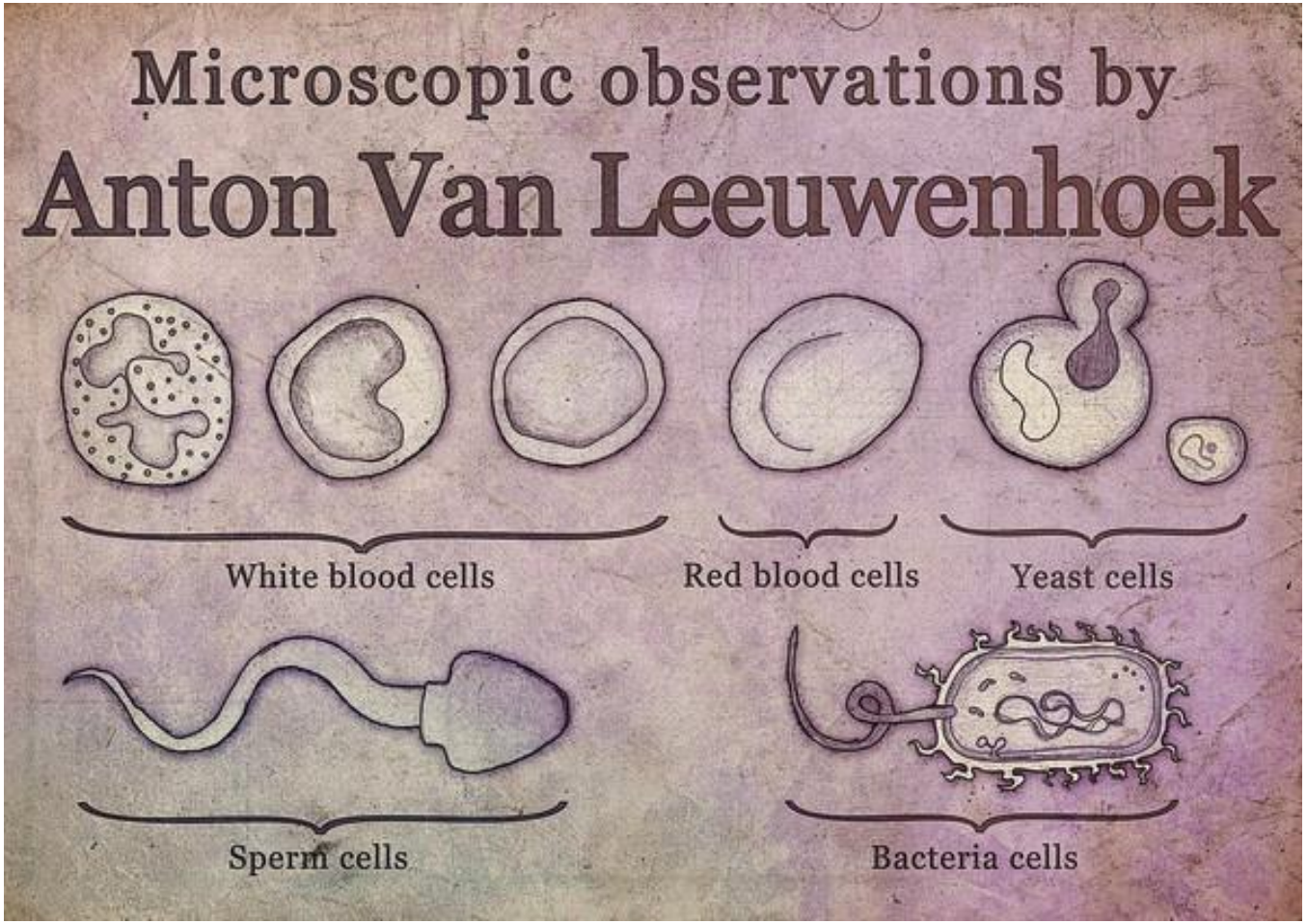
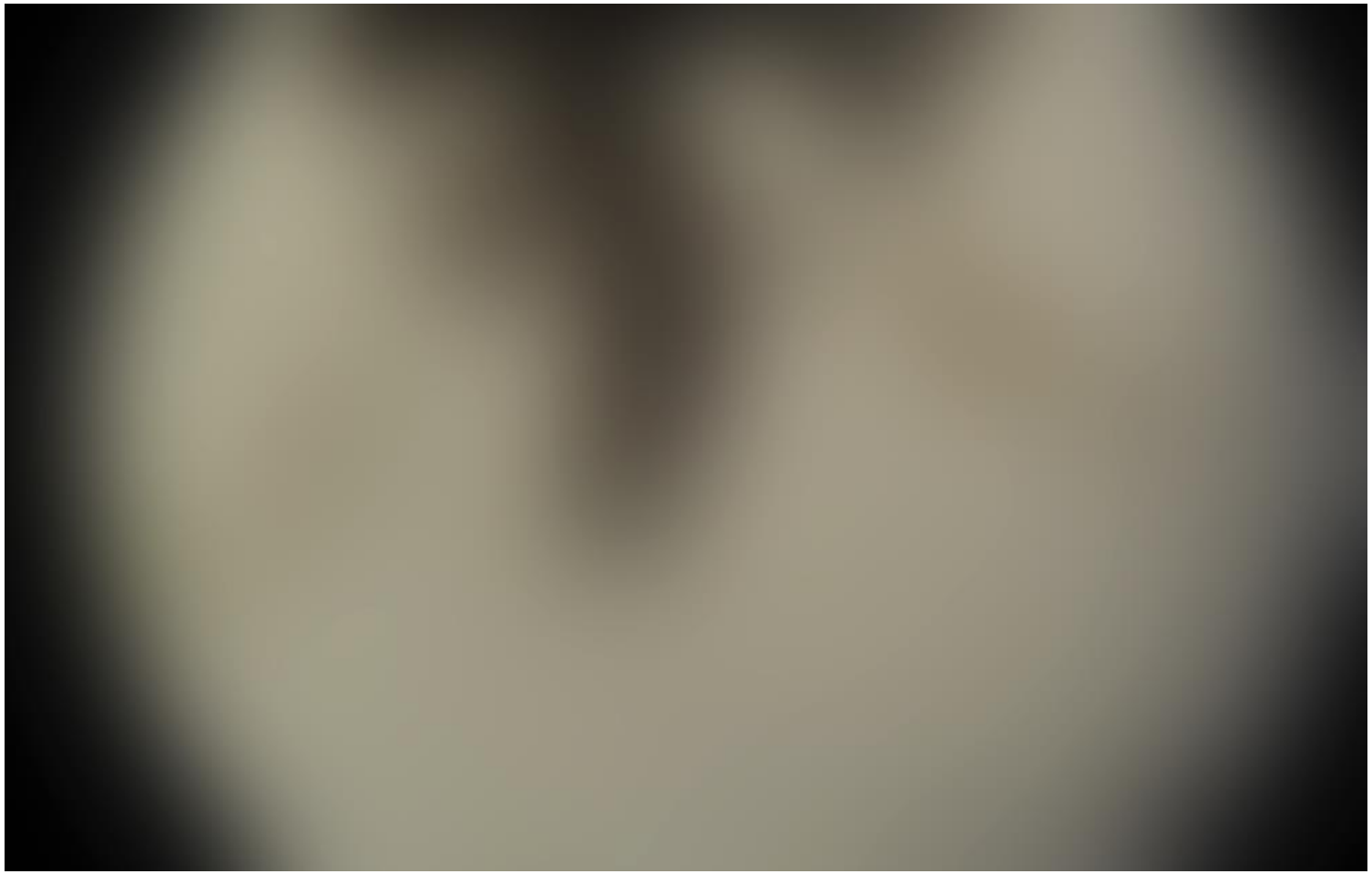
Anton Van Leeuwenhoek (1623-1723)

The “father of microbiology”

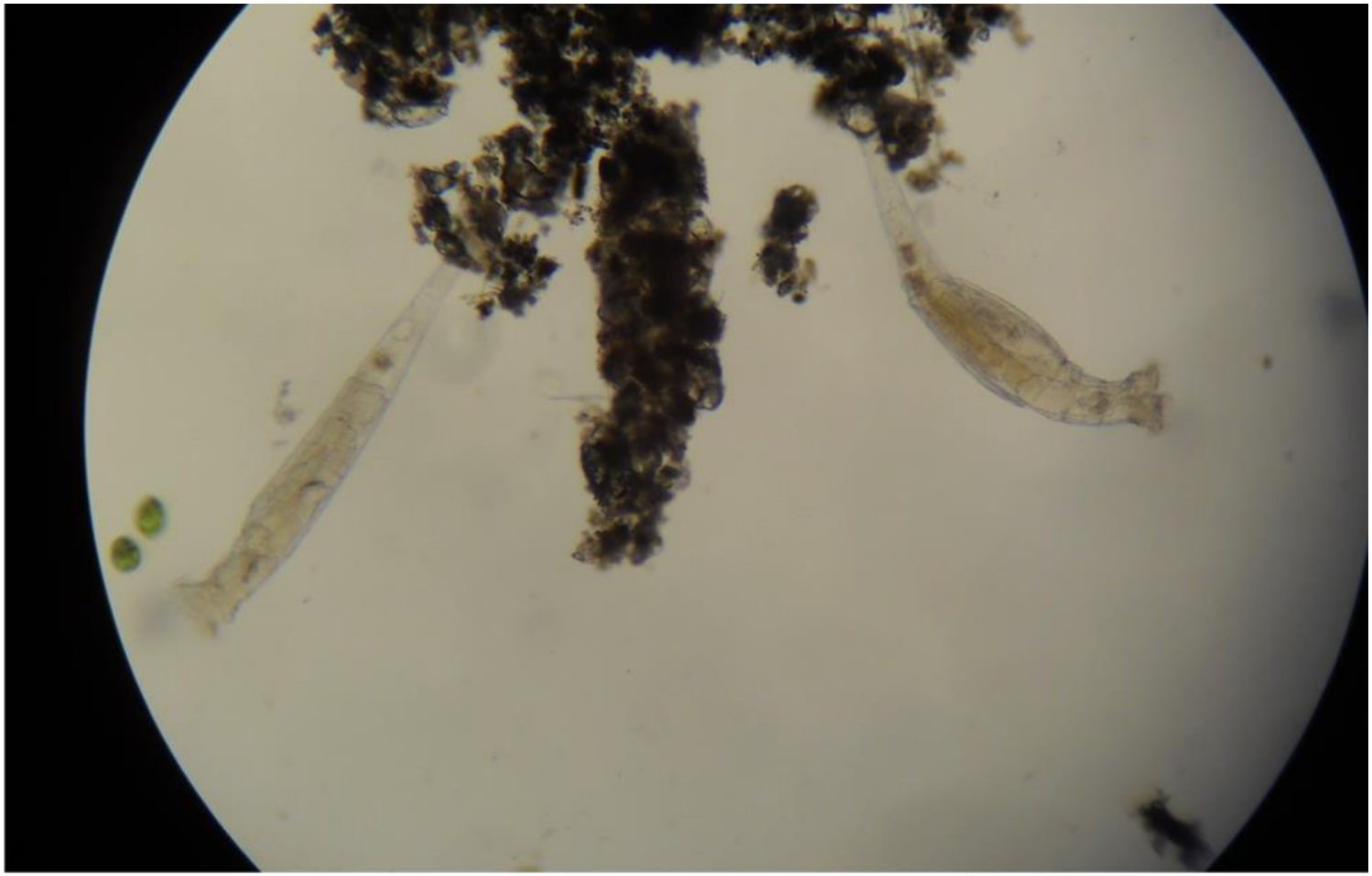
Some of his discoveries



By improving the microscope

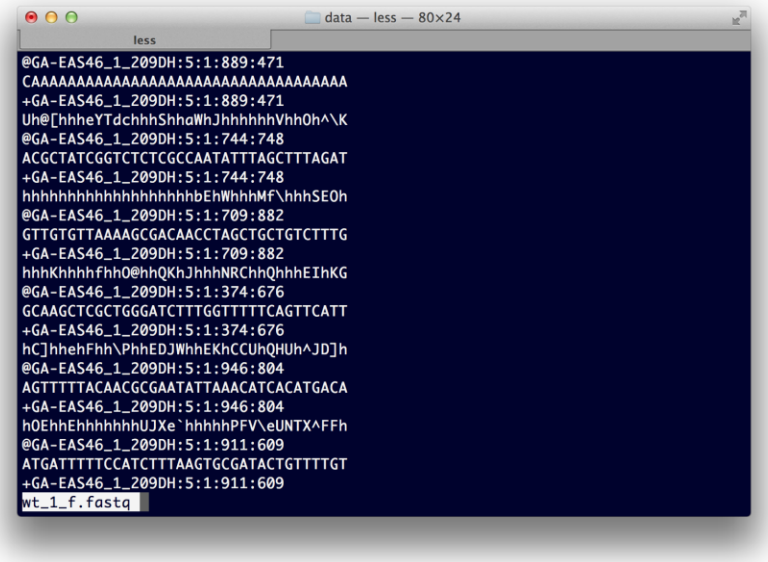


he saw what others could not



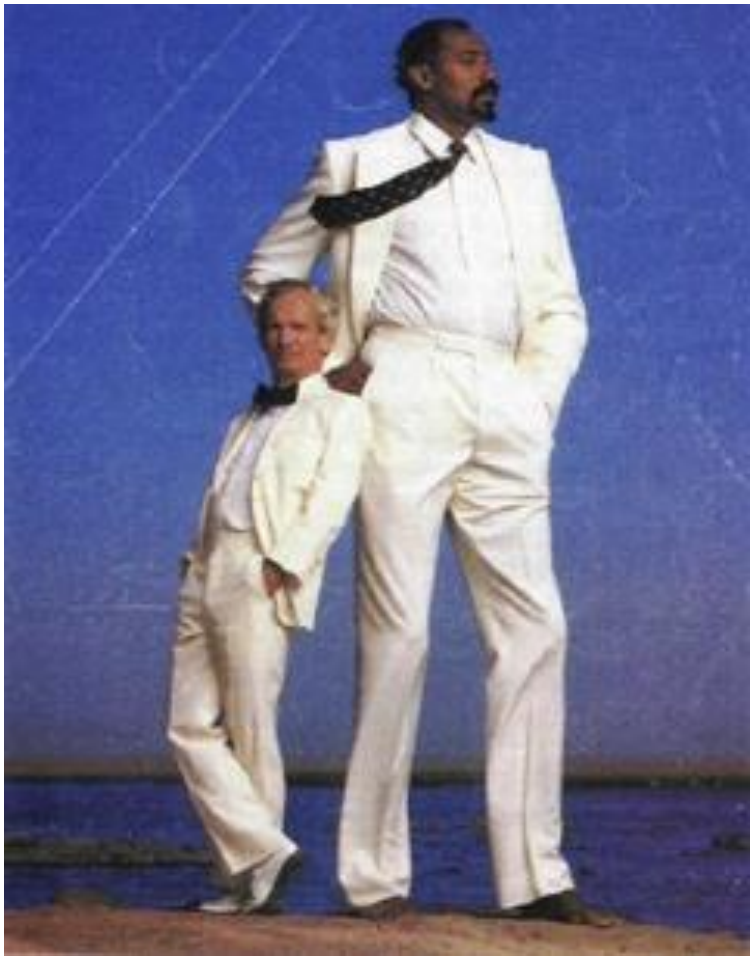
21st century version

Modern high-throughput technology

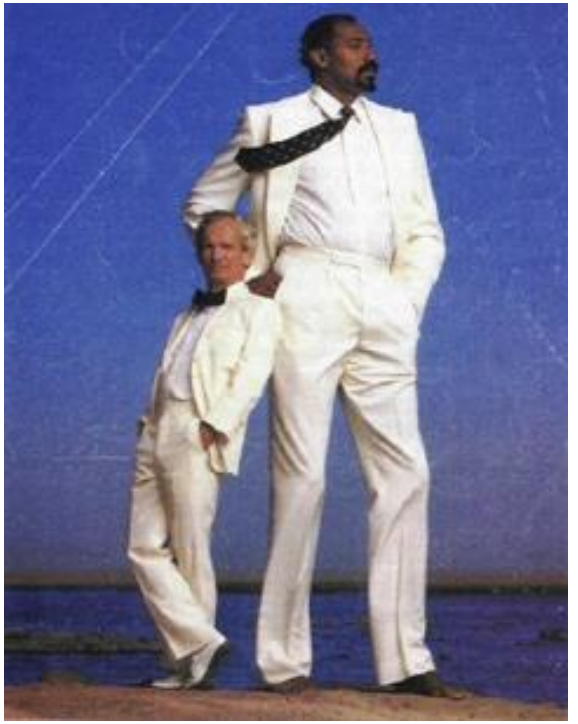


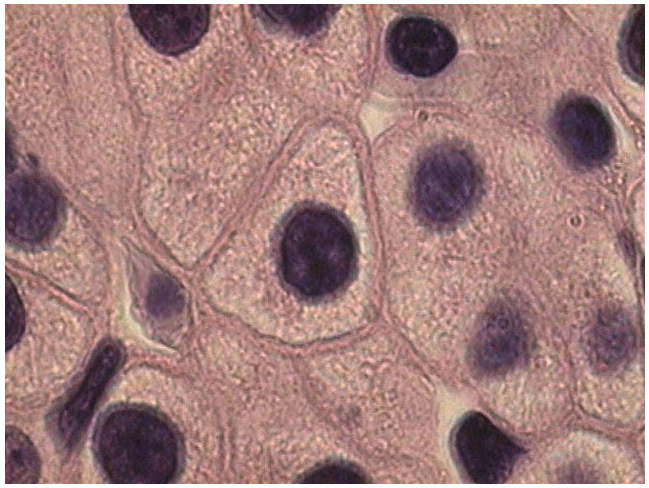
Produces complex data, not images

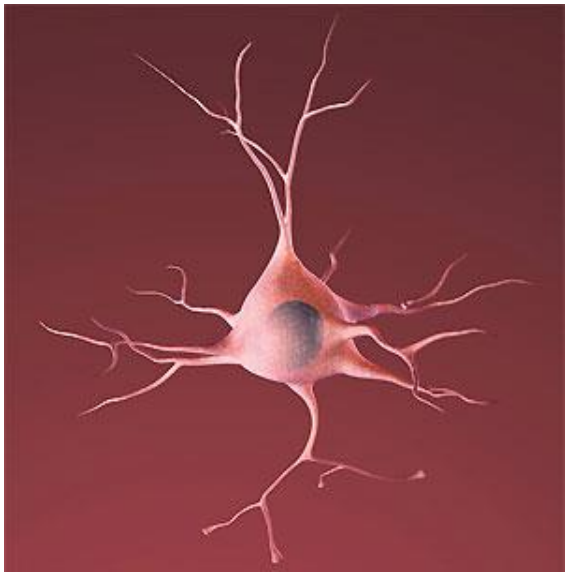
What is the basis of phenotypic variation?



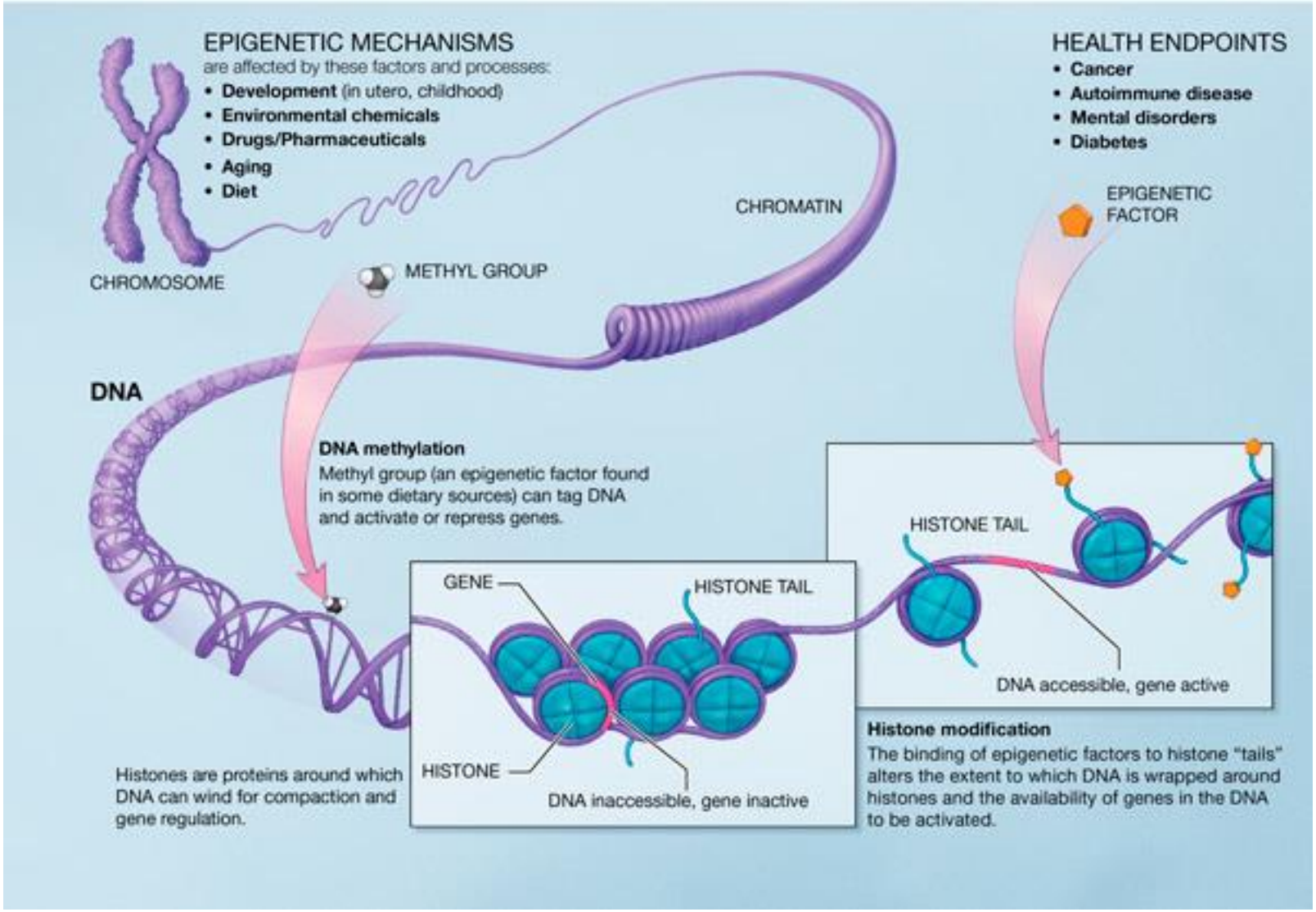
What is the basis of phenotypic variation?





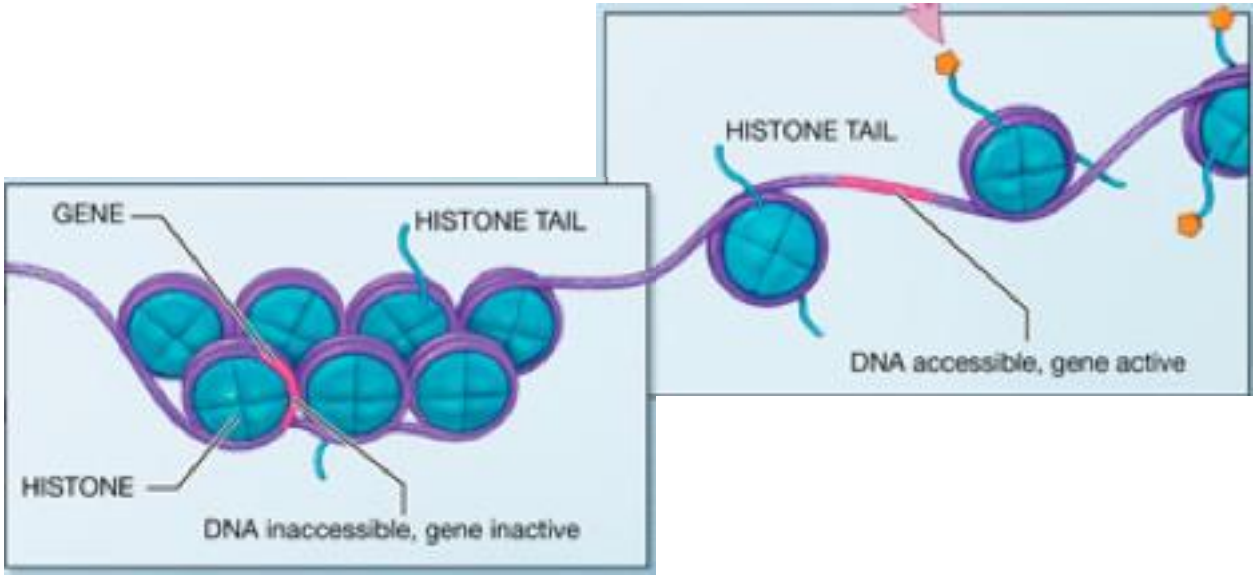


Epigenetics



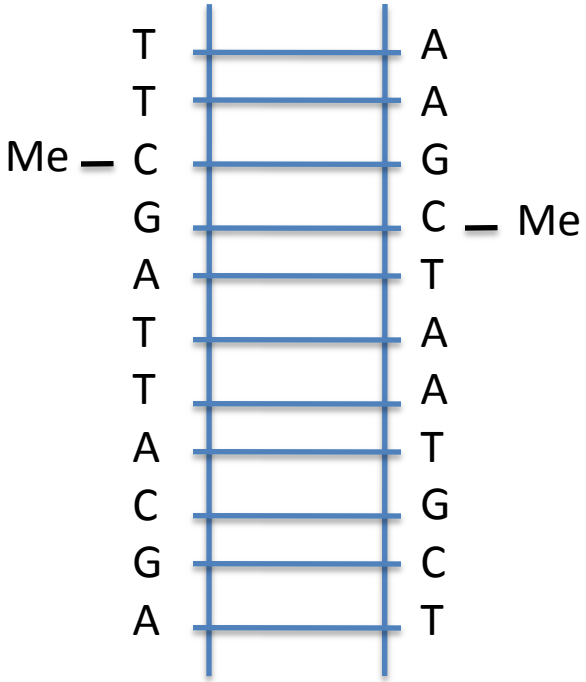
http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg

Epigenetics

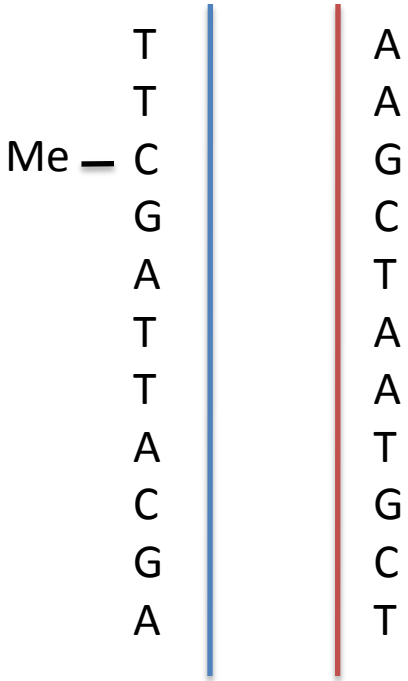


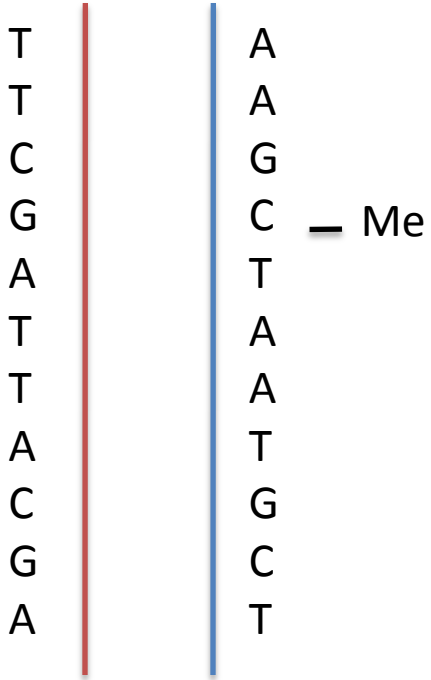
http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg

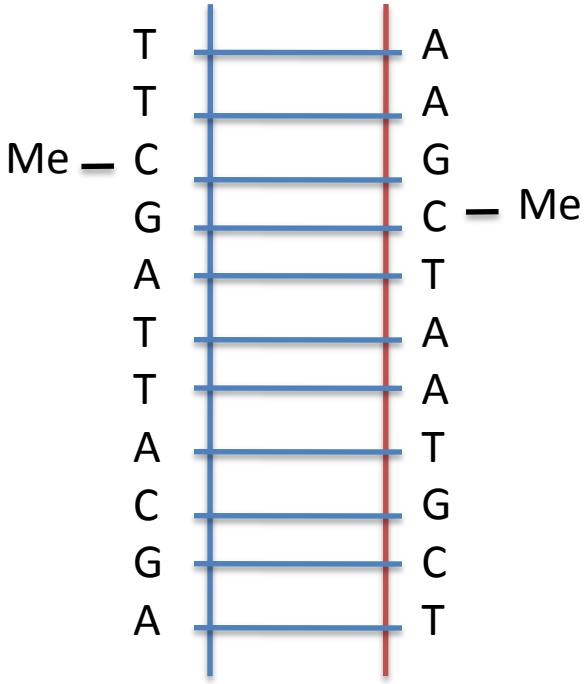
DNA Methylation

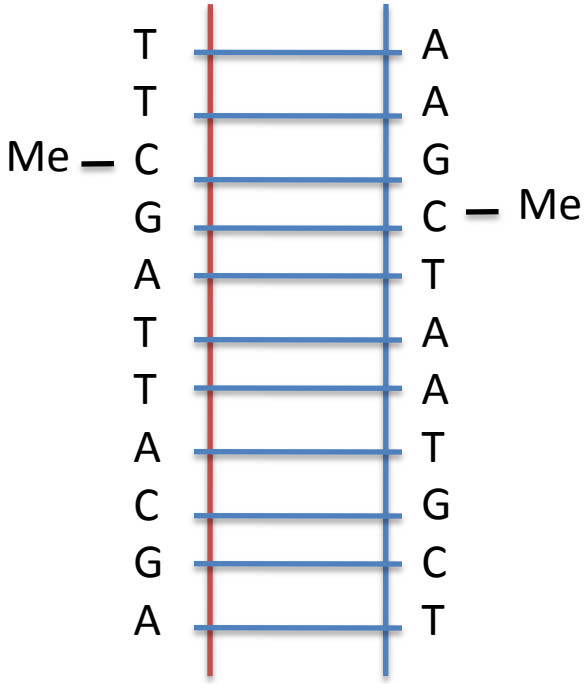



T
T
Me C
G
A
T
T
A
C
G
A
A
A
G
C Me
T
A
A
T
G
C
T

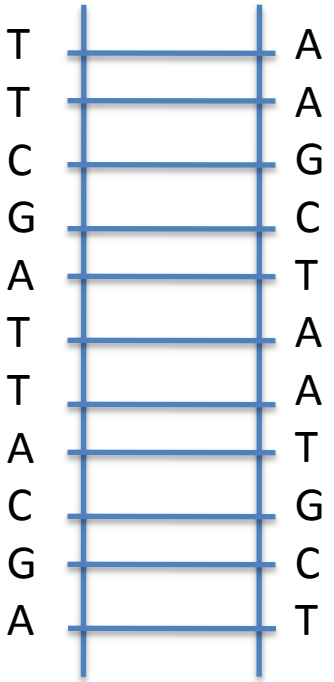

Me
T
T
C
G
A
T
T
A
C
G
A
A
A
G
C
T
A
A
T
G
C
T


T
T
C
G
A
T
T
A
C
G
A
A
A
G
C
Me
T
A
A
T
G
C
T

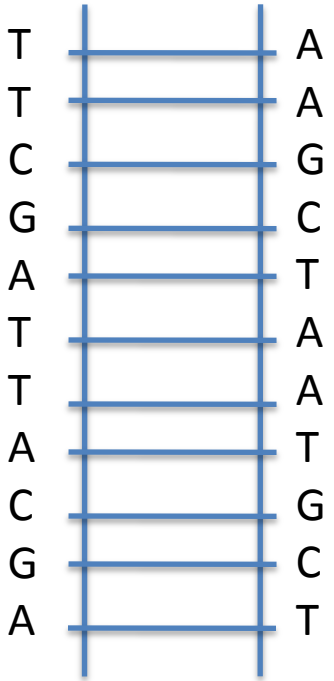

T A
T A
Me C G
G C Me
A T
T A
T A
A T
C G
G C
A T


T A
T A
Me C G
G C Me
A T
T A
T A
A T
C G
G C
A T

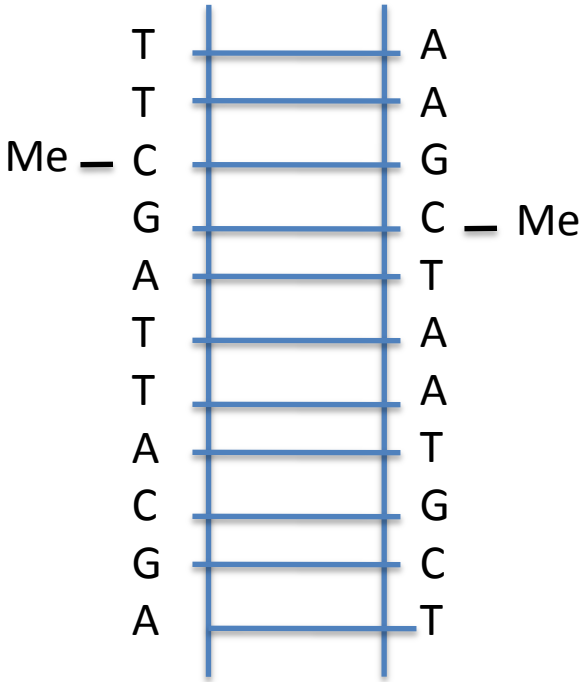
Liver



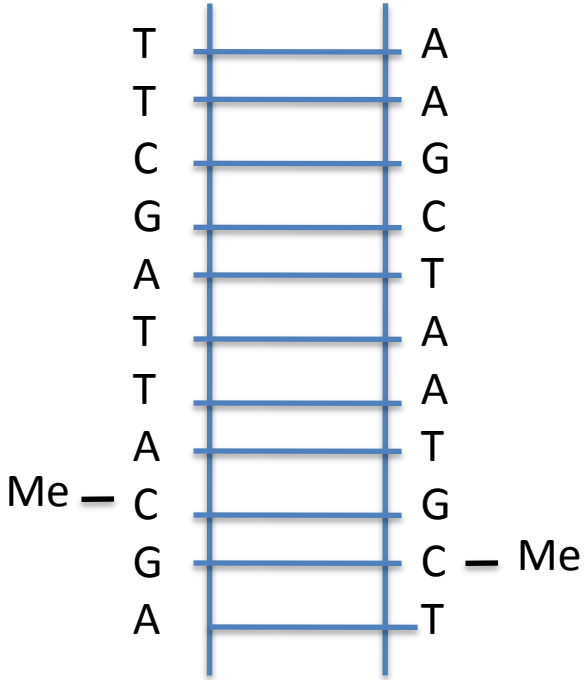
Colon



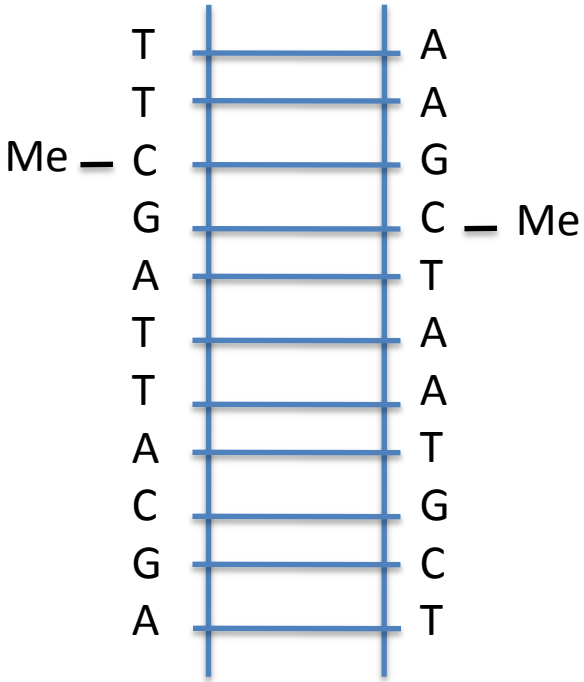
Liver

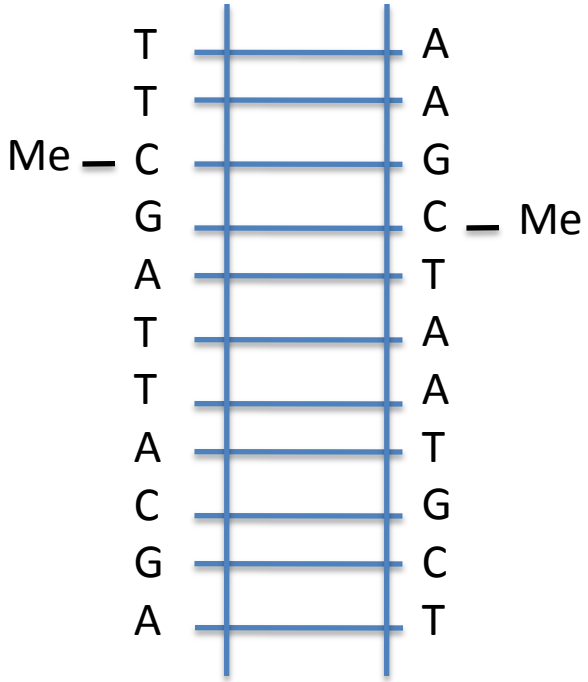

T A
T A
Me C G
G C Me
A T
T A
T A
A T
C G
G C
A T

Colon

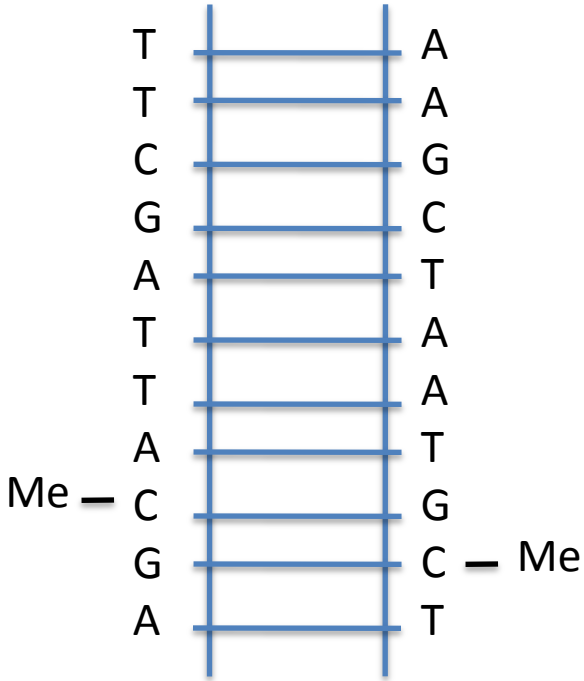

T A
T A
C G
G C
A T
T A
T A
A T
Me C G
G C Me
A T

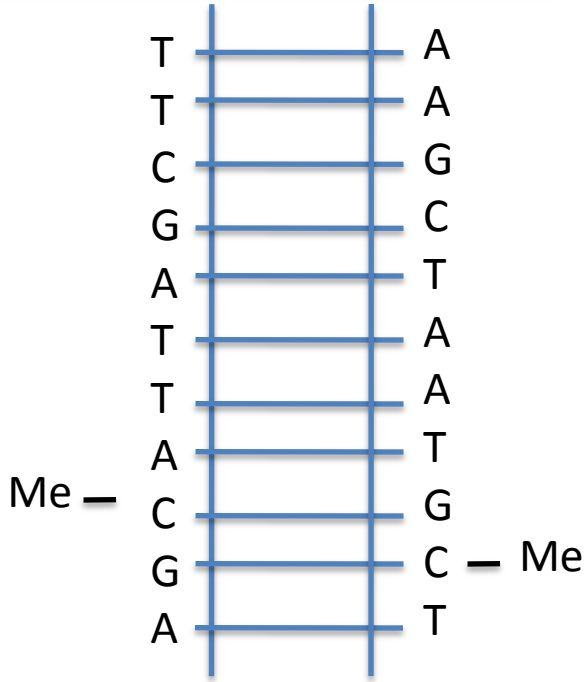
Liver


T
T
Me
C
G
A
T
T
A
C
G
A
A
A
G
C
T
A
A
T
G
C
T
Me
Me

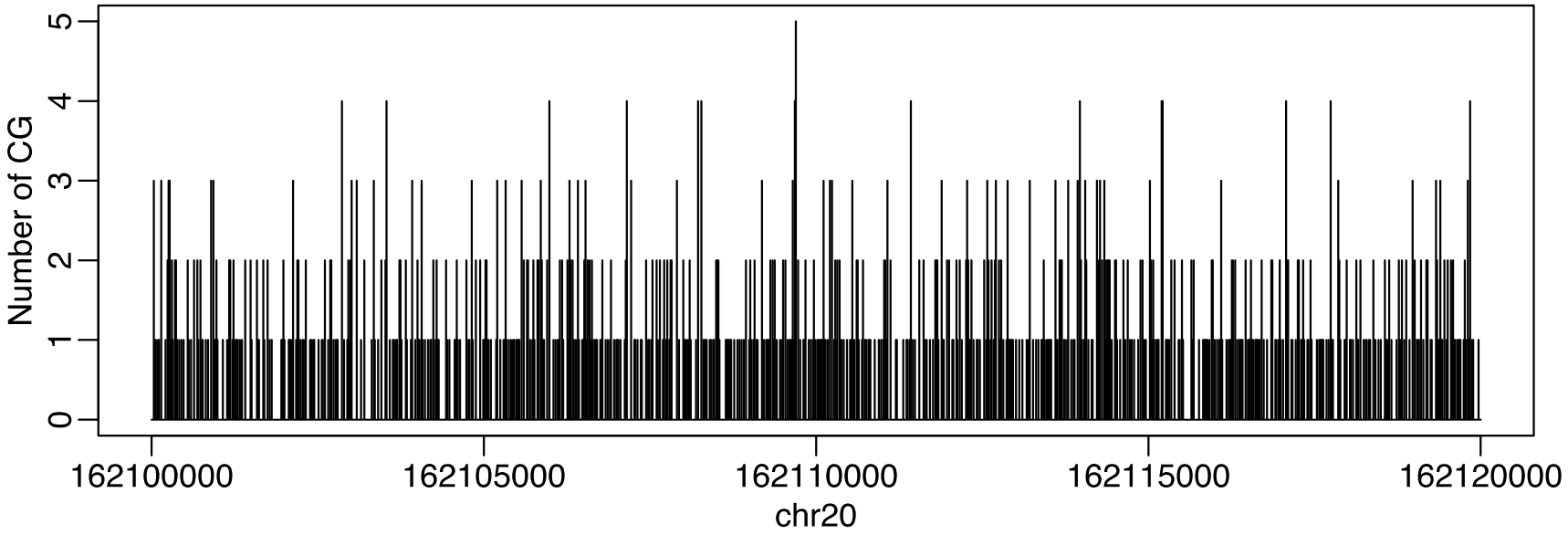

T
T
Me
C
G
A
T
T
A
C
G
A
A
A
G
C
T
A
A
T
G
C
T
Me
Me

Colon


T
T
C
G
A
T
T
A
Me
C
G
A
A
A
G
C
T
A
A
T
G
C
T
Me
Me

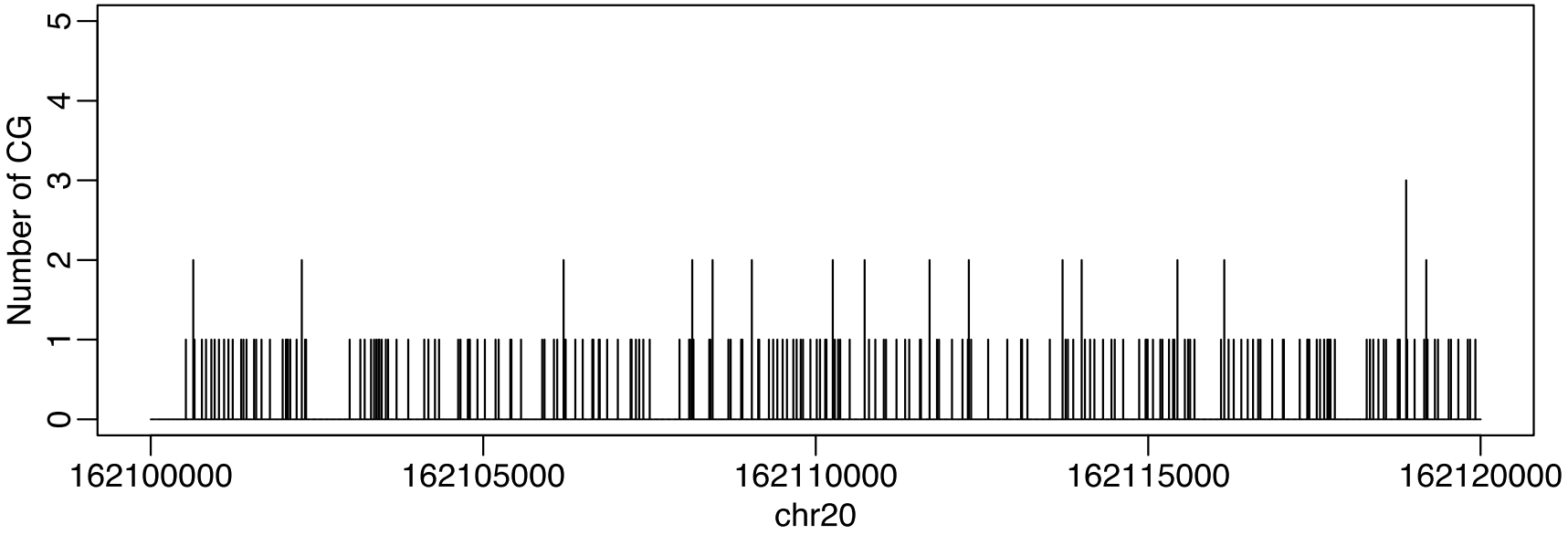

T
T
C
G
A
T
T
A
Me
C
G
A
A
A
G
C
T
A
A
T
G
C
T
Me
Me

GC counts on the genome



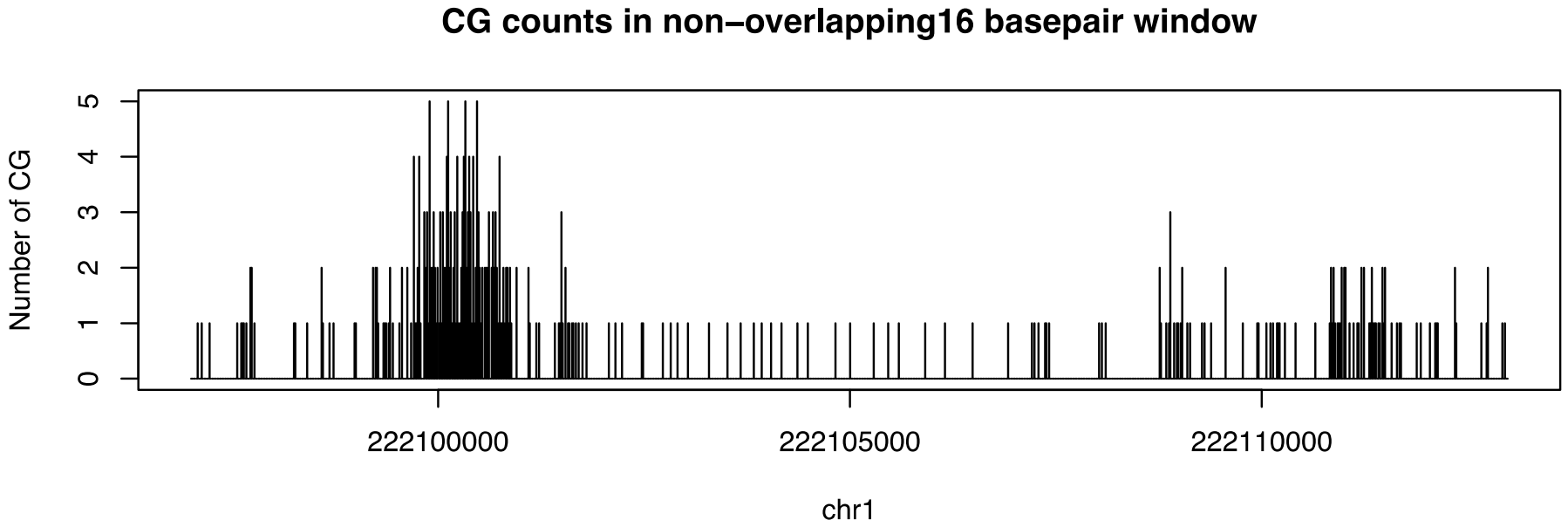
These are counts in 16 basepair bins

CpG are depleted



- These are counts in 16 basepair bins
- We see rate of about 1 in 100

CpG Islands



- But CpGs cluster into *islands* enriched near promoter

Irizarry et al. (2009) Mammalian Genome
Wu et al (2010) Biostatistics,
New illumina CpG array will use our CGI

Conventional wisdom in 2004

- Higher methylation in cancer CpG islands silence tumor suppressor genes

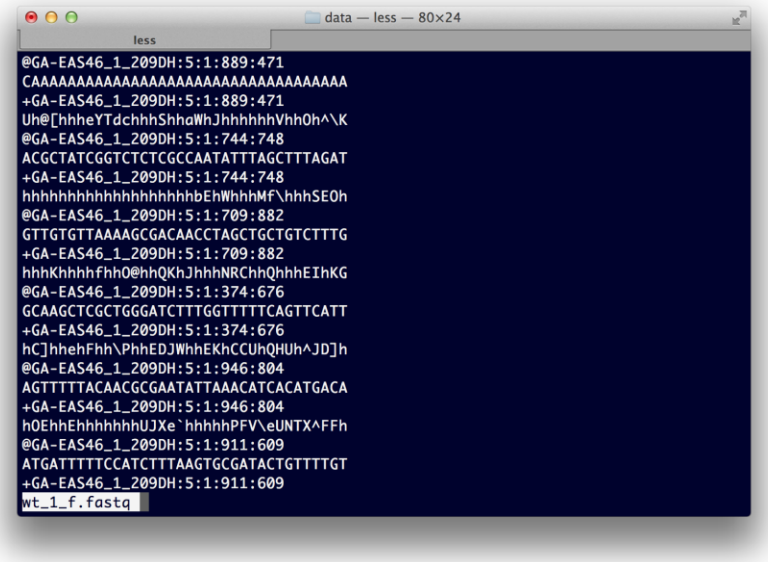
- Globally cancer cells have less methylation

  High throughput measurement permitted a more comprehensive view

High-throughput technologies

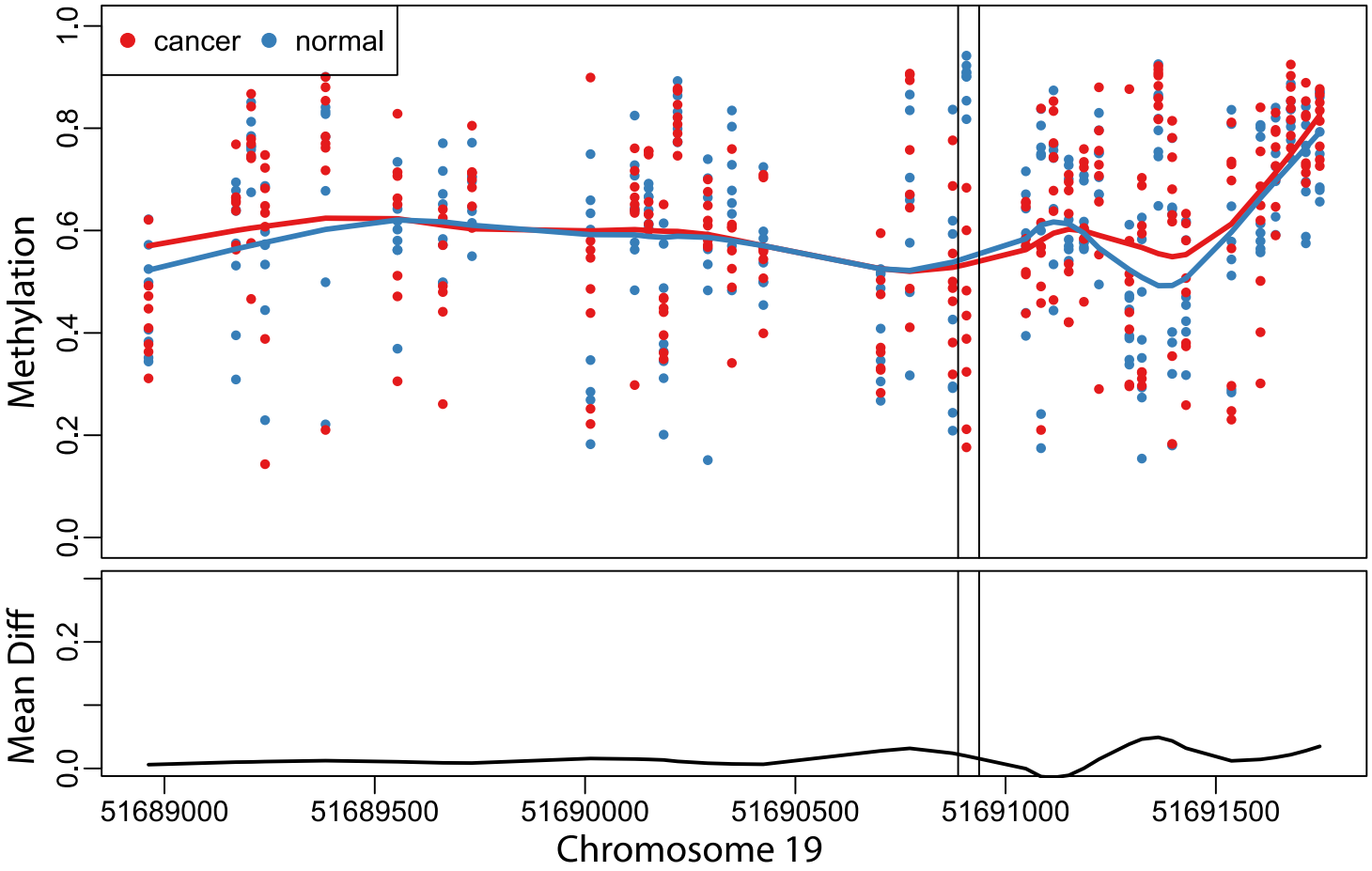
21st century version

Modern high-throughput technology



Produces complex data, not images

Genomic traceplot



Microarray data after much preprocessing

General Model

Baseline methylation level

Effect at j-th position

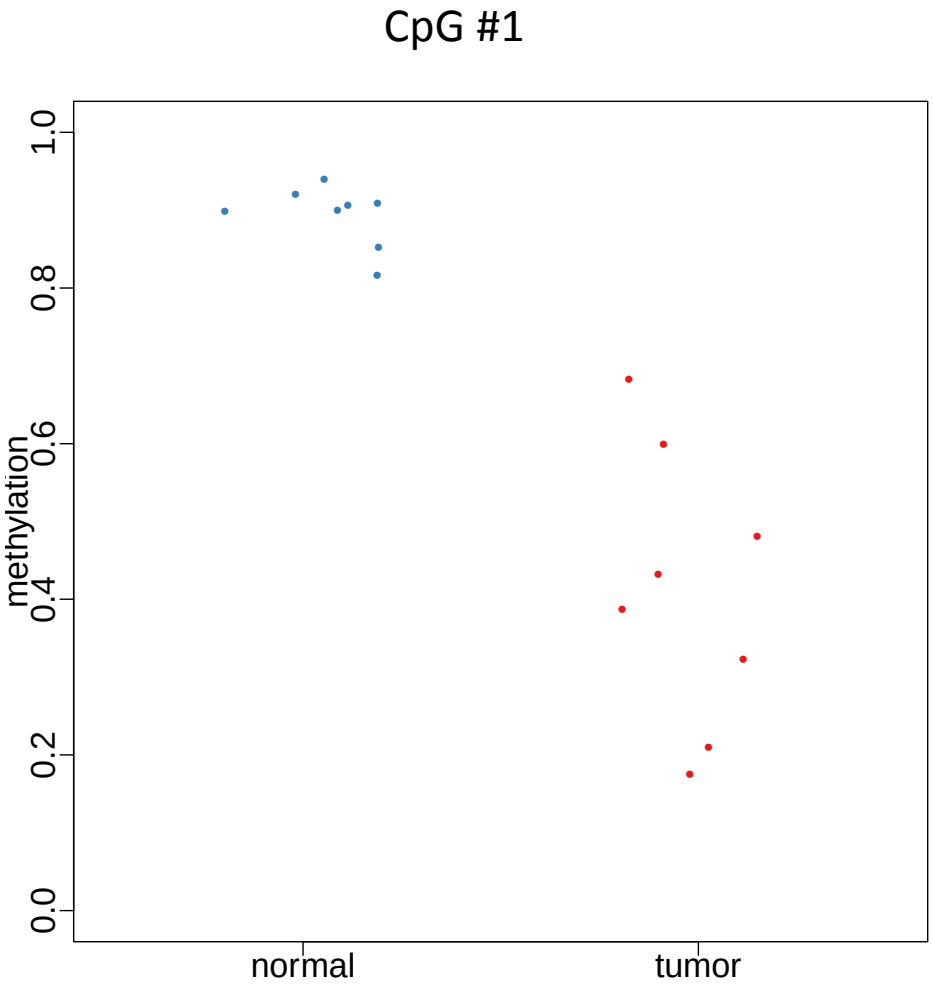
Measurement error

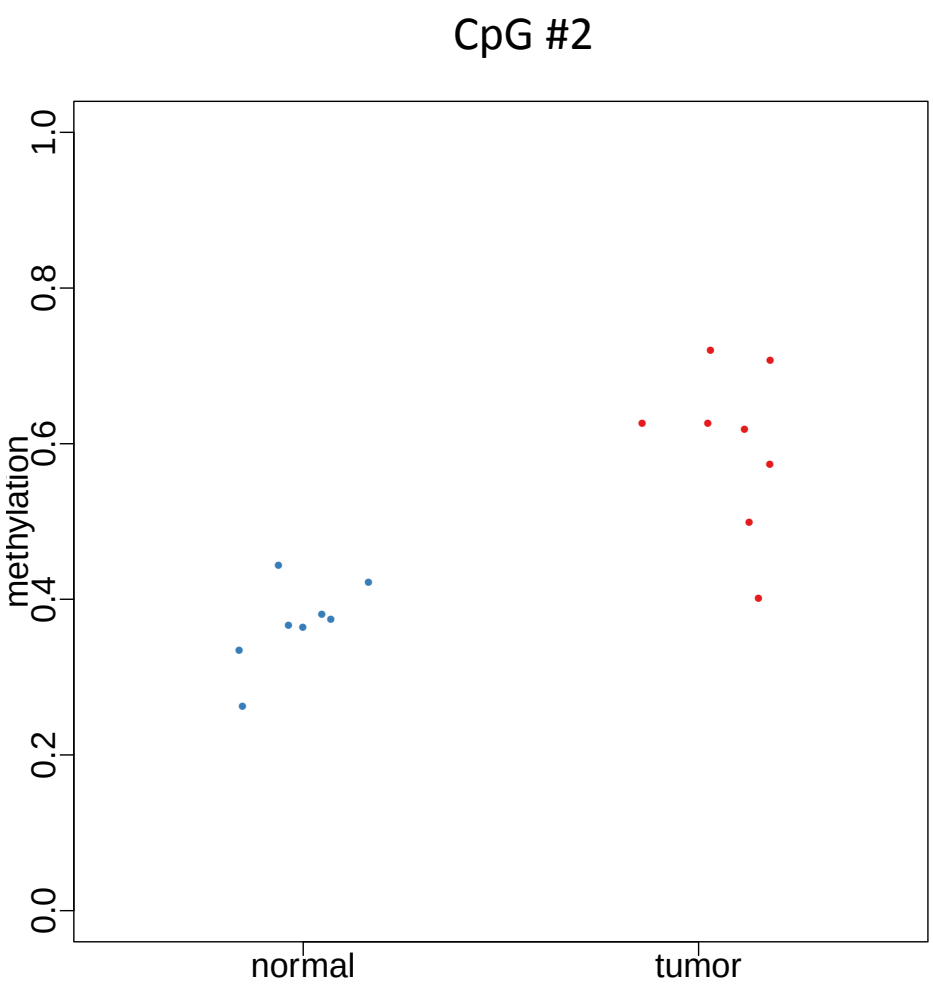
$$Y_{ij} = b_0(l_j) + X_i b_1(l_j) + e_{ij}$$

Observed Data

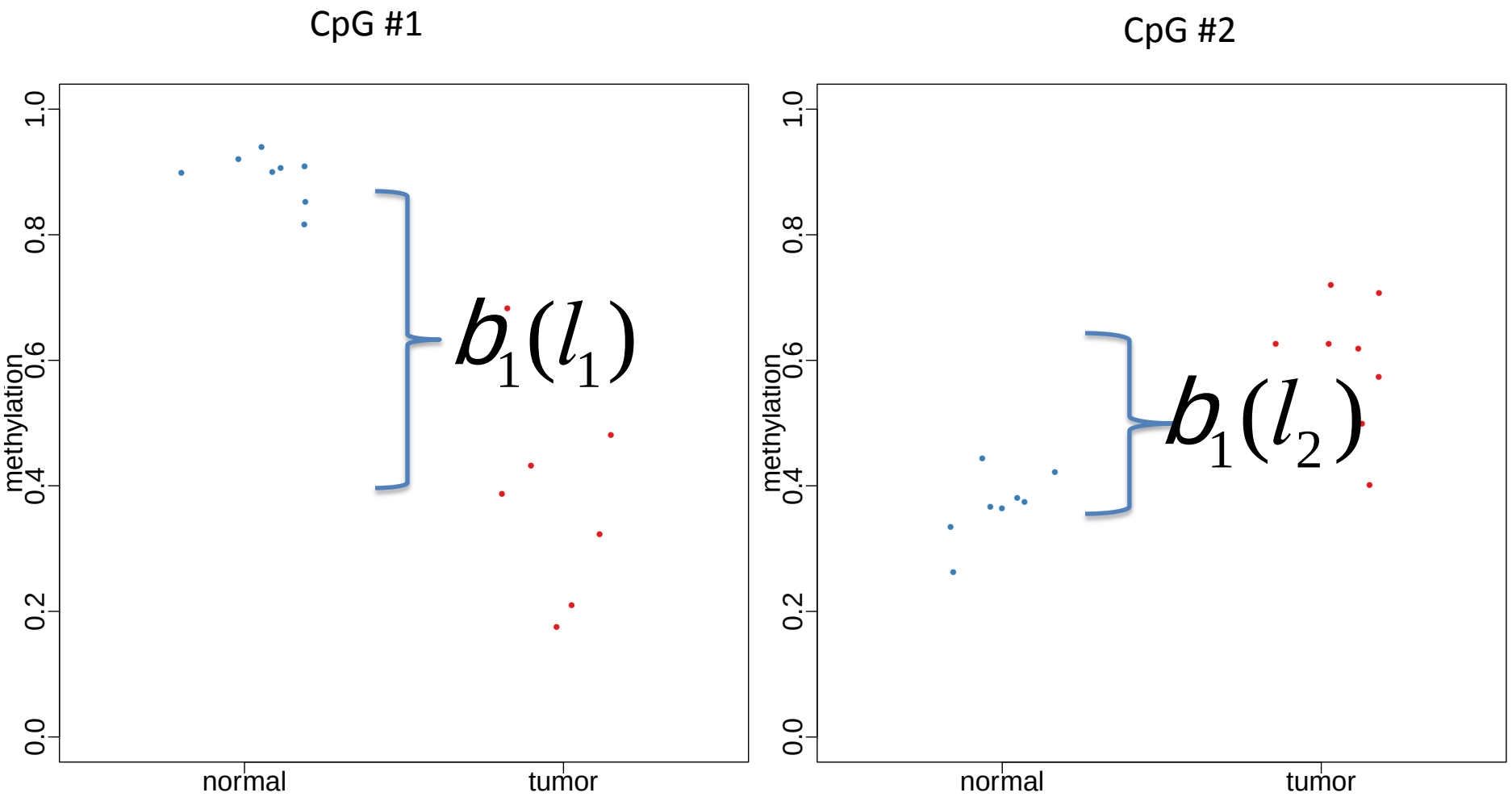
Outcome of interest

Do we trust single measurements?

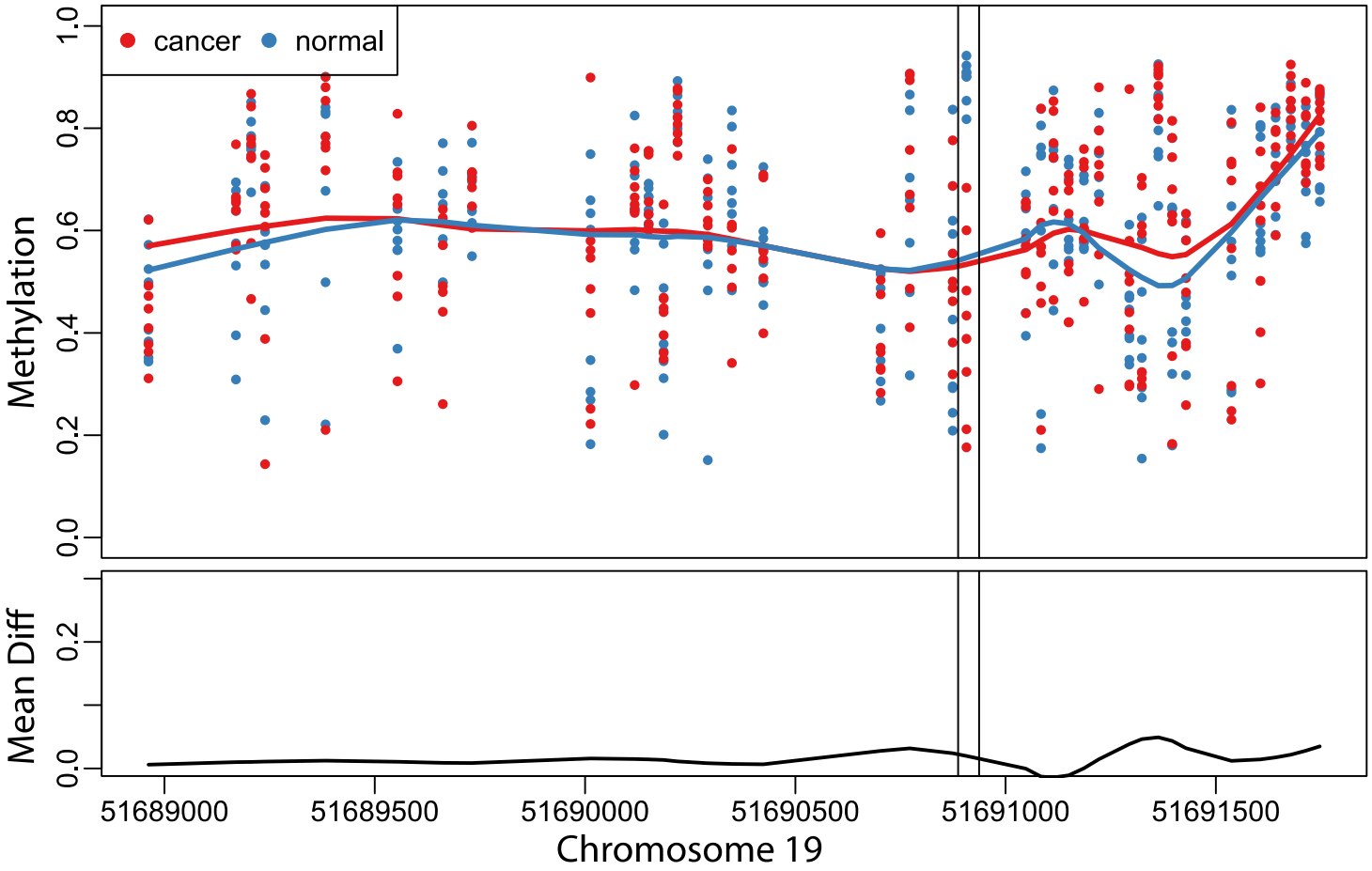




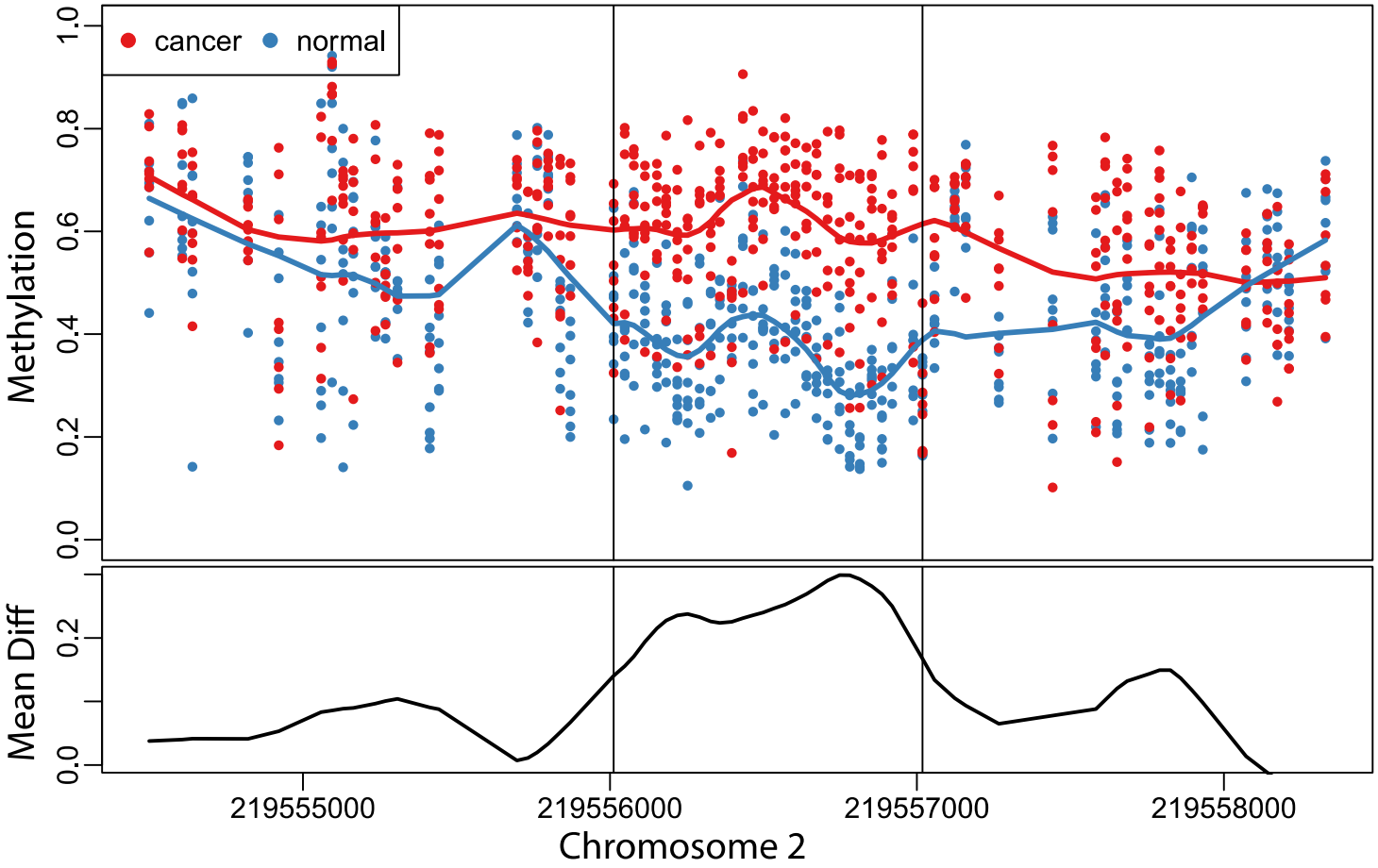
Do we trust single measurements? Note X is 1 (cancer) or 0 (normal)



CpG #1


cancer
normal
Methylation
0.0
0.2
0.4
0.6
0.8
1.0
Mean Diff
0.0
0.2
51689000
51689500
51690000
51690500
51691000
51691500
Chromosome 19

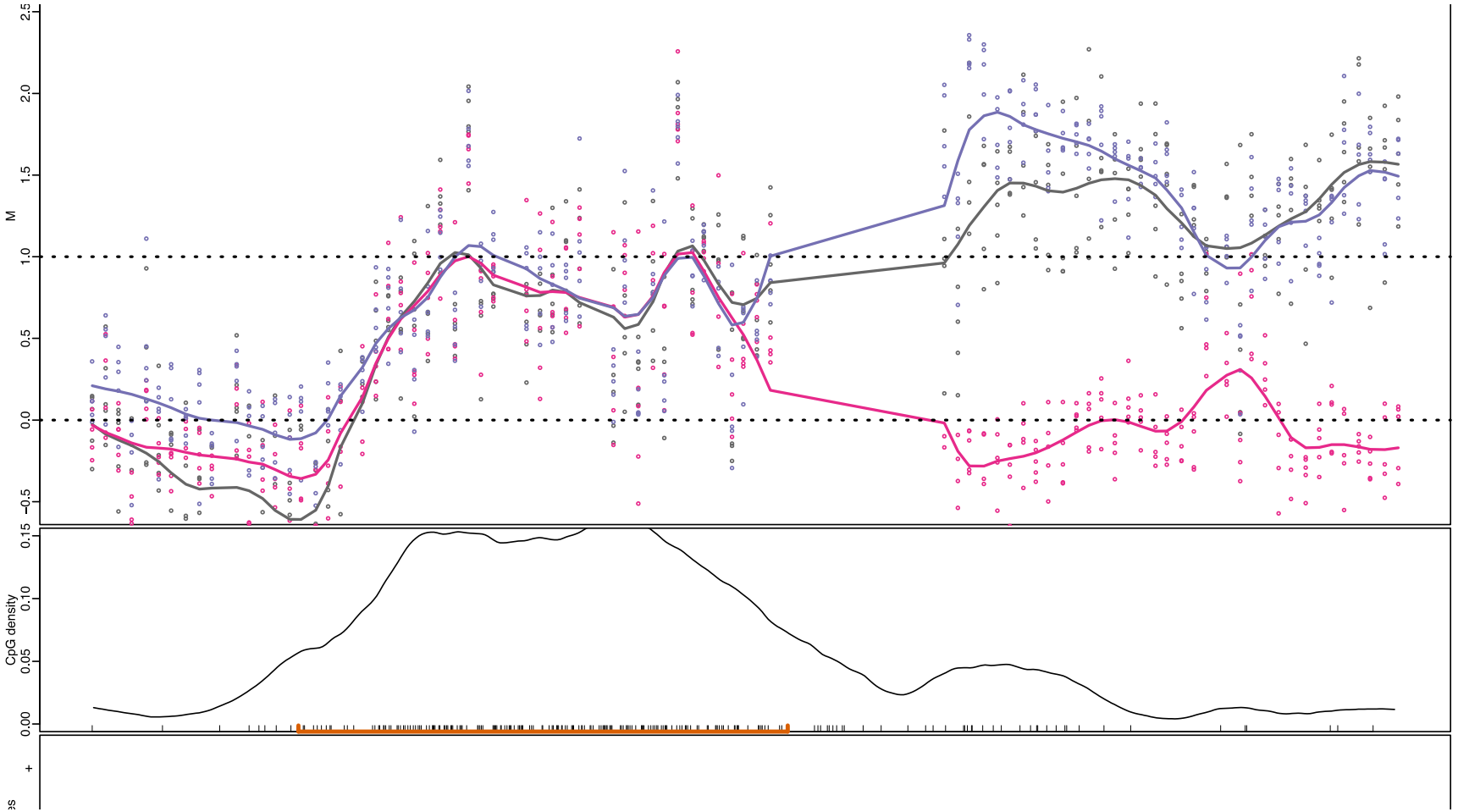
CpG #2


cancer
normal
Methylation
0.0
0.2
0.4
0.6
0.8
1.0
Mean Diff
0.0
0.2
219555000
219556000
219557000
219558000
Chromosome 2

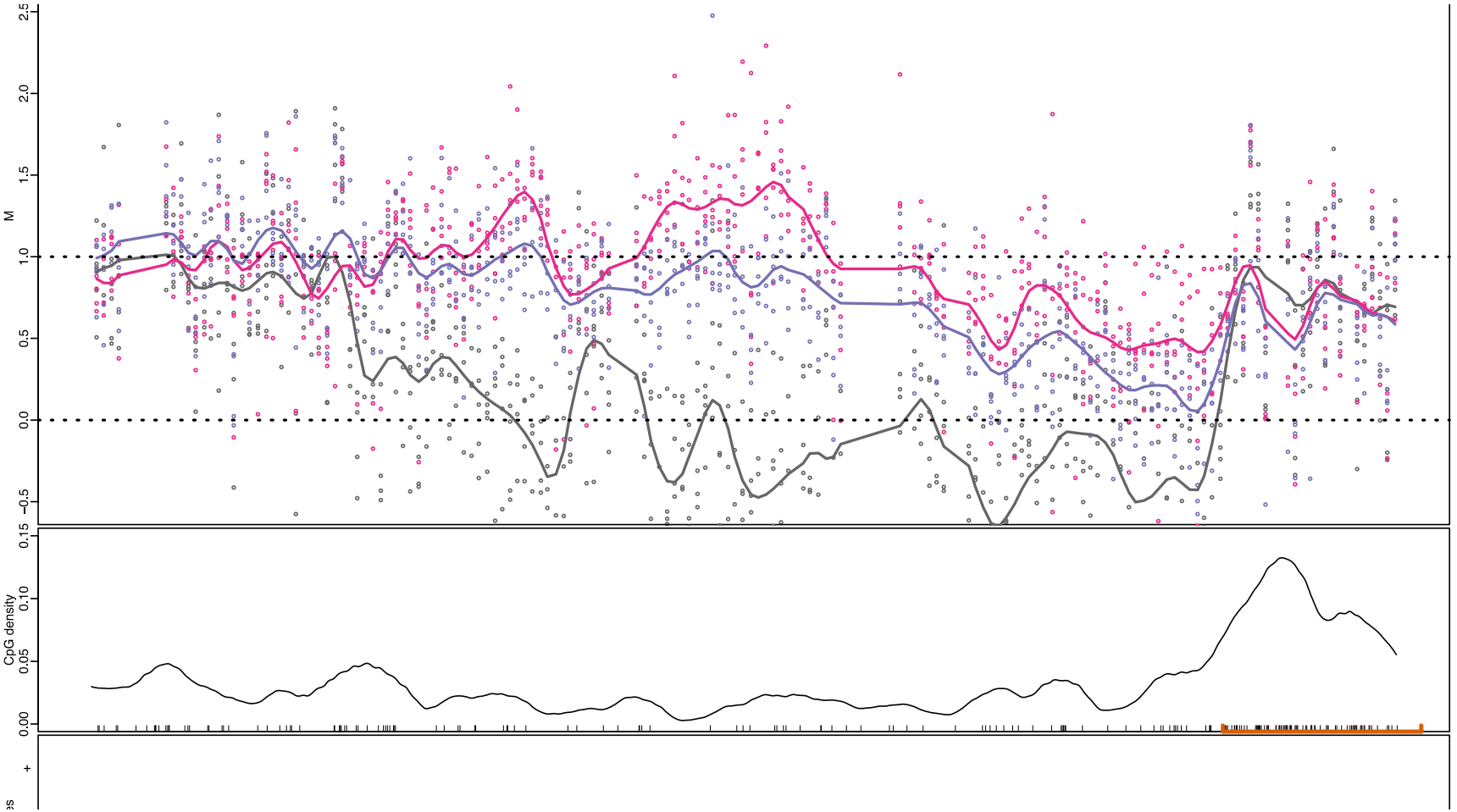
Scanning the genome lead to biological discoveries

Irizarry et al (2009) Nature Genetics

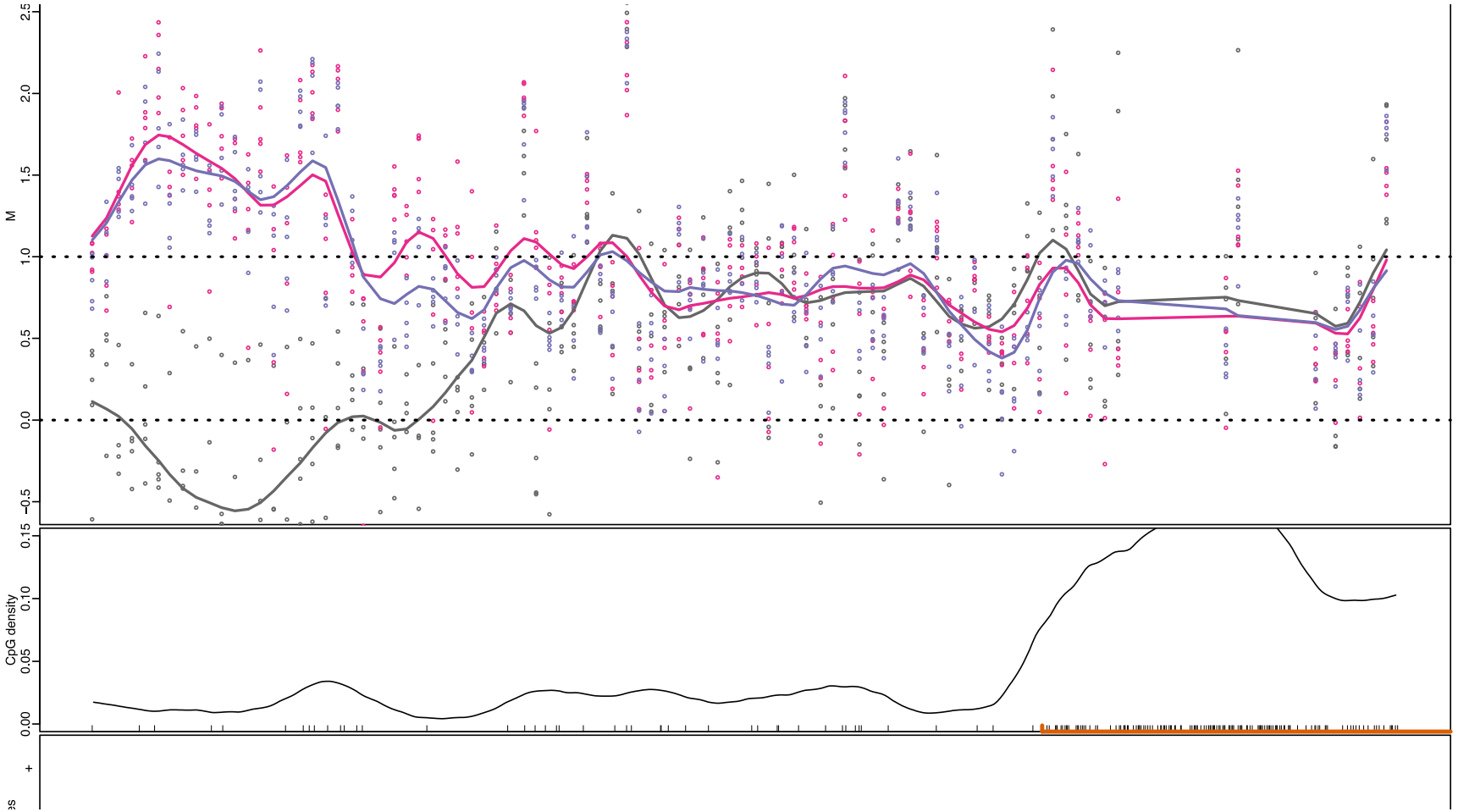
Visual examination



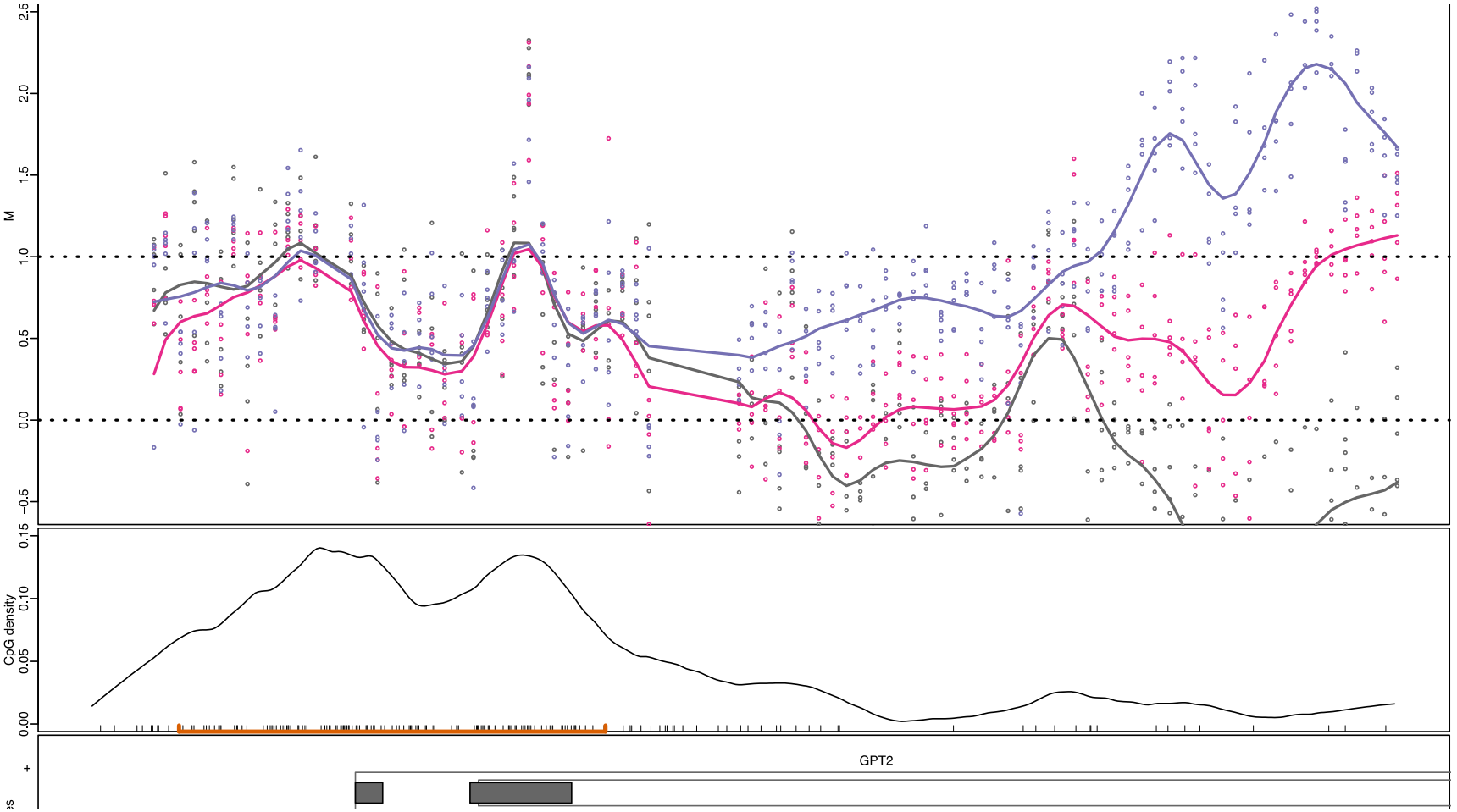
Visual examination



Visual examination

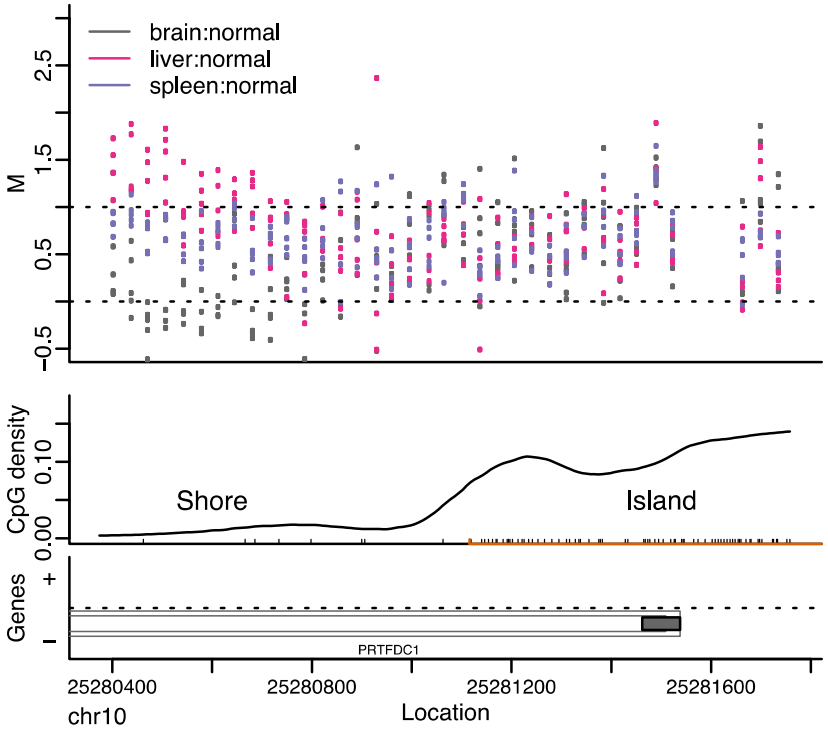


Visual examination



21st century version

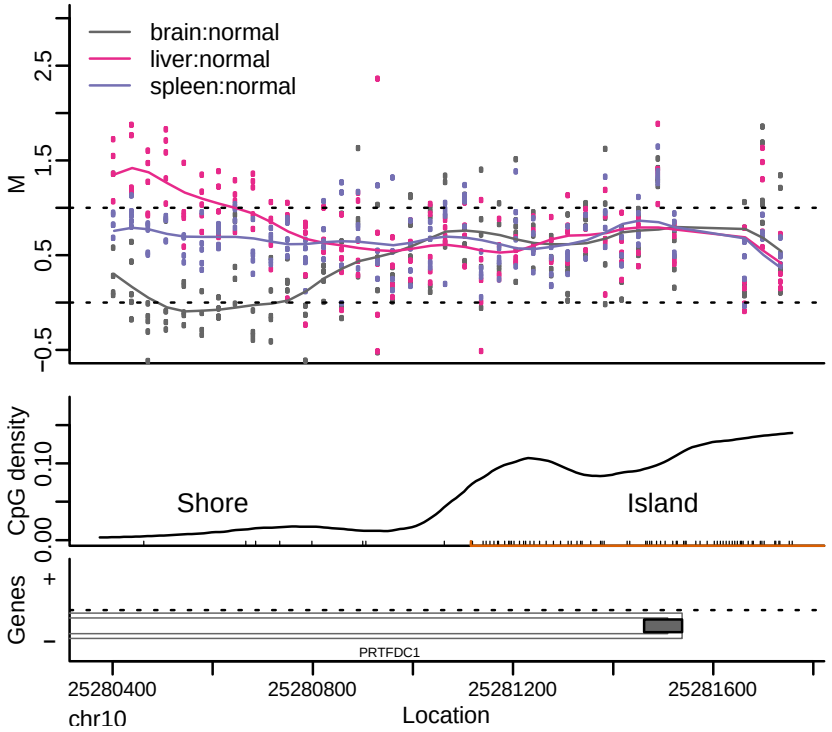
Modern high-throughput technology



My work has helped bring data into focus

21st century version

Modern high-throughput technology

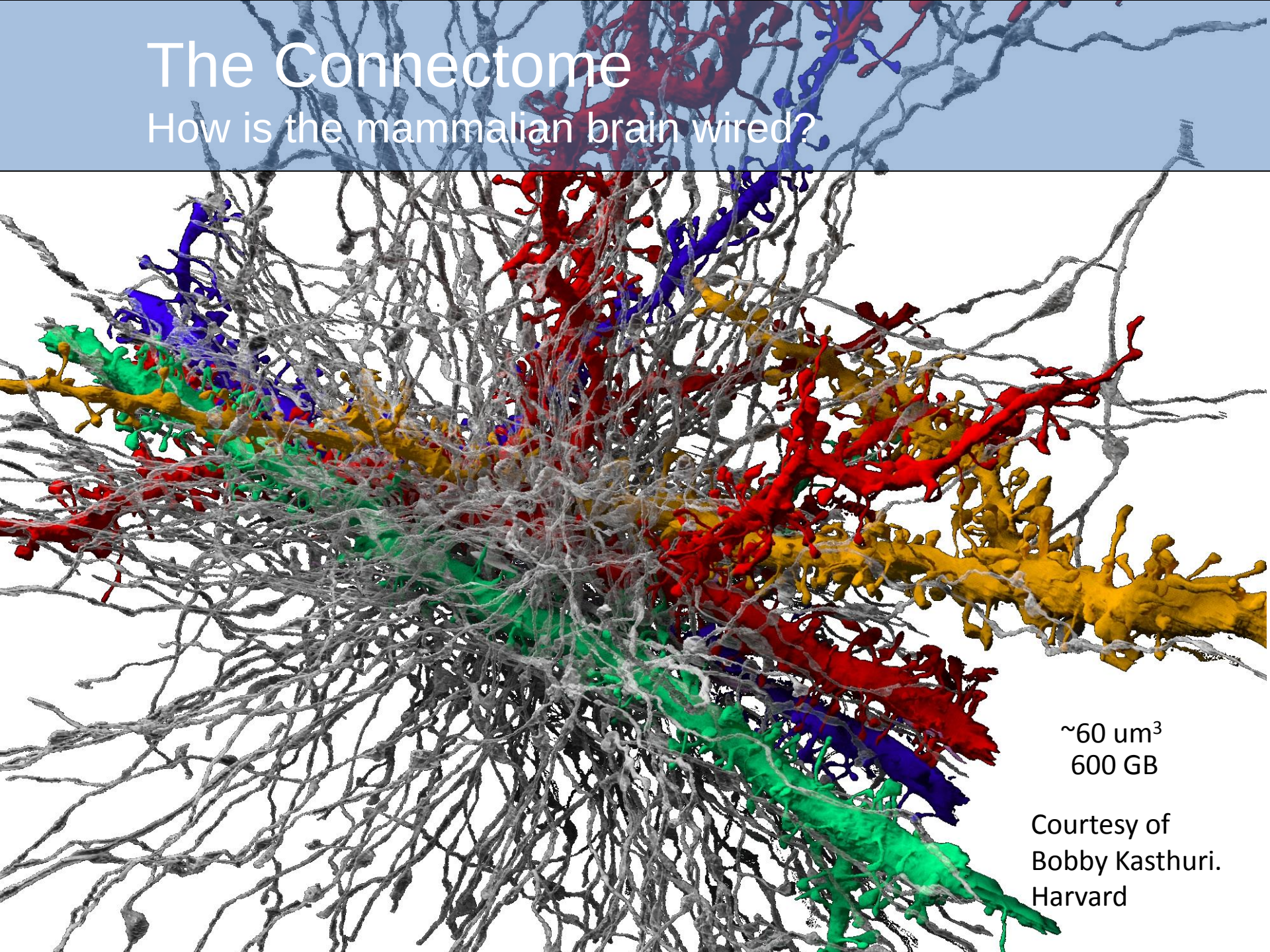


My work has helped bring data into focus

Connectomics

The Connectome

How is the mammalian brain wired?

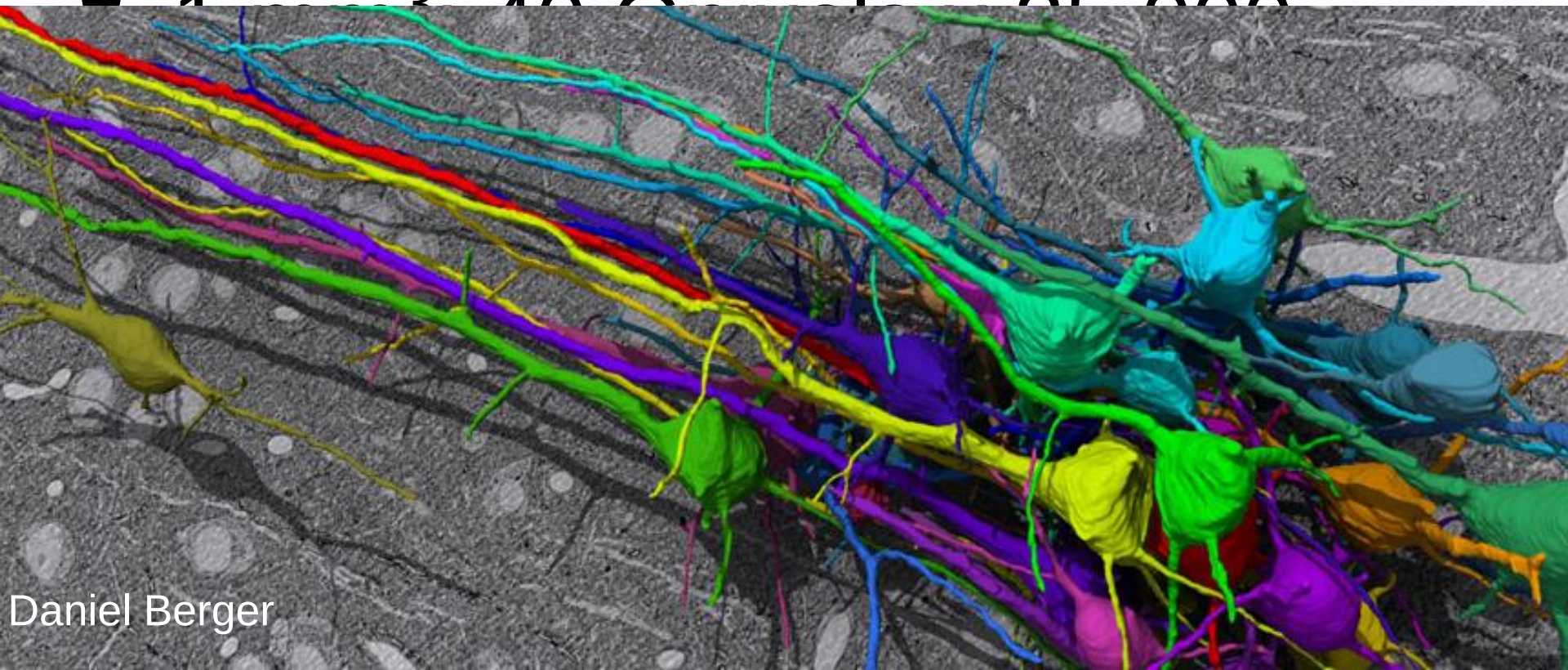


~60 um^3
600 GB

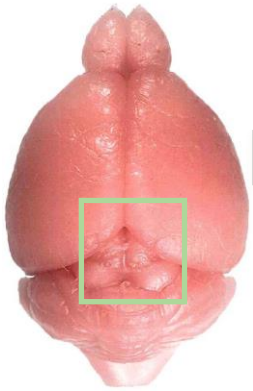
Courtesy of
Bobby Kasthuri.
Harvard

The Data Challenge

- Pixel resolution: 3-5 nm; Slice thickness: 30-50 nm
- 1 mm^3, 40 Gvoxels, ~25,000

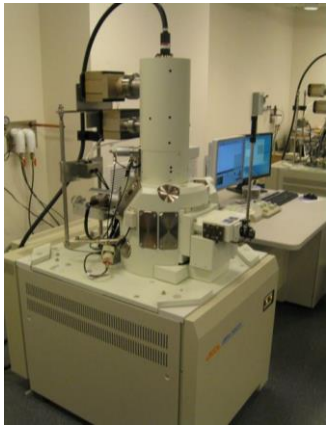


Connectome Workflow

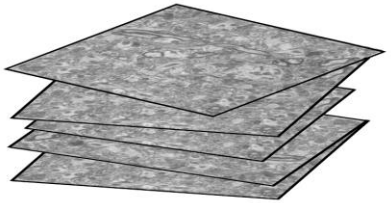


Cutting

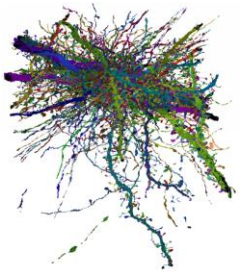
Imaging



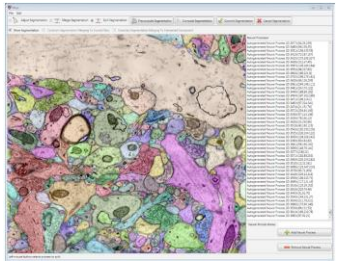
Alignment & Registration



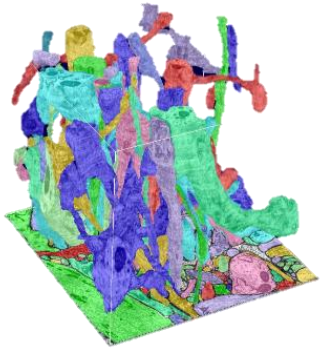
Analysis



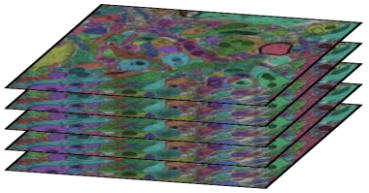
Proof Reading



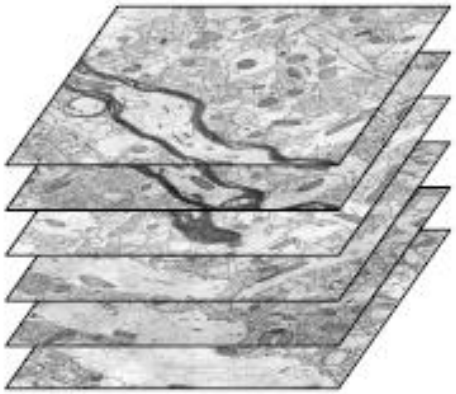
Visualization



Segmentation

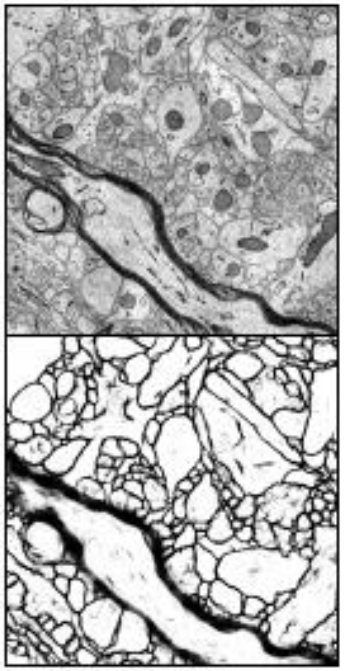


Automatic Segmentation



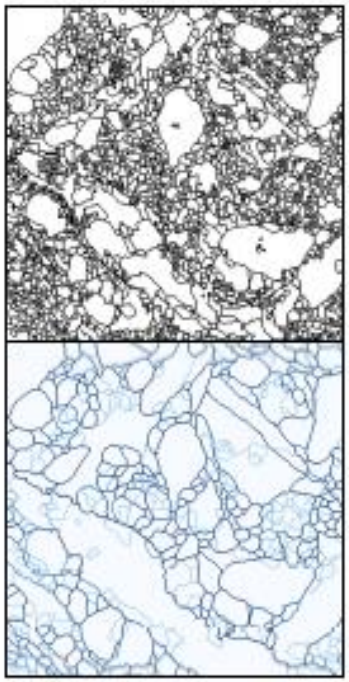
EM Image Block

Aligned input images.



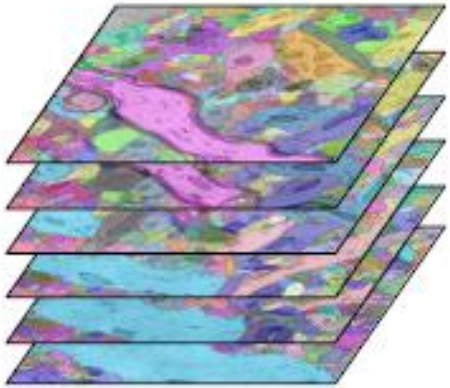
Classification

Pylearn2 [5] and Maxout [6] deep learning networks are used to classify membrane pixels.



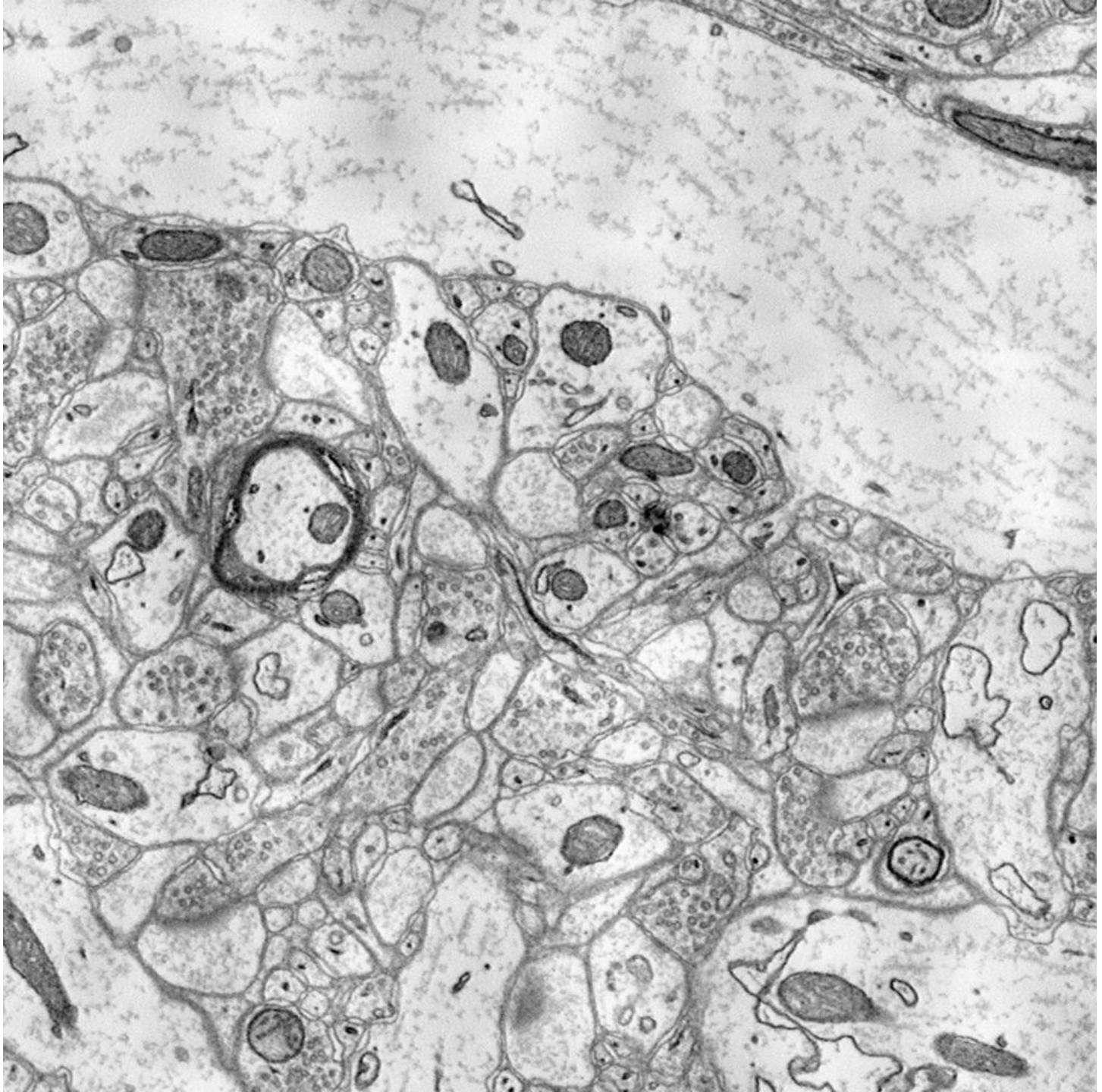
2d Segmentation

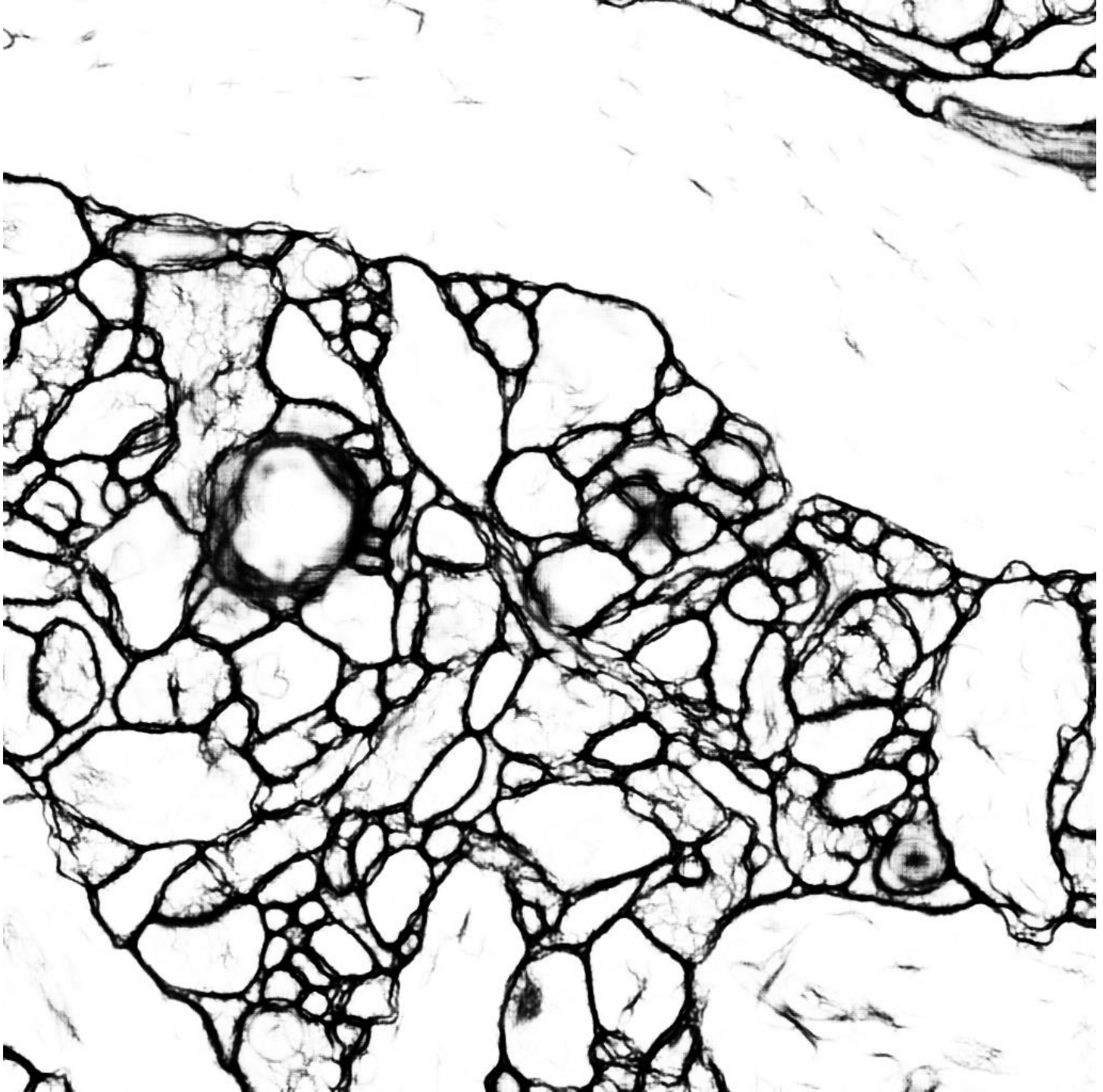
Segmentations (upper) or multi-segmentations (lower) are generated with the watershed algorithm.

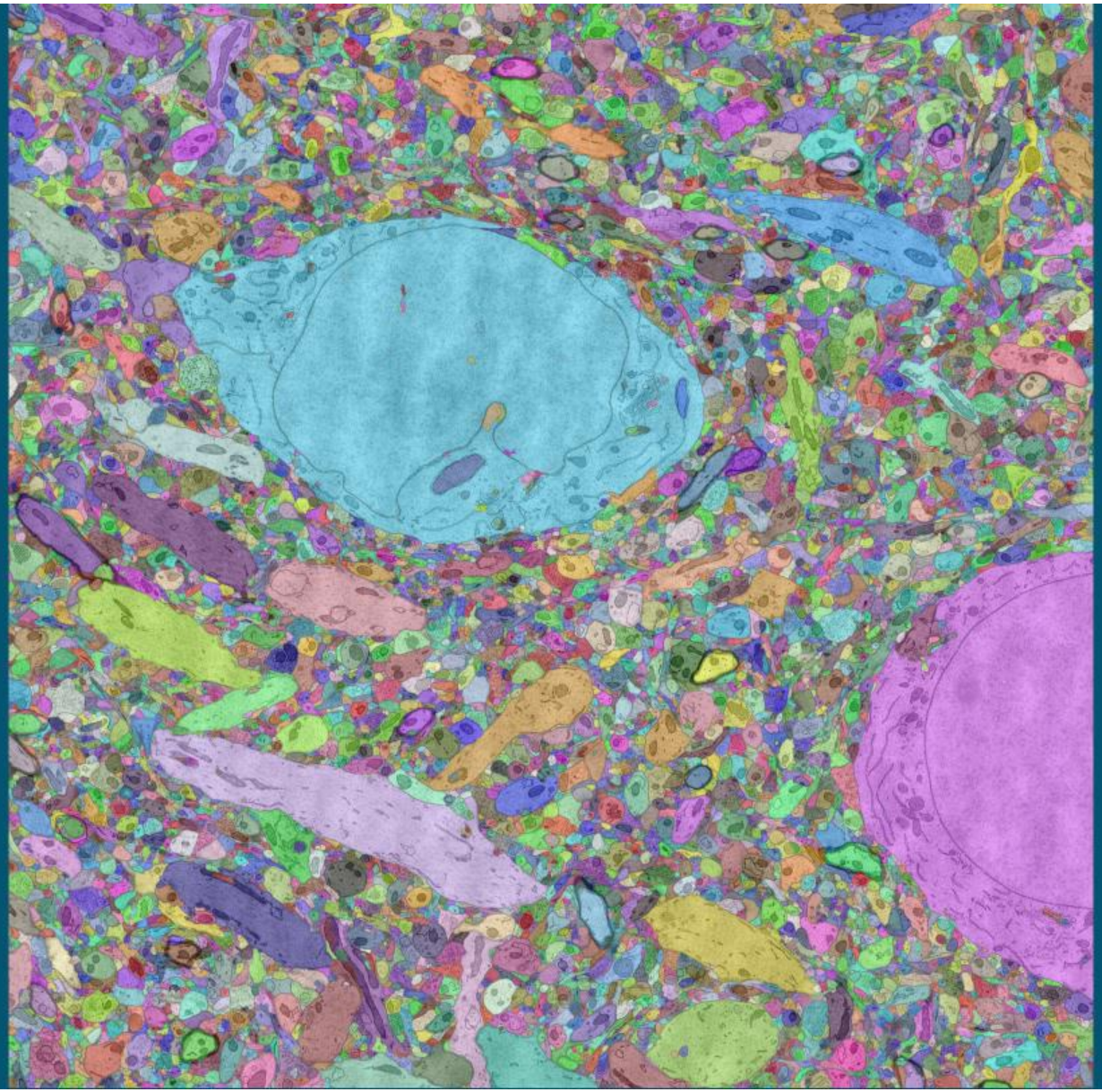


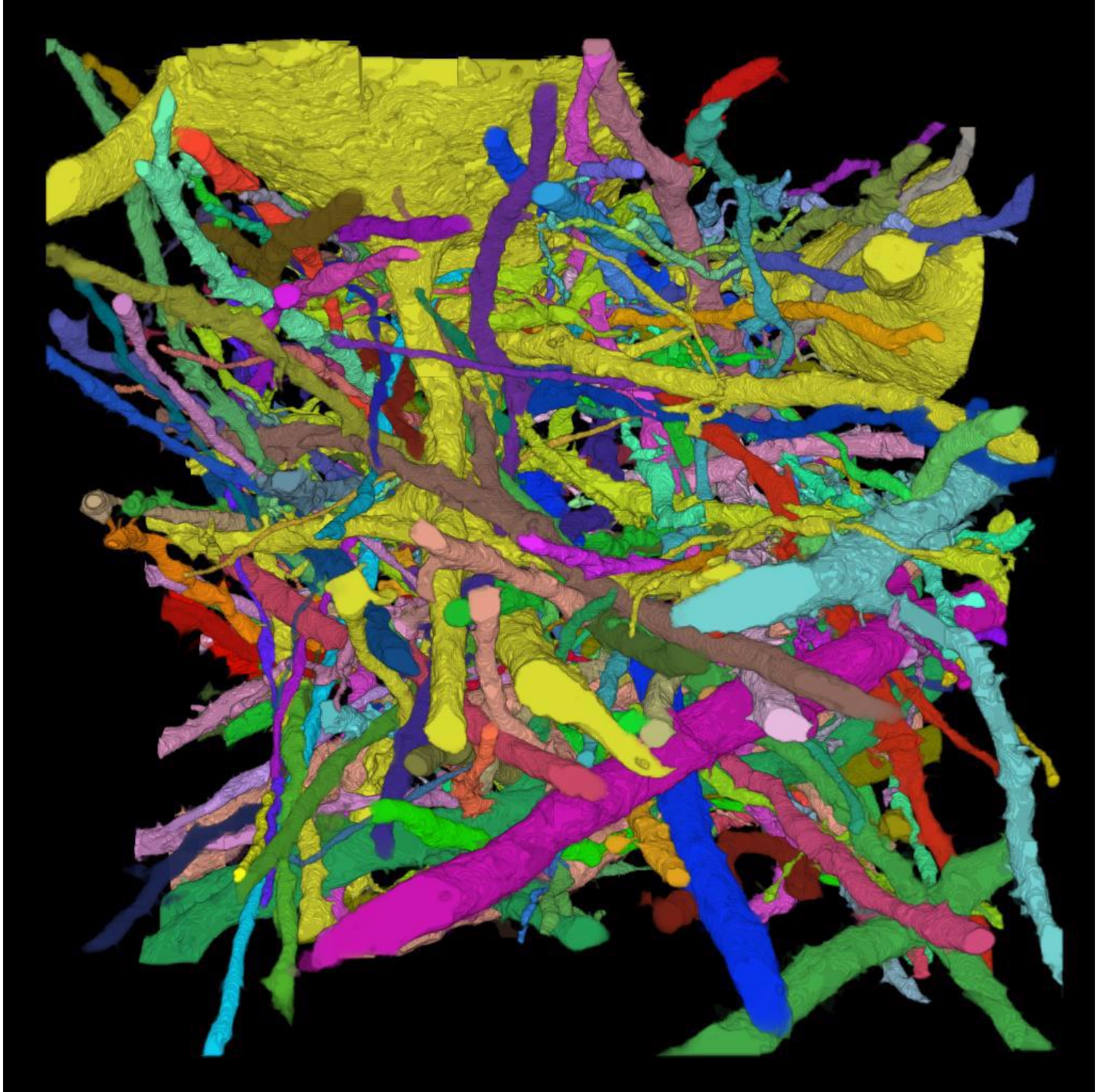
3d Segmentation

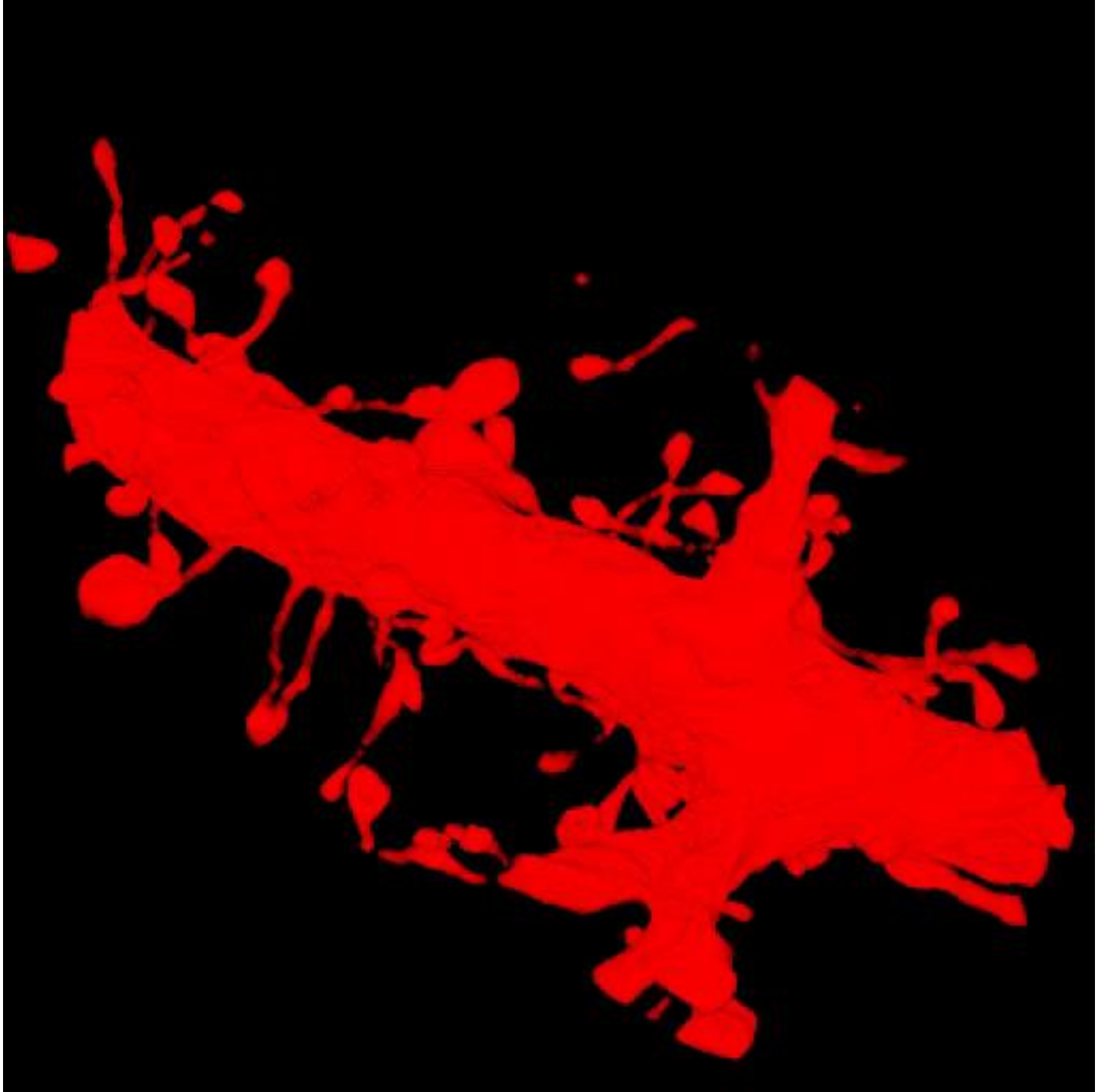
Blocks are processed in parallel by Segmentation Fusion [4] or Gala [7].






1000 nm





Thank You

Computer Vision

Ray Jones

Seymour Knowles-Barley

Hanspeter Pfister

EM Data

Bobby Kasthuri

Josh Morgan

Alyssa Wilson

Dongil Lee

Richard Schalek

Jeff Lichtman

Annotations

Daniel Berger

Bobby Kasthuri

Alyssa Wilson

Dongil Lee

Linda Xu

Summer Interns

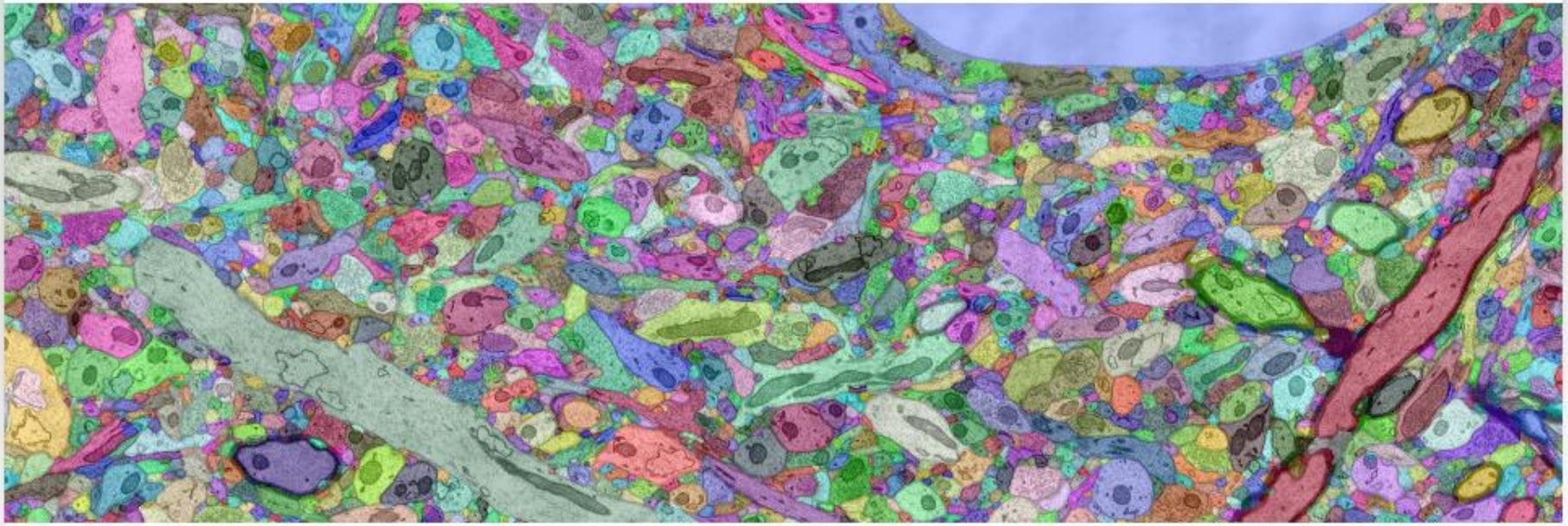
Students from Masconomet Regional High School

rhoana.org

RhoANA

Dense Automatic Neural Annotation

HARVARD
School of Engineering and Applied Sciences



RhoANA is software for dense automatic annotation of neurons in EM serial sections. It includes a processing pipeline, as well as *Mojo*, a proofreading and annotation tool.

A preprint describing our work is available.

Our code is available on github.

Links:

Source code
Pfister group
Lichtman group

Final Projects

- ~~Team Registration due 10/23~~
- Project proposals due 11/17 **2pm!!!!!**
- Project review - Week of 11/17-21
- IPython Process book due 12/10
- Project webpage and 2 minute screencast due 12/12
- Project presentations and best project prizes 12/16

Office hours

- Rafael office hours during normal lecture time
- by appointment
- CLSB 3 Blackfan circle

- Verena in NW235.1
- by appointment

Website Description

Home Piazza Syllabus Schedule Homework Readings Projects Resources

http://cs109.github.io/2014/pages/projects.html

Project Proposal Form

- 1 per team
- Please do this soon!
- Time schedule is tight.

Project Review

- Your TF will contact you via email.
- You meet with your TF during the week.
- We reserved the lecture time.
- You are welcome to make other arrangements.
- Lecture time has benefits of scheduling and me and Rafael there.
- Online students can schedule Skype meetings.

IPython Process Book

- important part of your project
- overview and motivation
- related Work
- initial Questions
- data: Source, scraping method, cleanup, etc.
- exploratory Analysis
- final Analysis
- standalone Document that **fully** describes your project in detail
- Use visualizations!

Code – Zen (Picks)

- >>> import this

- The Zen of Python, by Tim Peters
- **Explicit is better than implicit.**
- **Simple is better than complex.**
- Flat is better than nested.
- **Readability counts.**
- Errors should never pass silently.
- Now is better than never.
- **If the implementation is hard to explain, it's a bad idea.**

Website

- Use Github, Google Sites, Personal Sites
- Open to the public
- Links to Ipython notebook and data
- Visualizations
- Video

Screencast

- 2 minute limit
- We have seen cool examples.
- You have two days to work on the video and the website, make them count!

Peer Assessment

- Preparation - were they prepared during team meetings?
- Contribution - did they contribute productively to the team discussion and work?
- Respect for others' ideas - did they encourage others to contribute their ideas?
- Flexibility - were they flexible when disagreements occurred?

- You will also assess your own performance.

Grading

- Part I: Final Project Proposal
  - 10% of your CS109 grade
- Part II:
  - 25% of your CS109 grade
  - Ipython Process Notebook (80%)
  - Website and Screencast (20%)

What we learned

Data Wrangling

- Python

CS164 – Software Engineering

- David Malan

- Introduction to principles of software engineering and best practices, including code reviews, source control, and unit tests. Topics include Ajax, event handling, HTTP, MVC, object-oriented design, relational databases, and user experience. Projects include web apps with front-end UIs (mobile and desktop) and back-end APIs. Languages include JavaScript and PHP.

CS124 -Data Structures and Algorithms

- Fundamentals
- Graph Algorithms
- Greedy Algorithms
- Dynamic Programming
- Divide and Conquer
- Hashing
- Linear Programming
- Randomized Algorithms
- NP-completeness review
- Novel approaches to NP-complete problems

CS 165 – Data Systems

- **Expected learning outcomes**
- Become familiar with the history and evolution of data systems design over the past 4-5 decades.
- Understanding the basic tradeoffs in designing and implementing modern data systems.
- Being able to design a new data system given a data-driven scenario and to built a prototype.
- Being able to understand which data system is a good fit given the needs of an application.
- Advanced C programming and debugging skills.

CS207- Systems Development for Computational Science

- Apply basic software development tool-chains, including source-code control, testing frameworks, and documentation tools, to the process of designing and implementing large software systems;
- Apply design principles to the decomposition of software into re-usable components, and to the production of those components;
- Know how to approach an existing piece of software for maintenance, extension, and modification;
- Design, develop, and deploy a set of software components to produce a scalable, reliable, and reproducible experimental system for scientific investigation;
- Use a variety of approaches to software development team organization, and select techniques that are appropriate in different circumstances.

CS205 – Computing Foundations for Computational Science

- Apply basic computer science concepts such as modularity, abstraction, and encapsulation to scientific problems
- Recognize and recall computer architectures, algorithms, and data structures that are relevant to computational science
- Apply concepts of parallel programming and “parallel thinking” to computational science
- Analyze and visualize large scientific data and implement data-intensive computations on cluster and cloud infrastructures
- Use open-source tools for large- and fine-grain parallel computations, cloud computing, and visualization

CS262 - Introduction to Distributed Computing

- design and implementation of large systems
- running on multiple computers connected by a network.
- investigate the fundamental characteristics of distributed systems
- investigate how to build systems that exploit those fundamental characteristics.

Basic Visualization / EDA

- Distributions and Statistical Summaries
- Visualization Goals, Data Types, Statistical Graphs
- Boxplots, histograms, scatterplots

- Data sets:
  - Baseball
  - World Income

STAT 110 210

- A comprehensive introduction to probability.
- sample spaces and events
- conditional probability
- Bayes' Theorem
- Univariate distributions
- Multivariate distributions
- Limit laws
- Markov chains

Dimensionality Reduction

- Distance
- Clustering
- Linear Algebra
- Singular Value Decomposition

- Visualization of Multi-Dim Data, Maps and Text
- Big Data Visualization

Statistical Inference

- Polls
- Binomial distribution, CLT
- Modeling

STAT 111 211

- Basic concepts of statistical inference from Frequentist and Bayesian perspectives. Topics include, but
- not limited to: maximum likelihood methods, confidence and Bayesian interval estimation, hypothesis
- testing, least squares methods and categorical data analysis.

STAT 139/149

- Statistical Sleuthing Through Linear Models
- A serious introduction to statistical inference with linear models and related methods. Topics include t-tools and permutation-based alternatives, multiple-group comparisons, analysis of variance, linear regression, model checking and refinement, and causation versus correlation. Emphasis on thinking statistically, evaluating assumptions, and developing tools for real-life applications.

Regression

- Summaries
- Detecting Associations

- Data set: Baseball

Caveats

- Correlation is not causation
- Regression fallacy
- Multiple comparisons (we did not cover this)

Bayesian Stats

- Predicting performance in baseball
- Modeling elections

AM207 - Monte Carlo Methods for Inference and Data Analysis

- BASIC MONTE CARLO METHODS
- BAYESIAN FORMALISM
- MARKOV CHAIN MONTE CARLO (MCMC)
- ADVANCED MCMC METHODS
- OPTIMIZATION
- DYNAMIC SYSTEMS
- Gaussian Processes
- PROBABILISTIC MODELS
- EM mixture model
- Probabilistic Graphical Models
- (if time permits) Bayesian Averaging

STAT 220 - Bayesian Data Analysis

STAT 221 - Applied Bayesian Statistical Computing

- Basics of the Bayesian inference
- Multi-parameter models
- Inference from large samples
- Hierarchical models
- Computation I (non-iterative methods
- Missing data
- Computation II (iterative sampling methods
- Model checking
- Topics on Bayesian modeling and computation

Machine Learning

- Concepts
  - Error matrix
  - False positives and false negatives
  - Evaluation procedures: cross-validation
  - Train, Test, Validate

Machine Learning

- Supervised learning:
  - K-nearest neighbors
  - Decision Trees
  - Random Forest
  - SVM
- Unsupervised learning:
  - SVD
  - Clustering

CS181 – Machine Learning

- *Introduction and Course Overview*
- *Clustering with K-Means and K-Medoids*
- *Hierarchical Clustering*
- *Principal Component Analysis*
- *Supervised Learning*
- *Linear Regression*
- *Model Selection*
- *Linear Classification*
- *Classification and CV Review*
- *Probabilistic Classification*
- *Neural Networks*
- *Regression and Classification Trees*
- *Max-Margin Classification*
- *Support Vector Machines*
- *Markov Decision Processes*
- *Reinforcement Learning*
- *Partially-Observable Markov Decision Processes*
- *Expectation Maximization*
- *Hidden Markov Models*

CS281 – Advanced Machine Learning

- Introduction to Inference and Learning
- Simple Discrete Models
- Simple Gaussian Models
- Bayesian Statistics
- Linear Regression
- Linear Classifiers
- Generalized Linear Models
- Directed Graphical Models
- Mixture Models
- Sparse Linear Models
- Exact Inference
- Variational Inference
- Loopy Belief Propagation
- Monte Carlo Basics
- Markov Chain Monte Carlo
- Advanced MCMC
- Latent Dirichlet Allocation
- State Space Models
- Graph Models
- Kernels
- Gaussian Processes
- Dirichlet Processes
- Boltzmann Machines
- Neural Networks

STAT 183

- **Learning from Big Data** - Kaggle style!
- 3-5 competitions open on Kaggle
- learn form the data in teams

Advanced Visualization

- Graph Visualization and Story telling

CS171-Visualizations

- **After successful completion of this course, you will be able to…**
- Critically evaluate visualizations and suggest improvements and refinements
- Use JavaScript and other tools to scrape, clean, and process data
- Use standalone visualization applications to quickly explore data
- Apply a structured design process to create effective visualizations
- Conceptualize ideas and interaction techniques using sketching
- Use principles of human perception and cognition in visualization design
- Create web-based interactive visualizations using JavaScript and D3
- Use storytelling principles to design coherent and clear visualizations

CS105 – Privacy and Technology (GOV1430)

- Privacy Concepts
- Surveillance
- Tapping and tracing
- Data aggregation, analytics, and privacy
- Data-intensive Science
- Biometrics and DNA

Institute for
Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

- Master of Science in CSE
  - 1 year, 8 courses
- Master of Engineering in CSE
  - 2 years, 8 courses, 1 thesis

- For PhD students:
  - secondary field in CSE
  - 4 courses

Compute Fest

- Skill building workshops
  - Monday, January 12 - Friday, January 16

- Student challenge!
  - Tuesday, January 20 - Thursday, January 22