ChIP-seq hands-on

Iros Barozzi, Campus IFOM-IEO (Milan)

Saverio Minucci, Gioacchino Natoli Labs

Main goals

- Becoming familiar with essential tools and formats
- Visualizing and contextualizing raw data

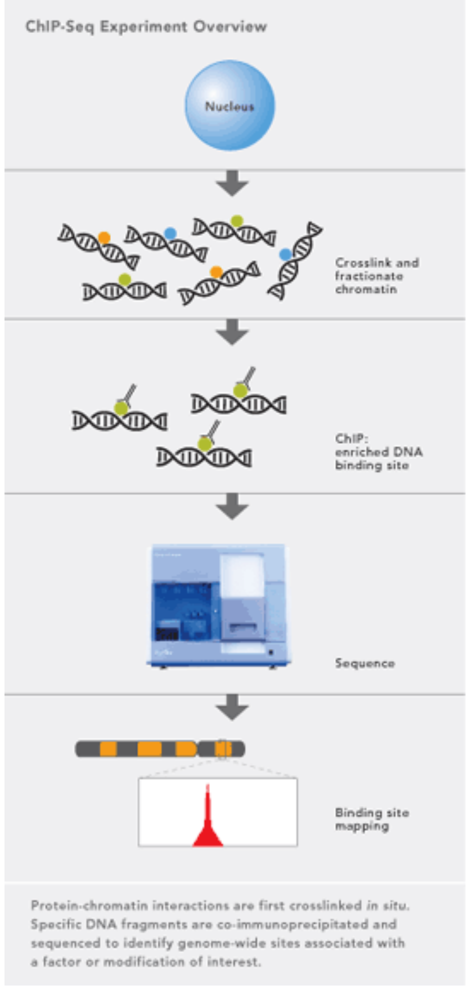
- Understand biases at each step of the analysis
- If something went wrong, identify which experimental step could have risen the issue

- FAQs

Overview

- Quality control
- Alignment
- Raw data visualization
- Peak calling
- Experimental validation

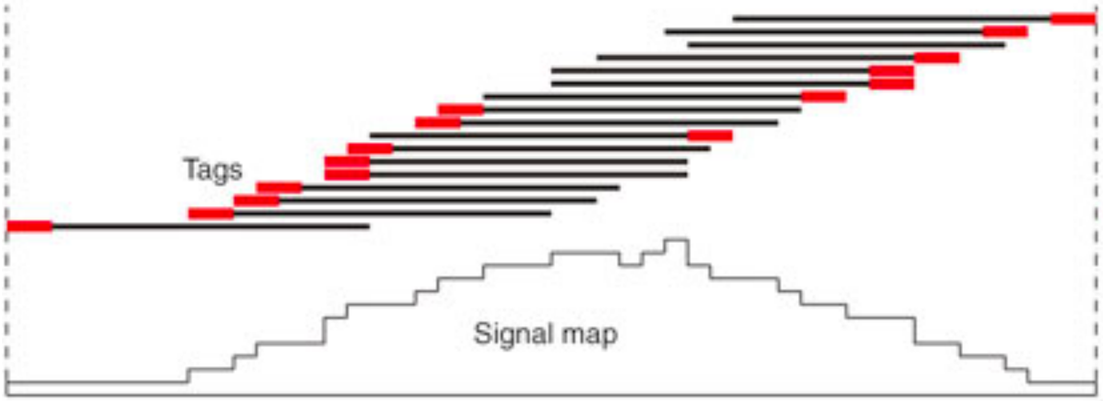
- At each step:
  - critical evaluation
  - understanding possible issues


ChIP-Seq Experiment Overview
Nucleus
Crosslink and fractionate chromatin
ChIP: enriched DNA binding site
Sequence
Binding site mapping

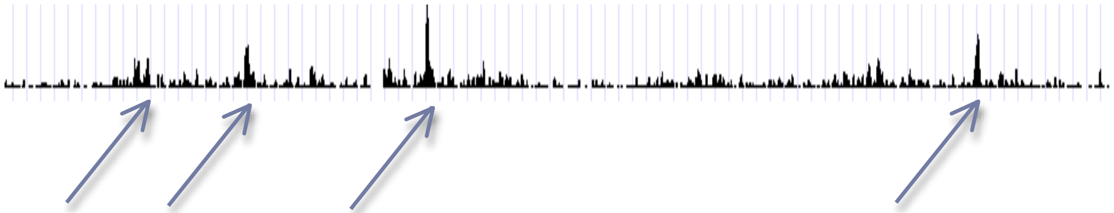
Protein-chromatin interactions are first crosslinked *in situ*. Specific DNA fragments are co-immunoprecipitated and sequenced to identify genome-wide sites associated with a factor or modification of interest.

Illumina website

Overview



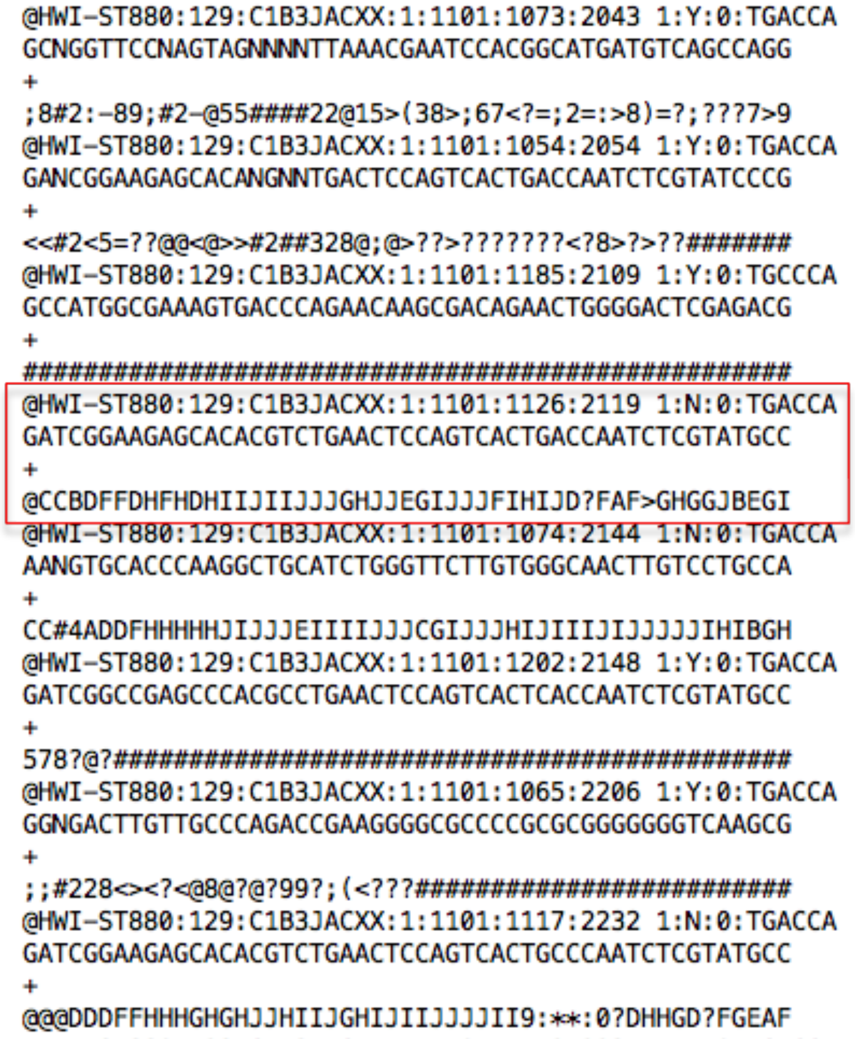
Alignment



Raw data visualization

Peak calling

FASTQ format



```
@HWI-ST880:129:C1B3JACXX:1:1101:1073:2043 1:Y:0:TGACCA
GCNGGTTCCNAGTAGNNNNTTAAACGAATCCACGGCATGATGTCAGCCAGG
+
;8#2:-89;#2-@55####22@15>(38>;67<?=;2=:>8)=?;????7>9
@HWI-ST880:129:C1B3JACXX:1:1101:1054:2054 1:Y:0:TGACCA
GANCGGAAGAGCACANGNNTGACTCCAGTCACTGACCAATCTCGTATCCCG
+
<<#2<5=??@@<@>>#2##328@;@>??>???????<?8>?>??#######
@HWI-ST880:129:C1B3JACXX:1:1101:1185:2109 1:Y:0:TGCCCA
GCCATGGCGAAAGTGACCCAGAACAAGCGACAGAACTGGGGACTCGAGACG
+
###################################################
@HWI-ST880:129:C1B3JACXX:1:1101:1126:2119 1:N:0:TGACCA
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCC
+
@CCBDFFDHFHDHIIJIIJJJGHJJEGIJJJFIHIJD?FAF>GHGGJBEGI
@HWI-ST880:129:C1B3JACXX:1:1101:1074:2144 1:N:0:TGACCA
AANGTGCACCCAAGGCTGCATCTGGGTTCTTGTGGGCAACTTGTCCTGCCA
+
CC#4ADDFHHHHHJIJJJEIIIIJJJCGIJJJHIJIIIJIJJJJJIHIBGH
@HWI-ST880:129:C1B3JACXX:1:1101:1202:2148 1:Y:0:TGACCA
GATCGGCCGAGCCCACGCCTGAACTCCAGTCACTCACCAATCTCGTATGCC
+
578?@?#############################################
@HWI-ST880:129:C1B3JACXX:1:1101:1065:2206 1:Y:0:TGACCA
GGNGACTTGTTGCCCAGACCGAAGGGGCGCCCCGCGCGGGGGGGTCAAGCG
+
;;#228<><?<@8@?@?99?;(<???#########################
@HWI-ST880:129:C1B3JACXX:1:1101:1117:2232 1:N:0:TGACCA
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGCCCAATCTCGTATGCC
+
@@@DDDFFHHHGHGHJJHIIJGHIJIIJJJJII9:**:0?DHHGD?FGEAF
```

Read/Tag

Qscores

@description

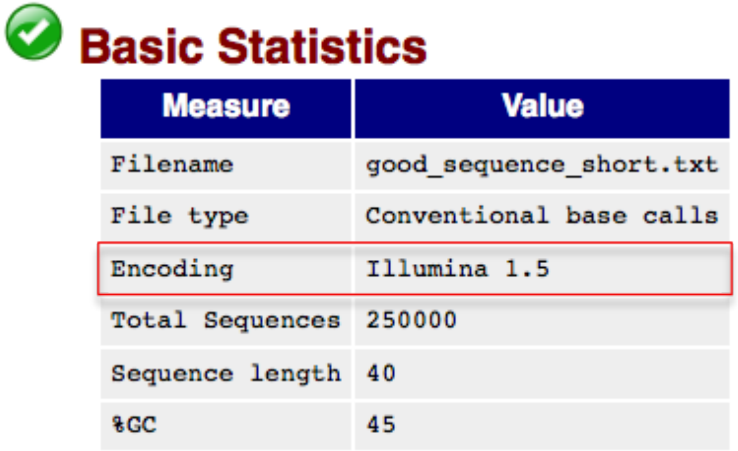
+description

Quality control

- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- critical evaluation of good/bad fastqc output
- what to really expect from a HiSeq lane:
  - trimming
  - contaminants evaluation

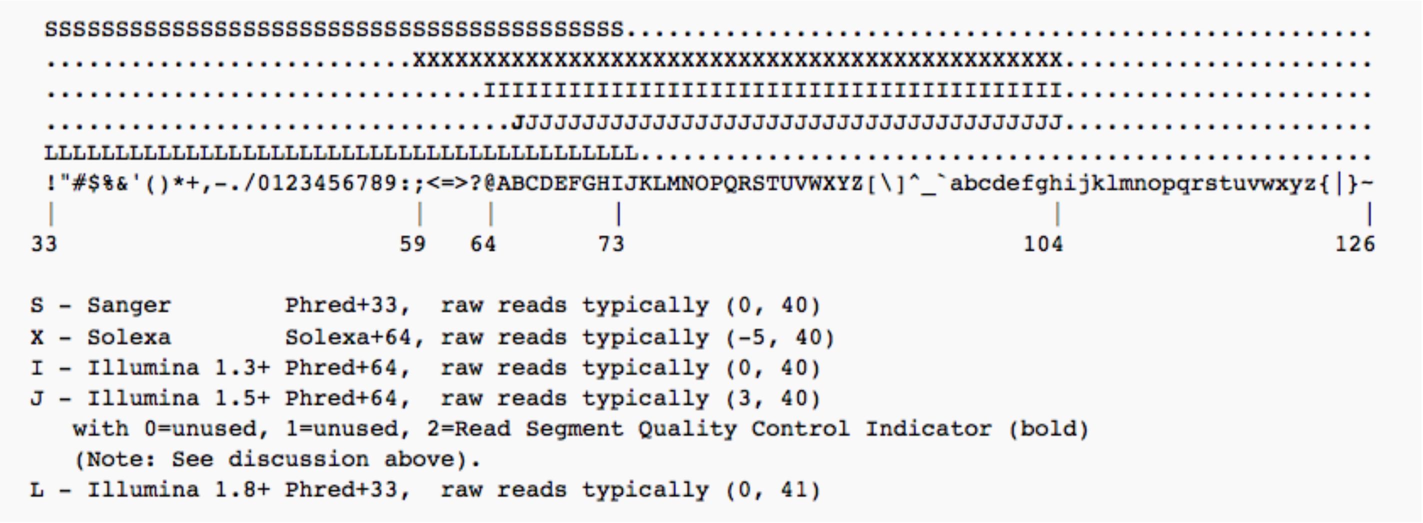
- Before we start: don't get scared, some biases can be intrinsic to the regions of the DNA you are IPing and not a technical problem

Quality control



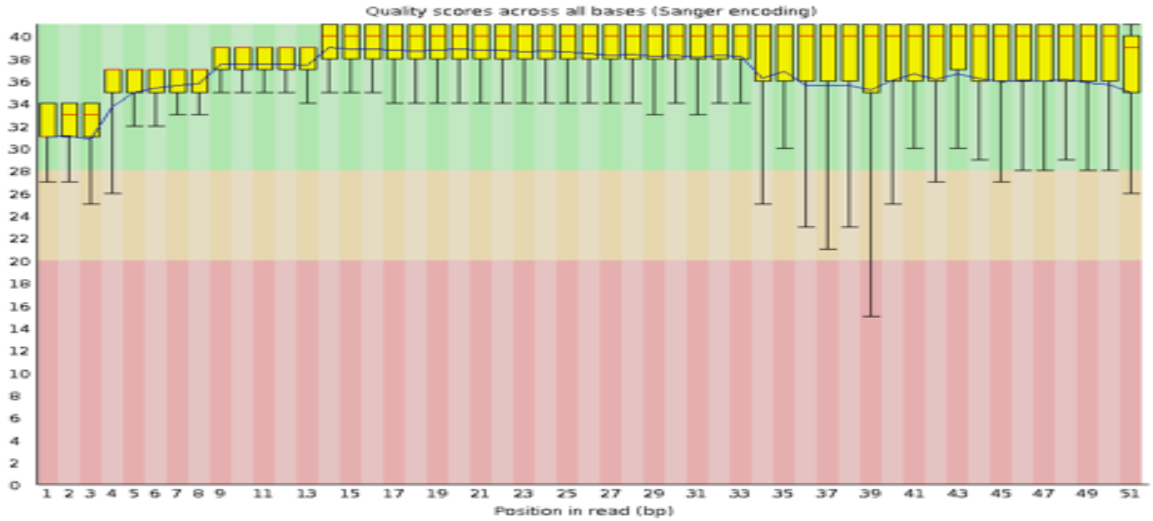
Basic Statistics

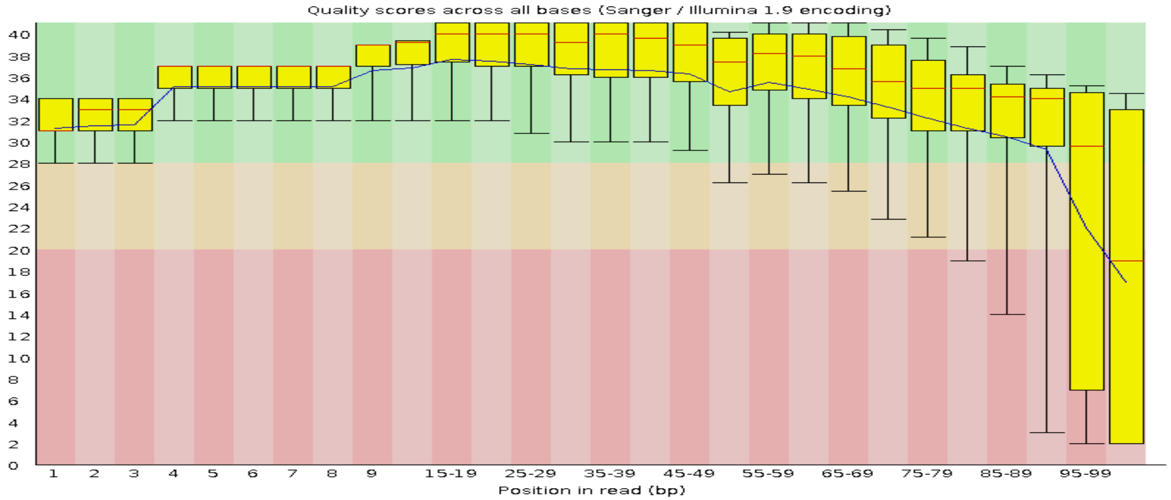
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequence length	40
%GC	45



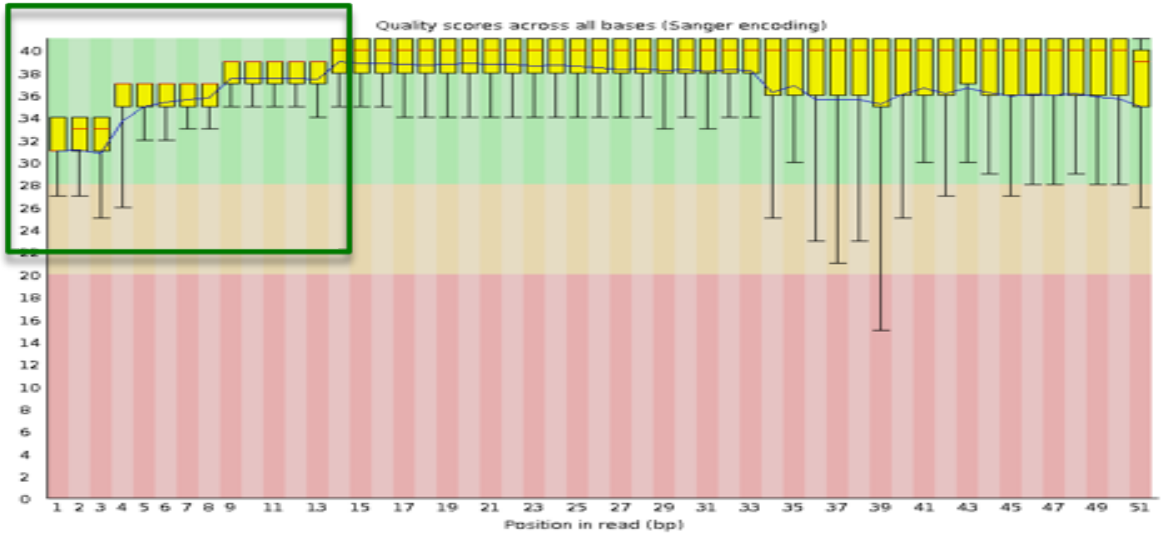
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

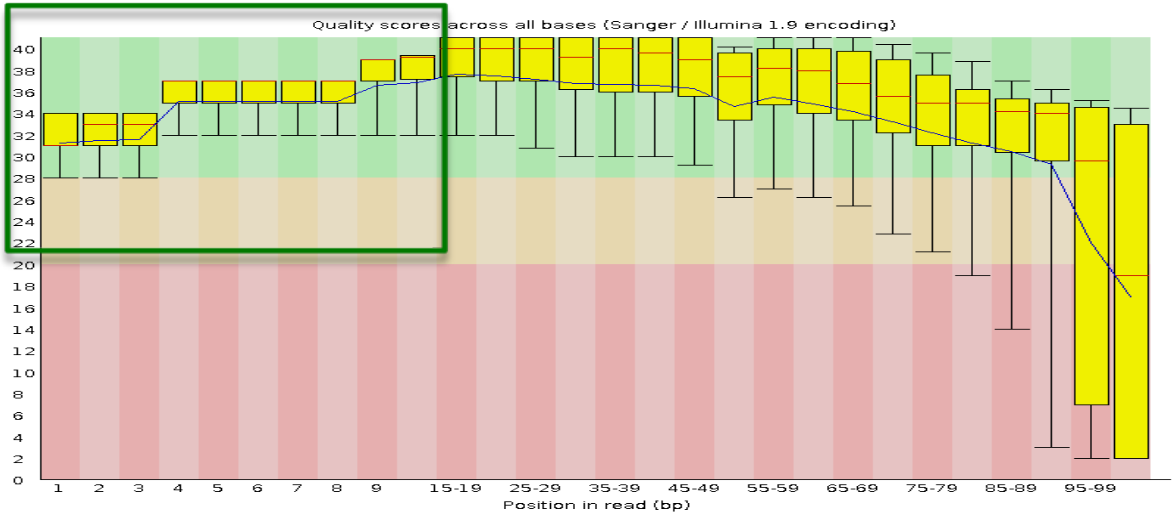
Quality control: HiSeq



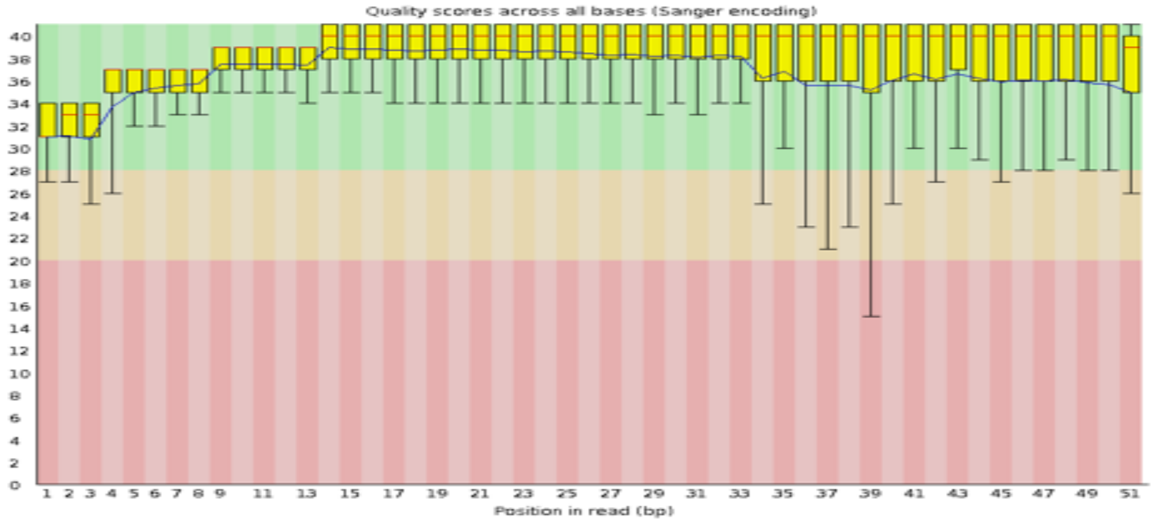


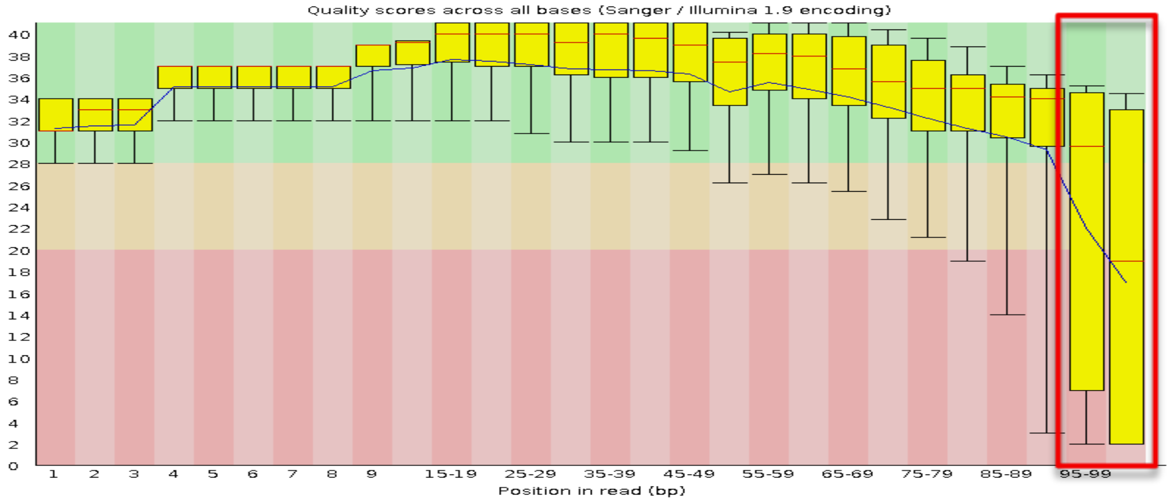
Quality control: HiSeq





Quality control: HiSeq

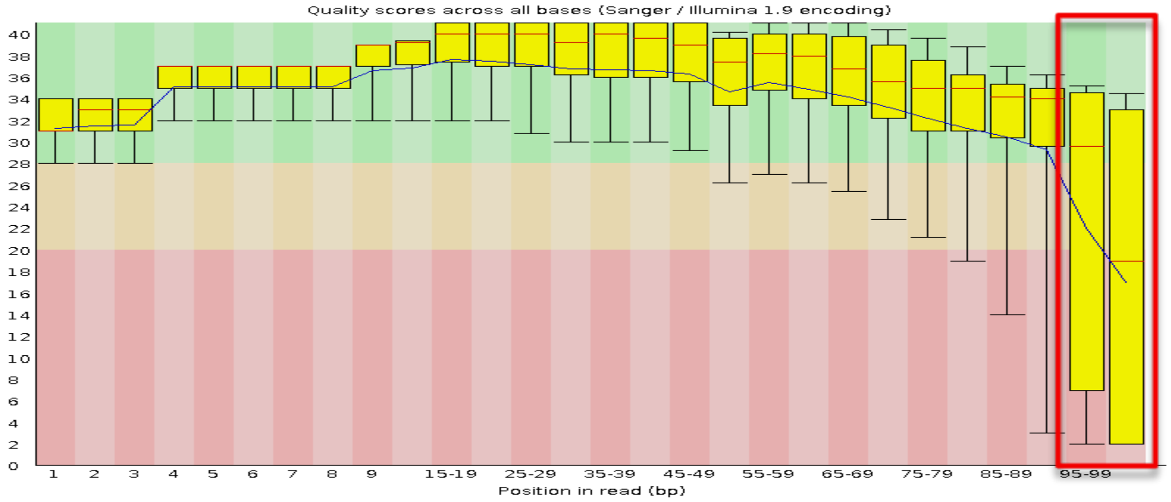




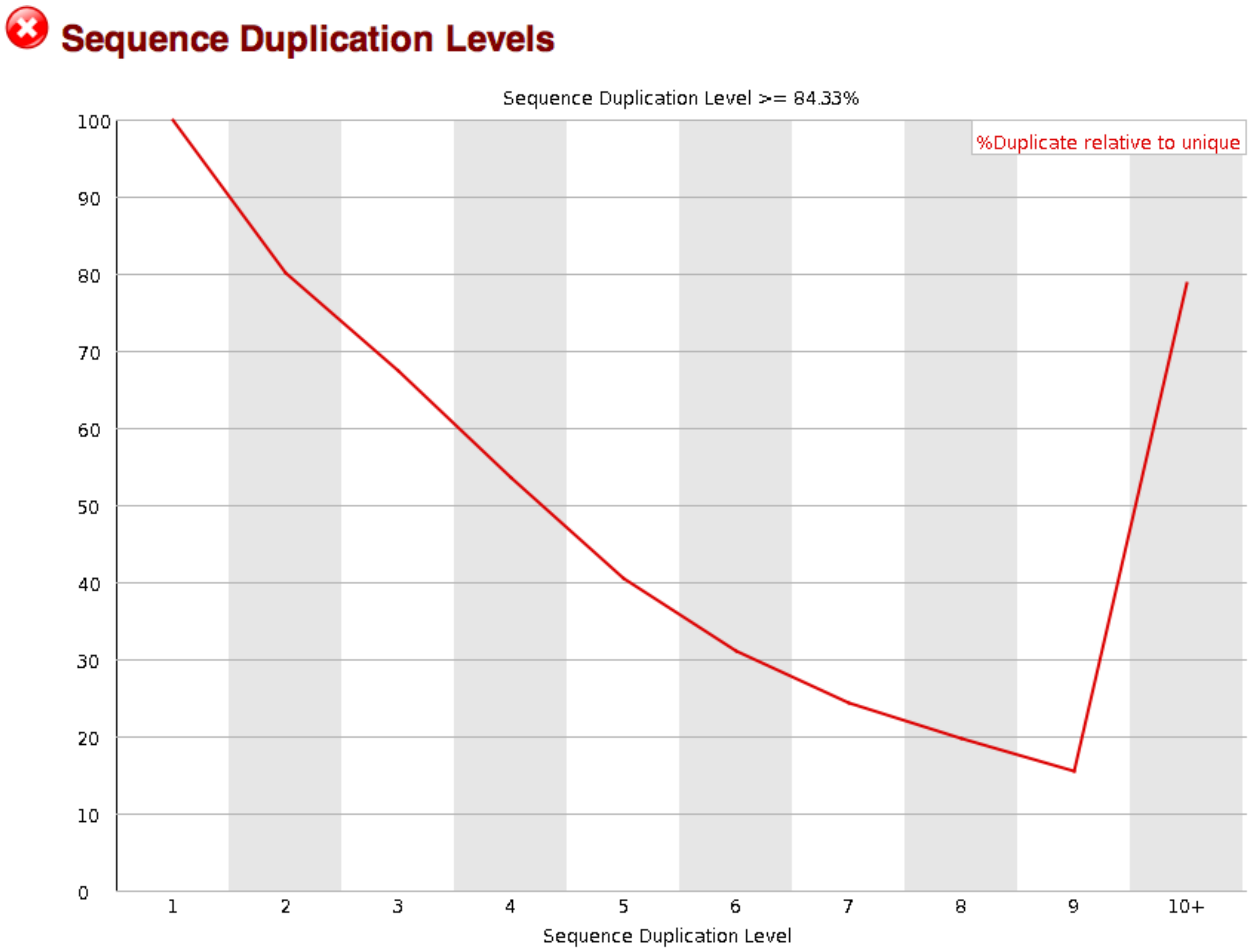
Quality control: HiSeq

Align only these substrings

TTAATTGATTTTATGGATAAAAGTATTGTTTAGATTTGTTTTTTTAATATTGGATTTATATAAAAAAAGATGTTTTTTTTTGTATTTTTTT~~TGTAGTTTAA~~
TGGAGGAATGAGAGGAAAGGTGGAAGTAATTTTGTGTTAATTAAAGGTATTTTTAAAAATTTAATAATATGTATTGAGAAGGAGGTAATTT~~TTCTTTAATT~~
TTTATTTTTAATTTTTTTTTTGTTTTTTGGTGTTTTGTTTTTTTTTGGGTATTTGAGTTTAAATTTTTTAGAATAATTGTAACTGTGAGAA~~AATTGGCATT~~
GTTTTGGGTTTTTAAGGATTTTTTTTTATTTATGTTGTAATTTTTTTAAATGTTAGATTATTTTAGTGATGAGTTTAGGATATATGTGTAT~~TAAGGTGGAT~~
GAAAGAAAAAAAGAAAGAAAGAAATAAAGAAAGAAAGAAGAAAGGAAGGAAGGTTAAGGGAGTGGAGTAGGTAGAGTGAGAAATGTAGAGAT~~CGGAATAGTT~~

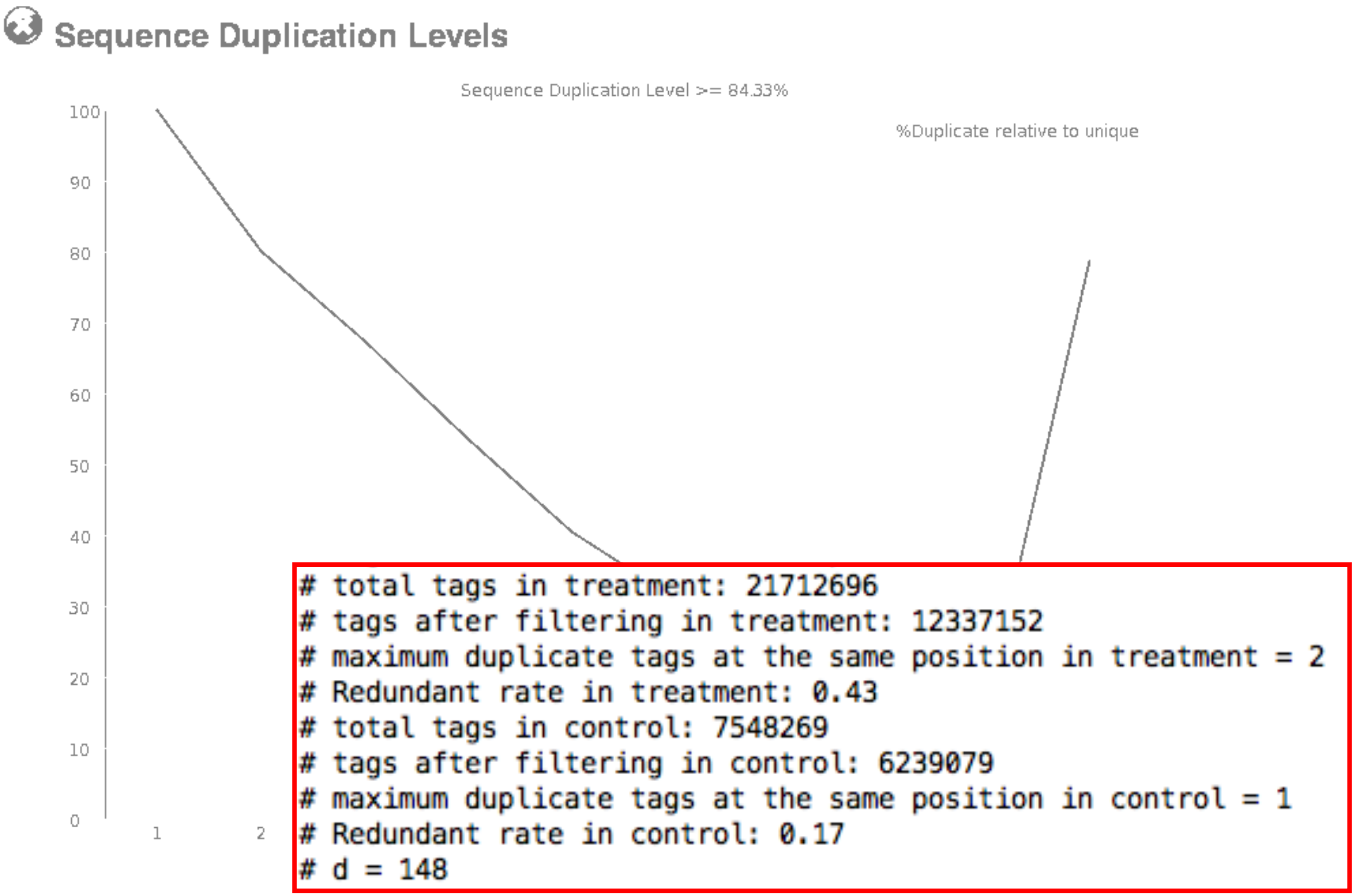


Quality control



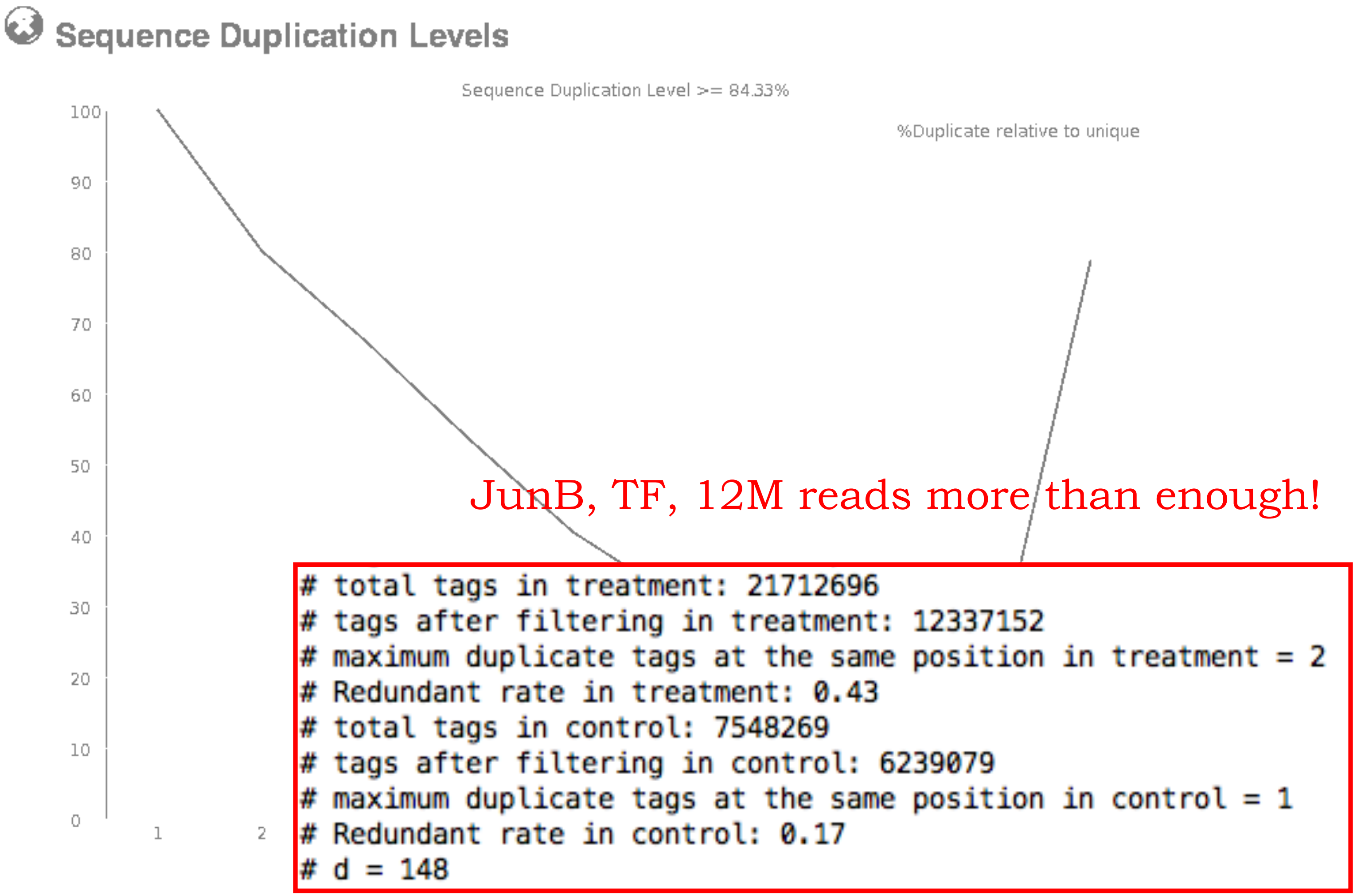
☒ Sequence Duplication Levels

Quality control



```
# total tags in treatment: 21712696
# tags after filtering in treatment: 12337152
# maximum duplicate tags at the same position in treatment = 2
# Redundant rate in treatment: 0.43
# total tags in control: 7548269
# tags after filtering in control: 6239079
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.17
# d = 148
```

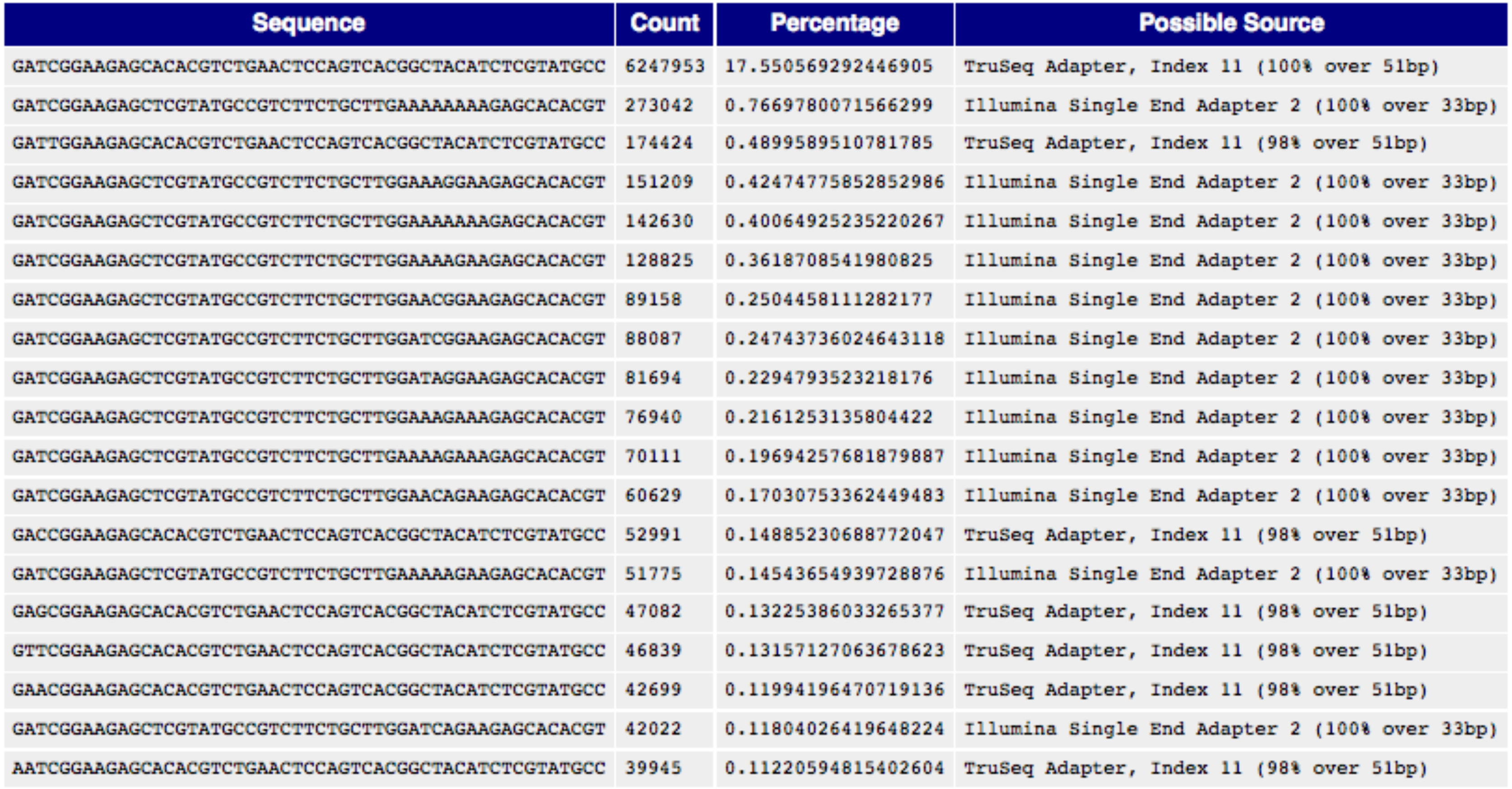
Quality control



JunB, TF, 12M reads more than enough!

```
# total tags in treatment: 21712696
# tags after filtering in treatment: 12337152
# maximum duplicate tags at the same position in treatment = 2
# Redundant rate in treatment: 0.43
# total tags in control: 7548269
# tags after filtering in control: 6239079
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.17
# d = 148
```

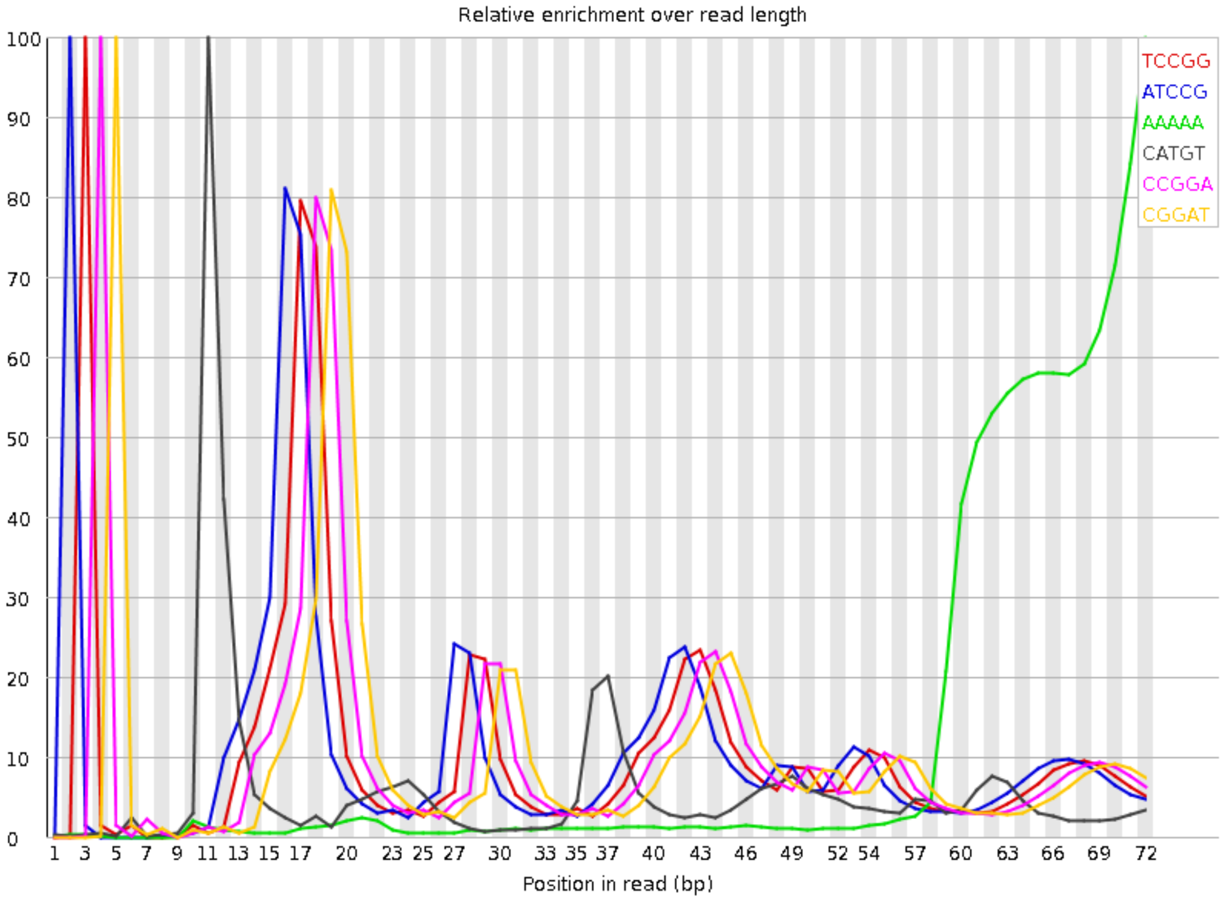
Quality control



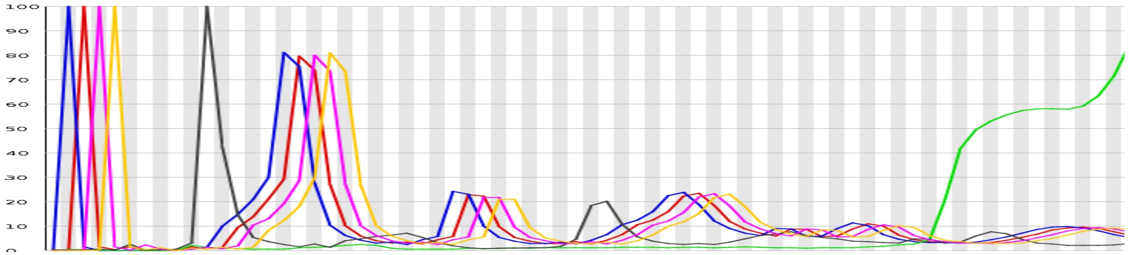
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	6247953	17.550569292446905	TruSeq Adapter, Index 11 (100% over 51bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAAAAAAAGAGCACACGT	273042	0.7669780071566299	Illumina Single End Adapter 2 (100% over 33bp)
GATTGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	174424	0.4899589510781785	TruSeq Adapter, Index 11 (98% over 51bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAAAGGAAGAGCACACGT	151209	0.42474775852852986	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAAAAAAAGAGCACACGT	142630	0.40064925235220267	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAAAAGAAGAGCACACGT	128825	0.3618708541980825	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAACGGAAGAGCACACGT	89158	0.2504458111282177	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGATCGGAAGAGCACACGT	88087	0.24743736024643118	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGATAGGAAGAGCACACGT	81694	0.2294793523218176	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAAAGAAAGAGCACACGT	76940	0.2161253135804422	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAAAGAAAGAGCACACGT	70111	0.19694257681879887	Illumina Single End Adapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAACAGAAGAGCACACGT	60629	0.17030753362449483	Illumina Single End Adapter 2 (100% over 33bp)
GACCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	52991	0.14885230688772047	TruSeq Adapter, Index 11 (98% over 51bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAAAAGAAGAGCACACGT	51775	0.14543654939728876	Illumina Single End Adapter 2 (100% over 33bp)
GAGCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	47082	0.13225386033265377	TruSeq Adapter, Index 11 (98% over 51bp)
GTTCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	46839	0.13157127063678623	TruSeq Adapter, Index 11 (98% over 51bp)
GAACGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	42699	0.11994196470719136	TruSeq Adapter, Index 11 (98% over 51bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGATCAGAAGAGCACACGT	42022	0.11804026419648224	Illumina Single End Adapter 2 (100% over 33bp)
AATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCC	39945	0.11220594815402604	TruSeq Adapter, Index 11 (98% over 51bp)

Quality control

☒ Kmer Content



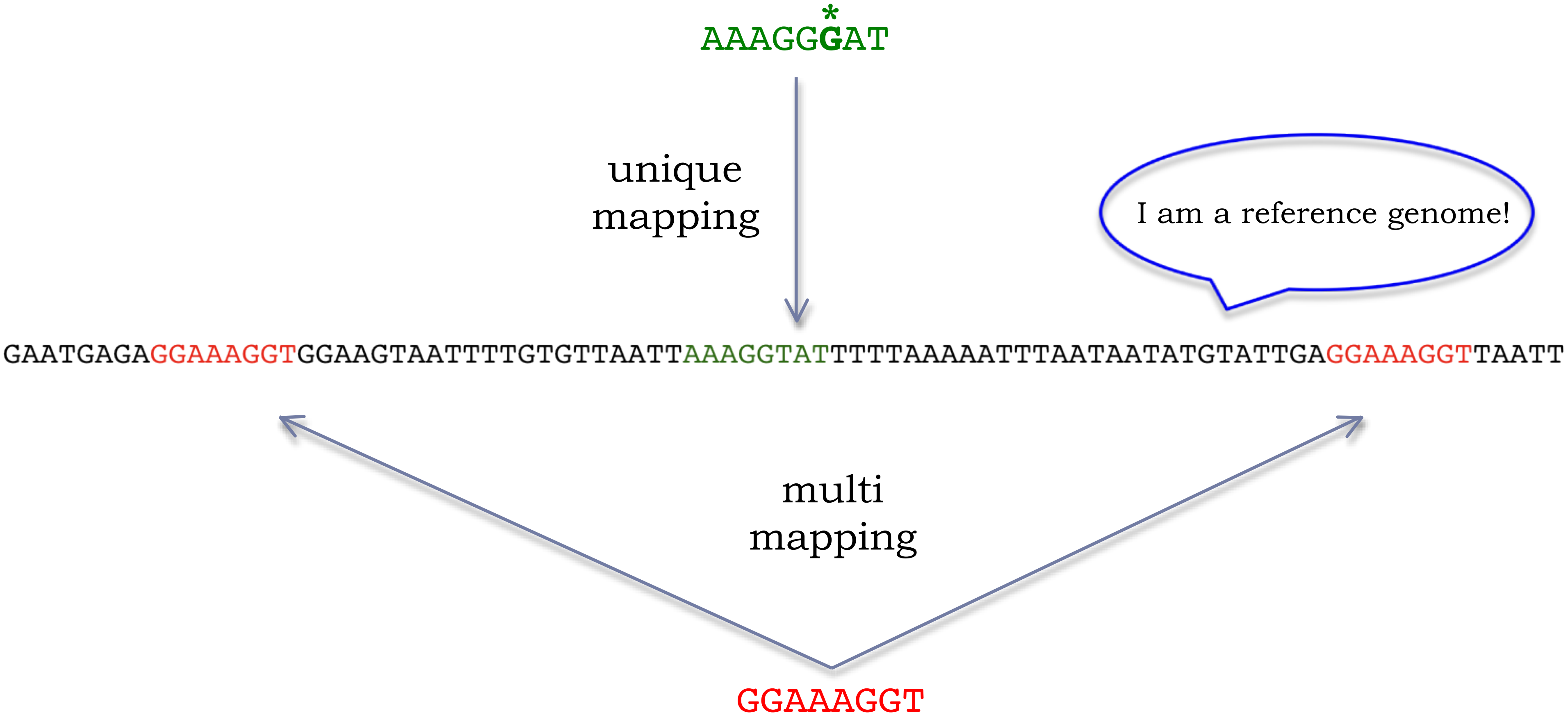
Quality control



NTGTTGAT**GACATGTATCCGGATGTTTGGAGTGGGACATGTATCCGGATGTG**TTTTGTTGACATGTATCCGGATGTT
NTTATGGTG**GACATGTATCCGGATGTTGTGAATTGGACATGTATCCGGATGT**CCTATAGTGGACATTATCCGGATGT
NAGAGAATGTGGACATGTA**TCCGG**ATGTGTTGGCTTGACACATATCCGGATGTTCCACACAGGACAAGATCGGAAGG
NTTATGCAGAAACGGACATGTA**TCCGG**ATGTGTCTTGTGTGACATGTATCCGGATGTCTGCAACGTCCGCTGAGCCA
NTTCATAGGTGCGGACATGTA**TCCGG**ATGTGTCGCAATGTGATATGATCCGGATGTCGAGGCTCTACATCCGGATAC
NTGATGTGGGGTGACATGTA**TCCGG**ATGGATTTGGGGACATGTATCCGGATGCATCCGGATACATGTCATGTCACAG
NTAATGCTGGACATGTA**TCCGG**ATGTCTTTCTGTGGGGCAGCCACTGGCTCATCGAAAGATCGGAAGAGCTCGTATG
NATGGTACGGACATGTA**TCCGG**ATGTCATCCGGATACATGTCCAGGACGACAGATCGGAAGAGCTCGTATGCCGTCT
NTGGTGCTGGACATGTA**TCCGG**ATGTATGTTCTCGATAAAACATCCGGATACATGACAACACCAGAGATCGGAAGAG
NTGGTGGTTGGGACATGTA**TCCGG**ATGTTTACAGTATCTGCAACGTCCGCTGAGCCAGTGGCTGCCCCACCAGTGTT
NACATCATGGACATGTA**TCCGG**ATGAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAATA

Some primer contamination!

Alignment



*
AAAGGGAT

unique mapping

I am a reference genome!

GAATGAGAGGAAAGGTGGAAGTAATTTTGTGTTAATTAAAGGTATTTTTAAAAATTTAATAATATGTATTGAGGAAAGGTTAATT

multi mapping

GGAAAGGT

Alignment

- uniqueness depends on which regions are enriched for the protein you are immuno-precipitating

```
H3K27ac:
# reads processed: 39’897’199
# reads with at least one reported alignment: 33’984’279 (85.18%)
# reads that failed to align: 2’568’674 (6.44%)
# reads with alignments suppressed due to -m: 3’344’246 (8.38%)
```

```
H3K9me3
# reads processed: 82’402’674
# reads with at least one reported alignment: 29’278’881 (35.53%)
# reads that failed to align: 11’211’423 (13.61%)
# reads with alignments suppressed due to -m: 41’912’370 (50.86%)
```

Alignment

- I use **Bowtie**

- no gaps
- no clipping

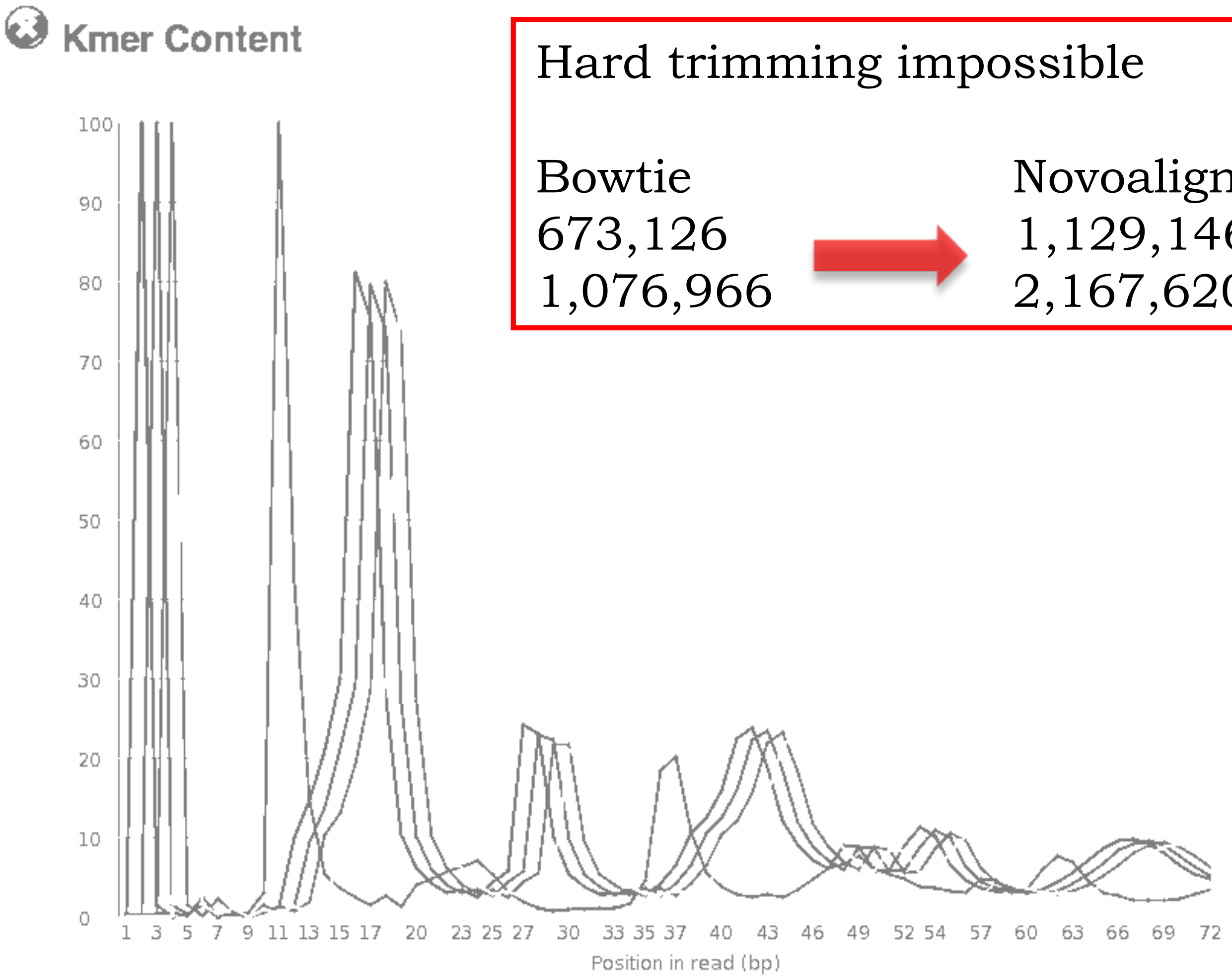
```
REF: AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGA
READ:       AAAGTGTCGCC-GACTAGCTCC
```

- ~35-40' for 25 millions reads using 4 cores

Alignment

	Gaps	Clipping
Bowtie	no	no
Bowtie2	yes	yes
Bwa	yes	yes
Novoalign	yes	yes

Alignment



Hard trimming impossible

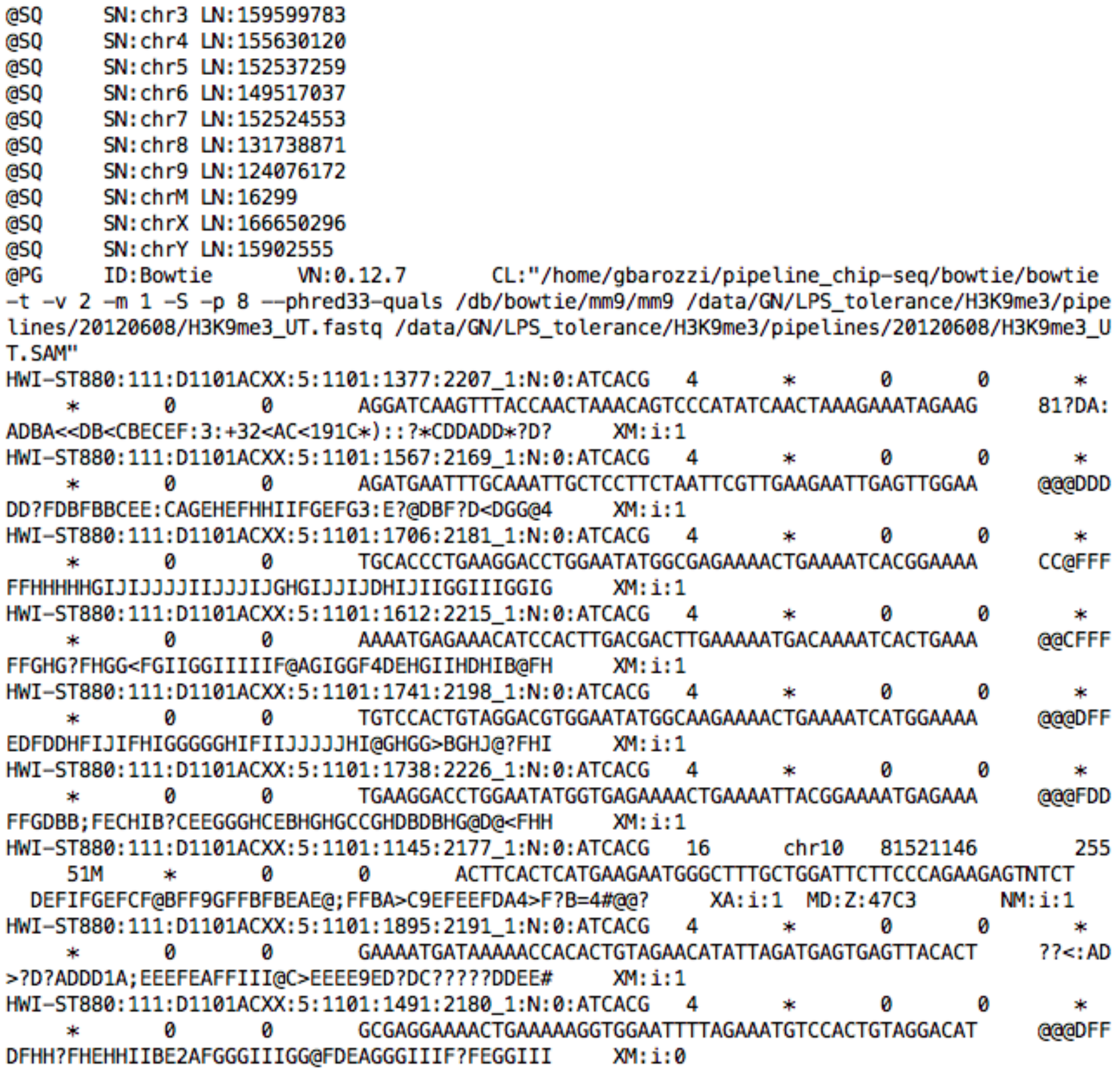
Bowtie	Novoalign
673,126	1,129,146
1,076,966	2,167,620

Alignment

- `bowtie -t -v 2 -m 1 -S -p 8 --phred33-quals /db/bowtie/mm9/mm9 sample.fastq sample.SAM`

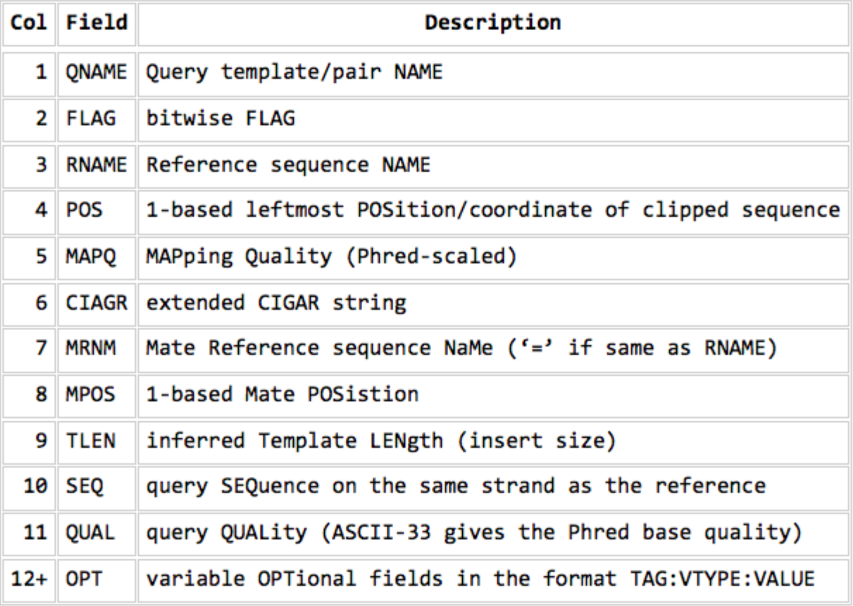
- `-m 1` uniquely alignable reads only
- `-v 2` up to two mismatches
- `-S` SAM output
- `-p 8` uses 8 CPUs
- `--phred33-quals` quality scores
- `/db/bowtie/mm9/mm9` the genomic index

SAM/BAM formats



```
@SQ     SN:chr3 LN:159599783
@SQ     SN:chr4 LN:155630120
@SQ     SN:chr5 LN:152537259
@SQ     SN:chr6 LN:149517037
@SQ     SN:chr7 LN:152524553
@SQ     SN:chr8 LN:131738871
@SQ     SN:chr9 LN:124076172
@SQ     SN:chrM LN:16299
@SQ     SN:chrX LN:166650296
@SQ     SN:chrY LN:15902555
@PG     ID:Bowtie       VN:0.12.7       CL:"/home/gbarozzi/pipeline_chip-seq/bowtie/bowtie
-t -v 2 -m 1 -S -p 8 --phred33-quals /db/bowtie/mm9/mm9 /data/GN/LPS_tolerance/H3K9me3/pipe
lines/20120608/H3K9me3_UT.fastq /data/GN/LPS_tolerance/H3K9me3/pipelines/20120608/H3K9me3_U
T.SAM"
HWI-ST880:111:D1101ACXX:5:1101:1377:2207_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       AGGATCAAGTTTACCAACTAAACAGTCCCATATCAACTAAAGAAATAGAAG     81?DA:
ADBA<<DB<CBECEF:3:+32<AC<191C*)::?*CDDADD*?D?     XM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1567:2169_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       AGATGAATTTGCAAATTGCTCCTTCTAATTCGTTGAAGAATTGAGTTGGAA     @@@DDD
DD?FDBFBBCEE:CAGEHEFHHIIFGEFG3:E?@DBF?D<DGG@4     XM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1706:2181_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       TGCACCCTGAAGGACCTGGAATATGGCGAGAAAACTGAAAATCACGGAAAA     CC@FFF
FFHHHHHGIJIJJJJIIJJJIJGHGIJJIJDHIJIIGGIIIGGIG     XM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1612:2215_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       AAAATGAGAAACATCCACTTGACGACTTGAAAAATGACAAAATCACTGAAA     @@CFFF
FFGHG?FHGG<FGIIGGIIIIIF@AGIGGF4DEHGIIHDHIB@FH     XM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1741:2198_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       TGTCCACTGTAGGACGTGGAATATGGCAAGAAAACTGAAAATCATGGAAAA     @@@DFF
EDFDDHFIJIFHIGGGGGHIFIIJJJJJHI@GHGG>BGHJ@?FHI     XM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1738:2226_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       TGAAGGACCTGGAATATGGTGAGAAAACTGAAAATTACGGAAAATGAGAAA     @@@FDD
FFGDBB;FECHIB?CEEGGGHCEBHGHGCCGHDBDBHG@D@<FHH     XM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1145:2177_1:N:0:ATCACG   16      chr10   81521146        255
     51M     *       0       0       ACTTCACTCATGAAGAATGGGCTTTGCTGGATTCTTCCCAGAAGAGTNTCT
  DEFIFGEFCF@BFF9GFFBFBEAE@;FFBA>C9EFEEFDA4>F?B=4#@@?     XA:i:1  MD:Z:47C3       NM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1895:2191_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       GAAAATGATAAAAACCACACTGTAGAACATATTAGATGAGTGAGTTACACT     ??<:AD
>?D?ADDD1A;EEEFEAFFIII@C>EEEE9ED?DC?????DDEE#     XM:i:1
HWI-ST880:111:D1101ACXX:5:1101:1491:2180_1:N:0:ATCACG   4       *       0       0       *
     *       0       0       GCGAGGAAAACTGAAAAAGGTGGAATTTTAGAAATGTCCACTGTAGGACAT     @@@DFF
DFHH?FHEHHIIBE2AFGGGIIIGG@FDEAGGGIIIF?FEGGIII     XM:i:0
```

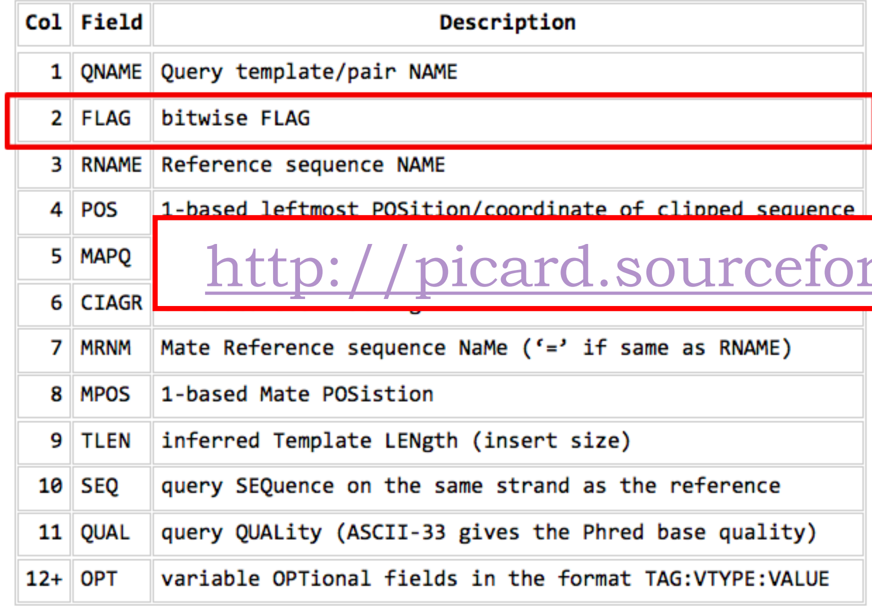
SAM/BAM formats



Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

```
HWI-ST880:111:D1101ACXX:5:1101:1145:2177_1:N:0:ATCACG
16
chr10
81521146
255
51M
*
0
0
ACTTCACTCATGAAGAATGGGCTTTGCTGGATTCTTCCCAGAAGAGTNTCT
DEFIFGEFCF@BFF9GFFBFBEAE@;FFBA>C9EFEEFDA4>F?B=4#@@?
XA:i:1
MD:Z:47C3
NM:i:1
```

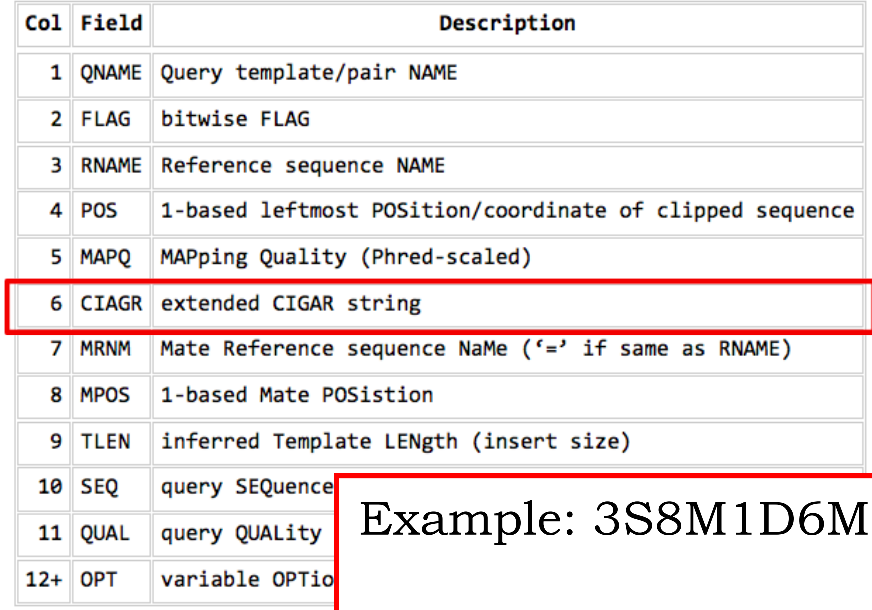
SAM/BAM formats



Col	Field	Description	
1	QNAME	Query template/pair NAME	HWI-ST880:111:D1101ACXX:5:1101:1145:2177_1:N:0:ATCACG
2	FLAG	bitwise FLAG	16
3	RNAME	Reference sequence NAME	chr10
4	POS	1-based leftmost POSition/coordinate of clipped sequence	81521146
5	MAPQ	[illegible]	[illegible]
6	CIAGR	[illegible]	[illegible]
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)	*
8	MPOS	1-based Mate POSistion	0
9	TLEN	inferred Template LENgth (insert size)	0
10	SEQ	query SEQuence on the same strand as the reference	ACTTCACTCATGAAGAATGGGCTTTGCTGGATTCTTCCCAGAAGAGTNTCT
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)	DEFIFGEFCF@BFF9GFFBFBEAE@;FFBA>C9EFEEFDA4>F?B=4#@@?
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE	XA:i:1 MD:Z:47C3 NM:i:1

http://picard.sourceforge.net/explain-flags.html

SAM/BAM formats



Col	Field	Description	
1	QNAME	Query template/pair NAME	HWI-ST880:111:D1101ACXX:5:1101:1145:2177_1:N:0:ATCACG
2	FLAG	bitwise FLAG	16
3	RNAME	Reference sequence NAME	chr10
4	POS	1-based leftmost POSition/coordinate of clipped sequence	81521146
5	MAPQ	MAPping Quality (Phred-scaled)	255
6	CIAGR	extended CIGAR string	**51M**
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)	*
8	MPOS	1-based Mate POSistion	0
9	TLEN	inferred Template LENgth (insert size)	0
10	SEQ	query SEQuence	CTTCCCAGAAGAGTNTCT
11	QUAL	query QUALity	EFEEFDA4>F?B=4#@@?
12+	OPT	variable OPTio	

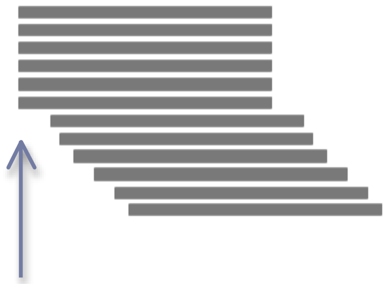
NM:i:1

Example: 3S8M1D6M4S

3 soft, 8 match, 1 deletion, 6 match and 4 soft

PCR duplicates

GAATGAGAGGAAAGGTGGAAGTAATTTTGTGTTAATTAAAGGTATTTTTAAAAATTTAATAATATGTATTGAGGAAAGGTTAATT



These reads are likely to have been generated by a non-random amplification process (PCR) rather than random fragmentation

(unless you have a very low genomic coverage or a very high sequencing depth)

Consider just one (or a number estimated using a statistics)

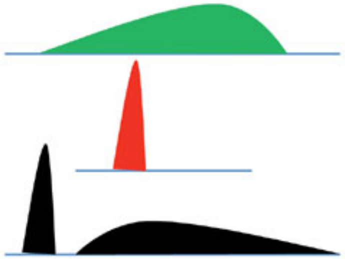
Data visualization

- UCSC genome browser: http://genome.ucsc.edu
- Very important to get acquainted with your data!

- http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Irosbarozzi&hgS_otherUserSessionName=Epigen_20121031

Peak calling

- I use MACS and RSEG



Peak calling (choose the right tool)

Type of peak	Example
Broad	H3K27me3
Sharp	CTCF
Sharp & broad	Pol II

Adapted from PMID: 21934668

Peak calling

- MACS is very good in:
  - finding SHARP signals in IP versus a control (e.g. TFs)
  - finding small but reproducible differences among IPs (e.g. changes in chromatin marks after challenging the cells with toxin or drugs that span down to less than 1 kbp, Nos2 example: chr11:78,683,143-78,694,401)

- RSEG is very good at:
  - finding BROAD signals in IP versus a control (e.g. H3K9me3)
  - finding domain-level chromatin changes between different IPs

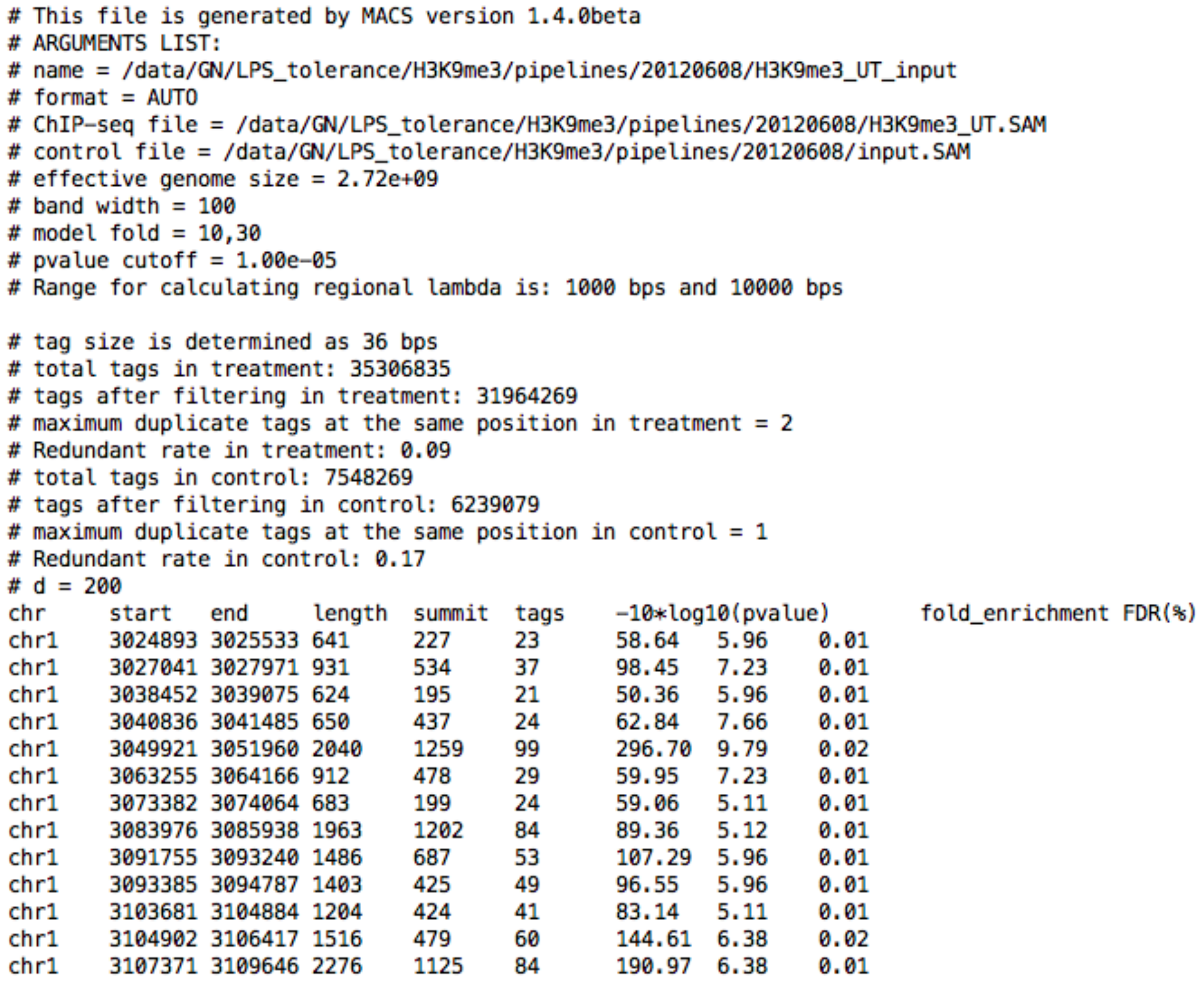
Peak calling: MACS

- `macs14 –t IP.SAM –c input.SAM --name=IP_vs_input --format=SAM --tsize=51 --gsize=2.72e9 --wig --single-wig --pvalue=1e-5 --format=SAM --space=10`

- P-value threshold: `--pvalue=1e-5`
- `--wig --single-wig`
- `--format=SAM`

- If you do not have any input/IgG: `--nolambda`

Peak calling: MACS



```
# This file is generated by MACS version 1.4.0beta
# ARGUMENTS LIST:
# name = /data/GN/LPS_tolerance/H3K9me3/pipelines/20120608/H3K9me3_UT_input
# format = AUTO
# ChIP-seq file = /data/GN/LPS_tolerance/H3K9me3/pipelines/20120608/H3K9me3_UT.SAM
# control file = /data/GN/LPS_tolerance/H3K9me3/pipelines/20120608/input.SAM
# effective genome size = 2.72e+09
# band width = 100
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Range for calculating regional lambda is: 1000 bps and 10000 bps

# tag size is determined as 36 bps
# total tags in treatment: 35306835
# tags after filtering in treatment: 31964269
# maximum duplicate tags at the same position in treatment = 2
# Redundant rate in treatment: 0.09
# total tags in control: 7548269
# tags after filtering in control: 6239079
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.17
# d = 200
chr     start   end     length  summit  tags    -10*log10(pvalue)       fold_enrichment FDR(%)
chr1    3024893 3025533 641     227     23      58.64   5.96    0.01
chr1    3027041 3027971 931     534     37      98.45   7.23    0.01
chr1    3038452 3039075 624     195     21      50.36   5.96    0.01
chr1    3040836 3041485 650     437     24      62.84   7.66    0.01
chr1    3049921 3051960 2040    1259    99      296.70  9.79    0.02
chr1    3063255 3064166 912     478     29      59.95   7.23    0.01
chr1    3073382 3074064 683     199     24      59.06   5.11    0.01
chr1    3083976 3085938 1963    1202    84      89.36   5.12    0.01
chr1    3091755 3093240 1486    687     53      107.29  5.96    0.01
chr1    3093385 3094787 1403    425     49      96.55   5.96    0.01
chr1    3103681 3104884 1204    424     41      83.14   5.11    0.01
chr1    3104902 3106417 1516    479     60      144.61  6.38    0.02
chr1    3107371 3109646 2276    1125    84      190.97  6.38    0.01
```

Peak calling: RSEG

- IP versus random expectation (no control):

```
rseg -c mouse-mm9-size.bed -o $PWD -i 20 -v -d deadzones-k36-
mm9.bed H2AK5ac_UT.tags.bed
```

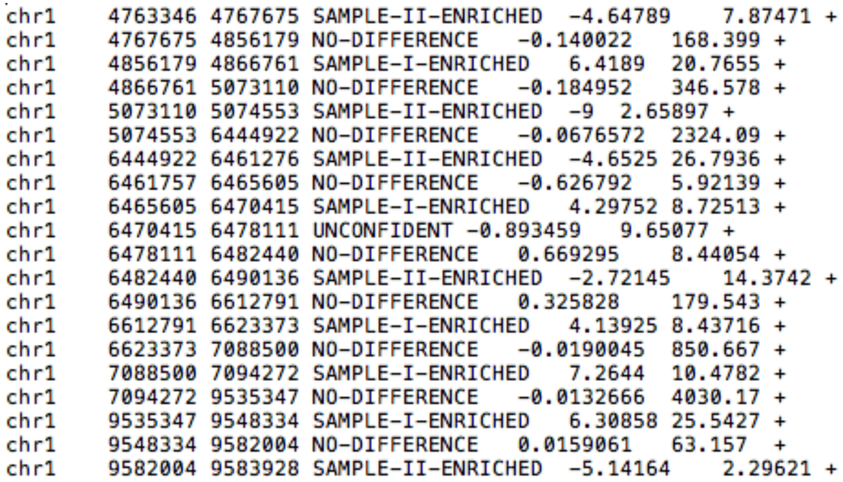
- IP versus control:

```
rseg-diff -c mouse-mm9-size.bed -o $PWD -i 20 -v -mode 2 -d
deadzones-k36-mm9.bed H2AK5ac_UT.tags.bed input.tags.bed
```

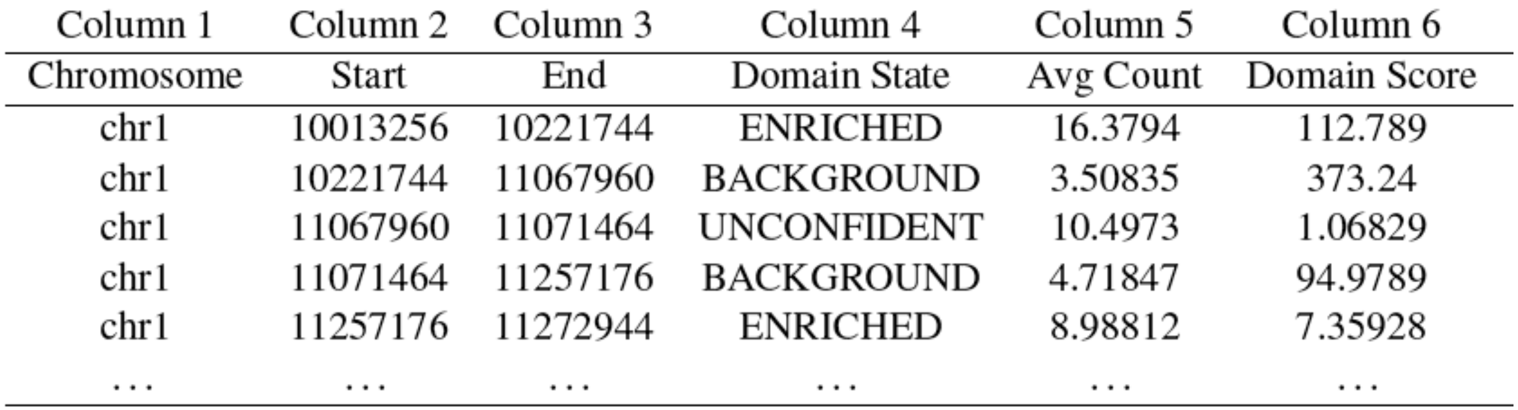
- IP versus IP:

```
rseg-diff -c mouse-mm9-size.bed -o $PWD -i 20 -v —mode 3 -d
deadzones-k36-mm9.bed H2AK5ac_LPS_2h.tags.bed
H2AK5ac_UT.tags.bed
```

Peak calling: RSEG

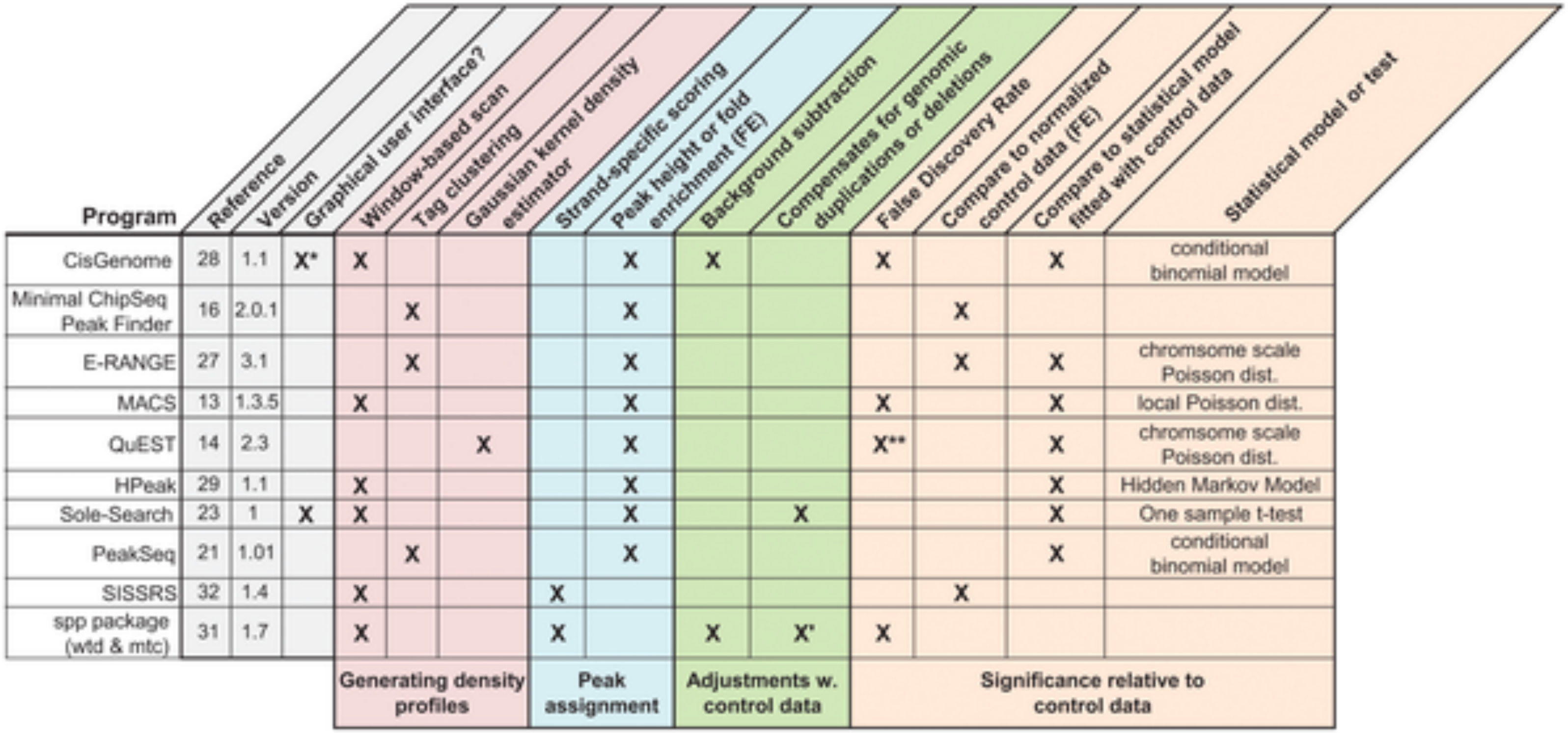


```
chr1    4763346 4767675 SAMPLE-II-ENRICHED  -4.64789    7.87471 +
chr1    4767675 4856179 NO-DIFFERENCE   -0.140022   168.399 +
chr1    4856179 4866761 SAMPLE-I-ENRICHED   6.4189  20.7655 +
chr1    4866761 5073110 NO-DIFFERENCE   -0.184952   346.578 +
chr1    5073110 5074553 SAMPLE-II-ENRICHED  -9  2.65897 +
chr1    5074553 6444922 NO-DIFFERENCE   -0.0676572  2324.09 +
chr1    6444922 6461276 SAMPLE-II-ENRICHED  -4.6525 26.7936 +
chr1    6461757 6465605 NO-DIFFERENCE   -0.626792   5.92139 +
chr1    6465605 6470415 SAMPLE-I-ENRICHED   4.29752 8.72513 +
chr1    6470415 6478111 UNCONFIDENT -0.893459   9.65077 +
chr1    6478111 6482440 NO-DIFFERENCE   0.669295    8.44054 +
chr1    6482440 6490136 SAMPLE-II-ENRICHED  -2.72145    14.3742 +
chr1    6490136 6612791 NO-DIFFERENCE   0.325828    179.543 +
chr1    6612791 6623373 SAMPLE-I-ENRICHED   4.13925 8.43716 +
chr1    6623373 7088500 NO-DIFFERENCE   -0.0190045  850.667 +
chr1    7088500 7094272 SAMPLE-I-ENRICHED   7.2644  10.4782 +
chr1    7094272 9535347 NO-DIFFERENCE   -0.0132666  4030.17 +
chr1    9535347 9548334 SAMPLE-I-ENRICHED   6.30858 25.5427 +
chr1    9548334 9582004 NO-DIFFERENCE   0.0159061   63.157  +
chr1    9582004 9583928 SAMPLE-II-ENRICHED  -5.14164    2.29621 +
```



Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Chromosome	Start	End	Domain State	Avg Count	Domain Score
chr1	10013256	10221744	ENRICHED	16.3794	112.789
chr1	10221744	11067960	BACKGROUND	3.50835	373.24
chr1	11067960	11071464	UNCONFIDENT	10.4973	1.06829
chr1	11071464	11257176	BACKGROUND	4.71847	94.9789
chr1	11257176	11272944	ENRICHED	8.98812	7.35928
...	...	...	...	...	...

Peak calling



				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			
Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			

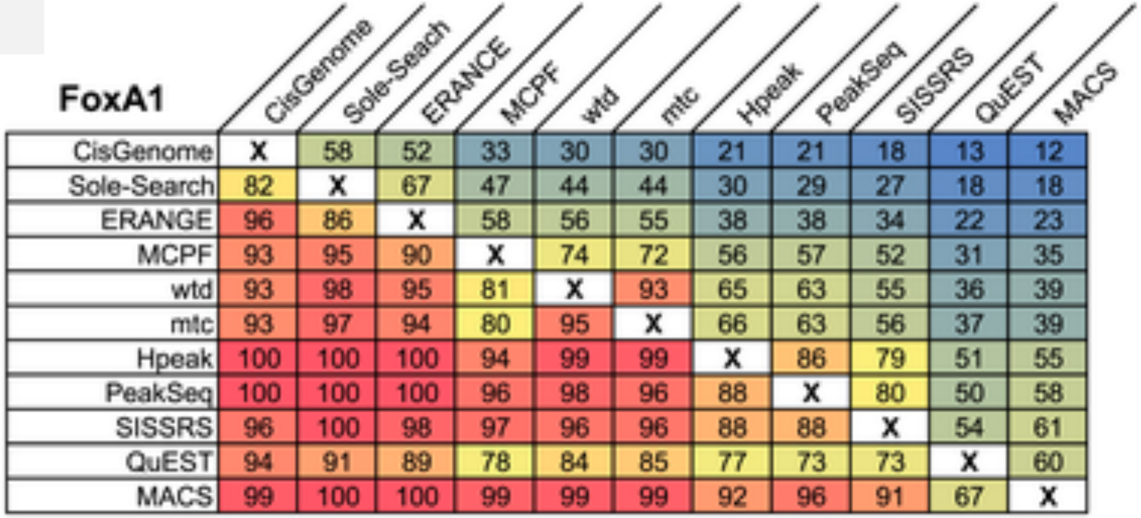
X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

PMID: 20628599

Peak calling



FoxA1	CisGenome	Sole-Seach	ERANCE	MCPF	wtd	mtc	Hpeak	PeakSeq	SISSRS	QuEST	MACS
CisGenome	X	58	52	33	30	30	21	21	18	13	12
Sole-Search	82	X	67	47	44	44	30	29	27	18	18
ERANGE	96	86	X	58	56	55	38	38	34	22	23
MCPF	93	95	90	X	74	72	56	57	52	31	35
wtd	93	98	95	81	X	93	65	63	55	36	39
mtc	93	97	94	80	95	X	66	63	56	37	39
Hpeak	100	100	100	94	99	99	X	86	79	51	55
PeakSeq	100	100	100	96	98	96	88	X	80	50	58
SISSRS	96	100	98	97	96	96	88	88	X	54	61
QuEST	94	91	89	78	84	85	77	73	73	X	60
MACS	99	100	100	99	99	99	92	96	91	67	X

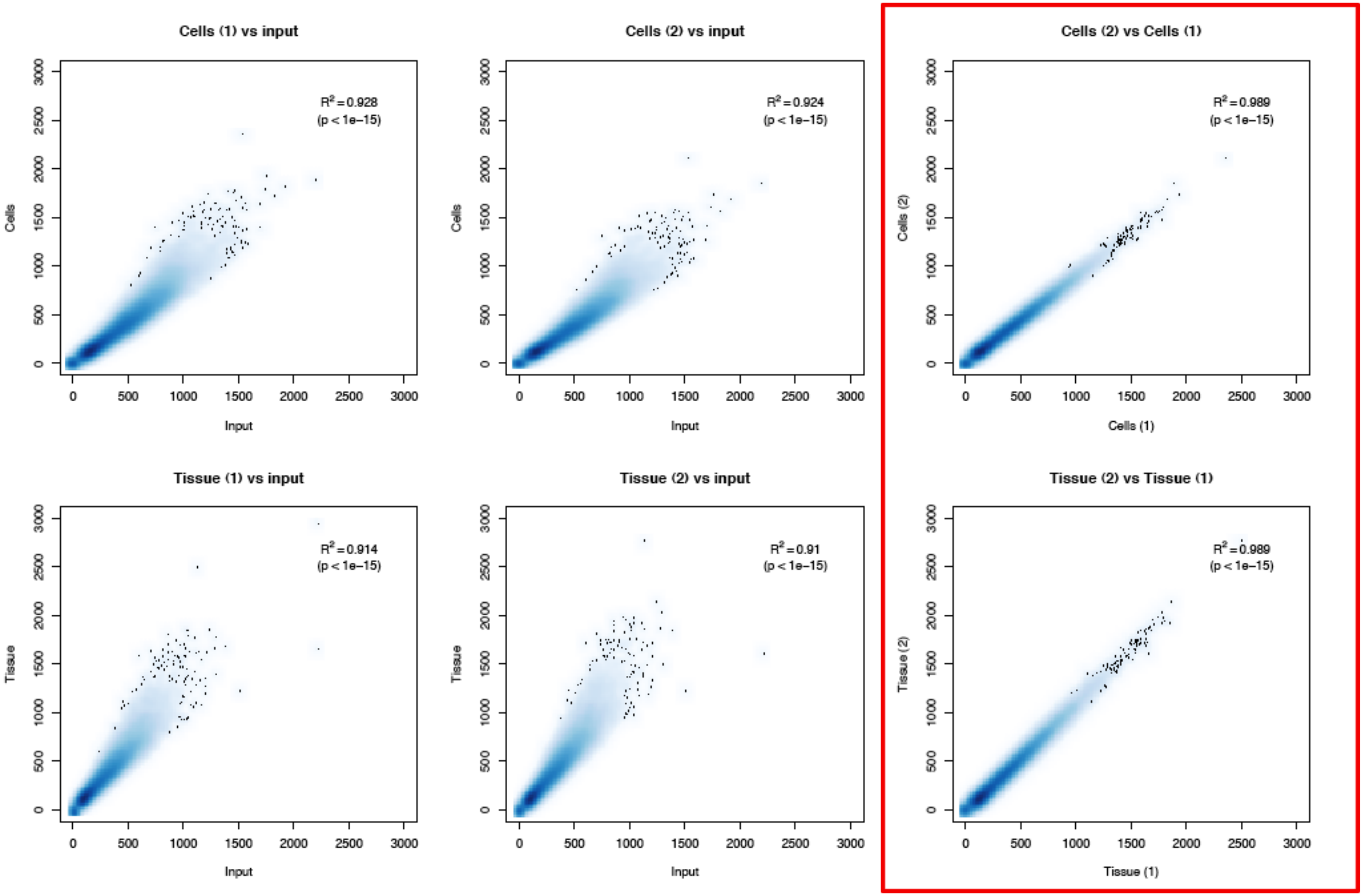
PMID: 20628599

Normalization

- Can we compare datasets with a different sequencing depth?

- How do we normalize on sequencing depth?

Assumption: linearity



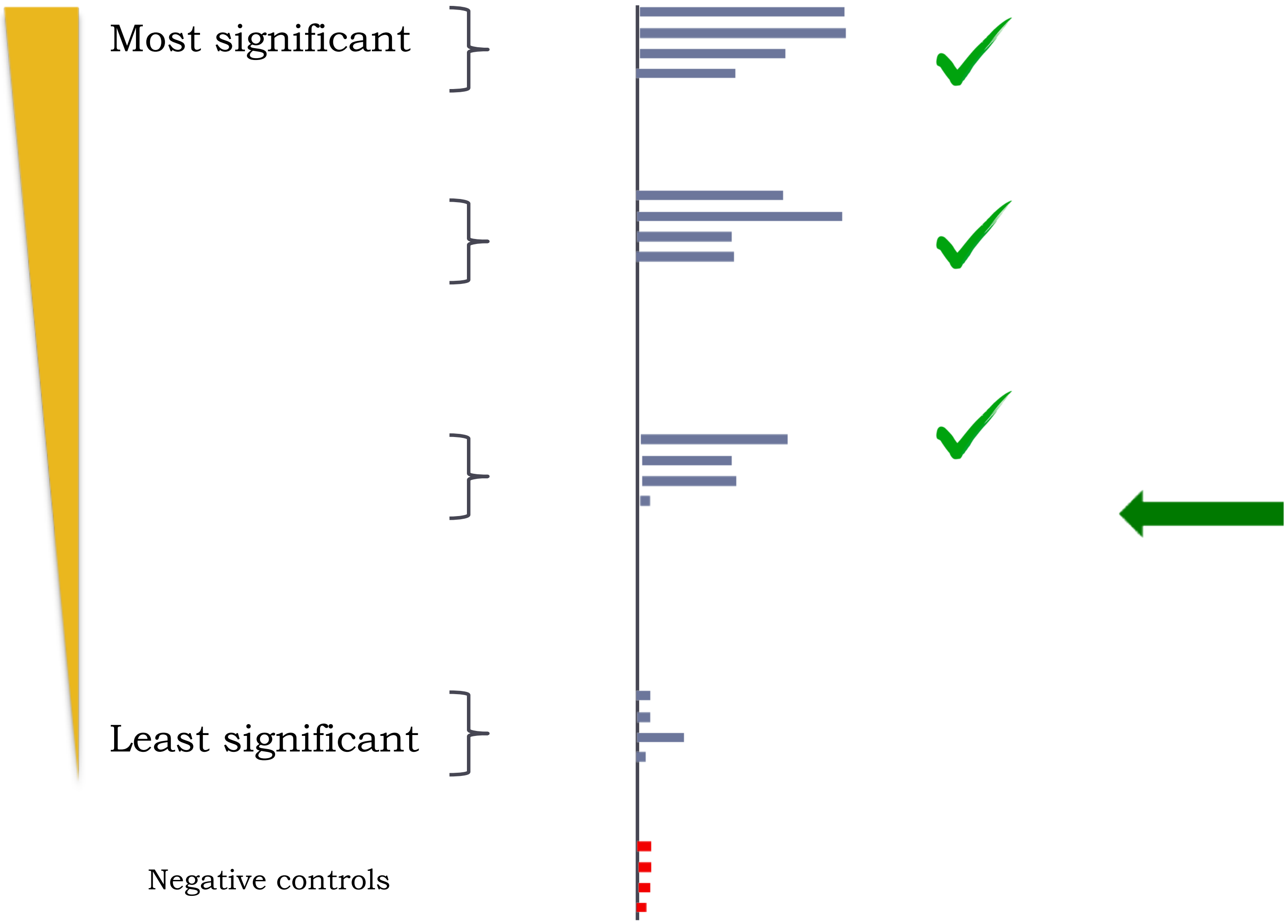
Summary

- Quality control
- Alignment to the reference genome
- Dealing with PCR duplicates
- Data visualization
- Peak calling

Summary

```
H3K27ac:
# reads after quality filtering:               39’897’199
# reads with a unique alignment on the genome: 33’984’279 (85.18%)
# reads after PCR duplicates:                  31’069’231 (77.87%)


H3K9me3
# reads after quality filtering:               82’402’674
# reads with a unique alignment on the genome: 29’278’881 (35.53%)
# reads after PCR duplicates:                  26’273’699 (31.88%)


H3K4me3 (with strong PCR bias)
# reads after quality filtering:               28’681’583
# reads with a unique alignment on the genome: 16’928’963 (59.02%)
# reads after PCR duplicates:                  2’715’994  (9.47%)
```

Experimental (qPCR) validation



Most significant

Least significant

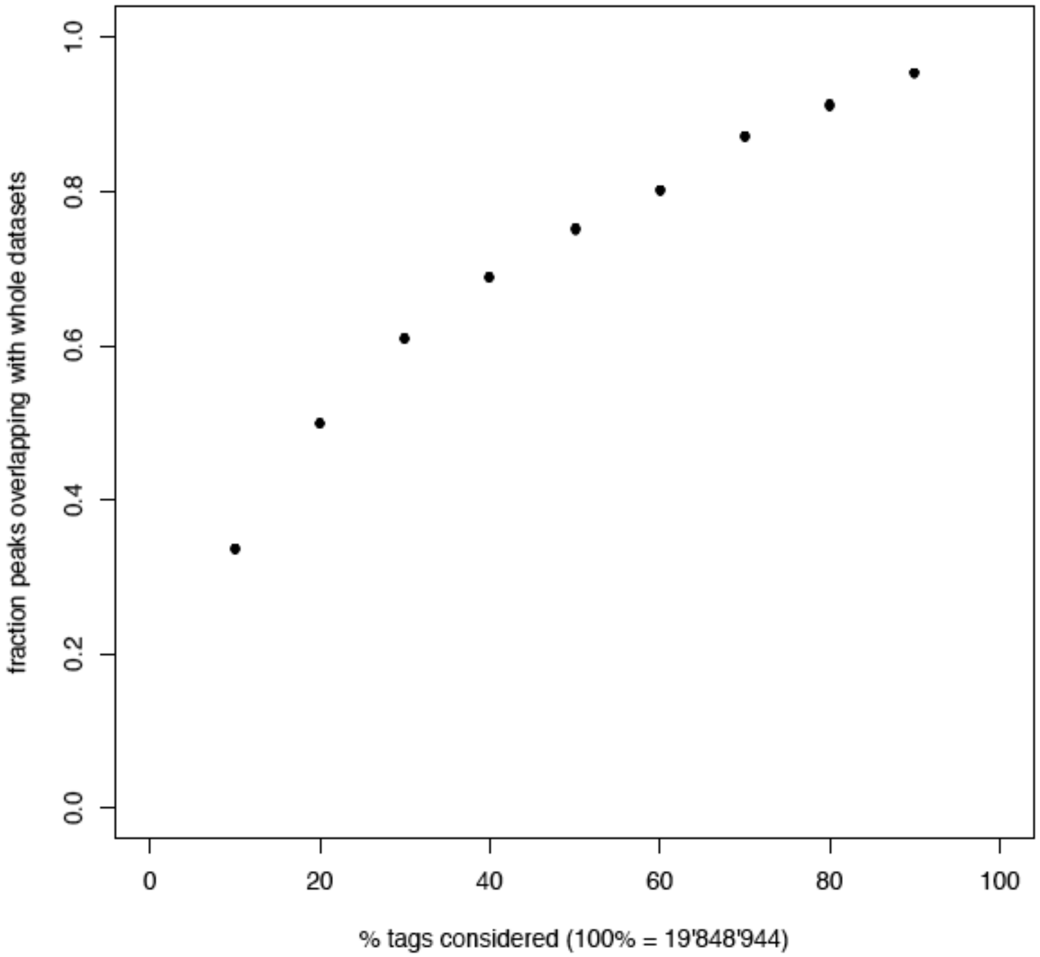
Negative controls

FAQs: how deep?

- What is the efficiency of your antibody (SNR)?
- What is the fraction of the genome potentially covered by the protein of interest?

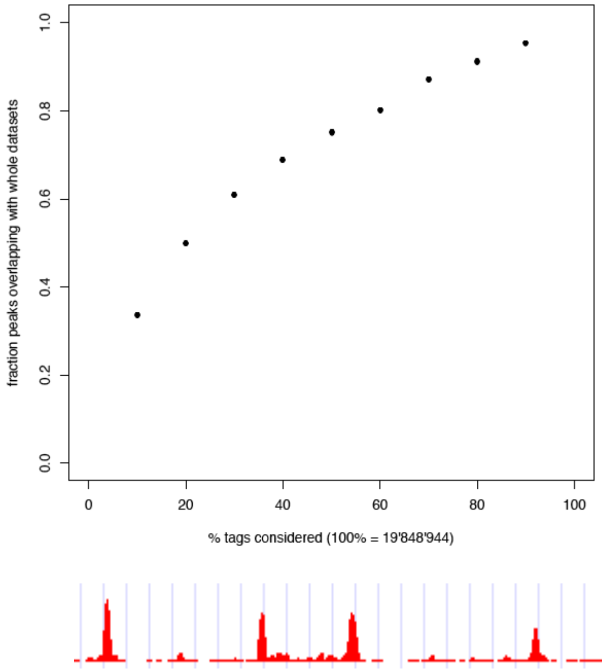
- Saturation plots

FAQs: how deep?



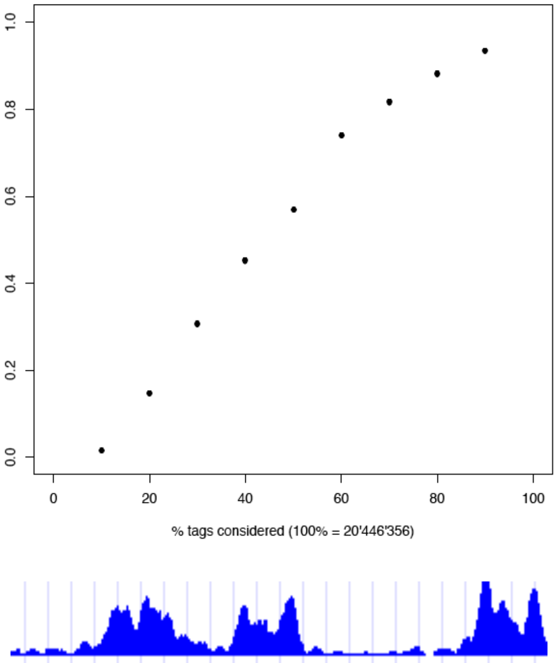
FAQs: how deep?

PU.1 (TF)



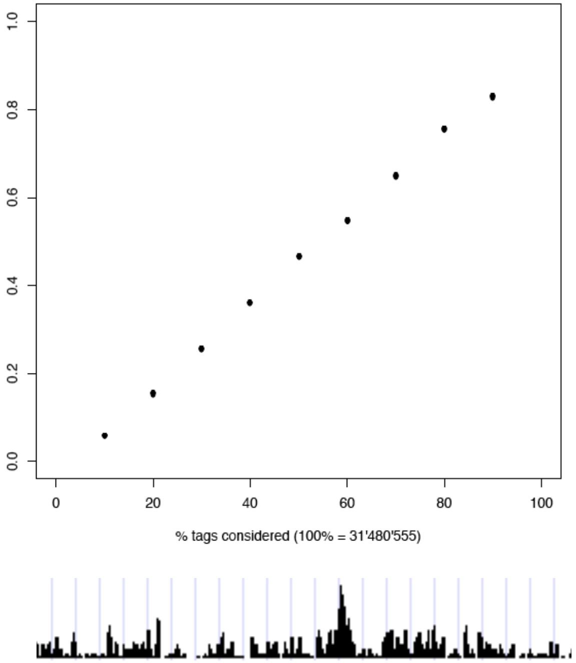
~50 Mbp

H3K4me1



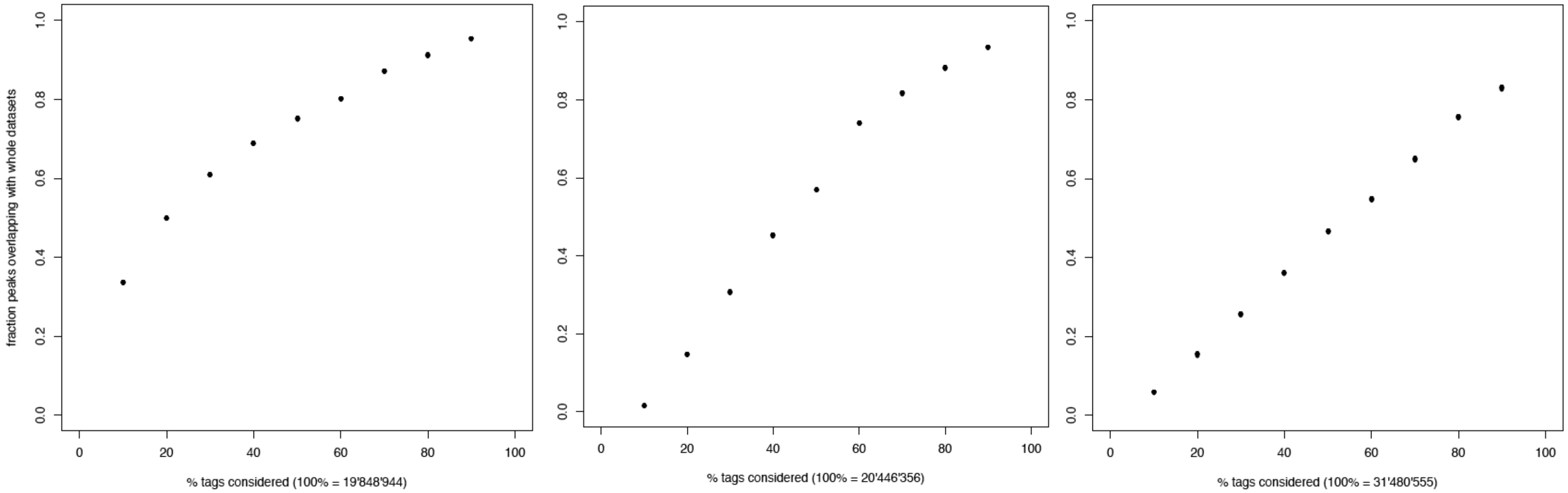
~140 Mbp

H3K9me3



~150 Mbp
(>250 Mbp)

FAQs: how deep?



PU.1 (TF)

100% ~ 19M
(PCR bias < 15%)
(30M raw reads)

H3K4me1

100% ~ 20M
(PCR bias < 15%)
(30M raw reads)

H3K9me3

100% ~ 30M
(PCR bias < 15%)
(60M raw reads)

- You always have the opportunity to sequence your library again and increase the depth!

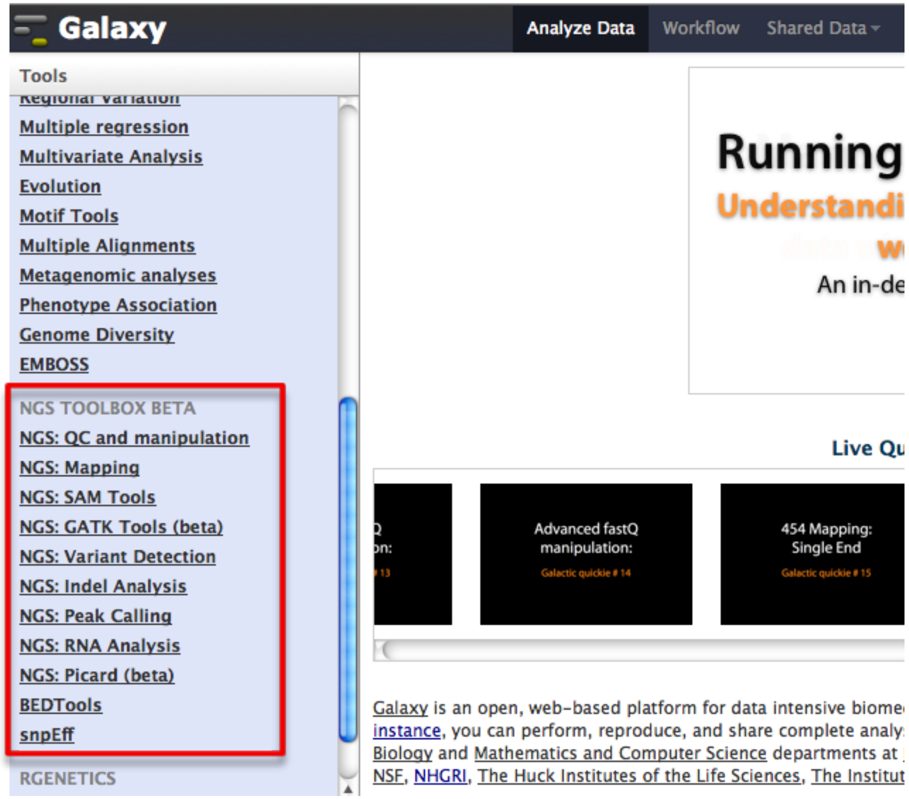
FAQs: what about the control?

- Which is the best control?
  Ideally the best control is the IP performed in the same cells in which the protein is not expressed. This is rarely feasible.
  Input of the IP and IgG are equally good control.

- What if no experimental control is available?
  Don’t worry you can run your analysis without estimating local biases. In most cases artifacts are a small fraction on the total number of enriched regions and won’t dramatically affect the results.

Galaxy

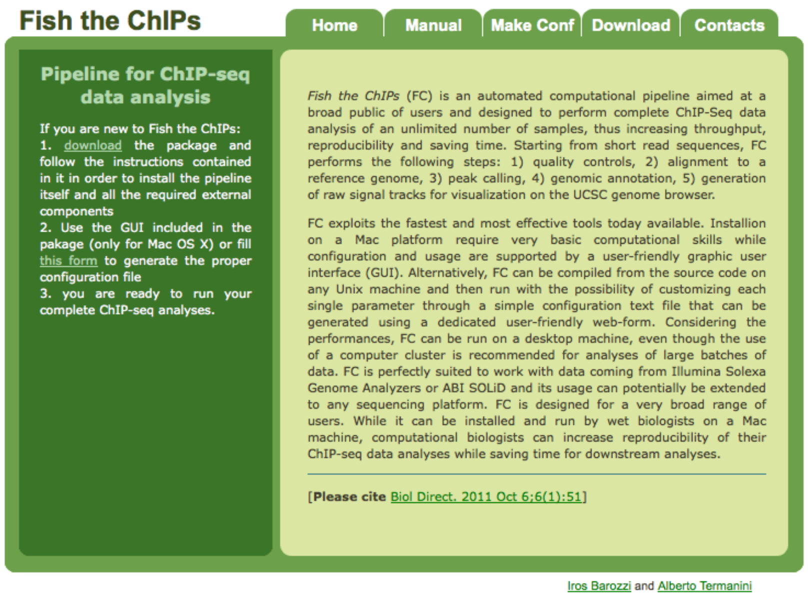
- https://main.g2.bx.psu.edu/



Galaxy is an open, web-based platform for data intensive biome
instance, you can perform, reproduce, and share complete analy
Biology and Mathematics and Computer Science departments at
NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institut

Fish the ChIPs

- A Mac GUI for ChIP-seq analysis
- http://bio.ifom-ieo-campus.it/ftc/



Pipeline for ChIP-seq data analysis

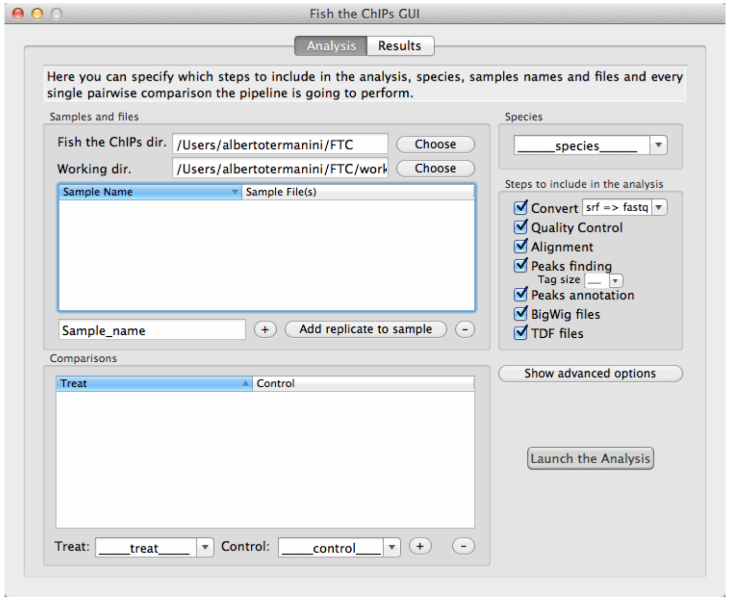
If you are new to Fish the ChIPs:
1. download the package and follow the instructions contained in it in order to install the pipeline itself and all the required external components
2. Use the GUI included in the pakage (only for Mac OS X) or fill this form to generate the proper configuration file
3. you are ready to run your complete ChIP-seq analyses.

Fish the ChIPs (FC) is an automated computational pipeline aimed at a broad public of users and designed to perform complete ChIP-Seq data analysis of an unlimited number of samples, thus increasing throughput, reproducibility and saving time. Starting from short read sequences, FC performs the following steps: 1) quality controls, 2) alignment to a reference genome, 3) peak calling, 4) genomic annotation, 5) generation of raw signal tracks for visualization on the UCSC genome browser.

FC exploits the fastest and most effective tools today available. Installion on a Mac platform require very basic computational skills while configuration and usage are supported by a user-friendly graphic user interface (GUI). Alternatively, FC can be compiled from the source code on any Unix machine and then run with the possibility of customizing each single parameter through a simple configuration text file that can be generated using a dedicated user-friendly web-form. Considering the performances, FC can be run on a desktop machine, even though the use of a computer cluster is recommended for analyses of large batches of data. FC is perfectly suited to work with data coming from Illumina Solexa Genome Analyzers or ABI SOLiD and its usage can potentially be extended to any sequencing platform. FC is designed for a very broad range of users. While it can be installed and run by wet biologists on a Mac machine, computational biologists can increase reproducibility of their ChIP-seq data analyses while saving time for downstream analyses.

[**Please cite** Biol Direct. 2011 Oct 6;6(1):51]

Iros Barozzi and Alberto Termanini



Here you can specify which steps to include in the analysis, species, samples names and files and every single pairwise comparison the pipeline is going to perform.

Samples and files

Fish the ChIPs dir. /Users/albertotermanini/FTC

Working dir. /Users/albertotermanini/FTC/work

Sample Name	Sample File(s)

Sample_name

Species

____species____

Steps to include in the analysis

- ☑ Convert srf => fastq
- ☑ Quality Control
- ☑ Alignment
- ☑ Peaks finding
  - Tag size __
- ☑ Peaks annotation
- ☑ BigWig files
- ☑ TDF files

Comparisons

Treat	Control

Treat: ____treat____ Control: ____control___

UCSC session

- UCSC session at: http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Irosbarozzi&hgS_otherUserSessionName=Epigen_20121031

- Murine macrophages
- Untreated and LPS 4h
- Pu.1, H3K4me1, H3K4me3

- Peaks coordinates in BED for Pu.1 UT can be downloaded from: http://www.zeroidee.org/iros/bws_rome/Pu.1_UT/

Supplementary: manipulating files

- Samtools
  http://samtools.sourceforge.net/samtools.shtml

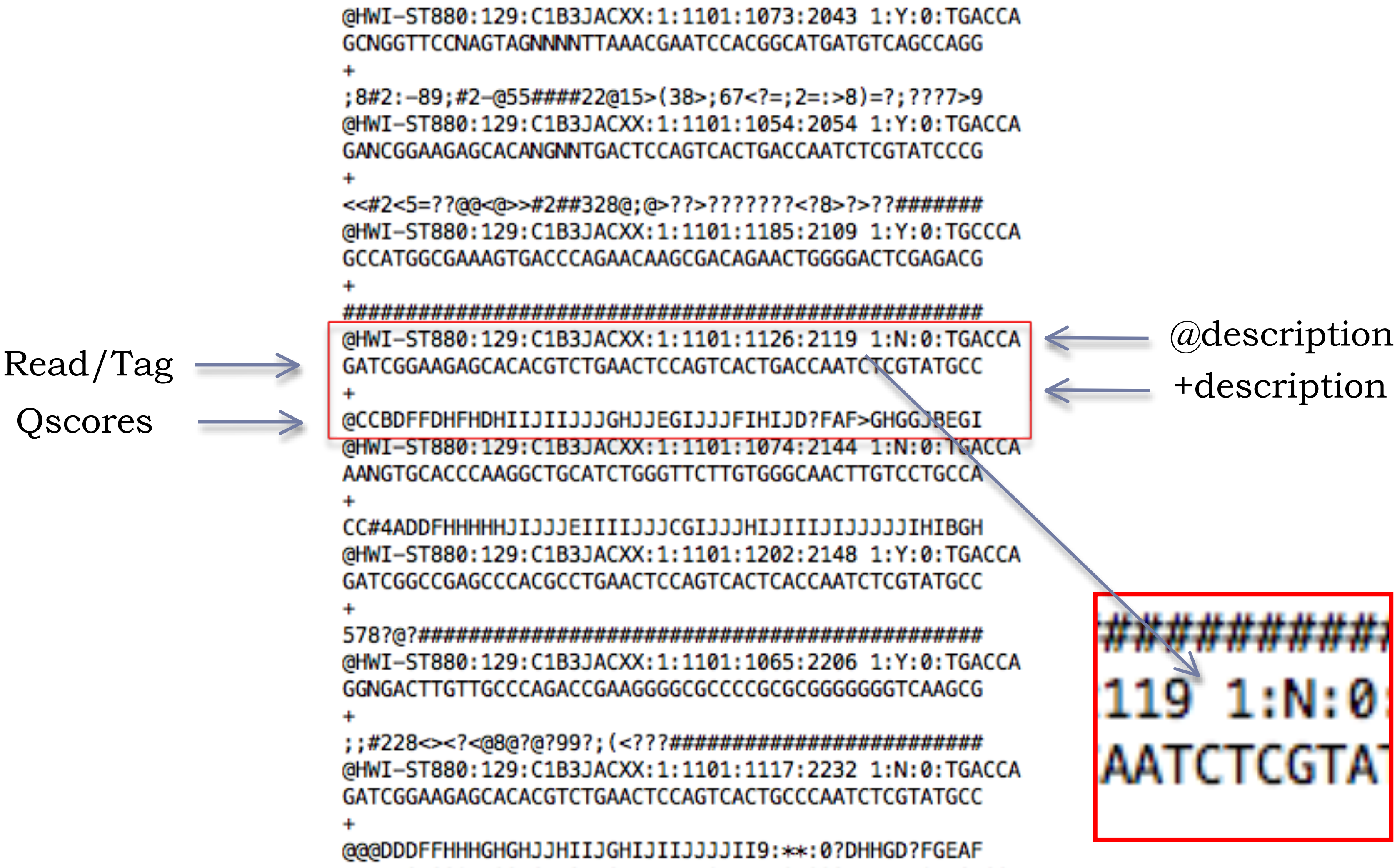
- Bedtools
  http://code.google.com/p/bedtools/

- Picard
  http://picard.sourceforge.net/

Supplementary: useful literature

- Nature Methods 6, S22 - S32 (2009)
  Computation for ChIP-seq and RNA-seq studies
  Shirley Pepke, Barbara Wold & Ali Mortazavi
- Nature Reviews Genetics 10, 669-680 (October 2009)
  ChIP–seq: advantages and challenges of a maturing technology
  Park PJ
- Nat Immunol. 2011 Sep 20;12(10):918-22. doi: 10.1038/ni.2117.
  ChIP-Seq: technical considerations for obtaining high-quality data.
  Kidder BL, Hu G, Zhao K.
- PLoS One. 2010 Jul 8;5(7):e11471.
  Evaluation of algorithm performance in ChIP-seq peak detection.
  Wilbanks EG, Facciotti MT.

Supplementary: FASTQ format - replace the spaces



```
@HWI-ST880:129:C1B3JACXX:1:1101:1073:2043 1:Y:0:TGACCA
GCNGGTTCCNAGTAGNNNNTTAAACGAATCCACGGCATGATGTCAGCCAGG
+
;8#2:-89;#2-@55####22@15>(38>;67<?=;2=:>8)=?;????7>9
@HWI-ST880:129:C1B3JACXX:1:1101:1054:2054 1:Y:0:TGACCA
GANCGGAAGAGCACANGNNTGACTCCAGTCACTGACCAATCTCGTATCCCG
+
<<#2<5=??@@<@>>#2##328@;@>??>???????<?8>?>??#######
@HWI-ST880:129:C1B3JACXX:1:1101:1185:2109 1:Y:0:TGCCCA
GCCATGGCGAAAGTGACCCAGAACAAGCGACAGAACTGGGGACTCGAGACG
+
##################################################
@HWI-ST880:129:C1B3JACXX:1:1101:1126:2119 1:N:0:TGACCA
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCC
+
@CCBDFFDHFHDHIIJIIJJJGHJJEGIJJJFIHIJD?FAF>GHGGJBEGI
@HWI-ST880:129:C1B3JACXX:1:1101:1074:2144 1:N:0:TGACCA
AANGTGCACCCAAGGCTGCATCTGGGTTCTTGTGGGCAACTTGTCCTGCCA
+
CC#4ADDFHHHHHJIJJJEIIIIJJJCGIJJJHIJIIIJIJJJJJIHIBGH
@HWI-ST880:129:C1B3JACXX:1:1101:1202:2148 1:Y:0:TGACCA
GATCGGCCGAGCCCACGCCTGAACTCCAGTCACTCACCAATCTCGTATGCC
+
578?@?############################################
@HWI-ST880:129:C1B3JACXX:1:1101:1065:2206 1:Y:0:TGACCA
GGNGACTTGTTGCCCAGACCGAAGGGGCGCCCCGCGCGGGGGGTCAAGCG
+
;;#228<><?<@8@?@?99?;(<???########################
@HWI-ST880:129:C1B3JACXX:1:1101:1117:2232 1:N:0:TGACCA
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGCCCAATCTCGTATGCC
+
@@@DDDFFHHHGHGHJJHIIJGHIJIIJJJJII9:**:0?DHHGD?FGEAF
```

Read/Tag

Qscores

@description

+description