
Contents

CHAPTER	1 ■ Introduction to ChIP-seq	1
	BORBALA MIFSUD	
1.1	CHIP-SEQ EXPERIMENT	1
1.2	IMPROVED DETECTION PROTOCOLS	4
1.2.1	ChIP-exo	4
1.2.2	ChIP-nexus	4
1.2.3	CUT&RUN	4
1.2.4	DamID	5
1.3	CHIP-SEQ DATA ANALYSIS WORKFLOW	5
1.4	DESIGNING A CHIP-SEQ EXPERIMENT	6
1.4.1	ChIP-seq controls	6
1.4.2	Sources of bias	6
1.4.3	Antibody quality	7
1.4.4	Read depth	8
1.4.5	Read properties	8
1.4.6	Replicates	9
CHAPTER	2 ■ Getting Started	11
	ANAÏS BARDET	
2.1	CHIP-SEQ DATASETS	11
2.2	COMPUTATIONAL REQUIREMENTS	13
2.2.1	Computing environment	13
2.2.2	Data	13
2.2.3	Software	14
2.2.4	File formats	15
2.3	DATA RETRIEVAL FROM GEO	15

2.4	CODING TIPS	16
2.5	GRAPHICAL USER INTERFACE TOOLS	19
CHAPTER	3 ■ General Quality Control	21
	KATHI ZARNACK	
3.1	INTRODUCTION	21
3.1.1	FASTQ files	21
3.1.2	Available tools	22
3.2	MEASURES OF HTS DATA QUALITY	22
3.2.1	Selected quality metrics	22
3.2.2	FastQC	25
3.3	TRIMMING AND FILTERING	25
3.3.1	Adapter removal	25
3.3.2	Low-quality trimming	26
3.3.3	Trim Galore!	26
CHAPTER	4 ■ Genomic Alignment	27
	KATHI ZARNACK	
4.1	INTRODUCTION	27
4.1.1	Alignment concepts	27
4.1.2	Available tools	28
4.2	PARAMETERS AND CONSIDERATIONS	28
4.2.1	Mismatches	28
4.2.2	Multi-mapping	29
4.2.3	Other parameters	29
4.2.4	Output format	31
4.3	GENOMIC ALIGNMENT WITH BOWTIE 2	31
CHAPTER	5 ■ ChIP-seq-specific Quality Control	35
	BORBALA MIFSUD	
5.1	CHIP-SEQ-SPECIFIC QUALITY METRICS	35
5.1.1	Signal enrichment	35
5.1.2	Forward and reverse read distribution	38
5.1.3	Duplicate reads	38

5.2	CHIPQC	39
CHAPTER 6	Peak Calling	41
	ANAÏS BARDET	
6.1	CHIP-SEQ SIGNAL TYPES	41
6.1.1	Sharp signal for transcription factors	41
6.1.2	Broad signal for histone marks	42
6.1.3	Mixed signal for RNA polymerase II	42
6.2	GENERAL PEAK CALLING STRATEGY	42
6.2.1	Estimation of fragment size	42
6.2.2	Enrichment of reads	44
6.2.3	Significance score	44
6.2.4	Multiple testing correction	44
6.2.5	Choice of thresholds	44
6.3	EXISTING TOOLS AND CONSIDERATIONS	45
6.3.1	Single-end versus paired-end libraries	45
6.3.2	Sequencing depth and library complexity	46
6.3.3	Experimental resolution	46
6.3.4	New generation of peak callers	47
6.3.5	Post-processing	47
6.4	PEAKZILLA FOR TRANSCRIPTION FACTOR DATA	47
6.5	MACS2 FOR HISTONE MARK DATA	49
6.6	SATURATION ANALYSIS	50
CHAPTER 7	Data Visualisation	53
	ANAÏS BARDET	
7.1	READ DENSITIES	53
7.2	PEAK REGIONS	57
7.3	GENOME BROWSER	57
CHAPTER 8	Comparative Analysis	59
	ANAÏS BARDET AND BORBALA MIFSUD	
8.1	OVERLAP OF PEAK REGIONS	59

8.2	IRREPRODUCIBLE DISCOVERY RATE (IDR)	61
8.2.1	Peak calling for IDR	63
8.2.2	Calculating IDR	64
8.3	COMPARISON OF READ DENSITIES	64
8.3.1	Merging peak regions	65
8.3.2	Counting reads for each sample	66
8.3.3	Normalising read counts	66
8.3.4	Comparing read counts	67
8.4	DIFFERENTIAL BINDING ANALYSIS	70
8.4.1	Using DESeq2	70
8.4.2	DiffBind	73
CHAPTER	9 ■ Downstream Analyses	77
	ANAÏS BARDET AND KATHI ZARNACK	
9.1	GENOMIC CONTEXT	77
9.1.1	Genomic location	77
9.1.2	Distance to genes	80
9.2	FUNCTIONAL ANALYSES	83
9.2.1	Assignment to target genes	83
9.2.2	Gene ontology analysis	83
9.2.3	Other gene-set enrichment analyses	84
9.3	SEQUENCE ANALYSIS	85
9.3.1	Motif analysis	85
9.3.2	Sequence conservation	87
9.4	INTEGRATION WITH OTHER DATASETS	89
9.4.1	Additional ChIP-seq datasets	89
9.4.2	Expression data	91
9.4.3	Other types of data	92
	Bibliography	93
	Index	99