Making NGS Data Analysis Clinically Practical: Repeatable and Time-Effective Workflows

Autumn Laughbaum, Biostatistician
with introduction by Dr. Andreas Scherer, President & CEO
Questions during the presentation

Use the Questions pane in your GoToWebinar window
Introduction

Researcher

Clinician
Hands-on Time Savings

- **2 weeks**: 1 trio using Excel
- **2 hours**: 100 trios using SVS
- **2 minutes**: Unlimited trios using automated workflow
Today’s Agenda

1. Status Quo
   - Moving from Excel > SVS > Automated Workflows
   - General sequencing workflow

2. Example Workflow: Ogden Syndrome

3. Example Workflow: Trio Analysis

4. Discussion
Analyzing Data from a secondary analysis pipeline: Using Excel
Analyzing Data from a secondary analysis pipeline: Using SVS
Analyzing Data from a secondary analysis pipeline: Automated Workflow

1) Import VCF File(s) including:
   - Genotypes
   - Read Depths
   - Genotype Qualities
2) Filter to Exon Regions using the following options:
   - Gene Tracks RefSeqGenes-UCSC_GRCh37_Homo_sapiens.idf.
   - Expand regions by 10 bp
3) Annotate and filter based on additional VCF fields:
   - Drop if Read Depth < 10.0
   - Drop if Genotype Quality < 20.0
4) Annotate based on Genomic Super Dups, UCSC
5) Annotate and filter based on dbNSFP_NS_Functional_Predictions-v2.0-2013-03-22-GRCh37_Homo_sapiens.idf
6) Perform Variant Classification:
   - Activate only variants with the following classifications:
     - u'Frameshift Sub', u'Stoploss', u'Splicing', u'Init Codon', u'Ins', u'Nonsyn SNV', u'Stopgain', u'Del', u'Frameshift Del', u'Frameshift Ins', u'Sub'
7) Filter to de Novo Variants [ENDPOINT 1]
8) Annotate and filter based on NHLBI ESP6500SV-V2 Exomes - Variant Frequencies 2013_03, GRCh37:
   - Inactivate if European American MAF > 0.01
9) Filter to Variants that fit Recessive Inheritance Pattern [ENDPOINT 2]
10) Filter to Variants that fit Specific Inheritance Pattern:
    - NA12878 in [All Ref]
11) Score Compound Heterozygous Variants [ENDPOINT 3]
Analyzing Sequencing Data

Data from secondary analysis pipeline (VCF) – 2 million variants

- Filter to coding regions
- Filter on read depth & quality score
- Filter based on population frequencies
- Non-synonymous variants
- Functional prediction
- Inheritance pattern

- 40,000 variants
- 35,000 variants
- 30,000 variants
- 10,000 variants
- 3,000 variants
- 2 variants
Workflow Example – Ogden Syndrome
SVS Demo
Ogden Syndrome

5 Samples
107,000 variants

- Filter based on VCF quality metrics
- Filter based on population frequencies
- Annotate based on functional predictions
- Filter based on variant classification
- Filter based on inheritance pattern

1 damaging variant

Ref_Alt

Alt
Workflow Example – Trio Analysis
Trio Analysis – 3 Distinct Mutation types

- de Novo Mutations
- Rare Recessive Mutations
- Compound Heterozygous Mutations

**Additional Annotations**

- Find de Novo Variants
- Filter on Variant Frequencies
- Heterozygous Child
- Rare Recessive Mutations
de Novo Mutation

Diagram illustrating genetic relationships: Father (Ref_Ref) to Child (Ref_Alt) through Mother (Ref_Ref)
Compound Heterozygous Mutation

Within a Gene

- Ref_Alt
- Ref_Ref
- Ref_Alt
- Ref_Alt
- Ref_Ref
- Ref_Alt
- Ref_Ref
- Ref_Alt
This tool will automate a workflow with the following steps:

1) Import VCF File(s) including
   - Genotypes
   - Read Depths
   - Genotype Qualities

2) Filter to Exon Regions using the following options:
   * Gene Track=RefSeqGenes-UCSC_GRCh_37_Homo_sapiens.idf:1
   * Expand regions by 10bp

3) Annotate and Filter based on additional VCF fields:
   * Drop if Read Depth <= 10.0
   * Drop if Genotype Quality <= 20.0

4) Annotate based on Genomic Super Dups, UCSC

5) Annotate and Filter based on dbNSFP_NS_Functional_Predictions-v2.0-2013-03-22-GRCH_GRCh_37_Homo_sapiens.idf

6) Perform Variant Classification
   * Activate only variants with the following classifications:
     - [u'Frameshift Del', u'Stoploss', u'Splicing', u'Init Codon', u'Ins', u'Nonsyn SNV', u'Stopgain', u'Del', u'Frameshift Del', u'Frameshift Ins', u'Sub']

7) Filter to de Novo Variants [ENDPOINT 1]

8) Annotate and Filter based on NHLBI ESP6500SI-V2 Exomes - Variant Frequencies 2013_03, GHI
   * Inactivate if European American MAF >= 0.01

9) Filter to Variants that fit Recessive Inheritance Pattern [ENDPOINT 2]

10) Filter to Variants that fit Specific Inheritance Pattern
    * NA12878 in ['Alt_Ref']

11) Score Compound Heterozygous Variants [ENDPOINT 3]
Trio Analysis – Automated Workflow

Import Sorted VCF Files -- Select Files

- Pedigree Spreadsheet:
  - Pedigree Data - Sheet 1

Base Dataset Name: SVS_CEP trio

Input VCF files(s):
- NA12878.mapped.ILLUMINA.bwa.CEU.exome.2...
- NA12891.mapped.ILLUMINA.bwa.CEU.exome.2...
- NA12892.mapped.ILLUMINA.bwa.CEU.exome.2...

Buttons:
- Add Files
- Add Directory
- Remove
- Start Workflow
- Cancel
- Help

Golden Helix
Accelerating the Quest for Significance™
Trio Analysis – Automated Workflow

- 2.3 Million
  - 38K
  - 36K
  - 30K
  - 10K
    - de Novo: 2
    - Compound Heterozygous: 85
    - Rare Recessive: 12

SVS Information:

This tool will automate a workflow with the following steps:

1) Import VCF File(s) including
   - Genotypes
   - Read Depths
   - Genotype Qualities

2) Filter to Exon Regions using the following options:
   - Gene Track=RefSeqGenes-UCSC_GRCh_37_Homo_sapiens.idf
   - Expand regions by 10bp

3) Annotate and Filter based on additional VCF fields:
   - Drop if Read Depth <= 10.0
   - Drop if Genotype Quality <= 20.0

4) Annotate based on Genomic Super Dups, UCSC

5) Annotate and Filter based on dbNSFP_NS_Functional_Predictions-
v2.0-2013-03-22-GRCh_37_Homo_sapiens.idf

6) Perform Variant Classification
   - Activate only variants with the following classifications:
     - [u'Frameshift Sub', u'Stoploss', u'Splicing', u'Init Codon', u'Ins', u'Nonsyn SNV', u'Stopgain', u'Del', u'Frameshift Del', u'Frameshift Ins', u'Sub']

7) Filter to de Novo Variants [ENDPOINT 1]

8) Annotate and Filter based on NHLBI ESP650CS-V2 Exomes - Variant Frequencies 2013_03, GHI
   - Inactivate if European American MAF > 0.01

9) Filter to Variants that fit Recessive Inheritance Pattern [ENDPOINT 2]

10) Filter to Variants that fit Specific Inheritance Pattern
    - NA28978 in "[Alt,Ref]"

11) Score Compound Heterozygous Variants [ENDPOINT 3]
Golden Helix SVS Workflow Automation

Results from automated workflow "NGS Trio Complete Workflow" in project SVS_CEPHtrio at 2013-09-04 15:51:57

Final deNovo Variant Count: 2
*17:18647625-SNV at 17:18647625, Ref/Alt=T/A, gene=FBXW10, transcript=NM_031456, p.Ile23Asn

Final Rare Recessive Variant Count: 12
*1:52306064-Del at 1:52306064, Ref/Alt=TCT/-, gene=NRD1, transcript=NM_001101662, NM_001242361, NM_002525, p.Glu22del
*1:145302704-SNV at 1:145302704, Ref/Alt=A/G, gene=NBPF10, transcript=NM_001039703, p.Lys381Arg
*3:190106072-Del at 3:190106072, Ref/Alt=G/-, gene=CLDN16, transcript=NM_006580, p.Ala56fs
*3:190106074-SNV at 3:190106074, Ref/Alt=G/C, gene=CLDN16, transcript=NM_006580, p.Ala56Pro
*5:139931630-Ins at 5:139931630, Ref/Alt=-/G, gene=SRA1, transcript=NM_001035235, p.Val110fs
*5:140531592-SNV at 5:140531592, Ref/Alt=C/T, gene=PCDH6, transcript=NM_018939, p.Thr585lle
*5:140563579-SNV at 5:140563579, Ref/Alt=C/T, gene=PCDH16, transcript=NM_020957, p.Thr482lle
*5:140564088-SNV at 5:140564088, Ref/Alt=C/T, gene=PCDH16, transcript=NM_020957, p.Arg652Cys
*5:140581220-SNV at 5:140581220, Ref/Alt=C/A, gene=PCDH11, transcript=NM_018931, p.Arg625Ser
*12:10573094-SNV at 12:10573094, Ref/Alt=C/G, gene=KLRC3, transcript=NM_002261, NM_007333, p.Trp129Ser
*15:43925134-SNV at 15:43925134, Ref/Alt=T/A, gene=CATSPER2, transcript=NM_172095, ?
*17:18565350-SNV at 17:18565350, Ref/Alt=G/C, gene=ZNF286B, transcript=NM_001145045, p.Thr490Ser

Final Compound Het Variant Count: 85
*1:1423281-SNV at 1:1423281, Ref/Alt=G/A, gene=ATAD3B, transcript=NM_031921, p.Arg418Gln
*1:65146995-SNV at 1:65146995, Ref/Alt=A/C, gene=CACHD1, transcript=NM_020925, p.Asp1103Ala
*1:65157120-SNV at 1:65157120, Ref/Alt=C/G, gene=CACHD1, transcript=NM_020925, p.Pro1183Arg
*1:70460304-SNV at 1:70460304, Ref/Alt=A/G, gene=LRRC7, transcript=NM_020794, p.Asn293Ser
*1:70504932-SNV at 1:70504932, Ref/Alt=C/A, gene=LRRC7, transcript=NM_020794, p.Ala1104Glu
Automated Workflow Results
Additional Annotations
Example in GenomeBrowse
Discussion

WORKFLOW EXAMPLE – TRIO ANALYSIS
<table>
<thead>
<tr>
<th>Clinicians</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Running well-defined workflow on additional samples</td>
<td>- Building and testing workflows</td>
</tr>
<tr>
<td>- Minimum user-interface knowledge</td>
<td>- More complex but intuitive interface</td>
</tr>
<tr>
<td>- Small learning curve</td>
<td>- Larger learning curve</td>
</tr>
<tr>
<td>- Limited hands-on time</td>
<td>- Power to investigate and manipulate data</td>
</tr>
</tbody>
</table>
Conclusion – The GHI Approach

Work closely with clients to learn about needs and develop workflow specification

Create document and workflow diagram outlining specifications and requirements

Build automated-workflow prototype

Thorough internal testing and complete documentation

Finished product works seamlessly within SVS
Questions during the presentation

Use the Questions pane in your GoToWebinar window