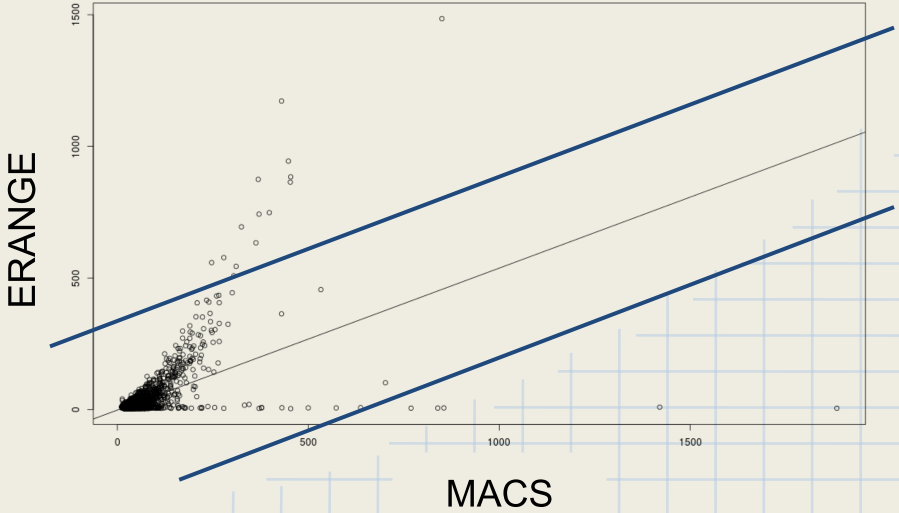


Framework - Correlations

- Structural Correlation
  - Genomic sites detected by more than 1 tool

- Intensity Correlation
  - Co-localized peaks often differ greatly in their intensity
  - Intensities are subjected to linear regression model and outliers are removed





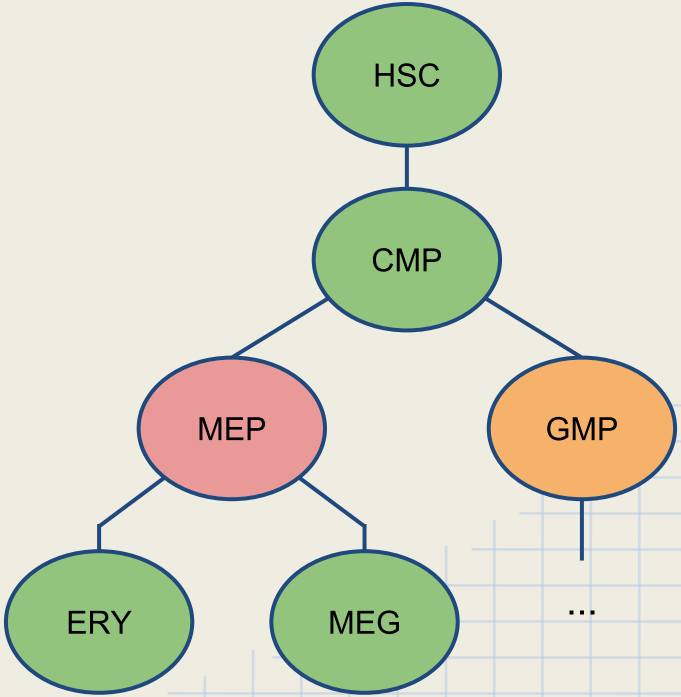
Case Study - Overview

System

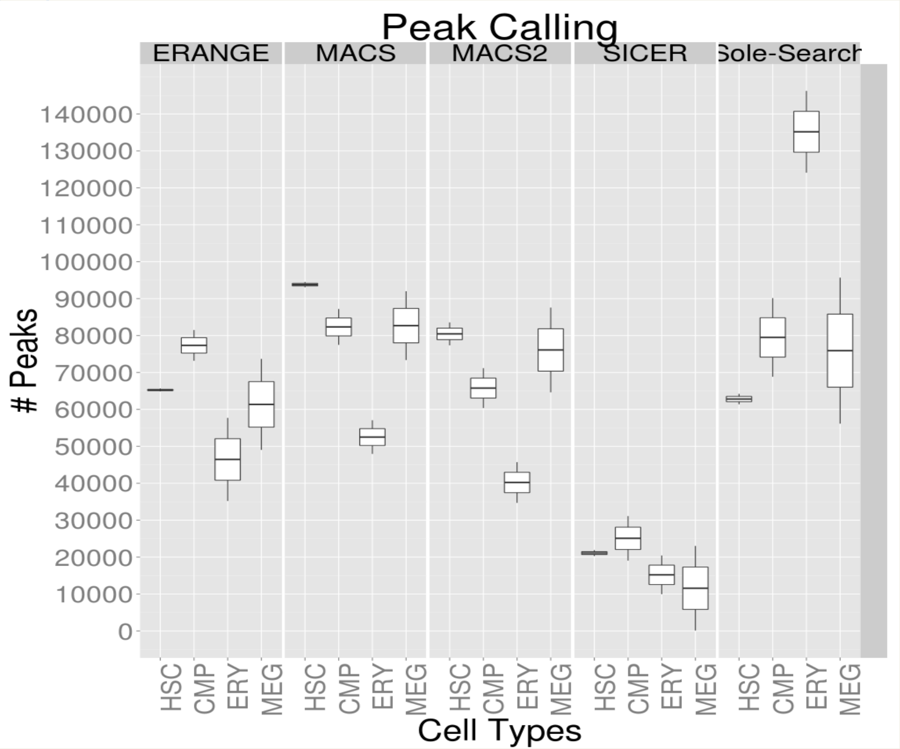
- Hematopoiesis is the formation of blood cellular components from a stem cell
- Biological system containing roughly 60 different cell types
- Cell types can be enriched through flow cytometry
- Foundation for important medical insights (e.g. Leukemia research)

Experiment

- DNA Methylation via MBD2 pulldown
- 4 different cell types (HSC, CMP, ERY, MEG)
- 2 replicates each
- 1 supernatant samples as control



Case Study - Results



- Observation of HSC > CMP > ERY (*Hogart 2012*)
- Generally MEG > ERY
- MACS > MACS2 > ERANGE > SICER
- Sole-Search produces strong outlier for ERY

Case Study - Correlations

HSC	ERY	MACS	SICER	Peaks	Upstream	Promoter	RefSeq	Downstream
x		x		86733	10433	1278	55566	14325
x			x	61196	7191	992	38793	10123
x		x	x	60452	7096	991	38289	9996
	x	x		47202	5622	681	29528	8031
	x		x	36712	4368	531	22907	6321
	x	x	x	35099	4160	512	21770	5982
x	x	x		44794	5294	621	27482	7506
x	x		x	34566	4030	476	21105	5837
x	x	x	x	33351	3900	466	20271	5589

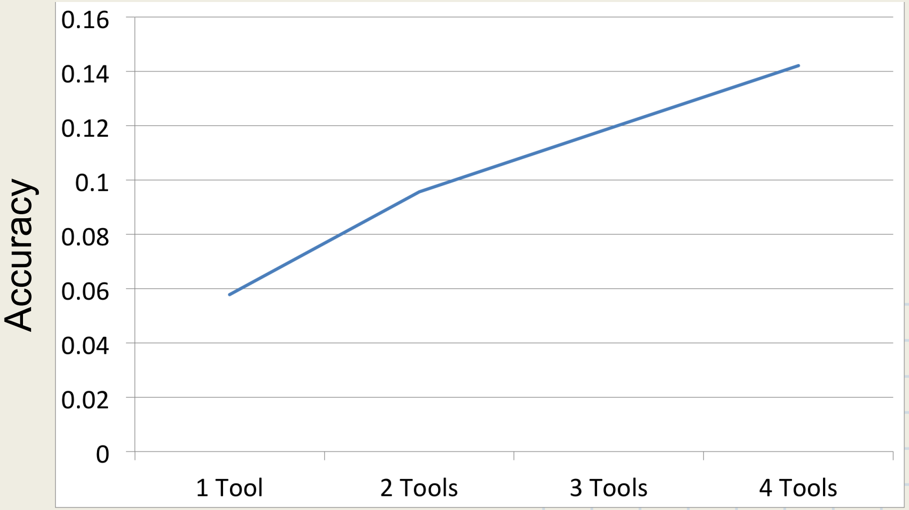
Framework - Benchmarks

Setup

- Positive and negative sites established for a subset of predicted peaks in well studied epigenetics data sets
  - Rye et al.
    - Transcription Factor Binding
  - Micsinai et al.
    - Histone Modification

Accuracy

- Closeness of measurements to true value
- Average ACC increases by about 150%
- Similar observations hold true for all benchmarking metrics



Conclusions

- SigSeeker
  - Framework for the analysis of epigenetics data
  - Ensemble of peak calling techniques
- Insights
  - Adding a tool to the ensemble seems to have a similar power as adding another replicate
  - Different backgrounds have strong impact on number of peaks called
- Applications
  - Hematopoiesis case study through large methylation pulldown experiment
  - Global decline in the number of predicted methylated sites during hematopoietic stem cell differentiation

Acknowledgements

Bodine Lab at NHGRI

- David M. Bodine
- Elisabeth F. Heuston
- Amber Hogart
- Stephanie Battle

NISC

- Jim Mullikin
- Baishali Maskeri

Flow Cytometry Core

- Stacie Anderson

Penn State University

- Ross Hardison
- Tejaswini Mishra

Johns Hopkins University

- Michael McDevitt

NHGRI

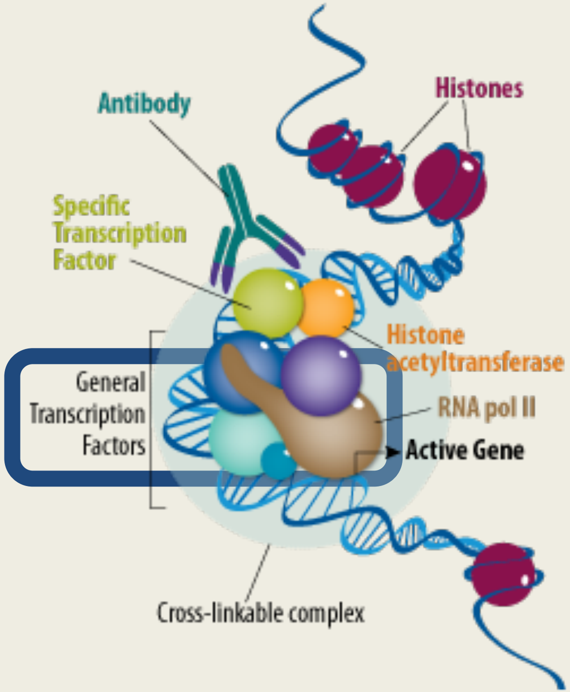
- Laura N. Elnitski
- Nigel Crawford

Ohio University

- Lonnie R. Welch

Thank you !

Background - Epigenetics


Antibody
Histones
Specific Transcription Factor
Histone acetyltransferase
General Transcription Factors
RNA pol II
Active Gene
Cross-linkable complex

- Transcription Factors
  - Punctuated binding events that facilitates the recruitment of RNA pol II and subsequent transcription

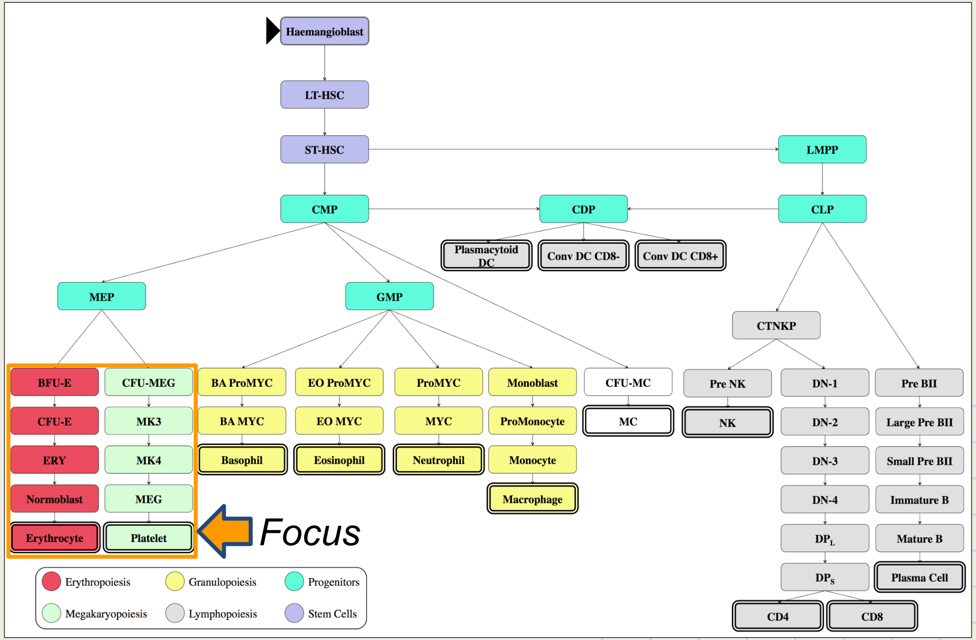
- Antibodies have been designed for the recognition of DNA methylation, Histone modification and a large number of Transcription factors

Analysis - Algorithms

Tool	Citation	Peak Finding Categorization	Background Model
MACS	Zhang2008	Sliding Window	Poisson tag count distribution
cisGenome	Ji2008	Sliding Window	Negative binomial distribution
BayesPeak	Spyrou2009	Sliding WIndow	Negative binomial distribution
BroadPeak	Wang2013	Sliding Window, Predefined Regions	Poisson tag count distribution
Sole-Search	Blahnik2009	Sliding Window	Duplicate and Deletion Compensation
SICER	Zang2009	Binning	Poisson tag count distribution
E-Range	Mortazavi2008	Clustering Distance XSETs	Fold Enrichment
FSeq	Boyle2008	Gaussian KDE	Std. Dev. Threshold
FindPeaks	Fejes2008	Overlapping XSETs	Monte Carlo Simulation

Results - Case Study

- Hematopoiesis is a biological system containing roughly 60 different cell types



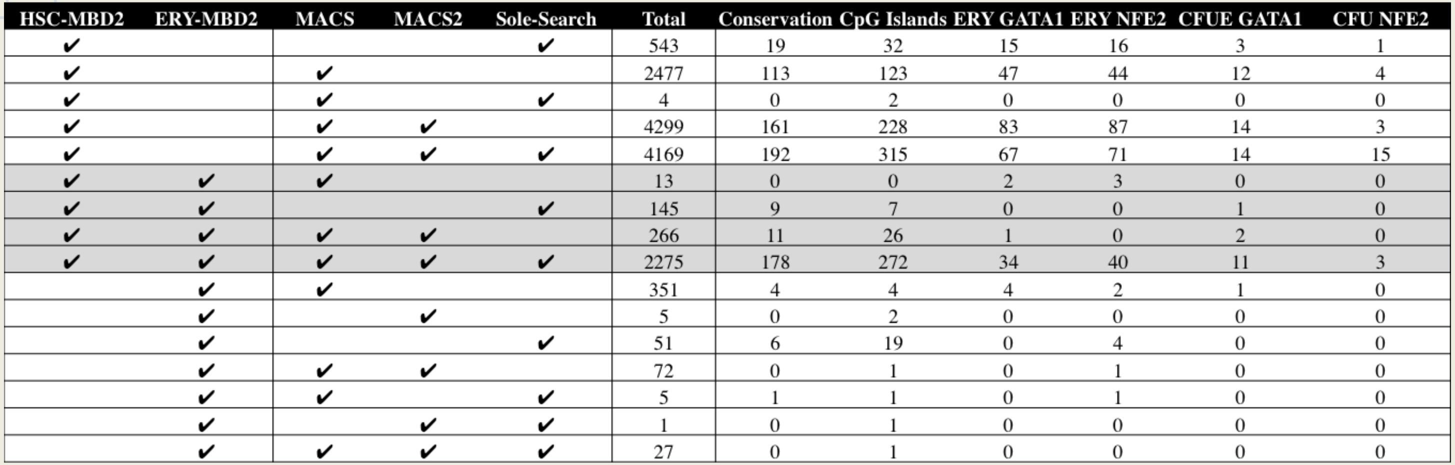
- Cell types can be enriched through flow cytometry

- Can be studied in adult specimen

- Good in-vitro vs in-vivo correlation

- Foundation for important medical insights (e.g. Leukemia research)

Results - Case Study



HSC-MBD2	ERY-MBD2	MACS	MACS2	Sole-Search	Total	Conservation	CpG Islands	ERY GATA1	ERY NFE2	CFUE GATA1	CFU NFE2
✔				✔	543	19	32	15	16	3	1
✔		✔			2477	113	123	47	44	12	4
✔		✔		✔	4	0	2	0	0	0	0
✔		✔	✔		4299	161	228	83	87	14	3
✔		✔	✔	✔	4169	192	315	67	71	14	15
✔	✔	✔			13	0	0	2	3	0	0
✔	✔			✔	145	9	7	0	0	1	0
✔	✔	✔	✔		266	11	26	1	0	2	0
✔	✔	✔	✔	✔	2275	178	272	34	40	11	3
	✔	✔			351	4	4	4	2	1	0
	✔		✔		5	0	2	0	0	0	0
	✔			✔	51	6	19	0	4	0	0
	✔	✔	✔		72	0	1	0	1	0	0
	✔	✔		✔	5	1	1	0	1	0	0
	✔		✔	✔	1	0	1	0	0	0	0
	✔	✔	✔	✔	27	0	1	0	0	0	0

- Novel peaks in ERY rarely conserved
- Majority of CpG Island methylation is lost during differentiation
- Peaks retained during differentiation discovered by all tools, ERY specific peaks are best described by MACS
- ERY methylation leads to lack of transcription factor binding in ERY