A Data Analytics View of Genomics

Cheng Soon Ong

Data61, CSIRO
Australian National University
University of Melbourne

28th Australasian Joint Conference on Artificial Intelligence 2015
1 December 2015

www.ong-home.my
The bottleneck in genome sequencing is no longer data generation – the computational challenges around data analysis, display and integration are now rate limiting. New approaches and methods are required to meet these challenges.

Green, Guyer and National Human Genome Research Institute
Charting a course for genomic medicine from base pairs to bedside, Nature 2011.

Why machine learning?

- A lot of data
- Data is noisy
- Large number of features
- No precise biological theory
- Complex relationships

Let the data do the talking!
Genome wide association studies
   Find genetic variation corresponding to an attribute of interest.

The search for genes
   A very brief overview of molecular biology

Biological sequencing
   The big data revolution in life sciences
SNP

Single Nucleotide Polymorphisms or single nucleotide variations (SNVs) are mutations on a single nucleotide (A,C,T or G) in the genome.
For example: AAGCCTA to AAGCTTA.

Alleles

There are two alleles: e.g. C and T.

Major/Minor allele

The nucleotide that occurs commonly in the population is called the major allele (denoted by a capital $B$) and the nucleotide that occurs more rarely is called the minor allele (denoted by a small letter $b$).

Diploid

haploid $\implies$ one chromosome set
diploid $\implies$ two chromosome sets
hexaploid $\implies$ six chromosome sets
Genotype
The genotype is the specific combination of alleles.

Phenotype
The phenotype is the observable trait or characteristic of an individual, for example whether the individual is healthy or sick.

Case-control studies
A cohort of sick individuals (cases) and healthy individuals (controls) are genotyped and their corresponding binary phenotype are recorded.

We use the framework of hypothesis testing
Why Hypothesis Tests?

- Given a case control study, test whether a particular SNP is associated with the phenotype.
- Look through each SNP one by one, and test to see if there is a difference in the frequency of the alleles seen in cases versus controls.
- If difference is statistically significant

⇒ SNP is associated with the phenotype.
null hypothesis $\mathcal{H}_0$
- genotype is independent of the phenotype

alternative hypothesis $\mathcal{H}_1$
- SNP is associated with the disease state

hypothesis test can be stated as follows

$$\mathcal{H}_0 : \theta \in \Theta_0 \quad \text{and} \quad \mathcal{H}_1 : \theta \in \Theta_1.$$ 

Important design choices

- How to represent intuition as a probabilistic model?
- How to decide on a test statistic?
- What is the distribution of the random variable?
- What is the level of significance ($\alpha$)?

Sinsheimer, “Statistics 101” – A Primer for the Genetics of Complex Human Disease, 2011
Agresti, “Categorical Data Analysis”, 2002
Wasserman, “All of Statistics”, 2004
Hypothesis test

- Let $X$ be a random variable with range $X$.
- $R \subset X$ called the rejection region
- If $X \in R$ then we reject the null hypothesis, otherwise we do not reject the null hypothesis.

$$R = \{ x : T(x) > c \}$$

where $T$ is a test statistic and $c$ is a critical value.

- The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
Outcomes of hypothesis tests

<table>
<thead>
<tr>
<th>$H_0$ true</th>
<th>Accept $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>correct</td>
<td>type I error</td>
</tr>
<tr>
<td>$H_1$ true</td>
<td>type II error</td>
<td>correct</td>
</tr>
</tbody>
</table>

Significance level

The probability of a rejecting $H_0$ when it is true is called the **significance level**.

p-value vs significance

- Reject $H_0$ when p-value < significance level
- p-value is computed from observation
- significance level is chosen by expert
Allelic test of association

- Single locus, haploid genome
- 200 individuals: 100 cases, 100 controls
- \( B \) and \( b \) are equally common in the population

**Null hypothesis**
No association between the allele and the phenotype

<table>
<thead>
<tr>
<th></th>
<th>allele ( B )</th>
<th>allele ( b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>50 ( (E_B,1) )</td>
<td>50 ( (E_b,1) )</td>
</tr>
<tr>
<td>Control</td>
<td>50 ( (E_B,0) )</td>
<td>50 ( (E_b,0) )</td>
</tr>
</tbody>
</table>
Experimental Observation

<table>
<thead>
<tr>
<th></th>
<th>allele $B$</th>
<th>allele $b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>50 ($E_{B,1}$)</td>
<td>50 ($E_{b,1}$)</td>
</tr>
<tr>
<td>Control</td>
<td>50 ($E_{B,0}$)</td>
<td>50 ($E_{b,0}$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>allele $B$</th>
<th>allele $b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>23 ($O_{B,1}$)</td>
<td>77 ($O_{b,1}$)</td>
</tr>
<tr>
<td>Control</td>
<td>68 ($O_{B,0}$)</td>
<td>32 ($O_{b,0}$)</td>
</tr>
</tbody>
</table>

**Pearson $\chi^2$ test of independence**

$$X^2 = \sum_{i \in \{0,1\}} \sum_{v \in \{B,b\}} \frac{(O_{v,i} - E_{v,i})^2}{E_{v,i}}.$$
Chi squared distribution

\[ f(x; k) = \begin{cases} 
  x^{(k/2-1)} \exp\left(-\frac{x}{2}\right) \frac{2^{(k/2)} \Gamma\left(\frac{k}{2}\right)}{k!}, & x \geq 0 \\
  0 & \text{otherwise}
\end{cases} \]

http://en.wikipedia.org/wiki/Chi-squared_distribution
\begin{align*}
\chi^2 &= \frac{(23 - 50)^2}{50} + \frac{(77 - 50)^2}{50} + \frac{(68 - 50)^2}{50} + \frac{(32 - 50)^2}{50} \\
&= 42.12
\end{align*}

What is the probability of observing a value greater than 42.12 of a $\chi^2$ random variable given that the null hypothesis is true?

\[ P(\chi^2 > 42.12) < 10^{-10} \]
The p-value is not ...

... the probability that the null hypothesis is true.

... the probability that a finding is “merely a fluke”.

... the probability of falsely rejecting the null hypothesis.

... the probability that a replicating experiment would not yield the same conclusion.

... indicating the size or importance of the observed effect.

The significance level of the test is not determined by the p-value.
$M$ hypothesis tests

$H_{0m}$ versus $H_{1m}$, $m = 1, \ldots, M$

and let $p_1, \ldots, p_M$ denote the $M$ p-values for these tests.

**Bonferroni Method**

Reject null hypothesis $H_{0m}$ if

$$p_m < \frac{\alpha}{M}.$$ 

**Outcome**

The probability of falsely rejecting any null hypothesis is less than or equal to $\alpha$. 
False discovery proportion

Let $M_0$ be the number of null hypotheses that are true.

\[ M_1 = M - M_0 \]

<table>
<thead>
<tr>
<th>$\mathcal{H}_0$ accepted</th>
<th>$\mathcal{H}_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{H}_0$ True</td>
<td>$U$</td>
<td>$V$</td>
</tr>
<tr>
<td>$\mathcal{H}_0$ False</td>
<td>$T$</td>
<td>$S$</td>
</tr>
<tr>
<td>Total</td>
<td>$M-R$</td>
<td>$R$</td>
</tr>
</tbody>
</table>

Define the *false discovery proportion* (FDP)

\[ \text{FDP} = \begin{cases} 
  V/R & \text{if } R > 0 \\
  0 & \text{if } R = 0.
\end{cases} \]
False discovery rate

**M hypothesis tests**

We order the p-values in increasing order.

**Benjamini-Hochberg Method**

1. For a given $\alpha$, find the largest $k$ such that

$$p_k \leq k \frac{\alpha}{M}.$$ 

2. Then reject all $H_{0m}$ for $m = 1, \ldots, k$.

**Theorem**

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \frac{M_0}{M} \alpha \leq \alpha.$$ 

**Outcome**

For a given significance level $\alpha$, the Benjamini Hochberg method bounds the false discovery rate.
Multiple testing

Suppose 800 of 500,000 variants are significant at 0.05 level.

**p-value < 0.05**
Expect $0.05 \times 500000 = 25000$ false positives

**false discovery rate < 0.05**
Expect $0.05 \times 800 = 40$ false positives

**family wise error rate < 0.05**
The probability of at least 1 false positive < 0.05
The basics of hypothesis testing applied to GWAS

Some Genomics Nomenclature
GWAS, SNPs, Allele, Diploid, Genotype, Phenotype

Hypothesis Testing
- $H_0$ vs $H_1$
- Design test statistic and compute p-value
- Reject $H_0$ if p-value $< \alpha$.

Multiple Testing
- Bonferroni correction
- Benjamini Hochberg method

Genome Wide Interaction Search (GWIS)
Consider the association of all pairs of genotypes to phenotypes

Large search space
- 5000 individuals, 500,000 SNPs
- Need to tabulate 125 billion contingency tables

Classification based analysis
- Focus on SNPs in case control studies
- New statistical tests
- Consider specificity and sensitivity
- Gain over univariate ROC
- CPU ($\approx$ days) and GPU ($\approx$ hours)

Web service
Goudey et. al. BMC Genomics, 2013
What is a biomarker?

How to measure?
- Clinical observations
- Whole genome sequencing
- Probes (arrays) for large studies

Looking at shadows

What to measure?
- Assumption: genetic cause
- DNA, RNA, Protein
- SNP, INDEL, CNV, Methylation, ...

Where to measure?
- Non-invasive diagnostic test
- Does tissue show variation?
Outline

Genome wide association studies
  Find genetic variation corresponding to an attribute of interest.

The search for genes
  A very brief overview of molecular biology

Biological sequencing
  The big data revolution in life sciences
The world is round

https://www.nasa.gov/image-feature/nasa-captures-epic-earth-image
The world is round

Genomics has given us a new perspective that has demanded a complete recasting and expansion of the material on molecular genetics ...

The traditionally explanatory cartoons that we show on nearly every page of the book generally represent only the primitive first step toward an explanation.

Preface: Alberts et. al. 2002
Tree of life

Bacteria, Archea, Eukaryote

Diagram showing the differences between eukaryotic and prokaryotic cells.
Inside a cell

https://en.wikipedia.org/wiki/DNA
**DNA** Positive strand, written 5’ to 3’.
e.g. AATCGAAGTGA

**RNA** T ⇒ U
e.g. AAUCGAAGUUA

**Amino acid** 3 letters of RNA (codon) ⇒ amino acid,
20 letter alphabet.

Lewin, Genes
Classification of Sequences

Example: Recognition of splice sites

- Every 'AG' is a possible acceptor splice site
- Computer has to learn what splice sites look like
given some known genes/splice sites . . .
- Prediction on unknown DNA

```
ATCCCGGATTGGATG
AGGGTCCCCTTGAGAGG
CCGGGTATATATATAGG
TTAGGTTCCCTCCGCGC
```
Many algorithms depend on numerical representations.

Each example is a vector of values (features).

Use background knowledge to design good features.

```
<table>
<thead>
<tr>
<th></th>
<th>x_1</th>
<th>x_2</th>
<th>x_3</th>
<th>x_4</th>
<th>x_5</th>
<th>x_6</th>
<th>x_7</th>
<th>x_8</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC before</td>
<td>0.6</td>
<td>0.2</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>...</td>
</tr>
<tr>
<td>GC after</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.6</td>
<td>0.3</td>
<td>0.4</td>
<td>0.7</td>
<td>0.6</td>
<td>...</td>
</tr>
<tr>
<td>AGAGAAG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>TTTAG</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>...</td>
</tr>
</tbody>
</table>
```
Recognition of Splice Sites

Given: Potential acceptor splice sites

Goal: Rule that distinguishes true from false ones

e.g. exploit that exons have higher GC content

or

that certain motifs are located nearby
Given the DNA, predict resulting mRNA and protein

- Requires very accurate identification of
  - splice sites, translation & transcription starts & stops
  - sites of regulation (transcription, splicing, etc.)

- Develop methods to integrate single site predictions
  - usually HMMs
  - Novel learning methods for structured outputs
POL II binds to a rather vague region of $\approx [-20, +20]$ bp

Upstream of TSS: promoter containing transcription factor binding sites

Downstream of TSS: 5’ UTR, and further downstream coding regions and introns (different statistics)

3D structure of the promoter must allow the transcription factors to bind
Gene Finding II: Transcr. Termination

Polyadenylation signal (AATAAA or variants) 10-30 bp upstream

T-rich or GT-rich elements 20-40 bp downstream

Transcription end is several hundreds of bp after 3’ cleavage site, mechanism not yet understood
Finding Intron-Exon junctions

Window of \( \approx 150 \) nucleotides

\[
\begin{align*}
\text{CT...GTAGAG} & \quad \text{TGTA..GAAGCT AG GAGCGC..ACCGT} \\
& \quad \text{ACGCGT...GA} \\
\end{align*}
\]

known splice site

- **true sites**: fixed-length window around splice site
- **decoys sites**: generated by shifting the window

\[
\begin{align*}
\text{AAAATAAGTAACATTACATCTTTTAGAGAAAGACGTTTCAACCATTGGAG} \\
\text{AAGATTAACAAAACAAATTTTAGCATTACAGATATAATCTAATT} \\
\text{CAGTCACCACAGATATTTTACGTTTCAACACACATCCGTCTGTGCC} \\
\text{TTAATTCACCTCCACATACTCCAGATCATCAATTCTCCAACACACAC}
\end{align*}
\]

⇒ Very unbalanced problem (1:200)
⇒ Millions of points from EST databases
⇒ Large scale methods necessary
Predict a sequence of binary decisions

www.mgene.org/web
**Goal:** Find sites of alternative splicing, conditions and regulating genes

- Understand differences between alternative and constitutive splicing
- Predict yet unknown alternative splicing events
- Predict on newly sequenced organisms
- Experimentally verify predictions via RT-PCR.

Genes are regulated by proteins called transcription factors.

Environment, e.g., metabolism (internal), temperature (external)

https://en.wikipedia.org/wiki/Transcription_factor

Alon, An Introduction to Systems Biology, 2007

Lawrence et. al. Learning and Inference in Computational Systems Biology, 2010
Chromatin structure

- DNA packed tightly in nucleus
- DNA wrapped around histones to form nucleosomes
- Nucleosomes organised into chromatin fibres
- Transcription accessibility
- DNA repair

http://dx.doi.org/10.1103/PhysRevLett.114.178102

https://en.wikipedia.org/wiki/Nucleic_acid_structure
Methylation

https://theconversation.com/explainer-what-is-epigenetics-13877
DNA 
Positive strand, written 5’ to 3’.
e.g. AATCGAAGTTA

RNA 
T ⇒ U
e.g. AAUCGAAGUUA

Amino acid 
3 letters of RNA (codon) ⇒ amino acid,
20 letter alphabet.

Splicing 
pre-mRNA to mature mRNA

Transcription factor 
Regulate expression of gene,
through promoters and repressors

Epigenetics 
Methylation, Chromatin marks
Outline

Genome wide association studies
Find genetic variation corresponding to an attribute of interest.

The search for genes
A very brief overview of molecular biology

Biological sequencing
The big data revolution in life sciences
- Identifying biomarkers
- Bottleneck: data analysis
- Open area of research
Sequencing
History of sequencing

1960s: DNA - properties, proto sequencing
70s-90s: Manual sequencing - Sanger, Maxam-Gilbert
90s: Automated Sanger - flourescent, clones, colony picking
2003: Human genome - 25 cents per 1000 bases
00s: NGS, Clusters - 454-Roche, Solexa-Illumina, Ion Torrent

Illumina HiSeq X Ten: 6 billion 150 base sequences in 3 days

USD 1000 genome

Data volume
- HiSeq X Ten: 12 GB per hour
- 700MB per human genome
  ≈ 200GB reads

Work in progress
- Multiplexing - tag sequences
- Capture: Enrich a particular set
- Paired Ends: sequence from both ends
- Small amounts of DNA
- Longer reads

Small sequencers
- Single cell sequencing: PacBio
- Real time sequencing: Oxford Nanopore
6 billion 150 base sequences

SO WHAT?
Apply cheap sensor
DNA Sequencing

Analogy: Shotgun sequencing
Take many copies of a text, split at random points, reconstruct.

Alignment
  - Dynamic programming
  - Needleman-Wunsch and Smith-Waterman
    https://en.wikipedia.org/wiki/List_of_sequence_alignment_software

Assembly
  - reference genome vs de-novo
  - grouping: reads → contigs → scaffold
  - Bridges of Königsberg → de Bruijn graphs
Cohort

Single Nucleotide Variation
  - Recall two copies of chromosomes at every location: AA, AB, BB
  - Noise free, high coverage ⇒ frequency = probability
  - Probabilistic methods for maximum a posteriori estimation
  - Correlations along the genome

https://en.wikipedia.org/wiki/SNV_calling_from_NGS_data

Structural variation
  - copy number variation
  - insertions, deletions
  - inversion, translocation

http://www.ncbi.nlm.nih.gov/dbvar/content/overview/

Study cohort germline vs somatic mutations

http://www.bioplanet.com/gcat
Replication

Multiple samples to estimate noise

Technical
- Effect of measurement instrument
- Different days, researcher
- Usually same biological sample

Biological
- Effect of biological development
- Different individuals of same “species”

Reproducibility crisis?
- Psychology: https://osf.io/ezcuj/
- Cancer biology: underway
  http://elifesciences.org/collections/reproducibility-project-cancer-biology
**confounding**
Common variable affecting two variables of interest.

- salt water
- red skin
- sun

**batch effect**
There is a hidden confounding variable for the effect, e.g. time

- Randomisation: randomly allocate samples to cases/controls
- Stratification: age, gender, group, geography

Lambert, Black: Learning from our GWAS mistakes, 2011
Bisulfite Seq

Methylation - epigenetics
- Identify methylated bases
- Regulates gene expression

Chemistry
- Bisulfite conversion converts unmethylated C to U
  - \( \text{AAC}^M \text{GGTC}^M \text{CCAGT} \)
  - \( \text{AAC}^M \text{GGTC}^M \text{UUAGT} \)

Algorithm
- Align converted sequence to reference
- Need to disambiguate unmethylated C from T
  - \( \text{AAC}^M \text{GGU}^M \text{CU}^M \text{UUAGU} \)
- E.g. latent variable models

http://www.epigenome.org/
**RNA Seq**

**Chemistry** Convert RNA to DNA

**Gene Expression**

- Recall: mRNA translated to proteins
- Which genes are expressed in what tissues at which levels?
- What are the regulators of a particular gene?
- How does treatment change expression (differential expression)?

https://www.encodeproject.org/

**Splicing**

Align expressed RNA to reference genome

[Diagram of splicing](http://biorxiv.org/content/early/2015/03/26/017095)
*-Seq

dsRNA-Seq
FRAG-Seq
SHAPE-Seq
PARTE-Seq
PARS-Seq
DMS-Seq

Nucleo-Seq
DNase-Seq
Sono-Seq
ChIA-PET-Seq
FAIRE-Seq
NOMe-Seq
ATAC-Seq

GRO-Seq
Quartz-Seq
CAGE-Seq
Nascent-Seq
Cel-Seq
3P-Seq

https://liorpachter.wordpress.com/seq/
Putting things together

Association Study

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3023</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID4454</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID7675</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID2283</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sequence Analysis

Variation
- SNP
- Structural
- Methylation
- Expression
- ...

Reference Genome Sequence
What is a biomarker?

How to measure?
Use adaptive experimental design to identify important time series.
Busetto et. al. Near-optimal experimental design for model selection in systems biology, 2013

What to measure?
Combine various sources of information for robust decision making.
Macintyre et. al. Associating disease-related genetic variants in intergenic regions to the genes they impact, 2014

Where to measure?
Use expert domain knowledge to construct dynamical models.
Brodersen et. al. Generative embedding for model-based classification of fMRI data, 2011
Open Source

Machine Learning Open Source Software
mloss.org  mldata.org
Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?
jmlr.org/papers/v15/delgado14a.html
Spoiler: No

Usability and Reproducibility
- (too much) focus on new algorithms
- Documentation, modularity issues
- Literate programming
  rmarkdown.rstudio.com  yihui.name/knitr  jupyter.org
- Scientific computing workflows
  galaxyproject.org  www.taverna.org.uk

Dream: App Bazaar for data science
Summary

A Data Analytics View of Genomics

**Genome wide association studies**
- Find genetic variation corresponding to an attribute of interest
- Hypothesis testing framework
- Batch effects and experimental design

**The search for genes**
- Glimpse of molecular biology
- Machine learning on sequences

**Biological sequencing**
- Bottleneck is analysis
- Sequence assignment and deconvolution

Please make your research open
Any questions?