Comparison of two H3K27me3 ChIP-seq data sets from mouse CD8+ T cells responding to influenza infection

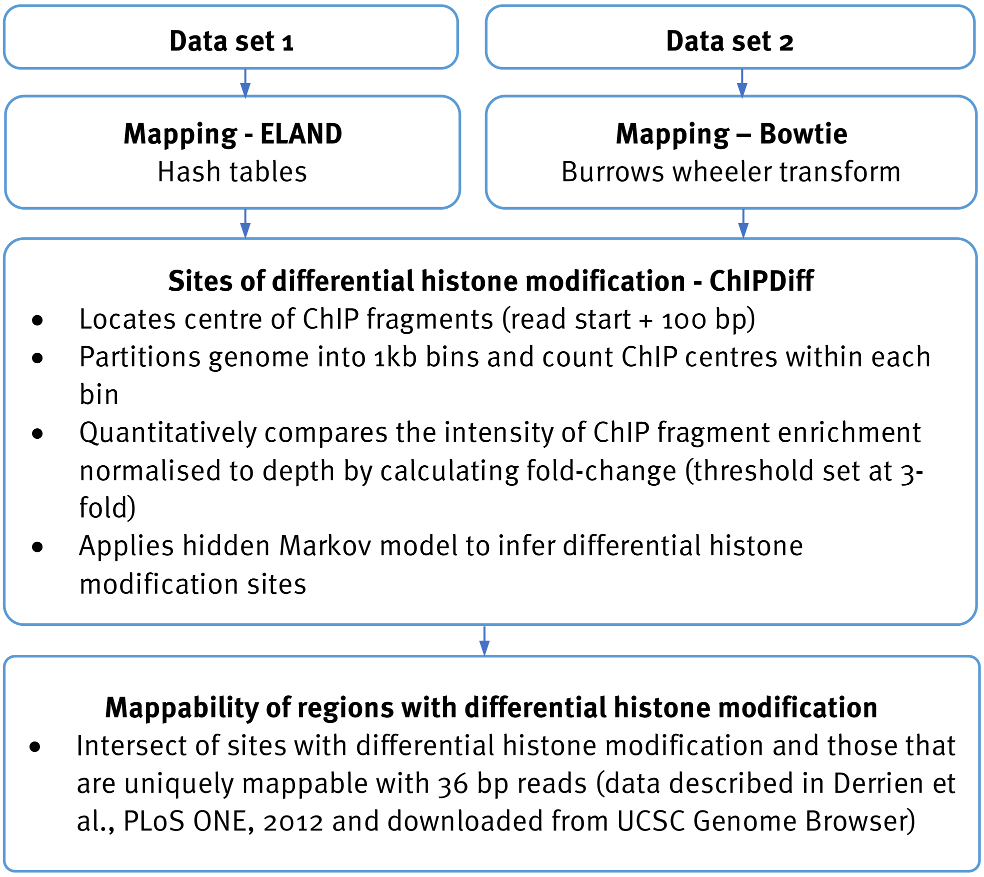
Table 1. Features of ChIP-seq data sets

	Data set 1	***Data set 2***
Mouse genotype	WT	CD4+ T cell deficient
Date	May 2010	April 2012
Sequencer	Illumina Genome Analyser II	Illumina HiSeq 2000
Read length	35 bp	100 bp
Depth (million reads)	52	130

Data sets 1 and 2 are subject to biological and technical variation as well as differences in read length and sequencing depth. Moreover, mapping of sequencing reads to the mouse genome was initially performed using different tools (ELAND vs Bowtie). In spite of these limitations an overview of differences between the data was performed.

Aim

Quantify and compare H3K27me3 enrichment in data sets 1 and 2



Data set 1

Data set 2

Mapping - ELAND
Hash tables

Mapping – Bowtie
Burrows wheeler transform

Sites of differential histone modification - ChIPDiff

- Locates centre of ChIP fragments (read start + 100 bp)
- Partitions genome into 1kb bins and count ChIP centres within each bin
- Quantitatively compares the intensity of ChIP fragment enrichment normalised to depth by calculating fold-change (threshold set at 3-fold)
- Applies hidden Markov model to infer differential histone modification sites

Mappability of regions with differential histone modification

- Intersect of sites with differential histone modification and those that are uniquely mappable with 36 bp reads (data described in Derrien et al., PLoS ONE, 2012 and downloaded from UCSC Genome Browser)

Figure 1. Schematic outlining the processing steps applied to data sets 1 and 2

Results

Table 2. Outcome of data processing and comparison

Total number of differentially enriched regions	***2399***
Enriched in data set 2	97%
Occurring in regions of genome unique by 36mers to no greater than 2 sites	15%
Occurring in regions of genome unique by 36mers to one site	6.4%

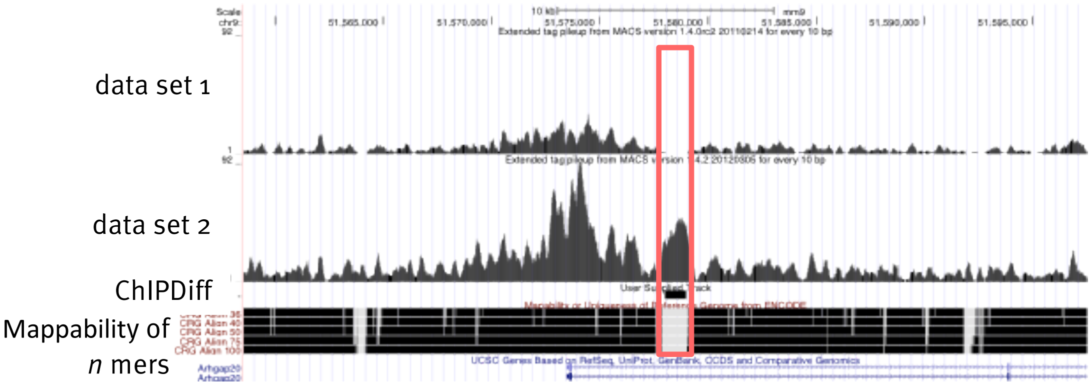

Scale
chr9:
10 kb
mm9
51,565,000
51,570,000
51,575,000
51,580,000
51,585,000
51,590,000
51,595,000
Extended tag pileup from MACS version 1.4.0rc2 20110214 for every 10 bp
data set 1
Extended tag pileup from MACS version 1.4.2 20120305 for every 10 bp
data set 2
ChIPDiff
User Supplied Track
Mappability or Uniqueness of Reference Genome from ENCODE
Mappability of
n mers
CRG Align 36
CRG Align 40
CRG Align 50
CRG Align 75
CRG Align 100
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Arhgap20
Arhgap20

Figure 2. Peak of H3K27me3 detected in region of low mappability in data set 2. ChIP-seq data and regions called by ChIPDiff were viewed on the UCSC Genome Browser alongside a track showing mappability for reads of different length. The box highlights a region flagged by ChIPDiff, where a peak is observed only in data set 2 in a region of low mappability.

Comparison of data sets 1 and 2 using ChIPDiff revealed 2399 regions of the genome with a greater than 3-fold difference in the level of H3K27me3, 97% of which were enriched in data set 2. The data were viewed using the UCSC Genome Browser and a pattern of enrichment in data set 2 over regions with low uniqueness emerged (Figure 1). To quantify this difference, regions called by ChIPDiff overlapping genomic regions of low mappability were counted. Only 15% of the enriched regions fell within areas of the genome mappable to no more than 2 locations by 36mers, and only 6.4% were uniquely mappable. This is a low proportion given that 79.92% of the mouse genome as a whole is uniquely mappable by 36mers (Derrien *et al.*, *PLOS One*, 2012) and suggested that one of the contributors to the difference in peaks called between the data sets was read length bias. The data were also sequenced to different depth and mapped using different tools, though ChIPDiff normalizes for depth reducing the impact of this factor.

A second analysis approach was applied to eliminate read length bias and variability introduced by mapping algorithm. Reads in data set 2 were first trimmed at the 3` end to 35 bp length, and both data sets were then mapped using Bowtie and compared using ChIPDiff (Figure 2). This analysis revealed only 4 differentially enriched regions.

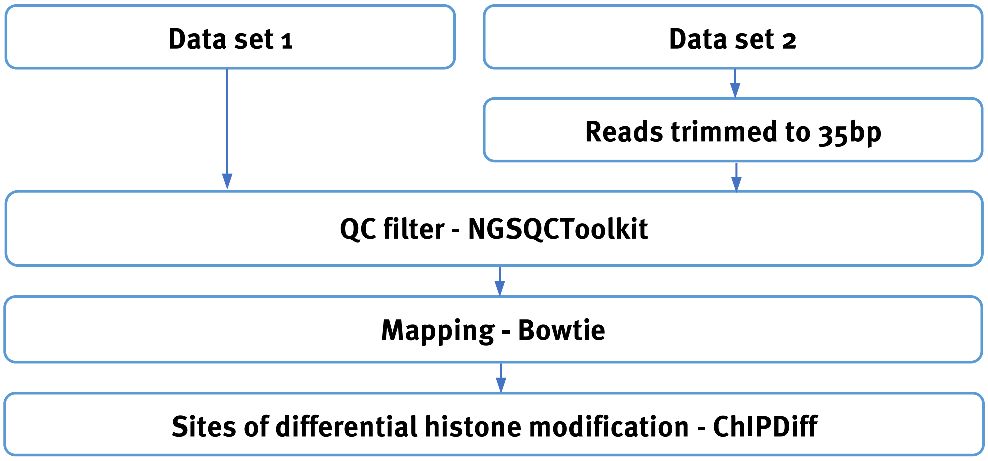


Figure 2. Schematic outlining follow-up processing steps applied to data sets 1 and 2

Concluding remarks

Although the contributions of biological and other factors to the differences between data sets 1 and 2 were not differentiated, controlling for read length greatly reduced the number of H3K27me3 peaks called as differentially enriched by ChIPDiff.