Application Modules for:
SNP/Indel/Structural Variant Analysis
Large Genome Alignment and Variant Discovery
Exome Assembly and Variant Discovery
RNA-Seq Analysis
Target Assembly
de novo & Paired End Assembly
Paired End Merging
Digital Gene Expression
Transcriptome Analysis
Metagenomic Analysis
miRNA Discovery and Quantification
Rare Disease Analysis

Compatible with:
Ion PGM™
Roche Genome Sequencer FLX™
Illumina® Genome Analyzer
Life Technologies SOLiD™ System
NextGENe Benefits
● Instant Knowledge...Single Annotated data review screen
● Increased Accuracy...Unique technologies increase accuracy by
  Removing sequencing errors
  Elongating short read sequences
● Compatible with all major sequencing systems...Automated format conversion tool
● Biologist Friendly Windows interface...
  Application Driven
  Automated inspection of input files to set analysis parameters
  Requires no scripting
  Reduces Bioinformatics requirements
  Unattended Batch Processing Capabilities
● Annotated results in single easy-to-navigate view...
  Multiple integrated, exportable reports
  Analysis filters
● Low-Cost Hardware Requirements

NextGENe Software is a complete, “free-standing” analysis package designed for use by biologists in the analysis of data from Next Generation Sequencing systems. The icon driven, easy-to-use Windows® interface significantly reduces bioinformatics requirements, provides annotated analysis review, while reducing sequencing errors to improve analysis accuracy and speed.

Instant Knowledge
● Annotation
● Easy Navigation
● Exportable

NextGENe’s analysis browser provides a highly interactive review of annotated analysis results in a single view. Navigation is as simple as drawing “boxes” to zoom in or out, graphics and text reports are hyperlinked to speed data review and “hot” keys ease navigation.

Example of annotated Whole Human Genome data review with NextGENe browser. Navigation is simple either using Hot Keys or by dragging mouse over screen to move across the genome or zoom in on selected areas. Text Reports are linked to browser for quick, easy data review.

Mutation Report is hyperlinked to graphical NextGENe browser and dbSNP database. Several filtering options are available to speed and ease analysis review:

Paired End Merging
● Increases Sequence Accuracy to 99%
● Improves Assemblies
● Increases read length to 250bp-350bp

The Paired End Merging application of NextGENe software uses overlapping reads to merge paired end reads creating more unique and accurate sequences. The Condensation Tool™ clusters similar reads using a 12bp anchor sequence as well as the flanking shoulder sequences (of varying lengths).

The Paired End linking application works by elongating paired reads to the point that there is an overlap between the pairs. This allows two paired reads to be merged together to form one continuous longer read. 75bp reads from a 200 bp library, for example, in a single elongation cycle will overlap and allow for merging of the paired reads. Each cycle of elongation typically elongates the reads 160%. Enough elongation cycles are used until at least 15% of the paired reads overlap.
Condensation Tool™
- Identifies identical anchor sequences
- Statistically polishes short reads to reduce instrumental error
- Increases read length and accuracy

The Condensation Tool is used to statistically polish and lengthen short sequence reads into fragment sizes that are more manageable. Short reads such as those from the Illumina GA and Life Technologies SOLID System™ are often not unique within the genome being analyzed. By clustering similar reads containing a unique anchor sequence, data of adequate coverage is condensed and the short reads are lengthened. The unique anchor sequence, or index, is a 12 base fragment that is found in several of the reads. All reads containing this exact sequence are clustered together. Many of the reads within a cluster contain several homologous nucleotides both upstream and downstream of the index sequence. This read cluster can then be sorted by the flanking shoulder regions into sub groups based upon similarity. The consensus of these groups is much larger in length, and these elongated base pair fragments are more unique within the genome, with exceptions such as homopolymeric regions, repeats and duplications.

NextGENe Software’s Condensation Tool can also be used to remove errors in association with Ion PGM™ and Roche systems. By clustering similar keywords within several reads that are flanked by homopolymers, errors at homopolymers and within the remainder of the reads can be corrected.

NextGENe offers several Condensation options, allowing biologists to select the error correction methodology that works best for the data sets.

NextGENe’s Condensation Tool clustered similar reads containing the same anchor sequence of CTGGGGTTACAG. The right shoulder of 8 nucleotides is used to subdivide the groups differing in sequences of GTGTGAGC and GTGCCTGC. A consensus sequence is generated for each group, almost doubling the read lengths. Several condensation cycles can be employed to further lengthen reads for larger Indel discovery.

SNP/INDEL Detection
- SNP Detection
- Indel Detection (up to 33% of elongated read length)
- Low False Positive Rate
- Low False Negative Rate
- Biologist friendly reporting
- Export results to data base or LIMS system
- Scoring of Variants

SNP’s and Micro Indels, up to 1/3 of elongated read length, can be detected in sequencing data from both longer sequence reads (Roche Titanium FLX and Ion PGM™) and short reads from the Illumina and SOLID System sequencing technology. Use of the Condensation Tool elongates short reads, increasing their uniqueness probability while polishing the data to remove chemistry and instrumental errors. NextGENe software automatically calculates a confidence score for each found variant.

In the region of aligned sequence reads, novel mutation calls are highlighted in blue, previously reported in purple.

The Whole Genome Pane is located at the top of the display – coverage is indicated by gray lines, red tick marks indicate the breakpoints between the transcripts within reference, blue tick marks identify the location of novel SNPs, previously reported SNPs are indicated in purple.

A Mutation Report was generated for the run, showing a list of all variations marked as mutation calls. Calls can be manually reviewed, and this report allows for calls to be edited, deleted or added. Options are available for customizing the view of this information, in addition to further filtering. The calls within this report are organized by position within the reference, and each line contains the position within reference, the reference nucleotide, coverage, percentages for each allele found, and percentages of reads containing indels, amino acid changes, gene and/or chromosome location and dbSNP identification. In the figure, a mixture of alleles is observed at position 7753. The reference nucleotide is G while 73.13% of aligned read have a C at the position and the remaining 26.87% of aligned have a G at the position. The dbSNP identification is shown since this is a known SNP. The genotype is shown as CG indicating that the mutation is heterozygous.
Structural Variant Detection

- Discovery of large Insertion & Deletion; Translocation; Gene Fusion
- Flexible Alignment technology allows for large mismatches
- Creation of “Pseudo Pairs” allows for detection and mapping of structural variations

Structural variants (SVs) include insertions, deletions, inversions, and gene fusions that frequently occur across the human genome with over 1000 segmental deletions and over 200 copy number polymorphisms having been reported. These genomic structures have been shown to be important in a number of diseases, usually referred to as genomic disorders.

There are several ways to detect and map SVs, but there are limitations. Microarrays are useful for detecting differences in copy number but are unable to detect smaller SVs and cannot map boundaries. It is possible to detect SVs smaller than 1 kb with sequencing. Paired-end read mapping (PEM) has been used to detect shorter deletions and to hone in on breakage sites but is unable to detect structure variations larger than the library size. NextGENe makes it easy to both find and map structural variants with sequence data from the Roche Genome Sequencer FLX Titanium System. Targeted sequencing methods such as exon capture with Roche Nimblegen or Agilent SureSelect™ assays can be used to significantly reduce the cost and time requirements of these experiments.

NextGENe allows large mismatches when aligning to the genome in order to find SVs and display information about those regions in a structural variation report. The structural variation report uses a specialized algorithm to list regions with high variant frequency. Interference from false positives caused by sequencing errors is rarely detected in this report since multiple errors are unlikely to occur in a local region.

NextGENe then generates pseudo-paired reads for the sequences aligned to these regions by breaking the original reads into pairs. These can be aligned to the reference genome in order to map the SVs, as seen in figure. Detailed information on where these reads align is available in the Paired Read Reports.

A detailed view shows how NextGENe highlights the mismatched portion of a read. This is an example of fusion gene discovery using NextGENe software. In this example Pseudo-paired reads have been created and split with each half aligning to their appropriate gene. The paired read report identifies the location of each half.

RNA-Seq Analysis

- Detect multiple transcripts- Insertions, Deletions, Fusions, Un-annotated Transcripts including new genes
- Expression Analysis – Actual coverage (min, max, avg) or normalized (Reads Per Thousand and RPKM)
- SNP detection including RNA editing
- Use GBK files or provided whole-genome references.
- Strand-specific analysis

NextGENe includes a comprehensive, proprietary algorithm for alignment of transcriptome data. It can be selected as the “Transcriptome” application in the project wizard. It has four main steps:

1. Perform whole genome alignment
2. Generate a junction reference based on annotated and newly detected exons and align unmatched reads to it
3. Call transcripts based on aligned data and annotated information
4. Align original data to called transcripts to ensure the best possible alignment

After alignment all three types of analysis results (SNPs, Expression, and Transcript calls) are available from the same project. The project view button switches back and forth between the aligned reads view and the called transcripts view. Strand-specific sequencing can be used to resolve overlapping transcripts, while paired-end sequencing is useful for more sensitive fusion detection.

Images below are from Applied Biosystems Strand-Specific SOLID™ System Sequencing.
Digital Gene Expression Studies, CNV, ChIPSeq & miRNA Analysis

- Align to entire Genome or to specific references
- Identifies binding sites and transcription sites
- Reports sequences, expressions levels and information for each identified peak
- Available comparison report compares multiple individuals or time based analysis
- Removes Duplicate Reads
- Expression Reporting
- Search Tool
- Displays Biological Information for each tag
- Lists new gene separately

All 2nd generation DNA sequencing technologies generate millions to hundreds of millions of the short sequence reads per run, providing powerful solutions for analyzing gene expression. However, the high inherent error rates of these systems, as well as the sheer volume of data produced, pose significant challenges for analysis.

NextGENe is an excellent tool to take advantage of the hundreds of millions of short sequence reads provided by the Illumina GA or the Life Technologies SOLID System. NextGENe’s unique statistical polishing capabilities remove chemistry and instrumental artifacts, providing accurate results, with a low false positive and negative rate.

The Sequence Alignment Tool has a Whole Genome View at the top of the screen, which shows each sequence of the library. Placing the mouse over the library while holding down control activates a yellow box containing the biological information for the tag that is currently at the cursor. The bottom of the screen contains all reads as they have been aligned to the library.

Mutation Confidence Scoring

- Overall mutation confidence score provided for every mutation
- Any penalty scores can be disabled
- Quickly view the distribution of scores in a project
- Filter based on the overall score or on penalty sub-scores

NextGENe includes a proprietary mutation confidence scoring system designed to make it easier to find the called mutations that are most likely to represent true variations. The overall score is the product of the coverage score and several penalty sub-scores:

- **Coverage score** – starts at 0 and has no upper limit, but is rarely higher than 32. It is calculated as $8 \times \log_{10}(\text{adjusted coverage})$. The adjusted coverage gives greater weight to the higher quality 5’ end of reads and less weight to the lower quality 3’ end.

- **Read Balance Score (0 to 1)** – A score of 1 indicates perfect or near-perfect balance between the number of forward and reverse reads. Unbalanced data may indicate misalignment or may allow basecalling biases to cause false positives. If the data is expected to be biased (as it is for some targeted sequencing applications) then this score should be disabled.

- **Allele Balance Score (0 to 1)** – Measures the major and minor allele balance and compares it to the read balance. The calculation is similar to a chi-square test. If the allele balance is different from the read balance then there is strong evidence that the mutation may be an error.

- **Homopolymer Score (0 to 1)** – Penalizes indels occurring in homopolymer regions for data that has that error profile

- **Mismatch Score (0 to 1)** – Penalizes mutations when many mismatches occur in a small area. This usually indicates untrimmed adapter sequence or misalignment.

- **WrongAllele Score (0 to 1)** – Penalizes mutations when a third allele is found which makes more than one possible mutation call possible (such as 60% A, 20%C, 20%T). This score is especially helpful for targeted capture data.

NextGENe Pipeline Automation

Streamline analysis of your NGS data with the NextGENe pipeline tool, which will sequentially perform analysis of multiple projects, by querying the sequencer platform for data availability; perform necessary format conversion; de-multiplex bar coded data; perform alignment against selected, annotated references; create chosen reports, apply filtering, and store results all on an unattended basis. While NextGENe is running, more projects can be added to the queue and NextGENe will begin processing these when it has completed previous jobs in queue. The NextGENe pipeline tool wizard quickly allows quick and easy setup of multiple analyses in minutes.

NextGENe Pipeline Automation tool is driven through a simple setup Wizard. Multiple analyses can be scheduled and run on an unattended basis.

Pipeline Tool will:

- Seek Data Availability
- Perform Format Conversion
- De-multiplex bar coded data sets
- Obtain Reference or Disease Panel(s)
- Perform Analysis
  - Re-Sequencing
  - Amplicon SNP/Indel
  - De novo Assembly
  - RNA-Seq
- Create Custom Reports
- Log and Save project file
Rare Disease Tool

- Advanced Comparison between multiple projects - 3 options
  - Manually set expected SNP types
    - Homozygous, Heterozygous, Present, Negative, Undetermined, etc
  - Load an inheritance template (Autosomal recessive, X-linked dominant, etc)
  - Compound Heterozygous
    - Includes a report listing all valid pairs of mutations
- Many advanced filtering options
  - Annotation (mRNA, CDS, Splice sites)
  - Mutation Confidence Score
  - dbSNP mutations
  - Substitutions (Silent, Missense, Nonsense) or Indels
  - Advanced ROI filtering
- View multiple projects side-by-side

NextGENe includes an advanced mutation comparison tool. When searching for causative mutations of rare diseases it can be used to narrow down the list of mutations from tens of thousands to a few dozen that can then be examined for possible clinical significance. The tool can also be used to compare unrelated samples in a very specific way- multiple samples with undetermined genotypes can be compared against positive and negative control samples.

Analysis of Pooled Samples using Barcode/index tags

- Automatically parses samples according to tags
- Flexible Tool selects bar codes based on stringent or loose fit criteria
- Works effectively with all major methodologies

NextGENe's Barcode Sorting Tool is able to accurately determine the number of tags used in sample preparation. Tags that are found at low frequency, often the result of sequencing errors, are not used for parsing the sample file.

NextGENe's Barcode Sorting Tool is able to allow for a small amount of error in matching tags by comparing the barcode to 3 segments within the read tags and requiring only one segment to match perfectly with the expected tag. Users can also choose to require a perfect match between the expected tag and the read tag. Some portions with same sequence in different barcodes will be ignored in the determination of the sample.
Fast Whole Large Genome Alignment
- Annotated References furnished
- Fast Alignment
- Variant Discovery

SoftGenetics has developed a modified Burrows-Wheeler transform (BWT) alignment method that includes several improvements over other methods to generate fast alignment of sequence reads to a whole large genome reference such as the human genome with high accuracy.

NextGENe’s whole genome alignment method is the first to align reads from the Roche Genome Sequencer FLX System, which often contain many indels due to homopolymer errors, to a whole genome reference with high speed. The whole genome alignment algorithm is also capable of quickly aligning SOLID™ System, Ion PGM™ and Illumina® Genome Analyzer data. Additionally, NextGENe’s whole genome alignment tool features complete annotation of the reference.

Alignment of high throughput short sequence reads to a large reference genome like the human genome is a difficult challenge. The Burrows-Wheeler transform is a widely accepted data compression algorithm that has been in use since 1994. Massively parallel sequencers such as the Illumina Genome Analyzer (Solexa sequencing technology), the Life Technologies SOLID System, Ion PGM™ and the Roche Genome Sequencer FLX System are capable of producing 1-200 million reads per run which has led to an interest in the usage of BWT algorithm to align this large volume of sample reads to entire genomes. In May 2009 researchers from The Wellcome Trust Sanger Institute published a paper detailing their novel BWT alignment method, BWA which improved upon previous methods by allowing for the alignment of 51bp Illumina reads as well as SOLID System color-space reads.

NextGENe's whole genome alignment algorithm aligns reads to the whole genome by matching seeds smaller than the read length and then extending the alignment to find the best matching position for the whole read. This allows for the alignment of long reads and reads with indels.

SoftGenetics furnishes several annotated large genomes including Human, Mouse, Rat, C. elegans, Dog, Rice, Horse and Cow. Additional genomes can be furnished upon request and NextGENe also contains a special tool for indexing additional large genomes.

<table>
<thead>
<tr>
<th></th>
<th>Total Reads</th>
<th>Read Length</th>
<th>Seed Size</th>
<th>Matched Reads</th>
<th>% Matched</th>
<th>Time (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long Reads</td>
<td>1.46M</td>
<td>100-200 bp</td>
<td>17 bp</td>
<td>1.32M</td>
<td>90.4%</td>
<td>1.5 3 10.5</td>
</tr>
<tr>
<td>Short Reads</td>
<td>24.4M</td>
<td>51 bp</td>
<td>17 bp</td>
<td>20.7M</td>
<td>84.8%</td>
<td>1.5 2 5</td>
</tr>
</tbody>
</table>

Whole genome alignment performance for short read and long read data. Data shown is for alignment to the whole human genome. Seed size indicates the number of bases used to find perfect matching positions in the reference.

Please open disk to review your applications of interest and to install a 30-day trial of NextGENe software.

Disk Contains:
- Application Notes
- 30-day free NextGENe Trial
- User Manual

Minimum Computer Recommendations:
DeskTop PC
- 64 bit, Windows XP, Vista, or 7 Operating System
- Processor: Dual Quad Core Processors
- RAM: 12GB
- 2TB Hard Drive

Intel Powered Macintosh
- OS: 10.4.6, with Parallels desktop for MAC or Apple Boot Camp
- Windows 64 bit XP, Vista, or 7 Operating System
- Processor: Dual Quad Core (2.4 GHz)
- RAM: 12GB DDR2-800MH
- 2TB Hard Drive

Note: Some applications will require additional memory