Tutorial

Batching of Multi-Input Workflows

November 21, 2017

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

Batching of Multi-Input Workflows

This tutorial will demonstrate how workflows that have more than one variable input can be run in batch mode. This is the case for many Ready-to-Use workflows available in the toolbox of Biomedical Genomics Workbench, including all workflows that analyze a trio or a family of fours and workflows that compare tumor to normal samples. In this tutorial we will use the "Identify Causal Inherited Variants In Trio" workflow to illustrate how multiple families' genomes can be analyzed in batches. Note that you must be working with the Biomedical Genomics Workbench 3.0 or higher to be able to batch workflows with multiples inputs.

Download and import data

To allow you to run this tutorial within minutes, we have generated for each family member a subset of whole exome data containing only chromosome 21. We also introduced a clinically relevant variant in the proband and one of its parent. The Biomedical Genomics Workbench offers a reference data set specific to chromosome 21 and available from the Data Manager.

The data set you will download includes:

- The reads from the father, mother and proband for each of the two families A and B.
- A target regions file called *SeqCap_EZ_Exome_v2_BED_CHR21*. Please note that when working with your own data, you can obtain the relevant target region tracks from the vendor of the amplicon or hybridization kit.
- An Excel spreadsheet with information about the samples, such as which reads belong to which family, and which family members are affected by a particular disease.

Go through the following steps to download and import the data into the Workbench.

1. Download the sample data from our website: `http://resources.qiagenbioinformatics.com//testdata/batching_workflows_tutorial.zip`.
2. Unzip the zip file on your local machine.
3. Start the *Biomedical Genomics Workbench*.
4. Launch **File** | **Import** (📥) | **Standard Import** (📥) and choose the subfolder called "reads and target region" that contains the sequencing reads and the target sequence via the toolbar. Leave the import type set to Automatic and save the imported data in a folder you create in the Navigation Area called "tutorial".
5. You do not need to import the Excel file at this point.

Once the data has been downloaded and imported, you should see the folder and files in the Navigation Area as shown in figure 1.

Data management configuration

As the data used in this tutorial is from chromosome 21 only, you will need to specify a Reference Data Set that contains only chromosome 21. This Reference Data Set can be found and selected in the Data Manager.

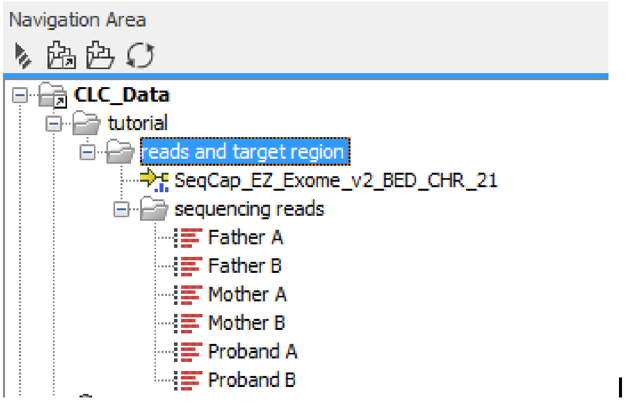
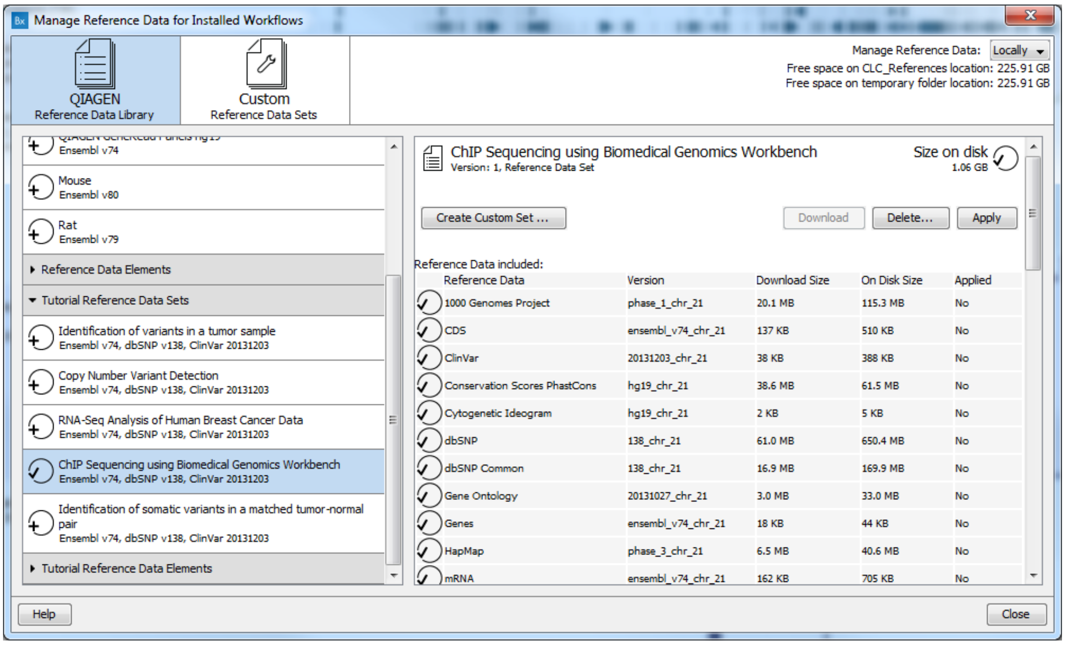


Figure 1: *The data set has been imported.*

1. Click on the button labeled **Data Management** ( ) in the top right corner of the Workbench.
2. In the top right corner of the Data Manager, you can see whether you are managing the reference data locally or on a server (if you have one). Make sure you are downloading the Reference Data Set in the location where you will run the workflow, i.e., "locally" if you want to run the workflow on your machine, and "on the server" if you intend to run the analyses on a server or on a grid.
3. Click on Tutorial Reference Data Sets and choose "Batching of Multi-Input Workflows".
4. In the new window (figure 2) click on the **Download** button to import in your CLC_References folder the elements that were not yet downloaded (marked with an (+) icon). After downloading, check that all elements are marked with an (✓) icon.



Manage Reference Data: Locally
Free space on CLC_References location: 225.91 GB
Free space on temporary folder location: 225.91 GB

Ensembl v74

Mouse
Ensembl v80

Rat
Ensembl v79

Reference Data Elements

Tutorial Reference Data Sets

Identification of variants in a tumor sample
Ensembl v74, dbSNP v138, ClinVar 20131203

Copy Number Variant Detection
Ensembl v74, dbSNP v138, ClinVar 20131203

RNA-Seq Analysis of Human Breast Cancer Data
Ensembl v74, dbSNP v138, ClinVar 20131203

ChIP Sequencing using Biomedical Genomics Workbench
Ensembl v74, dbSNP v138, ClinVar 20131203

Identification of somatic variants in a matched tumor-normal pair
Ensembl v74, dbSNP v138, ClinVar 20131203

Tutorial Reference Data Elements

ChIP Sequencing using Biomedical Genomics Workbench
Version: 1, Reference Data Set

Size on disk
1.06 GB

Reference Data included:

	Reference Data	Version	Download Size	On Disk Size	Applied
✓	1000 Genomes Project	phase_1_chr_21	20.1 MB	115.3 MB	No
✓	CDS	ensembl_v74_chr_21	137 KB	510 KB	No
✓	ClinVar	20131203_chr_21	38 KB	388 KB	No
✓	Conservation Scores PhastCons	hg19_chr_21	38.6 MB	61.5 MB	No
✓	Cytogenetic Ideogram	hg19_chr_21	2 KB	5 KB	No
✓	dbSNP	138_chr_21	61.0 MB	650.4 MB	No
✓	dbSNP Common	138_chr_21	16.9 MB	169.9 MB	No
✓	Gene Ontology	20131027_chr_21	3.0 MB	33.0 MB	No
✓	Genes	ensembl_v74_chr_21	18 KB	44 KB	No
✓	HapMap	phase_3_chr_21	6.5 MB	40.6 MB	No
✓	mRNA	ensembl_v74_chr_21	162 KB	705 KB	No

Figure 2: *Apply the Reference Data Set called "Batching of Multi-Input Workflows" in the Data managemet.*

5. Once all elements have been downloaded, click the **Apply** button in the top right corner, and then click **Close** to exit the Data Manager.

Note: remember to change back the set of references applied once you are done with this tutorial by following the instructions given at the end of this tutorial.

Running a Ready-to-Use workflow in batch mode

In this tutorial we will analyze two families at once using the batch option using a workflow containing a Trio. The reads for each family member are now available in the Navigation Area, but to be able to run both families at the same time, you need to provide additional information (organized in an Excel spreadsheet) about the origin and the grouping of the reads: which family member do the samples belong to, who is affected by the disease, etc.

To find the workflow, go to:

> **Toolbox** | **Ready-to-Use Workflows** | **Whole Exome Sequencing** ( ) | **Hereditary Disease** ( ) | **Identify Casual Inherited Variants In Trio (WES)**

1. Right click on the **Identify Casual Inherited Variants In Trio (WES)** tool. This will open a menu as shown on figure 3. Choose "Run in Batch Mode".

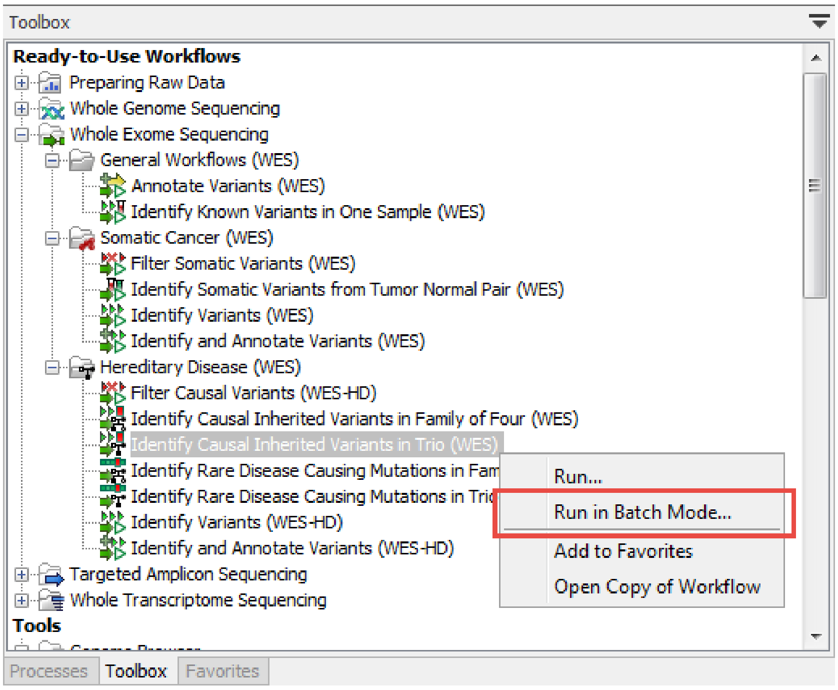
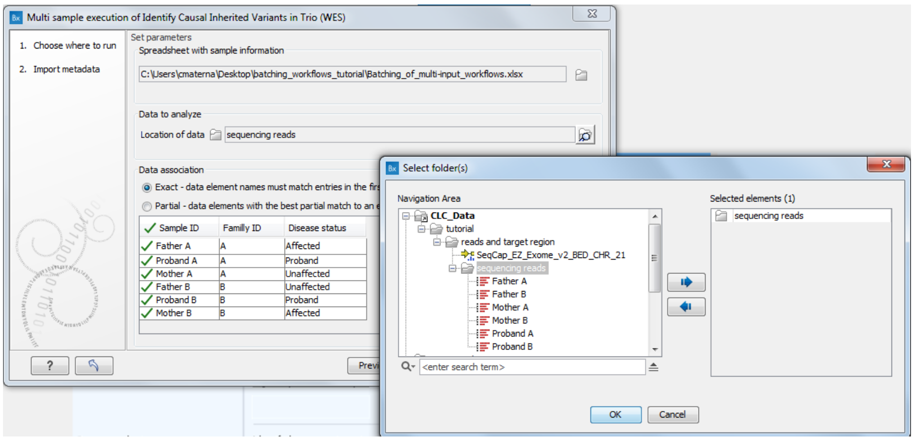

Toolbox
Ready-to-Use Workflows
Preparing Raw Data
Whole Genome Sequencing
Whole Exome Sequencing
General Workflows (WES)
Annotate Variants (WES)
Identify Known Variants in One Sample (WES)
Somatic Cancer (WES)
Filter Somatic Variants (WES)
Identify Somatic Variants from Tumor Normal Pair (WES)
Identify Variants (WES)
Identify and Annotate Variants (WES)
Hereditary Disease (WES)
Filter Causal Variants (WES-HD)
Identify Causal Inherited Variants in Family of Four (WES)
Identify Causal Inherited Variants in Trio (WES)
Identify Rare Disease Causing Mutations in Fam
Identify Rare Disease Causing Mutations in Tri
Identify Variants (WES-HD)
Identify and Annotate Variants (WES-HD)
Targeted Amplicon Sequencing
Whole Transcriptome Sequencing
Tools
Run...
Run in Batch Mode...
Add to Favorites
Open Copy of Workflow
Processes
Toolbox
Favorites

Figure 3: *Right clickling the Identify Casual Inherited Variants In Trio (WES) tool will open a window giving the posibility to choose Run in Batch Mode.*

2. Depending on your local setup you might be asked to select where to run the workflow. Choose the appropriate option for you and click **Next**.

3. In the window called "Import metadata":

   - In the "Spreadsheet with sample information" section, select the Excel spreadsheet with the sample information that you downloaded on your computer earlier.
   - In the "Data to analyze" section select the folder containing the reads for the two families as shown on figure **??**. Note that to analyze multiple families in one batch run, the reads must be stored together in one or several folders, as it is only possible to select folders, and not individual files.
   - In the "Data association" section, all samples should now be marked with a green check mark and the wizard window should look like figure 4. Click **Next**.



Multi sample execution of Identify Causal Inherited Variants in Trio (WES)

1. Choose where to run
2. Import metadata

Set parameters

Spreadsheet with sample information

C:\Users\cmaterna\Desktop\batching_workflows_tutorial\Batching_of_multi-input_workflows.xlsx

Data to analyze

Location of data: sequencing reads

Data association

● Exact - data element names must match entries in the firs

○ Partial - data elements with the best partial match to an e

✓ Sample ID	Familly ID	Disease status
✓ Father A	A	Affected
✓ Proband A	A	Proband
✓ Mother A	A	Unaffected
✓ Father B	B	Unaffected
✓ Proband B	B	Proband
✓ Mother B	B	Affected

Select folder(s)

Navigation Area

```
CLC_Data
  tutorial
    reads and target region
      SeqCap_EZ_Exome_v2_BED_CHR_21
      sequencing reads
        Father A
        Father B
        Mother A
        Mother B
        Proband A
        Proband B
```

Selected elements (1)

sequencing reads

Figure 4: *The Import metadata wizard window after the spreadsheet and the reads were specified and correctly associated.*

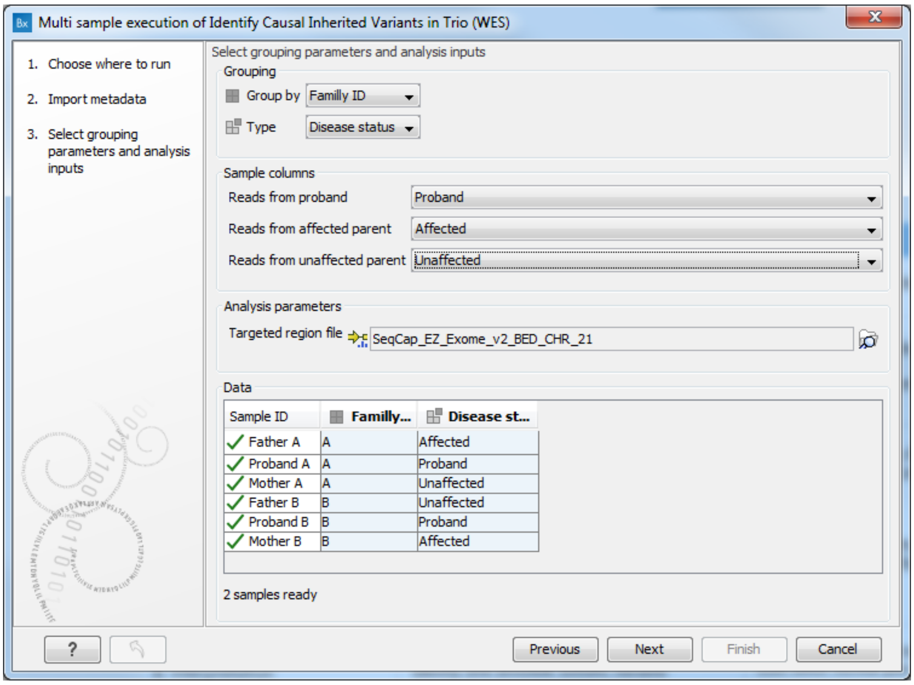
4. In the next wizard window (figure 5):

   - In the "Select grouping parameters and analysis inputs" section, choose for "Group by" the option **Family ID** from the dropdown menu, as we are here grouping the samples by families (one batch is one family). For the "Type", choose the option **Disease status**, as the "Type" indicates who is the affected parent, the unaffected parent and the proband.
   - In the "Sample columns" section, select successively "Proband", "Affected" and "Unaffected" in the relevant fields.
   - Finally select *SeqCap_EZ_Exome_v2_BED_CHR21* as target region. The wizard window should now look like figure 5.

   Click **Next**.

5. In the "Fixed Ploidy Variant Detection" windows, keep "Required variant probability" and "Minimum frequency" at their default values of 50.0 and 10.0 , but set **Minimum coverage** and **Minimum count** to 40 for all family members independently, i.e, in the three successive windows specific to proband, affected parent and unaffected parent (figure 6). We usually recommend to use default settings but have chosen here more stringent parameters for demonstration purposes.

6. In the "QC for Target Sequence" wizard window leave the parameters as default with Minimum coverage set to 30 and leave the options "Ignore non-specific matches" and "Ignore broken pairs" unchecked (figure 7).



Multi sample execution of Identify Causal Inherited Variants in Trio (WES)

1. Choose where to run
2. Import metadata
3. Select grouping parameters and analysis inputs

Select grouping parameters and analysis inputs

Grouping

Group by: Familly ID

Type: Disease status

Sample columns

Reads from proband: Proband

Reads from affected parent: Affected

Reads from unaffected parent: Unaffected

Analysis parameters

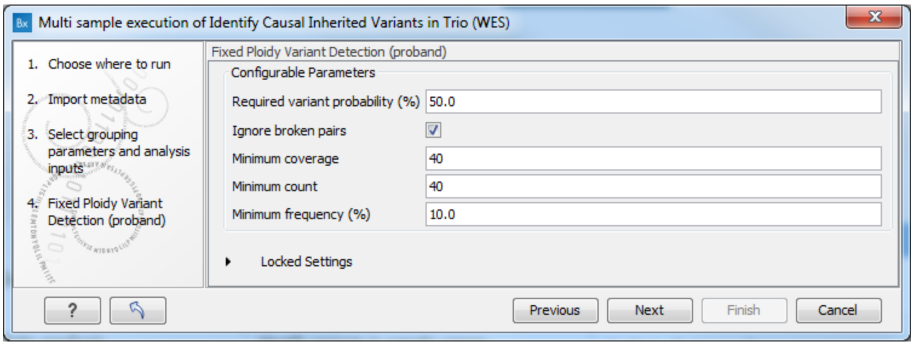
Targeted region file: SeqCap_EZ_Exome_v2_BED_CHR_21

Data

Sample ID	Familly...	Disease st...
✓ Father A	A	Affected
✓ Proband A	A	Proband
✓ Mother A	A	Unaffected
✓ Father B	B	Unaffected
✓ Proband B	B	Proband
✓ Mother B	B	Affected

2 samples ready

Figure 5: *The Multi sample execution of Identify Casual Inherited Variants in Trio (WES) after the parameters have been set.*



Multi sample execution of Identify Causal Inherited Variants in Trio (WES)

1. Choose where to run
2. Import metadata
3. Select grouping parameters and analysis inputs
4. Fixed Ploidy Variant Detection (proband)

Fixed Ploidy Variant Detection (proband)

Configurable Parameters

Required variant probability (%): 50.0

Ignore broken pairs: ☑

Minimum coverage: 40

Minimum count: 40

Minimum frequency (%): 10.0

Locked Settings

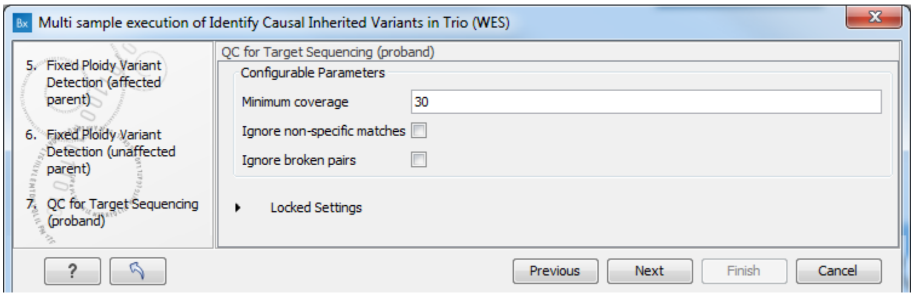
Figure 6: *Set Minimum count and Minimum coverage to 40 for the "Fixed Ploidy Variant Detection" tool.*

7. In the "Remove Variants Found in HapMap" window, leave the 12 elements as they are already selected (figure 8) and click **Next**.

8. Finally, choose to save the results in a separate folder: for example create a new folder called Results.

Analyzing the results

Once the analysis has completed, the results are organized by batch, with one subfolder per family in the Results folder as can be seen on figure 9. Each folder contains a Genome Browser View that can be opened by double clicking on it from the Navigation Area. The View opens together with a variant table listing the variants identified in the proband. These variants are the putative disease causing variants as they are shared with the affected parent, and not with the


Multi sample execution of Identify Causal Inherited Variants in Trio (WES)

5. Fixed Ploidy Variant Detection (affected parent)
6. Fixed Ploidy Variant Detection (unaffected parent)
7. QC for Target Sequencing (proband)

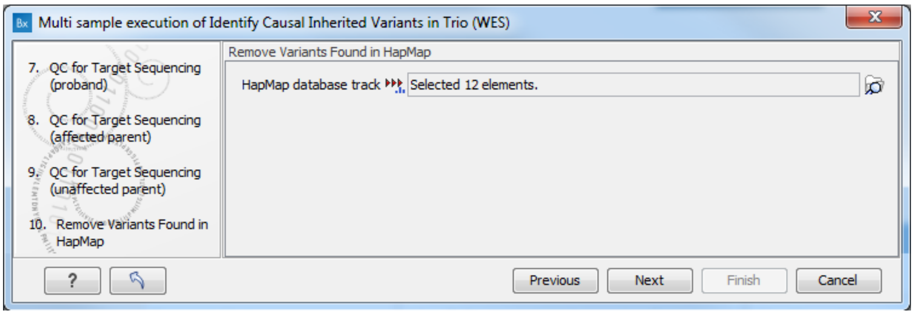
QC for Target Sequencing (proband)

Configurable Parameters

Minimum coverage: 30
Ignore non-specific matches: □
Ignore broken pairs: □

Locked Settings

Figure 7: *Set the Minimum coverage in the "QC for Target Sequence" wizard window as 30.*


Multi sample execution of Identify Causal Inherited Variants in Trio (WES)

7. QC for Target Sequencing (proband)
8. QC for Target Sequencing (affected parent)
9. QC for Target Sequencing (unaffected parent)
10. Remove Variants Found in HapMap

Remove Variants Found in HapMap

HapMap database track: Selected 12 elements.

Figure 8: *We willwork with the 12 populations from the hapmap database.*

unaffected parent.

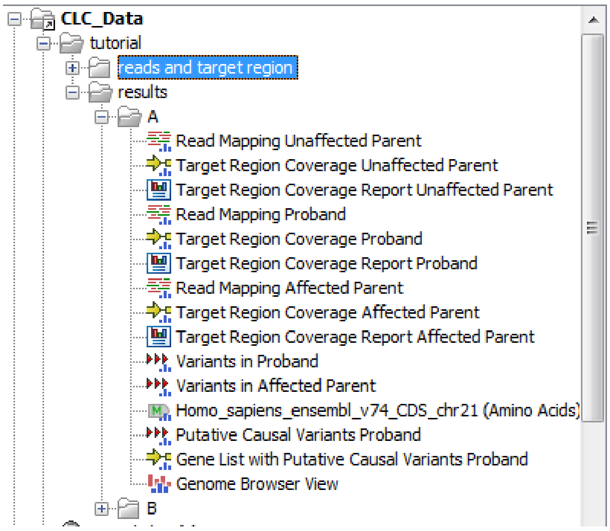
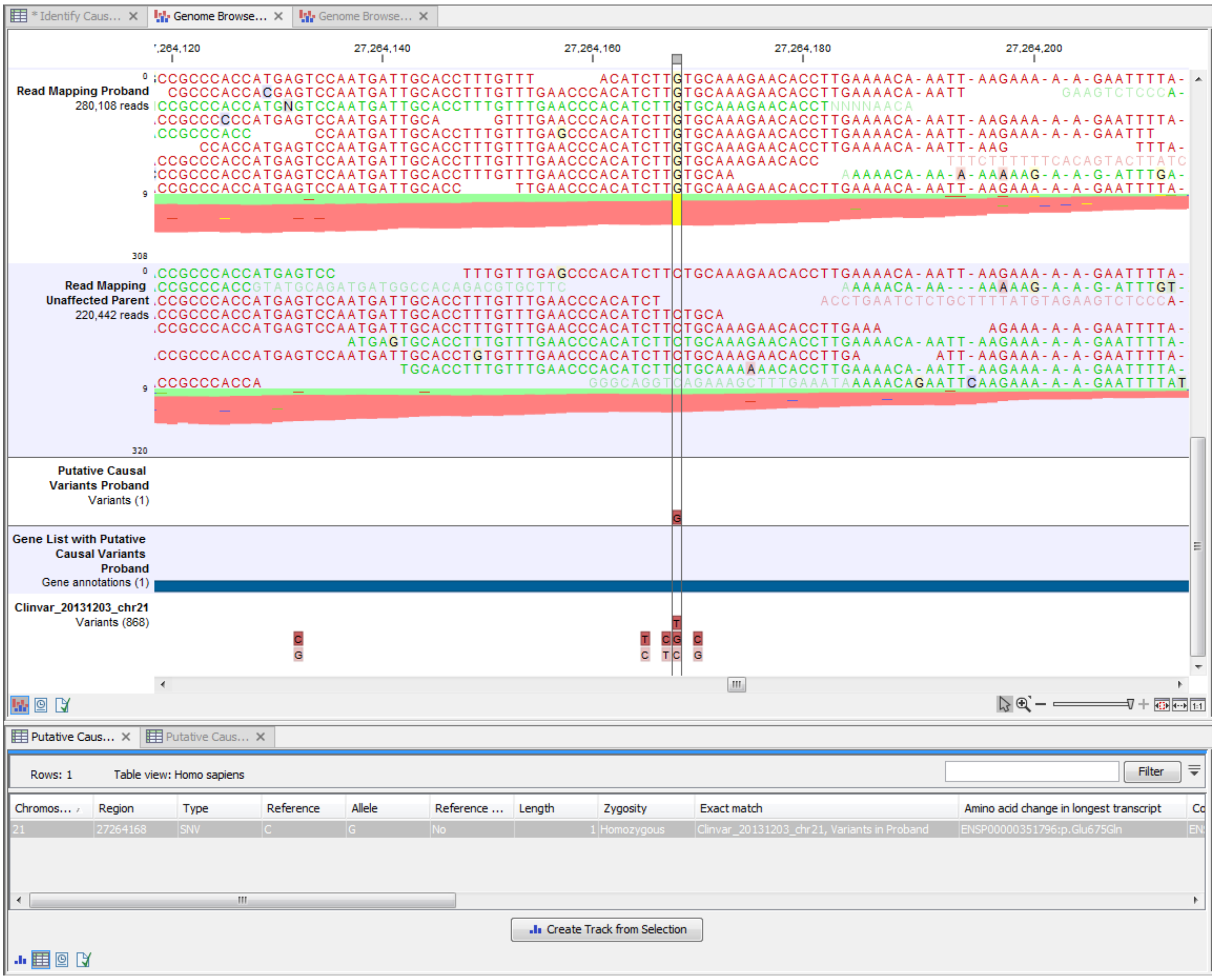

CLC_Data
tutorial
reads and target region
results
A
Read Mapping Unaffected Parent
Target Region Coverage Unaffected Parent
Target Region Coverage Report Unaffected Parent
Read Mapping Proband
Target Region Coverage Proband
Target Region Coverage Report Proband
Read Mapping Affected Parent
Target Region Coverage Affected Parent
Target Region Coverage Report Affected Parent
Variants in Proband
Variants in Affected Parent
Homo_sapiens_ensembl_v74_CDS_chr21 (Amino Acids)
Putative Causal Variants Proband
Gene List with Putative Causal Variants Proband
Genome Browser View
B

Figure 9: *When running a workflow in a batch mode, one folder per batch is generated.*

The Genome Browser View and the table below are linked, so clicking on a variant in the table will display that variant in the Browser View at the nucleotide level, allowing you to look closely at the read mapping and coverage at a particular position. You get the opportunity to compare the reads in the unaffected parent with the ones in the affected parent and the proband (figure 10).

It also allows you to take a look at the amino acids and see if the variant has resulted in any change in the amino acid sequence (figure 11).

The table below also shows information about changes in the allele compared to the reference,



Rows: 1 Table view: Homo sapiens

Chromos...	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	Amino acid change in longest transcript	Co
21	27264168	SNV	C	G	No	1	Homozygous	Clinvar_20131203_chr21, Variants in Proband	ENSP00000351796:p.Glu675Gln	EN

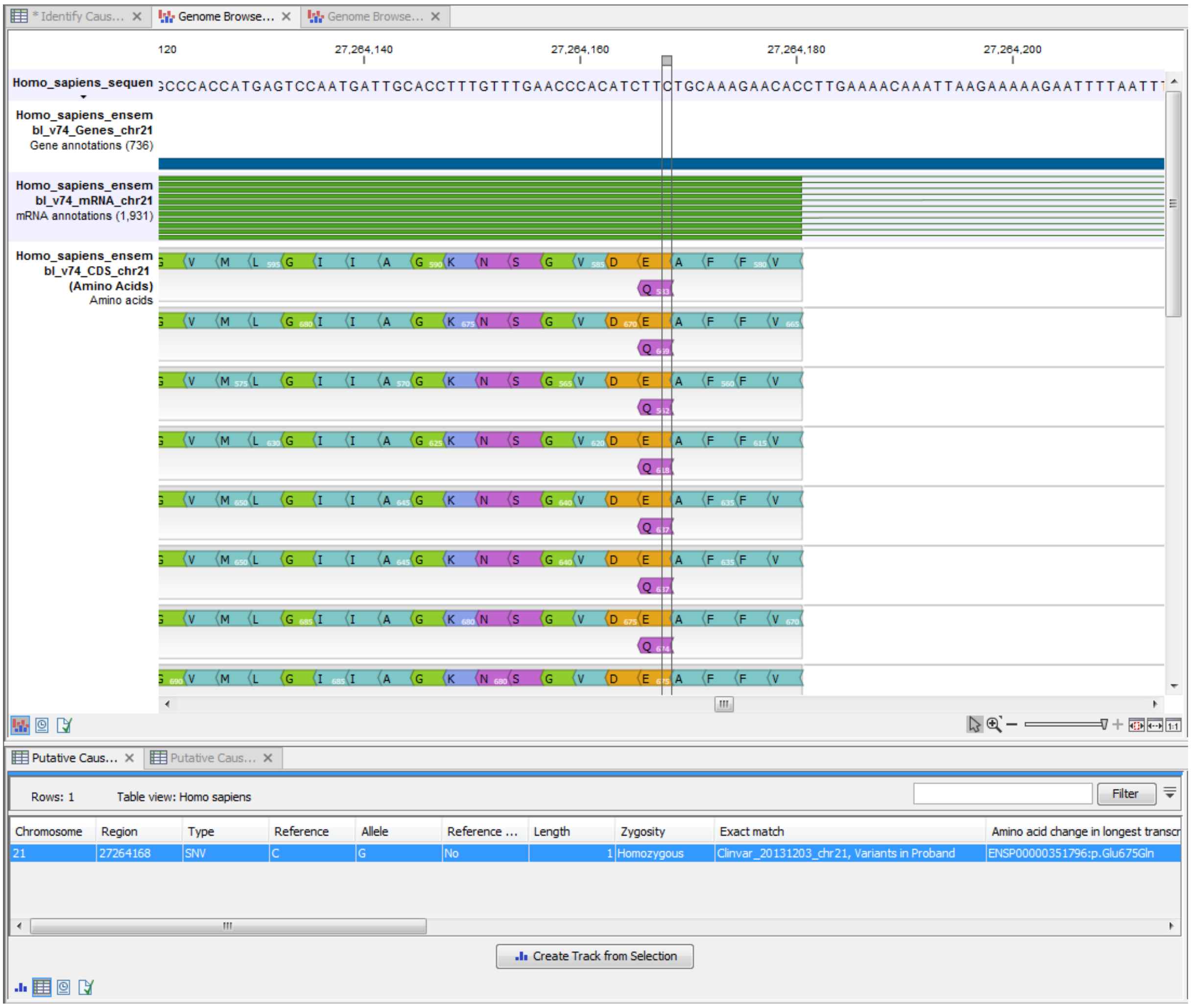
Figure 10: *When viewing the variant at the nucleotide level, it is possible to compare the reads in the different family members and get an overview of whether the variant is already known in a particular database.*

amino acid changes, information about which gene carries the change, and suggest which disease the change might be involved in. In this tutorial we have identified one mutation per family: for family A, the mutation is in the amyloid precursor protein gene and causing hereditary cerebral hemorrhage with amyloidosis - Dutch type; for family B, the variant is in the alphaA-crystallin gene CRYAA and linked to inherited cataract-microcornea.

Data Management reminder

Before you leave this tutorial, remember to change the applied Reference Data Set back to a full genome version.

1. Open the **Data Management** tool by clicking on the button with that name in the top right corner of the Workbench.
2. If you have a server, choose whether you want to apply the Reference dataset locally or on the server.



* Identify Caus...
Genome Browse...
Genome Browse...
27,264,140
27,264,160
27,264,180
27,264,200
Homo_sapiens_sequen
Homo_sapiens_ensembl_v74_Genes_chr21 Gene annotations (736)
Homo_sapiens_ensembl_v74_mRNA_chr21 mRNA annotations (1,931)
Homo_sapiens_ensembl_v74_CDS_chr21 (Amino Acids) Amino acids
Putative Caus...
Putative Caus...
Filter
Create Track from Selection

Rows: 1 Table view: Homo sapiens

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	Amino acid change in longest transcr
21	27264168	SNV	C	G	No	1	Homozygous	Clinvar_20131203_chr21, Variants in Proband	ENSP00000351796:p.Glu675Gln

Figure 11: *The amino acids trackin the Genome Browser View shows potential amino acids changes that have occurred as the result of the variant.*

3. Select the **hg19** or **hg38** Reference Data Set.

4. Click on the button labeled **Apply** before clicking on the button labeled **Close**.