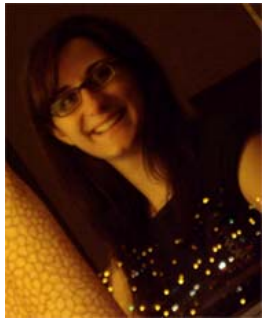
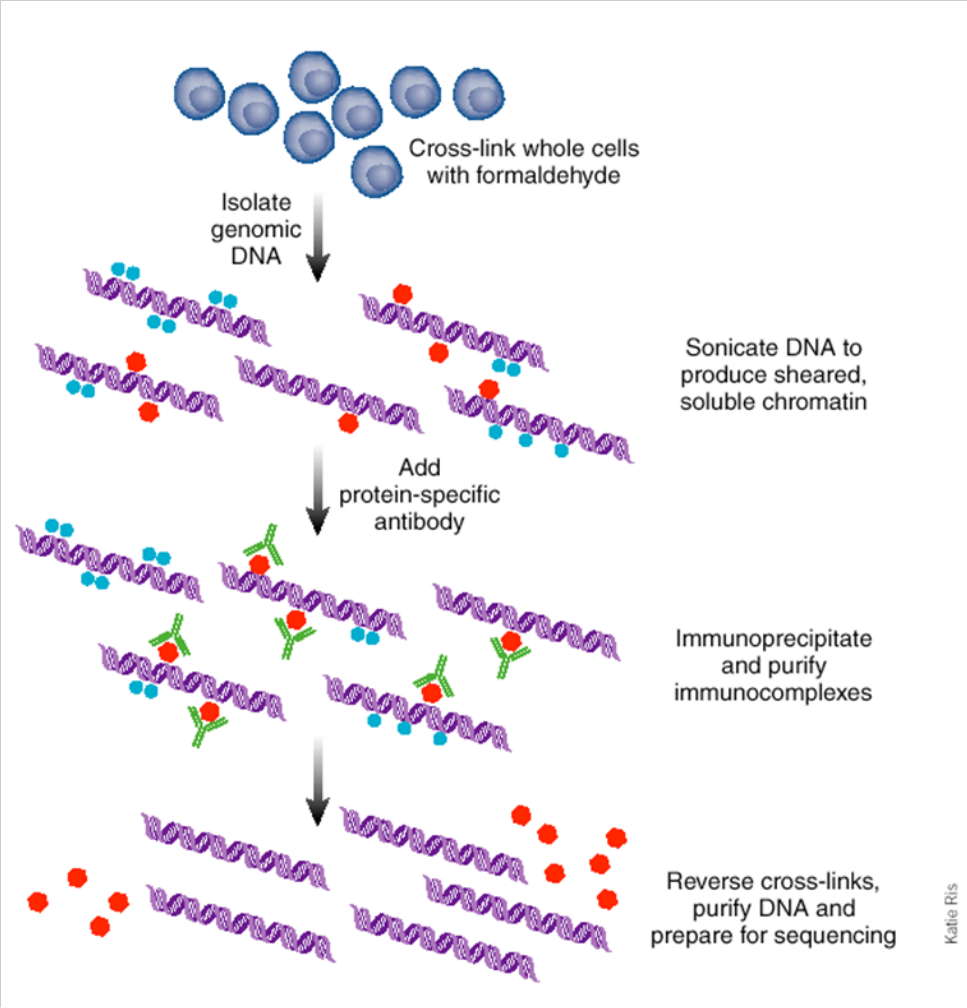
Integrative modeling of transcription factor binding and histone marks predicts gene expression in lymphoma

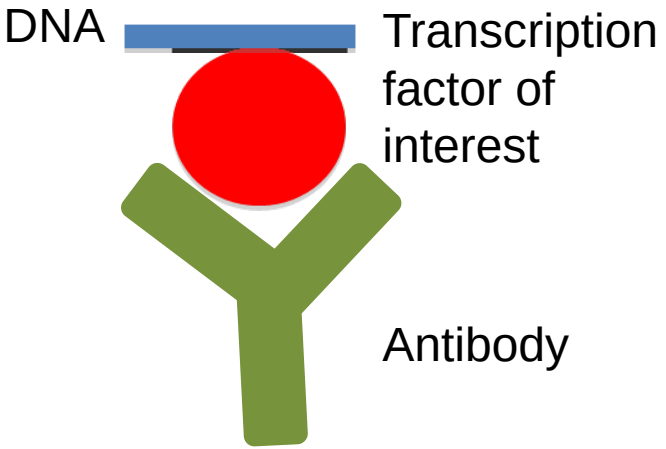


Jenny Giannopulou

Yanwen Jiang

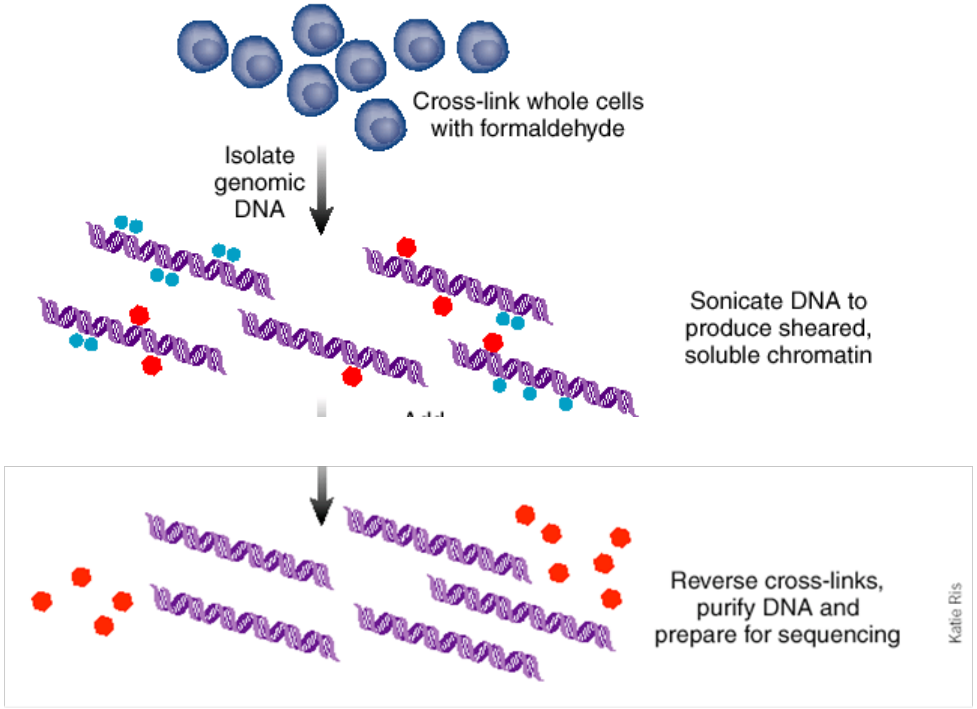
ChIP-seq to map where transcription factors bind



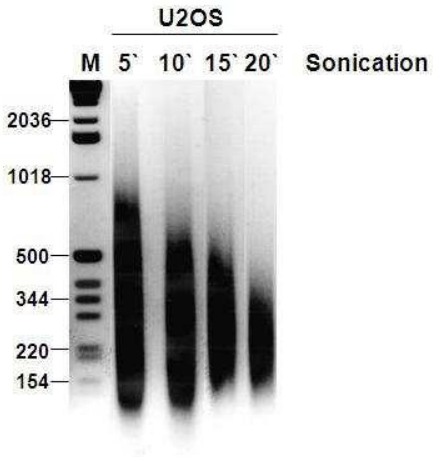


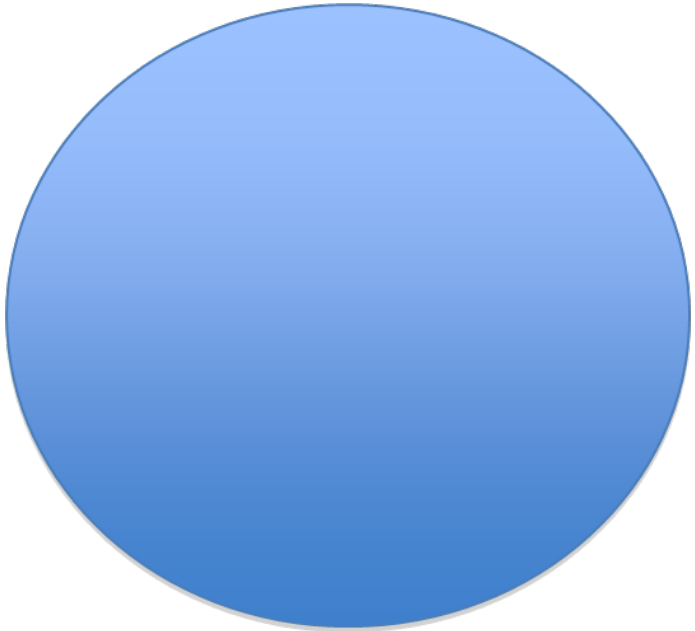
Genome Analyzer II (Illumina)

Control: input DNA



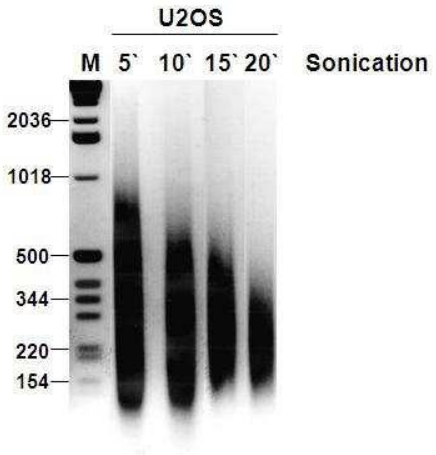
Genome Analyzer II (Illumina)

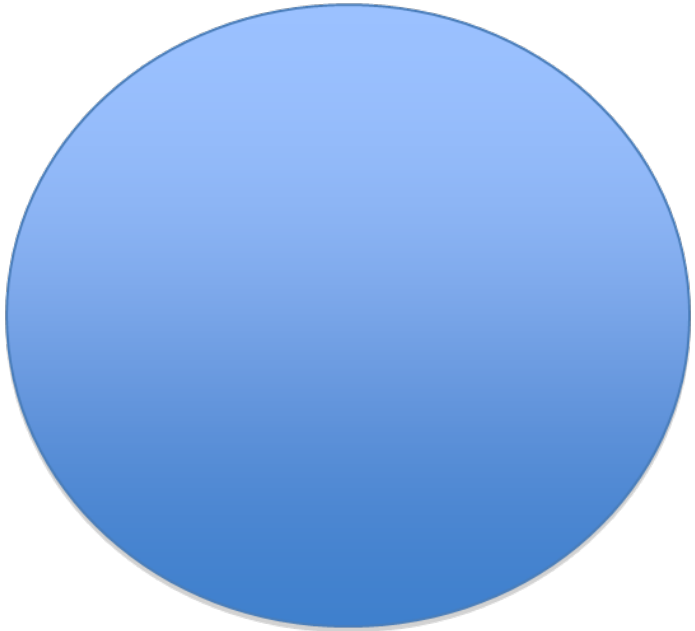




ACCAATAACCGAGGCTCATGCTAAGGCGTTAGCCACAGATGGAAGTCCGACGGCTTGATCCAGAATGGTGTGTGGATTGCCTTGGAACTGATTAGTGAATTC
TGGTTATTGGCTCCGAGTACGATTCCGCAATCGGTGTCTACCTTCAGGCTGCCGAACTAGGTCTTACCACACACCTAACGGAACCTTGACTAATCACTTAAG

Average length ~ 250bp

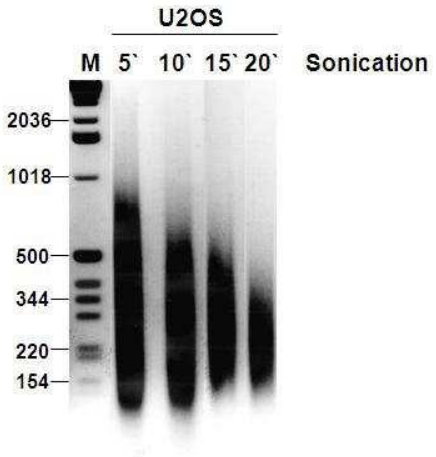


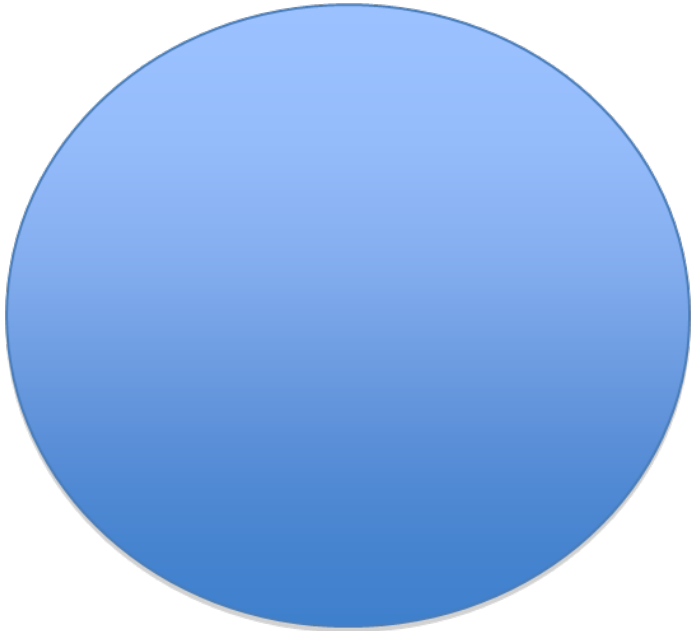


25-40bp

ACCAATAACCGAGGCTCATGCTAAGGCGTTAGCCACAGATGGAAGTCCGACGGCTTGATCCAGAATGGTGTGTGGATTGCCTTGGAACTGATTAGTGAATTC
TGGTTATTGGCTCCGAGTACGATTCCGCAATCGGTGTCTACCTTCAGGCTGCCGAACTAGGTCTTACCACACACCTAACGGAACCTTGACTAATCACTTAAG

Average length ~ 250bp





25-40bp

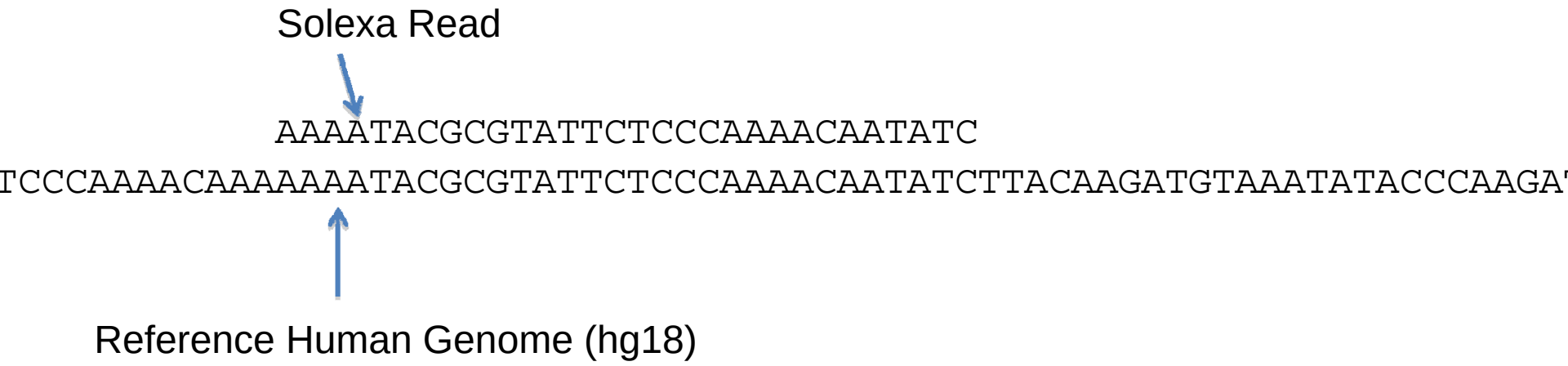
ACCAATAACCGAGGCTCATGCTAAGGCGTTAGCCACAGATGGAAGTCCGACGGCTTGATCCAGAATGGTGTGTGGATTGCCTTGGAACTGATTAGTGAATTC
TGGTTATTGGCTCCGAGTACGATTCCGCAATCGGTGTCTACCTTCAGGCTGCCGAACTAGGTCTTACCACACACCTAACGGAACCTTGACTAATCACTTAAG

Average length ~ 250bp

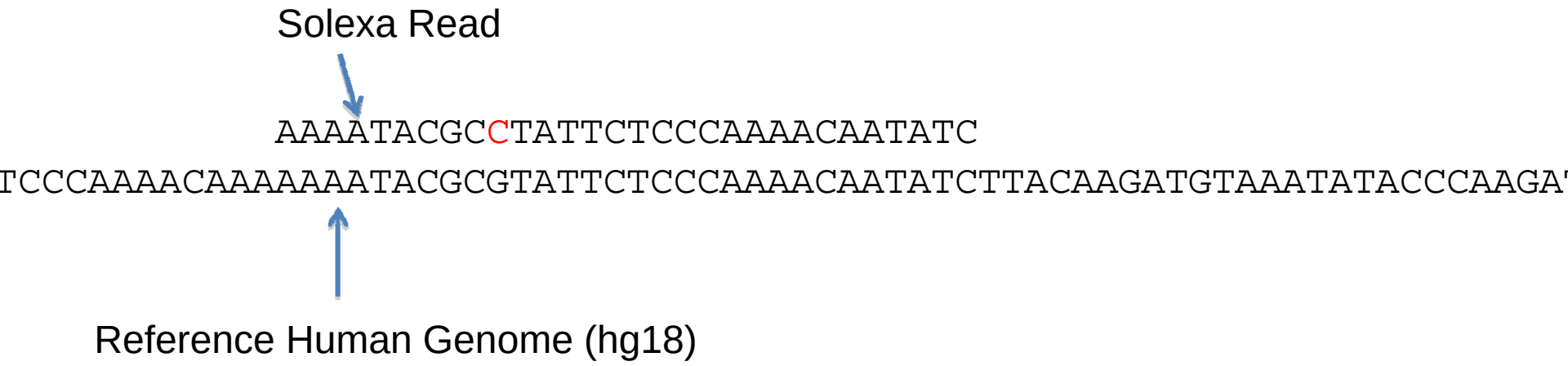
BCL6 ChIP-seq

- Lymphoma cell line (OCI-Ly1)
- 1 lane for ChIP, 1 for input DNA, 1 for QC
- 36nt long sequences
- 30 Million reads
- Aligned/mapped to hg18 with Eland

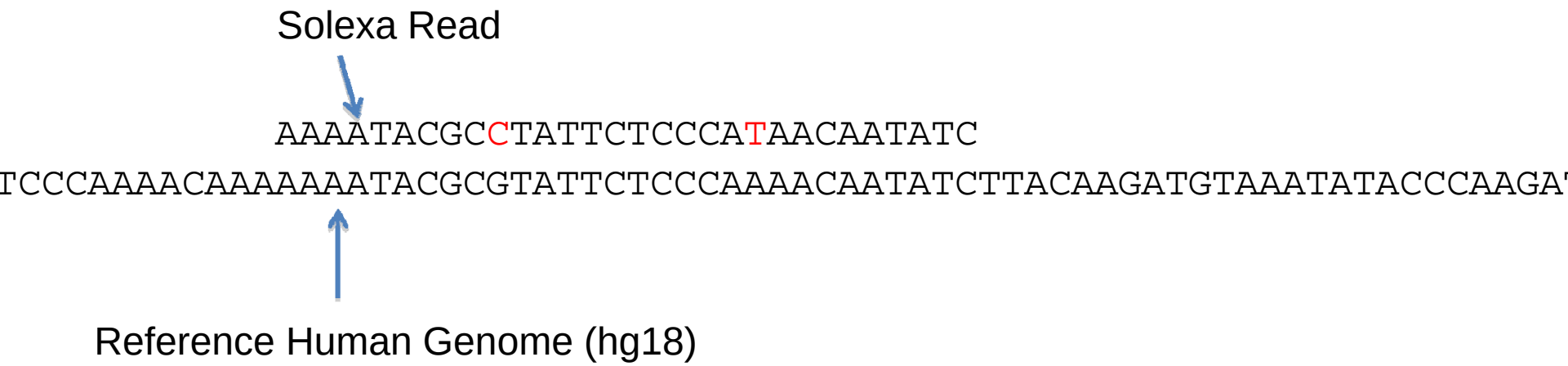
Read mapping with Eland



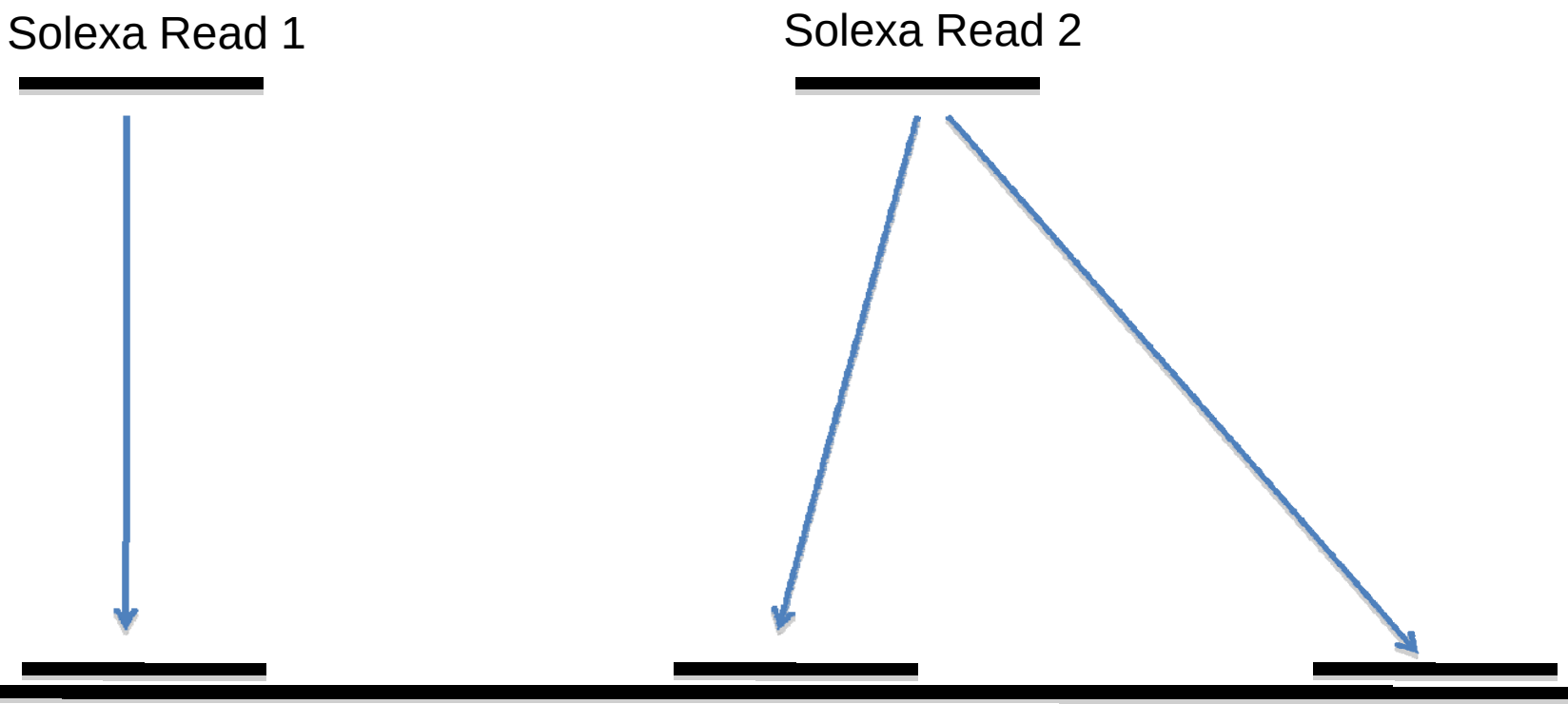
Read mapping with Eland



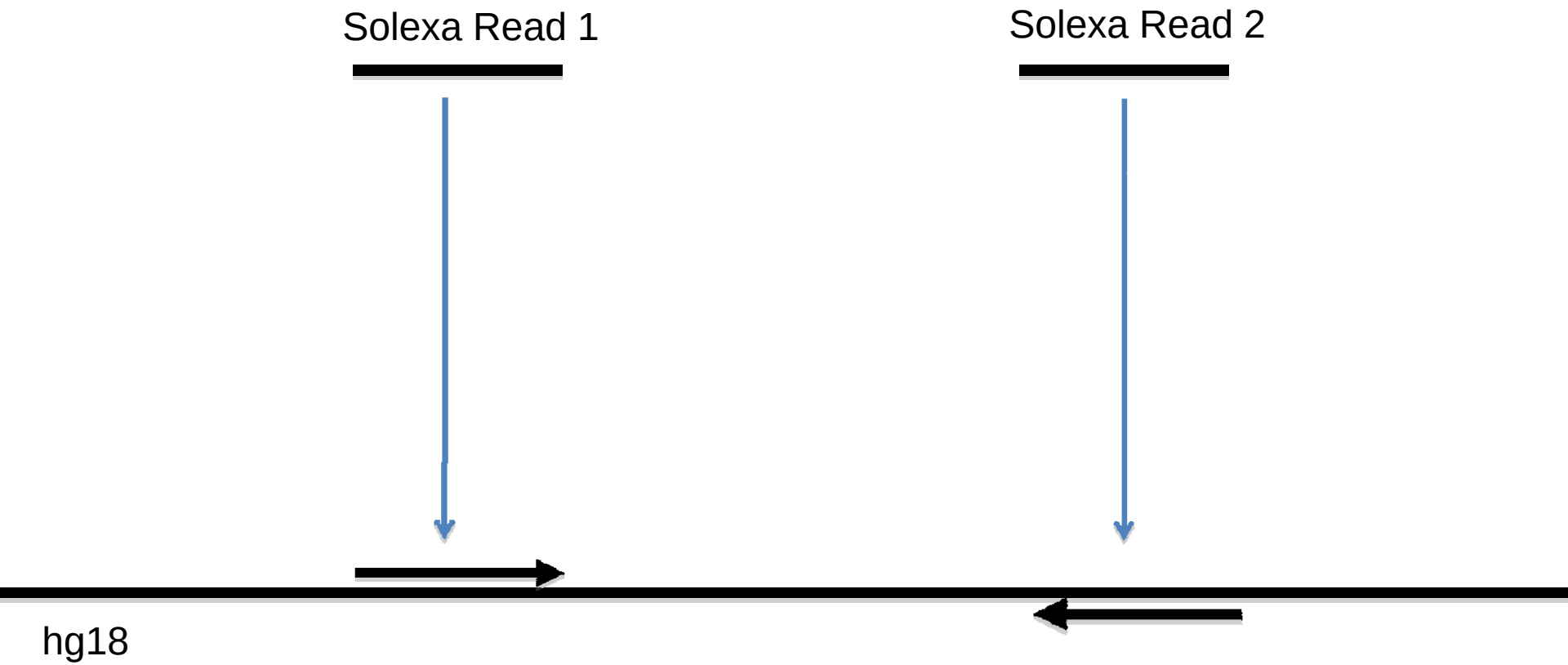
Read mapping with Eland



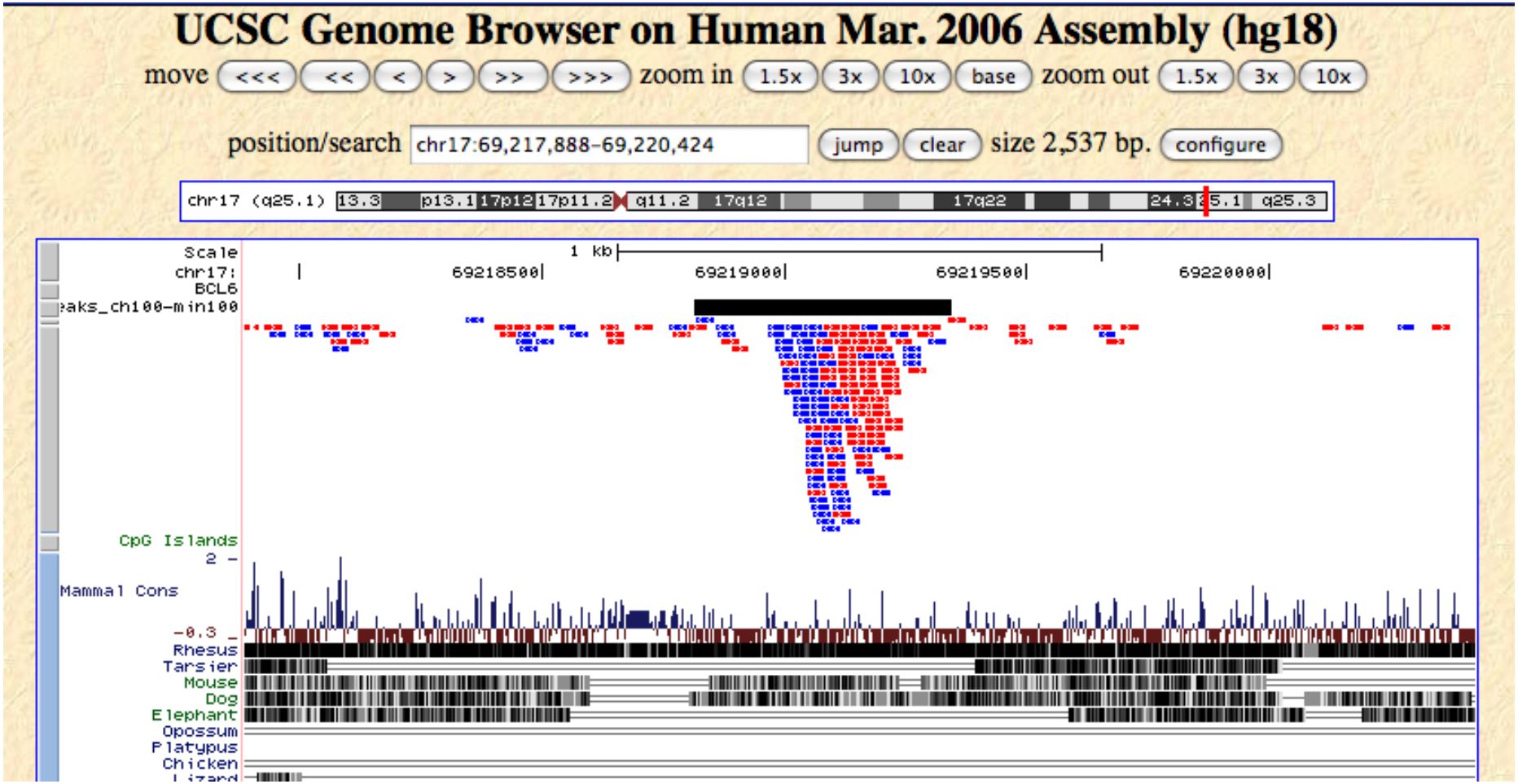
Reads can map to multiple locations/chromosomes



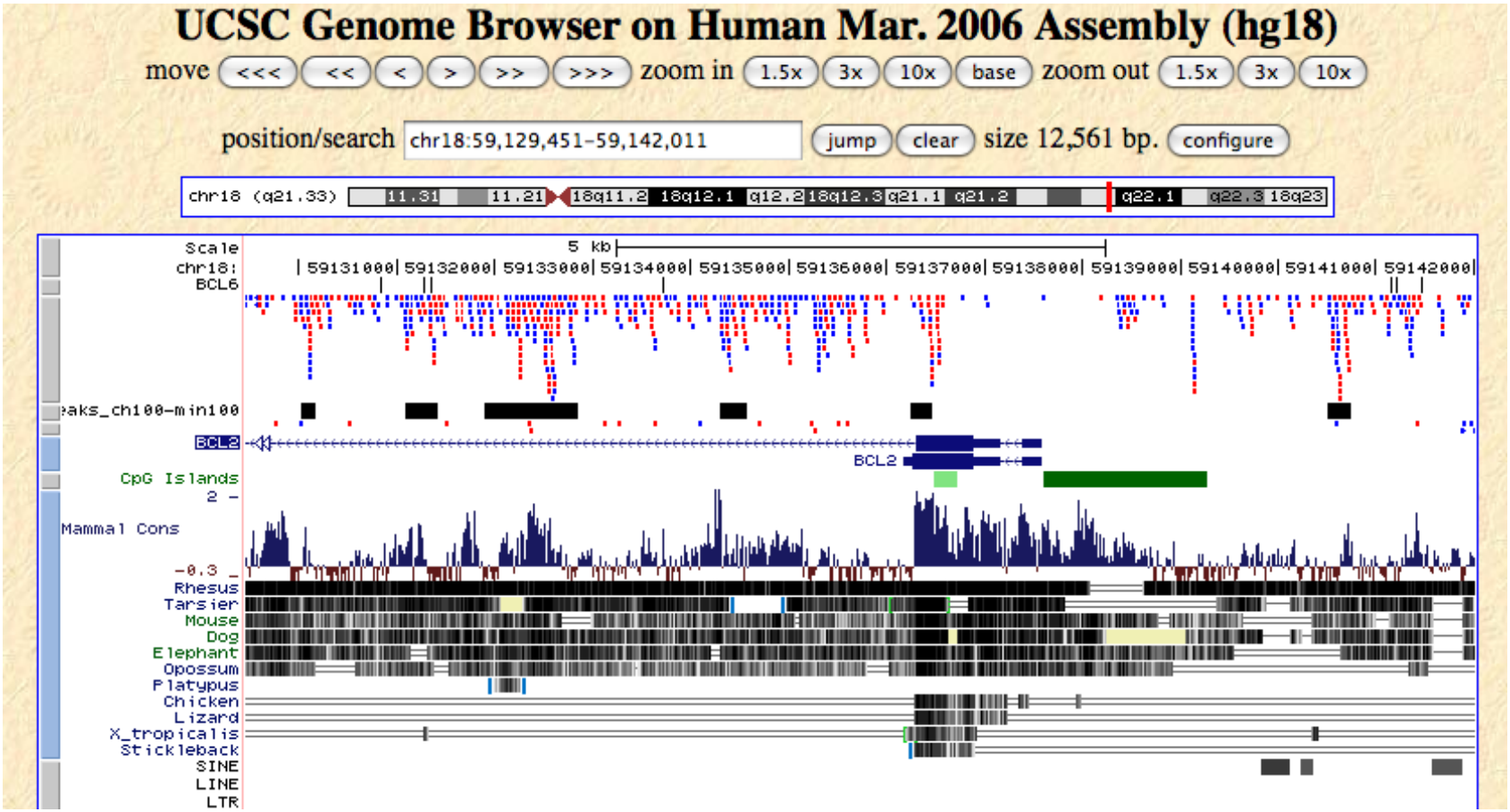
Reads map to one strand or the other



```
TCACAAAACAAGTCCTAACAAATTTAAGAGTAT    U0    1    13    62    chr8.fa      59699745    R    DD
AGAAAAATCCTTTTTATTATATAAACAATACAT    U2    0    0     1     chr5.fa      121195098   F    DD    15G   20G
   GTCATCAAACTCCAAGGATTCTGTTTTCAACATACT    U0    1    1    0    chr18.fa    8914049     R    DD
   GAAAGTGATTAGCAGATTGTCATTTAATAATTGTCT    U2    0    0    1    chr1.fa     97496963    F    DD    18G   28G
   GATAAATTTTTTCCTACAATCTTAAATTATTACACA    U1    0    1    0    chr3.fa     95643444    R    DD    10C
   AAAAATTAAACAATTCTAAAAATATTTTTATCTTAA    U2    0    0    1    chr2.fa     177727639   R    DD    18C   31G
   GCACATGTCATACTCTTTCTAGCTCTCTTATTTTTC    U0    1    0    0    chr8.fa     79132719    R    DD
TTAATGTAAAAAATAGGATACTGAATTGTGATA    U1    0    1    0    chr10.fa    69774166    F    DD    30G
   GTAGTTAACAATAATTTATTTTATACTTCAAAATTC    U1    0    1    17    chrX.fa    26496842    R    DD    7A
AGAATTAATTAATCAAAACACCAAATGTACTTC    U0    1    0    0    chr12.fa    72700465    F    DD
TTGACTTTATTATTTTTTCTTCAATGTTTTTAA    NM    0    0    0
AGTACATCAAATACATATTATATACTTTACATA    R2    0    0    2
   AATCCATATACATTTCTTTTTAATCATTTCCTCTTT    U1    0    1    0    chr11.fa    94204222    F    DD    20G
AGTTTCTTAATCCTGAGTTCTAATTTTATTTCA    R0    29    255    255
TTTTATAAATTTTTAATTTCATTTTAATTTATA    NM    0    0    0
   GTTTTTAAAATCAACACTTTTATTATAGAAGTAGCA    U0    1    0    1    chr12.fa    62166701    R    DD
CTGATGTAAACTTGGTAAAAACATTGACATAAA    U0    1    0    0    chr14.fa    65160857    F    DD
GAAAATGACTATGTCAAAATATTATCTCTCAAT    U0    1    0    0    chr5.fa     97782464    F    DD
   GTTTTACTGATTTTCTTACTTACTAAACTACCTGTT    U0    1    0    0    chr7.fa    133200265    F    DD
GATACGGCGACCACCGACAGGTTCAGAGTTCTA    NM    0    0    0
AATTATTCAGAAGTCAAATCTGTGCTTAGTTTA    U2    0    0    1    chr5.fa     162472124   R    DD    3G    7C
TGTATCATATATATTTATGTATCATATATATTT    R1    0    3    2
   GATTGCTCCATTATTTGTTAAAAACATAGTAAAATA    NM    0    0    0
AGATCAGTACTTCAAAGAGATATCTGCACTCCC    U0    1    1    9    chr12.fa    33830898    R    DD
AGTCCCAATATTCCATTAATCCCAATAAATATA    U2    0    0    1    chr6.fa     110722427   F    DD    15G   19G
ATAATAATAGCAGTTATGGCATCGAGATAATTT    U0    1    0    0    chr2.fa     47305609    R    DD
GAGGGCACACATCACAAACAAGTTTCTGAGAAT    R2    0    0    3
TATCCACTTGCAGACTTTACAAACAAATTTTTT    R2    0    0    4
   GGCAGATGAAACTTCTATACACTATATTTTAGCCAG    U0    1    0    0    chr13.fa    90021137    F    DD
AGAAAAACTATTGAAAAAATAGTTACTTTCCAA    U0    1    0    0    chr1.fa     74303257    R    DD
TAGATGATATCGAGGGCATTAGAAGTAAATAGC    U0    1    0    0    chr5.fa     16031200    F    DD
AGGAAATAATAAAGATAAAAGTAGAAAAAGTGA    U0    1    0    0    chr1.fa     187326417   F    DD
AATTATGTTGTTGTAATTATTGTTTGTTTTTTT    U0    1    0    0    chr15.fa    46739015    R    DD
GACAATCCAGCTGTCATAGAAACTGACTATTTT    U0    1    0    0    chr12.fa    38910133    R    DD
AATTCTCCCAAAACAACAAGATGTAAATATACC    U0    1    0    0    chr3.fa     101625712   R    DD
CTTACACTGATATGAAGAAATACCTGAGACTGG    U0    1    2    67   chr2.fa     214128537   R    DD
   GAGAAACACACATATTTTTGTAAGTGCCATCACATC    U1    0    1    0    chr7.fa    13668652    R    DD    18C
TTATCTAACACACAAGATGATGTTTGTTTTTAT    NM    0    0    0
TGTAGAAAATTTTCTGCCCTAAAATATTTGTTA    U1    0    1    0    chr6.fa     74625385    F    DD    13G
TCCTAAAGTGTATCTTATGTTTTTTCATCTTCT    U1    0    1    0    chr12.fa    7400023     R    DD    9C
   AATAAAACAAATTCCAATGGCTTAGATTCTACTTAA    U2    0    0    1    chr10.fa    98020799    R    DD    15C   20C
TGGTCATACTTCCCAAAGCGATCTACAGATTCA    U1    0    1    29   chr3.fa     50834510    R    DD    19C
TTTCCACATTTCTGTGGAAGCCTCACAATCATT    R2    0    0    2
   ATTAATCAACAGCAACATTAATCAACTGAATCAACA    U0    1    0    0    chr2.fa     46078825    R    DD
   GAATAAATAATCAAAACATATAATACATTTTTTTAT    U1    0    1    0    chr5.fa     41496935    F    DD    32G
TACACATATATATACATATATATATACACATAT    R0    47    255    255
   GAGAAGGAAATGTGTTTTCTAAGTTTCTTTATCTTC    U1    0    1    0    chr4.fa     188020201   F    DD    32G
```


UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr17:69,217,888-69,220,424 jump clear size 2,537 bp. configure
chr17 (q25.1) 13.3 p13.1 17p12 17p11.2 q11.2 17q12 17q22 24.3 25.1 q25.3
Scale
chr17:
1 kb
69218500
69219000
69219500
69220000
BCL6
aks_ch100-min100
CpG Islands
Mammal Cons
Rhesus
Tarsier
Mouse
Dog
Elephant
Opossum
Platypus
Chicken


position/search
chr17:7,506,319-7,545,306
jump
clear
size 38,988 bp.
configure
chr17 (p13.1)
chr17:
7510000
7515000
7520000
7525000
7530000
7535000
7540000
7545000
BCL6
BCL6_chr8_reads_r1
BCL6_input_chr8_reads_r1
BCL6_peaks_ch100-min100
BCL6_reads_chr17
sidow_peaks
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
CpG Islands
CpG Islands (Islands < 300 Bases are Light Green)


UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr18:59,129,451-59,142,011 jump clear size 12,561 bp. configure
chr18 (q21.33) 11.31 11.21 18q11.2 18q12.1 q12.2 18q12.3 q21.1 q21.2 q22.1 q22.3 18q23
Scale
5 kb
chr18:
59131000 59132000 59133000 59134000 59135000 59136000 59137000 59138000 59139000 59140000 59141000 59142000
BCL6
BCL2
CpG Islands
Mammal Cons
Rhesus
Tarsier
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
SINE
LINE
LTR

Peak detection

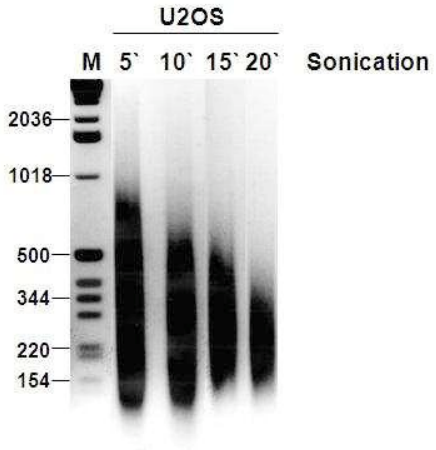
- Calculate read count at each position (bp) in genome
- Determine if read count is greater than expected

Peak detection

- We need to correct for input DNA reads (control)
-

  - non-uniformaly distributed (form peaks too)

  - vastly different numbers of reads between ChIP and input



genome

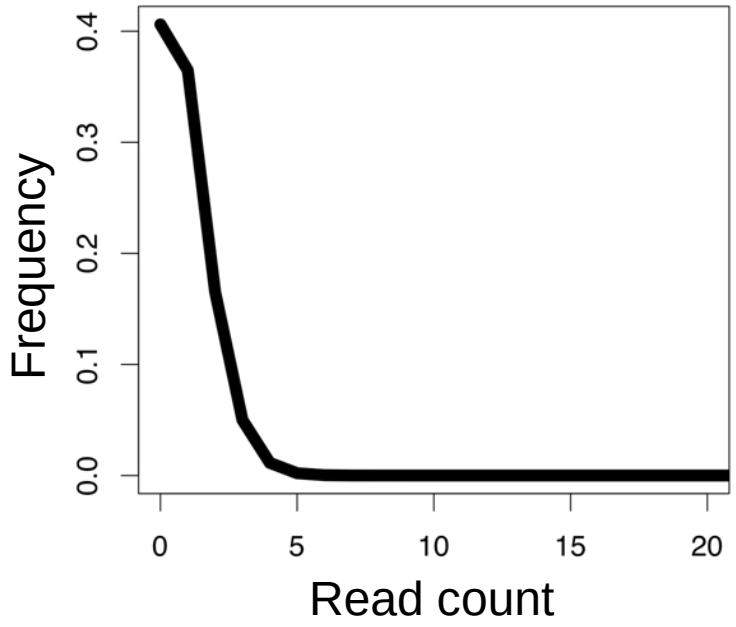
Read count

Expected read count

T A T T A A T T A T C C C C A T A T A T G A T A T genome

Expected read count = total number of reads * extended fragment length / chr length

Is the observed read count at a given genomic position greater than expected ?



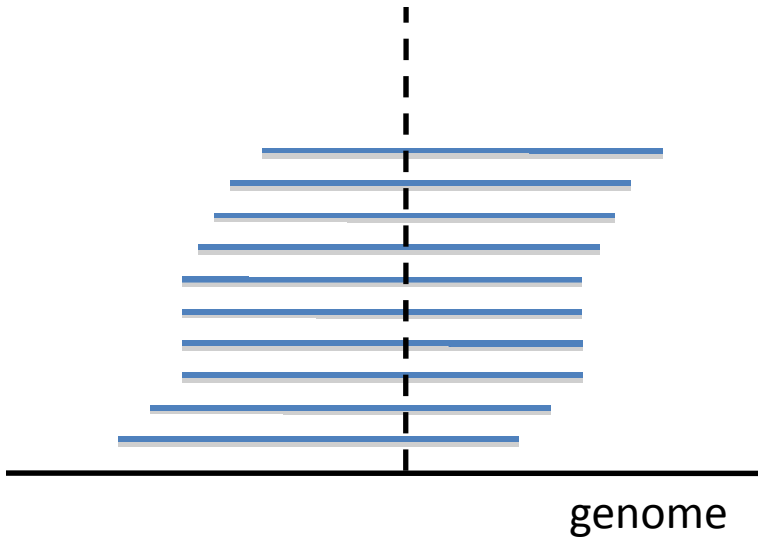
$$P(X \geq x) = 1 - \sum_{0}^{x-1} \frac{\lambda^x e^{-\lambda}}{x!}$$

x = observed read count

λ = expected read count

The Poisson distribution

Is the observed read count at a given genomic position greater than expected ?



$$P(X \geq x) = 1 - \sum_{0}^{x-1} \frac{\lambda^x e^{-\lambda}}{x!}$$

x = 10 reads (observed)

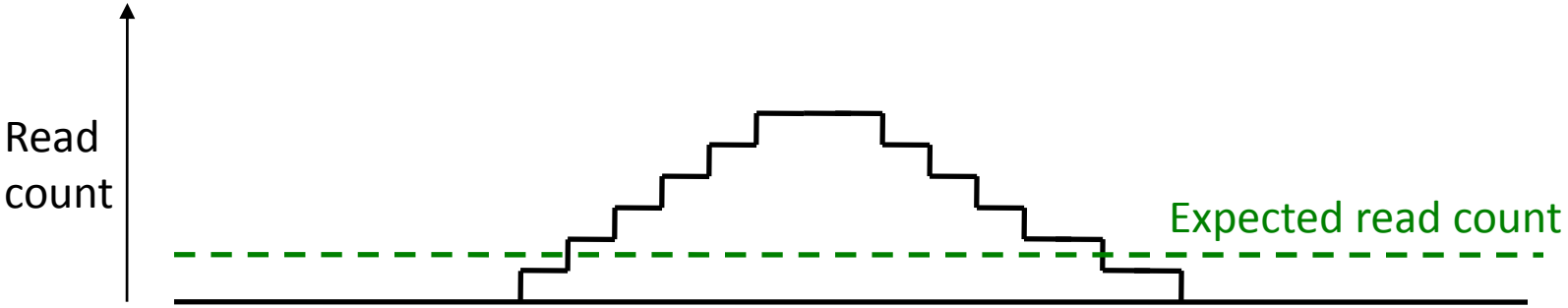
λ = 0.5 reads (expected)

$P(X>=10) = 1.7 \times 10^{-10}$

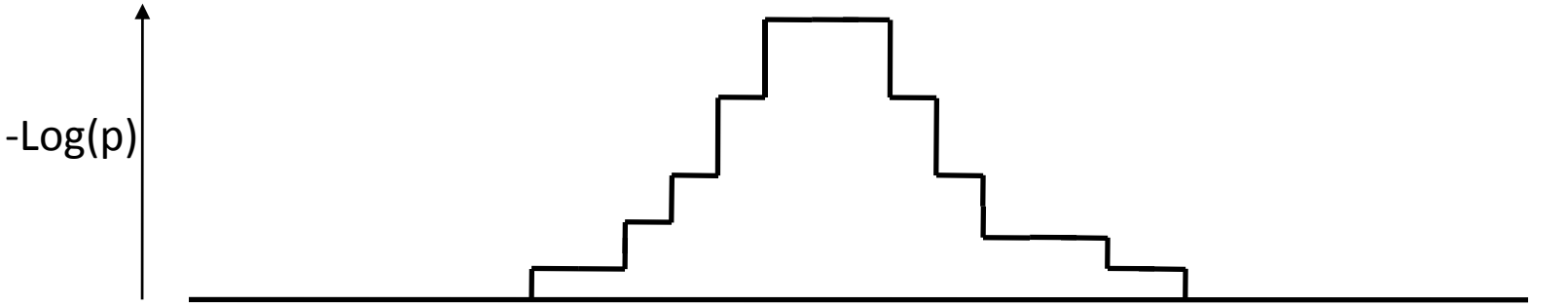
log10 P(X>=10) = -9.77

-log10 P(X>=10) = 9.77

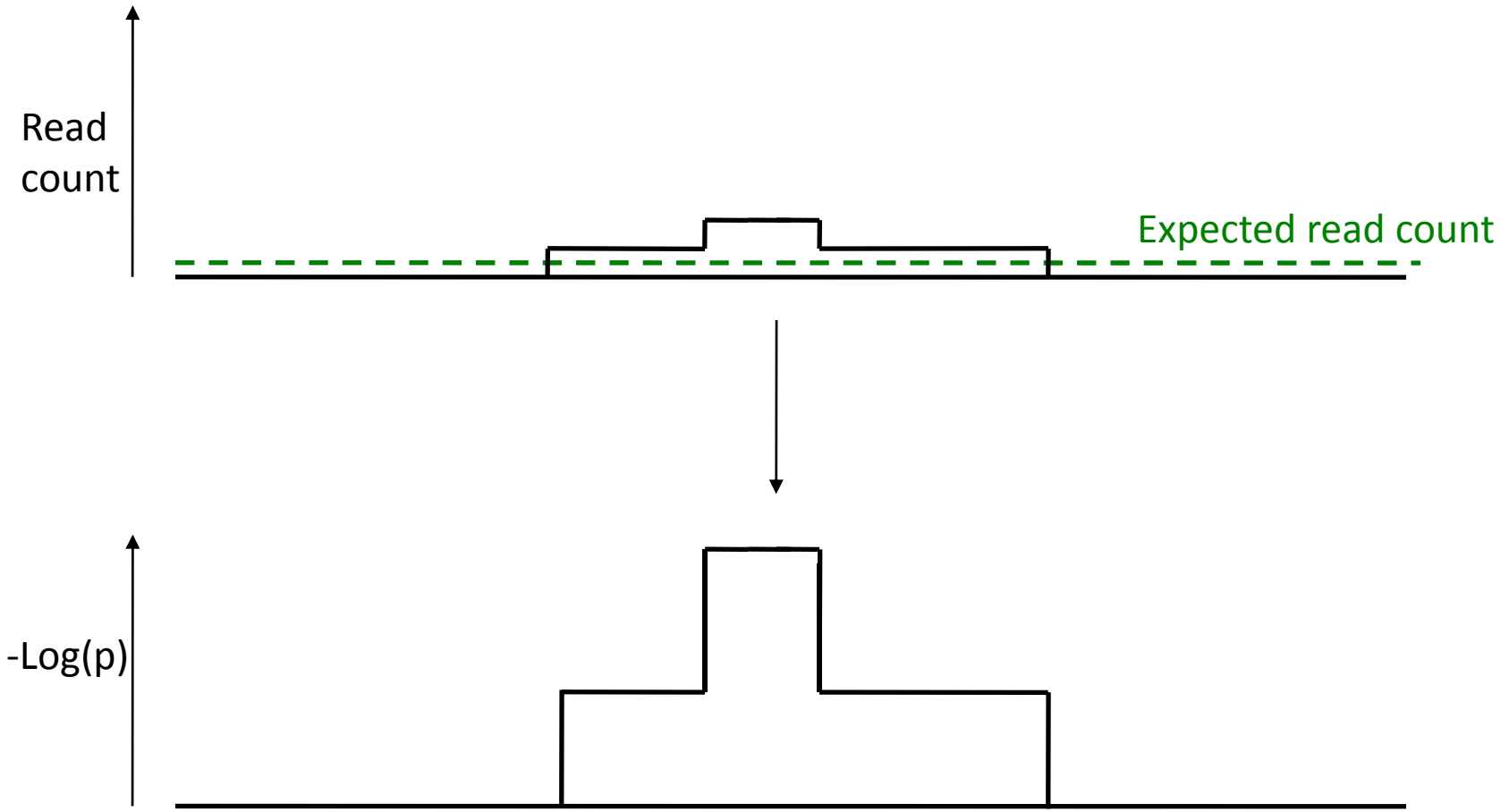
The Poisson distribution



$$P_c(X \geq x) = 1 - \sum_{0}^{x-1} \frac{\lambda_c^x e^{-\lambda_c}}{x!}$$

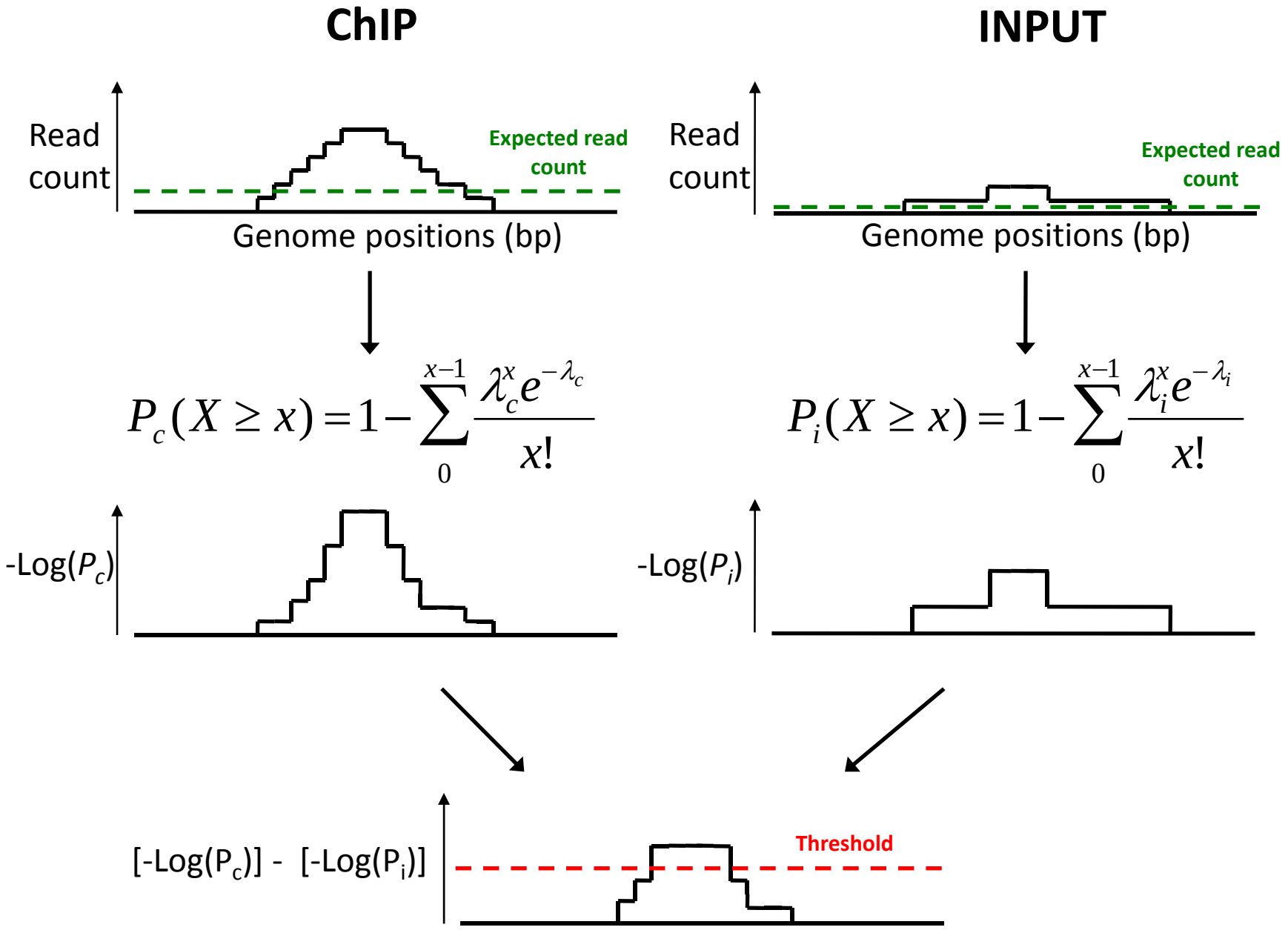


Expected read count = total number of reads * extended frag len / chr len



Expected read count = total number of reads * extended frag len / chr len

Input reads


ChIP
INPUT
Read count
Expected read count
Genome positions (bp)
Read count
Expected read count
Genome positions (bp)
$P_c(X \geq x) = 1 - \sum_{0}^{x-1} \frac{\lambda_c^x e^{-\lambda_c}}{x!}$
$P_i(X \geq x) = 1 - \sum_{0}^{x-1} \frac{\lambda_i^x e^{-\lambda_i}}{x!}$
-Log(P_c)
-Log(P_i)
[-Log(P_c)] - [-Log(P_i)]
Threshold

Normalized Peak score (at each bp)

$$R = -\log 10 \frac{P(X_{ChIP})}{P(X_{input})}$$

Will detect peaks with high read counts in ChIP, low in Input

Works when no input DNA !

$$P_i(X \geq x) = 1 - \sum_{0}^{x-1} \frac{\lambda_i^x e^{-\lambda_i}}{x!}$$

Non-mappable fraction of the genome

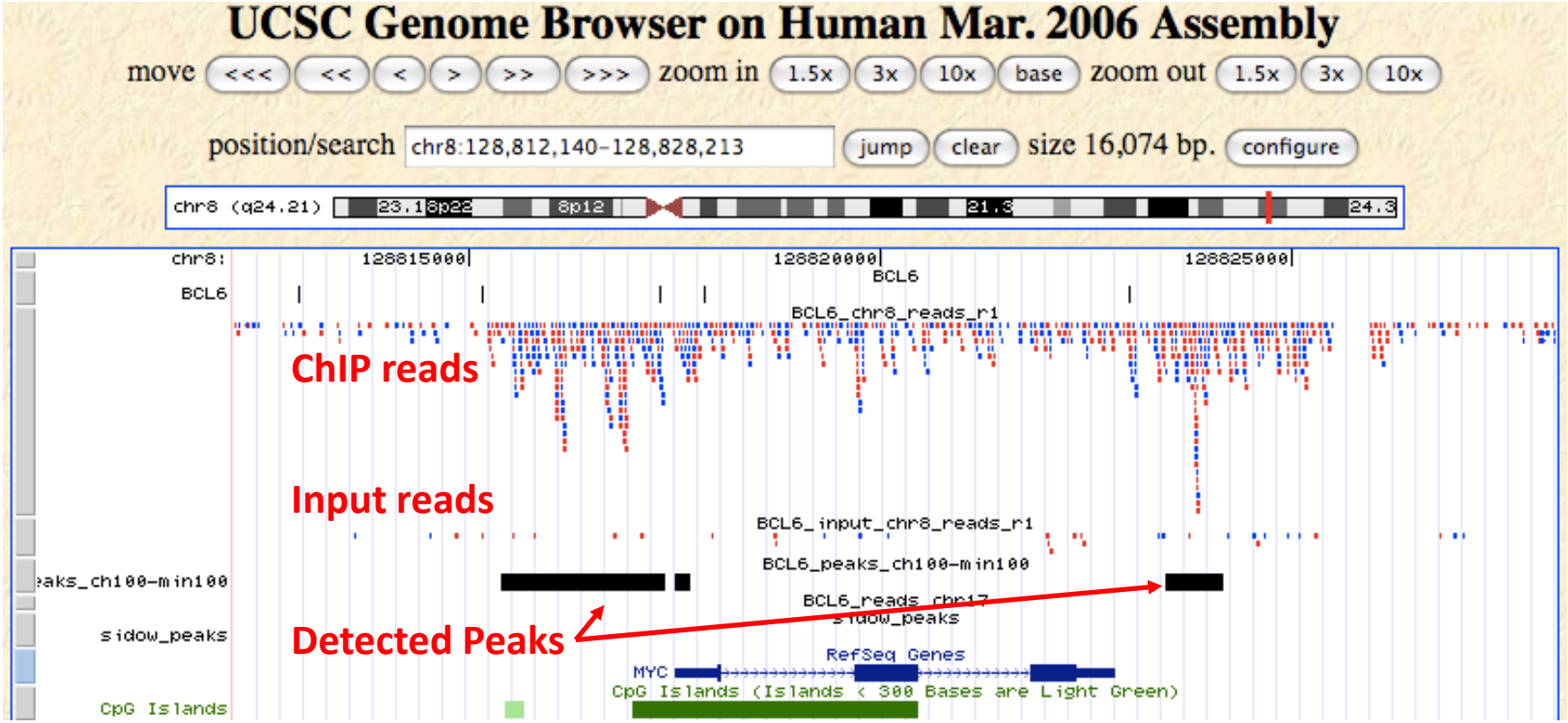
We enumerated all 30-mers, counted # occurrences, calculated non-unique fraction of genome

chr18	9369067/76117153	0.123087459668913 (=12%)
chr2	33849240/242951149	0.139325292921335
chr3	27854877/199501827	0.139622164963933
chr4	27090014/191273063	0.141630052737745
chr6	24330283/170899992	0.142365618132972
chr8	20932821/146274826	0.143106107677065
chr5	26029902/180857866	0.143924633059643
chr12	19382853/132349534	0.14645199279659
chr11	20039443/134452384	0.149044906485258
chr20	10017788/62435964	0.160449000194824
chr7	26182588/158821424	0.164855517225434
chr10	22968951/135374737	0.169669404417753
chr17	14496284/78774742	0.184021980040252
chrX	31269270/154913754	0.201849540099583
chr1	55186693/247249719	0.223202247602959
chr13	28668063/114142980	0.251159230291692
chr16	23552340/88827254	0.265147676410215
chr14	29689825/106368585	0.279122120502026
chrM	4628/16571	0.279283084907368
chr9	43125838/140273252	0.307441635415995
chr19	20251255/63811651	0.317359834491667
chr15	31877970/100338915	0.317702957023205
chr21	16867677/46944323	0.359312392256674
chr22	21176578/49691432	0.426161556382597
chrY	43209644/57772954	0.747921665906161 (=74%)

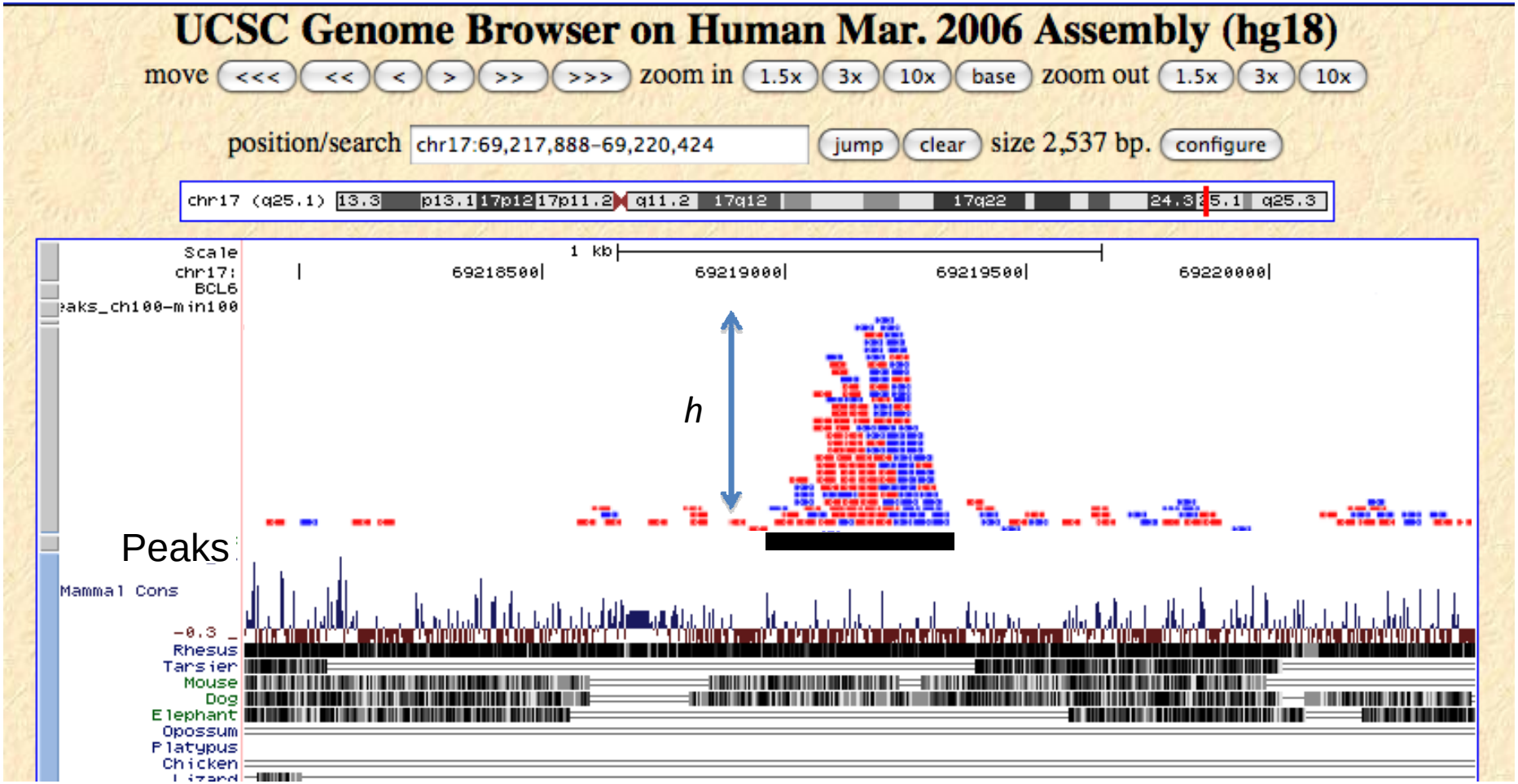
Peak detection

- Determine all genomic regions with R>=15
- Merge peaks separated by less than 100bp
- Output all peaks with length >= 100bp
- Process 23M reads in <7mins

BCL6: 18,814 peaks



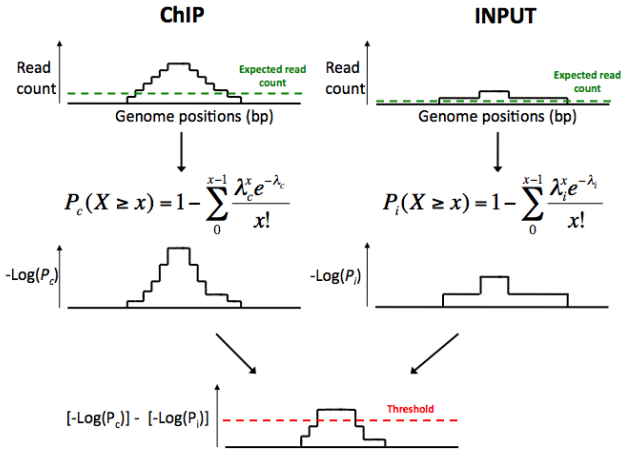
80% are within <20kb of a known gene

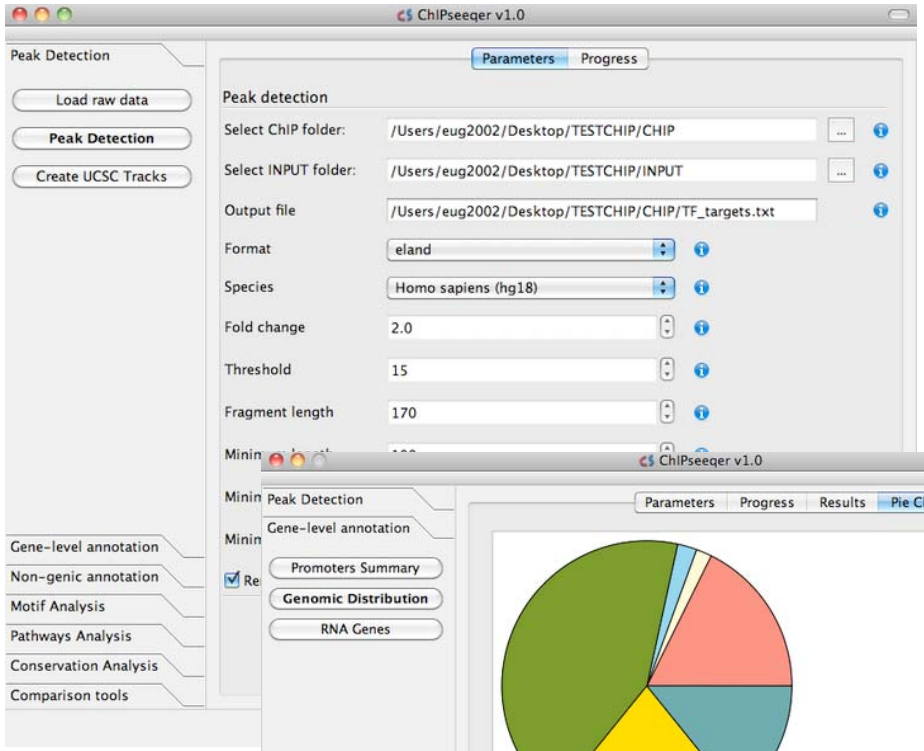


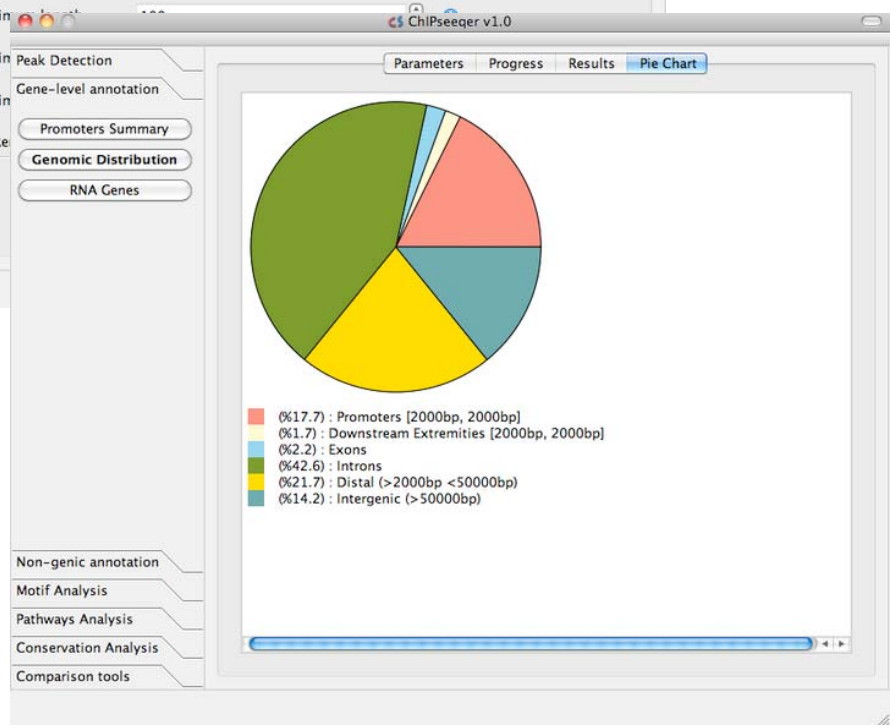
Transcription factor binding

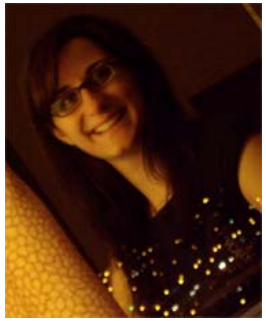
Human genome

ChIPseeqer









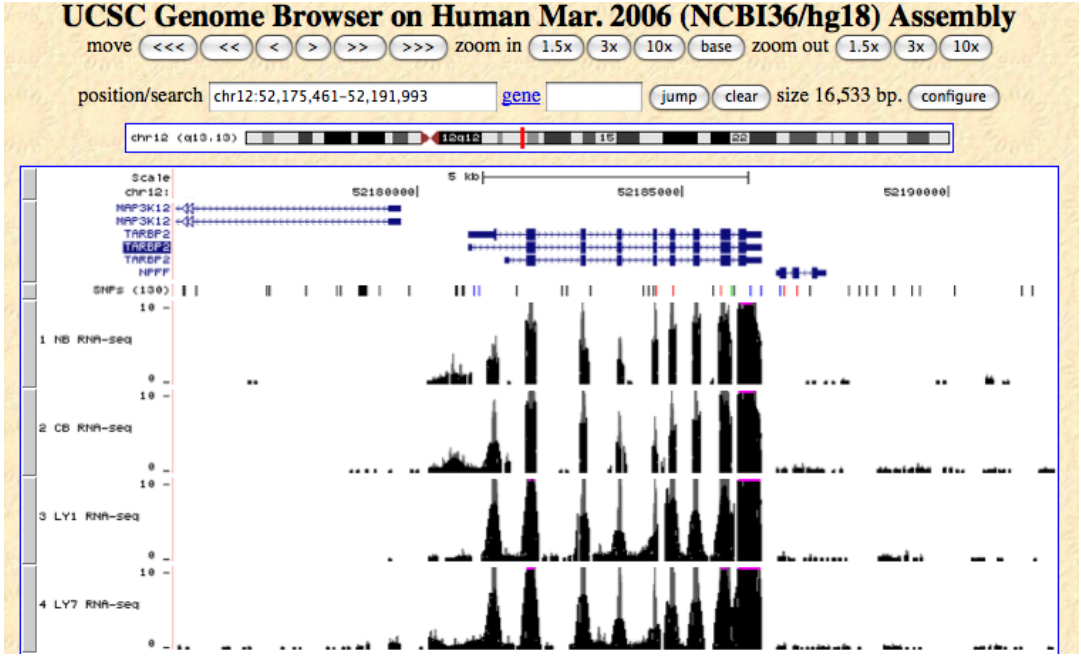
Jenny Giannopulou

ChIP-seq in lymphoma cells (LY1 cell line)

Transcription Factor / Histone modification	B cell/lymphoma function	Current status			
		LY1	LY3	CB	NB
BCL6	Oncogene and master regulator of germinal center phenotype	X		X	N/A
PU.1	Myeloid and B cell development	X			
PAX5	B cell lineage commitment	X			
CTCF	Insulator and enhancer blocking	X		X	X
BCOR	BCL6 co-repressor	X			
MTA3	BCL6 co-repressor	X			
EZH2	Catalytic subunit of Polycomb			X	N/A
H3K4me3	Marks transcriptionally active promoters	X	X	X	X
H3K4me1	Marks active promoters and enhancers	X		X	X
H3K9Ac	Marks active promoters	X			
H3K27Ac	Marks active promoters and enhancers	X			
H3K27me3	Marks silenced promoters	X		X	X
H3K79me2	Marks elongating promoters	X			
H3K79me3	Marks active promoters	X			
DNA methylation	Epigenetic Promoter Mark	X	X	X	X

Yanwen Jiang

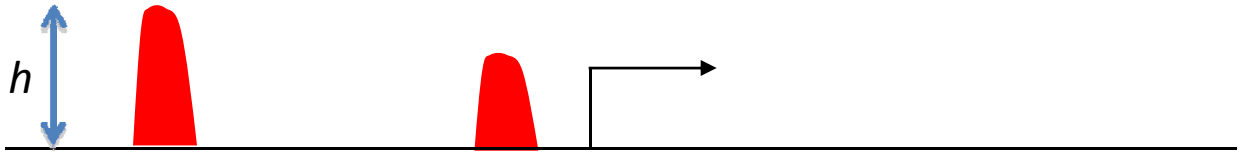
RNA-seq in LY1 cells

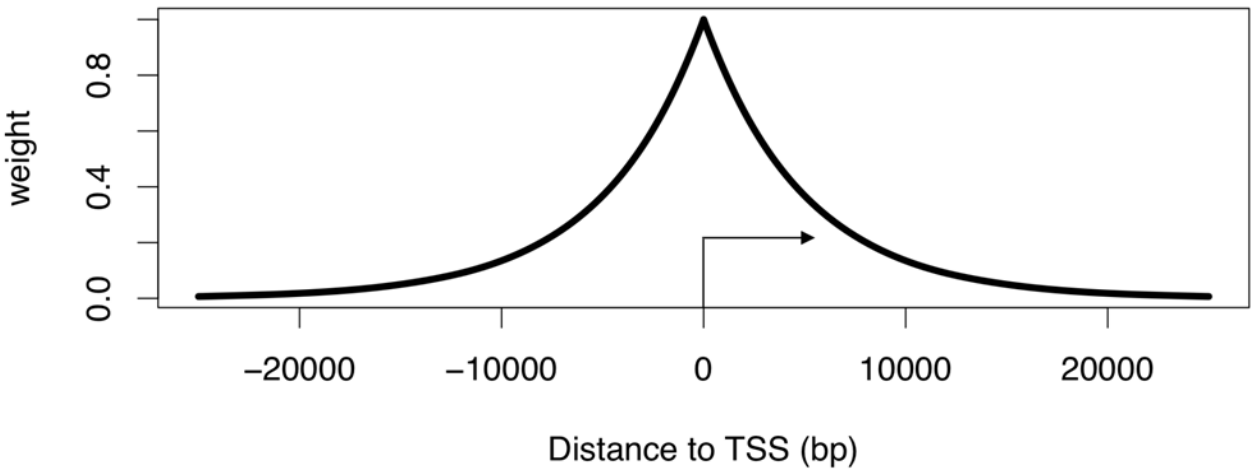


	RPKM
GENE	LY1-RNAseq
NM_018117	24.9
NM_001130845	45.9
NM_021107	32.9
NM_173803	1.2
NM_006528	1.6
NM_182607	1.7
NM_017722	20.2
NM_018283	26.9
NM_014068	0.2
NM_006228	20.1
NM_183377	0.0
NM_002115	0.0
NM_004504	2.9
NM_004358	35.3
NM_022114	0.0
NM_032125	40.1
NM_001011666	1.6
NM_018905	0.0
NM_080746	1.2
NM_001145155	0.0
NM_001040167	0.2
NM_001144994	0.0
NM_017812	63.1

RPKM = # reads per kilobase per million reads

Modeling the influence of a TF's binding on a promoter





$$x_{TF=i,promoter=j} = \sum_k h_k e^{-d_k/d_0}$$

h_k: Peak height

d_k/d_0: Weight

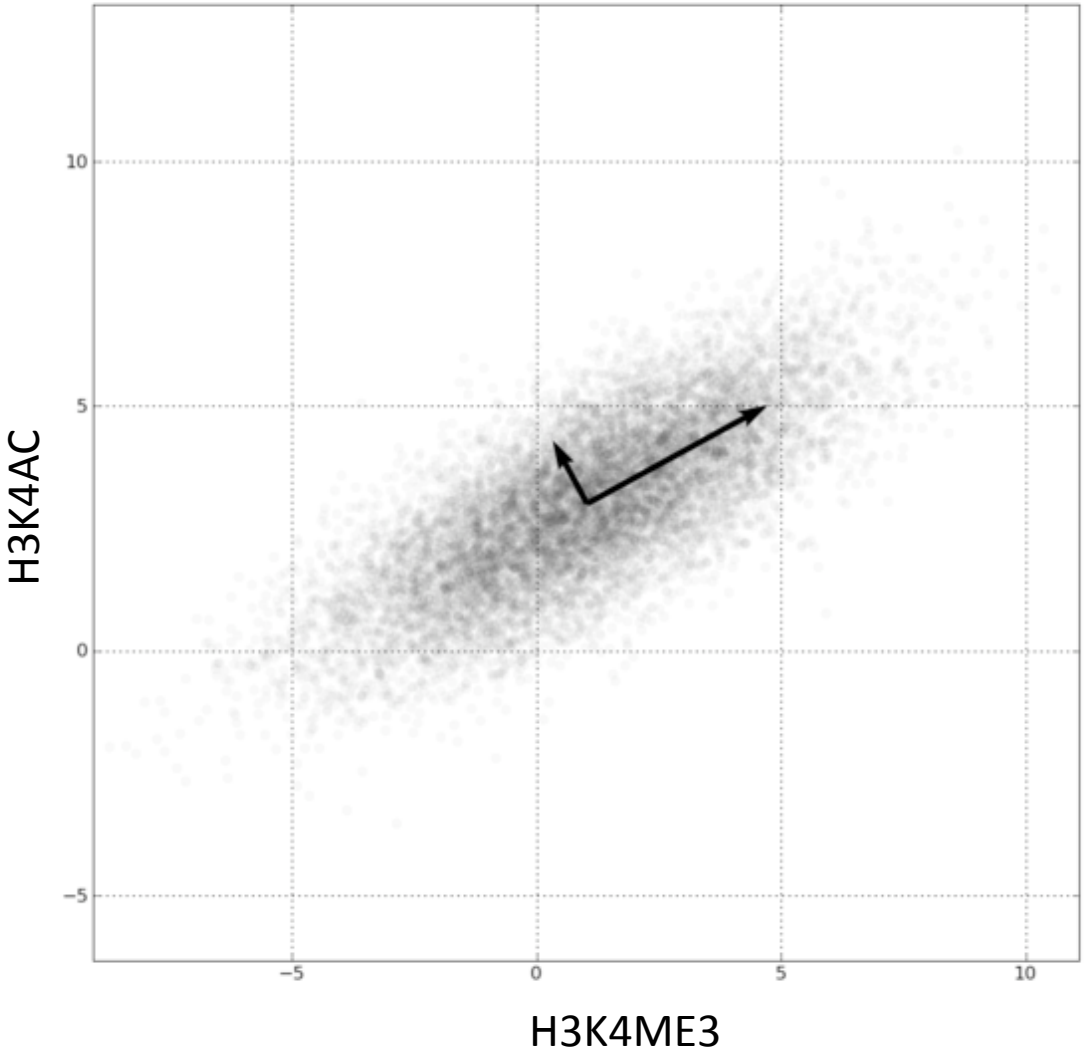
Transcription factors / histone modifications

Transcripts

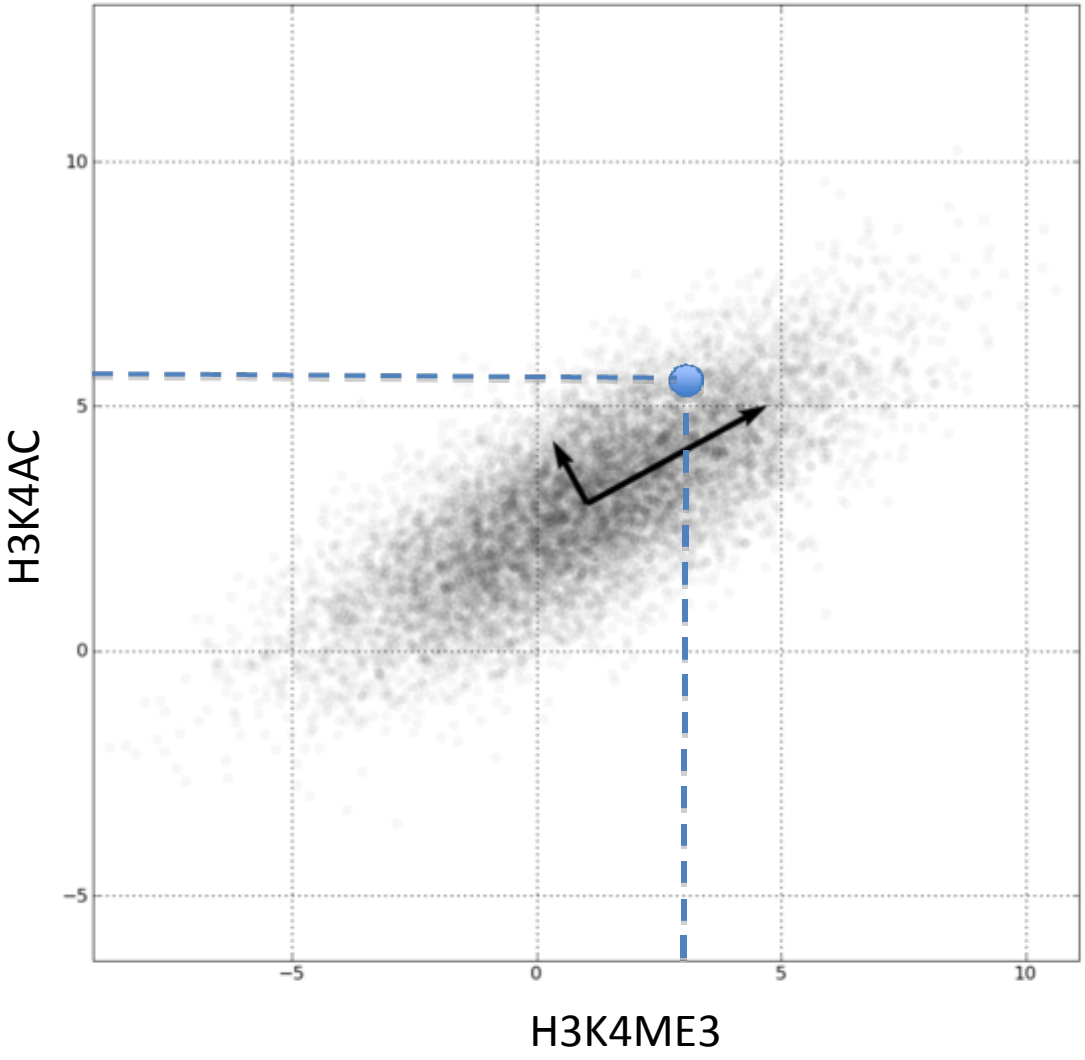
GENE	BCL6	MTA3	BCOR	K4ME1	K4ME3	K79ME2	K79ME3	K79AC	...	DNAm
NM_018117	17.54	23.69	29.20	56.05	100.25	0.00	38.81	49.35	...	-0.17
NM_001130845	126.7	203.7	373.2	113.4	58.08	104.7	148.5	117.3	...	0.56
NM_021107	0.00	0.00	0.05	41.03	222.2	18.53	48.24	87.66	...	0.92
NM_173803	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	-0.27
NM_006528	0.00	16.19	35.06	40.42	113.3	0.00	0.00	0.00	...	0.56
NM_182607	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	...	-0.74
NM_017722	89.05	3.96	66.30	59.98	183.1	10.06	114.6	37.54	...	0.30
NM_018283	0.00	19.48	28.95	53.16	85.51	0.02	0.06	105.3	...	0.25
NM_014068	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.43
NM_006228	16.98	0.19	0.58	8.33	0.87	0.00	0.01	1.19	...	-0.23
NM_183377	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	-1.11
NM_002115	0.09	0.00	0.30	0.00	0.00	0.00	0.00	0.00	...	-0.12
NM_004504	2.38	0.00	22.70	49.26	64.23	20.06	122.1	7.80	...	0.08
NM_004358	0.00	73.31	78.10	101.7	109.0	36.74	44.79	90.73	...	0.67
NM_022114	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	-0.24
NM_032125	0.00	0.32	23.41	57.10	157.6	49.68	121.2	29.71	...	-0.30
NM_001011666	21.05	0.00	0.00	0.48	0.00	0.00	0.00	0.00	...	-0.17
NM_018905	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.72
NM_080746	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.06
NM_001145155	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.37
NM_001040167	2.33	31.89	209.1	5.17	40.27	0.00	0.00	0.03	...	0.34
NM_001144994	0.00	0.87	1.98	2.09	6.81	1.05	1.10	1.66	...	0.29
NM_017812	31.99	0.00	17.73	50.97	120.3	87.4	166.9	29.53	...	-0.08

TF and histone marks are very highly correlated !

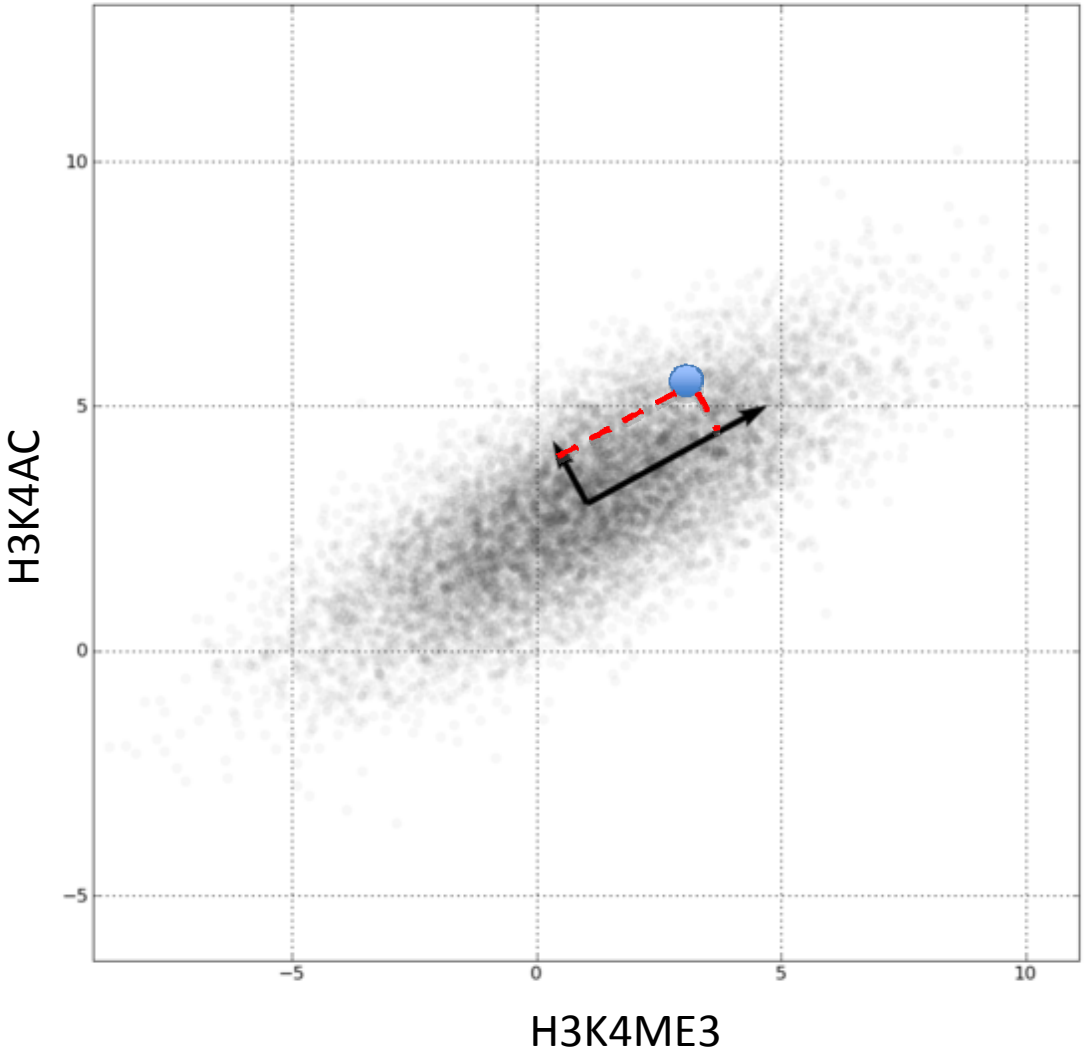
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



Promoter	PC1	PC2	PC3	PC4	PC5	PC6	PC7	...	RNAseq-RPKM
NM_018117	2.65	0.81	0.12	-0.0	0.62	0.91	-0.4	...	24.9
NM_001130845	13.9	11.4	2.89	-1.9	4.10	-0.6	0.75	...	45.9
NM_021107	4.47	2.22	-1.5	0.43	-0.3	-3.4	0.74	...	32.9
NM_173803	-0.1	0.04	0.09	-0.1	0.31	0.01	-0.4	...	1.2
NM_006528	2.03	2.74	2.18	1.46	-0.2	-0.5	-2.0	...	1.6
NM_182607	-0.3	0.05	0.23	-0.3	0.89	0.01	-1.8	...	1.7
NM_017722	5.12	0.96	1.41	0.54	0.07	1.33	1.07	...	20.2
NM_018283	3.32	2.31	-0.5	0.07	0.96	-2.0	0.57	...	26.9
NM_014068	-0.3	1.56	0.18	0.38	-1.2	0.16	-1.2	...	0.2
NM_006228	0.02	1.18	0.79	0.08	0.11	-0.1	-0.9	...	20.1
NM_183377	-0.4	0.01	0.34	-0.5	1.33	0.02	-1.7	...	0.0
NM_002115	-0.2	0.67	0.17	0.00	-0.1	0.08	-1.1	...	0.0
NM_004504	2.75	1.27	0.16	0.52	0.15	-2.1	-1.2	...	2.9
NM_004358	6.30	3.30	2.27	-8.3	-2.9	0.31	-0.9	...	35.3
NM_022114	-0.1	0.05	0.08	-0.1	0.27	0.01	-0.4	...	0.0
NM_032125	3.48	-0.7	-0.9	-0.0	-0.0	-0.7	-1.7	...	40.1
NM_001011666	-0.1	1.10	0.86	0.11	-0.1	0.41	-1.6	...	1.6
NM_018905	0.08	1.46	-0.0	0.47	-1.0	-0.7	0.03	...	0.0
NM_080746	0.04	1.07	0.05	0.07	0.08	-1.0	-0.4	...	1.2
NM_001145155	-0.1	1.80	0.10	0.31	-0.6	-1.1	-0.8	...	0.0
NM_001040167	2.28	4.86	-0.0	-1.5	-0.1	2.35	-2.5	...	0.2
NM_001144994	-0.2	1.46	0.14	0.28	-0.8	0.02	-1.4	...	0.0
NM_017812	4.16	-1.3	-0.0	0.38	-0.9	-0.4	-1.5	...	63.1
NM_194249	0.58	-0.2	-0.0	-0.2	0.60	-0.0	-0.7	...	18.7
NM_199005	-0.1	-0.1	0.04	-0.1	0.36	-0.0	-0.7	...	3.3
NM_182626	0.21	2.35	0.00	0.41	-0.8	-0.9	-0.6	...	1.0
NM_001170689	2.62	2.67	-1.0	-0.1	0.49	0.64	-0.0	...	4.2
NM_152312	4.30	2.24	-0.2	-0.6	-0.4	1.28	-1.2	...	11.0
NM_006863	-0.2	0.34	0.11	-0.0	0.02	0.04	-0.2	...	0.0

… (~25,000 unique RefSeq promoters)

PCs are orthogonal !

Model:

$$RPKM_i = \beta_0 + \sum_{j=1}^{m} \beta_j PC_{ij}$$

Model fitting

Find $\hat{\beta}_j$ that minimize

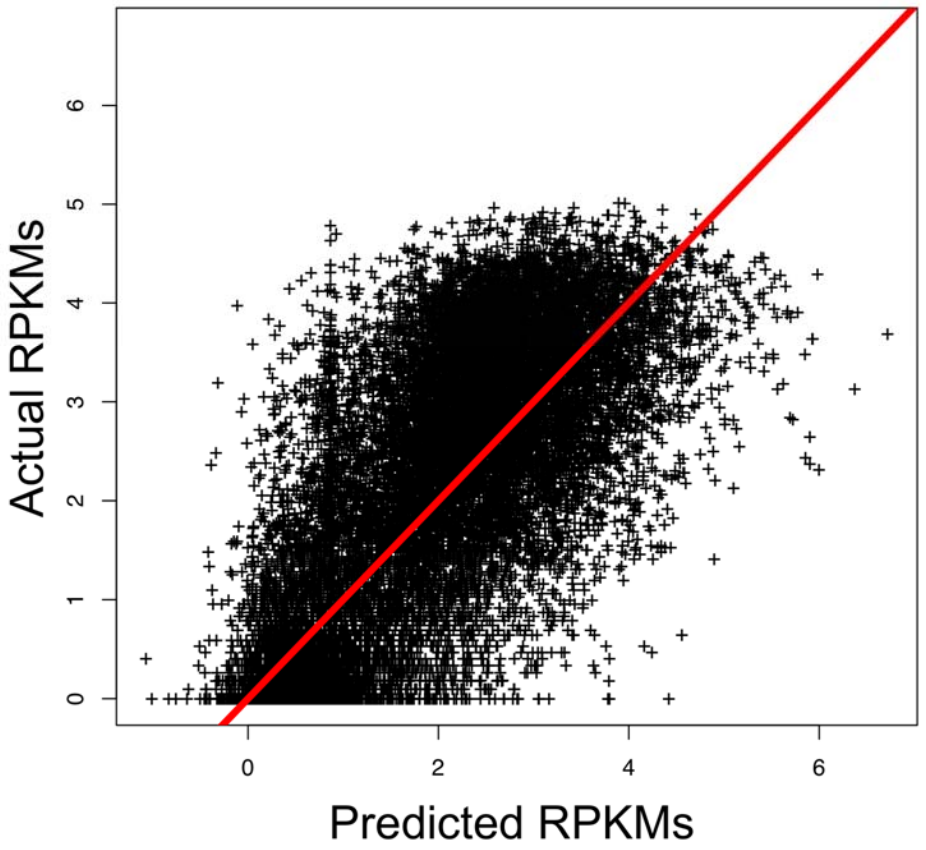
$$\sum_{i=1}^{n}(RPKM_i - \hat{\beta}_0 + \sum_{j=1}^{m}\hat{\beta}_j PC_{ij})^2$$

Model assessment

$$R^2 = 1 - \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2$$

(y = RPKM)

7 TF and 8 histone modifications predict 65% of variance in gene expression levels



Transcription Factor / Histone modification	B cell/lymphoma function
BCL6	Oncogene and master regulator of germinal center phenotype
PU.1	Myeloid and B cell development
PAX5	B cell lineage commitment
CTCF	Insulator and enhancer blocking
BCOR	BCL6 co-repressor
MTA3	BCL6 co-repressor
EZH2	Catalytic subunit of Polycomb
H3K4me3	Marks transcriptionally active promoters
H3K4me1	Marks active promoters and enhancers
H3K9Ac	Marks active promoters
H3K27Ac	Marks active promoters and enhancers
H3K27me3	Marks silenced promoters
H3K79me2	Marks elongating promoters
H3K79me3	Marks active promoters
DNA methylation	Epigenetic Promoter Mark

R^2=0.65, Spearman=0.804

Assessing individual coefficients

$$RPKM_i = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j PC_{ij}$$

Calculate t-statistic

$$t_j = \hat{\beta}_j / se(\hat{\beta}_j)$$

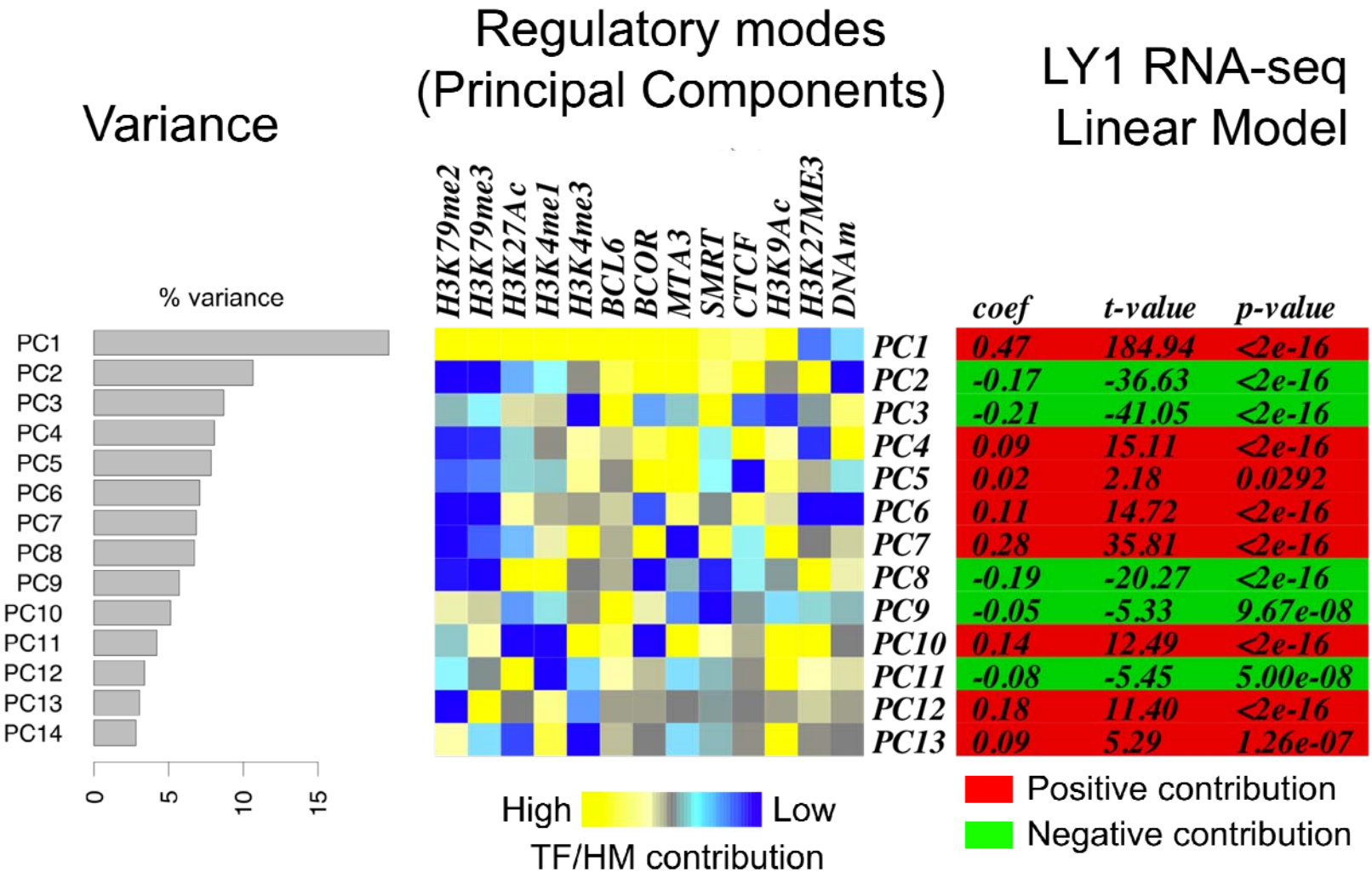
Calculate p-value using t-distribution with n-p degrees of freedom

$$P(X \geq t_j)$$

```
attach(m)
fit <- lm(log(RPKM+1) ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 )
print(summary(fit))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.278557   0.025975  49.222  < 2e-16 ***
PC1          0.407975   0.006672  61.146  < 2e-16 ***
PC2          0.416035   0.013539  30.728  < 2e-16 ***
PC3          0.159544   0.012484  12.780  < 2e-16 ***
PC4          0.010444   0.011371   0.918    0.358
PC5          0.304946   0.014746  20.680  < 2e-16 ***
PC6         -0.081407   0.014956  -5.443 5.34e-08 ***
PC7          0.071195   0.014813   4.806 1.56e-06 ***
PC8          0.098751   0.015496   6.372 1.93e-10 ***
PC9         -0.366861   0.018438 -19.897  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.161 on 11760 degrees of freedom
  (3543 observations deleted due to missingness)
Multiple R-squared: 0.3812,     Adjusted R-squared: 0.3808
F-statistic: 805.1 on 9 and 11760 DF,  p-value: < 2.2e-16
```



$$RPKM_i = \beta_0 + \sum_{j=1}^{m} \beta_j PC_{ij}$$

BCL6 *In silico* knockdown

1. Set all BCL6 values to 0.0

2. Project new binding data into original PCs

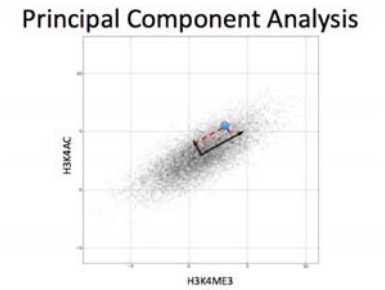
3. Predict RPKMs using original fitted model

4. Compare RPKMs to RPKMs predicted by original binding data and original model

Transcription factors / histone modifications

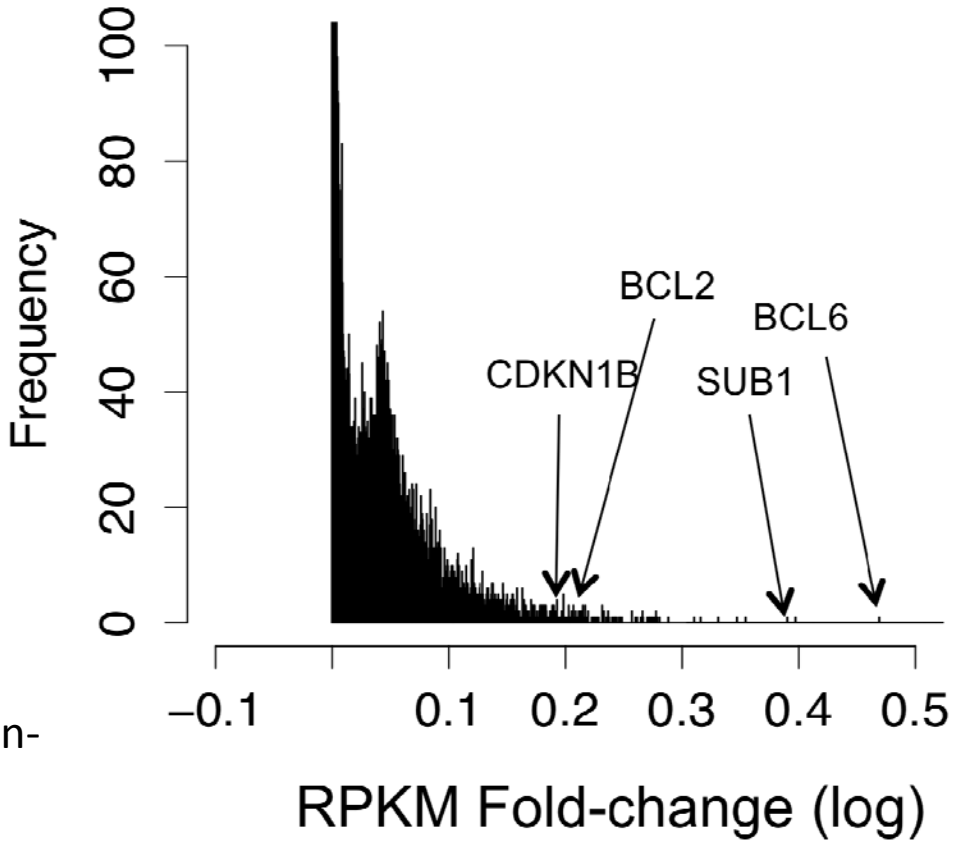
Transcripts

GENE	BCL6	MTA3	BCOR	K4ME1	K4ME3	K79ME2	K79ME3	K79AC	...
NM_018117	0	23.69	29.20	56.05	100.25	0.00	38.81	49.35	...
NM_001130845	0	203.7	373.2	113.4	58.08	104.7	148.5	117.3	...
NM_021107	0	0.00	0.05	41.03	222.2	18.53	48.24	87.66	...
NM_173803	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_006528	0	16.19	35.06	40.42	113.3	0.00	0.00	0.00	...
NM_182607	0	0.00	0.00	0.01	0.01	0.00	0.00	0.00	...
NM_017722	0	3.96	66.30	59.98	183.1	10.06	114.6	37.54	...
NM_018283	0	19.48	28.95	53.16	85.51	0.02	0.06	105.3	...
NM_014068	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_006228	0	0.19	0.58	8.33	0.87	0.00	0.01	1.19	...
NM_183377	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_002115	0	0.00	0.30	0.00	0.00	0.00	0.00	0.00	...
NM_004504	0	0.00	22.70	49.26	64.23	20.06	122.1	7.80	...
NM_004358	0	73.31	78.10	101.7	109.0	36.74	44.79	90.73	...
NM_022114	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_032125	0	0.32	23.41	57.10	157.6	49.68	121.2	29.71	...
NM_001011666	0	0.00	0.00	0.48	0.00	0.00	0.00	0.00	...
NM_018905	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_080746	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_001145155	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_001040167	0	31.89	209.1	5.17	40.27	0.00	0.00	0.03	...
NM_001144994	0	0.87	1.98	2.09	6.81	1.05	1.10	1.66	...
NM_017812	0	0.00	17.73	50.97	120.3	87.4	166.9	29.53	...



$$RPKM_i = \beta_0 + \sum_{j=1}^{m} \beta_j PC_{ij}$$

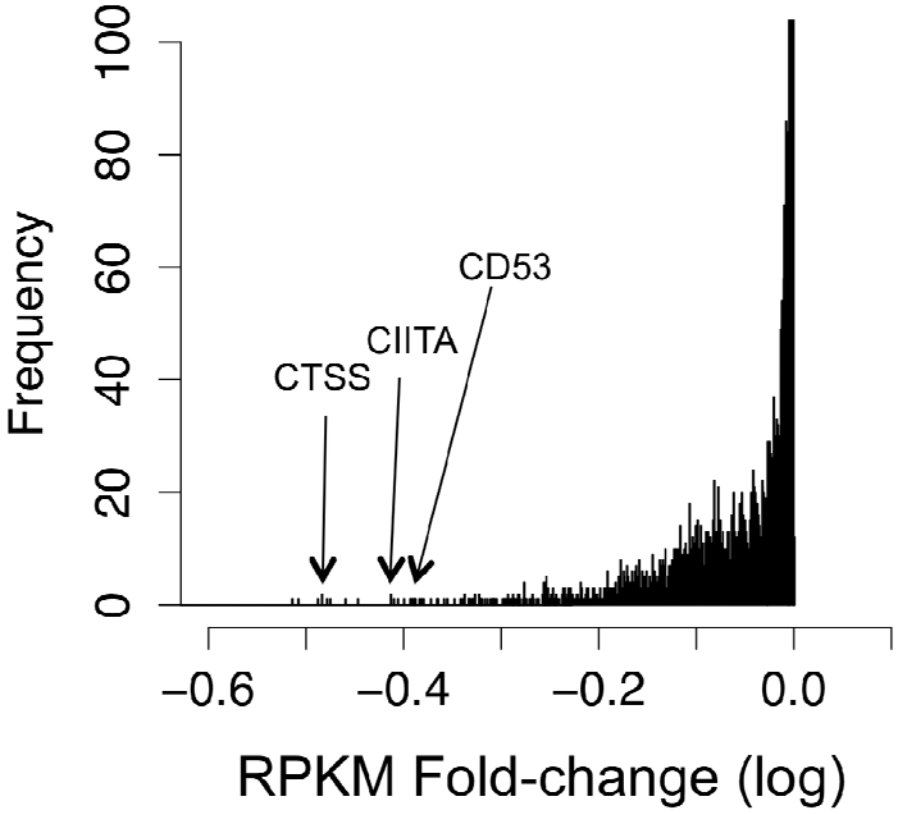
Simulated BCL6 knockdown predicts (correctly) that BCL6 is an obligate repressor

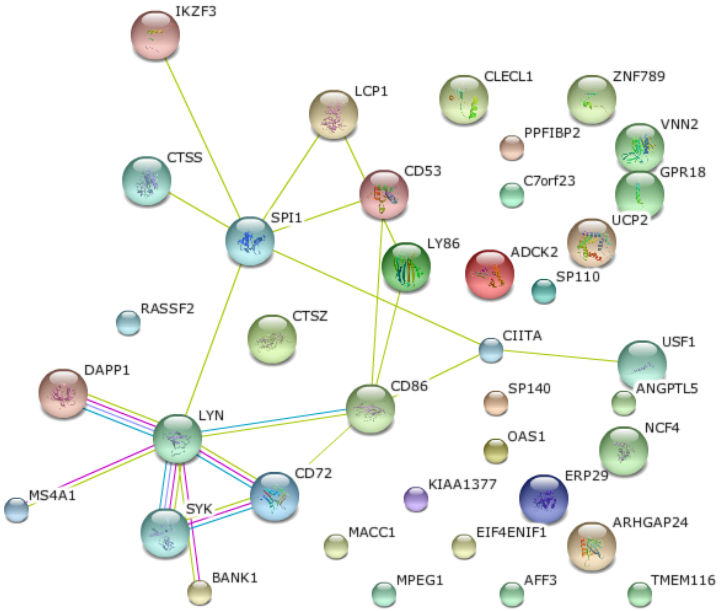


No genes are significantly down-regulated !

Top 250 up-regulated genes are enriched with genes with expression higher in NB compared to LY1 ($p<0.005$)

Simulated PU.1 knockdown (activator) rediscovers important PU.1 targets

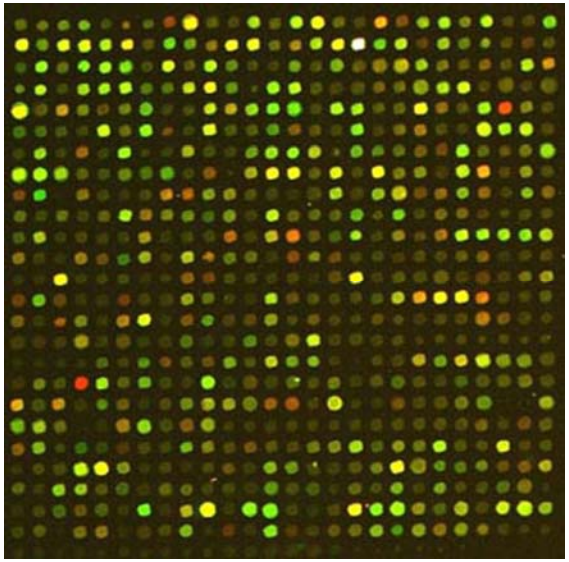


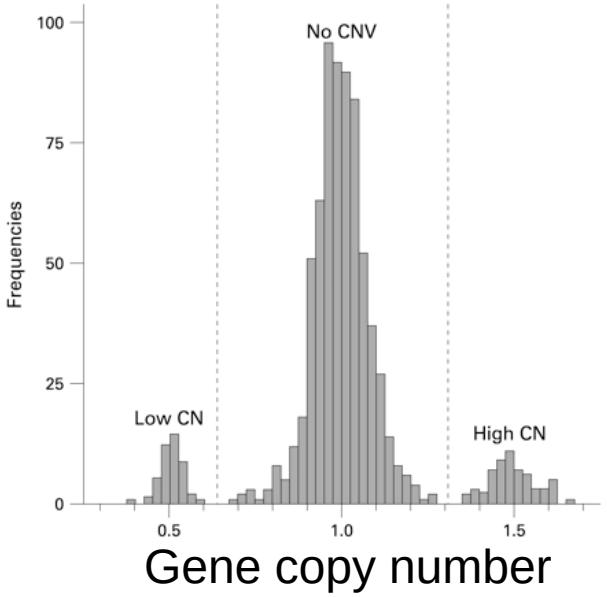


STRING analysis of top 40 down-regulated genes + PU.1

Do CNVs contribute to the model ?

$$RPKM_i = \beta_0 + \sum_{j=1}^{m} \beta_j PC_j + \beta_{m+1} CNV_i$$





We generated Affymetrix 6.0 SNP array data in LY1

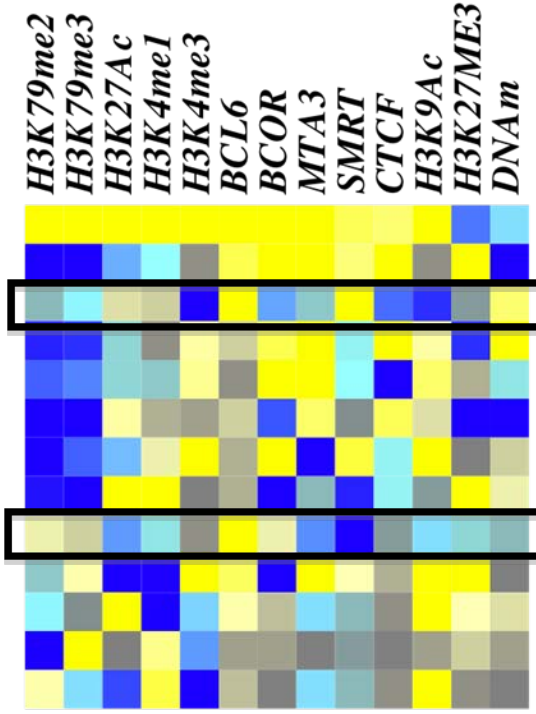
Yanwen Jiang, Huimin Geng

Ongoing work

- Using non-negative matrix factorization instead of PCA
- Integration of DNA looping
- Integration of post-transcriptional regulation (can we improve over 65%?)
- Experimental validation (knockdowns)

Can the binding patterns predict what will happen upon siRNA knockdown ?

$$\log(siBCL6/NT)_i = \beta_0 + \sum_{j=1}^{m} \beta_j PC_{ij}$$



Ly1 expression

siBCL6/NT 24h

	coef	t-value	p-value	coeff	t-value	p-value
PC1	0.47	184.94	<2e-16	-0.00	-2.30	0.0215
PC2	-0.17	-36.63	<2e-16	-0.00	-1.68	0.0923
PC3	-0.21	-41.05	<2e-16	0.04	11.63	<2e-16
PC4	0.09	15.11	<2e-16	0.01	2.25	0.0248
PC5	0.02	2.18	0.0292	0.01	2.06	0.0399
PC6	0.11	14.72	<2e-16	0.01	1.80	0.0715
PC7	0.28	35.81	<2e-16	0.04	9.18	<2e-16
PC8	-0.19	-20.27	<2e-16	0.03	4.61	4.02e-06
PC9	-0.05	-5.33	9.67e-08	0.09	15.02	<2e-16
PC10	0.14	12.49	<2e-16	0.04	4.95	7.49e-07
PC11	-0.08	-5.45	5.00e-08	-0.11	-11.83	<2e-16
PC12	0.18	11.40	<2e-16	-0.02	-2.06	0.0390
PC13	0.09	5.29	1.26e-07	-0.01	-0.70	0.4851

BCL6 *In silico* knockdown

1. Set all BCL6 values to 0.0

2. Project new binding data into original PCs

3. Predict RPKMs using original fitted model

4. Compare RPKMs to RPKMs predicted by original binding data and original model

GENE	BCL6	MTA3	BCOR	K4ME1	K4ME3	K79ME2	K79ME3	K79AC	...
NM_018117	17.54	23.69	29.20	56.05	100.25	0.00	38.81	49.35	...
NM_001130845	126.7	203.7	373.2	113.4	58.08	104.7	148.5	117.3	...
NM_021107	0.00	0.00	0.05	41.03	222.2	18.53	48.24	87.66	...
NM_173803	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_006528	0.00	16.19	35.06	40.42	113.3	0.00	0.00	0.00	...
NM_182607	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	...
NM_017722	89.05	3.96	66.30	59.98	183.1	10.06	114.6	37.54	...
NM_018283	0.00	19.48	28.95	53.16	85.51	0.02	0.06	105.3	...
NM_014068	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_006228	16.98	0.19	0.58	8.33	0.87	0.00	0.01	1.19	...
NM_183377	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_002115	0.09	0.00	0.30	0.00	0.00	0.00	0.00	0.00	...
NM_004504	2.38	0.00	22.70	49.26	64.23	20.06	122.1	7.80	...
NM_004358	0.00	73.31	78.10	101.7	109.0	36.74	44.79	90.73	...
NM_022114	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_032125	0.00	0.32	23.41	57.10	157.6	49.68	121.2	29.71	...
NM_001011666	21.05	0.00	0.00	0.48	0.00	0.00	0.00	0.00	...
NM_018905	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_080746	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_001145155	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
NM_001040167	2.33	31.89	209.1	5.17	40.27	0.00	0.00	0.03	...
NM_001144994	0.00	0.87	1.98	2.09	6.81	1.05	1.10	1.66	...
NM_017812	31.99	0.00	17.73	50.97	120.3	87.4	166.9	29.53	...

… (~25,000 unique RefSeq promoters)

BCL6 *In silico* knockdown

1. Set all BCL6 values to 0.0

2. Project new binding data into original PCs

3. Predict RPKMs using original fitted model

4. Compare RPKMs to RPKMs predicted by original binding data and original model