# Replication of ChiP-Seq Analysis

George Rouvalis

# Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes

Piu Wong

It is unclear how epigenetic changes regulate the induction of erythroid-specific genes during terminal erythropoiesis. In this project the researchers used global mRNA sequencing (mRNA-seq) and chromatin immunoprecipitation coupled to high-throughput sequencing (CHIP-seq) to investigate the changes that occur in mRNA levels, RNA polymerase II (Pol II) occupancy, and multiple posttranslational histone modifications when erythroid progenitors differentiate into late erythroblasts. Among genes induced during this developmental transition, there was an increase in the occupancy of Pol II, the activation marks H3K4me2, H3K4me3, H3K9Ac, and H4K16Ac, and the elongation methylation mark H3K79me2. In contrast, genes that were repressed during differentiation showed relative decreases in H3K79me2 levels yet had levels of Pol II binding and active histone marks similar to those in erythroid progenitors.

They also found that relative changes in histone modification levels, in particular, H3K79me2 and H4K16ac, were most predictive of gene expression patterns. The results suggest that in terminal erythropoiesis both promoter and elongation-associated marks contribute to the induction of erythroid genes, whereas gene repression is marked by changes in histone modifications mediating Pol II elongation. The data map the epigenetic landscape of terminal erythropoiesis and suggest that control of transcription elongation regulates gene expression during terminal erythroid differentiation.

**Contents**

**Data**

| Sample Name | GEO Accession | Chip Anti-Body | Project Report Shortname |
|---|---|---|---|
| SRR1157326 | GSM688808 | H3K4me2 | Sample0 |
| SRR1157329 | GSM688811 | H3K27me3 | Sample1 |
| SRR1157333 | GSM688815 | RNA Pol II | Sample2 |
| SRR1157341 | GSM688824 | RNA Pol II | Sample3 |

No control or input/baseline was among the samples selected.

In general, the mouse genome (mm9 version) was used for reference during the stages of Bowtie, MACS, IGV and MEME.

**Tools**

Python 2.7.12

Samtools 1.11 (Utility)

Bedtools 2.27.1 (Utility)

FastQC 0.11.9 (Quality Control)

Minion (Adapter Prediction)

Cutadapt 1.9.1 (Adapter Trimming)

Bowtie 1.0.0 (Alignment)

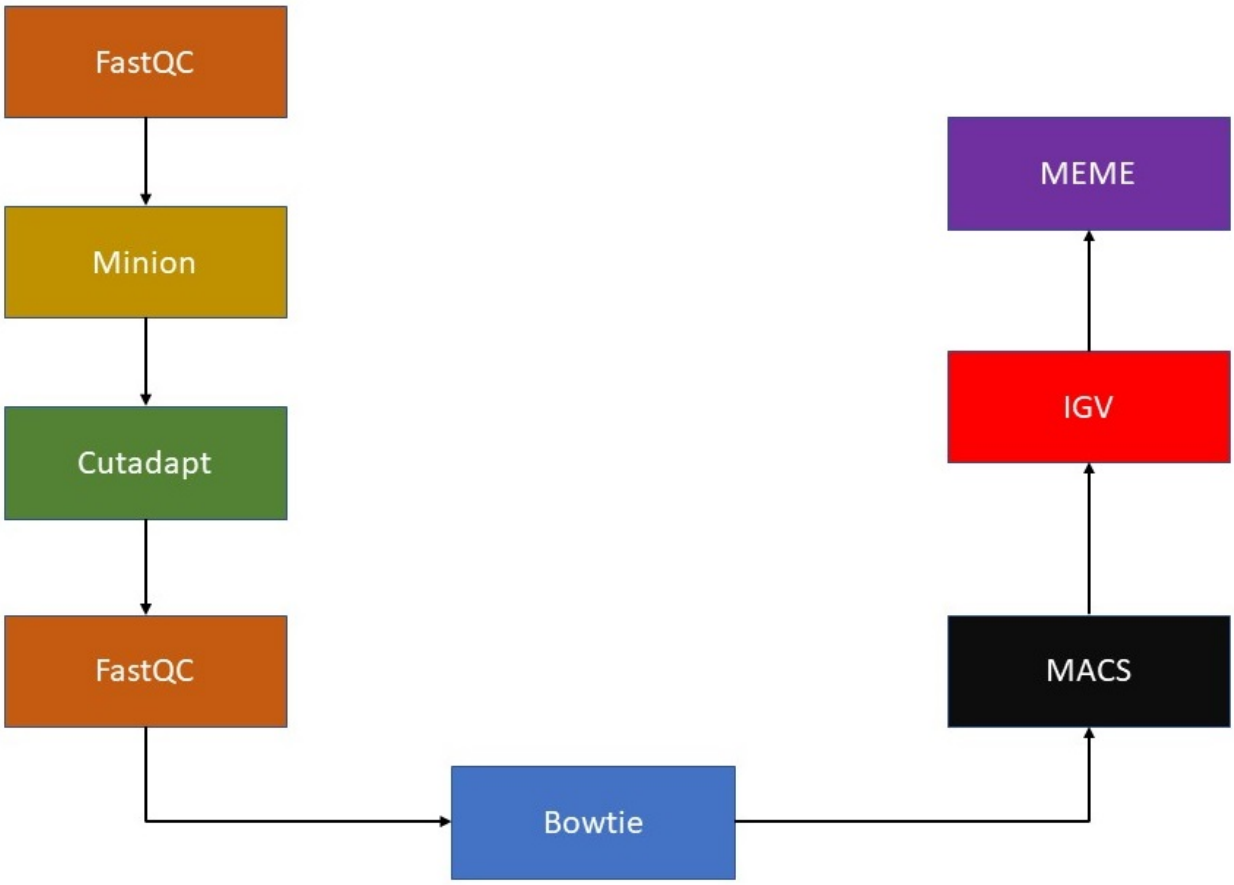
MACS 1.4.1 (Peak Calling)

MEME 5.0.2 (Motif analysis)

**Protocol**

The following steps took place as part of the ChiP-seq analysis:

1. Quality Control (FastQC)
2. Adapter Prediction (Minion)
3. Adapter Trimming (Cutadapt)
4. Quality Control (FastQC)
5. Alignment (Bowtie)
6. Peak Calling (MACS)
7. Visualization (IGV)
8. Motif Analysis (MEME)



You can view the log file of all commands by clicking the link:

Commands Log

These commands were executed for each sample.

**FastQC**

At first, there needed to be some quality control checks on the raw sequence data of the four samples, which would give us a quick impression of whether the data has any problems of which I should be aware before doing any further analysis.

- For Sample1 & Sample3, I observed a failure implying that something is happening in the proportion of each base in each position. The lines should be parallel in normal conditions.
- Furthermore, there was a failure in “Adapter Content”, suggesting that there are adapters in the samples.
- For Sample3, also, the Per base-Per sequence quality was poor.

Below, by clicking, you can see the extended fastQC analysis for all samples:

Sample0 Sample1

Sample2 Sample3

**Minion**

This tool was used for searching for the 3' adapter without prior knowledge of the sequence. Minion outputs one or two candidates, accompanied with metadata and selected from a longer list of candidates. If the adapter sequence is known, minion will compare this adapter with all the candidates it found and output the best match.

Unfortunately, one these four samples, minion detects sequences that don’t appear to be adapters. The expected length is supposed to be ~12, and the results is more than this. Probably the results are just some overexpressed sequences that can be found throughout the samples.

After this failed attempt to detect the adapters, Illumina Genome Analyzer was noticed to be used for these samples. Heading over the prior fastQC analysis and examining the “Adapter Content”, it revealed the presence of Illumina Universal Adapters.

**Cutadapt**

Cutadapt was used to find and remove adapter sequences. In this case, as derived from the previous analysis stage, the removal of Illumina Universal Adapter was necessary.

*Illumina Universal Adapter: AGATCGGAAGAG*

**FastQC #2**

A second quality control check was repeated so the changes derived from the cutadapt step can be further inspected.

Although, all the adapters were successfully removed, the existence of zero length sequences was noticed. These can be filtered out the cutadapt stage.

**Bowtie**

The indexes were built from the mouse genome and the alignment process started.

Those are the results from the alignment:

Sample0

# reads processed: 19396915
# reads with at least one reported alignment: 11646200 (60.04%)
# reads that failed to align: 2953663 (15.23%)
# reads with alignments suppressed due to -m: 4797052 (24.73%)
Reported 11646200 alignments to 1 output stream(s)

Sample1

# reads processed: 3572026
# reads with at least one reported alignment: 2728529 (76.39%)
# reads that failed to align: 262188 (7.34%)
# reads with alignments suppressed due to -m: 581309 (16.27%)
Reported 2728529 alignments to 1 output stream(s)

Sample2

# reads processed: 3446127
# reads with at least one reported alignment: 2365739 (68.65%)
# reads that failed to align: 359016 (10.42%)
# reads with alignments suppressed due to -m: 721372 (20.93%)
Reported 2365739 alignments to 1 output stream(s)

Sample3

# reads processed: 22997145
# reads with at least one reported alignment: 4089861 (17.78%)
# reads that failed to align: 13577205 (59.04%)
# reads with alignments suppressed due to -m: 5330079 (23.18%)
Reported 4089861 alignments to 1 output stream(s)

Finally, the sam file was converted to bam, then sorted and finally indexed so it can be loaded into IGV and be used on MACS as well.

**MACS**

The peaks.bed file was then sorted based on their score and the top three were selected.

See, for every sample, the top regions:

Sample0

| chr | start | end | length | -10*log10(pvalue) |
|---|---|---|---|---|
| chr12 | 112996046 | 112998500 | 2455 | 3118.2 |
| chr17 | 25251616 | 25253639 | 2024 | 3100 |
| chr9 | 35112724 | 35113245 | 522 | 3100 |

Sample1

| chr | start | end | length | -10*log10(pvalue) |
|---|---|---|---|---|
| chr17 | 39979830 | 39983741 | 3912 | 3100 |
| chr11 | 108872874 | 108873447 | 574 | 2998.72 |
| chr5 | 147072479 | 147073018 | 540 | 1664.64 |

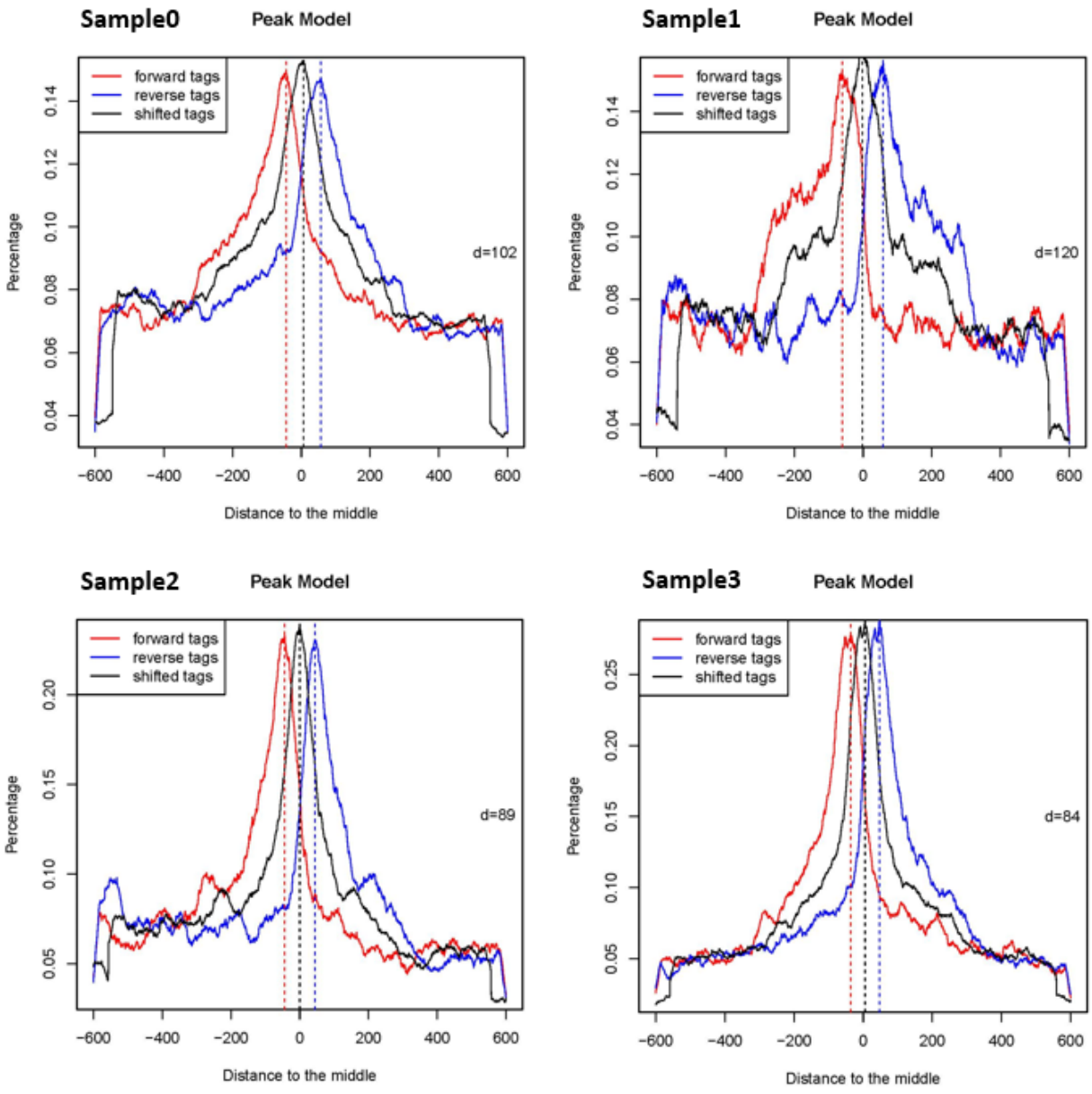
Sample2

| chr | start | end | length | -10*log10(pvalue) |
|---|---|---|---|---|
| chr13 | 21869842 | 21872771 | 2930 | 3184.03 |
| chr13 | 23646580 | 23648688 | 2109 | 3100 |
| chr13 | 23712010 | 23714673 | 2664 | 3100 |

Sample3

| chr | start | end | length | -10*log10(pvalue) |
|---|---|---|---|---|
| chr9 | 35112744 | 35113179 | 436 | 2475.71 |
| chrUn_random | 5495488 | 5497315 | 1828 | 1819.11 |
| chr13 | 23830529 | 23832826 | 2298 | 1462.54 |

The peak model for each sample:

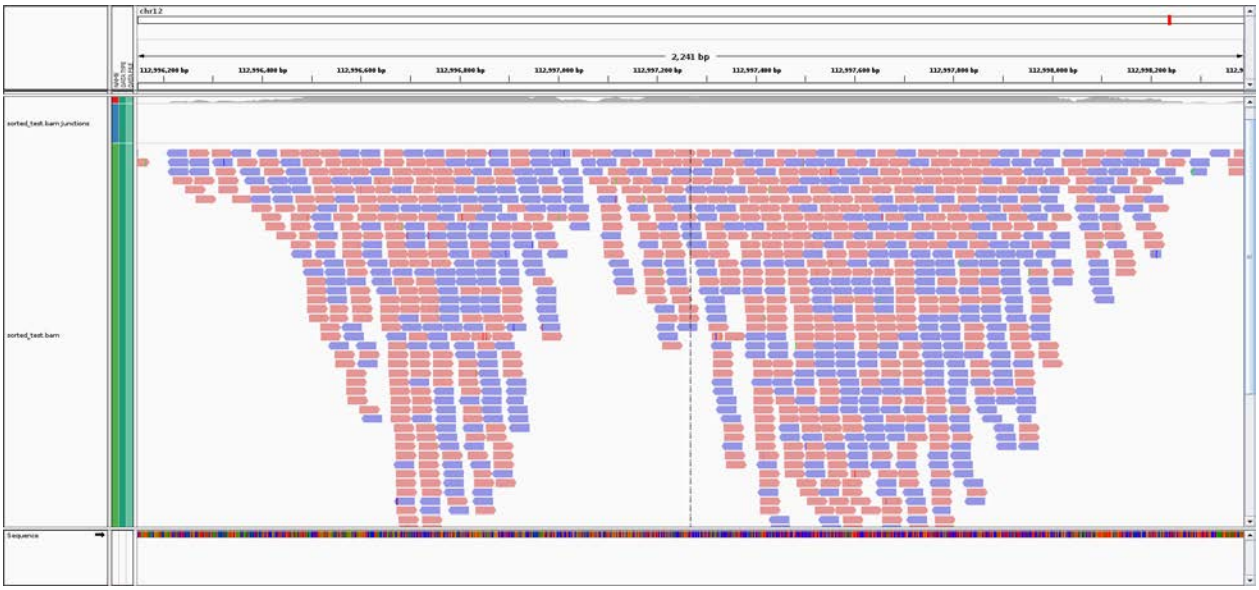


**IGV**

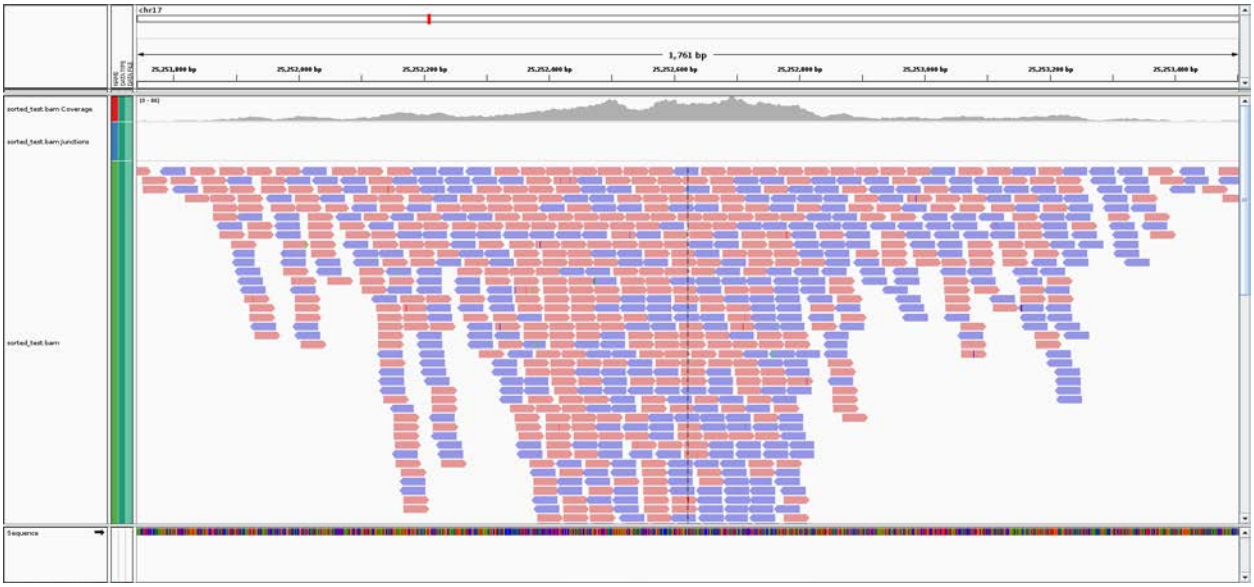
The mouse genome was, firstly loaded, and afterwards each sorted bam file from each sample.

Sample0

-Peak1




-Peak2

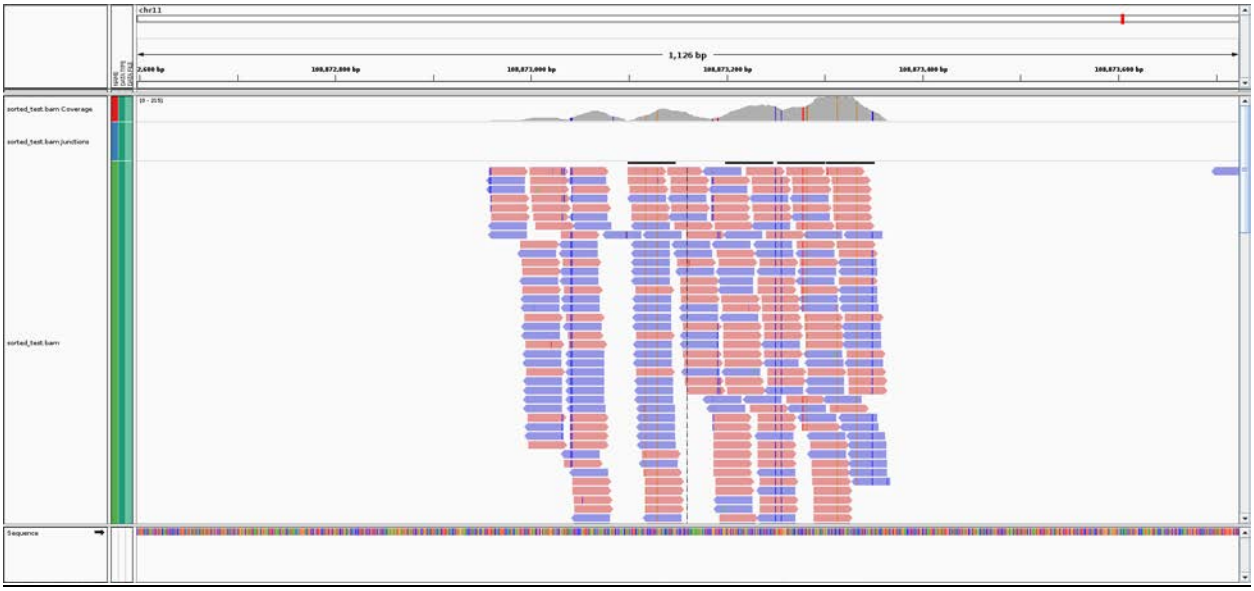


-Peak3

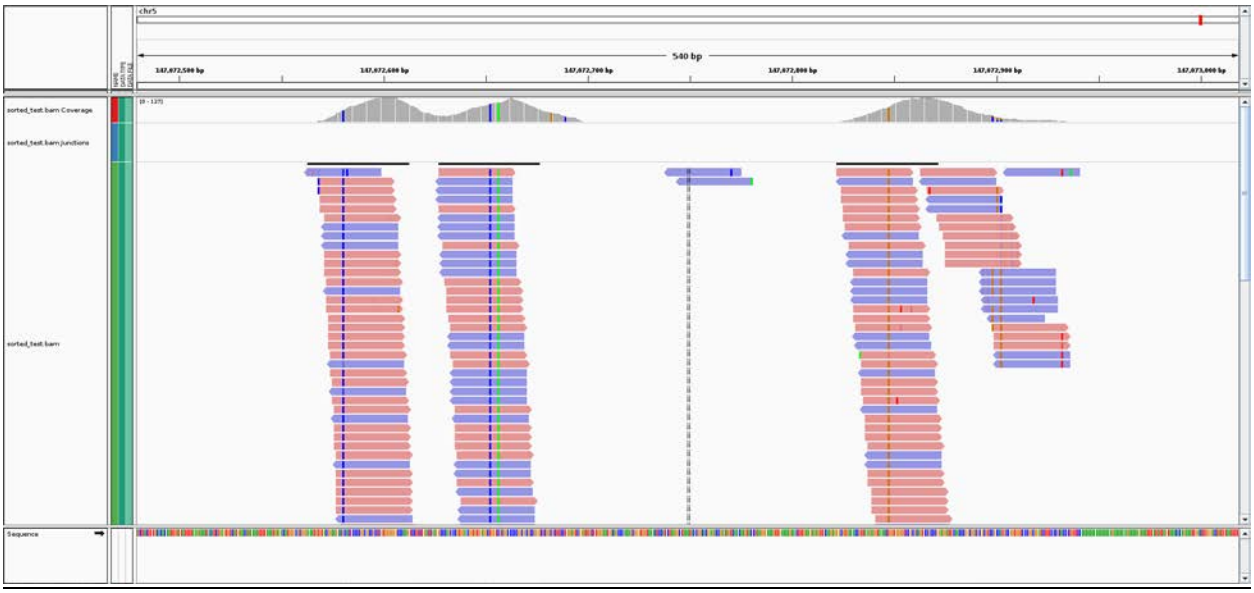

Sample1

-Peak1

-Peak2


chr11
1,126 bp
sorted_test.bam Coverage
sorted_test.bam Junctions
sorted_test.bam
Sequence

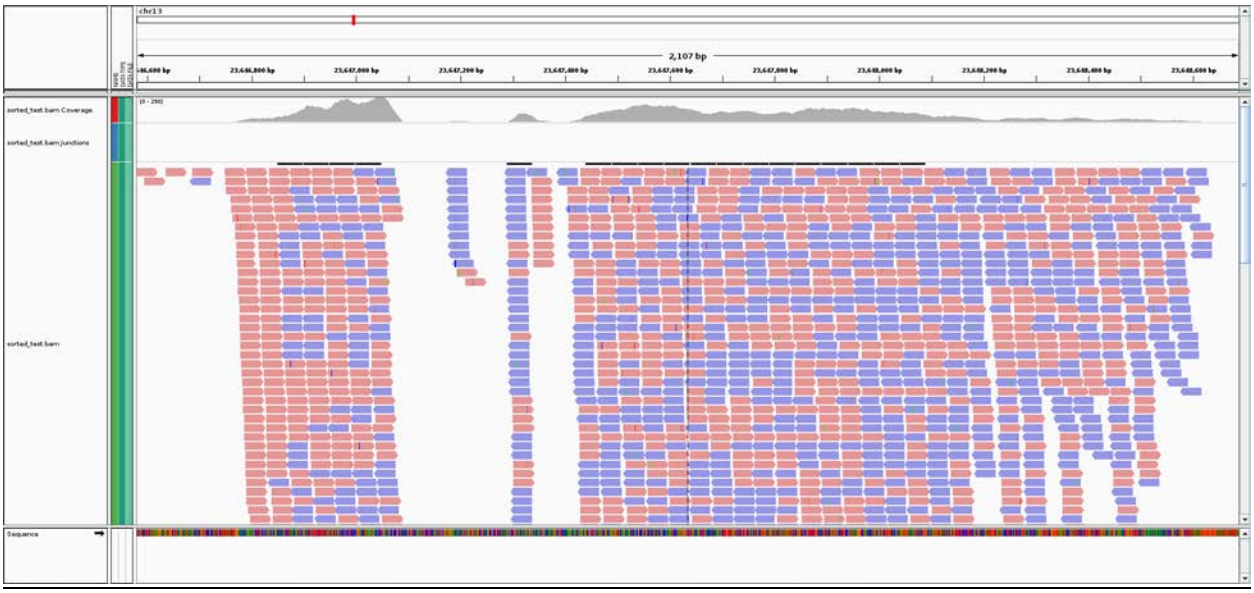
-Peak3


chr5
540 bp
sorted_test.bam Coverage
sorted_test.bam Junctions
sorted_test.bam
Sequence

Sample2

-Peak1

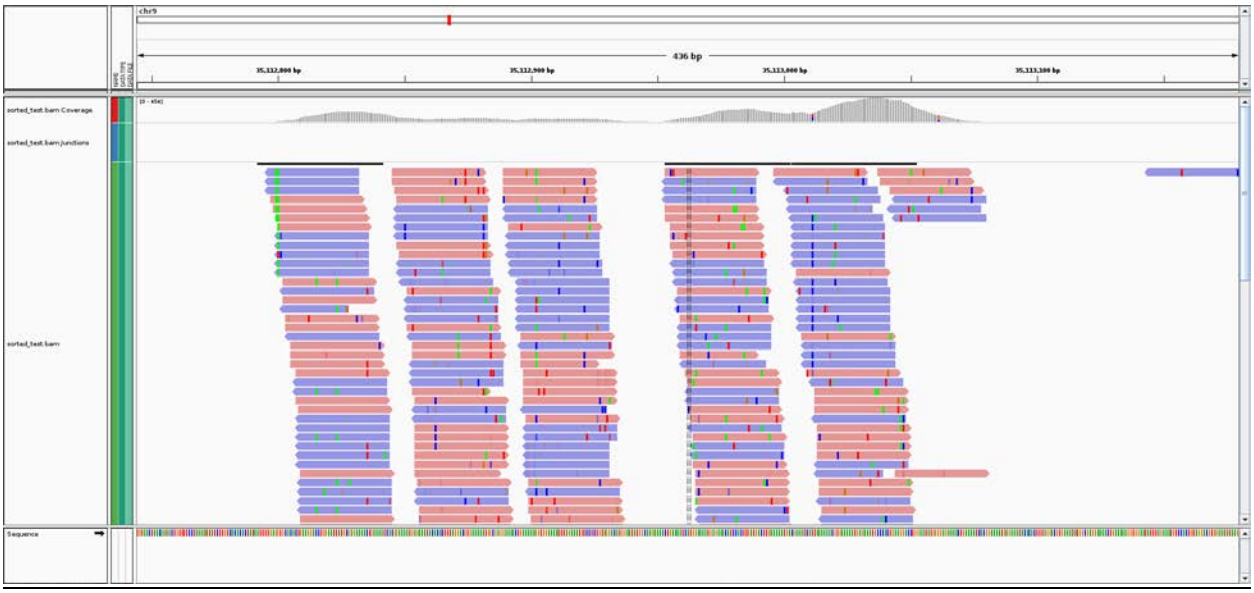

-Peak2



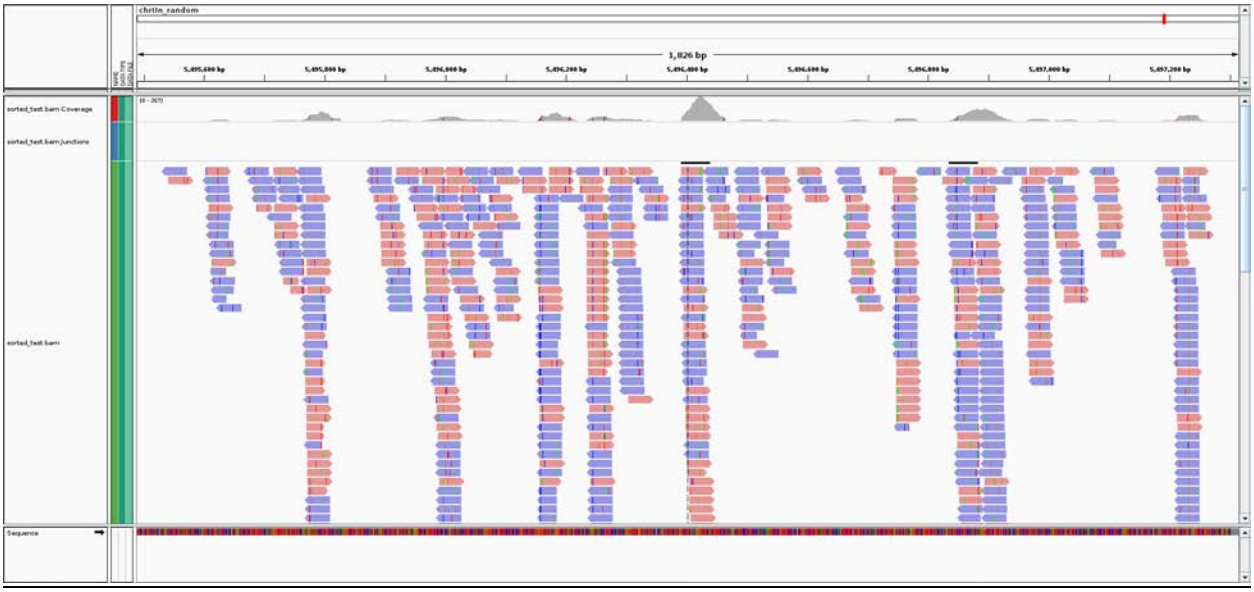
-Peak3


Sample3

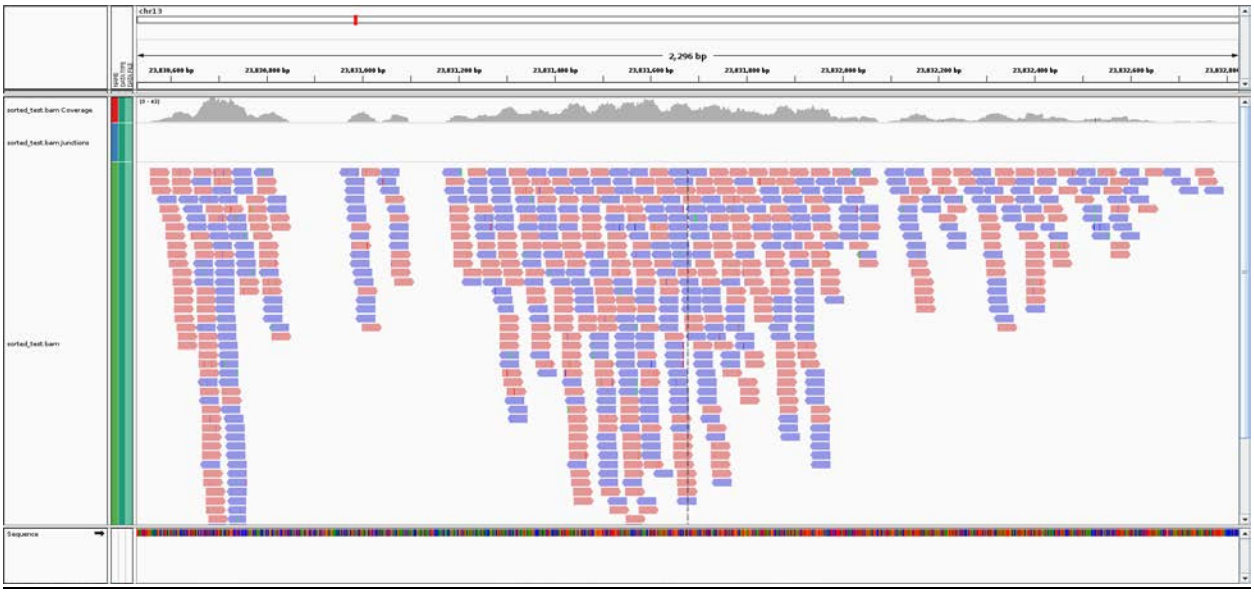
-Peak1



-Peak2


chr1b_random
1,826 bp
sorted_test.bam Coverage
sorted_test.bam Junctions
sorted_test.bam
Sequence

-Peak3


chr1.3
2,296 bp
sorted_test.bam Coverage
sorted_test.bam Junctions
sorted_test.bam
Sequence

**MEME**

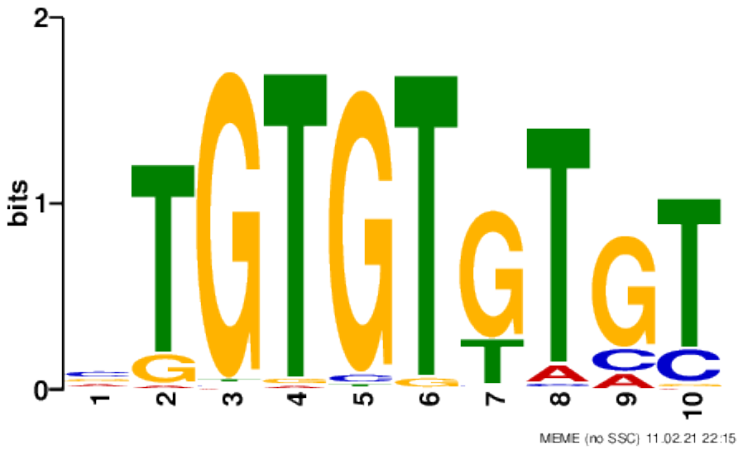
MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in the sequences and splits variable-length patterns into two or more separate motifs.

A motif is an approximate sequence pattern that occurs repeatedly in a group of related sequences. MEME represents motifs as position-dependent letter-probability matrices that describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

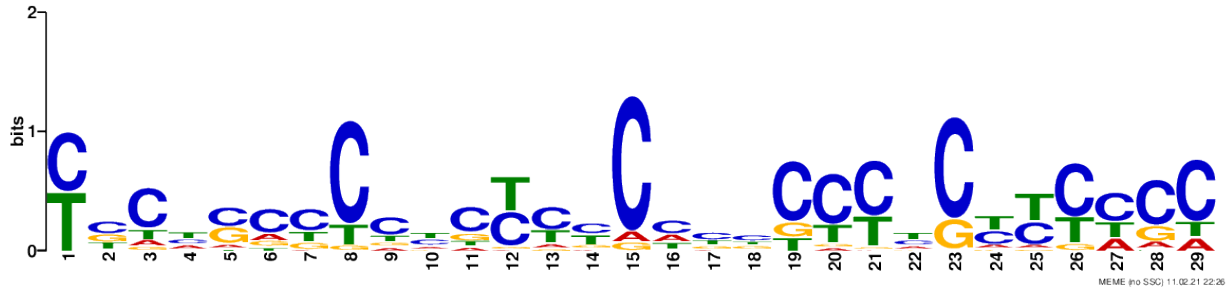
MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif.

These are the discovered motifs for each sample:

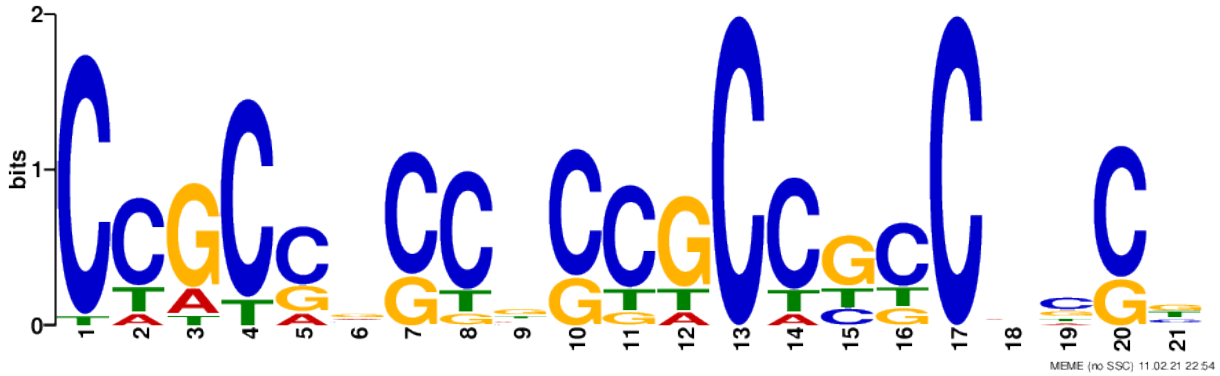
Sample0



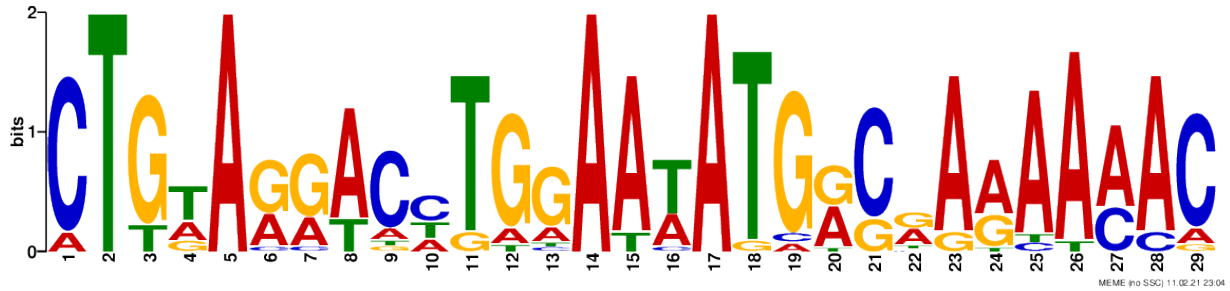
Sample1




Sample2




Sample3



## References

1. Cantor AB, Orkin SH. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene.* 2002;21(21):3368–3376. [PubMed] [Google Scholar]
2. Kelly TK, De Carvalho DD, Jones PA. Epigenetic modifications as therapeutic targets. *Nat Biotechnol.* 2010;28(10):1069–1078. [PMC free article] [PubMed] [Google Scholar]
3. Kiefer CM, Hou C, Little JA, Dean A. Epigenetics of beta-globin gene regulation. *Mutat Res.* 2008;647(1):68–76. [PMC free article] [PubMed] [Google Scholar]
4. Heintzman ND, Ren B. Finding distal regulatory elements in the human genome. *Curr Opin Genet Dev.* 2009;19(6):541–549. [PMC free article] [PubMed] [Google Scholar]
5. Henikoff S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet.* 2008;9(1):15–26. [PubMed] [Google Scholar]
6. Berger SL. The complex language of chromatin regulation during transcription. *Nature.* 2007;447(7143):407–412. [PubMed] [Google Scholar]
7. Cui K, Roh TY, Schones DE, et al. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell.* 2009;4(1):80–93. [PMC free article] [PubMed] [Google Scholar]
8. Orford K, Kharchenko P, Lai W, et al. Differential H3K4 methylation identifies developmentally poised hematopoietic genes. *Dev Cell.* 2008;14(5):798–809. [PubMed] [Google Scholar]
9. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell.* 2007;130(1):77–88. [PMC free article] [PubMed] [Google Scholar]
10. Sawado T, Halow J, Im H, et al. H3 K79 dimethylation marks developmental activation of the beta-globin gene but is reduced upon LCR-mediated high-level transcription. *Blood.* 2008;112(2):406–414. [PMC free article] [PubMed] [Google Scholar]
11. Vakoc CR, Mandat SA, Olenchock BA, Blobel GA. Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol Cell.* 2005;19(3):381–391. [PubMed] [Google Scholar]
12. Fromm G, de Vries C, Byron R, et al. Histone hyperacetylation within the beta-globin locus is context-dependent and precedes high-level gene expression. *Blood.* 2009;114(16):3479–3488. [PMC free article] [PubMed] [Google Scholar]
13. Cheng Y, Wu W, Kumar SA, et al. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res.* 2009;19(12):2172–2184. [PMC free article] [PubMed] [Google Scholar]
14. Yu M, Riva L, Xie H, et al. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell.* 2009;36(4):682–695. [PMC free article] [PubMed] [Google Scholar]
15. Fujiwara T, O'Geen H, Keles S, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell.* 2009;36(4):667–681. [PMC free article] [PubMed] [Google Scholar]
16. Song SH, Kim A, Ragoczy T, Bender MA, Groudine M, Dean A. Multiple functions of Ldb1 required for beta-globin activation during erythroid differentiation. *Blood.* 2010;116(13):2356–2364. [PMC free article] [PubMed] [Google Scholar]
17. Soler E, Andrieu-Soler C, de Boer E, et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev.* 2010;24(3):277–289. [PMC free article] [PubMed] [Google Scholar]
18. Muse GW, Gilchrist DA, Nechaev S, et al. RNA polymerase is poised for activation across the genome. *Nat Genet.* 2007;39(12):1507–1511. [PMC free article] [PubMed] [Google Scholar]
19. Zeitlinger J, Stark A, Kellis M, et al. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet.* 2007;39(12):1512–1516. [PMC free article] [PubMed] [Google Scholar]
20. Rahl PB, Lin CY, Seila AC, et al. c-Myc regulates transcriptional pause release. *Cell.* 2010;141(3):432–445. [PMC free article] [PubMed] [Google Scholar]
21. Gilchrist DA, Dos Santos G, Fargo DC, et al. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell.* 2010;143(4):540–551. [PMC free article] [PubMed] [Google Scholar]
22. Welboren WJ, van Driel MA, Janssen-Megens EM, et al. ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.* 2009;28(10):1418–1428. [PMC free article] [PubMed] [Google Scholar]
23. Gilchrist DA, Nechaev S, Lee C, et al. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes Dev.* 2008;22(14):1921–1933. [PMC free article] [PubMed] [Google Scholar]
24. Bai X, Kim J, Yang Z, et al. TIF1gamma controls erythroid cell fate by regulating transcription elongation. *Cell.* 2010;142(1):133–143. [PMC free article] [PubMed] [Google Scholar]
25. Pokholok DK, Harbison CT, Levine S, et al. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell.* 2005;122(4):517–527. [PubMed] [Google Scholar]
26. Zhang J, Socolovsky M, Gross AW, Lodish HF. Role of Ras signaling in erythroid differentiation of mouse fetal liver cells: functional analysis by a flow cytometry-based novel culture system. *Blood.* 2003;102(12):3938–3946. [PubMed] [Google Scholar]
27. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008;18(11):1851–1858. [PMC free article] [PubMed] [Google Scholar]
28. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008;5(7):621–628. [PubMed] [Google Scholar]

29. Huang da W, Sherman BT, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* 2007;35:W169–W175. Web Server issue. [PMC free article] [PubMed] [Google Scholar]
30. Hattangadi SM, Burke KA, Lodish HF. Homeodomain-interacting protein kinase 2 plays an important role in normal terminal erythroid differentiation. *Blood.* 2010;115(23):4853–4861. [PMC free article] [PubMed] [Google Scholar]
31. Guenther MG, Frampton GM, Soldner F, et al. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell.* 2010;7(2):249–257. [PMC free article] [PubMed] [Google Scholar]
32. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell.* 2007;129(4):823–837. [PubMed] [Google Scholar]
33. Rao B, Shibata Y, Strahl BD, Lieb JD. Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide. *Mol Cell Biol.* 2005;25(21):9447–9459. [PMC free article] [PubMed] [Google Scholar]
34. Bannister AJ, Schneider R, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T. Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem.* 2005;280(18):17732–17736. [PubMed] [Google Scholar]
35. Schuettengruber B, Chourrout D, Vervoot M, Leblanc B, Cavalli G. Genome regulation by polycomb and trithorax proteins. *Cell.* 2007;128(4):735–745. [PubMed] [Google Scholar]
36. Lim PS, Hardy K, Bunting KL, et al. Defining the chromatin signature of inducible genes in T cells. *Genome Biol.* 2009;10(10):R107. [PMC free article] [PubMed] [Google Scholar]
37. Araki Y, Wang Z, Zang C, et al. Genome-wide analysis of histone methylation reveals chromatin state-based regulation of gene transcription and function of memory CD8+ T cells. *Immunity.* 2009;30(6):912–925. [PMC free article] [PubMed] [Google Scholar]
38. Barski A, Jothi R, Cuddapah S, et al. Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res.* 2009;19(10):1742–1751. [PMC free article] [PubMed] [Google Scholar]
39. Dou Y, Milne TA, Tackett AJ, et al. Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF. *Cell.* 2005;121(6):873–885. [PubMed] [Google Scholar]
40. Bernstein BE, Humphrey EL, Erlich RL, et al. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A.* 2002;99(13):8695–8700. [PMC free article] [PubMed] [Google Scholar]
41. Li B, Carey M, Workman JL. The role of chromatin during transcription. *Cell.* 2007;128(4):707–719. [PubMed] [Google Scholar]
42. Bernstein BE, Kamal M, Lindblad-Toh K, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell.* 2005;120(2):169–181. [PubMed] [Google Scholar]
43. Pekowska A, Benoukraf T, Ferrier P, Spicuglia S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.* 2010;20(11):1493–1502. [PMC free article] [PubMed] [Google Scholar]
44. Ng HH, Robert F, Young RA, Struhl K. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell.* 2003;11(3):709–719. [PubMed] [Google Scholar]
45. Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T. Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat Cell Biol.* 2004;6(1):73–77. [PubMed] [Google Scholar]
46. Bottardi S, Aumont A, Grosveld F, Milot E. Developmental stage-specific epigenetic control of human beta-globin gene expression is potentiated in hematopoietic progenitor cells prior to their transcriptional activation. *Blood.* 2003;102(12):3989–3997. [PubMed] [Google Scholar]
47. Kim A, Dean A. Developmental stage differences in chromatin subdomains of the beta-globin locus. *Proc Natl Acad Sci U S A.* 2004;101(18):7028–7033. [PMC free article] [PubMed] [Google Scholar]
48. Schubeler D, MacAlpine DM, Scalzo D, et al. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* 2004;18(11):1263–1271. [PMC free article] [PubMed] [Google Scholar]
49. Feng Y, Yang Y, Ortega MM, et al. Early mammalian erythropoiesis requires the Dot1L methyltransferase. *Blood.* 2010;116(22):4483–4491. [PMC free article] [PubMed] [Google Scholar]