SOFTGENETICS®

NextGENe®

Next Generation Sequencing Software for Biologists

Application Modules for:

SNP/Indel/Structural Variant Analysis
CNV Analysis
Somatic Mutation Mining
Large Genome Alignment and Variant Discovery
Exome Analysis and Variant Discovery
RNA-Seq/Transcriptome Analysis
Targeted Sequencing and Resequencing Analysis
HLA Analysis
STR Human Identity Analysis
de novo & Paired End Assembly
Paired End Merging
Digital Gene Expression
Metagenomic Analysis
miRNA Discovery and Quantification
Rare Disease Analysis and Prediction
Multiple Project Comparison

Compatible with:

Ion Torrent Platforms
Roche Sequencing Platforms
Illumina Sequencing Platforms
Life Technologies SOLiD™System

SOFTGENETICS®
Software PowerTools for Genetics Analysis

NextGENe Benefits

Instant Knowledge…

Single Annotated data review screen

Increased Accuracy…

Unique technologies increase accuracy by

- Removing sequencing errors
- Elongating short read sequences
- Platform specific technologies

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…

- Automated pipeline tool speeds analysis
- Multiple integrated, exportable reports
- Analysis filters

Automated Analysis Pipeline

- Automated linkage to Geneticist Assistant™ NGS interpretation workbench

Low-Cost Hardware Requirements

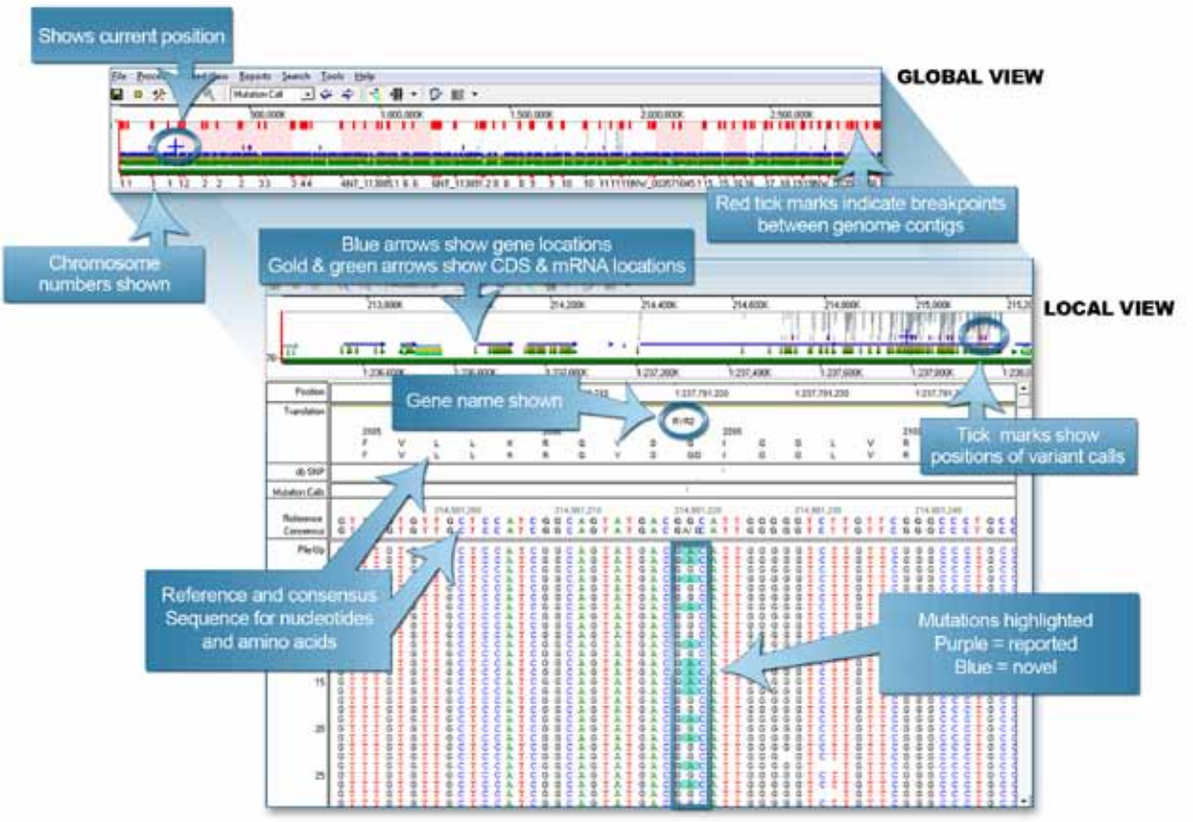
NextGENe Software is a complete, “free-standing” analysis package designed for use by biologists in the analysis of data from Next Generation Sequencing systems. The icon driven, easy-to-use Windows® interface significantly reduces bioinformatics requirements, provides annotated analysis review, while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

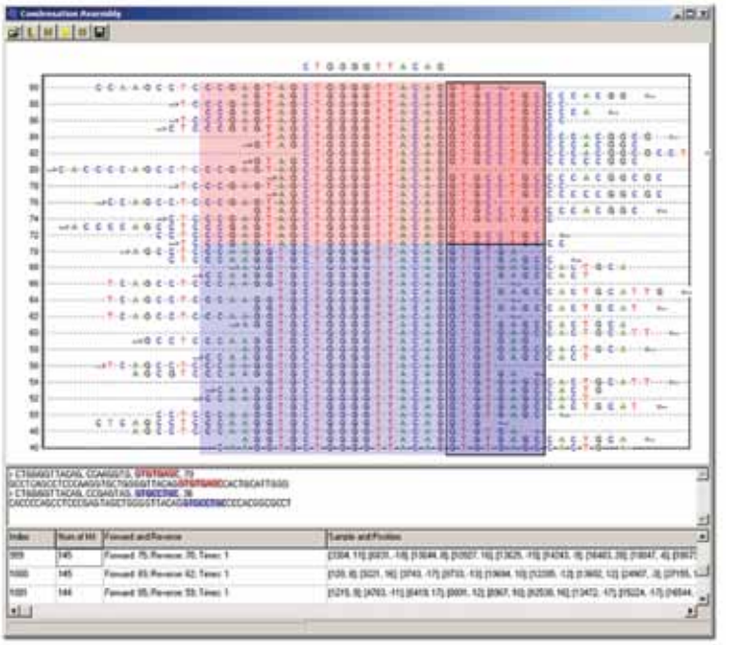
Condensation Tool™ (US Patent Number 8,271,206)

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® platform and Life Technologies SOLiD Systems™ 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 Torrent 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.

Optimized Analysis Performance

NextGENe has been optimized to perform on any Windows® 64 bit operating system from Vista through Windows 8, or Windows based server 2008R2 and forward. NextGENe also performs well on Mac® hardware when utilizing VMware, boot camp or similar boot managers along with Windows OS. NextGENe is multi-threaded, utilizing all available processors.

Below are performance examples on various fasta data sets of NextGENe version 2.3.3 forward:

Computer Type	Data Set Type	Number of Cores Used	Total Analysis Time	Million Reads per Hour
Intel® Core™ i7 2.8 GHz 8 core, 16GB RAM	10 GB Illumina, single end 100 bp reads	6	1 hour	61
Intel® Core™ i7 2.8 GHz 8 core, 16GB RAM	1GB Illumina single end 100bp reads	6	5 Min	119
Intel® Core™ i7 2.8 GHz 8 core, 16GB RAM	5GB x2 Illumina paired end 100bp reads	6	1.4 hours	48
Intel® Xeon™ 2.3 GHz 16 core, 60GB RAM	10GB x2 Illumina paired end 100bp reads	14	2.6 hours	46
Intel® Xeon™ 2.3 GHz 16 core, 60GB RAM	0.8 GB Ion PGM AmpliSeq™, 117bp reads	4	2 Min	19
Intel® Xeon™ 2.3 GHz 16 core, 60GB RAM	4.5GB Ion Torrent Proton WES 126bp reads	14	1.3 hours	26
Intel® Xeon™ 2.3 GHz 16 core, 60GB RAM	4.5 GB Ion Torrent Proton WES 126 bp reads	4	2.7 hours	13

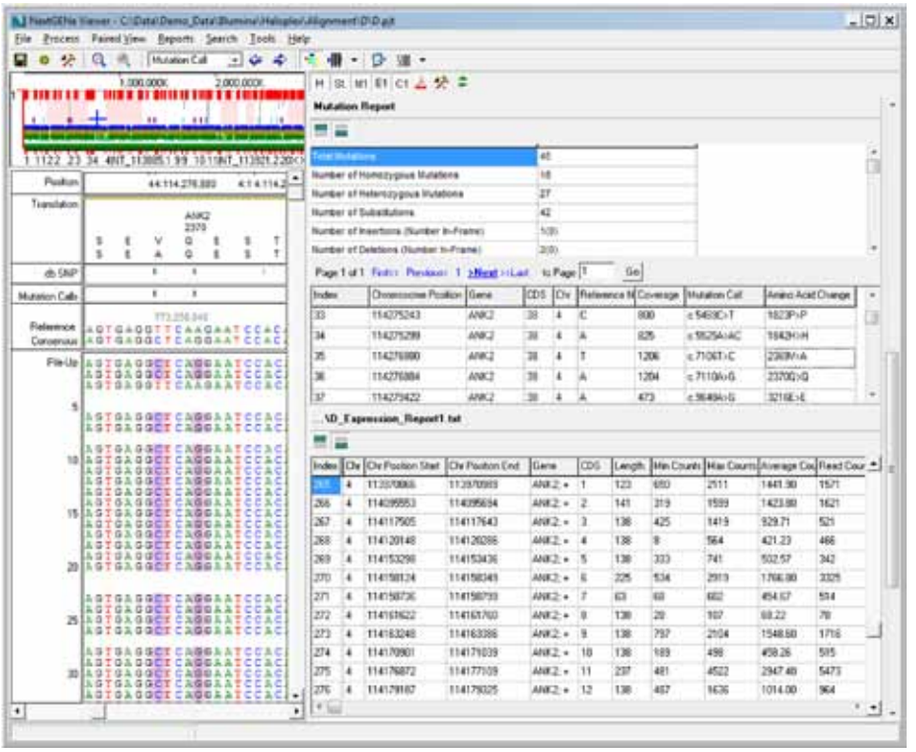
Reporting of Results and Quality Control Metrics

Customizable

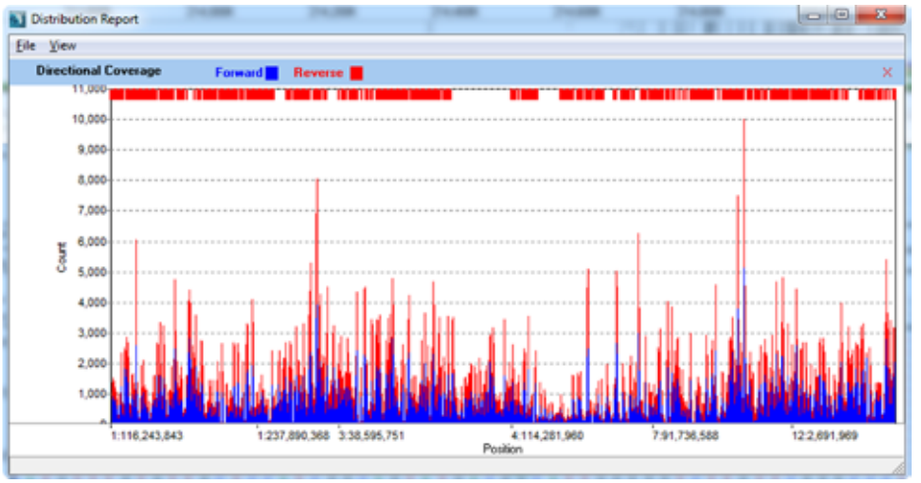
Various Formats including VCF & SIFT

NextGENe produces several reports, including Mutation Report, Coverage Curve Report, Distribution Report, Expression Report, Paired Read Reports, and more. Each report can be saved in various formats. For example, the Mutation Report can be saved in VCF, SIFT or a tab delimited text format with your choice of columns of information. NextGENe Viewer's main display contains several panes of information. The reporting pane can display several different reports, one of which is the Summary Report. The Summary Report can show several statistics and reports in a single view, wrapping up the results of a single project to show quality, coverage, mutation calls and more in one report.The Summary Report is tied in to the Post Processing step of the Project Wizard, allowing you to set up each report's configuration prior to the analysis. The Summary Report can be customized after the project is completed also, allowing you to tailor the report to each project's specific needs.

Customized Clinical Research Report



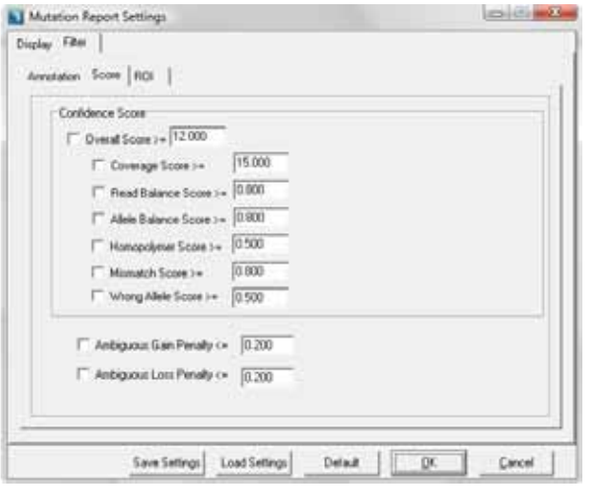
Coverage Plot



Run information, such as sample and reference files, processing time, and more are shown in the Summary Report. Statistics such as the number of matched reads, mismatched bases, etc. are also displayed. Reports that can be included in this summary are the Coverage Curve Report, Expression Report, Mutation Report, as well as others. Each report can be included in multiple formats within the Summary Report, enabling you to categorize the information. For example, you can configure one mutation report to identify the unreported mutation calls that are predicted to be pathogenic while having another mutation report to identify the reported mutation calls. In addition to charts and tables that each report offers, many of these reports bring with them their own summaries and statistics, including target coverage statistics, low coverage regions, mutation call summary, etc.

Mutation Confidence Scoring

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



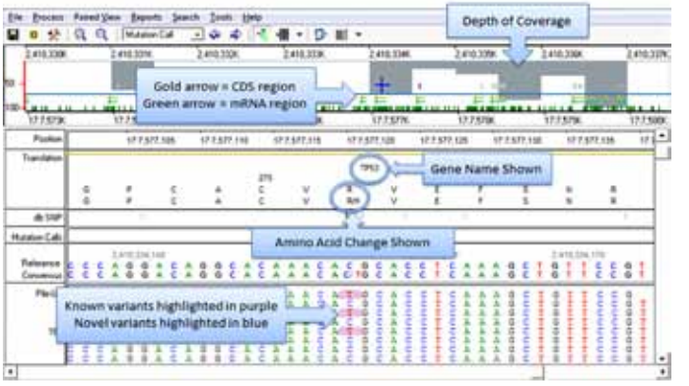
NextGENe software includes a proprietary mutation confidence scoring system designed to make it easier to find the called mutations that are most likely to represent true variations. The overall score is the product of the coverage score and several penalty sub-scores:

- **Coverage score** – starts at 0 and has no upper limit, but is rarely higher than 32. It is calculated as 8 * log10 (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)** – Penalized 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.

SNP/INDEL Detection

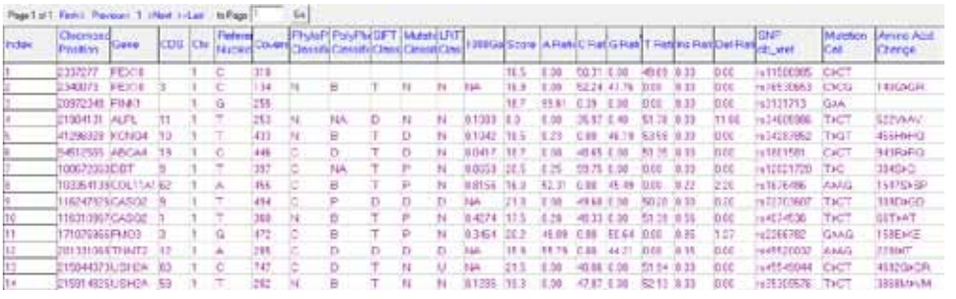
SNP Detection
Indel Detection (up to 33% of elongated read length)
Low False Positive Rate
Biologist friendly reporting
Export results in spreadsheet form or VCF format to data base or LIMS system
Scoring of Variants

SNP's and Micro Indels, up to 1/3 of elongated read length, can be detected in sequencing data from all sequencing technologies.Use of the Condensation® Tool elongates short reads, increasing read uniqueness probability in the genome, 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, 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, causative prediction by several databases, 1000 genomes frequency percentages for each allele found, and percentages of reads containing indels, amino acid changes, gene and/or chromosome location and dbSNP identification.

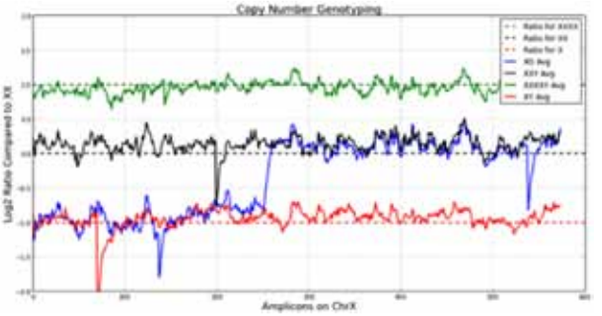
Copy Number Variation (CNV) Tool

- **Provides a confidence interval for the ratio between sample and control.**
- **Classifies regions as potential deletions, insertions, or normal copy number based on confidence interval**
- **Processes individual CNV changes (per exon, or per amplicon) into multi-locus calls that can be made more confidently.**

NextGENe includes a new Copy Number Variation (CNV) analysis tool designed to make variant calls on a case-control basis. It uses a proprietary normalization method, and works well with targeted sequencing data such as Ion AmpliSeq™ panels or the HaloPlex™ Target Enrichment System from Agilent Technologies or consistent depth of coverage whole Exome Sequencing data. Ideally the two samples used in a comparison will be as close as possible in experimental conditions. No special processing is needed to use the tool - any aligned NextGENe projects can be utilized for CNV analysis.

Sample	E.pjt											
Control	F.pjt											
Index	Description	Chr	Chr Start	Chr End	Gene	CDS	Length	Log2 Ratio	Score	Original Coverage (Sample:Control)	Normalized Coverage (Sample:Control)	
1	Amplicon206	chr7	91623915	91624109	AKAP9: +	6	195	0.90	3.72	100:55	100:54	
2	Amplicon255	chr7	150645512	150645650	KCNH2: -	11	139	-1.03	8.26	197:413	197:402	
3	Amplicon358	chr19	35521705	35521783	SCN1B: +	1	79	-2.73	11.00	38:258	38:251	

Figure indicates the detection of a known deletion in the KCNH2 gene using HaloPlex™ Cardiac Panel data. The log2 ratio (-1.03) was very close to the expected value of -1.0.

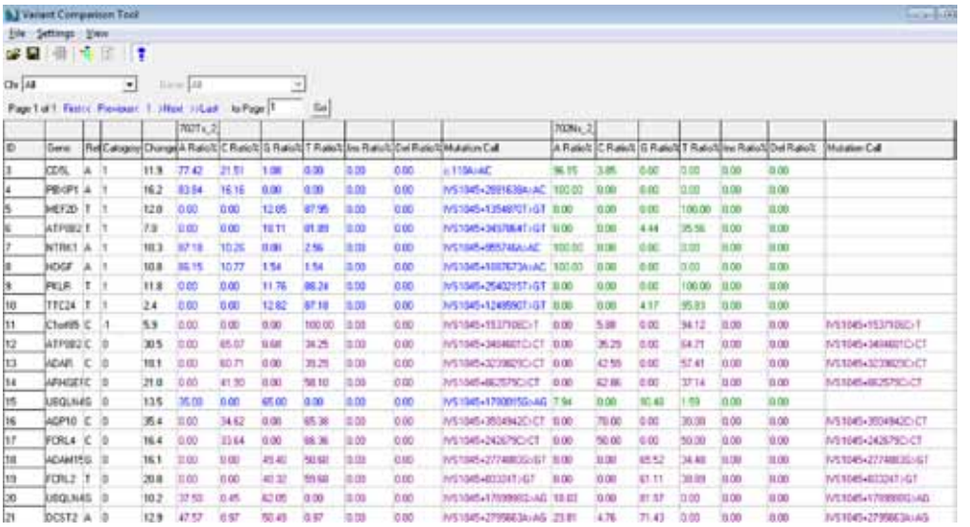


***AmpliSeq™ Comprehensive Cancer Panel** results for chrX, allowing for detection of duplications and deletions. The control sample was female, so the male sample (red) appears to be a deletion (1/2 the normalized coverage). The XO sample appears to have a deletion in the beginning of chrX.*

Mining Tool for Somatic Mutations and Loss of Heterozygosity

NextGENe software includes a quick and accurate method to detect somatic mutations by comparing mutant and normal sequences, such as those from tumor and blood samples. This comparison makes it possible to select a few dozen mutations for validation by Sanger sequencing with less than 20 minutes of review after alignment and mutation calling have been performed on the individual projects.

Detecting somatic mutations using next-generation sequencing is difficult because of high background noise. Background noise is created by the sequencing process. It can also be introduced at other steps, such as contamination of a tumor sample by healthy cells and (much more rarely) contamination of normal cells (such as a blood sample) by tumor cells. NextGENe's variant comparison tool has advanced filtering options that make it easy to remove many false positive and biologically irrelevant mutations. It now provides a one-click ranking function which ranks up to 100 low frequency mutations in order of potential importance.

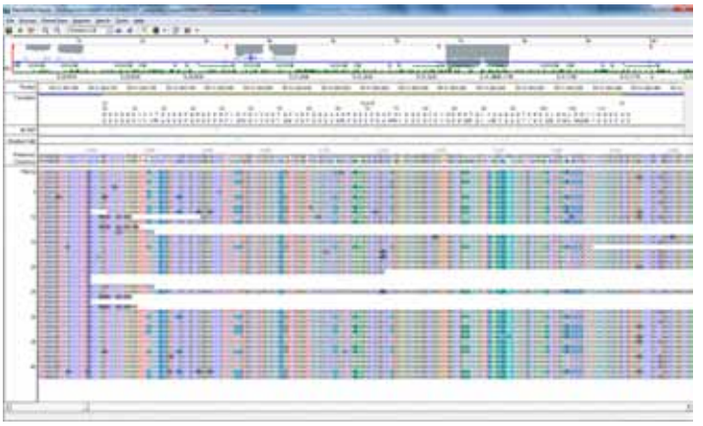


Comparison of tumor and normal samples in NextGENe software Variant Comparison Tool, after sorting with the Top List function.

HLA Applications

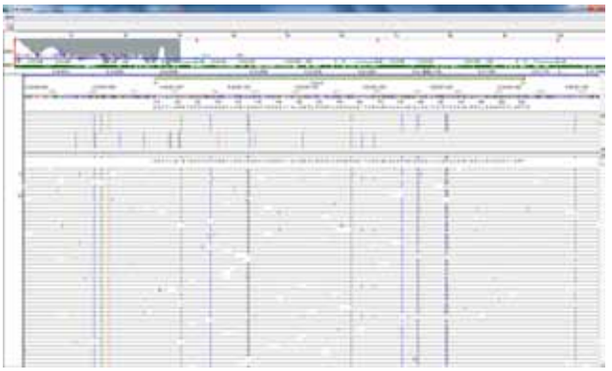
HLA typing using next-generation sequencing allows for resolving multiple alleles because of the high depth of coverage. When using Sanger sequencing, multiple alleles are combined into one signal and minor alleles (due to PCR amplification bias or indel frame shift) are harder to detect. NGS provides a more distinct signal - each read is separate from the others. This allows for detection even in cases of strong PCR bias and insertion and deletion. It is possible to determine HLA genotypes across all exons of an HLA gene using whole gene amplification via long-range PCR. Currently most next-generation sequencers are capable of producing reads that are several hundred bases long, which allows for coverage of exon 2 and 3 from MHC Class I genes without issue.

NextGENe software processes HLA data by first aligning reads to a reference of selected HLA genes. For each exon the software generates the two most common consensus sequences and sorts aligned reads into three categories based on them- allele one, allele two, and "other". The consensus sequences are then compared to an HLA library to determine to most likely HLA genotypes according to amino acid differences, silent SNP changes, and then intron changes.

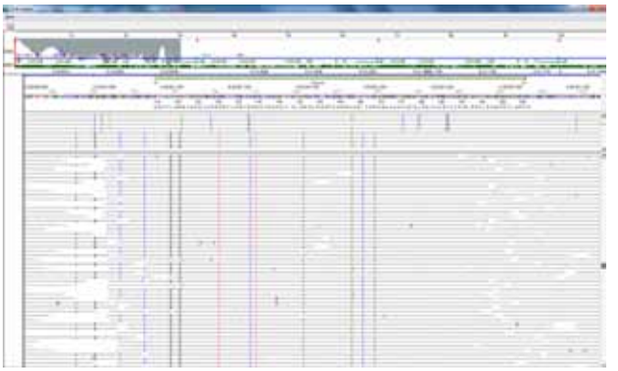


Raw data aligned to reference, grey shadow at top indicates depth of coverage

Best match Genotype

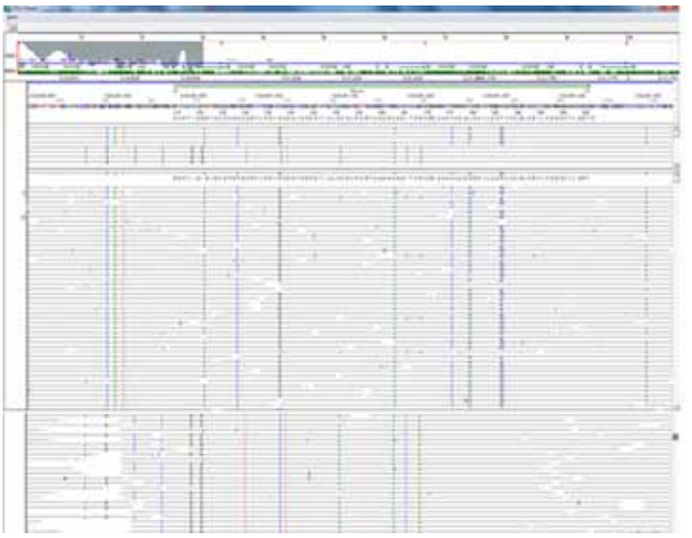


Aligned reads of Allele 1



Best match Genotype

Aligned reads of Allele 2



Combined view of Alleles 1 and 2

The top panel shows the coverage of the HLA genes. Blue cross indicates the location of the sequence text in the 2nd panel. 3rd panel shows HLA genotyping calls for both alleles. Panel 4 and 5 shows the reads in two alleles.

Analysis of Targeted Sequencing Panels

NextGENe software is able to very rapidly process samples from Ion Torrent™, Roche, SOLiD System and Illumina platforms in order to find potentially important mutations in AmpliSeq™, HaloPlex™, Multiplicom™, Roche GS G Type Assay and all other targeted Amplicon Panels. NextGENe software includes sorting and trimming tools, alignment, and mutation calling on low cost Window PC's.

NextGENe software is also able to annotate the mutations that were found using dbSNP, the dbNSFP database, the COSMIC database, or custom databases. Alignments and variant calling is typically accomplished in minutes, with pre-alignment processing such as bar code sorting taking only a few minutes. All of these steps can be fully automated in order to make processing samples even faster and easier.

Results are shown in NextGENe's comprehensive viewer which also allows comparison of up to 20 individual samples and prediction filtering.

A 12 bp deletion in TET2 detected at approximately 41% frequency from Roche GS G Type TET2/CBL/ KRAS Panel. Data courtesy of 454 Life Sciences.

Comparison of two Illumina HaloPlex™ datasets. Data courtesy of Agilent Technologies

Variant Comparison and Prediction Tool

Advanced Comparison between multiple projects 4 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
- Shows shared or differences between all individuals

View multiple projects side-by-side

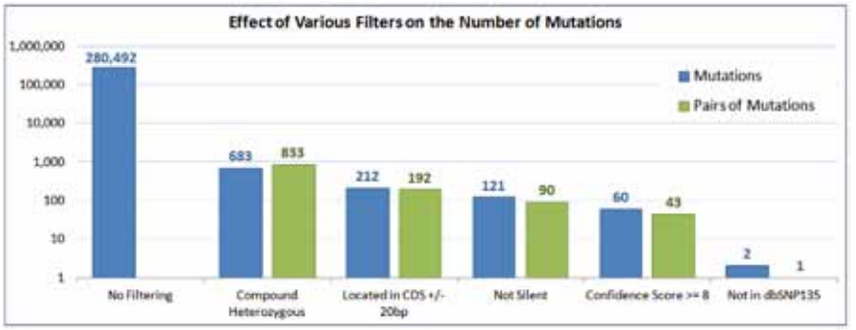
Many advanced filtering options

- Annotation (mRNA, CDS, Splice sites)
- Mutation Confidence Score
- dbSNP mutations
- Substitutions (Silent, Mis-sense, Nonsense) or Indels
- Advanced ROI filtering

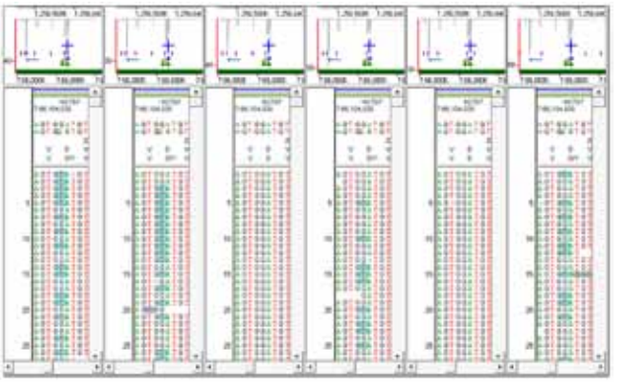
Prediction database integration

- dbNSFP, includes 1000 genomes frequency, PhyloP, PolyPhen-2, Mutation Taster, SIFT
- COSMIC
- Custom, allows import of proprietary or other public databases

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.



Number of Mutations Left After Each Filtering Step Within NextGENe software



One of two causative mutations found as compared across all six projects. The projects (left to right) are the two affected children, the mother, the father, and the two unaffected children. The second unaffected child has this mutation but does not have the other mutation in the same gene.

Analysis of Paired End data

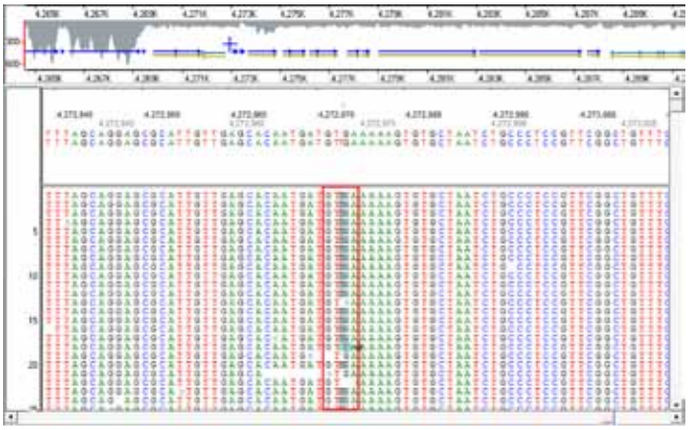
Unique technologies improve sequence analysis accuracy.

Analysis of Ion Torrent Paired End Data

The Ion PGM system now provides a paired end sequencing option in which DNA fragments can be sequenced from both directions. Paired end sequences can be merged into a single, high-quality sequence using NextGENe's "Overlap Merger" tool. This process greatly improves accuracy, especially at the 3' end, reducing the need for trimming. This approach makes it possible to improve the accuracy of the Ion Torrent platform to greater than 99.8%.

	314	316	316 Long-Read	318 Long-Read
Aligned Bases Before	34,702,328	421,924,059	443,882,848	1,247,900,768
Aligned Bases After	35,118,735	427,449,703	463,832,320	1,281,760,871
Error Rate Before	0.6463%	0.5169%	1.4603%	2.0155%
Error Rate After	0.2319%	0.2207%	0.3985%	0.7005%

Read Length and Accuracy Improvements of Ion PGM Paired End data

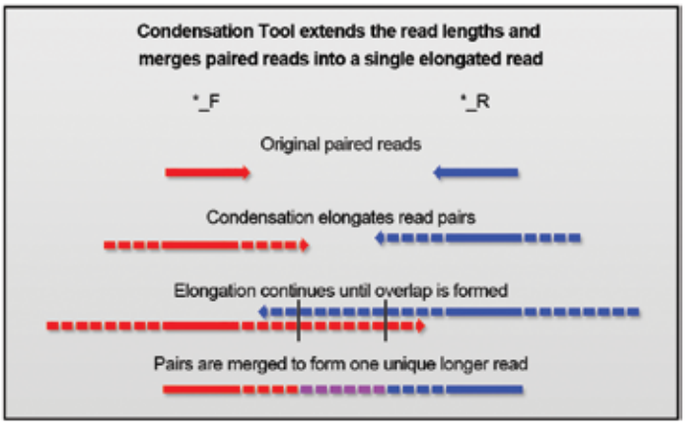


Alignment results for a merged 316 dataset displayed in the NextGENe Viewer. The highlighted position contains an insertion relative to the reference.

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).



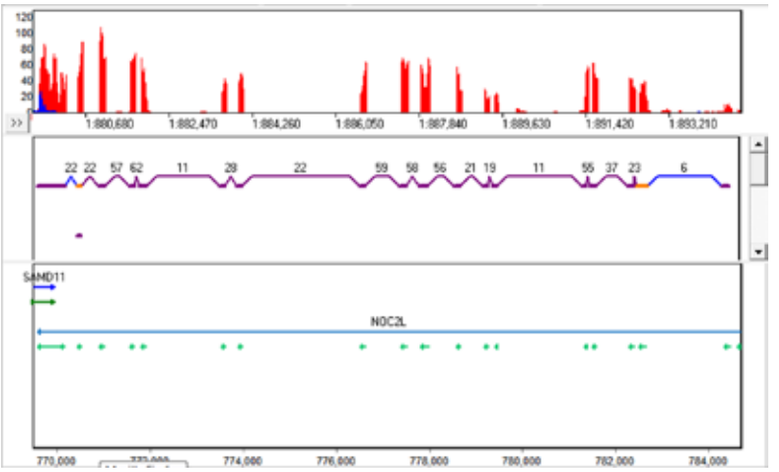
The Paired End linking application works by elongating paired reads to the point that there is an overlap between the pairs. This allows two paired reads to be merged together to form one continuous longer read. 75bp reads from a 200 bp library, for example, in a single elongation cycle will overlap 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.

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

Due to reference sequence difficulties associated with alternative splicing and fusion genes, alignment of RNA-seq data is more challenging than alignment of DNA sequences. Short reads - especially those that fall within large exons - are able to align normally since they will generally match the reference with very few mismatches. Reads that span an exon-exon junction are more difficult because they must be split at the correct position and each part of the read must align correctly. Fusion genes provide even more of a challenge because the partial reads can align almost anywhere in the genome.

Different solutions to these challenges have been implemented in various software packages. Q-PALMA uses a machine learning algorithm and training datasets in order to identify splice junctions [1]. SuperSplat divides sequence reads at multiple positions and tries to find mapping sites where the sub-reads are separated by an intron in a certain size range[2]. TopHat is a software package that first finds potential exons based on coverage and then finds splice sites and links using canonical splice site sequence information [3]. NextGENe uses a novel algorithm to correctly align reads belonging to annotated and novel transcripts while providing the added benefit of a highly graphical interface that doesn't require use of scripting or the command line. Analysis can be performed on a desktop PC in just a few hours without any training datasets or pre-filtering of the reads.



The new RNA-Seq Transcript View. The purple links and exons are known, the blue links are novel, and the orange exons are alternative splicing. The figure of Ion Torrent 318 chip RNA-Seq data set, available on Ion Torrent Development community website.

1. Fabio De Bona et al., "Optimal spliced alignments of short sequence reads," Bioinformatics 24, no. 16 (2008): i174 -i180.
2. Douglas W. Bryant et al., "Supersplat --spliced RNA-seq alignment," Bioinformatics 26, no. 12 (June 15, 2010): 1500 -1505.
3. Cole Trapnell, Lior Pachter, and Steven L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," Bioinformatics 25, no. 9 (May 1, 2009): 1105 -1111.

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

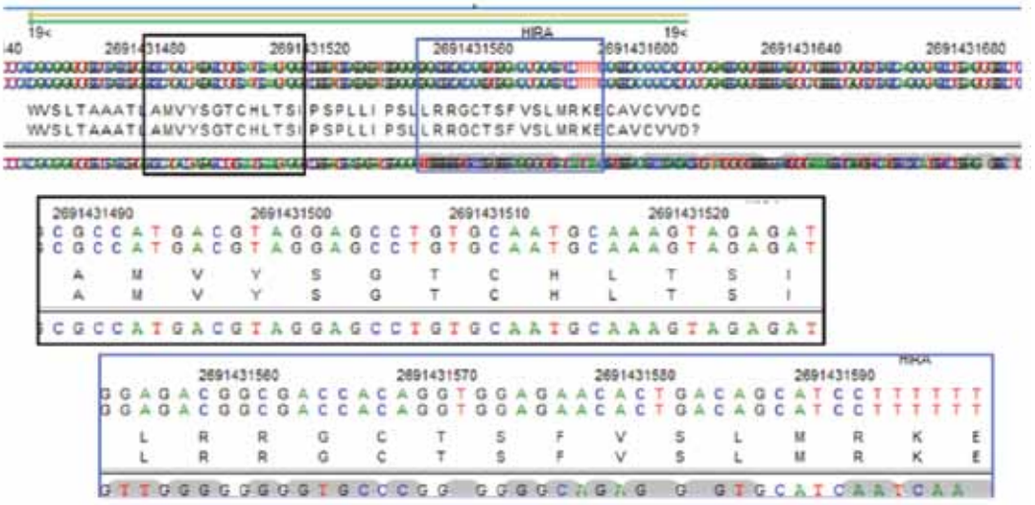
Structural variants (SVs) include insertions, deletions, inversions, and gene fusions that frequently occur across the human genome with over 1000 segmental deletions and over 200 copy number polymorphisms having been reported. These genomic structures have been shown to be important in a number of diseases,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

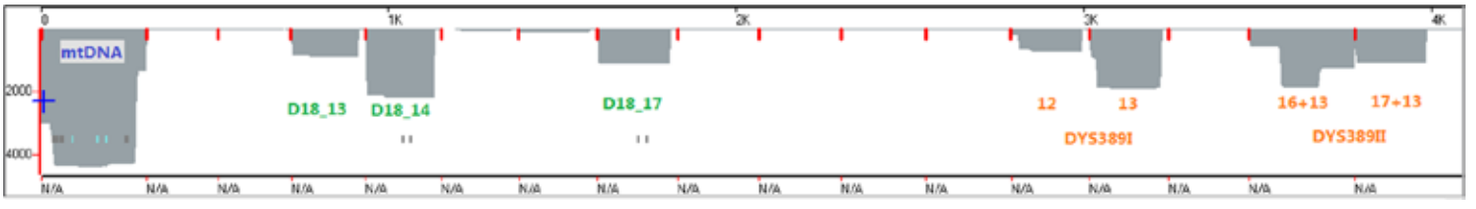
NextGENe then generates pseudo-paired reads for the sequences aligned to these regions by breaking the original reads into pairs. These can be aligned to the reference genome in order to map the SVs, as seen in figure. Detailed information on where these reads align is available in the Paired Read Reports.

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 would be created and split with each half aligning to their appropriate gene. The paired read report identifies the location of each half.



STR Human Identity Analysis

There is tremendous potential for the use of 2nd generation sequencing in forensic analysis. It promises to make analysis faster and cheaper because electrophoresis is no longer necessary and because samples can be barcoded with multiplex identifier (MID) tags and then combined to be analyzed at the same time. It also increases the amount of information available. Comparing sequence data rather than electrophoresis results allows for the comparison of Single Nucleotide Polymorphisms (SNPs) between samples.



NextGENe is an incredibly useful tool for working with next generation sequencing for forensic analysis. It is able to accurately align thousands of reads in seconds. The built-in barcode sorting tool makes it easy to work with multiplexed samples. SNPs are conveniently displayed in the mutation report while STR polymorphisms are counted and the results are displayed in the expression report.

de novo Assembly

NextGENe software includes several options to assist in creating fast and accurate assemblies. These include traditional assemblers and several new technologies detailed below.

Floton™ Assembler

Exclusively for Ion Torrent and Roche technology
Reduces homopolymer errors to substitutions, greatly improving assembly capability
Fast, Accurate assemblies

Ion Torrent systems are fundamentally different from most other sequencing systems. The Ion PGM and Proton systems use a “post-light” technology because it doesn’t depend on detection of light emission from nucleotide incorporation. Instead, it uses a silicon chip containing millions of individual pH meters. Its flow-based approach detects pH changes caused by release of hydrogen ion during incorporation of unmodified nucleotides in DNA replication. Because of this different approach to sequencing, the instrument and reagents are much less expensive and it has a unique error profile- most errors are indels rather than substitutions, especially in homopolymer regions.

These errors are more problematic for assembly than substitution errors because of the increased complexity of gapped comparison. However, NextGENe software now includes the newly developed Floton assembler, which is able to treat these homopolymer errors as substitution errors. In doing so, it is possible to correct the errors during assembly. This method condenses the sequence into flow calls of individual bases and the number of bases in each flow (see figure). By converting the sequence data into this format, the indels are essentially converted into substitution errors (different base count numbers), allowing for faster computation time and correction of most homopolymer errors. When adjusting the index size, it is important to note that the index size is based on a number of flows, rather than a number of bp.

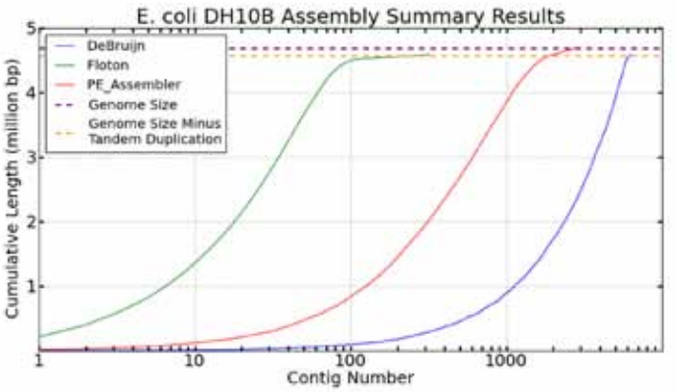
Conversion of base calls into flow calls. In this example a flow size of 9 corresponds to a 17 bp sequence.

Assembly of STO-409 (316 chip) E. coli DH10B

Quality-filtered data:
- Average Coverage = 18x
- Reads = 2,068,174
- Length (Average) = 113.8
- Length (Range) = 25 to 275

Results:

	Floton	PE Assembler	DeBruijn
Contigs	320	2,825	6,309
Max Contig Length	212,035	14,571	9,188
Average Contig Length	14,339	1,662	727
N50	64,441	3,750	1,230



Comparison of 3 different assemblers in NextGENe software

The NextGENe software's Floton Assembler is specially designed to handle data from flow-based sequencing technologies that tend to have indel errors rather than substitution errors. When paired with the Ion Torrent or Roche systems it provides a very fast and inexpensive method for *de novo* sequencing of bacterial and other small genomes. It is designed to run on relatively inexpensive hardware- these projects were run on a typical laptop - and to be easy to use. Often the default settings do not need to be adjusted at all, and in this case only the index size and minimum contig length were adjusted.

Assembly of Illumina MiSeq™ data

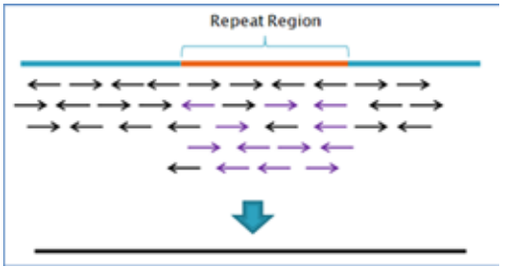
NextGENe software Stepwise Assembler

Sequences that repeat throughout the genome can pose a problem for the assembly of short reads. Normally the repeat reads would assemble within the incorrect contig. This causes two problems - the repeat regions elsewhere in the genome have reduced coverage and contigs terminate at repeat regions due to the formation of multiple ambiguous contigs.

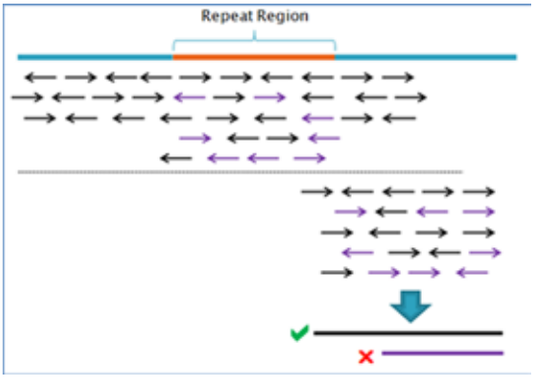
NextGENe software solves this problem with a stepwise paired-end assembly. The end of a contig produced by assembly may indicate a repeat region. The software first calls in the reads paired to those assembled (using overlaps) at the end of the contig (up to 1 ½ times the library size from the end). These reads are then assembled and the software chooses the most complete assembly to continue the contig. Shorter assemblies are not used because they are formed by repetitive sequences erroneously assembled together due to their similarity.

Processing Time	1 hour, 48 min
Total Reads	7,267,665
Reads Used	7,101,508
% Reads Used	97.7%
Number of Contigs	65
Average Length	70,747 bp
Minimum Length	363
Maximum Length	494,530 bp
N50	212,730 bp

E. coli data courtesy Illumina Inc.



Incorrect assembly of reads in a repeat region. Purple reads are from the same repeat sequence elsewhere in the genome.



NextGENe calls in the paired reads and assembles them. Reads paired to those that do not belong in the original assembly will form a separate contig (purple). Only the longest contig is used in the assembly.

Assembly Condensation® Tool

NextGENe now includes a tool to assist in rapid and accurate assemblies of data from all NGS systems. This condensation based tool first scans the short reads removing the excess coverage including repeat regions, retaining longer high-quality reads. This is followed by trimming low quality bases from 3'sequences and finally removes low frequency reads prior to commencing with assembly operation.

Benefits:

Removes excess coverage (including in repeat regions)

- Longer, higher-quality reads are retained
- Assembly speed and accuracy are improved
- Less RAM is required
- Works without a reference
- Trims low frequency 3' sequence caused by sequencing errors

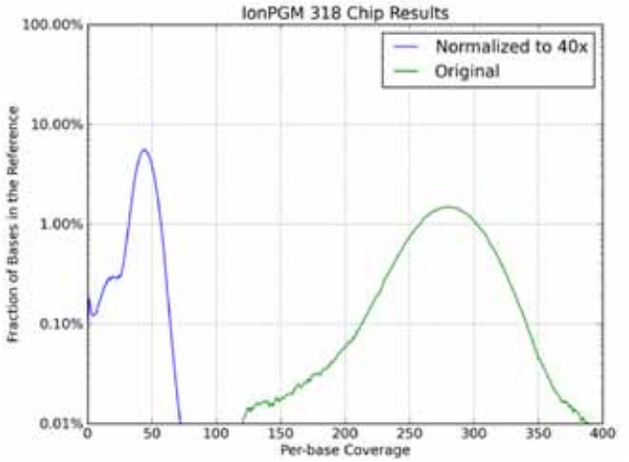
Operation:

1. Calculate "flow-mer" frequencies (32 flows each)
2. Remove reads from high-coverage regions. Reads are kept randomly based on the product of three fractions:
   - Normalization Factor (Desired Coverage / Estimated Coverage)
   - Quality adjustment
   - Read Length adjustment
3. Trim low frequency (< 1/10 of desired coverage) 3' flow-mers and remove low frequency reads

	Before	After	Difference
Reads	5,973,581	912,010	-84.7%
Bases	1,323,802,002	199,676,721	-84.9%

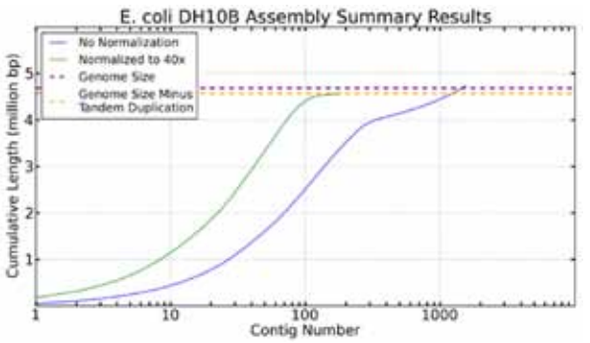
Results:

E coli DH10B sequence run on a 318 chip (C23-140)



Coverage distribution before and after.

Data courtesy of Ion Community



	No Condensation	With Condensation	Difference
Contigs	1,555	182	-88.3%
Maximum Length	54,303	177,056	3.3x
Average Length	3,051	25,084	8.2x
N50	16,290	57,186	3.5x
Process Time	69 min, 22 seconds	27 min, 26 seconds	-60%

Assembly Results.

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 software whole genome alignment method is the first to align reads from the Roche Genome Sequencer FLX System, which often contains 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 Torrent™ and Illumina® Platform data. Additionally, NextGENe software 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 Torrent PGM and Proton 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.

Digital Gene Expression Studies, CNV, ChIPSeq & miRNA Analysis

Align to entire Genome or to specific references
Indentifies binding sites and transcription sites
Reports sequences, expression levels and information for each identified peak
Available comparison report compares multiple individuals or time based analysis
Removes Duplicate Reads
Expression Reporting
Search Tool
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 Ion Torrent Proton, Illumina platforms 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.

Analysis of Pooled Samples using Barcode/index tags

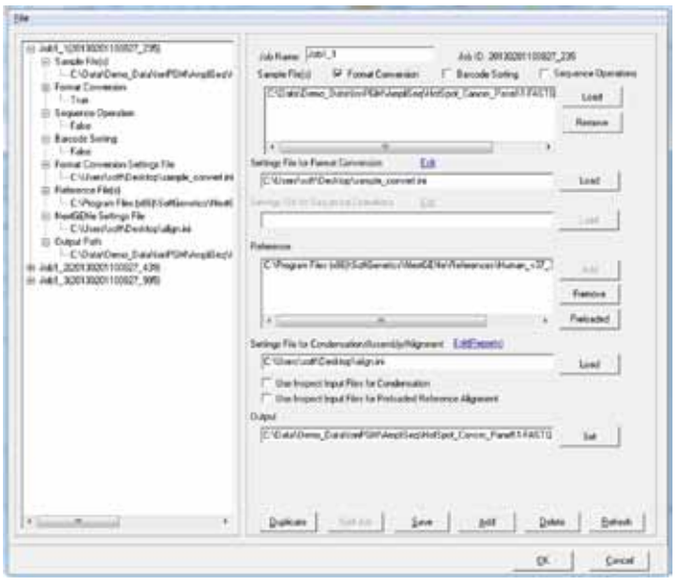
Automatically parses samples according to tags
Flexible Tool selects bar codes based on stringent or loose fit criteria
Works effectively with all major methodologies

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.

NextGENe Pipeline Automation

- Seeks Data Availability
- Performs Format Conversion
- De-multiplex bar coded data sets
- Obtain Reference or Disease Panel(s)
- Performs Analysis
  - Re-Sequencing
  - Amplicon SNP/Indel
  - De novo Assembly
  - RNA-Seq
- Creates Custom Reports
- Logs and Saves project file



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.

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 software Trial
User Manual

Minimum Computer Recommendations:
Desktop PC
64 bit, Windows® XP, Vista, 7 or 8 Operating System, Windows® Server 2008 R2
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, 7 or 8 Operating System, Windows® Server 2008 R2
Processor: Dual Quad Core (2.4 GHz)
RAM: 12GB DDR2-800MH
2TB Hard Drive

Note: Some applications will require additional memory

If trial disc is not present please email info@softgnetics.com
for a free 30-day trial

SOFTGENETICS®
Software PowerTools for Genetics Analysis

www.softgenetics.com

SOFTGENETICS®

Software PowerTools for Genetics Analysis

SoftGenetics
Oakwood Centre
100 Oakwood Avenue
Suite 350
State College PA 16803 USA
info@softgenetics.com
www.softgenetics.com

For Clinical Research