Analysis of 3 Israeli families with Keratoconus

9 June 2016

By

Wes Goar
What is Keratoconus?
Review of data
Linkage Analysis
Exome Sequencing
Autozygosity
CNV analysis
Linkage using only exome data
Keratoconus

- Most common corneal dystrophy and a major cause of blindness
  - Cornea thins and ocular pressures causes corneal bulge
  - Causes myopia (nearsightedness), irregular astigmatism, and corneal scarring
  - Incidence rate: 1 in 2000 in the general population
  - Age of onset is *generally* during the late teens/early twenties
  - Progresses for a period of 10-20 years
  - Similar incidence of prevalence among genders

- The pattern of inheritance is poorly understood

- Corneal transplantation only treatment when visual acuity is no longer correctable by contact lenses
History of Keratoconus Genetics

- Inheritance patterns
  - 1938- J.B. Hamilton strongly supported an autosomal recessive inheritance pattern
  - 1969- Falls & Allen suggested irregular autosomal dominance inheritance pattern
  - 1986- Ihalainen suggested both autosomal dominance and incomplete penetrance
  - 2000- Wang et al. suggested an autosomal recessive inheritance pattern
Pathogenesis

- 2005 - Lema & Duran found an increase in levels of IL6, TNFA, and MMP9 in the tears of KT patients compared to non-KT patients
- 2015 - Shetty et al. found a reduction in the protein levels of LOX and COL4A1 in the basement membrane of KT patients compared to healthy donor corneas.

Molecular genetics

- 2002 - VSX1 - has both evidence for and against a pathogenic role in KT
- SOD1, COL4A3, COL4A4, SPARC have both evidence for and against a pathogenic role in KT
- 2014 - Lechner et al. found significant enrichment of potentially pathogenic ZNF469 alleles in KT patients compared to controls
Diagnosis

- 4 Stages of Progression
- Each stage measures
  - Level of myopia
  - Level of astigmatism
  - K-reading- measures the curvature of the anterior surface of the cornea
  - Pachymetry- measures the thickness of the eye’s cornea
Astigmatism
3 Israeli families with Keratoconus for a total of 16 samples
DNA samples collected by the Ruti Pavari lab in Israel
Limited phenotypic and family history information given
Family 2

N/A

N/A

Exome

33

20

58

N/A
Family 3

- 43
- 12
- N/A
- N/A
- Exome
Data...

- An Affymetrix Genome-Wide SNP 5.0 array was performed to genotype the samples.
- Merlin was used on the Affymetrix SNP data to determine regions of the genome consistent with segregation using both dominant and recessive inheritance models.
- 1 sample from each family was submitted for whole-exome sequencing through the DNA core at the University of Iowa (HiSeq 2500)
Results from Linkage...

- Familial information didn't allow us to receive a high enough positive LOD score to have genome-wide significance.
- They did however allow us to eliminate regions inconsistent with linkage.
- Extracted locations that had LOD scores above 0.
Filtering using dominant model

<table>
<thead>
<tr>
<th>Family</th>
<th>Mbp</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family 1</td>
<td>379 (11.7%)</td>
<td>3750</td>
</tr>
<tr>
<td>Family 2</td>
<td>728.786 (22.5%)</td>
<td>6241</td>
</tr>
<tr>
<td>Family 3</td>
<td>1685 (52%)</td>
<td>13929</td>
</tr>
<tr>
<td>Shared</td>
<td>25 (0.77%)</td>
<td>245</td>
</tr>
</tbody>
</table>
Filtering using dominant model
Agilent SureSelect with additional baits (xGen Lockdown, IDT, Coralville, IA) added to capture undercovered and select intronic regions in retinal degeneration genes

Illumina Hiseq 2000/2500 – 4 samples per high output lane

BWA

PicardTools

GATK indel realigner

GATK Unified Genotyper - SNVs and small indels
Conifer – large CNVs (4+ exons)

Small Variant Calling

CNV Calling

Annotation

Filtering

Exome Capture

Sequencing

Alignment

Remove Duplicates

Realignment

Custom Scripts

Custom Scripts
Filtering steps after pipeline...

Variant: 3:97510664 C / T

**Annotations**
This variant falls on 5 transcripts in 1 genes:

- **missense**
  - ARL6

- **intron**
  - ARL6 - ENST00000476753

**Population Frequencies**

<table>
<thead>
<tr>
<th>Population</th>
<th>Allele Count</th>
<th>Allele Number</th>
<th>Number of Homozygotes</th>
<th>Allele Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latino</td>
<td>4</td>
<td>11546</td>
<td>0</td>
<td>0.0003464</td>
</tr>
<tr>
<td>African</td>
<td>0</td>
<td>10402</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>East Asian</td>
<td>0</td>
<td>8638</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>European (Finnish)</td>
<td>0</td>
<td>6614</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>European (Non- Finnish)</td>
<td>0</td>
<td>66714</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>906</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>South Asian</td>
<td>0</td>
<td>16510</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>121330</td>
<td>0</td>
<td>3.297e-08</td>
</tr>
</tbody>
</table>

**Filtering using Minor Allele Frequencies to remove variants seen too commonly in the population**

- 1000 Genomes, Exome Aggregation Consortium and our own local set of 1000+ exomes

- Remove variants found in areas inconsistent with familial segregation (1 file for Dominant & 1 for Recessive Inheritance)

- Remove variants of low quality & synonymous mutations
### Progression in filtering...

<table>
<thead>
<tr>
<th>Family</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>KT/Non-KT Individuals</td>
<td>3/4</td>
<td>3/1</td>
<td>2/2</td>
</tr>
<tr>
<td>Total number of variants</td>
<td>144,183</td>
<td>144,472</td>
<td>153,708</td>
</tr>
<tr>
<td># Plausible variants (PV)</td>
<td>451</td>
<td>459</td>
<td>458</td>
</tr>
<tr>
<td># PV segregating in the family</td>
<td>55</td>
<td>55</td>
<td>198</td>
</tr>
<tr>
<td># PV in shared regions</td>
<td>10</td>
<td>13</td>
<td>9</td>
</tr>
</tbody>
</table>
Developing a candidate list...

- **OMIM**
  - Catalog of human gene & genetic disorders and traits, with a particular focus on gene-phenotype relationship

- **WebGestalt**
  - Designed for functional genomic, proteomic, and large-scale genetic studies from which a large number of gene lists are generated

- **SMART (Simple Modular Architecture Research Tool)**
  - Allows the identification & annotation of genetically mobile domains and the analysis of domain architectures

- **Ocular Tissue Database**

- **STRING**
  - Biological database & web resource of known & predicted protein-protein interactions
Recessive candidate...

- PRSS53 (Polyserase 3)
  - Codes a serine-type endopeptidase (breaks peptide bonds of nonterminal amino acids)
  - Compound Heterozygote Mutation
  - Pro460Leu
    - Probably damaging according to PolyPhen2 (.997)
  - Pro212His
    - Possibly damaging according to PolyPhen2 (.873)
  - Both mutations lie within the Trypsin-like Serine Protease Domain
Dominant Candidates...

- IGFBP2 (Insulin like growth factor binding protein 2)
- Seen in all three families
  - Heterozygous
  - Leu25Pro
    - Probably damaging according to PolyPhen2 (.971)
- After further examination, the variant is within other non-disease exomes and is an artifact from sequencing errors in an area of low complexity (Visual inspection of the reads performed using IGV)
IGV of IGFBP2...
Dominant Candidates...

- THSD4 (Thrombospondin Type-1 Domain-Containing Protein 4)
  - Heterozygous mutation
  - Glu382Lys
  - Probably damaging according to PolyPhen2 (.996)
  - Mutations within the TSP1 domain
- Passed IGV inspection
- Known expression in the eye
- Involved with assembly of micorfibrils; incorrect assembly could lead to degradation of the cornea
IGV of THSD4...
Autozygosity

- Homozygosity in which the 2 alleles are identical by descent
- Using the SNP array data and Plink, we determined areas of autozygosity among members within each family.
- Using BEDtools we identified regions of shared autozygosity between the 3 families (4.15Mbp)
To identify copy number variants, we ran Conifer on our 3 exomes. We identified 4 duplications and 2 deletions in the 3 samples. One region had a duplication in the sample from family 1 and a deletion in the sample from family 3. Further inspection of that region led us to discover a 2 exon deletion in the sample from family 2 in that same region. These variants are the subject of ongoing investigation, including their confirmation in the samples.
We have developed a protocol to identify regions of the genome consistent with segregation using exome sequencing data. We first extract informative markers from the exome sequencing data, then Merlin is used to identify regions consistent with segregation. We have used this protocol on families with Pigment Dispersion Syndrome and have identified regions of the genome with significant LOD scores (>3).
Conclusions & Future Directions

- Limitations:
  - Need more phenotypic data
  - We are working on obtaining more samples with Dr. Pavari in Israel
  - We are currently performing further investigation for a few of the variants we identified in our current samples
    - KT could be a complex disease
  - Lots of different tools and techniques available to help identify candidates for your disease of interest.
Acknowledgments

- Val Sheffield
- Todd Scheetz
- Charles Searby
- Adam DeLuca
- Scott Whitmore
- Sheffield Lab
- Ruti Pavari
- Rob Mullins
- Mike Anderson
- John Fingert
- Michael Wagoner

The Stephen A. Wynn Foundation
This research was supported in part through computational resources provided by The University of Iowa, Iowa City, Iowa.

NIH Predoctoral Training Grant
T32GM008629