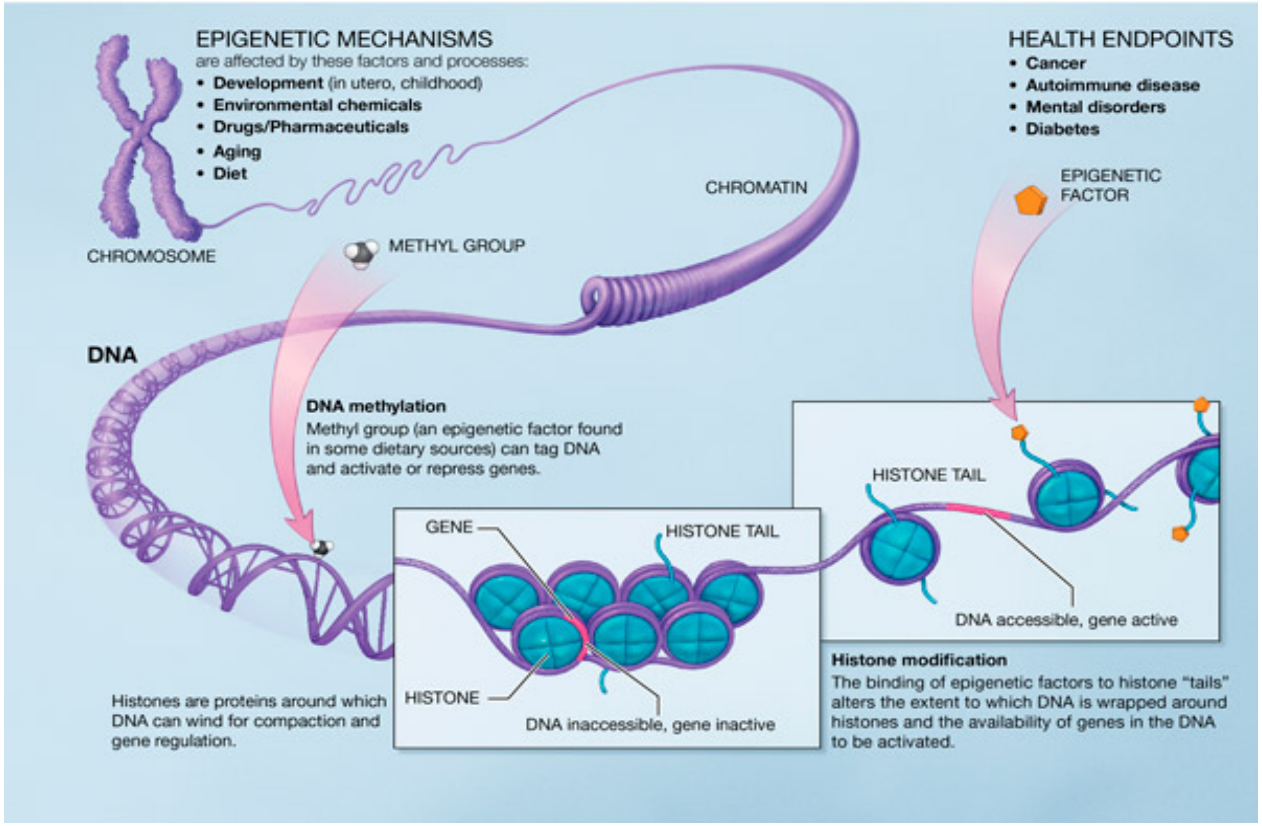
Other applications of second-generation sequencing

Review

- We have covered for second-generation sequencing:
  - Overview technologies.
  - Data and statistical issues.
  - RNA-seq, ChIP-seq and their analysis strategies.
- Today we will introduce some other applications of sequencing, mainly
  - For DNA methylation: bisulfite sequencing (BS-seq).
  - Hi-C for 3-dimensional chromatin structures.

Epigenetics

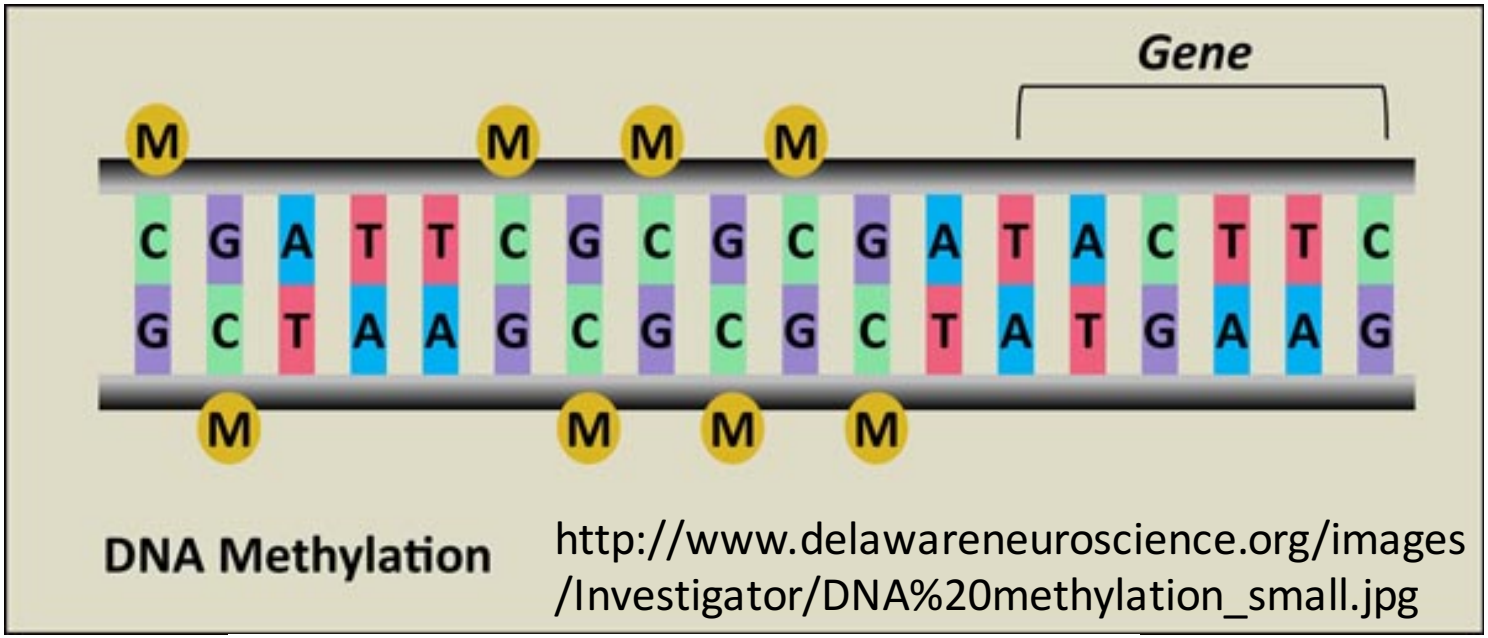
Non-DNA sequence related, heritable mechanisms to control gene expressions. Examples: DNA methylation, histone modifications.



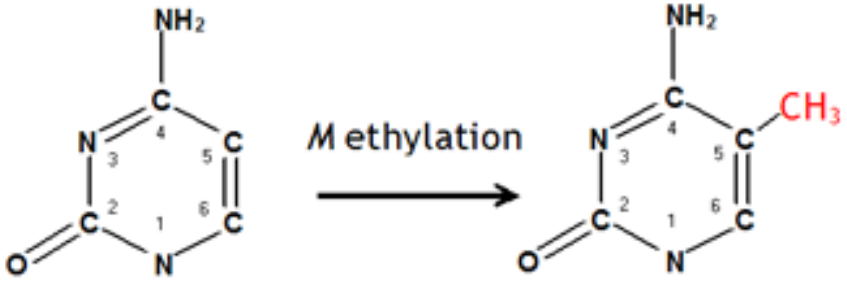
wikipedia

DNA methylation

- An epigenetic modification of the DNA sequence.
- Involves adding a methyl group to cytosine.
- Primarily happens at the CpG sites (when C and G are at consecutive bases), although non-CG methylation exists.
- Mostly detected in higher organisms:
  - In human genome, most CpG sites are fully methylated(over 90%) except at CpG island where the methylation level is minimal.
  - Methylation are detected in some plants, insects and bacteria, but the levels are low.



http://www.delawareneuroscience.org/images/Investigator/DNA%20methylation_small.jpg

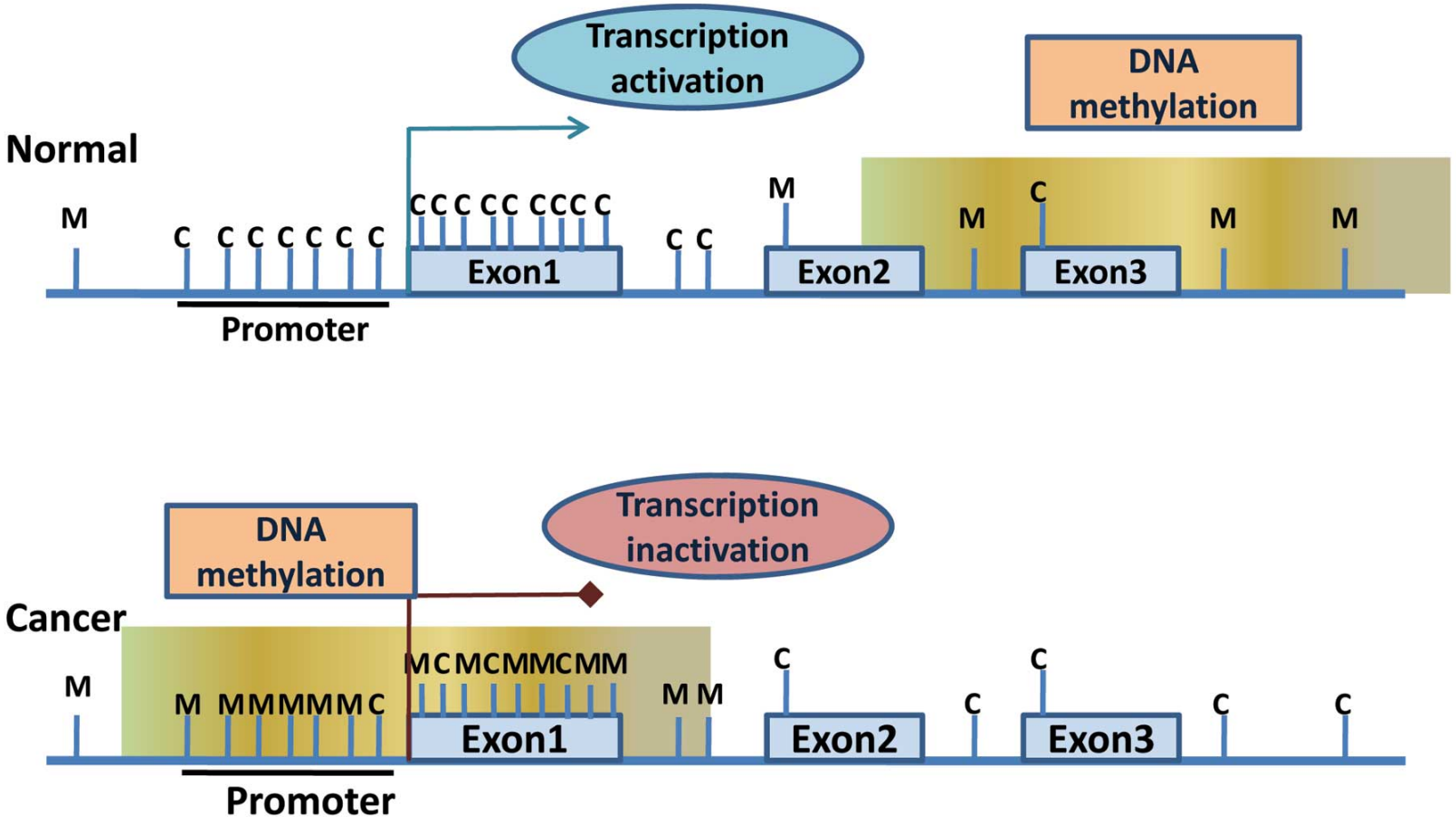


http://www.bio.miami.edu/dana/pix/cytosine.bmp

Function of DNA methylation

- Important in gene regulation: methylation at TSS suppress gene expression.
- Play crucial role in development and differentiation: help cells establish identity.
- Believed to be interacting with environment exposures. So it is being used to explain GxE interactions.
- Often referred to as the "5th base".
- Recent researches found different types of methylation, e.g., hydroxyl methylation.

DNA Methylation regulates gene expression



http://www.spandidos-publications.com/article_images/or/31/2/OR-31-02-0523-g00.jpg

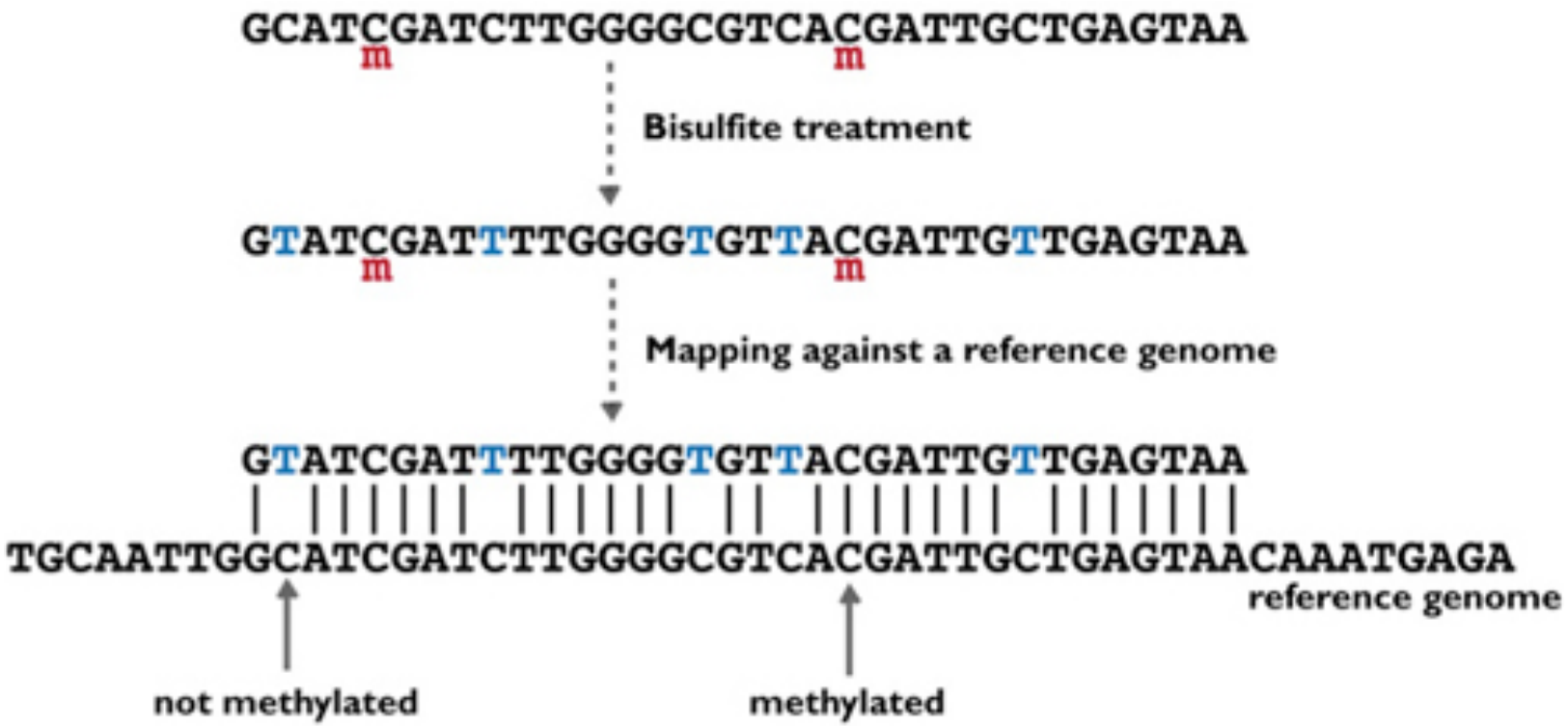
Detecting DNA methylation

- Capture based: MeDIP-seq (Methylated DNA immunoprecipitation followed by sequencing).
  - Same as ChIP-seq, but use antibody against methylated DNA.
  - Analysis methods are the same as ChIP-seq.
  - Resolution is low: can roughly quantify the amount of DNA methylation in a few hundred bps.
- Bisulfite sequencing (BS-seq): bisulfite conversion of DNA followed by sequencing:
  - Base pair resolution: measures the methylation status of each nucleotide.

Bisulfite sequencing

- Technology in a nutshell:
  - First treat the DNA with bisulfite. As a result,
    - Unmethylated C will be turned into T.
    - Methylated C will be protected and still be C.
    - No change for other bases.
  - Amplify, then sequence the treated DNA segments.
    - The mismatches between C-T measures the methylation strength.
- Raw data: sequence reads, but not exactly from the reference genome.

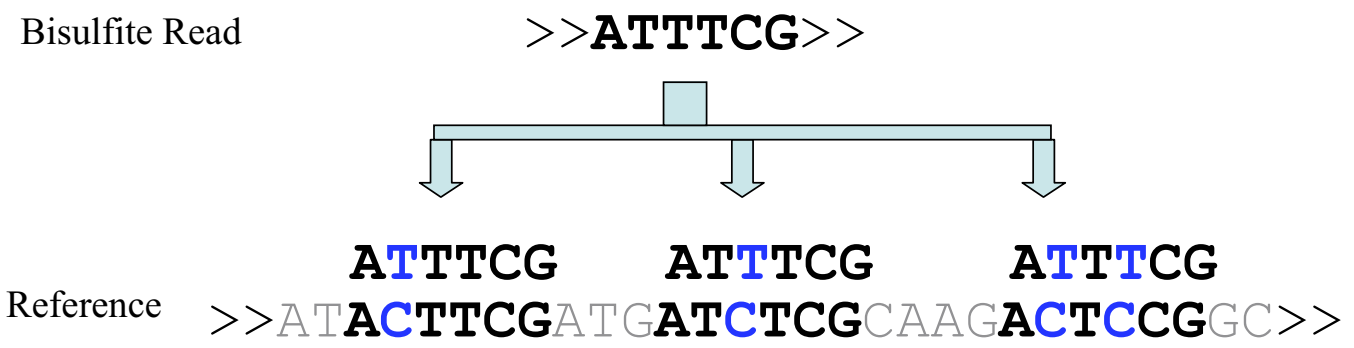
Bisulfite Sequencing



http://www.ecseq.com/services/EPIseq.html

Alignment of BS-seq

- The reads from BS-seq cannot be directly aligned to the reference genome.
  - There are four different strands after bisulfite treatment and PCR.
  - T could be aligned to T or C.
  - The search space for alignment is bigger.



Alignment strategy

- Use existing alignment software (eg, bowtie) as is:
  - Problem: C-T mismatches make some reads can't be aligned.
- Naïve method: change both the reference and reads to make all C's to T's, then align.
  - Problem: create other mismatches.
- Better ideas:
  - Consider the methylation status during alignment: create multiple versions of the reference "seed" (there will be four sets of references at each locations containing a C).
- Clever implementations needed.

Alignment tools

- See a list of available BS-seq aligner at http://www.mi.fu-berlin.de/w/ABI/ExistingBisulfiteMappers.
- Performances wise, they are usually slower:
  - in the rate of a few hundred reads per second.

Data after alignments

- Special software needed to process the alignment file.
- At each C position, report the total number of reads covering that site, and the number of reads with T:

```
chr1301087422  18
chr1301089431  27
chr1301092212  10
chr130109577   6
chr130109716   6
chr130110257   5
```

- These are usually inputs for downstream BS-seq analysis.

BS-seq data analysis

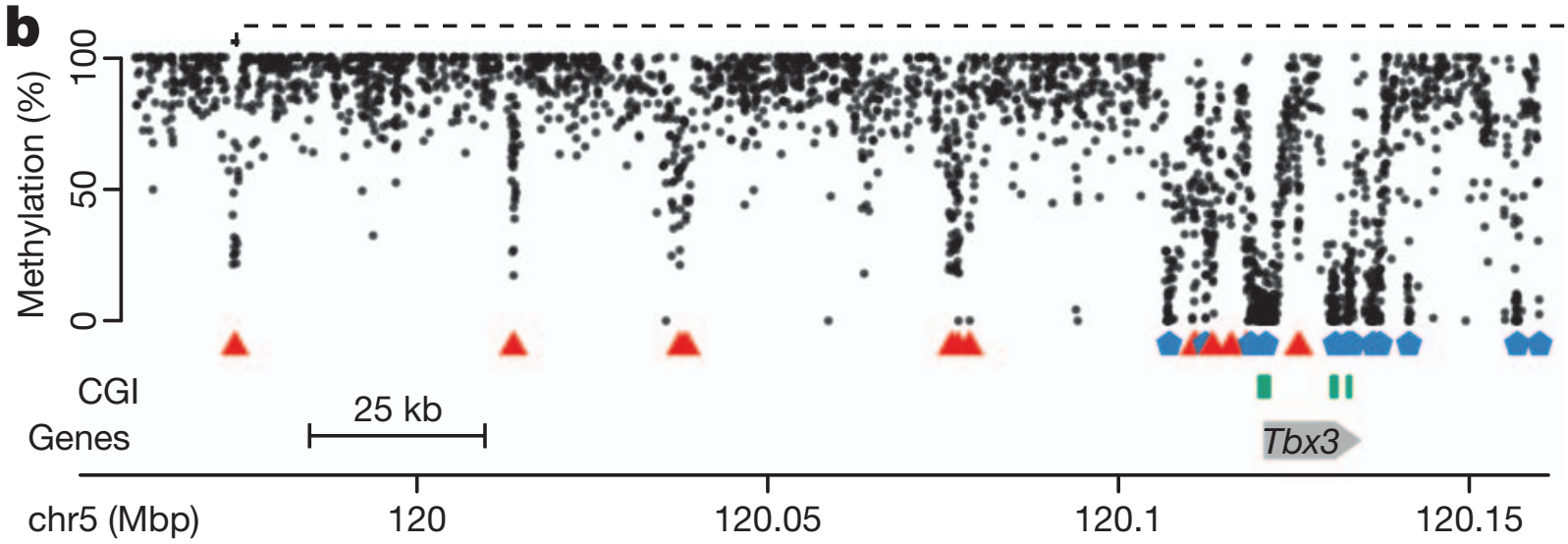
- Compared with ChIP-seq and RNA-seq, still in relatively early stage.
- Questions include:
  - Single dataset analysis:
    - Segment genome according to methylation status.
  - Comparison of multiple datasets:
    - Differential methylation (DM) analysis.

Single BS-seq dataset analysis

- Detecting the methylation loci/regions:
  - Estimate “methylation density” (percentage of cells have methylation) at each C position, which is simply #methy/#total at each CpG site, but:
    - Background error rates need to be considered.
    - Spatial correlation among nearby CpG sites can be utilized to improve estimation.
  - Methylated regions (or states) can be determined by smoothing based method (e.g., moving average, HMM) using the estimated percentage as input.

An HMM approach

- Stadler *et al.* (2012) ***Nature***:
  - Using the estimated percentages as input to fit a 3-state HMM: FMR, LMR and UMR.



Smoothing method

- Can directly smooth the percentages, but that doesn't consider the uncertainty in percentage estimates.
- A better approach: BSmooth model (Hansen *et al.* 2012 ***Genome Biology***).
  - Assumes the true methylation level is a smooth curve of genomic coordinates.
  - The observed counts follow a binomial distribution.

BSmooth smoothing

- Notations at position j:
  - N_j, M_j: total/methylated reads.
  - π_j: underlying true methylation level.
  - l_j: location.
- Model:

$$M_j \sim \mathrm{Bin}(N_j, \pi_j)$$

$$\log(\pi_j / (1 - \pi_j)) = \beta_0 + \beta_1 l_j + \beta_2 l_j^2$$

- Fitting: weighted glm in each 2kb window, where the weights depend on the variances of estimated π_j.

Bsmooth Bioconductor package: bsseq

- Mainly provide functions for smoothing and some visualization.
- Implemented in parallel computing environment to speed up the calculation.

```
M <- matrix(0:8, 3, 3)
Cov <- matrix(1:9, 3, 3)
BS1 <- BSseq(chr = c("chr1", "chr2", "chr1"),
             pos = c(1,2,3), M = M, Cov = Cov,
             sampleNames = c("A","B", "C"))
BS1 <- BSmooth(BS1)
```

Differential methylation analysis

- Comparison of methylation profiles under different biological conditions is of great interests.
  - Results from such analysis are: differentially methylated loci (DML) or regions (DMR).
- Strategy to detect DML:
  - Hypothesis testing at each CpG site.
- Strategy to detect DMR:
  - Need to combine data from nearby CpG sites because of the spatial correlation.

DML detection based on 2x2 table

- At each CpG site, summarize the counts from two samples into a 2x2 table:

	Total	Methylated
Sample 1	40	2
Sample 2	25	19

- Chi-square or Fisher's exact test can be applied.
- Bsseq has function `fisherTests` for this:

```
fisherTests(BSobj, group1, group2)
```

Wald-test based

- Can handle data with replicates.
- The key is to estimate within group variances.
- BSmooth approach (for two group comparison):
  - Denote the group assignment for ith sample by X_i.
  - Number of replicates in two groups are n_1 and n_2.
  - Frame the estimated values of into a two-group testing framework: $\pi_{ij}=a(l_j)+ b(l_j)X_i+\varepsilon_{i,j,}\ \varepsilon_{i,j}\sim N(0, \sigma_j^2)$.
  - Use SAM-alike method to estimate σ_j^2, then do Wald test.

Shrinkage based method
(Feng et al. 2014, NAR)

- Similar to that in RNA-seq DE analysis, the BS-seq data can be modeled as Beta-binomial distribution:

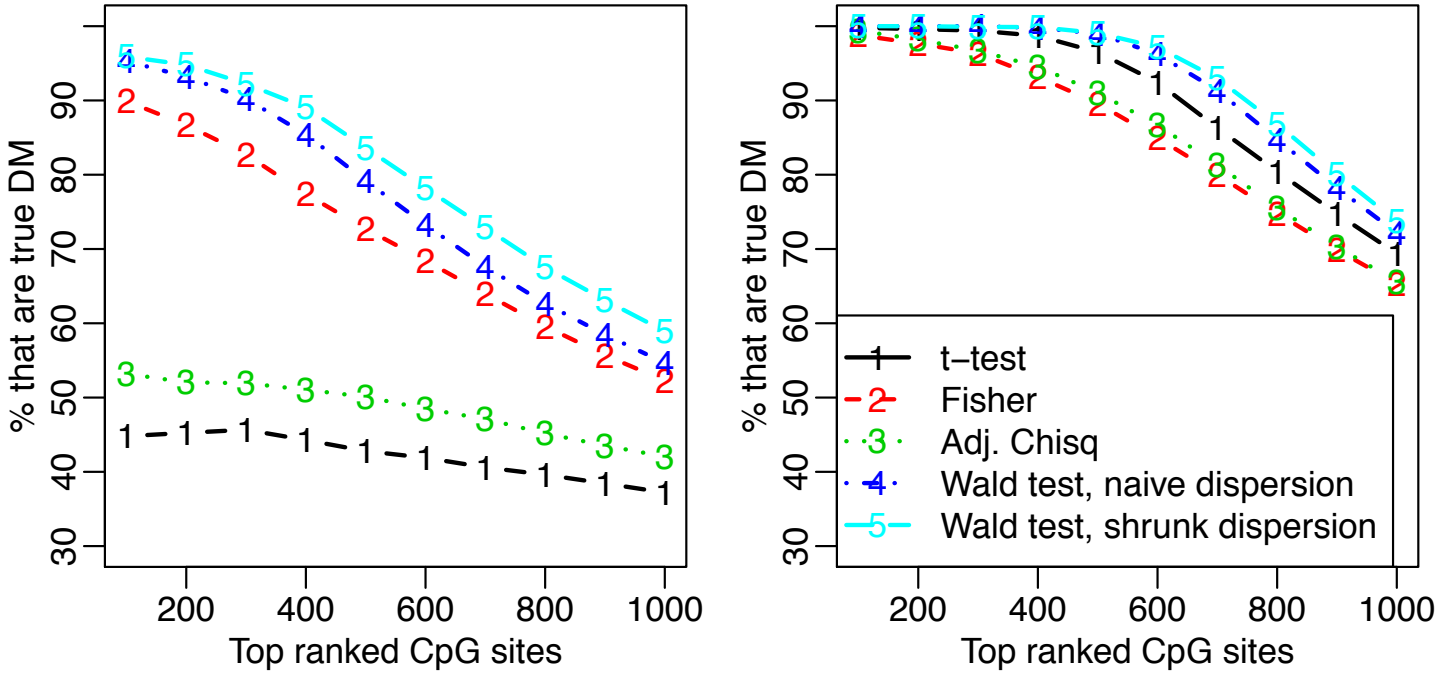
$$X_{ijk}|p_{ijk}, N_{ijk} \sim Binomial\ (N_{ijk}, p_{ijk})$$

$$p_{ijk} \sim Beta(\mu_{ij}, \emptyset_{ij})$$

- Beta distribution is parameterized by mean and dispersion, and impose a log-normal prior on dispersions. $\emptyset_{ij} \sim lognormal\ (m_{0j}, r_{oj}^2)$
- Wald test procedure can be derived.

Simulation results

- The Wald test with shrunk dispersion performs favorably compared with other methods.



Things to consider in DMR calling

- Coverage depth:
  - Should one filter out sites with shallower coverage?
- With biological replicates:
  - CpG specific biological variances.
  - Small sample estimate of the variance.
- Spatial correlation of methylation levels among nearby CpG sites.
  - Is smoothing appropriate?
  - What is data has low spatial correlation, like in 5hmC.

Existing methods for DML/DMR detection

- BSmooth (Hansen *et al.* 2012, *GB*):
  - Smoothing, then take the smoothed values and run two-group t-test.
- MethylKit(Akalin *et al.* 2012, *GB*):
  - Logsitic regression or Fisher's exact test.
  - Recently implemented DSS Wald test approach.
- BiSeq (Hebestreit *et al.* 2013, *Bioinformatics*):
  - Smoothing, then take the smoothed value and run beta glm.
- DSS (Feng et al. 2014, NAR):
  - Based on beta-binomial model. Empirical Bayesian estimate of dispersions, and Wald test.
  - Spatial correlations are ignored

- MOABS (Sun *et al.* 2014, *GB*):
  - Based on beta-binomial model to define *CDIF*, the lower bound of CI for methylation difference in two groups.
  - Spatial correlations are ignored.
- methylSig (Hebestreit et al. 2014, Bioinformatics)
  - Based on beta-binomial model. MLE based method to estimate dispersion.
  - Likelihood ratio test.
- DSS-single (Wu et al. 2015, NAR)
  - Works for single replicated data, use nearby CpG sites are "pseudo-replicates".
- RADMeth (Dolzhenko et al. 2014, BMC Bioinformatics)
  - Based on beta-binomial GLM, works for multiple factor design.

Useful bioc packages - bsseq

- First create `BSseq` objects
- Use `BSmooth` function to smooth.
- `fisherTests` performs Fisher's exact test, if there's no replicate.
- `BSmooth.tstat` performs t-test with replicates.
- `dmrFinder` calls DMRs based on BSmooth.tstat results.

```
BSobj = BSmooth(BSobj)
dmlTest=fisherTests(BSobj, group1=c("C1", "C2","C3"),
                    group2=c("N1","N2","N3"))
dmr <- dmrFinder(dmlTest)
```

Useful bioc packages - DSS

- Input data has the same format as `bsseq`.
- `DMLtest` performs Wald test at each CpG.
- `callDML/callDMR` calls DML or DMR.
- More options in DML/DMR calling.

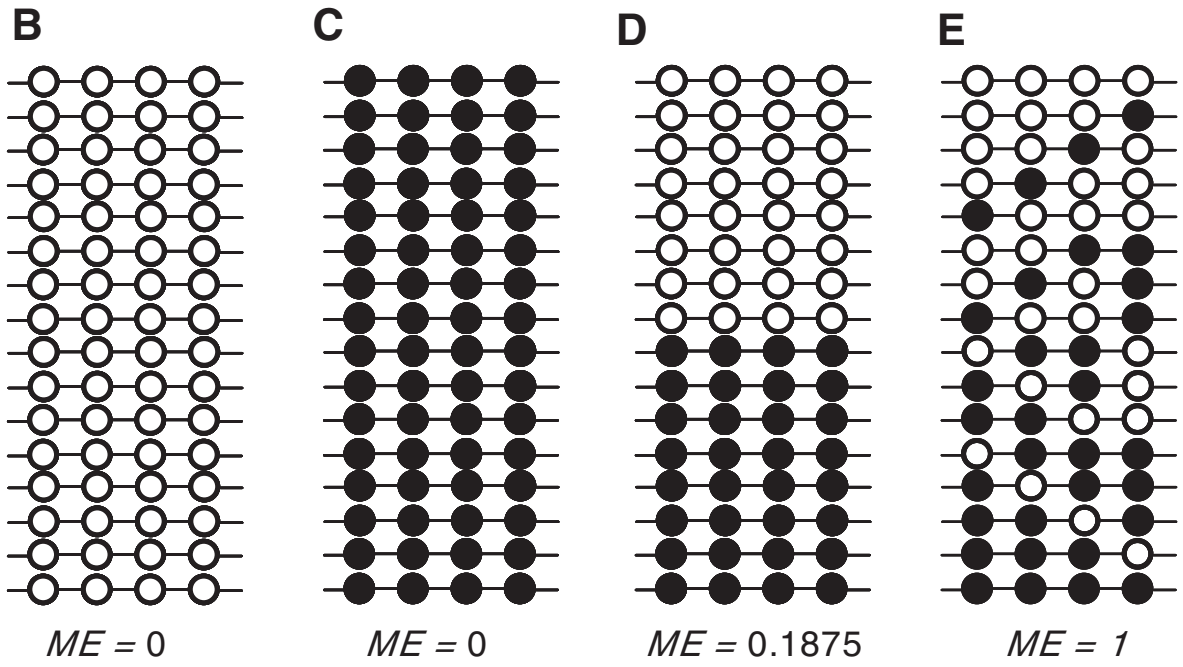
```
dmlTest <- DMLtest(BSobj, group1=c("C1", "C2", "C3"),
                   group2=c("N1","N2","N3"),
                   smoothing=TRUE, smoothing.span=500)
dmrs <- callDMR(dmlTest)
```

Another paradigm – single read BS-seq analysis

- So far we have focused on “marginal” methylation levels (aggregated information from all reads).
- Sometimes data at each single read provide additional information.
- Useful reads:
  - Xie et al. (2011) NAR.
  - Landan et al. (2012) Nat. Genetics.

Single read information

- Methylation entropy or polymorphism.

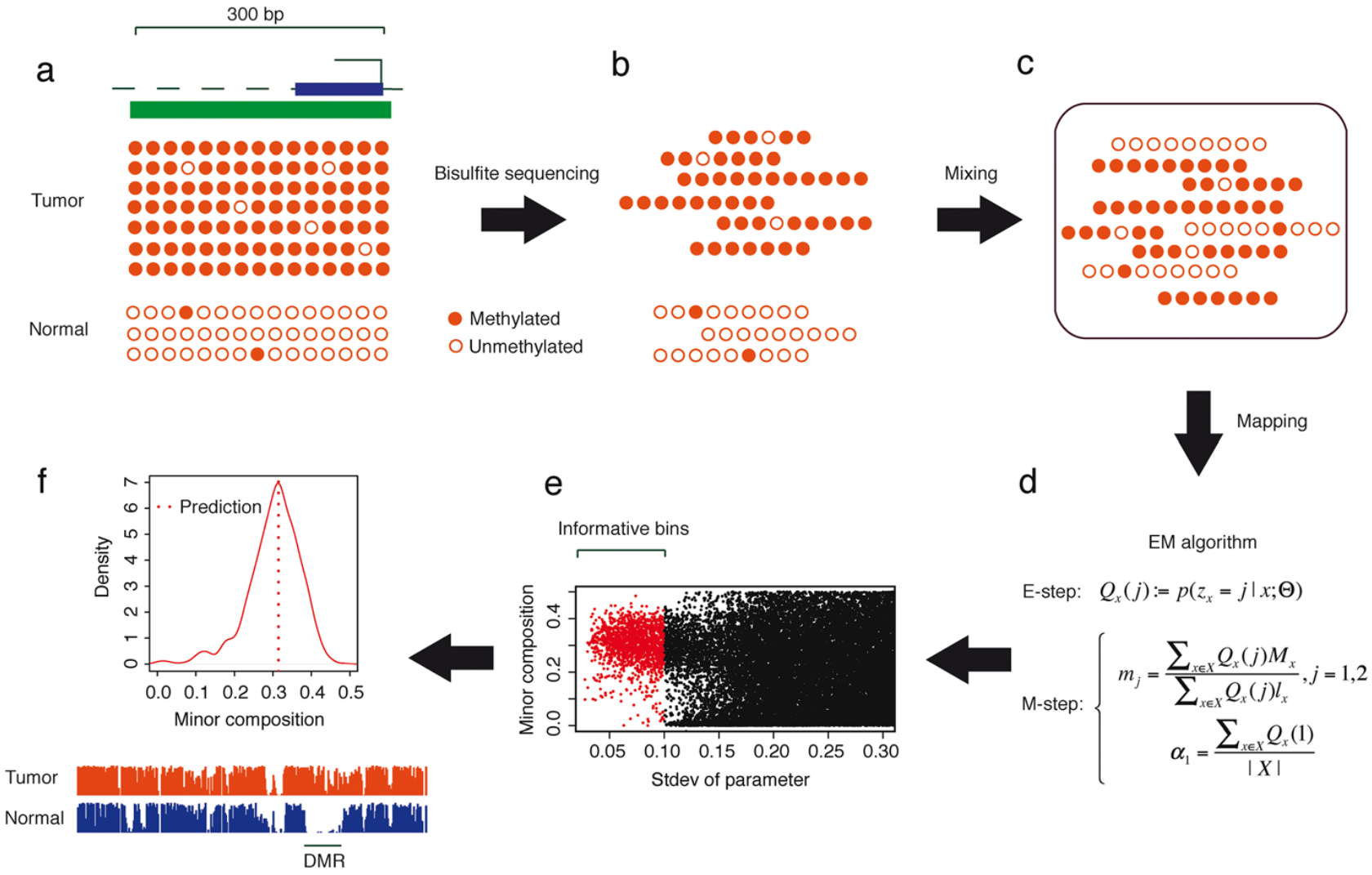


Xie et al. (2011) NAR

What single read tells us

- Comparison of methyl-entropy/polymorphism among different samples.
- Sample deconvolution
  - Zheng et al. (2014) GB: MethylPurify
  - estimate the proportion of cell types in a mixed sample (such as cancer), as well as calling DMRs.

MethylPurify



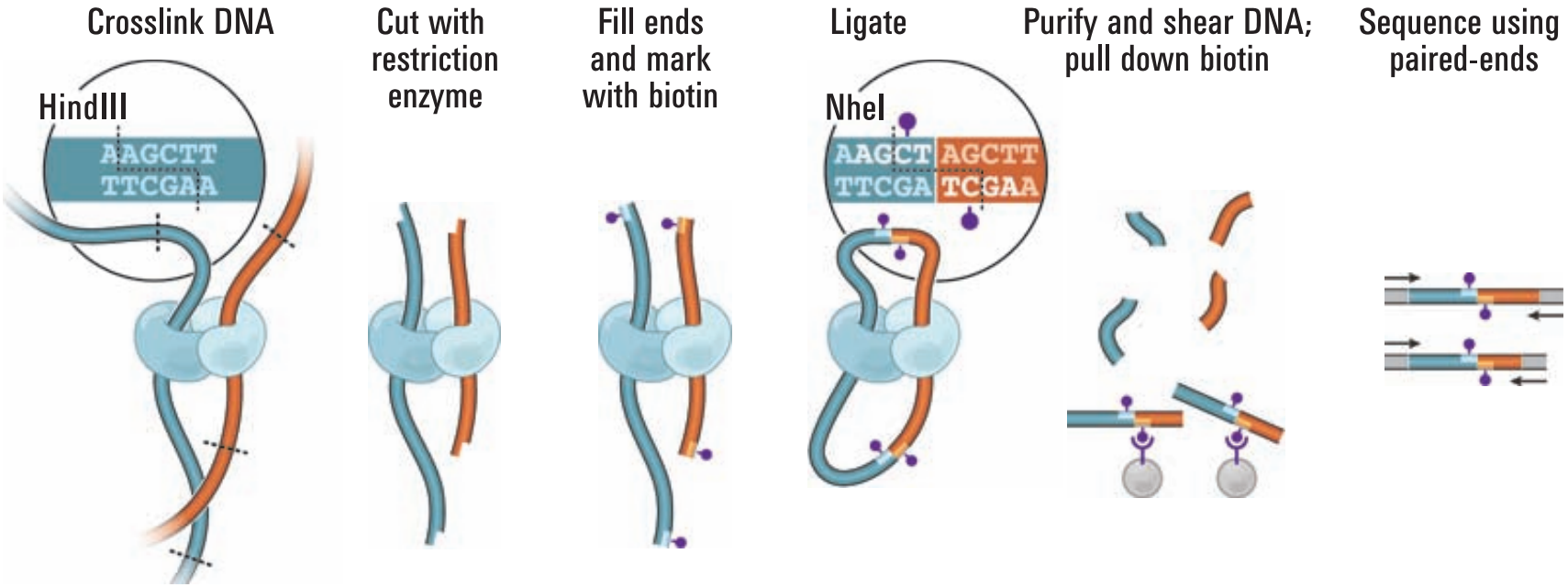
Conclusion on BS-seq analyses

- Careful in alignments.
- Data modeling is different from ChIP/RNA-seq: Poisson/NB vs. Binomial models.
- DMR calling needs to consider spatial correlation, coverage and biological variances.
- Single read analysis could be very useful.
- A lot of room for method development.

Detecting long-range interactions

- So far we have assumed the genome is a long line.
- In reality, chromosomes fold into complicated structures in nucleus. Implications:
  - Genomic loci far away on chromosome could be close spatially due to chromosome folding.
  - This is important for studying gene regulatory mechanisms, e.g., detecting enhancers.
- Traditional lower throughput methods:
  - 3C: Chromosome Conformation Capture.
  - 5C: Carbon-Copy Chromosome Conformation Capture.
- High-throughput: Hi-C

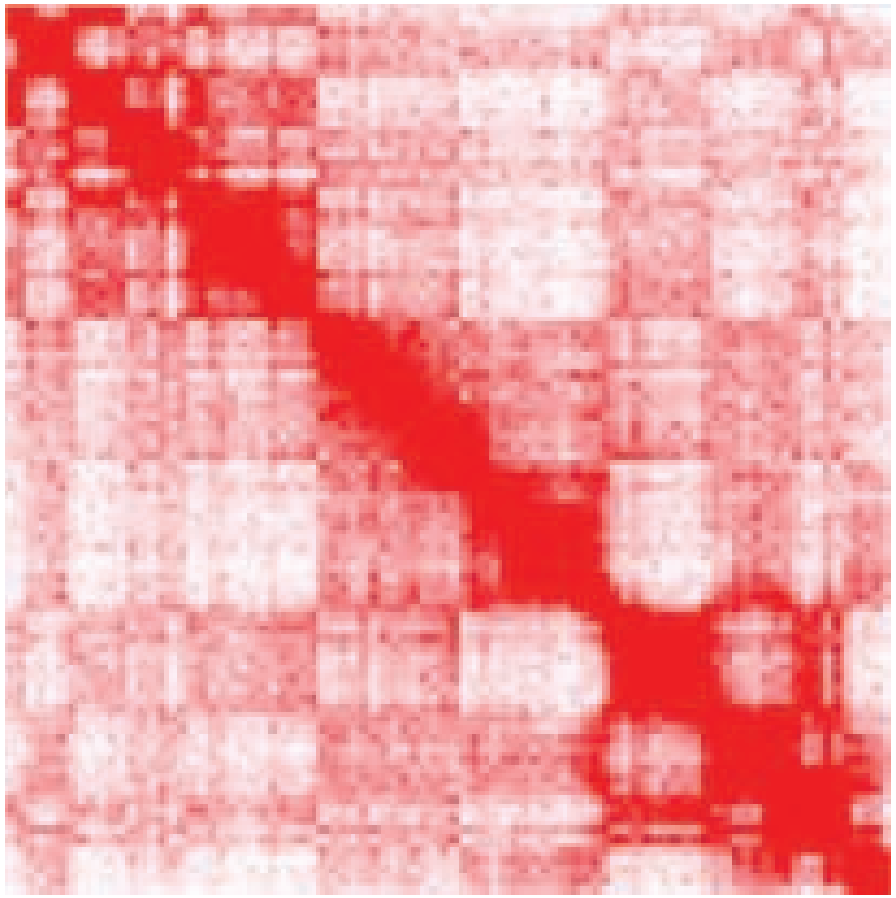
Hi-C experimental procedures



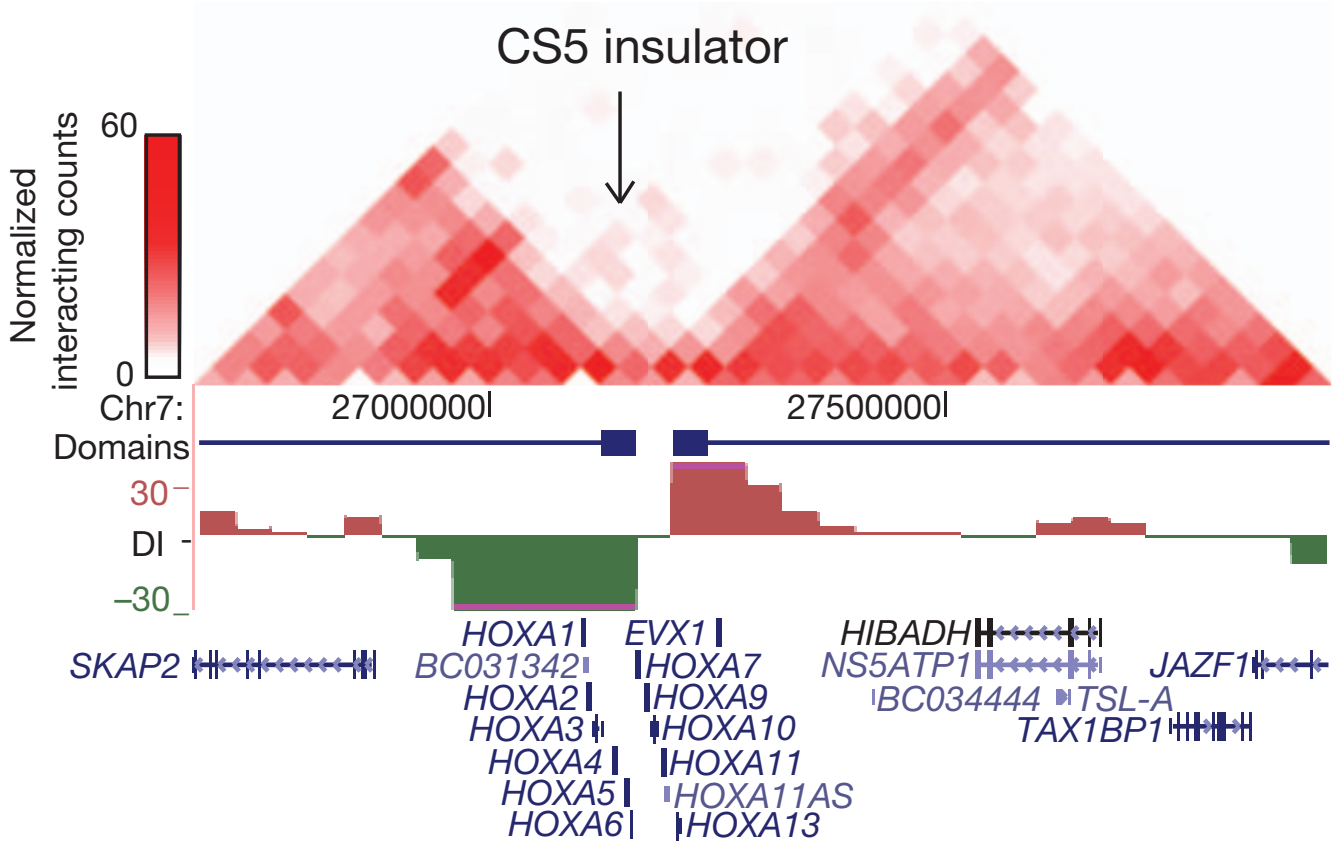
Hi-C data

- Paired end sequencing, each pair is for a pair of interacting regions.
- Usually summarized the counts into a 2D matrix:
  - First cut genome into N equal sized bins (size depends on sequence depth).
  - Summarize the read counts into NxN matrix. The element (*i, j*) represents the number of pairs with one end from the ith window and the other end from the jth window.
  - The counts represent the strength of interaction.
  - Usually the numbers on diagonal are greater.

Visualize Hi-C data in a heatmap



Overlay with other 1-D data



Data analysis

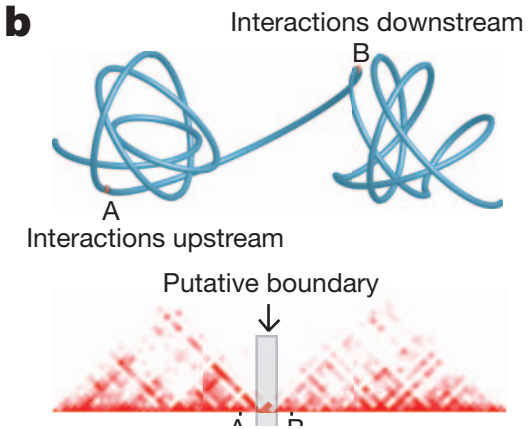
- Normalization.
- An easier one: defining domains (regions with higher level of self-interaction).
- Harder one: find long-range interaction.
- Others: infer 3D structures.
- Barely touched: comparison (differential domain).

Normalization

- Consider distance between read pairs, GC contents, mappability, etc. to create a baseline of counts (expected number of reads in each elements of the matrix).
- Subtract (or divide) the baseline from the observed counts to get the signals.
- A couple approaches:
  - Yaffe *et al.* (2011) ***Nature Genetics***: likelihood based.
  - Imakaev *et. al.* (2012) ***Nature Method***: assuming equal visibility at all loci and do median-polish type of correction (iteratively divide the row/column sums).
- Results: usually improved correlation among replicates.

Domain detection

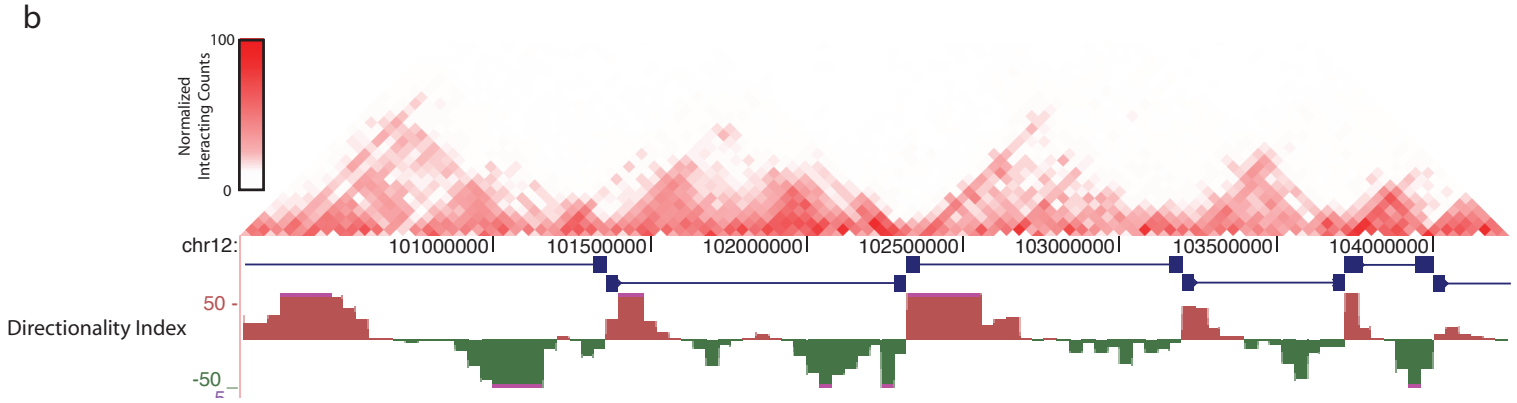
- The genome are organized into different “domains”.
- Can be seen as the blocks on diagonal of the heatmap.



- To detect, use the facts that the interactions are higher within a domain, and lower cross domains.
- Still an open statistical problem.

Domain detection by HMM (Dixon *et al.* 2012, *Nature*)

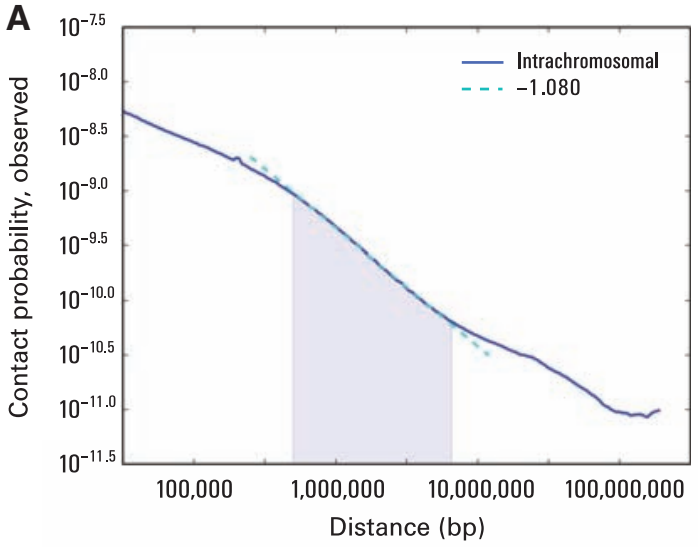
- Compute directionality index (DI).



- Run 2-state HMM on DI assuming Gaussian emission.
- Define domains based on HMM results: a domain starts from the beginning of a “up” region, and ends at the end of its next “down” region.

Detecting long-range interactions

- The interactions can be seen on the heatmap as bright, off-diagonal spots.
- A harder problem, partly because there are not enough reads.
- Still an open statistical problem. A simple method is a Poisson test, with the baseline rates computed from all data:



Comparison, e.g., differential interaction

- Barely touched (people still struggle with domains and interactions).
- Conceptually, one want to compare the interactions between different samples, e.g., locus A interacts with locus B in normal cell but not in cancer.
- For an element in the matrix, can we take the counts then use RNA-seq DE test methods?
  - No! Because the backgrounds could be different. This is similar to ChIP-seq differential binding problem.
  - Also neighboring elements in the matrix need to be combined to make inference (like in ChIP-seq, but combine in 2-D), so some (kernel) smoothing is needed.

Construct 3D structure

- BACH (Bayesian 3D constructor for Hi-C data), Hu et al. (2013) PloS CB
  - The read counts represent the physical distances between pairs of loci on the genome.
  - Given these distances the 3D structure can be estimated.
  - Based on a Poisson model, and with some constraints, the 3D coordinates of each bin on the genome can be estimated.
  - Estimation procedure is based on MCMC.

Conclusion on Hi-C data

- Technology to detect chromosomal interactions using sequencing.
- Usually requires more reads.
- Still in very early infancy in terms of analysis methods. A lot of room for development.

A grand overview of the class

- The technologies and statistical methods for:
  - Gene expression microarrays and a little bit ChIP-chip.
  - Second-generation sequencing: ChIP-seq and RNA-seq.
- Bioconductor tools for analyzing genomic data, including:
  - Biostrings, BSgenome, GenomicRanges, GenomicFeatures for general genomic data.
  - A little bit of Rsamtools for sequencing data.
  - Several Bioc packages for DE/DM analyses in:
    - microarray: siggenes, limma.
    - RNA-seq: DESeq, edgeR, DSS.
    - BS-seq: bsseq, DSS
- Some software tools for analyzing sequence data:
  - bowtie: alignment.
  - samtools: for manipulating SAM/BAM files.