Advancements in Genomic Analysis at Children’s Mercy Kansas City

Shane Corder
HPC Systems Engineer
Children’s Mercy Kansas City
Center for Pediatric Genomic Medicine
Children’s Mercy Research Institute
Advanced Research Computing Team
Survey
~8000 known genetic diseases

- affects 1 in 30 children in the US
- causes 1 in 6 children’s hospital admissions
- causes 1 in 5 deaths in the roughly 60,000 babies born in the Kansas City area
The hard facts

- We know the genetic cause of <5000 of these diseases
- Diagnosis often takes years
- Diagnosis often impacts treatment and always impacts families
- Imagine....
A critically ill newborn...
A sick 10 year old with muscle weakness...
A mother with no hope for an answer...
The Human Genome

6.4 billion letters in pairs
19,000 genes - coding for roughly 100,000 proteins
Decoding the Genome

- Human Genome Project generated first “draft” in 15 months
  - Generating the sequence draft cost $300 million
- Later released final sequence in 2003
  - Draft to final sequence cost an additional $150 million
- HGP was a 13 year project costing roughly $3 billion
- Today a HiSeq 4000 produces 16 human genomes in 3 days
  - Reagent costs of <$1,600 per genome
Data Deluge

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Transient Data</th>
<th>Permanent Data</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Analysis</td>
<td>1.62TB</td>
<td>104GB</td>
<td>1.62TB</td>
</tr>
<tr>
<td>Secondary Data</td>
<td>301GB</td>
<td>71GB</td>
<td>301GB</td>
</tr>
<tr>
<td>Fastq</td>
<td>1.2GB</td>
<td>1.2GB</td>
<td>1.2GB</td>
</tr>
<tr>
<td>BAM</td>
<td>71GB</td>
<td>825MB</td>
<td>71GB</td>
</tr>
<tr>
<td>VCF - variants only</td>
<td></td>
<td></td>
<td>177GB</td>
</tr>
<tr>
<td>Annotated Variant</td>
<td></td>
<td></td>
<td>per genome</td>
</tr>
</tbody>
</table>

The center is capable of generating 64 whole genomes every 6 days.
• **Compute**
  - Pinnacle Flex blades
  - 900 core

• **Storage**
  - DDN GS7K
  - WOS (In the works)
Genome Center Network

- CPGM resources on their own subnet with traffic between main hospital network and subnet passing through firewall
- Windows clients access the network filesystem through CIFS via MediaScalar
- backup/DR
- compute cluster
- batch queuing system
- database server
- web server
- cluster head nodes
- jobs
- advanced clustering technologies, inc.
- UNIVA
- CentOS
• Intel E5-2690 v3 @ 2.60GHz
  • w/ HT for 48 core
• 128GB of RAM
• 120GB Intel SSD
• 2x 1.6TB NVME drives
  • RAID0 @ 3.2TB for staging
• 10GbE
## Method

<table>
<thead>
<tr>
<th>Sample</th>
<th>Method</th>
<th>DNA Isolation, QC &amp; Shearing</th>
<th>Library Prep</th>
<th>Library QC</th>
<th>SBS</th>
<th>Alignment</th>
<th>Variant Calling</th>
<th>RUNES Variant Annotation</th>
<th>VIKING Provisional Diagnosis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published WGS$_{50}$</td>
<td>Multiple$^c$</td>
<td>2:30</td>
<td>3:15</td>
<td>1:30</td>
<td>25:30</td>
<td>14:40</td>
<td>2:30</td>
<td>0:05</td>
<td>50:00</td>
<td></td>
</tr>
<tr>
<td>SBS$_{18}$, GSNAP/GATK</td>
<td>5006-01</td>
<td>2:30</td>
<td>3:15</td>
<td>1:30</td>
<td>19:45</td>
<td>22:30</td>
<td>0:29</td>
<td>na</td>
<td>49:59</td>
<td></td>
</tr>
<tr>
<td>WGS$<em>{26}$, SBS$</em>{18}$ &amp; Dragen</td>
<td>UDT_173</td>
<td>2:30</td>
<td>3:02</td>
<td>1:30</td>
<td>17:58</td>
<td>0:15</td>
<td>0:15</td>
<td>0:34</td>
<td>0:04</td>
<td>26:08</td>
</tr>
<tr>
<td>WGS$<em>{26}$, SBS$</em>{18}$ &amp; Dragen</td>
<td>UDT_103</td>
<td>2:30</td>
<td>3:05</td>
<td>1:30</td>
<td>18:25</td>
<td>0:19</td>
<td>0:22</td>
<td>0:31</td>
<td>0:05</td>
<td>26:47</td>
</tr>
<tr>
<td>WGS$<em>{26}$, SBS$</em>{18}$ &amp; Dragen</td>
<td>NA12878</td>
<td>2:30</td>
<td>3:15</td>
<td>1:30</td>
<td>18:00</td>
<td>0:19</td>
<td>0:22</td>
<td>0:33</td>
<td>n.a.</td>
<td>26:28</td>
</tr>
<tr>
<td>WGS$<em>{26}$, SBS$</em>{18}$ &amp; Dragen</td>
<td>NA12878</td>
<td>2:30</td>
<td>3:15</td>
<td>1:30</td>
<td>18:36</td>
<td>0:10</td>
<td>0:11</td>
<td>0:35</td>
<td>n.a.</td>
<td>26:47</td>
</tr>
</tbody>
</table>

## Sample Yield (GB) Pipeline

<table>
<thead>
<tr>
<th>Sample</th>
<th>Yield (GB)</th>
<th>Pipeline</th>
<th>Reads Aligned</th>
<th>Alignments with mapQ &gt; 20</th>
<th>Variants Called</th>
<th>Analytic Sensitivity</th>
<th>Analytic Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA12878</td>
<td>133</td>
<td>DRAGEN</td>
<td>99.4%</td>
<td>95.48%</td>
<td>4,782,970</td>
<td>99.93%</td>
<td>99.87%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSNAP/GATK-1.6</td>
<td>98.5%</td>
<td>96.33%</td>
<td>5,343,988</td>
<td>99.54%</td>
<td>98.57%</td>
</tr>
<tr>
<td>NA12878</td>
<td>65$^a$</td>
<td>DRAGEN</td>
<td>97.7%</td>
<td>91.31%</td>
<td>4,633,357</td>
<td>99.42%</td>
<td>99.46%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSNAP/GATK-3.2</td>
<td>96.2%</td>
<td>92.86%</td>
<td>4,571,157</td>
<td>97.29%</td>
<td>95.35%</td>
</tr>
<tr>
<td>UDT_173$^b$</td>
<td>106</td>
<td>DRAGEN</td>
<td>99.5%</td>
<td>94.92%</td>
<td>4,742,150</td>
<td>96.13%</td>
<td>97.74%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSNAP/GATK-1.6</td>
<td>99.3%</td>
<td>96.88%</td>
<td>4,294,504</td>
<td>88.54%</td>
<td>98.06%</td>
</tr>
</tbody>
</table>
Stat-Seq - Rapid Medical Genome Sequencing

1. Identify candidate patient
2. Parental consent
3. DNA Sample

Sample Prep/Sequencing -> Alignment -> Variant Detection -> Characterization -> Interpretation

30 hours
15 hours
3 hours
1.5 hours
.5 hours

23
.50
.25
.5

26
= 50 hours
Other benefits:

- Sure it’s fast and gives great accuracy...
- We’ll lessen our development load - mainly on our variant detection pipeline
- The Dragen can take BCL or Fastq files
- Our current max sequencing load can’t touch this thing
- Frees up our compute cluster to develop new things and for other compute heavy jobs to be scheduled
Does it make a difference?
Patient CMH000487

- Fetal MRI: Several congenital anomalies
- Delivery in the CMH materno-fetal health center
- Admitted to the NICU
- Acute liver failure @ 2 months of age
- Cause unknown despite extensive testing
Diagnosis and treatment change

- Following testing and confirmation - IV corticosteroids & immunoglobulin
- Liver function returned to normal and baby got to go home
What are we really accomplishing here?

- Rapidly diagnose critically ill patients
- Create a paradigm shift
- End the diagnostic odyssey
- Enable powerful treatment options
- Identify genes for which no other test exists
- Discover new disorders and disease genes
Majority of variants are novel

Majority of variants not observed in other public databases

- **Novel**: 42
- **dbSNP**: 1
- **other**: 1

Novel 58%
The nordic suite

SSAGA - clinical presentation

- Maps clinical symptoms to diseases to genes using standardized vocabularies - SNOMED, Human Phenotype Ontology
- Nominates superset of clinically relevant genes
- Shortens list of variants and candidate genes
- Standardizes clinical information to assist in interpretation of results
- Archival of patient clinical features
13 mo. male with muscle weakness, developmental regression, startle reflex, ataxia and leukodystrophy on brain MRI…
The nordic suite

RUNES - variant characterization

What does that mean???

• Variant calls must be evaluated to determine their functional consequence
• Characterization done through prediction tools and cross-referencing with external databases
• Final ACMG variant score

 annotations

affected gene(s)?
known disease gene?
transcript context?
cause loss of initiation
cause loss/gain of stop?
change amino acid?
disrupt translation frame?
disrupt splicing?
has variant been observed before?
known disease causing mutation?
known to be benign?
population allele frequency?
The nordic suite

The CMH Variant Warehouse

- database recording characterized results of every variant observed in the CPGM population
- 140M variants (as of 2016-03-28)
- 4584 patients and family members
  - 1803 TaGSCAN
  - 2315 exome
  - 447 whole genome
- Searchable by gene, category, allele frequency
- Curation tools based on ACMG recommendations for capturing in depth manual analysis
The nordic suite

VIKING - Interpretation

• Whole Genome Scale

• SSAGA and RUNES integration

• Trio/familial/set analysis

• Dynamic filtering
  – Relevant clinical features
  – Variant classification
  – Allele frequency
Free for Academic/Research Use

Variant Warehouse - https://warehouse.cmh.edu
RUNES/VIking - https://www.childrensmercy.org/genomesoftwareportal
SSAGA - https://ssaga.cmh.edu

bioinformatics@cmh.edu
“The best way to find yourself is to lose yourself in the service of others.”

–Mahatma Gandhi

Thank you.