Lecture 6: Various!

Matt McQueen | Associate Professor

Department of Integrative Physiology
Institute for Behavioral Genetics
Institute of Behavioral Science
University of Colorado Boulder

Department of Epidemiology (secondary)
Colorado School of Public Health
University of Colorado
Wrap up topics

- Interpreting GCTA output
- Next Generation Sequencing
- Copy Number Variants
- Meta-Analysis
Interpreting GCTA Output
## GCTA Output

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>V(1)</td>
<td>8.460930</td>
<td>5.852812</td>
</tr>
<tr>
<td>V(e)</td>
<td>9.985167</td>
<td>5.369622</td>
</tr>
<tr>
<td>Vp</td>
<td>18.446097</td>
<td>0.989077</td>
</tr>
<tr>
<td>V(1)/Vp</td>
<td>0.458684</td>
<td>0.304386</td>
</tr>
<tr>
<td>logL</td>
<td>-1791.054</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>923</td>
<td></td>
</tr>
</tbody>
</table>

**Genetic Variance**

**Residual (error)**

**Phenotypic Variance**

**“heritability”**
NOTE: This is the narrow sense heritability (additive effects)
Next Generation Sequencing
Sequencing

A.

100 - 400 bp [1990s]

500-900 bp [2000s]

400-700 bp [2000s]

100-200 bp [2010s]

2000 – 10^5 (?) bp [2010s-]

T G A C

Sanger Sequencing
Slab gel electrophoresis
1 read/4 lanes

Massively Parallel Sequencing
Consensus, long read
10^4 reads/picotiter plate

Massively Parallel Sequencing
Consensus, short read
10^4 reads/flow cell, slide
10^4-10^5 reads/ion chip

Massively Parallel Sequencing
Single molecule, long read
10^4 reads/ZMW chip
10^5 reads/GridION node
Sequencing coverage vs depth

High Coverage

High Depth
Next Generation Sequencing

• Moving fast
  – High depth, high coverage now possible
  – Prices falling
What are we expecting to find?

• Is this a looking under the lamp post issue?
  – More and more precise measurement
• Is there something new that we haven’t seen?
Next Generation Sequencing

• Will this provide more answers than GWAS?
Sequencing

• Objective
  – Find **rare/common variants** associated with disease

• Design
  – Cohort, case-control, family-based

• Molecular information
  – 3B base-pair

• Desired outcome
  – Find genetic variation underlying disease
Disease and DNA Variation

Penetrance: $P(D \mid G)$

2012 Nature Reviews | Genetics
GWAS: Common Disease / Common Variant

Higher disease prevalence associated with T allele
Sequencing: Rare Variant Hypothesis

Diseased

Non-Diseased
Inherited vs de novo mutation

Offspring

Offspring
Inherited vs de novo mutation

**Inherited**

- Dad
- Mom

**Offspring**

**de novo (private)**

- Dad
- Mom

**Offspring**
Tumor genomes

Gerlinger et al (2012) | NEJM
Paternal Age, Autism and Mutations

Kong et al., 2012
Disease characteristic vs prediction

• Mutations and genetic variation may be part of the disease process
• However, can we use our DNA to predict future disease?
  – Using “clones” (monozygotic twins) might help us answer the question...
<table>
<thead>
<tr>
<th>Disease/Condition</th>
<th>Sex</th>
<th>Number of MZ Twin Pairs</th>
<th>Number MZ Disease Concordant Pairs</th>
<th>Number MZ Disease Discordant Pairs</th>
<th>Disease Prevalence in Cohort (CR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder Cancer</td>
<td>Male &amp; Female</td>
<td>15668</td>
<td>5</td>
<td>189</td>
<td>0.6%</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>Female</td>
<td>8437</td>
<td>42</td>
<td>505</td>
<td>3.5%</td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>Male &amp; Female</td>
<td>15668</td>
<td>30</td>
<td>416</td>
<td>1.5%</td>
</tr>
<tr>
<td>Leukemia</td>
<td>Male &amp; Female</td>
<td>15668</td>
<td>2</td>
<td>103</td>
<td>0.3%</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>Male &amp; Female</td>
<td>15668</td>
<td>18</td>
<td>296</td>
<td>1.1%</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>Female</td>
<td>8437</td>
<td>3</td>
<td>125</td>
<td>0.8%</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>Male &amp; Female</td>
<td>15668</td>
<td>3</td>
<td>123</td>
<td>0.4%</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>Male</td>
<td>7231</td>
<td>40</td>
<td>299</td>
<td>2.6%</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td>Male &amp; Female</td>
<td>15668</td>
<td>11</td>
<td>223</td>
<td>0.8%</td>
</tr>
<tr>
<td>Thyroid Autoimmunity</td>
<td>Male &amp; Female</td>
<td>284</td>
<td>7</td>
<td>17</td>
<td>5.5%</td>
</tr>
<tr>
<td>Type 1 Diabetes</td>
<td>Male &amp; Female</td>
<td>4307</td>
<td>3</td>
<td>20</td>
<td>0.3%</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>Male &amp; Female</td>
<td>4307</td>
<td>29</td>
<td>113</td>
<td>2.0%</td>
</tr>
<tr>
<td>Alzheimer's Disease</td>
<td>Male &amp; Female</td>
<td>398</td>
<td>2</td>
<td>8</td>
<td>1.5%</td>
</tr>
<tr>
<td>Dementia</td>
<td>Male &amp; Female</td>
<td>398</td>
<td>3</td>
<td>16</td>
<td>2.8%</td>
</tr>
<tr>
<td>Parkinson Disease</td>
<td>Male &amp; Female</td>
<td>3477</td>
<td>7</td>
<td>60</td>
<td>1.1%</td>
</tr>
<tr>
<td>Chronic Fatigue</td>
<td>Female</td>
<td>1803</td>
<td>133</td>
<td>526</td>
<td>22.0%</td>
</tr>
<tr>
<td>Chronic Fatigue</td>
<td>Male</td>
<td>1426</td>
<td>48</td>
<td>266</td>
<td>12.7%</td>
</tr>
<tr>
<td>Gastro Esophageal Reflux Disorder (GERD)</td>
<td>Female</td>
<td>1260</td>
<td>63</td>
<td>284</td>
<td>16.3%</td>
</tr>
<tr>
<td>Gastro Esophageal Reflux Disorder (GERD)</td>
<td>Male</td>
<td>918</td>
<td>32</td>
<td>185</td>
<td>13.6%</td>
</tr>
<tr>
<td>Irritable Bowel Syndrome</td>
<td>Male &amp; Female</td>
<td>1252</td>
<td>14</td>
<td>97</td>
<td>5.0%</td>
</tr>
<tr>
<td>Coronary heart disease (CHD) Death</td>
<td>Female</td>
<td>2004</td>
<td>97</td>
<td>424</td>
<td>15.4%</td>
</tr>
<tr>
<td>Coronary heart disease (CHD) Death</td>
<td>Male</td>
<td>1640</td>
<td>153</td>
<td>451</td>
<td>23.1%</td>
</tr>
<tr>
<td>Stroke-related Death</td>
<td>Male &amp; Female</td>
<td>3852</td>
<td>35</td>
<td>316</td>
<td>5.0%</td>
</tr>
<tr>
<td>General Dystocia</td>
<td>Female</td>
<td>928</td>
<td>40</td>
<td>173</td>
<td>13.6%</td>
</tr>
<tr>
<td>Pelvic Organ Prolapse</td>
<td>Female</td>
<td>3376</td>
<td>34</td>
<td>157</td>
<td>3.3%</td>
</tr>
<tr>
<td>Stress Urinary Incontinence</td>
<td>Female</td>
<td>3376</td>
<td>13</td>
<td>87</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

MZ: Monozygotic. Disease prevalence in cohort (cohort risk, CR) was determined as described in the Materials and Methods.

Roberts et al., 2012
NGS Analytic Considerations

• Common variation
  – GWAS pipeline applies

• Rare variation
  – Might require new methods/thinking
Analysis of rare variants

• Effectively count data
  – Number of mutations/variants

• Accumulation of rare variants
  – Genome-wide
  – Genic region
  – Pathway/system
Analysis of rare variants

• Counts follow a Poisson distribution
  – “rate” of mutational load

• Weight variants
  – Prior biological information
  – Up-weight specific variants
Better prediction of functional effects for sequence variants

Maximilian Hecht\textsuperscript{1*}, Yana Bromberg\textsuperscript{2,3,4}, Burkhard Rost\textsuperscript{1,4}

From Varl-SIG 2014: Identification and annotation of genetic variants in and disease
Boston, MA, USA. 12 July 2014

Figure 1 SNAP2 performs best for the ALL data set. This figure shows performance estimates for the ALL data set. Our new method SNAP2 (dark blue, AUC = 0.905) outperforms its predecessor SNAP (light blue, AUC = 0.880), PolyPhen-2 (orange, AUC = 0.853) and SIFT (green, AUC = 0.838) over the entire spectrum of the Receiver Operating Characteristic (ROC) curve. Curves are significantly different from each other at a significance level of $P < 10^{-4}$ as measured by the DeLong method [59]. All SNAP2 results were computed on the test sets not used in training after a rigorous split into training, cross-training and testing. Results for PolyPhen-2 and our original SNAP included some of those proteins in their training, suggesting over-estimated performance.
Watch this space

- Methods are changing fast
Copy Number Variation
CNV

A B C D

Reference
CNV

Reference

Deletion
CNV

Reference

Deletion

Duplication
CNV

Reference

Deletion

Duplication

Multi-allelic
CNV

Reference

Deletion

Duplication

Multi-allelic

Inversion
How do we measure CNVs?

- GWAS platforms
- RT PCR and dPCR methods
- Next Gen Sequencing
GWAS Platform

• PennCNV is a common tool designed to harness Illumina and Affy data
  – Reliable and well-documented
Analysis of copy number variations at 15 schizophrenia-associated loci

Elliott Rees, James T. R. Walters, Lyudmila Georgieva, Anthony R. Isles, Kimberly D. Chambert, Alexander L. Richards, Gerwyn Mahoney-Davies, Sophie E. Legge, Jennifer L. Moran, Steven A. McCarroll, Michael C. O’Donovan, Michael J. Owen and George Kirov
# CNV Analysis

## Table 1: Findings from our data-set for previously implicated copy number variation (CNV) loci in schizophrenia

<table>
<thead>
<tr>
<th>Locus</th>
<th>Position in Mb</th>
<th>Case group (n = 6882)</th>
<th>Control group (n = 6316)</th>
<th>OR (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CNVs, n</td>
<td>Frequency, %</td>
<td>CNVs, n</td>
<td>Frequency, %</td>
</tr>
<tr>
<td>1q21.1 del</td>
<td>chr1:146,57-147,39</td>
<td>12</td>
<td>0.17</td>
<td>1</td>
<td>0.016</td>
</tr>
<tr>
<td>1q21.1 dup</td>
<td>chr1:146,57-147,39</td>
<td>8</td>
<td>0.12</td>
<td>5</td>
<td>0.079</td>
</tr>
<tr>
<td>NRXN1 del</td>
<td>chr2:50,15-51,26</td>
<td>11</td>
<td>0.16</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>3q29 del</td>
<td>chr3:195,73-197,34</td>
<td>4</td>
<td>0.058</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>WBS dup</td>
<td>chr7:72,74-74,14</td>
<td>3</td>
<td>0.044</td>
<td>1</td>
<td>0.016</td>
</tr>
<tr>
<td>VIPR2 dup</td>
<td>chr7:158,82-158,94</td>
<td>1</td>
<td>0.015</td>
<td>6</td>
<td>0.095</td>
</tr>
<tr>
<td>15q11.2 del</td>
<td>chr15:22,80-23,09</td>
<td>44</td>
<td>0.64</td>
<td>26</td>
<td>0.41</td>
</tr>
<tr>
<td>AS/PWS dup</td>
<td>chr15:24,82-28,43</td>
<td>8</td>
<td>0.12</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>15q13.3 del</td>
<td>chr15:31,13-32,48</td>
<td>4</td>
<td>0.058</td>
<td>2</td>
<td>0.032</td>
</tr>
<tr>
<td>16p13.11 dup</td>
<td>chr16:15,51-16,30</td>
<td>24</td>
<td>0.35</td>
<td>12</td>
<td>0.19</td>
</tr>
<tr>
<td>16p11.2 distal del</td>
<td>chr16:28,82-29,05</td>
<td>0</td>
<td>0.00</td>
<td>2</td>
<td>0.032</td>
</tr>
<tr>
<td>16p11.2 dup</td>
<td>chr16:29,64-30,20</td>
<td>27</td>
<td>0.39</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>17p12 del</td>
<td>chr17:14,16-15,43</td>
<td>4</td>
<td>0.058</td>
<td>3</td>
<td>0.047</td>
</tr>
<tr>
<td>17q12 del</td>
<td>chr17:34,81-36,20</td>
<td>1</td>
<td>0.01</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>22q11.2 del</td>
<td>chr22:19,02-20,26</td>
<td>20</td>
<td>0.29</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Totals**

171          2.48       58          0.92

---

del, deletion; dup, duplications; NA, not applicable; WBS, Williams-Beuren syndrome; AS/PWS, Angelman/Prader-Willi syndrome.

a. Copy number variation positions are in UCSC Build 37. Significant results are in bold (using Fisher exact test, 1-tailed).
Meta-Analysis
Aggregating the evidence

• Often, we are interested in combining evidence across independent studies
• There are a variety of ways to do this
Differing approaches...

- Mega-Analysis
- Combining Significance
- Meta-Analysis
- Weighted Hypothesis Testing
Mega-Analysis

• Combine two or more samples
• Requires access to raw data
• Many consortia utilize this approach
Mega-Analysis

• **Strengths**
  – Unprecented statistical power

• **Weaknesses**
  – Combining across heterogeneous samples
  – Ignore variation between studies
Combining significance

• Rather than combine raw data, you combine test statistics and/or p-values
• Simplest approach
  – Fisher’s Method

\[ X_{2k}^2 \sim -2 \sum_{i=1}^{k} \ln(p_i) \]
Fisher’s Method

• Strengths
  – Simple approach
  – Does not require raw data

• Weaknesses
  – Assumptions
    • Independent tests
    • Uniform distribution of p-values
  – Lack of effect size (only p-values)
Meta-Analysis

• Combining effect size estimates across studies
  – Odds ratios, risk ratios, etc.

• Important distinction
  – Random vs Fixed Effects
Fixed vs Random Effects

• Fixed Effects Meta-Analysis
  – Ignores between-study variance

• Random Effects Meta-Analysis
  – Incorporates between-study variance
  – More conservative (wider confidence intervals)
Conducting a meta-analysis

• Requirements
  – Proper extensive literature search
  – Parameter estimate (i.e. odds ratio)
  – Standard error

• Various tools to conduct a meta-analysis
  – R packages
    • Metafor is a good option
    • Provides graphics
Examples

• See alzgene.org