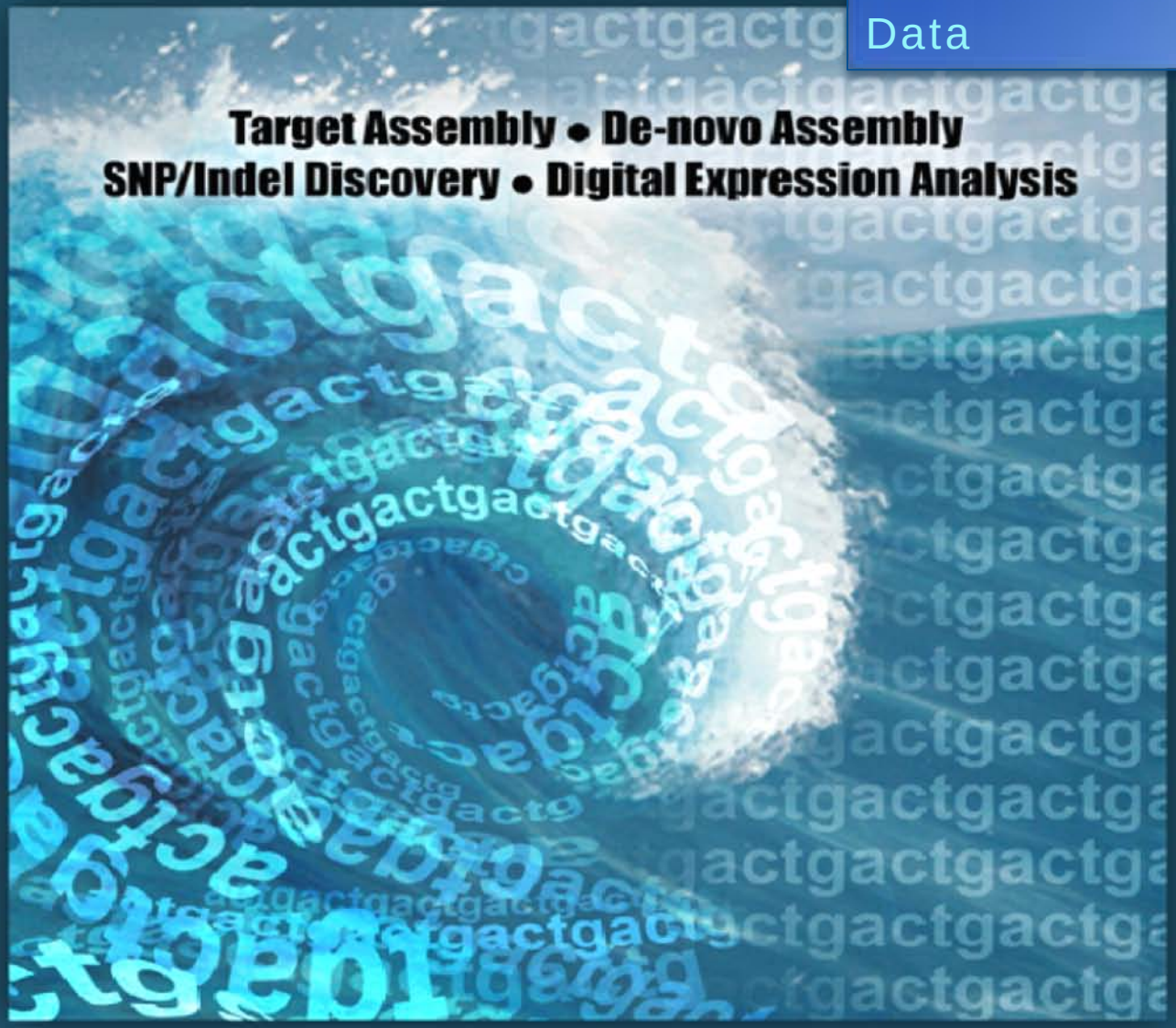
NextGENe™

Next Generation Sequencing Software

Highly Accurate, Biologist-friendly Software for the analysis of Next Generation Sequencing Data



SOFTGENETICS®

Software PowerTools for Genetic Analysis

NextGENe™
Next Generation Sequencing Software

NextGENe Advantages

- Instrumentation
  - GS FLX System – Roche/454
  - Genome Analyzer – Illumina/Solexa
  - SOLiD – Applied Biosystems
- Computation
  - Easy-to-use Windows® Interface
    - Analysis on a 64-bit PC (less than $3500)
  - Optimized for Performance and Accuracy
  - Technical Support/Training

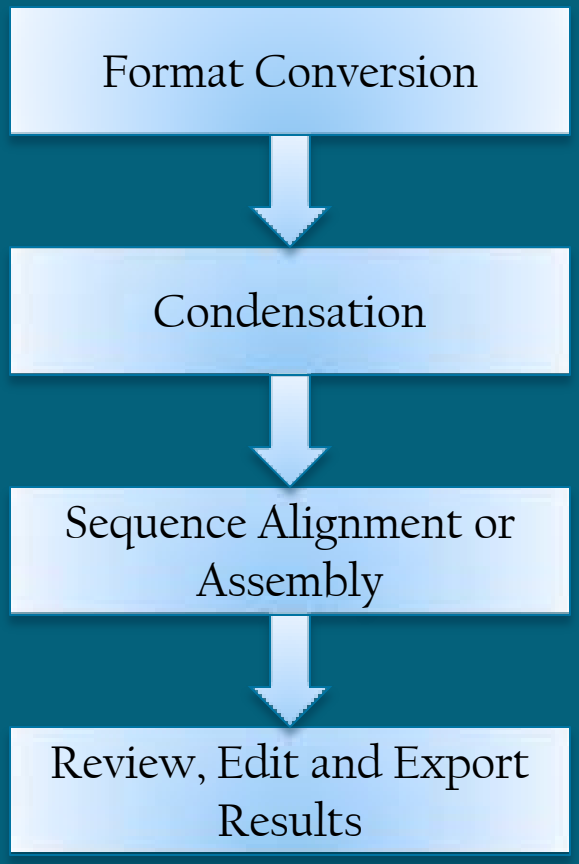
NextGENe™
Next Generation Sequencing Software

Applications

- SNP/Indel Discovery
- *De novo* Assembly
- Expression Studies
  - Transcriptome/miRNA
  - DGE/SAGE
  - ChIP-Seq

NextGENe™
Next Generation Sequencing Software

Workflow



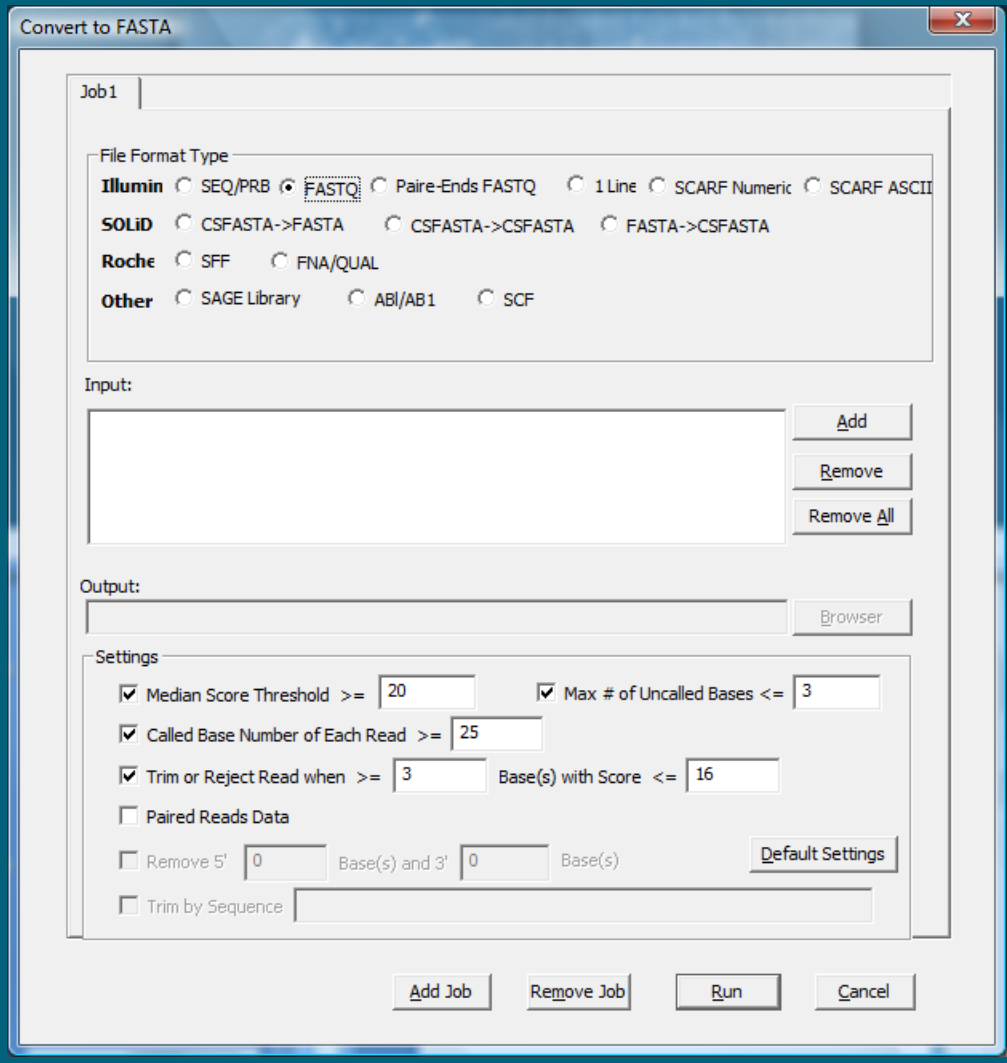
Point-and-click Run Wizard guides users through project set-up step-by-step

NextGENe™

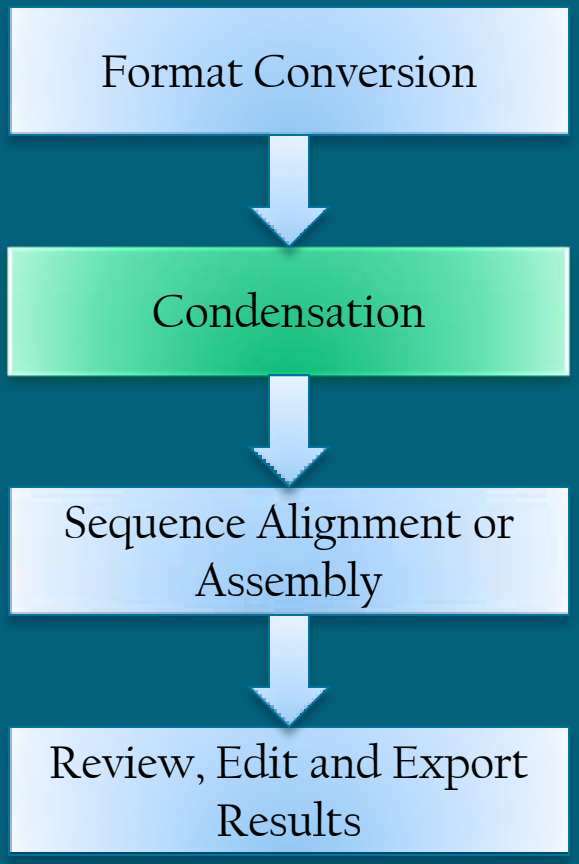
Next Generation Sequencing Software

Convert sample files to fasta format. Use quality scores to trim low quality data

- Format Conversion
- Condensation
- Sequence Alignment or Assembly
- Review, Edit and Export Results



NextGENe™
Next Generation Sequencing Software



Multiple Condensation Methods Available:

- **Consolidation**: Corrects errors, Lengthens reads and Reduces read count
- **Elongation**: Corrects errors and Lengthens reads (recommended for paired end data)
- **Error Correction**: Corrects errors only
  - 454 Error Correction
  - SOLiD/Illumina Error Correction

NextGENe™
Next Generation Sequencing Software

Condensation

SOLiD System & Illumina

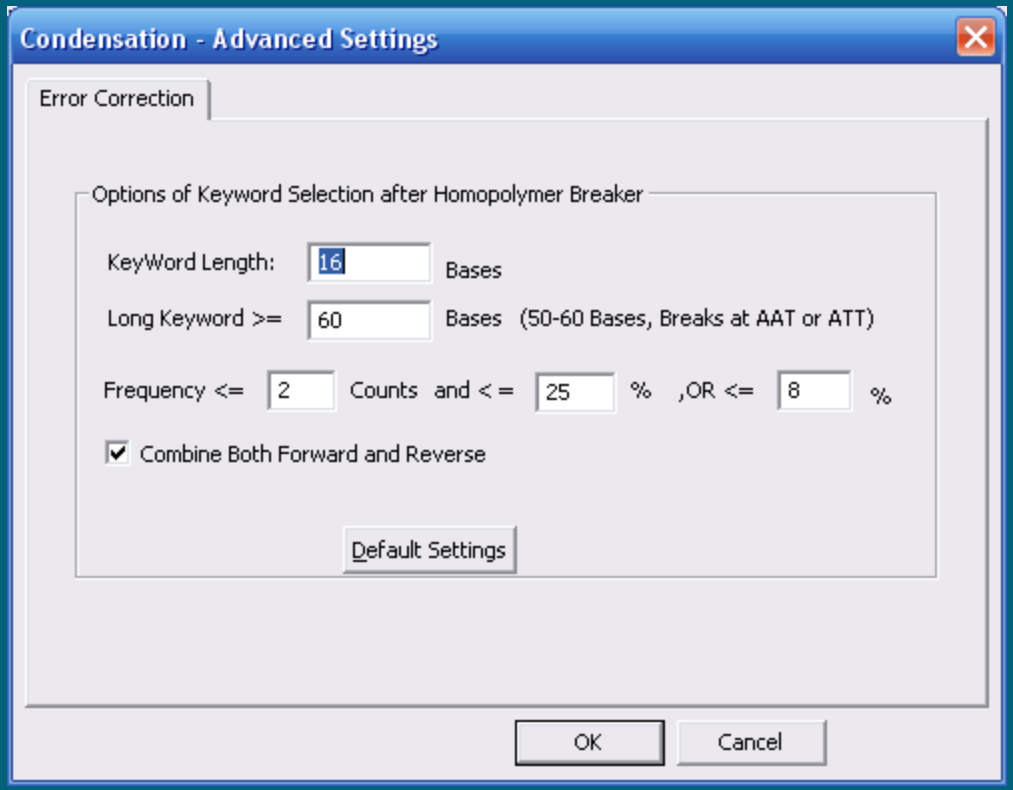
- Reads grouped by 12 bp anchor sequences
- Sub-grouped by shoulder sequences
- Consensus is used to correct errors and, when applicable, lengthen reads and reduce read count

NextGENe™
Next Generation Sequencing Software

Condensation

Roche/454

- Reads are broken into fragments at homopolymers of >=3 bps.
- Keywords (fragments between homopolymers above set length) are used to cluster reads
- Consensus is used to correct errors

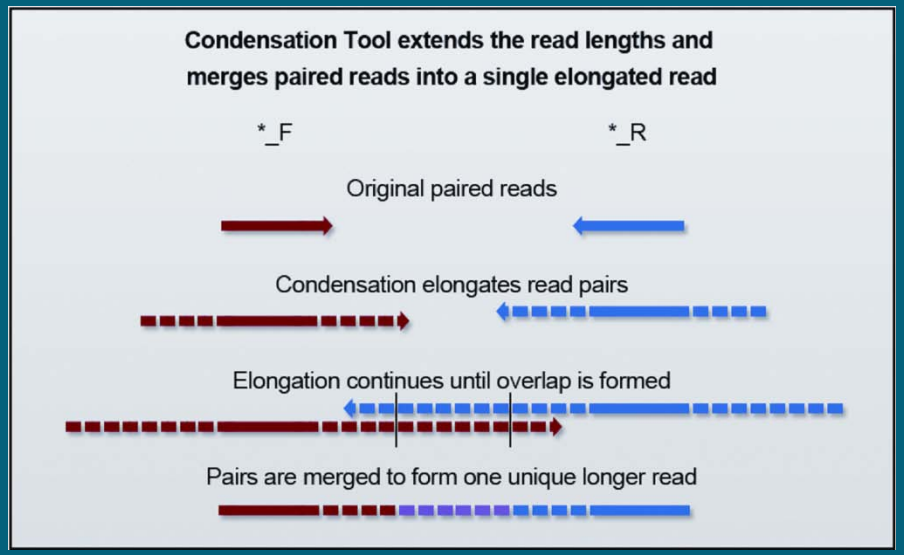


>000066_3814_1605 length=101 uaccno=FCC1VVV04JK5NB
AATGTATCTTACA3,3AGTATTCACCTGTAGCGTAATCGGACT
TCTCA4,4AGCCT3,3TAAATGGT4,4TCCTGT3,3TCT3,3TAGC
CATTAACATAGATGTCTTAATAG

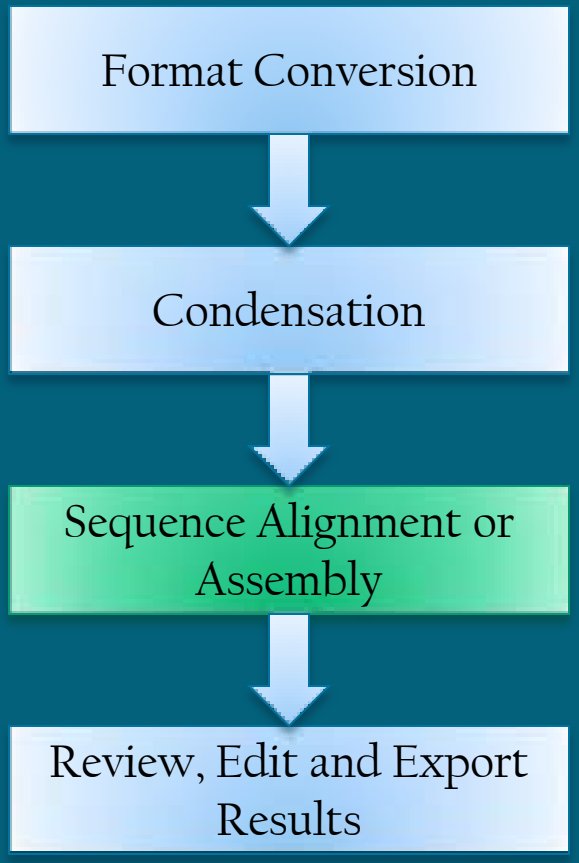
NextGENe™
Next Generation Sequencing Software

Paired End Merging

- Reads are elongated using the Condensation Tool to form overlap between paired reads (multiple cycles can be used as needed)
- Overlapping paired reads can be merged if overlapping portion matches



NextGENe™
Next Generation Sequencing Software



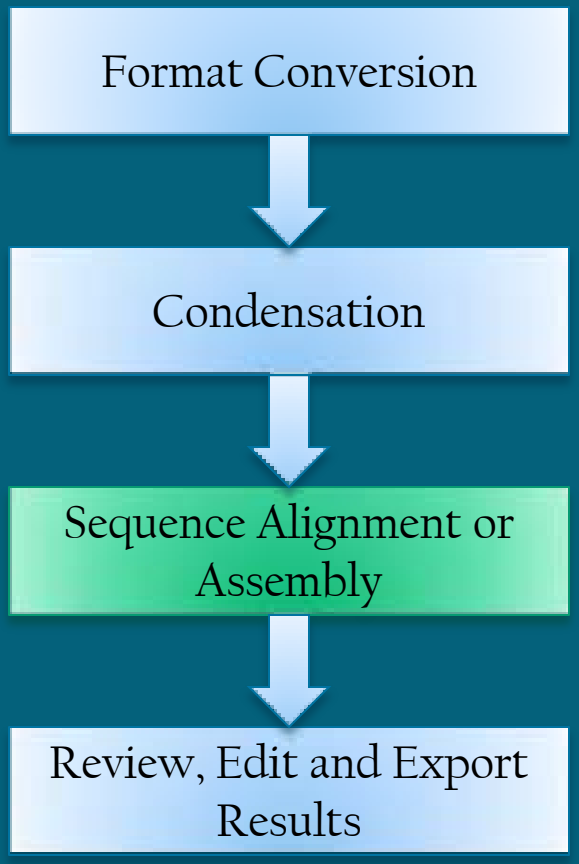
- Possible matching positions are found using 12 bp sequences of reads.
- Best alignment position is determined by greatest number of matched bases and, when two positions have same number of matched bases, highest uniqueness score.

  Uniqueness score = sqrt(1/n)

  n = number of hits in reference

NextGENe™

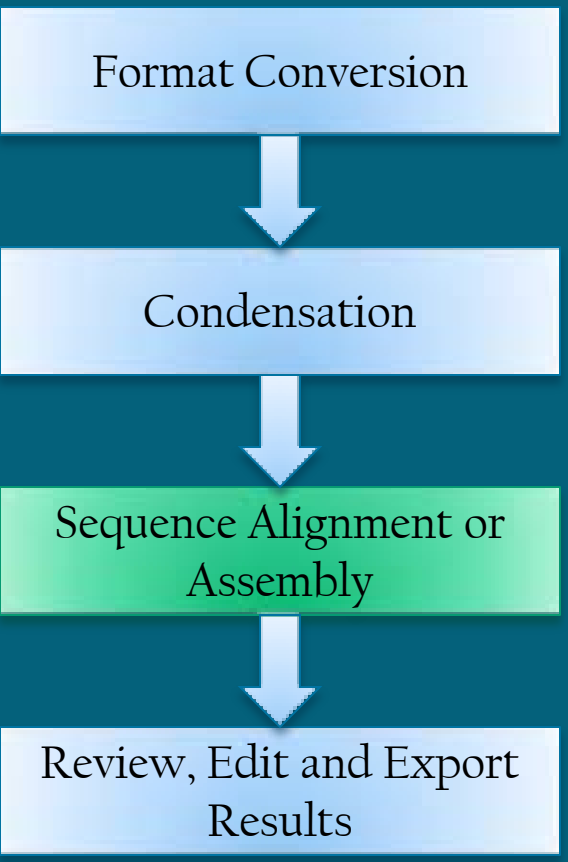
Next Generation Sequencing Software



- Aligns to whole large genomes (i.e. human) using Burrows-Wheeler Transform
- Indexes reference sequence prior to alignment, resulting in rapid mapping of reads
- Alignment by:
  1. Matching entire read perfectly
  2. Matching entire read with mismatches
  3. Matching using seed sequences

Indexes can be supplied by SoftGenetics or create your own using NextGENe tool

NextGENe™
Next Generation Sequencing Software



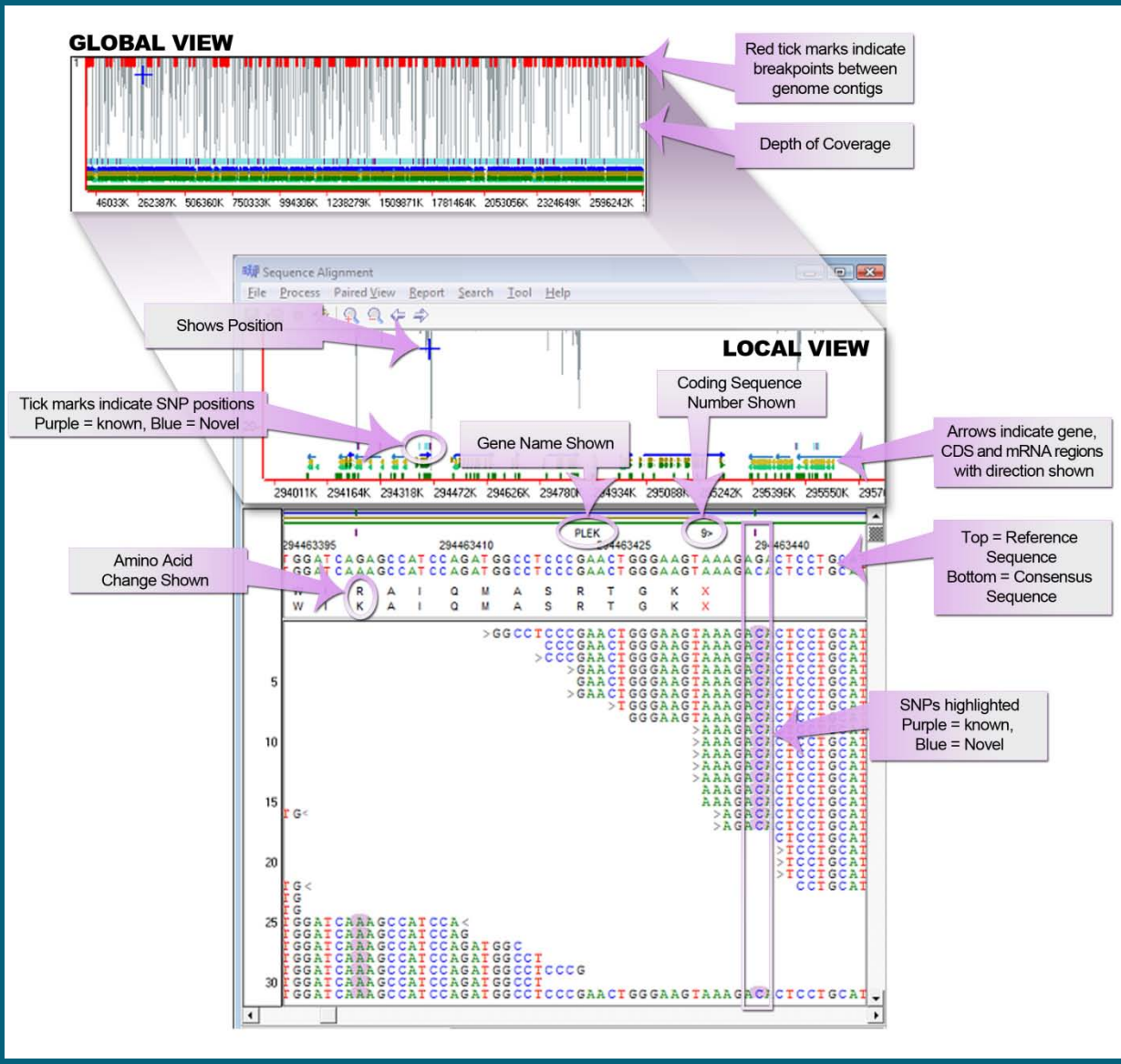
Multiple Assembly Methods Available:

- **De Bruijn**: uses de Bruijn graph, able to utilize paired reads
- **Maximum Overlap**: looks for overlaps considering each position (for Illumina reads)
- **PE Assembler**: Uses paired read data to assemble across repeats
- **Greedy**: looks for overlaps by considering each position (for 454 reads)
- **Skeleton™**: uses seed keys (between homopolymers) to determine overlaps between reads (for 454 reads, faster than Greedy method)

NextGENe™

Next Generation Sequencing Software

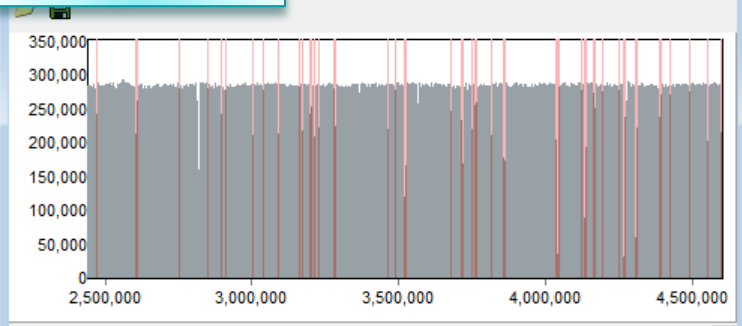
- Format Conversion
- Condensation
- Sequence Alignment or Assembly
- Review, Edit and Export Results



NextGENe™
Next Generation Sequencing Software

Reports: Data Characteristics

Coverage Curve



Region	Description	Start	End
1		211068	211079
2		252589	252779
3		252900	252919
4		253031	253096
5		253166	253223

Distribution Report

Matched\Unmatched Report

	Matched:4380144	Unmatched:120345
Index	Title	Sequence
1	>>14_CGGCGAAACACA_23_GCTGATCGTT_GGA	AGCATTTTCAGGCGCTGATCGTTCGGCGAAACACAGGAGCAGCCAATCGCCAGA
2	>>20_CGGCGAAACAGC_22_TCGGCACG_GCCAC	AGAAATGATCGGAATCGGCACGCGGCGAAACAGCGCCACCACCGCAAACCAGTCCGC
3	>>23_AAAAAAAACAGC_23_ATTAAGAAAC_GCC	CTTACAACACATTATTAAGAAACAAAAAAAACAGCGCCCAACACTTCCCCGGACGCT
4	>>42_AGTTCGCTGTCG_23_TCTCGACCGC_GTA	CGGTAATCAATTATCTCGACCGCAGTTCGCTGTCGGTAGCAAATTTAACGATTCGTGA
5	>>44_AAAAAAAACCAG_23_GCGAGAAT_CAGGT	TCCCTTTTTAGCGCGGCGAGAATAAAAAAAACCAGCAGGTATAATCTGCTGGCGGGTG
6	>>65_CGGCGAAACGGC_23_TTTATCTT_GGTCA	AGTTGCCAGCGCAGGTTTATCTTCGGCGAAACGGCGGTCAATTTCGTCCATTCGATCC
7	>>123_AAAAAAAAGAGT_23_TTTTGAGC_TCGC	GGGCAACTTGGGCTATTTTGAGCAAAAAAAAGAGTTCGCCAGATACCATTTTGATGC
8	>>237_AAAAAAAATCGC_23_AATCAATATC_AAA	ATAATATTCACCAAATCAATATCAAAAAAAATCGCAAAACATATAATTCAATACAAAT
9	>>243_CGGCGAAATGTC_21_GCGAATGT_CGGT	CTTATTTAATGTTGCGAATGTCGGCGAAATGTCCGGTCACAGCAGCAGCGGCAGC
10	>>266_CGGCGAACAAAA_23_GGCCAAAG_TCAG	CTACATCCTCGAAGAGGCCAAAGCGGCGAACAAAATCAGCCCGGTCAGTGAGTCACCC

NextGENe™
Next Generation Sequencing Software

Reports: Variant Detection

Blue – Novel SNPs
Purple – Reported SNPs
Various Display and Filtering Options

Mutation Report

File

Index	Reference Position	Gene	CDS	Chr	Reference Nucleotide	Coverage	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	SNP db_xref	Genotype	Mutation Call	AminoAcid Change
34	2388440598	CDC27	11	17	T	6	0.00	100.00	0.00	0.00	0.00	0.00		CC	T>C	414K>R
35	2388440640	CDC27	11	17	T	9	0.00	88.89	0.00	11.11	0.00	0.00		CC	T>C	400K>R
36	2388440642	CDC27	11	17	G	7	0.00	85.71	14.29	0.00	0.00	0.00		CC	G>C	399S>R
37	2388440653	CDC27	0	17	G	4	0.00	75.00	25.00	0.00	0.00	0.00		CC	G>C	
38	2388455191	CDC27	7	17	T	13	0.00	0.00	92.31	7.69	0.00	0.00	rs3208659	GG	T>G	260N>H
39	2388455325	CDC27	7	17	A	10	20.00	80.00	0.00	0.00	0.00	0.00		CC	A>C	215L>W
40	2443697788	GNAL	0	18	A	4	25.00	0.00	75.00	0.00	0.00	0.00	rs1647555	GG	A>G	
41	2562494785	MYO1F	0	19	C	5	80.00	20.00	0.00	0.00	0.00	0.00		AA	C>A	
42	2562494787	MYO1F	0	19	C	5	0.00	0.00	0.00	0.00	0.00	100.00		delCC	delCC	
43	2562494788	MYO1F	0	19	C	5	0.00	0.00	0.00	0.00	0.00	100.00		-	-	
44	2562494808	MYO1F	0	19	C	4	25.00	0.00	75.00	0.00	0.00	0.00		GG	C>G	

Hyperlink to NCBI dbSNP database

NextGENe™
Next Generation Sequencing Software

Reports: Compare SNP Projects

SNP Comparison Report

File Setting

			s_1_align_									s_2_align_								
Index	Reference Pos	Gene	Coverage	A Ratio%	C Ratio%	G Ratio%	T Ratio%	Del Ratio%	Ins Ratio%	Mutation Call	AminoAcid Change	Coverage	A Ratio%	C Ratio%	G Ratio%	T Ratio%	Del Ratio%	Ins Ratio%	Mutation Call	AminoAcid Change
28	68382	murE	33	0.00	0.00	96.97	3.03	0.00	0.00	A>G	371A>A	22	0.00	0.00	100.00	0.00	0.00	0.00	A>G	371A>A
29	108360		17	0.00	0.00	100.00	0.00	0.00	0.00	T>G		17	0.00	0.00	100.00	0.00	0.00	0.00	T>G	
30	111113		28	0.00	0.00	78.57	21.43	0.00	0.00	G>GT		34	n/a	n/a	n/a	n/a	n/a	n/a		
31	142348	fhuA	32	0.00	0.00	100.00	0.00	0.00	0.00	R>G	254X>W	25	0.00	0.00	100.00	0.00	0.00	0.00	R>G	254X>W
32	145989	fhuB	16	n/a	n/a	n/a	n/a	n/a	n/a			29	0.00	0.00	20.69	79.31	0.00	0.00	T>GT	142S>SA
33	180966	dnaE	22	0.00	0.00	22.73	77.27	0.00	0.00	T>GT	579N>NK	21	14.29	0.00	23.81	61.90	0.00	0.00	T>GT	579N>NK
34	198078	rrsH	161	n/a	n/a	n/a	n/a	n/a	n/a			115	0.00	0.00	80.00	20.00	0.00	0.00	G>GT	
35	207222	mltD	21	n/a	n/a	n/a	n/a	n/a	n/a			29	0.00	24.14	0.00	75.86	0.00	0.00	T>CT	280I>VI
36	218669	yafJ	24	100.00	0.00	0.00	0.00	0.00	0.00	T>A	80L>Q	28	100.00	0.00	0.00	0.00	0.00	0.00	T>A	80L>Q
37	220851	yafL	20	0.00	0.00	100.00	0.00	0.00	0.00	T>G	12S>S	18	5.56	0.00	88.89	5.56	0.00	0.00	T>G	12S>S
38	223592	lfhA	8	0.00	25.00	0.00	75.00	0.00	0.00	T>CT		14	n/a	n/a	n/a	n/a	n/a	n/a		

Compare mutations detected in two or more samples aligned to the same reference

- Negative mutations shown in green
- Called mutations shown in blue or purple for known SNPs

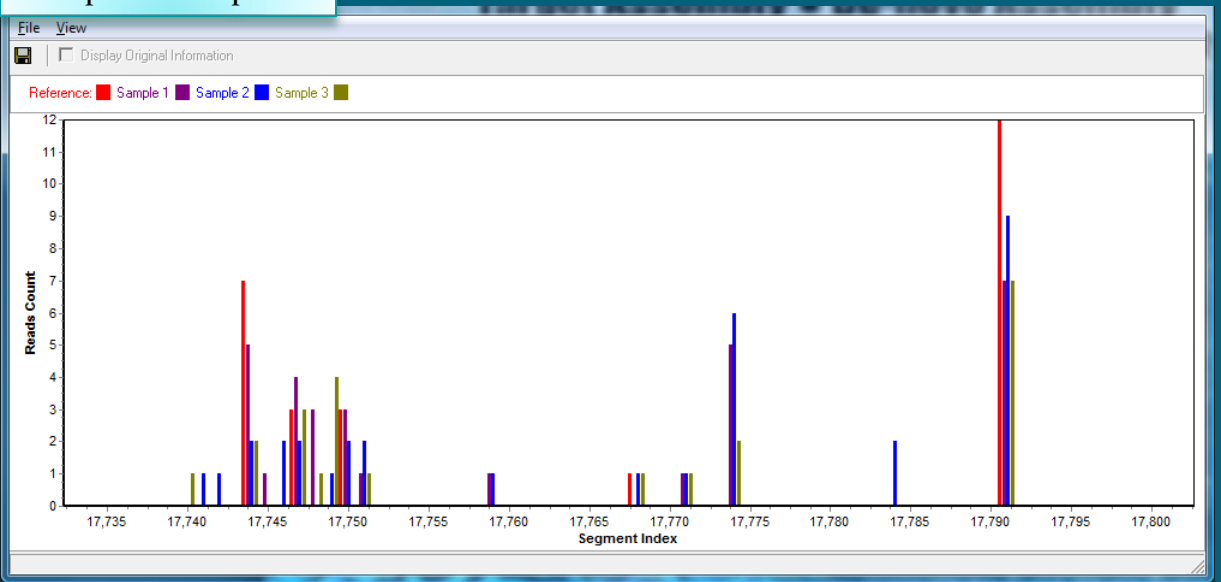
NextGENe™
Next Generation Sequencing Software

Reports: Expression

Expression Report

Segment	Description	Start	End	Length	Max Counts	Average Counts	Reads Counts	Forward Reads	RPKM
1	NC_010473; Unassigned1	1	190	190	28	19.97	107	60	207.1589
2	NC_010473; ECDH10B_0001; thrL; +	190	255	66	43	31.80	61	25	339.9846
3	NC_010473; Unassigned2	256	337	82	32	21.74	53	30	237.7581
4	NC_010473; ECDH10B_0002; thrA; +	337	2799	2463	34	19.74	1385	683	206.8515
5	NC_010473; Unassigned3	2800	2801	2	18	17.00	2	2	367.8522

Comparison Report



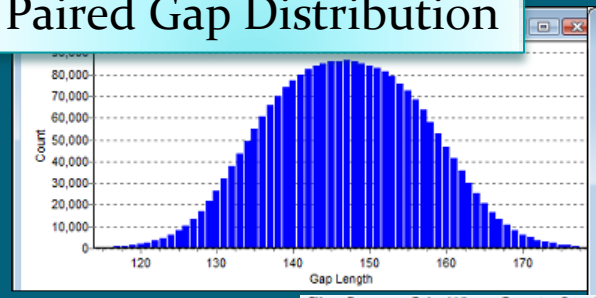
NextGENe™
Next Generation Sequencing Software

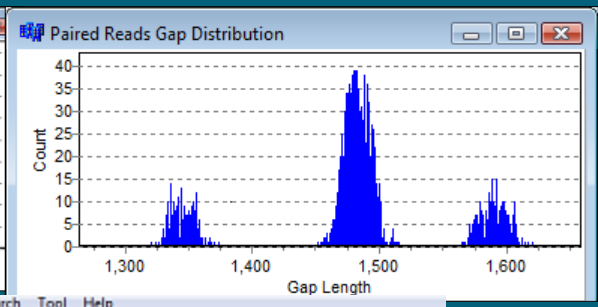
Paired End & Mate Pair Data

Paired information can be utilized for alignment and reported to detect large structural rearrangements

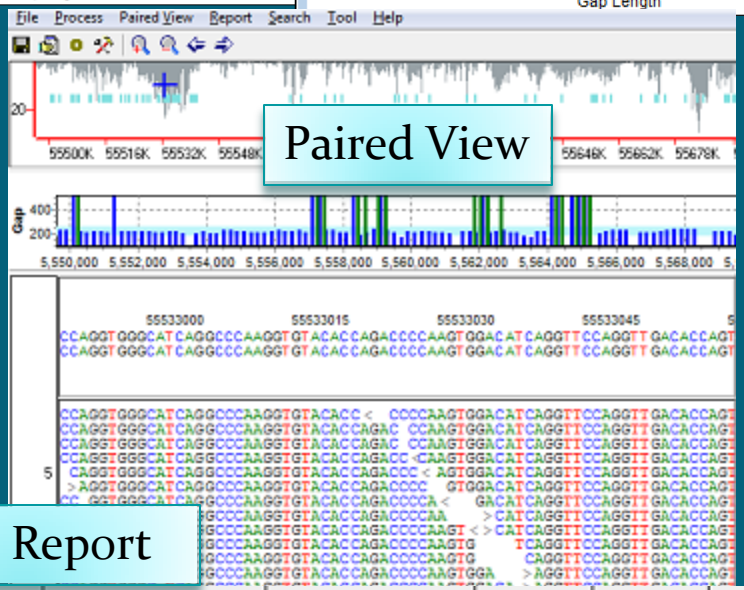
- Paired View
- Specialized Reports

Paired Gap Distribution





Paired View



Paired Report

Index	Read1 Name	Read2 Name	Read1 Start	Read2 Start	Gap Distance
1	>1378_824_691_F3	>1378_824_691_R3	34154	2723377	2689200
2	>1378_969_78_F3	>1378_969_78_R3	247948	842224	594253
3	>1378_991_574_F3	>1378_991_574_R3	437004	3946011	3508984
4	>1378_1046_612_F3	>1378_1046_612_R3	780171	858707	78513

NextGENe™
Next Generation Sequencing Software

Peak Identification

- Useful for applications such as ChIP-Seq and microRNA analyses
- Identifies regions of high coverage
- Each region is listed in the Peak Identification Report

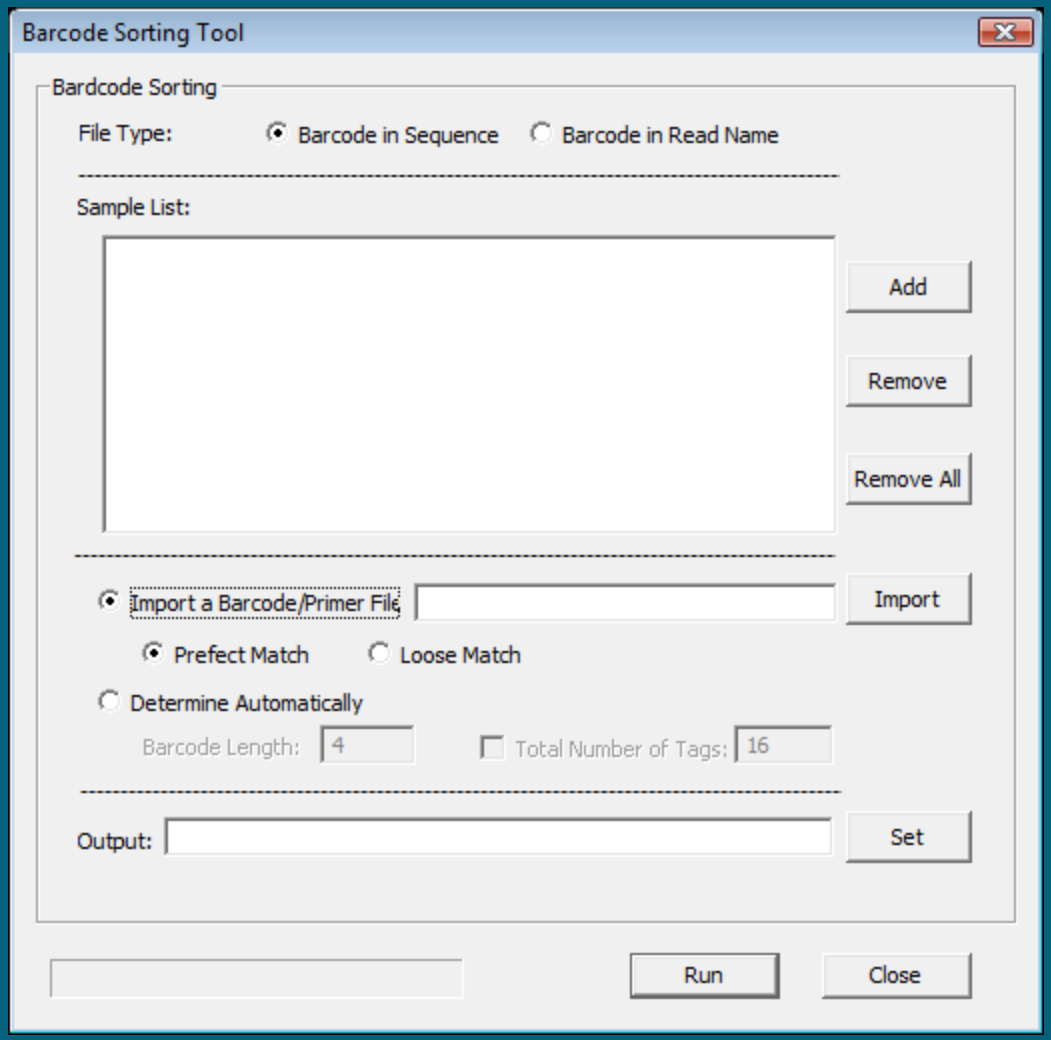
Peak Identification Report

Index	Chr	Reference Region	Chromosome Region	Length	Coverage(75%)	Transcript site	Gene Distance	Sequence
1113	IV	2618598..2618778	2618598..2618778	181	10	2618638..261873	SPT3(0)	AACGAAAAGTAAAAAGTAA(
1114	IV	2620870..2621208	2620870..2621208	339	9	2620989..262108	SHE9(+974)	AGTAGGATCGTTTGAAGAC
1115	IV	2621881..2622341	2621881..2622341	461	8	2622061..262216	RPT3(+205)	AGGACGAACAACGCCATTT
1116	IV	2622532..2622804	2622532..2622804	273	8	2622618..262271	RPT3(+856)	CGGGTTCCGATCGTGAAGT
1117	IV	2625018..2625366	2625018..2625366	349	12	2625142..262524	SXM1(+1699)	ATGAACTATCCCCATTTGC1
1118	IV	2627373..2628049	2627373..2628049	677	12	2627660..262776	UTP5(0)	ATGTATGTAAAGATAGATG

NextGENe™
Next Generation Sequencing Software

Barcoded Data

Multiplexed samples can be sorted by barcode tags



NextGENe™
Next Generation Sequencing Software

NextGENe Software

- Performance
  - Unique Condensation Tool to reduce sequencing errors
  - Rapid processing
  - High accuracy
- Flexibility
  - Data accepted from multiple systems in various formats
  - Specialized modules designed for several applications
  - Results are easily edited and exported
- Ease of Use
  - No scripting required
  - Run Wizard guides users step-by-step
  - Technical Support/Training