Visualization of NGS Data: ngsplot

Typical NGS Work Flow

- Short reads (eg. fastq files)
- Use available QC tools to filter (eg. FastQC)
- Align using bowtie or tophat
- Make sense of aligned reads

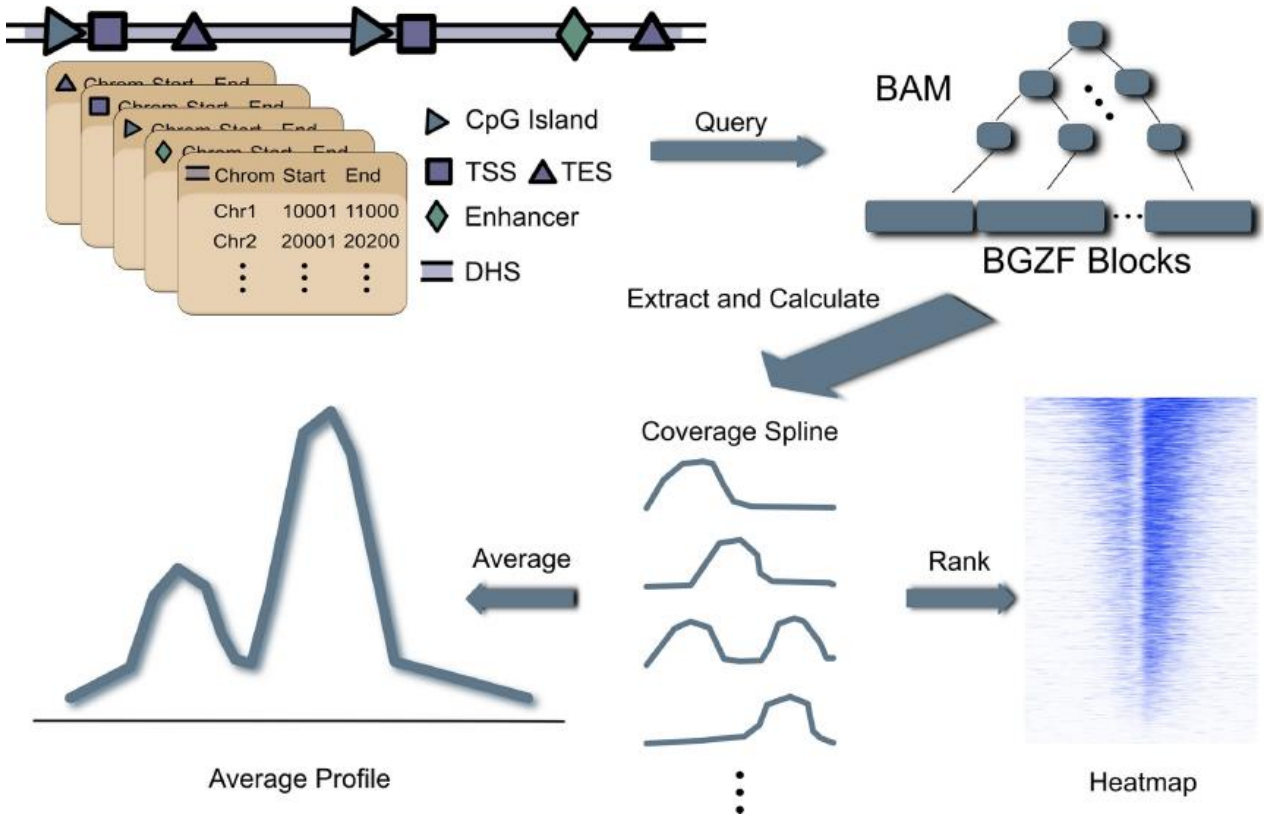
Why ngsplot?

- Summarize data *visually* at functional genomic regions (eg. TSSs, exons, enhancers, etc.)
- Epigenetics:
  - histone modifications or marks enriched near TSS
  - co-occurrence of histone marks
- ChIP-Seq:
  - enrichment of TF in the promoter or gene body
- Genome browsers may not capture enrichment information since they're good at viewing *slices* of the genome.

Overview: ngsplot

Step 1: Defining region(s) of interest
Input: species and regions from ngsplot database, BAM file from alignment

Step 2: Meaningful plots in the defined region(s)
Output: Average profile plot and heatmap images



Shen, L., et al. (2014)

ngsplot: Database

Item	Command-line Flag	Count	Description
Annotation sources	-D	4	**Refseq**, **Ensembl**, ENCODE, muENCODE
Species	-G	17	**human**, chimpanzee, macaque, **mouse**, rat , cow, horse, chicken, **zebrafish**, **drosophila**, **C.elegans**, **S. cerevisiae**, S.pombe , H.pylori, S.acidocaldarius, **A.thaliana**, Z.mays
Biotypes	-R	7	TSS, TES, genebody, exon, CGI, DHS, enhancer
Gene type	-F	5	Protein coding, lincRNA, miRNA, pseudogene, misc
Exon types	-F	7	canonical, promoter, polyA, variant, altDonor, altAcceptor, altBoth
CGIs	-R	10	Hg18, hg19, mm9, mm10, rn4, rn5, bosTau6, galGal4, panTro4, rheMac2
Enhancers	-R, -F	9 (hg19)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/ Cell types: H1hesc (default), Gm12878, Hepg2, Hmec, Hsmm, Huvec, K562, Nhek, Nhlf.
		15 (mm9)	http://chromosome.sdsc.edu/mouse/download.html. Cell types: mESC, bone marrow, cerebellum, cortex, heart, intestine, kidney, liver, lung, MEF, olfactory bulb, placenta, spleen, testes, thymus.
Dnase I hypersensitive sites (DHS)	-R, -F	125(hg19)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgDnaseUniform/ Cell types: H1hesc, A549, Gm12878, Helas3, Hepg2, Hmec, Hsmm, Hsmmtube, Huvec, K562, Lncap, Mcf7, Nhek, Th1,
Region analysis	-F	8	ProximalPromoter, Promoter1K, Promoter3K, Genebody, Genedesert, OtherIntergenic, Pericentromere, Subtelomere

Modified from Shen, L., et al. (2014)

Hands-on

Listing available genomes using `ngsplotdb` command

ngsplot: Implementation

- Find genomic coordinates on regions of interest
  - use ngsplot database with predefined regions
  - custom regions in BED file
- Normalization and coverage based on aligned reads
- Generate plots: average profile and heatmaps

ngsplot: Two-step Normalization

- Length normalization since regions are of different lengths
  - Extend genomic region by the expected fragment length (-FL option, default value is 100) on both sides and overlap aligned reads (extended to fragment length)
  - Calculate single base resolution coverage using two options,
    - spline fit through all data points and sample 101 points at equal interval
    - Make 101 equal-sized bins and calculate averaged value in each bin
- Library size normalization (eg. RPM)

Running ngsplot

- Main script for plotting: `ngs.plot.r`
- Written in R and Python
- Open source
- Enter `ngs.plot.r` at the command prompt to get usage

ngs.plot.r arguments

- Mandatory arguments

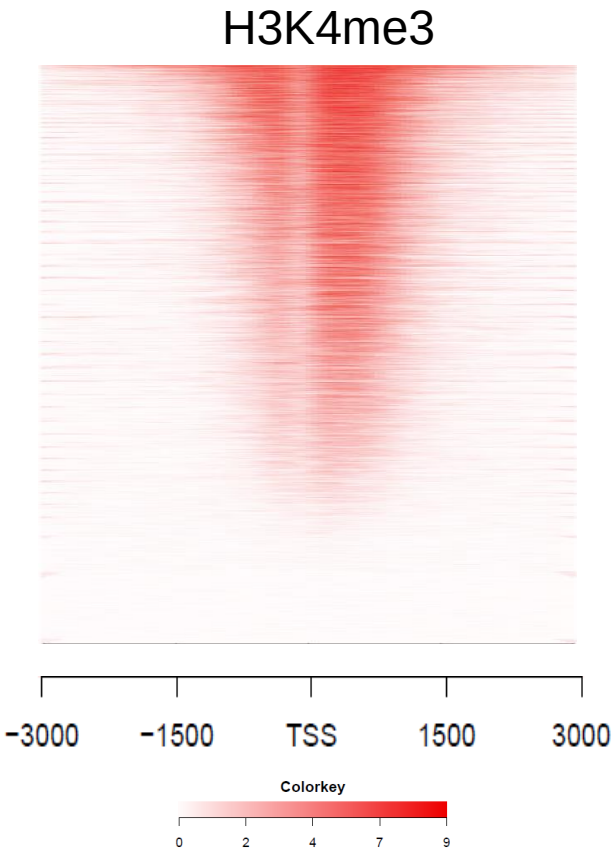
Argument	Explanation
-G	Genome name (hg19, mm9,...)
-R	Genomic regions to plot (tss, tes, genebody, exon,…)
-C	Bam file or a configuration file for multiple plot
-O	Name of output

- Optional arguments (incomplete list)

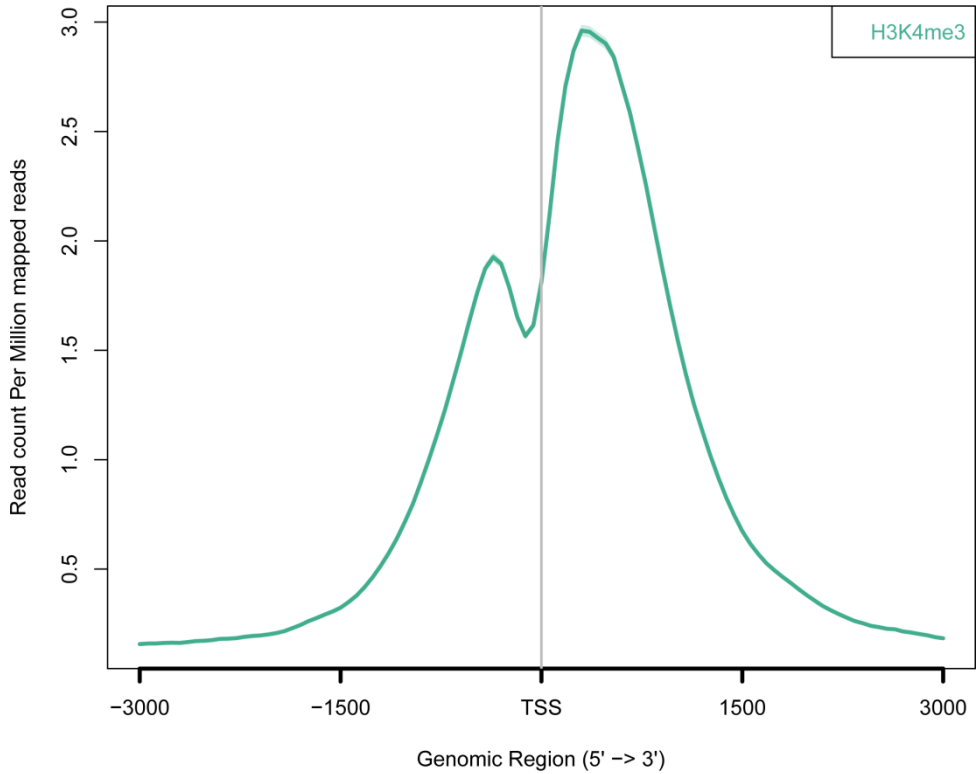
Argument	Explanation
-AL	Algorithm to normalize coverage vectors (spline or bin)
-GO	Gene order algorithm (total, hc, max,…)
-FL	Fragment length (eg. fragment size from experiment)
-D	Gene database (ensembl, refseq)
-E	Gene list to subset regions
-L	Flanking region size (in bases)
-T	Image title/name

Hands-on

- Examining enrichment near TSS



Heatmap showing H3K4me3 enrichment (by color intensity and region) near TSS, where each row is a gene.



Average profile plot summarizing the heatmap (left), note: all genes/features are now collapsed. H3K4me3 enrichment can be clearly seen near the TSS. The two peaks can be also seen on the heatmap (left) by the two distinct banding pattern separated by the TSS.

ngsplot: Numbers Behind the Plots

- ngsplot creates a zip file output which contains this information.

Average Profile Plot

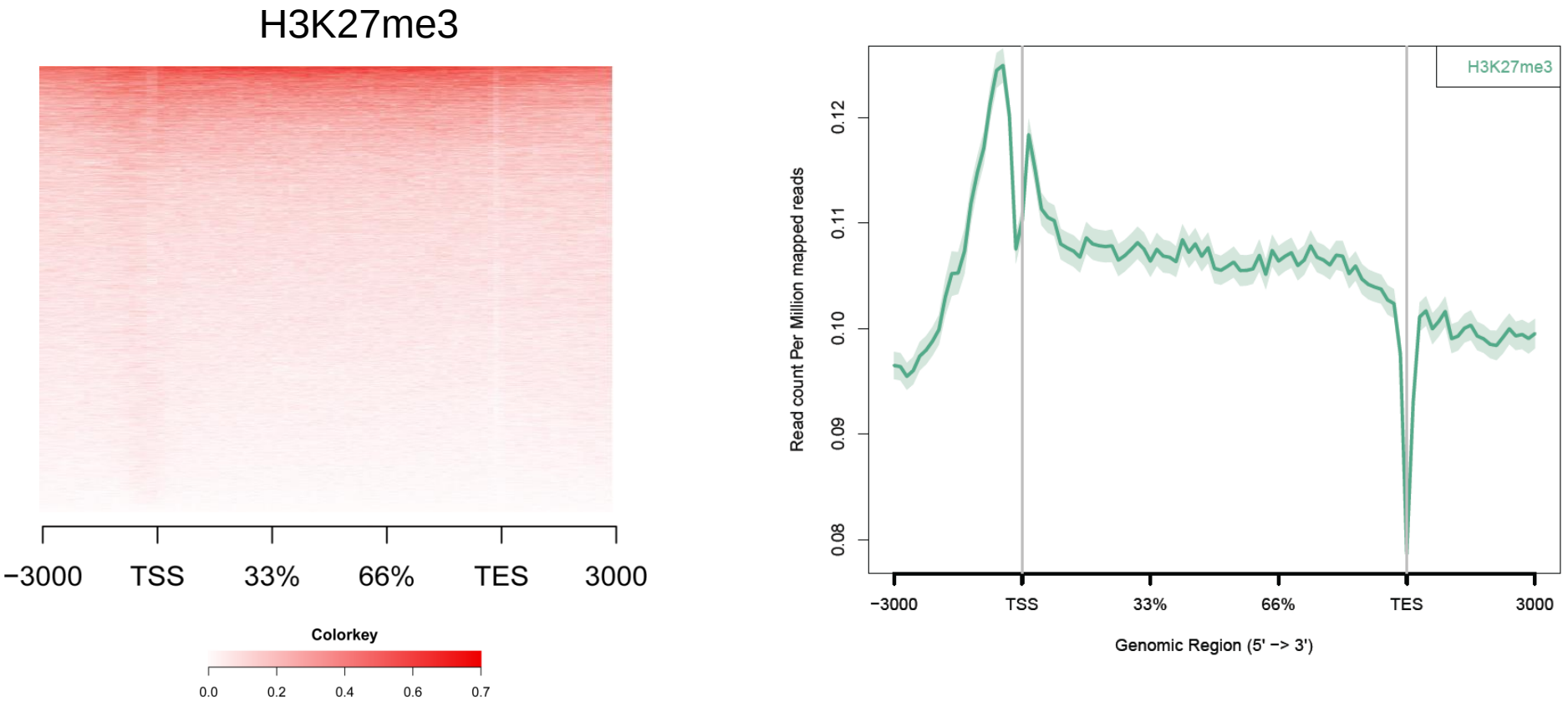
Position	RPM
1	0.157436
2	0.159852
3	0.160182
4	0.162543
.	.
.	.
.	.
101	0.184014

Heatmap

gid	gname	tid	strand	1	2	3		100	101
ENSG00000121410	A1BG	ENST00000263100	-	0	0	0	. . .	0	0
ENSG00000215277	C14orf164	ENST00000399910	+	3.3264	5.544	6.652	. . .	0	0
	.								
	.								
	.								
ENSG00000074755	ZZEF1	ENST00000381638	-	0	0	0	. . .	0	0
ENSG00000036549	ZZZ3	ENST00000370798	-	0	0	0	. . .	0	0

Hands-on

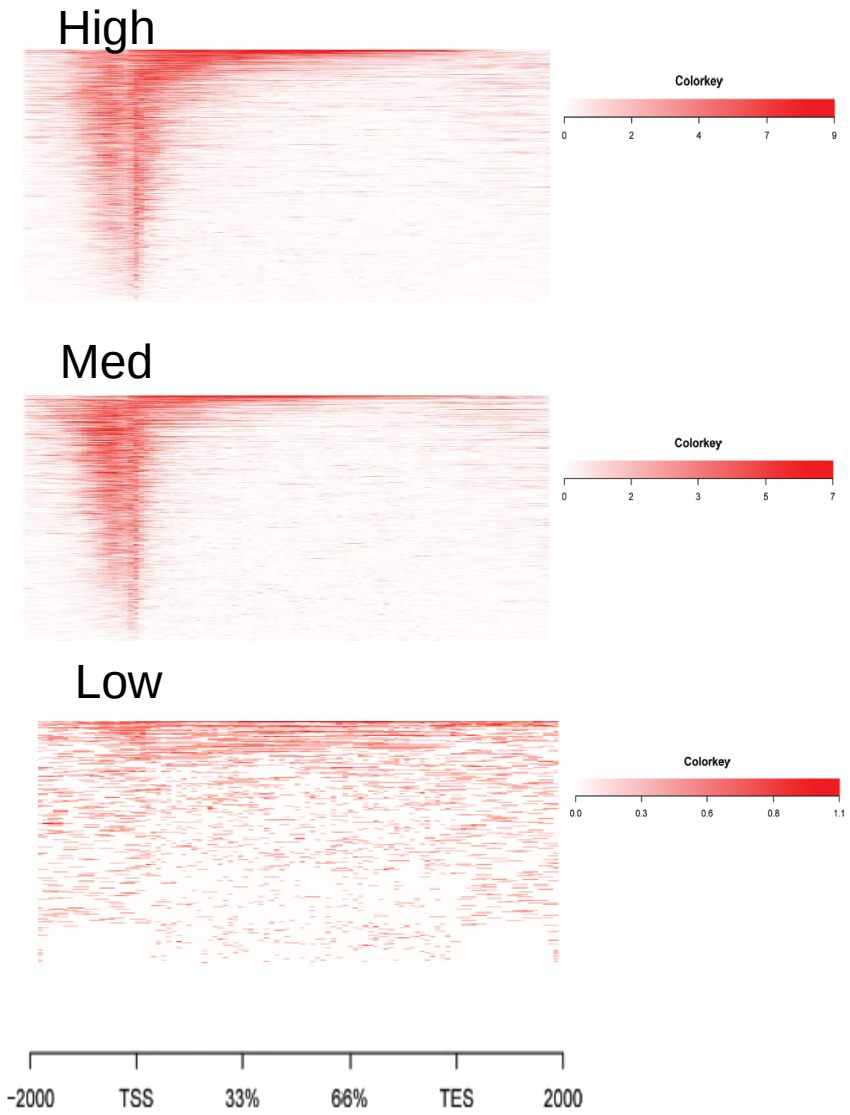
- H3K27me3 distribution across the gene body

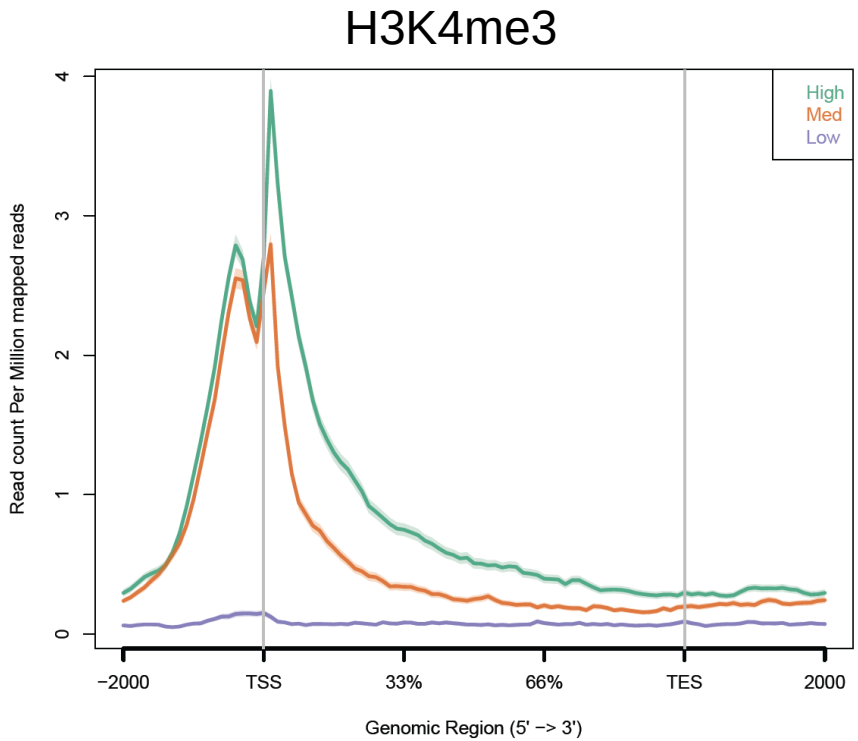


Heatmap (left) and average profile (right) showing H3K27me3 modifications across the gene body. The enrichment near the TSS may not be obvious from the heatmap. On the average profile plot, the lighter green shade represents the standard error (SEM).

Hands-on

- Multiple plots on the same graph with different subsets of genes

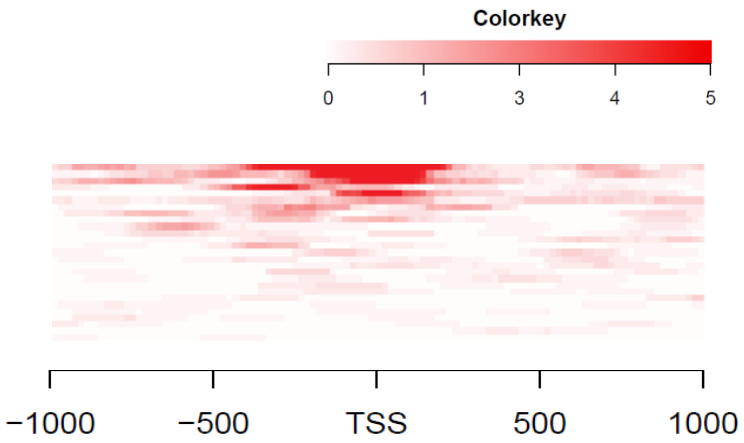


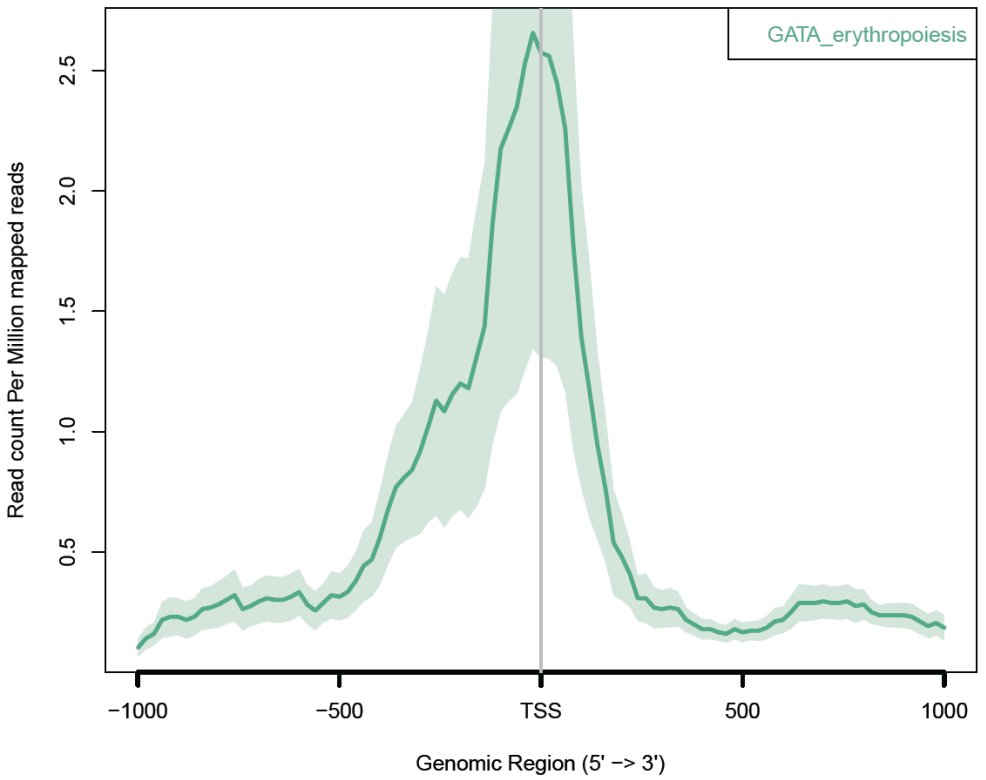


Heatmaps(left) and average profile plot with three different subsets of genes (low, medium, high) based on expression level. Genes that are highly/moderately expressed have an enrichment for H3K4me3 near the TSS that's not seen in lowly expressed genes.

Hands-on

- TF binding site near the TSS

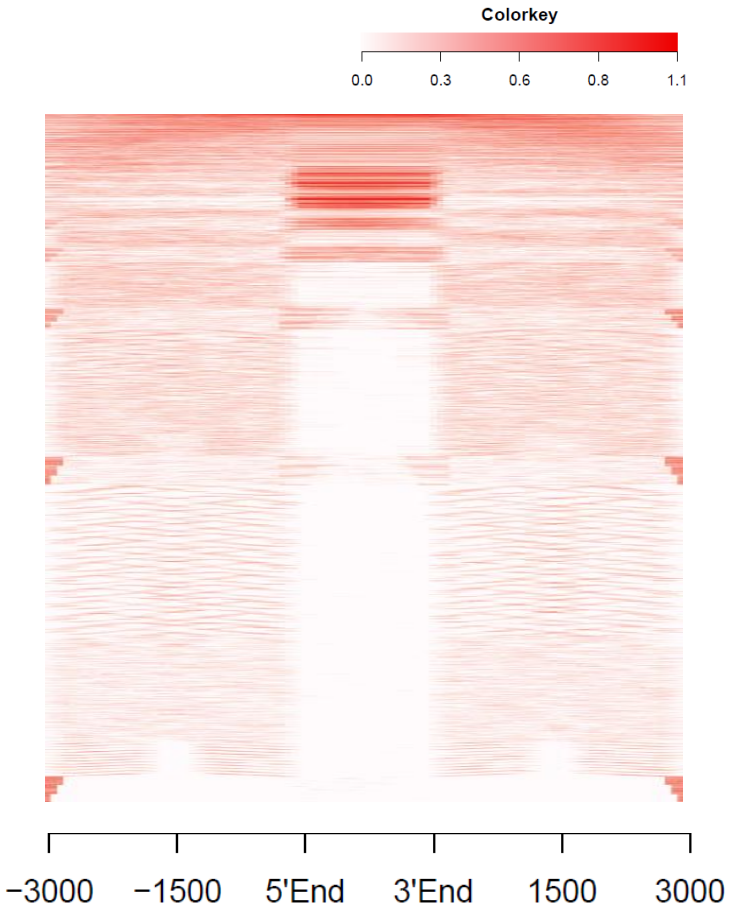


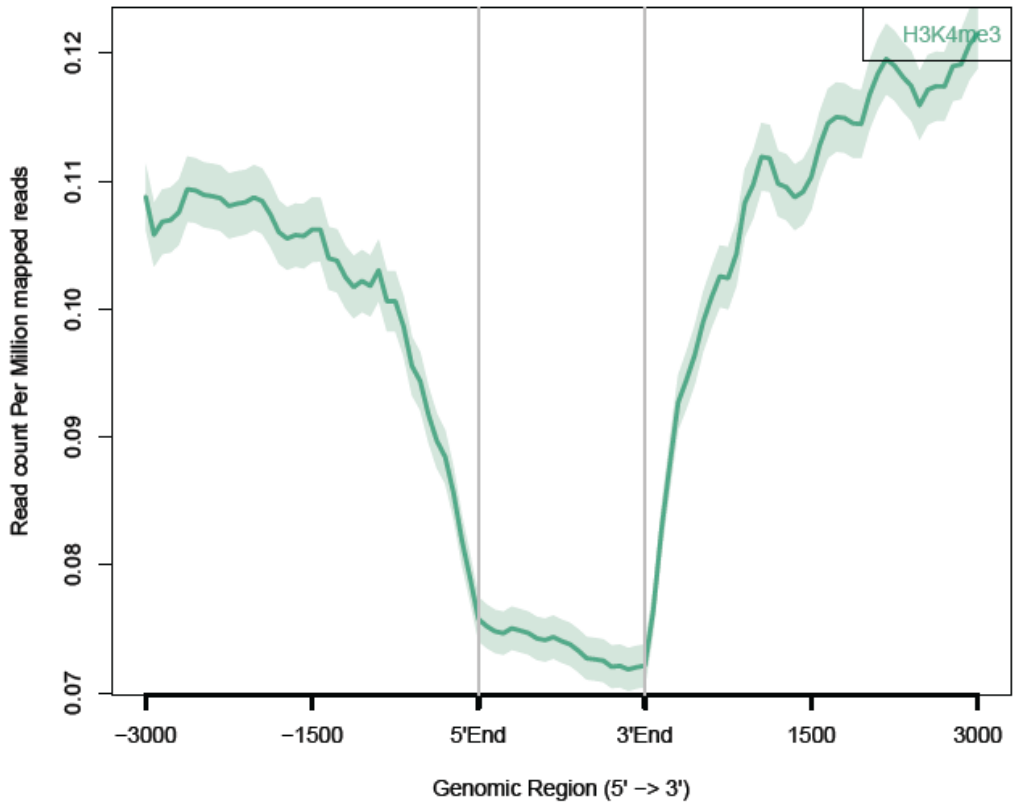


GATA2 binding site (peak) can be clearly seen in the subset of erythropoiesis genes.

Hands-on

- H3K4me3 in repeats

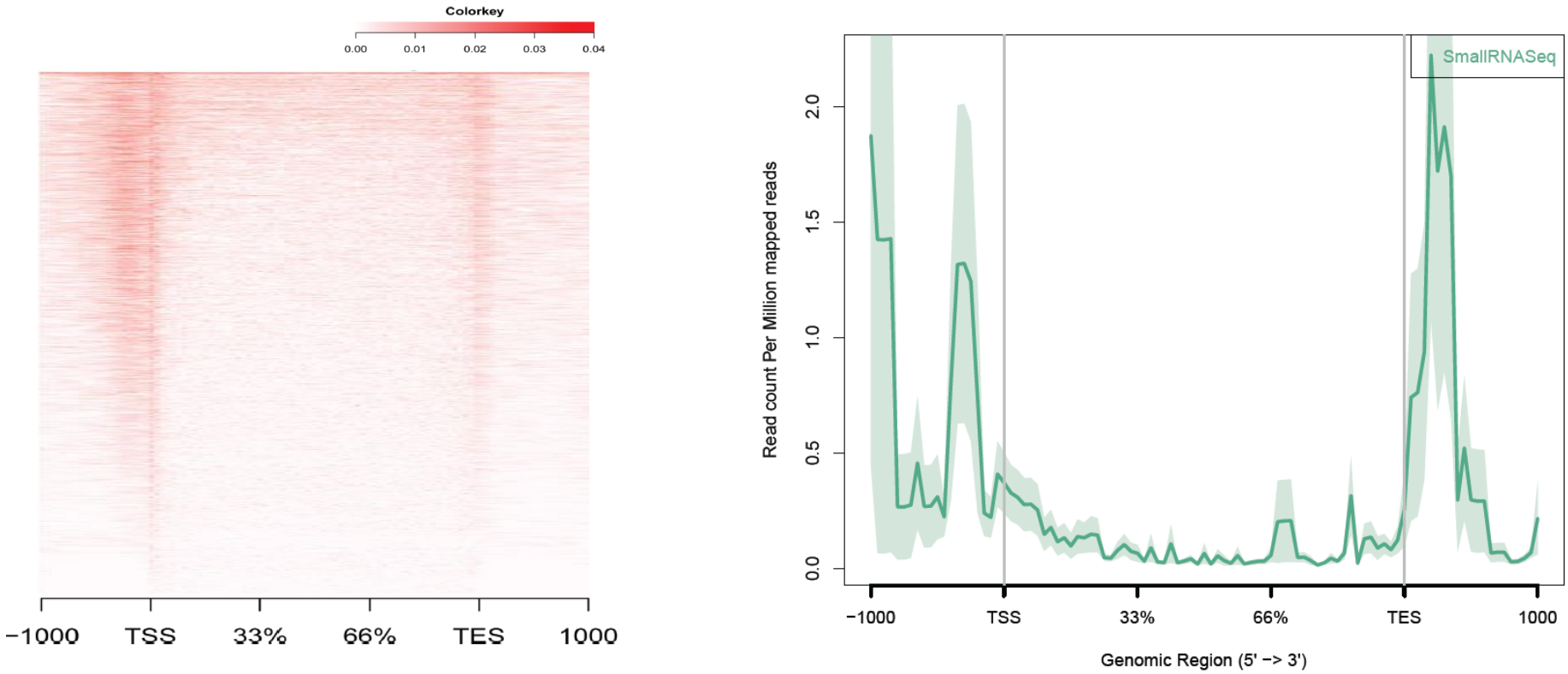




Both the heatmap and average profile plot showing a possible depletion of H3K4me3 marks in the custom region of repeats. Alternatively, it could be reads are unable to map to the repeat regions.

Hands-on

- Small RNA-Seq coverage across the gene body



Heatmap (left) and average profile plot (right) showing small non-coding RNAs found outside gene bodies.

ngsplot: Ranking Genes

- Genes on the heatmap can be ranked in different orders (-GO option):
  - Total (default)
  - Hierarchial clustering
  - Max
  - Product
  - Difference
  - Principal Component Analysis (PCA)
  - none

More Information

- ngsplot wiki: https://code.google.com/p/ngsplot/
  - contains additional information and features in ngsplot
  - more examples and datasets
- Shen, L., et al. *ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases* BMC Genomics 15:284 (2014)