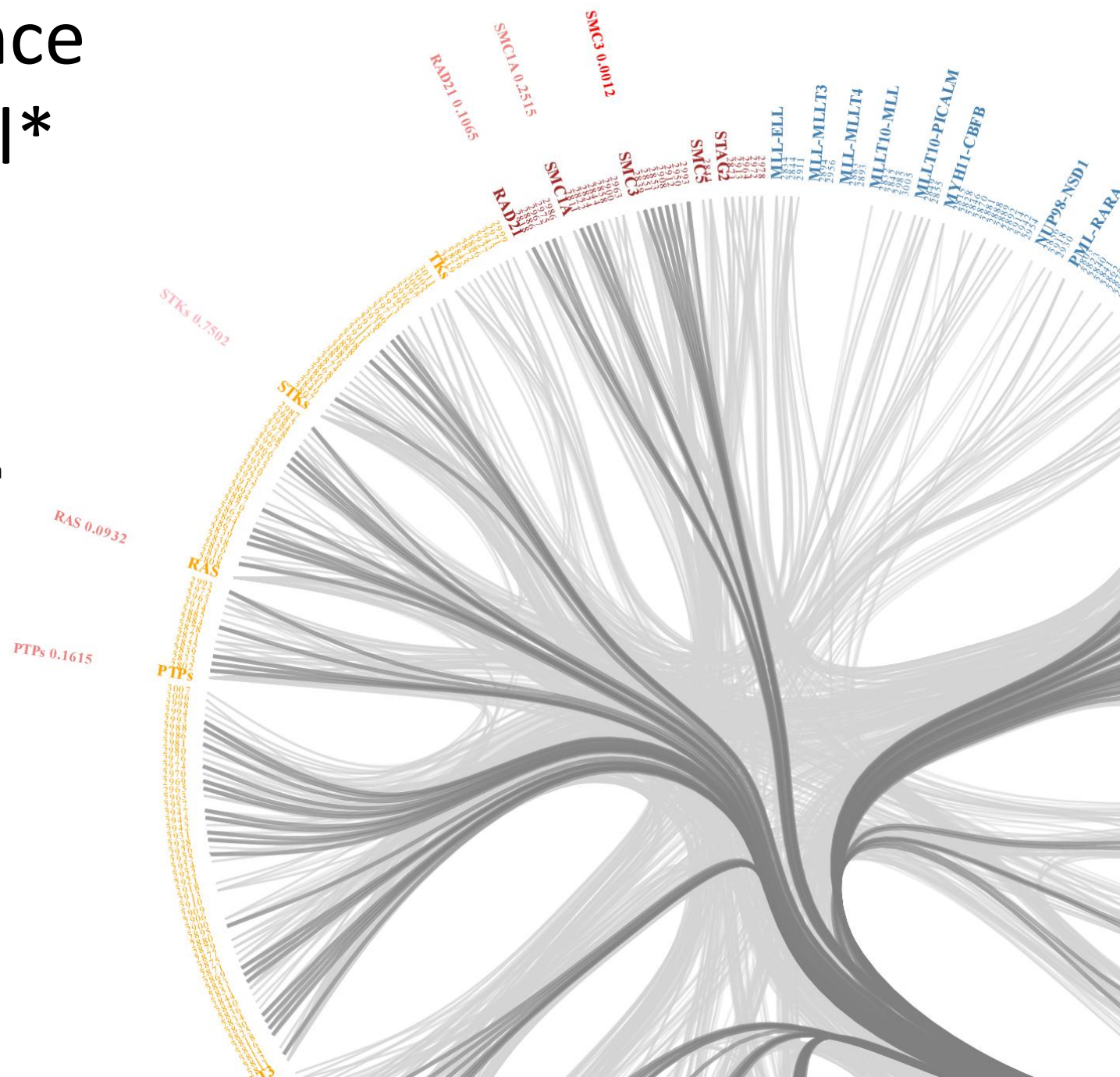
Big* science on a small* budget

Tim Triche, Jr.
University of Southern California



* relatively

(by biology standards) *(ibid)*

Big ^ Science on a Small ^ Budget

Major data-generating projects

- 1000 Genomes Project & Cancer Genome Atlas
- ENCODE & the Reference Epigenome Mapping Consortium

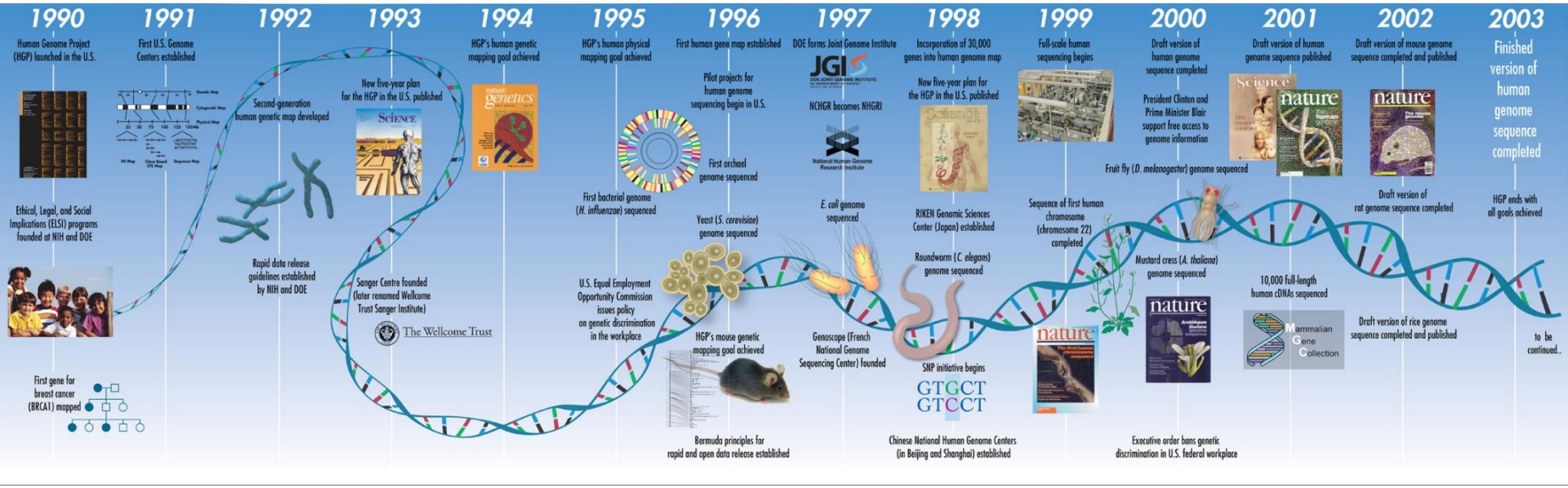
Case studies

- Chromatin state models & environmental epigenetics
- Bayesian change points, broad peaks & two-way streets

BioC workflows

- Exploring chromatin states: chromophobe , GenometriCorr
- Digesting histone mark ChIP-seq data: Rsubread, BCPeakR

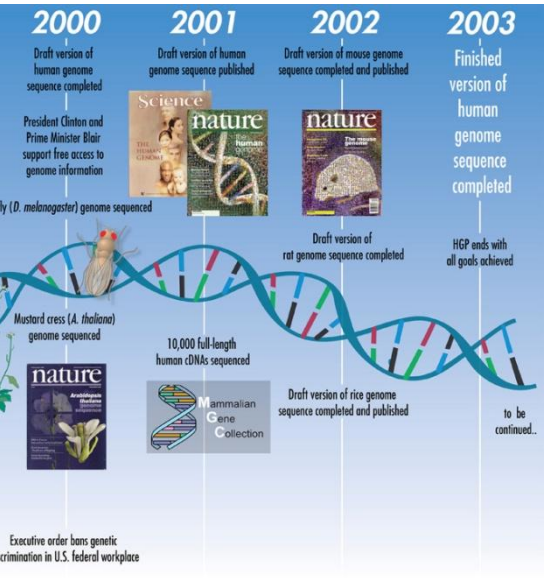
One human genome is useful

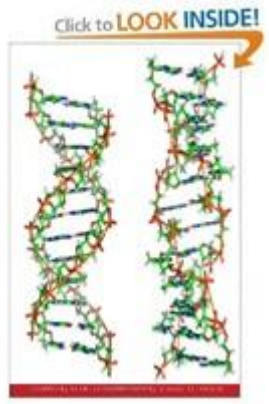


Nature

- Genomes are fairly consistent across tissues in humans
- The genome is *nearly* identical across human somatic cells
- A reference genome allows compact notation for changes

1000 human genomes are more useful

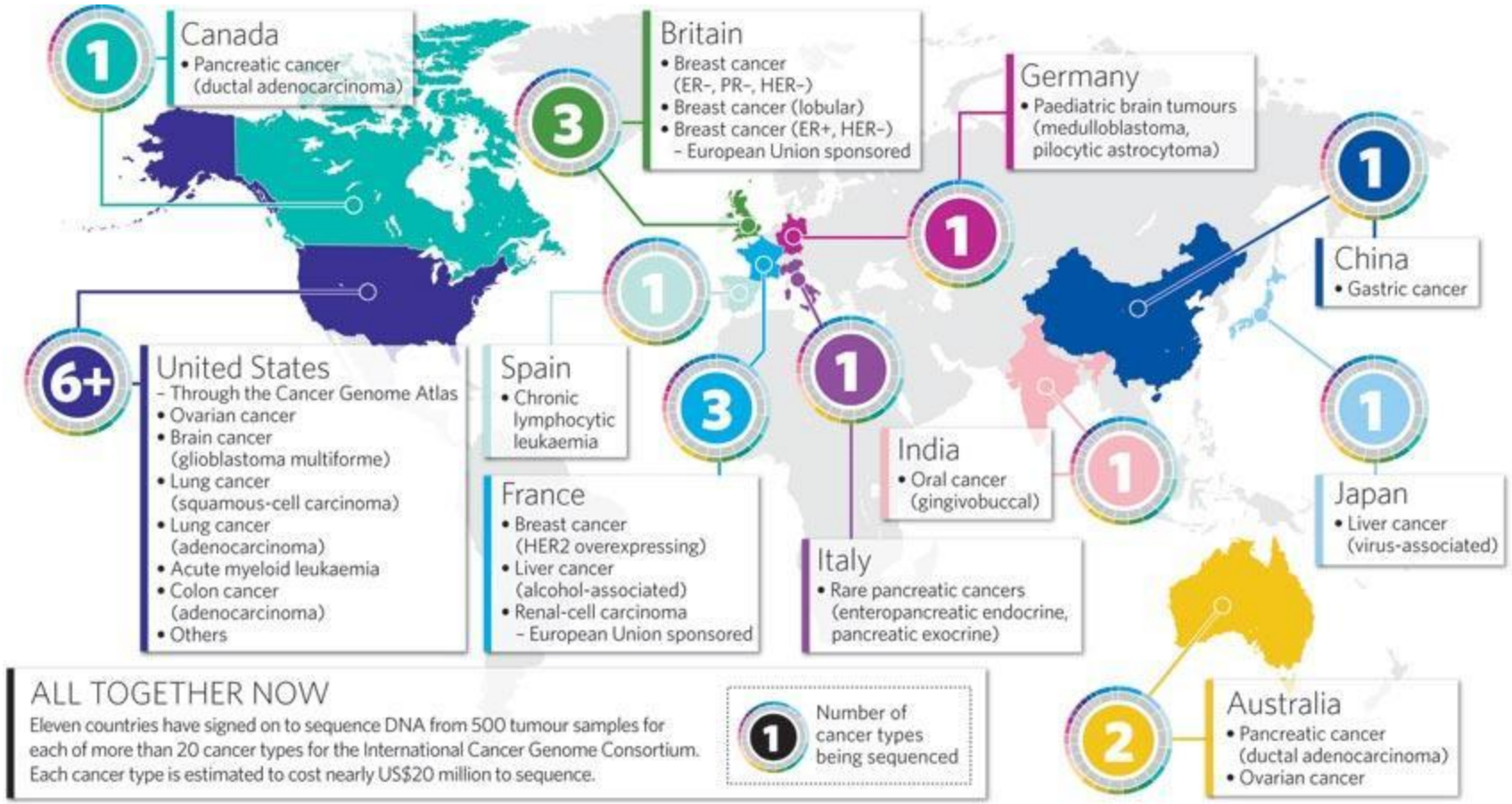




Sharif Sakr

- Despite aggregate genomic similarities, various populations are more and less susceptible to various maladies & risks.
- All else being equal, more (representative) data is better.
- Germline variation can also inform functional inference.

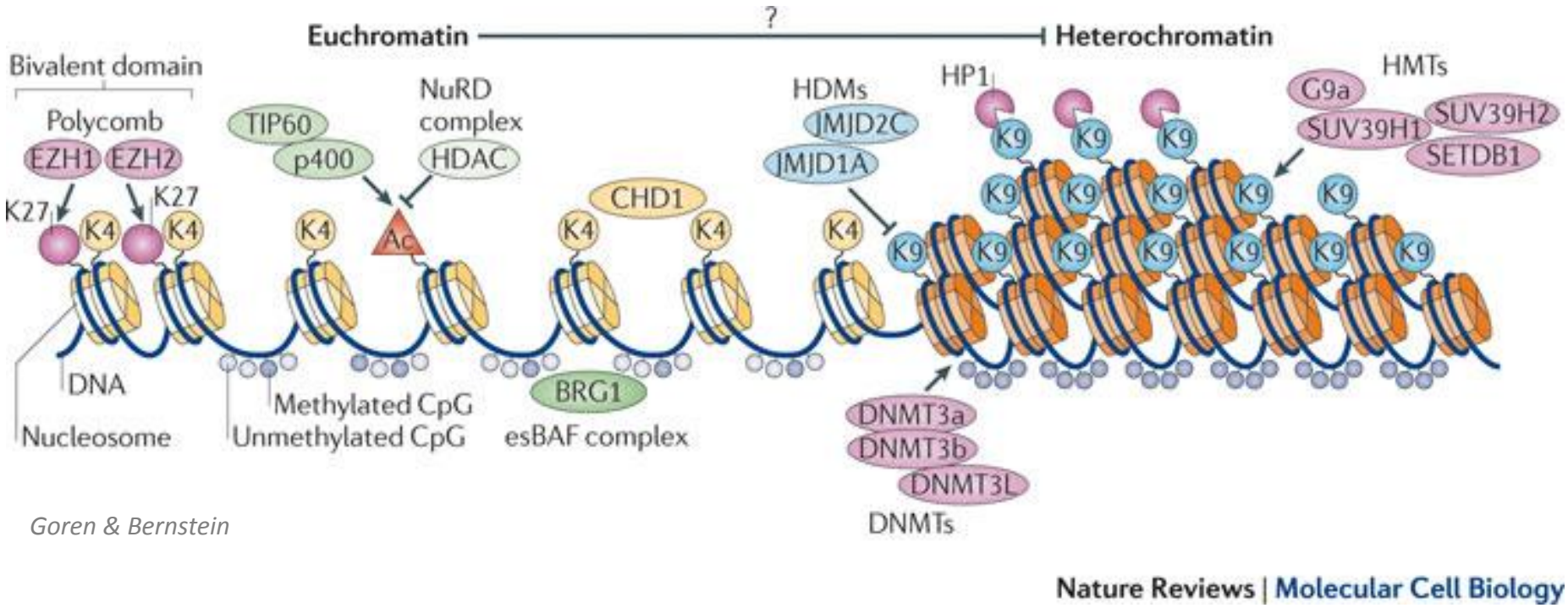
10000 tumor/normal genomes are also useful



Nature

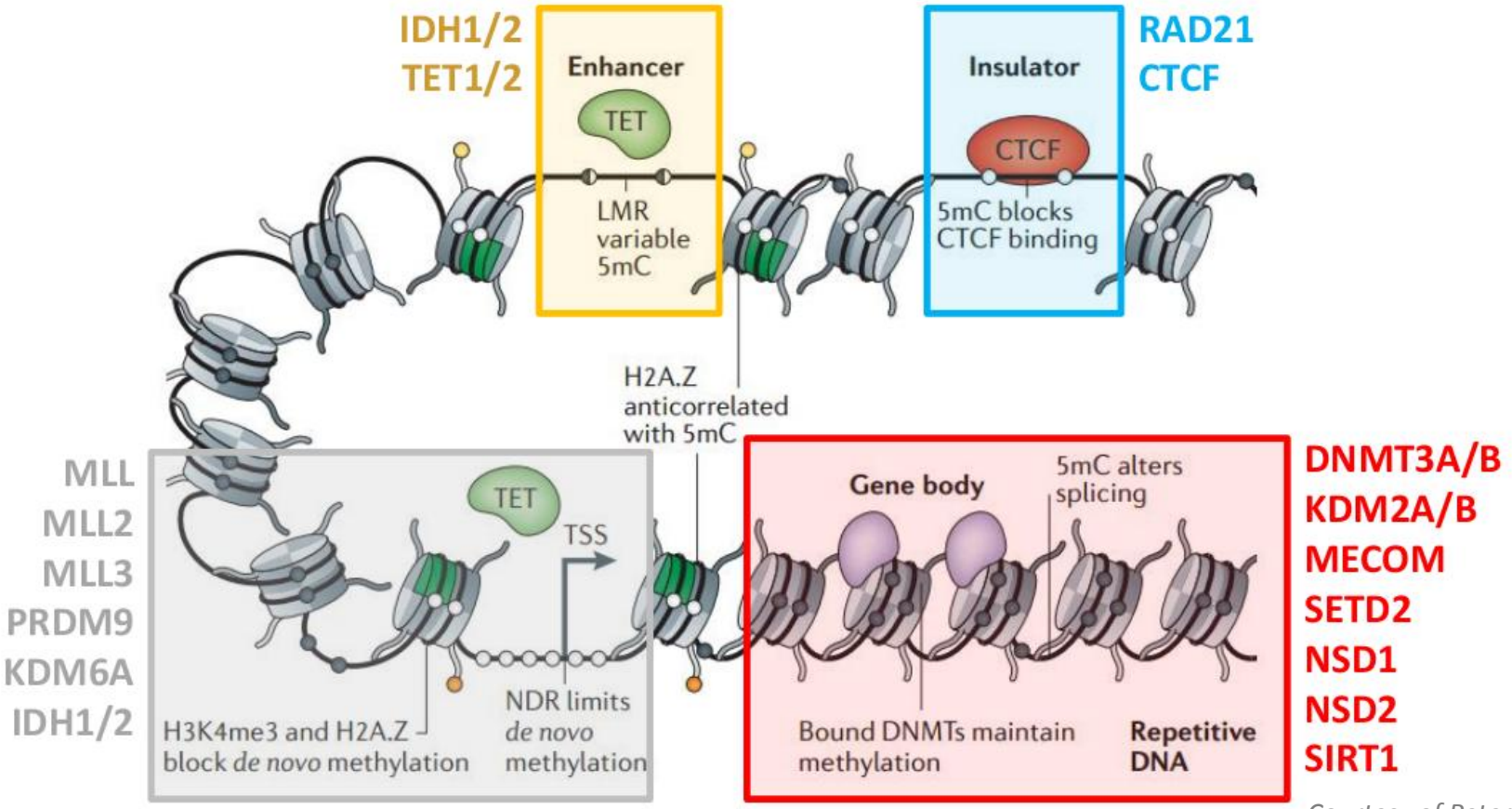
But neither normal nor cancer cells are homogeneous.

Epigenetic marks link the genome and transcriptome



- These marks differ from cell to cell, and also “drift” with age.
- Many non-coding genetic variants with disease risk have been found to confer epigenetic consequences; many recurrent mutations across cancers impact epigenetic machinery:

Recurrent mutations across cancers interfere both directly and indirectly with the epigenetic machinery

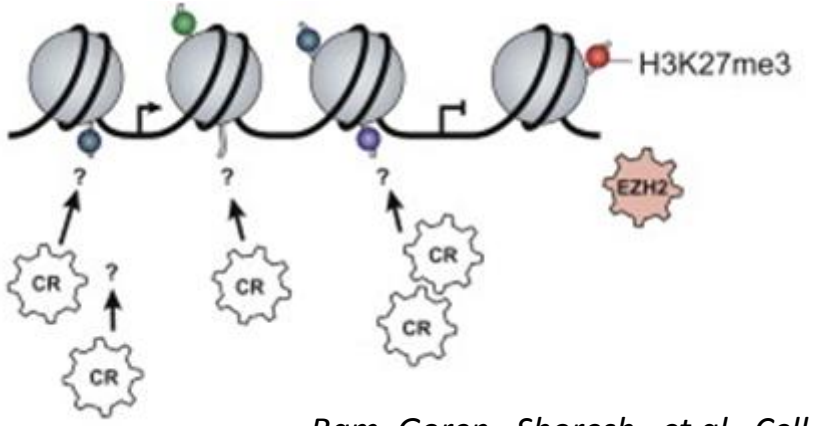


Courtesy of Peter Jones

- In myeloid malignancies, the *majority* of cases are affected.

ENCODE provides a model for the “histone code”

Jason Ernst (formerly MIT, now UCLA) built a hidden Markov model for multiple histone marks as multivariate Bernoulli emissions from hidden biological states. These states group regulatory factors and define contexts for the impact of both genetic and epigenetic changes.



Ram, Goren, Shoresh , et al. Cell 2011

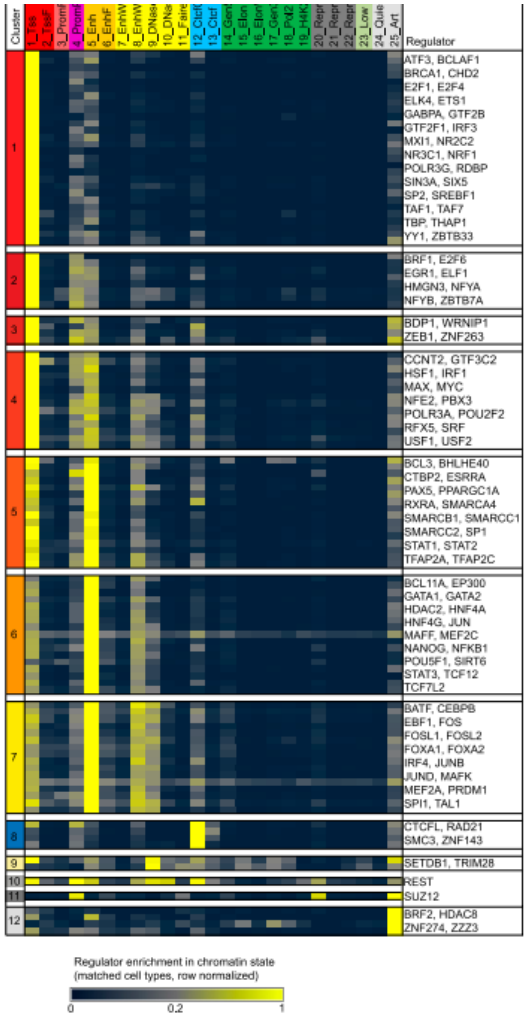
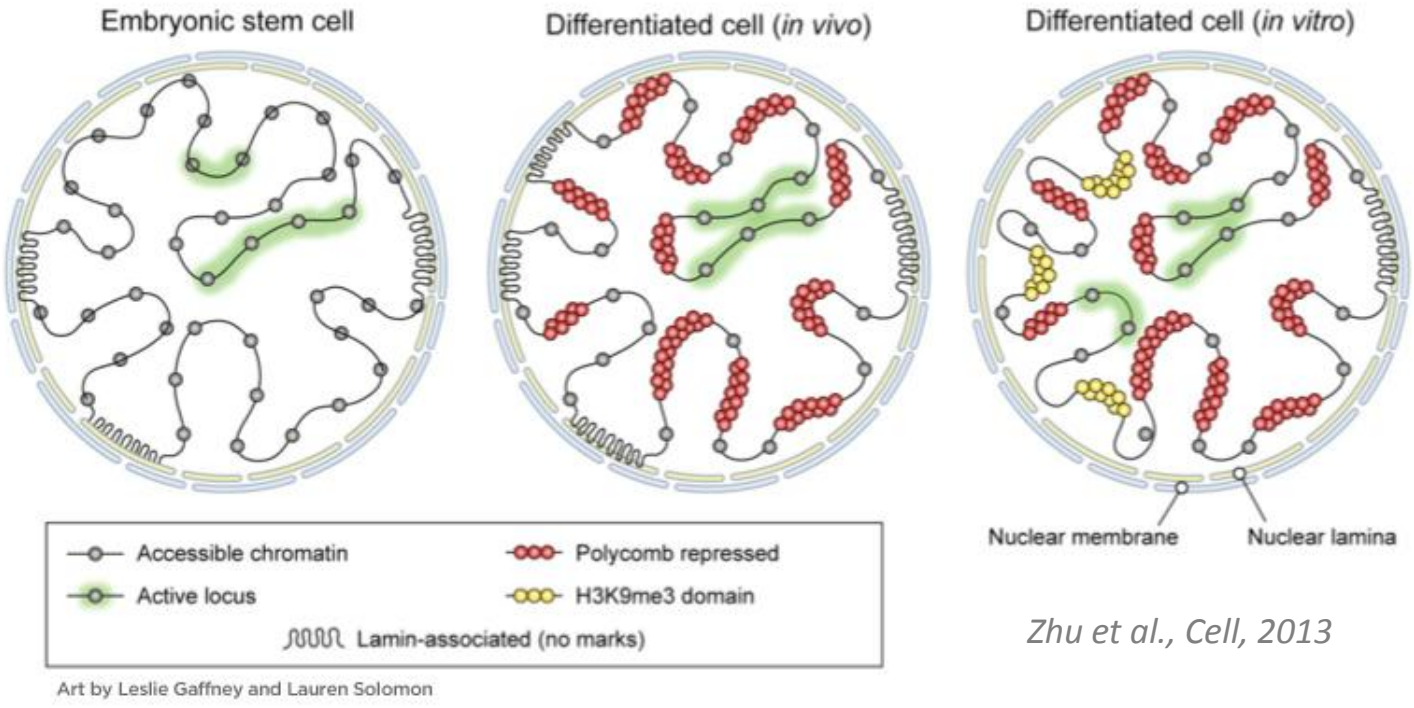


Figure 1. Regulator enrichments for each chromatin state in matched cell types. Different regulators show distinct chromatin state preferences. For each regulator with matching chromatin data, the average enrichment is shown for each chromatin state (columns). Enrichments have been row-normalized, scaling by the largest enrichment value for each experiment. K-means clustering with 12 clusters produced the clusters labeled C1–C12.

However, ENCODE's cell lines do not necessarily represent primary tissues



Zhu et al., Cell, 2013

- Zhu, Bernstein, and colleagues broadly observed such phenomena.
- The NIH Epigenomics Roadmap project (REMC) exists to collect and distribute such data for representative human primary tissues.
- These primary tissues are critical reference points for many studies.

(by biology standards) *(ibid)*

Big ^ Science on a Small ^ Budget

Major data-generating projects

- 1000 Genomes Project & Cancer Genome Atlas
- ENCODE & the Reference Epigenome Mapping Consortium

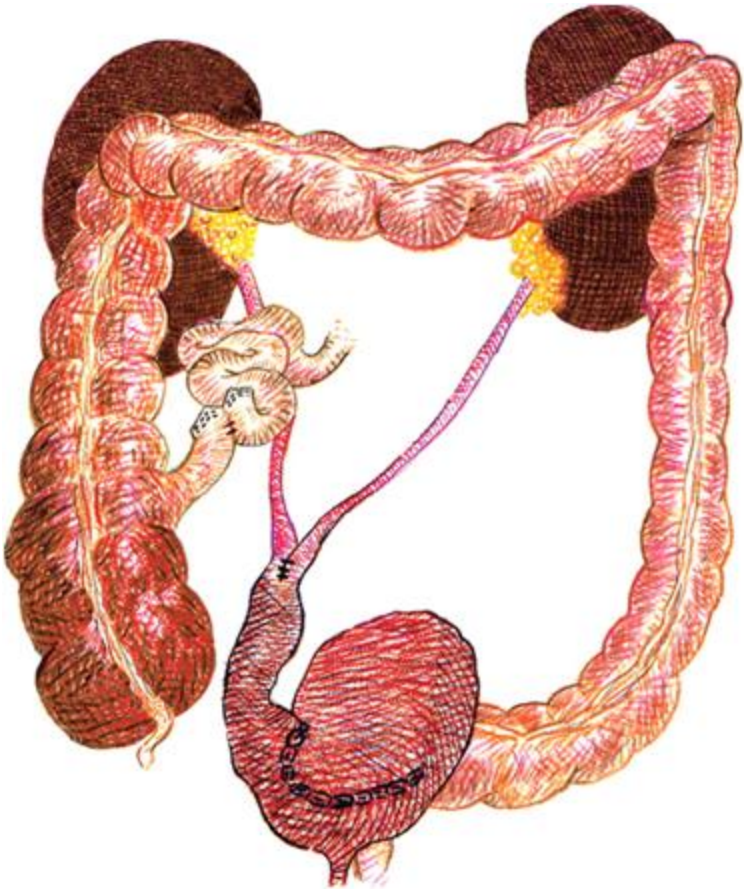
Case studies

- Chromatin state models & environmental epigenetics
- Bayesian change points, broad peaks & two-way streets

BioC workflows

- Exploring chromatin states: chromophobe , GenometriCorr
- Digesting histone mark ChIP-seq data: Rsubread, BCPeakR

Interpreting environmental changes in DNA methylation via chromatin states



The natural experiment:

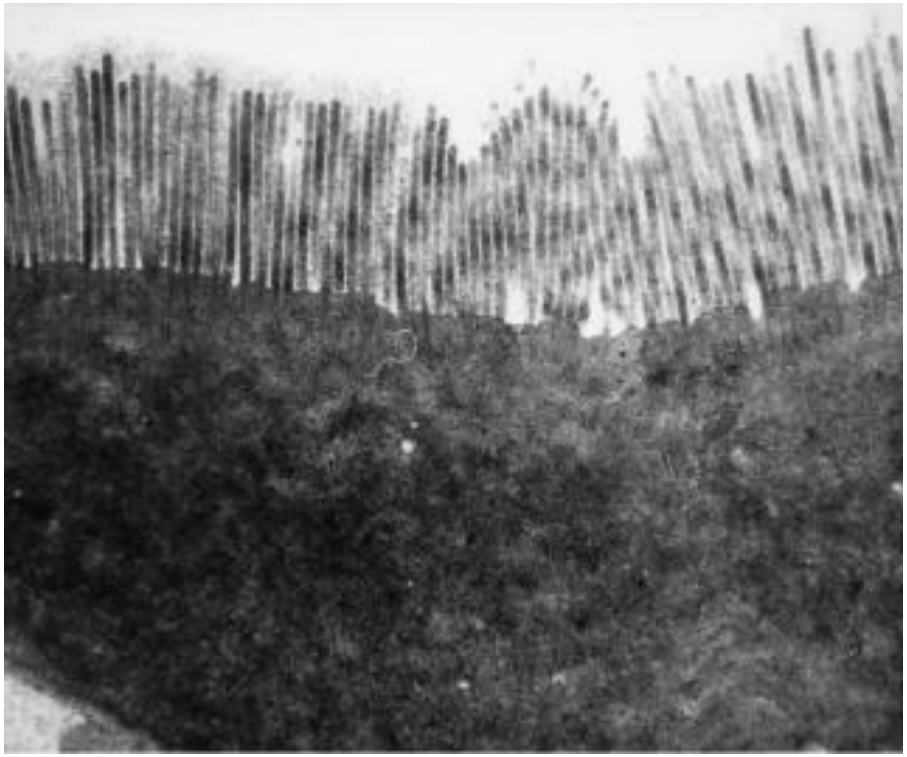
Radical cystectomy is followed by surgery to create a neobladder from a patient's ileum.

The question:

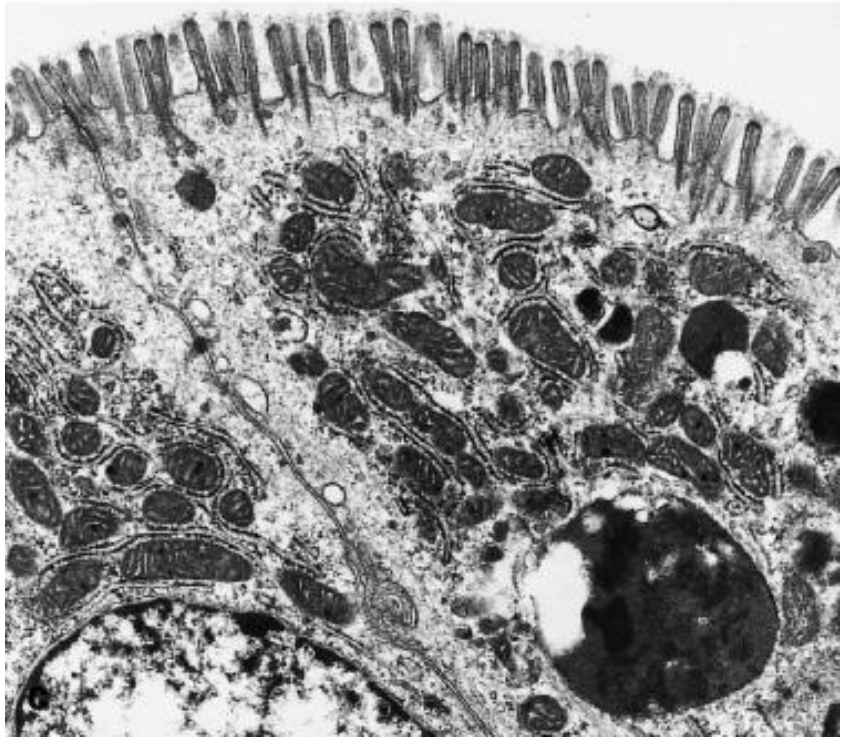
Changing only the niche, will we the adaptations in DNA methylation serve as a proxy of epigenetic state?

What regulates this (smooth) transition?

BEFORE



AFTER



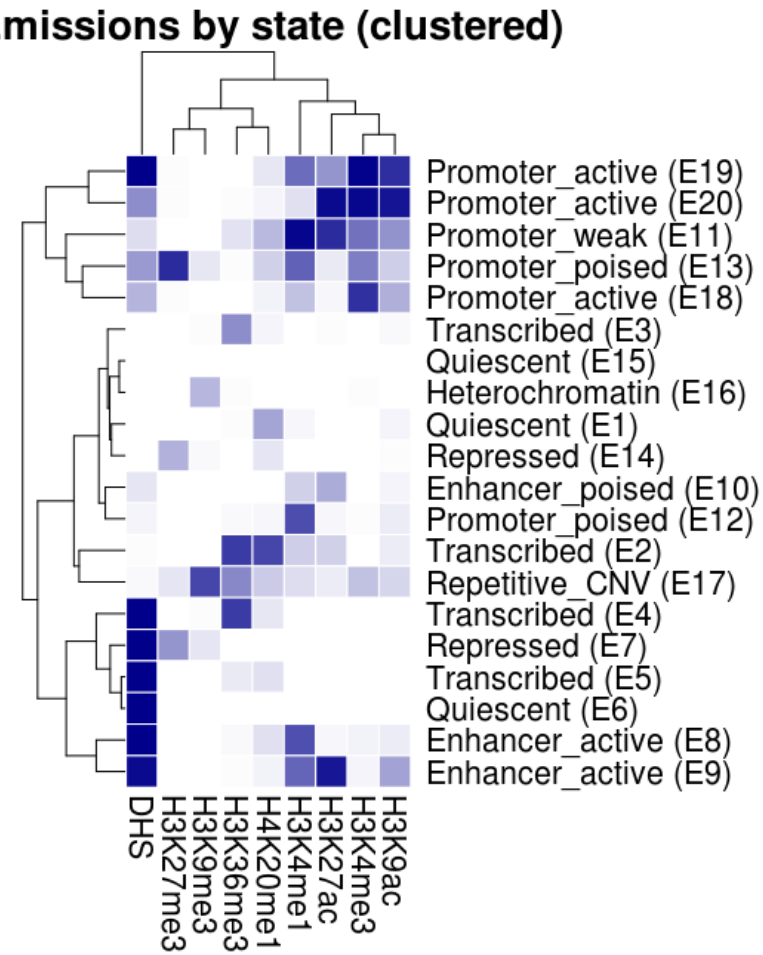
Intestinal epithelium adopts a urothelium-like structure in response to environmental changes, over the span of several years.

Aragona et al, 1997

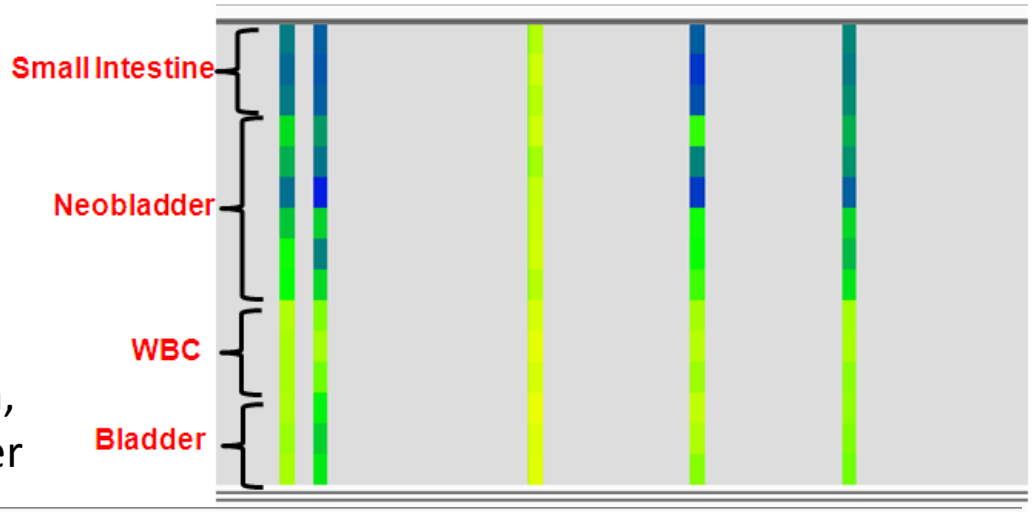
A use for Roadmap data in an ENCODE model

Our experiment arises in healthy, primary tissues, so cell lines are not informative. However, the ChromHMM model helps us to make sense of our results, so we applied it the set of marks available for the tissues that served as our endpoints.

After some fiddling and testing, we arrived at a sensible model that fit observed correlates (e.g. RNAseq data for each tissue).



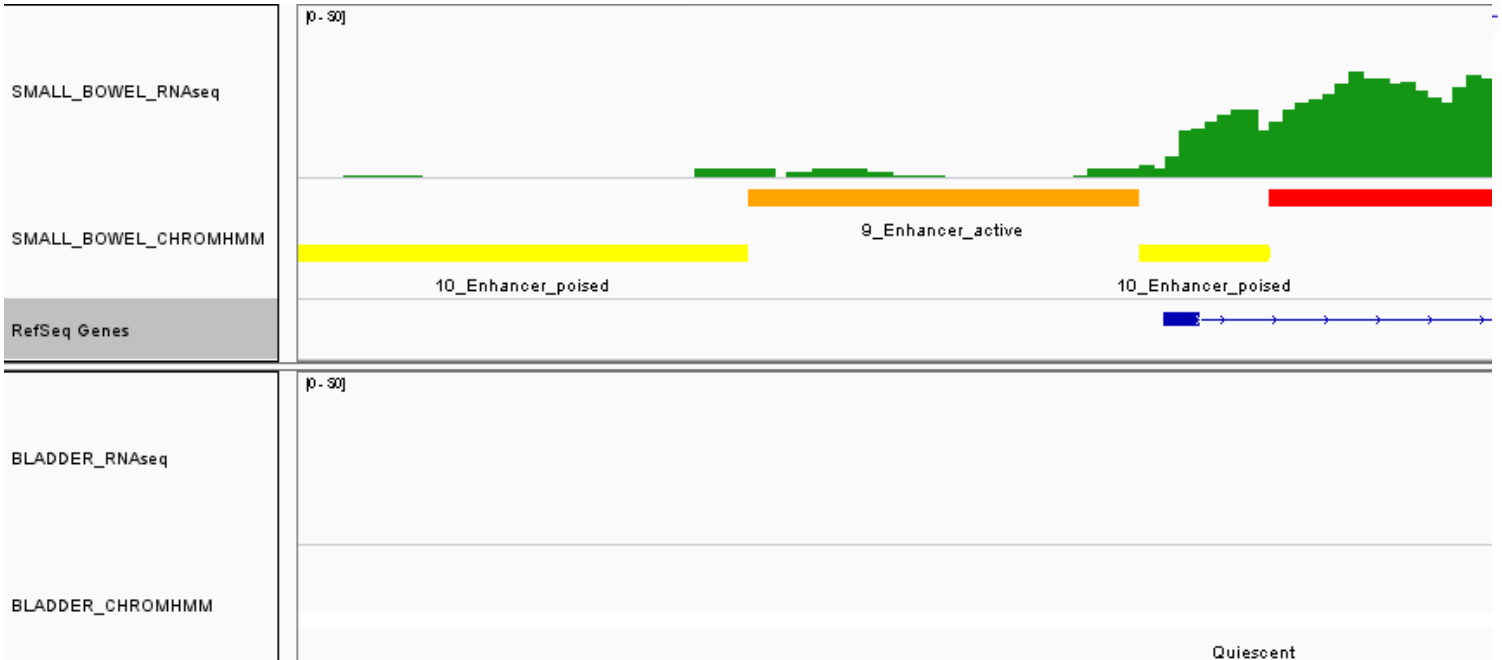
Intestine-specific genes are repressed…



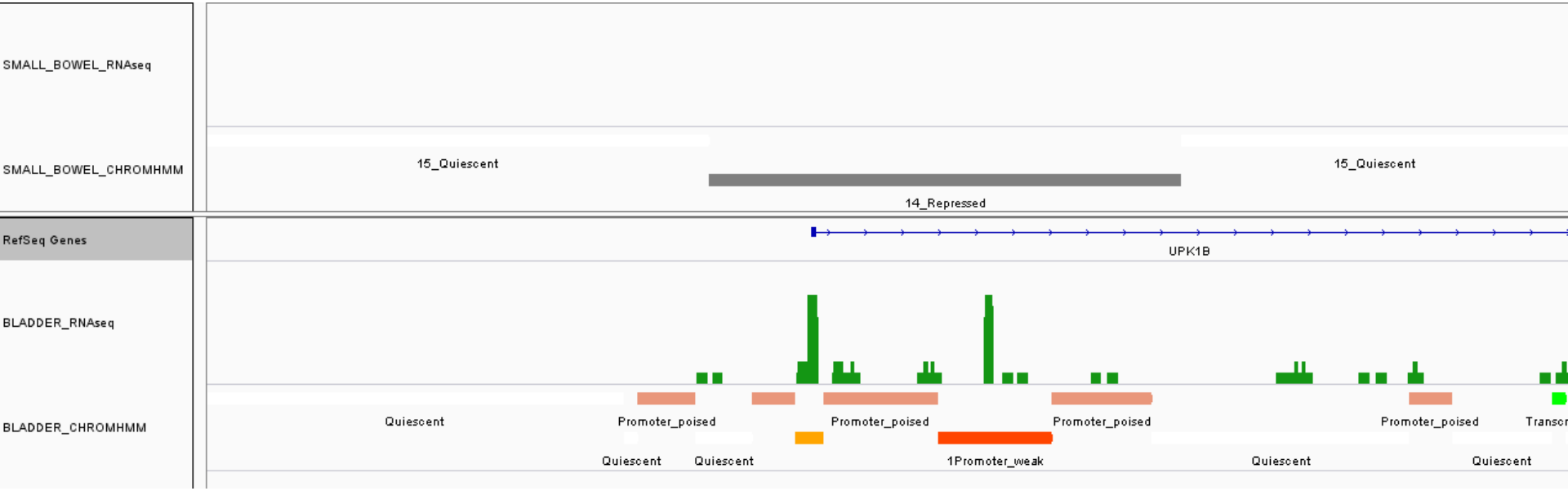
Time Since Surgery

Methylation
0 ---------->1

VIL1 is active in ileum, but inactive in bladder



and bladder-specific genes are activated.

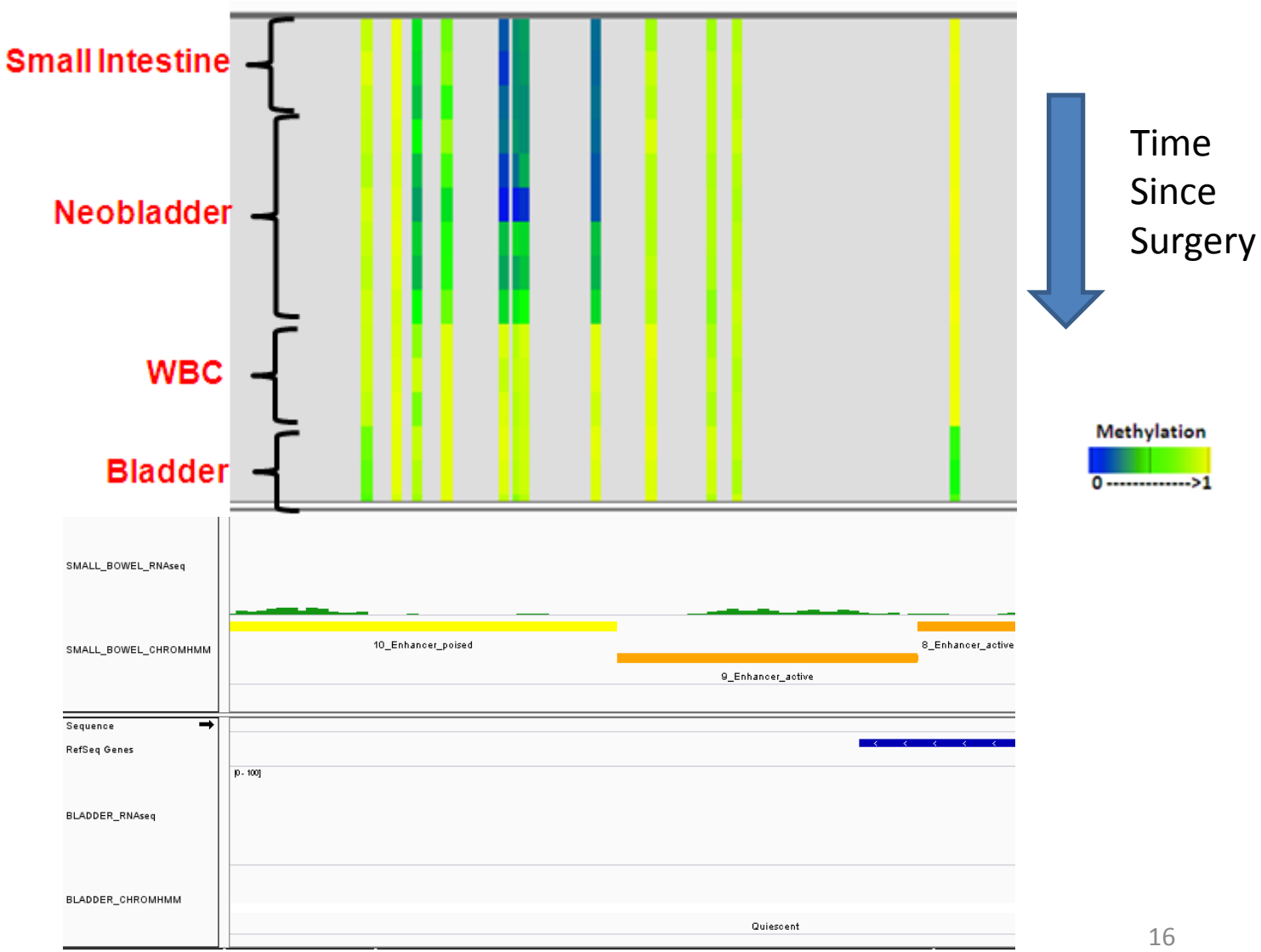


As the repressive state of the UPK1B promoter disappears, so also does DNAm.

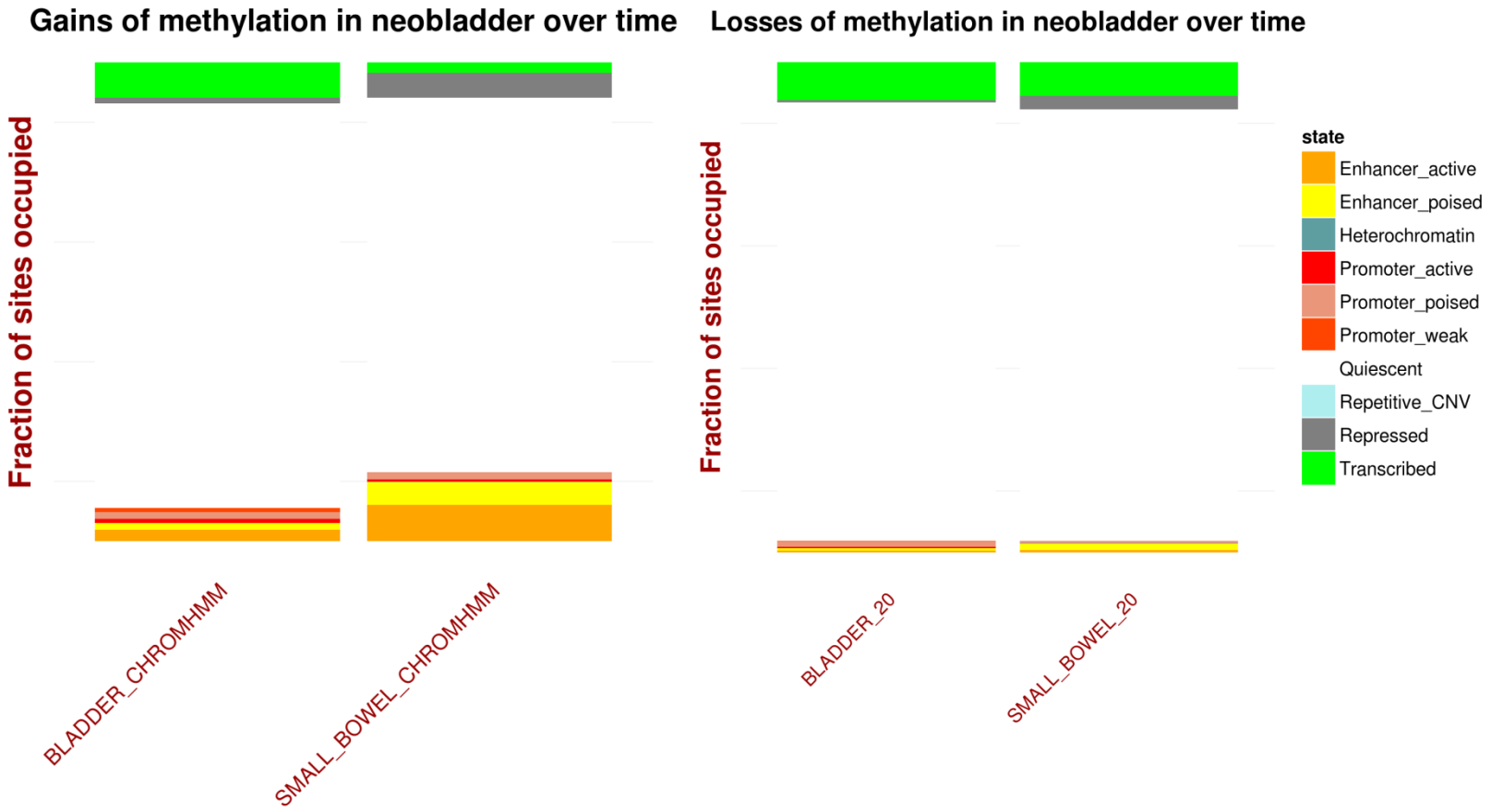
As the machinery to scan for DMRs using limma and bumphunter is now more stable, we intend to re-fit the data, aiming at discovering more coherent drivers.

Nonetheless, a tabulation of the changes suggests that even disparate loci capture biologically significant regulatory events, especially at tissue-specific enhancer sites.

Novel distal enhancers change state…



…and changes take on a coherent form

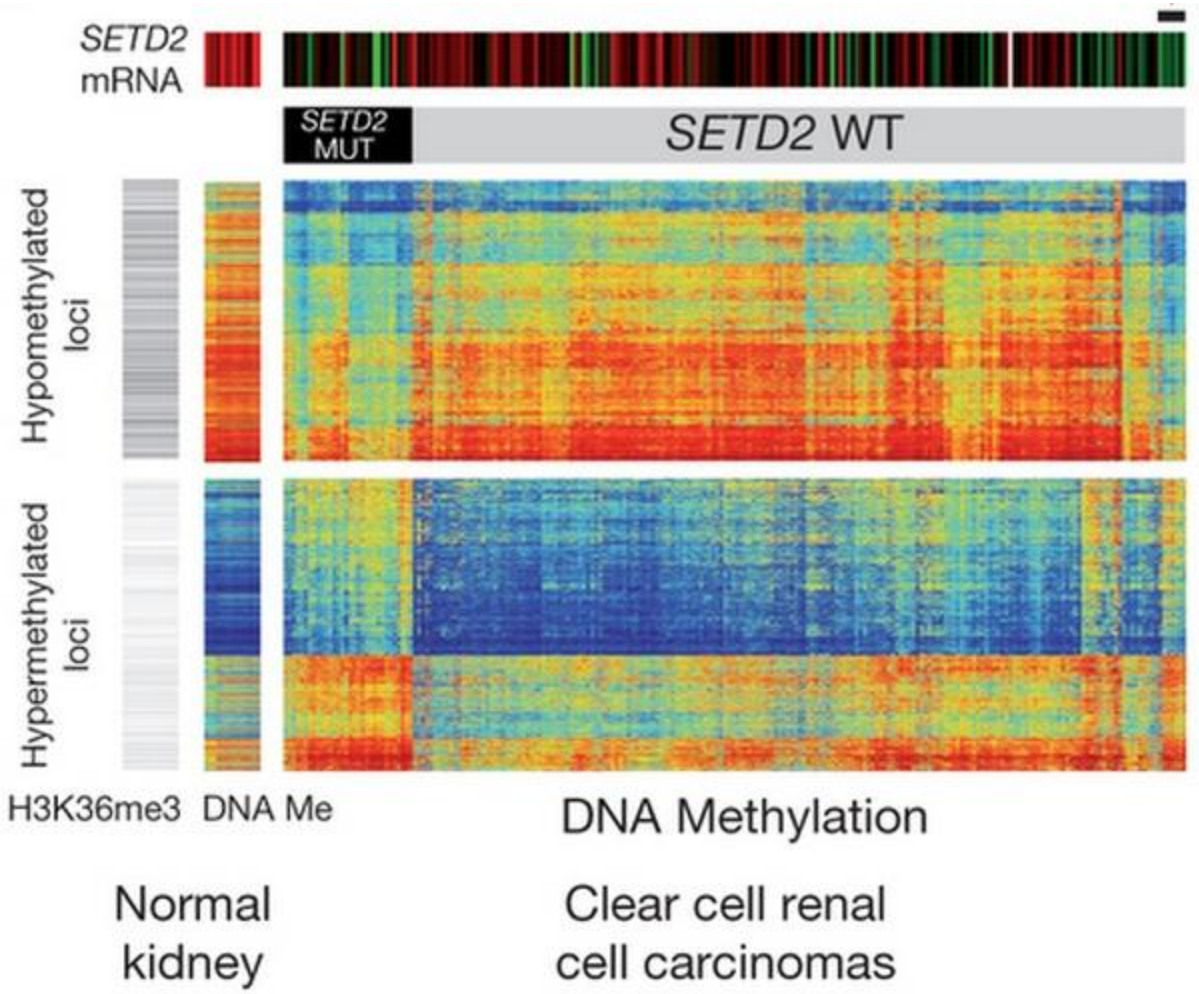


The major (regulatory) changes focus on enhancers.

Compared to expression, overall DNAm moves slowly.

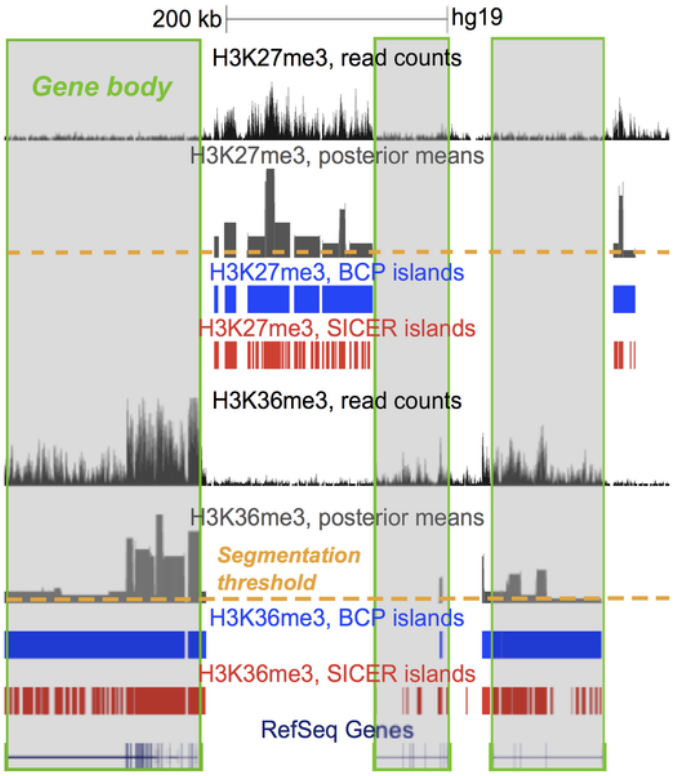
Another use for raw Roadmap data: SETD2 mutant-specific DNAm changes

- In the recent KIRC (renal cell carcinoma) paper, we sought to show SETD2 mutation impact on H3K36me3-marked sites.
- Broad histone marks are notoriously balky to work with.



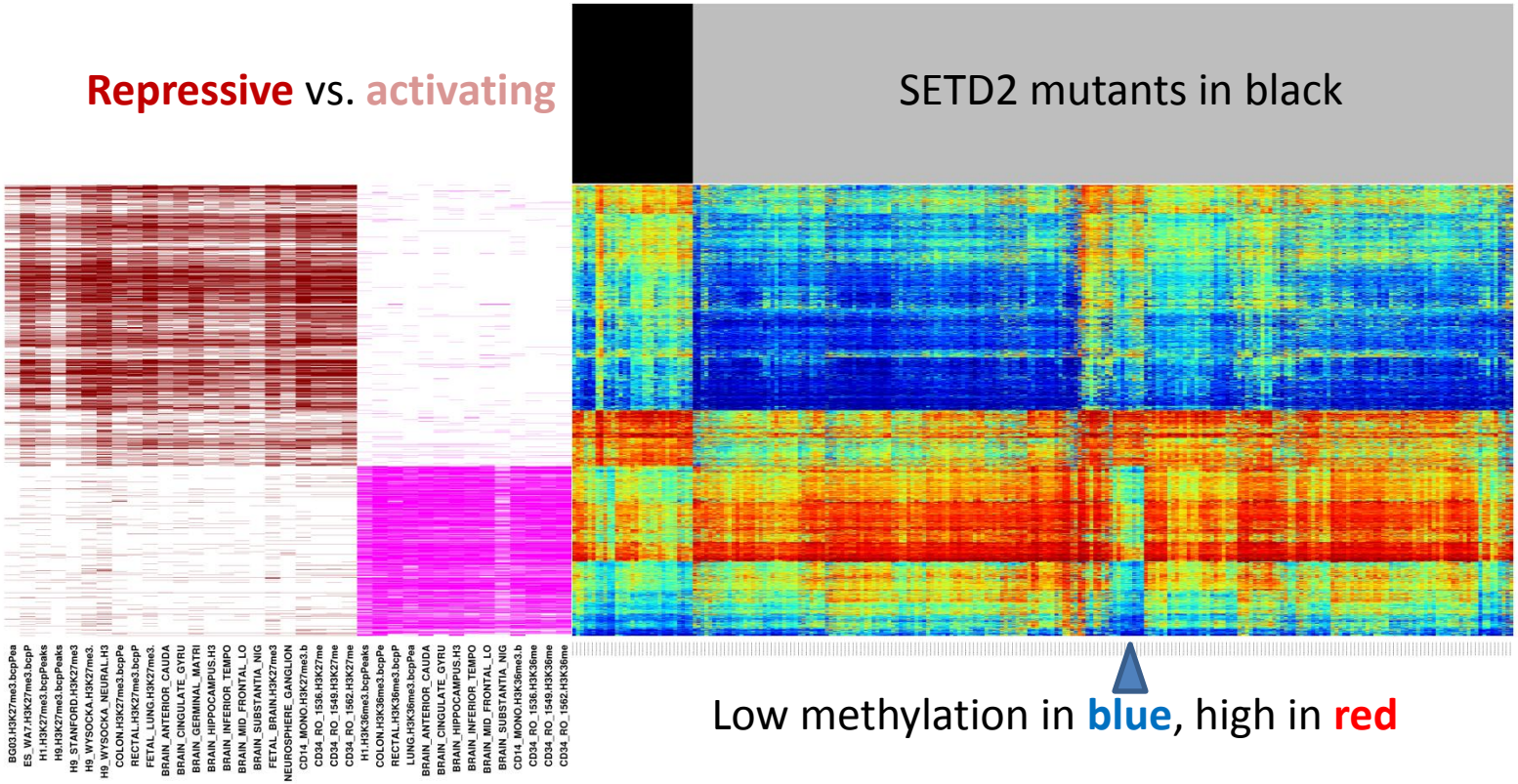
A better solution: Bayesian changepoint model for broad peaks (histone marks)

- Using results from infinite HMMs, apply a bounded-complexity mixture model to within-peak read counts.
- Recursively merge states into ‘islands’ until threshold FDR is exceeded given the posterior read counts in each island.



- Result: vastly improved predictive ability for correlated marks (e.g. DNAm and H3K36/K27me3)
- Original code: Haipeng Xing, Yifan Mo, Wiley Liao

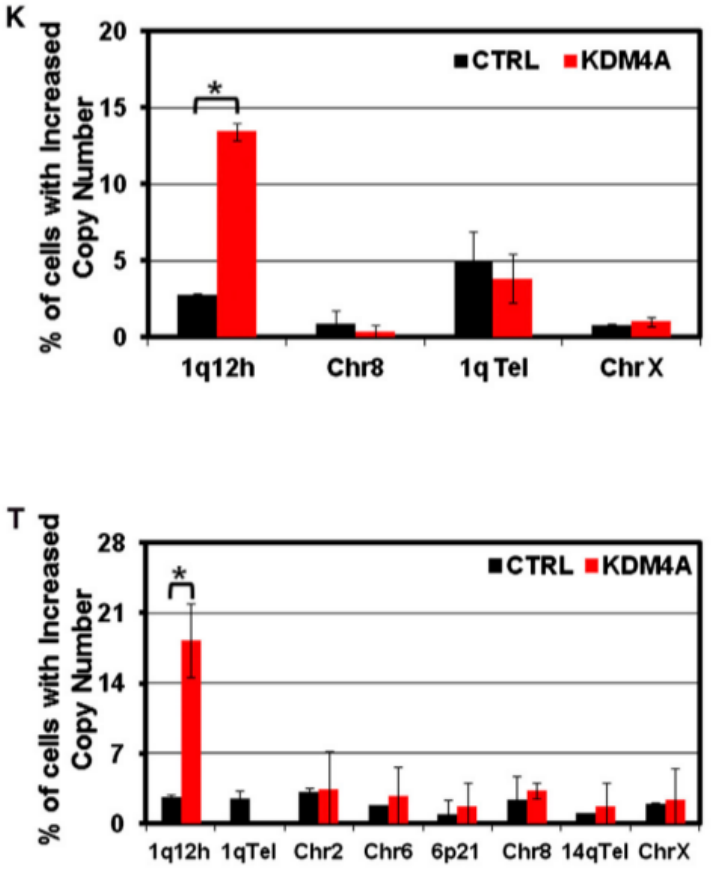
Same data, better calls, better results



- When we apply this logic for islands of repressive and activating marks, the nature of the changes indicates the mutations and reveals a phenocopy.

Recent work suggests a 2-way street

- In Black et al. (Cell 2013), Gad Getz' group has shown that KDM4A amplification & overexpression is associated with recurrent focal copy number gains in ovarian tumors.
- Suv39h1 or HP1γ overexpression suppresses the copy gain; H3K9/K36 methylation dysfunction promotes it
- SETD2 mutants interfere with H3K36 trimethylation, as do MMSET/WHSC1 and NSD1 mutants. This suggests one common mechanism by which epigenetic dysregulation can act in a feedback loop to promote focal genetic aberrations across tumors.



(by biology standards) *(ibid)*

Big ^ Science on a Small ^ Budget

Major data-generating projects

- 1000 Genomes Project & Cancer Genome Atlas
- ENCODE & the Reference Epigenome Mapping Consortium

Case studies

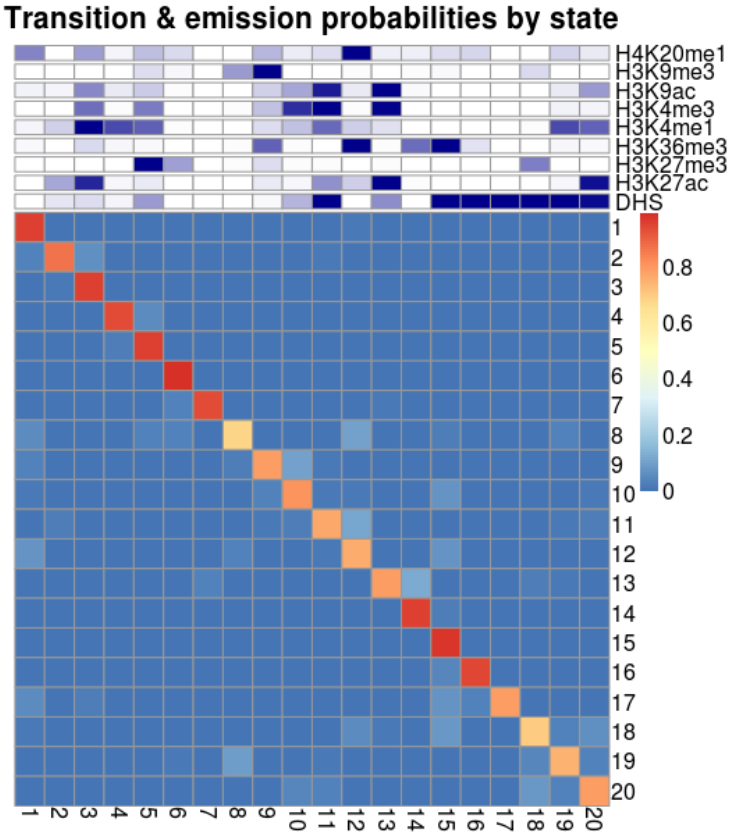
- Chromatin state models & environmental epigenetics
- Bayesian change points, broad peaks & two-way streets

BioC workflows

- Exploring chromatin states: chromophobe , GenometriCorr
- Digesting histone mark ChIP-seq data: Rsubread, BCPeakR

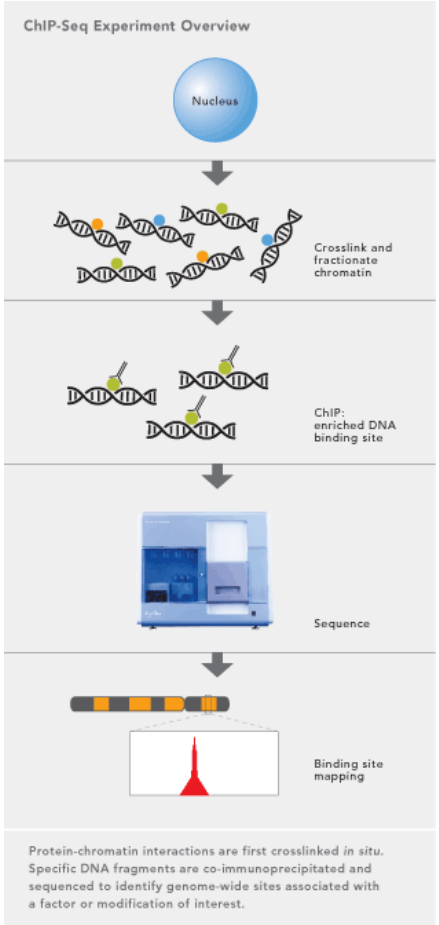
Segmentation exploration: chromophobe

- GenomicRanges and rtracklayer do most of the hard work for this job!
- GenometriCorr makes testing for spatial correlation quite simple
- chromophobe eases model import, exploration, validation, and export
- MethylSeekR segmentations are also supported in more recent releases
- Goals: automate chromatin state and methylation state segmentations from pre-processed data (WIG and BED files); farm out visualization to shiny and/or Gviz



ChIP-seq realignment, extension & calls

- Realignment of third-party experiments (e.g. Rick Young or Joanna Wysocka's ChIP-seq data on SRA) is greatly abetted by a simple SRAdb + Rsubread wrapper.
- PICS is already fine for *sharp* (TF) peaks
- BCPeakR wraps BCP via Rcpp to offer a performant *broad* peak caller via R (*)
- Resulting segmented islands can be processed with chromophobe & genometricorr just like any others.



** BCPeakR, chromophobe & several other packages on github to be submitted to BioC*

Thank you

The Bioconductor core developers and the community:
Martin Morgan, Marc Carlson, Sean Davis, Herve Pages, Val Obenchain, Wei Shi, Michael Lawrence, Paul Shannon, Kasper Hansen, Evan Johnson

My colleagues, mentors, and friends at USC and afield:
Peter W. Laird, Kim Siegmund, Hui Shen, Fides Lay, Peggy Farnham, Ben Berman, Moiz Bootwalla, Toshinori Hinoue, Peter A. Jones, Giridharan Ramsingh, Akil Merchant, Preet Chaudhary, Huy Dinh, Jason Ernst, Anshul Kundaje, Leslie Cope, Jim Herman, Steve Baylin

My family: my wife Catherine and my daughter Isabel.

this list is not meant to be exhaustive, but merely exhausting!

fin