Cancer Epigenetics Study Using Next-Generation Sequencing Data

July 29, 2010
Big Data For Science

Sun Kim and Heejooon Chae
School of Informatics and Computing
Center for Bioinformatics Research
Indiana University, Bloomington, Indiana, USA

Overview of The Talk

- Background on epigenomics and DNA methylation
- OSU-IU Center for Cancer Systems Biology
- Mapping sequence reads
- Data
- BioVLAB-mCpG

Part I: Epignomics and DNA Methylation

Epigenetics

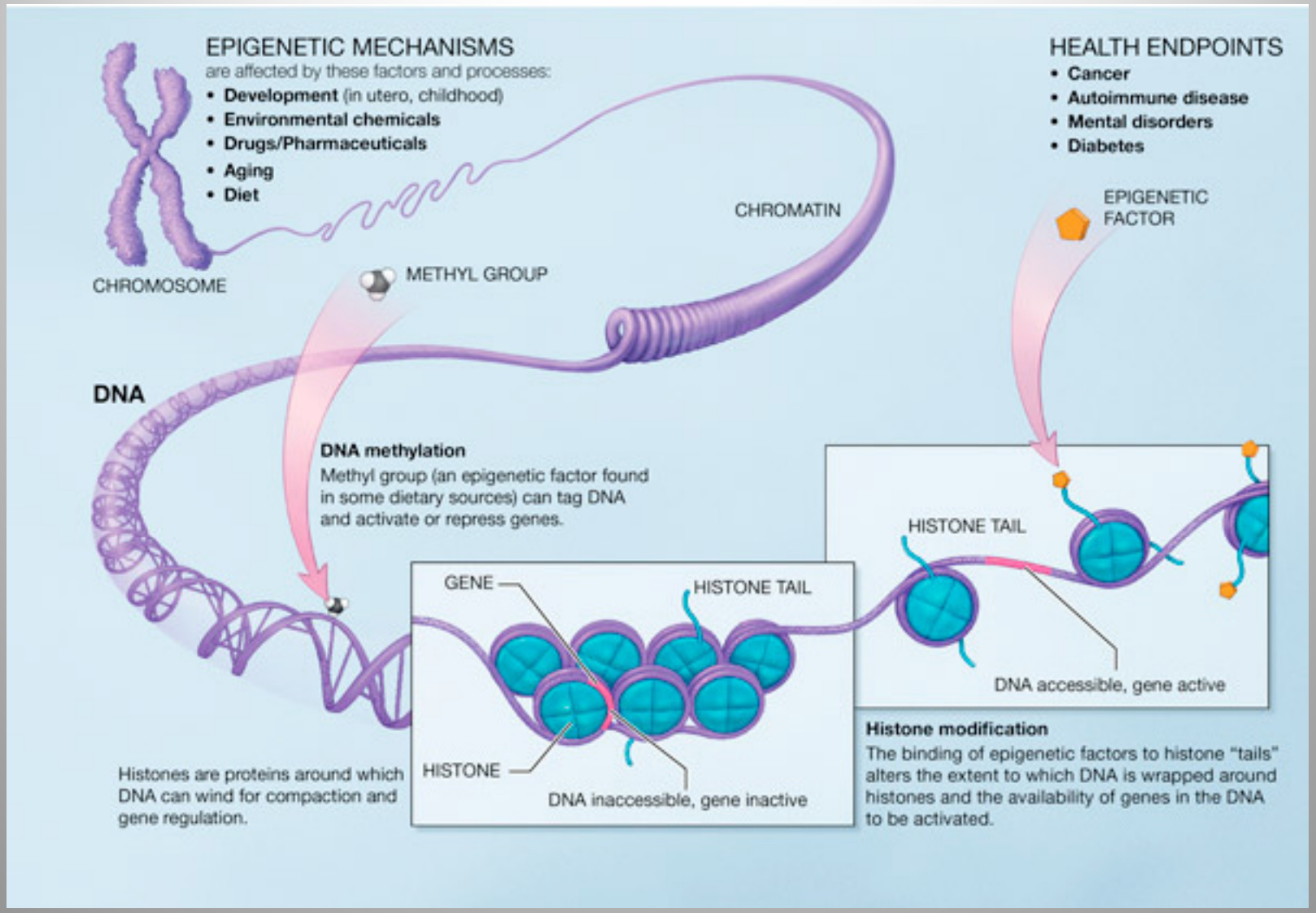
- Epigenetics is the study of ***heritable*** changes in gene function that occur without a change in DNA sequence.

- Summarizes mechanisms and phenomena that affect the phenotype of a cell or an organism without affecting the genotype.

- Modifications of DNA (cytosine methylation) and proteins (histones) define the epigenetic profile.

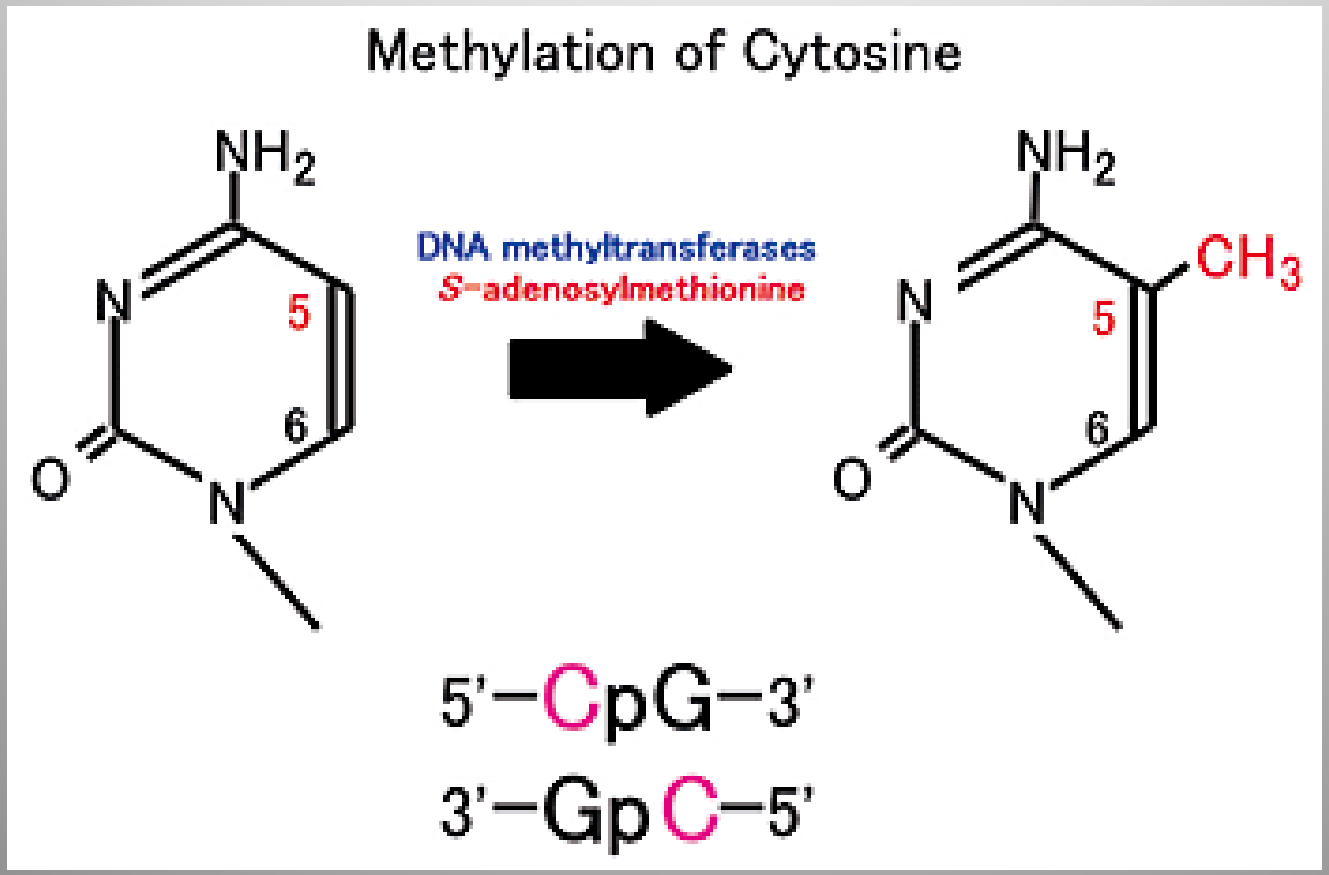
- *Epigenomics* is the study of these epigenetic changes on a genome-wide scale.

This slide is from Ken Nepthew at IU.



http://nihroadmap.nih.gov/epigenomics/epigeneticmechanisms.a

DNA Methylation



Normal Cellular Functions Regulated by Epigenetic Mechanisms

- **Correct organization of chromatin**
  -Controls active and inactive states of embryonic and somatic cell- Epigenetic components contribute to plasticity and stability during development.
  -Involved in maintenance of differentiated cells.

- **Specific DNA methylation patterns, chromatin modifications**
  -Controls gene- and tissue-specific epigenetic patterns.

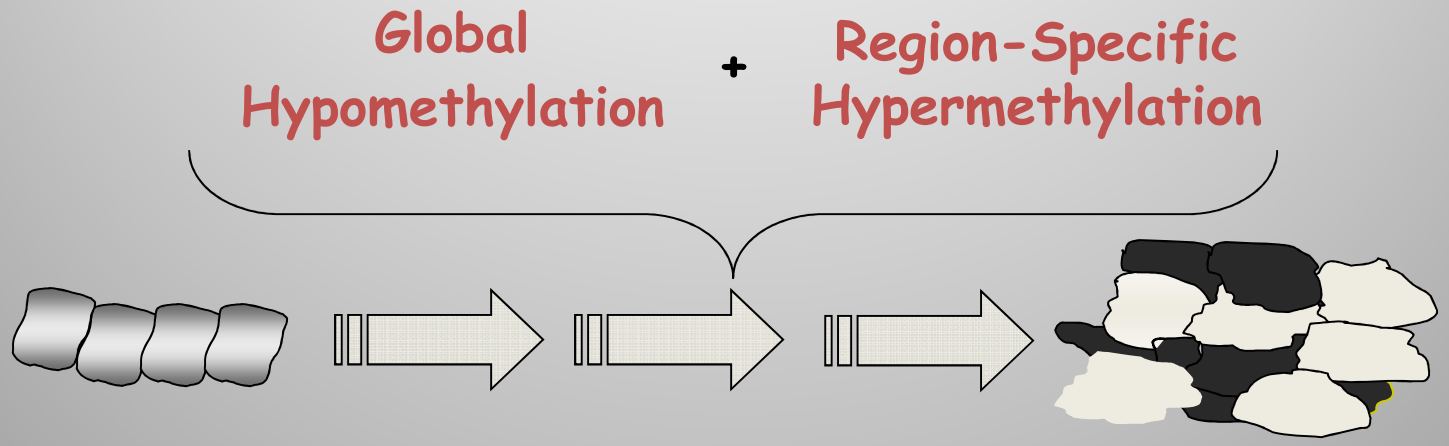
- **Genomic imprinting**- Essential for development

- **Silencing of repetitive elements**
  -Maintains chromatin order, proper gene expression patterns

- **X chromosome inactivation-** Balances gene expression

This slide is from Ken Nepthew at IU.

Progressive Accumulation of DNA Methylation in Cancer



Accumulation of Genetic and Epigenetic Abnormalities

This slide is from Ken Nephew at IU.

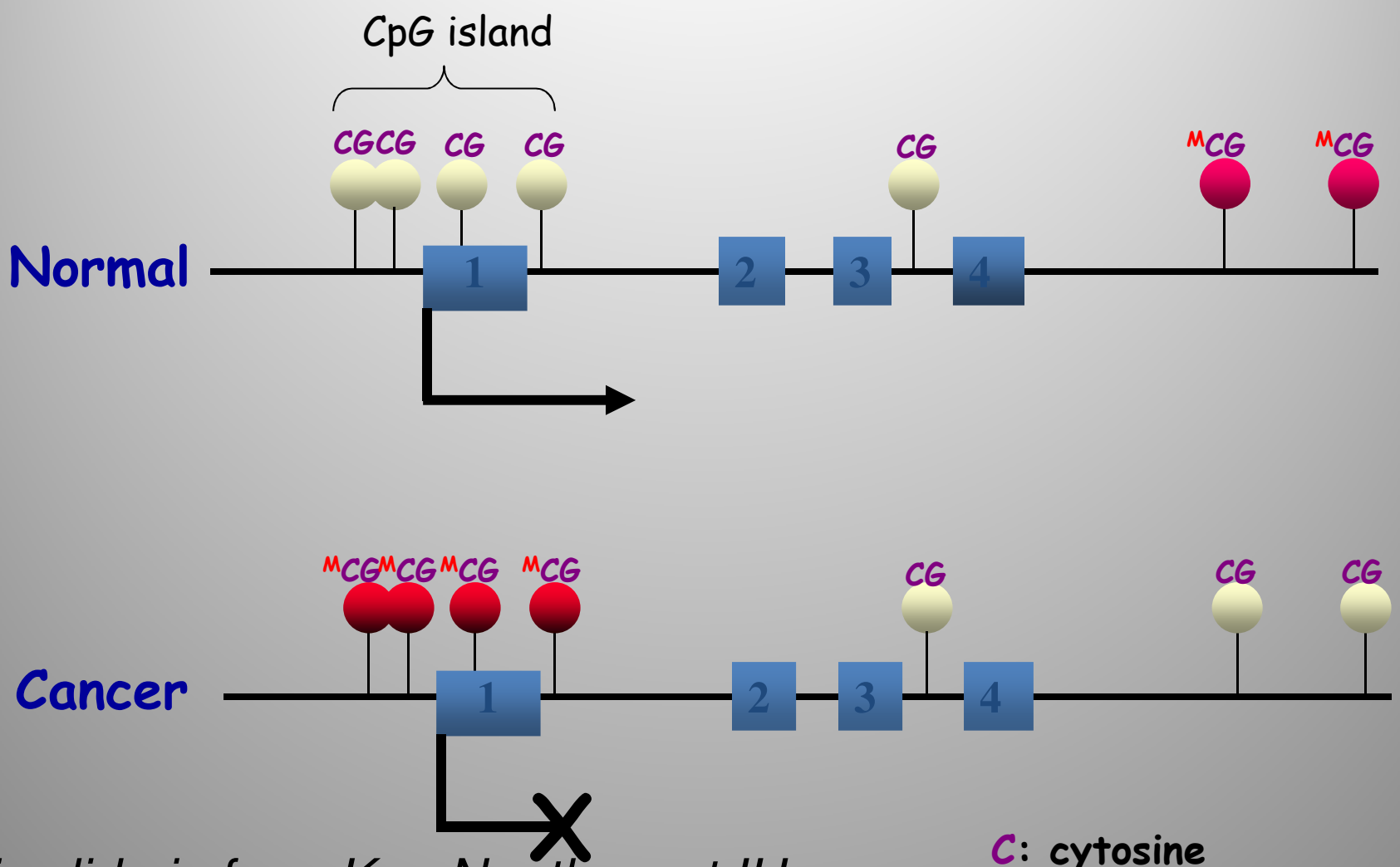
CpG Islands

- CpG island: a cluster of CpG residues often found near gene promoters (sequences ~1000 base pairs in length with a GC content of over 60%)

- ~29,000 CpG islands in human genome (~60% of all genes are associated with CpG islands)

- Most CpG islands are <u>un</u>methylated in normal cells.

DNA Methylation and Gene Silencing in Cancer Cells



This slide is from Ken Nepthew at IU.

-- Sun Kim group at IU --

C: cytosine

mC: methylcytosine

Histone modifications: Histone Code

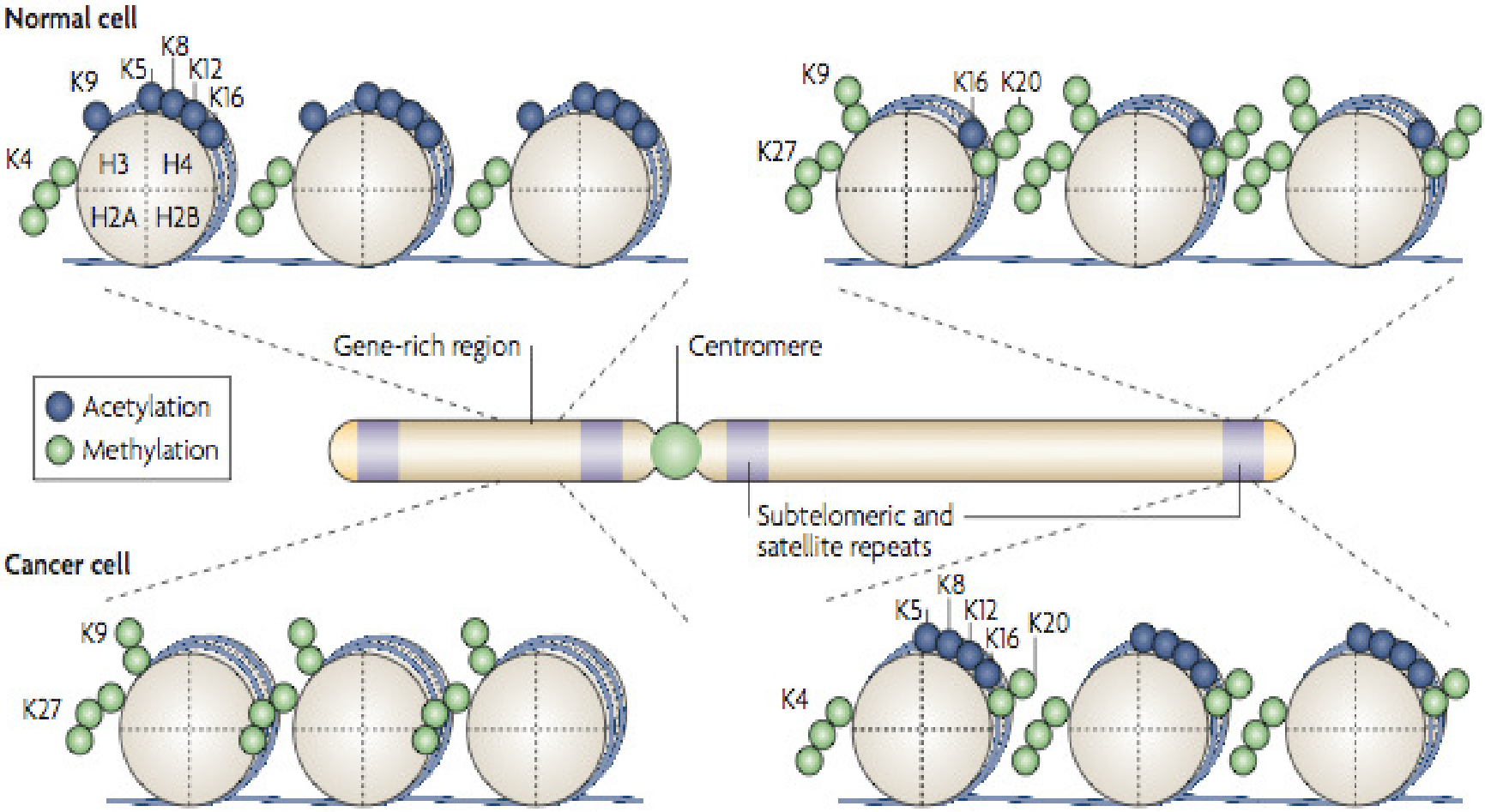
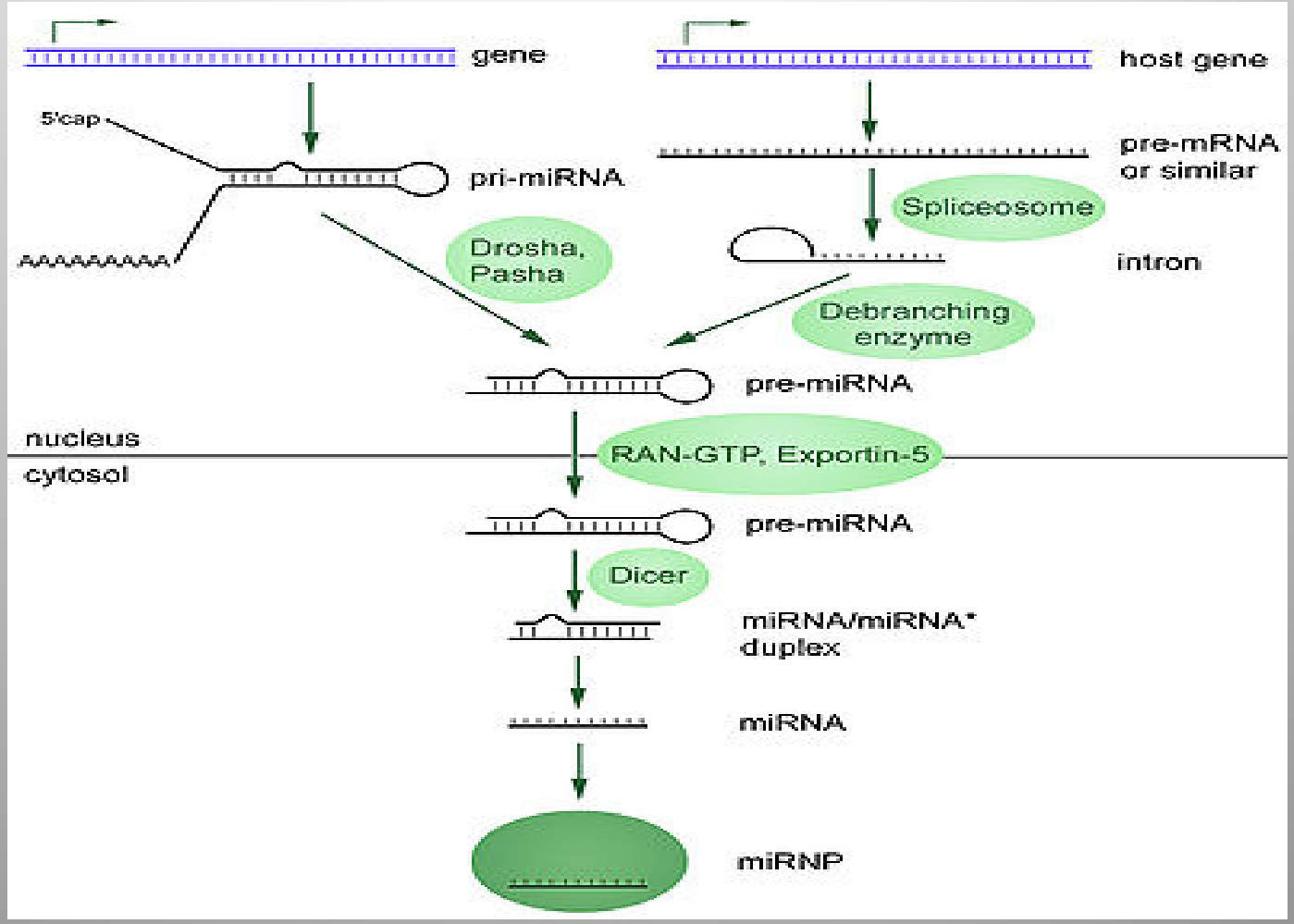


Figure 4 | **Histone-modification maps for a typical chromosome in normal and cancer cells.** Nucleosomal arrays

Nature Reviews Genetics 8, 286-298 (April 2007)

MicroRNA



http://en.wikipedia.org/wiki/MicroRNA

PART 2: OSU-IU Center for Cancer Systems Biology

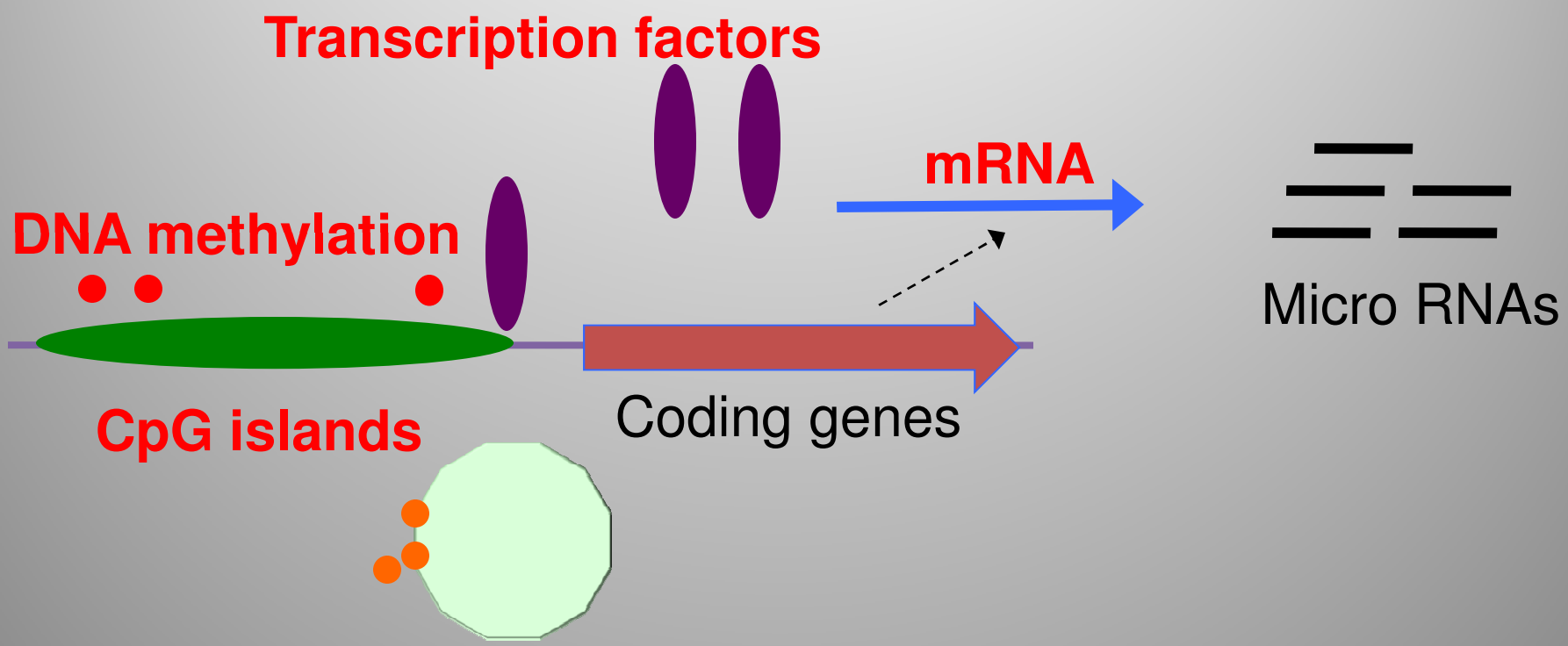
OSU-IU Integrated Cancer Biology Program (ICBP) Center

- The Integrative Cancer Biology Program (http://icbp.nci.nih.gov/) is a program launched by US National Cancer Institute in 2004.
- OSU-IU ICBP Center aims to characterize the role of epigenomics in the development of drug resistance in human cancer for a period of 2004 – 2015.

Drug Resistance in Human Cancer

- The OSU-IU Center has been investigating the mechanism of developing drug resistance in breast, prostate, and ovarian cancer.
- In particular, we are interested in investigating changes in *epigenetic mechanisms* in terms of gene regulation and pathway activation while in transition to a hormone-/chemo-sensitive to ***a hormone-/chemo-insensitive phenotype*** in cancer.

DNA Methylation vs. Transcription Factor



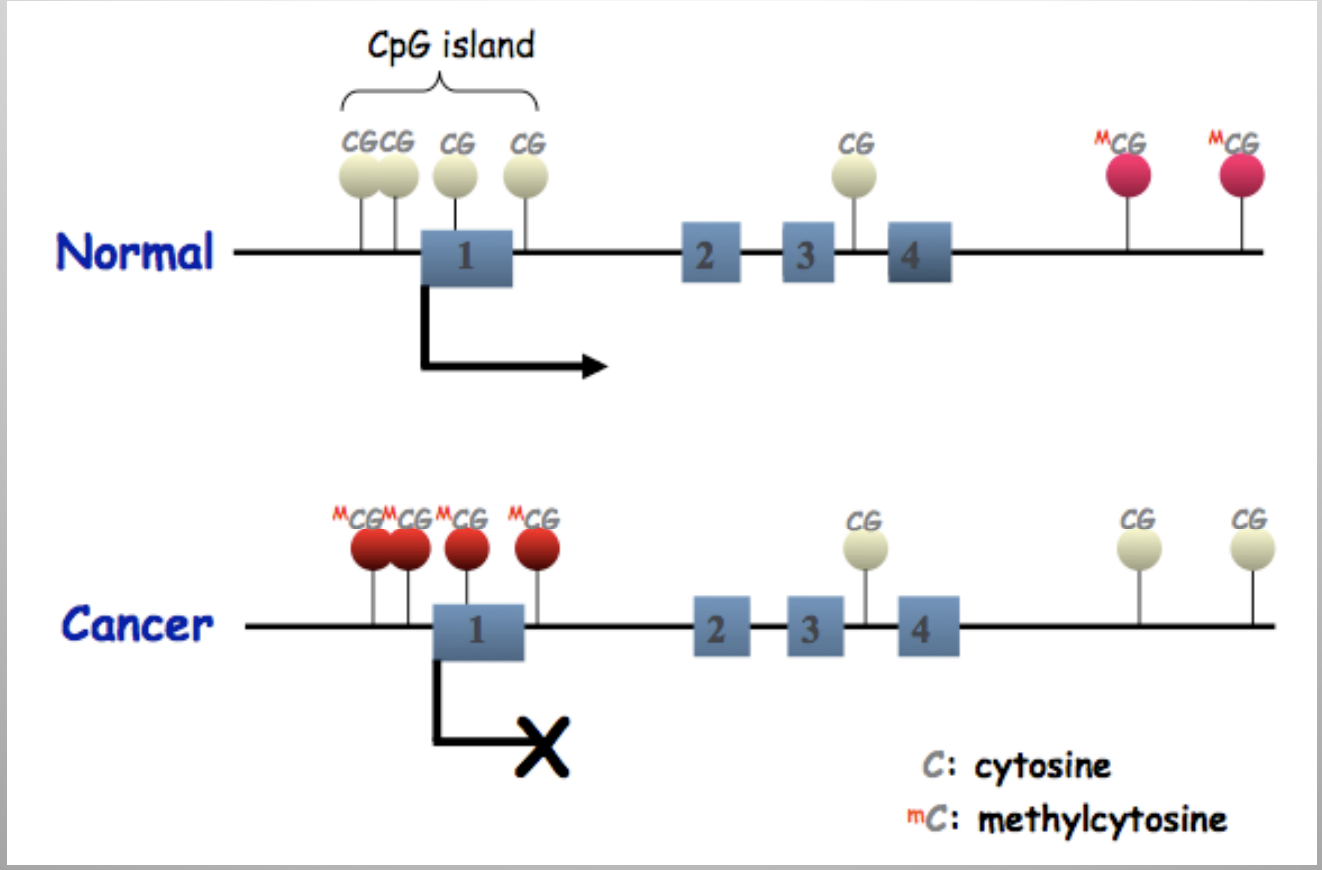
6 Methylome Projects

- To investigate the effect of DNA methylation in drug-resistance cancer phenotype, we sequence and study 6 cell lines:
  1. Breast cancer: 2 cell lines before and after drug resistance phenotype.
  2. Prostate cancer: 2 cell lines before and after drug resistance phenotype.
  3. Ovarian cancer: 2 cell lines before and after drug resistance phenotype.

Basic Data Analysis

- Comparing methylation difference in two cell lines (e.g., before and after drug-resistance phenotype).
- Integrated analysis with histone modification, microRNA, gene expression, and phenotypes.

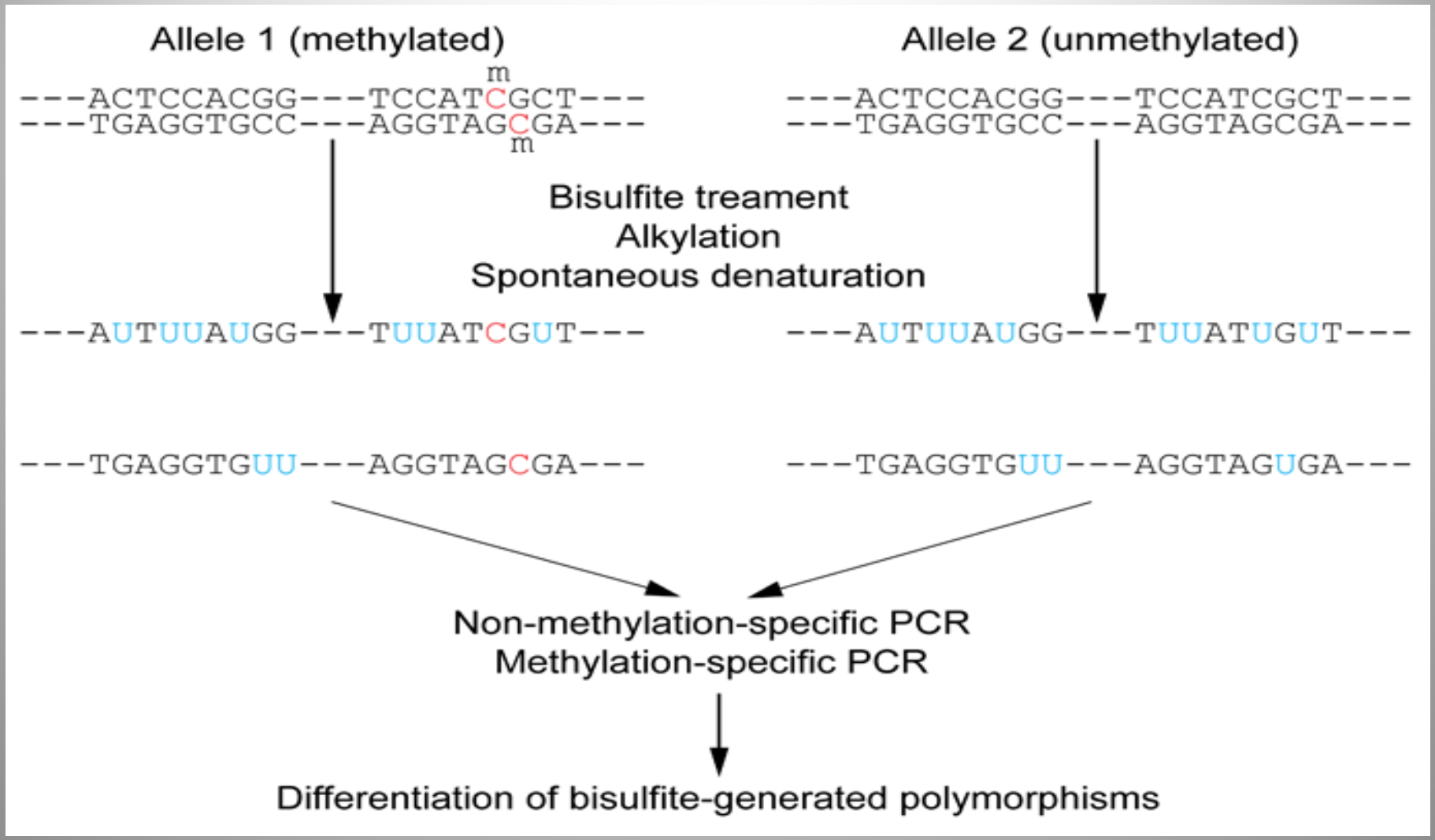
Comparative Analysis of Methylation in Two Cell Lines



- Promotor methylation analysis and expression of downstream genes.
- Promotor methylation and transcription factors and their binding sites.
- Intergenic methylation and alternative splicing.
- Methylation in non-CpG context.

PART 3: Sequence read mapping

Bisulfite Sequencing to Identify Methylated Cytosines

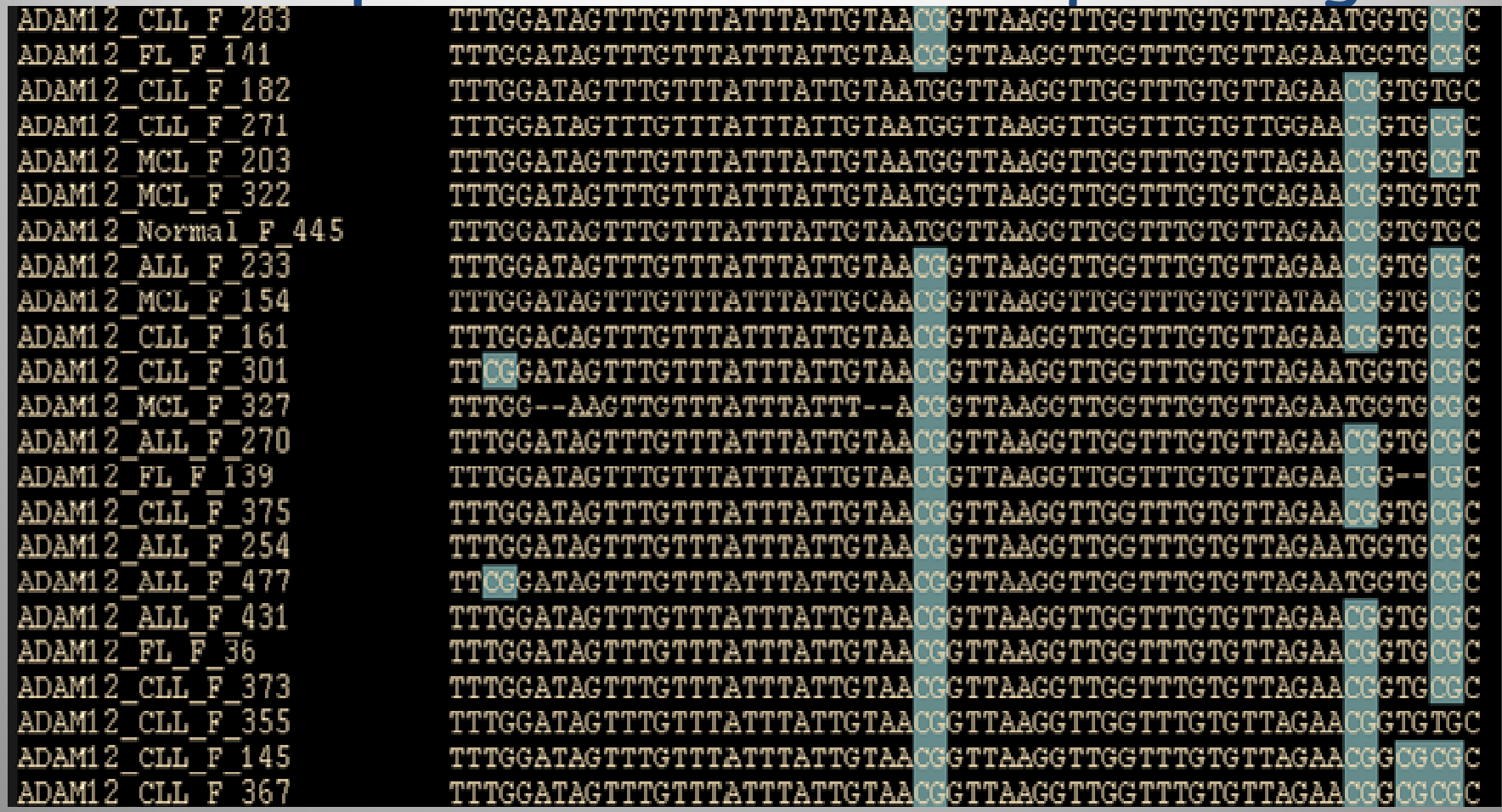


http://en.wikipedia.org/wiki/Bisulfite_sequencing

Challenges in Mapping Sequence Reads from Bisulfite Treated DNA

- A lot of reads should be mapped: several hundred millions to several billions.
- To know which cytosines are methyated, we need to sequence bisulfite treated DNA. This results in dealing with sequences of alphabet size 3, thus it takes more time.

Example of Bisulfite Sequencing



Methylation status of ADAM12 gene promotor region: courtesy by Huidong Shi at Medical College of Georgia.

Performance Comparison of Mapping Algorithms

	Benchmark data		Human genome	
Program	Time	# mapped	Time	# mapped
GNUMap	47.9 s	**71 262**	985 m 14 s	**7 739 321**
Bowtie	7.0 s	62 298	14 m 43 s	6 699 526
SOAP	11.7 s	62 208	32 m 20 s	6 764 050
MAQ	46.5 s	62 208	*3488 m 28 s	6 764 054
Slider	16 m 31 s*	58 551	Crashed	Crashed
SeqMap	81.2 s	56 326	1703 m 04 s	5 455 538
Novocraft	24.4 s	56 238	*920 m 25 s	5 306 782
RMAP	9.2 s	1202	*295 m 54 s	3 447 086

Bold values show that GNUMAP achieves the best performance.

From *Bioinformatics*. 2010 Jan 1;26(1):38-45

PART 4: Data

Two data sets

- 6 methylome data sets from our center
- 2 cell line data from

  Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009 Nov 19;462(7271):315-2

Data and Runtime Estimation

Estimate of CPU & storage requirement

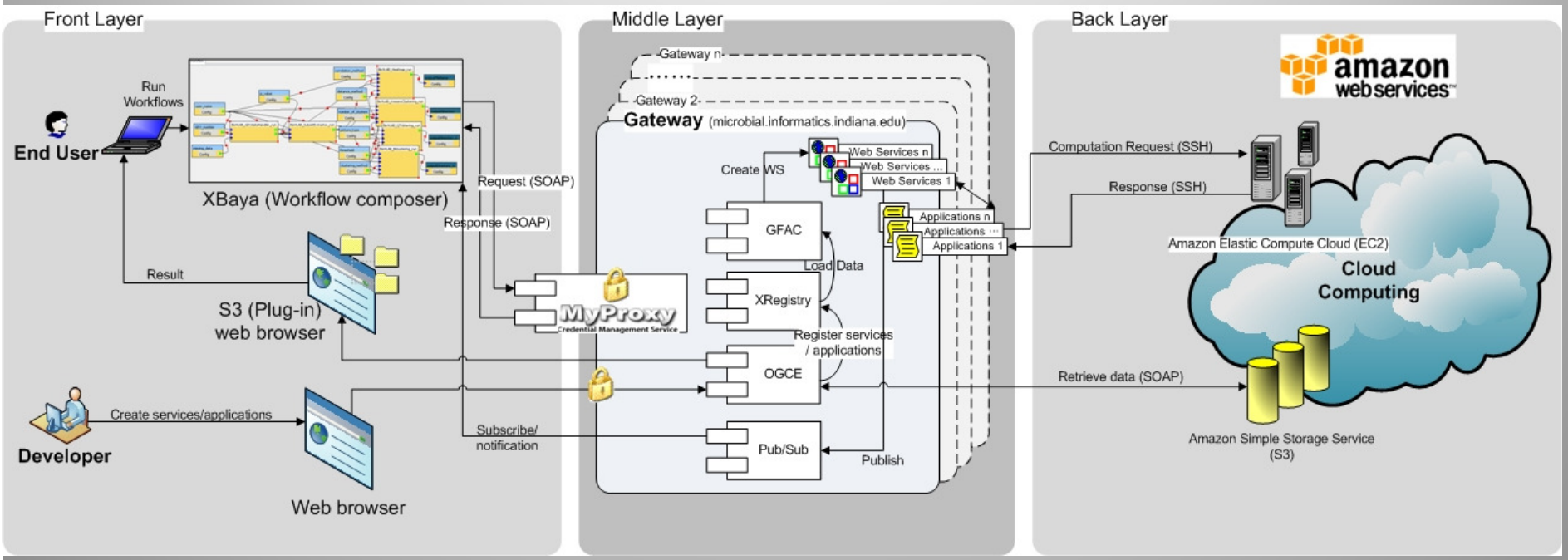
Aligner	Data set	# of read/ cell line	CPU hrs/ cell line	# of cell line	Total CPU hrs	Temporal Max Disk
GNUMAP	OSU	80 million	950	6	5700	1000GB
	Nature	4 billion	47500	2	95000	13TB
	TBD	4 billion	47500	2	95000	13TB
Bowtie	OSU	80 million	150	6	900	1000GB
	Nature	4 billion	750	2	1500	13TB
	TBD	4 billion	750	2	1500	13TB
Sub total					199000	13TB
Bisulfite treatment reads need 4times run per each read					X4	
Total					**796000**	**13TB**

PART 5: BioVLAB-mCpG

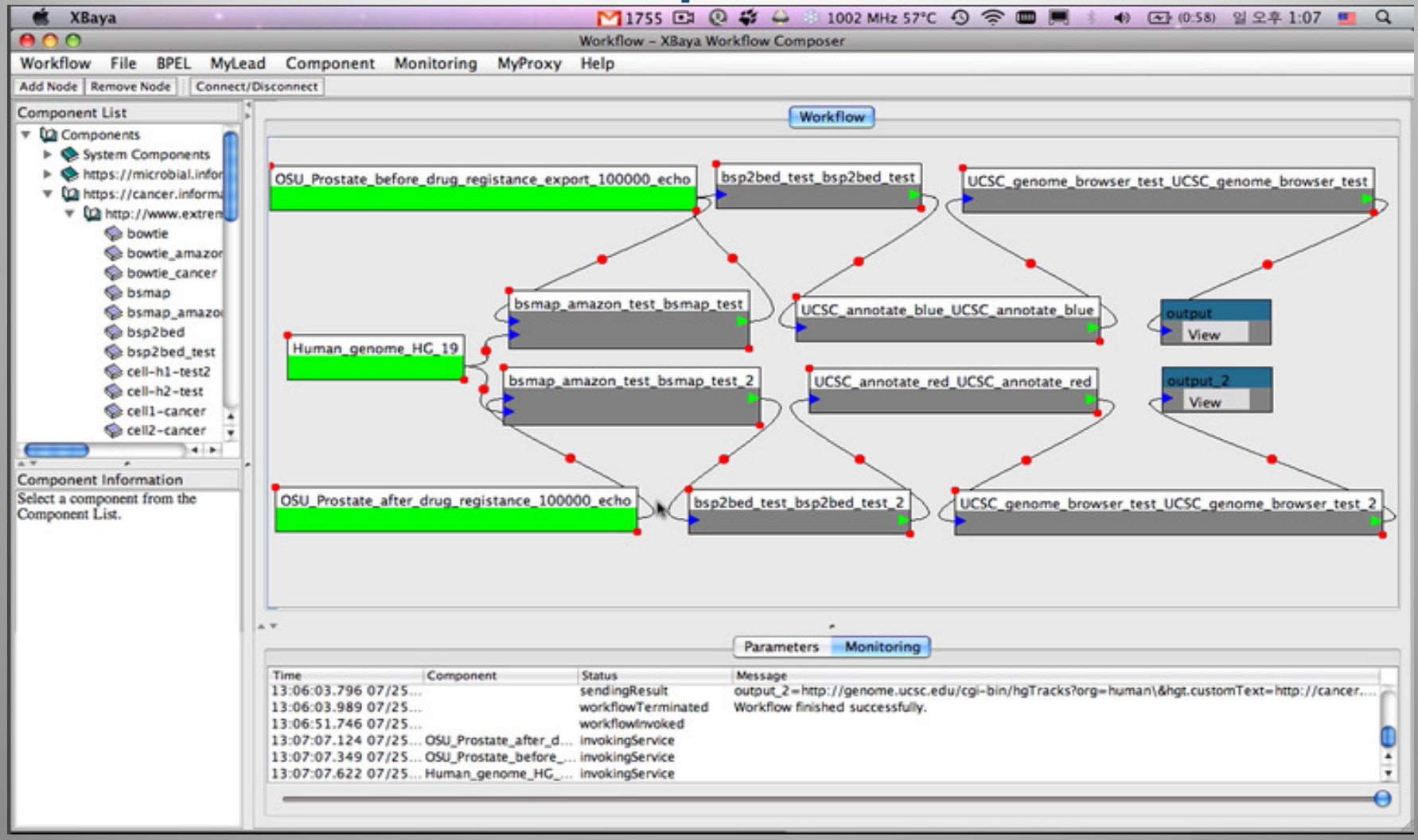
BioVLAB: Motivation

- We have developed a computational infrastructure, called BioVLAB, for the analysis of molecular biology data utilizing Amazon Cloud Computing (or any high performance computing machines) and a graphical workflow composer, XBaya.
- Easy to perform computational analysis:
  1. Set up an account
  2. Download a precomposed workflow
  3. (Modify workflow if needed: application-specific cloud)
  4. Run it

BioVLAB Architecture

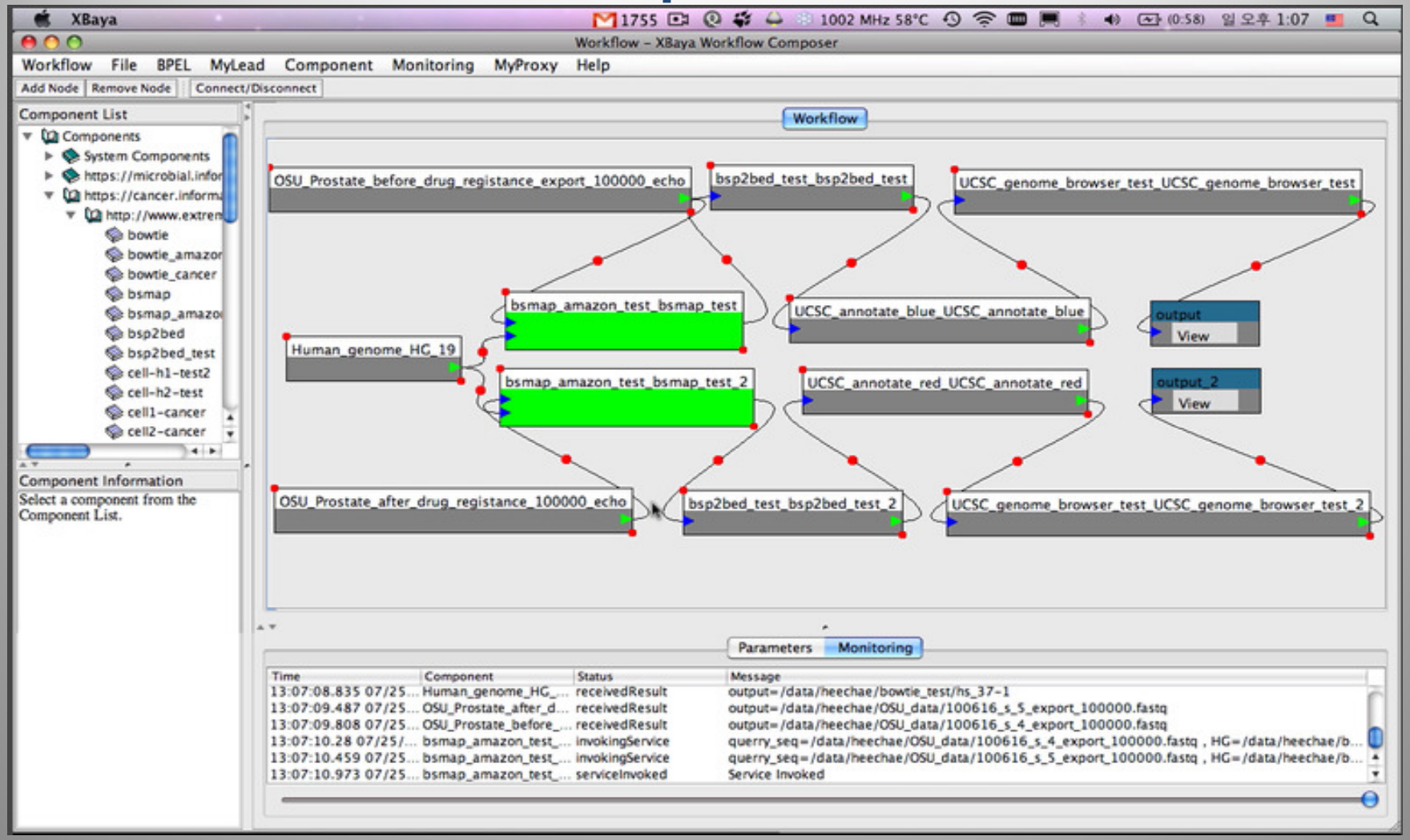


BioVLAB-mCpG Screenshots



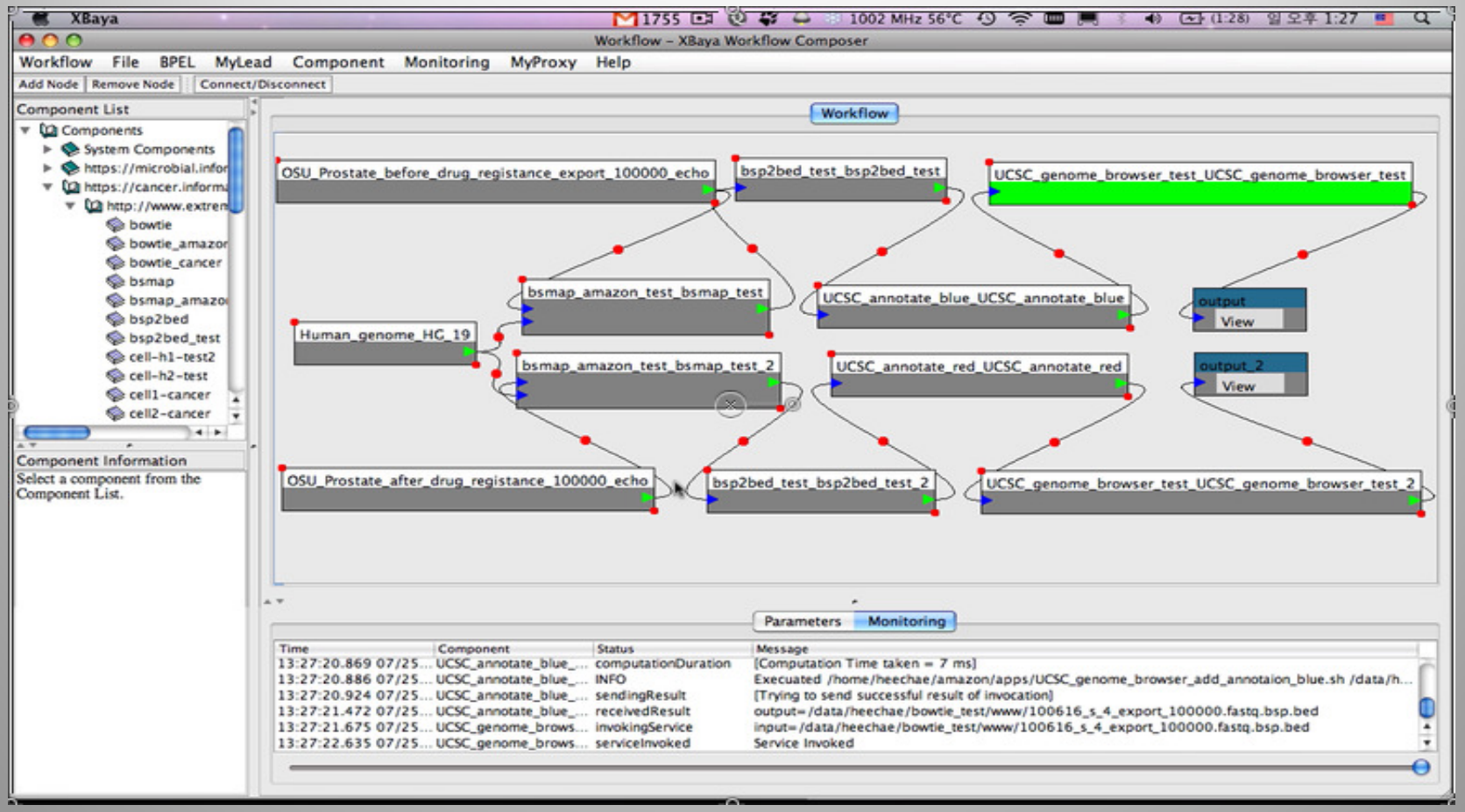
Data (in green color) is ready.

BioVLAB-mCpG Screenshots



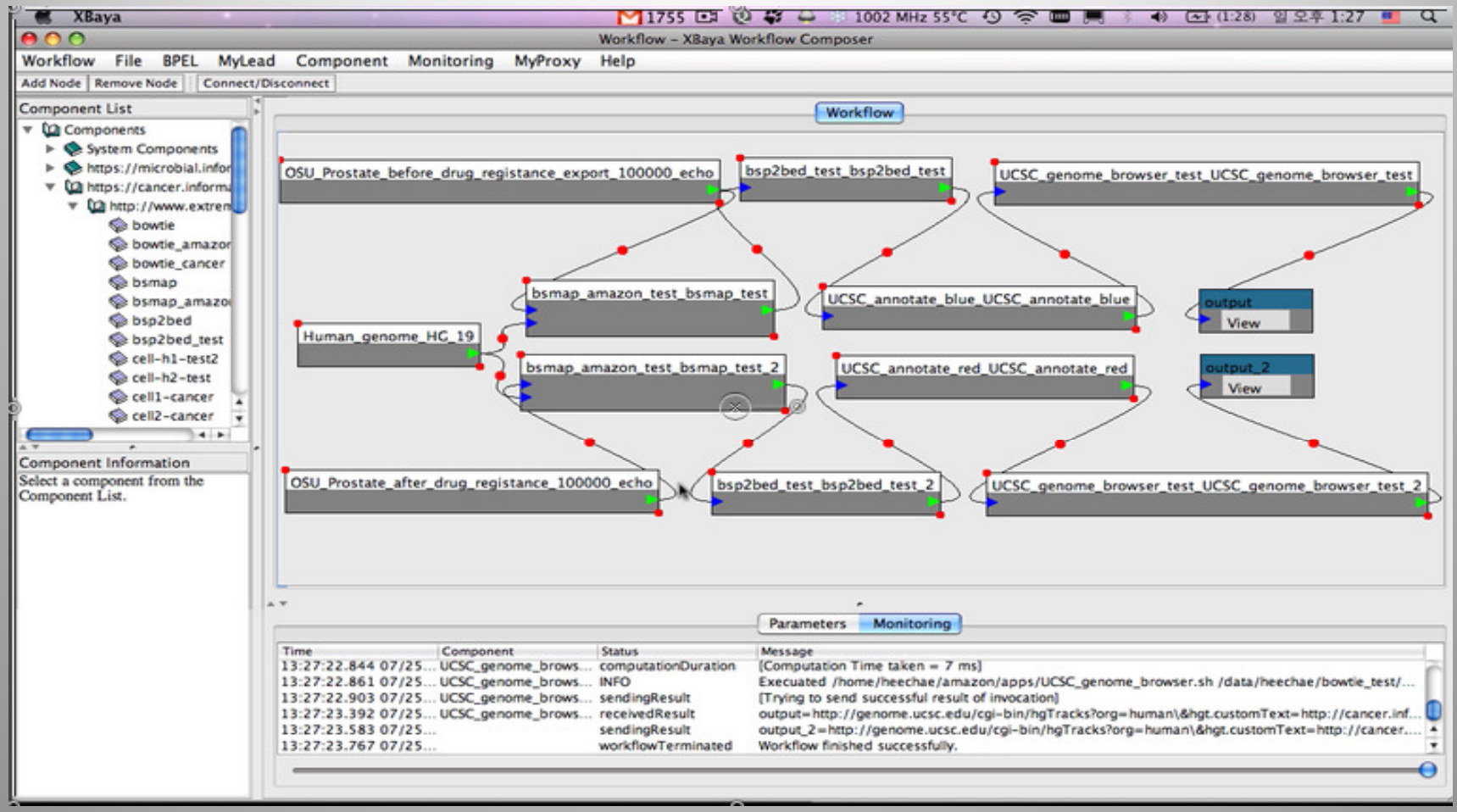
Sequence reads are being mapped by BSmap (green color).

BioVLAB-mCpG Screenshots



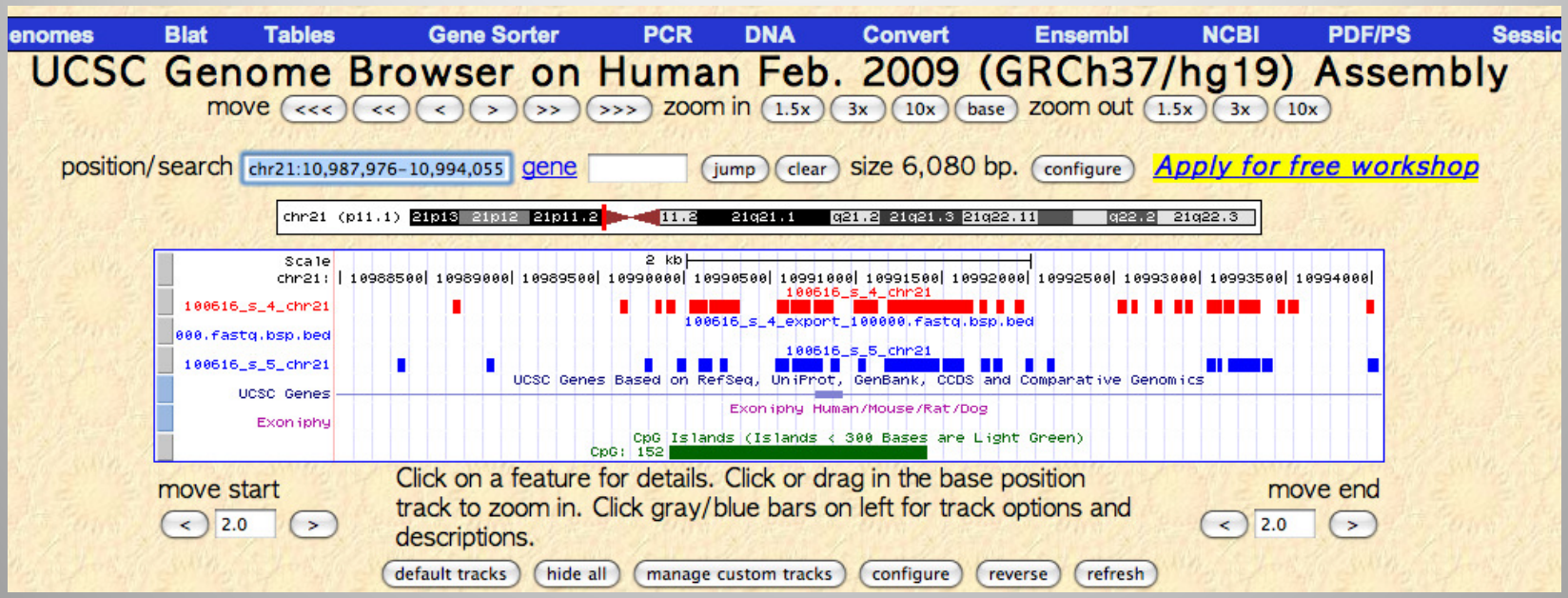
Uploading the result to the UCSC Genome Browser. (green color).

BioVLAB-mCpG Screenshots



Finished! Let's look at visualized data.

BioVLAB-mCpG Screenshots



Two lines (in red and blue colors) show DNA mthylation status in the context of exon and a CpG Island.

Acknowledgements

- Heejoon Chae, Youngik Yang, Hyungro Lee, Jong Yul Choi
- Suresh Marru, Chathura Herath, Marlon Pierce
- Ken Nephew at IU, Tim Huang at OSU and OSU-IU CCSB members
- NCI ICBP
- TeraGrid
- IU UITS

Thank you!!