Lessons Learned from Sequencing and Genomics

Richard Gibbs, Baylor College of Medicine
Human Genome Sequencing Center

COI: Co-investment with Life Technology; Co-Founder SeqWright
Lessons (and outline):

1. TECHNOLOGY: Mind-boggling
   (quality not what is used to be)
2. BACKGROUND VARIATION: Enormous
   (hard to find *important* variation)
3. GENETICS RULES: Sequencing does not solve everything
   (genetics and functional studies most important)
4. ANALYSIS: Building robust pipelines
   (It takes a village....)
4. UBIQUITY: Social trend will be for ubiquitous sequencing
   (cohorts replaced by medical records?)
5. THIS PROJECT: How it might look?
   (a strawman....)
So Far:

SCD a ‘monogenic disorder’ \textbf{but}:

1. Exceptions to the primary mutation
2. Modifiers of severity
3. Differential response to primary therapy
4. Different predisposition to other effectors e.g. pain, infection etc
5. Other

\textit{Each of these ‘secondary genetic’ influences represents the same challenge as more complex phenotypes!}
Raw sequence production…..<< $50/Gb!!

Wild guess – today 4,000 genomes

Mind boggling!!

5,000 2011 Genomes?
30,000 2012 Genomes?

1. TECHNOLOGY
EXOMES INSTEAD OF GENOMES??.

**Whole-Genome Sequencing (WGS)**
- **Cost**: Still costly, but decreasing rapidly
- **Technical**: No capture step, automatable
- **Variation**: Uncovers ALL genetic and genomic variation (SNVs + CNVs)
  - Discovery of functional coding and non-coding variation
  - ~3.5 Million variants
- **Disease**: Suitable for mendelian and complex trait gene identification, as well as sporadic phenotypes caused by de novo SNVs or CNVs

**Exome Sequencing**
- **2011, ‘The CORE’ With Supplements**: Aim >99% of exons
- **Variation**: Limited to coding and splice-site variants in annotated genes
  - ~20,000 variants
- **Disease**: Good for highly penetrant mendelian disease gene identification
Current methods have lower quality……
Sequencing for discovery:

Candidate Genes (10–100)

Candidate Genes (100–1,000)

Whole Exome Sequencing

Regional Sequencing

Whole Genome Sequencing

Cost

Known genes

Gene interactions

Coding mutation burden

Exons + flanks

Whole genome variation
What does our knowledge of population variation tell us about the challenges for generally solving the challenges of common diseases?

What is the nature of BACKGROUND rare variation?
Rare Variants in the Site-Frequency Spectrum

- \(~10,000\) ns vs reference
- 1 - 300 novel ns variants per person
- More *novel* functional variants than previously expected
- *Huge* impact on assessing functional significance

Micortex:
\(~11,000\) samples, \(~25\) kb

Rare variation more likely to be deleterious!!
3: GENETICS Vignette I: NIMH/NHGRI Autism Sequencing

- Collaboration BI and BCM (Sequencing)
- Multiple analysis centers
- Aimed for all exonic variation
- Statistical differences between mutation burden?

New ‘Autism Genes’?
Two years later.........a lot of data!!:

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th></th>
<th></th>
<th>Control</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Sum</td>
<td>Male</td>
<td>Female</td>
<td>Sum</td>
<td></td>
</tr>
<tr>
<td>BCM</td>
<td>440</td>
<td>65</td>
<td>505</td>
<td>240</td>
<td>251</td>
<td>491</td>
<td>996</td>
</tr>
<tr>
<td>BI</td>
<td>344</td>
<td>86</td>
<td>430</td>
<td>177</td>
<td>202</td>
<td>379</td>
<td>809</td>
</tr>
</tbody>
</table>

> 500,000 variant sites
Large QA/QC effort
Battery of tests
**No new loci (yet)**
Testing for *de novo* mutations..
3: GENETICS Vignette II: Epilepsy Candidate Gene Sequencing

• 237 ion channel genes
• Case vs control study’
• Few familys
• No definitive ‘hits’
• Lots of hypotheses!

3: GENETICS Vignette III: Autism Candidate Gene Sequencing

Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders

Christian P. Schaaf1,†, Aniko Sabo2,‡, Yasunari Sakai1,¶, Jacy Crosby5, Donna Muzny2, Alicia Hawes2, Lora Lewis2, Humeira Akbar2, Robin Varghese2, Eric Boerwinkle5, Richard A. Gibbs1,2,∗ and Huda Y. Zoghbi1,3,4,6,∗

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oligogenic event</td>
<td>18</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>No oligogenic event</td>
<td>321</td>
<td>370</td>
<td>691</td>
</tr>
<tr>
<td>Total</td>
<td>339</td>
<td>376</td>
<td>715</td>
</tr>
</tbody>
</table>
Contrast to: Solving Mendelian diseases:

- Dozens are being reported each month
  - > 50 underway at BCM
- 10% new genes
- 50% ‘oops’
- remainder in process
Mendelian Disease Score Card:

(Approximately; n~30):
10% ‘new’ genes
50% ‘retrospective ‘insight’
40% unsolved so far!

Nice Editorial: Les Biesecker

1. Unambiguous
2. Less expensive
3. Tackles de novo’s
4. Better use of physician time…
5. SCALABLE!!!
Conclusions from Genetic Discovery so far:

*New methods make simple genetic problems easy to solve!!*

AND

*Family studies most tractable....*
LESSON IV: Analysis Networks: Many participants, group efforts:

- Dozens of conference calls, multiple centers and individuals, redundancy as well as single dedicated individuals.

  e.g. the ‘1000 Phone Calls Project’

  Approximately equal resources are needed post-data accumulation, as needed for sequencing.
LESSON V: UBIQUITY

Clan Genomics: Interest in Family Genetic Health will drive diagnostics:
How might this project look:

Thought ‘project’: $5M:
- 5,000 genomes? / 20,000 exomes?

Programs:
1. Rare non-HbS cases?
2. Variable HU responders
3. Other ‘variant phenotype’ issues?
How might this project look:

Thought ‘project’: $5M:
- 5,000 genomes? / 20,000 exomes?

Programs:

1. Rare non-HbS cases? (25%)

2. Variable HU responders (25%?)

3. Other ‘variant phenotype’ issues? (50%)
DESIGN:

1. Rare non-HbS cases? (25%)
   - Sequence probands, parents, sibs etc
DESIGN: Variable therapeutic response?

Whole exomes of low and high segments of distribution?
DESIGN: Other ‘variant phenotype’ issues?

- Multiple designs
- Phenotyping critical
- Family studies optional
- Power calculations needed (>1,000 cases)
- GWAS data to be considered
WORK PLAN:

• Complete design
• Identify samples, data generators, analysts
• Revisit ELSI/Consent issues
• Power calculations needed (>1,000 cases)
• GWAS data to be considered
• BUT……..
• Devise central data management/sharing protocols
• Establish management structure, mission milestones
• Manage timelines, integrate and respond to other efforts
• etc
For Meeting:

- Do we need a ‘centralized effort’
- Can we establish community-led ‘buy in’ to a mission objective.
- Will this compromise distributed science approach
- Is it the time to think BIG, challenge the status quo?
Acknowledgements

James Lupski*
Jeff Reid
David Deiros
Donna Muzny
Lynne Nazareth
Christie Kovar
Aniko Sabo
Matthew Bainbridge
Claudia Gonzaga-Jauregui
David Wheeler

Micortex Team
Eric Boerwinkle
Charlie Sing
Aleks Coventry
Andy Clark
Lara Bull

Sequencing Discovery
Huda Zoghbi
Jeff Noebels
Melissa Ramocki
Rui Chen
Shalini Jhangiani

ENCODE/HapMap3
Fuli Yu

Other HGSC
Steve Scherer
Lynne Nazareth
Mike Metzker
Rui Chen

ARRA AUTISM
Joe Buxbaum
Bernie Devlin
Jerry Schellenberg
Jim Sutcliffe
Kathryn Roeder
Benjamin Neale
Eric Boerwinkle

African Genomes
Stephan Schuster
Vanessa Hayes
Desmond Tutu

454/James Watson – Watson Project

Lifel
Linh Hoang*
Kevin McKernan
Francisco de la Vega
Tim Hunkapiller
Michael Rodes
Gina Costa
Quynh Dong
Fiona Hyland
Heather Peckham
Yutao Fu

All HGSC Staff

Jeff Reid
Donna Muzny
Yuan-Qing Wu
Fuli Yu
Aniko Sabo
Lisa Trevino
Eric Boerwinkle
James Lupski
Claudia Gonzaga-Jauregui