Tackling data analysis challenges in genome informatics using HPC

Michael Mueller
Physiological Genomics & Medicine Group
Clinical Genome Informatics Facility
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*
The Human Genome

The Human Genome Project
1990 - 2003

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

3.1 billion nucleotides (A,C,T,G)
The Human Genome
The Human Genome

Functional annotation of the genome sequence

The Human Genome Project identify and map genes
The Human Genome

Functional annotation of the genome sequence

The Human Genome Project identify and map genes

Encode project (2003) find functional elements
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

<table>
<thead>
<tr>
<th>T/A</th>
<th>G/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/G</td>
<td>G/C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATGATGAATTCATTTAACCCTTGAGTTTCCATTTCTTCACCTGCTCCGTGGGGCACT</td>
</tr>
<tr>
<td>GACCTGAAACTGGGAGAGCTTTAATAAAATTTTGCAGAGAGTTTAGATGACCTTAC</td>
</tr>
<tr>
<td>AAGCTGCTCCCATGGAAGTGCTCTGCTGTAAAGCCTAATGAGATTGTGATGCTTAC</td>
</tr>
<tr>
<td>CATGAGGCAGAAAATACGCTGATCAGGACTAGTTTCCCTGCCCCGGCCGAGAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
<tr>
<td>AGAGTATGGCCTTCTAGGGTCTCGTCTTTGCTGTGGTCACTGGGATTTGCTGTAGCGGCCATTGT</td>
</tr>
<tr>
<td>TCGAAGAAGAATCTCAGACATTTTTTTCATACGCTGACCTGCTGCAACTGAGG</td>
</tr>
<tr>
<td>ATACAGTAAATTTTTGTCACATTTCCATGATTCCTGTCACACAGGAAAGG</td>
</tr>
<tr>
<td>CTTATCATATAGCAGACTCCTGGCCACACTGGAATTTCTCCTTGCTCGACATAG</td>
</tr>
</tbody>
</table>

Sequence Diagram: 
- **T/A** changes to **G/T**
- **C/G** changes to **G/C**
Variant detection by next-generation sequencing

genomic DNA

fragmented DNA

paired-end sequencing

read mapping

Illumina HiSeq 2000

...CTACAT GAGAA...CACA--A ACTGGA...........TGGCGT GGCTAA............ACGAGC GCCTACATGGAGGTCTGAGAAAGGCCACAGCACAAAATTGGGCACTGGAGCAAGAGGAGTGGTGTGGAGGCCTGGCTAAGTATTGACCAATGAGCATCTTTC...

©2010, Illumina Inc. All rights reserved.
Variant detection by next-generation sequencing

**genomic DNA**

**fragmented DNA**

**paired-end sequencing**

**read mapping**

```plaintext
GCC..............GAGAAA GCCACA............TGGTCA TGGAGC............GCCGTGG GCCTGG............CAATGA
.............GTTCTG AAAGGC............AAAATT CACTGG............AGTGGC TGGAGG............TTGACC
.....TAGATG GTCTGA............CA--AC AATTGG............AAGAGG TGGGTG............AGTATT ACCAAT.....
.....ACATGG AGGTCT............AGGCCA A--ACAA............GCACTG GCMAAG............GTGGAG CCAAGGC............CAATGA
GCCTA ATGGAG............GAAAGG ATTTGG............AAAGAG AGTGGG............GCCTGG AAGTAT............TGAGCA
..CTACAT GAGAAT...CACA--A ACTGGA............TGGCGT GGCTAA............ACGAGC
GCCTACATGGAGG6TCTGAGAAAGGCCACAGCACAATGTTGGGACTGGAGCAAGAGAGGAGTGGTGGAGGCCTGGCTAAGTATGGCAATCTGAGCA
```
Variant detection by next-generation sequencing

30X

read mapping

GCC. GAGAAA GCCACA. TGGTCA TGGAGC. GCGTGG GCCTGG. CAATGA
GGTCTG AAAGGC. AAAATT CACTGG. AGTGGC TGGAGG. TTGACC
TAGATG GTCTGA. CAAC AATTGG. AAGAGG TGGCCT. AGTTCT ACCAAT
ACATGG AGGTCT. AGGCCA A--ACAA. GCACTG GCAAAG. GTGGAG CCTGGC. CAATGA
GCCTA ATGGAG. GAAAGG ATTGGG. AAAGAG AGTGGC. GCCTGG AAGTAT. TGAGCA
CTACAT GAAAAA.CACA--A. ACTGGG. TGGCCT. GGCTAA. ATGGAGC
GCC. GAGAAA GCCACA. TGGGCA TGGAGC. GCGTGG GCCTGG. CAATGA
GGTCTG AAAGGC. AAAATT CACTGG. AGTGGC TGGAGG. TTGACC
TACATG GCCTGA. CAAC AATTGG. AAGAGG TGGCCT. AGTTACC ACCAAT
ACATGG AGGTCT. AGGCCA A--ACAA. GCACTG GCAAAG. GTGGAG CCTGGC. CAATGA
GCCTA ATGGAG. GAAAGG ATTGGG. AAAGAG AGTGGC. GCCTGG AAGTAT. TGAGCA
CTACAT GAAAAA.CACA--A. ACTGGG. TGGCCT. GGCTAA. ATGGAGC
GCC. GAGAAA GCCACA. TGGGCA TGGAGC. GCGTGG GCCTGG. CAATGA
GGTCTG AAAGGC. AAAATT CACTGG. AGTGGC TGGAGG. TTGACC
TACATG GCCTGA. CAAC AATTGG. AAGAGG TGGCCT. AGTTACC ACCAAT
ACATGG AGGTCT. AGGCCA A--ACAA. GCACTG GCAAAG. GTGGAG CCTGGC. CAATGA
GCCTA ATGGAG. GAAAGG ATTGGG. AAAGAG AGTGGC. GCCTGG AAGTAT. TGAGCA
CTACAT GAAAAA.CACA--A. ACTGGG. TGGCCT. GGCTAA. ATGGAGC
GCC. GAGAAA GCCACA. TGGGCA TGGAGC. GCGTGG GCCTGG. CAATGA
GGTCTG AAAGGC. AAAATT CACTGG. AGTGGC TGGAGG. TTGACC
TACATG GCCTGA. CAAC AATTGG. AAGAGG TGGCCT. AGTTACC ACCAAT
ACATGG AGGTCT. AGGCCA A--ACAA. GCACTG GCAAAG. GTGGAG CCTGGC. CAATGA
GCCTA ATGGAG. GAAAGG ATTGGG. AAAGAG AGTGGC. GCCTGG AAGTAT. TGAGCA
CTACAT GAAAAA.CACA--A. ACTGGG. TGGCCT. GGCTAA. ATGGAGC
GCC. GAGAAA GCCACA. TGGGCA TGGAGC. GCGTGG GCCTGG. CAATGA
GGTCTG AAAGGC. AAAATT CACTGG. AGTGGC TGGAGG. TTGACC
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
Next-Generation Sequencing (NGS)

Reduction in DNA sequencing costs

Cost per Genome

National Human Genome Research Institute
genome.gov/sequencingcosts
Informatics Challenges

- **data storage & I/O**
  - ~7TB raw data per sequencer per year

- **mapping algorithm efficiency**
  - 1\textsuperscript{st} generation: hash-based
  - 2\textsuperscript{nd} 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)
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
From genome to epigenome
From genome to epigenome

DNA methylation

histone modification
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
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
Measuring DNA methylation by bisulphite sequencing

- Genomic DNA: C C C C C C C
- Converted genomic DNA: C C T C C C T
- Sequencing read: C C T C C T

**Bisulphite conversion**

**Sequencing**
Measuring DNA methylation by bisulphite sequencing

genomic DNA

bisulphite sequencing reads
Measuring DNA methylation by bisulphite sequencing
Measuring DNA methylation by bisulphite sequencing

genomic DNA

bisulphite sequencing reads
Measuring DNA methylation by bisulphite sequencing

Genomic DNA with methylated cytosines (Me) is treated with bisulphite, converting unmethylated cytosines to uracils (U) and leaving methylated cytosines as thymines (T). The resulting bisulphite sequencing reads then reflect the methylation status of the original DNA sequence.
Measuring DNA methylation by bisulphite sequencing

Genomic DNA

- Me

Bisulphite sequencing reads

- Me
- Me
- Me
Measuring DNA methylation by bisulphite sequencing
Measuring DNA methylation by bisulphite sequencing
Measuring DNA methylation by bisulphite sequencing
Measuring DNA methylation by bisulphite sequencing

genomic DNA

bisulphite sequencing reads
Measuring DNA methylation by bisulphite sequencing

**genomic DNA**

```
Me  Me  Me  Me
```

**bisulphite sequencing reads**

```
Me  Me  Me  Me

Me  Me  Me  Me

Me  Me  Me  Me
```

% methylation

<table>
<thead>
<tr>
<th>% methylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>20</td>
</tr>
</tbody>
</table>

```
20  40  60  80  100
```
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
### Sequencing the rat methylome

<table>
<thead>
<tr>
<th>sample</th>
<th>reads [million]</th>
<th>mapped reads [million]</th>
<th>depth of coverage [%]</th>
<th>depth of coverage after filtering [x-fold]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN2</td>
<td>258</td>
<td>230</td>
<td>89</td>
<td>7</td>
</tr>
<tr>
<td>BN4</td>
<td>413</td>
<td>361</td>
<td>87</td>
<td>13</td>
</tr>
<tr>
<td>BN5</td>
<td>454</td>
<td>404</td>
<td>89</td>
<td>11</td>
</tr>
<tr>
<td>BN7</td>
<td>370</td>
<td>340</td>
<td>92</td>
<td>12</td>
</tr>
<tr>
<td>BN total</td>
<td>1,495</td>
<td>1,336</td>
<td>89</td>
<td>42</td>
</tr>
<tr>
<td>SHR6</td>
<td>497</td>
<td>448</td>
<td>91</td>
<td>15</td>
</tr>
<tr>
<td>SHR7</td>
<td>452</td>
<td>398</td>
<td>89</td>
<td>12</td>
</tr>
<tr>
<td>SHR8</td>
<td>421</td>
<td>369</td>
<td>90</td>
<td>11</td>
</tr>
<tr>
<td>SHR9</td>
<td>391</td>
<td>356</td>
<td>91</td>
<td>12</td>
</tr>
<tr>
<td>SHR total</td>
<td>1,762</td>
<td>1,571</td>
<td>89</td>
<td>50</td>
</tr>
</tbody>
</table>

* filtering by read clonality, uniqueness, mapping quality and strand specificity
Inter- vs intra-strain differences in CpG methylation

Variability in CpG methylation greater between than within inbred strains

Distance measure: Euclidean
Clustering method: Ward
Distribution of methylation levels and inter-strain differences in methylation

The rat genome is globally highly methylated

The majority of methylation differences are small

Analysis of 10.6 million CpG dinucleotides identified 77,088 differentially methylated CpGs (<1%)
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
Methylation and gene expression

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

Promoters of differentially expressed genes are enriched in differentially methylated CpGs

Methylation and gene expression differences are negatively correlated
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
## 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)
- Korinna Northwood (Research Assistant)

**Imperial College High Performance Computing**
- Simon Burbidge (HPC Co-ordination Manager)