Deciphering human non-coding DNA using machine learning approaches

Guillaume Bourque
Department of Human Genetics, McGill University and McGill University and Genome Quebec Innovation Center

MLPM Summer School
Paris, September 17th 2014
Outline

• Applications of next-generation sequencing
• Functional genomics
• Example of machine learning approaches in functional genomics
• Role of transposable elements in gene regulation
Outline

• Applications of next-generation sequencing
• Functional genomics
• Example of machine learning approaches in functional genomics
• Role of transposable elements in gene regulation
DNA bases sequenced at the Innovation Center

300 Trillions!

~3000 human genomes at 30X
Sequencing human genomes

2001

The Human Genome

~ 3 Billion $

2011

1000 Genomes Project

~ 10 000 $

2015 (?)

Your Genome

< 1000 $
Big Data

2014 → 2 X 10 TBytes → 1 TByte

Intensity files

2014 

300 TBytes → 15 TBytes

Reads + qualities

30 TB of raw data / month 
360 TB of raw data / year
Large NGS project

Cancer project with whole genome data:

500 tumors

- 125 TB raw
- 500 X 3 lanes = 500 X 250GB

500 matched-normal

- 125 TB raw
- 500 X 3 lanes = 500 X 250GB
Applications (I)

- *De novo* sequencing
  - From the human genome... To all model organisms... To all relevant organisms (e.g. extreme genomes)... To “all” organisms?
Applications (II)

• Genome re-sequencing
  – Map genomic structural variations across individuals (to understand genetic disorders and also susceptibility factors)
  – Cancer genome sequencing
  – Agricultural crops

The Cancer Genome Atlas

1000 Genomes Project
Exome sequencing as a tool for Mendelian disease gene discovery

Michael J. Bamshad*†, Sarah B. Ng‡, Abigail W. Bigham*§, Holly K. Tabor*‖, Mary J. Emond*, Deborah A. Nickerson† and Jay Shendure‡

Abstract | Exome sequencing — the targeted sequencing of the subset of the human genome that is protein coding — is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Over the past 2 years, experimental and analytical approaches relating to exome sequencing have established a rich framework for discovering the genes underlying unsolved Mendelian disorders. Additionally, exome sequencing is being adapted to explore the extent to which rare alleles explain the heritability of complex diseases and health-related traits. These advances also set the stage for applying exome and whole-genome sequencing to facilitate clinical diagnosis and personalized disease-risk profiling.
Exome sequencing for Mendelian disease

“... about one-half to one-third (~3,000) of all known or suspected Mendelian disorders (for example, cystic fibrosis and sickle cell anaemia) have been discovered. However, there is a substantial gap in our knowledge about the genes that cause many rare Mendelian phenotypes.”

“Accordingly, we can realistically look towards a future in which the genetic basis of all Mendelian traits is known, ...”
Advances in understanding cancer genomes through second-generation sequencing

Matthew Meyerson, Stacey Gabriel and Gad Getz

Abstract | Cancers are caused by the accumulation of genomic alterations. Therefore, analyses of cancer genome sequences and structures provide insights for understanding cancer biology, diagnosis and therapy. The application of second-generation DNA sequencing technologies (also known as next-generation sequencing) — through whole-genome, whole-exome and whole-transcriptome approaches — is allowing substantial advances in cancer genomics. These methods are facilitating an increase in the efficiency and resolution of detection of each of the principal types of somatic cancer genome alterations, including nucleotide substitutions, small insertions and deletions, copy number alterations, chromosomal rearrangements and microbial infections. This Review focuses on the methodological considerations for characterizing somatic genome alterations in cancer and the future prospects for these approaches.
Cancer genome sequencing

Can obtain a full catalogue of mutations
Applications (III)

• Quantitative biology of complex systems
  – New high-throughput technologies in functional genomics: ChIP-Seq, RNA-Seq, ChIA-PET, RIP-Seq, ...
  – From single-gene measurements, to thousands of probes on arrays, to profiles covering all 3B bases of the genome
Transcription Regulation

Lenhard, Nat Rev Genet, 2012
Functional genomics

Locate regulatory regions in the human genome in different cell types, describe their functions, and identify how they differ between different groups (i.e. “disease” vs “healthy”).
Somatic mutations in 100 kidney tumors

- 1000 mutations (Total 575693)
- 1000 coding mutations (Total 6172)
Chromatin immunoprecipitation – Sequencing (ChIP-Seq)
RNA-Seq: digital expression and much more

Some of the ENCODE ChIP-Seq
International Human Epigenome Consortium

Welcome to the IHEC Data Portal

You may select IHEC datasets in these charts to view them in the Data Grid. Alternatively you can download or display them in a Genome Browser.

By Consortium

5011 datasets

By Tissue

5011 datasets

By Assay Category

5011 datasets

http://epigenomesportal.ca/ihec
IHEC Data Grid

Data Grid

Track Hubs

- McGill EMC (CEPH/PG) 575
- CEMT (CEPH/PG) 62
- Blueprint 259
- ENCODE 1595
- NIH Roadmap 2550
- Multiple Institutions

Click here for instructions.

Order: by Consortium

Visualize in Genome Browser  Get track link  Download tracks  Reset
Outline

• Applications of next-generation sequencing
• Functional genomics
• Example of machine learning approaches in functional genomics (work by Toby Hocking)
• Role of transposable elements in gene regulation
Motivation: visual differences between ChIP-seq profiles
Challenges with peak calling

• Different types of profiles (narrow peaks or broad signal)
• Many peak caller algorithms...
• Lots of parameters...

How to choose the right peak caller and optimal parameters?
Previous work in computer vision: look and add labels to...

Photos

Cell images

Copy number profiles

Labels: names

phenotypes

alterations

CVPR 2013

246 papers

CellProfiler

873 citations

SegAnnDB

H, et. al. 2014.

Sources: http://en.wikipedia.org/wiki/Face_detection
Annotated regions for possible peak starts and ends
Implicit specification of negative regions

No peaks in monocytes
No other peaks in T cells

Peak start Peak end T cell peak
Visual peak annotation
A peak detection function or peak caller $c: \mathbb{R}^d \rightarrow \{0,1\}^d$ takes a coverage profile $x \in \mathbb{R}^d$ as input, and returns a binary peak call prediction $y = c(x) \in \{0,1\}^d$ (0 is background noise, 1 is a peak).

The goal is to learn how to call peaks $c(x_i)$ which agree with the annotated regions $R_i^+, R_i^-, \bar{R}_i$ for some test samples $i$. To quantify the error of the peak calls with respect to the annotation data, we define the annotation error as the sum of false positive (FP) and false negative (FN) regions:

$$E(y, R_i, \bar{R}_i, R_i^+, R_i^-) = \text{FP}(y, R_i, \bar{R}_i, R_i^-) + \text{FN}(y, R_i, \bar{R}_i, R_i^+).$$

(1)

The supervised machine learning problem can be formalized as the following optimization problem. Find the peak caller $c$ with minimal annotation error on a set of test samples:

$$\min_{c} \sum_{i \in \text{test}} E[c(x_i), R_i, \bar{R}_i, R_i^+, R_i^-].$$

(2)
Choosing the threshold

\[ \hat{\lambda} = \arg \min_{\lambda} \sum_{i \in \{1, \ldots, n\}} E\left[ c_{\lambda}(x_i, R_i, \overline{R_i}, R_i^+, R_i^-) \right]. \]
Creating a benchmark dataset

<table>
<thead>
<tr>
<th>mark</th>
<th>H3K36me3</th>
<th>H3K4me3</th>
<th>H3K36me3</th>
<th>H3K36me3</th>
<th>H3K4me3</th>
<th>H3K4me3</th>
<th>H3K4me3</th>
<th>H3K4me3</th>
</tr>
</thead>
<tbody>
<tr>
<td>annotator</td>
<td>TDH</td>
<td>TDH</td>
<td>AM</td>
<td>TDH</td>
<td>PGP</td>
<td>TDH</td>
<td>immune</td>
<td>immune</td>
</tr>
<tr>
<td>sample set</td>
<td>other</td>
<td>other</td>
<td>immune</td>
<td>immune</td>
<td>immune</td>
<td>immune</td>
<td>immune</td>
<td>immune</td>
</tr>
<tr>
<td>windows</td>
<td>4</td>
<td>29</td>
<td>23</td>
<td>4</td>
<td>30</td>
<td>27</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>noPeaks</td>
<td>72</td>
<td>536</td>
<td>752</td>
<td>230</td>
<td>1653</td>
<td>1656</td>
<td>702</td>
<td></td>
</tr>
<tr>
<td>peakStart</td>
<td>68</td>
<td>305</td>
<td>294</td>
<td>200</td>
<td>813</td>
<td>796</td>
<td>216</td>
<td></td>
</tr>
<tr>
<td>peakEnd</td>
<td>60</td>
<td>311</td>
<td>294</td>
<td>200</td>
<td>730</td>
<td>933</td>
<td>216</td>
<td></td>
</tr>
<tr>
<td>peaks</td>
<td>218</td>
<td>403</td>
<td>638</td>
<td>287</td>
<td>243</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tcell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>monocyte</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>bcell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>kidney</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kidneyCancer</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>skeletalMuscleCtrl</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>skeletalMuscleMD</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>leukemiaCD19CD10BCells</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7 genomic regions, 2 histone marks, 4 expert annotators, 8 different cell types

Total of 12,286 annotated regions
## Training and testing

<table>
<thead>
<tr>
<th>Same</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator, cell types</td>
<td>Chr1</td>
<td>Chr2</td>
</tr>
<tr>
<td>Annotator, regions</td>
<td>T cells</td>
<td>Kidney cells</td>
</tr>
<tr>
<td>Cell types, regions</td>
<td>Annotator 1</td>
<td>Annotator 2</td>
</tr>
</tbody>
</table>
Performance of different peak callers

![Graph showing the performance of different peak callers for H3K4me3 (peaks) and H3K36me3 (broad regions). The graph compares the true positive rate against the false positive rate for various tools such as macs, macs.broad, homer, sicer, hmcan, hmcan.broad, and rseg. The parameter column indicates selected and default settings.]
Example region
Performance is stable across annotation datasets
Test error comparable for different annotators
Impact of the size of the training set
Conclusions

• The macs algorithm showed the minimum train error of 9–20% across the 4 annotated “peak” data sets.
• The hmcan.broad algorithm showed the minimum train error of 7–20% across the 3 annotated “broad” data sets.
• Error of the different algorithms and model ordering was independent of the annotator
• Each algorithm quickly achieves its model-specific minimum error, after only about 4 annotated windows in the train set
• For some models the trained model parameters were clearly better than the default model parameters
Future directions

• Calling peaks is an “easy” problem
• The real question that we would like to address is: Can we identify genetic variants that have an impact on the “phenotype”?
• This requires combining genetic and epigenomic data
• Previously, there wasn’t enough data the use of ML approaches to answer this question
• But now, we are generating matched genetic and epigenomic data on 100s of individuals...
• Can we learn from this data and build a predictive model?
Outline

• Applications of next-generation sequencing
• Functional genomics
• Example of machine learning approaches in functional genomics
• Role of transposable elements in gene regulation
30-50% of every mammalian genome is ‘Dark matter’

Repeats
Transposable Elements
Transposable Elements
Transposable Elements
Transposable Elements
Transposable Elements
Transposable Elements
Transposable Elements
Transposable Elements

time
Transposable Elements
Transposable Elements
Transposable Elements
Transposable Elements

(time)

---

McGill

GenomeQuébec
TEs have changed mammalian genomes
Transposable Elements (TEs) in the human genome

Cordeaux and Batzer,
*Nat Rev Genet*, 2009

Goodier and Kazazian,
*Cell* 2008

About 25% of the human genome consists of lineage-specific repeats. For the mouse genome it’s 30%.
Two models for regulatory site acquisition

McClintock 1950, Britten & Davidson 1971, Brosius 1991, ...
Genome-wide occupancy maps reveal a strong association to repeats for many TFs.

28% of CTCF sites in B2 repeats
Repeat Associated Binding Sites (RABS)

Human versus mouse ES cells

Collaboration with Huck-Hui Ng’s lab at GIS
OCT4 and NANOG RABS in human ES cells

<table>
<thead>
<tr>
<th>Repeat Class</th>
<th>Repeat Family</th>
<th>Repeats</th>
<th>Bound Repeats</th>
<th>Expected</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low_complex</td>
<td>GC_rich</td>
<td>13426</td>
<td>1215</td>
<td>624.6</td>
<td>1.9</td>
</tr>
<tr>
<td>ERV1</td>
<td>LTR7</td>
<td>2357</td>
<td>813</td>
<td>32.7</td>
<td>24.9</td>
</tr>
<tr>
<td>MER1_type</td>
<td>MER5A</td>
<td>35661</td>
<td>806</td>
<td>333.1</td>
<td>2.4</td>
</tr>
<tr>
<td>MER1_type</td>
<td>MER5A1</td>
<td>14897</td>
<td>529</td>
<td>136.9</td>
<td>3.9</td>
</tr>
<tr>
<td>L1</td>
<td>L1PA2</td>
<td>4789</td>
<td>361</td>
<td>241.6</td>
<td>1.5</td>
</tr>
<tr>
<td>ERVK</td>
<td>LTR5_Hs</td>
<td>610</td>
<td>340</td>
<td>27.8</td>
<td>12.2</td>
</tr>
<tr>
<td>MER1_type</td>
<td>MER1B</td>
<td>5131</td>
<td>323</td>
<td>80.0</td>
<td>4.0</td>
</tr>
<tr>
<td>MER1_type</td>
<td>MER5B</td>
<td>24784</td>
<td>321</td>
<td>213.8</td>
<td>1.5</td>
</tr>
<tr>
<td>ERV1</td>
<td>MER4A1</td>
<td>2190</td>
<td>319</td>
<td>34.0</td>
<td>9.4</td>
</tr>
<tr>
<td>ERV1</td>
<td>MER21C</td>
<td>4915</td>
<td>235</td>
<td>84.8</td>
<td>2.8</td>
</tr>
</tbody>
</table>

**NANOG**

<table>
<thead>
<tr>
<th>Repeat Class</th>
<th>Repeat Family</th>
<th>Repeats</th>
<th>Bound Repeats</th>
<th>Expected</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple_repeat</td>
<td>(TG)n</td>
<td>54553</td>
<td>600</td>
<td>141.6</td>
<td>4.2</td>
</tr>
<tr>
<td>Simple_repeat</td>
<td>(CA)n</td>
<td>54275</td>
<td>548</td>
<td>141.0</td>
<td>3.9</td>
</tr>
<tr>
<td>Low_complex</td>
<td>GC_rich</td>
<td>13426</td>
<td>315</td>
<td>159.4</td>
<td>2.0</td>
</tr>
<tr>
<td>ERV1</td>
<td>LTR9B</td>
<td>757</td>
<td>255</td>
<td>3.1</td>
<td>81.7</td>
</tr>
<tr>
<td>MER1_type</td>
<td>MER5A</td>
<td>35661</td>
<td>249</td>
<td>105.6</td>
<td>2.4</td>
</tr>
<tr>
<td>ERV1</td>
<td>LTR7</td>
<td>2357</td>
<td>208</td>
<td>10.6</td>
<td>19.7</td>
</tr>
<tr>
<td>MER1_type</td>
<td>MER5A1</td>
<td>14897</td>
<td>204</td>
<td>47.4</td>
<td>4.3</td>
</tr>
<tr>
<td>Low_complex</td>
<td>GA-rich</td>
<td>23936</td>
<td>206</td>
<td>80.9</td>
<td>2.5</td>
</tr>
<tr>
<td>ERV1</td>
<td>LTR9</td>
<td>1962</td>
<td>176</td>
<td>14.9</td>
<td>11.8</td>
</tr>
<tr>
<td>Low_complex</td>
<td>CT-rich</td>
<td>23713</td>
<td>181</td>
<td>77.2</td>
<td>2.3</td>
</tr>
</tbody>
</table>

**OCT4**
RABS in human and mouse ES cells

Different families of repeats have contributed a significant fraction of binding sites in both species.

Kunarso et al. Nat Genet, 2010
Human versus mouse ES cells

Human

Mouse

OCT4
NANOG
CTCF

Oct4
Nanog
Ctcf

Occupancy profile
Expression profile

OCT4 RNAi

Oct4 RNAi
OCT4 sites are associated with genes that are down regulated following RNAi treatment.
OCT4 RABS are enriched in proximity of down regulated genes
Human versus mouse ES cells

Human

Mouse

OCT4
NANOG

CTCF

Oct4
Nanog

Ctcf

OCT4 RNAi

Oct4 RNAi

Occupancy profile

Expression profile
Binding sites around genes with conserved expression profiles

Conserved Targets (72)
- 11 (15%)
- 61 (85%)

8% RABS
13% RABS

Human-Specific Targets (160)

22% RABS
Human-specific target driven by RABS

Human locus

Mouse locus
Transposable Elements have Rewired the Core Regulatory Network of Human ES Cells
Transposable Elements have Rewired the Core Regulatory Network of Human ES Cells

Britten & Davidson 1971
Conclusions (Are not)

Gene 'Jumps' Serve a Purpose, Study Shows
by Brian Thomas, M.S. *

In the tiny world of the cell, segments of DNA called transposons copy and reinsert themselves into the DNA. They eventually produce large repetitive sequences that have for many years been considered useless "junk" or remnants of ancient viral infections. But a new study has uncovered an important function for transposons.

Researchers at the Genome Institute of Singapore and other institutions suspected that transposons played an important role in the embryonic development of mammals. They decided to explore how the transposons interact with other key genetic pieces called "transcription factors" during the development of mice and humans.
“However, a biblical perspective predicts high functionality throughout genomes, with traces of degradation having accumulated since the curse that God placed on the earth, ... the high level of functionality of transposons is more consistent with creation.”

Researchers at the Genome Institute of Singapore and other institutions suspected that transposons played an important role in the embryonic development of mammals. They decided to explore how the transposons interact with other key genetic pieces called "transcription factors" during the development of mice and humans.
Conclusions

• Up to 25% of new regulatory sites have been contributed by transposable elements in both human and mouse.
• Only 15% of the genes with conserved expression profiles have a conserved binding site.
• A number of human-specific targets have been wired into the core regulatory network of ES cells by repeats.
Repeats and gene regulation

Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals

Vincent J Lynch, Robert D Leclerc, Gemma May & Günter P Wagner

Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages

Dominic Schmidt,1,2,6 Petra C. Schwaiger,3,6 Michael D. Wilson,1,2 Benoit Ballester,3 Ângela Gonçalves,3 Claudia Kutter,1,2 Gordon D. Brown,1,2 Aileen Marshall,1,6 Paul Flice,3,4,* and Duncan T. Odom1,2,4,*

The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements

Pierre-Étienne Jacques1,2, Justin Jeyakani1, Guillaume Bourque3,4*

1 Computational and Systems Biology, Genome Institute of Singapore, Singapore, Singapore; 2 Département de Biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada; 3 Department of Human Genetics, McGill University, Montréal, Québec, Canada; 4 McGill University and Génomique Québec Innovation Center, Montréal, Québec, Canada
Acknowledgements

Lab
Galih Kunarso
Pierre-Etienne Jacques
LeeAnn Ramsay
Justin Jeyakani
Jean Monlong
Simon Girard
Toby Hocking
Eric Audemard

EDCC team
David Bujold
Bryan Caron
David Morais (UdeS)
Carol Gauthier (UdeS)
Alain Veilleux (UdeS)
ME Rousseau

Analysis team
Louis Letourneau
Mathieu Bourgey
Maxime Caron
Gary Lévesque
Robert Eveleigh
Francois Lefebvre
Johanna Sandoval
Pascale Marquis
Joel Fillon
Julien Tremblay

IT team
Terrance Mcquilkin
Marc-André Labonté
Genevieve Dancausse
Andras Frankel
Alexandru Guja

Development team
Nathalie Émond
Francois Cantin
Catherine Côté
Daniel Guertin
Louis Dumond Joseph
Francois Korbuly
Marc Michaud
Thuong Ngo
Francois Massé

Open positions for students and postdoc!

guil.bourque@mcgill.ca