Development of Genomics Plugins in i2b2

Lori Phillips, MS
AUG Meeting
June 18, 2013
Big Picture - Data flow of next-gen sequencing

- base calls from the sequencer
- FASTQ files with base calls
- SAM with standard alignment
- VCF digests variants
- GVF maps to ontologies

De-identified Data Warehouse
Importing NGS variant output into i2b2

- Variant Call Format
- Gene Annotated VCF
- Genome Variation Format

Diagram:
- VCF
- ANNOVAR
- GVF
- i2b2

Observation fact
Pipeline - VCF to VCF-ANNO

VCF

ANNOVAR*

VCF-ANNO

**Pipeline - VCF-ANNO to GVF**

<table>
<thead>
<tr>
<th>Exonic</th>
<th>TTLL10</th>
<th>1</th>
<th>1105366</th>
<th>1105366</th>
<th>T</th>
<th>C</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1105366</td>
<td>.</td>
<td>T</td>
<td>C</td>
<td>.</td>
<td>PASS</td>
<td></td>
</tr>
<tr>
<td>AA=T;AC=4;AN=114;DP=3251</td>
<td>GT:DP</td>
<td>1/0:54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*VCF-ANNO to GVF*

**GVF**

<table>
<thead>
<tr>
<th>chr1</th>
<th>VCF</th>
<th>SNV</th>
<th>1105366</th>
<th>1105366</th>
<th>.</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ID=1;Reference_seq=T;Variant_seq=C;Variant_feature=exonic;Gene=TTLL10; Genotype=heterozygous</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Kong, Sek-Won, Lee, Joon, Boston Children’s Hospital (perl script) modified for ANNOVAR by Lori Phillips*
Pipeline – GVF to I2B2 records

chr1  VCF   SNV      1105366  1105366  .   +

ID=1;Reference_seq=T;Variant_seq=C;Variant_feature=exonic;Gene=TTLL10;
Genotype=heterozygous

GVF2I2B2

1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"@"|1
"chr1"|"GVF2I2B2"

1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:Start"|1|"N"|"E"
1105366|"GVF2I2B2"

1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:End"|1|"N"|"E"
1105366|"GVF2I2B2"

1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:Zygosity"|1|"T"
"heterozygous"|"GVF2I2B2"

1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:HUGO"|1|"T"
"TTLL10"|"GVF2I2B2"

1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SO:0001791"|1
"GVF2I2B2"
Import NGS Variant Data

Analysis details
Information related to the NGS data

- Specify input file: [Browse]
- Input file format: VCF
- VCF mapping file: [Browse]
- I2B2 Patient number:
- I2B2 Encounter number:
- Date of encounter:
- Reference genome version: hg18

Sample details
Information related to the sample

- Sample ID:
- Sample Type: TISSUE
- Anatomical Source: Pericardium
- Collection Method: BIOPSY
- Additive: UNKNOWN

Sample Pathology
Information related to the sample pathology

- Pathology: TUMOR
- Tumor Grade: UNKNOWN
- Tumor Stage: UNKNOWN

Submit
Progress Bar:
## Mapping file

```plaintext
##genome-build hg18
##file-date 2010-07-07
#sample|patient_num|encounter_num
NA12878|1000000090|1880003090
NA12891|1000000093|1880003093
NA12892|1000000094|1880003094
```
Import NGS Variant Data

Analysis details
Information related to the NGS data

Specify input file: els\ANNOVAR\CEU.trio.2010_07.indel.txt
Input file format: VCF-ANNOVAR
VCF mapping file: els\ANNOVAR\CEU.trio.2010_07.map.txt

I2B2 Patient number:
I2B2 Encounter number:
Date of encounter:
Reference genome version: hg18

Submit

Sample details
Information related to the sample

Sample ID:
Sample Type: TISSUE
Anatomical Source: Pericardium
Collection Method: BIOPSY
Additive: UNKNOWN

Sample Pathology
Information related to the sample pathology

Pathology: TUMOR
Tumor Grade: UNKNOWN
Tumor Stage: UNKNOWN

VCF ANNOVAR to GVF step: Converting VCF line 8000
### i2b2 Workbench for Demo SQL Server

### Bulk Load Status

<table>
<thead>
<tr>
<th>Upload Id</th>
<th>Status</th>
<th>Rows Processed</th>
<th>Rows Loaded</th>
<th>Date Loaded</th>
<th>File Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>372</td>
<td>COMPLETED</td>
<td>2349938</td>
<td>2349938</td>
<td>6/13/2013</td>
<td>\phsinfra16\genomics\1000genomes\CEUInDels\ANNOVAR\NA12892.1880003094.i2b2</td>
</tr>
<tr>
<td>371</td>
<td>COMPLETED</td>
<td>2381626</td>
<td>2381626</td>
<td>6/13/2013</td>
<td>\phsinfra16\genomics\1000genomes\CEUInDels\ANNOVAR\NA12891.1880003093.i2b2</td>
</tr>
<tr>
<td>370</td>
<td>COMPLETED</td>
<td>2475996</td>
<td>2475996</td>
<td>6/13/2013</td>
<td>\phsinfra16\genomics\1000genomes\CEUInDels\ANNOVAR\NA12878.1880003090.i2b2</td>
</tr>
</tbody>
</table>
1. Send the i2b2 file to the FR

2. Tell the CRC the file is ready to load

3. SSIS package loads the i2b2 file to observation_fact table
Navigating NGS Variant Data

with Sequence Ontology

Combination of concepts and modifiers to identify:

- An SNV/SNP located on a 3’UTR
- An SNV/SNP associated with a certain gene
- An SNV/SNP of specified zygosity
The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.


Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts, USA.

Abstract
Genetic association studies are viewed as problematic and plagued by irreproducibility. Many associations have been reported for type 2 diabetes, but none have been confirmed in multiple samples and with comprehensive controls. We evaluated 16 published genetic associations to type 2 diabetes and related sub-phenotypes using a family-based design to control for population stratification, and replication samples to increase power. We were able to confirm only one association, that of the common Pro12Ala polymorphism in peroxisome proliferator-activated receptor-gamma(PPARgamma) with type 2 diabetes. By analysing over 3,000 individuals, we found a modest (1.25-fold) but significant ($P=0.002$) increase in diabetes risk associated with the more common proline allele (85% frequency). Moreover, our results resolve a controversy about common variation in PPARgamma. An initial study found a threefold effect, but four of five subsequent publications failed to confirm the association. All six studies are consistent with the odds ratio we describe. The data implicate inherited variation in PPARgamma in the pathogenesis of type 2 diabetes. Because the risk allele occurs at such high frequency, its modest effect translates into a large population attributable risk-influencing as much as 25% of type 2 diabetes in the general population.

PMID: 10973253 [PubMed - indexed for MEDLINE]
The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.


Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts, USA.

Abstract

Genetic association studies are viewed as problematic and plagued by irreproducibility. Many associations have been reported for type 2 diabetes, but none have been confirmed in multiple samples and with comprehensive controls. We evaluated 16 published genetic associations to type 2 diabetes and related sub-phenotypes using a family-based design to control for population stratification, and replication samples to increase power. We were able to confirm only one association, that of the common Pro12Ala polymorphism in peroxisome proliferator-activated receptor-gamma (PPARgamma) with type 2 diabetes. By analysing over 3,000 individuals, we found a modest (1.25-fold) but significant (P=0.002) increase in diabetes risk associated with the more common proline allele (85% frequency). Moreover, our results resolve a controversy about common variation in PPARgamma. An initial study found a threefold effect, but four of five subsequent publications failed to confirm the association. All six studies are consistent with the odds ratio we describe. The data implicate inherited variation in PPARgamma in the pathogenesis of type 2 diabetes. Because the risk allele occurs at such high frequency, its modest effect translates into a large population attributable risk-influencing as much as 25% of type 2 diabetes in the general population.
Gene Association Modifier
### Quick Gene Search

**Search symbols, keywords or IDs for:**

Results that □ equal □ begin □ contain

Display □ 50 □ hits

**Total hits: 11**

<table>
<thead>
<tr>
<th>Approved Symbol</th>
<th>Approved Name</th>
<th>Location</th>
<th>Best Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPAR~withdrawn</td>
<td>symbol withdrawn, see PPARA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPARA</td>
<td>peroxisome proliferator-activated receptor alpha</td>
<td>22q12-q13.1</td>
<td>Previous Symbols: PPAR</td>
</tr>
<tr>
<td>PPARD</td>
<td>peroxisome proliferator-activated receptor delta</td>
<td>6p21.2</td>
<td>Approved Symbol: PPARD</td>
</tr>
<tr>
<td>PPARG</td>
<td>peroxisome proliferator-activated receptor gamma</td>
<td>3p25</td>
<td>Approved Symbol: PPARG</td>
</tr>
<tr>
<td>PPARGC1A</td>
<td>peroxisome proliferator-activated receptor gamma, coactivator 1 alpha</td>
<td>4p15.1</td>
<td>Approved Symbol: PPARGC1A</td>
</tr>
<tr>
<td>PPARGC1B</td>
<td>peroxisome proliferator-activated receptor gamma, coactivator 1 beta</td>
<td>5q33.1</td>
<td>Approved Symbol: PPARGC1B</td>
</tr>
<tr>
<td>MED1</td>
<td>mediator complex subunit 1</td>
<td>17q12</td>
<td>Previous Symbols: PPARBP</td>
</tr>
<tr>
<td>PPARAL~withdrawn</td>
<td>entry withdrawn</td>
<td></td>
<td>Approved Symbol: PPARAL~withdrawn</td>
</tr>
<tr>
<td>ANGPTL4</td>
<td>angiopoietin-like 4</td>
<td>19p13.3</td>
<td>Name Synonyms: PPAR gut related protein</td>
</tr>
<tr>
<td>FAM120B</td>
<td>family with sequence similarity 120B</td>
<td>6q27</td>
<td>Name Synonyms: PPAR gamma constitutive coactivator 1</td>
</tr>
</tbody>
</table>
Specifying Gene Association Modifier

Choose modifier value of SNV/SNP

You are allowed to search within the narrative text associated with the term SNV/SNP.

- No Search Requested
- By abnormal flag
- Search within Text

Exact: PPARG

OK  Cancel  Gene Assist
Building a Translational Genomic Query

- Group 1: SNV/SNP with HGNC Gene Symbol modifier of “PPARG”
The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.


Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts, USA.

Abstract
Genetic association studies are viewed as problematic and plagued by irreproducibility. Many associations have been reported for type 2 diabetes, but none have been confirmed in multiple samples and with comprehensive controls. We evaluated 16 published genetic associations to type 2 diabetes and related sub-phenotypes using a family-based design to control for population stratification, and replication samples to increase power. We were able to confirm only one association, that of the common Pro12Ala polymorphism in peroxisome proliferator-activated receptor-gamma (PPARgamma) with type 2 diabetes. By analysing over 3,000 individuals, we found a modest (1.25-fold) but significant ($P=0.002$) increase in diabetes risk associated with the more common proline allele (85% frequency). Moreover, our results resolve a controversy about common variation in PPARgamma. An initial study found a threefold effect, but four of five subsequent publications failed to confirm the association. All six studies are consistent with the odds ratio we describe. The data implicate inherited variation in PPARgamma in the pathogenesis of type 2 diabetes. Because the risk allele occurs at such high frequency, its modest effect translates into a large population attributable risk-influencing as much as 25% of type 2 diabetes in the general population.
Building a Translational Genomic Query

- Group 2: SNV/SNP with exon variant modifier
  - Note that “Items instance will be same” is selected on the panels
Building a Translational Genomic Query

- **Group 3: Diabetes Mellitus**
  - Select “Treat Independently” for this panel
Run the query
Summary

- A Genomics plug-in was created to create observation-fact files from VCF files.

- A bulk loader was written in native (SQL Server) code to allow for the rapid loading of 2-5 million rows / patient into observation-fact table.

- Sequence Ontology (available at NCBO) that is associated with GVF format can be used to query the next generation sequencing data that was imported into i2b2.