User's Guide to DNASTAR® SeqMan NGen® 12.0

For Windows®, Macintosh® and Linux®

DNASTAR, Inc. 2014

Contents

SeqMan NGen Overview

Note: For Customer Support contact information, and for a link to the most up-to-date version of this help, please see Before You Begin.

DNASTAR's SeqMan NGen® software gives you the ability to assemble and analyze large genomes with unsurpassed speed using sequence data from any major next-gen sequencing platform.

Features of the software allow you to:

- Assemble human or other large eukaryotic genomes against a genomic template quickly and easily on a desktop computer.
- Assemble nearly any type of next-gen data, including Ion Torrent, Illumina, Pacific Biosciences, Roche 454, and SOLiD.
- Perform reference-guided assemblies of billions of sequence reads and *de novo* assemblies of up to 30 million sequence reads (genome sizes up to 50 megabases).

  Note: The upper limit for project size depends on many factors, including your computer's level of RAM and its processor speed. See our recommended technical requirements for more information.

- Assemble multi-sample data, such as MID-tagged 454 samples, for later analysis in SeqMan Pro.
- Utilize an interactive, data-rich SNP report, including probabilities and genotypes determined with Bayesian statistics as well as dbSNP, GERP and COSMIC associations.
- Detect potential structural variation regions (displayed as a table in SeqMan Pro).
- Align reads against a database of genomic templates.
- Recalculate SNPs for a BAM-based assembly project.

If you plan to create a templated assembly and wish to use the dbSNP, GERP or COSMIC association features, you'll first need to download a free DNASTAR genome template package that contains the template sequence, annotations, and associated dbSNP entries. Packages are available for a variety of model organisms.

SeqMan NGen utilizes scripts to run an assembly. SeqMan NGen's wizard allows effortless generation of scripts with no programming required. For video tutorials on using the SeqMan NGen wizard, or access to a command line scripting manual, please click here and choose the **Resources** tab on the left.

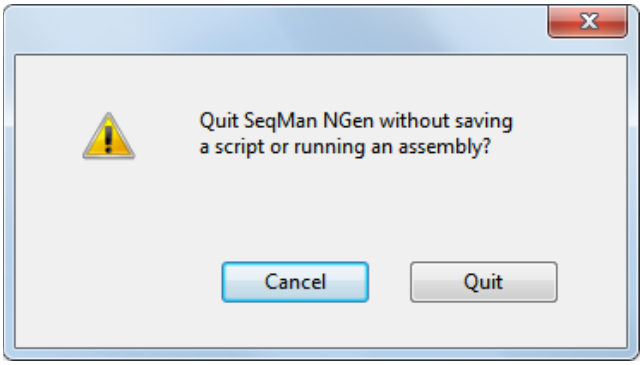
Wizard Navigation

Navigate through the SeqMan NGen wizard using the buttons at the bottom of each dialog.

- Click the **Help** button (Win) or the question mark icon (Mac) to launch the user's guide topic for the current panel.

  Note for Linux Users: The wizard **Help** button is not active on the Linux platform. Instead, wizard help is provided in this PDF. Scripting command help is available via the *SeqManNGen_ScriptingManual.txt* file that was included with your SeqMan NGen software. For further assistance, please visit the Training and Support section of our website or contact us at support@dnastar.com.

- Click **< Back** and **Next >** to navigate to the previous or next panel.
- Click **Quit** to exit SeqMan NGen. If you have not yet saved a script or run an assembly, the following confirmation prompt will appear:



Choose **Quit** to exit without saving changes, or **Cancel** to return to the wizard.

Non-English Keyboards

SeqMan NGen recognizes only standard English-keyboard characters as input. If you are using a non-English keyboard, we recommend that you switch to a "virtual" English keyboard. Click a link for instructions: Windows 8, Windows 7, Macintosh OS X 10.8, Macintosh OS X 10.9, Linux.

Before You Begin

We're here to help! If you have any difficulties with or questions about this application, please contact a DNASTAR support representative:

- E-mail: support@dnastar.com
- Phone (Madison, WI, USA): 608-258-7420
- In the USA and Canada, call toll free: 1-866-511-5090
- In the UK, call free on: 0-808-234-1643
- In Germany, call free on: 0-800-182-4747

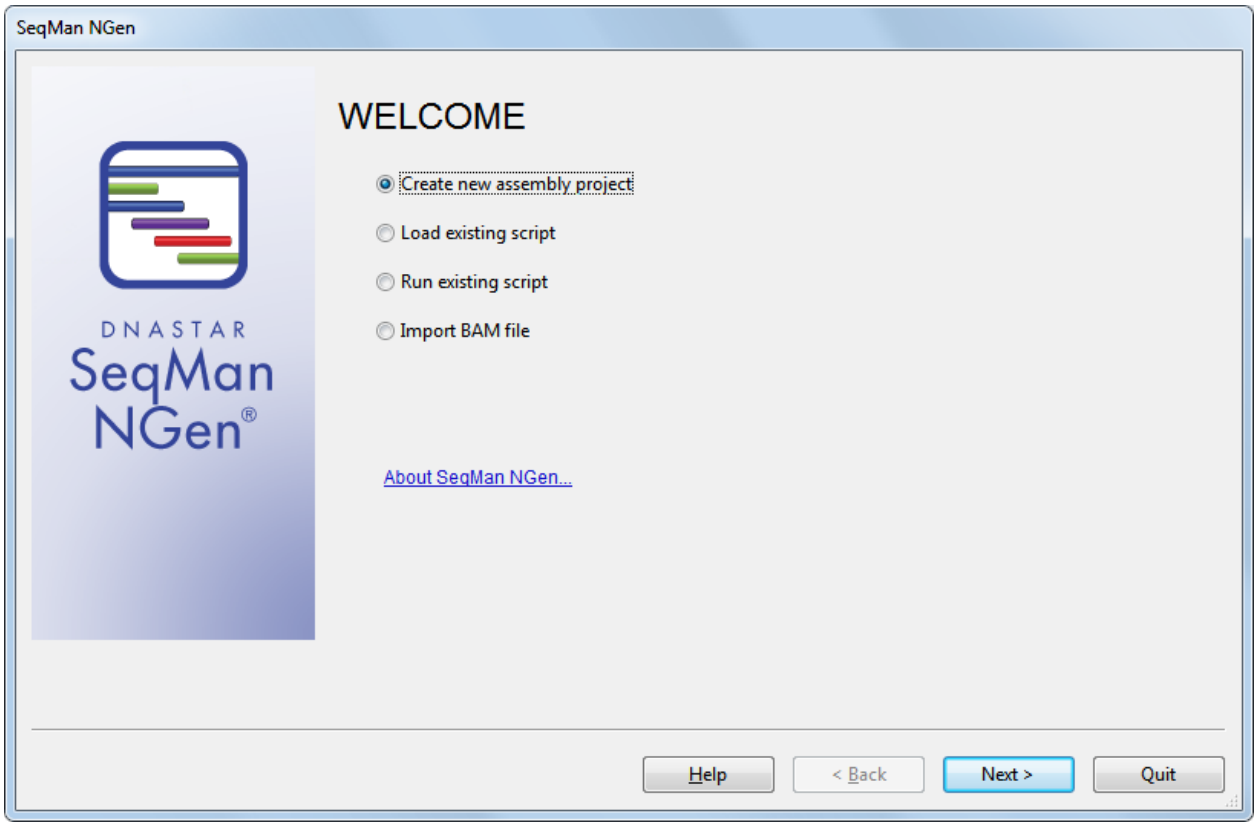
This help document pertains to SeqMan NGen® version 12, and was last updated on June 12, 2014.

If you accessed this help from within a Lasergene application: The help installed with your application was current at the time of the version release. To view the most recent help online, please visit our Training & Support page and click the appropriate application link.

To access free video tutorials for this product, please visit www.dnastar.com and use the tabs at the top to choose **Support > Videos**.

The Welcome Screen

The Welcome screen is the initial SeqMan NGen wizard screen for all workflows.



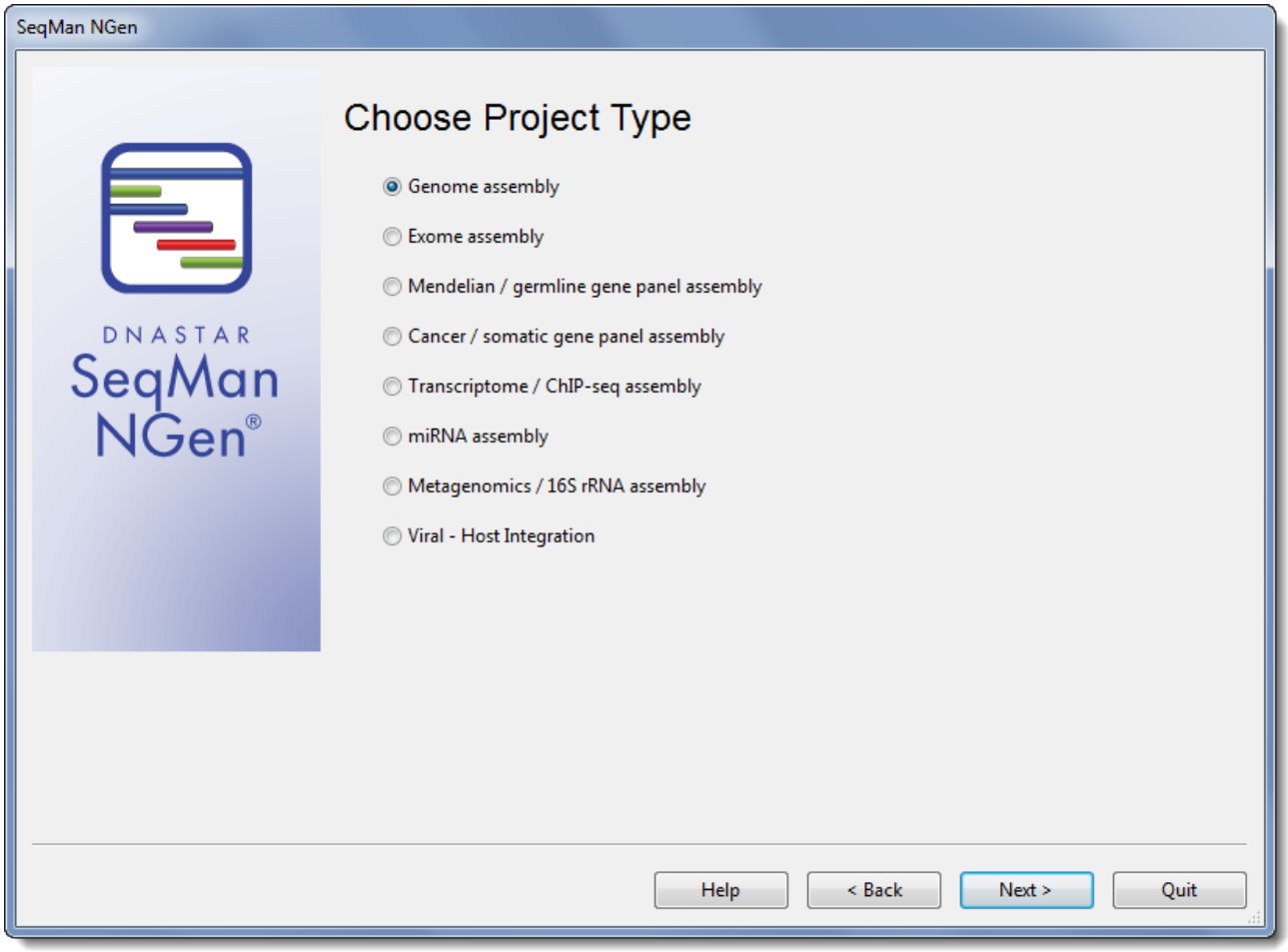
Choose from the following options:

- Select **Create new assembly project** to create a new project step-by-step using the wizard. Then click **Next** to proceed to the Choose Project Type screen.
- Choose **Load existing script** to load parameters from a past project into the wizard; this is similar to the "**File > Open**" command in other DNASTAR applications. Then click **Next** to open a file browser. After loading the script, you will be taken automatically to the Choose Project Type screen.
- Select **Run existing script** to load a past project and proceed to the end of the wizard without needing to forward through intervening screens. Then click **Next** to open a file browser. After loading the script, you will be taken automatically to the "Your assembly is ready to begin" dialog.
- Choose **Import BAM file** to go directly to the BAM Import screen, from which you can import the desired BAM file (*.bam or *.assembly). This choice also leads to the Recalculate SNPs workflow.

Click the **About SeqMan NGen** link to view basic information about this application, such as its version number. You must click **OK** to close the information window before continuing with the wizard.

Choose Project Type

The Choose Project Type dialog lets you select from several workflow types. SeqMan NGen will then populate the rest of the wizard with appropriate default parameters for your project.



Your selection in this screen determines which assembly types are enabled in the next screen, entitled Choose Assembly Type.

This selection in Choose Project Type…		…enables these options in Choose Assembly Type					
	Normal Templated	Templated with Host Removal	Templated with Control	*De Novo*	*De Novo* with Host Removal	Reference-Guided	Special Templated
Genome assembly	x		x	x		x	x
Exome assembly	x		x				
Mendelian / germline gene panel assembly	x		x				
Cancer / somatic gene panel assembly	x		x				
Transcriptome / ChIP-seq assembly	x		x	x			
miRNA assembly	x		x	x			
Metagenomics/16S rRNA assembly	x	x		x	x		
Viral-Host Integration		(The Choose Assembly Type screen is not present in this workflow; host removal is performed automatically)					

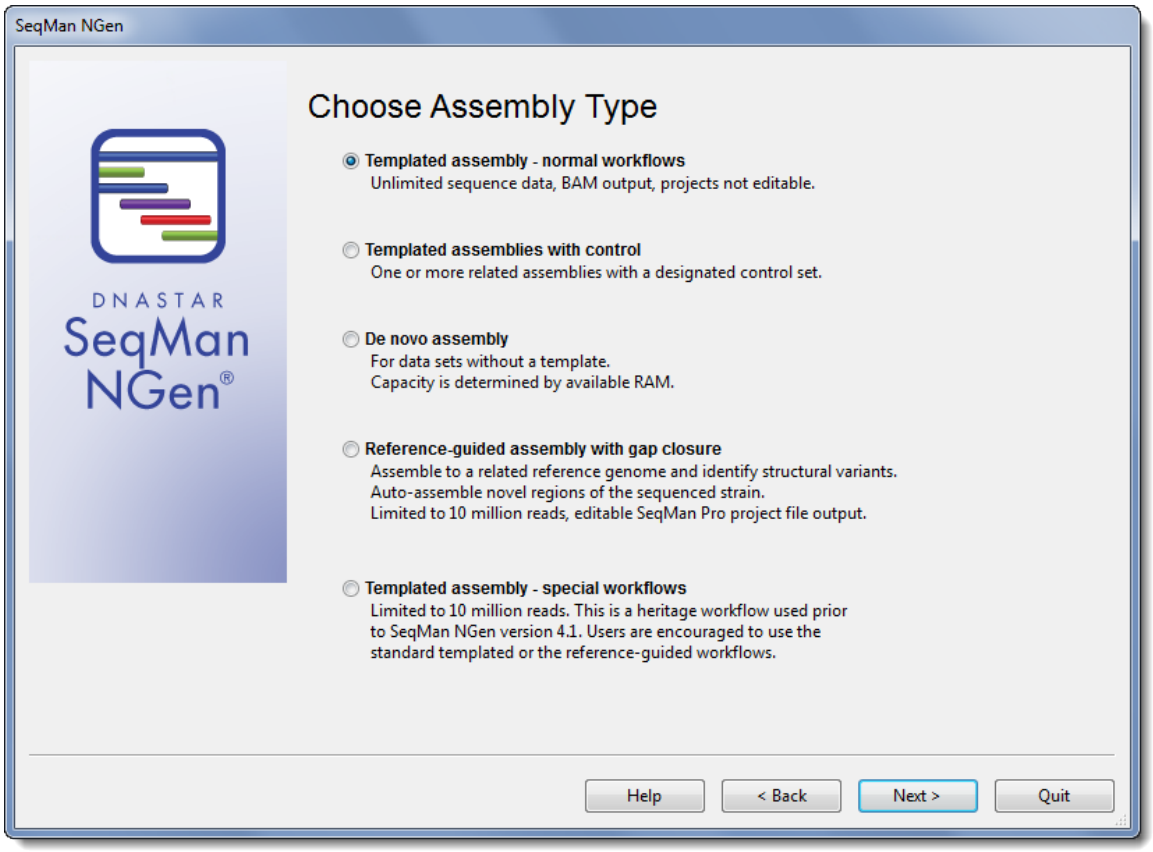
Once you are finished, click **Next >** to continue to the next wizard screen.

Choose Assembly Type

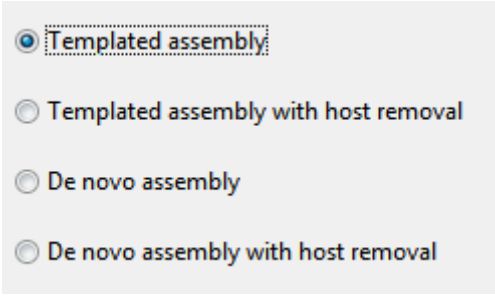
The Choose Assembly Type dialog allows you to choose between several types of templated and/or *de novo* assemblies. The text below each selection can assist you in determining the suitability of a particular assembly type for your data set.

The options available in this dialog depend upon what you selected in the previous screen. See the table in Choose Project Type for details.

Below is the version of the screen that appears for the Genome Assembly workflow.



The following version is seen in the Metagenomics/16S rRNA assembly workflow.



- **Templated assembly – normal workflows** – (called **Templated assembly** in the Metagenomics/16S rRNA Assembly workflow) To assemble/align reads onto one or more reference sequences/templates. This type of assembly can include billions of reads and large eukaryotic genomes. The BAM-formatted assembly cannot be edited, but can be viewed and analyzed using a utility such as DNASTAR's SeqMan Pro.
- **Templated assemblies with control** – To assemble one or more related assemblies with a designated control set.

- **Templated assembly with host removal** – To remove the DNA of a specified host before assembling/aligning the remaining reads onto one or more reference sequences/templates. This option is only available in the Metagenomics/16S rRNA assembly workflow.
- ***De novo* assembly** – To run a *de novo* (untemplated) assembly of up to 30 million sequence reads and up to a 50 Mbase genome. When assembling a data set *de novo*, we recommend using paired end data if available.

  Note: The ***De novo* assembly** option does not appear if you selected **Targeted resequencing / Exome assembly** in the Choose Project Type wizard dialog.
- ***De novo* assembly with host removal** – To remove the DNA of a specified host before running a *de novo* (untemplated) assembly. This option is only available in the Metagenomics/16S rRNA assembly workflow.
- **Reference-guided assembly with gap closure** – To assemble/align reads onto one or more reference sequences/templates. This option automates the assembly of indels using mate-pair data, and can include up to 10 million reads and up to a 100 Mbase genome. The SQD-formatted assembly can be edited at a later time using SeqMan Pro. For more information about this assembly type, see Reference-Guided Assembly with Gap Closure.
- **Templated assembly – special workflows** – To assemble/align reads onto one or more reference sequences/templates. This type of assembly can include up to 10 million reads and up to a 100 Mbase genome. It can be edited at a later time using a utility like SeqMan Pro.

  Note: The **Templated assembly - special workflows** selection was eliminated in SeqMan NGen 4.1, but was reintroduced in 4.1.2 as a heritage/legacy workflow for use only with the **Genome assembly** project type. We encourage you to use the standard templated or the reference-guided workflows whenever possible.

Once you are finished, click **Next >** to continue to the next wizard screen. SeqMan NGen will populate the rest of the wizard with appropriate default parameters for your assembly.

Reference-Guided Assembly with Gap Closure

The **Reference-guided assembly with gap closure**, one of the selections in the Choose Assembly Type screen, is a semi-automated assembly option that utilizes both reference-guided ("templated") and *de novo* assembly steps to resolve three types of structural variation (SV): insertions, deletions and replacements (indels) with minimal user intervention.

The following conditions apply if you wish to follow this workflow:

- In the Choose Project Type screen, you should choose **Genome Assembly**.
- In the Choose Assembly Type screen, you should choose **Reference-guided assembly with gap closure**.

- Your data should be from a haploid genome with at least one mate pair data set with read lengths of 100 bases or greater.
- Your total number of reads should be 10 million or less. If you use a larger data set, only the first 10 million reads will be used. For mate pair data, equal numbers of matching forward and reverse reads are processed.

Steps 1-3: Assembly in SeqMan NGen

During assembly, data is processed in several stages (see figures below):

Step 1) Data is mapped and aligned to a user-defined reference genome and then analyzed for characteristic SV motifs.

Step 2) The reference sequence is split at the detected SV sites, forming a series of ordered contigs.

Step 3a) Mate pair and split reads from each SV event are collected in site-specific pools and assembled *de novo*. Deletions are detected using three types of data: split reads, spanning paired-end reads, and sequence coverage information. For insertions and replacements, mate pair reads corresponding to the new sequence are collected from the unassembled read pool. Only reads anchored by mates flanking the SV in the main assembly are used at this stage.

Step 3b) The *de novo* assembled contigs are then brought into the main assembly and positioned consistently with the mate pair information.

Step 3c) For SVs where the gap is not completely covered by the *de novo* assembled contigs (e.g. insertions longer than twice the size of the insert library), additional reads from the unassembled read pool matching and extending the ends of the joining contigs are added in an attempt to "walk" across the gap. This walk is terminated when either no new reads are found or when a repeated element is encountered.

At the end of assembly in SeqMan NGen, two types of output files are produced. These allow the project to be evaluated and further processed in SeqMan Pro:

- An *.assembly package with a non-editable BAM formatted alignment file of the initial reference-guided assembly without further processing.
- The fully processed assembly in an editable SeqMan *.sqd file format.

Step 4: Further Processing in SeqMan Pro

Step 4) The editable *.sqd document, containing the fully processed assembly, is used for gap closure to complete the new sequence.

In SeqMan Pro's Project Summary window, the contigs will appear in a single ordered scaffold with the *de novo* generated contigs. The values in the contig position column of the window roughly correspond to the 5' position of the first base of that contig in the reference genome, although the position is generally shifted 20 bp downstream to accommodate positions for the gap filling contigs. The ordered contigs can be merged using SeqMan Pro's **Contig > Align Contigs End-to-End** option. This option ensures that only adjacent contigs are considered for merging, mitigating against false joins caused by repeat elements. With sufficient depth of coverage, this step should close a significant number of gaps. However, some gaps may remain. These may be caused, for example, by long insertions that could not be reliably walked across in this automated fashion. The remaining gaps that require more manual intervention can be closed using the suite of tools in SeqMan NGen and SeqMan Pro.

Evaluation in SeqMan Pro

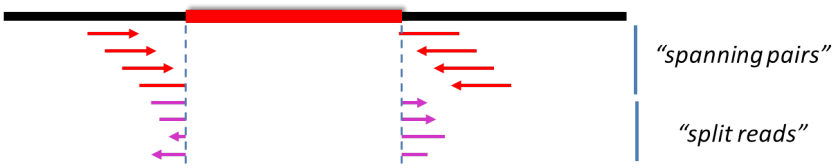
The *.assembly package allows you to inspect the initial reference-guided assembly and detected SV events via SeqMan Pro's Structural Variation table. Single nucleotide polymorphisms (SNPs) and small insertions and deletions can also be inspected using the SNP table.

The following images show Steps 1-4 for deletions and three types of insertions. The terms "split reads" and "junction reads" are used in some of these images and are defined as:

- **Split reads** – Reads in which the first portion matches one location in the genome, and the adjacent portion matches a downstream location on the same strand. The endpoint of the first segment and the start point of the second segment generally define the breakpoints of the deletion, although the exact positions may vary by a few bases in some cases. The presence of multiple "split reads" at a given position is required to avoid spurious splits caused by, for example, micro-repeats in the genome.
- **Junction reads** – Mate pair reads where one read aligns either upstream or downstream of the structural variant, and its mate aligns either on the other side ("spanning pairs") or within the new sequence (in the case of insertions and indels). In the latter case, reads within the inserted sequence are identified from the unassembled read pool by virtue of their mates in the assembly.

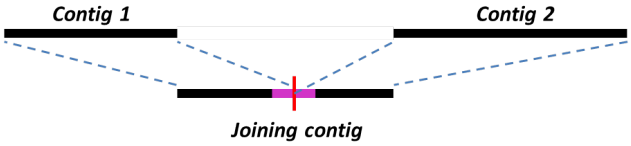
Deletions

1) Detect deletion site



2) Split at breakpoints ("edges")

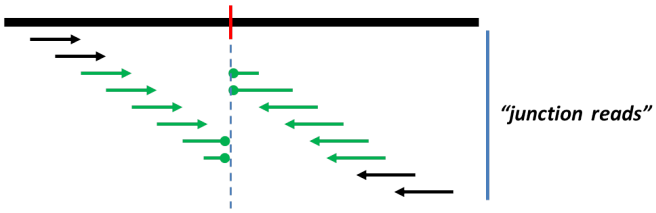
3a) Collect and *de novo* assemble spanning paired and split reads into joining contig
3b) Position joining contig in main assembly



4) Align contigs end-to-end in SeqMan Pro

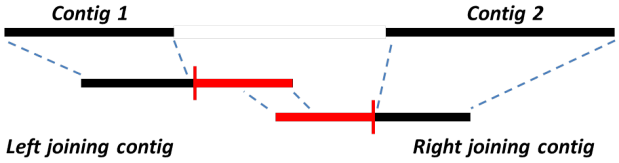
Insertions shorter than the insert size

1) Detect insertion site



2) Split reference at breakpoint ("edge(s)")

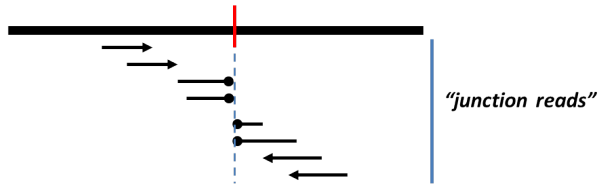
3a) Collect and *de novo* assemble junction reads and their mates into joining contigs
3b) Position joining contigs in main assembly



4) Align contigs end-to-end in SeqMan Pro

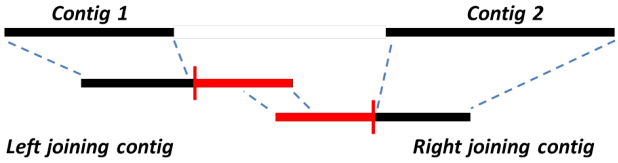
Insertions shorter than twice the insert size

1) Detect insertion site



2) Split reference at breakpoint ("edge(s)")

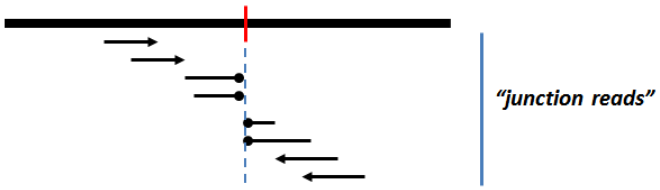
3a) Collect and *de novo* assemble junction reads and their mates into joining contigs
3b) Position joining contigs in main assembly



4) Align contigs end-to-end in SeqMan Pro

Insertions longer than twice the insert size

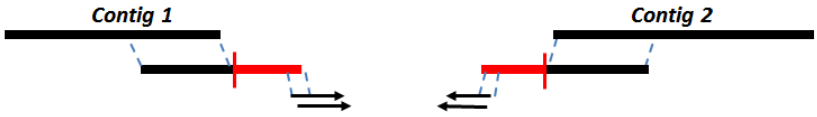
1) Detect insertion site



2) Split reference at breakpoint ("edge(s)")

3a) Collect and *de novo* assemble junction reads and their mates into joining contigs
3b) Position joining contigs in main assembly

3c) Utilize reads from unassembled read pool to "walk" across gap



4) Align contigs end-to-end in SeqMan Pro

Metagenomics/16S rRNA Workflows

The **Metagenomics/16S rRNA** workflow, one of the selections in the Choose Project Type screen, offers both reference-guided ("templated") and *de novo* assembly options, with and without removal of host DNA. The default parameters for this workflow have been optimized to take into account the short read lengths and presence of repetitive DNA sequences common to metagenomic and 16S rRNA data.

To follow this workflow:

- In the Choose Project Type screen, select **Metagenomics/16S rRNA assembly**. When you make this selection, SeqMan NGen automatically pre-filters Metagenomics/16S data prior to assembly by removing redundant, low-quality sequences.
- In the Choose Assembly Type screen, select from any of the four options: **templated assembly**, **templated assembly with host removal**, ***de novo* assembly**, or ***de novo* assembly with host removal**.
- If you select one of the two options involving host removal, the workflow will include the Input Host Files screen. Many genome packages are available for free download from the DNASTAR website. SeqMan NGen will remove host DNA first, then assemble the remaining data using the method you specified (templated or *de novo*).
- If you select a templated option, the workflow will include the Input Biome Genomes screen. Reference sequences can be downloaded from any 16S rRNA database, such as Silva, Greengenes or the Ribosomal Database Project (RDP).
- In the Input Sequences screen, you may enter either single-end or paired-end reads. If available, paired-end reads are recommended for highest accuracy.

Viral-Host Integration Workflows

The **Viral-Host Integration** workflow, chosen via the Choose Project Type screen, is a special type of assembly used to locate putative viral insertion sites.

To follow this workflow:

- In the Choose Project Type screen, select **Viral-Host Integration** . When you make this selection, SeqMan NGen automatically sets up a templated assembly that is optimized for locating viral insertion sites.
- In the Input Host Files screen, input one or more host files. Many genome packages are available for free download from the DNASTAR website.
- In the Input Viral Genomes screen, add the viral genome(s) of interest.

- In the Input Sequence Files screen, input your sequencing reads. These should consist of the virus-infected host DNA for which you wish to determine likely viral insertion sites.

Since chimeric reads (sequences consisting of both host and viral DNA) usually indicate viral insertion sites, SeqMan NGen looks for chimeric reads in a three-step process:

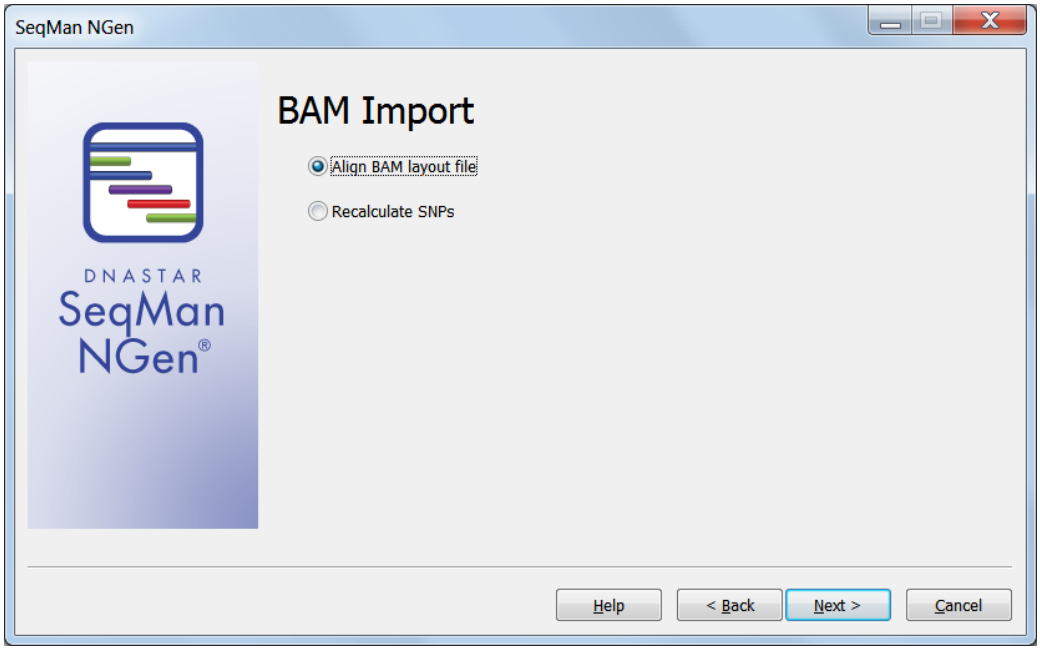
1) The viral genome is used as the initial assembly template.
2) The sub-set of reads that mapped to the viral genome is then re-assembled against the host template.
3) The host template assembly results are output in BAM file format.

To explore possible viral insertion sites, launch SeqMan Pro and view the Coverage Reports for the individual contigs (**Contig > Coverage Report**).

During both templated assembly steps, SeqMan NGen "masks" (trims) whichever half of the chimeric read does not match the template for that step. Use the Coverage Report to navigate to positions with multiple reads, as evidenced in the depth column. The reads at these positions should be trimmed to the same base indicating the insertion site. You may "untrim" the reads to verify that they also contain viral sequence.

BAM Import

The BAM Import dialog allows you to choose whether you wish to align against a BAM layout file or recalculate SNPs for an existing alignment. After you make a selection, SeqMan NGen will populate the rest of the wizard with appropriate default parameters for your assembly.

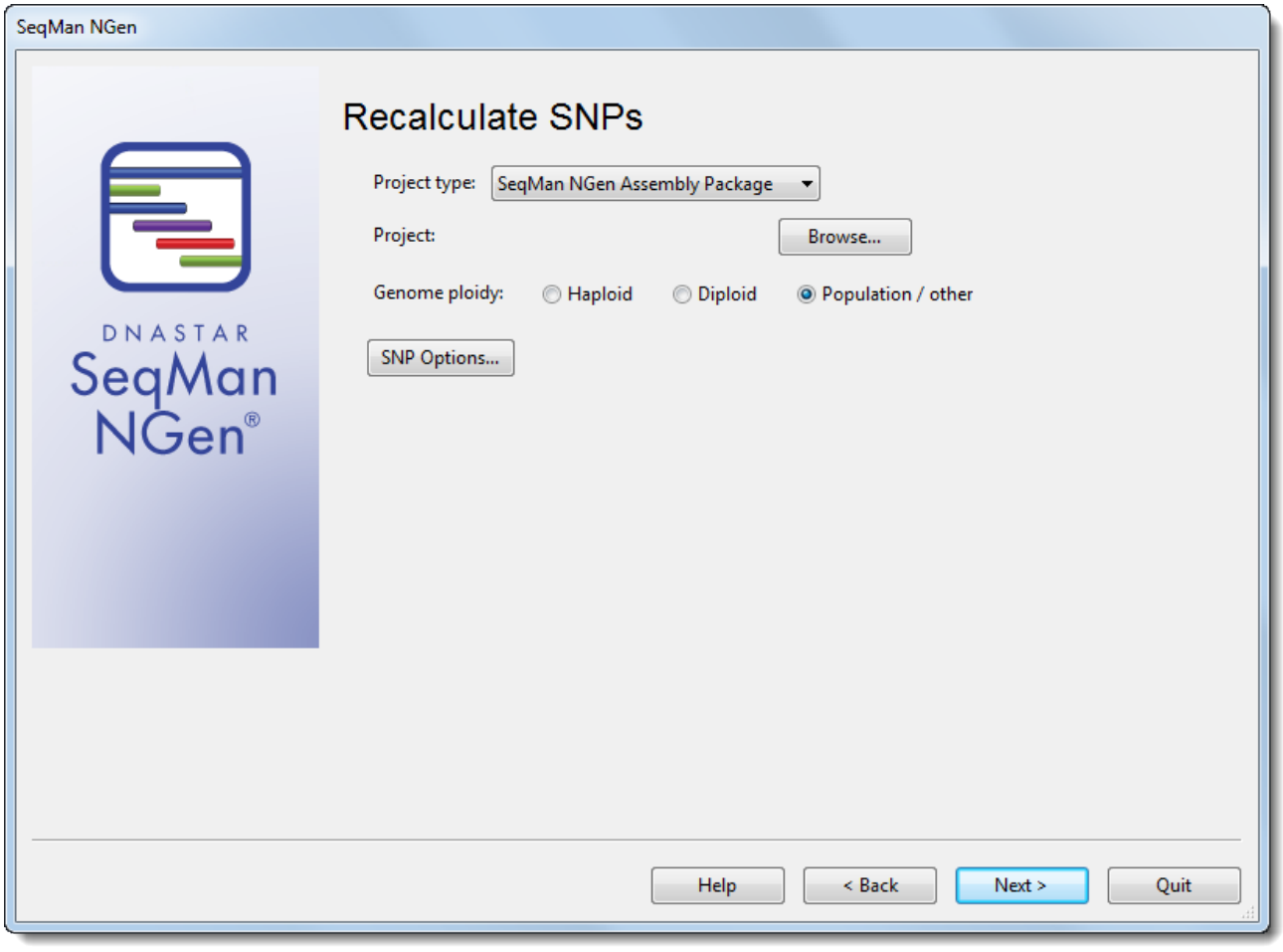


- **Align BAM layout file** – To assemble reads against a BAM-format template. This also gaps a BAM file that is not already gapped. SNPs are calculated automatically as part of the alignment process.
- **Recalculate SNPs** – To calculate SNPs for a finished assembly.

Once you are finished, click **Next >** to continue to the next wizard screen.

Recalculate SNPs

If you are recalculating SNPs for a BAM assembly, you must choose a project type and location from the Recalculate SNPs dialog.



- **Project type** – Choose **SeqMan NGen Assembly Package** if you wish to recalculate SNPs on an existing assembly package. SeqMan NGen will recalculate SNPs into the existing package. Otherwise, choose **BAM Project** and use the **Browse** button to designate a BAM file.

- **Project** – If you select a **SeqMan NGen Assembly Package** as the project type, you must use the **Browse** button to select the existing assembly. If you selected **BAM Project** in the drop-down menu, you must use the **Browse** button to select the BAM file.

  Note: The BAM file *must* be fully gapped.

- **Genome ploidy** – Select the type of ploidy for your project. Choosing **Haploid** or **Diploid** establishes the statistical model SeqMan NGen will use in estimating probabilities during SNP calls. Selecting **Population / other** (e.g. for a polyploid genome or if doing a population study) causes SeqMan NGen not to calculate probabilities.

If desired, click the **SNP Options** button to open the SNP Options dialog. This dialog allows you to view and edit options for recalculating SNPs.

Once you are finished, click **Next >** to continue to the next wizard screen.

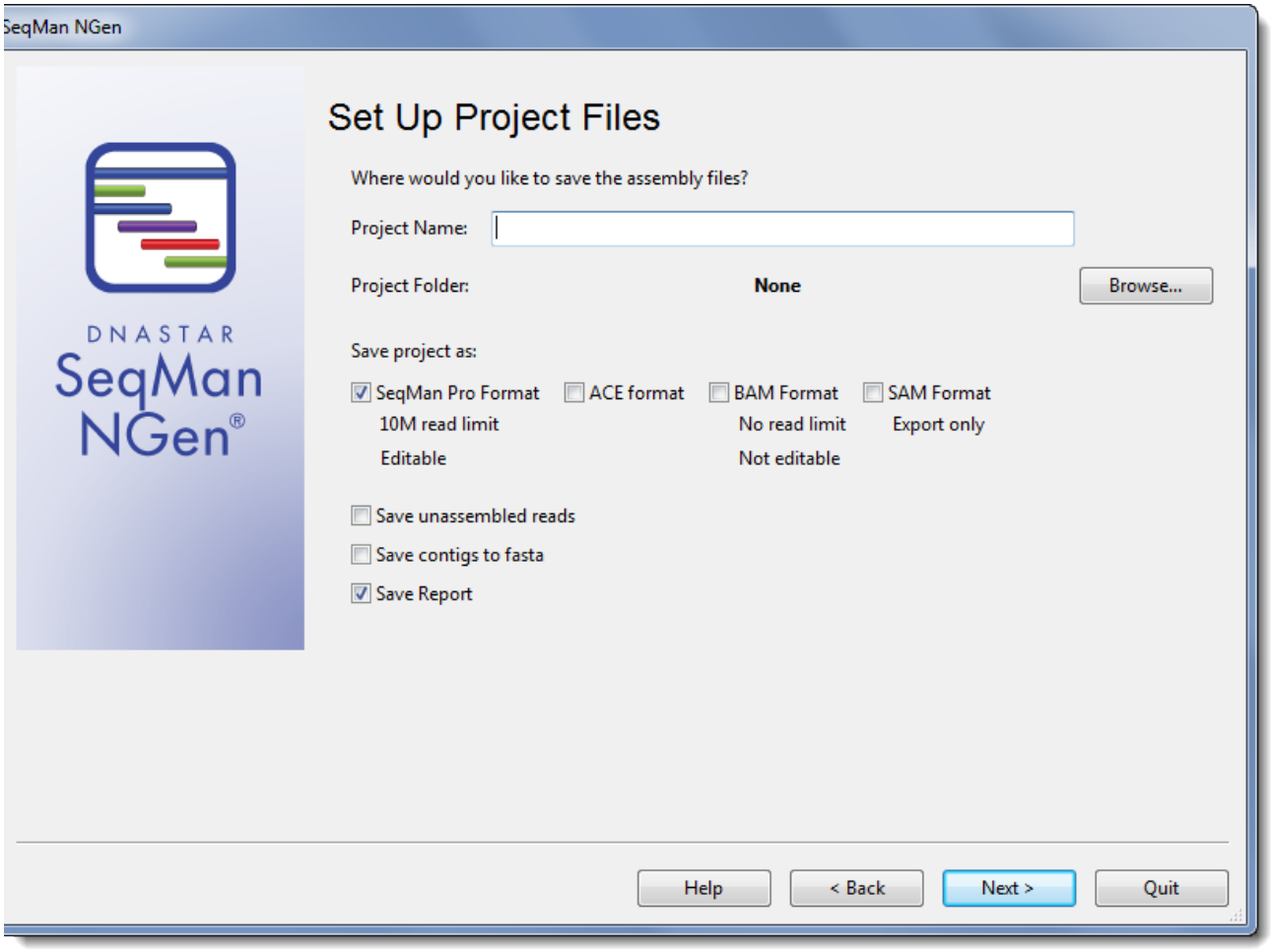
Set Up Project Files

You must select a name and location for your project in the Set Up Project Files dialog before proceeding further in the wizard. There are two versions of this dialog depending upon your choices in previous wizard screens. See the links below for detailed information.

Set Up Project Files (*De Novo*, Special Templated)

You must select a name and location for your project in the Set Up Project Files dialog before proceeding further in the wizard. The following version of the dialog is shown only when you are following the *de novo* or special templated workflows.

Note: For other workflows, see Set Up Project files (All Others).



There are two mandatory fields in this dialog:

- **Project name** – Enter a name for all output files, including the finished assembly. By default, alignment files are saved in SeqMan Pro (*.sqd) format.
- **Project folder** – Use the **Browse** button to select a location for your assembly output files. Required disk space may range from 1 GB to 5 TB, depending on a variety of factors. See our technical requirements page for more information.

Note: Never save the assembly output files directly to the desktop, as the many intermediate files and folders created during assembly may hamper or prevent further computer operations. However, files may be saved to a folder on the desktop.

The following checkboxes let you request additional output files. These will all have the name and location specified above, but different file extensions:

- **Save project as** – Check one or more boxes to save assembly output files in these formats:

Format Type	Extension	Editable?	Viewable in SeqMan Pro?	Read Limit
SeqMan Pro format [1]	*.sqd	Yes	Yes	10 million
ACE format	*.ace	Yes	Yes	10 million
BAM format	*.bam	No	Yes	None
SAM format [2]	*.sam	Export only	No	None

[1] If your assembly exceeds the read limit size for SeqMan Pro format, it will automatically be saved in BAM format instead.

[2] The SAM format option is grayed out (unavailable) if you selected ***De novo* assembly** in the Choose Assembly Type dialog.

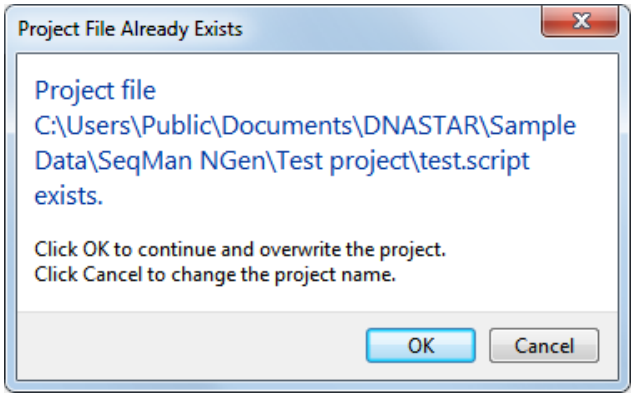
- **Save unassembled reads** – Check the box to save all sequences that were not assembled in the project as a multi-sequence *.fastq file. A default quality score of 15 will be given to each base.
- **Save contigs to fasta** – Check the box to save the consensus sequences from each contig in the assembly as a multi-sequence *.fasta file.
- **Save Report** – Check the box to save an assembly report text file.

Note for all users: Choosing SeqMan Pro format in **Save project as** causes all report information to be saved within the SeqMan Project file (*.sqd), even if you do not check the box to save the report separately. To view the report in SeqMan Pro, choose **Project > Report**.

Note for Windows users: To open a text report with the correct formatting displayed, we recommend using Wordpad, Notepad++, or Microsoft Excel®, and not the default Windows text editor, Notepad.

Once you are finished, click **Next >** to continue to the next wizard screen.

If you choose a name that already exists in the chosen location, you will receive the following warning.

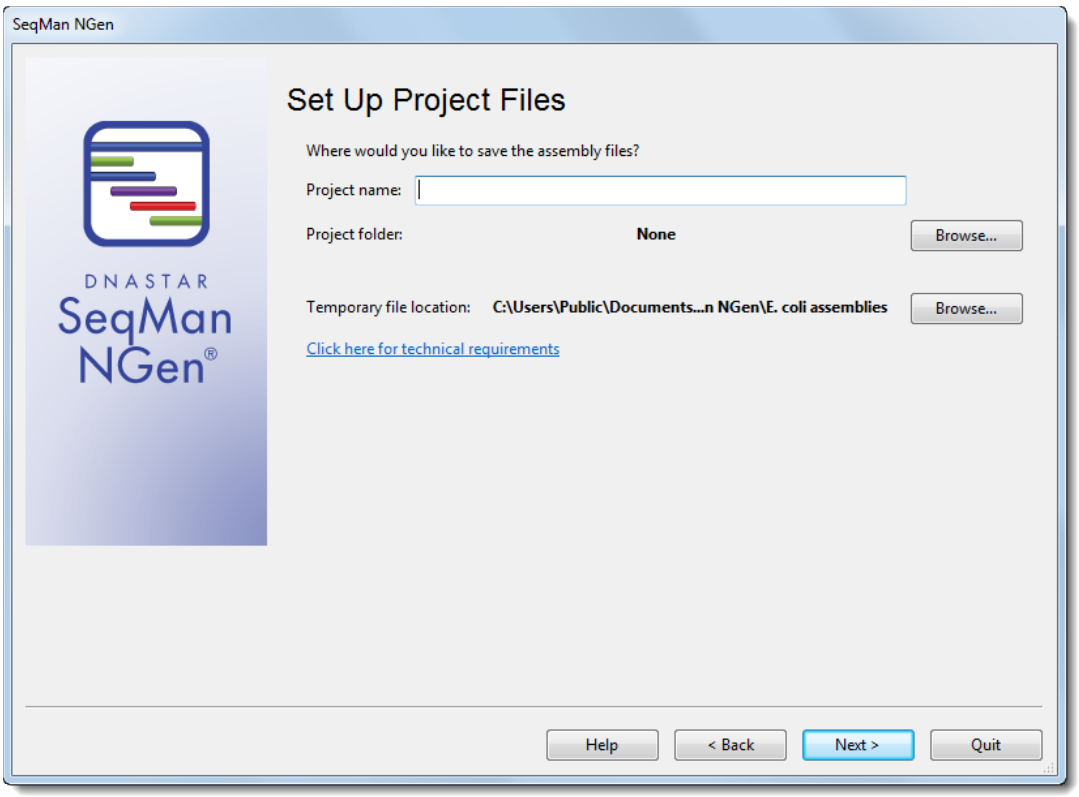


Click **OK** to continue and over-write the earlier project; or **Cancel** to return to the wizard screen, where you may change the project name and/or location.

Set Up Project Files (All Others)

You must select a name and location for your project in the Set Up Project Files dialog before proceeding further in the wizard. The following version of the dialog is shown for all workflows other than the *de novo* or special templated workflows.

Note: For *de novo* or special templated workflows, see Set Up Project Files (*De Novo*, Special Templated).



- **Project name** – Enter a name for all output files, including the finished assembly. The finished assembly will be saved in BAM format.
- **Project folder** – Use the **Browse** button to select a location for your assembly output files. Click the link below for information about disk space requirements.
- **Temporary file location** – Use the **Browse** button to designate a location for the intermediate files produced during assembly. We recommend using an external hard drive as the temporary file location. SeqMan NGen will remember and use the temporary file location for future assemblies.

Note 1: Never save the assembly output files or temporary files directly to the desktop, as the many intermediate files and folders created during assembly may hamper or prevent further computer operations. However, files may be saved to a folder on the desktop.

Note 2: By default, most temporary files are deleted when the assembly is complete. Other files (e.g., [template_name].FasInfo.sqlite and [template_name].mer) may remain in the temporary file location in order to facilitate efficient reassembly of data in the future.

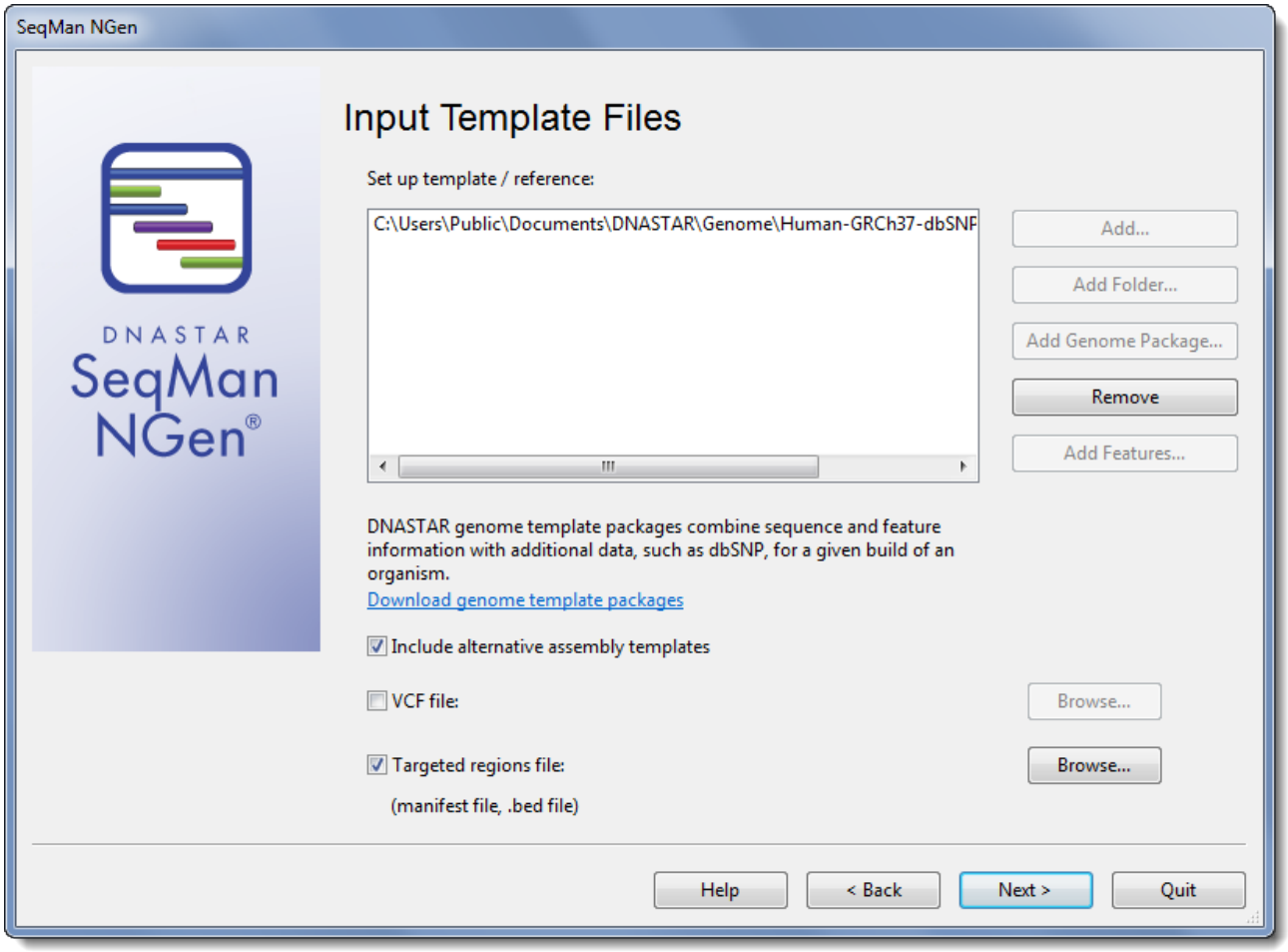
Use the link "Click here for technical requirements" to open a DNASTAR web page describing technical requirements for reference-guided and *de novo* assemblies.

Once you are finished, click **Next >** to continue to the next wizard screen.

Input Template/Host Files

If you are doing a Viral-Host Integration or a Metagenomics/16S rRNA workflow, this screen is called "Input Host Files." For other templated assembly workflows, it is named "Input Template Files." You must enter one or more template (reference) or host sequences here before proceeding further in the wizard.

Note: If you wish to manually specify an isoform to use in SNP calling, you will need to perform a minor edit to the template sequence before adding it here. See Manually Specify an Isoform for more information.



Depending upon your workflow, there will be between three and five buttons on the right of the dialog.

- **Add** – Click to navigate to and select one or more individual sequence(s) to use as template or host sequences.

  Note: Before adding a reference file to certain projects, you may wish to first annotate it in SeqBuilder for known SNPs/variations and other features.

  SeqMan NGen supports the following file formats for reference and host sequences.

Format	Extension	Notes
DNASTAR SEQ files	*.seq	Feature-containing files can be used directly as input for both the sequence and feature annotation.
GenBank files	*.gbk	GenBank flat files be used directly as input for both the sequence and feature annotation.
FASTA	*.fas, *.fna, *.fasta, *.txt	Can be used with or without an associated *.gff annotation file.
General Feature Format sequence files	*.gff	Sequence-containing GFF files can be used directly as input for both the sequence and feature annotation.
Genome template packages	*.genometemplatepackage	This option is available for templated assemblies only. See "**Note**" (at beginning of this topic) and "**Add Genome Package**" (below) for details on downloading and working with genome packages.

- **Add Folder** – Click to navigate to and select an entire folder of sequences to use as template or host sequences. All sequences within the specified folder will be added. After adding files using the **Add** or **Add Folder** buttons, the **Add Genome Package** button will be grayed out.
- **Add Genome Package** – Click this button to browse to an extracted genome package on your hard drive. If you are following a templated workflow and wish to use DNASTAR's database association features, you must input one of the DNASTAR genome packages at this step.

  Note: If you haven't yet downloaded and/or extracted a genome package, learn how to do so in Downloading and Extracting a Genome Package.

  Your downloaded genome package will be utilized by SeqMan NGen, but the actual genome files will remain in their original locations on your hard drive. After adding files using the **Add** or **Add Folder** buttons, the **Add Genome Package** button will be grayed out.

- **Remove** – Click to remove a selected (highlighted) file from the list.
- **Add Features** – (This option is not available in the viral-host integration or Metagenomics/16S rRNA workflows.) Click to add separate *.gff feature files. Feature files are not displayed in the reference sequence window.

- **Include alternative assembly templates** check box - If you added a genome package as your template, this new option will appear near the bottom of the dialog. Check the box to include alternate sequence (alt loci) representation for variant regions. See the "Variation" section of this Genome Reference Consortium announcement for details. Alternate sequences include those known to be in a particular chromosome, but whose exact position is unknown; and sequences with known positions, but otherwise incomplete entries.
- **VCF file** check box (not visible in all workflows) - Certain SeqMan NGen workflows allow you to import a custom VCF SNP file (e.g., created in SeqMan Pro or ArrayStar) with data from one or more assemblies. To add a VCF file, check the box next to **VCF file** and then use the corresponding **Browse** button to navigate to the file. If you elect to do this, positions within the VCF file will be given a VCF SNP ID during the assembly process. After assembly, information about each position can be viewed in the SeqMan Pro SNP Report. (Within SeqMan Pro, choose **SNP > SNP Report**.)

Note 1: SeqMan NGen only supports one VCF file per assembly project. If you have multiple VCF files (e.g., one per chromosome), you will need to merge the information into a single VCF file before browsing to the file. For more information, see Make a Custom VCF File.
Note 2: If you selected **Templated assemblies with control** from the Choose Assembly Type screen, the VCF import option is not visible in this screen. Instead, VCF files are imported from the Set Up Experiments screen.

- **Targeted Regions file** check box - If you chose **Exome assembly**, **Mendelian/germline gene panel assembly**, or **Cancer/somatic gene panel assembly** from the Choose Project Type screen, you cannot proceed to the next screen until you have specified a file containing targeted region information. Click the **Browse** button to the right of the **Targeted Regions file** checkbox (only visible in the workflows listed) and navigate to a BED or manifest file. BED files must have the extension *.bed. Manifest files can have various extensions, but must be in the correct format.

Once you are finished, click **Next >** to continue to the next wizard screen.

Downloading and Extracting a Genome Package

Genome template packages can be downloaded and added via the Input Template/Host Files wizard screen.

To download a genome package:

Click the Download genome template packages link at the bottom of the wizard screen. This opens a DNASTAR web page from which you can download packages from a variety of genomes for free. Each package contains the template sequence, annotations, and associated dbSNP linking information. Human genome packages also contain GERP scores and COSMIC linking information.

Downloaded genome packages are saved on your computer as ZIP files, and must be extracted prior to use.

To extract a downloaded genome package:

- On **Macintosh**: Double-click on the ZIP file. The files will be automatically extracted via the Archive Utility.
- On **Windows 7** & **Windows 8**: Double-click on the ZIP file. In the ensuing Explorer window, click **Extract all files** from the top left. Choose a location for the files and select **Extract**.

See Input Template/Host Files for instructions on adding the genome package to SeqMan NGen.

Note: SeqMan NGen can read and produce output using a variety of common chromosome naming conventions, including "chr1" and "ch1," as well as Arabic and Roman numerals. Chromosome names are captured from genome template packages and used to assign contig IDs to entries from BED, VCF and manifest files.

Annotating Template Sequence Prior to Assembly

Using annotated template sequences in SeqMan NGen may enable you to better analyze the identified putative SNPs when viewing your assembled project in SeqMan Pro. If desired, annotate your template sequence in SeqBuilder (the Lasergene application for sequence editing and visualization) prior to adding it to the Input Template/Host Files dialog.

1) Launch SeqBuilder.
2) Go to **File > Open** and select your template sequence.
3) Select the range of sequence where a feature will be added. (Use **Edit > Go to Position** to navigate quickly up and down your sequence.)
4) Go to **Features > New Feature.** A new "misc_feature" will be added to your sequence and displayed in the Feature List.

5) Click on “misc_feature” from within the Feature List and select the appropriate feature type from the list provided**.** For example:

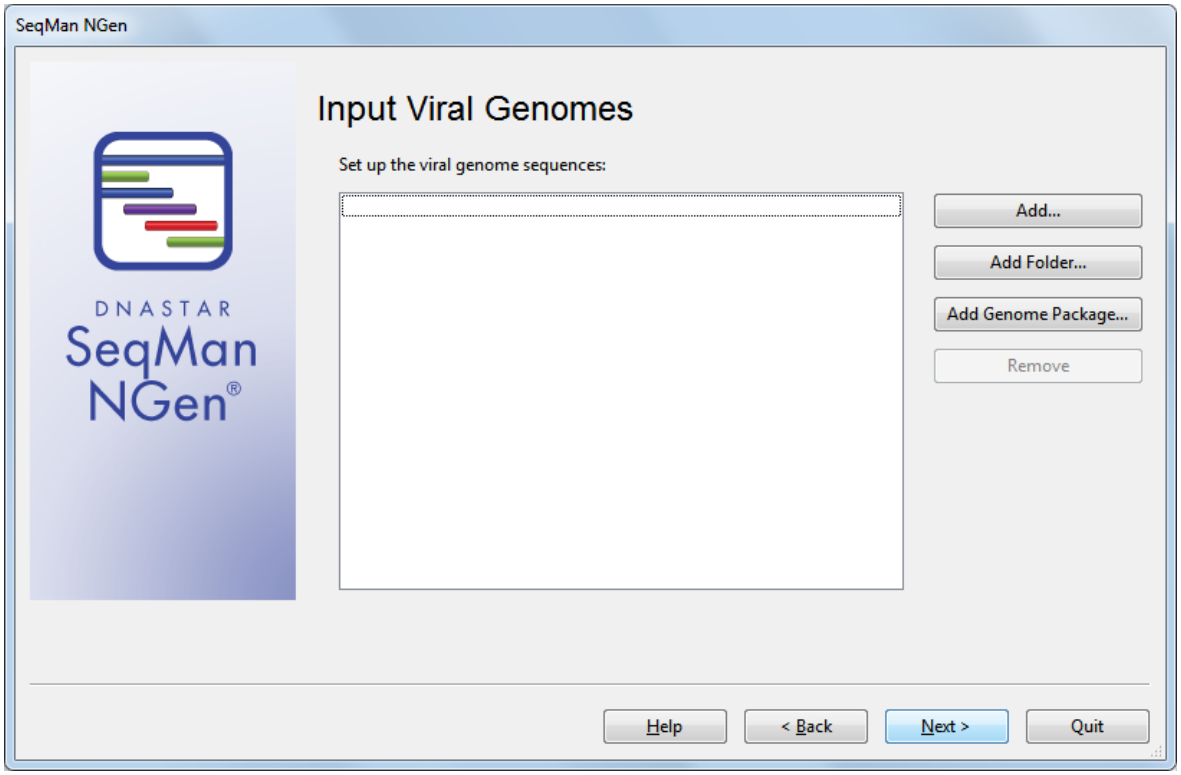
   - For SNPs, choose **Variation > variation**.
   - For exons, choose **Gene > exon**.
   - For CDS features, choose **Transcript > CDS**.
   - For origin of replication, choose **Structure > rep_origin**.

   Note: The next feature you create will automatically be of the same feature type you just selected, enabling you to create all the features of one type more quickly.

6) Repeat steps 3-5 until all of your features have been added. Then go to **File > Save As** and save your sequence in *.sbd, *.seq or *.gbk format. Your annotated template sequence is now ready for assembly in SeqMan NGen and subsequent analysis in SeqMan Pro.

Input Viral/Biome Genomes

In viral-host integration workflows, this screen is named "Input Viral Genomes." In Metagenomics/16S rRNA workflows, it is called "Input Biome Genomes." Other than the name, the wizard screens are identical. You must enter one or more reference sequences before proceeding further in the wizard.



- **Add** – Click to navigate to and select one or more individual genomes. SeqMan NGen supports the following file formats: *.seq, *.gbk, *.fas, *.fna, *.fasta, *.txt, *.gff and *.genometemplatepackage.

  Note: If you are following the Metagenomics/16S rRNA workflow, we recommend inputting a biome genome that is in Fasta, rather than GenBank, format. If you use a GenBank file, SeqMan NGen may run out of memory parsing all the features in all the templates. If your biome genome is currently in GenBank format, use the Convert File Type template in DNASTAR's SeqNinja utility to convert it to Fasta format before importing it into SeqMan NGen.

- **Add Folder** – Click to navigate to and select an entire folder of genome sequences. All sequences within the specified folder will be added.

  Note: Once you have added files using the **Add** or **Add Folder** buttons, the **Add Genome Package** button will be grayed out.

- **Add Genome Package** – Click this button to browse to the extracted genome package on your hard drive (see Downloading and Extracting a Genome Package). Your downloaded genome package will be utilized by SeqMan NGen, but the actual genome files will remain in their original locations on your hard drive.

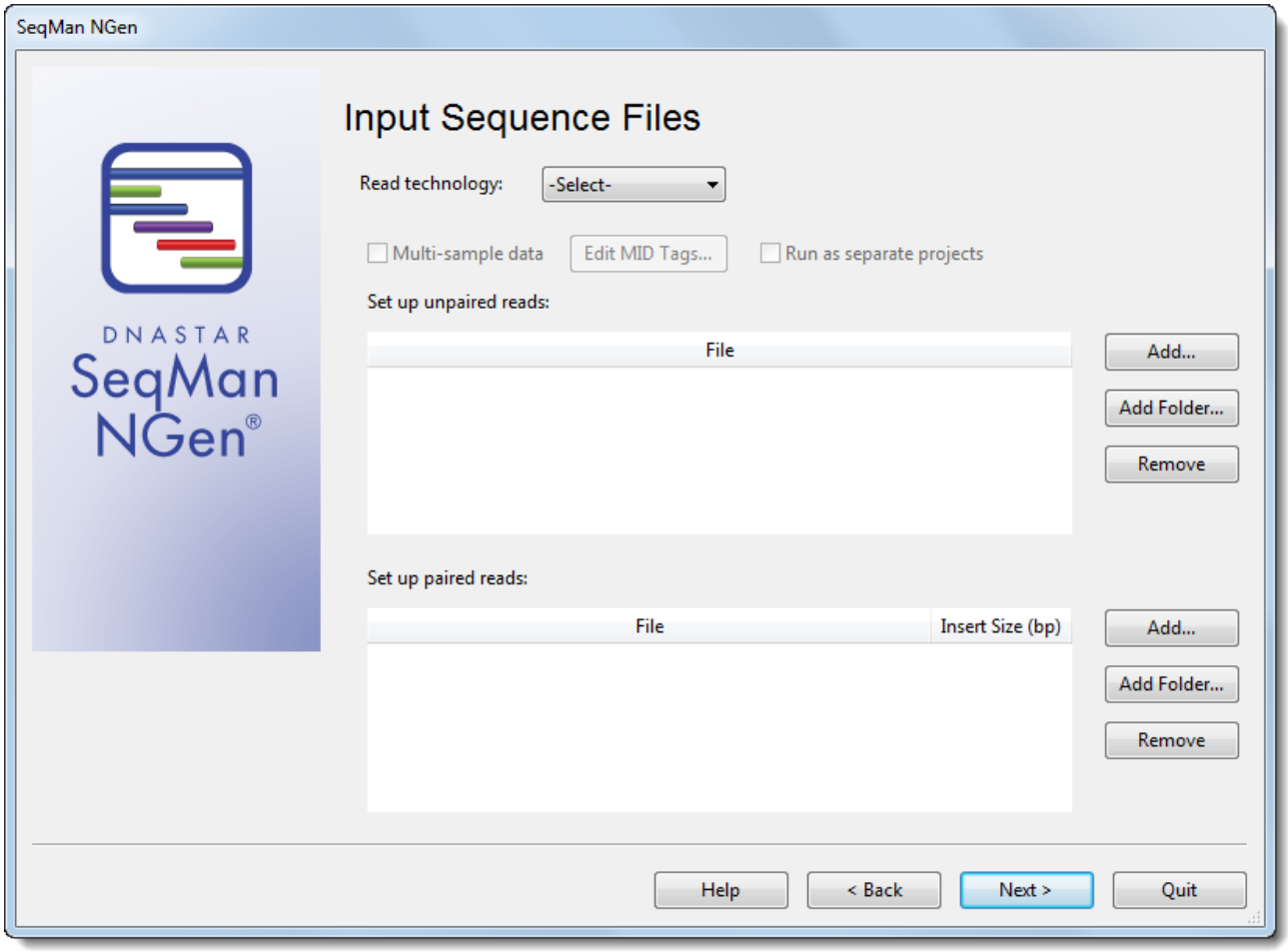
  Note: Once you have added an extracted genome package using the **Add Genome Package** button, the **Add** and **Add Folder** buttons will be grayed out.

- **Remove** – Click to remove a selected (highlighted) file from the list.

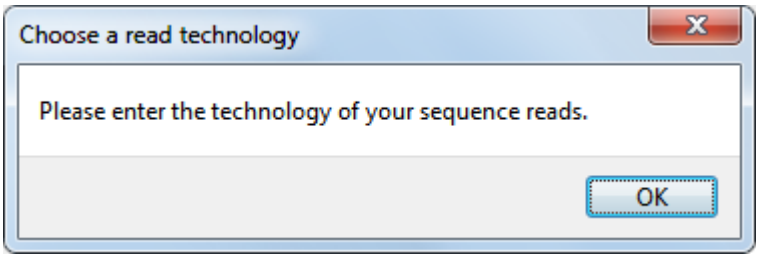
Once you are finished, click **Next >** to continue to the next wizard screen.

Input Sequence Files

You must choose a sequencing technology and enter one or more read files in the Input Sequence Files dialog before continuing with the wizard.



1) You must select a **Read technology** from the drop-down menu before proceeding to the next screen. If you do not make a selection here before clicking **Next**, you will receive the following reminder:



Choose from **Illumina < 50 nt**, **Illumina > 50 nt**, **454**, **Ion Torrent**, **Pac Bio, Sanger** (not available in all workflows) or **Other**. The default values for parameters and other assembly options in subsequent panels will be based on your selection.

The following notes refer to specific workflows or data types:

- If you are doing a reference-guided assembly with gap closure, you must select either **Illumina > 50 nt**, **454** or **Ion Torrent**.

- For *de novo* assemblies, if you select **Illumina > 50 nt** and enter an insert size of 150 bp or less in the **Set Pair Information** dialog, the assembler will assume the reads overlap and will attempt to create a single "super-read" from each pair. Read pairs that cannot be merged, either because they do not overlap or have numerous errors in the overlapping region, will not be included in the assembly.
- **Sanger** is only included in the menu if you are doing a *de novo* assembly.
- Both types of **Ion Torrent** paired reads—"mate pairs" and "paired ends"—are supported.

2) If you are using multi-sample data, check the box to the left of **Multi-sample data**, then click **Edit MID Tags** to open a related customization dialog: Edit Group Names for Illumina, PacBio and Ion Torrent data; Edit MID Tags for 454 data. (Note that **Multi-sample data** and **Edit MID Tags** are disabled for certain project types). After you enter customization information, SeqMan NGen will then produce combined assemblies with data organized and analyzed on a per sample basis.

Note: In assembling multi-sample data, SeqMan NGen considers all samples together. This can affect the final gapped alignment and therefore potentially yield slightly different results than assembling each sample individually.

---

3) If you want to separate the multi-sample data and run them as separate projects, check the **Run as separate projects** box. The same reference sequence(s) and parameters will be used for all the projects. This option is disabled for 454 read technology and for certain project types.

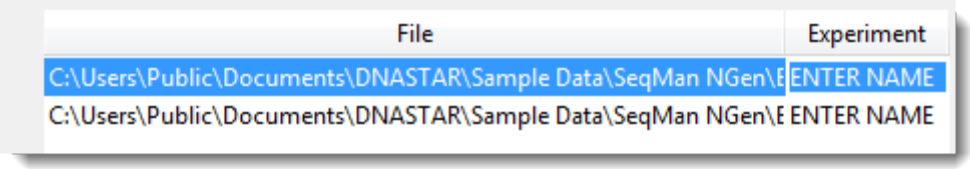
4) Set up unpaired or paired reads using the buttons on the right of the screen. Assemblies can include both single and paired end read files. Single ended files should be added to the top pane of the window, while paired end files should be added to the bottom pane. In both cases, files may be added individually or in folders.

   - **Add** – Click to navigate to and select one or more individual sequence(s) for your assembly project. See the SeqMan NGen Supported File Types page for supported file formats.
   - **Add Folder** – Click to navigate to and select an entire folder of sequences. All sequences within the specified folder will be added.
   - **Remove** – Click to remove a selected (highlighted) file from the list.

   If you are doing a reference-guided assembly with gap closure, you must enter at least one set of paired end read files before you can proceed to the next wizard screen.

   If paired end data are added, the Set Pair Information dialog pops up automatically. See Set Pair Information (Certain Sanger Data) or Set Pair Information (All Others) for more information. Once you have entered pair information, the insert size that you input will appear in the "Insert Size (bp)" column.

If you chose **Templated assemblies with control** from the Choose Assembly Type screen, adding either unpaired or paired reads will cause a new column, "Experiment," to appear in the sequence file area.



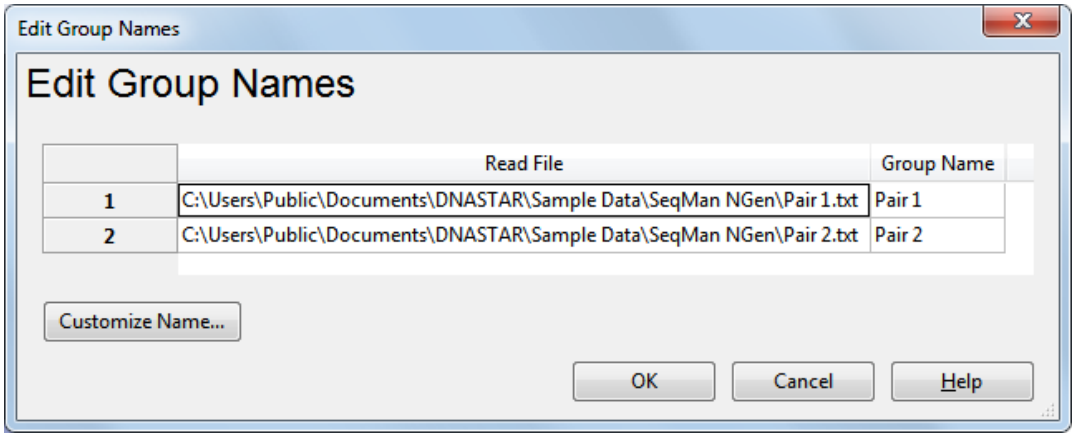
Each of the Experiment cells initially contains the text "ENTER NAME." Double-click in each cell and type in a name for that experiment. Data files with the same Experiment name will be assembled together. You will not be allowed to proceed to the next wizard screen until you have entered the experiment information.

Once you are finished, click **Next >** to continue to the next wizard screen.

Edit Group Names

The Edit Group Names dialog opens when you specify a multi-sample data set in the Input Sequence Files dialog, and then click the **Customize Sample Names** button. Use this dialog to define custom names for sorting and displaying individual sample sets in SeqMan Pro.

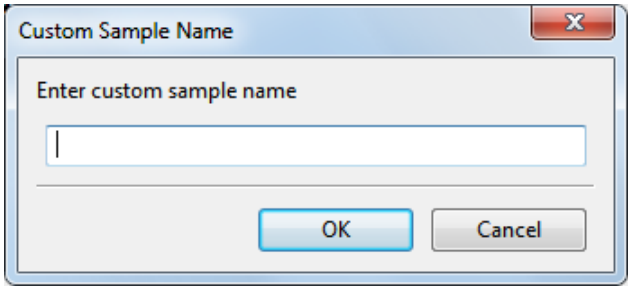
Note: If you chose **454** as the **Read Technology** in the Input Sequence Files dialog, you will instead see the Edit MID Tags dialog.



The **Read File** column contains individual read file names, while the **Group Name** column contains your custom group names.

Ion Torrent, Illumina and PacBio data all require you to enter a **Group Name** for each sequence file before leaving the dialog. Read files will be separated into individual samples files based on these custom names.

To add a custom group name, type it directly in the **Group Name** column. Alternatively, you can select the row to be edited and click **Customize Name**. Type the desired name and click **OK**.

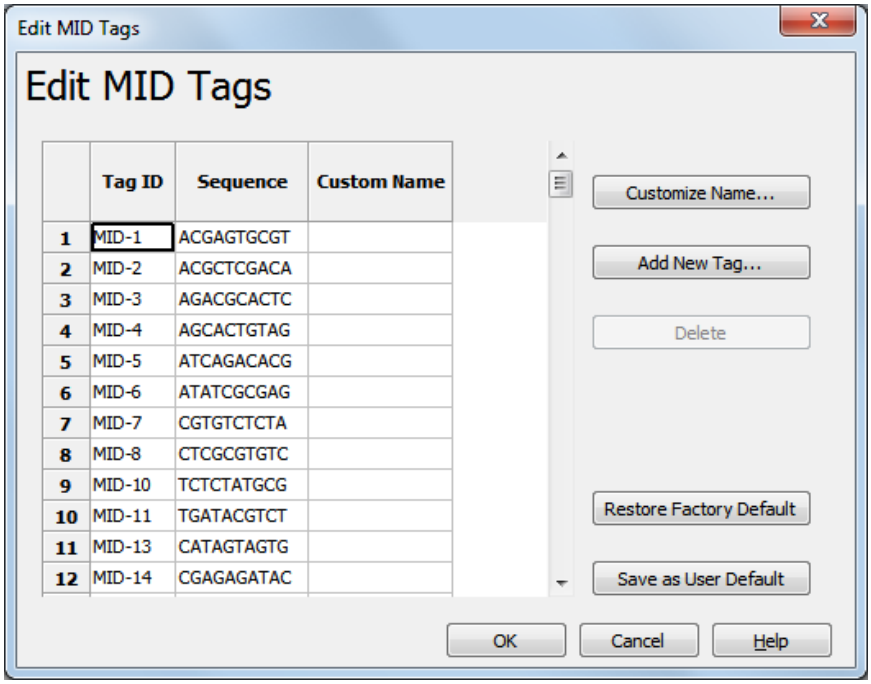


Within a particular project, all the tags must be the same length.

Edit MID Tags

The Edit MID Tags dialog is accessed when you specify a multi-sample 454 data set in the Input Sequence Files dialog, and then click the **Customize Sample Names** button. This dialog allows you to define custom names for sorting and displaying individual sample sets in SeqMan Pro.

Note: If you chose anything other than **454** as the **Read Technology** in the Input Sequence Files dialog, you will instead see the Edit Group Names dialog.



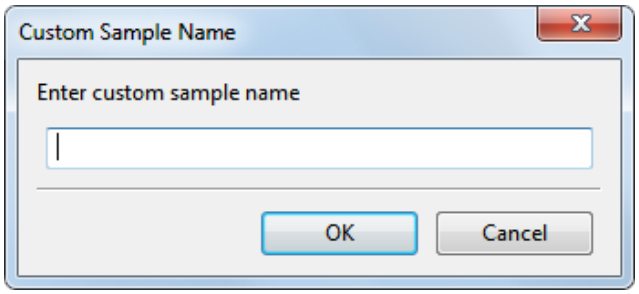
MID, Indexing, and Barcode tags are synonyms for short stretches (5-10 bases) of unique DNA sequence that are added to samples, allowing the samples to be amplified and sequenced as a

pool. After sequencing, the data for individual samples are separated by identifying the tag sequence and sorting the reads into different bins based on their tags.

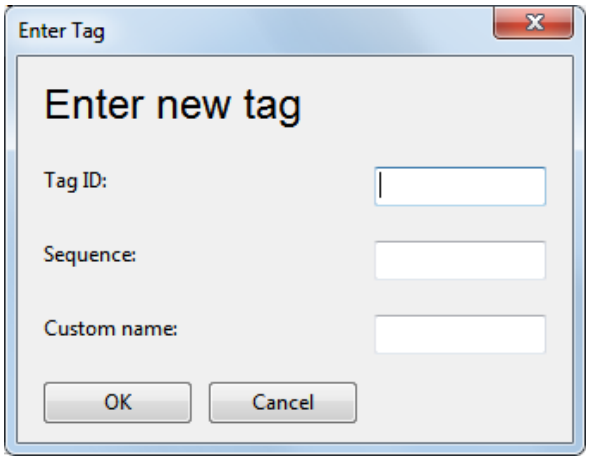
Prior to assembly, SeqMan NGen splits all the reads in a specified file(s) into individual fastq files based on the information in this table. The three columns contain the following information:

- **Tag ID** –contains the 151 Titanium MID identifiers (Roche Technical Bulletin 005-2009).
- **Sequence** – the sequence of the tag.
- **Custom Name** – displays any custom sample names.

To add a custom name for a MID group, select the row to be edited and click **Customize Name**. Type the desired name and click **OK**. Note that within a particular project, all the tags must be the same length.



To add a non-MID tag to the bottom of the list, click **Add New Tag**. Type in the information for all three columns of the table, then click **OK**.



To remove a row from the table, select the row you wish to remove, then press **Delete**.

To restore the default MID tag table, click **Restore Factory Default**. Clicking the Restore Factory Default settings always restores factory defaults, not the user default table.

To save a modified table as the default, click **Save as User Default**. All information in the table, including any custom names, will be saved and used as the initial default table the next time you set up a multi-sample project with the same read technology. Choosing this button will over-write any previously saved default table.

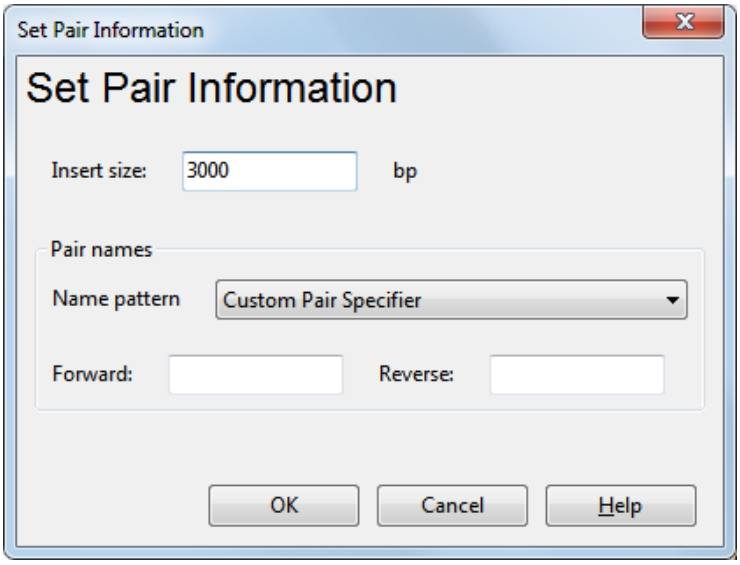
When you are done, click **OK**. During assembly, read files will be separated into individual samples files based on tag information and your own custom names, if provided.

Set Pair Information (Certain Sanger Data)

In the Input Sequence Files dialog, the following version of the Set Pair Information dialog pops up automatically when you fulfill these criteria:

- Any workflow except normal templated + Sanger read technology + paired data

  Note: If you are doing a normal templated workflow **and/or** using non-Sanger paired data, a different version of the dialog appears. See Set Pair Information (All Others) for details.



- **Insert size** – Enter the anticipated distance between paired end reads across the library. SeqMan NGen will use this value to automatically calculate the minimum and maximum insert distances. The default value is 3000 bp.

  Note: During assembly, SeqMan NGen lists this value as a range. If, for example, you enter an **Insert size** of 300, the Assembly Log will list the value as "0 to 450." This convention does not impact assembly results.

- **Name pattern** – In order for NGen to identify Sanger pairs using a sequence naming convention, the convention must systematically distinguish between different pair reads while specifying which pair reads are associated. Forward and reverse sequences must have identical names except for the unique portion that determines the direction of the clone.

  If applicable, select one of the following predefined file naming patterns from the **Name Pattern** dropdown list:

  - sample_f.abi < > sample_r.abi
  - sample100.f_abc.abi < > sample100.r_abc.abi
  - sample_n100.abi <: > sample_f100.abi
  - SAMPL0D1234.abi <: > SAMPL0E1234.abi

  If none of the predefined patterns matches your file naming convention, you may select **Custom Pair Specifier** from the dropdown list, and then manually enter the appropriate expressions for **Forward** and **Reverse** naming conventions.

  Note: Naming conventions should use a subset of regular expressions which utilize elements of the Grep language. For more information, see Example Regular Expressions.

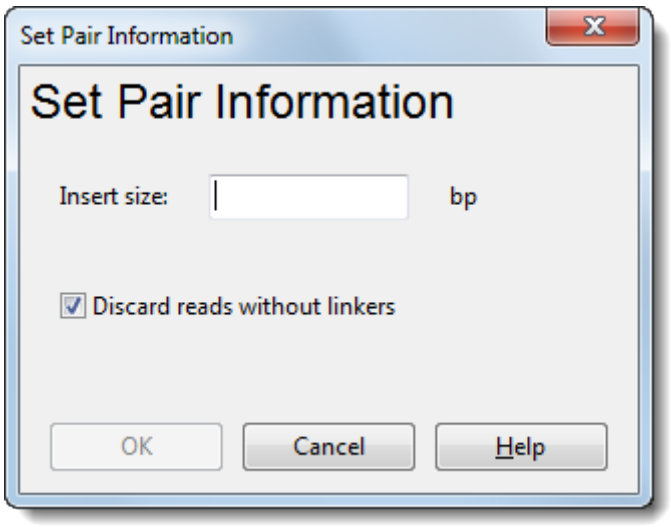
Once you are finished, click **OK** to return to the Input Sequence Files dialog.

Set Pair Information (All Others)

In the Input Sequence Files dialog, the following version of the Set Pair Information dialog pops up automatically when you fulfill either of these criteria:

- Normal templated workflow + any read technology + paired data
- Any workflow except normal templated + non-Sanger read technology + paired data

Note: If you are doing any workflow other than normal templated **and** are using Sanger data, a different version of the dialog appears. See Set Pair Information (Certain Sanger Data) for details.



- **Insert size** – This box may originally be blank or may contain a changeable default value, depending on the read technology you chose. Enter the anticipated distance between paired end reads across the library. SeqMan NGen will use this value to automatically calculate the minimum and maximum insert distances.
- **Discard reads without linkers** – This option only appears in the dialog if you chose **Templated assembly – special workflows** in the Choose Assembly Type screen, and specified a **Read Technology** of **Ion Torrent** in the Input Sequence Files screen. If you input an **Insert size** and leave the box checked, clicking **OK** will launch the following dialog.

Choose between **Standard Linker** and **Custom Linker**. If you choose the latter, you must paste or type the junction linker in the box provided. Click **OK** to return to the Input Sequence Files dialog.

The following information should be considered when making choices in the Set Pair Information dialog:

- During assembly, SeqMan NGen lists this value as a range. If, for example, you enter an **Insert size** of 300, the Assembly Log will list the value as "0 to 450." This convention does not impact assembly results.
- For short inserts containing fewer than 1000 bases, SeqMan NGen sets the minimum size to 0 to catch smaller outliers, which tend to be common. For larger inserts, it sets the minimum to half of the **Insert size**, with the exception of Illumina data, which is set to 0. Long insert Illumina reads have a minimum of 0 because only half the reads consist of long inserts. The other half consist of short inserts (~300 bp), with the short inserts pointing towards one other, and the long inserts pointing away. The 0 value is used by SeqMan NGen's small genome assembler as a flag to account for the undetermined insert size.
- If you specified **Ion Torrent** read technology in the Input Sequence Files dialog and enter a value of 0-799 in the Set Pair Information dialog, SeqMan NGen assumes the library is paired end (small insert). For values ≥ 800, the library is assumed to be mate pair (long insert).

Using Paired End Data

Note: The following information does not apply to the normal templated workflow.

Paired end reads are typically in two files with the forward reads in one file and the reverse reads in the other. SeqMan NGen assumes the pair will be from opposite ends of the same DNA fragment, and sequenced from the end of the fragment inwards.

To add paired reads, go to the Input Sequence Files dialog and **Add** your read files to the lower pane ("Set up paired reads").

To enable SeqMan NGen to identify pairs, a sequence naming convention must systematically distinguish between different pair reads while specifying which pair reads are associated. Forward and reverse sequences must have identical names except for the unique portion that determines the direction of the clone. Expressions for these naming conventions are created using a subset of *regular expressions*, which utilize elements of the Grep language. The following rules apply:

- Two parallel files must use standard naming convention (e.g. s_7_1_sequence and s_7_2_sequence).
- "Forward" and "reverse" reads must be in *exactly* the same order in the two files.
- Both forward and reverse reads must be present for every pair, including pairs where one of the reads failed or is of very low quality.

As an example, forward and reverse Sanger pair files are named as follows: 01f.abi and 01r.abi, where "01" distinguishes that they are members of the same pair. The "f" and "r" at the end of each sequence name distinguishes the orientation.

In Grep, the naming convention would be written as follows:

- Forward convention: (.*)f\..*$
- Reverse convention: (.*)r\..*$

Note: For more information on Grep name patterns, see Example Regular Expressions.

See the links below for read technology-specific information about using paired-end data.

Illumina Pairs

Paired end reads are typically in two files, or a small number of files if they are from multiple runs or lanes. These pairs are specified by a naming convention used in the *.fasta file comment line.

For SNG assemblies (called SMNG in Linux) with paired end reads, SeqMan NGen automatically adds the following information to the script:

```
setPairSpecifier pairs:
{ {
   forward: "(.*)/1"
   reverse: "(.*)/2"
   min: 0
   max: 750
   key: Illumina
} }
```

If reads do not match one of the pair specifiers, or if the forward and reverse specifiers are represented by empty strings (""), SNG will attempt to match using the whole name of the sequence. If exactly two reads have the same name, they will be considered a match.

For XNG assemblies, SeqMan NGen adds the following information:

```
{
   is Pair: true
   file: "****"
```

```
   SeqTech: "Illumina"
   minDist: 0
   maxDist: 750
}
```

For XNG assemblies with paired-end reads, SeqMan NGen recognizes the pairs by their file names. The following examples demonstrate some of the filename formats that SeqMan NGen supports for XNG pairs. Large-bold text in the examples is used to highlight the region of each filename that specifies the forward and reverse reads:

"R_2011_11_21_11_06_08_user_C29-100_PE_DH10B_11_Auto_C29-100_PE_DH10B_11_4120_reverse_**pe2**.fastq",

"R_2011_11_21_11_06_08_user_C29-100_PE_DH10B_11_Auto_C29-100_PE_DH10B_11_4120_forward_**pe1**.fastq",

"Strain1234_L7_**R1**_ATCACG_Index1.fastq",

"Strain1234_L7_**R2**_ATCACG_Index1.fastq",

"K12-1-B_TGACCA_L006_**R1**.fastq",

"K12-1-B_TGACCA_L006_**R2**.fastq",

"GBBC920_GGCTAC_L008_**R1**.filt.50bp.fastq",

"GBBC920_GGCTAC_L008_**R2**.filt.50bp.fastq"

"tiny_**1**.txt",

"tiny_**2**.txt",

"tiny_**1**_sequence.txt",

"tiny_**2**_sequence.txt",

"**tiny1**._qseq",

"**tiny2**._qseq",

"s_1_**1**_sequence.txt"

"s_1_**2**_sequence.txt"

"C29-129_forward_**pe1**.fastq"

"C29-129_forward_**pe2**.fastq"

The Grep used to match the **pairFileNames** is shown below:

```
"(?'name'.*?)_R1_(?'ext'.*)\\.fastq",
"(?'name'.*?)_R2_(?'ext'.*)\\.fastq",
"(?'name'.*?)_R1\\.(?'ext'.*)\\.fastq",
"(?'name'.*?)_R2\\.(?'ext'.*)\\.fastq",
"(?'name'.*?)_forward_pe1(?'ext_p'\\.fastq)",
"(?'name'.*?)_reverse_pe2(?'ext_p'\\.fastq)",
"(?'name'.*?)1\\.fastq",
"(?'name'.*?)2\\.fastq",
"(?'name'.*?)1_sequence\\.txt",
"(?'name'.*?)2_sequence\\.txt",
"(?'name'.*?)1\\.txt",
"(?'name'.*?)2\\.txt",
"(?'name'.*?)1\\._qseq",
"(?'name'.*?)2\\._qseq",
```

The following script command can be used to add support for a new filename format. The command must be executed before assembly. The pattern will be used for all subsequent **assembleTemplate** commands for that run of XNG

```
pairFilePattern
     forward: "(?'name'.*?)_R1_(?'ext'.*)\.fastq"
     reverse: "(?'name'.*?)_R2_(?'ext'.*)\.fastq"
```

Roche 454 Pairs

Paired end reads are provided as a single read containing the pair joined by a linker sequence. When assembling 454 paired end reads, SeqMan NGen will check for the presence of a linker defining the paired end reads. Reads with an identifiable linker are split into forward and reverse reads with the forward read flipped so the traditional orientation is maintained. These reads are

then put into parallel fastq files. SeqMan NGen appends each file name with ***_1*** or ***_2***, following the Illumina paired end convention. The read names themselves are appended with ***__for*** or ***__rev***. In cases where the linker occurs at the end of the read, the linker is removed and a single end read is placed in a file with ***_unpaired*** appended to the name. Reads where no read is detected are also placed in the ***_unpaired*** file. 454 paired end splitting can also be specified through scripting.

Sanger Pairs

Paired end reads are typically all in multiple files with the forward pairs having an "f" or "forward" in the name and the reverse pairs having "r" or "reverse" in the name.

Example Regular Expressions

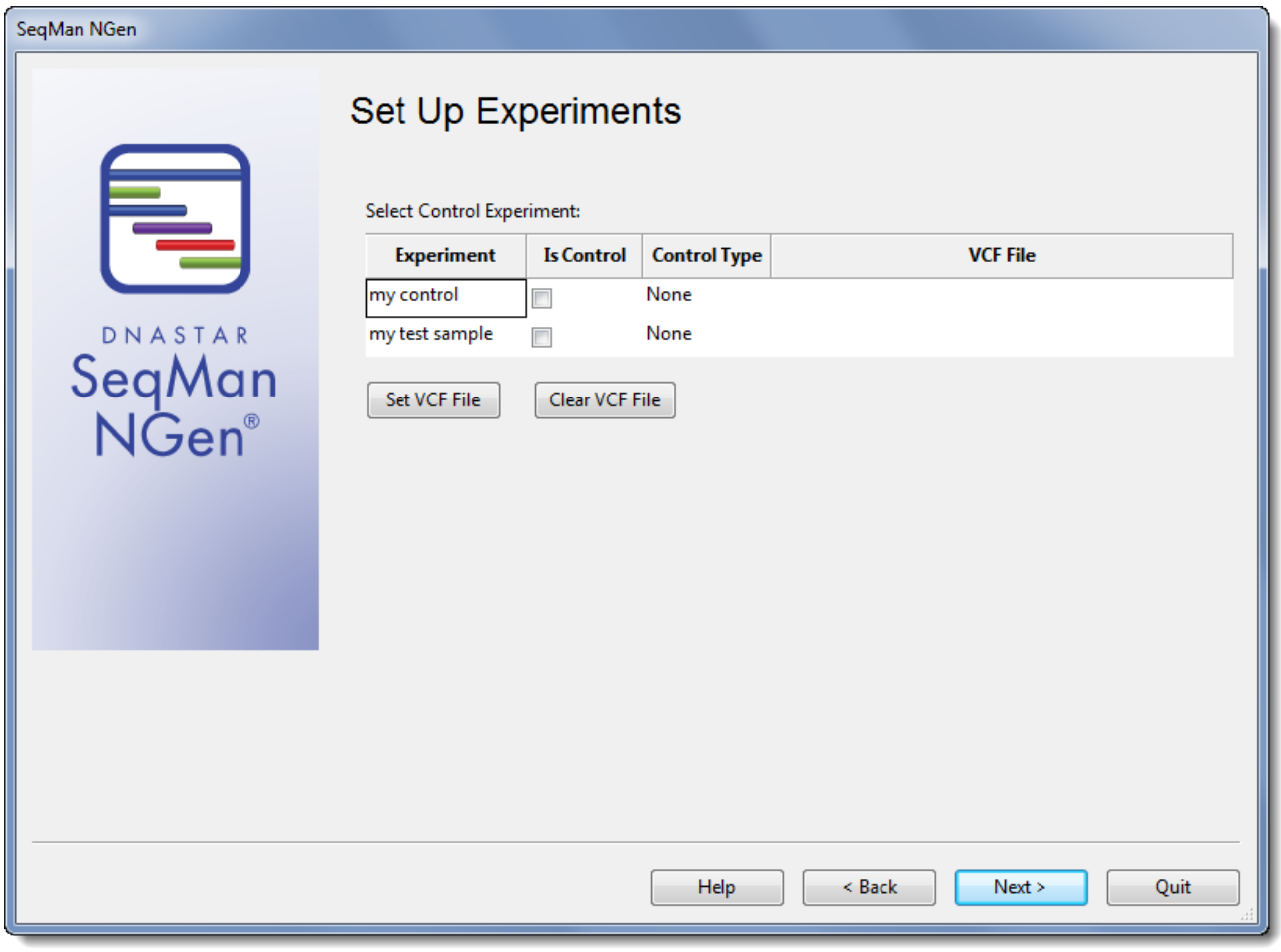
Examples of expressions you may find useful regarding paired end naming specifications follow. Please note this is not a complete list of regular expressions, and the definitions of the terms used are limited to their application to SeqMan NGen paired end naming specifications.

Special Characters	
`[ ]`	Character class--used to enclose a list of alternatives
`\`	A switch that makes special characters literal and literal characters special
`( )`	Grouping--used to delimit a string comprising a "phrase." Phrases are necessary in paired end specification so you can match a pair of forward and reverse reads while still distinguishing their orientation. In SeqMan NGen, phrases in parentheses must match for two reads to qualify as a pair; phrases outside the parentheses are used to distinguish members of the same pair.
`\d`	Any digit (0-9)
`\D`	Any non-digit character
`\w`	Any alphanumeric "word" character (including "_")
`.`	Any character
`|`	Alternate--either the term before "|" or after "|"
`^`	Match at the beginning of the line only
`$`	Match at the end of the line only
Numerical Modifiers	
`*`	0 or more
`+`	1 or more
`?`	1 or 0
`{n}`	Exactly n
`{n,}`	At least n
`{n,m}`	At least n but not more than m

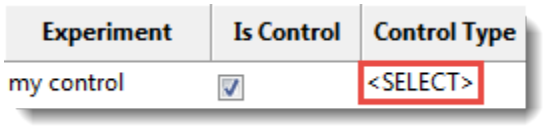
Example Expressions and Their Meanings	
d	Literally the letter d
\d	Any digit (0-9)
\d*	Zero or more digits
\d+	One or more digits
(\d+)	A phrase comprising one or more digits--same as "\d+", but causes SeqMan NGen to match the names from the string inside the phrase when other characters in the name may not match.
\.	Literally the period symbol (.)
.	Any character
.+	One or more of any characters
.*	Zero or more of any characters
a|b	a OR b
ab[i1]	abi or ab1
abi$	Ends with abi
[\.\d]	A period OR a digit
[abc]	a OR b OR c
[abc]+	One or more characters from the set a, b, c
.*f	Any number of any characters followed by the letter "f"
(.*)f	A phrase comprising any number of any characters, followed by the letter "f"--same as "*.*f", but causes SeqMan NGen to match the phrase in parentheses without matching the "f" in a read name
(\D+)r(\d+)	One or more non-digit characters followed by "r" followed by one or more digits.
(\d{2,4})f(\.abi)	Two, three or four digits followed by "f" followed by "*.abi"

Set Up Experiments

If you chose **Templated assemblies with control** from the Choose Assembly Type screen, you were prompted to enter experiment names in the Input Sequence Files dialog. The screen that follows in this situation is Set Up Experiments, which allows you to enter further information about controls and to specify a VCF file.



The upper part of this wizard screen contains four columns:

Column	Description
Experiment	This column is pre-loaded with the experiment names specified in the previous screen. If you wish to edit an experiment name, you must go back to that screen (Input Sequence Files) using the **< Back** button and edit them there.
Is Control	Use the checkboxes to indicate which experiments are controls. When a box is checked, the adjacent **Control Type** cell changes from **None** to **<SELECT>**. 
Control Type	By default, each **Control Type** is listed as **None**, signifying that the experiment is not a control. If you check a box in the **Is Control** column, the **Control Type** changes from **None** to **<SELECT>**. Click on any cell in the Control Type column to activate a dropdown menu from which you can select one of three options: • **None** – Not a control. • **Baseline** – A normal tissue control used in combination with diseased (e.g., tumor) tissue from the same individual. If you select **Baseline**, and the **Is Control** box was not previously checked, it will be checked automatically. • **Validation** – A control sample for assessing the quality and accuracy of the capture, sequencing, assembly and variant calling. For example, highly curated variant files of the HapMap NA12878 genome are available from the Genome in a Bottle consortium, with corresponding reference materials available from the Coriell Institute. Validity of the completed assembly is performed within DNASTAR's ArrayStar application. See the ArrayStar online help for details. If you select **Validation**, and the **Is Control** box was not previously checked, it will be checked automatically. If you specify a **Validation** control, you must specify a VCF file for that experiment before you can proceed to the next wizard screen. Otherwise, an error message will appear: "A VCF file must be specified for the experiment marked as a validation control."
VCF File	Select any cell in this column and then click the **Set VCF File** button to specify a VCF file. See "The **Set VCF File** button" section just below this table.

The **Set VCF File** button:

This button is used to launch a file browser from which you may select a file in VCF format. To associate a VCF file with a particular experiment or control, select the empty cell in the **VCF File** column before using the button. However, the button may also be used without first selecting a cell.

After you select the VCF file, the Associate VCF File dialog opens, offering three options.

- **VCF file for validation control experiment**
- **VCF file for non-validation experiments**
  VCF file for all experiments

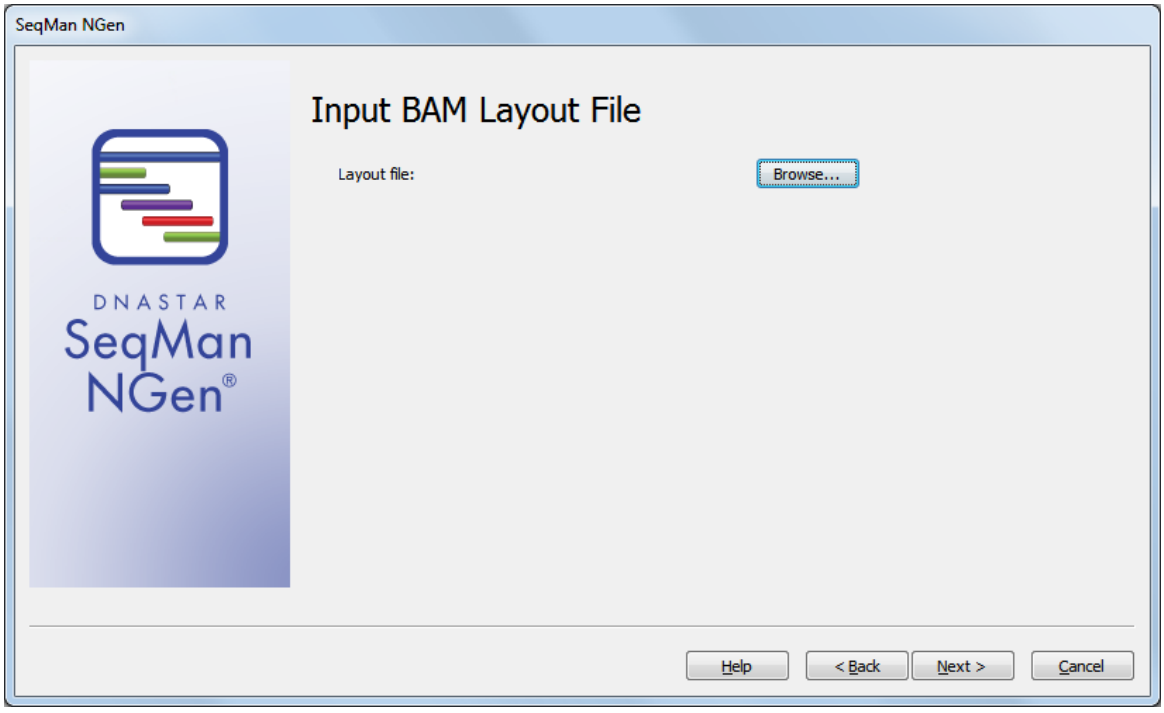
The option selected in this dialog overrides the cell (if any) that was selected in the **VCF File** column. For example, if you select a **VCF File** cell in the table corresponding to a **Validation** control sample, but then choose **VCF file for non-validation experiments**, the non-validation experiments cells in the table will be populated with the selected VCF file.

The **Clear VCF File** button:

To remove a VCF file from a cell in the **VCF File** column, select the cell and then click **Clear VCF File**.

Input BAM Layout File

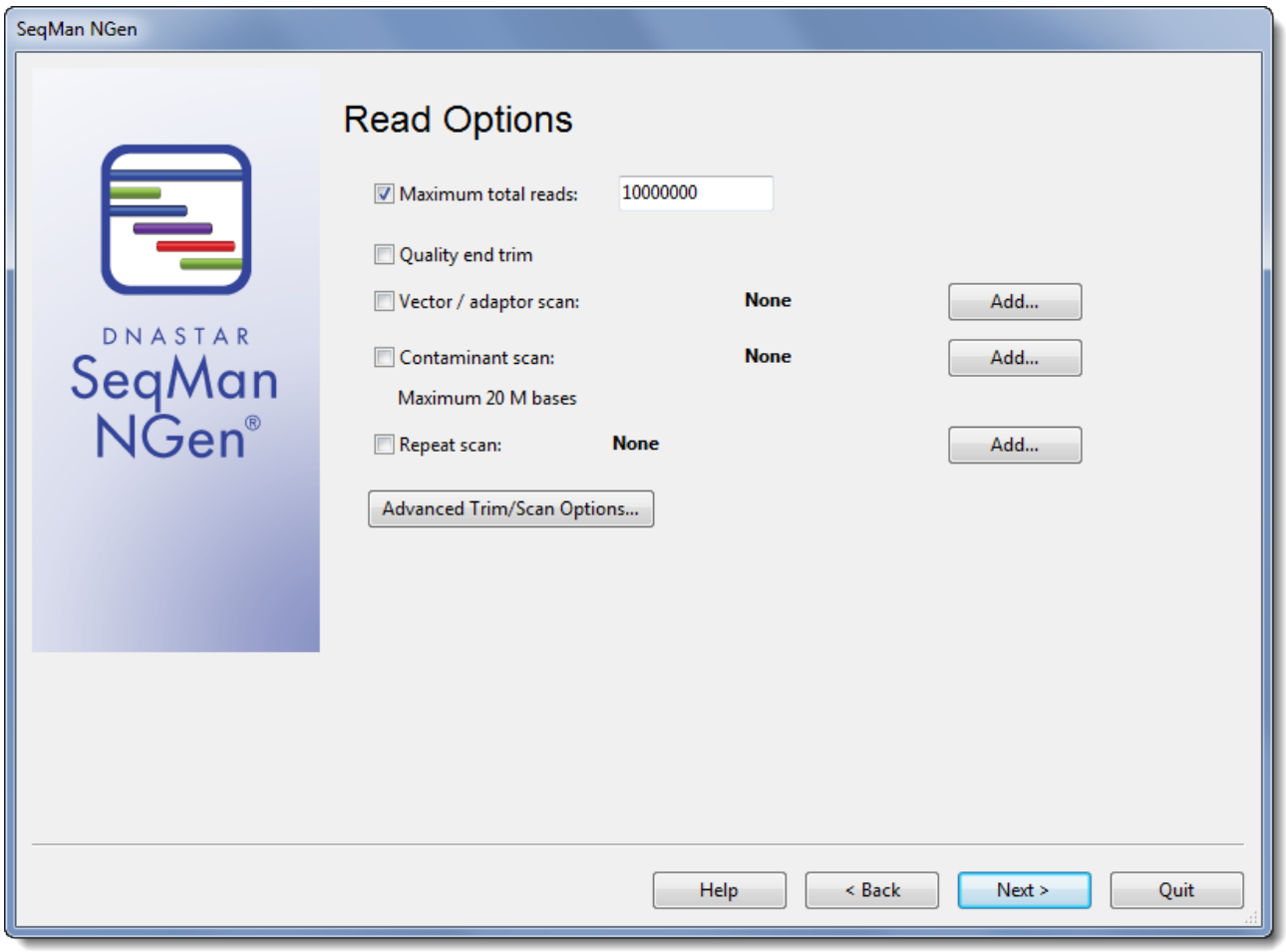
Click the **Browse** button in this dialog to choose the BAM reference sequence (*.bam).



Once you are finished, click **Next >** to continue to the next wizard screen.

Read Options

The Read Options dialog displays the parameters used for running pre-assembly scans and allows you to adjust their values. This dialog is only available for *de novo* and special templated workflows.



- **Maximum Total Reads** – By default, this box is checked and the number 10,000,000 is entered. When using Illumina technology, we recommend leaving the box checked and specifying a value to limit the number of reads used in the assembly. For 454 technologies, we suggest unchecking the box and leaving the field blank.

  Note: If you check **Maximum Total Reads**, be sure to add individual read files rather than folders in the Input Sequence Files dialog. Adding files individually causes SeqMan NGen to use an equal amount of reads from each file. If you instead add a folder, SeqMan NGen may potentially use reads from only the first file(s).

Specify whether you would like SeqMan NGen to perform any of the following pre-assembly tasks:

- **Quality end trim** – To automatically trim reads prior to assembly based on quality scores and specified quality end trimming parameters.
- **Vector/adapter scan** – To use specified vector/adaptor scan parameters to scan and trim reads for the vector or adapter.

- **Contaminant scan** – To use specified contaminant scan parameters to scan and remove reads that contain contaminant sequences. This option is not available in the Metagenomics/16S rRNA workflow.
- **Repeat scan** – To use specified repeat scan parameters to scan reads for known repetitive sequences. All sequences identified as repeats will be added to the assembly last, after all non-repeats have been assembled.

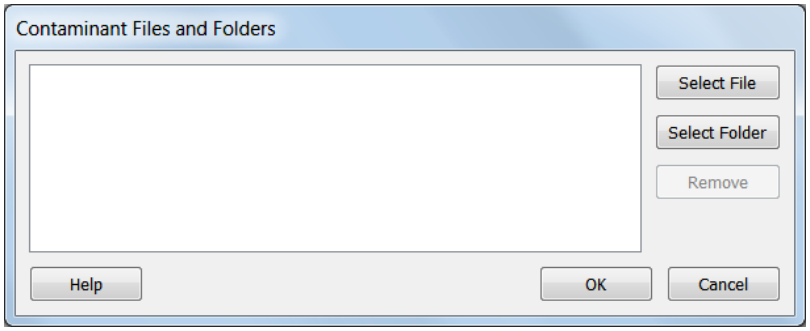
Click an **Add** button to the right of these last three options to select the desired vector, contaminant, or repetitive sequence(s) from the corresponding Files and Folders dialog.

To edit options for any of the above tasks, or to set up parameters for fixed end trimming, click the **Advanced Trim/Scan Options** button to open the Advanced Trim/Scan Options dialog. The option for fixed end trimming is *only* accessible by clicking this button.

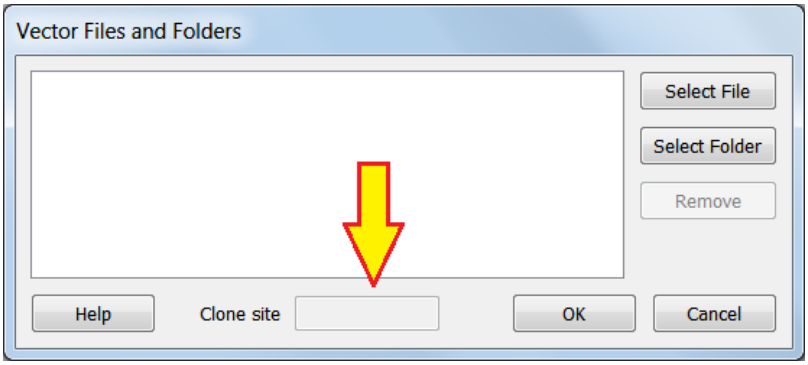
Once you are finished, click **Next >** to continue to the next wizard screen.

Files and Folders Dialogs

The Read Options dialog allows you to access Vector, Contaminant, and Repeat Files and Folders dialogs via the three associated **Add** buttons. These nearly identical Files and Folders dialogs are used to add files for the functions of vector trimming, contaminant scanning, and removal of known repeats.



- **Select File** – Click to navigate to and select individual sequence(s). See the SeqMan NGen Supported File Types page for supported file formats
- **Select Folder** – Click to navigate to and select an entire folder of sequences.
- **Clone site** (Vector dialog only) – Enter the position of the cloning site where insertion occurs.

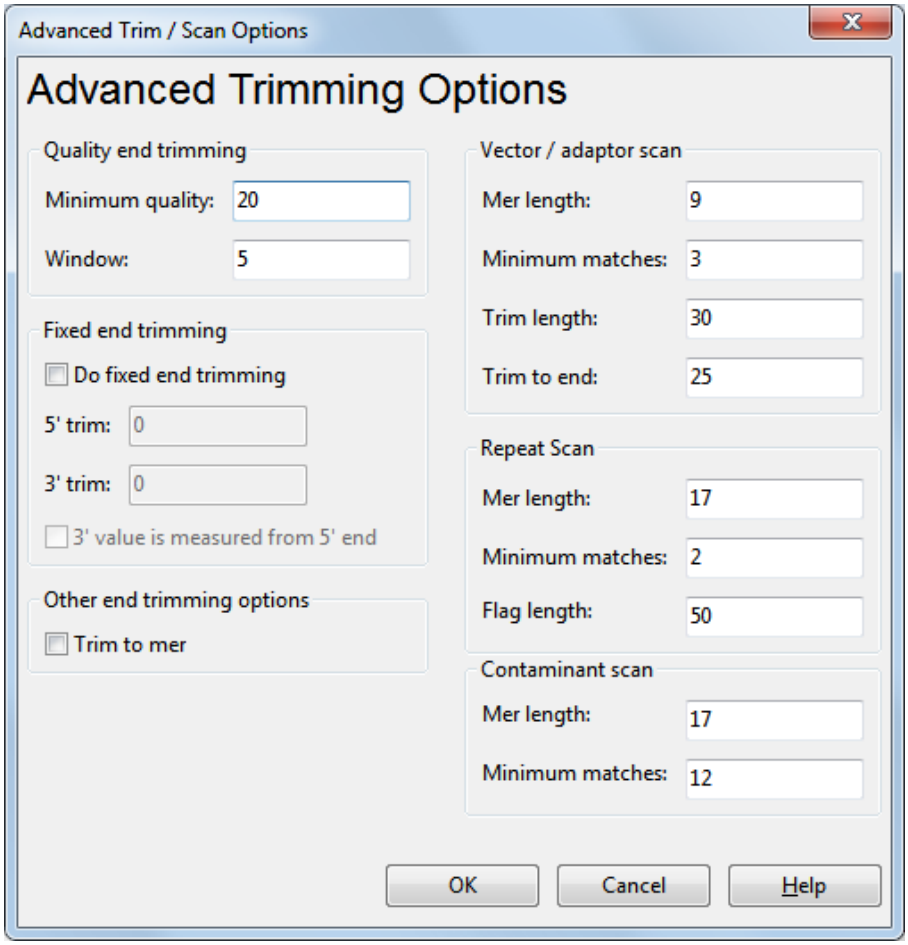


- **Remove** – Click to remove a selected (highlighted) file from the list.

Once you are finished, click **OK** to save your changes and return to the Read Options dialog, or **Cancel** to discard changes before returning.

Advanced Trim/Scan Options

From the Read Options dialog, clicking the **Advanced Trim/Scan Options** button brings you to the Advanced Trim/Scan Options dialog. This dialog allows you to view and modify trimming parameters and vector, repeat and contaminant scanning parameters.



Quality End Trimming Settings:

- **Minimum quality** – The minimum averaged quality score of the evaluated window that is required in order to be considered low-quality.
- **Window** – The length of the window to be used for averaging quality scores.

Fixed End Trimming Settings:

- **Do fixed end trimming** – Check this box to implement pre-assembly fixed end data trimming. Enter the number of base pairs you wish to trim in the **5' trim** and **3' trim** fields.

Note: The values entered for **5' trim** and **3' trim** are used differently, depending on whether **3' value is measured from 5' end** is selected. If it is *not* selected, then the **5' trim** and **3' trim** values will indicate the number of bases for SeqMan NGen to trim from the respective ends of each read. If it *is* selected, then the **5' trim** and **3' trim** values will indicate the specific coordinates to which reads should be trimmed.

Other End Trimming Options:

- **Trim to mer** – Check this box to trim the reads to the matching mer within the read. For each read, SeqMan NGen looks for mers that exist in the template (for templated assemblies) or in any other read in the assembly (for *de novo* assemblies). It then sets the trimming for the read to the start of the first mer found and the end of the last mer found. Trimming to mer may be useful when assembling data without accurate quality scores or data with very short linkers.

Vector/Adapter Scan Settings:

- **Mer length** – The minimum length of a mer required to be considered an exact match when searching for vector.
- **Minimum matches** – The minimum number of matching mers required to start an alignment.
- **Trim length** – The minimum length required for a mer to be considered as a match for vector trimming.
- **Trim to end** – The distance to the endpoint where trimming will go all the way to the end of the sequence.

Repeat Scan Settings:

- **Mer length** – The minimum length of a mer required to be considered an exact match when scanning for repeats.
- **Minimum matches** – The minimum number of matching mers required to be considered a repeat.
- **Flag length** – The minimum length required for a mer to be flagged as a repeat.

Contaminant Scan Settings:

- **Mer length** – The minimum length of a mer required to be considered an exact match when scanning for contaminants.
- **Minimum matches** – The minimum number of matching mers required to mark the sequence as a contaminant.

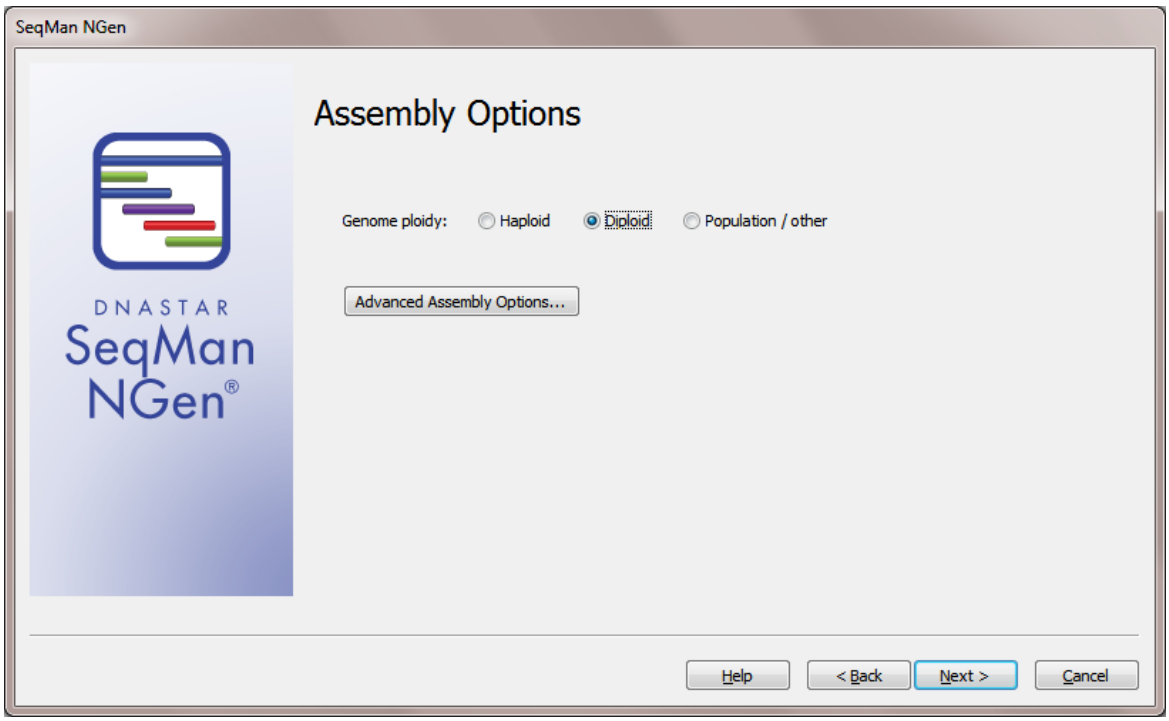
Once you are finished, click **OK** to save your changes and return to the Read Options dialog, or **Cancel** to discard changes before returning.

Assembly Options

The Assembly Options dialog allows you to specify the parameters to use for your assembly. There are several versions of this dialog depending upon your choices in previous wizard screens. Follow the links below to go to the appropriate help topic for your workflow.

Assembly Options (BAM Layout)

The Assembly Options dialog allows you to specify the parameters to use for your assembly. If you are following the BAM layout workflow, the following version of the dialog appears.



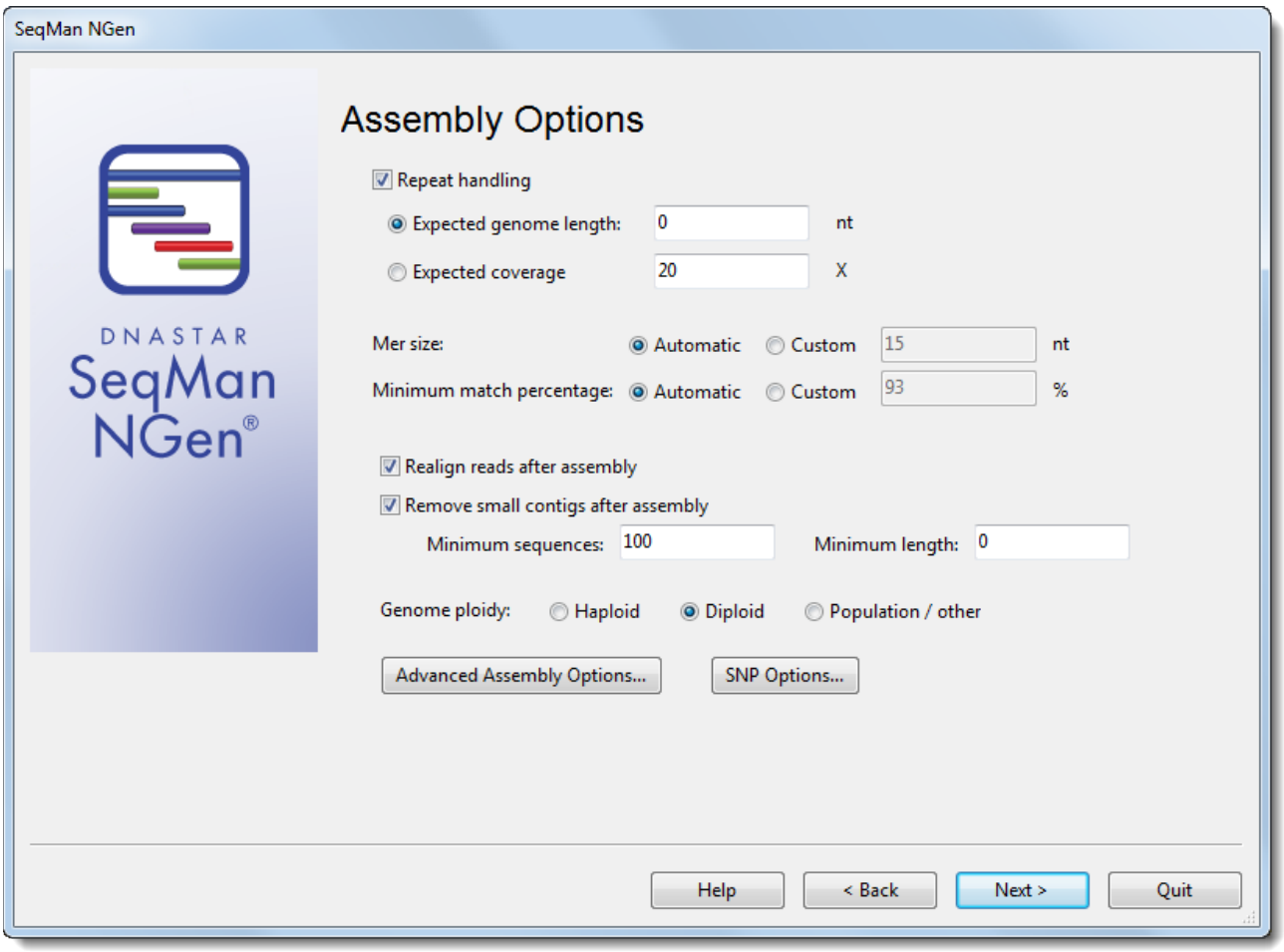
Select the type of **Genome ploidy** for your project. Choosing **Haploid** or **Diploid** establishes the statistical model SeqMan NGen will use in estimating probabilities during SNP calls. Selecting **Population / other** (e.g. for a polyploid genome) causes SeqMan NGen not to calculate probabilities.

If desired, click the **Advanced Assembly Options** button to open the Advanced Assembly Options dialog. This dialog allows you to view and edit additional assembly parameters.

Once you are finished, click **Next >** to continue to the next wizard screen.

Assembly Options (*De Novo*, Special Templated)

The Assembly Options dialog allows you to specify the parameters to use for your assembly. If you are following the *de novo* or special templated workflows, you will see the following version of the dialog.



- **Repeat Handling** – Checking this box automatically computes a threshold for determining the number of identical subsequences of bases, or mers, used to indicate a putative repeat. (For more information, see the Repeat Handling section.)

  Note: The Repeat Handling section is not included in this dialog if you are performing a special templated assembly or a transcriptome assembly.

- **Expected genome length** – If you know the approximate length of the genome/fragment being assembled, select this button and specify a length. SeqMan NGen will then calculate the expected average coverage empirically from the amount of data. This, in turn, allows repeat regions to be identified and handled more accurately, resulting in a better assembly. If the approximate genome length is not known, use the **Expected coverage** option.
- **Expected coverage** – If you do not know the length of the genome/fragment, select this button and provide an estimate of the depth of the sequencing. The default value for this field is 20, and the maximum allowable value is 65,535. If you enter a value larger than the maximum, you may receive an error message and be prevented from continuing until you choose a value less than or equal to the maximum.

Note: Use caution when estimating the value for **Expected coverage**. If the value you use is significantly lower than the actual depth, the assembly may take a much longer time to complete and may have too many mers flagged as repeats. We recommend using **Expected genome length** whenever possible.

- **Mer size** – The minimum length of a mer (overlapping region of a fragment read), in bases, required to be considered a match when arranging reads into contigs. Mer size information is used to identify matches during the assembly layout phase. The default mer size is determined by the selected read technology and is shown in the window. For more information, see the Mer Tags section.
- **Automatic** – Select this button to automatically set the size based on assembly type and sequencing technology.
- **Custom** – Select this button to choose the size yourself. You must enter the desired number of base pairs in the field at right. Lowering the mer size increases the sensitivity of finding matches, but also increases the likelihood of finding spurious matches in addition to the correct match. Lowering the mer size can also greatly increase the requirements for storing intermediate and temporary files with large projects.

- **Minimum match percentage** – Specifies the minimum percentage of matches in an overlap that are required to include a sequence read in the final alignment. For more information, see the Match Percentage section.
  - **Automatic** – Select this button to automatically set the percentage based on assembly type and sequencing technology.
  - **Custom** – Select this button to designate the percentage yourself. You must enter a number in the field at right.
- **Realign reads after assembly –** Check this box to include a realignment step after the assembly. This step analyzes each sequence at the nucleotide level to determine the exact position of each sequence in the alignment and realigns contigs as needed. For templated assemblies, this option may improve the accuracy of the final assembly by correcting occasional misalignments that can occur in gapped regions. However, this step may significantly increase the time needed to assemble.

- **Remove small contigs after assembly** – Check this box and type values in one or both boxes:
  - **Minimum sequences** – to disassemble any untemplated contigs with fewer than the specified number of sequences.
  - **Minimum length** – to disassemble any untemplated contigs shorter than the specified length.

  Note: Both options affect only untemplated contigs. No templated contigs will be removed.

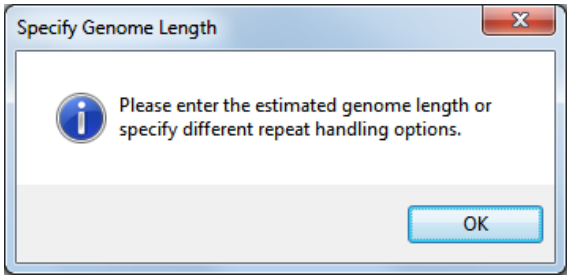
- **Genome ploidy** – Select the type of ploidy for your project. Choosing **Haploid** or **Diploid** establishes the statistical model SeqMan NGen will use in estimating probabilities during SNP calls. Selecting **Population / other** (e.g. for a polyploid genome) causes SeqMan NGen not to calculate probabilities.

  Note: The **Genome ploidy** option is only displayed if **BAM Format** is checked in the **Save project as** section of the Set Up Project Files dialog.

If desired, click the **Advanced Assembly Options** button to open the Advanced Assembly Options dialog. This dialog allows you to view and edit additional assembly parameters. Or click the **SNP Options** button to open the SNP Options dialog, where you can change SNP-related parameters.

Note: The **SNP Options** button is only displayed if you checked "**BAM Format"** in the **Save project as** section of the Set Up Project Files dialog.

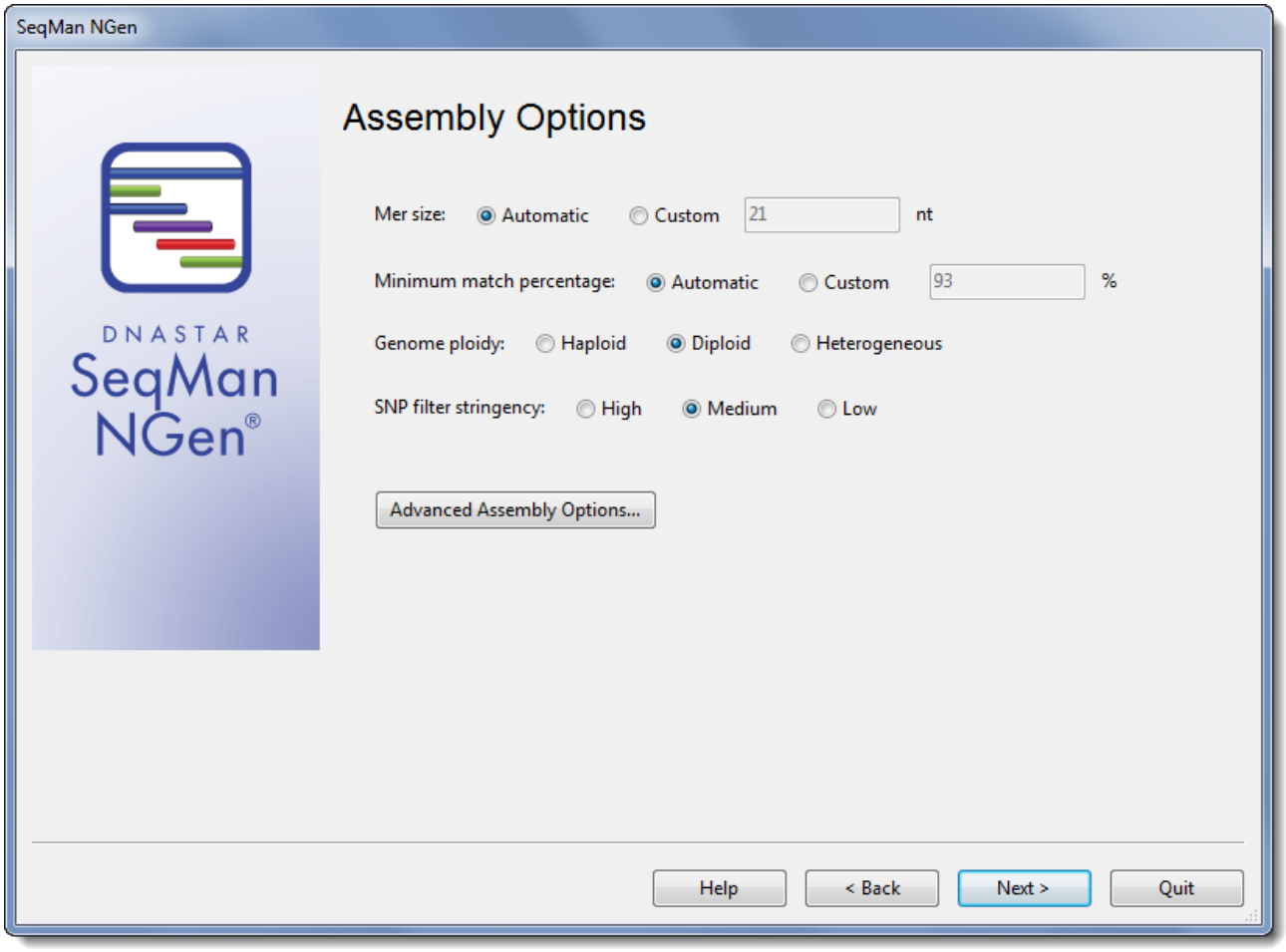
Once you are finished, click **Next >** to continue to the next wizard screen. Note that if you check **Repeat handling** without specifying an **Expected genome length**, you will receive the following error message after clicking **Next**.



Click **OK** and adjust the dialog parameters before again clicking **Next**.

Assembly Options (All Others)

The Assembly Options dialog allows you to specify the parameters to use for your assembly. If you are following the normal templated, reference-guided or viral-host workflows, the following version of the dialog appears.



- **Mer size** – The minimum length of a mer (overlapping region of a fragment read), in bases, required to be considered a match when arranging reads into contigs. Mer size information is used to identify matches during the assembly layout phase. The default mer size is determined by the selected read technology and is shown in the window. For more information, see the Mer Tags section.
    - **Automatic** – Select this button to automatically set the size based on assembly type and sequencing technology.

- **Custom** – Select this button to choose the size yourself. You must enter the desired number of base pairs in the field at right. Lowering the mer size increases the sensitivity of finding matches, but also increases the likelihood of finding spurious matches in addition to the correct match. Lowering the mer size can also greatly increase the requirements for storing intermediate and temporary files with large projects.

- **Minimum match percentage** – Specifies the minimum percentage of matches in an overlap that are required to join two sequences in the same contig. (For more information, see the Match Percentage section.)
  - **Automatic** – Select this button to automatically set the percentage based on assembly type and sequencing technology.
  - **Custom** – Select this button to designate the percentage yourself. You must enter a number in the field at right.

- **Genome ploidy** – Select the type of ploidy for your project. Choosing **Haploid** or **Diploid** establishes the statistical model SeqMan NGen will use in estimating probabilities during SNP calls. Selecting **Population / other** (e.g. for a polyploid genome) causes SeqMan NGen not to calculate probabilities.

Note: The **Genome ploidy** option is only displayed if you checked "**BAM Format"** in the **Save project as** section of the Set Up Project Files dialog.

- **SNP filter stringency** – (not available in all workflows) The three radio buttons specify **High**, **Medium** or **Low** stringency levels for "soft" filtering of SNPs. This means that SNPs of the least interest to you will be automatically hidden when SNP reports/tables are viewed in SeqMan Pro or ArrayStar. However, these filtered SNPs are *not* removed from the assembly, and can be made visible again by changing the SNP filtering parameters in either SeqMan Pro or ArrayStar.

Note: "Hard" filtering of SNPs can be done through the SNP tab of the Advanced Options dialog.

SNP Filter Parameter	SNP Filter Stringency		
	High	Medium (default)	Low
Depth	10	10	2
P not ref	90	75	50
Min SNP%	15	15	0 (i.e., > 0)

****

If desired, click the **Advanced Assembly Options** button to open the Advanced Assembly Options dialog. This dialog allows you to view and edit additional assembly parameters.

Once you are finished, click **Next >** to continue to the next wizard screen.

Advanced Assembly Options

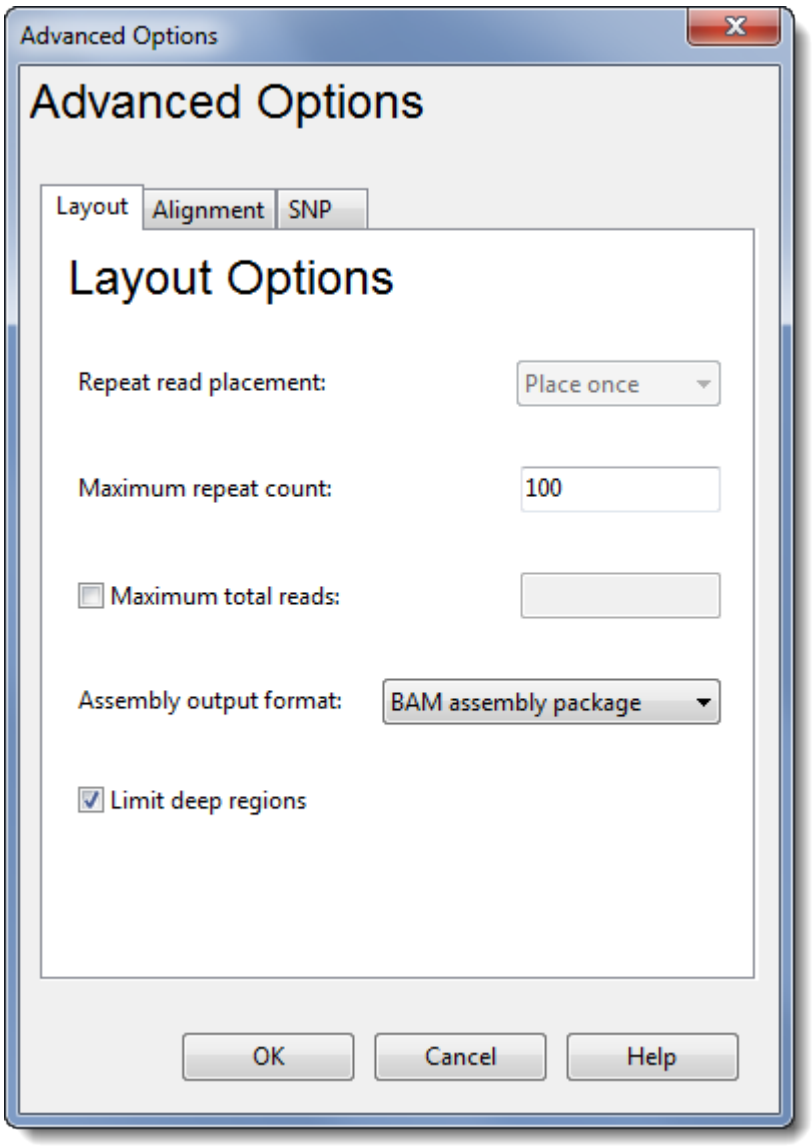
Clicking the **Advanced Assembly Options** button from the Assembly Options or SNP Options dialog opens an Advanced (Assembly) Options dialog. The dialog may be plain or tabbed, depending upon your choices in previous wizard screens. Follow the links below to go to the appropriate help topic for your workflow.

Advanced Options (Normal Templated, Reference-Guided)

In normal templated and reference-guided workflows, clicking the **Advanced Assembly Options** button from the Assembly Options dialog opens a tabbed Advanced (Assembly) Options dialog. There are several versions of this dialog depending upon your choices in previous wizard screens.

Layout Options

Clicking the **Advanced Assembly Options** button from certain Assembly Options dialogs opens a tabbed Advanced Options dialog. The Layout tab allows you to view and edit Layout Options. Default parameters vary according to the sequencing technology and project type specified elsewhere in the wizard, and values seldom need to be changed.



- **Repeat read placement** – (currently disabled).
- **Maximum repeat count** – Enter the maximum number of occurrences for any given mer in the reference sequence for it to be used in matching. Mers exceeding this value are flagged as repeats and are not used as mer tags in determining overlaps.
- **Maximum total reads** – Check the box and enter a value if you wish to limit the read depth. Utilizing this option can make the assembly proceed faster.
- **Assembly output format** – Use the drop-down menu to choose a format for the assembly output.

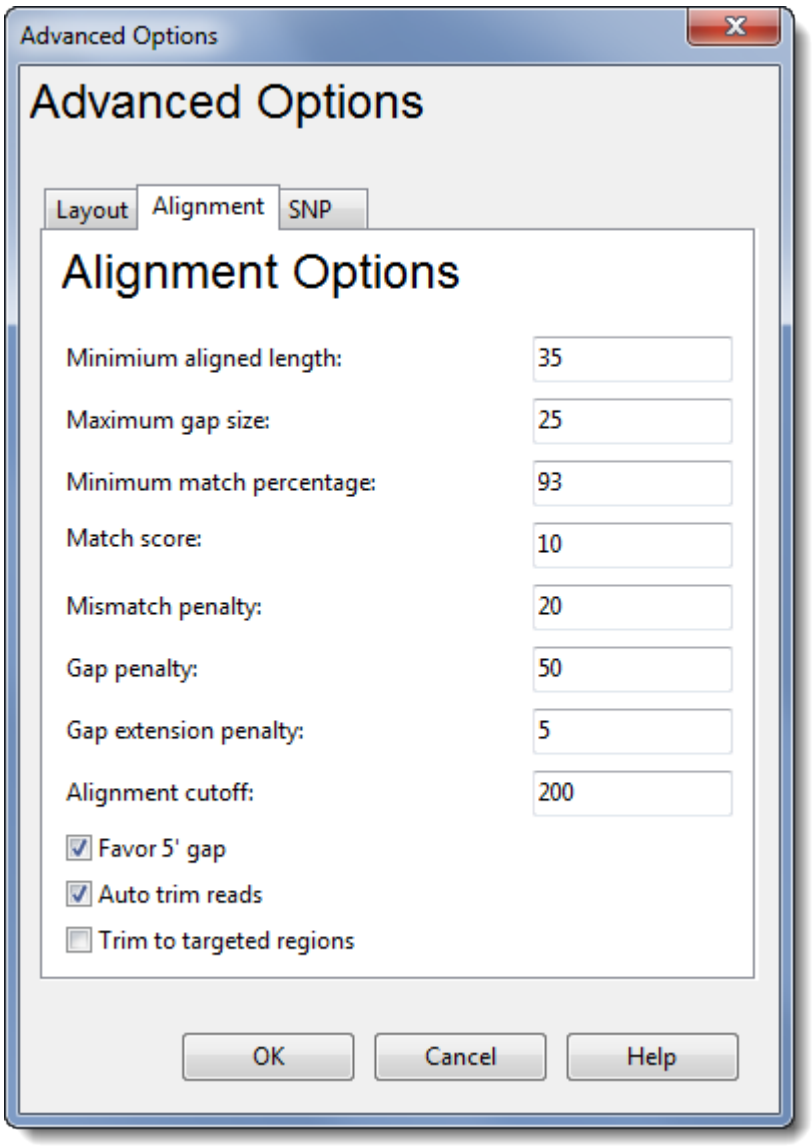
- **BAM assembly package** – To save the output as a *.bam file.
- **SeqMan Pro document (.sqd)** – To save the output in both *.bam and *.sqd formats.
- **Unassembled seqs only** – This option can be used as a filter to remove reads from an unwanted source in a mixed sample (e.g. removing host DNA from a viral sample).

Note: If you are doing a reference-guided assembly with gap closure, the only option enabled is *.sqd.

- **Limit deep regions** – If this box is checked, areas of the assembly where an extreme number of reads (> 10000) are laid out to the same area of the template will be filtered before alignment. This is not an exact filter and the maximum depth will typically be between 10,000-20,000 reads. This filter can improve performance, sometimes significantly.

Alignment Options

Clicking the **Advanced Assembly Options** button from certain Assembly Options dialogs opens a tabbed Advanced Options dialog. The Alignment tab allows you to view and edit Alignment Options for the gapped alignment phase. Default parameters vary according to the sequencing technology and project type specified elsewhere in the wizard, and values seldom need to be changed.



Enter values for:

- **Minimum aligned length** – The minimum length of at least one aligned segment of a read after trimming. The default value varies depending on the read technology you selected.
- **Maximum gap size** – The maximum number of gaps allowed per 1000 bases in the alignment.
- **Minimum match percentage** – The minimum percentage of matches in an overlap required to join two sequences in the same contig. SeqMan NGen determines the percentage to use based on the sequencing technology you specified in the Assembly Options dialog.
- **Match score** – The score for a base match during an alignment. This score contributes to the pairwise score used to calculate match percentage. Increasing this value will allow for longer or more frequent gaps, thus forcing bases that match to be assembled together.

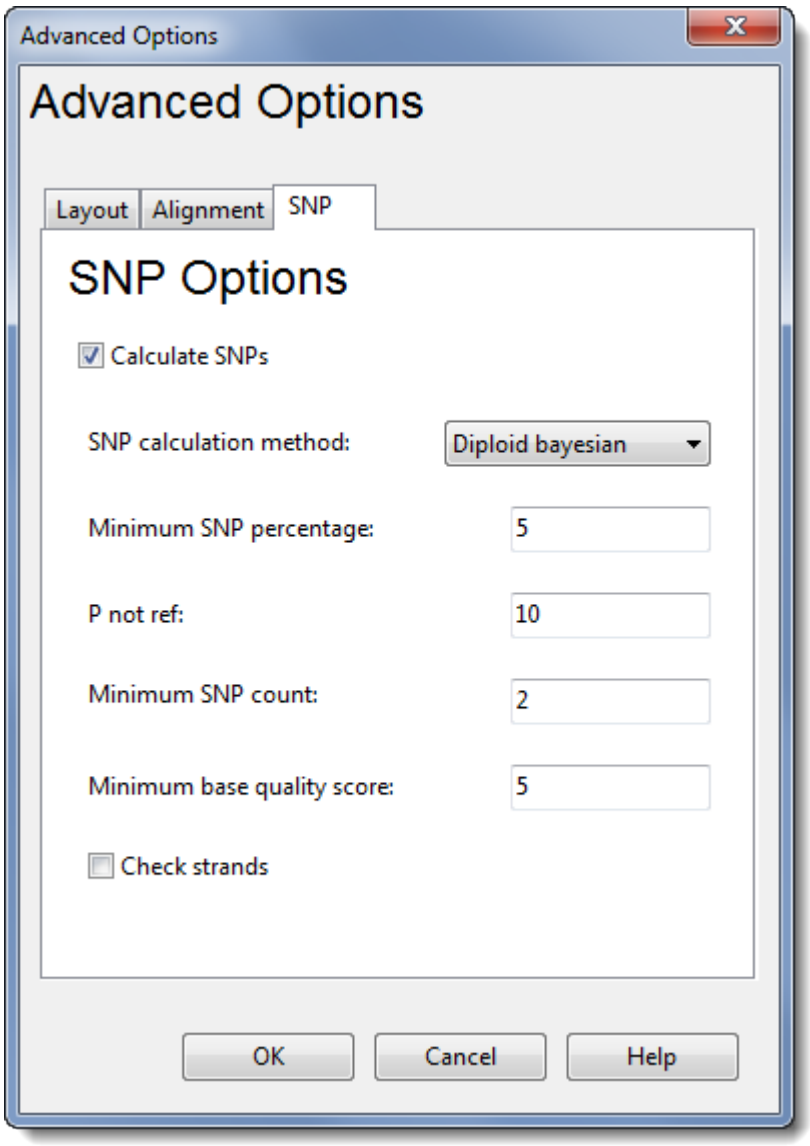
- **Mismatch penalty** – The penalty for a base mismatch during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage.
- **Gap penalty** – The penalty for opening a gap during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. A high gap penalty suppresses gapping, while a low value promotes gapping.
- **Gap extension penalty** – The penalty to the alignment score for extending a new or existing gap by one base. This is in contrast to the gap penalty, which is the penalty to the alignment score for opening up a new gap.
- **Alignment cutoff** - Determines if the accumulation of gap openings, gap extensions and mismatches causes the alignment score to drop below the maximum alignment score. If so, the alignment will stop and will be trimmed to the point where the alignment score was at its maximum.
- **Favor 5' gap** – If this box is checked, insertions and deletions in homopolymeric runs or simple sequence repeats will preferentially occur on the 5' end (top strand) of the run/repeat. The box is checked by default.
- **Auto trim reads** – If this box is checked, the ends of reads are trimmed to best match alignment to the template. SeqMan NGen will mark the portion of the read that aligns well to the template, and will set the trimming to skip any of the poorly aligning parts of the read. Checking this option optimizes the end trimming of reads to maintain as much of the read as possible, while still meeting the minimum match percentage threshold. However, checking the box can also lead to the removal of true variant bases located near the ends of reads. The box is checked by default.
- **Trim to targeted regions** – If this box is checked, reads extending beyond the 5' or 3' end of a targeted region will be trimmed to the target boundary. The box is unchecked by default.

SNP Options

Launch the SNP Options dialog by clicking the **Advanced Assembly Options** button from certain Assembly Options dialogs (you may then need to click the SNP tab) or the **SNP Options** button from the Recalculate SNPs dialog . The SNP Options dialog allows you to view and edit options related to SNP calculation.

The options chosen in this dialog affect the "hard" filtering of SNPs. This means that SNPs of the least interest to you will be automatically and permanently removed from the assembly.

Note: For information on reversible "soft" filtering of SNPs, see the description of **SNP filter stringency** in the topic Assembly Options (All Others).



Default parameters vary according to the sequencing technology and project type specified elsewhere in the wizard, and values seldom need to be changed.

- **Calculate SNP's** – Check to turn on in-line SNP detection. This box is checked by default.
- **SNP calculation method** – Use the drop-down menu to select the desired calculation method.
    - **Simple percentage** – To detect only whether the column is most likely to have the same base or a different base from the reference. Reference bases are not reported. The most frequent base in the column which is not the reference is treated as the potential SNP call. The "SNP%" output value is the percentage of the column which corresponds to the base chosen as the alternative to the reference. The direction and quality (weight) of the bases in the column are not considered. You may choose a minimum threshold for this value.

- **Diploid bayesian** – To use a Bayesian statistical model very similar to MAQ (Li *et al.*, 2008, *Genome Res.* 18:1851) to call SNPs between three potential genotypes: homozygous reference, homozygous variant (some other base), and heterozygous (two bases, which may include the reference). This menu choice is optimized for diploid genomes. Before applying this model, the simple SNP caller is run to more quickly establish a percentage with which the column is screened. If the column passes a minimum percentage screen, it is then checked against a minimum variant depth: the most frequent variant base must meet or exceed this threshold. Putative SNP-containing columns are then evaluated with a statistical model that considers the two most frequent bases in the column as possible alleles. If there is only one base, the reference is used as the other base, regardless of its depth in the column. The model then calculates the **P not ref** of each set of bases, meaning the probability that they occurred by random chance. This is based on the base frequency, combined frequency of the two bases, the quality scores (weights), and the directions of the reads. A putative SNP base must have at least one read on each strand. The heterozygous call's probability is based on simple permutations and a constant modifier, with the strands considered separately. Since they are the only possible genotypes, probabilities are normalized against one another, and the highest probability is called.
- **Haploid bayesian** – (default) A Bayesian method similar to the one above, but optimized for haploid genomes.

Enter values for:

- **Minimum SNP percentage** – The minimum percent of non-reference bases required to call an SNP. When it performs SNP passes, SeqMan NGen will include regions in an assembly that have coverage less than or equal to the specified value. The default value is 5. A non-zero value is recommended when using Ion Torrent data, or working with larger genomes or doing population studies. Very low values will lead to larger files, but do not necessarily result in better SNP calls.
- **P not ref** - The minimum SNP quality score (Q_{call}) required to include a position as a putative SNP.

Note: If you chose **Cancer / somatic gene panel assembly** in the Choose Project Type screen, **P not ref** is disabled. That workflow uses a "simple percentage" SNP caller and the **P not ref** statistic is not calculated.

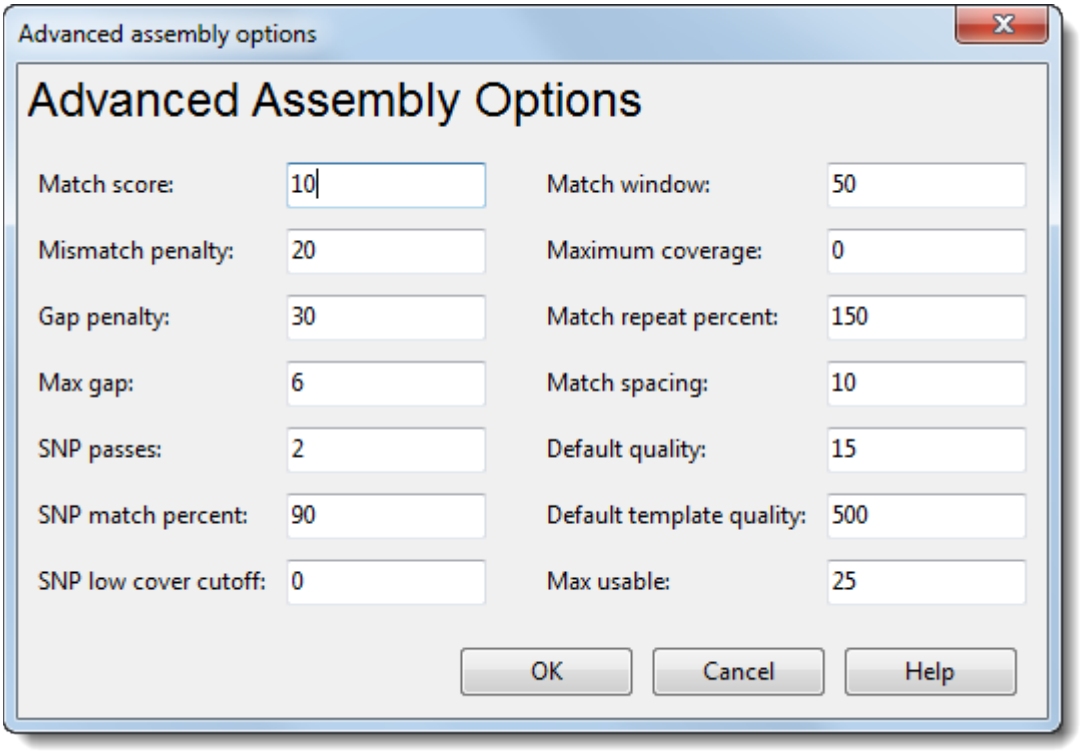
- **Minimum SNP Count** – The minimum number of non-reference bases required to call an SNP. When it performs SNP passes, SeqMan NGen will include regions in an assembly that have coverage less than or equal to the specified value.
- **Minimum base quality score** – The minimum quality score below which a base will not be considered.
- **Check strands** – Check this box to consider the strandedness of each read during SNP calculation. By default, the box is unchecked.

Note: Minimum SNP percentage and **Minimum SNP Count** can be used in tandem to control the number of reportable SNPs, and by extension, the size of the SNP table.

Once you are finished, click **OK** to save changes and return to the previous dialog (either Assembly Options or SNP Options), or **Cancel** to return without saving changes.

Advanced Assembly Options (De Novo)

Clicking the **Advanced Assembly Options** button from the Assembly Options dialog in a *de novo* workflow opens the Advanced Assembly Options dialog. Default parameters vary according to the sequencing technology and project type specified elsewhere in the wizard, and values seldom need to be changed.



Enter values for:

- **Match score** – The score for a base match during an alignment. This score contributes to the pairwise score used to calculate match percentage. Increasing this value will allow for longer or more frequent gaps, thus forcing bases that match to be assembled together.
- **Mismatch penalty** – The penalty for a base mismatch during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage.
- **Gap penalty** – The penalty for opening or extending a gap during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. A high gap penalty suppresses gapping, while a low value promotes gapping.

- **Max gap** – The maximum number of gaps allowed per 1000 bases in the alignment.
- **SNP passes** – The number of times SeqMan NGen will cycle through a templated assembly, attempting to fill in regions with zero or low coverage due to SNPs.
- **SNP match percent** – The minimum match percentage required during passes to fill in SNP regions. The default value will change depending on the type of assembly and the read technology you selected.
- **SNP low cover cutoff** – The minimum coverage required in an assembly to be excluded from SNP passes. SeqMan NGen will include regions in an assembly that have coverage less than the value specified as well as regions with zero coverage when it performs SNP passes. (See the **SNP passes** parameter above.)
- **Match window** – The size of the window used to calculate match percentage.
- **Maximum coverage** – The maximum depth of coverage allowed in a templated assembly. SeqMan NGen will not exceed the coverage specified by this threshold. The default value of "0" equals unlimited coverage.

  Note: This parameter is only available for templated assemblies, and should be used with caution, as it will limit the number of sequences included in the assembly.

- **Match repeat percent** – The percent frequency a mer occurs compared to its expected frequency. Mers exceeding this value are flagged as repeated and not used as mer tags in determining overlaps.
- **Match spacing** – The length of the window of a sequence read where at least one mer tag will be chosen. The default value will change depending on the read technology you selected.
- **Default quality** – The value used for the base quality of sequences without quality scores.
- **Default template quality** – The value used for the base quality of template sequences without quality scores.
- **Max usable** – Any mers occurring more frequently than the Repeat Handling **expected coverage** value multiplied by this value are disregarded as mer tags from the assembly.

Once you are finished, click **OK** to save changes and return to the Assembly Options dialog, or **Cancel** to return without saving changes.

Match Percentage

By default, SeqMan NGen uses a local match percentage which requires that the match percentage threshold be met in each overlapping window of 50 bases. The size of this window can be adjusted by specifying a different value for the match window parameter.

An example containing a repeated region follows.

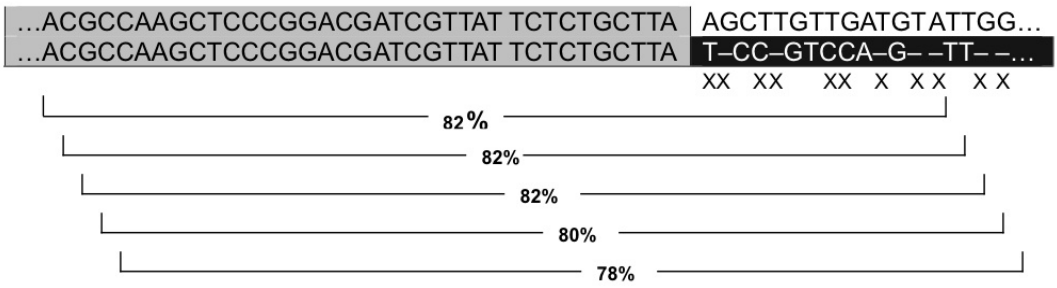
A genome fragment has repeated regions labeled A and A', and two unique regions labeled B and C.

When the fragment is sequenced, one of the sequences contains parts of regions A and B, and another contains parts of regions A' and C:

In this example, a minimum match percentage of 80% is used. When the two sequences are aligned, the 400 bases in the overlapping A and A' regions match 100%. The 200 bases in the overlapping B and C regions match 42%. Over the entire alignment, 484 out of 600 bases match, yielding a global match percentage of 81%.

However, SeqMan NGen checks the match percentage for every alignment of 50 bases. The alignment below shows the last 36 overlapping bases of A and A' and the first 18 overlapping bases of B and C. Each mismatch in the overlap is marked by an X below the alignment. In the first 50 bases shown, there are 41 matches, and the match percentage is 82%. This is above the threshold of 80%, so the match percentage of the next 50 bases is checked and is also found to be 82%.

Each fifty bases are checked along the overlap as long as the match percentage is at or above the threshold. In this case, the alignment fails once it gets far enough into the overlap of the unique regions, B and C, that the match percentage drops to 78%. The sequences will not be assembled together into a contig, which is correct for this data set.



Mer Tags

The SeqMan NGen layout algorithm relies on unique subsequences of bases, or *mers*, which occur in overlapping regions of fragment reads. Mers that are common to two or more fragment reads are aligned to determine the overall layout of reads. Overlapping reads have many mers in common, but only a few mers per overlapping region are needed to identify the overlap. These mers are called *mer tags*. The use of mers to tag fragments and identify overlaps is illustrated in the following figure:

Original DNA Sequence:

CGAATGTCATATGGCAGTACACGGCGTACGTTAGGTTTCTGAGGGATTTTCGAG

Fragment Reads:

1. CGAATGTCATATGGCAGTA
2. TATGGCAGTACACGGCGTACGT
3. GGCGTACGTTAGGTTT
4. TTAGGTTTCTGAGGGATT
5. AGGTTTCTGAGGGATTTTCGAG

Fragment Read Layout:

1. CGAATGTCATATGGCAGTA
2. TATGGCAGTACACGGCGTACGT
3. GGCGTACGTTAGGTTT
4. TTAGGTTTCTGAGGGATT
5. AGGTTTCTGAGGGATTTTCGAG

Note: As shown in the above figure, a 54bp original DNA sequence is covered by five overlapping fragment reads. The 6-mer tags for each fragment read are underlined. Matching mer tags are aligned to determine the layout of the reads.

The power of using mer tags relies on the ability of SeqMan NGen to choose mers that are most likely to occur only once in the original DNA sequence. It is important to avoid choosing mers that occur in repeated regions since the result may be fragment reads that are incorrectly aligned together.

Three parameters are involved in choosing mer tags: **Match Size**, **Repeat Handling**, and **Match Spacing**. All of these parameters can be adjusted in the Advanced assembly options dialog.

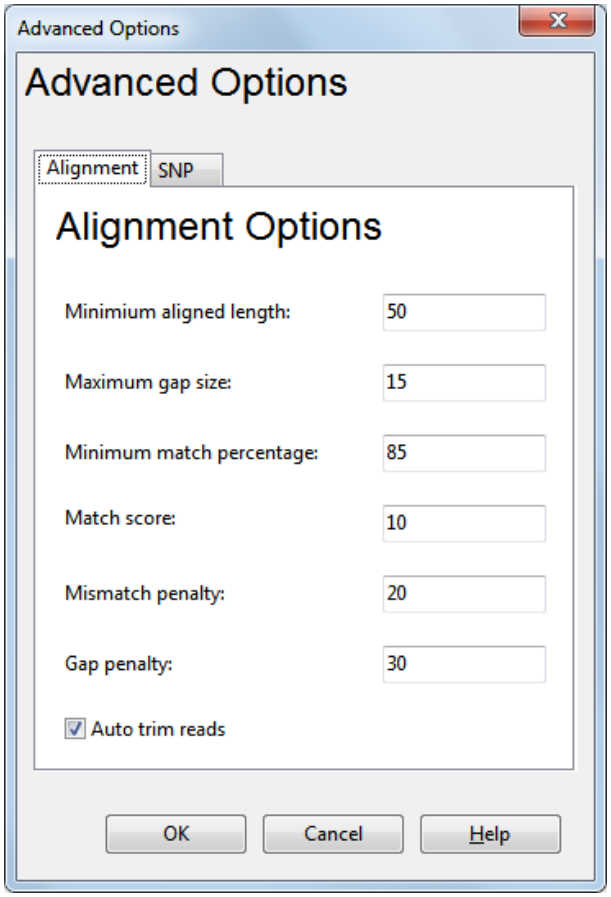
The **Match Size** and **Repeat Handling** parameters help to choose tags that are most likely to be unique in the original DNA sequence. **Match Size** sets the length of the mers. The longer the mer, the higher the probability that it is unique. **Repeat Handling** parameters help to identify which mers are not likely to be unique. If a mer occurs more often than expected in the dataset, the mer may be part of a repeated region.

Match Spacing specifies the preferred distance between mer tags. The smaller the **Match Spacing** parameter value, the more memory and more time the assembly will take. If a fragment read is shorter than the **Match Spacing** value, multiple mer tags are still chosen for the read.

Note: During assembly, any given read will only be assigned to one contig, even if it matches the hit criteria for more than one contig. If there is no information linking the read to a specific contig (e.g. a unique SNP or a paired-end constraint), SeqMan NGen will assign the sequence randomly to one of the contigs for which it meets the criteria.

Advanced Options (BAM Layout)

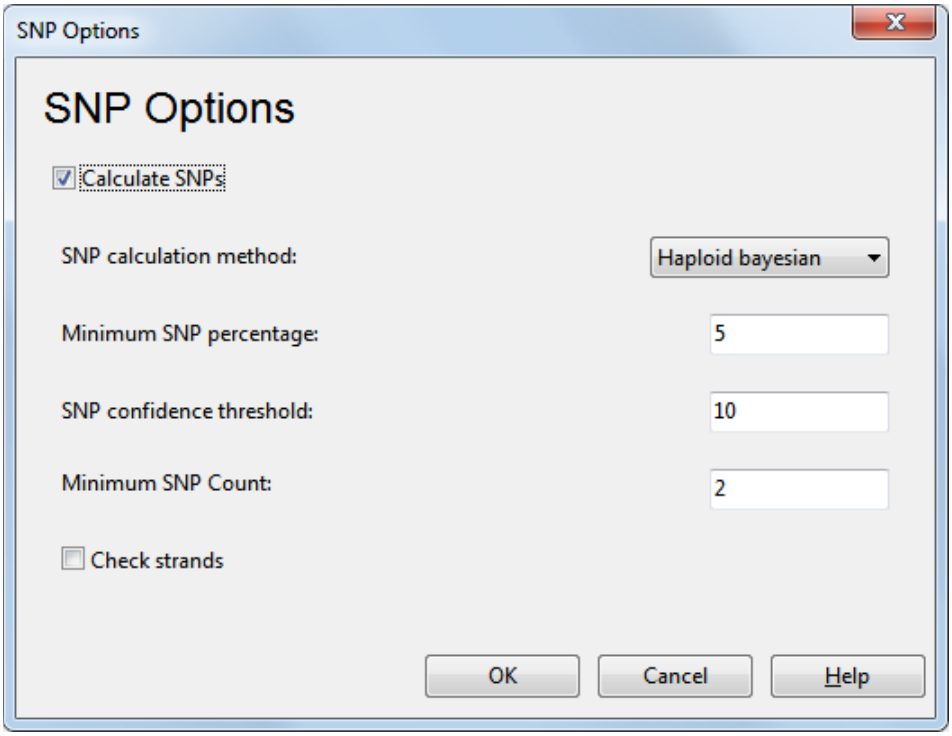
Clicking the **Advanced Assembly Options** button from the Assembly Options (BAM Layout) dialog opens the Advanced Options dialog. For information about this dialog, see the topics Alignment Options and SNP Options.



Once you are finished, click **OK** to save changes and return to the Assembly Options (BAM Layout) dialog, or **Cancel** to return without saving changes.

SNP Options Dialog

Clicking the **SNP Options** button from the Recalculate SNPs dialog opens the SNP Options dialog.

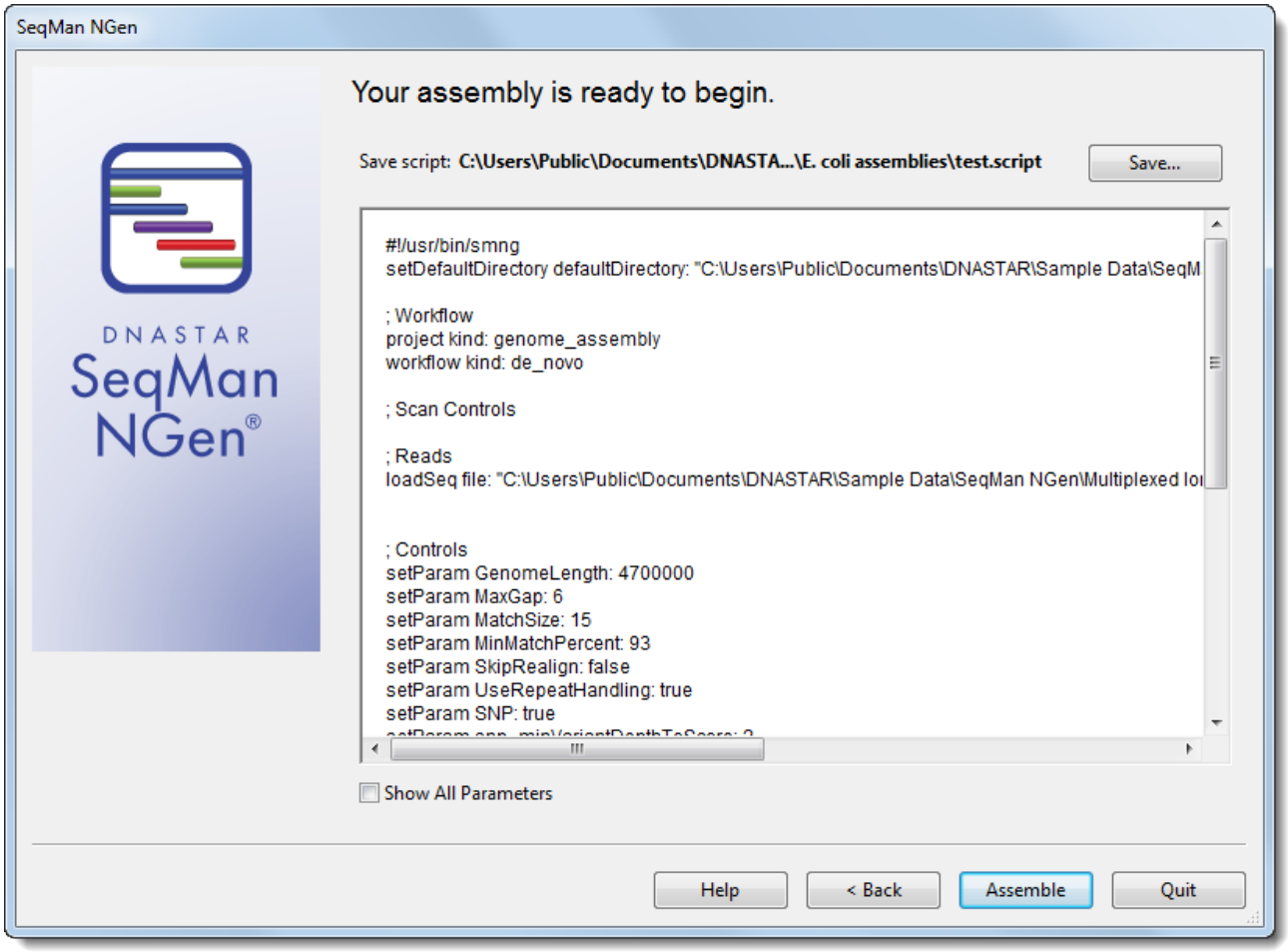


For more information, refer to the SNP Options tab of the Advanced Options dialog. Both SNP Options dialogs are identical except that one is a standalone dialog, while the other is part of a tabbed dialog.

Once you are finished, click **OK** to save changes and return to the Recalculate SNPs dialog, or **Cancel** to return without saving changes.

The "Your assembly is ready to begin" Dialog

This dialog is the final pre-assembly wizard dialog for all workflows:



The main part of this dialog shows the current script: a snapshot of the assembly set up and parameters.

Note: This text in this dialog is not editable. Changes can be made by returning to a previous page of the wizard and making alterations there.

If available, you may check **Show All Parameters** to view all parameters, rather than only the user-edited parameters. This can be useful if you want to keep a record of all of the parameter values used for an assembly. This checkbox is only present for certain types of workflows (e.g., *De Novo*, Special Templated).

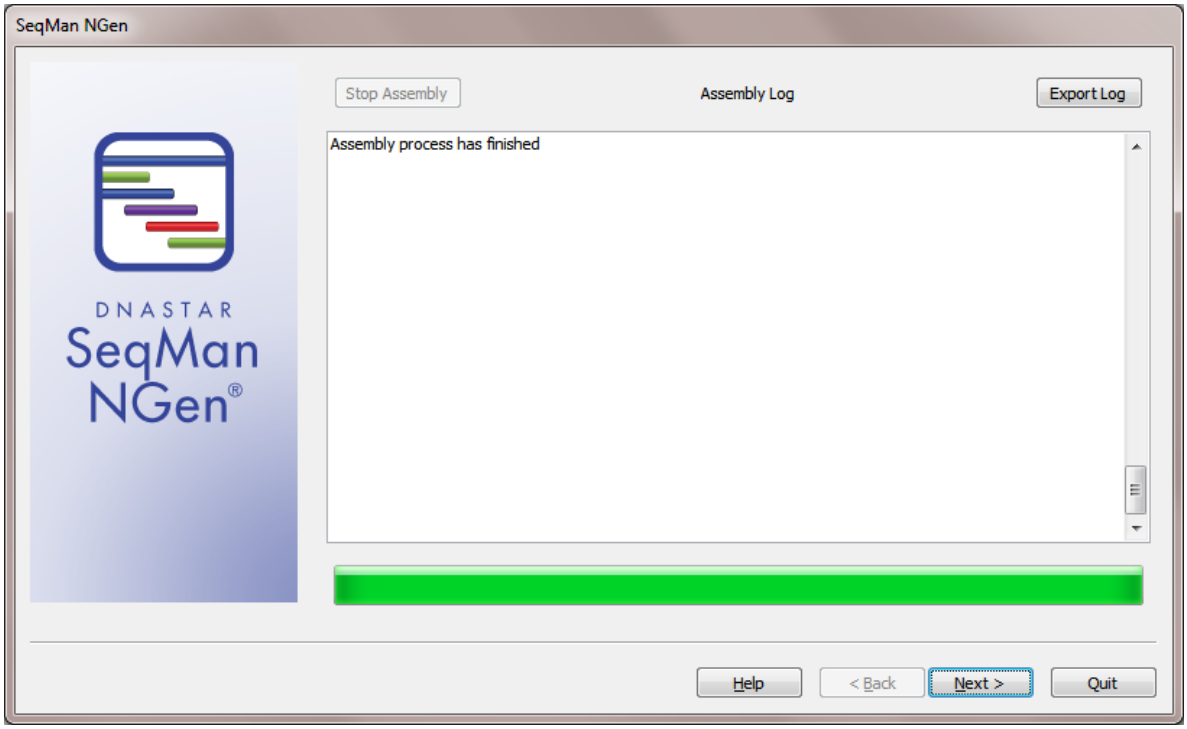
Click **Save** to save your project and convert your wizard choices into a SeqMan NGen assembly script (*.script). The resulting assembly script is an editable text file that can be modified and re-run if desired.

Note: When you **Save** after having checked the **Run as separate projects** box in the Input Sequence Files screen, a set of three separate scripts is saved for the project. If you save one or more of these scripts to a location other than the main project folder, any attempt to run the assemblies from the SeqMan NGen project script will fail. Moving the projects back to the main project folder will allow assembly to proceed.

Click the **Assemble** button to activate the script. The Assembly Log will open, displaying the status of the assembly.

The Assembly Log

After pressing the Assemble button from the "Your assembly is ready to begin" dialog, the Assembly Log opens, showing the status of the assembly. Once assembly has successfully completed, the following text will be displayed: "Assembly process has finished."



The Assembly Log has two buttons:

- **Export Log** – exports the progress information as a text file. A save dialog will open prompting you to choose a location in which to save the file.

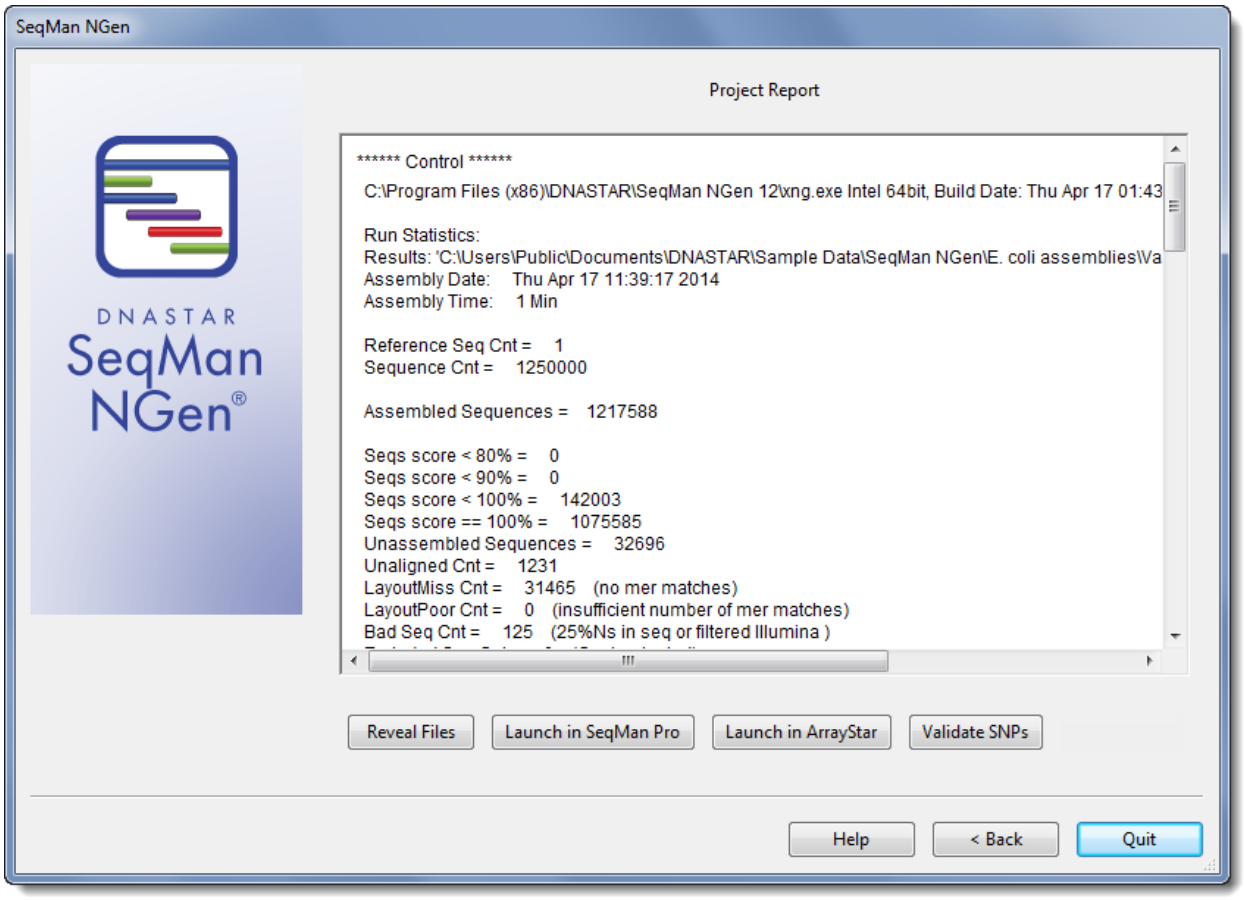
  Note for Windows users: To open a text report with the correct formatting displayed, we recommend using Wordpad, Notepad++, or Microsoft Excel®, and not the default Windows text editor, Notepad.

- **Stop Assembly** – aborts an assembly that is still in progress. Additional ways to halt the assembly include clicking **Ctrl+C** (Win) or **Cmd+C** (Mac). After halting assembly, you will see the message "Assembly process was not successful!"

Whether the process finishes on its own or is stopped manually, click **Next** to proceed to the Project Report dialog.

The Project Report Dialog

After assembly has finished in the Assembly Log, you will automatically be transported to the Project Report dialog.



If assembly failed, the dialog displays the message "Assembly failed. No report available." Otherwise, you will see the Assembly Report information in the body of the screen.

Between two and four buttons are displayed in the row under the report. The availability of a particular button depends on the workflow, and sometimes on the type of machine being used (e.g., Linux vs. Windows).

- **Reveal Files** – To open the folder where the assembly output and associated files are stored.
- **Launch in SeqMan Pro** – To launch the completed assembly in SeqMan Pro. This button is not available for assemblies performed using Linux. Instead, you will need to move the completed assembly to a Windows or Macintosh computer in order to view the assembly in SeqMan Pro.
- **Launch in ArrayStar –**To launch ArrayStar with the selected project open. This button is not available on Macintosh systems or for Metagenomics/16S rRNA or Viral - Host Integration workflows.
- **Validate SNPs -** To launch ArrayStar with the selected project open and to automatically run the SNP validity tests (i.e., **Statistics > Validation Control Accuracy**). This button is only available for the Templated assemblies with control workflow.

Note: In the case of multiple assembly projects, some of the buttons below will open a list of the projects. Choose the one you wish to open in SeqMan Pro or ArrayStar, and then click **OK**.

In addition to the usual **Help** and **< Back** buttons at the bottom of the dialog, there is one button that is unique to this dialog. The **Quit** button can be used to close SeqMan NGen, whether or not the assembly completed successfully.

The Assembly Report

A post-assembly text report is viewable in the Project Report dialog. This report summarizes your assembly statistics, including the parameters used, the number of assembled/unassembled sequences and contigs in your project, and the average quality scores. If you do any of the following, the report will be exported as a text file:

- Save the assembly in SeqMan Pro format (*.sqd) in the Set Up Project Files dialog.
- Check **Save Report** in the Set Up Project Files dialog.

Note for all users: The same information contained within this report is also saved within each SeqMan Project file (*.sqd) regardless of whether you choose to export the report by setting this parameter. The report can be viewed in SeqMan Pro by going to **Project > Report**.

Note for Windows users: To open a text report with the correct formatting displayed, we recommend using Wordpad, Notepad++, or Microsoft Excel®, and not the default Windows text editor, Notepad.

Some of the terms used in the Assembly Report are defined below:

Assembly Totals	
Contigs	Total number of contigs assembled.
Contigs > 2K	Total number of assembled contigs that are more than 2000 base pairs in length.
Contigs to Reach Genome Length x	Number of contigs needed to cover the genome length specified in the Workflow pane.
Assembled Sequences	Number of sequences utilized in the assembly.
Unassembled Sequences	Number of sequences excluded from the assembly.
All Sequences	Total number of sequences in the project.
Contig N50	Contig size at which 50% of the sequence data are represented.*
Average Coverage	Average depth of coverage in the assembly.
Average Totals	
Sequences Per Contig	Average number of sequences used for each contig.
Average Lengths	
Contigs	Average contig length.
Assembled Sequences	Average length of sequences used in the assembly.
Unassembled Sequences	Average length of sequences excluded from the assembly.
All Sequences	Average length of all sequences in the project.
Average Quality	
Assembled Sequences	Average quality score of sequences used in the assembly.
Unassembled Sequences	Average quality score of sequences excluded from the assembly.
All Sequences	Average quality score of all sequences in the project.
Assembly Parameters	The values specified in the Workflow, Reads, Controls and Actions tabs prior to assembly.

*In a typical microbial genome assembly, Contig N50 values exceed 80K base pairs and genome coverage is attained in less than 100 contigs. In many assemblies, contig N50 exceeds 100K with genome coverage attained in 25 contigs. If paired-end Roche 454 Life Sciences data are used, contigs can be ordered into a handful of large scaffolds to attain genome coverage that greatly facilitates gap closure and completion of the genome assembly.

Output Files for Different Workflows

The output file structure varies depending upon your workflow and on the assembler used for that workflow. SeqMan NGen uses two powerful assemblers: XNG and SNG (called SMNG in Linux).

- The XNG assembler (patent pending) is used for all templated assemblies, including reference-guided assembly. This assembler features an algorithm for fast, accurate assembly of extremely large genomes, and creates BAM-based outputs (e.g., *.assembly files).
- The SNG/SMNG assembler is used in both reference-guided assembly and *de novo* assembly. The SNG/SMNG assembler generates finished assemblies in any of four formats: SeqMan Pro (SQD), ACE, SAM or BAM.

CPU usage Note: The XNG assembler uses multiple cores, but the exact number varies over the course of the assembly. The SNG/SMNG assembler uses one core during assembly.

Click the links below to see a list of the output files for a given workflow.

XNG Workflow Output

This topic describes the outputs of XNG workflows. These include:

- Templated workflows.
- BAM alignment workflow (Welcome screen = **Import BAM file**; BAM Import = **Align BAM layout file**).
  SNP recalculation workflow (Welcome screen = **Import BAM file**; BAM Import = **Recalculate SNPs**).
- Reference-guided assembly with gap closure (Choose Project Type = **Genome assembly**; Choose Assembly Type dialog = **Reference-guided assembly with gap closure**). This workflow uses both the XNG and SNG assemblers, but the output files are most similar to XNG outputs.

Each workflow varies in the number and contents of output files and folders. Only a subset of items in the table below may appear for a particular workflow.

Note: In the table below, it should be understood that the project name precedes any hyphen (-) or period (.) used at the beginning of file and folder names.

Single assemblies	All gene panel projects Multiple sample assemblies run as separate assemblies
The project folder has the name specified in Set Up Project Files and contains: **.script** file (if saved in the "Your assembly is ready to begin" dialog). .assembly folder -noSplit.assembly folder (Reference-guided assembly with gap closure workflows only) -Reports folder -zinternal folder info folder	The project folder _D2HLink_49587 contains the name specified in Set Up Project Files followed by the suffix "**_assemblies**." This folder contains: **.script** file (if saved in the "Your assembly is ready to begin" dialog). **Results.txt** file - Overview information and statistics for each assembly. **.table.txt** file **.template.script** file **_arstar.script** file – A script to load all assemblies as a SNP project in ArrayStar. **_arstarValidation.script** file – (only if validation control was present) A script to load the validation control assembly and associated VCF file as a SNP project in ArrayStar and to automatically calculate the accuracy statistics. .assembly folders (one per sample) -Reports folders (one per sample) -zinternal folder info folder

Contents of the .assembly Folder

Note: The contents of the **-noSplit.assembly** folder are similar to those of the **.assembly** folder.

The **.assembly** folder is part of the output for XNG workflows.
In the table below, it should be understood that the project name precedes any hyphen (-) or period (.) used at the beginning of file and folder names.

File Extension	Description
It is intended that the entire **.assembly** folder be opened in SeqMan Pro for viewing and analysis of the assembly. However, the following individual files also contain useful information.	
.vcf	The VCF file for the assembly, if one was specified.
.bed, **.txt**, etc.	The target region file (.bed or manifest) for the assembly, if one was specified.
.templateInfo	Contains general information for each contig in the assembly.
.enrichment_Summary.txt	Contains the textual information for the **Project > Show coverage of target regions** option in SeqMan Pro.
***.sqd**	This file is only created when the ***.assembly** is first opened in SeqMan Pro. It contains saved display specific information such as SNP filtering criteria. Doubling clicking on this file will open the ***.assembly** package in SeqMan Pro.
There is normally no reason to open the following files.	
.auxPair	(internal use only)
.bam	The BAM formatted alignment file.
.bam.bai	The BAM index file.
.capture.userSNP.vcf	(internal use only)
.combined.snpExt	(internal use only)
.coverage	Contains information at each position along the contig where the coverage changes.
.coverage2	Contains information for the maximum coverage of 100 base pair intervals across the contig.
.coverage4	Contains information for the maximum coverage of 10,000 base pair intervals across the contig.
.coverage.missingSNP	Contains information about positions in dbSNP that had coverage and were called the reference base in the assembly.
.exomeCapture-features	(internal use only)
.info	Contains information used by SeqMan Pro in displaying the assembly.
.midinfo	(internal use only)
missing.fas	A fasta file of reads with no mers matching the reference.
missing.fas.qual	A base quality file of reads with no mers matching the reference.
.nocoverage.missingSNP	Contains information about positions in dbSNP that had no coverage in the assembly.
outofOrder.txt	A text file of sequence reads not included in the final assembly due to excessive trimming during the alignment phase.
.pair	(internal use only)

.pairDist	Contains information about the position and distance between paired end reads.
pairSpecifiers.txt	(internal use only)
poor.fas	A fasta file of reads rejected at the layout phase due to match scores below the threshold.
poor.fas.qual	A base quality file of reads rejected at the layout phase due to match scores below the threshold.
.quant	Reprises information in the .coverage4 .coverage2 and /or .coverage files.
.region_capture.bed	(internal use only)
report.txt	Contains the textual information for the **Project > Report** option in SeqMan Pro.
.snp	Contains all the information for SNPs called using the "Simple" method.
.snpExt	Contains all the information for SNPs called using either the "Diploid" or "Haploid" method.
SNPs.log	An optional text form of the .snpExt table that contains information on how each was calculated. If you encounter a problem, this file is useful for DNASTAR Support to help you with trouble-shooting.
.splitExt	(internal use only)
.template-comment	Contains the comment information for that contig.
.template-features	Contains the feature information for that contig.
.template-features2	(internal use only)
.template.fof	A file-of-files containing the path and file names of the reference sequences.
.template-gapped-seq	A .seq file of the template containing gaps.
.template-gaps	A binary file of the template gap information.
.template-seq	A .seq file of the template without gaps.
unaligned.fas.qual	A base quality file of reads rejected at the alignment phase.

Note for Windows users: To open text reports with the correct formatting displayed, we recommend using Wordpad, Notepad++, or Microsoft Excel®, and not the default Windows text editor, Notepad.

Contents of the -Reports Folder

The **-Reports** folder is part of the output for XNG workflows.
In the table below, it should be understood that the project name precedes any hyphen (-) or period (.) used at the beginning of file and folder names.

File Suffix or Extension	Description
-zinternal	(click link at left for details)
-enrichment_Summary.txt	(internal use only)
-perTemplateResults.txt	Overview information and assembly statistics per contig.
-projectReport.txt	Overview information of the assembly. The same report can be viewed within SeqMan Pro using the **Project > Report** menu command.
-unassembled.fastq	The unassembled reads from the assembly in Fastq format. If production of this file is not specified in the script, three files are created instead: • missing.fastq – unassembled reads with no hits to any template. • poor.fastq - unassembled reads with scores too low to include in the layout. • unaligned.fastq - unassembled reads included in the layout, but rejected by the aligner.

Contents of the -zinternal Folder

The **-zinternal** folder is part of the output for XNG workflows.
In the table below, it should be understood that the project name precedes any hyphen (-) or period (.) used at the beginning of file and folder names.

File Suffix or Extension	Description
info	(click link at left for details)
bamToSQD.script	for converting the assembly to .sqd format.
pairScheme.info	(internal use only)
results.txt	source of _perTemplateResults.txt.
The following files instruct SeqMan NGen to convert unassembled reads into a separate SQD project:	
batchUnassembled	An UNIX executable file for *de novo* assembly of unassembled reads with v. and vi.
batchUnassembled.table.txt	A table of values for running an SNG assembly (called SMNG in Linux) of the missing.fas reads against the template sequence.
batchUnassembled.template.script	The SNG (called SMNG on Linux) script containing variables that are specified by the batchMissing.table.txt.

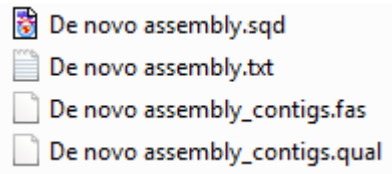
Contents of the info Folder

The **info** folder is part of the output for XNG workflows.
In the table below, it should be understood that the project name precedes any hyphen (-) or period (.) used at the beginning of file and folder names.

File Extension	Description
-[templateID].insertion2	Contains structural variation information.
-[templateID].sV_Edges.txt	Contains structural variation information.

SNG Workflow Output

The *de novo* workflows use SeqMan NGen's SNG assembler. For SNG workflows, the results folder contains the following files:



File Suffix or Extension	Description
.sqd	The main assembly output. To view and analyze the assembly, open this file with SeqMan Pro.
.txt	(internal use only)
-contigs.fas	Created when contigs are saved in FASTA format.
-contigs.qual	Created when contigs are saved in FASTA format. The values in the file are the sum of the base qualities at each position in the contig, up to a maximum of 90.

Note for Windows users: To open text reports with the correct formatting displayed, we recommend using Wordpad, Notepad++, or Microsoft Excel®, and not the default Windows text editor, Notepad.

How To…

View Assembly Results in SeqMan Pro

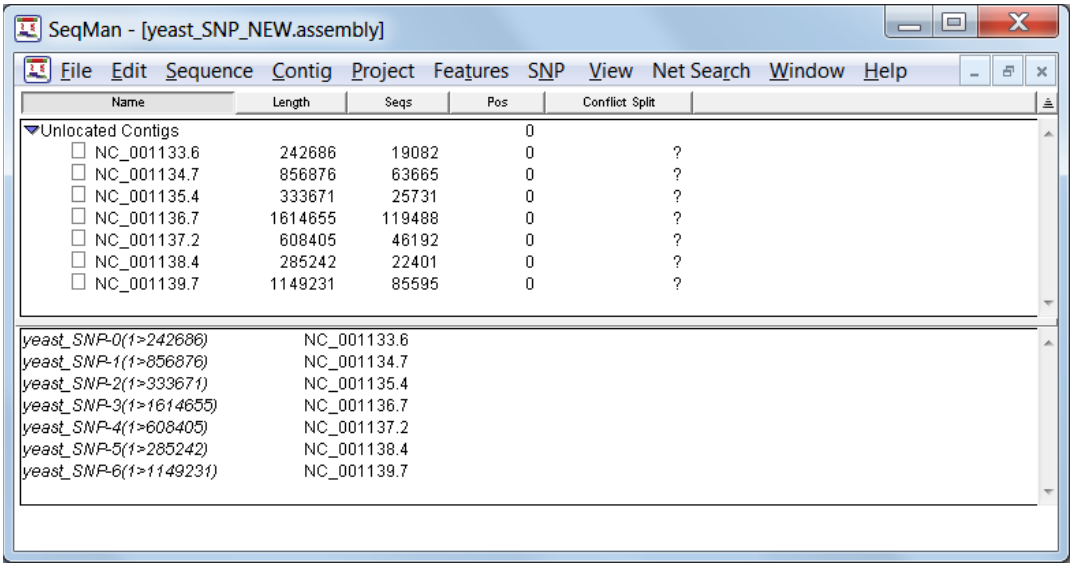
After assembly is complete, click the **Launch in SeqMan** button in the SeqMan NGen wizard to open the assembly in SeqMan Pro. Alternatively, click the **Reveal Files** button to locate the folder of assembly results, which can be dragged and dropped on the SeqMan Pro application to view the assembly.

Note for all users: If you've opted to create both an *.sqd and an *.assembly output, you may notice that the files do not exactly match. That's because *.sqd files—unlike *.assembly files—allow sequences both before and after the template sequence.

Note for Macintosh users: Macintosh only allows one copy of SeqMan Pro to be open at a time. If you are a Mac user who has opted to save your SeqMan NGen assembly in both *.sqd and *.assembly formats, the *.assembly file will be the first to open in SeqMan Pro. Once the *.sqd file has been created by SeqMan NGen, SeqMan Pro will prompt you to save the *.assembly file so it can open the *.sqd file instead.

Note for Linux users: SeqMan Pro is not available on Linux. Linux users must move the finished assembly to a Windows or Macintosh computer in order to view it in SeqMan Pro.

Assembly results will appear in the SeqMan Project Window.



The contigs in your project will be named as follows:

- If you assembled your data using a template sequence, the resulting contig will take the name of the template sequence name.
- If you used repeat handling, the contigs in your project made up of sequences flagged as possible repeats will be named: Repeat-00001, Repeat-00002, Repeat-00003, etc.
- If you scanned your assembly for known repeats, then the contigs containing the known repeated sequences will take the name of the repeated sequence.
- If none of the above applies, the contigs in your project will be named Contig 00001, Contig 00002, Contig 00003, etc.

Some SeqMan Pro menu options may be grayed out when working with large assembly projects such as BAM assemblies.

To open your project in SeqMan Pro at a later time, use **File > Open** to open a SeqMan Project (*.sqd) or SeqMan NGen assembly (*.assembly); or use **File > Import** to open an ACE (*.ace) file. You may also drag and drop the assembly project on SeqMan Pro. If your project file or original sequence files have been moved, SeqMan Pro will prompt you to locate the sequence files.

Detailed descriptions of all of SeqMan Pro's features can be accessed within SeqMan Pro by going to **Help > Contents**. The SeqMan Pro Help topics *Discovering SNPs* and *Working with Features* may be particularly useful.

Create a SeqMan NGen Assembly to Use with ArrayStar

ArrayStar® is DNASTAR's software for gene expression analysis, gene characterization and large set analysis. ArrayStar can use SeqMan NGen assemblies as input. Here are some common workflows which involve SeqMan NGen assemblies being analyzed and viewed in ArrayStar:

- Import transcriptome assemblies into ArrayStar as RNA-Seq projects to perform RPK, RPM or RPKM normalization on your data.
- Import SNP tables from your SeqMan NGen assemblies in order to compare large sets of SNPs and affected genes using ArrayStar's visualization tools.
- Import SeqMan NGen assemblies into ArrayStar to perform Validation Control Accuracy testing. See Create an Assembly for Validation Control Accuracy Testing.

Create an Assembly for Validation Control Accuracy Testing

In SeqMan NGen, create a project with the following settings:

1) Welcome screen: **Create new assembly project**.
2) Choose Project Type: **Exome assembly** or **Mendelian / germline gene panel assembly**.
3) Choose Assembly Type: **Templated assemblies with control**.
4) Input Template Files: Use the **Add Genome Package** to add the associated genome (in most cases, this will be "human"). In the bottom of the screen, leave both boxes checked and click the **Browse** button. Navigate to and open the targeted regions (*.txt or *.bed) file.
5) Input Sequence Files: Choose the **Read technology**, then add the read data for the Validation Control to the unpaired or paired sections. In the **Experiment** column, name each sample (e.g., "control"). Paired reads should be given the same names.
6) Set Up Experiments: In the row with the Validation Control experiment (there may only be one row total), check the **Is Control** box. Use the **Control Type** drop-down menu to select **Validation**. Press the **Set VCF File** button and Open the VCF (*.vcf) file corresponding to the experiment. In the ensuing dialog, select **VCF File for validation control experiment** and click **OK**.
7) Assembly Options: Settings can normally be left at their defaults.
8) Perform the assembly.

9) If you performed the assembly on a Windows machine and will be doing the ArrayStar analysis on the same machine, click the **Validate SNPs** button to launch ArrayStar and perform the analysis. Otherwise, see the ArrayStar help topic Validation Control Accuracy for further instructions.

Export ArrayStar Sequences to SeqMan NGen

ArrayStar® is DNASTAR's software for gene expression analysis, gene characterization and large set analysis.

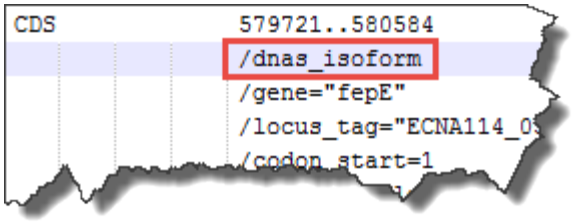
Within ArrayStar, launch the Export Sequences dialog via the **Data > Export Sequences** command. Within this dialog, be sure to check both **Export Matching Reads** and **Launch NGen afterwards**. The latter option will cause the template and matching reads to be loaded into SeqMan NGen for assembly and downstream analysis. In this case, the "template" sequence will be the single GenBank or FASTA file that contains the fragments of the original template sequences that match your selection. Click the **Export** button to export the sequences and automatically launch SeqMan NGen.

Manually Specify an Isoform

By default, SeqMan NGen chooses the longest CDS as the isoform for SNP calling. If desired, you may override the automated choice by specifying the preferred isoform manually in the template sequence.

To do so, follow these steps prior to importing the template sequence into SeqMan Wizard's Input Template Files screen:

1) Open the template (reference) sequence in a text editor.
2) Locate the feature of interest. Just below its location coordinates, type in **/dnas_isoform**.



3) Save the edited template sequence.
4) Input the saved version of the template sequence into SeqMan Wizard's Input Template Files screen.

Make a Custom VCF File

VCF files can be custom-made or automatically generated by sequencing software. For a description of various VCF version specifications, see the Sourceforge VCF Specification page.

Tables in VCF files must follow all of the rules below:

- The first two columns must be included, and in the same order as shown below.
- All cells in the first two columns must be filled.

- Four optional columns are allowed. If optional columns are present, the assembler will check the length of the string and compare against the length of the called variant. The base identities will not be checked. Further columns are allowed, but will be ignored.

- **IMPORTANT**: The table portion of the file must be sorted numerically, first by #CHROM, and then by POS. Make sure to sort the columns numerically (1, 2, 3…) and not alphabetically (1, 11, 12…). If you attempt to run the assembly after loading an improperly-sorted VCF file, multiple red error messages will be displayed during the assembly.

  Note: When you try to open extremely large VCF files in a spreadsheet program or text editor for sorting purposes, you may receive an "insufficient memory" warning. If you need to sort a VCF file that is too big to open on your machine, we recommend using Sourceforge's VCFTools.

If quotation marks are used anywhere in the VCF file, they must be straight quotes, not curly or "smart" quotes. In addition, quotation marks should not be used in lines beginning with `##contig`, `##UnifiedGenotyper`, or `##INFO`. If these rules are not followed, an error message will appear during assembly stating that "the VCF file has an incorrect or missing header." Though the assembly will continue, the VCF SNP file that is output will be empty.

#CHROM (required)	POS (required)	ID (optional)	REF (optional)	ALT (optional)	INFO (optional)	Column 7 and beyond (ignored)
Chromosome identifier. Numbers are preferred, but **chr** or **ch** prefixes are allowed. All cells in this column must be filled.	Position in the reference sequence. All cells in this column must be filled.	All cells must contain either: • For known dbSNP entries, the rs ID • For unknown or nonexistent IDs, a period (.)	The reference base(s) For unknown bases, a period (.)	The variant base(s) For unknown bases, a period (.)	User ID and source assembly information For unknown bases, a period (.)	These columns may contain data, but they will be ignored by the SeqMan NGen assembler

Note: Chromosome names are captured from genome template packages and used to assign contig IDs to entries from BED, VCF and manifest files. SeqMan NGen can read and produce output using common naming conventions (i.e., "chr" and "ch") and Arabic numerals. It understands that chr1, ch1, or 1 can all be used to represent "the first template in the index," and so on.

In addition, Genome Template Packages sometimes internally define "short names" for particular chromosomes. For example, the *C. elegans* template package names its chromosomes using the standard convention for that organism: "I", "II", "III", "IV", "V", "X", "M." SeqMan NGen does not normally recognize Roman numerals, but can in this case, because the numbers are "short names" that have been mapped to specific chromosomes.

Make a Custom BED File

If you chose **Exome assembly**, **Mendelian/germline gene panel assembly**, or **Cancer/somatic gene panel assembly** from the Choose Project Type screen, you may import a targeted regions file in the Input Template Sequence screen.

BED files are used to define capture regions in the assembly, and can be generated by the sequence provider or made by hand. These files are basically tab-separated text files whose extension has been changed to *.bed. See the UCSC Genome Bioinformatics BED file page for detailed information.

The BED file can consist of multiple sections, each with a different track name. Text is allowed between the tables without restriction.

Tables must follow all of the rules below:

- A header row (the one below appears in blue and black) is optional and can contain any text.
- The first three columns must be included, and in the same order as shown below.
- All cells in the first three columns must be filled.
- Additional table columns are allowed, but will be ignored.
- **IMPORTANT**: Each table in the file must be primarily sorted by the first column, and secondarily sorted by the second column. Make sure to sort the columns numerically (1, 2, 3…) and not alphabetically (1, 11, 12…).

Note: If only chromosome 1 (and possibly 11) appears in SeqMan Pro's "Coverage of Targeted Regions" report (**Project > Show coverage of target regions**), this is indicative of incorrect sorting.

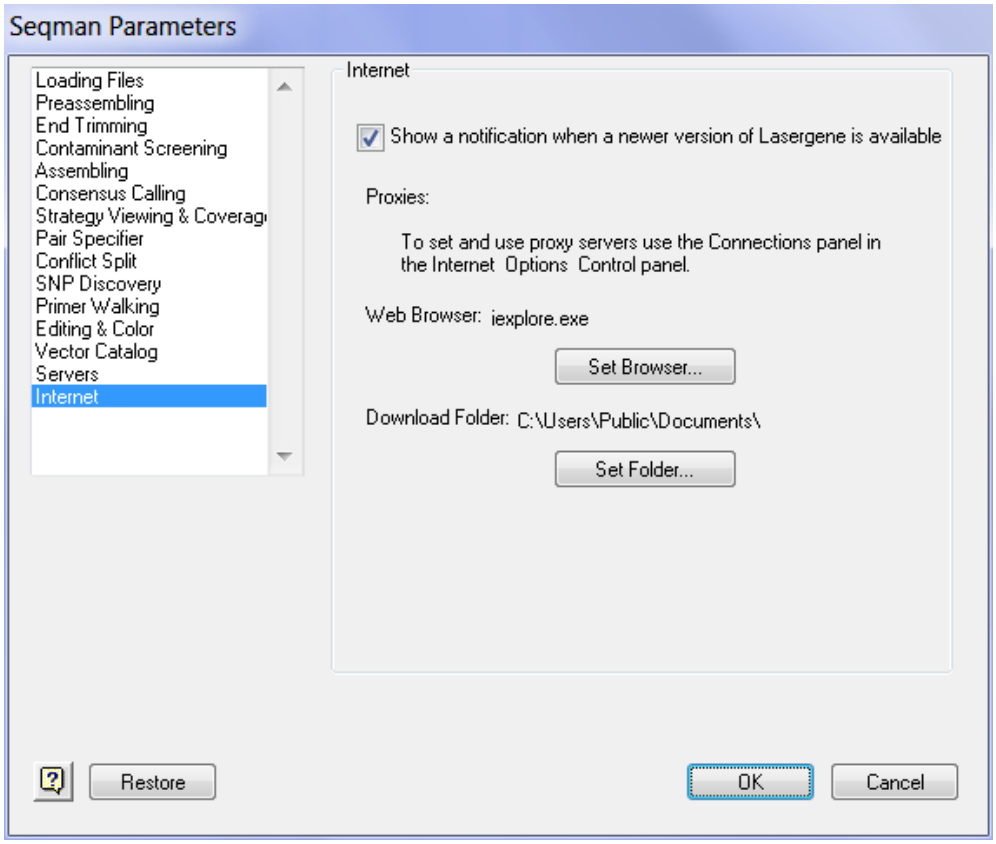
chrom (required)	chromStart (required)	chromEnd (required)	Column 4 and beyond (ignored)
The name of the chromosome or scaffold. Numbers are preferred, but **chr** or **ch** prefixes are allowed.	Starting position for the feature.	Ending position for the feature.	Data in these columns are ignored.

Note: SeqMan NGen can read and produce output using a variety of common chromosome naming conventions, including "chr1" and "ch1," as well as Arabic and Roman numerals. Chromosome names are captured from genome template packages and used to assign contig IDs to entries from BED, VCF and manifest files.

Control Automatic Software Updates

SeqMan NGen respects the options you choose in SeqMan Pro regarding the automatic updating of software. To check or change the setting:

1) Launch SeqMan Pro.
2) Choose **Project > Parameters** from the menu.
3) Select **Internet** from the list on the left.
4) Check the box next to **Show a notification when a newer version of Lasergene is available** if you want Lasergene to automatically check for software updates. Otherwise, uncheck the box.
5) Click **OK** to save your changes.



Frequently Asked Questions

Why doesn't SeqMan NGen run in the command line?

If you are a command line user and type: sng [or smng, or xng] "path to script," SeqMan NGen may not initially run. This happens when the application has not been added to your environment variable PATH. After installing SeqMan NGen on a computer for the first time, command-line users may want to restart the computer before using the application. Restarting will add the application to your environment variable PATH.

If you are unable to restart your computer, the following two options are available:

1) Use the wizard to create and run the script.
2) At the command-line, specify the path to the application followed by the path to the script.

Why isn't SeqMan NGen on Ubuntu's installed software list?

SeqMan NGen does not appear in the Ubuntu Software Center's **Installed Software** list. This issue has been observed on Linux machines running Ubuntu 10.04, but may occur with other versions, as well.

The issue can be resolved using either of the methods below:

- Perform at least one SeqMan NGen assembly prior to opening the Ubuntu Software Center.
- Open the Synaptic Package Center by choosing **SYSTEM > Administration > Synaptic Package Manager**. Search for SeqMan NGen. After confirming the presence of SeqMan NGen, close the Package Manager and re-launch the Software Center.

After using either method, SeqMan NGen Linux should now be listed in the Software Center's **Installed Software** list.

Why is the "Export Aligned" value higher than expected?

Since split reads are counted in each location where they align, it is possible for the value in the export aligned column ("Export_Aligned") to exceed the total number of sequences ("NumSeqs"). The number of exported split reads can be viewed in the Assembly Report's "Export_Split_Cnt" column.

Why is the MID column missing from the SeqMan Pro SNP Report?

In order for SeqMan Pro to display the MID column in the SNP Report, you must choose either the diploid bayesian or haploid bayesian method of SNP calculation prior to assembly in SeqMan NGen. This can be done from the following dialogs:

- **Assembly Options** dialog > **Advanced Assembly Options** button > **SNP** tab > SNP calculation method drop-down menu = **Diploid bayesian** or **Haploid bayesian**.
- **Recalculate SNPs** > **SNP Options** button > SNP calculation method drop-down menu = **Diploid bayesian** or **Haploid bayesian**.

If you instead choose the "simple percentage" method for SNP calculation from the drop-down menu, the MID column will be omitted from the SNP Report.

What file extensions are used for unassembled sequences?

In SeqMan NGen 2.2 and earlier, unassembled sequences were saved in FASTA and QUAL formats (*.fas., *.qual). In SeqMan NGen 3.0 or later, these sequences are instead saved in FASTQ format (*.fastq).

Note that FASTQ files created with the SeqMan NGen wizard will have a *.fastq extension, while those created via a command line script will have a *.fas extension.

Why can't I add a downloaded genome package as my template?

Downloaded genome packages are saved on your computer as ZIP files, and must be extracted prior to use.

To extract the genome package:

- On **Macintosh**: Double-click on the ZIP file. The files will be automatically extracted via the Archive Utility.
- On **Windows Vista** or **Windows 7**: Double-click on the ZIP file. In the ensuing Explorer window, click **Extract all files** from the top left. Choose a location for the files and select **Extract**.

To add the genome package to SeqMan NGen:

1) Open the SeqMan NGen wizard and follow the on-screen instructions.
2) In the "Add Template Files" dialog, select **Add Genome Package**.
3) Navigate to the extracted file, which has the extension *.genometemplate.
4) Click **Open**.

Why do assembly statistics vary from version 3.0 to 3.1?

When you perform the identical assembly in SeqMan NGen versions 3.0 and 3.1, some assembly statistics may appear “worse” in the newer version. This is actually due to an improvement in SeqMan NGen’s calculation algorithm. The criterion for calculating contig length is now more stringent and excludes positions that contain more than 50% gaps. In general, this causes the average contig length in version 3.1 to be shorter than in version 3.0. The statistics that may be affected by this improvement include:

- Contigs to Reach Genome Length
- Contig N50
- Average Coverage
- Average Lengths: Contigs

This change in the algorithm assures that you will get the most accurate assembly possible.

Appendix

Supported File Types

For a list of supported file formats, please see the SeqMan NGen Supported File Types page of our website.

Manifest File Formats

If you chose **Exome assembly**, **Mendelian/germline gene panel assembly**, or **Cancer/somatic gene panel assembly** from the Choose Project Type screen, you may import a targeted regions file in the Input Template Sequence screen.

Manifest files are tab-separated files used to define the chromosomal coordinates of gene targets in the assembly, and can be made by hand or automatically generated, e.g., by Illumina. (See Illumina's manifest file PDF for a description.) Manifest files can have various file extensions, though *.txt is commonly used.

The following examples show the most basic required columns for the manifest file, as well as two formats that are used by Illumina. These examples are provided in case you need to trouble-shoot problems with existing manifest files. If you want to create your own targeted regions file, we recommend making a BED file rather than a manifest file. See Make a Custom BED File for detailed instructions.

Note: The columns below can appear in any order. Columns with pink headers are required, while columns with green headers are optional

User-made file (most basic version):

[Regions]	Chromosome	Start	End
Name			
28324371	chrM	577	647

Illumina manifest file – format 1:

[Targets]		Target Number	Chromosome	Start Position	End Position	Probe Strand	Sequence	Species	Build ID
TargetA	TargetB								
chr1.43815008.43815009	chr1.43815008.43815009	1	chr1	43814982	43815163	+	GGT(…)GCC	Homo sapiens	hg19
chr1.115256528.115256531	chr1.115256528.115256531	1	chr1	115256500	115256680	+	TCT(…)TTA	Homo sapiens	hg19

Illumina manifest file – format 2:

[Regions]	Chromosome	Start	End	Upstream Probe Length	Downstream Probe Length
Name					
28324371	chrM	577	647	0	0

Note: SeqMan NGen can read and produce output using a variety of common chromosome naming conventions, including "chr1" and "ch1," as well as Arabic and Roman numerals. Chromosome names are captured from genome template packages and used to assign contig IDs to entries from BED, VCF and manifest files.

Repeat Handling

Repeat handling parameters compute a threshold for deciding the number of identical subsequences of bases (mers) used to indicate a putative repeat. Mers that are common to two or more fragment reads are aligned to determine the overall layout of reads. (For additional information, see the Mer Tags section.)

Repeat handling is controlled via the **Repeat read placement** command in the Layout Options dialog, or in the repeat handling parameters of the Advanced Assembly Options dialog.

The repeat threshold can also be computed by multiplying the **Match repeat percent** parameter value in the Advanced Assembly Options dialog by the **Expected coverage**. Any mer that occurs

more frequently than the computed threshold is not considered for use as a mer tag in determining overlaps. Coverage can be determined in two ways:

- By using the length of the genome/fragment being sequenced, as specified by the **Expected genome length** parameter in the Assembly Options dialog. If this option is used, SeqMan NGen calculates expected coverage by dividing the **Expected genome length** value by the total length of all sequences in the project.
- By using a fixed number for expected coverage, as specified by the **Expected coverage** parameter in the Assembly Options dialog.

Detection of Structural Variations

In addition to SNPs and small insertions and deletions, genetic variation can also involve large scale rearrangements. These rearrangements may include large insertions and deletions, inversions, and translocations—collectively known as *structural variations* (*SV's*).

During a templated data assembly, SeqMan NGen automatically detects insertions or deletions (*indels*) greater than 10 bp based on a combination of two data types:

Coverage – read depth can be suggestive of larger indels, including duplications or the collapse of a duplication.

Split reads – reads spanning a deletion in the new genome relative to the reference genome can be "split" into two segments based on matches to discontinuous regions on the reference. For example, the following split read alignment indicates there is a 35bp deletion in the new genome:

Ref: AGGCTGACCTCCTGGGCCTAAGACACTGAGTGCCCCAATATGACTCGACTAGCA

Read: AGGCTGACCTCGACTAGCA

Ref: AGGCTGACCTCCTGGGCCTAAGACACTGAGTGCCCCAATATGACTCGACTAGCA

Split: AGGCTGACCTC GACTAGCA

The SeqMan NGen algorithm requires that four criteria be met for splitting a read:

1) At least 25 bases on each "half" of the split must match the reference. This means that reads must be at least 50bp long, though in practice they should be > 75bp.
2) The first mer match must be within 10 bases of the start of the read, and the final mer match must be within 10 bases of the end of the read. This increases the likelihood that the entire read will align after splitting.

3) The distance between the two closest mers on either side of the split must be within 20% of the total read length. For example, in a 100 base read where bases 5-30 make up the mer match on the 5' "half" of the read, then the first mer match on the 3' half must start between bases 31 and 50 (30+(100*0.2)=50) of the read. This relatively simple requirement allows for SNPs or sequencing errors near the actual split to be tolerated and resolved during alignment.

4) The two "halves" must be aligned in the same orientation.

In practice, two copies of the read are given to the aligner: one seeded with the 5' mer match and the other with the 3' mer match. The aligner then extends the alignment on both sides of each copy, and then trims each copy to maximize the final alignment score. It is the final trimmed internal position for each copy that is reported in SeqMan Pro's Structural Variation Report.

Note: Because of the trimming and the flexibility in the location where the nearest internal mer match must begin (criteria #3 above), it is possible that base substitutions are present in the split region that will not be displayed in SeqMan Pro's Alignment View. Thus, while split reads have far greater resolution power than coverage, the breakpoints identified must be considered as provisional.

To view a tabular report with structural variation findings for an assembly, open the assembly in SeqMan Pro and select **Contig > Structural Variation Report**.

Equivalence Between Wizard Settings and SNG Scripting Commands

The three left-most columns of the table below show the applicable SNG wizard screen, setting name, and the default value for that setting.

The next two columns show the equivalent scripting command and parameter described in the SeqMan NGen Scripting Manual.

The right-most column shows the alphanumeric code under which the command appears in the Scripting Manual. All commands are found in the section entitled "IV. SNG Commands."

Wizard Screen	Wizard Setting	SNG Wizard Default	Equivalent SNG Scripting Command	Equivalent SNG Scripting Parameter	Code in Scripting Manual v. 12
Advanced Assembly Options (De Novo)	Default quality	15	setParam	DefaultQuality	P4
	Default template quality	500	setParam	TemplateDefaultQuality	P36
	Gap penalty	30	setParam	GapPenalty	P6
	Match repeat percent	150	setParam	MatchRepeatPercent	P11
	Match score	10	setParam	MatchScore	P12
	Match spacing	75	setParam	MatchSpacing	P14
	Match window	50	setParam	MatchWindowLength	P15
	Max gap	15	setParam	MaxGap	P18
	Max usable	25	setParam	MaxUsableCount	P19
	Maximum coverage	0	setParam	MaxAssemblyCoverage	P16
	Mismatch penalty	20	setParam	MismatchPenalty	P23
	SNP low cover cutoff	0	setParam	LowCoverageThreshold	P10
	SNP match percent	90	setParam	SNPMatchPercentage	P31
	SNP passes	2	setParam	SNPPasses	P33
Advanced Trim/Scan Options	3' trim	0	FixedTrim	end3	U1
	3’ value is measured from 5’end	false	trimRelative	(N/A)	U3

	5’ trim	0	FixedTrim	end5	U2
	Flag length	50	setRepeatParam	MinFlagLength	R5
	Mer length	17	setContaminantParam	MerLength	O1
	Mer length	17	setRepeatParam	MerLength	R3
	Mer length	9	setVectorParam	MerLength	S7
	Minimum matches	12	setContaminantParam	MinMerMatch	O2
	Minimum matches	2	setRepeatParam	MinMerMatch	R6
	Minimum matches	3	setVectorParam	MinMerMatch	S9
	Minimum quality	20	setQualityParam	MinAveLowQual	Q5
	Trim length	30	setVectorParam	MinTrimLength	S10
	Trim to end	25	setVectorParam	EndCutOff	S2
	Window	5	setQualityParam	WinLength	Q8
Assembly Options (De Novo, Special Templated)	Mer size	21	setParam	MatchSize	P13
	Minimum length	0	removeSmallContigs	minLength	W
	Minimum match percentage	85	setParam	MinMatchPercent	P22
	Minimum sequences	10	removeSmallContigs	minSeqs	W2
	Realign reads after assembly	false	setParam	SkipRealign	P24
	Remove small contigs after assembly	true	removeSmallContigs	(N/A)	W1

	Repeat handling	true	setParam	UseRepeatHandling	P38
Input Sequence Files	(choose 454 and enter paired data)	false	DiscardLinkerLess	(N/A)	JJ2
Read Options	Contaminant scan	false	contamScan	(N/A)	T2
	Maximum total reads	10,000,000	maxSeqs	(N/A)	K10
	Quality end trim	true	assemble	trimEnds	T5
	Repeat scan	false	repeatScan	(N/A)	T4
	Vector/adaptor scan	false	vectScan	(N/A)	T6
Set Up Project Files (De Novo, Special Templated)	Save unassembled reads	false	setParam	Assemble Boneyard	P2
Your assembly is ready to begin	N/A	true	doAssemble	(N/A)	T3

Complete List of Parameters by Read Technology

Normal Templated Assembly – Metagenomics

Parameters	Illumina < 50 nt	Illumina > 50 nt	454	Ion Torrent	Pac Bio	Other
Set Pair Information, if paired	500	500	3000	user defined	No pairs allowed	5000
Assembly Options						
Mer Size	19	21	21	19	15	25
Minimum Match Percentage	93	93	85	90	80	90
Genome Ploidy	Population/ Other	Population/ Other	Population/ Other	Population/ Other	Population/ Other	Population/ Other
Advanced Options						
Layout Tab						
Repeat Read Placement	Place all	Place all	Place all	Place all	Place all	Place all
Maximum Total Reads	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)
Maximum Repeat Count	100	100	100	100	100	100
Assembly Output Format	BAM assembly package	BAM assembly package	BAM assembly package	BAM assembly package	BAM assembly package	BAM assembly package
Split Templates at No Coverage	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
Alignment Tab						
Minimum Aligned Length	25	35	50	25	45	50

Maximum Gap Size	6	6	15	15	20	15
Minimum Match Percentage	93	93	85	90	80	90
Match Score	10	10	10	10	10	10
Mismatch Penalty	20	20	20	20	20	20
Gap Penalty	30	30	30	30	30	30
Auto Trim Reads	True	True	True	True	True	True
SNP Tab						
Calculate SNPs	True	True	True	True	True	True
SNP Calculation Method	Simple percentage	Simple percentage	Simple percentage	Simple percentage	Simple percentage	Simple percentage
Minimum SNP Percentage	1	1	1	1	1	1
SNP Confidence Threshold	10	10	10	10	10	10
Minimum SNP Count	2	2	2	2	2	2
Minimum Base Quality Score	5	5	5	5	5	5
Check Strands	False	False	False	False	False	False

Normal Templated Assembly – All Others

Parameters	Illumina < 50 nt	Illumina > 50 nt	454	Ion Torrent	Pac Bio	Other
Set Pair Information, if paired	500	500	3000	User defined	No pairs allowed	5000
Assembly Options						
Mer Size	19	21	21	19	15	21
Minimum Match Percentage	93	93	85	90	80	90
Genome Ploidy	Diploid	Diploid	Diploid	Diploid	Diploid	Diploid
Advanced Options						
Layout Tab						
Repeat Read Placement	Place once	Place once	Place once	Place once	Place once	Place once
Maximum Repeat Count	100	100	100	100	100	100
Maximum Total Reads	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)	Unselected (10000000)
Assembly Output Format	BAM assembly package	BAM assembly package	BAM assembly package	BAM assembly package	BAM assembly package	BAM assembly package
Split Templates at No Coverage	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
Alignment Tab						
Minimum Aligned Length	25	35	50	25	45	50
Maximum Gap Size	6	6	15	15	20	15
Minimum Match Percentage	93	93	85	90	80	90
Match Score	10	10	10	10	10	10
Mismatch Penalty	20	20	20	20	20	20
Gap Penalty	30	30	30	30	30	30
Auto Trim Reads	True	True	True	True	True	True
SNP Tab						
Calculate SNPs	True	True	True	True	True	True
SNP Calculation Method	Diploid bayesian	Diploid bayesian	Diploid bayesian	Diploid bayesian	Diploid bayesian	Diploid bayesian

Minimum SNP Percentage	5	5	5	5	5	5
SNP Confidence Threshold	10	10	10	10	10	10
Minimum SNP Count	2	2	2	2	2	2
Minimum Base Quality Score	5	5	5	5	5	5
Check Strands (SNP_Checkstrandedness)	False	False	False	False	False	False

Special Templated Assembly - Genome

Save Project As	SeqMan Format 10M read limit Editable
Save Unassembled Reads	FALSE
Save Contigs To FASTA	FALSE
Save Report	TRUE

Parameters	Illumina < 50 nt	Illumina > 50 nt	454	Ion Torrent	PacBio	Sanger	Other
Set Pair Information, if paired	500	500	3000	User defined	No pairs allowed	3000	5000
Read Options							
Maximum Total Reads	10000000	10000000	10000000	10000000	10000000	10000000	10000000
Quality End Trim	True	True	True	True	True	True	True
Vector/Adaptor Scan	False	False	False	False	False	False	False
Contaminant Scan	False	False	False	False	False	False	False
Repeat Scan	False	False	False	False	False	False	False
Advanced Trim/Scan Options							
Quality End Trimming							
Minimum Quality	20	20	20	15	5	20	20
Window	5	5	5	5	5	5	5
Fixed End Trimming							

Do Fixed End Trimming	False	False	False	False	False	False	False
5' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
3' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
3' Value is Measured From 5' End	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
Other End Trimming Options							
Trim To Mer	False	False	False	False	False	False	False
Vector/Adaptor Scan							
Mer Length	9	9	9	9	9	9	9
Minimum Matches	3	3	3	3	3	3	3
Trim Length	30	30	30	30	30	30	30
Trim to End	25	25	25	25	25	25	25
Repeat Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	2	2	2	2	2	2	2
Flag Length	50	50	50	50	50	50	50
Contaminant Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	12	12	12	12	12	12	12
Assembly Options							
Mer Size	15	21	21	19	15	25	21
Minimum Match Percentage	93	93	85	90	80	90	90
Realign Reads After Assembly	True	True	True	True	True	True	True
De novo Assemble Remaining Unassembled Reads	False	False	False	False	False	False	False
Split Template at Zero Coverage	False	False	False	False	False	False	False

Remove Small Contigs After Assembly	True	True	True	True	True	True	True
___ Sequence Minimum	100	100	10	10	10	10	10
Advanced Assembly Options							
Match Score	10	10	10	10	10	10	10
Mismatch Penalty	20	20	20	20	20	20	20
Gap Penalty	30	30	30	30	30	30	30
Max Gap	6	6	15	15	20	15	15
SNP Passes	2	2	2	2	2	2	2
SNP Match Percent	90	90	90	90	90	90	90
SNP Low Cover Cutoff	0	0	0	0	0	0	0
Match Window	50	50	50	50	50	50	50
Maximum Coverage	0	0	0	0	0	0	0
Match Repeat Percent	150	150	150	150	150	150	150
Match Spacing	10	50	75	20	150	150	10
Default Quality	15	15	15	15	15	15	15
Default Template Quality	500	500	500	500	500	500	500
Max Usable	25	25	25	25	25	25	25

Only when "Save project as BAM Format" is checked	Special Genome
Assembly Options	
Genome Ploidy	Diploid
SNP Options	
Calculate SNPs	True
SNP Calculation Method	Diploid bayesian
Minimum SNP Percentage	5
SNP Confidence Threshold	10
Minimum SNP Count	2
Minimum Base Quality Score	5
Check Strands	False

De Novo Assembly - Transcriptome

	Illumina < 50 nt	Illumina > 50 nt	454	Ion Torrent	Pac Bio	Sanger	Other
Set Pair Information, if paired	500	500	3000	User defined	No pairs allowed	3000	5000
Read Options							
Maximum Total Reads	10000000	10000000	10000000	10000000	10000000	10000000	10000000
Quality End Trim	True	True	True	True	True	True	True
Vector/Adaptor Scan	False	False	False	False	False	False	False
Contaminant Scan	False	False	False	False	False	False	False
Repeat Scan	False	False	False	False	False	False	False
Advanced Trim/Scan Options							
Quality End Trimming							
Minimum Quality	20	20	20	15	5	20	20
Window	5	5	5	5	5	5	5
Fixed End Trimming							
Do Fixed End Trimming	False	False	False	False	False	False	False
5' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled

3' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
3' Value is Measured from 5' End	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
Other End Trimming Options							
Trim To Mer	False	False	False	False	False	False	False
Vector/Adaptor Scan							
Mer Length	9	9	9	9	9	9	9
Minimum Matches	3	3	3	3	3	3	3
Trim Length	30	30	30	30	30	30	30
Trim to End	25	25	25	25	25	25	25
Repeat Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	2	2	2	2	2	2	2
Flag Length	50	50	50	50	50	50	50
Contaminant Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	12	12	12	12	12	12	12
Assembly Options							
Mer Size	15	21	21	19	15	25	21
Minimum Match Percentage	93	93	85	90	80	90	90
Realign Reads After Assembly	True	True	True	True	True	True	True
Remove Small Contigs After Assembly	True	True	False	True	True	True	True
___ Sequence Minimum	100	100	10	10	10	10	10
Advanced Assembly Options							
Match Score	10	10	10	10	10	10	10
Mismatch Penalty	20	20	20	20	20	20	20

Gap Penalty	30	30	30	30	30	30	30
Max Gap	6	6	15	15	20	15	15
SNP Passes	2	2	2	2	2	2	2
SNP Match Percent	90	90	90	90	90	90	90
SNP Low Cover Cutoff	0	0	0	0	0	0	0
Match Window	50	50	50	50	50	50	50
Maximum Coverage	0	0	0	0	0	0	0
Match Repeat Percent	150	150	150	150	150	150	150
Match Spacing	10	50	75	20	150	150	10
Default Quality	15	15	15	15	15	15	15
Default Template Quality	500	500	500	500	500	500	500
Max Usable	25	25	25	25	25	25	25

Only when Save project as BAM Format is checked	De Novo Transcriptome
Assembly Options	
Genome Ploidy	Diploid
SNP Options	
Calculate SNPs	TRUE
SNP Calculation Method	Diploid bayesian
Minimum SNP Percentage	5
SNP Confidence Threshold	10
Minimum SNP Count	2
Minimum Base Quality Score	5
Check Strands	False

De Novo Assembly - Genome Assembly

Save Project As	SeqMan Format 10M read limit Editable
Save Unassembled Reads	FALSE
Save Contigs To FASTA	TRUE
Save Report	TRUE

	Illumina < 50 nt	Illumina > 50 nt	454	Ion Torrent	Pac Bio	Sanger	Other
Set Pair Information, if paired	500	500	3000	User defined	No pairs allowed	3000	5000
Read Options							
Maximum Total Reads	10000000	10000000	10000000	10000000	10000000	10000000	10000000
Quality End Trim	True	True	True	True	True	True	True
Vector/Adaptor Scan	False	False	False	False	False	False	False
Contaminant Scan	False	False	False	False	False	False	False
Repeat Scan	False	False	False	False	False	False	False
Advanced Trim/Scan Options							
Quality End Trimming							
Minimum Quality	20	20	20	15	5	20	20
Window	5	5	5	5	5	5	5
Fixed End Trimming							
Do Fixed End Trimming	False	False	False	False	False	False	False
5' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
3' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
3' Value is Measured from 5' End	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
Other End Trimming Options							
Trim to Mer	False	False	False	False	False	False	False
Vector/Adaptor Scan							

Mer Length	9	9	9	9	9	9	9
Minimum Matches	3	3	3	3	3	3	3
Trim Length	30	30	30	30	30	30	30
Trim To End	25	25	25	25	25	25	25
Repeat Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	2	2	2	2	2	2	2
Flag Length	50	50	50	50	50	50	50
Contaminant Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	12	12	12	12	12	12	12
Assembly Options							
Repeat Handling	True	True	True	True	True	True	True
Expected Genome Length	0	0	0	0	0	0	0
Expected Coverage	20	20	20	20	20	20	20
Mer Size	15	21	21	19	15	25	21
Minimum Match Percentage	93	93	85	90	85	90	90
Realign Reads After Assembly	True	True	True	True	True	True	True
Remove Small Contigs After Assembly	True	True	True	True	True	True	True
___ Sequence Minimum	100	100	10	10	10	10	10
Advanced Assembly Options							
Match Score	10	10	10	10	10	10	10
Mismatch Penalty	20	20	20	20	20	20	20
Gap Penalty	30	30	30	30	30	30	30
Max Gap	6	6	15	15	20	15	15

SNP Passes	2	2	2	2	2	2	2
SNP Match Percent	90	90	90	90	90	90	90
SNP Low Cover Cutoff	0	0	0	0	0	0	0
Match Window	50	50	50	50	50	50	50
Maximum Coverage	0	0	0	0	0	0	0
Match Repeat Percent	150	150	150	150	150	150	150
Match Spacing	10	50	75	20	150	150	10
Default Quality	15	15	15	15	15	15	15
Default Template Quality	500	500	500	500	500	500	500
Max Usable	25	25	25	25	25	25	25

Only when Save project as BAM Format is checked	***De Novo* Genome**	***De Novo* Metagenomics**
Assembly Options		
Genome Ploidy	Diploid	Population /Other
SNP Options		
Calculate SNPs	True	True
SNP Calculation Method	Diploid bayesian	Simple percentage
Minimum SNP Percentage	5	1
SNP Confidence Threshold	10	10
Minimum SNP Count	2	2
Minimum Base Quality Score	5	5
Check Strands	FALSE	FALSE

De Novo Assembly - Metagenomics

Save Project As	SeqMan Format 10M read limit Editable
Save Unassembled Reads	FALSE
Save Contigs To FASTA	TRUE
Save Report	TRUE

	Illumina < 50 nt	Illumina > 50 nt	454	Ion Torrent	Pac Bio	Sanger	Other
Set Pair Information, if paired	500	500	3000	User defined	No pairs allowed	3000	5000
Read Options							
Maximum Total Reads	10000000	10000000	10000000	10000000	10000000	10000000	10000000
Quality End Trim	True	True	True	True	True	True	True
Vector/Adaptor Scan	False	False	False	False	False	False	False
Contaminant Scan	False	False	False	False	False	False	False
Repeat Scan	False	False	False	False	False	False	False
Advanced Trim/Scan Options							
Quality End Trimming							
Minimum Quality	20	20	20	15	5	20	20
Window	5	5	5	5	5	5	5
Fixed End Trimming							
Do Fixed End Trimming	False	False	False	False	False	False	False
5' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
3' Trim	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
3' Value is Measured from 5' End	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
Other End Trimming Options							
Trim To End	False	False	False	False	False	False	False
Vector/Adapto							

r Scan							
Mer Length	9	9	9	9	9	9	9
Minimum Matches	3	3	3	3	3	3	3
Trim Length	30	30	30	30	30	30	30
Trim to End	25	25	25	25	25	25	25
Repeat Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	2	2	2	2	2	2	2
Flag Length	50	50	50	50	50	50	50
Contaminant Scan							
Mer Length	17	17	17	17	17	17	17
Minimum Matches	12	12	12	12	12	12	12
Assembly Options							
Repeat Handling	False	False	False	False	False	False	False
Expected Genome Length	N/A	0	0	0	0	0	0
Expected Coverage	N/A	20	20	20	20	20	20
Mer Size	15	21	21	19	15	25	21
Minimum Match Percentage	93	93	85	90	85	90	90
Realign Reads After Assembly	True	True	True	True	True	True	True
Remove Small Contigs After Assembly	True	True	True	True	True	True	True
___ Sequence Minimum	100	100	10	10	10	10	10
Advanced Assembly Options							
Match Score	10	10	10	10	10	10	10
Mismatch Penalty	20	20	20	20	20	20	20

Gap Penalty	30	30	30	30	30	30	30
Max Gap	6	6	15	15	20	15	15
SNP Passes	2	2	2	2	2	2	2
SNP Match Percent	90	90	90	90	90	90	90
SNP Low Cover Cutoff	0	0	0	0	0	0	0
Match Window	50	50	50	50	50	50	50
Maximum Coverage	0	0	0	0	0	0	0
Match Repeat Percent	150	150	150	150	150	150	150
Match Spacing	10	50	75	20	150	150	10
Default Quality	15	15	15	15	15	15	15
Default Template Quality	500	500	500	500	500	500	500
Max Usable	25	25	25	25	25	25	25

SeqMan NGen Scripting Manual

Note: Because this section is also provided as a separate text-only document for Linux users, it uses minimal formatting compared to the rest of the SeqMan NGen help.

SeqMan NGen Assemblers

SeqMan NGen contains two assemblers, XNG and SNG (called SMNG in Linux), with different capabilities and scripting languages. Therefore, it is essential to match the correct assembler with the type of assembly project to be done.

Reference-guided ("templated") assemblies:

The XNG assembler is used for nearly all reference-guided assemblies. This assembler is capable of assembling data sets of any size, given sufficient disk resources and modest RAM requirements (see http://www.dnastar.com/t-sub-support-technical-reqs-seqman-ngen.aspx for

details). The primary output is a BAM-formatted alignment file for each reference sequence. Note that BAM files cannot be edited.

For small genome (less than 30MB) reconstruction projects with fewer than 10 million reads, where editing is required, templated assemblies can also be performed using SNG/SMNG. The SNG/SMNG assembler generates finished assemblies in any of four formats: SQD, ACE, SAM or BAM. SeqMan (SQD) and ACE files are editable in SeqMan Pro, but the number of data reads is limited to 10 million or fewer. BAM files of any size can be created, but may not be edited.

De novo assemblies:

The SNG/SMNG assembler is used for all de novo assemblies. SNG/SMNG generates finished assemblies in SeqMan (SQD) or ACE format. Both are editable in SeqMan Pro, but the number of data reads is limited to 10 million or fewer.

Specifying XNG or SNG/SMNG When Running a Script

To specify which assembler to use to run your script, type xng or sng (smng in Linux) followed by the path and script file name after the command prompt. Alternatively, add either the #!/usr/bin/xng or #!/usr/bin/sng (#!/usr/bin/smng in Linux) command as first line of the script and execute through the command line.

Scripting Manual Conventions

Due to the constraints of TXT format, the following formatting conventions apply to Parts III and IV of this scripting manual:

"Commands" are listed alphabetically, and are denoted in the list by alphabetical characters (e.g., "A" or "CC"). One command is separated from the next using a long line of asterisks.

"Parameters" for a command are listed here in alphabetical order, not the order in which they are written in a script. Parameters are denoted in the list by the same letter(s) as the command, plus a number (e.g., "A35"). For the optimal organization and usage of parameters in a script, please refer to the Example sections.

"Properties" for a parameter are listed in alphabetical order, and are denoted in the list by the same letter(s) and number as the parameter, plus a lower-case letter (e.g., "A35c"). Properties are indented slightly, and are also bracketed between short lines of asterisks.

"Examples appear below the Commands/Parameters/Properties that they are intended to illustrate.

****

XNG Commands

A) assembleTemplate

(required) Initiates the assembly of the loaded sequences using the specified template as a reference.

Parameters for 'assembleTemplate':

A1) assemble: [matchContam|noMatchContam|all]

(optional) Specifies whether to use the part of the query that matches the contaminant sequence(s), the part that doesn't match, or both. Default is 'noMatchContam.'

A2) autoTrim: [true|false]

(optional) Specifies whether mismatching ends of reads should automatically be trimmed. Default is 'true.'

A3) boneyardAssembly: [true|false]

(optional) Specifies whether sequences not used in the original or incremental XNG assemblies should be added to the assembly project by the SNG assembler. This command pertains only to reference-guided assemblies with gap closure. By default, during this type of assembly, the XNG assembler first finds structural variations (SVs) then splits the contig after each SV. Elements of this process can be modified using this command. The default is 'true.'

A4) contaminant: [directory/filename enclosed in quotes]

Use of this parameter partitions the query data by running an additional mer-match (layout) against the specified contaminant sequence(s). A full assembly is then run using the part of the query that either matches or does not match the contaminant sequence(s). This parameter can be used for removing reads originating from an organism(s) that may have also been present in the query data set (e.g., reads from human DNA present in a metagenomic sample from the human gut).

A5) dbSNPTable: [directory/filename enclosed in quotes]

(Intended for internal use only).

A6) deleteIntermediates: [true|false]

(optional) Specifies whether intermediate files are saved or deleted. These files can be large with large scale projects. Default is 'false.'

A7) directoryMer: [directory/filename enclosed in quotes]

(optional) Specifies the path and directory where both the template and query data mer files will be stored. Alternatively, separate directories for the template and query mer files can be specified using the parameters below. If no directory is specified, the mer file will be created in the directory containing the sequence data.

A8) directoryQueryMer: [directory/filename enclosed in quotes]

(required) Specifies the path and directory where the query mer file will be stored.

A9) directoryTemplateMer: [directory/filename enclosed in quotes]

(required) Specifies the path and directory where the template mer file will be stored.

A10) filterDeepLayout: [true|false]

(optional) Specifies that XNG remove superfluous sequences in areas of deep coverage. Default is 'true.'

A11) forceMake: [true|false]

(optional) Specifies whether new intermediate mer files will be created. A value of false means that existing valid intermediate files will be used. Default is 'true.'

A12) format: [BAM|SQD|none]

(optional) Specifies the format of the alignment output file. If 'none' is entered, the assembly is run to include the alignment phase, but no alignment output is generated. This parameter can be used to remove reads from a contaminant source. Default is 'BAM.'

A13) gapPenalty: [number]

The penalty for opening or extending a gap during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. A high gap penalty suppresses gapping, while a low value promotes gapping. Default is 30.

A14) hits: [directory/filename enclosed in quotes]

(required) Specifies the path and name of the hit file. Incomplete paths will be appended to the default directory.

A15) layout: [directory/filename enclosed in quotes]

(required) Specifies the path and name of the layout file. Incomplete paths will be appended to the default directory.

A16) layout type: [unique|once|multiple]

Specifies how reads are to be laid out. Default is 'once.'

A17) matchScore: [number]

The score for a base match during an alignment. This score contributes to the pairwise score used to calculate match percentage. Increasing the matchScore value allows for longer or more frequent gaps, thus forcing bases that match to be assembled together. Default is 10.

A18) MaxGap: [number from 0-1000]

The maximum number of gaps allowed per 1000 bases in the alignment. Default is 6.

A19) maxSeqs: [number]

(optional) Specifies the maximum number of query sequences to add to an assembly. Use of this command can speed up assembly. This parameter does not have a default value.

A20) merLayoutMin: [number from 11-1000]

(optional) Specifies the minimum length (in bases) of at least one stretch of matching mers used to identify matches between the reference and query data. The minimum value is equal to the mer. The maximum value is the read length, which would require the entire read be an exact match. For example, with a merSize of 19 and a merLayoutMin of 21, at least one stretch of three consecutive mers in a read would have to match for the read in order to be included in the layout. Default is 25.

A21) merMinimizer: [number]

(Intended for internal use only)

A22) merSize: [number]

(required) Specifies the length (in bases) of mers used to identify matches between the reference and query data. This parameter does not have a default value.

A23) merSkip: [number]

(Intended for internal use only) Specifies the number of positions to ignore or "skip" when creating the template mer file. Normally, mers are only skipped in the query (see 'merSkipQuery,' below). The first and last mer of every read are always included. Increasing the value reduces the size of the intermediate files as well as the overall assembly time. However, larger values can also reduce the number of reads included in the assembly, especially with short read data. Default is 0.

0 = do not skip

2 = skip every second base

3 = skip every third base

etc.

A24) merSkipQuery: [number]

(optional) Specifies the number of positions to ignore or "skip" when creating the query mer file. The first and last mer of every read are always included. Increasing the value reduces the size of the intermediate files as well as the overall assembly time. However, larger values can also reduce the number of reads included in the assembly, especially with short read data. Default is 0.

0 = do not skip

2 = skip every second base

3 = skip every third base

etc.

A25) minAlignedLength: [number from 11-1000]

(optional) Specifies the minimum number of bases that must align after trimming for a read to be included in the assembly. Default is 25.

A26) minMatchPercent: [number]

The minimum percentage of matches in an overlap required to join two sequences in the same contig. Default is 93.

A27) mismatchPenalty: [number]

The penalty for a base mismatch during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. Default is 20.

A28) noSVPairSort: [true|false]

Specifies whether to turn off the calculation of pairs for structural variations. This may potentially reduce XNG assembly time. Default is 'false.'

A29) onePackage: [true|false]

(optional) Specifies whether an assembly containing multiple reference sequences should be bundled into a single .assembly package. If 'false' is entered, one .assembly package is created per contig. Default is 'true.'

A30) openInSeqman: [true|false]

(optional; not available for Linux users) Specifies whether the completed assembly should immediately be launched in SeqMan. Default is 'false.'

A31) output: [directory/filename enclosed in quotes]

(required) Specifies the path and directory of the output files. Incomplete paths are appended to the default directory.

A32) pairDist: [true|false]

(Intended for internal use only)

A33) placeHit: [true|false]

(Intended for internal use only)

A34) probe: [number]

(Intended for internal use only)

A35) query: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name(s) of the query data to be assembled. A folder with one or data files can also be used in place of individual file names.

****

Properties for 'query':

A35a) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder.

A35b) isPair: [true|false]

(optional) Specifies whether the query files contain paired end data. Default is 'false.'

A35c) minDist: [number]

(required if 'isPair' is 'true') Specifies the minimum expected distance in bases between paired end reads. Default is 0.

A35d) maxDist: [number]

(required if 'isPair' is 'true') Specifies the maximum expected distance in bases between paired end reads. Defaults are 500 for Illumina < 50nt; 3000 for Illumina > 50 nt; and 5000 for all others.

A35e) seqTech: [normalScore|IonTorrent|SOLiD|Illumina|454|unknown]

(optional) Specifies the offset to be used when converting compressed quality scores into numerical values. These are the offsets used for the technology specified:

normalScore 33

IonTorrent 33

SOLiD 33

Illumina 64

454 33; quality scores for homopolymeric runs of >= 2 are oriented from 5' to 3' on the top strand.

unknown determined automatically based on the first data file.

Example for 'query':

query: {{file: "/data/home/proj/Illumina_s_5_1.txt"}

{file: "/data/home/proj/Illumina_s_5_2.txt "}

isPair: true

minDist: 400

maxDist: 700

seqTech: Illumina}

****

A36) recordSplitsOnly: [true|false]

(optional) Functional only when used in the same program as 'splitTemplateContigs' or 'recordStructVariations.' Specifies whether or not to turn off contig splitting while still recording SVs for later inclusion in the Structural Variation Report. The default is 'false.'

A37) recordStructVariations: [integer between 0-3|true|false]

(optional) Specifies under which circumstances structural variations (SVs) should be calculated and recorded. Default is 2.

0|false Don't calculate SVs

1|true Calculate SVs at zero coverage

2 Calculate SVs at insertions and deletions

3 Calculate SVs at zero coverage and at insertions

A38) repeatCnt: [number from 1-10000]

(optional) Specifies the minimum number of occurrences of a mer in the reference sequence(s) for it to be considered repeated. Mers exceeding this number will not be used for identifying matches. The default is 100.

A39) results: [directory/filename enclosed in quotes]

(optional) Specifies the path and name of the result summary file. This file contains a compilation of assembly statistics and uses the extension fileSize.txt. Incomplete paths will be appended to the default directory.

A40) saveUnSplitAssembly: [true|false]

(optional) Specifies whether XNG should save both the normal assembly output, [filename].assembly, and the unsplit intermediate assembly, [filename]-noSplit.assembly. The latter file contains SVs but no SNPs, and can be used to validate splits in the final assembly. The default is 'false.'

A41) showCDSVariant: [true|false]

(optional) Specifies whether or not XNG should show all variants of a CDS feature contacted by a SNP. The version number for the CDS variant will then appear in brackets when viewed in the SNP report in SeqMan Pro. Default is 'true.'

A42) sngConvertOptions: [text string]

(Intended for internal use only)

A43) snp: [true|false]

(optional) Specifies whether or not a SNP detection pass of the gapped alignment should be made during the assembly. Default is 'true.'

A44) snp_checkStrandedness: [true|false]

(optional) Specifies whether or not the strand that each read comes from is considered in the SNP calculation. This is ignored by the Simple method. Default is 'false.'

A45) snp_limitEndPos: [number]

(optional) Specifies the 3' most coordinate of the specified template from which to stop calculating SNPs. A value between 1 and the length of the template must be entered.

A46) snp_limitStartPos: [number]

(optional) Specifies the 5' most coordinate of the specified template from which to begin calculating SNPs. A value between 1 and the length of the template must be entered. Default is 1.

A47) snp_limitTemplateID: [number]

(optional) Specifies a single template ID for which to calculate SNPs. By default, counting begins from 0.

A48) snp_logEndPos: [number]

(optional) Specifies the 3' most coordinate of the specified template from which to stop storing a detailed log of SNP information. A value between 1 and the length of the template must be entered. Default is 1.

A49) snp_logLevel: [whole number from 0-3]

(optional) Specifies the level of detailed logging to store in the "shared" project directory as "SNP.log." Level 0 specifies that no log will be stored. Level 1 stores detailed info on the SNPs which were called, level 2 also logs columns where the preliminary filtered passed but the final filtering failed, and level 3 logs all columns. This is ignored by the simple SNP calling method. Default is 0.

A50) snp_logStartPos: [number]

(optional) Specifies the 5' most coordinate of the specified template from which to begin storing a detailed log of SNP information. A value between 1 and the length of the template must be entered. Default is 1.

A51) snp_logTemplateID: [number]

(optional) Specifies a single template from which to store a detailed log of SNP information. By default, counting begins from 0.

A52) snp_minPctToScore: [number from 0-1]

(optional) Specifies minimum percentage of reads in a column which must differ from the reference in order to score the column. For the Simple method, this is the only criteria used to call a SNP. For the Diploid and Haploid methods, this is a filter applied before the other parameters. Default is 0.05.

A53) snp_minProbNonrefToCall: [number from 0-1]

(optional) Specifies the minimum probability of a SNP column which is required to call a SNP, expressed as a number from 0 and 1. The probabilities of all genotypes other than Homozygous Reference are totaled and checked against this number. This is the final filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 0.1, requiring a minimum 10% change.

A54) snp_minVariantDepthToScore: [number from 0-100]

(required if "snp" is true) Specifies the minimum depth required for a specific base (or deletion) in a column before it is considered usable for SNP calling. This is the second filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 2.

A55) snp_minWeight: [number]

(optional) Specifies the minimum quality score for a base to be considered in the SNP calculation. Default 5.

A56) snp_showAllFeatures: [true|false]

(optional) Specifies whether XNG should count SNPs multiple times if the SNP contacts different versions (variants) of a CDS feature. Default is 'true.'

A57) snp_writeExtended: [true|false]

(optional) Specifies whether the additional values produced by the Haploid or Diploid SNP calculation methods are included in the SNP table. Default is 'true.'

A58) snpMethod: [simple|haploid|diploid|population]

(optional) Specifies the SNP detection method to use. Simple produces a count of each type of base in the column and calculates the percent of non-reference bases. Haploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base called at that position. Diploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base(s) called at that position. Based on the scores, it also calls the genotype at each position. Default is 'diploid.'

A59) splitTemplateContigs: [integer between 0-3|true|false]

(optional) Specifies under which circumstances contigs should be cut after a templated assembly. Any split contigs will be grouped into scaffolds with a defined position to allow for easy sorting when the project is viewed in SeqMan Pro. This command pertains only to reference-guided assemblies with gap closure. By default, during this type of assembly, the XNG assembler first finds structural variations (SVs) then splits the contig after each SV. Elements of this process can be modified using this command. Default is 2.

0|false Don't split

1|true Split at locations with zero coverage

2 Split at insertions and deletions

3 Split at zero coverage and at insertions

****

A60) template: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the template file. A folder with one or more template files can also be used in place of individual file names. Each entry must also be enclosed by brackets. If more than template entry is used, the list must also be enclosed by an additional set of brackets.

****

Properties for 'template':

A60a) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder.

A60b) feature: [directory/filename enclosed in quotes]

optional) Specifies the directory and file name for annotated features when the reference sequence and feature annotations are in separate files.

Examples for 'template':

Sequence and annotation in one file:

```
AssembleTemplate
 template: {{file: "/data/home/proj/MG1655.gbk"}
           {file: "/data/home/proj/W3110.gbk"}}
```

Sequence and annotation in separate files:

```
AssembleTemplate
 template: {file: "/Library/ABC_proj/references/MG1655.fas"
           feature: "/Library/ABC_proj/references/MG1655.gff"}
```

****

A61) templateHitCntThresh: [number]

(Intended for internal use only)

A62) unassembled: [directory/filename enclosed in quotes]

A63) verify: [true|false]

****

B) computeSNP

(optional) Sets parameters for the SNP computation phase of the assembly. The command is designed for use with existing BAM files that have not been analyzed for SNPs, or to re-analyze an existing file with different parameters. The associated parameters are also available for full assemblies under the 'assembleTemplate' command.

Parameters for 'computeSNP':

B1) calcJunctionSeqs: [true|false]

(optional) In the structural variation workflow, specifying 'false' prevents junction sequences from being calculated. Default is 'true.'

B2) concurrentAligns: [number]

(Intended for internal use only)

B3) file: [directory/filename enclosed in quotes]

(required) Specifies the path and name of one or more .assembly projects from which to compute SNPs.

B4) showCDSVariant: [true|false]

(optional) Specifies whether XNG should show all variants of a CDS feature contacted by a SNP. The version number for the CDS variant will then appear in brackets when viewed in the SNP report in SeqMan Pro. Default is 'true.'

B5) snp_checkStrandedness: [true|false]

(optional) Specifies whether the strand that each read comes from is considered in the SNP calculation. This is ignored by the Simple method. Default is 'false.'

B6) snp_limitEndPos: [number]

(optional) Specifies the 3' most coordinate of the specified template from which to stop calculating SNPs. A value between 1 and the length of the template must be entered. Default is 1.

B7) snp_limitStartPos: [number]

(optional) Specifies the 5' most coordinate of the specified template from which to begin calculating SNPs. A value between 1 and the length of the template must be entered. Default is 1.

B8) snp_limitTemplateID: [number]

(optional) Specifies a single template ID for which to calculate SNPs. By default, counting begins from 0.

B9) snp_logEndPos: [number]

(optional) Specifies the 3' most coordinate of the specified template from which to stop storing a detailed log of SNP information. A value between 1 and the length of the template must be entered. Default is 1.

B10) snp_logLevel: [number]

(optional) Specifies the level of detailed logging to store in the "shared" project directory as "SNP.log". Level 0 specifies that no log will be stored. Level 1 stores detailed info on the SNPs which were called, level 2 also logs columns where the preliminary filtered passed but the final filtering failed, and level 3 logs all columns. This is ignored by the simple SNP calling method. Default is 0.

B11) snp_logStartPos: [number]

(optional) Specifies the 5' most coordinate of the specified template from which to begin storing a detailed log of SNP information. A value between 1 and the length of the template must be entered. Default is 1.

B12) snp_logTemplateID: [number]

(optional) Specifies a single template from which to store a detailed log of SNP information. By default, counting begins from 0.

B13) snp_minPctToScore: [number from 0-1]

(optional) Specifies minimum percentage of reads in a column which must differ from the reference in order to score the column. For the Simple method, this is the only criteria used to call a SNP. For the Diploid and Haploid methods, this is a filter applied before the other parameters. Default is 0.05.

B14) snp_minProbNonrefToCall: [number from 0-1]

Specifies the minimum probability of a SNP column which is required to call a SNP. The probabilities of all genotypes other than Homozygous Reference are totaled and checked against this number. This is the final filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 0.1, requiring a minimum 10% change.

B15) snp_minVariantDepthToScore: [number from 0-100]

Specifies the minimum depth required for a specific base (or deletion) in a column before it is considered usable for SNP calling. This is the second filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 2.

B16) snp_minWeight: [number]

(optional) Specifies the minimum quality score for a base to be considered in the SNP calculation. Default is 5.

B17) snp_showAllFeatures: [true|false]

(optional) Specifies whether XNG should count SNPs multiple times if the SNP contacts different versions (variants) of a CDS feature. Default is 'true.'

B18) snp_writeExtended: [true|false]

(optional) Specifies whether the additional values produced by the Haploid or Diploid SNP calculation methods are included in the SNP table. Default is 'true.'

B19) snp_writeMissingDBSnps: [true|false]

(optional) In a SNP assembly, specifying 'false' causes missing SNPs not to be recorded, saving time and file space. Default is 'true.'

B20) snpMethod: [simple|haploid|diploid|population]

(optional) Specifies the SNP detection method to use. Simple produces a count of each type of base in the column and calculates the percent of non-reference bases. Haploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base called at that position. Diploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base(s) called at that position. Based on the scores, it also calls the genotype at each position. Default is 'diploid.'

B21) userSNP: [directory/filename enclosed in quotes]

(optional) Specifies a location for storing the VCF SNP table.

****

C) createGenomeTemplate

(Intended for internal use only)

Parameters for 'createGenomeTemplate'

C1) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder of the input file.

C2) output: [directory/filename enclosed in quotes]

The path and name of the output file.

****

D) diskPath

(required) Defines the default directory where temporary intermediate files from the assembly will be stored. The files can be large with large scale projects. Visit http://www.dnastar.com/t-sub-support-technical-reqs-seqman-ngen.aspx to view space requirements for a range of representative projects.

Parameters for 'diskPath':

D1) clean: [true|false]

Specifies whether or not to clean the merge disk. When automated scripts are being run simultaneously or sequentially, this command can be useful for emptying the merge disk between assemblies. Default is 'false.'

D2) pathMac: [directory/filename enclosed in quotes]

Specifies the default path and file name for Macintosh.

D3) pathWin: [directory/filename enclosed in quotes]

Specifies the default path and file name for Windows.

D4) path: [directory/filename enclosed in quotes]

(required) Specifies the default path and file name.

Example for 'diskPath':

```
  diskPath
   path: "/data/proj/"
```

****

E) dumpConsensus

(Intended for internal use only). To convert the binary consensus file created during assembly into a text file.

Parameters for 'dumpConsensus':

E1) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder.

****

F) dumpSNP

(Intended for internal use only). Creates a tab delimited text file from one or more SNP containing binary files generated during assembly. SNP binary files include those with the .snpExt suffix contained in an .assembly package as well as those with either the .coverage.missingSNP or .nocoverage.missingSNP suffix contained in the _shared folder. To convert all the .snpExt files in a package simply use the .assembly name.

Parameters for 'dumpSNP':

F1) file: [directory/filename enclosed in quotes]

(required) Specifies the path and name of .assembly package (all SNP files will be included), one or more individual .snpExt files or either/both of the missingSNP files.

F2) output: [directory/filename enclosed in quotes]

(required) Specifies the path and name of the output file.

F3) refPos_end: [number]

F4) refPos_start: [number]

F5) snp_maxProbNonrefToCall: [number]

F6) snp_minProbNonrefToCall: [number]

F7) snp_type: [simple|SNP|missing|user]

****

G) dumpSplits

(Intended for internal use only). To convert the binary splits file created during assembly into a text file.

Parameters for 'dumpSplits':

G1) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder.

G2) output: [directory/filename enclosed in quotes]

The path and name of the output file.

****

H) execute

(optional) Executes any shell script command.

Parameters for 'execute':

H1) command: [text string]

Text for any shell script command.

****

I) extractPairs

(optional) Creates a tab delimited table of pair end information.

Parameters for 'extractPairs':

I1) file: [directory/filename enclosed in quotes]

The path and name of any pair distance file (.pairdist file) from within a project's shared folder.

I2) output: [directory/filename enclosed in quotes]

The path and name of the output file.

****

J) include

(optional) When building a script, this command can be used to call up additional lines of script previously stored in a text file. In this way, a group of commands can be shared between two or more scripts.

Parameters for 'include':

J1) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder.

****

K) loadAssembly

(Intended for internal use only)

Parameters for 'loadAssembly':

K1) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder.

****

L) loadBAM

(optional) Sets parameters for analyzing existing BAM files. It allows ungapped BAM files to be converted into a fully gapped assembly file or to re-gap an existing file with different parameters. The command also permits SNPs to be calculated or re-calculated with different parameters starting with an existing BAM file. The associated parameters are also available for full assemblies under the 'assembleTemplate' command.

Parameters for 'loadBAM':

L1) align: [true|false]

(optional) Specifies whether a gapped alignment will be done. Default is 'false.'

L2) format: [BAM|SQD|none]

(optional) Specifies the format of the alignment output file. If 'none' is entered, the assembly will be run including the alignment phase, but no alignment output is generated. This can be used to remove reads from a contaminant source. Default is 'BAM.'

L3) gapPenalty: [number]

The penalty for opening or extending a gap during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. A high gap penalty suppresses gapping, while a low value promotes gapping. Default is 30.

L4) layout: [directory/filename enclosed in quotes]

(required) Specifies the path and name of the BAM file.

L5) matchScore: [number]

The score for a base match during an alignment. This score contributes to the pairwise score used to calculate match percentage. Increasing the matchScore value will allow for longer or more frequent gaps, thus forcing bases that match to be assembled together. Default is 10.

L6) minAlignedLength: [number]

The minimum length of aligned sequence that must be attained between the read and reference for the read to be included in the assembly. Default is 25.

L7) minMatchPercent: [number]

The minimum percentage of matches in an overlap required to join two sequences in the same contig. Default is 93.

L8) mismatchPenalty: [number]

The penalty for a base mismatch during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. Default is 20.

L9) output: [directory/filename enclosed in quotes]

(required) Specifies the path and directory of the output files.

L10) snp: [true|false]

(optional) Specifies whether a SNP detection pass of the gapped alignment is made during the assembly. Default is 'false.'

L11) snp_checkStrandedness: [true|false]

(optional) Specifies whether the strand that each read comes from is considered in the SNP calculation. This is ignored by the Simple method. Default is 'false.'

L12) snp_limitEndPos: [number]

(optional) Specifies the 3' most coordinate of the specified template from which to stop calculating SNPs. A value between 1 and the length of the template must be entered. Default is 1.

L13) snp_limitStartPos: [number]

(optional) Specifies the 5' most coordinate of the specified template from which to begin calculating SNPs. A value between 1 and the length of the template must be entered. Default is 1.

L14) snp_limitTemplateID: [number]

(optional) Specifies a single template ID for which to calculate SNPs. By default, counting begins from 0.

L15) snp_logEndPos: [number]

(optional) Specifies the 3' most coordinate of the specified template from which to stop storing a detailed log of SNP information. A value between 1 and the length of the template must be entered. Default is 1.

L16) snp_logLevel: [number]

(optional) Specifies the level of detailed logging to store in the "shared" project directory as "SNP.log". Level 0 specifies that no log will be stored. Level 1 stores detailed info on the SNPs which were called, level 2 also logs columns where the preliminary filtered passed but the final filtering failed, and level 3 logs all columns. This is ignored by the simple SNP calling method. Default is 0.

L17) snp_logStartPos: [number]

(optional) Specifies the 5' most coordinate of the specified template from which to begin storing a detailed log of SNP information. A value between 1 and the length of the template must be entered. Default is 1.

L18) snp_logTemplateID: [number]

(optional) Specifies a single template from which to store a detailed log of SNP information. By default, counting begins at 0.

L19) snp_minPctToScore: [number from 0-1]

(optional) Specifies minimum percentage of reads in a column which must differ from the reference in order to score the column. For the Simple method, this is the only criteria used to call a SNP. For the Diploid and Haploid methods, this is a filter applied before the other parameters. Default is 0.05.

L20) snp_minProbNonrefToCall: [number from 0-1]

(optional) Specifies the minimum probability of a SNP column which is required to call a SNP. The probabilities of all genotypes other than Homozygous Reference are totaled and checked against this number. This is the final filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 0.1, requiring a minimum 10% change.

L21) snp_minVariantDepthToScore: [number]

Specifies the minimum depth required for a specific base (or deletion) in a column before it is considered usable for SNP calling. This is the second filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 2.

L22) snp_minWeight: [number]

(optional) Specifies the minimum quality score for a base to be considered in the SNP calculation. Default is 5.

L23) snp_writeExtended: [true|false]

(optional) Specifies whether the additional values produced by the Haploid or Diploid SNP calculation methods are included in the SNP table. Default is 'true.'

L24) snpMethod: [simple|haploid|diploid|population]

(optional) Specifies the SNP detection method to use. Simple produces a count of each type of base in the column and calculates the percent of non-reference bases. Haploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base called at that position. Diploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base(s) called at that position. Based on the scores, it also calls the genotype at each position. Default is 'diploid.'

L25) template: [directory/filename enclosed in quotes]

(required) Specifies the path and name of the reference sequence file(s).

****

M) mergeIonTorrentShortReads

(optional) When using Ion Torrent data, use of this command merges overlapping short reads into mini-contigs.

Parameters for 'mergIonTorrentShortReads':

M1) output: [directory/filename enclosed in quotes]

(required) Specifies the path and directory of the output files.

M2) query: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name(s) of the query data to be assembled. A folder with one or data files can also be used in place of individual file names.

****

N) pairFilePattern

Allows you to specify the pattern for pair files using the GREP language.

Parameters for 'pairFilePattern':

N1) forward: [text string enclosed in quotes]

A naming pattern to match forward clones.

N2) reverse: [text string enclosed in quotes]

A naming pattern to match reverse clones.

Example for 'pairFilePattern':

```
  pairFilePattern
    forward: "(?'name'.*)_R1_(?'ext'.*)\fastq
    reverse: "(?'name'.*)_R2_(?'ext'.*)\fastq
```

****

O) pause

(optional) Creates a pause and can be used when running table scripts to stop at any point.

Parameters for 'pause':

O1) prompt : [text string enclosed in quotes]

Text to appear in the console. The pause is terminated by hitting the Enter key.

Example for 'pause':

```
 pause
  prompt: "Table script paused. Press enter to continue."
```

****

P) quit

(optional) Terminates a script.

****

Q) runScript

(optional) Allows batching of multiple projects of the same type (e.g. assembly, computeSNPs). There are required three file: 1) a runScript file with variables, 2) a file with a table of values for the variables, and 3) a script file specifying the action to be carried out.

Parameters for 'runScript':

Q1) script: [directory/filename enclosed in quotes]

The filename and location of the script.

Q2) table: [directory/filename enclosed in quotes]

The filename and location of the file containing text strings and numbers values for each variable.

Example for 'runScript' (runScript file):

```
 setDefaultDirectory directory: "."

  set $force: false

  set $DataDisk: "/Volumes/Raid/DataDisk"

  set $ResultDisk: "/Volumes/ResultDisk"

  set $MergeDisk: "/Volumes/MergeDisk0"

  set $snp:true

  set $snpMethod:"Diploid"

  set $repCnt:100

  set $merLayoutMin:19

  diskPath path: {"${MergeDisk}/mergeSort Data"}}

  runScript table: "testAssembly.txt" script: "testAssembly.template.script"
```

Example for 'runScript' (table file):

```
defaultDir      template        query   isPair  seqTech         project merSize         snp
        snpMethod

"${ResultDisk}/rice"  ${DataDisk}/rice.genome     ${DataDisk}/rice       FALSE
        Illumina        rice    21      TRUE Diploid

"${ResultDisk}/ecoli"${DataDisk}/Ecoli.gbk        ${DataDisk}/ecoli     TRUE Illumina
        Ecoli   21      TRUE Diploid

"${ResultDisk}/Exome"       ${DataDisk}/GRCh37.gbk    ${DataDisk}/Sample1
        FALSE           454     HuEx    19      TRUE Diploid
```

Example for 'runScript' (script file):

```
 ; "assembly.template.script"

  setMachineMemory memory:32
```

```
  setDefaultDirectory directory:        $defaultDir
  compareSeqs template:       $template
  query:          {file: $query
  isPair: $isPair
  seqTech: $seqTech}
  directoryMer:          "intermediateFiles"
 ; directoryQueryMer: "intermediateFiles"
  hits: "intermediateFiles/${project}.hits"
  layout:         "intermediateFiles/${project}.layout"
  output:         "results_${mersize}_${merSkipQuery}/${project}"
 ; results per project
 ; results: "${project}.results.txt"
 ; aggregate all results
  results: "${ResultDisk}/assembly.results.txt"
  merSize: $mersize
  merSkipQuery: $merSkipQuery
  repeatCnt: $repCnt
  merLayoutMin: $merLayoutMin
  layoutType: once
  maxGap: 6
  format: BAM
  onePackage: true
  snp: $snp
  snpMethod: $snpMethod
 ; snp_writeExtended: true
  forceMake: $force
**************************************************************
```

R) set

(optional) Used to set variables. See the example below and those under the 'runScript' command.

Example for 'set':

```
set $snp:true
set $snpMethod:"Diploid"
```

****

S) setDefaultDirectory

(required) Defines the default directory for the project. When a default directory is specified, files located in that directory only need to be identified by their subfolder and/or file name in subsequent commands.

Parameters for 'setDefaultDirectory':

S1) directory: [directory/filename enclosed in quotes]

(required) Specifies the default directory. Previously called 'defaultDirectory.'

S2) directoryMac: [directory/filename enclosed in quotes]

Specifies the default directory for Macintosh. Previously called 'defaultMacDirectory.'

S3) directoryWin: [directory/filename enclosed in quotes]

Specifies the default directory for Windows. Previously called 'defaultWinDirectory.'

Example for 'setDefaultDirectory':

```
setDefaultDirectory
 directory: "/data/home/proj/"
```

****

T) setMachineMemory

(optional) Defines the amount of random access memory (RAM) that the program will use. Limiting the amount of RAM available to the assembler allows you to use the computer for other purposes while an assembly is running. However, this will likely slow down the assemblies and is not recommended for large projects.

Parameters for 'setMachineMemory':

T1) memory: [number that is a multiple of 4]

(required) Amount of RAM (in GB) to be used, entered in multiples of four. Entering a value greater than the available RAM causes all RAM to be used. There is no default value.

Example for 'setMachineMemory':

```
 setMachineMemory

  memory: 32
```

****

U) setParam

Allows you to adjust the stringency of one or more of the assembling parameters for the project. SeqMan NGen will use the default values for any parameter that is not specified within the script.

Parameters for 'setParam':

U1) gapPenalty: [number]

The penalty for opening or extending a gap during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. A high gap penalty suppresses gapping, while a low value promotes gapping. Default is 30.

U2) matchScore: [number]

The score for a base match during an alignment. This score contributes to the pairwise score used to calculate match percentage. Increasing the matchScore value will allow for longer or more frequent gaps, thus forcing bases that match to be assembled together. Default is 10.

U3) minAlignedLength: [number]

The minimum length of aligned sequence that must be attained between the read and reference for the read to be included in the assembly. Default is 25.

U4) minMatchPercent: [number]

The minimum percentage of matches in an overlap required to join two sequences in the same contig. Default is 93.

U5) mismatchPenalty: [number]

The penalty for a base mismatch during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. Default is 20.

V) trimToTargetRegions: [true|false]

Controls whether reads are trimmed, by default, to the boundaries of the targeted regions, as defined by the .bed or manifest file. Default is true, meaning that the reads are trimmed to the

stated boundaries. Selecting “true” is equivalent to checking the “Trim to targeted regions” box in the Alignment tab of the SeqMan NGen wizard’s Advanced Options dialog.

****

SNG Commands

Note: To see how SNG commands and parameters map to equivalent SeqMan NGen wizard settings, open the appendix of the SeqMan NGen help (http://www.dnastar.com/t-help-seqman-ngen.aspx) and select the topic "Equivalence Between Wizard Settings and SNG Scripting Commands."

Part I. Project Management Commands

A) closeProject

(optional) Closes the current project and frees the memory in use so that the system is ready for additional assemblies. This can be useful if you want to run multiple assemblies in one script.

****

B) runScript

(optional) Allows you to run a table script within the current script. A table script references variable values for specified parameters and other elements in a script. This enables you to run multiple projects from the same script, substituting new parameter values and other variables each time. SeqMan NGen will run the table script repeatedly, using the variable values from one row of the table for each iteration of the script until all of the rows have been used. For more information, see the Using Table Scripts in SeqMan NGen section.

Parameters for 'runScript':

B1) file: [directory/filename enclosed in quotes]

Specifies the directory and file/folder.

B2) script: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the table script you wish to run.

B3) table: [directory/filename enclosed in quotes]

(required) Specifies the delimited text file containing the variable values.

Example for 'runScript':

```
runScript
```

script: "/Library/abc_Project/abc_script.script"

table: "/Library/abc_Project/table.txt"

****

C) saveProject

This command saves the assembly to a project file. By default, the SeqMan Pro project file format (*.sqd) is used. Phrap (*.ace) and FASTA (*.fas) formats may also be specified by using the format parameter, and specifying the desired file extension using the file parameter.

Note: As a command-line tool, SeqMan NGen will not prompt you if you try to save a new project file with the same name as an existing file in the same location. When you run a script multiple times, be sure to change the file name of the project to be saved each time to prevent existing project files from being overwritten.

Parameters for 'saveProject':

C1) file: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the project file to be saved.

C2) format [SeqMan|SeqMan8|SeqMan7|Phrap|Fasta|BAM|SAM]

(optional) Specifies the output file format. Default is 'SeqMan.'

SeqMan Saves a 64-bit SeqMan Pro project file (*.sqd) that is compatible with SeqMan Pro version 8.1 and higher (default).

SeqMan8 Saves a 32-bit SeqMan Pro project file (*.sqd) that is compatible with SeqMan Pro version 8.0 and higher.

SeqMan7 Saves a 32-bit SeqMan Pro project file (*.sqd) that is compatible with SeqMan Pro version 7.2 and higher. Note that this project file will be much bigger than the same project created in either of the SeqMan formats listed above.

Phrap Saves an .ace file.

Fasta Saves .fas and .qual files of the consensus sequence for each contig.

BAM Saves a BAM file (SNG/SMNG templated assemblies only).

SAM Saves a SAM file (SNG/SMNG templated assemblies only).

C3) onePackage: [true|false]

(optional) Specifies whether an assembly containing multiple reference sequences should be bundled into a single .assembly package. If 'false' is entered, one .assembly package is created per contig. Default is 'true.'

C4) openInSeqMan: [true|false]

(not available for Linux users) Specifies whether to automatically launch SeqMan Pro and open the completed assembly once the script has completed. Default is 'true.'

Example for 'SaveProject':

```
SaveProject
 file: "/Library/My projects/ABC_project.sqd"
 format:seqman
 openInSeqMan:true
```

****

D) saveReport

(optional) Exports a report as a text file that summarizes assembly statistics, including the parameters used, the number of assembled/unassembled sequences and contigs, average quality scores, and the number of sequences excluded from the assembly due to exceeding the maxAssemblyCoverage parameter. The same information contained within this report is also saved within the SeqMan Pro project file (*.sqd) regardless of whether you choose to export the report by setting this parameter. The report can be viewed in SeqMan Pro using the Project>Report command.

Parameters for 'saveReport':

D1) file: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the report to be saved. .

Example for 'saveReport':

```
saveReport
 file: "/Library/abc_Project/abc_report.txt"
```

****

E) WriteUnassembledSeqs

(optional) Saves all sequences that were not assembled in the project as *.fas and *.qual files.

Parameters for 'WriteUnassembledSeqs':

E1) file: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the unassembled sequences to be saved.

E2) saveTrimmed: [true|false]

Specifies whether to save only the trimmed portion of the unassembled sequences. Default is 'false.'

Part II. File Loading Commands and Parameters

F) load454PairedEnd

Loads a file of Roche 454 sequences and checks for the presence of a linker defining the paired end sequences. If the linker is found, the linker is removed and the remaining portion is split into two sequences linked with a paired end constraint.

Parameters for 'load454PairedEnd':

F1) DiscardLinkerless: [true|false]

Specifies whether to discard any read where no portion of the mate pair linker was found. In this way, reads that do not have a linker sequence will be discarded from the assembly. Default is 'false.'

F2) file: [directory/filename enclosed in quotes]

The directory and file name of the .fas, .fna, or .sff file containing the 454 sequences.

F3) linker: [directory/filename enclosed in quotes]

The directory and file name of the .fas, fna, or .sff file containing the 454 linker sequences. If not specified, SeqMan NGen will use its default 454 linker sequence: GTTGGAACCGAAAGGGTTTGAATTCAAACCCTTTCGGTTCCAAC

F4) max: [number]

The maximum distance for the paired end constraint. Default is 10000. (Also called 'maxDistance').

F5) min: [number]

The minimum distance for the paired end constraint. Default is 0. (Also called 'minDistance').

Example for load454PairedEnd':

```
load454PairedEnd
  file: "/Library/454 data/123_Pairedend.fas"
  linker: "/Library/454 data/123_linkerseqs.fas"
  min: 0
  max: 10000
```

DiscardLinkerless: false

****

G) LoadConstraint

Loads a constraint file. The file can be in the NCBI ancillary file format, or in the CAP3 constraint file format. SeqMan NGen uses constraint files to identify paired end reads, similar to using the 'setPairSpecifier' command. Constraint files in the NCBI ancillary file format also contain trimming information, which SeqMan NGen will load and use. SeqMan NGen will create a CAP3 file when saving a Phrap project (*.ace) that used paired end constraints.

Parameters for 'LoadConstraint':

G1) file: [directory/filename enclosed in quotes]

The directory and file name of the constraint sequence file.

Example for 'LoadConstraint':

loadConstraint

file: "/Library/constraints/123_xyz.con"

****

H) LoadContaminant

Loads a contaminant sequence file to be used to identify known contaminants, such as primers, in the assembly. Sequences that contain at least 12 matching 17-mers are flagged as contaminant sequences and will be removed from the assembly. See our website (http://www.dnastar.com/t-smgafileformats.aspx) for a list of supported file types.

Parameters for 'LoadContaminant':

H1) file: [directory/filename enclosed in quotes]

The directory and file name of the contaminant sequence file. A folder may also be specified, in which case all of the sequence files within that folder will be loaded and used for contaminant screening.

Example for 'loadContaminant':

loadContaminant

file: "/Library/contaminants/123_abc.seq"

****

I) loadLayout

Loads a layout file to be used for an assembly. The format may be either a SOLiD General Feature Format file (*.gff) or a File of Filenames file (*.fof). When this command is used, SeqMan NGen still aligns each read from the file to the template, but uses the information contained within the specified file to determine the overall layout of reads.

Parameters for 'loadLayout':

I1) layoutFile: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the layout file. Both *.gff and *.fof formats are accepted.

I2) templateFile: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the template file.

Example for 'loadLayout':

```
loadLayout
 templateFile: "/Library/123_project/template.seq"
 layoutFile: "/Library/123_project/layoutfile.gff"
```

****

J) LoadRepeat

Loads a sequence file to be used to identify repeat sequences in the assembly. All sequences identified as repeats will be added to the assembly last, after all non-repeats have been assembled. See our website (http://www.dnastar.com/t-smgafileformats.aspx) for a list of supported file types.

Parameters for 'LoadRepeat':

J1) file: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the repeat sequence file. A folder may also be specified, in which case all of the sequence files within that folder will be loaded and used as repetitive sequences.

Example for 'loadRepeat':

```
loadRepeat
 file: "/Library/repetitive_seqs/123_repeat.seq"
```

****

K) loadSeq

Loads a sequence file or files for assembly. See our website (http://www.dnastar.com/t-smgafileformats.aspx) for a list of supported file types.

Parameters for 'loadSeq':

K1) blockContig: [text string]

(optional) Used in the reference-guided workflow.

K2) blockName: [text string]

(optional) Used in the reference-guided workflow.

K3) blockPos: [number]

(optional) Used in the reference-guided workflow.

K4) DiscardLinkerless: [true|false]

Specifies whether reads that do not have a linker sequence should be discarded from the assembly. Default is 'false.'

K5) file: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the sequence file(s) to be loaded. A folder may also be specified, in which case all of the sequence files within that folder will be loaded.

K6) groupName: [text string]

Used to identify the multi-sample group name for a read file.

K7) isPair: [true|false]

(optional) Specifies whether the query files contain paired end data. Default is 'false.'

K8) linker: [directory/filename enclosed in quotes]

The directory and file name of the .fas, fna, or .sff file containing the 454 linker sequences. If not specified, SeqMan NGen will use its default 454 linker sequence: GTTGGAACCGAAAGGGTTTGAATTCAAACCCTTTCGGTTCCAAC

K9) max: [number]

The maximum distance for the paired end constraint. Default is 10000.

K10) maxSeqs: [number]

Specifies the maximum number of reads to load from a file. There is no default value.

K11) mergePairs: [true|false]

Specifies whether the reads are paired end data that overlap and should therefore be merged. Default is 'false.'

K12) min: [number]

The minimum distance for the paired end constraint. Default is 0.

K13) multi-sample: [true|false]

Specifies whether reads are from a multi-sample run. Default is 'false.'

K14) seqTech: [normalScore|IonTorrent|SOLiD|Illumina|454|unknown]

(optional) Specifies the offset to be used when converting compressed quality scores into numerical values. These are the offsets used for the technology specified:

normalScore 33

IonTorrent 33

SOLiD 33

Illumina 64

454 33; quality scores for homopolymeric runs of >= 2 are oriented from 5' to 3' on the top strand.

unknown determined automatically based on the first data file.

K15) templateFragment : [number]

(optional) Used in reference-guided assemblies with gap closure.

Example for 'loadSeq':

```
  loadSeq
   file: "/Library/ABC_project/ABC_sequences.fas"
**********************************************************************
```

L) LoadTemplate

Loads a sequence file to be used as a template for all other sequences to be assembled to. The template sequence will be displayed as a "reference" sequence in SeqMan Pro for SNP analysis. See our website (http://www.dnastar.com/t-smgafileformats.aspx) for a list of supported file types.

Parameters for 'LoadTemplate':

L1) file: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the template sequence file to be loaded. A folder may also be specified, in which case all of the sequence files within that folder will be loaded and treated as template sequences.

Example for 'loadTemplate':

```
loadTemplate
 file: "/Library/abc_Project/abc_template.seq"
```

****

M) LoadVector

Loads a vector sequence file to be used for vector trimming. See our website (http://www.dnastar.com/t-smgafileformats.aspx) for a list of supported file types.

Parameters for 'LoadVector':

M1) cloneSite: [number]

This parameter specifies the position of the cloning site on the vector where insertion occurs. There is no default value.

M2) file: [directory/filename enclosed in quotes]

(required) Specifies the directory and file name of the vector sequence file to be used for vector trimming.

Example for 'loadVector':

```
loadVector
 file: "/Library/vectors/123_vector.seq"
 cloneSite:826
```

****

N) setDefaultDirectory

(required) Defines the default directory for the project. When a default directory is specified, files located in that directory only need to be identified by their subfolder and/or file name in subsequent commands.

Parameters for 'setDefaultDirectory':

N1) directory: [directory/filename enclosed in quotes]

(required) Specifies the default directory. Previously called 'defaultDirectory.'

N2) directoryMac: [directory/filename enclosed in quotes]

Specifies the default directory for Macintosh. Previously called 'defaultMacDirectory.'

N3) directoryWin: [directory/filename enclosed in quotes]

Specifies the default directory for Windows. Previously called 'defaultWinDirectory.'

Examples for 'setDefaultDirectory':

setDefaultDirectory: "/Library/ABC_proj/"

Once you have set a default directory, you may use two periods .. before a file name to specify that the file you wish to use is located in the parent folder of the default directory you specified.

loadVector file: "../123Vector.fas"

This specifies that the vector file, 123Vector.fas, is located in the ABC Data folder, the parent folder of the default directory.

Part III. Parameter Settings Commands

O) setContaminantParam

Allows you to adjust the parameters used for scanning for contaminant sequences. In order to be applied, this command must appear in the script before the 'loadContaminant' command, and the 'contamScan' parameter for the 'assemble' command must be set to 'true.'

Parameters for 'setContaminantParam':

O1) MerLength: [number from 5-50]

The minimum length of a mer required to be considered an exact match when scanning for contaminants. Default is 17.

O2) MinMerMatch: [number from 1-50]

The minimum number of matching mers required to mark the sequence as a contaminant. Default is 12.

Example for 'setContaminantParam':

setContaminantParam MerLength:17

setContaminantParam MinMerMatch:12

****

P) setParam

Allows you to adjust the stringency of one or more of the assembling parameters for the project. SeqMan NGen will use the default values for any parameter that is not specified within the script.

Parameters for 'setParam':

P1) AllowConstraintBased: [true|false]

Specifies whether the assembler should use constraints during assembly. Default is 'true.'

P2) AssembleBoneyard: [true|false]

Specifies whether, after a templated assembly has been completed, the unassembled sequences remaining should be assembled into contigs. If the template has been split, SeqMan NGen will attempt to join the split contigs together in new arrangements. Default is 'false.'

P3) CoverageType: [genome|fixed]

Specifies the type of coverage to be used for repeat handling. 'Genome' uses the length of the genome being assembled to calculate the expected coverage. 'Fixed' uses a fixed value as the expected coverage. If you know the length of the genome/fragment being assembled, we recommend using 'genome' for this parameter and then specifying the length using the 'genomeLength' parameter. If you do not know the genome/fragment length, used 'fixed' and provide the most accurate estimate of expected coverage for the 'FixedCoverage' value. Default is 'genome.' (Note: this parameter was called "Coverage" prior to SeqMan NGen 2.0.)

P4) DefaultQuality: [number from 5-100]

The value used for the base quality of sequences without quality scores. Default is 15.

P5) FixedCoverage: [number from 1-65535]

The estimated depth of the sequencing, which can be used instead of the genome length for repeat handling. Use caution when estimating the value for fixedCoverage. If the value you use is significantly lower than the actual depth, the assembly may take a much longer time to complete and may have too many mers flagged as repeats. Default is 20.

P6) GapPenalty: [number from 0-1000]

The penalty for opening or extending a gap during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. A high gap penalty suppresses gapping, while a low value promotes gapping. Default is 30.

P7) GenomeLength: [number from 0-1015 ULL]

Specifies the length of the genome or fragment being assembled. This is used to calculate expected coverage in determining repeat handling. Default is 0. (Note: this parameter was called "setGenomeParam" prior to SeqMan NGen 2.0.)

P8) HaploidSNP : [true|false]

Specifies whether to use the second most common base at a position when performing SNP passes. (See the 'snpPasses' parameter). Using this parameter will increase the SNP percentage for SNPs occurring on one allele of a diploid genome in a templated assembly. When haploidSNP is set to 'true', the lowCoverageThreshold parameter value should be greater than zero. Default is 'false.'

P9) HaploidThreshold: [number from 0-100]

The minimum number of times that the second most common base must occur at a position in order for it to be used to find SNPs during haploid SNP passes. (See the haploidSNP parameter above). Default is 0.

P10) LowCoverageThreshold: [number from 0-10000]

The minimum coverage required in an assembly to be excluded from SNP passes. SeqMan NGen will include regions in an assembly that have coverage less than the value specified as well as regions with zero coverage when it performs SNP passes. (See the snpPasses parameter). Default is 0.

P11) MatchRepeatPercent: [number from 100-1000]

The percent frequency a mer occurs compared to its expected frequency. Mers exceeding this value are flagged as repeated and not used as mer tags in determining overlaps. Default is 150. (Note: this parameter was called "maxCoverageRatio" prior to SeqMan NGen 2.0.)

P12) MatchScore: [number from 1-1000]

The score for a base match during an alignment. This score contributes to the pairwise score used to calculate match percentage. Increasing the matchScore value will allow for longer or more frequent gaps, thus forcing bases that match to be assembled together. Default is 10.

P13) MatchSize: [odd whole number]

The minimum number of matching consecutive bases required to determine the overlap of sequence reads. If an even number is entered, SeqMan NGen will automatically increase the value to the next odd number. Default is 21. (Note: this parameter was called "setParam MerLength" prior to SeqMan NGen 2.0.)

P14) MatchSpacing: [number from 1- 1000000]

The length of the window of a sequence read where at least one mer tag will be chosen. Default is 50. (Note: this parameter was called "merTagWindow" prior to SeqMan NGen 2.0.)

P15) MatchWindowLength: [number from 10-1000]

The size of the window used to calculate the match percentage. Default is 50.

P16) MaxAssemblyCoverage : [number from 0-65535]

The maximum depth of coverage allowed in the templated assembly. SeqMan NGen will not exceed the coverage specified by this threshold. This parameter is only available for templated assemblies, and should be used with caution as it will limit the number of sequences included in the assembly. The default value of 0 indicates unlimited coverage.

P17) MaxContigs: [number]

The maximum number of contigs to write to an .assembly project. This command is not generally needed due to SeqMan's capacity to handle a very large number of contigs. There is no default value.

P18) MaxGap: [number from 0-1000]

The maximum number of gaps allowed per 1000 bases in the alignment. Default is 6.

P19) MaxUsableCount: [number from 1-65535]

Any mers occurring more frequently than FixedCoverage multiplied by MaxUsableCount are disregarded as mer tags from the assembly. Default is 25.

P20) MinContigSeqs: [number from 0-10000]

The minimum number of sequences in a contig. After an assembly has been completed, any contigs without a template sequence will be disassembled if they contain fewer sequences than the number specified. The use of this parameter is recommended when performing de novo assemblies using data from Next Generation sequencing technologies, such as Illumina, as these types of assemblies can produce tens of thousands of very small contigs. Default is 0.

P21) Minimizer: [number]

(Intended for internal use only). An experimental way of choosing mer tags that may save time and memory. The accuracy of this parameter has not been verified by DNASTAR.

P22) MinMatchPercent: [number from 0-100]

The minimum percentage of matches in an overlap required to join two sequences in the same contig. Default is 93. (Note: this parameter was called "minMatchPercentage" prior to SeqMan NGen 2.0.)

P23) MismatchPenalty: [number from 0-1000]

The penalty for a base mismatch during an alignment. This penalty is deducted from the pairwise score used to calculate match percentage. Default is 20.

P24) SkipRealign: [true|false]

This parameter only affects de novo assemblies, and specifies whether to skip the realignment step of the assembly. The realignment step will then analyze each sequence at the nucleotide level to determine the exact position of each sequence in the alignment. Default is 'false.'

P25) SNP: [true|false]

(optional) Specifies whether a SNP detection pass of the gapped alignment is made during the assembly. Default is 'true.'

P26) snp_checkStrandedness: [true|false]

(optional) Specifies whether the strand that each read comes from is considered in the SNP calculation. This is ignored by the Simple method. Default is 'false.'

P27) snp_minPctToScore: [number from 0-1]

(optional) Specifies minimum percentage of reads in a column which must differ from the reference in order to score the column. For the Simple method, this is the only criteria used to call a SNP. For the Diploid and Haploid methods, this is a filter applied before the other parameters. Default is 0.05.

P28) snp_minProbNonrefToCall: [number from 0-1]

(optional) Specifies the minimum probability of a SNP column which is required to call a SNP, expressed as a number from 0 and 1. The probabilities of all genotypes other than Homozygous Reference are totaled and checked against this number. This is the final filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 0.1, requiring a minimum 10% change.

P29) snp_minVariantDepthToScore: [number from 0-100]

(required if "snp" is true) Specifies the minimum depth required for a specific base (or deletion) in a column before it is considered usable for SNP calling. This is the second filter applied during the Diploid and Haploid SNP calling methods, and is ignored by the Simple method. Default is 2.

P30) snp_minWeight: [number]

(optional) Specifies the minimum quality score for a base to be considered in the SNP calculation. Default 5.

P31) SNPMatchPercentage: [number from 0-100]

The minimum match percentage required during passes to fill in SNP regions. See the snpPasses parameter. Default is 90.

P32) snpMethod: [simple|haploid|diploid|population]

(optional) Specifies the SNP detection method to use. Simple produces a count of each type of base in the column and calculates the percent of non-reference bases. Haploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base called at that position. Diploid uses a Bayesian statistical model to calculate a probability score that the position contains a polymorphism and give a quality score for the base(s) called at that position. Based on the scores, it also calls the genotype at each position. Default is 'diploid.'

P33) SNPPasses: [number from 0-10]

The number of times SeqMan NGen will cycle through a templated assembly, attempting to fill in regions with low coverage or no coverage due to SNPs. Default is 2.

P34) SplitFalseJoins: [true|false]

Specifies whether the assembler should identify and splits false joins based on the set of false join parameters indicated. Default is 'false.'

P35) SplitTemplateContigs: [true|false]

Specifies whether, after a templated assembly has been completed, the template should be split into contigs at areas where there is zero coverage. Split contigs will be grouped into scaffolds with a defined position to allow for easy sorting when the project is viewed in SeqMan Pro. Annotations on the template sequence will also be split, and any /codon_start qualifiers will be adjusted to stay in frame. Default is 'false.'

P36) TemplateDefaultQuality: [number from 5-50000]

The value used for the base quality of template sequences without quality scores. Default is 500.

P37) TrimToMer: [true|false]

Specifies whether to trim the reads to the matching mer tags within the read. For each read, SeqMan NGen looks for mers that exist in the template (for templated assemblies) or in any other read in the assembly (for de novo assemblies). It then sets the trimming for the read to the start of the first mer found and the end of the last mer found. Trimming to mer may be useful when assembling data without accurate quality scores, data with very short linkers, or when assembling SOLiD data. Default is 'false.'

P38) UseRepeatHandling: [true|false]

Specifies whether to use the repeat probabilities to determine if a mer occurs too frequently to use. This parameter should only be used for de novo assemblies, unless the assembleBoneyard parameter is set to 'true' for the templated assembly. Default is 'true.'

Example for 'setParam':

setParam useRepeatHandling:true

setParam coverageType:fixed

setParam fixedCoverage:20

setParam matchSize:15

setParam minMatchPercent:90

setParam matchSpacing:10

setParam matchRepeatPercent:150

setParam maxUsableCount:25

setParam maxGap:15

setParam matchWindowLength:50

setParam matchScore:10

setParam maxAssemblyCoverage:0

setParam gapPenalty:30

setParam mismatchPenalty:20

setParam defaultQuality:15

setParam templateDefaultQuality:500

setParam splitFalseJoins:true

setParam allowConstraintBased:true

setParam skipRealign:false

setParam splitTemplateContigs:false

setParam assembleBoneyard:false

setParam minContigSeqs:0

setParam snpPasses:2

setParam snpMatchPercentage:90

setParam lowCoverageThreshold:0

setParam haploidSNP:false

setParam haploidThreshold:0

****

Q) setQualityParam

Allows you to adjust the parameters used for quality trimming. In order to be applied, the 'trimEnds' parameter for the 'assemble' command must be set to 'true.'

Parameters for 'setQualityParam':

Q1) EndRegion: [number from 1-100]

The number of bases at the end of a sequence considered to be the "end region" which is used by other quality parameters. Default is 5.

Q2) MaxN: [number from 1-100]

The maximum number of "N" bases permitted in the window used for N-based quality trimming. Default is 2.

Q3) MaxNHiQual: [number from 0-100]

The maximum number of "N" bases permitted in the window used for N-based quality trimming to meet the high-quality threshold. Default is 1.

Q4) MinAveHiQual: [number from 10-40]

The minimum averaged quality score of the evaluated window required to be considered high-quality. Default is 22.

Q5) MinAveLowQual: [number from 5-40]

The minimum averaged quality score of the evaluated window required to be considered low-quality. Default is 20.

Q6) MinEndBaseQual: [number from 5-40]

The minimum quality base score required in the specified end region. Default is 15.

Q7) NTrimWinLength: [number from 5-100]

The length of the window used for "N-based" quality trimming. N-based quality trimming trims bases that are called "N" and is used only when quality scores are not available. Default is 7.

Q8) WinLength: [number from 2-100]

The length of the window used for averaging quality scores. Default is 5.

Example for 'setQualityParam':

```
setQualityParam winLength:30
```

setQualityParam minAveLowQaul:14

setQualityParam minAveHiQaul:18

setQualityParam minEndBaseQaul:15

setQualityParam endRegion:15

setQualityParam nTrimWinLength:50

setQualityParam maxN:2

setQualityParam maxNHiQual:1

****

R) setRepeatParam

Allows you to adjust the parameters used for scanning for repetitive sequences. In order to be applied, this command must appear in the script before the 'loadRepeat' command, and the 'repeatScan' parameter for the 'assemble' command must be set to 'true.'

R1) AlignCutoff: [number from 10-1000000]

The minimum acceptable alignment score. When the alignment score drops below the specified value, this indicates that the end of the alignment between the read and the repeat has been reached, and the alignment will stop. Default is 100.

R2) MaxMerGap: [number from 0-50]

The maximum distance between two mers required to be considered a matching pair. Default is 10.

R3) MerLength: [number from 5-50]

The minimum length of a mer required to be considered an exact match when scanning for repeats. Default is 17.

R4) MinEndFlagLen: [number from 5-1000000]

The minimum length required for a mer to be flagged as a repeat if the segment is bound by the end of the read. Default is 25.

R5) MinFlagLength: [number from 5-1000000]

The minimum length required for a mer to be flagged as a repeat. Default is 50.

R6) MinMerMatch: [number from 2-25]

The minimum number of matching mers required to start an alignment. Default is 2.

Example for 'setRepeatParam':

setRepeatParam merLength:17

setRepeatParam minMerMatch:2

setRepeatParam maxMerGap:10

setRepeatParam minFlagLength:50

setRepeatParam alignCutoff:100

setRepeatParam minEndFlagLength:25

****

S) setVectorParam

Allows you to adjust the parameters used for vector trimming. In order to be applied, this command must appear in the script before the 'loadVector' command, and the 'vectScan' parameter for the 'assemble' command must be set to 'true.'

Parameters for 'setVectorParam':

S1) AlignCutoff: [number from 10-1000000]

The minimum acceptable alignment score. When the alignment score drops below the specified value, this indicates that the end of the alignment between the read and the vector has been reached, and the alignment will stop. Default is 100.

S2) EndCutOff: [number from 0-1000000]

The distance to the endpoint where trimming will go all the way to the end of the sequence. Default is 25.

S3) EndMerMatch: [number from 1-25]

The minimum number of mer matches required to start an alignment in the specified end region. Default is 1.

S4) EndRegion: [number from 0-1000000]

The number of bases at the end of a sequence where a lower stringency for matching and trimming is used. Default is 15.

S5) MaxMerGap: [number from 0-50]

The maximum distance between two mers required to be considered a matching pair. Default is 5.

S6) MergeTrimGap: [number from 0-1000000]

Maximum distance between two trim segments that will cause the segments to be merged
MergeTrimGap limits trimming to the ends of sequence reads, while EndCutOff doesn't.
Controls how sensitive trimming should be in areas where some portions of the sequence match a vector and other portions don't. The higher the number the more likely the vector trimmer will find all the vector sequence in a region of poor quality. The smaller the number, the more confidence there is that the bases trimmed are actually vector and not a spurious match. Default is 7, which is suitable for trimming linkers from the ends of sequences.

S7) MerLength: [number from 5-25]

The minimum length of a mer required to be considered an exact match when searching for vector. Default is 9.

S8) MinEndTrimLength: [number from 5-1000000]

The minimum length to be trimmed when a vector matches the end of a read. This parameter can be useful in preventing small spurious matches from being trimmed, which may be significant with short read technologies. Default is 5.

S9) MinMerMatch: [number from 1-25]

The minimum number of matching mers required to start an alignment. Default is 3.

S10) minTrimLength: [number from 5-1000000]

The minimum length required for a mer to be considered as a match for vector trimming. Default is 30.

Example for 'setVectorParam':

```
setVectorParam merLength:9
setVectorParam minMerMatch:3
setVectorParam MerGap:5
setVectorParam minTrimLength:30
setVectorParam minEndTrimLength:5
setVectorParam alignCutoff:100
setVectorParam endRegion:15
setVectorParam endCutoff:25
setVectorParam endMerMatch:1
```

Part IV. Preprocessing and Assembling Commands and Parameters

T) assemble

(required) Reprocesses and assembles the sequences that have been loaded. Preprocessing may include quality trimming, and scanning for vector, repetitive, and contaminant sequences.

Parameters for 'assemble':

T1) assembleBlocks: [true|false]

Specifies whether the assembly is a reference guided assembly. Default is 'false.'

T2) contamScan: [true|false]

If true, sequences will be scanned for the specified contaminant sequences before assembling. Also see loadContaminant. Default is 'false.'

T3) doAssemble: [true|false]

If false, only the preprocessing will be done, and the sequences will not be assembled. Default is 'true.'

T4) repeatScan: [true|false]

If true, sequences will be scanned for the specified known repetitive sequences before assembling. Also see loadRepeat. Default is 'false.'

T5) trimEnds: [true|false]

If true, the sequences will be trimmed based on quality scores before assembling. Default is 'false.'

T6) vectScan: [true|false]

If true, the sequences will be scanned and trimmed for vector before assembling. Also see loadVector. Default is 'false.'

Example for 'assemble':

```
 assemble
  trimEnds:false
  vectScan:false
  repeatScan:false
  contamScan:false
  doAssemble:true
**********************************************************
```

U) FixedTrim

Trims reads prior to assembly using fixed values. Based on the parameter settings for this command, SeqMan NGen will trim reads either by a specified number of bases from each end, or to a specified range.

U1) end3: [number from 0-1000000]

If trimRelative (see below) is set to 'true,' then this value indicates the number of bases for SeqMan NGen to trim from the 3' end of each read. If trimRelative is set to 'false,' then this value indicates the specific 3' coordinate to which reads should be trimmed. Default is 0.

U2) end5 : [number from 0-1000000]

If trimRelative (see below) is set to 'true,' then this value indicates the number of bases for SeqMan NGen to trim from the 5' end of each read. If trimRelative is set to 'false,' then this value indicates the specific 5' coordinate to which reads should be trimmed. Default is 0.

U3) trimRelative: [true|false]

Specifies whether the value for the end3 and end5 parameters should indicate the number of bases for SeqMan NGen to trim from the 3' or 5' end of each read. When 'false,' the value specified for the end3 or end5 parameter indicates the specific coordinate to which reads should be trimmed. Default is 'true.'

Example for 'fixedTrim':

```
 fixedTrim
  end5:10
  end3:20
  trimRelative:true
```

****

V) RealignContigs

(optional) Does another pass through a templated assembly once the initial assembly is complete, and realigns contigs as needed. (This step occurs automatically for de novo assemblies.) Using this command may improve the accuracy of the final assembly by correcting occasional misalignments that can occur in gapped regions, however note that this step may significantly increase the time to assemble. This command must appear in the script after the 'assemble' command.

****

W) RemoveSmallContigs

This command disassembles any contigs without template sequences that have fewer than the specified number of sequences.

Parameters for 'removeSmallContigs':

W1) minLength: [number]

Specifies the minimum length of a contig to prevent it from being disassembled. Default is 0.

W2) minSeqs: [number]

(required) Specifies the minimum number of sequences necessary in a contig to prevent it from being disassembled. Default is 0.

****

X) SetPairSpecifier

Defines the paired end pair specifier for the paired Sanger and Illumina sequences in the assembly. This command must appear in the script before the assemble command, but after sequences have been loaded (loadSeq). For more information on assembling 454 paired end data, see the 'load454PairedEnd' command. Pair specifiers define the naming convention for sequence pairs, as well as requirements for a minimum and maximum distance between the opposite ends of the inserts. Expressions for forward and reverse naming conventions should be created using the paired end specification language. Forward and reverse sequences must have identical names except for the unique portion that determines the direction of the clone.

Parameters for 'SetPairSpecifier':

X1) pairs: [forward|reverse|min|max]

This parameter lists the paired end constraints, specified by the following four values. Each value should be separated by a space and the list of values enclosed in double brackets {}. An additional set of brackets is required around all of the paired end constraints, regardless of whether one or multiple pair constraints are specified.

X2) forward: [text string enclosed in quotes]

A naming pattern to match forward clones.

X3) max: [number]

The maximum distance for the paired end sequences to be separated. There is no default value.

X4) min: [number]

The minimum distance for the paired end sequences to be separated. There is no default value.

X5) reverse: [text string enclosed in quotes]

A naming pattern to match reverse clones.

Example for 'setPairSpecifier':

(defines 2 pair specifiers each with different size ranges)

```
  setPairSpecifier
   pairs:{{forward:"(.*)(2kb)(.*)-FP.*$"reverse:"(.*)(2kb)(.*)-RP.*$" min: 1500 max: 2500}
        {forward:"(.*)(8kb)(.*)-FP.*$" reverse:"(.*)(8kb)(.*)-RP.*$" min: 7000 max: 9000}}
```

****

Y) SplitLinkerReads

Splits specified reads based on their match to given linker sequences. Reads that align to the linker and include the linker site (as specified by the linkerSite parameter or by the cloneSite option in an *.fof file) will be split into two reads. The two newly split reads will be designated by _A and _B appended to the name.

Parameters for 'SplitLinkerReads':

Y1) linkerFile: [directory/filename enclosed in quotes]

The directory and file name of the linker file.

Y2) linkerSite: [number]

The position indicating where reads should be split. There is no default value.

Y3) seqFile: [directory/filename enclosed in quotes]

The directory and file name of the sequence reads.

Example for 'splitLinkerReads':

```
  splitLinkerReads
   seqFile: "/Library/123_project/reads.fas"
   linkerFile: "/Library/123_project/linker.fas"
   linkerSite:30
```

****

Z) SplitTemplates

Splits template contigs into multiple contigs in areas where there is zero coverage. Split contigs will be grouped into scaffolds with a defined position to allow for easy sorting when the project

is viewed in SeqMan Pro. Annotations on the template sequence will also be split, and any /codon_start qualifiers will be adjusted to stay in frame.

****

AA) appendToAssembly

(This command is for the reference-guided workflow and is intended for internal use only).

****

BB) convertReads

(optional) Converts a sequence from one file format to another. This command is particularly useful for converting SOLiD .csfasta files into .fastq files that can be used by the XNG assembler.

Parameters for 'convertReads':

BB1) destination: [directory/filename enclosed in quotes]

The location and filename for the output.

BB2) file: [directory/filename enclosed in quotes]

The input file containing the reads. (Synonym for 'reads').

BB3) format: [genbank|fastq]

(optional) Specifies the format of the output file. If 'genbank' is entered, the output will be in .gbk format. If 'fastq' is entered, the output will be in .fastq format. Default is 'fastq.'

BB4) reads: [directory/filename enclosed in quotes]

The input file containing the reads. (Synonym for 'file').

****

CC) extendContigs

(Intended for internal use only).

Parameters for 'extendContigs':

CC1) extendPasses: [number]

CC2) mergeContigsInScaffold: [true|false]

****

DD) include

(optional) When building a script, this command can be used to call up additional lines of script previously stored in a text file. In this way, a group of commands can be shared between two or more scripts.

Parameters for 'include':

DD1) file: [directory/filename enclosed in quotes]

Specifies a directory and name for the file.

****

EE) MakeSeqNamesUnique

(Intended for internal use only).

****

FF) RemoveDuplicates [true|false]

Coalesces multiple identical reads at the same position into a single read, provided the reads match the template exactly. If this feature is active, at the end of assembly, XNG will print the message: "Coalesced $lld identical reads that matched the template exactly." Default is 'false.'

****

GG) set

(optional) Used to set variables. See the example below and those under the 'runScript' command.

Example for 'set':

```
set $snp:true
set $snpMethod:"Diploid"
```

****

HH) setAssemblyReport

(Intended for internal use only). Used to designate a file for a tab delineated report, similar to a report that XNG generates. This is useful during development to test how code changes impact results.

Parameters for 'setAssemblyReport':

HH1) file: [directory/filename enclosed in quotes]

Specifies the folder and file name. (Synonym for 'name').

HH2) name: [directory/filename enclosed in quotes]

Specifies the folder and file name. (Synonym for 'file').

****

II) SplitMIDSeqs

(optional) Used to split 454 MID reads into individual files with one file per MID tag.

Parameters for 'SplitMIDSeqs':

II1) destination: [directory/filename enclosed in quotes]

The location and filename for the output.

II2) file: [directory/filename enclosed in quotes]

The location and filename for the input. (Synonym for 'reads').

II3) reads: [directory/filename enclosed in quotes]

The location and filename for the input. (Synonym for 'file').

****

JJ) SplitPairs

(optional) Used to split 454 or ion torrent mate pair files into forward and reverse (and singleton) files.

Parameters for 'SplitPairs':

JJ1) destination: [directory/filename enclosed in quotes]

The location and filename for the output.

JJ2) DiscardLinkerless: [true|false]

Specifies that reads without a linker sequence should be discarded from the assembly. Default is 'false.'

JJ3) file: [directory/filename enclosed in quotes]

The location and filename for the input. (Synonym for 'reads').

JJ4) reads: [directory/filename enclosed in quotes]

The location and filename for the input. (Synonym for 'file').

JJ5) seqTech: [text string]

(optional) Specifies the offset to be used when converted compressed quality scores into numerical values. Values of normalScore, IonTorrent (for IonTorrent data), or SOLiD (for Applied Biosystems SOLiD data) will use an offset of 33. A value of Illumina (for Illumina data) will use an offset of 64. A value of 454 (for Roche 454 data) will use an offset of 33 and orient quality scores for all homopolymeric runs of two or more to be descending from 5' to 3' on the top strand. If a value of unknown is entered, the assembler will determine the offset from the first data file.

Example for 'SplitPairs':

```
SplitPairs
  destination:"c:data\splitReads\"
  reads: {
    { file:"C:data\reads\file1.fas"   format: IonTorrent }
    { file: "C:data\reads\file2.fas"   format:454   discardLinkerless: true}
    }
```

Research References

Li H, Ruan J, and Durbin R (2008). “Mapping short DNA sequencing reads and calling variants using mapping quality scores.” Genome Res. 2008 Nov;18(11):1851-8. doi: 10.1101/gr.078212.108. Epub 2008 Aug 19.

Index