Imperial College London

Tackling data analysis challenges in genome informatics using HPC

Michael Mueller
Physiological Genomics & Medicine Group
Clinical Genome Informatics Facility
Imperial College London

MRC Clinical Sciences Centre

THE MAGDI YACOUB INSTITUTE

British Heart Foundation

Imperial College London

Overview

- Genome sequencing: *why & how?*
- Informatics challenges in genome data analysis
- From genome to epigenome:

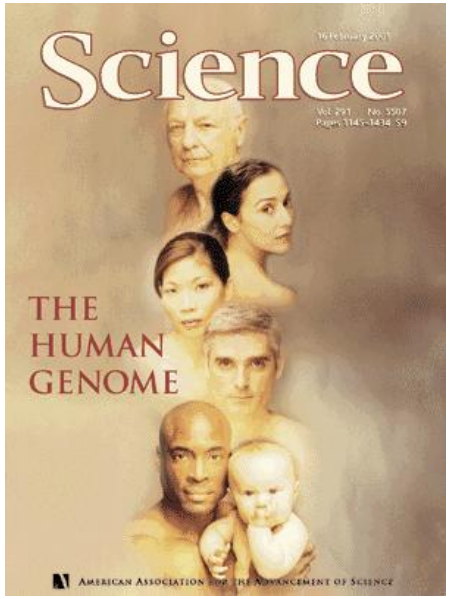
  Genetic control of inter-individual variation in DNA methylation

Imperial College London

The Human Genome

The Human Genome Project
1990 - 2003

determine the sequence of chemical bases (nucleotides) which make up DNA



articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

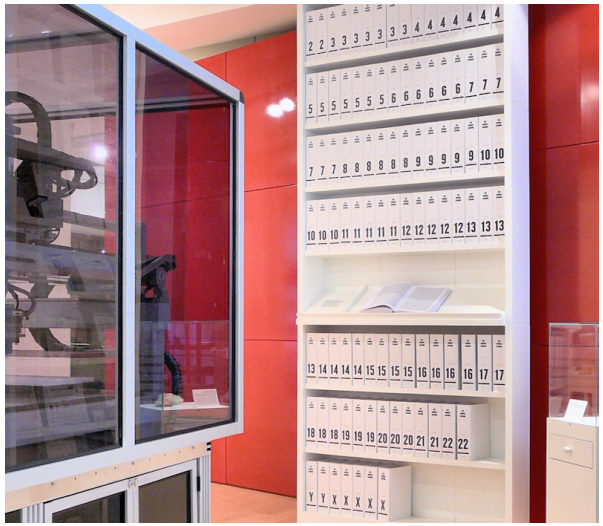
2001

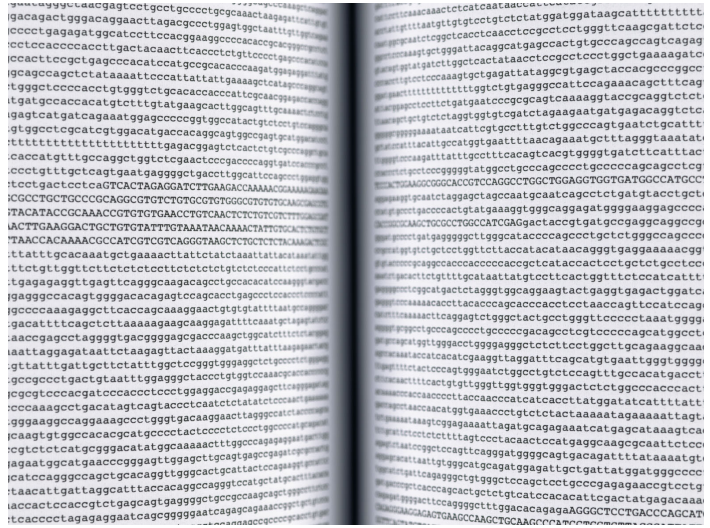
2003

Imperial College London

The Human Genome

3.1 billion nucleotides (A,C,T,G)





Imperial College
London

The Human Genome

CATGATGAATTCATTTAACCCTTGAGTTTCCATTTCTTCACCTGCTCCGTGGGGCACT
AACGCCTCCCTCAGAGGCTTCTGGTGAGAATCAGTGTTTCCCTGCCCCCGCCCCGCCC
TCCATGCCCCTTCTCCACGTTCTCACTGTGCTAGGTGCTCTTCTCTGTCTTTCTCTTC
CACCAGCCTGTGGGAAACCTGAGATGAAAGTCGTGTCTTACCCATCTTTGTATTTCCA
GCATCTGAAACTGGGCAGAGCTTAATAAATATTTTGCTGGAGAGGTTGATGATCTTAC
AAAGCTCCCATTGAAAGGTGGCTCTCTGTAAAGCAAAGTTACAATGAGATTGTGATGA
ACATTGTCCTTGTGGCTTTTCACTTAGTCCCCTCCCTTCACCTGAAGAGCAAATTTTC
CTCAAAAGTACACAGCAAACAAATGACCCACTGGTGACACTGCTGCCTTTAGACCCTG
CTGGAAAGAATCTCCACATTTATTAACATTCCCAAAGTAAATTTATCAGGTAGCATTC
ATCAGGTAACATTTGTTGCACATTCATGACTTTTCTACTGTCCACAAAGGCATATGTC
CTTATCATATGCAGACTCCTTGGCCACACTGGATTCCTCCTTCCCTCCTCGACATGGA
AGAGATGGCATCTTAGGGTCTCTTGTGTTCTTCCTGAAGAGGCCTGTCGGGCAGGAAA
AGGCTGCAGCTGCCTTCCTGGGAGAAGGAGGAGATGAGTGTATCCTGAACACCTATTA
TGTGCTAGGGGCTATTGTAGATACATGACATTATCATGCTCATTTTCACAAATGAGGA
AACTGAGGCTCAGAAGACTTAAATTATTTGCCCAAGAGTTCATAAATGATAGAGCCAG
CATTAGAGTCCAGGACTGTCTGATTTCAGACCTAAGCTGTTCCCTCTGCACATCATGT
CCCACCAGTAAGGAAGATCTGGGTCTCAGAGCTGAGCCAAGACCTCCCGGGTCCTCTG
GGGTTTTCTGTGTCTTTCAGAGTGGCCTACATGGAGGTCTGAGAAAGGCCACAGCACA
AAATTGGGCACTGGAGCAAAGAGGAGTGGTGTGGAGGCCTGGCTAAGTATTGACCAAT
GAGCAGGAGTAGGGGCCAGAGCTTGGAGCCCTCAGGTGATAGGTGACCAGGCTGTTGT
TCCACTTTGAAATGCAGGCCCCAGACTAAGGACGGCAGCGAAGCAGAGCTCCCTCGTT
GGTGCAGAGGTTCCCTAAGCTCCTGGGCATTCGTAAGAACTGAGGTTTGCCTTTATTC
TTCTCATGGCTCATGTTATAGCCATTCACTCCAGAAAGCCTGGCACGTCATGGACTGT
TCAAGGCTGTGCTCCATAGAGTAATGATGCCTCCAGCTATGCGAGGCTTTGGGCCCAC
CCTTCCCACTGCCCCTGAGGGCTAAGGGGAGCCCTTCCTTCTGCTCCTGCCTGCTCAC
CAGTGTGCCTTTATTAGTTCGGTGGAGAAACCTCGGTAGGCAGGAGGCATAAGTCCAG
CCACAGAAACCCTGTGCAGATGAGGCTGGGGATGTAGTGAGTGCTGCAGAAGTGAGTG
ACTCAGACACAGAAGAGCTTCGGGTGACAAGCACTAGGACATAGCATTGGATGGGGGG
GAGGTGGGACAAGGAGAGTACTGCCTGTCCTGATGTCTGTCTTCCCTAGCTCCCAGCT
CTTCACCGATGGGGATCATTGTGGCTGTGGTCACTGGGATTGCTGTAGCGGCCATTGT
TGCTGCTGTAGTGGCCTTGATCTACTGCAGGAAAAAGCGGATTTCAGGTTTGTAGCTC
CTCCCGGTCCCTTTGGTTATCAGTTTCCACTTGGCCCAGGCCCTAACCCCAGACATTG
CCAGAATCCCTCTCTTTGGGCTAGATACACATTCAGATCTAGGCCCGTATTGTATTAT
AGTCATTCATTCGTTTATCTTGCTGAGTGGATGACAAAAAGAGGGGAATTGTTAAAGG
AAAATTTAAATGGAGACTGGAAAAATTCCTGAGCAAACAAAACCACCTGGCCCTTAGA
AATAGCTTTAACTTTGCTTAAACTACAAACACAAGCAAAACTTCACGGGGTCATACTA
CATACAAGCATAAGCAAAACTTAACTTGGATGATTTCTGGTAAATGCTTATGTTAGAA
ATAAGACAACCCCAGCCAATCACAAGCAGCCTACTAACATATAATTAGGTGACTAGGG

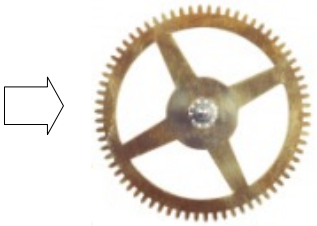
Imperial College London

The Human Genome

Functional annotation of the genome sequence

The Human Genome Project
identify and map genes

CATGATGAATTCATTTAACCCTTGAGTTTCCATTTCTTCACCTGCTCCGTGGGGCACT
AACGCCTCCCTCAGAGGCTTCTGGTGAGAATCAGTGTTTCCCTGCCCCCGCCCCGCCC
TCCATGCCCCTTCTCCACGTTCTCACTGTGCTAGGTGCTCTTCTCTGTCTTTCTCTTC
CACCAGCCTGTGGGAAACCTGAGATGAAAGTCGTGTCTTACCCATCTTTGTATTTCCA
GCATCTGAAACTGGGCAGAGCTTAATAAATATTTTGCTGGAGAGGTTGATGATCTTAC
AAAGCTCCCATTGAAAGGTGGCTCTCTGTAAAGCAAAGTTACAATGAGATTGTGATGA
ACATTGTCCTTGTGGCTTTTCACTTAGTCCCCTCCCTTCACCTGAAGAGCAAATTTTC
CTCAAAAGTACACAGCAAACAAATGACCCACTGGTGACACTGCTGCCTTTAGACCCTG
CTGGAAAGAATCTCCACATTTATTAACATTCCCAAAGTAAATTTATCAGGTAGCATTC
ATCAGGTAACATTTGTTGCACATTCATGACTTTTCTACTGTCCACAAAGGCATATGTC
CTTATCATATGCAGACTCCTTGGCCACACTGGATTCCTCCTTCCCTCCTCGACATGGA
AGAGATGGCATCTTAGGGTCTCTTGTGTTCTTCCTGAAGAGGCCTGTCGGGCAGGAAA
AGGCTGCAGCTGCCTTCCTGGGAGAAGGAGGAGATGAGTGTATCCTGAACACCTATTA
TGTGCTAGGGGCTATTGTAGATACATGACATTATCATGCTCATTTTCACAAATGAGGA
AACTGAGGCTCAGAAGACTTAAATTATTTGCCCAAGAGTTCATAAATGATAGAGCCAG
CATTAGAGTCCAGGACTGTCTGATTTCAGACCTAAGCTGTTCCCTCTGCACATCATGT
CCCACCAGTAAGGAAGATCTGGGTCTCAGAGCTGAGCCAAGACCTCCCGGGTCCTCTG
GGGTTTTCTGTGTCTTTCAGAGTGGCCTACATGGAGGTCTGAGAAAGGCCACAGCACA
AAATTGGGCACTGGAGCAAAGAGGAGTGGTGTGGAGGCCTGGCTAAGTATTGACCAAT
GAGCAGGAGTAGGGGCCAGAGCTTGGAGCCCTCAGGTGATAGGTGACCAGGCTGTTGT
TCCACTTTGAAATGCAGGCCCCAGACTAAGGACGGCAGCGAAGCAGAGCTCCCTCGTT
GGTGCAGAGGTTCCCTAAGCTCCTGGGCATTCGTAAGAACTGAGGTTTGCCTTTATTC
TTCTCATGGCTCATGTTATAGCCATTCACTCCAGAAAGCCTGGCACGTCATGGACTGT
TCAAGGCTGTGCTCCATAGAGTAATGATGCCTCCAGCTATGCGAGGCTTTGGGCCCAC
CCTTCCCACTGCCCCTGAGGGCTAAGGGGAGCCCTTCCTTCTGCTCCTGCCTGCTCAC
CAGTGTGCCTTTATTAGTTCGGTGGAGAAACCTCGGTAGGCAGGAGGCATAAGTCCAG
CCACAGAAACCCTGTGCAGATGAGGCTGGGGATGTAGTGAGTGCTGCAGAAGTGAGTG
ACTCAGACACAGAAGAGCTTCGGGTGACAAGCACTAGGACATAGCATTGGATGGGGGG
GAGGTGGGACAAGGAGAGTACTGCCTGTCCTGATGTCTGTCTTCCCTAGCTCCCAGCT
CTTCACCGATGGGGATCATTGTGGCTGTGGTCACTGGGATTGCTGTAGCGGCCATTGT
TGCTGCTGTAGTGGCCTTGATCTACTGCAGGAAAAAGCGGATTTCAGGTTTGTAGCTC
CTCCCGGTCCCTTTGGTTATCAGTTTCCACTTGGCCCAGGCCCTAACCCCAGACATTG
CCAGAATCCCTCTCTTTGGGCTAGATACACATTCAGATCTAGGCCCGTATTGTATTAT
AGTCATTCATTCGTTTATCTTGCTGAGTGGATGACAAAAAGAGGGGAATTGTTAAAGG
AAAATTTAAATGGAGACTGGAAAAATTCCTGAGCAAACAAAACCACCTGGCCCTTAGA
AATAGCTTTAACTTTGCTTAAACTACAAACACAAGCAAAACTTCACGGGGTCATACTA
CATACAAGCATAAGCAAAACTTAACTTGGATGATTTCTGGTAAATGCTTATGTTAGAA
ATAAGACAACCCCAGCCAATCACAAGCAGCCTACTAACATATAATTAGGTGACTAGGG



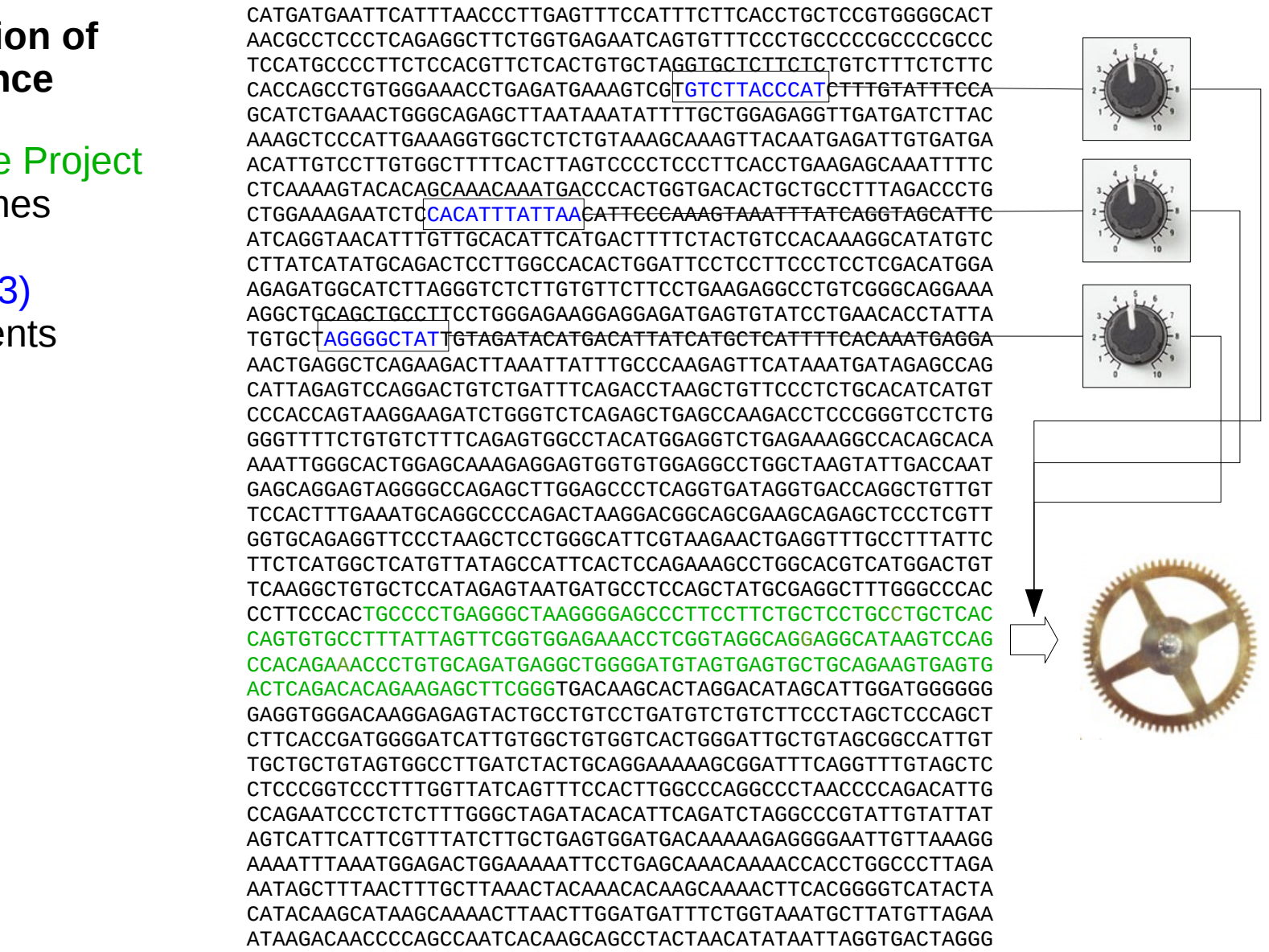
Imperial College London

The Human Genome

Functional annotation of the genome sequence

The Human Genome Project
identify and map genes

Encode project (2003)
find functional elements



Imperial College London

The Human Genome

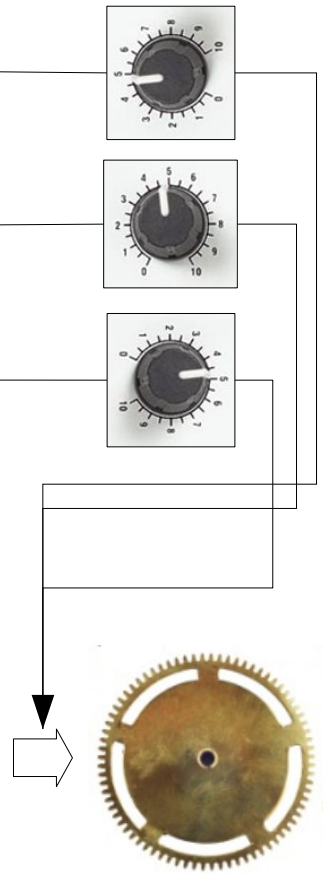
Functional annotation of the genome sequence

The Human Genome Project
identify and map genes

Encode project (2003)
find functional elements

1000 genomes project (2008)
identify inter-individual sequence variation

CATGATGAATTCATTTAACCCTTGAGTTTCCATTTCTTCACCTGCTCCGTGGGGCACT
AACGCCTCCCTCAGAGGCTTCTGGTGAGAATCAGTGTTTCCCTGCCCCCGCCCCGCCC
TCCATGCCCCTTCTCCACGTTCTCACTGTGCTAGGTGCTCTTCTCTGTCTTTCTCTTC
CACCAGCCTGTGGGAAACCTGAGATGAAAGTCGTGTCTTACCCATCTTTGTATTTCCA
GCATCTGAAACTGGGCAGAGCTTAATAAATATTTTGCTGGAGAGGTTGATGATCTTAC
AAAGCTCCCATTGAAAGGTGGCTCTCTGTAAAGCACAATGAGATTGTGATGA
ACATTGTCCTTGTGGCTTTTCACTTAGTCCCCTCCTGAAGAGCAAATTTTC
CTCAAAAGTACACAGCAAACAAATGACCCACTGGTGACACTGCTGCCTTTAGACCCTG
CTGGAAAGAATCTCCACATTTATTAACATTCCCAAAGTAAATTTATCAGGTAGCATTC
ATCAGGTAACATTTGTTGCACATTCATGACTTTTCTACTGTCCACAAAGGCATATGTC
CTTATCATATGCAGACTCCTTGGCCACACTGGATTCCTCCTTCCCTCCTCGACATGGA
AGAGATGGCATCTTAGGGTCTCTTGTGTTCTTCCTGAAGAGGCCTGTCGGGCAGGAAA
AGGCTGCAGCTGCCTTCCTGGGAGAAGGAGGAGATGAGTGTATCCTGAACACCTATTA
TGTGCTAGGGGCTATTGTAGATACATGACATTATCATGCTCATTTTCACAAATGAGGA
AACTGAGGCTCAGAAGACTTAAATTATTTGCCCAAGAGTTCATAAATGATAGAGCCAG
CATTAGGACTGTCTGATTTCAGACCTAAGCTGTTCCCTCTGCACATCATGT
CCCACCAAGATCTGGGTCTCAGAGCTGAGCCAAGACCTCCCGGGTCCTCTG
GGGTTTTCTGTGTCTTTCAGAGTGGCCTACATGGAGGTCTGAGAAAGGCCACAGCACA
AAATTGGGCACTGGAGCAAAGAGGAGTGGTGTGGAGGCCTGGCTAAGTATTGACCAAT
GAGCAGGAGTAGGGGCCAGAGCTTGGAGCCCTCAGGTGATAGGTGACCAGGCTGTTGT
TCCACTTTGAAATGCAGGCCCCAGACTAAGGACGGCAGCGAAGCAGAGCTCCCTCGTT
GGTGCAGAGGTTCCCTAAGCTCCTGGGCATTCGTAAGAACTGAGGTTATTC
TTCTCATGGCTCATGTTATAGCCATTCACTCCAGAAAGCCTGGCACGCTGT
TCAAGGCTGTGCTCCATAGAGTAATGATGCCTCCAGCTATGCGAGGCTTTGGGCCCAC
CCTTCCCACTGCCCCTGAGGGCTAAGGGGAGCCCTTCCTTCTGCTCCTGCCTGCTCAC
CAGTGTGCCTTTATTAGTTCGGTGGAGAAACCTCGGTAGGCAGGAGGCATAAGTCCAG
CCACAGAAACCCTGTGCAGATGAGGCTGGGGATGTAGTGAGTGCTGCAGAAGTGAGTG
ACTCAGACACAGAAGAGCTTCGGGTGACAAGCACTAGGACATGGATGGGGGG
GAGGTGGGACAAGGAGAGTACTGCCTGTCCTGATGTCTGTCAGCTCCCAGCT
CTTCACCGATGGGGATCATTGTGGCTGTGGTCACTGGGATTGCTGTAGCGGCCATTGT
TGCTGCTGTAGTGGCCTTGATCTACTGCAGGAAAAAGCGGATTTCAGGTTTGTAGCTC
CTCCCGGTCCCTTTGGTTATCAGTTTCCACTTGGCCCAGGCCCTAACCCCAGACATTG
CCAGAATCCCTCTCTTTGGGCTAGATACACATTCAGATCTAGGCCCGTATTGTATTAT
AGTCATTCATTCGTTTATCTTGCTGAGTGGATGACAAAAAGAGGGGAATTGTTAAAGG
AAAATTTAAATGGAGACTGGAAAAATTCCTGAGCAAACAAAACCACCTGGCCCTTAGA
AATAGCTTTAACTTTGCTTAAACTACAAACACAAGCAAAACTTCACGGGGTCATACTA
CATACAAGCATAAGCAAAACTTAACTTGGATGATTTCTGGTAAATGCTTATGTTAGAA
ATAAGACAACCCCAGCCAATCACAAGCAGCCTACTAACATATAATTAGGTGACTAGGG



Imperial College London

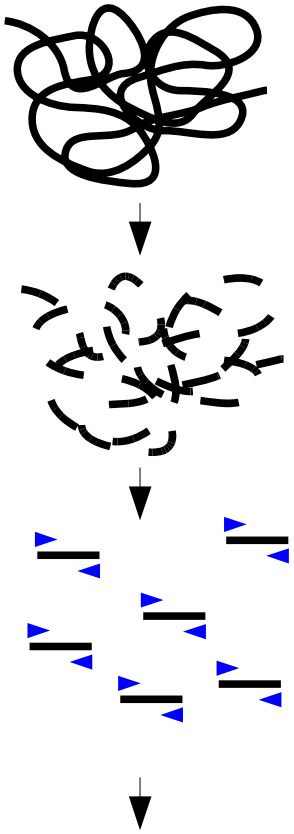
Variant detection by next-generation sequencing

genomic DNA

fragmented DNA

paired-end sequencing

read mapping



Illumina HiSeq 2000

```
..CTACAT        GAGAAT...CACA--A          ACTGGA..........TGGCGT        GGCTAA.........ACGAGC
GCCTACATGGAGGTCTGAGAAAGGCCACAGCACAAAATTGGGCACTGGAGCAAAGAGGAGTGGTGTGGAGGCCTGGCTAAGTATTGACCAATGAGCA
```

Imperial College London

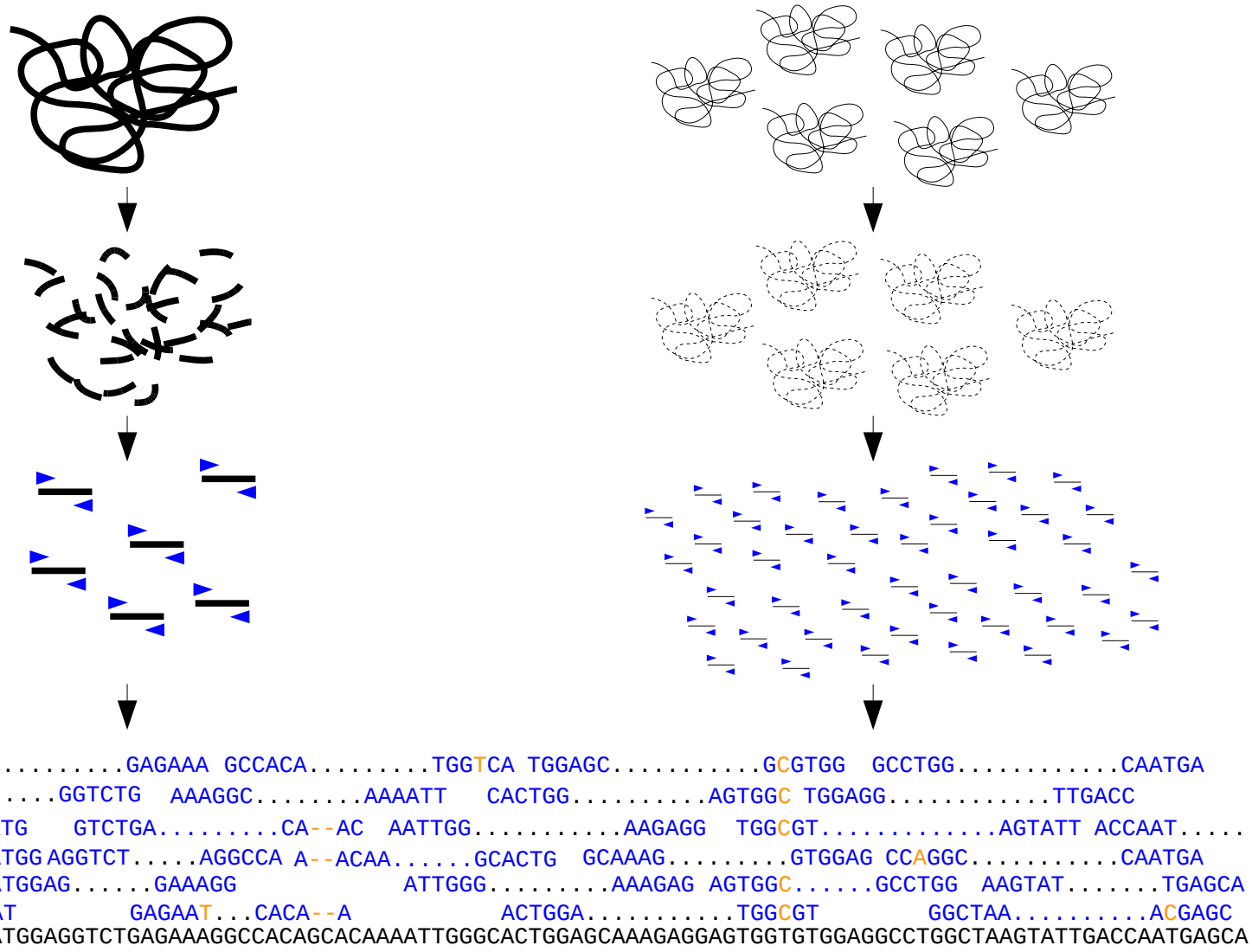
Variant detection by next-generation sequencing

genomic DNA

fragmented DNA

paired-end sequencing

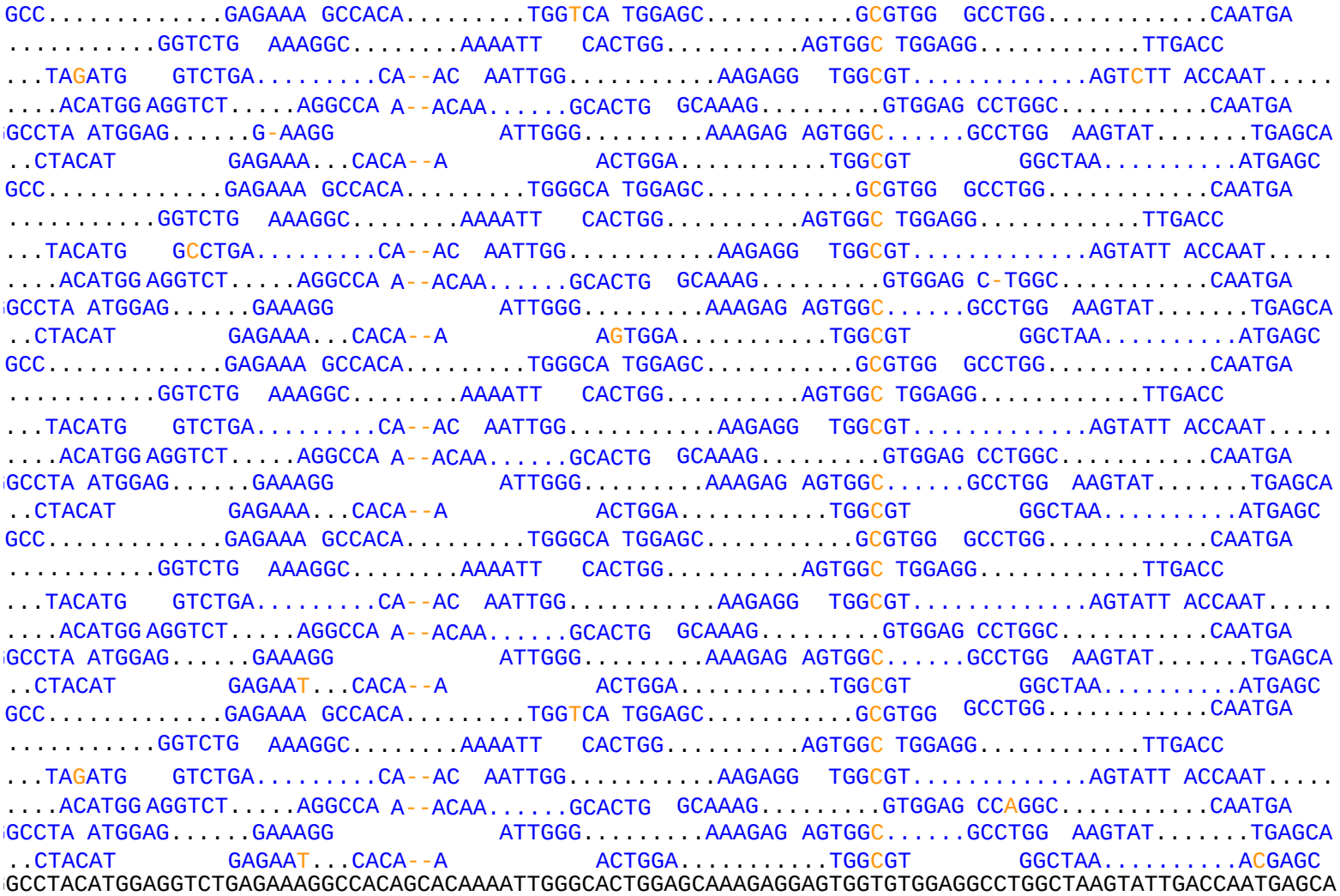
read mapping



Imperial College London

Variant detection by next-generation sequencing

read mapping



Imperial College London

Next-Generation Sequencing (NGS)

Throughput

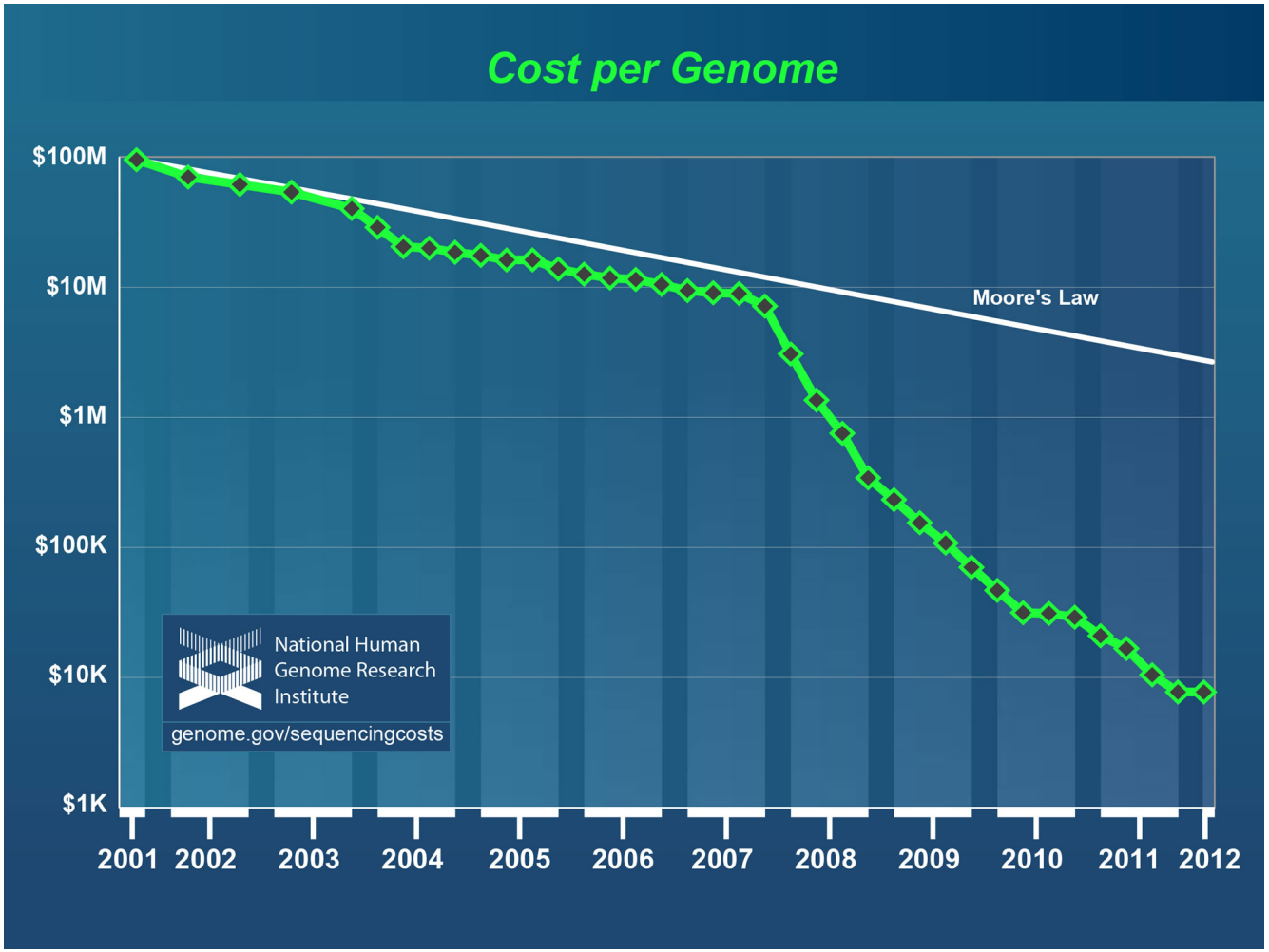
Illumina HiSeq 2000 run

- 10 days runtime
- ~6 billion 100bp paired-end reads
- ~30x coverage of 6 human genome
- ~220Gb of (compressed) raw data

Imperial College London

Next-Generation Sequencing (NGS)

Reduction in DNA sequencing costs



Imperial College London

Next-Generation Sequencing (NGS)

Informatics Challenges

- **data storage & I/O**
  - ~7TB raw data per sequencer per year
- **mapping algorithm efficiency**
  - 1st generation: hash-based
  - 2nd generation: Burrows-Wheeler transform
    - 10 times faster at similar sensitivity
- **CPU time**
  - 720 CPU hours (30 days) for ~1 billion reads (30x coverage of the human genome)

Imperial College London

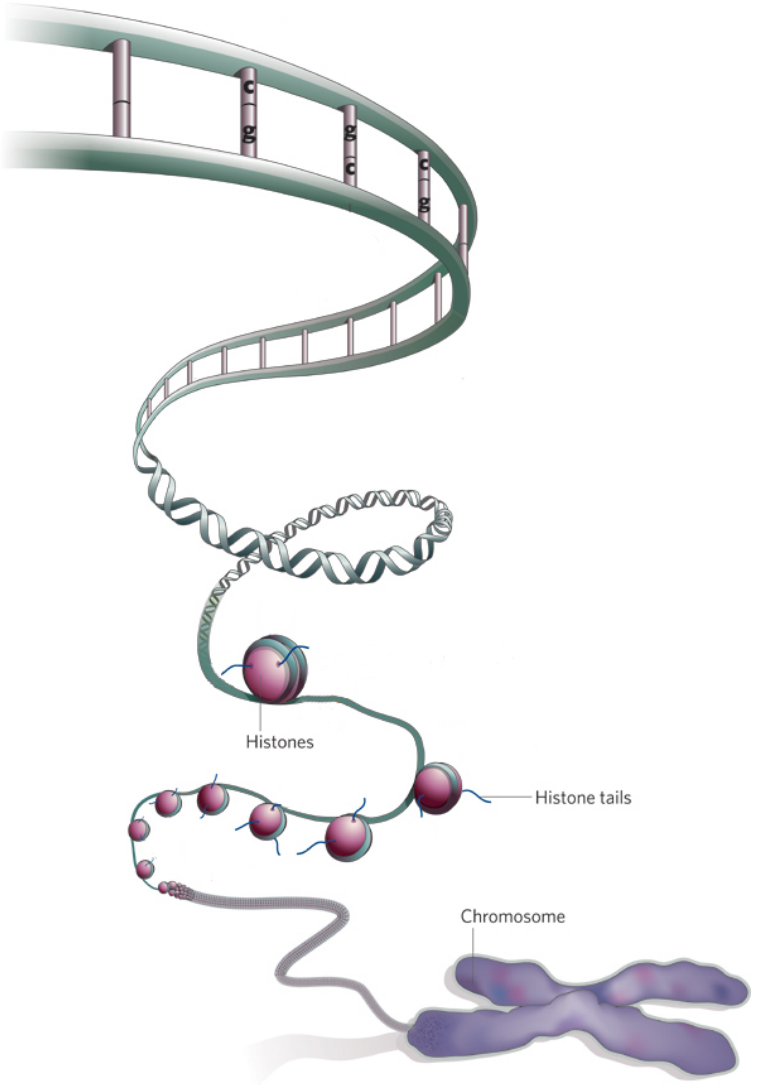
Next-Generation Sequencing (NGS)

NGS data analysis using Imperial College HPC service

- dedicated access to 112 CPU cores on PC cluster
- access to SGI Altix UV system with 256 CPU cores and up to 3TB of shared memory
- 60TB storage

Imperial College
London

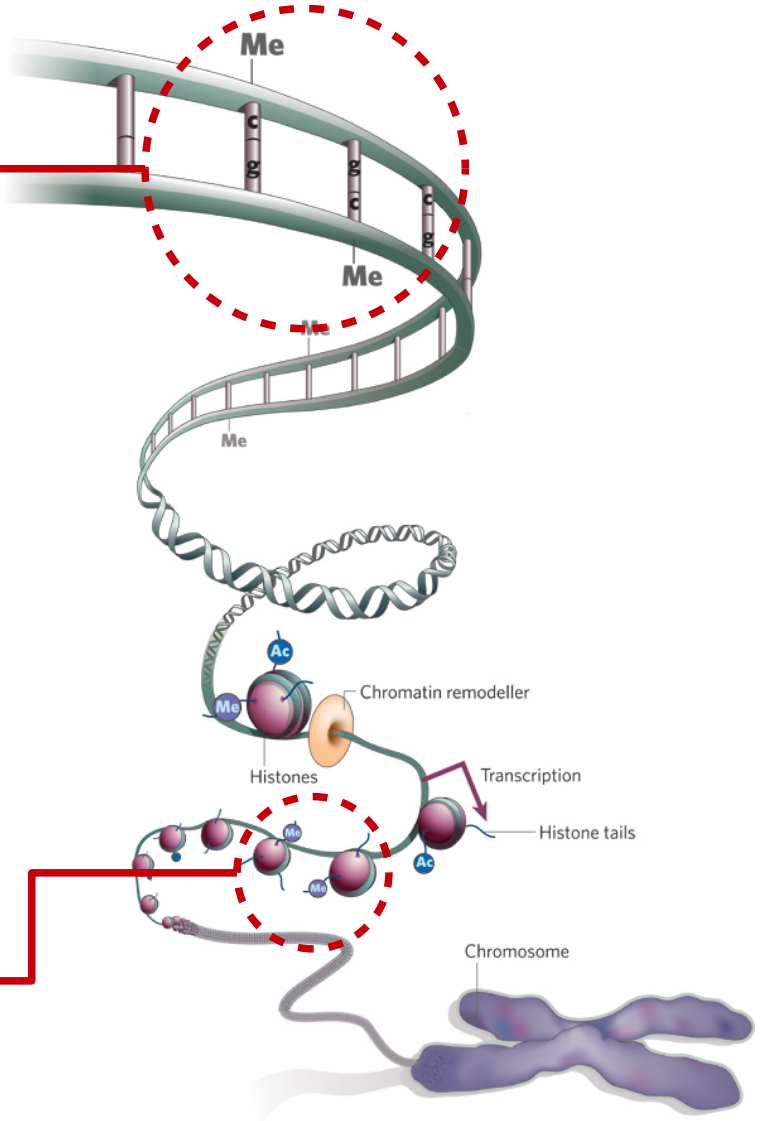
From genome to epigenome



Imperial College
London

From genome to epigenome

DNA methylation



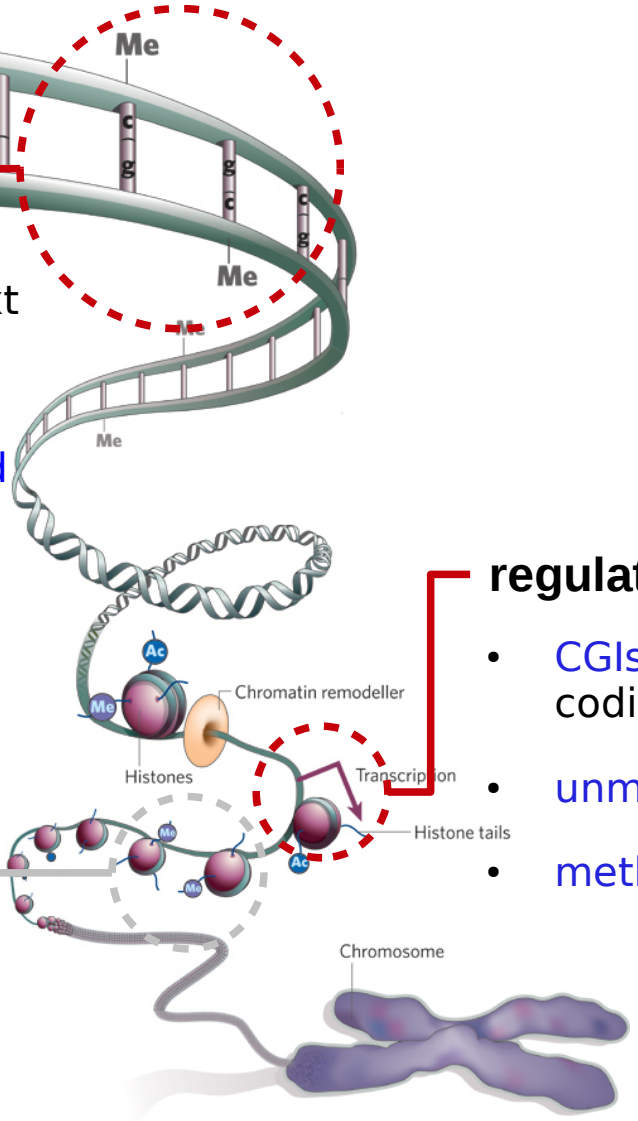
histone modification

Imperial College London

DNA methylation & gene expression

DNA methylation

- occurs in CpG dinucleotides context
- CpGs cluster in CpG Islands (CGI)
- genome globally highly methylated
- CGIs mostly unmethylated



regulation of transcription

- CGIs found in 50% of protein coding gene promoters
- unmethylated promoters active
- methylated promoters repressed

histone modification

Imperial College London

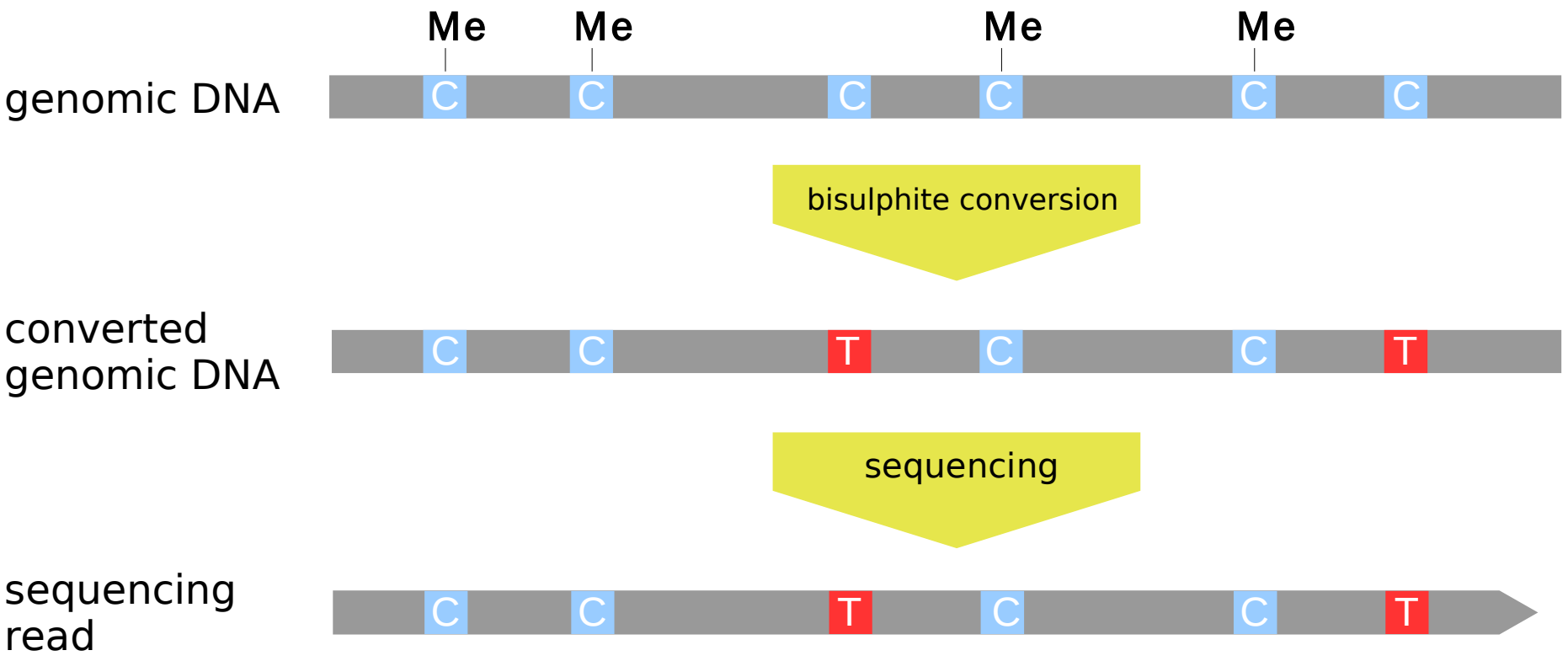
Genetic control of inter-individual variation in DNA methylation

Aims

- measure DNA methylation levels at single nucleotide resolution in two rat inbred strains (Brown Norway & Spontaneously Hypertensive Rat)
- identify inter-strain differences in DNA methylation
- determine inheritance and genetic regulators of DNA methylation

Imperial College London

Measuring DNA methylation by bisulphite sequencing



Imperial College London

Measuring DNA methylation by bisulphite sequencing

bisulphite sequencing reads

Imperial College London

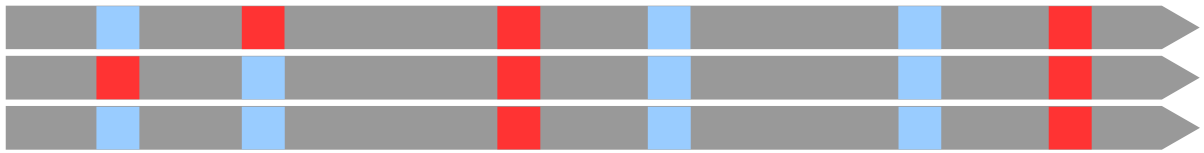
Measuring DNA methylation by bisulphite sequencing

bisulphite sequencing reads

Imperial College London

Measuring DNA methylation by bisulphite sequencing

bisulphite sequencing reads

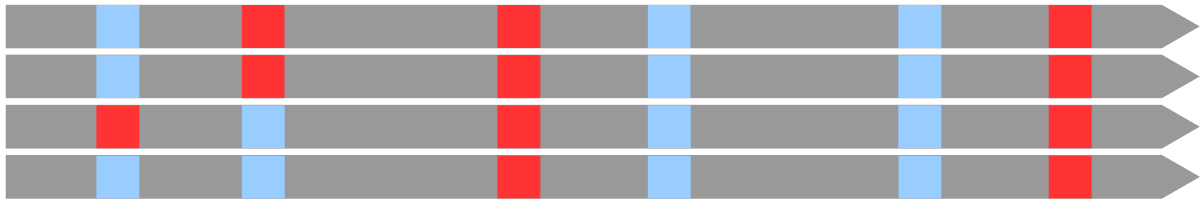


Imperial College London

Measuring DNA methylation by bisulphite sequencing

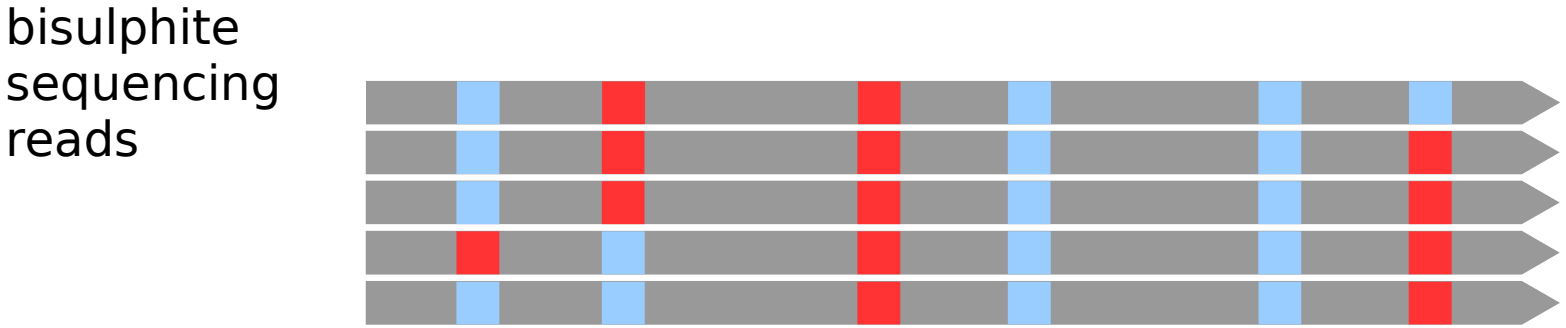
genomic DNA

bisulphite sequencing reads



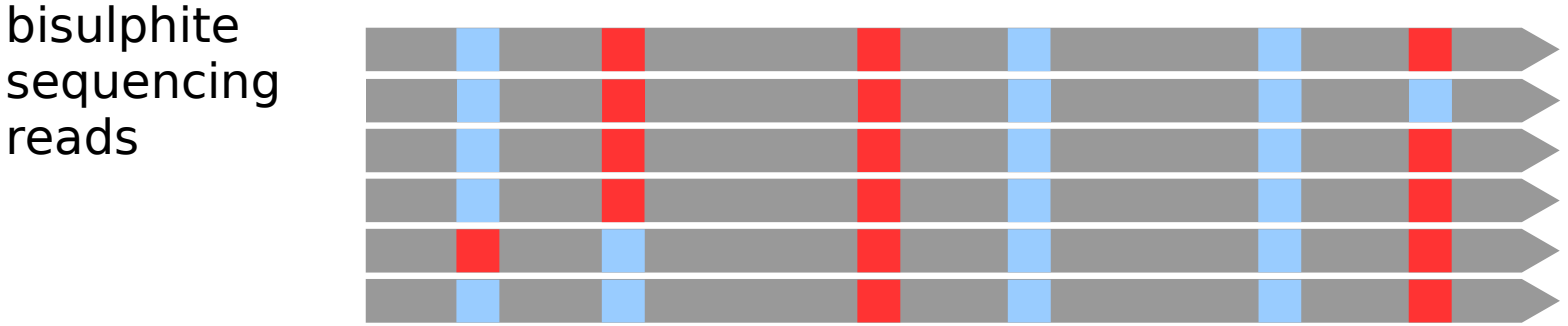
Imperial College London

Measuring DNA methylation by bisulphite sequencing



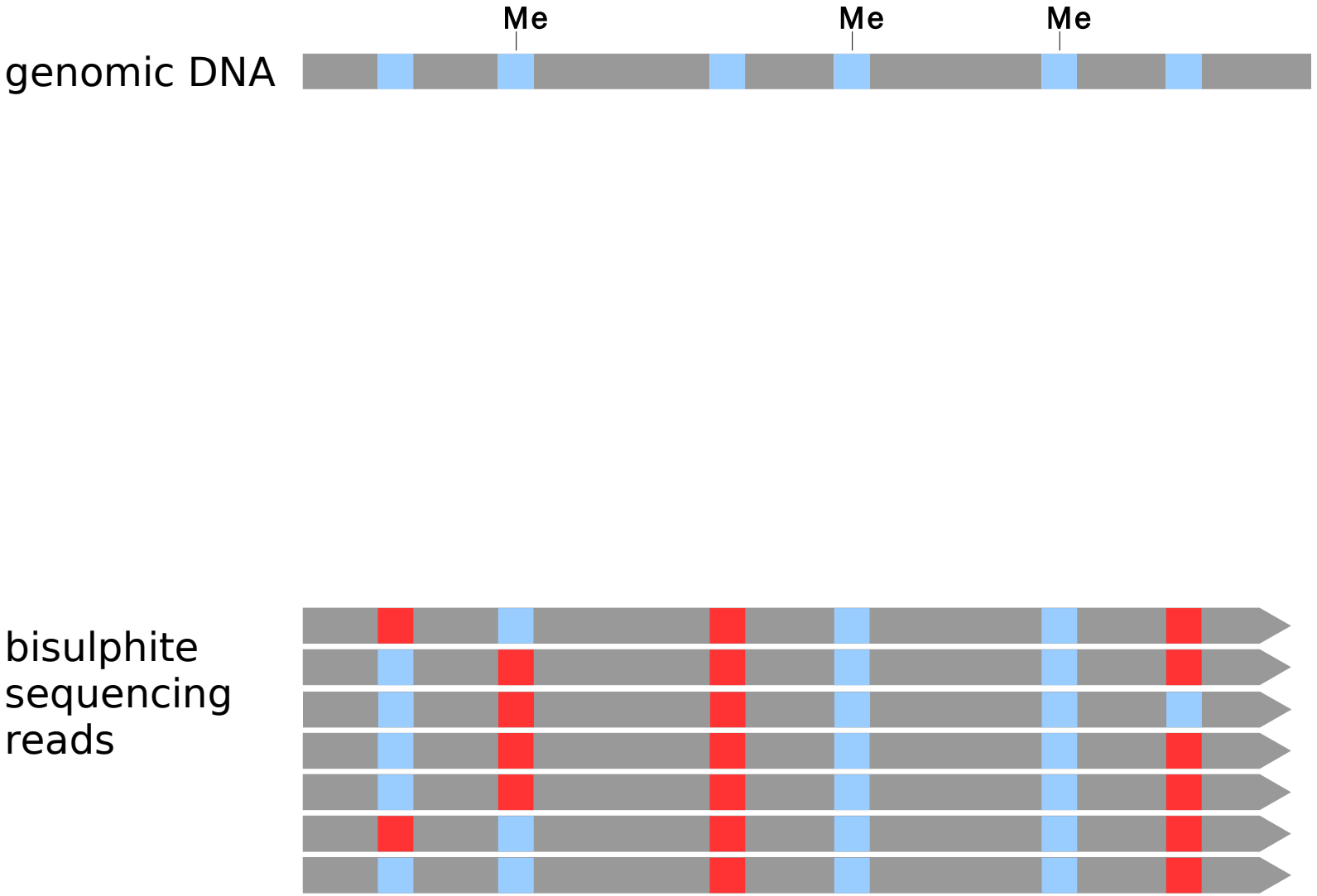
Imperial College London

Measuring DNA methylation by bisulphite sequencing



Imperial College London

Measuring DNA methylation by bisulphite sequencing

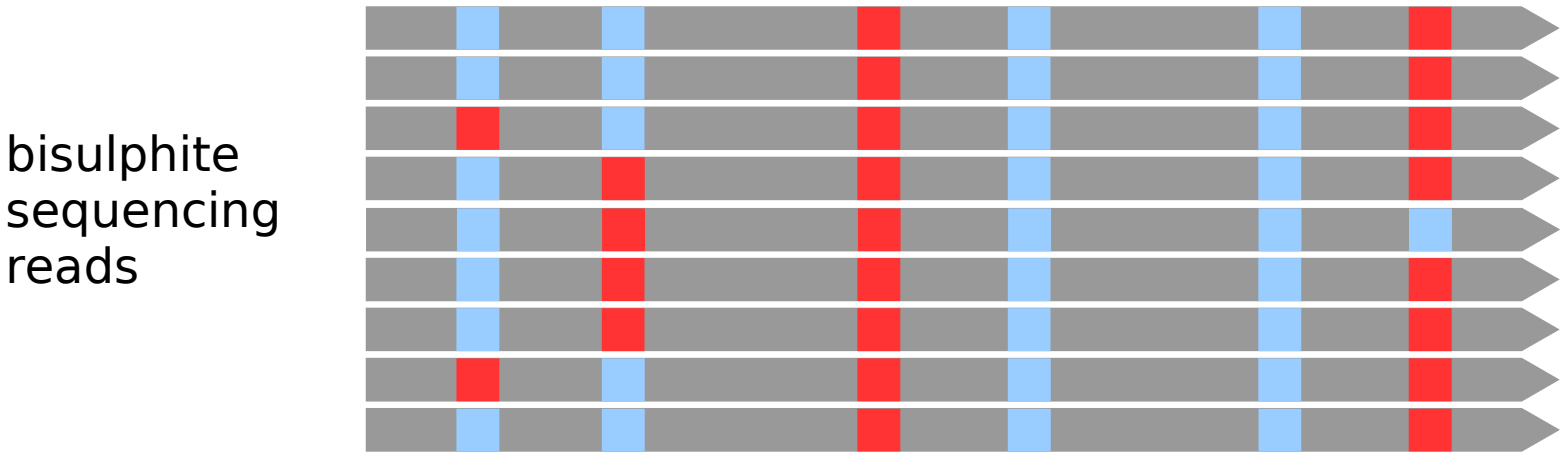


Imperial College London

Measuring DNA methylation by bisulphite sequencing

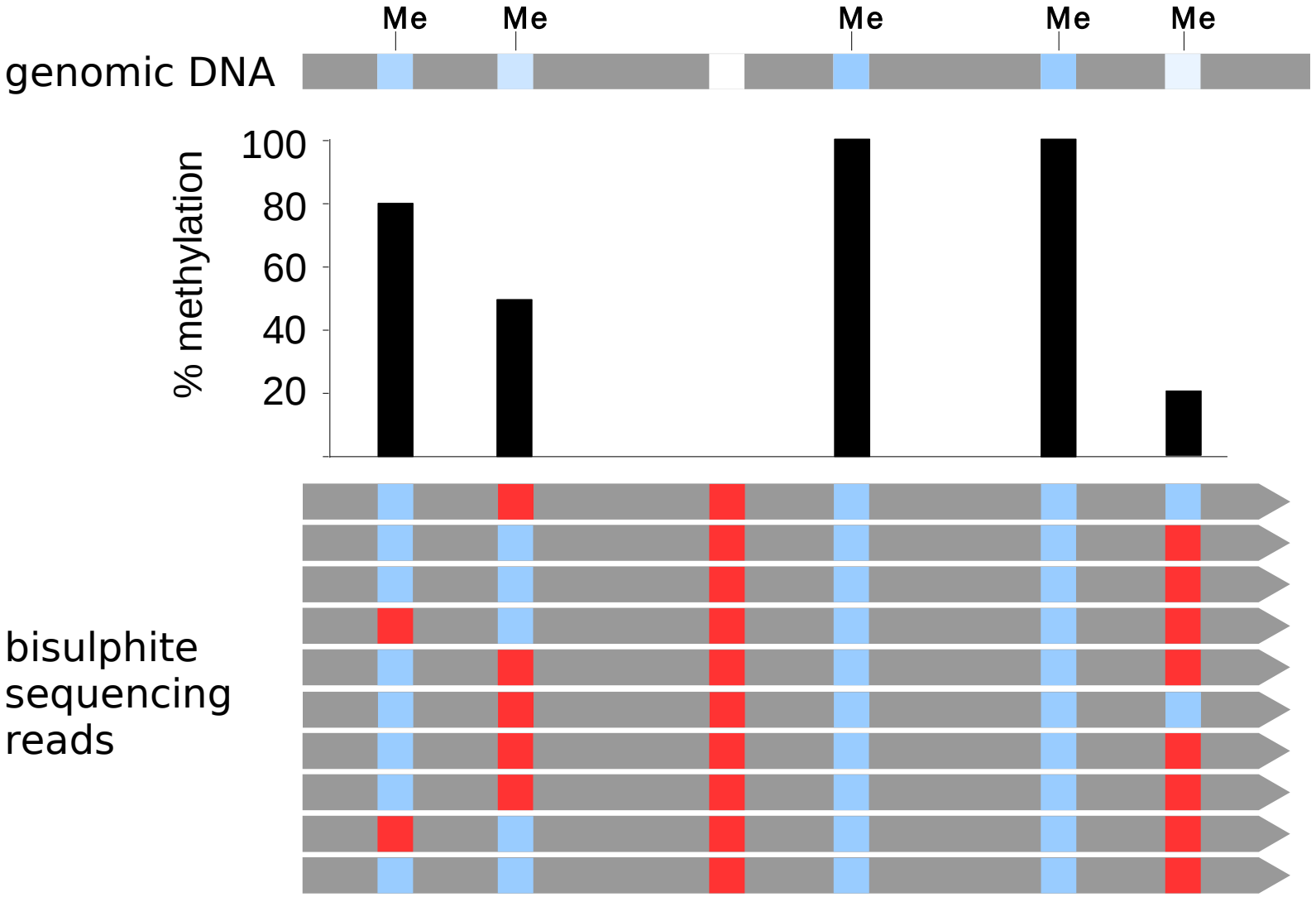
Imperial College London

Measuring DNA methylation by bisulphite sequencing



Imperial College London

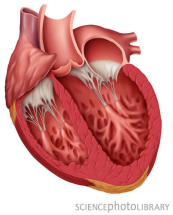
Measuring DNA methylation by bisulphite sequencing



Imperial College London

Sequencing the SHR & BN methylomes

- Illumina paired-end sequencing
- 100 bp read length
- 4 lanes per sample



- left ventricle of the heart

- SHR/Ola, BN-Lx, reciprocal F1 crosses
- 4 biological replicates
- six weeks old

Imperial College London

Sequencing the rat methylome

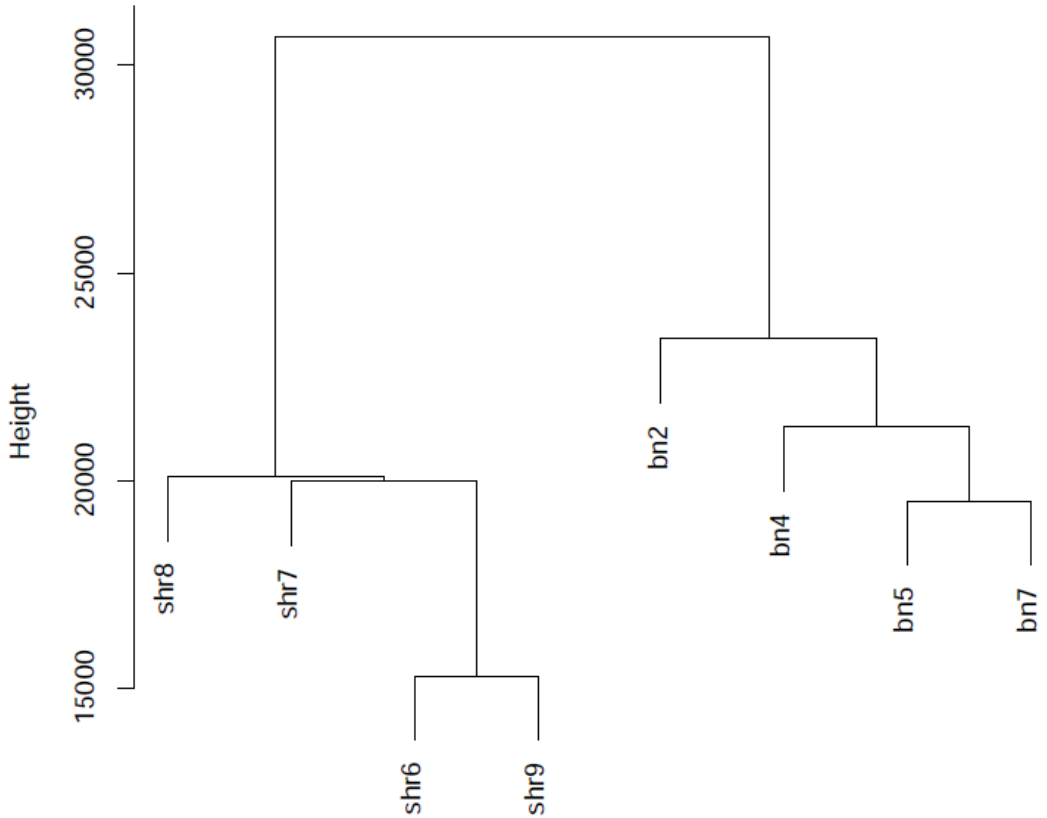
sample	reads	mapped reads		depth of coverage	depth of coverage after filtering*
	[million]	[million]	[%]	[x-fold]	[x-fold]
BN2	258	230	89	7	4
BN4	413	361	87	13	7
BN5	454	404	89	11	8
BN7	370	340	92	12	9
BN total	**1,495**	**1,336**	**89**	**42**	**29**
SHR6	497	448	91	15	10
SHR7	452	398	89	12	8
SHR8	421	369	90	11	8
SHR9	391	356	91	12	8
SHR total	**1,762**	**1,571**	**89**	**50**	**34**

* filtering by read clonality, uniqueness, mapping quality and strand specificity

Imperial College London

Inter- vs intra-strain differences in CpG methylation

Variability in CpG methylation greater between than within inbred strains



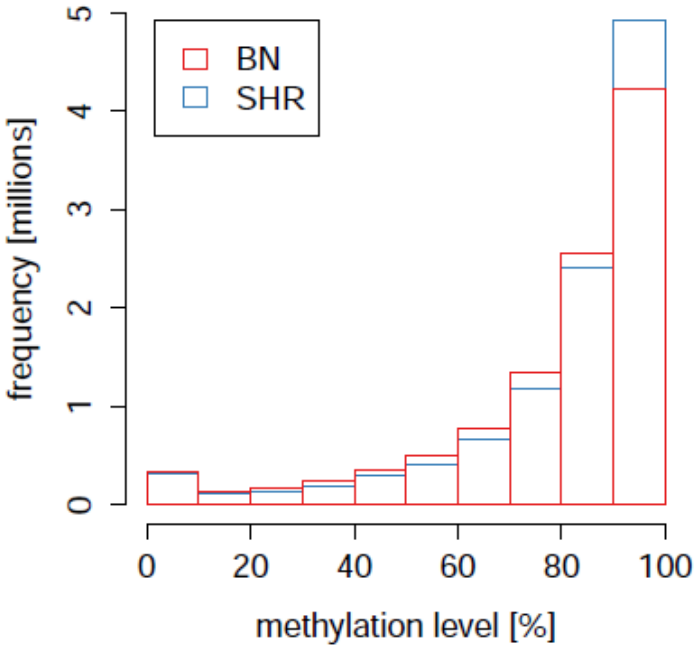
Distance measure: Euclidean
Clustering method: Ward

Imperial College London

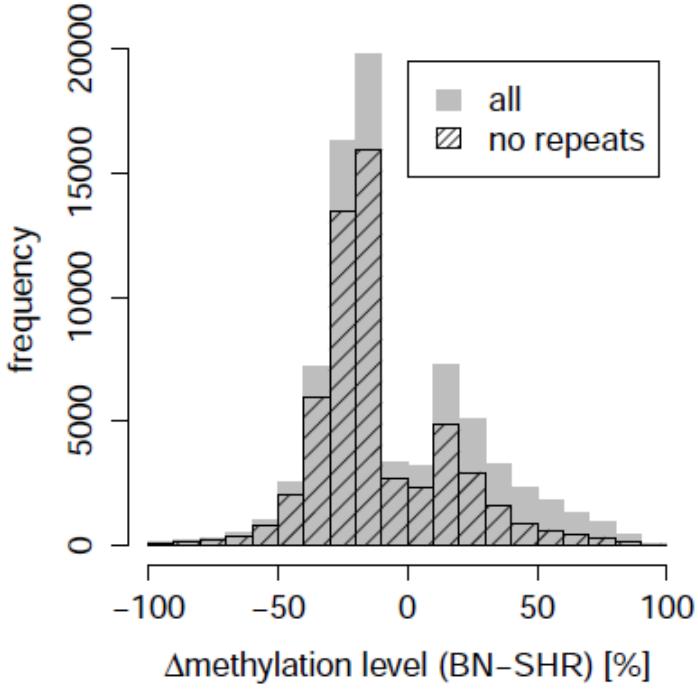
Distribution of methylation levels and inter-strain differences in methylation

analysis of 10.6 million CpG dinucleotides identified
77,088 differentially methylated CpGs (<1%)

The rat genome is globally highly methylated



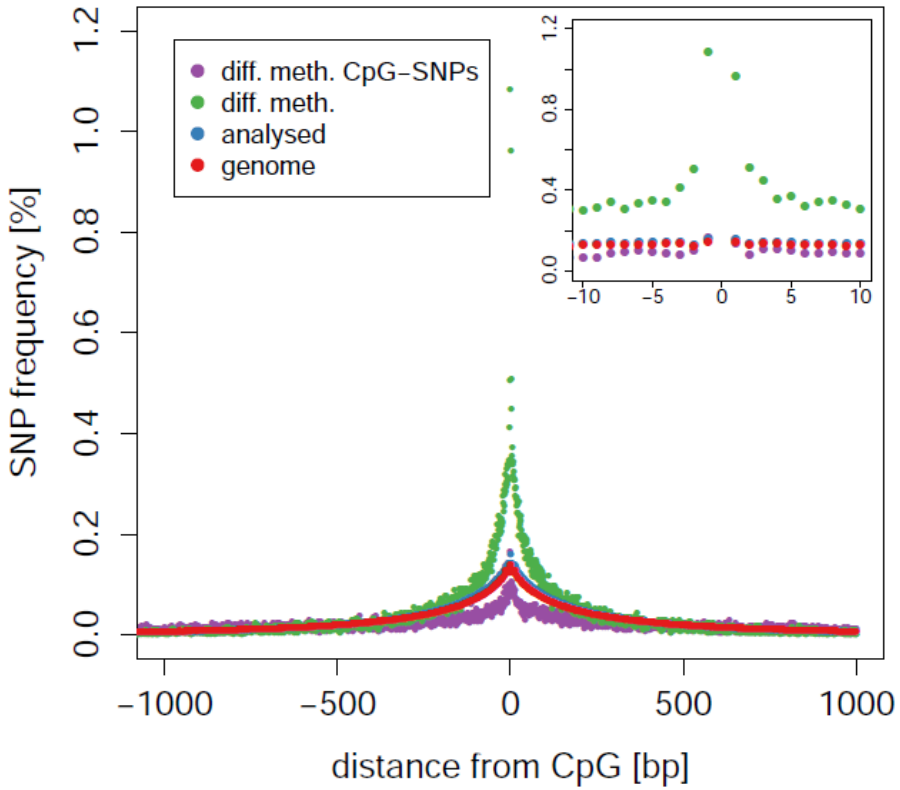
The majority of methylation differences are small



Imperial College London

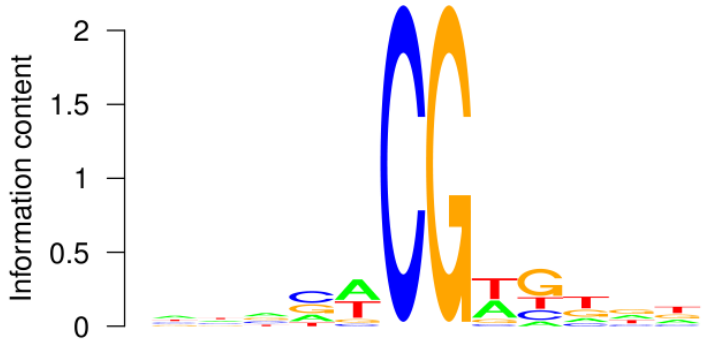
Sequence variation at differentially methylated CpGs

Sequence variation is higher in the proximity of differentially methylated CpGs

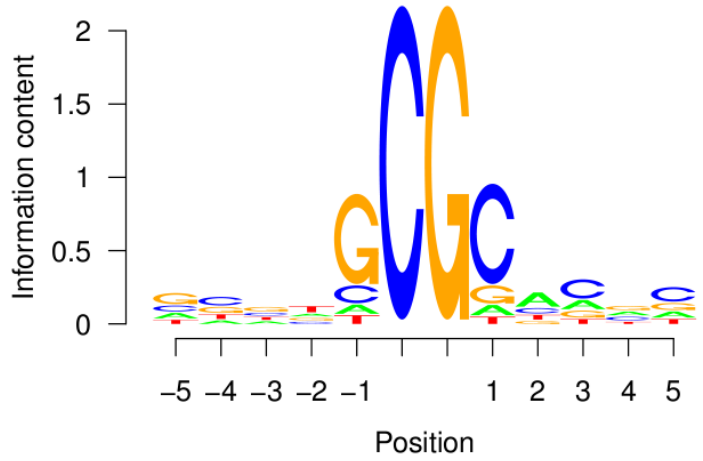


Increased and decreased methylation is associated with specific sequence motifs

increased methylation



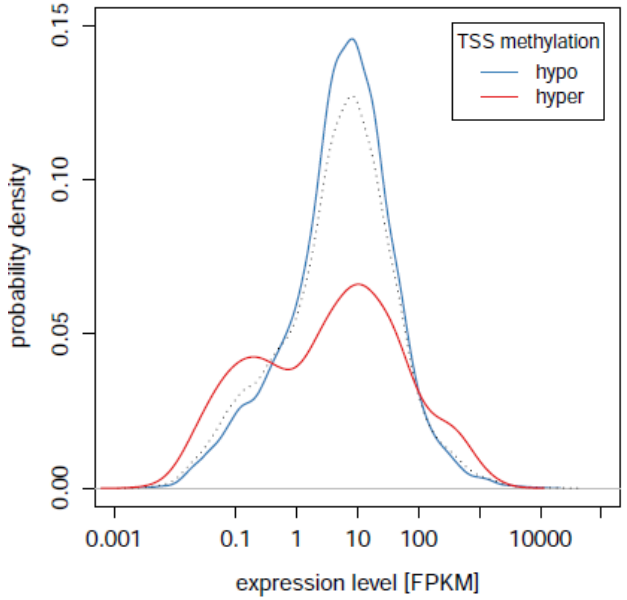
decreased methylation



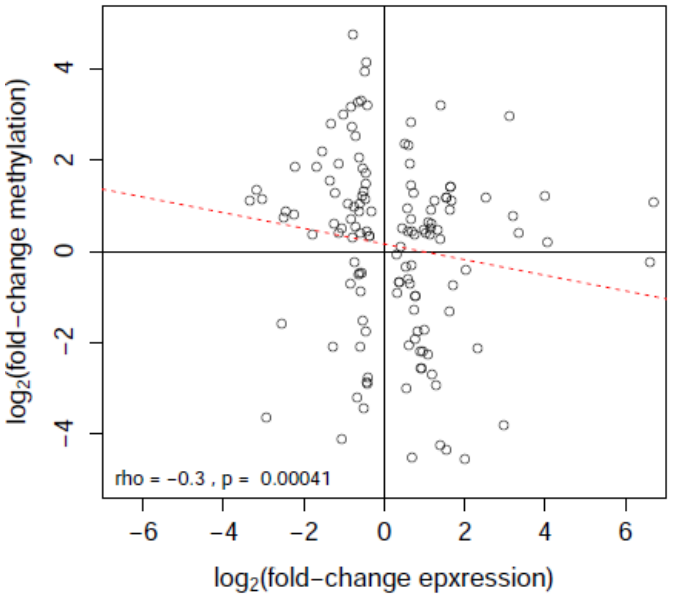
Imperial College London

Methylation and gene expression

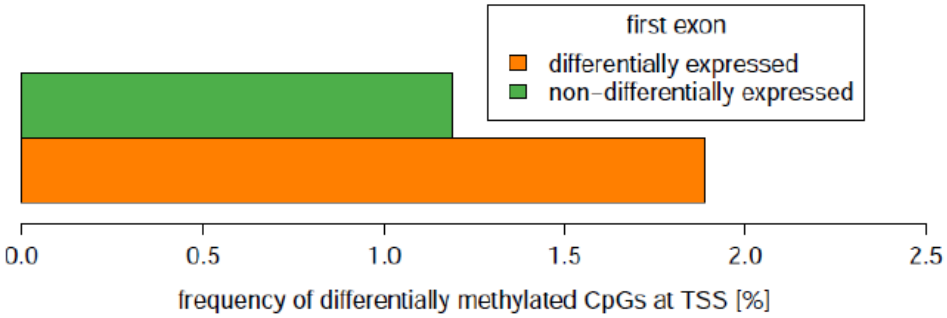
Genes with highly and lowly methylated promoters have different distribution of expression levels



Methylation and gene expression differences are negatively correlated



Promoters of differentially expressed genes are enriched in differentially methylated CpGs



Imperial College London

Genetic control of inter-individual variation in DNA methylation

Summary

- generated whole-genome methylation profiles at single nucleotide resolution in two rat strains (SHR & BN)
- quantified methylation differences between SHR & BN
  - inter-strain methylation differences are greater than intra-strain differences
  - majority of methylation differences are small
  - SNP frequency in proximity of differential methylation CpGs is increase
  - increased/decreased methylation associated with specific sequence motifs
- integrated methylation and gene expression profiles
  - clear but complex association of differential methylation and differential expression

Imperial College London

Acknowledgments

Physiological Genomics & Medicine Group

Michelle Johnson
(Post-doctoral Research Associate)

Klio Maratou
(Post-doctoral Research Associate)

Prashant Shrivastava
(Post-doctoral Research Associate)

Prof. Timothy J. Aitman
(Group Head)

Cardiac Regeneration Group National Heart & Lung Institute

Prof. Nadia A. Rosenthal
(Head of the Heart Science Centre)

MRC CSC Genomics Laboratory

Laurence Game
(Facility Head)

Adam Giess
(Bioinformatician)

Korinne Northwood
(Research Assistant)

Imperial College High Performance Computing

Simon Burbidge
(HPC Co-ordination Manager)

British Heart Foundation

THE MAGDI YACOUB INSTITUTE