Quality Challenges for NGS Implementation in the Clinical Laboratory

Liz Worthey Ph.D.
Assistant Professor
Pediatric Genomics
Department of Pediatrics
Human and Molecular Genetics Center
Medical College of Wisconsin
The following relationship(s) exist related to this presentation:

- CHW/MCW offers fee for service genetic counseling and whole genome sequencing.
- Presenter holds a specific patent related to the tertiary analysis of WGS data
There are many aspects related to quality when performing WGS MDx in the clinic.

In this talk I will focus on:

• A discussion of the types of sequencing and mapping quality issues we see when performing high throughput whole genome sequencing
• A more in depth review of quality issues that can impact interpretation
• Challenges and lessons learned from our experience at MCW/CHW
Strategies for identifying sequencing errors

Based on known error rates we expect ~300,000 sequencing errors per WGS

- In regions with low sequence coverage (<8x) it is hard to distinguish between sequencing errors and heterozygous variants
- Based on extracted coverage data we expect that ~75,000 of these errors are in low coverage regions
- Of these we expect that ~800 will be in coding regions

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Total Variants</th>
<th>Exonic Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=20%</td>
<td>~60,000</td>
<td>~400</td>
</tr>
<tr>
<td>&lt;=15%</td>
<td>~25,000</td>
<td>~100</td>
</tr>
<tr>
<td>&lt;=10%</td>
<td>~1,700</td>
<td>~10</td>
</tr>
</tbody>
</table>

Within repeat >2 alleles, lowest support

Low quality Low % supporting reads

~22,000 variants

We can filter to attempt to exclude these errors; for example by excluding variants where a low percentage of reads support the variant call
Problems with the quality of the mapping can also confound the analysis

Copy differentiating nucleotide

A /T alleles identified; heterozygous variant called
Strategies for identifying mapping errors

Some regions are littered with mapping errors; these are often easy to spot by eye, but for WGS we must automate analysis.

These criteria can be used to flag likely mapping errors:

- Within repeat
- Low complexity
- Is or has pseudogene
- Low % supporting reads <25%
- Flagged as possible assembly issues
- >2 alleles
- Many paralogs
- Neighbouring gap
- Overcovered
- Within segmental duplication

All of these annotations are available in our tools for selection/exclusion.

~45,000
But of course we can’t just exclude genes that have or are pseudogenes

Many well characterised disease associated genes have many pseudogenes

<table>
<thead>
<tr>
<th>PSEUDOGENES</th>
<th>GENES</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>28,120</td>
</tr>
<tr>
<td>1</td>
<td>882</td>
</tr>
<tr>
<td>2</td>
<td>264</td>
</tr>
<tr>
<td>3</td>
<td>115</td>
</tr>
<tr>
<td>4</td>
<td>71</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>12+</td>
<td>86</td>
</tr>
</tbody>
</table>

1,288 of genes with known pseudogenes are disease associated in OMIM

401 are disease associated in HGMD

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Gene</th>
<th>#Pseudo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thrombocytopenia 4</td>
<td>CYCS</td>
<td>37</td>
</tr>
<tr>
<td>Iron overload, autosomal dominant</td>
<td>FTH1</td>
<td>19</td>
</tr>
<tr>
<td>Mucopolysaccharidosis VII</td>
<td>GUSB</td>
<td>12</td>
</tr>
<tr>
<td>Leukodystrophy, hypomyelinating, 4</td>
<td>HSPD1</td>
<td>22</td>
</tr>
<tr>
<td>Cirrhosis, cryptogenic</td>
<td>KRT18</td>
<td>54</td>
</tr>
<tr>
<td>Oral-facial-digital syndrome 1</td>
<td>OFD1</td>
<td>18</td>
</tr>
<tr>
<td>Spermatogenic failure, Y-linked, 2</td>
<td>USP9Y</td>
<td>15</td>
</tr>
</tbody>
</table>
Presence of functional paralogues can also lead to missmapping errors.

In this sample we identified an apparent premature stop in the PABPC3 (Poly(A) binding protein, cytoplasmic 3) gene.

Many paralogs
Many pseudogenes
The quality of the sequencing data is not evenly distributed across the genome

• In WGS large (and many smaller) regions are missed:
  – Problems in the reference being used to map the sequence reads are transferred to the new analysis
  – Problems with regions that are hard to sequencing in the lab; repeats that form secondary structures that the polymerase cannot navigate
  – Presence of repetitive regions that lead to difficulties in accurately determining the correct map location for the read within the genome

• At an average 40x coverage:
  – ~75% of genes are well covered (at least 7x across at least 95% of length)
  – 100-150 genes are very poorly covered (only 5% covered to at least 7x)
  – ~100 genes have no coverage (at a threshold of at least 7x)

• It is important to know which regions have not been sequenced
  – When attempting to exclude regions of interest; for example a gene family where members are already associated with the phenotype
  – Lack of sufficient depth of coverage may result in uncertain zygosity calls
Potentially important regions are missing at high (all levels of?) coverage
Sequencing gaps in known disease genes

Two unrelated individuals share this particular gap; one has developmental delay the other is immunodeficient; it is also found in other samples.

We see gaps in the exonic proportion of many other disease associated genes that are shared across the genomes of individuals with diverse phenotypes (as well as being seen in our healthy control samples).
This sequence coverage information ends up in the medical report

“All exons of all genes were analyzed for depth of coverage. This analysis includes the genes in table 1. Those genes in table 1 with insufficient depth of coverage to be fully analyzed are listed in table 2. Because of insufficient depth of coverage, the genes in table 2 may harbor variants that were not detected by this test. Depth of coverage of introns is not documented in this table.”

These findings can direct strategies for additional testing
For example whether single gene re-sequencing should be performed
Clinical interpretation requires use of various types of annotation data

- Input variants from exome sequencing
- Protein coding or possible splice site altering
- Non-synonymous
- Novel (versus dbSNP129)
- Homozygous or hemizygous
- Likely to be damaging
- Altering highly conserved nucleotides
- Absent in other reference genomes

* GSTM1 transferase – many null mutations known

Image from C. Sun et al 1999
An ever increasing amount of annotation data is used in the interpretation of variants.

**Input**
- 4.5 million SNVs
- 0.5 million indels

**Variant Features**
- Frameshift
- Splice donor/acceptor
- Abrogation of start codon
- Readthrough
- Premature Stop
- PolyPhen damaging
- Insertion or deletion
- SIFT ortholog deleterious
- Low frequency in 1000 Genomes
- Low frequency in dbSNP
- Highly conserved (if appropriate) *
- Is or has pseudogene
- Within segmental duplication
- Within repeat or low complexity
- Within low complexity
- Within segmental duplication
- Within low complexity
- In low complexity
- In segmental duplication

**Annotation Data**
- SIFT deleterious
- PolyPhen damaging
- Low frequency in dbSNP
- Low frequency in 1000 Genomes
- Low frequency in Valcrie *
- Low frequency in house data *
- Low frequency in publication *
- Commonly Null: in house data *
- Commonly Null: in publication *
- Damaging variants w/o phenotype *
- Abrogation of start codon
- Flagged as possible assembly issues
- Expression fits Phenotype *
- GO term matches
- Many paralogs

**Output**
- Prioritised list of variants

**Variant Association**
- Variant assoc with similar phenotype
- Gene assoc with similar phenotype
- Animal model shares phenotype
- Gene assoc with similar phenotype
- GO term matches

**Variants with Phenotype**
- Variant assoc with similar phenotype
- Animal model shares phenotype
- Gene assoc with similar phenotype
- GO term matches

**Low Frequency Variants**
- Low frequency ESP Genomes
- Low frequency in 1000 Genomes
- Low frequency in dbSNP
- Low frequency in Valcrie *
A sophisticated tool is required to accurately manage this data

There are close to 100 annotations loaded or produced within the tool.

Since analysis is automated this requires careful and comprehensive QA/QC.

QA/QC of externally derived algorithms and the datasets they use is also required.

PolyPhen algorithm update QA/QC took 3 weeks and required working closely with the Harvard development team.

QA of an update takes 2 months; the first Illumina based release took 6 months and generated more than 300 pages of QA and SOP CLIA documentation.

Quality assurance, the process or set of processes used to measure and assure the quality of a product.

Quality control, the process of meeting products and services to consumer expectations.
A variety of quality issues exist with this annotation data

- Incorrectly reported primary data
- Incorrect mappings made between data
- Errors in IDs
- Typos/human entry errors
- Data being used inappropriately

Many of the datasets being used were not initially built for clinical use. A number of them started as PhD student or PostDoc projects. The data is now “grandfathered in”

I want to give a few examples of issues uncovered in our cases.
Existing databases contain lots of incorrect data

Often the mutation data has been manually entered into the source databases – typos are not uncommon.

The V339A substitution does not exhibit a shift in polarity and displays a decrease in Kyte-Doolittle hydrophobicity from 4.2 to 1.8. Approximately 0.69% of nonsense mutations in HGMD are Val-Ala. The mutation occurs 1863 amino acids from the end of the protein.
Some reported rare mutations are actually not so rare polymorphisms

- Many polymorphisms have been annotated as disease mutations in the literature and are now propagated into the mutation DBs
- If the literature is amended, info does not always make it to the DBs

<table>
<thead>
<tr>
<th>Red</th>
<th>Orange</th>
<th>Green</th>
<th>Olive</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLB1</td>
<td>ATP7B</td>
<td>HESX1</td>
<td>ALG6</td>
</tr>
<tr>
<td>ATP7B</td>
<td>ETFB</td>
<td>HGSNAT</td>
<td>DPYD</td>
</tr>
<tr>
<td>ETFB</td>
<td>NHLRC1</td>
<td>ACADM</td>
<td>HADHA</td>
</tr>
<tr>
<td>NHLRC1</td>
<td>IGHMBP2</td>
<td>MTHFR</td>
<td>GAA</td>
</tr>
<tr>
<td>IGHMBP2</td>
<td>SERPINA1</td>
<td>GALC</td>
<td>AHI1</td>
</tr>
<tr>
<td>SERPINA1</td>
<td>NPHS1</td>
<td>MEFV</td>
<td>AMPD1</td>
</tr>
<tr>
<td>NPHS1</td>
<td>LAMA2</td>
<td>ARSB</td>
<td>SBDS</td>
</tr>
<tr>
<td>LAMA2</td>
<td>ADA</td>
<td>CPT1A</td>
<td>CDH23</td>
</tr>
<tr>
<td>ADA</td>
<td>MTHFR</td>
<td>CYP21A2</td>
<td>ATP7B</td>
</tr>
<tr>
<td>MTHFR</td>
<td>GALC</td>
<td>ARSB</td>
<td>SBDS</td>
</tr>
<tr>
<td>GALC</td>
<td>MEFV</td>
<td>CPT1A</td>
<td>CDH23</td>
</tr>
<tr>
<td>MEFV</td>
<td>ARSB</td>
<td>CYP21A2</td>
<td>ATP7B</td>
</tr>
<tr>
<td>ARSB</td>
<td>CPT1A</td>
<td>LAMA2</td>
<td>DVD</td>
</tr>
<tr>
<td>CPT1A</td>
<td>GLB1</td>
<td>ETFB</td>
<td>NEFL</td>
</tr>
</tbody>
</table>

- We are curating potentially unreliable associations as we come across them either in our data or from the literature
- These are stored in our tool and used for querying (deprioritisation)
Allele frequency is extremely useful in variant filtering

But the allele frequency reported for a particular variant can vary widely amongst commonly used data sources

These variants are not randomly selected – these are 421 highly divergent variants (between 1000 genomes project and dbSNP) associated with disease in HGMD
Sometimes the quality of the data is fine but the data is being used inappropriately.

Filtering variants by conservation is not always appropriate.

<table>
<thead>
<tr>
<th>Variants</th>
<th>16,124</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genic (within genes)</td>
<td>16,012</td>
</tr>
<tr>
<td>Protein coding</td>
<td>15,272</td>
</tr>
<tr>
<td>Non-synonymous (alter protein sequence)</td>
<td>7,156</td>
</tr>
<tr>
<td>Novel (versus dbSNP129)</td>
<td>878</td>
</tr>
<tr>
<td>Homozygous or hemizygous</td>
<td>70</td>
</tr>
<tr>
<td>Predicted to be damaging</td>
<td>17</td>
</tr>
<tr>
<td>Altering highly conserved position</td>
<td>4</td>
</tr>
<tr>
<td>Not found in reference sequences</td>
<td>2*</td>
</tr>
<tr>
<td>Excluding GSTM1 *</td>
<td>1</td>
</tr>
</tbody>
</table>

This is from my paper.
Generic use of conservation scores leads to exclusion of excellent candidate variants

- Extremely poor conservation based on multispecies derived scores
  - Most variants would have a low (close to background) score
- But we know the gene is primate specific
  - If we look at primates only we see very high conservation

Apply filtering by conservation very conservatively
Match conservation score selection to conservation of the affected system or organ
Many polymorphisms that look like deleterious mutations have already been described

- Lots of time wasted following up these candidates
- To be more efficient we:
  - Have started manually curating these cases from the literature
  - Have extracted candidates from in house and publicly available data
    - e.g. premature stops that show up in a high percentage of available samples with diverse or no known phenotypes

<table>
<thead>
<tr>
<th>Red</th>
<th>Yellow</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUT2</td>
<td>GSTT1</td>
<td>AGT</td>
</tr>
<tr>
<td>OVCH2</td>
<td>HPS4</td>
<td>APOC4</td>
</tr>
<tr>
<td>POLE2</td>
<td>IL17RB</td>
<td>CDHI5</td>
</tr>
<tr>
<td>SERPINB11</td>
<td>KRTAP1-1</td>
<td>CLCA3</td>
</tr>
<tr>
<td>SMUG1</td>
<td>LCE5A</td>
<td>CYP2C19</td>
</tr>
<tr>
<td>SPTBN5</td>
<td>LIG4</td>
<td>DSCR8</td>
</tr>
<tr>
<td>TAAR9</td>
<td>LPL</td>
<td>EPHX1</td>
</tr>
<tr>
<td>TAP2</td>
<td>MAGEE2</td>
<td>TRPM1</td>
</tr>
<tr>
<td>TLR5</td>
<td>MS4A12</td>
<td>OAS2</td>
</tr>
</tbody>
</table>

The evidence and source are stored in our tools and can be queried.
But don’t use this list as a “kill list”

- Presence of one common null mutation with an apparently benign effect does not mean that all mutations in the same gene should be regarded as benign.

- If the common premature stop is at the end of the gene, the novel premature stop at the beginning may well be disease causing.

**TAP2 example**

- Common and benign premature stop at 687 (of 703 amino acids)

- Type I Bare lymphocyte syndrome premature stop at 210

- HLA class I deficiency premature stops at 220, 273, and 623
Take home message from this section

• A lot (all?) of the reference data was initially collected for research purposes and not for clinical use

• In the “old days” researchers often focused on a small set of data (e.g. one or a family of genes or sequences)
  – They really knew the data, what the issues and errors are, and the fixes
  – They knew what cannot be fixed and what should be used with caution

• Today we have vast quantities of data, and the user performs analysis very broadly across many objects (the whole genome) but also very deeply using all the data for each of the genomic features
  – No one can know all the data
  – Individual users cannot track all the issues and fixes
  – Requires sophisticated tool but this necessitates lengthy QA periods

• The use of the data needs to be carried out with knowledge of all of the issues with the data
Acknowledgements

- Families and Patients
- Referring Physicians
- MCW- HMGC
  - Howard Jacob
  - David Dimmock
  - David Bick
  - Mary Shimoyama
  - Jill Northup
  - Jenny Guerts
- CHW Genetics Center
  - Angela Pickart
  - Regan Veith
  - William Rhead
- AGEN-Seq
  - Mike Tschannen
  - Dennis Schauer Jr
  - Daniel Helbling
  - Brett Chirempes
  - Jayme Wittke
  - Jamie Wendt -Andrae
- Bioinformatics/Curation/Systems support
  - Brandon Wilk
  - Jeremy Harris
  - Wendy Demos
  - George Kowalski
  - Weihong Liu
  - Jeff DePons
  - Sharon Tsaih
  - Brad Taylor
  - Stacy Zacher
  - Marek Tutaj
  - Greg McQuestion
  - Kent Brodie
  - Bryce Schuler
  - Florence Yeo
  - Sasha Zhang
  - Stan Laulederkind
  - Victoria Petri
  - Jennifer Smith
  - Alex Stoddard
- CHW
  - Juliet Kersten
  - Paula North
  - Tara Schmit
  - Altheia Roquemore-Goins
  - Gail Bernadi
  - Michael Gutzeit
  - Steven Leuthner
  - Rodney Willoughby
  - Thomas May
  - Robert Kliegman
- Funding/Support:
  - MCW Children’s Research Institute
  - Jeffrey Modell Foundation
  - Private Donors
  - Roche (provided sequencing and capture for case 1)