QIAGEN

Guide

Guide to Biomedical Genomics Workbench

August 5, 2016

Sample to Insight

Install Biomedical Genomics Workbench

The workbench will handle large amount of data, so your computer needs to have at least 16 GB RAM (24 GB RAM is recommended), and a minimum of 100 GB free disk space in the tmp directory as well as a minimum 90 GB free disk space required for the CLC_References directory.

- Go to: https://www.qiagenbioinformatics.com/products/biomedical-genomics-workbench/.
- Click on the "Take a test drive" button.
- Fill in the required information so you can receive an installer per email.
- Once you received the email, click on the link to download the installer.
- Open the installer and go through the steps using the default options.
- Your workbench is installed!

Downloading the reference data

To be able to process your data through the Ready to Use workflows, you need to download the relevant reference data on your computer.

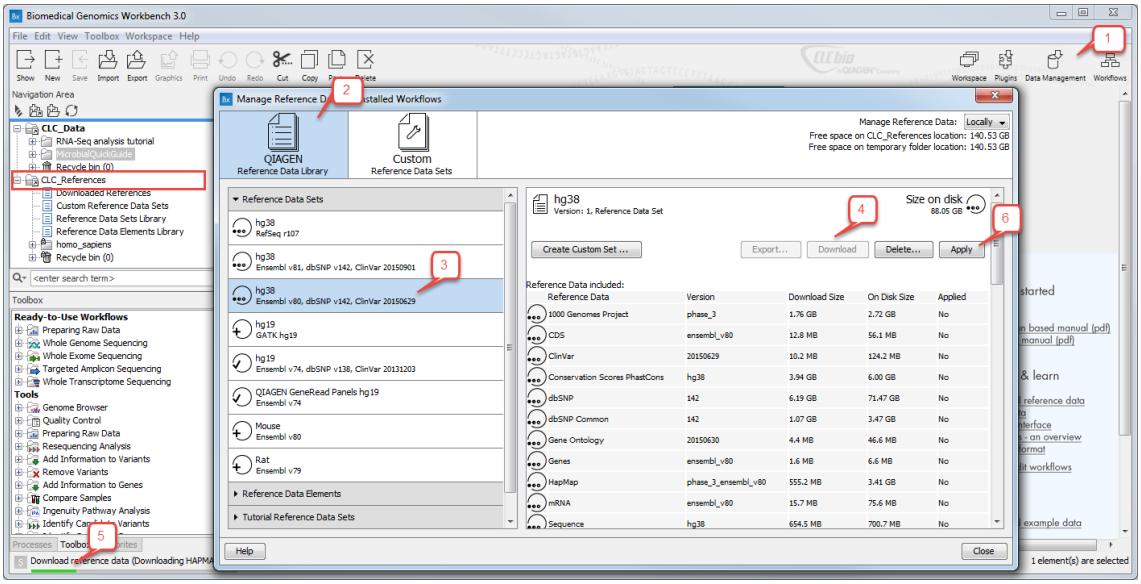

Figure 1: *Downloading the reference data.*

- Click on the Data Management icon in the top right corner of the workbench.
- In the QIAGEN Reference Data Library, select the Reference Data Set you want to download, and click on the button "Download". The download may take some time depending on the speed of your internet connection. A green process bar in the lower left corner of the workbench will indicate the progress.

- Once the data is downloaded (visualized by a check mark next to the data set), click on the button "Apply".
- By clicking "Apply", the Ready-to-Use workflows are now linked to the reference data and can be used.

The reference data is stored in a folder called CLC_References located in the Navigation Area, on the left hand side of the workbench.

Import the Target region files for Whole Exome Sequencing applications

In the case of targeted sequencing, you need to import in the workbench the target region file. If you do not have such a *.bed file, contact the company from which you bought the panel. Once the file is saved on your computer, click on the **Import** button on the top left of the workbench, and use the **Tracks** option. The file is then saved in the Navigation Area of your workbench.

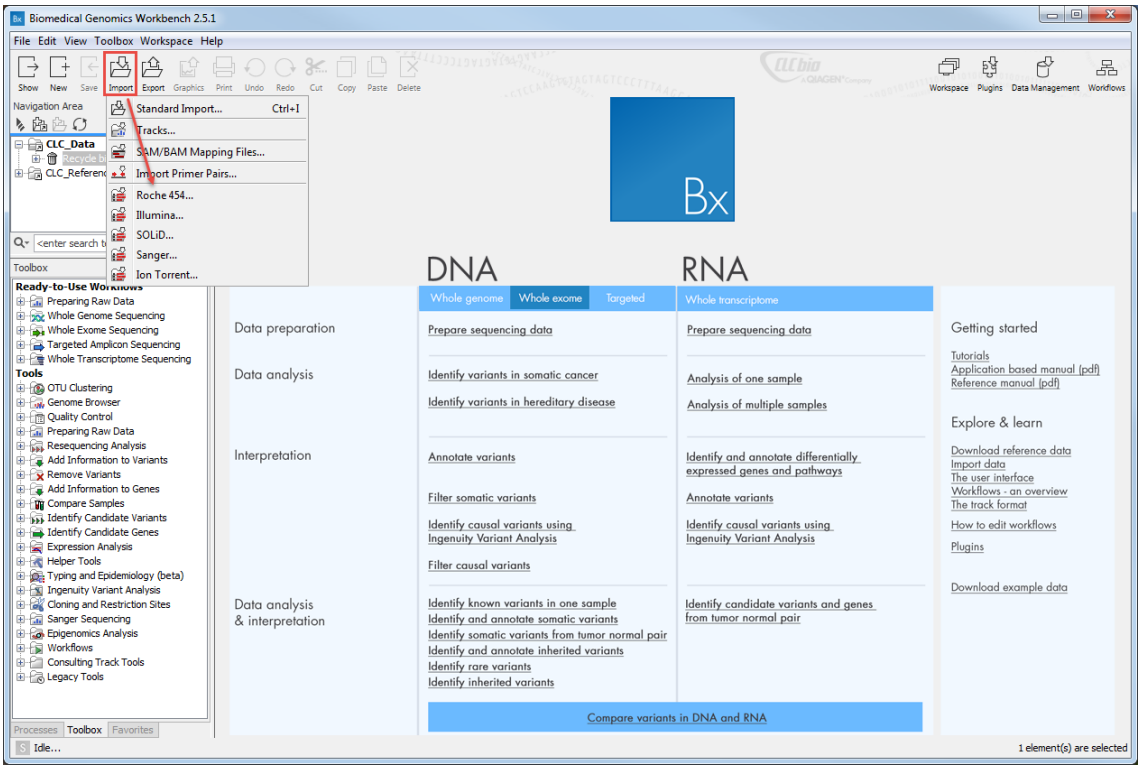

Figure 2: *Select the relevant option depending on what you are importing.*

Import your NGS data

Click on the Import button on the top left of the workbench, and use the option relevant to the type of data you are importing.

For Illumina fastq and fastq.gz files

- Select all relevant read files for all samples you want to import into the workbench. There is no need to unzip your data prior to the import.

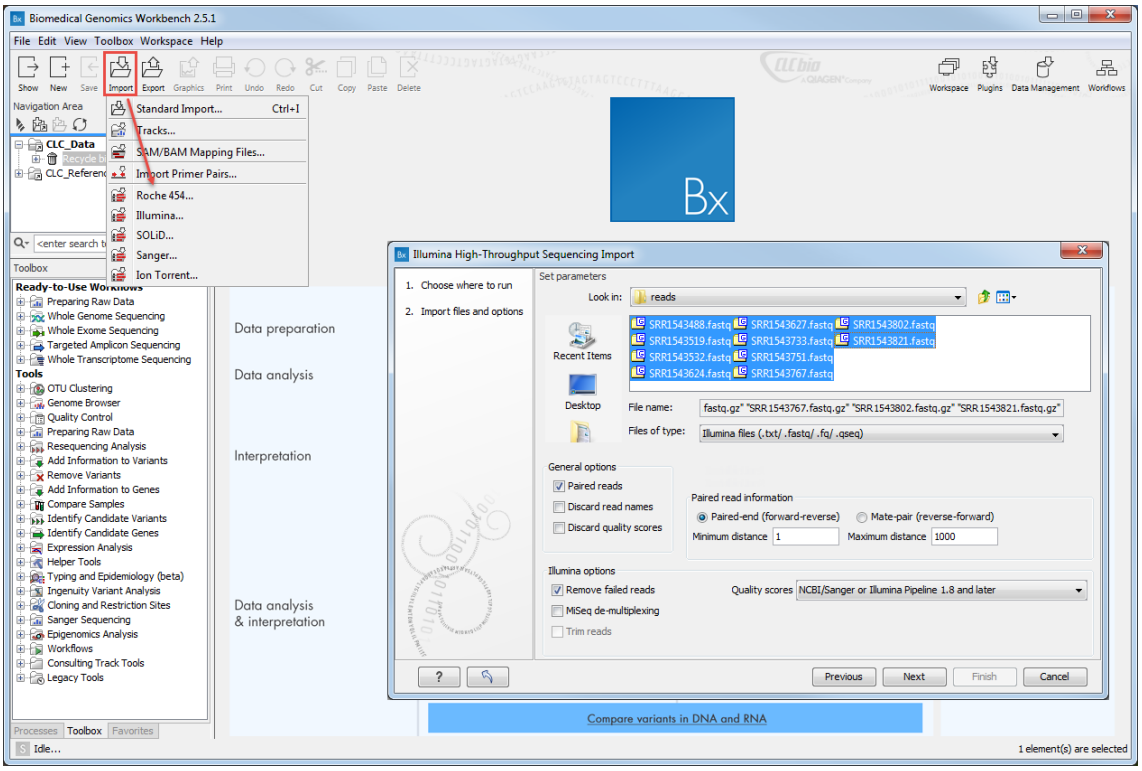

Figure 3: *Select the Illumina fastq and fastq.gz files.*

- If you have paired-end or mate-pair reads, check the Paired reads parameter under General options. This will activate the Paired read information section (see next step). If you have single reads, you can leave Paired reads unchecked and skip the following step.
- In the Paired read information section, select either Paired-end (forward-reverse) or Mate-pair (reverse-forward) and set up the Minimum and Maximum distances for your sequenced fragments using the following default values: between 1 and 1000 bp for paired-end distances and between 1000-5000 bp for mate-pair reads. Note that the tools included in Ready-to-Use workflows have an "Auto-detect paired distances" option that will overwrite and correct the information provided during import.
- Leave the other parameters as they are set by default.
- Click on the button "Next" and select a location in the Navigation Area where you would like to save your reads. Depending on the amount of data, the import might take some time.

Prepare raw data using a Ready-to-Use workflow

Create an adapter trimming list As remaining adaptor sequences in your data might lead to bias in downstream data analysis, we recommend trimming them off your reads. In the case of Illumina reads, the adapter sequences list can be found here: http://support.illumina.com/downloads/illumina-customer-sequence-letter.html.

- In the workbench, go to New | Trim Adapter List.
- Click on the button **Add Row** found at the bottom of the View Area in the New Adapter Trim List.
- Type or copy-paste the name and sequence of the first adapter, i.e. TruSeq Universal.
- Set the strand to **Minus**, choose to **Remove Adapter**, and leave the other options at their default values.
- Click on the button **Finish** to create the adapter trim list.
- In the Illumina letter, copy the Index Adapter sequence from the beginning (5' end) up to the underlined index. In this way you only need to add one adapter sequence and not all the index adapters.
- Reverse complement this common part of the Index Adapter sequence. You can use the following website `http://www.bioinformatics.org/sms/rev_comp.html` to do so.
- In the Adapter list, click on the **Add Row** button and type or copy paste in the name of the adapter, i.e. TruSeq Index.
- Then type or copy-paste the reverse complement sequence of the TruSeq Index adapter.
- Set the strand to **Minus**, choose to **Remove Adapter**, and leave the other options at their default values.
- Click **Finish**.
- Save the generated adapter trim list as New_Trim_Adapter_List in the Navigation Area using the Save button in the upper left corner of the workbench.

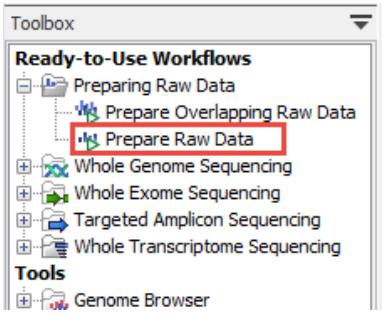

Figure 4: *Select the Prepare Raw Data workflow from the Toolbox.*

The "Prepare Raw Data" workflow

- Go to Toolbox | Ready-to-Use Workflows | Prepare Raw Data.
- This opens a wizard window. On the left side of the wizard, it is stated what you have to do at each step: select a sample to be run in the workflow, select the New_Trim_Adapter_List you just created, and save your results in the Navigation Area. Click **Previous** or **Next** to navigate between wizard windows, and **Finish** to start the workflow.

The Prepare Raw Data workflow will generate several reports, and up to two outputs of trimmed sequences files per samples. All outputs can be used as input for the Ready-to-Use workflow covered in the next paragraph, but we do not recommend including the broken pairs (orphans) as they may be of sub-optimal quality.

Analyze your data with a Ready-to-Use workflow

- To choose the relevant workflow, go to Toolbox | Ready-to-Use Workflows and select the folder that corresponds to the type of sequencing that was performed: Whole Genome or Whole Exome Sequencing.
- Then select the subfolder that corresponds to the research area you are investigating: Somatic Cancer or Hereditary Disease. The workflows from these two folders differ in the algorithm used to call variants: in somatic cancer workflows, variants can be detected at very low frequency, so these workflows could also be applicable to the study of mosaic diseases and in cases when allelic drop out occurs.
- Run any Ready-to-Use workflow with all parameters set as default for the first time.
- After looking at the outputs, you can if necessary modify the parameter settings of the workflow to optimize it to your need.

In the Somatic Cancer folder, the *Identify Somatic Variants from Tumor Normal Pair* workflow generates tumor-specific variants. You can refer to our tutorial to learn how to run this particular workflow: http://www.clcbio.com/files/tutorials/cancer-identify-somatic-variants.pdf. You can use the *Identify and Annotate Variants* workflow if you do not have matched samples.

In the Hereditary Disease folder, you can use one of the *Identify Causal Inherited Variants* or *Identify Rare Disease Causing Mutation* workflow. You have the choice between 2 versions of each workflow depending on the size of the family for which you have data: it can be a Trio, with 2 parents and the proband, or a Family of Four if you have access to data from a sibling or another related family member.

To explore in detail how to modify a Ready-to-Use workflow or how to customize a Reference Data Set to fit your needs, check our tutorial called "Modification of an Existing Workflow": http://www.clcbio.com/files/tutorials/cancer-modification_existing_workflow.pdf.

And finally, learn how to run any Ready-to-Use workflows in batching mode using our "Batching Ready-to-Use Workflows" tutorial: http://www.clcbio.com/files/tutorials/Batching_of_multi-input_workflows.pdf.

Inspecting results in the Genome Browser View

Once the analysis is done, open the file called Genome Browser View in the Navigation Area. In the View Area of the workbench, The Genome Browser view opens in a split view with a table listing the variants generated by the workflow. The table and the tracks are linked, which means that clicking on a variant in the table will display the selected variant in the context of mapped sequencing reads and diverse reference databases.

In read mapping tracks, paired reads still in proper pairs are blue. Single forward reads are green while single reverse reads are represented in red. In a paired-end experiment, red and green reads represent broken pairs: either the other read was not mapped, or it was mapped to another chromosome, or the distance between the reads is shorter or longer than specified or automatically calculated. When a read would have matched equally well another place in the mapping, it is considered a non-specific match and will be represented in yellow. Unaligned ends are shown with a faded color. Below the reads is an overflow graph in the same colors as the sequences (forward = green and reverse = red), with mismatches in reads shown as narrow horizontal lines (red = A, blue = C, yellow = G, and green = T). You can scroll through the read mapping to see the detailed reads hidden in the overflow graph by holding Alt and using the Scroll wheel of your mouse.
