DNASTAR®

DNASTAR Lasergene®

Assembling and Analyzing SNPs, Genes and Gene Ontology for Multiple Next-Gen Sequencing Samples on a Desktop Computer

Thomas Schwei[1]; Timothy Durfee PhD[1]; Katherine Maxfield[1]; Amber Pollack-Berti PhD[1]; Matthew Keyser[1]; Daniel Nash[1]; Jennifer Stieren[1]; Richard Nelson PhD[1]; Schuyler Baldwin[1]; Christopher Stern[1]; Kenneth Dullea[1]; John Schroeder[1]; Pavel Pinkas[1]; Guy Plunkett III PhD[1,2]; Frederick Blattner PhD[1,2,3]

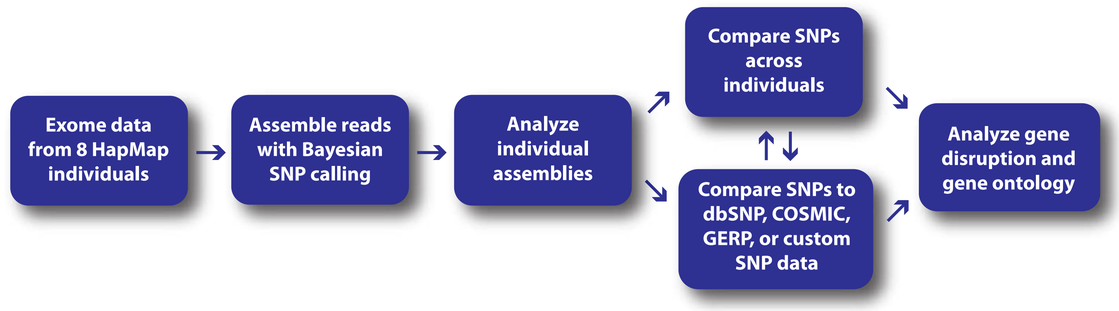
[1] DNASTAR, Inc. 3801 Regent Street, Madison, Wisconsin 53705
[2] University of Wisconsin Department of Genetics, Madison, Wisconsin 53706
[3] Scarab Genomics LLC, 1202 Ann Street, Madison, Wisconsin 53713

DNASTAR Software Pipeline

DNASTAR offers an integrated suite of software for assembling and analyzing sequence data from all major next-generation sequencing platforms supporting key workflows on a desktop computer. Supported workflows include reference-guided and *de novo* genome and transcriptome assembly and analysis, metagenomics sample assembly, targeted resequencing, exome assembly, and RNA-Seq, ChIP-Seq and miRNA alignment and analysis.

An advanced application that integrates some of the most powerful functionality in the software includes assembling and analyzing multiple samples using one reference template; probabilistic identification of SNPs, small indels and genotype calls with known variants correlated to their dbSNP and COSMIC IDs and GERP scores; review of SNPs from multiple samples within a single project; identification of structural variations; and, for large multi-sample projects with hundreds of individual data sets, tools for SNP quantitation, filtering, set comparison, clustering and indication of the gene disruption impact from called SNPs, as well as gene ontology.

Interactive views within the software facilitate fast, comprehensive analysis, helping scientists move quickly from raw next-gen sequencing data to genetic and genomic impact, including gene ontology. By using innovative algorithms within the software, scientists can have all of the assembly and analysis capabilities available to them on their desktop computer, supporting large data sets generated by high-throughput next-gen sequencing instruments and large numbers of small data sets beginning to be produced by bench-top next-gen sequencers.



Assemble reads with Bayesian SNP calling

Table 1. Statistics for human exome assemblies*

Data†	Description	Reads (in Millions)‡	Depth of Coverage§	Time to Assemble (hrs)
NA12156 (SRX005923)	CEPH/UTAH PEDIGREE 1408	139	58X	1.8
NA12878 (SRX005924)	CEPH/UTAH PEDIGREE 1463	187	90X	1.9
NA18507 (SRX005925)	YORUBA IN IBADAN, NIGERIA	167	86X	1.8
NA18517 (SRX005926)	YORUBA IN IBADAN, NIGERIA	150	78X	1.7
NA18555 (SRX005927)	HAN CHINESE IN BEIJING, CHINA	147	73X	1.7
NA18956 (SRX005928)	JAPANESE IN TOKYO, JAPAN	156	87X	1.7
NA19129 (SRX005929)	YORUBA IN IBADAN, NIGERIA	159	85X	1.8
NA19240 (SRX005930)	YORUBA IN IBADAN, NIGERIA	163	90X	1.8

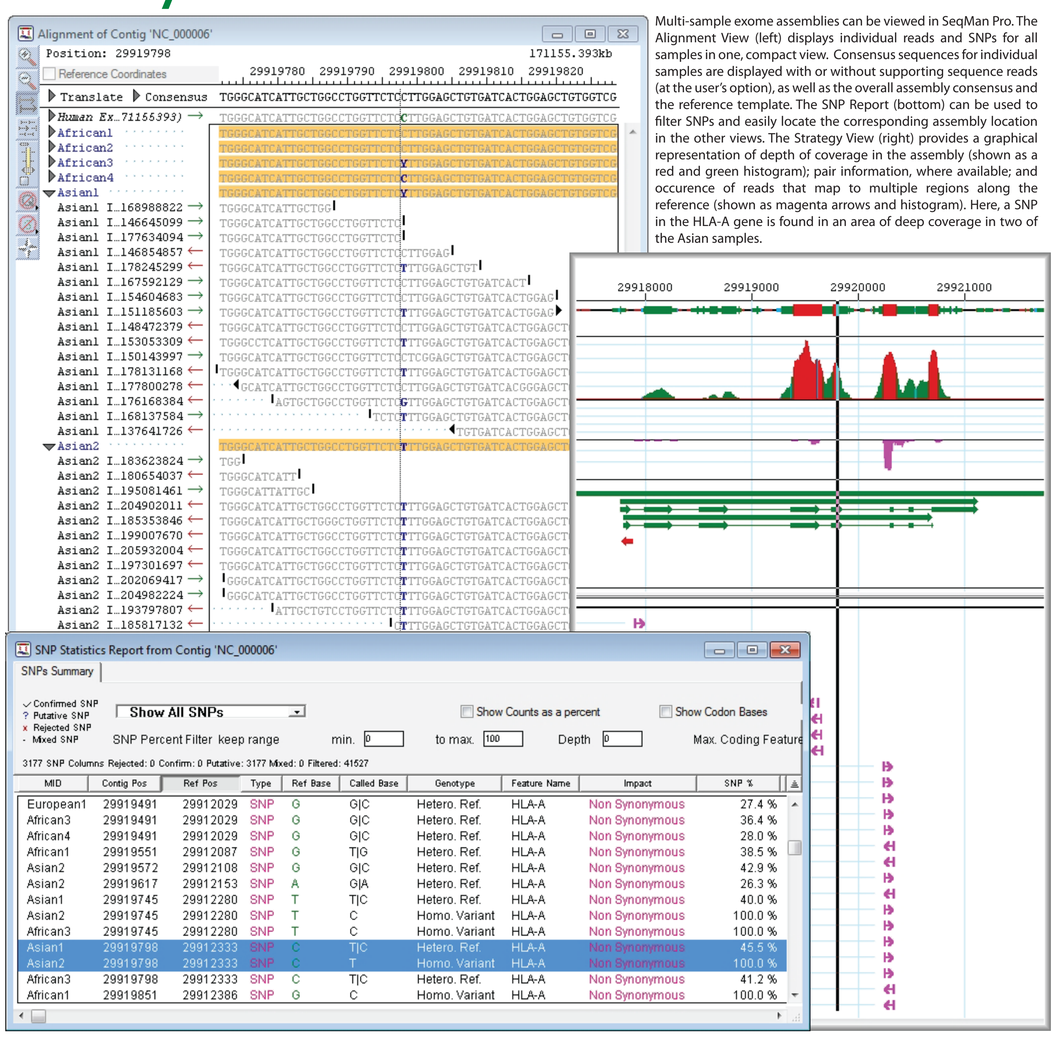
*All alignments performed on Windows 7 x64; 16 GB RAM, 6 x 2 TB 7200 RPM hard drives.

†Ng, SB, et al. Targeted capture and massively parallel sequencing of twelve human exomes. Nature 2009; 461(7261):272-6.

‡Illumina reads 76bp in length, unpaired.

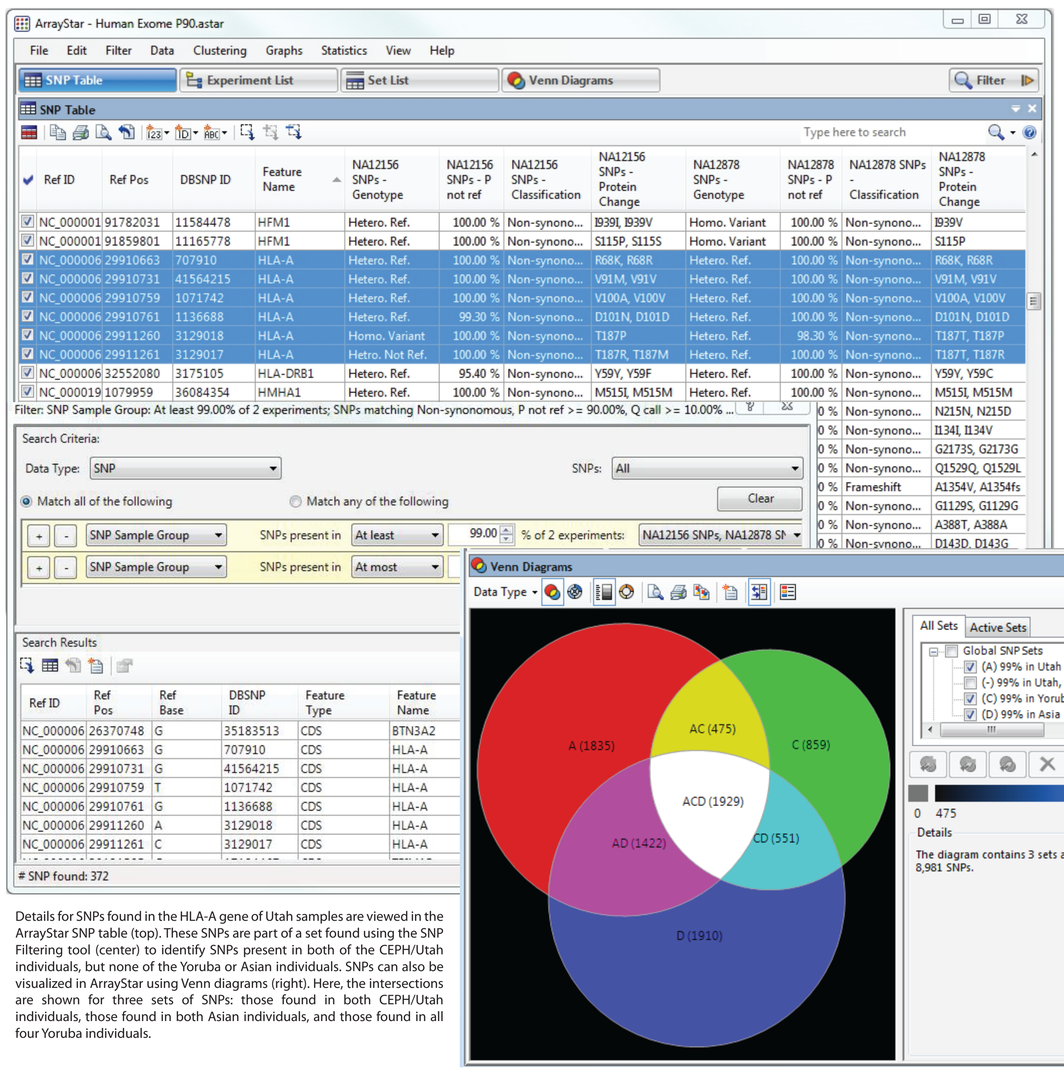
§Coverage calculated by summing the average coverage across the 20604 CDS entries and dividing by 20604. Assembly parameters included mer size = 19 , mer skip =2.

Analyze individual assemblies



Multi-sample exome assemblies can be viewed in SeqMan Pro. The Alignment View (left) displays individual reads and SNPs for all samples in one, compact view. Consensus sequences for individual samples are displayed with or without supporting sequence reads (at the user's option), as well as the overall assembly consensus and the reference template. The SNP Report (bottom) can be used to filter SNPs and easily locate the corresponding assembly location in the other views. The Strategy View (right) provides a graphical representation of depth of coverage in the assembly (shown as a red and green histogram); pair information, where available; and occurence of reads that map to multiple regions along the reference (shown as magenta arrows and histogram). Here, a SNP in the HLA-A gene is found in an area of deep coverage in two of the Asian samples.

Compare identified SNPs across individuals and to known SNPs



Details for SNPs found in the HLA-A gene of Utah samples are viewed in the ArrayStar SNP table (top). These SNPs are part of a set found using the SNP Filtering tool (center) to identify SNPs present in both of the CEPH/Utah individuals, but none of the Yoruba or Asian individuals. SNPs can also be visualized in ArrayStar using Venn diagrams (right). Here, the intersections are shown for three sets of SNPs: those found in both CEPH/Utah individuals, those found in both Asian individuals, and those found in all four Yoruba individuals.

Analyze gene disruption and gene ontology

ArrayStar also provides filtering tools to create Gene Sets based on SNP data for multiple samples. Filtering criteria may include SNP type, occurence of SNPs in one or more samples, occurence of splice site changes, and GERP values. Here the Gene Table (left) is shown for the genes with coding change SNPs in all four Yoruba individuals, including the HLA-A gene. ArrayStar also displays ontology information for selected genes (top).