Whole Genome Sequencing of African American Family Identifies Novel Susceptibility Variants in Kawasaki Disease

Jihoon Kim

February 3, 2015

University of California, San Diego
Disclosure Information

• Presenter
  – Jihoon Kim (UC San Diego)

• Title:
  – Whole Genome Sequencing of African American Family Identifies Novel Susceptibility Variants in Kawasaki Disease

• Financial Disclosure
  – No relevant financial relationship exists
About iDASH
Molecular Phenotyping for Kawasaki Disease

Recent Webinars

In the Future Will a Biological Database Really be Different than a Biological Journal?
Host: Philip Bourne
University of California, San Diego

Defining and Organizing Medical NLP Tasks
Host: Kiyotaka Taira
University of California, Los Angeles

iDASH News

Workshop on Mobile Data Repository and Analysis Platform
September 18, 2013, UCSD, La Jolla, CA

iDASH Third Annual All Hands Symposium
September 16-17, 2013, UCSD, La Jolla, CA

iDASH 2013 Summer Internship Symposium
August 9, 2013, Calit2 Auditorium, UCSD, La Jolla, CA
# KD incidence by race in US

<table>
<thead>
<tr>
<th>Year</th>
<th>Location</th>
<th>Race</th>
<th>Rate/100,000 children &lt;5 yrs.</th>
<th>Relative risk</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987-1989</td>
<td>Washington</td>
<td>White</td>
<td>12.7 23.4 33.3</td>
<td>Reference 1.8 2.6</td>
<td>(Davis et al. 1995)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997-1998</td>
<td>Georgia</td>
<td>White</td>
<td>8.0 18.4 32.0</td>
<td>Reference 2.3 4.0</td>
<td>(Gibbons et al. 2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992-1999</td>
<td>Detroit</td>
<td>White</td>
<td>7.7 15.0</td>
<td>Reference 1.9</td>
<td>(Abuhammour et al. 2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995-1999</td>
<td>California</td>
<td>White</td>
<td>14.7 24.6 35.3</td>
<td>Reference 1.7 2.4</td>
<td>(Chang 2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>U.S.</td>
<td>White</td>
<td>11.4 19.7 39.0</td>
<td>Reference 1.7 3.4</td>
<td>(Holman et al. 2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pedigree chart for a family of six

[Diagram showing a pedigree chart with symbols indicating affected and not affected individuals with KD.]
Whole genome sequencing
Sequencing

• Sample type: blood
• Whole Genome Sequencing
• Sequencer: Illumina HiSeq
• Sequencing location: Illumina
• Average depth of coverage: ~ 30
Big Data transfer

Illumina sent 500 GB hard drives to UC San Diego by FedEx.
WGS
- raw WGS reads from a family of six
  (6 FASTQ files)
- 6,712,158 variants
  (6 VCF files)
- 303 variants
  (117 genes)
  Prioritizing variants with deleteriousness
  - gain of function: literature, gene fusion, activating mutation, BSIFT, miRNA binding
  - loss of function: frameshift, stop-codon change, SIFT, PolyPhen-2, splice site loss, ENCODE TFBS, enhancer, conservation with phyloP, UTR

GWAS
- keep the variants that show significant association with disease in imputed GWAS from published studies
  under allelic or recessive inheritance modes

mRNA
- keep the promoter-region variants that shows significant genotype-disease interaction effect on mRNA expression of regulated genes from published studies

6 short reads (FASTQ)

6 aligned reads (BAM)

6.7 M variants (VCF)

21K variants (VCF)

303 variants (117 genes)

23 variants (10 genes)

1 variant (1 gene)
Alignment and Variant Calling

raw WGS reads from a family of six (6 FASTQ files)

WGS

Alignment and post-processing QC
- align short reads to a reference genome
- mark and remove PCR duplicates originated from DNA prep method
- realign reads around InDels to uncover hidden InDels and eliminate false positive SNPs
- recalibrate base quality scores based on covariates; reported qual, machine cycle, and dinucleotide

6 aligned reads (6 BAM files)

Variant calling
- call SNPs and InDels

6,712,158 variants (6 VCF files)
Input: short-read sequence data in FASTQ format

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTTGTGGGAACCGAAAGG
GTTTTGAATTTCAAAACCCTTTTCGTTTCCAAACCTTCAAA
AGCAATGCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""""""""7FQ71,'"";C?,B;?6B;:EA1EA
1EA5’9B:?:#9EA0D@2EA5’:>5?:%A;A8A;?9B;D@;
/=<?7=9<2A8==
```

Phred-like Quality Score

Output: a list of SNPs

- chromosome: 8
- position: 108390076
- reference allele: A
- alternative allele: G
- gene: ANGPT1
- region: intronic
- reference snp id: rs35899249
Filtering with genetic inheritance

Filtering with genetic inheritance model
- keep the variants at which all affected children have homozygous ALT allele
- drop the variants at which any unaffected child has the identical ALT homozygote to an affected child
- drop the variants at which any parent has the identical ALT homozygote to an affected child
- keep the variants at which at least one parent has the heterozygous genotypes

6,712,158 variants
(6 VCF files)

20,943 variants
(597 genes)
Prioritizing variants with deleteriousness

20,943 variants
(597 genes)

Prioritizing variants with deleteriousness
- gain of function: literature, gene fusion, activating mutation, BSIFT, miRNA binding
- loss of function: frameshift, stop-codon change, SIFT, PolyPhen-2, splice site loss, ENCODE TFBS, enhancer, conservation with phyloP, UTR

303 variants
(117 genes)
Intersection with Imputed GWAS

• 405 Kawasaki Disease subjects
• 6,252 normal controls
• 4,060,864 imputed genotypes (Illumina HumanRef-12 V4 BeadChip)
• Found 438,343 SNPs by PLINK (nominal P-value < 0.05)

Genome-wide association study identifies FCGR2A as a susceptibility locus for Kawasaki disease.


¹Corresponding author.
WGS Discovery Confirmed in GWAS

23 variants in 10 genes

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Position</th>
<th>Reference Allele</th>
<th>Sample Allele</th>
<th>dbSNP ID</th>
<th>Gene Region</th>
<th>Gene Symbol</th>
<th>Regulatory Site</th>
<th>Regulator</th>
<th>1000 Genomes Frequency</th>
<th>Imputed GWAS pvalue with recessive inheritance model</th>
<th>Imputed GWAS pvalue with allelic inheritance model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>113615193</td>
<td>T</td>
<td>C</td>
<td>386544236</td>
<td>Promoter</td>
<td>LRIG2</td>
<td>Promoter Loss</td>
<td>ENCODE TNF-kappaB; E2F1; F0</td>
<td>57.39</td>
<td>0.0163</td>
<td>0.002328</td>
</tr>
<tr>
<td>4</td>
<td>38825552</td>
<td>G</td>
<td>A</td>
<td>12650224</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>21.33</td>
<td>0.02011</td>
<td>0.008175</td>
</tr>
<tr>
<td>4</td>
<td>38826076</td>
<td>C</td>
<td>A</td>
<td>6822503</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>63.3</td>
<td>0.02001</td>
<td>0.008177</td>
</tr>
<tr>
<td>4</td>
<td>38826429</td>
<td>C</td>
<td>T</td>
<td>12645200</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>63.3</td>
<td>0.02001</td>
<td>0.008177</td>
</tr>
<tr>
<td>4</td>
<td>38827855</td>
<td>C</td>
<td>T</td>
<td>5743826</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>4.15</td>
<td>0.0423</td>
<td>0.01684</td>
</tr>
<tr>
<td>8</td>
<td>108261808</td>
<td>C</td>
<td>G</td>
<td>2507799</td>
<td>3'UTR</td>
<td>ANGPT1</td>
<td></td>
<td></td>
<td>34.5</td>
<td>0.0423</td>
<td>0.01434</td>
</tr>
<tr>
<td>10</td>
<td>104624072</td>
<td>G</td>
<td>A</td>
<td>4917985</td>
<td>3'UTR</td>
<td>C10orf32</td>
<td></td>
<td></td>
<td>42.73</td>
<td>0.01434</td>
<td>0.005483</td>
</tr>
<tr>
<td>10</td>
<td>104629465</td>
<td>T</td>
<td>G</td>
<td>3740400</td>
<td>Intronic</td>
<td>AS3MT</td>
<td>ENCODE TFBS</td>
<td>POLR2A</td>
<td>42.73</td>
<td>0.01434</td>
<td>0.005483</td>
</tr>
<tr>
<td>10</td>
<td>104677887</td>
<td>A</td>
<td>C</td>
<td>10883801</td>
<td>Promoter</td>
<td>CNNM2</td>
<td>Promoter Loss</td>
<td>ENCODE TEGR1; E2F1; E2F6; R</td>
<td>28.41</td>
<td>0.005483</td>
<td>0.004337</td>
</tr>
<tr>
<td>10</td>
<td>104837816</td>
<td>G</td>
<td>A</td>
<td>1046411</td>
<td>3'UTR</td>
<td>CNNM2</td>
<td></td>
<td></td>
<td>26.28</td>
<td>0.005483</td>
<td>0.004337</td>
</tr>
<tr>
<td>10</td>
<td>104848123</td>
<td>G</td>
<td>A</td>
<td>8139</td>
<td>3'UTR</td>
<td>NT5C2</td>
<td></td>
<td></td>
<td>42.47</td>
<td>0.005483</td>
<td>0.004337</td>
</tr>
<tr>
<td>10</td>
<td>104953547</td>
<td>T</td>
<td>C</td>
<td>10748839</td>
<td>Promoter</td>
<td>NT5C2</td>
<td>Promoter Loss</td>
<td>ENCODE TEGF1; ZEB1; TCF7L2;</td>
<td>44.43</td>
<td>0.01138</td>
<td>0.002193</td>
</tr>
<tr>
<td>10</td>
<td>105726812</td>
<td>A</td>
<td>G</td>
<td>10786779</td>
<td>Promoter</td>
<td>SLK</td>
<td>Promoter Loss</td>
<td>ENCODE TEGR1; CCNT2; E2F1;</td>
<td>79.29</td>
<td>0.01138</td>
<td>0.002193</td>
</tr>
<tr>
<td>11</td>
<td>102661080</td>
<td>G</td>
<td>A</td>
<td>2239008</td>
<td>3'UTR</td>
<td>MMP1</td>
<td></td>
<td></td>
<td>23.14</td>
<td>0.03223</td>
<td>0.004546</td>
</tr>
<tr>
<td>16</td>
<td>81590971</td>
<td>C</td>
<td>T</td>
<td>9888752</td>
<td>Intronic</td>
<td>CMIP</td>
<td></td>
<td></td>
<td>47.54</td>
<td>0.03443</td>
<td>0.03443</td>
</tr>
</tbody>
</table>
Re-discovered Known KD Susceptibility Variants

TLR6, ANGPT1, and MMP1

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Position</th>
<th>Reference Allele</th>
<th>Sample Allele</th>
<th>dbSNP ID</th>
<th>Gene Region</th>
<th>Gene Symbol</th>
<th>Regulatory Site</th>
<th>Regulator</th>
<th>1000 Genomes Frequency</th>
<th>Imputed GWAS pvalue with recessive inheritance model</th>
<th>Imputed GWAS pvalue with allelic inheritance model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>113615193</td>
<td>T</td>
<td>C</td>
<td>386544236</td>
<td>Promoter</td>
<td>LRIG2</td>
<td>Promoter Loss; ENCODE</td>
<td>TLR6</td>
<td>57.39</td>
<td>0.0163</td>
<td>0.002328</td>
</tr>
<tr>
<td>4</td>
<td>38825552</td>
<td>G</td>
<td>A</td>
<td>12650224</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>21.33</td>
<td>0.02011</td>
<td>0.008175</td>
</tr>
<tr>
<td>4</td>
<td>38826076</td>
<td>C</td>
<td>A</td>
<td>6822503</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>63.3</td>
<td>0.02001</td>
<td>0.008177</td>
</tr>
<tr>
<td>4</td>
<td>38826429</td>
<td>C</td>
<td>T</td>
<td>12645200</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>63.3</td>
<td>0.02001</td>
<td>0.008177</td>
</tr>
<tr>
<td>4</td>
<td>38827855</td>
<td>C</td>
<td>T</td>
<td>5743826</td>
<td>3'UTR</td>
<td>TLR6</td>
<td></td>
<td></td>
<td>4.15</td>
<td>0.04751</td>
<td>0.002853</td>
</tr>
<tr>
<td>8</td>
<td>108261808</td>
<td>C</td>
<td>G</td>
<td>2507799</td>
<td>3'UTR</td>
<td>ANGPT1</td>
<td></td>
<td></td>
<td>34.5</td>
<td>0.04751</td>
<td>0.002853</td>
</tr>
<tr>
<td>10</td>
<td>104624072</td>
<td>G</td>
<td>A</td>
<td>4917985</td>
<td>3'UTR</td>
<td>C10orf32</td>
<td></td>
<td></td>
<td>42.73</td>
<td>0.0423</td>
<td>0.01434</td>
</tr>
<tr>
<td>10</td>
<td>104629465</td>
<td>T</td>
<td>G</td>
<td>3740400</td>
<td>Intronic</td>
<td>AS3MT</td>
<td>ENCODE TFBS</td>
<td>POLR2A</td>
<td>42.73</td>
<td>0.01434</td>
<td>0.005483</td>
</tr>
<tr>
<td>10</td>
<td>104677887</td>
<td>A</td>
<td>C</td>
<td>10883801</td>
<td>Promoter</td>
<td>CNNM2</td>
<td>Promoter Loss; ENCODE</td>
<td>TEGR1;</td>
<td>28.41</td>
<td>0.005483</td>
<td>0.005483</td>
</tr>
<tr>
<td>10</td>
<td>104837816</td>
<td>G</td>
<td>A</td>
<td>1046411</td>
<td>3'UTR</td>
<td>CNNM2</td>
<td></td>
<td></td>
<td>26.28</td>
<td>0.004337</td>
<td>0.004337</td>
</tr>
<tr>
<td>10</td>
<td>104848123</td>
<td>G</td>
<td>A</td>
<td>8139</td>
<td>3'UTR</td>
<td>NT5C2</td>
<td></td>
<td></td>
<td>42.47</td>
<td>0.01072</td>
<td>0.005483</td>
</tr>
<tr>
<td>10</td>
<td>104953547</td>
<td>T</td>
<td>C</td>
<td>10748938</td>
<td>Promoter</td>
<td>NT5C2</td>
<td>Promoter Loss; ENCODE</td>
<td>T2F1; ZEB1; TCF7L2;</td>
<td>44.43</td>
<td>0.01138</td>
<td>0.005483</td>
</tr>
<tr>
<td>10</td>
<td>105726812</td>
<td>A</td>
<td>G</td>
<td>10786779</td>
<td>Promoter</td>
<td>SLK</td>
<td>Promoter Loss; ENCODE</td>
<td>TEGR1; CCNT2; T2F1;</td>
<td>79.29</td>
<td>0.02193</td>
<td>0.005483</td>
</tr>
<tr>
<td>11</td>
<td>102661080</td>
<td>G</td>
<td>A</td>
<td>2239008</td>
<td>3'UTR</td>
<td>MMP1</td>
<td></td>
<td></td>
<td>23.14</td>
<td>0.03223</td>
<td>0.004546</td>
</tr>
<tr>
<td>16</td>
<td>81590971</td>
<td>C</td>
<td>T</td>
<td>9888752</td>
<td>Intronic</td>
<td>CMIP</td>
<td></td>
<td></td>
<td>47.54</td>
<td>0.03443</td>
<td>0.004546</td>
</tr>
</tbody>
</table>
Gene Expression Analysis

• Same sample source
  – individuals participated in GWAS

• For each variant,
  – grouped subjects by their genotypes
  – tested if there is difference in mRNA expression level between acute and convalescent in each genotype


Global gene expression profiling identifies new therapeutic targets in acute Kawasaki disease.

Hoang LT\textsuperscript{1}, Shimizu C\textsuperscript{2}, Ling L\textsuperscript{1}, Naim AN\textsuperscript{1}, Khor CC\textsuperscript{1}, Tremoulet AH\textsuperscript{2}, Wright V\textsuperscript{3}, Levin M\textsuperscript{3}, Hibberd ML\textsuperscript{1}, Burns JC\textsuperscript{2}.
rs10786779 in SLK gene

• serine-threonine protein kinase 2 (SLK)
• Located in promoter of SLK genes
• 16 Genes regulated by SLK
  – EGR1  CCNT  E2F1  TAF1  E2F4
  – E2F6  PAX5  NFKB1  TBP  USF1
  – USF2  POLR2A  NRF1  HEY1  TFAP2A
  – TFAP2C
• Want to test if presence of SNP rs10786779 changes mRNA expression in two groups (acute vs. convalescent)
Differential expression of acute and convalescent only observed in genotype AA group of TFAP2C

TFAP2C
10786779

yellow: acute
blue: convalescent

Expression
-0.8 -0.7 -0.6 -0.5 -0.4 -0.3
GG GG GA GA AA AA

Genotype
87, 34, 4

P=0.04

sample size
SLK survived all three methods.
Serine Threonine Protein Kinase 2 (SLK)

- **Function:**
  - required in cell motility by phosphorylating paxillin at focal adhesion
  - up-regulated by scratch wounding of fibroblast monolayers (Wagner et al. 2008; Quizi et al. 2013).

- **Regulation**
  - Chip-seq data (Regulome, Encode) suggested this SNP most likely to affect binding of transcription regulator such as Myc, MAX, and NFkB that are implicated in cell proliferation, differentiation and apoptosis.

- **Allele frequency**
  - Affected siblings are G allele homozygote.
  - G allele frequencies are 65% in African, 90% in Asian and 83% in European according to 1000genome.
Discussion

• Difference in cohort ethnicity exist.
• Need more family WGS data for KD study.
• Sequenced a KD patient with severe aneurysm
  – External to the family in this study.
• Currently we are trying to explain newly discovered gene by WGS in KD context.
• Manuscript is in preparation for this WGS study.
Big Resource Needed for Big Data

<table>
<thead>
<tr>
<th>Step</th>
<th>Execution time (minutes)</th>
<th>Output size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean BAM</td>
<td>480</td>
<td>620</td>
</tr>
<tr>
<td>Split BAM to two FASTQ files</td>
<td>780</td>
<td>400</td>
</tr>
<tr>
<td>Align reads to the ref genome</td>
<td>1240</td>
<td>700</td>
</tr>
<tr>
<td>Merge the alignment to a single SAM</td>
<td>980</td>
<td>450</td>
</tr>
<tr>
<td>Convert SAM to BAM</td>
<td>560</td>
<td>150</td>
</tr>
<tr>
<td>Sort BAM</td>
<td>420</td>
<td>150</td>
</tr>
<tr>
<td>Remove PCR duplicates</td>
<td>540</td>
<td>140</td>
</tr>
<tr>
<td>Recalibrate the base quality</td>
<td>820</td>
<td>140</td>
</tr>
<tr>
<td>Re-align reads locally around INDELs</td>
<td>1350</td>
<td>140</td>
</tr>
<tr>
<td>Call variants</td>
<td>2150</td>
<td>3</td>
</tr>
<tr>
<td>Filter variants</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Annotate variants</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Multi-sample analysis</td>
<td>650</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>10010</strong></td>
<td><strong>2908</strong></td>
</tr>
</tbody>
</table>

1 week and 3TB storage per sample using
iDASH / San Diego Supercomputer Center (SDSC)
Acknowledgement

- NIH
  - Roadmap for Medical Research, Grant U54HL108460
- UC San Diego, USA
  - Jane Burns
  - Lucila Ohno-Machado
  - Chisato Shimizu
  - Olivier Harismendy
  - Eric Levy
  - Hai Yang
- Illumina
  - Jay Flatley
- University of Florence, Italy
  - Rolando Cimaz
- Murdoch Children’s Research Institute, Australia
  - David Burgner
- Sainte-Justine University Hospital Center, Canada
  - Nagib Dahdah
- The Hospital for Sick Children, Canada
  - Rae Young
- University of Amsterdam, The Netherlands
  - Taco Kuijpers
- Imperial College London, United Kingdom
  - Michael Levin
- Genome Institute of Singapore, Singapore
  - Chiea Chuen Khor
  - Sonia Davila
  - Long Truong Hong
  - Martin Hibberd
- International Kawasaki Disease Genetics Consortium (IKDGC)