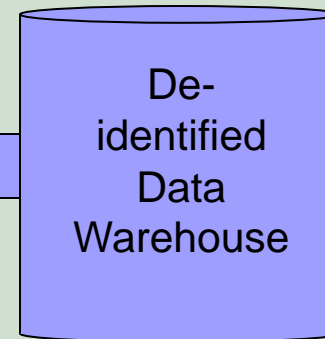
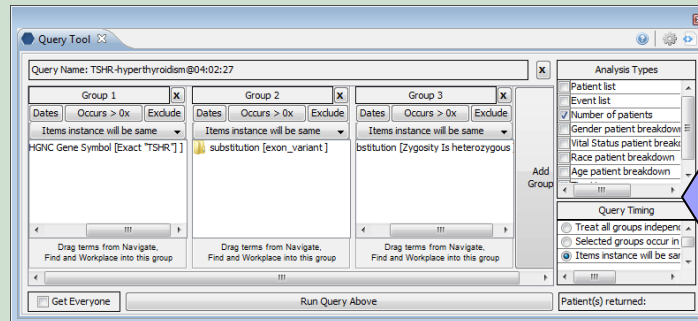
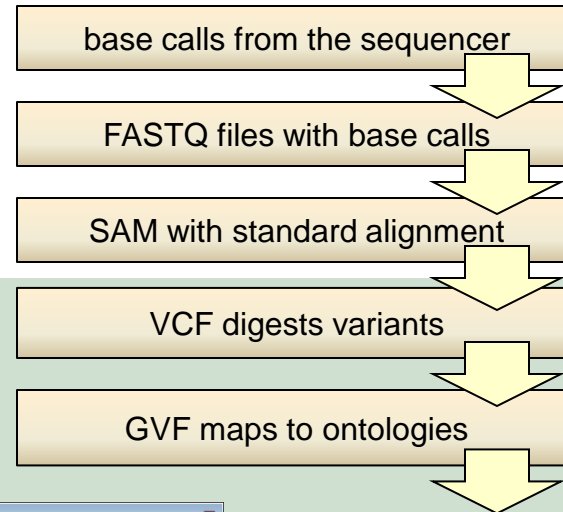


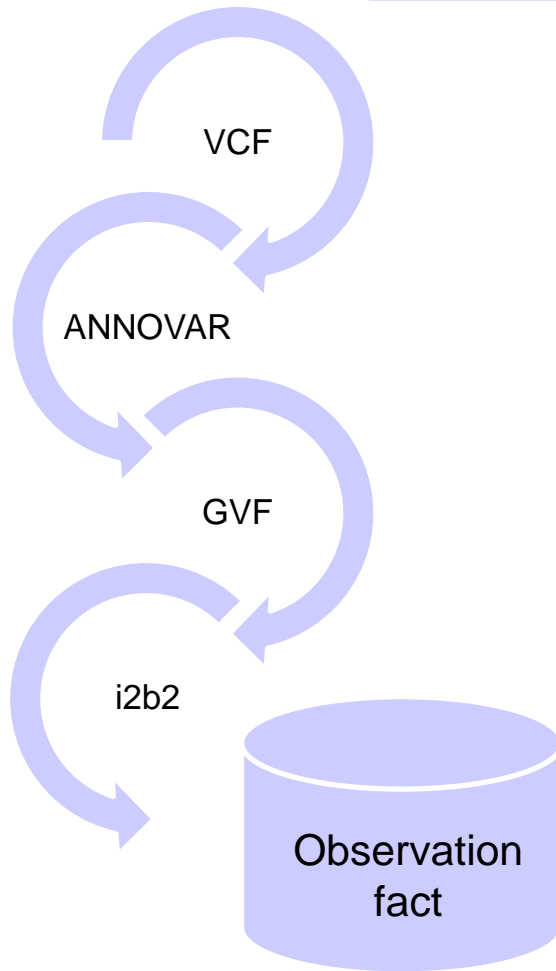
Development of Genomics Plugins in i2b2

Lori Phillips, MS
AUG Meeting
June 18, 2013

Big Picture - Data flow of next-gen sequencing



Importing NGS variant output into i2b2



Variant Call Format

Gene Annotated VCF

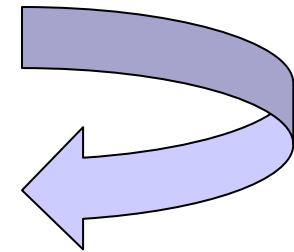
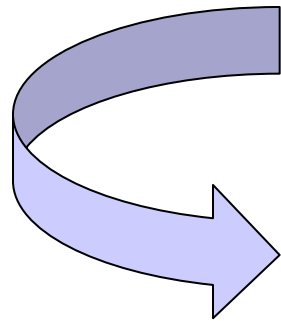
Genome Variation Format

Pipeline - VCF to VCF-ANNO

```
1      1105366      .      T      C      .      PASS
      AA=T;AC=4;AN=114;DP=3251      GT:DP      1/0:54
```

VCF

ANNOVAR*

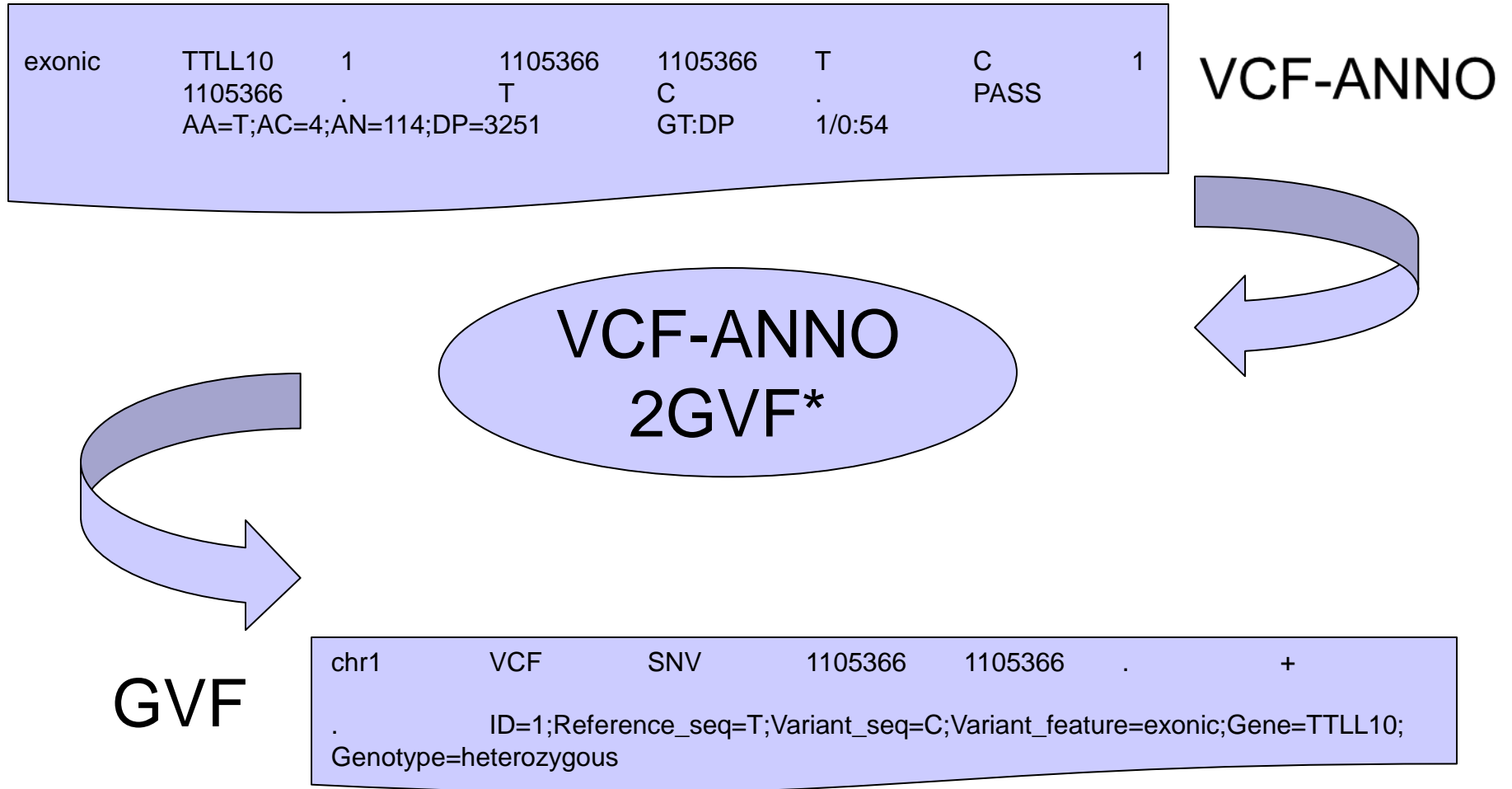


VCF-ANNO

```
exonic  TTLL10      1      1105366      1105366      T      C      1
      1105366      .      T      C      .      PASS
      AA=T;AC=4;AN=114;DP=3251      GT:DP      1/0:54
```

*Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38:e164, 2010 (www.openbioinformatics.org/annovar)

Pipeline - VCF-ANNO to GVF



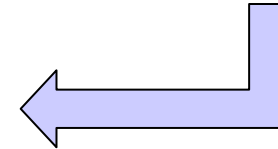
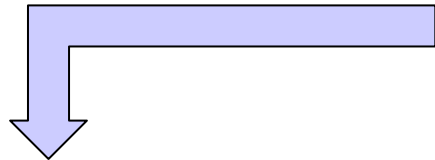
*Kong, Sek-Won, Lee, Joon, Boston Children's Hospital (perl script) modified for ANNOVAR by Lori Phillips

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
```

GVF

GVF2I2B2



I
2
B
2

- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"@"|1|||||||||||||"GVF2I2B2"|
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SO:0000340"|1|"T"|
"chr1"|||||||||||||"GVF2I2B2"| (chr1)
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:Start"|1|"N"|"E"|
1105366|||||||||||||"GVF2I2B2"| (start position)
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:End"|1|"N"|"E"|
1105366|||||||||||||"GVF2I2B2"| (end position)
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SO:0001029"|1|"T"|
"+"|||||||||||||"GVF2I2B2"| (+ strand)
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:Zygoty"|1|"T"|
"heterozygous"|||||||||||||"GVF2I2B2"| (heterozygous)
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SEQ:HUGO"|1|"T"|
"TTLL10"|||||||||||||"GVF2I2B2"| (associated gene)
- 1880001024|1000000024|"SO:0001483"|"@"|"2010-03-03 00:00:00"|"SO:0001791"|1||
|||||||||||||"GVF2I2B2"| (exonic variant)

Import NGS Variant Data

Analysis details

Information related to the NGS data

Specify input file:

Input file format:

VCF mapping file:

I2B2 Patient number:

I2B2 Encounter number:

Date of encounter:

Reference genome version:

Progress Bar:

Sample details

Information related to the sample

Sample ID:

Sample Type:

Anatomical Source:

Collection Method:

Additive:

Sample Pathology

Information related to the sample pathology

Pathology:

Tumor Grade:

Tumor Stage:

Import NGS Variant Data

Analysis details

Information related to the NGS data

Specify input file:

Browse

Sample details

Information related to the sample

Sample ID:

Sample Type:

TISSUE

Open

Organize New folder

- Libraries
 - Documents
 - Music
 - Pictures
 - Videos

- Computer
 - Local Disk (C:)
 - SMART (\\phsinfra16) (W:)
 - Documentation (\\infra1.partners.org) (X:)
 - build (\\infra1.partners.org) (Y:)

Name	Date modified	Type
CEU.trio.2010_07.indel.genotypes	2/20/2013 10:19 AM	vCa
CEU.trio.2010_07.indel.genotypes.vcf.exonic_variant_fu...	2/20/2013 10:23 AM	EXC
CEU.trio.2010_07.indel.genotypes.vcf.invalid_input	2/20/2013 10:23 AM	INV
CEU.trio.2010_07.indel.genotypes.vcf	2/20/2013 10:23 AM	LOI
CEU.trio.2010_07.indel.genotypes.vcf.variant_function	2/20/2013 10:23 AM	VAI
CEU.trio.2010_07.map	5/2013 10:24 AM	Tex
NA12878.1880003090	2013 2:57 PM	I2B.
NA12878.gvf	2013 2:51 PM	GVI
NA12878.gvf.i2b2log	6/4/2013 2:57 PM	I2B.
NA12891.1880003093	6/4/2013 2:58 PM	I2B.

Type: VARIANT_FUNCTION File
 Size: 66.2 MB
 Date modified: 2/20/2013 10:23 AM

File name: CEU.trio.2010_07.indel.genotypes.vcf.variant_function

**

Open Cancel

Import NGS Variant Data

Analysis details

Information related to the NGS data

Specify input file:

nes\CEUExon\ANNOVAR\CEU.exon.20

Browse

Input file format:

VCF

VCF-ANNOVAR

VCF

GVF

I2B2

VCF mapping file:

I2B2 Patient number:

I2B2 Encounter number:

Date of encounter:

Reference genome version:

hg18

Submit

Progress Bar:

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

Import NGS Variant Data

Analysis details

Information related to the NGS data

Specify input file: nes\CEUExon\ANNOVAR\CEU.exon.20

Input file format: VCF-ANNOVAR

VCF mapping file: ANNOVAR\CEU.exon.2010_03.map.txt

I2B2 Patient number:

I2B2 Encounter number:

Date of encounter:

Reference genome version: hg18

Progress Bar:

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

Mapping file

```
##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:

Input file format:

VCF mapping file:

I2B2 Patient number:

I2B2 Encounter number:

Date of encounter:

Reference genome version:

Progress Bar:

VCF ANNOVAR to GVF step: Converting VCF line 8000

Sample details

Information related to the sample

Sample ID:

Sample Type:

Anatomical Source:

Collection Method:

Additive:

Sample Pathology

Information related to the sample pathology

Pathology:

Tumor Grade:

Tumor Stage:

Import NGS Variant Data

Analysis details

Information related to the NGS data

Specify input file:

Input file format:

VCF mapping file:

I2B2 Patient number:

I2B2 Encounter number:

Date of encounter:

Reference genome version:

Progress Bar: 

VCF ANNOVAR to GVF step: VCF-ANNOVAR to GVF complete

NA12878.gvf to i2b2 step: Converting GVF line 4000

Sample details

Information related to the sample

Sample ID:

Sample Type:

Anatomical Source:

Collection Method:

Additive:

Sample Pathology

Information related to the sample pathology

Pathology:

Tumor Grade:

Tumor Stage:

Import NGS Variant Data

Analysis details

Information related to the NGS data

Specify input file:

Input file format:

VCF mapping file:

I2B2 Patient number:

I2B2 Encounter number:

Date of encounter:

Reference genome version:

Progress Bar:

VCF ANNOVAR to GVF step: VCF-ANNOVAR to GVF complete

GVF to i2b2 step: All GVF to i2b2 complete

Sample details

Information related to the sample

Sample ID:

Sample Type:

Anatomical Source:

Collection Method:

Additive:

Sample Pathology

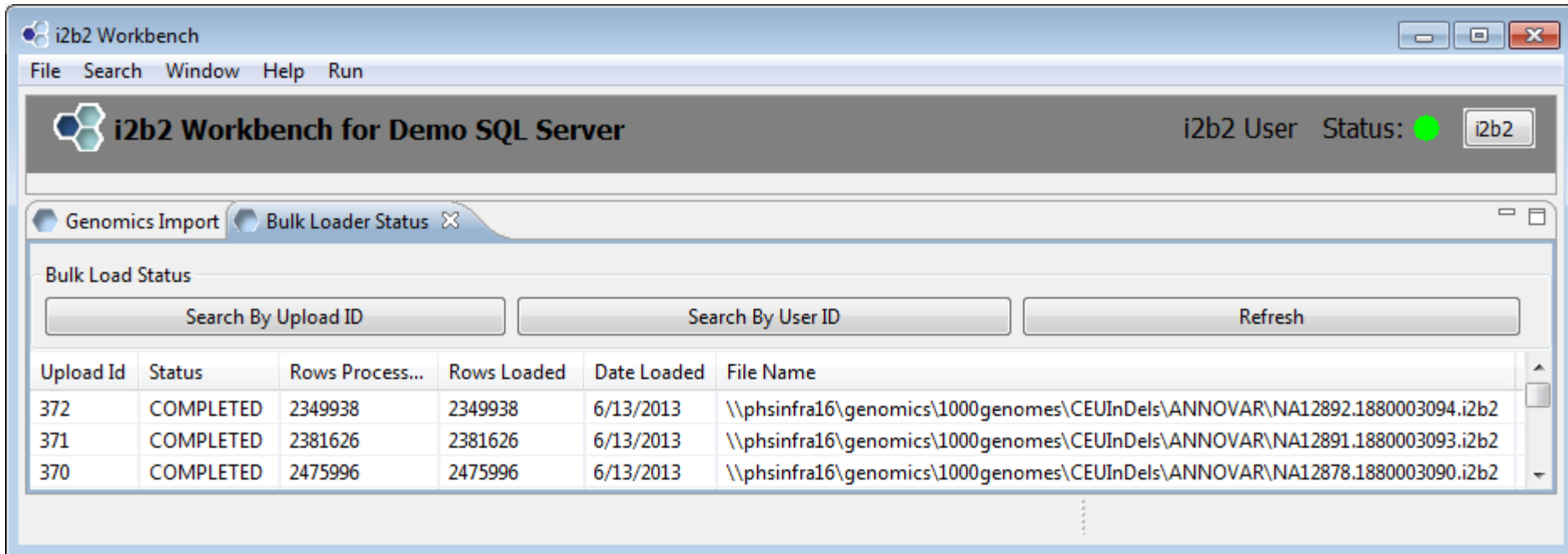
Information related to the sample pathology

Pathology:

Tumor Grade:

Tumor Stage:

Bulk Loader Status



The screenshot shows the i2b2 Workbench interface. The main window title is "i2b2 Workbench" and the application title is "i2b2 Workbench for Demo SQL Server". The user is logged in as "i2b2 User" with a green status indicator. The "Bulk Loader Status" window is active, displaying a table of upload records. The table has columns for Upload Id, Status, Rows Processed, Rows Loaded, Date Loaded, and File Name. Three records are shown, all with a status of "COMPLETED".

Bulk Load Status

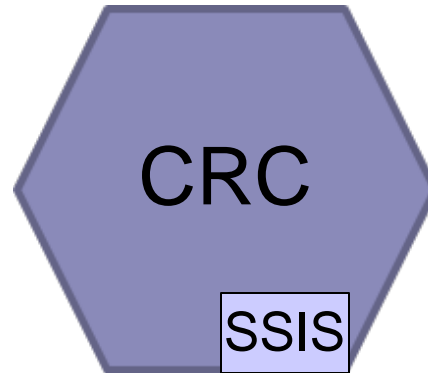
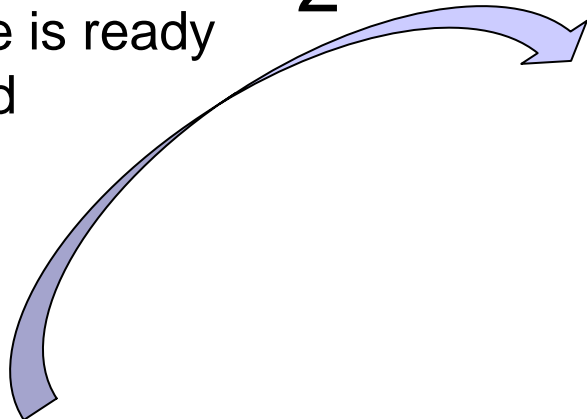
Search By Upload ID Search By User ID Refresh

Upload Id	Status	Rows Process...	Rows Loaded	Date Loaded	File Name
372	COMPLETED	2349938	2349938	6/13/2013	\\phsinfra16\genomics\1000genomes\CEUInDels\ANNOVAR\NA12892.1880003094.i2b2
371	COMPLETED	2381626	2381626	6/13/2013	\\phsinfra16\genomics\1000genomes\CEUInDels\ANNOVAR\NA12891.1880003093.i2b2
370	COMPLETED	2475996	2475996	6/13/2013	\\phsinfra16\genomics\1000genomes\CEUInDels\ANNOVAR\NA12878.1880003090.i2b2

Bulk Loading Observations

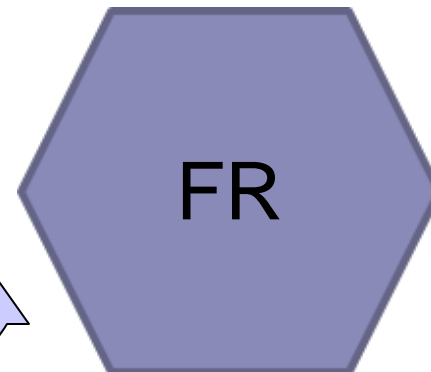
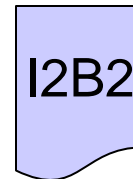
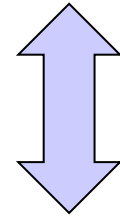
2. Tell the CRC
the file is ready
to load

2

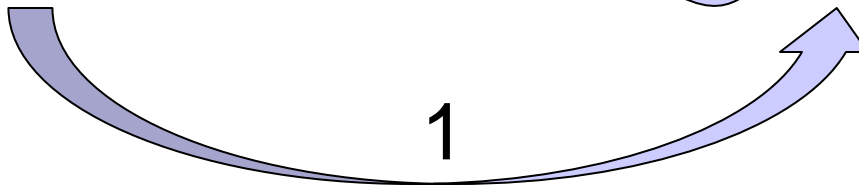


3. SSIS package
loads the i2b2 file to
observation_fact table

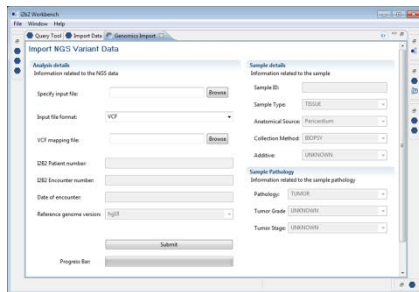
3



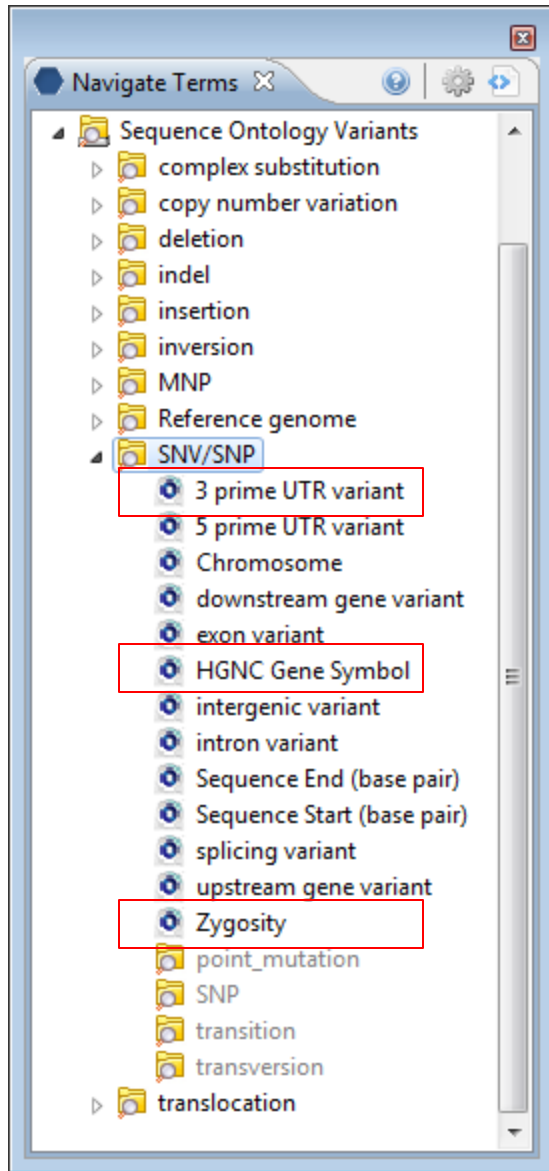
1



1. Send the i2b2 file to the FR



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 zygoty

Display Settings: Abstract

Send to:



Nat Genet. 2000 Sep;26(1):76-80.

The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES.

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]

Publication Types, MeSH Terms, Substances



LinkOut - more resources



Display Settings: Abstract

Send to:



Nat Genet. 2000 Sep;26(1):76-80.

The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES.

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]

Publication Types, MeSH Terms, Substances



LinkOut - more resources



Gene Association Modifier

The screenshot displays the i2b2 Workbench interface for a Demo SQL Server. The main window is titled "i2b2 Workbench for Demo SQL Server" and shows a "Query Tool" with three groups (Group 1, Group 2, Group 3) for defining search criteria. A dialog box titled "Choose modifier value of SNV/SNP" is open, allowing users to select a search modifier for the term "SNV/SNP".

Query Tool Configuration:

Group 1	Group 2	Group 3
Dates	Dates	Dates
Occurs > 0x	Occurs > 0x	Occurs > 0x
Exclude	Exclude	Exclude
Treat Independently	Treat Independently	Treat Independently
SNV/SNP [HGNC Gene Symbol]		

Analysis Types:

- Patient list
- Event list
- Number of patients
- Gender patient breakdown
- Vital Status patient break
- Race patient breakdown
- Age patient breakdown
- TimeLine

Query Timing:

- Treat all groups independ
- Selected groups occur in
- Items instance will be sar

Choose modifier value of SNV/SNP Dialog:

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

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

Containing []

Buttons: OK, Cancel, Gene Assist

Input File Information:

Specify input file: els\ANNOVAR\CEU.trio.2010_07.inc

Input file format: VCF-ANNOVAR

Sample Type: TISSUE

Anatomical Source: ericardium

Quick Gene Search

Search symbols, keywords or IDs for:

Results that equal begin contain

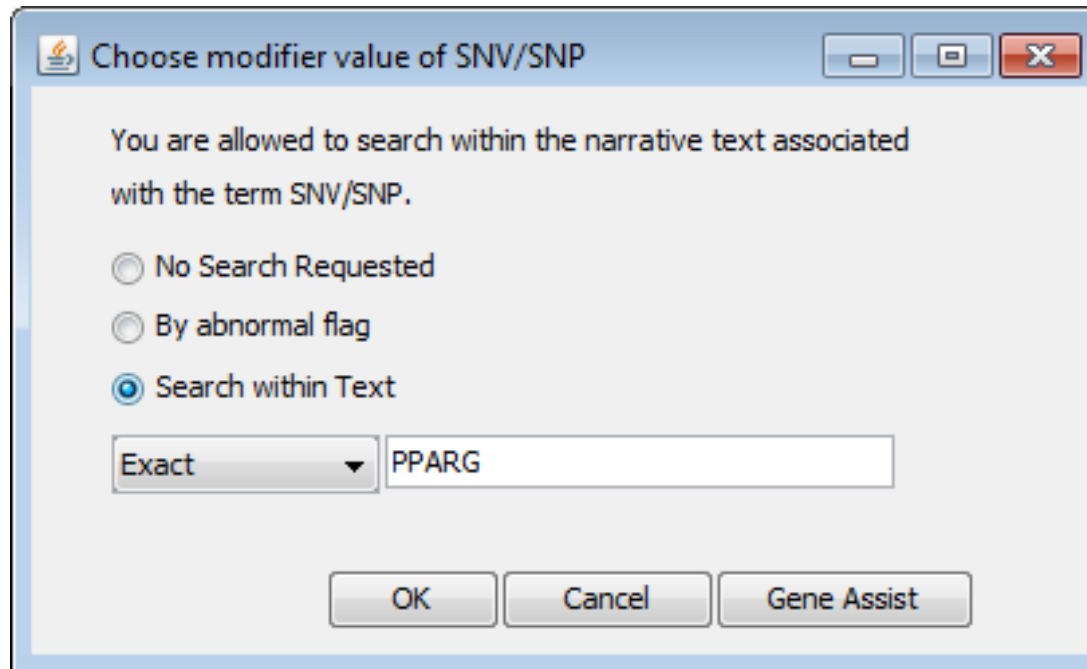
Display hits



Total hits: 11

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

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

OK Cancel Gene Assist

Building a Translational Genomic Query

- Group1: SNV/SNP with HGNC Gene Symbol modifier of “PPARG”

The screenshot displays the 'Query Tool' interface. At the top, there is a 'Query Name:' field. Below it, three query groups are defined:

- Group 1:** Contains the query 'Gene Symbol [LIKE[exact] "PPARG"]'. It has buttons for 'Dates', 'Occurs > 0x', and 'Exclude'. A dropdown menu is set to 'Treat Independently'. Below the query field, a scrollable area contains the text: 'The terms of this group are joined then intersected with other groups'.
- Group 2:** Contains an empty query field. It has buttons for 'Dates', 'Occurs > 0x', and 'Exclude'. A dropdown menu is set to 'Treat Independently'. Below the field, a scrollable area contains the text: 'Drag terms from Navigate, Find and Workplace into this group'.
- Group 3:** Contains an empty query field. It has buttons for 'Dates', 'Occurs > 0x', and 'Exclude'. A dropdown menu is set to 'Treat Independently'. Below the field, a scrollable area contains the text: 'Drag terms from Navigate, Find and Workplace into this group'.

At the bottom left, there is a checkbox labeled 'Get Everyone'. In the center, there is a 'Run Query Above' button. At the bottom right, there is a field labeled 'Patient(s) returned:'.

On the right side of the interface, there is a panel titled 'Analysis Types' with the following options:

- Patient list
- Event list
- Number of patients
- Gender patient breakdown
- Vital Status patient breakdown
- Race patient breakdown
- Age patient breakdown
- TimeLine

Below this panel is a section titled 'Query Timing' with the following options:

- Treat all groups independent
- Selected groups occur in the
- Items instance will be same

Display Settings: Abstract

Send to:



Nat Genet. 2000 Sep;26(1):76-80.

The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES.

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]

Publication Types, MeSH Terms, Substances

LinkOut - more resources

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

The screenshot displays the 'Query Tool' interface with the following configuration:

- Query Name:** (Empty field)
- Group 1:** Contains 'Gene Symbol [LIKE[exact] "PPARG"]'. The 'Items instance will be same' dropdown is selected.
- Group 2:** Contains 'SNV/SNP [exon variant]'. The 'Items instance will be same' dropdown is selected.
- Group 3:** (Empty)
- Analysis Types:** Includes 'Number of patients' (checked), 'TimeLine' (checked), and several other options like 'Patient list', 'Event list', etc., which are unchecked.
- Query Timing:** Includes 'Items instance will be same' (selected), 'Treat all groups independent', and 'Selected groups occur in the'.
- Buttons:** 'Get Everyone', 'Run Query Above', and 'Patient(s) returned:'.

Building a Translational Genomic Query

- Group 3: Diabetes Mellitus
 - Select “Treat Independently” for this panel

The screenshot shows the 'Query Tool' interface with three groups defined:

- Group 1:** Contains the term 'Gene Symbol [LIKE[exact] "PPARG"]'. The 'Items instance will be same' dropdown is selected.
- Group 2:** Contains the term 'SNV/SNP [exon variant]'. The 'Items instance will be same' dropdown is selected.
- Group 3:** Contains the term 'Diabetes mellitus'. The 'Treat Independently' dropdown is selected.

On the right side, the 'Analysis Types' panel is visible with the following options:

- Patient list
- Event list
- Number of patients
- Gender patient breakdown
- Vital Status patient breakdown
- Race patient breakdown
- Age patient breakdown
- TimeLine

The 'Query Timing' panel at the bottom right shows:

- Treat all groups independent
- Selected groups occur in the
- Items instance will be same

At the bottom of the window, there is a 'Get Everyone' checkbox, a 'Run Query Above' button, and a 'Patient(s) returned:' field.

Run the query

Query Name: SNV/S-SNV/S-Diabe@11:55:23

Group 1	Group 2	Group 3
Dates Occurs > 0x Exclude	Dates Occurs > 0x Exclude	Dates Occurs > 0x Exclude
Items instance will be same	Items instance will be same	Treat Independently
Gene Symbol [LIKE[exact] "PPARG"]	SNV/SNP [exon variant]	Diabetes mellitus
The terms of this group are joined then intersected with other groups	The terms of this group are joined then intersected with other groups	The terms of this group are joined then intersected with other groups

Add Group

Analysis Types

- Patient list
- Event list
- Number of patients
- Gender patient breakdown
- Vital Status patient breakdown
- Race patient breakdown
- Age patient breakdown

Query Timing

- Treat all groups independent
- Selected groups occur in
- Items instance will be same

Get Everyone

Run Query Above

Patient(s) returned: 7

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](http://ncbo.org)) that is associated with GVF format can be used to query the next generation sequencing data that was imported into i2b2.