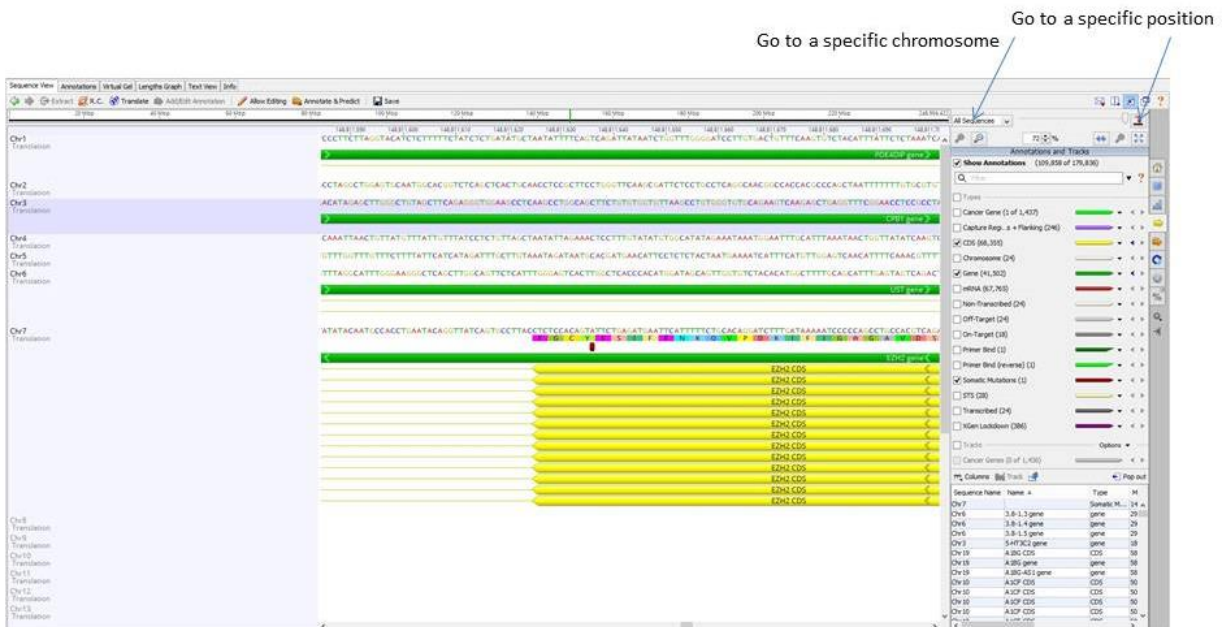


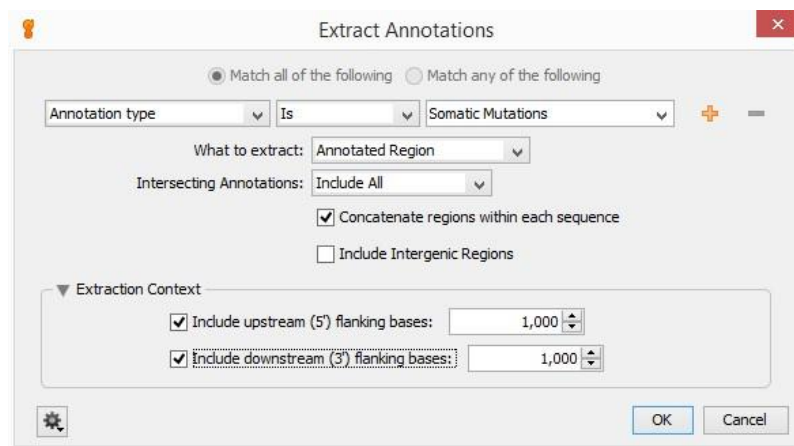
# GENEIOUS DETAILED WORKFLOW FOR THE ANALYSIS OF CELL-FREE DNA LIBRARIES SUBJECTED TO TARGETED HYBRIDIZATION CAPTURE USING PERSONALIZED BIOTINYLATED BAITS AND NGS

## 1. Import Fast Q files into Geneious.

- 1.1. Drag FastQ files into Geneious or
- 1.2. Go to File > Import > From File
- 1.3. Set the two sequence lists (\_R1 and \_R2) as paired reads. Go to Sequence > Set Paired Reads. Set expected distance/insert size to 300 bp. Delete the two original FastQ files
- 1.4. Set up Reference
- 1.5. Download Human Reference Genome and import it into Geneious (e.g. [http://support.illumina.com/sequencing/sequencing\\_software/igenome.html](http://support.illumina.com/sequencing/sequencing_software/igenome.html))
- 1.6. Label targeted positions (e.g. those genomic positions mutated in the primary tumors) as “Somatic Mutations”. Select a base or group of bases in the reference and then click on the “Add Annotation” button.

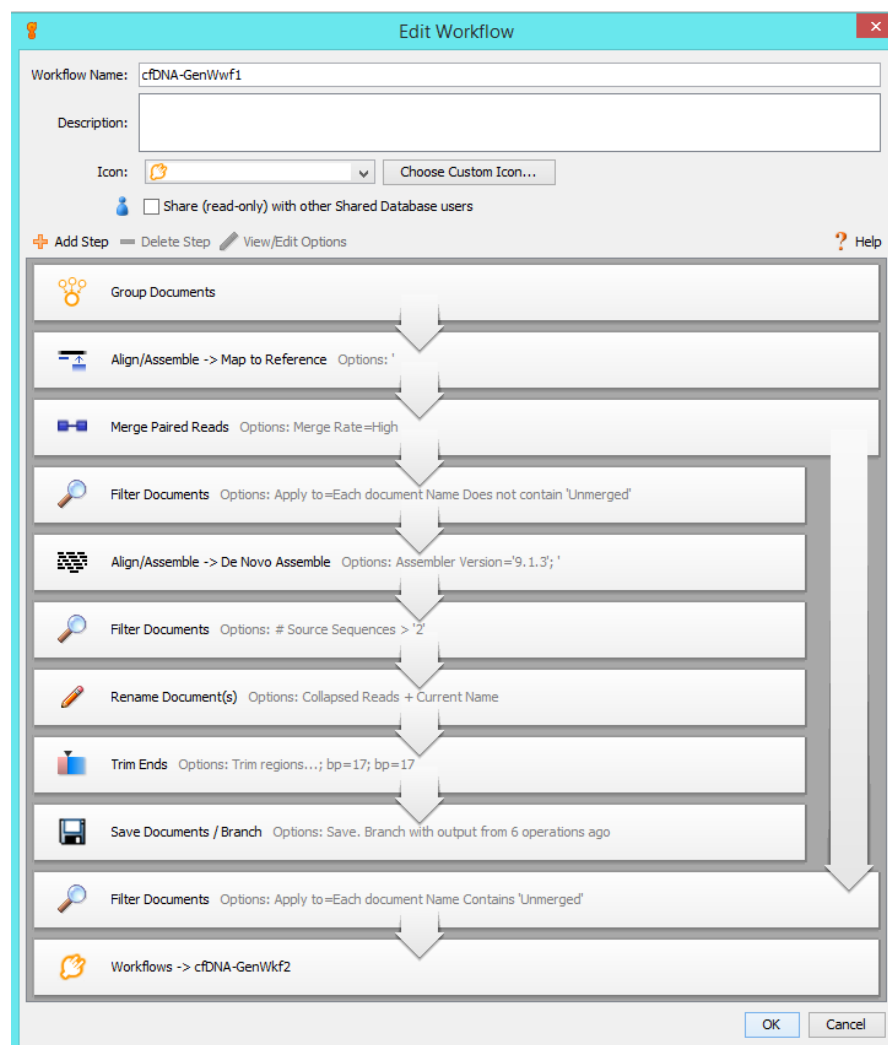


## 1.7. Go to Tools > Extract Annotations and use the following settings:



- 1.7. Select all extractions (Ctrl+A) and then go to Tools>**Concatenate Sequences** or Alignments. Add 500 bp spacers between loci. Do not circularize sequences.
- 1.8. If selected loci display high sequence complexity (there are no other regions of the genome with >90% sequence similarity) the reference is finished. Otherwise, similar loci must be included in the reference to avoid mapping artifacts. Running local **MegaBLAST** in Geneious or other platforms (i.e. UCSC Blat tool, <https://genome.ucsc.edu/cgi-bin/hgBlat?command=start>) can help ascertaining weather loci are composed of unique sequences or if similar loci (i.e. paralogs, pseudogenes etc.) exist. If so, extract those loci and concatenate them into the reference. Use the % Pairwise Identity and the Query Coverage Scores, which should be <90%.

2. Select reference and list of paired reads and go to **Workflows > cfDNA-GenWkf1**



\*Customizable options for the *Align/Assemble – Map to Reference* step of the workflow

The screenshot shows the 'Edit Align/Assemble -> Map to Reference' window. The main window has several sections: 'Options to expose to user when workflow is run', 'All Operation Options', 'Method', 'Trim Before Mapping', and 'Results'. A 'Trim Options' dialog box is overlaid on the right side, showing settings for trimming reads, such as 'Annotate new trimmed regions', 'Trim vectors', 'Minimum BLAST alignment score', and 'Trim 5' End' and 'Trim 3' End' options.

**Options to expose to user when workflow is run**

- Expose no options
- Expose all options
- Expose some options

Optionally label exposed options as:   Access exposed options via button

Expose:  With Alternative Label:

**All Operation Options (those not exposed to workflow user and default values for options that are exposed)**

**Data**

- Dissolve contigs and re-assemble
- Reference Sequence:
- Assemble by:  part of name, separated by
- Assemble each sequence list separately

**Method**

- Mapper:
- Sensitivity:
- Find structural variants and deletions of any size
- Find large deletions up to  bp
- Fine Tuning:
- Memory Required: 1.0 GB of 26 GB

*Note: Paired reads can be set up or changed using Sequence > Set Paired Reads*

**Trim Before Mapping**

- Use existing trim regions
- Remove existing trim regions from sequences
- Re-trim sequences

**Results**

Assembly Name:

- Save assembly report
- Save list of unused reads
- Save list of used reads  Include mates
- Save in sub-folder

**Mapping Options**

- Minimum mapping quality:
- Map multiple best matches:
- Trim paired read overhangs
- Only map paired reads which
- Minimum support for structural variant discovery:  reads
- Allow Gaps Maximum Per Read:  %
- Maximum Gap Size:
- Minimum Overlap:  %
- Minimum Overlap Identity:  %
- Word Length:
- Index Word Length:
- Ignore words repeated more than  times
- Maximum Mismatches Per Read:  %
- Maximum Ambiguity:
- Accurately map reads with errors to repeat regions
- Search more thoroughly for poor matching reads

**Trim Options Dialog**

- Annotate new trimmed regions (ignored by assembly)
- Remove new trimmed regions from sequences
- Trim vectors:
- Minimum BLAST alignment score:
- 
- Trim primers:  Allow Mismatches:
- Minimum Match Length:
- Error Probability Limit:  (decrease to trim more)
- Trim regions with more than a 5% chance of an error*
- Maximum low quality bases:
- Maximum ambiguities:
- Trim 5' End  At least  bp
- Trim 3' End  At least  bp
- Maximum length after trim:  (Trim excess from 3' end)

**Visualization**

- Lock annotations
- Compress annotations
- Show arrow
- Hide excess
- Sizes**

\*Customizable options for the *De Novo Assemble* step of the workflow

The screenshot displays the 'Edit Align/Assemble -> De Novo Assemble' window. The main interface is divided into several sections:

- Options to expose to user when workflow is run:** Includes radio buttons for 'Expose no options', 'Expose all options', and 'Expose some options'. It also features a text input for 'Optionally label exposed options as:' and a dropdown for 'Expose:'.
- All Operation Options (those not exposed to workflow user and default values for options that are exposed):** Contains checkboxes for 'Dissolve contigs and re-assemble', 'Assemble by: 1st part of name, separated by: -(hyphen)', 'Assemble each sequence list separately', and 'Use: 100% of data. Suitable for genome size between 0 KB and 0 KB'.
- Method:** Shows 'Assembler: Geneious' and 'Sensitivity: Custom Sensitivity'. It also indicates 'Memory Required: 84 MB of 26 GB' and includes a note: 'Note: Paired reads can be set up or changed using Sequence > Set Paired Reads'.
- Trim Before Assembly:** Includes radio buttons for 'Use existing trim regions', 'Remove existing trim regions from sequences', 'Re-trim sequences', and 'Do not trim (discard trim annotations)'. There is an 'Options' button next to 'Re-trim sequences'.
- Results:** Shows 'Assembly Name: {Reads Name} Assembly' and checkboxes for 'Save assembly report', 'Save list of unused reads', 'Save in sub-folder', 'Save contigs (Maximum: 1,000)', and 'Save consensus sequences'.
- Bottom Section:** Contains various assembly parameters such as 'Don't merge variants with coverage over approximately: 6', 'Merge homopolymer variants', 'Produce scaffolds', 'Allow Gaps', 'Minimum Overlap: 90%', 'Word Length: 12', 'Ignore words repeated more than: 100 times', 'Maximum Mismatches Per Read: 1%', 'Circularize contigs with matching ends', 'Maximum Gap Size: 1', 'Minimum Overlap Identity: 99%', 'Index Word Length: 12', 'Reanalyze threshold: 16', 'Maximum Ambiguity: 4', and 'Only use paired hits during assembly'. There is also a 'Low Memory Use' slider and a 'Speed' indicator.

The 'Consensus Options' dialog box is open, showing:

- Radio buttons for 'Save consensus used by assembler' and 'Generate new consensus from contig'.
- 'Threshold: 0% - Majority' and 'Threshold for sequences without quality: 75%'.
- Checkboxes for 'Ignore Gaps' and 'Assign Quality' (set to 'Total').
- 'If no coverage call' dropdown.
- 'Call' checkbox and 'Coverage' dropdown.
- 'Split into separate sequences around ? calls' checkbox.
- 'Reset to Defaults', 'OK', and 'Cancel' buttons.


On the right side of the main window, there is a table with the following columns: 'nces', '% Pairwise I...', 'Bit-Score', 'E Value', and 'Description'. The table contains several rows of data, including percentages and assembly descriptions like '1 - CTG', '2 - TCA', and '3 - AGT'.

\*Word Length and Index Word Length are the parameters that may stronger influence the speed of this operation.

**Edit Workflow**


Workflow Name: cfDNA-GenWkf2


Description:


Icon:  Choose Custom Icon...


Share (read-only) with other Shared Database users


+ Add Step - Delete Step View/Edit Options ? Help


 Workflows -> Trim and Filter Options: Error probability...=0.001; Trim regions... Post-Trim > '75'

 Set Paired Reads Options: Detach existing...

 Align/Assemble -> De Novo Assemble Options: '

 Rename Document(s) Options: Collapsed Reads + Current Name

 Filter Documents Options: # Source Sequences > '2' + Sequence Length '160'

 Trim Ends Options: Trim regions...; bp=17; bp=2

OK Cancel

\*Customizable options for the *De Novo Assemble* step of the workflow

The screenshot displays the 'Edit Align/Assemble -> De Novo Assemble' window. The main interface is titled 'Assemble reads (eg. Sanger or NGS) without using a reference'. It features several sections:

- Options to expose to user when workflow is run:** Includes radio buttons for 'Expose no options', 'Expose all options', and 'Expose some options'. There is also a field for 'Optionally label exposed options as:' and a checkbox for 'Access exposed options via button'.
- All Operation Options (those not exposed to workflow user and default values for options that are exposed):** Includes a 'Data' section with 'Disolve contigs and re-assemble' checkbox, 'Assemble by:' dropdown (set to '1st'), and 'Assemble each sequence list separately' checked. A 'Method' section shows 'Assembler: Genieous' and 'Sensitivity: Custom Sensitivity'. A note states 'Memory Required: 84 MB of 26 GB' and 'Note: Paired reads can be set up or changed using Sequence > Set Paired Reads'.
- Trim Before Assembly:** Includes radio buttons for 'Use existing trim regions', 'Remove existing trim regions from sequences', 'Re-trim sequences', and 'Do not trim (discard trim annotations)'.
- Results:** Includes 'Assembly Name: {Reads Name} Assembly' and checkboxes for 'Save assembly report', 'Save list of unused reads', 'Save in sub-folder', 'Save contigs', and 'Save consensus sequences'.
- Other options:** Includes 'Don't merge variants with coverage over approximately', 'Produce scaffolds', 'Circularize contigs with matching ends', 'Allow Gaps', 'Minimum Overlap', 'Word Length', 'Ignore words repeated more than', 'Maximum Mismatches Per Read', 'Low Memory Use' (1-5), 'Merge homopolymer variants', 'Circularize contigs with matching ends', 'Maximum Gap Size', 'Minimum Overlap Identity', 'Index Word Length', 'Reanalyze threshold', and 'Maximum Ambiguity'.

The 'Consensus Options' dialog box is open, showing:

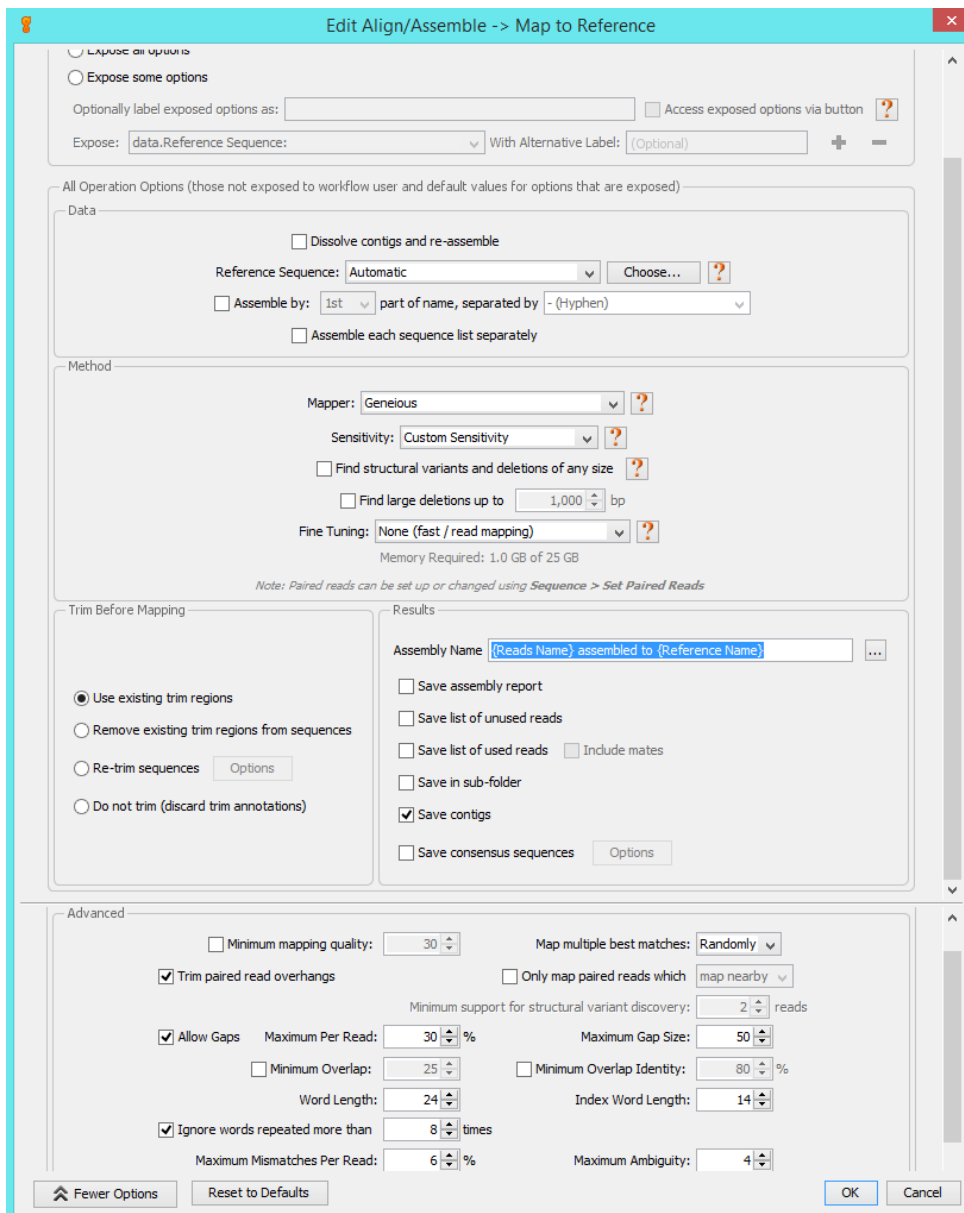
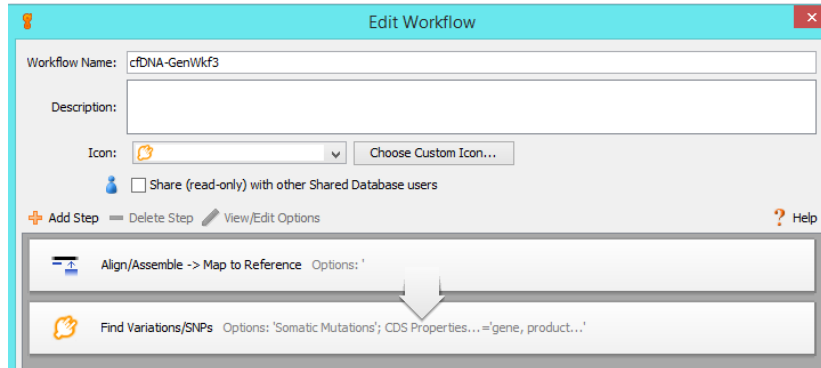
- Radio buttons for 'Save consensus used by assembler' and 'Generate new consensus from contig' (selected).
- 'Threshold: 0% - Majority' dropdown.
- 'Threshold for sequences without quality: 75%' dropdown.
- 'Ignore Gaps' checkbox.
- 'Assign Quality' dropdown (set to 'Total').
- 'If no coverage call' dropdown.
- 'Call' dropdown and 'if Coverage < 2' dropdown.
- 'Split into separate sequences around ? calls' checkbox.
- 'Reset to Defaults', 'OK', and 'Cancel' buttons.

The background interface also shows a table of assembly statistics and an 'Annotation' panel on the right.

nces	% Pairwise I...	Bit-Score	E Value	Description
	99.8%	-	-	18,636 reads 1
	98.5%	-	-	Assembly of 2
	-	-	-	1 - CTG
	-	-	-	2 - TCA
	-	-	-	3 - AGT
	-	-	-	Sequences f
	-	-	-	-
	-	-	-	-
	94.7%	-	-	Assembly of 2
	89.2%	-	-	Assembly of 3
	95.3%	-	-	Assembly of 2
	95.9%	-	-	Assembly of 2
	-	-	-	-
	-	-	-	Paired reads o
	-	-	-	-
	-	-	-	-
	99.2%	-	-	447,766 reads

\*Word Length and Index Word Length are the parameters that may stronger influence the speed of this operation.

3. Select the two lists of collapsed reads generated by the two previous workflows (i.e. **GenWkf1 + GenWkf2**) and the reference sequence and go to **Workflows > cfDNA GenWkf3**



\*Customizable options for the *Find Variations/SNPs* step of the workflow

**Edit Find Variations/SNPs**

Find variations such as SNPs and INDELs in nucleotide alignments and contigs

Options to expose to user when workflow is run

- Expose no options
- Expose all options
- Expose some options

Optionally label exposed options as:   Access exposed options via button  ?

Expose:  With Alternative Label:

All Operation Options (those not exposed to workflow user and default values for options that are exposed)

Find Polymorphisms

- Minimum Coverage:
- Minimum Variant Frequency:
- Maximum Variant P-value:  $10^{-3}$  (0.1% to see variant by chance) ?
- Minimum Strand-Bias P-value:  $10^{-5}$  when exceeding  % bias ?

Find Polymorphisms   In selected region only

Analyze Effects on Translations

- Analyze effect of polymorphisms on translations

Default Genetic Code:

Calculate Variant P-values

Assumed quality of bases without quality:  (99.0% correct)

P-value calculation method:

Homopolymer quality reduction for 454 / Ion Torrent:  % ?

Only Find SNPs

- Merge adjacent variations
- Ignore reference sequence (only find variations within the sample)
- Exclude paired reads over  % from their expected distance
- Use separate annotations for each variant at a position
- Record names of all contributing sequences for each variant
- Ignore reads mapped to multiple locations
- Don't find variations in annotation types:
- Only find variations in annotation types:

Also find variations within  bases of those types

CDS Properties to Copy:





directly from the supplemental material that accompanies this paper. Workflows intended to work on large captured regions have the suffix -lp added to the names of the three main workflows described above.