SOFTGENETICS®

NextGENe®

Next Generation Sequencing Software for Biologists

Application Modules for:

Variant Analysis, Targeted, WES, WGS
- SNPs
- Indels
- Structural Variations
- Somatic Mutation Mining

Copy Number Variation (CNV Analysis)

Non-Invasive Prenatal Testing (NIPT)

HLA Analysis

Patient Comparison
- Groups
- Families
- Trios

Functional Prediction Calling

RNA-Seq/Transcriptome Analysis

Digital Gene Expression

miRNA Discovery and Quantification

Forensic Human Identity
- STR
- Mitochondrial

Assembly
- *de novo*
- Paired end

Compatible with:

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

SOFTGENETICS®
Software PowerTools for Genetic 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
Track Manager

- ✓ Functional Prediction
- ✓ Disease Association
- ✓ Conservation Scores
- ✓ Population Frequencies

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 Interpretive Workbench

Low-Cost Hardware Requirements

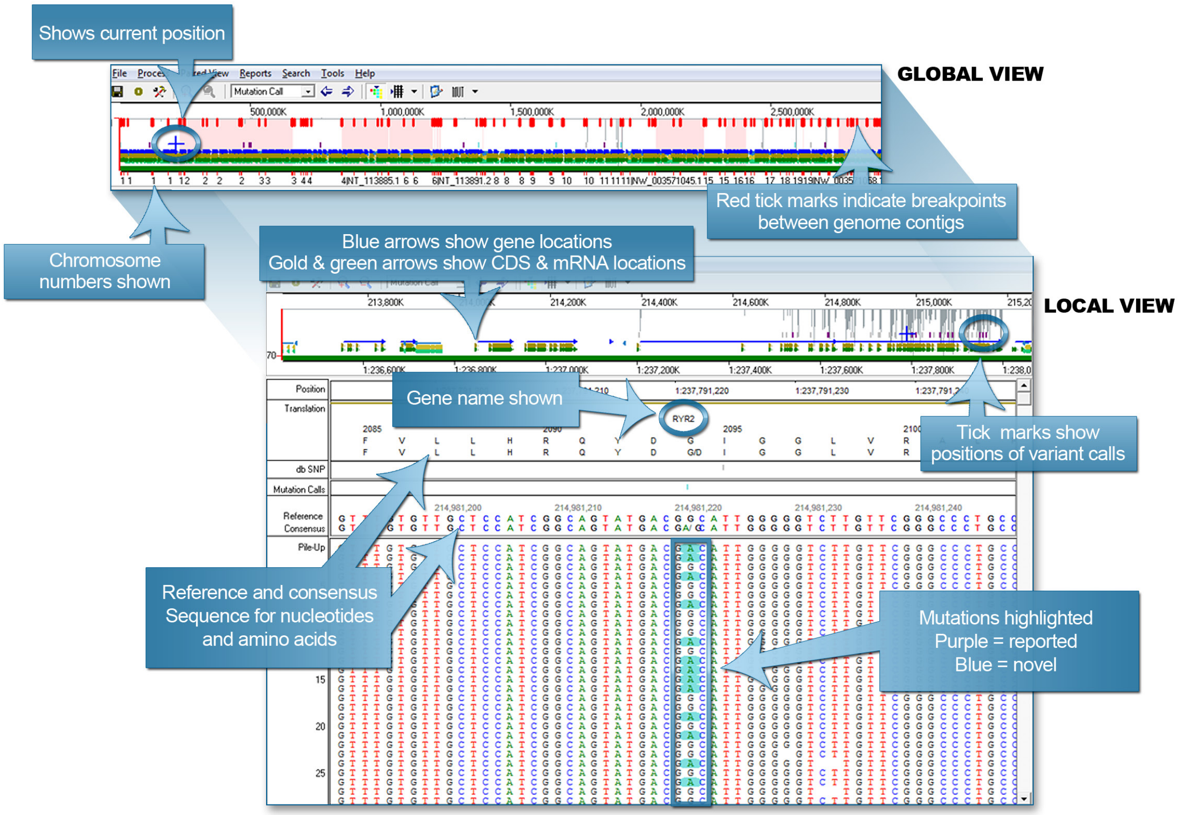
NextGENe Software is a complete, "free-standing" analysis package designed for use by biologists in the analysis of data from Next Generation Sequencing systems. The icon driven, easy-to-use Windows® interface significantly reduces bioinformatics requirements, provides annotated analysis review, while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

Features

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.



GLOBAL VIEW

Shows current position

Chromosome numbers shown

Red tick marks indicate breakpoints between genome contigs

Blue arrows show gene locations
Gold & green arrows show CDS & mRNA locations

LOCAL VIEW

Gene name shown

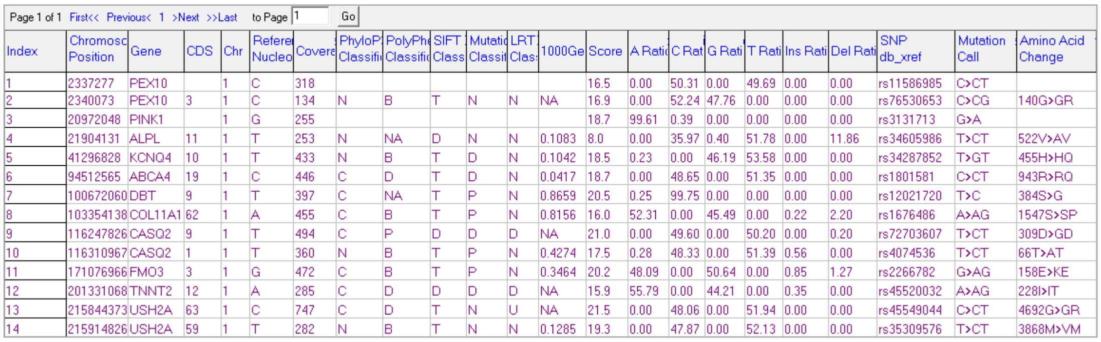
Tick marks show positions of variant calls

Reference and consensus Sequence for nucleotides and amino acids

Mutations highlighted
Purple = reported
Blue = novel

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 databases including dbSNP and COSMIC. Several filtering options are available to speed and ease analysis review:

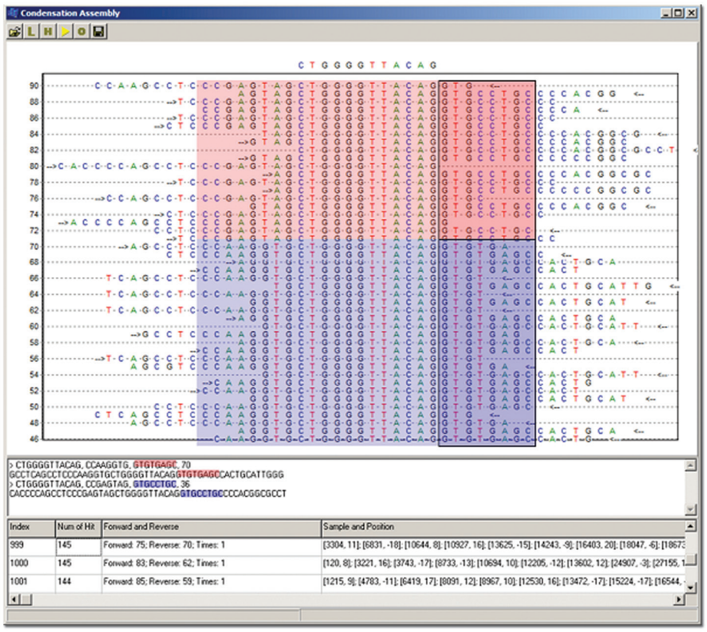


Page 1 of 1 First<< Previous< 1 >Next >>Last to Page 1 Go

Index	Chromosc Position	Gene	CDS	Chr	Referei Nucleo	Covera	PhyloP Classifi	PolyPhe Classific	SIFT Class	Mutatic Classit	LRT Clas:	1000Ge	Score	A Ratic	C Rat	G Rati	T Rati	Ins Rati	Del Rati	SNP db_xref	Mutation Call	Amino Acid Change
1	2337277	PEX10		1	C	318							16.5	0.00	50.31	0.00	49.69	0.00	0.00	rs11586985	C>CT	
2	2340073	PEX10	3	1	C	134	N	B	T	N	N	NA	16.9	0.00	52.24	47.76	0.00	0.00	0.00	rs76530653	C>CG	140G>GR
3	20972048	PINK1		1	G	255							18.7	99.61	0.39	0.00	0.00	0.00	0.00	rs3131713	G>A	
4	21904131	ALPL	11	1	T	253	N	NA	D	N	N	0.1083	8.0	0.00	35.97	0.40	51.78	0.00	11.86	rs34605986	T>CT	522V>AV
5	41296828	KCNQ4	10	1	T	433	N	B	T	D	N	0.1042	18.5	0.23	0.00	46.19	53.58	0.00	0.00	rs34287852	T>GT	455H>HQ
6	94512565	ABCA4	19	1	C	446	C	D	T	D	N	0.0417	18.7	0.00	48.65	0.00	51.35	0.00	0.00	rs1801581	C>CT	943R>RQ
7	100672060	DBT	9	1	T	397	C	NA	T	P	N	0.8659	20.5	0.25	99.75	0.00	0.00	0.00	0.00	rs12021720	T>C	384S>G
8	103354138	COL11A1	62	1	A	455	C	B	T	P	N	0.8156	16.0	52.31	0.00	45.49	0.00	0.22	2.20	rs1676486	A>AG	1547S>SP
9	116247826	CASQ2	9	1	T	494	C	P	D	D	D	NA	21.0	0.00	49.60	0.00	50.20	0.00	0.20	rs72703607	T>CT	309D>GD
10	116310967	CASQ2	1	1	T	360	N	B	T	P	N	0.4274	17.5	0.28	48.33	0.00	51.39	0.56	0.00	rs4074536	T>CT	66T>AT
11	171076966	FMO3	3	1	G	472	C	B	T	P	N	0.3464	20.2	48.09	0.00	50.64	0.00	0.85	1.27	rs2266782	G>AG	158E>KE
12	201331068	TNNT2	12	1	A	285	C	D	D	D	D	NA	15.9	55.79	0.00	44.21	0.00	0.35	0.00	rs45520032	A>AG	228I>IT
13	215844373	USH2A	63	1	C	747	C	D	T	N	U	NA	21.5	0.00	48.06	0.00	51.94	0.00	0.00	rs45549044	C>CT	4692G>GR
14	215914826	USH2A	59	1	T	282	N	B	T	N	N	0.1285	19.3	0.00	47.87	0.00	52.13	0.00	0.00	rs35309576	T>CT	3868M>VM

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**



```
> CTGGGGTTACAG, CCAAGGTG, GTGTGAGC, 70
GCCTCAGCCTCCCAAGGTGCTGGGGTTACAGGTGTGAGCCACTGCATTGGG
> CTGGGGTTACAG, CCGAGTAG, GTGCCTGC, 36
CACCCCAGCCTCCCGAGTAGCTGGGGTTACAGGTGCCTGCCCCACGGCGCCT
```

Index	Num of Hit	Forward and Reverse	Sample and Position
999	145	Forward: 75; Reverse: 70; Times: 1	[3304, 11]; [6831, -18]; [10644, 8]; [10927, 16]; [13625, -15]; [14243, -9]; [16403, 20]; [18047, -6]; [18673
1000	145	Forward: 83; Reverse: 62; Times: 1	[120, 8]; [3221, 16]; [3743, -17]; [8733, -13]; [10694, 10]; [12205, -12]; [13602, 12]; [24907, -3]; [27155, 1
1001	144	Forward: 85; Reverse: 59; Times: 1	[1215, 9]; [4783, -11]; [6419, 17]; [8091, 12]; [8967, 10]; [12530, 16]; [13472, -17]; [15224, -17]; [16544, -

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 of alignment and variant detection on various fasta data sets:

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

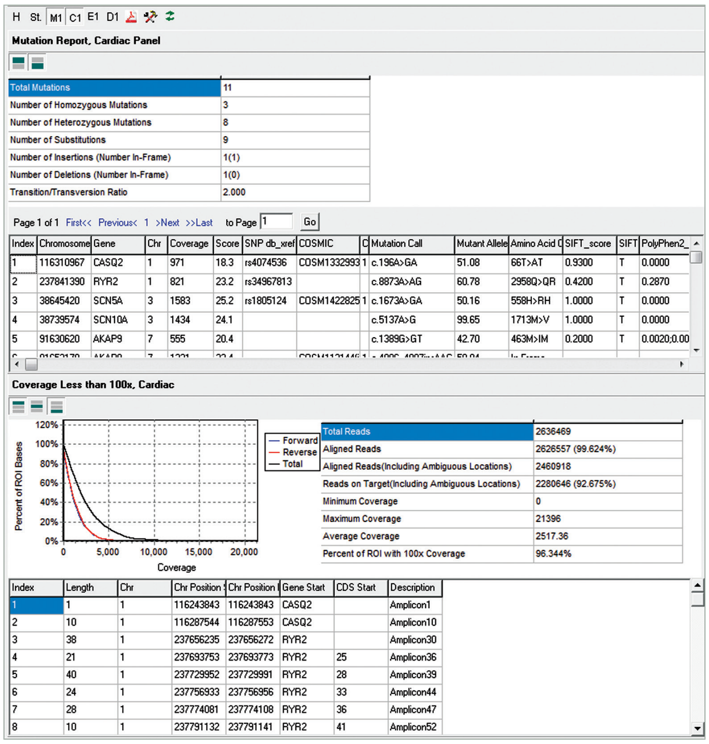
Reporting of Results and Quality Control Metrics

- **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.

NextGENe software's custom report builder allows users to create reports that concentrate only on the information required by their application or institution. In the above example the report contains total found mutations, type, chromosome position, gene, coverage levels by position, confidence score, dbSNP ID, COSMIC reference, mutation call, mutant allele %, amino acid change, SIFT and PolyPhen2 scores, as well as run quality data including read balance, total and total aligned reads, reads on target, percent of ROI covered and more.

Customized Clinical Research Report



Mutation Report, Cardiac Panel

Total Mutations	11
Number of Homozygous Mutations	3
Number of Heterozygous Mutations	8
Number of Substitutions	9
Number of Insertions (Number In-Frame)	1(1)
Number of Deletions (Number In-Frame)	1(0)
Transition/Transversion Ratio	2.000

Index	Chromosome	Gene	Chr	Coverage	Score	SNP db_xref	COSMIC	C	Mutation Call	Mutant Allele	Amino Acid C	SIFT_score	SIFT	PolyPhen2_
1	116310967	CASQ2	1	971	18.3	rs4074536	COSM1332993	1	c.196A>GA	51.08	66T>AT	0.9300	T	0.0000
2	237841390	RYR2	1	821	23.2	rs34967813			c.8873A>AG	60.78	2958Q>QR	0.4200	T	0.2870
3	38645420	SCN5A	3	1583	25.2	rs1805124	COSM1422825	1	c.1673A>GA	50.16	558H>RH	1.0000	T	0.0000
4	38739574	SCN10A	3	1434	24.1				c.5137A>G	99.65	1713M>V	1.0000	T	0.0000
5	91630620	AKAP9	7	555	20.4				c.1389G>GT	42.70	463M>IM	0.2000	T	0.0020;0.00

Coverage Less than 100x, Cardiac

Total Reads	2636469
Aligned Reads	2626557 (99.624%)
Aligned Reads(Including Ambiguous Locations)	2460918
Reads on Target(Including Ambiguous Locations)	2280646 (92.675%)
Minimum Coverage	0
Maximum Coverage	21396
Average Coverage	2517.36
Percent of ROI with 100x Coverage	96.344%

Index	Length	Chr	Chr Position S	Chr Position E	Gene Start	CDS Start	Description
1	1	1	116243843	116243843	CASQ2		Amplicon1
2	10	1	116287544	116287553	CASQ2		Amplicon10
3	38	1	237656235	237656272	RYR2		Amplicon30
4	21	1	237693753	237693773	RYR2	25	Amplicon36
5	40	1	237729952	237729991	RYR2	28	Amplicon39
6	24	1	237756933	237756956	RYR2	33	Amplicon44
7	28	1	237774081	237774108	RYR2	36	Amplicon47
8	10	1	237791132	237791141	RYR2	41	Amplicon52

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.

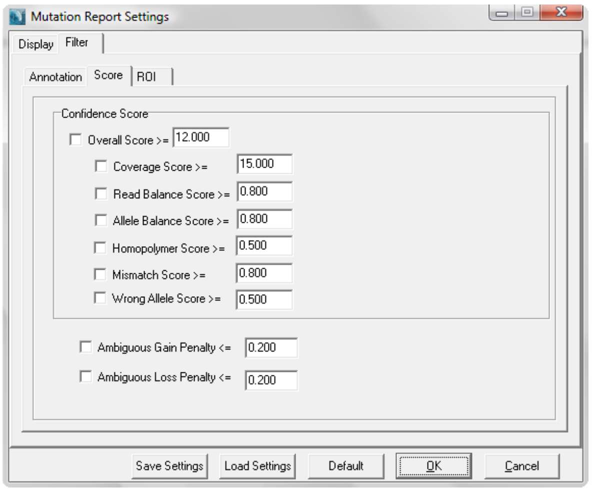
● **Read Balance Score (0 to 1)** – A score of 1 indicates perfect or near perfect balance between the number of forward and reverse reads. Unbalanced data may indicate misalignment or may allow basecalling biases to cause false positives. If the data is expected to be biased (as it is for some targeted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



Confidence Score

□ Overall Score >= 12.000

□ Coverage Score >= 15.000

□ Read Balance Score >= 0.800

□ Allele Balance Score >= 0.800

□ Homopolymer Score >= 0.500

□ Mismatch Score >= 0.800

□ Wrong Allele Score >= 0.500

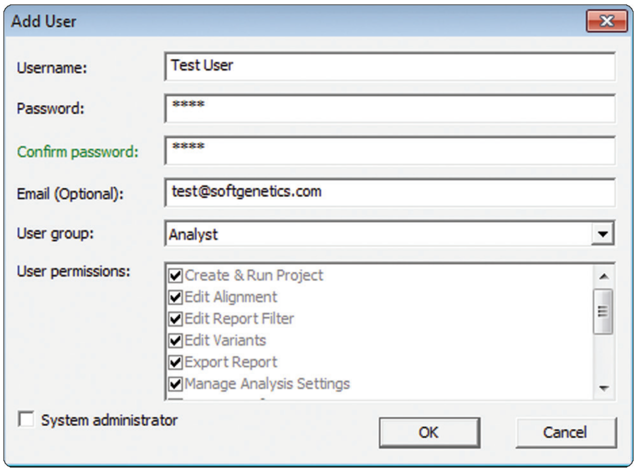
□ Ambiguous Gain Penalty <= 0.200

□ Ambiguous Loss Penalty <= 0.200

User Management

- **Enable to require user login**
- **Create and manage permissions for users and/or groups**
- **Audit trail records username of user who created or modified each project and/or report**

NextGENe's new User Management function provides the ability to control access for different users or groups of users and to track the user who created or modified any project or report. When enabled, user login is required to open NextGENe.



Add User

Username: Test User

Password: ****

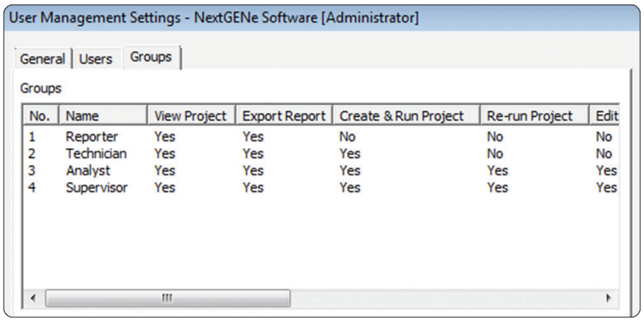
Confirm password: ****

Email (Optional): test@softgenetics.com

User group: Analyst

User permissions:
☑ Create & Run Project
☑ Edit Alignment
☑ Edit Report Filter
☑ Edit Variants
☑ Export Report
☑ Manage Analysis Settings

□ System administrator



User Management Settings - NextGENe Software [Administrator]

Groups

No.	Name	View Project	Export Report	Create & Run Project	Re-run Project	Edit
1	Reporter	Yes	Yes	No	No	No
2	Technician	Yes	Yes	Yes	No	No
3	Analyst	Yes	Yes	Yes	Yes	Yes
4	Supervisor	Yes	Yes	Yes	Yes	Yes

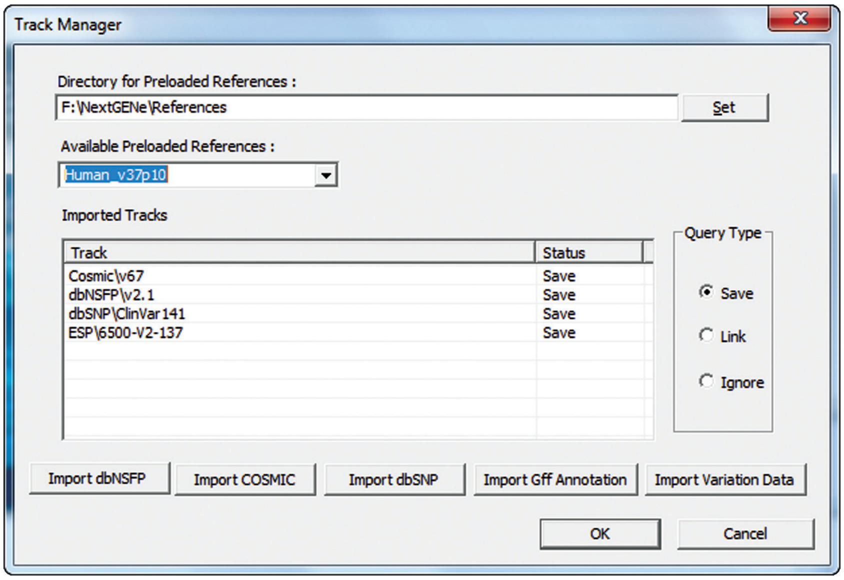
User accounts can be created for each user, with permissions assigned on an individual basis, or groups can be utilized to assign the same privileges to multiple users.

Variant Annotation Tracks and Pathogenicity Calling

- **Functional Prediction information**
  SIFT, PolyPhen-2, LRT, MutationTaster, FATHMM, CADD & MutationAssessor
- **Disease association**
  ClinVar & COSMIC
- **Conservation scores**
  phyloP, GERP++, phastCons & SiPhy
- **Population frequencies**
  1000 Genomes and Exome Variant Server
- **COSMIC - Catalogue of Somatic Mutations in Cancer from Sanger Institute**
- **dbSNP from NCBI**
- **Custom Tracks**

NextGENe's Track Manager Tool can be used to import variant databases as tracks for annotation. Information from imported tracks can be displayed and used for filtering in the Mutation Report. The Track Manager includes specialized support for 3 popular databases: the dbNSFP database, which contains pre-calculated functional prediction and conservation scores along with population frequencies from the 1000 Genomes and Exome Variant Server projects, the COSMIC database of cancer variants from the Sanger Institute and the dbSNP database from NCBI. There is also a custom track import feature that can be used to import other public or proprietary databases.

NextGENe's Track Manager can be used to easily import dbNSFP, COSMIC, dbSNP as well as custom databases as annotation tracks.



Track Manager

Directory for Preloaded References : F:\NextGENe\References

Available Preloaded References : Human_v37p10

Imported Tracks

Track	Status
Cosmic\v67	Save
dbNSFP\v2.1	Save
dbSNP\ClinVar141	Save
ESP\6500-V2-137	Save

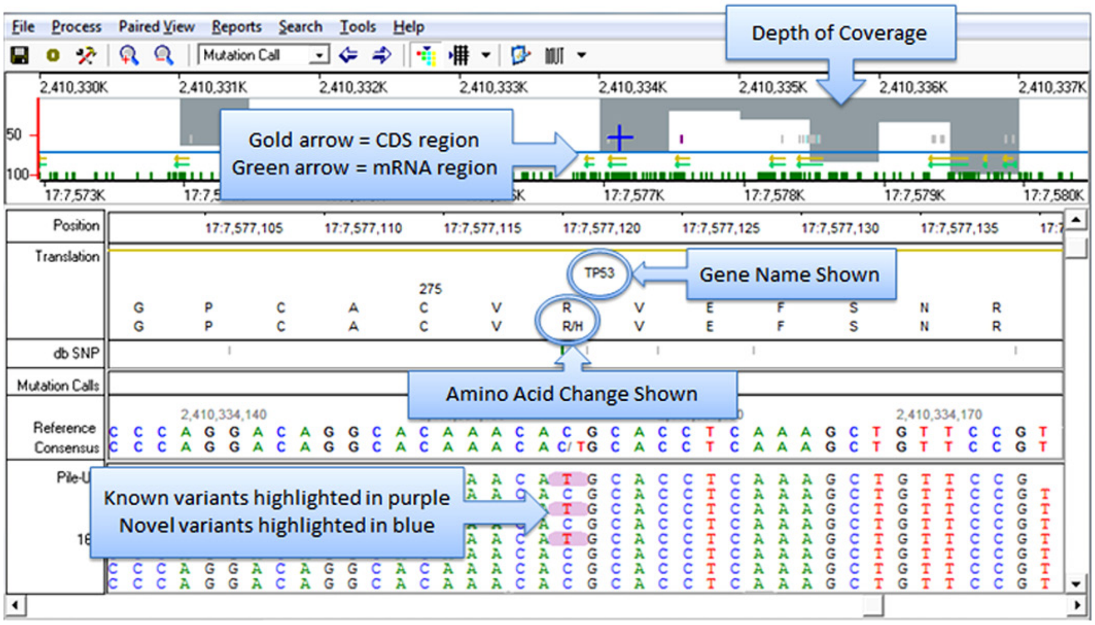
Query Type: ● Save ○ Link ○ Ignore

Applications

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 database 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.



Depth of Coverage

Gold arrow = CDS region
Green arrow = mRNA region

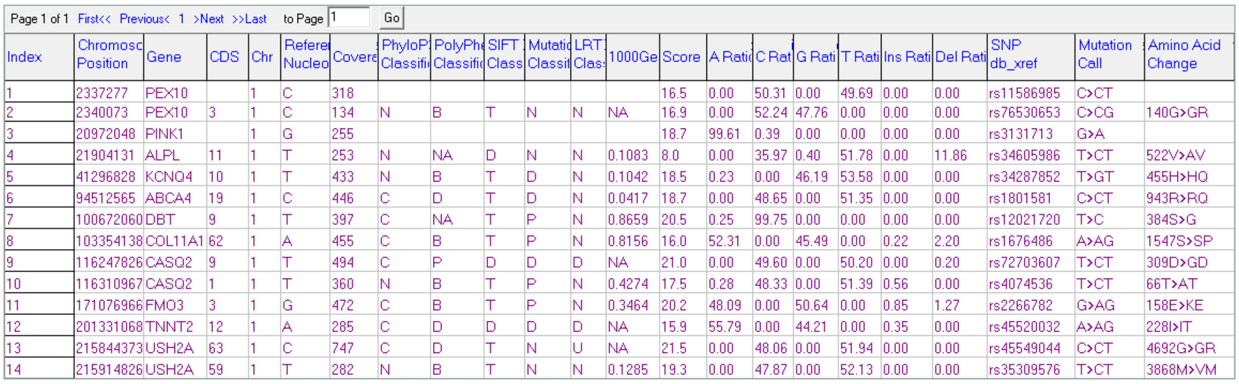
Gene Name Shown

Amino Acid Change Shown

Known variants highlighted in purple
Novel variants highlighted in blue

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.


Page 1 of 1 First<< Previous< 1 >Next >>Last to Page 1 Go

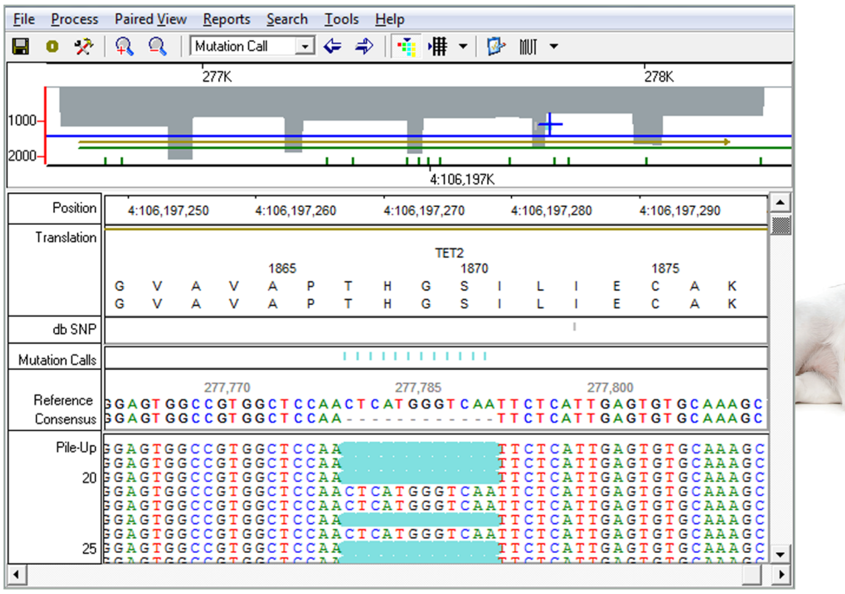
Index	Chromoso Position	Gene	CDS	Chr	Referei Nucleo	Covera	PhyloP Classifi	PolyPhe Classifi	SIFT Class	Mutatic Classit	LRT Clas:	1000Ge	Score	A Ratic	C Rat	G Rati	T Rati	Ins Rati	Del Rati	SNP db_xref	Mutation Call	Amino Acid Change
1	2337277	PEX10		1	C	318							16.5	0.00	50.31	0.00	49.69	0.00	0.00	rs11586985	C>CT	
2	2340073	PEX10	3	1	C	134	N	B	T	N	N	NA	16.9	0.00	52.24	47.76	0.00	0.00	0.00	rs76530653	C>CG	140G>GR
3	20972048	PINK1		1	G	255							18.7	99.61	0.39	0.00	0.00	0.00	0.00	rs3131713	G>A	
4	21904131	ALPL	11	1	T	253	N	NA	D	N	N	0.1083	8.0	0.00	35.97	0.40	51.78	0.00	11.86	rs34605986	T>CT	522V>AV
5	41296828	KCNQ4	10	1	T	433	N	B	T	D	N	0.1042	18.5	0.23	0.00	46.19	53.58	0.00	0.00	rs34287852	T>GT	455H>HQ
6	94512565	ABCA4	19	1	C	446	C	D	T	D	N	0.0417	18.7	0.00	48.65	0.00	51.35	0.00	0.00	rs1801581	C>CT	943R>RQ
7	100672060	DBT	9	1	T	397	C	NA	T	P	N	0.8659	20.5	0.25	99.75	0.00	0.00	0.00	0.00	rs12021720	T>C	384S>G
8	103354138	COL11A1	62	1	A	455	C	B	T	P	N	0.8156	16.0	52.31	0.00	45.49	0.00	0.22	2.20	rs1676486	A>AG	1547S>SP
9	116247826	CASQ2	9	1	T	494	C	P	D	D	D	NA	21.0	0.00	49.60	0.00	50.20	0.00	0.20	rs72703607	T>CT	309D>GD
10	116310967	CASQ2	1	1	T	360	N	B	T	P	N	0.4274	17.5	0.28	48.33	0.00	51.39	0.56	0.00	rs4074536	T>CT	66T>AT
11	171076966	FMO3	3	1	G	472	C	B	T	P	N	0.3464	20.2	48.09	0.00	50.64	0.00	0.85	1.27	rs2266782	G>AG	158E>KE
12	201331068	TNNT2	12	1	A	285	C	D	D	D	D	NA	15.9	55.79	0.00	44.21	0.00	0.35	0.00	rs45520032	A>AG	228I>IT
13	215844373	USH2A	63	1	C	747	C	D	T	N	U	NA	21.5	0.00	48.06	0.00	51.94	0.00	0.00	rs45549044	C>CT	4692G>GR
14	215914826	USH2A	59	1	T	282	N	B	T	N	N	0.1285	19.3	0.00	47.87	0.00	52.13	0.00	0.00	rs35309576	T>CT	3868M>VM

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.

Analysis of Targeted Sequencing Panels

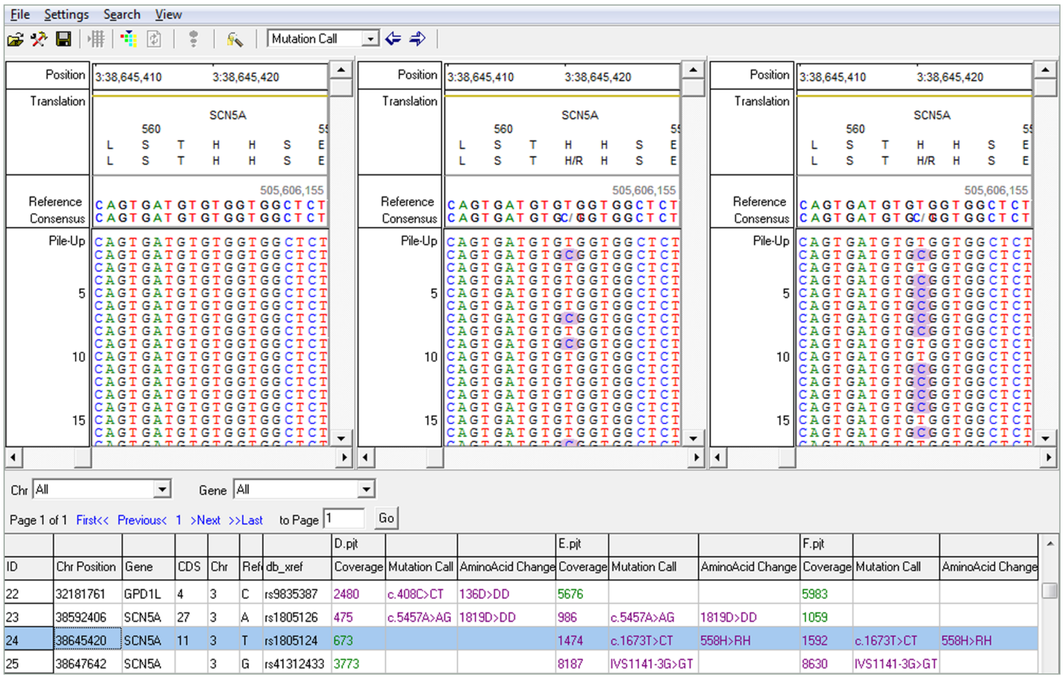
NextGENe software is able to very rapidly process samples from Ion Torrent™, Roche, SOLiD System and Illumina platforms in order to find potentially important mutations in AmpliSeq™, HaloPlex™, Multiplicom™, Roche GS G Type Assay and all other targeted sequencing panels. NextGENe software includes sorting and trimming tools, alignment, and mutation calling on low cost Windows PCs. 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 barcode sorting taking only a few minutes. All of these steps can be fully automated in order to make processing samples even faster and easier.

Trio 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.

Trio Analysis with NextGENe Viewer



							D.pjt			E.pjt			F.pjt		
ID	Chr Position	Gene	CDS	Chr	Ref	db_xref	Coverage	Mutation Call	AminoAcid Change	Coverage	Mutation Call	AminoAcid Change	Coverage	Mutation Call	AminoAcid Change
22	32181761	GPD1L	4	3	C	rs9835387	2480	c.408C>CT	136D>DD	5676			5983		
23	38592406	SCN5A	27	3	A	rs1805126	475	c.5457A>AG	1819D>DD	986	c.5457A>AG	1819D>DD	1059		
24	38645420	SCN5A	11	3	T	rs1805124	673			1474	c.1673T>CT	558H>RH	1592	c.1673T>CT	558H>RH
25	38647642	SCN5A		3	G	rs41312433	3773			8187	IVS1141-3G>GT		8630	IVS1141-3G>GT	

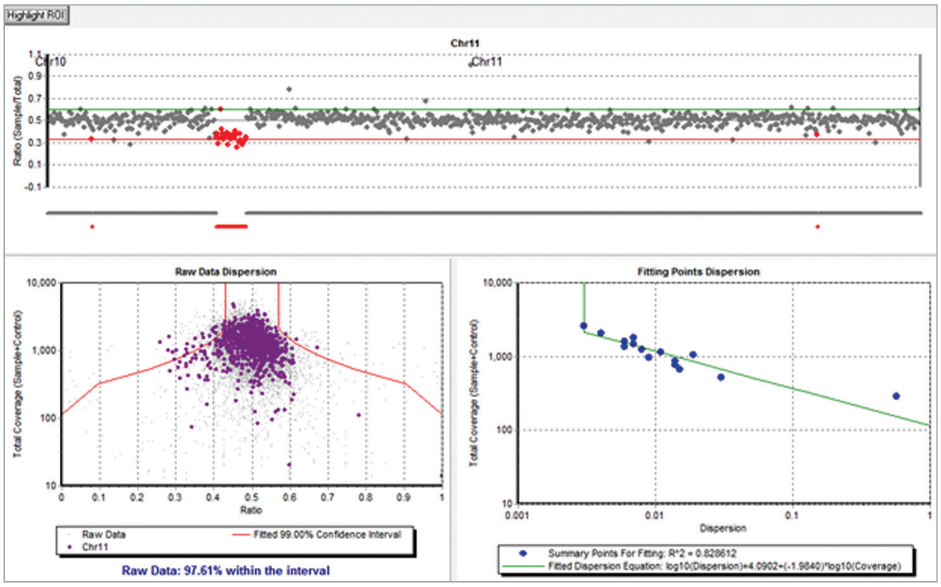
Trio results are shown in NextGENe's comprehensive viewer which also allows comparison of up to 20 individual samples and prediction filtering.

Copy Number Variation (CNV) Tool

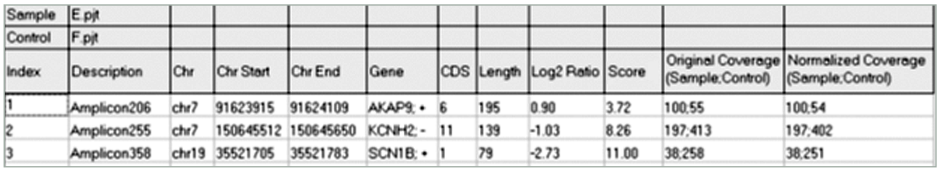
- **Two powerful CNV algorithms available**
- **Provides confidence scores for CNV calls**
- **Classifies regions as potential deletions, duplications, or normal copy number**
- **Processes individual CNV changes (per exon, or per amplicon) into multi-locus calls that can be made more confidently.**

NextGENe includes a new Copy Number Variation (CNV) analysis tool designed to make variant calls on a case-control basis. Two powerful CNV detection algorithms are included, using proprietary technology. The “SNP-Based” method uses a global coverage normalization and uses SNP locations to determine the median coverage for each region in both the sample and control. A log2 ratio is used to compare the median coverage values to make the CNV calls. The “Dispersion and Hidden Markov Model (HMM)” method utilizes coverage ratios between the sample and control, models the dispersion (noise) at varying coverage levels, and uses this information in conjunction with a Hidden Markov Model to make the CNV calls.

The CNV Tool 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.


Highlight ROI
Chr11
Chr10
Chr11
Ratio (Sample/Total)
1.1
0.9
0.7
0.5
0.3
0.1
-0.1
Raw Data Dispersion
Total Coverage (Sample+Control)
10,000
1,000
100
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Ratio
Raw Data
Chr11
Fitted 99.00% Confidence Interval
Raw Data: 97.61% within the interval
Fitting Points Dispersion
Total Coverage (Sample+Control)
10,000
1,000
100
10
0.001
0.01
0.1
1
Dispersion
Summary Points For Fitting: R^2 = 0.828612
Fitted Dispersion Equation: log10(Dispersion)=4.0902+(-1.9840)*log10(Coverage)

Specialized graphics are displayed for CNV analysis. This figure shows the results view with chromosome 11 selected with a deletion indicated in red



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.

Noninvasive Prenatal Testing Tool

NextGENe now includes in version 2.4.0 an easy to use tool for NIPT of chromosomal aneuploidies from maternal plasma that provides excellent sensitivity and specificity. It uses small whole-genome sequencing projects (10 to 20 million reads per project) so that multiple samples can be run on a single lane. Potential trisomy and monosomy in chr13, chr18, and chr21 is reported along with the most likely copy number values for chrX and chrY. This analysis requires a few dozen negative controls (from multiple runs) and some positive controls for testing sensitivity and specificity.

The tool is used in three stages:

1. Optimize the normalization technique and set up baseline values using the negative controls.
2. Optionally test positive and negative controls to calculate sensitivity and specificity. Use this information to fine-tune cutoff values for making calls based on the Normalized Chromosome Values (NCVs).
3. Test unknown samples one batch (run) at a time. Every sample will have results reported for all five tested chromosomes including inconclusive calls if the results were not confident enough.

Bianchi, Diana W., et al. “Genome-wide fetal aneuploidy detection by maternal plasma DNA sequencing.” Obstetrics & Gynecology 119.5 (2012): 890-901.

Somatic Mutation Comparison Tool

A new “Somatic Mutation Comparison Tool” has been added to the NextGENe Viewer tools menu in version 2.4.0. This specialized version of the variant comparison tool allows the user to load three exome projects- a tumor sample, a matched normal sample (such as a blood sample), and a pooled sample consisting of normal data from multiple sources. The tumor and normal projects are used to find somatic variants and the pooled project is used to eliminate artifacts due to library preparation and alignment.

Available settings include:
● **“Maximum Contamination”** – somatic variants may have up to this allele frequency in the normal project. This is to account for contamination of the normal sample with tumor sequence. Default setting is 5%. Any allele ratio greater than 5% in normal will be considered as germanline mutation.

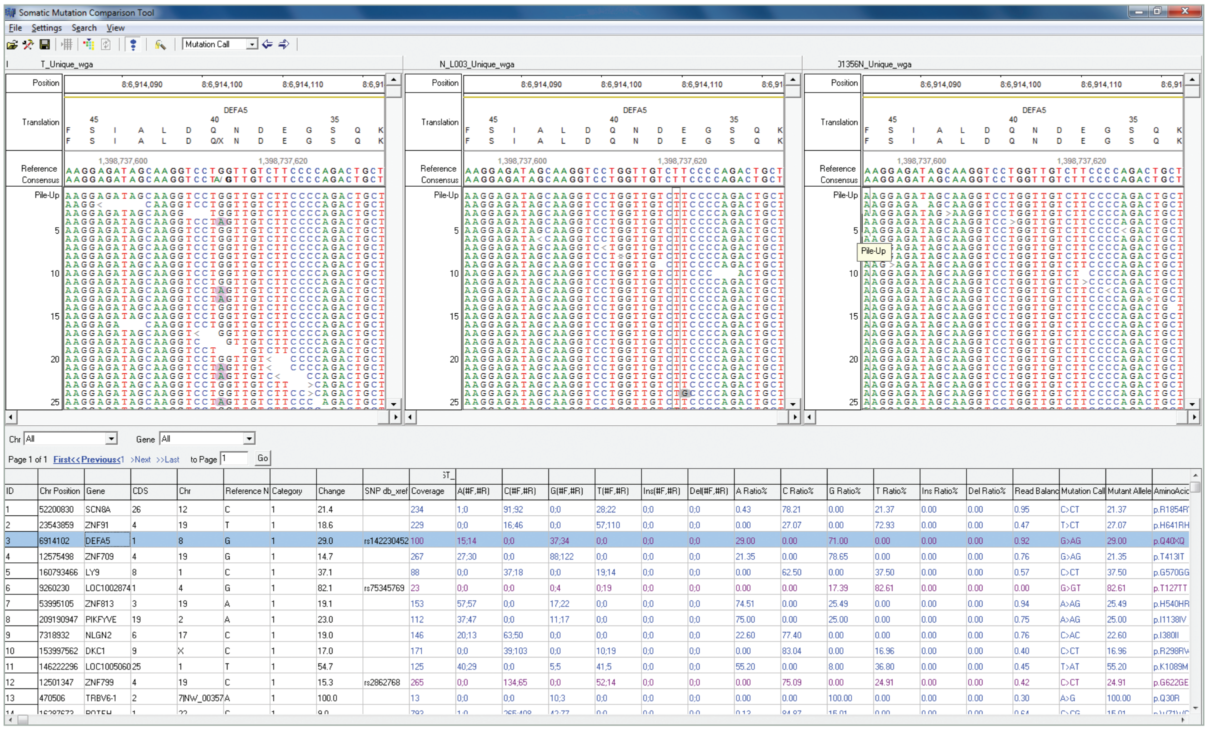
● **“Number of Pooled Samples”** – The number of samples used in the pooled project. The maximum contamination setting is divided by the square root of this number and used the same way, but for the pooled project instead of the normal project.

● **“Somatic Allele Count”** – The tumor sample must have at least this number of reads containing the variant. Default is 5 counts.

● **“Directional Balance Ratio (T/N)”** – The forward/reverse ratio in the tumor project divided by the same ratio in the normal project must not exceed this value. The default value is between(1/7x, 7x).

● **“Somatic Allele Frequency Ratio (T/N)”** – The ratio of the mutant allele percentage between the tumor and normal projects must be greater than this value. Default is set to 3.

● **“Pooled Allele Count Ratio (T/P)”** – This optional filter requires that the ratio of the number of reads with the mutant allele between the tumor and the pool projects must be greater than this value. Default is 3.



ID	Chr Position	Gene	CDS	Chr	Reference N	Category	Change	SNP db_xref	Coverage	A(#F,#R)	C(#F,#R)	G(#F,#R)	T(#F,#R)	Ins(#F,#R)	Del(#F,#R)	A Ratio%	C Ratio%	G Ratio%	T Ratio%	Ins Ratio%	Del Ratio%	Read Balanc	Mutation Call	Mutant Allele	AminoAcic
1	52200830	SCN8A	26	12	C	1	21.4		234	1;0	91;92	0;0	28;22	0;0	0;0	0.43	78.21	0.00	21.37	0.00	0.00	0.95	C>CT	21.37	p.R1854R
2	23543859	ZNF91	4	19	T	1	18.6		229	0;0	16;46	0;0	57;110	0;0	0;0	0.00	27.07	0.00	72.93	0.00	0.00	0.47	T>CT	27.07	p.H641RH
3	6914102	DEFA5	1	8	G	1	29.0	rs142230452	100	15;14	0;0	37;34	0;0	0;0	0;0	29.00	0.00	71.00	0.00	0.00	0.00	0.92	G>AG	29.00	p.Q40XQ
4	12575498	ZNF709	4	19	G	1	14.7		267	27;30	0;0	88;122	0;0	0;0	0;0	21.35	0.00	78.65	0.00	0.00	0.00	0.76	G>AG	21.35	p.T413IT
5	160793466	LY9	8	1	C	1	37.1		88	0;0	37;18	0;0	19;14	0;0	0;0	0.00	62.50	0.00	37.50	0.00	0.00	0.57	C>CT	37.50	p.G570GG
6	9260230	LOC1002874	1	4	G	1	82.1	rs75345769	23	0;0	0;0	0;4	0;19	0;0	0;0	0.00	0.00	17.39	82.61	0.00	0.00	0.00	G>GT	82.61	p.T127TT
7	53995105	ZNF813	3	19	A	1	19.1		153	57;57	0;0	17;22	0;0	0;0	0;0	74.51	0.00	25.49	0.00	0.00	0.00	0.94	A>AG	25.49	p.H540HR
8	209190947	PIKFYVE	19	2	A	1	23.0		112	37;47	0;0	11;17	0;0	0;0	0;0	75.00	0.00	25.00	0.00	0.00	0.00	0.75	A>AG	25.00	p.I1138IV
9	7318932	NLGN2	6	17	C	1	19.0		146	20;13	63;50	0;0	0;0	0;0	0;0	22.60	77.40	0.00	0.00	0.00	0.00	0.76	C>AC	22.60	p.I380II
10	153997562	DKC1	9	X	C	1	17.0		171	0;0	39;103	0;0	10;19	0;0	0;0	0.00	83.04	0.00	16.96	0.00	0.00	0.40	C>CT	16.96	p.R298RW
11	146222296	LOC1005060	25	1	T	1	54.7		125	40;29	0;0	5;5	41;5	0;0	0;0	55.20	0.00	8.00	36.80	0.00	0.00	0.45	T>AT	55.20	p.K1089M
12	12501347	ZNF799	4	19	C	1	15.3	rs2862768	265	0;0	134;65	0;0	52;14	0;0	0;0	0.00	75.09	0.00	24.91	0.00	0.00	0.42	C>CT	24.91	p.G622GE
13	470506	TRBV6-1	2	7|NW_00357	A	1	100.0		13	0;0	0;0	10;3	0;0	0;0	0;0	0.00	0.00	100.00	0.00	0.00	0.00	0.30	A>G	100.00	p.Q30R

Tests using 5 pooled controls have shown a reduction of up to 70% in the number of false positive calls. Most of these false positives were caused by common variants and systematic errors that appear in the tumor sample but were not called in the control sample due to low coverage.

Structural Variant Detection

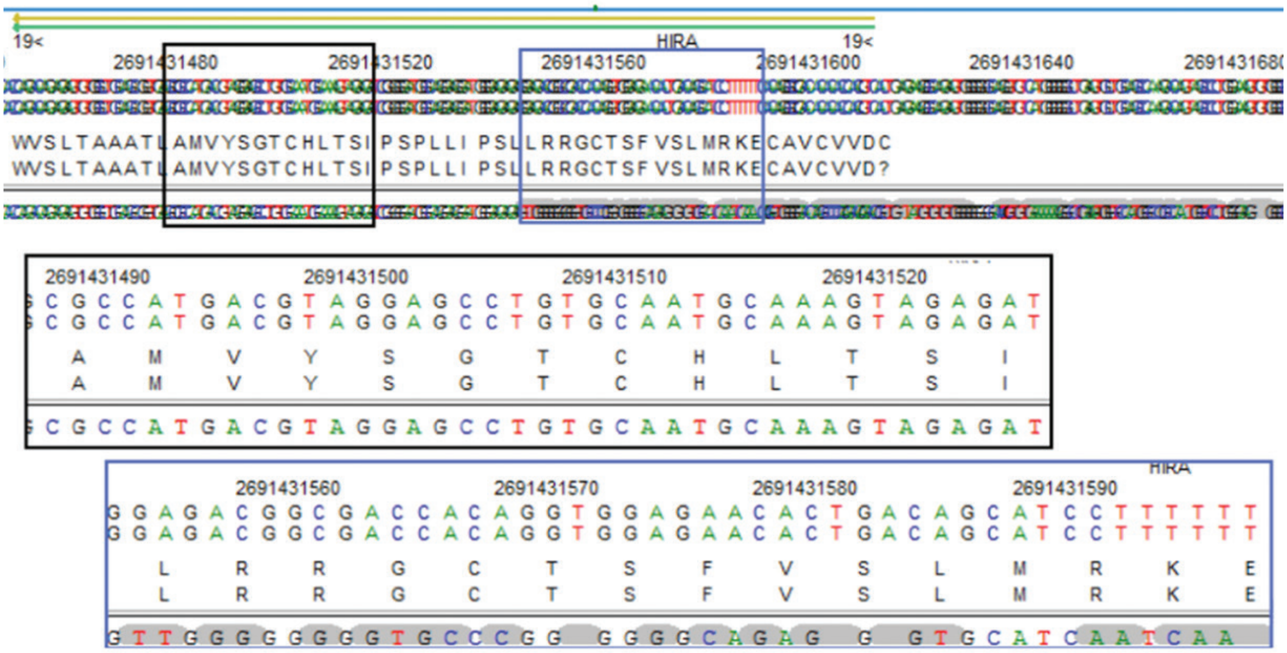
- **Discovery of large Insertion & Deletion; Translocation; Gene Fusion**
- **Flexible Alignment technology allows for large mismatches**
- **Creation of "Pseudo Pairs" allows for detection and mapping of structural variations**

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.


19<
HIRA
19<
2691431480
2691431520
2691431560
2691431600
2691431640
WVSLTAAATL AMVYSGTCHLTSI PSPLLI PSL LRRGCTSF VSLMRKE CAVCVVDC
WVSLTAAATL AMVYSGTCHLTSI PSPLLI PSL LRRGCTSF VSLMRKE CAVCVVD?
2691431490
2691431500
2691431510
2691431520
CGCCATGACGTAGGAGCCTGTGCAATGCAAAGTAGAGAT
CGCCATGACGTAGGAGCCTGTGCAATGCAAAGTAGAGAT
A M V Y S G T C H L T S I
A M V Y S G T C H L T S I
CGCCATGACGTAGGAGCCTGTGCAATGCAAAGTAGAGAT
HIRA
2691431560
2691431570
2691431580
2691431590
GGAGACGGCGACCACAGGTGGAGAACACTGACAGCATCCTTTTTTT
GGAGACGGCGACCACAGGTGGAGAACACTGACAGCATCCTTTTTTT
L R R G C T S F V S L M R K E
L R R G C T S F V S L M R K E
GTTGGGGGGGGTGCCCGG GGGGCAGAG G GTGCATCAATCAA

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.

Variant Comparison Tool, with functional prediction

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.

Excellent for comparing:

- Trios
- Families
- Groups

Advanced Comparison between multiple projects - 5 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
- Gene Association

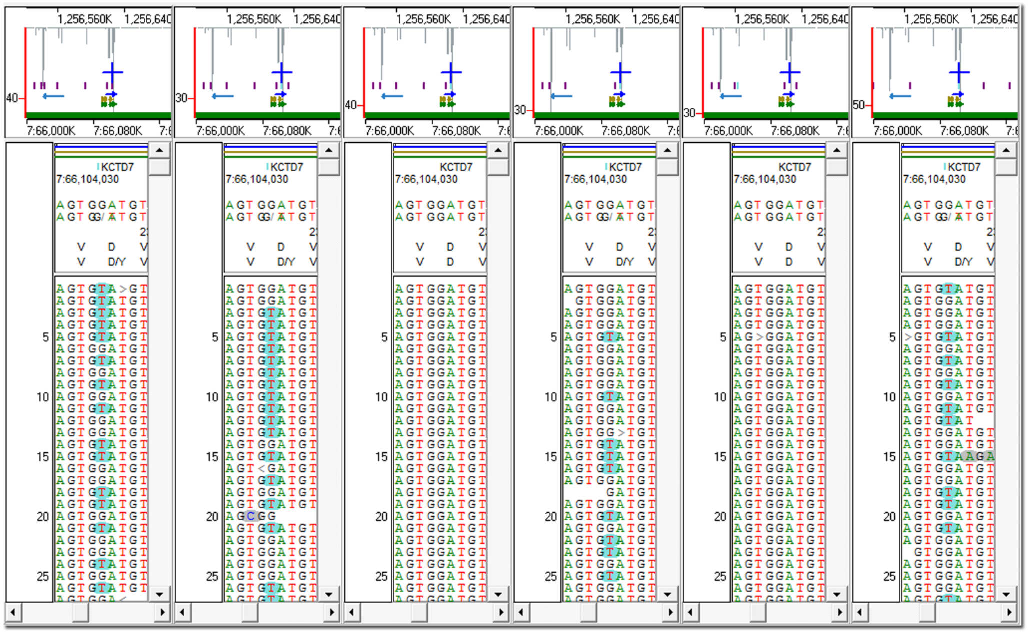
Many advanced filtering options

- Annotation (mRNA, CDS, Splice sites)
- Mutation Confidence Score
- dbSNP mutations
- Substitutions (Non-coding, Silent, Mis-sense, Nonsense) or Indels
- Advanced ROI filtering
- Imported Tracks such as Prediction Information
- Concordance

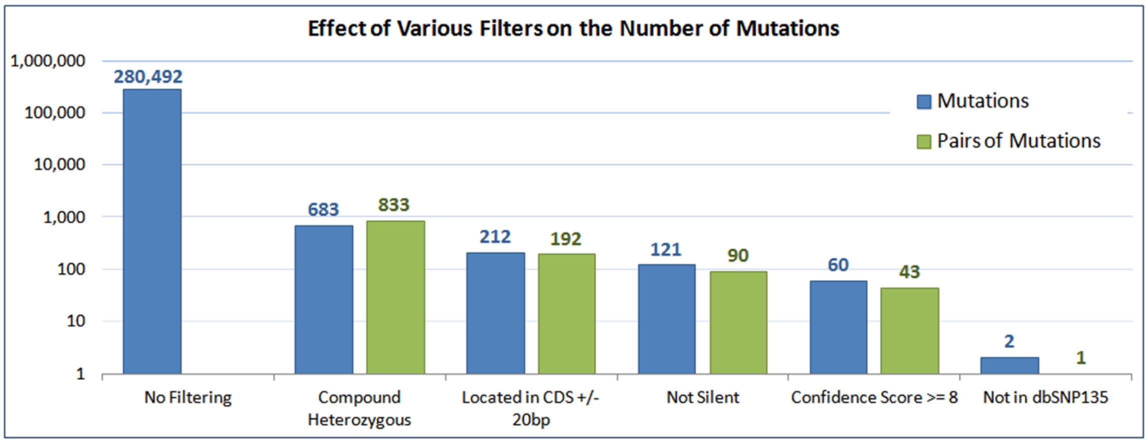
Prediction database integration

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

View multiple projects side-by-side



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.

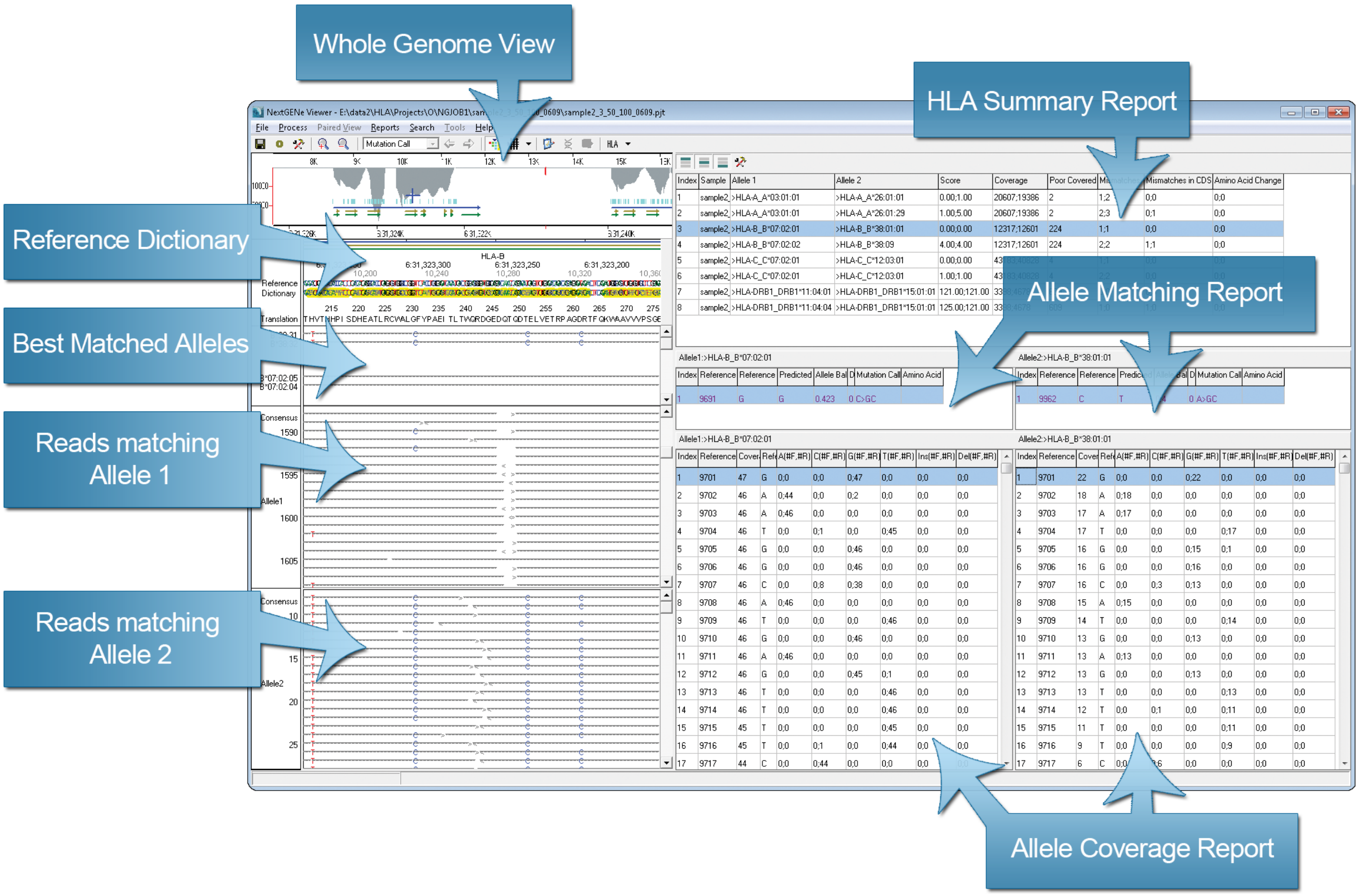


Number of Mutations Left After Each Filtering Step Within NextGENe software.

MHC/HLA Analysis

Typing of the Major Histocompatibility Complex (MHC/HLA) in humans and non-human primates using next-generation sequencing allows for phasing of multiple alleles. 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 and Class II 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 dictionary to determine to most likely HLA alleles according to amino acid differences, synonymous changes, and intronic changes.



Index	Sample	Allele 1	Allele 2	Score	Coverage	Poor Covered	Mismatches	Mismatches in CDS	Amino Acid Change
1	sample2_	>HLA-A_A*03:01:01	>HLA-A_A*26:01:01	0.00;1.00	20607;19386	2	1;2	0;0	0;0
2	sample2_	>HLA-A_A*03:01:01	>HLA-A_A*26:01:29	1.00;5.00	20607;19386	2	2;3	0;1	0;0
3	sample2_	>HLA-B_B*07:02:01	>HLA-B_B*38:01:01	0.00;0.00	12317;12601	224	1;1	0;0	0;0
4	sample2_	>HLA-B_B*07:02:02	>HLA-B_B*38:09	4.00;4.00	12317;12601	224	2;2	1;1	0;0
5	sample2_	>HLA-C_C*07:02:01	>HLA-C_C*12:03:01	0.00;0.00	43183;40828	4	1;1	0;0	0;0
6	sample2_	>HLA-C_C*07:02:01	>HLA-C_C*12:03:01	1.00;1.00	43183;40828	4	2;2	0;0	0;0
7	sample2_	>HLA-DRB1_DRB1*11:04:01	>HLA-DRB1_DRB1*15:01:01	121.00;121.00	3398;4678	[illegible]	[illegible]	[illegible]	[illegible]
8	sample2_	>HLA-DRB1_DRB1*11:04:04	>HLA-DRB1_DRB1*15:01:01	125.00;121.00	3398;4678	609	1;0	1;0	0;0

Allele1:>HLA-B_B*07:02:01

Index	Reference	Reference	Predicted	Allele Bal	D	Mutation Call	Amino Acid
1	9691	G	G	0.423	0	C>GC	

Allele2:>HLA-B_B*38:01:01

Index	Reference	Reference	Predicted	Allele Bal	D	Mutation Call	Amino Acid
1	9962	C	T	[illegible]	0	A>GC	

Allele1:>HLA-B_B*07:02:01

Index	Reference	Cover	Ref	A(#F,#R)	C(#F,#R)	G(#F,#R)	T(#F,#R)	Ins(#F,#R)	Del(#F,#R)
1	9701	47	G	0;0	0;0	0;47	0;0	0;0	0;0
2	9702	46	A	0;44	0;0	0;2	0;0	0;0	0;0
3	9703	46	A	0;46	0;0	0;0	0;0	0;0	0;0
4	9704	46	T	0;0	0;1	0;0	0;45	0;0	0;0
5	9705	46	G	0;0	0;0	0;46	0;0	0;0	0;0
6	9706	46	G	0;0	0;0	0;46	0;0	0;0	0;0
7	9707	46	C	0;0	0;8	0;38	0;0	0;0	0;0
8	9708	46	A	0;46	0;0	0;0	0;0	0;0	0;0
9	9709	46	T	0;0	0;0	0;0	0;46	0;0	0;0
10	9710	46	G	0;0	0;0	0;46	0;0	0;0	0;0
11	9711	46	A	0;46	0;0	0;0	0;0	0;0	0;0
12	9712	46	G	0;0	0;0	0;45	0;1	0;0	0;0
13	9713	46	T	0;0	0;0	0;0	0;46	0;0	0;0
14	9714	46	T	0;0	0;0	0;0	0;46	0;0	0;0
15	9715	45	T	0;0	0;0	0;0	0;45	0;0	0;0
16	9716	45	T	0;0	0;1	0;0	0;44	0;0	0;0
17	9717	44	C	0;0	0;44	0;0	0;0	0;0	0;0

Allele2:>HLA-B_B*38:01:01

Index	Reference	Cover	Ref	A(#F,#R)	C(#F,#R)	G(#F,#R)	T(#F,#R)	Ins(#F,#R)	Del(#F,#R)
1	9701	22	G	0;0	0;0	0;22	0;0	0;0	0;0
2	9702	18	A	0;18	0;0	0;0	0;0	0;0	0;0
3	9703	17	A	0;17	0;0	0;0	0;0	0;0	0;0
4	9704	17	T	0;0	0;0	0;0	0;17	0;0	0;0
5	9705	16	G	0;0	0;0	0;15	0;1	0;0	0;0
6	9706	16	G	0;0	0;0	0;16	0;0	0;0	0;0
7	9707	16	C	0;0	0;3	0;13	0;0	0;0	0;0
8	9708	15	A	0;15	0;0	0;0	0;0	0;0	0;0
9	9709	14	T	0;0	0;0	0;0	0;14	0;0	0;0
10	9710	13	G	0;0	0;0	0;13	0;0	0;0	0;0
11	9711	13	A	0;13	0;0	0;0	0;0	0;0	0;0
12	9712	13	G	0;0	0;0	0;13	0;0	0;0	0;0
13	9713	13	T	0;0	0;0	0;0	0;13	0;0	0;0
14	9714	12	T	0;0	0;1	0;0	0;11	0;0	0;0
15	9715	11	T	0;0	0;0	0;0	0;11	0;0	0;0
16	9716	9	T	0;0	0;0	0;0	0;9	0;0	0;0
17	9717	6	C	0;0	0;6	0;0	0;0	0;0	0;0

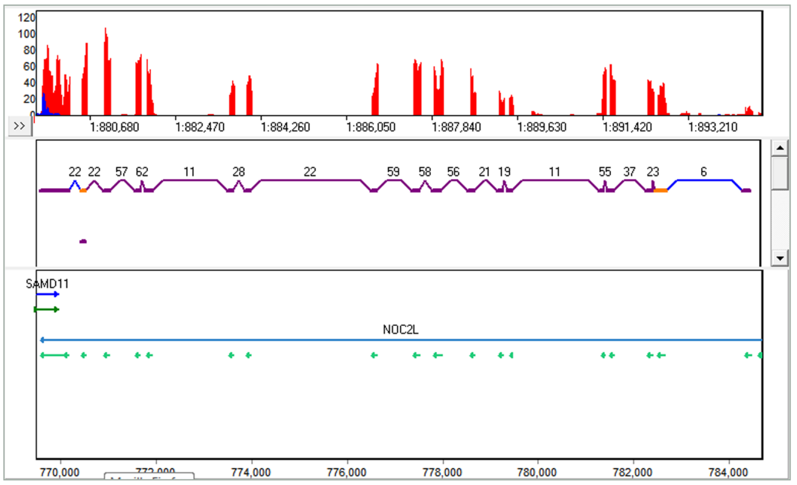
NextGENe Viewer includes several views and reports within a single screen. On the left side, the top pane shows a whole genome view of the HLA genes, the next pane shows the reference nucleotide sequence and a consensus sequence of all alleles, the middle pane shows the top matching alleles and the following two panes show the pile-up of reads grouped by alleles. On the right side, the top pane is the HLA Report listing the top allele picks for each gene and some summary statistics, the next pane down shows the Matching Report identifying the locations where there is a discrepancy between the allele call and the sequence within the reads and the last pane shows the coverage report listing all positions with inadequate coverage.

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, RPKM, FPKM)**
- **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. 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. 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. 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.



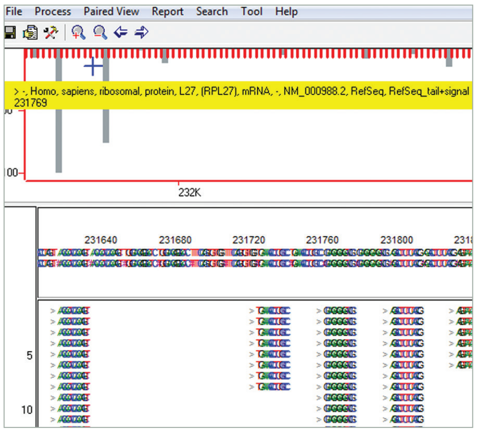
The new RNA-Seq Transcript View. The purple links and exons are known, the blue links are novel, and the orange exons are alternative splicing. The figure of Ion Torrent 318 chip RNA-Seq data set, available on Ion Torrent Development community website.

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.



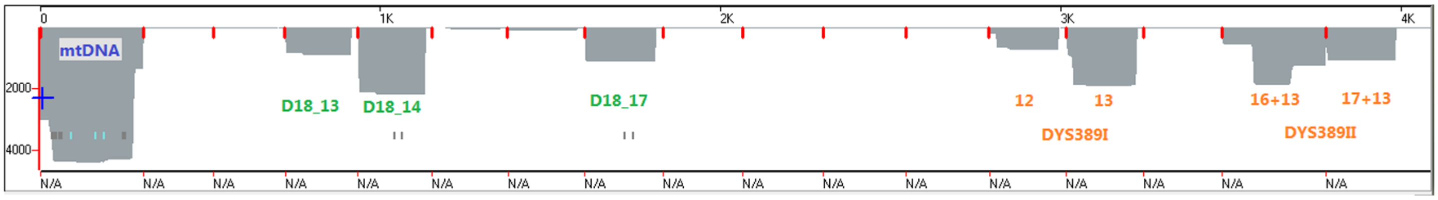
> -, Homo, sapiens, ribosomal, protein, L27, (RPL27), mRNA, -, NM_000988.2, RefSeq, RefSeq_tail+signal
231769

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.

Forensic, Human Identity Analysis

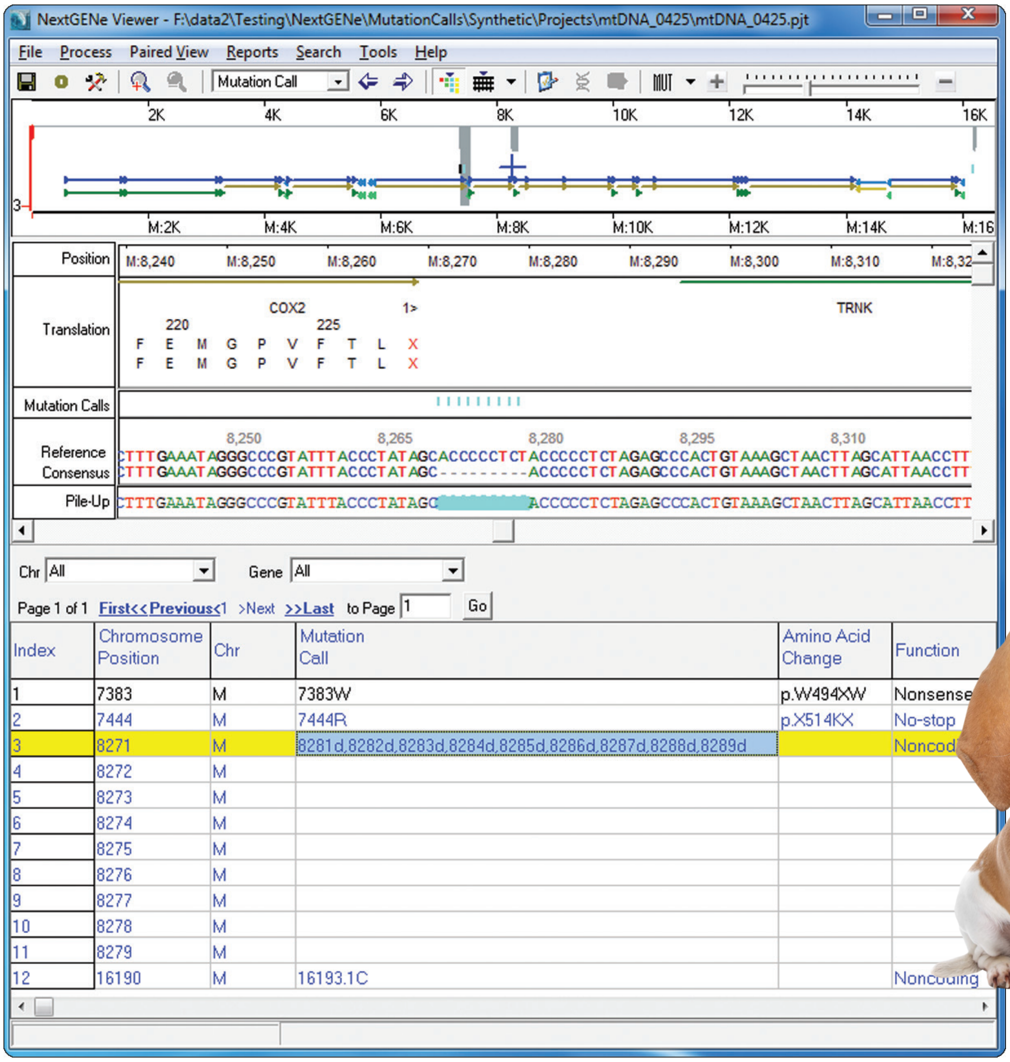
STR Analysis

There is tremendous potential for the use of 2nd generation sequencing in forensic STR 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.

Mitochondrial Analysis


NextGENe Viewer - F:\data2\Testing\NextGENe\MutationCalls\Synthetic\Projects\mtDNA_0425\mtDNA_0425.pjt
File
Process
Paired View
Reports
Search
Tools
Help
Mutation Call
2K
4K
6K
8K
10K
12K
14K
16K
M:2K
M:4K
M:6K
M:8K
M:10K
M:12K
M:14K
M:16
Position
M:8,240
M:8,250
M:8,260
M:8,270
M:8,280
M:8,290
M:8,300
M:8,310
M:8,32
Translation
COX2
220
225
1>
TRNK
F E M G P V F T L X
F E M G P V F T L X
Mutation Calls
8,250
8,265
8,280
8,295
8,310
Reference
Consensus
Pile-Up
CTTTGAAATAGGGCCCGTATTTACCCTATAGCACCCCCTCTACCCCCTCTAGAGCCCACTGTAAAGCTAACTTAGCATTAACCTT
CTTTGAAATAGGGCCCGTATTTACCCTATAGC---------ACCCCCTCTAGAGCCCACTGTAAAGCTAACTTAGCATTAACCTT
CTTTGAAATAGGGCCCGTATTTACCCTATAGC ACCCCCTCTAGAGCCCACTGTAAAGCTAACTTAGCATTAACCTT
Chr
All
Gene
All
Page 1 of 1
First<<Previous<1 >Next >>Last
to Page
1
Go

Index	Chromosome Position	Chr	Mutation Call	Amino Acid Change	Function
1	7383	M	7383W	p.W494XW	Nonsense
2	7444	M	7444R	p.X514KX	No-stop
3	8271	M	8281d,8282d,8283d,8284d,8285d,8286d,8287d,8288d,8289d		Noncod
4	8272	M			
5	8273	M			
6	8274	M			
7	8275	M			
8	8276	M			
9	8277	M			
10	8278	M			
11	8279	M			
12	16190	M	16193.1C		Noncoding

Several advantages are available within NextGENe software for forensic mitochondrial analyses. Heteroplasmy detection sensitivity can be enhanced with the use of NextGENe's patented Condensation Tool, which helps to distinguish instrumental error from low frequency variants. The vertebrate mitochondrial genetic code is automatically used by the NextGENe Viewer. Mutation calls are made by the NextGENe Viewer using interpretation guidelines for mitochondrial DNA analysis from the Scientific Working Group on DNA Analysis Methods.

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.

G G T C C G A A A A A A C G C C G

⇩

G T C G A C G C G

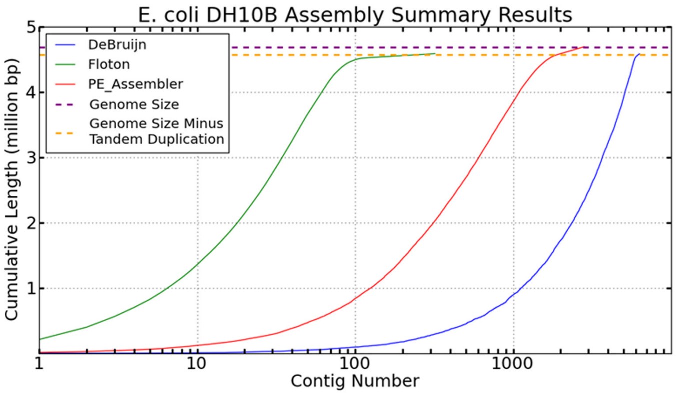
2 1 2 1 6 1 1 2 1

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

	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

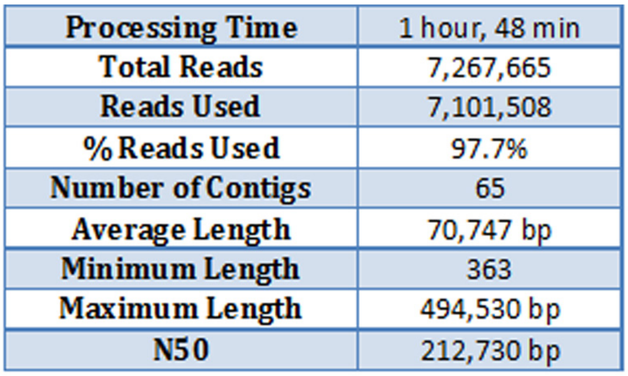
The NextGENe software’s Floton Assembler is specially designed to handle data from flow-based sequencing technologies that tend to have indel errors rather than substitution errors. When paired with the Ion Torrent or Roche systems it provides a very fast and inexpensive method for *de novo* sequencing 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**

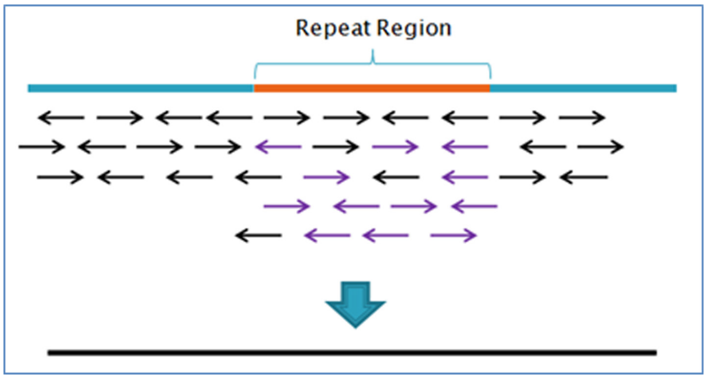
Sequences that repeat throughout the genome can pose a problem for the assembly of short reads. Normally the repeat reads would assemble within the incorrect contig. This causes two problems - the repeat regions elsewhere in the genome have reduced coverage and contigs terminate at repeat regions due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

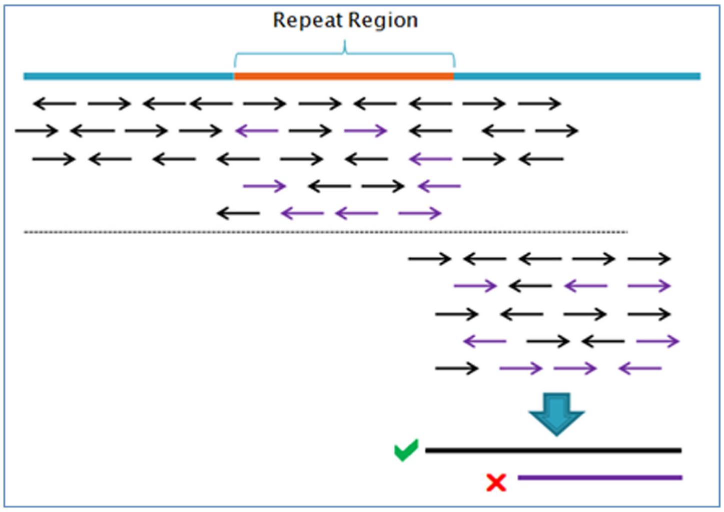


Processing Time	1 hour, 48 min
Total Reads	7,267,665
Reads Used	7,101,508
% Reads Used	97.7%
Number of Contigs	65
Average Length	70,747 bp
Minimum Length	363
Maximum Length	494,530 bp
N50	212,730 bp

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:

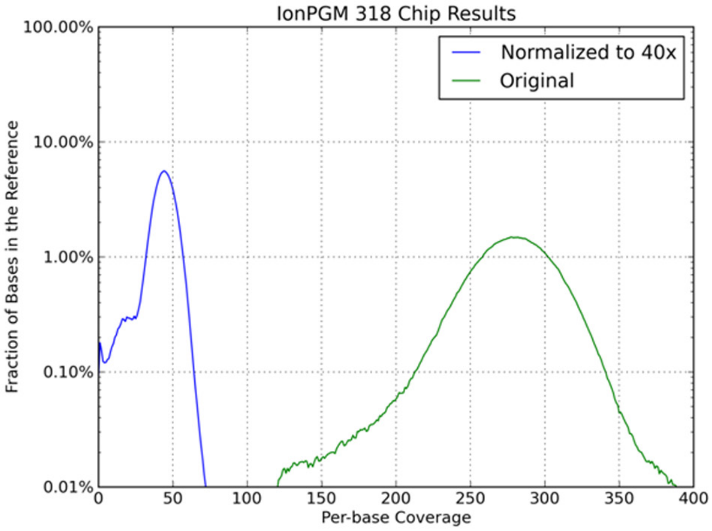
Removes excess coverage (including in repeat regions)
- Longer, higher-quality reads are retained
- Assembly speed and accuracy are improved
- Less RAM is required

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

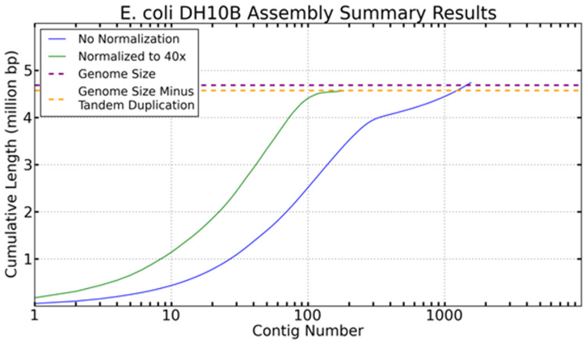
Results:

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



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

Coverage distribution before and after.



	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.

Data courtesy of Ion Community

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

Disc 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 Genetic Analysis

www.softgenetics.com

SOFTGENETICS®

Software PowerTools for Genetic Analysis

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

For Clinical Research