Other epigenomic sequencing

Mikhail Dozmorov

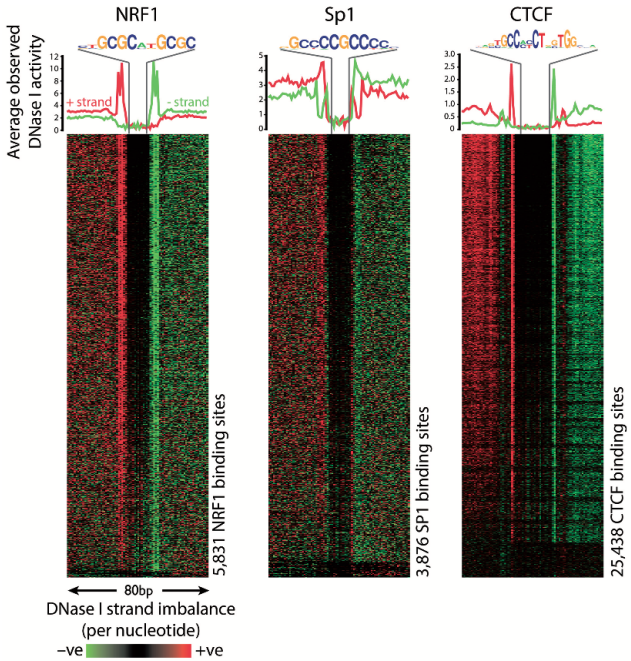
Spring 2018

Other "captured/targeted" sequencing technologies

- Enrich and then sequence selected genomic regions.
  - **MeDIP-seq**: measure methylated DNA.
  - **DNase-seq**: detect DNase I hypersensitive sites.
  - **FAIRE-seq**: detect open chromatin sites.
  - **Hi-C**: study 3D structure of chromatin conformation.
  - **GRO-seq**: map the position, amount and orientation of transcriptionally engaged RNA polymerases.
  - **Ribo-seq**: detect ribosome occupancy on mRNA. This is captured RNA-seq.

DNAse-seq

- A widely used approach in gene regulation studies uses DNase I as a tool to identify DNase I Hypersensitive Sites (DHSs) within chromatin
- DHSs represent open chromatin regions that are normally only accessible at sites of active regulatory elements such as transcriptional enhancers



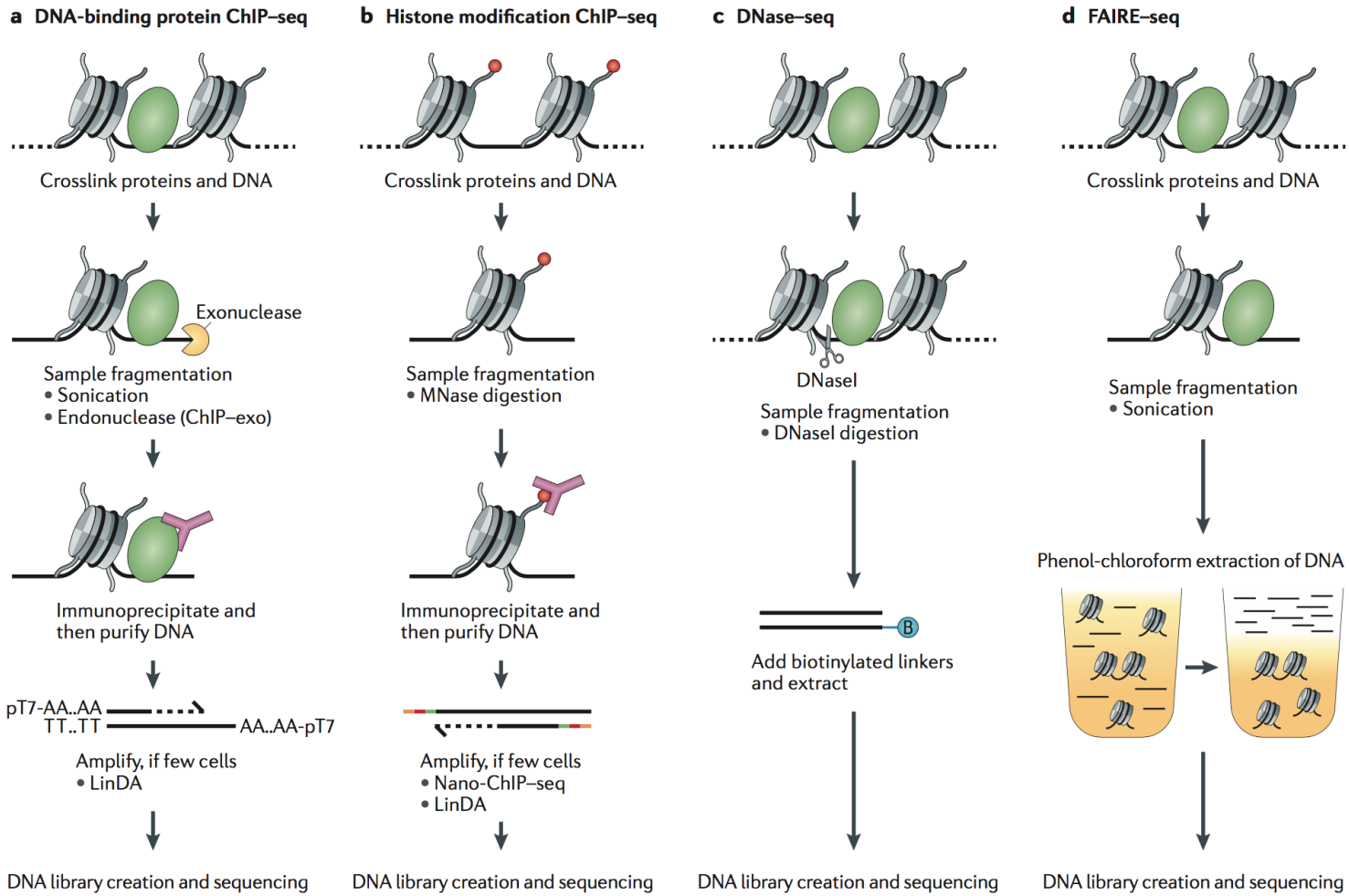
Cockerill,P.N. (2011) Structure and function of active chromatin and DNase I hypersensitive sites. FEBS J., 278, 2182–2210.

High-throughput chromatin organization techniques

TYPE OF ANALYSIS	2D STRUCTURE				3D STRUCTURE	
METHODOLOGY	**DNase1/MNase (Micrococcal nuclease)**	**FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements)**	**ChIP (Chromatin immunoprecipitation)**	**ATAC (Assay for Transposase-Accessible Chromatin)**	**HiC**	**ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)**
Sample input	Fresh/frozen tissues			Fresh cell lines	Fresh tissues	
Fragmentation	Endo- or exonuclease	Sonication	Endo- or exonuclease, sonication	No fragmentation step. Tn5 transposase tagmentation.	Endonuclease	
Enrichment	Size selection	Phenol-chloroform extraction	Antibody (Ab) capture	Specific adapters (tagmentation)	Biotin-streptavidin (fragments are ligated to biotin)	Antibody
Control input	Total genomic DNA with no enrichment					
Amplification	PCR-based					
Sequencing	4-letter based genome					
Advantages	Modest high-resolution	Does not depend on restriction enzyme or buffer composition	Single protein information	Low-input of sample (50,000 cells)	Single protein information	
Disadvantages	High-input of sample (>1 million cells)	Low resolution	Depends on Ab quality	Requires intact nuclei	Depends on Ab quality	
Array-based technologies	DNase1-chip MNase-chip	-	ChIP-chip	-	C3, C4, C5	-
Sequence-based technologies	DNase-Seq MNase-Seq	FAIRE-Seq	ChIP-Seq	ATAC-Seq	C4, C5, HiC	ChiA-PET
References	[62-66]	[67]	[68-71]	[72-74]	[75-79]	[80]

Kagohara, Luciane T., Genevieve L. Stein-O'Brien, Dylan Kelley, Emily Flam, Heather C. Wick, Ludmila V. Danilova, Hariharan Easwaran, et al. "Epigenetic Regulation of Gene Expression in Cancer: Techniques, Resources and Analysis." Briefings in Functional Genomics, August 11, 2017. https://doi.org/10.1093/bfgp/elx018.

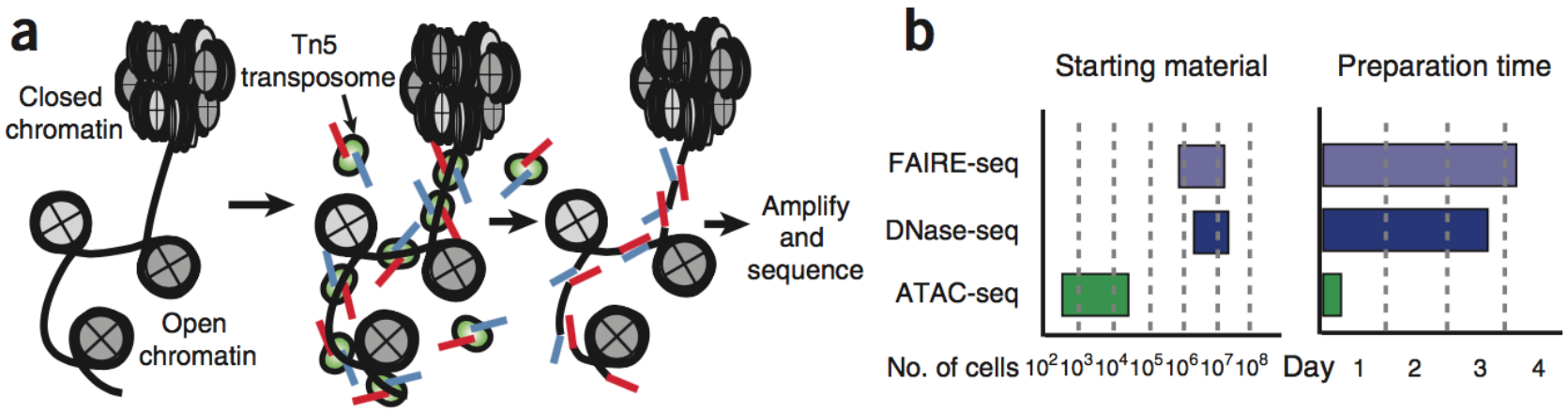
Comparison of experimental protocols



Furey, Terrence S. "ChIP–seq and beyond: New and Improved Methodologies to Detect and Characterize Protein–DNA Interactions." Nature Reviews Genetics 13, no. 12 (October 23, 2012): 840–52. doi:10.1038/nrg3306.

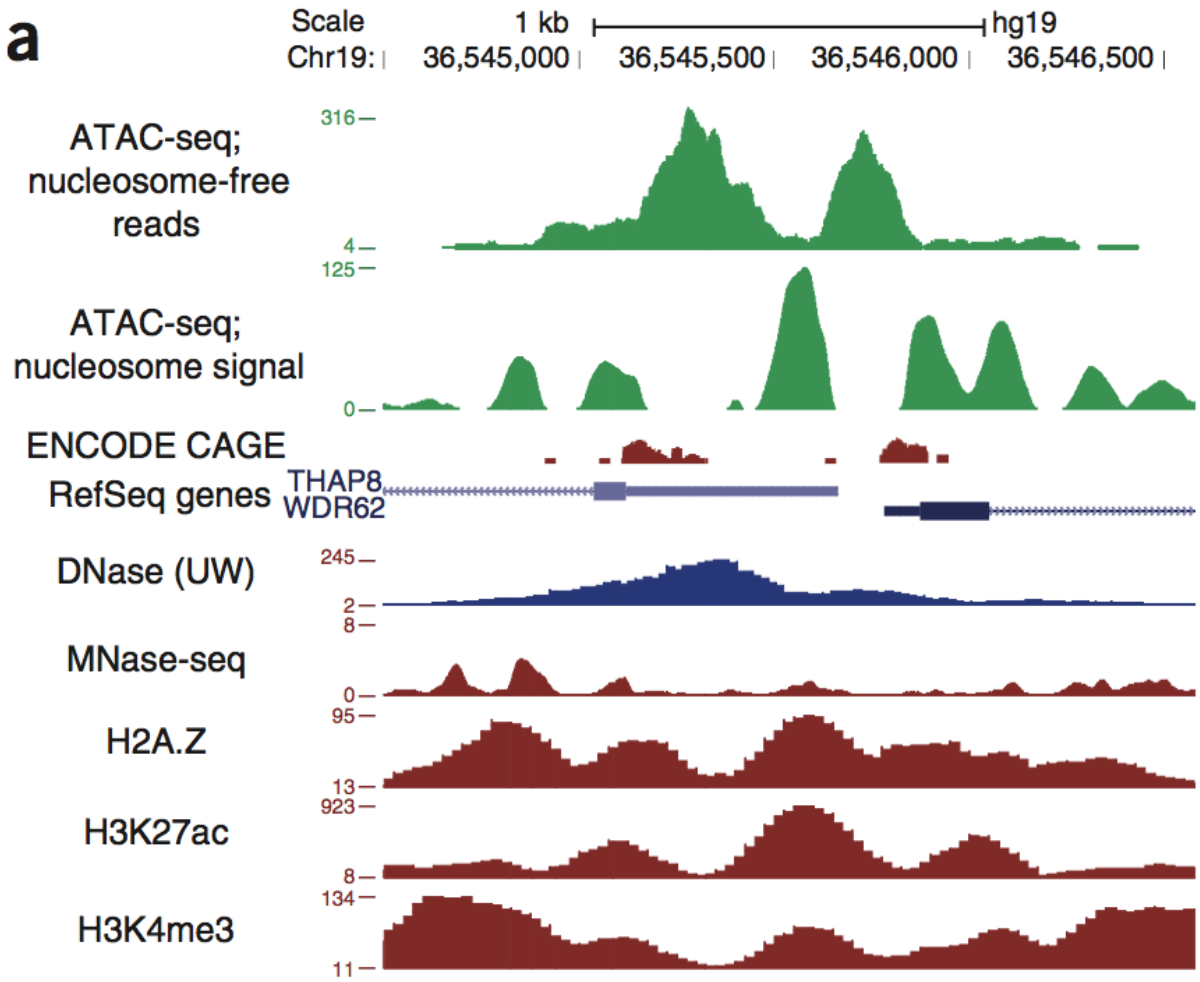
ATAC-seq: finding open chromatin regions

- ATAC-seq is an ensemble measure of open chromatin that uses the prokaryotic Tn5 transposase to tag regulatory regions by inserting sequencing adapters into accessible regions of the genome



Jason D Buenrostro et al., "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position," Nature Methods 10, no. 12 (December 2013): 1213–18, https://doi.org/10.1038/nmeth.2688.

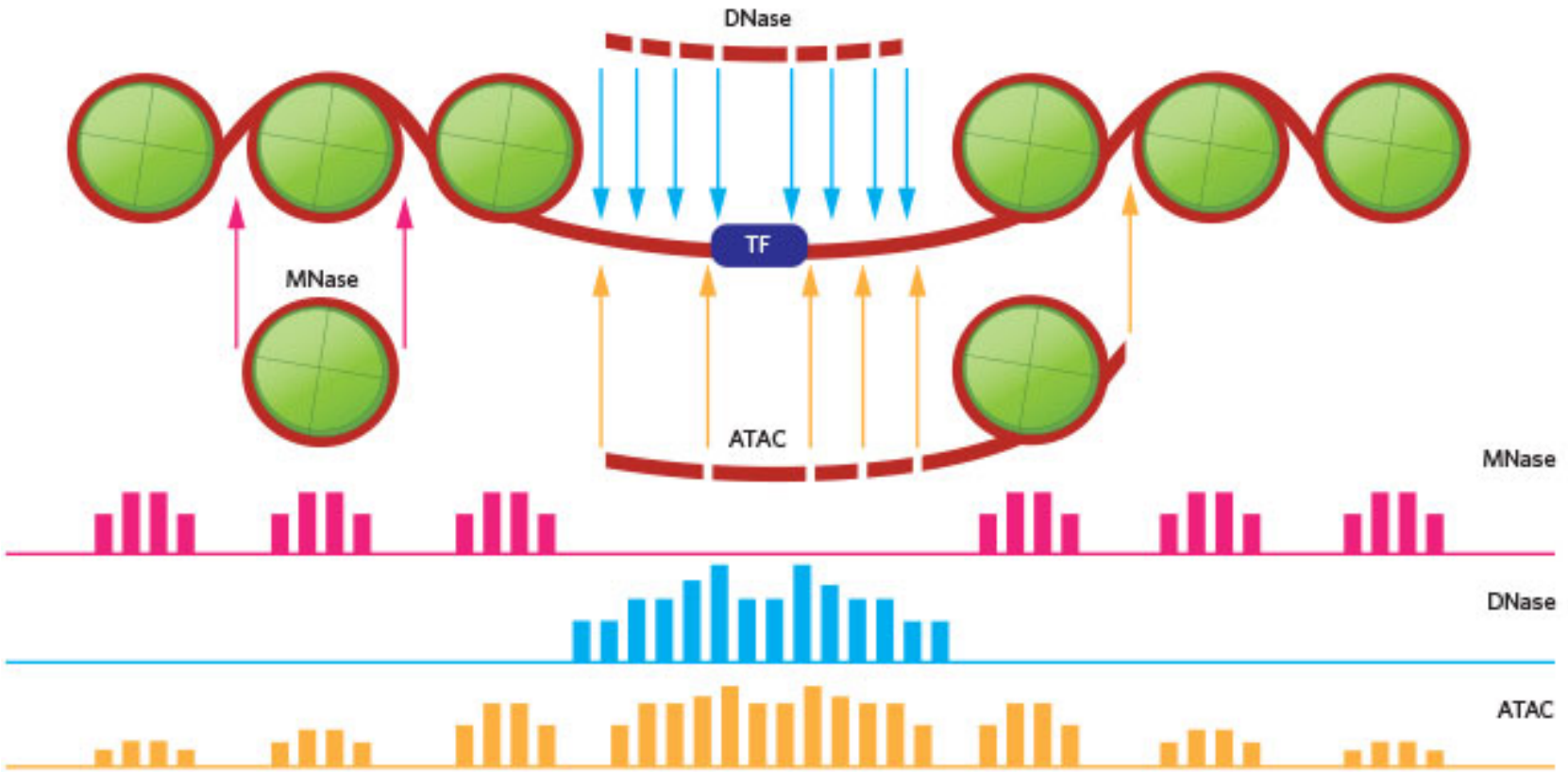
ATAC-seq: finding open chromatin regions



Jason D Buenrostro et al., "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position," Nature Methods 10, no. 12 (December 2013): 1213–18, https://doi.org/10.1038/nmeth.2688.

Technology-specific data

Peaks produced by different technologies are different - calling peaks should be adjusted



Calling peaks in any-seq

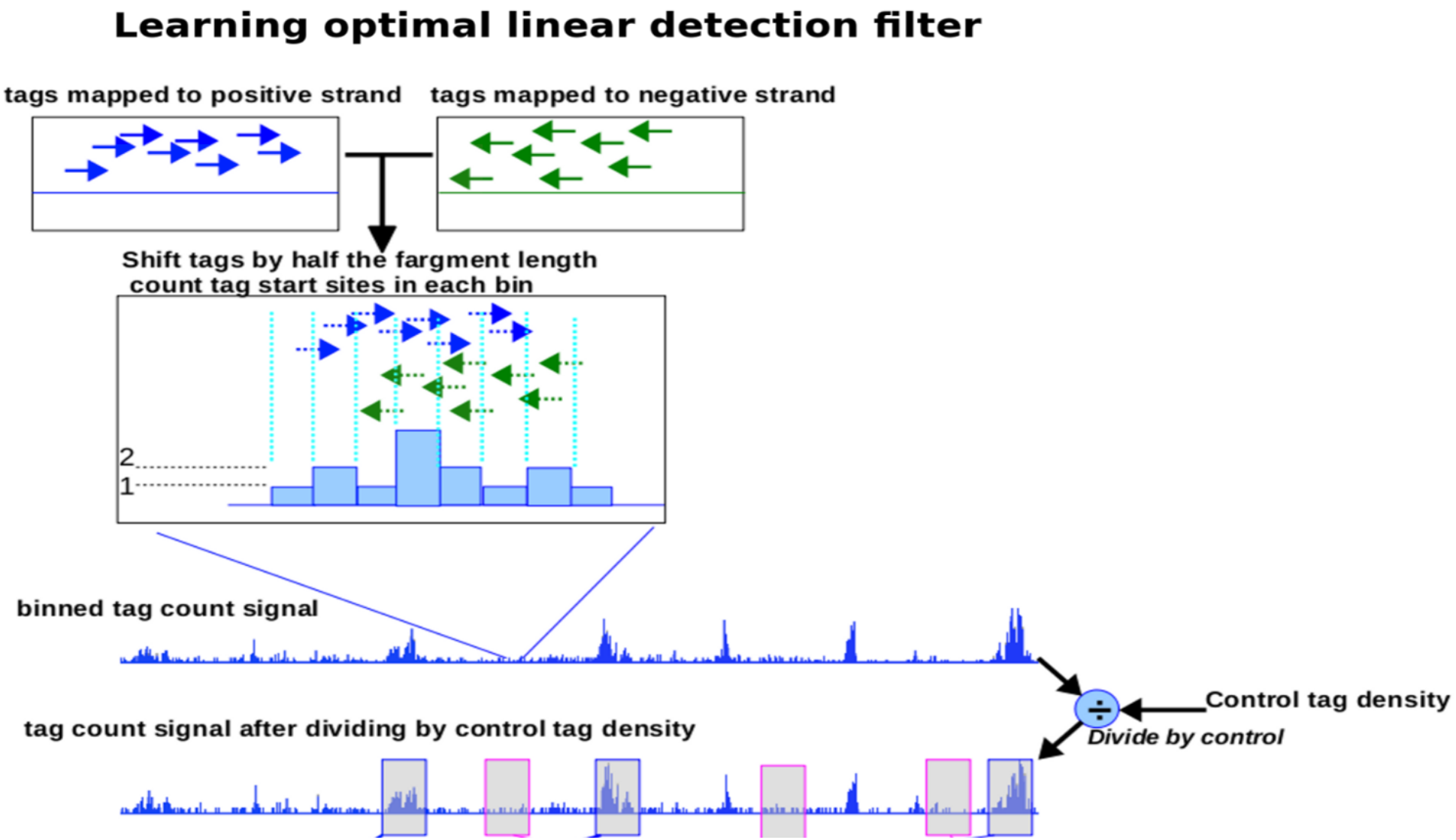
- General signal detection problem for peaks of arbitrary shape
- `DFilter` algorithm - a linear detection filter, known as a Hotelling observer, that provides mathematically optimal detection accuracy
- The objective of the Hotelling detection filter is to maximize the difference between filter outputs at true-positive regions and noise regions.
- More precisely, the Hotelling detection filter maximizes the ratio of the mean of this difference to its s.d.

Kumar, Vibhor, Masafumi Muratani, Nirmala Arul Rayan, Petra Kraus, Thomas Lufkin, Huck Hui Ng, and Shyam Prabhakar. "Uniform, Optimal Signal Processing of Mapped Deep-Sequencing Data." Nature Biotechnology 31, no. 7 (July 2013): 615–22. https://doi.org/10.1038/nbt.2596.

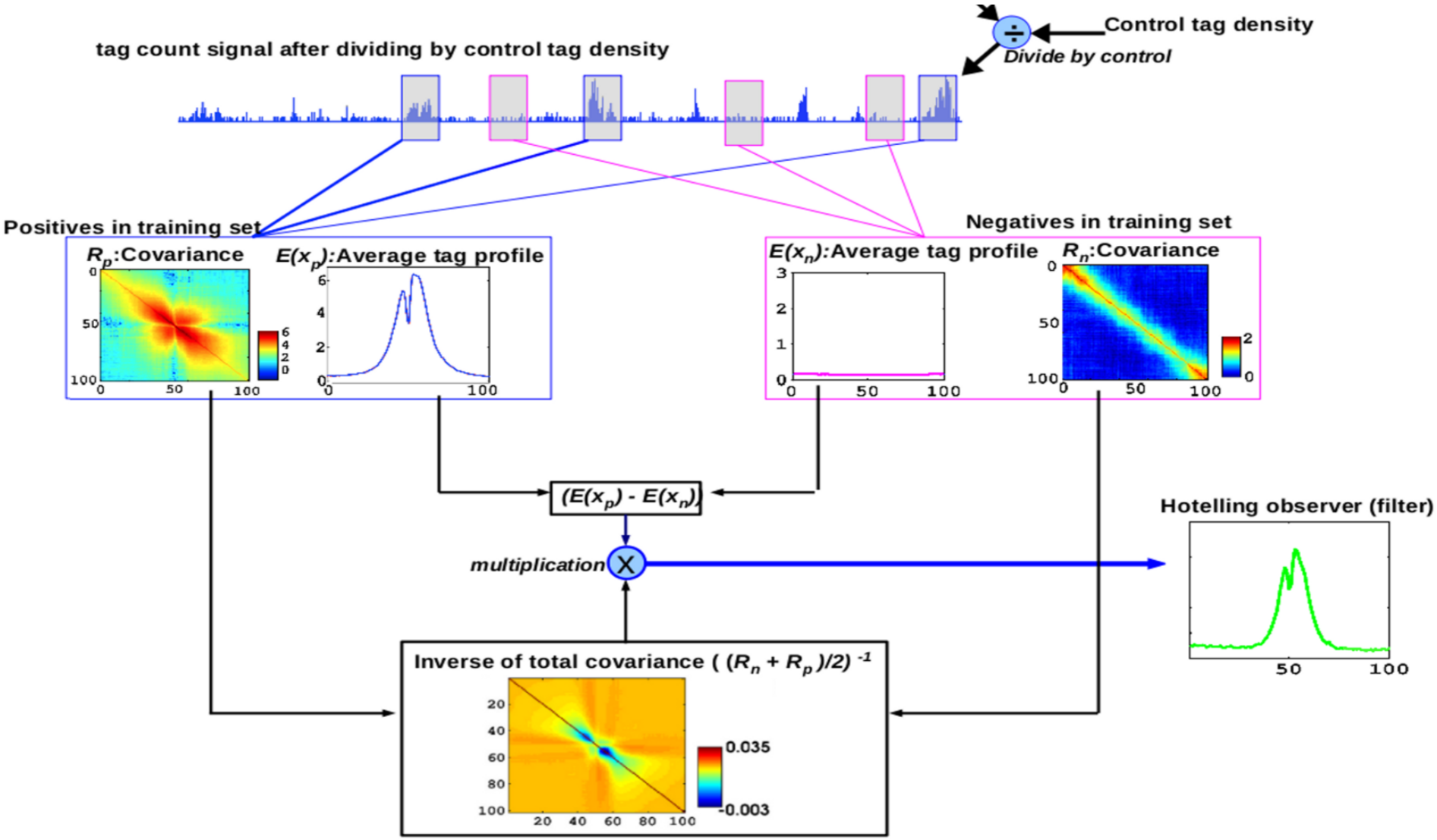
http://collaborations.gis.a-star.edu.sg/~cmb6/kumarv1/dfilter/

Hotelling, Harold. "The Generalization of Student's Ratio." The Annals of Mathematical Statistics 2, no. 3 (August 1931): 360–78. https://doi.org/10.1214/aoms/1177732979. https://projecteuclid.org/download/pdf_1/euclid.aoms/1177732979

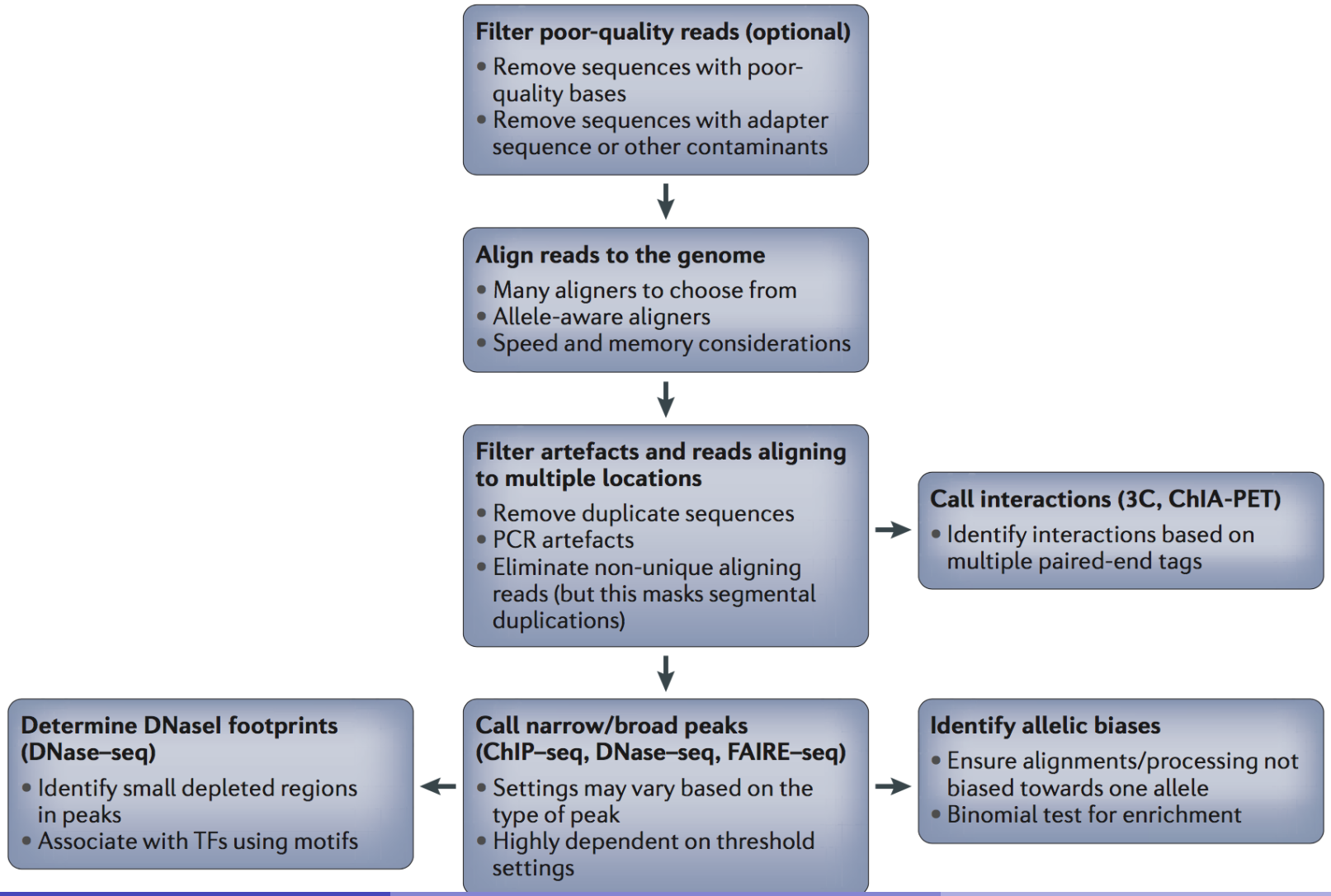
Hotelling detection filter



Hotelling detection filter



General analysis pipeline for sequence-tag experiments



Selected software tools available for three key steps in the analysis of sequence data

Software tool	Web address	Notes
Short-read aligners		
BWA	http://bio-bwa.sourceforge.net	Fast and efficient; based on the Burrows–Wheeler transform
Bowtie	http://bowtie-bio.sourceforge.net	Similar to BWA, part of suite of tools that includes TopHat and CuffLinks for RNA-seq processing
GSNAP	http://research-pub.gene.com/gmap	Considers a set of variant allele inputs to better align to heterozygous sites
Wikipedia list of aligners	http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment	A comprehensive list of available short-read aligners, with descriptions and links to download the software
Peak callers		
MACS	http://liulab.dfci.harvard.edu/MACS	Fits data to a dynamic Poisson distribution; works with and without control data
PeakSeq	http://info.gersteinlab.org/PeakSeq	Takes into account differences in mappability of genomic regions; enrichment based on FDR calculation
ZINBA	http://code.google.com/p/zinba	Can incorporate multiple genomic factors, such as mappability and GC content; can work with point-source and broad-source peak data
Differential peak calling		
edgeR	http://www.bioconductor.org/packages/2.9/bioc/html/edgeR.html	Uses negative binomial distribution to model differences in tag counts; uses replicates to better estimate significant differences
DESeq	http://www-huber.embl.de/users/anders/DESeq	Also uses negative binomial distribution modelling, but differs in the calculation of the mean and variance of the distribution
baySeq	http://www.bioconductor.org/packages/release/bioc/html/baySeq.html	Uses empirical Bayes approach to identify significant differences; assumes negative binomial distribution of data
SAMSeq	http://www.stanford.edu/~junli07/research.html#SAM	Based on the popular SAM software; a non-parametric method that uses resampling to normalize for differences in sequencing depth