Molecular Characterization of Tumors Using Next-Generation Sequencing

Using BaseSpace to visualize molecular changes in cancer.

Tumor-Normal Sequencing Data in BaseSpace

To enable researchers new to next-generation sequencing (NGS), Illumina has provided an example data set from tumor-normal whole genome sequencing. It is available for viewing in BaseSpace, Illumina's cloud computing platform. For more information, visit the BaseSpace page on the Illumina website

Project Description

The DNA was extracted from the HCC1187 breast ductal carcinoma cell line and a matching lymphoblastoid cell line from the same individual. Briefly, 500 ng of DNA was used to create libraries using the early access version of the TruSeq DNA PCR-Free Sample Prep Kit. Please note that the lymphoblastoid line HCC1187 BL does not represent a true matched normal from the same tissue type as the tumor, but serves as a reasonable surrogate.

The libraries were sequenced on four and eight lanes of a HiSeq 2000 flow cell for the normal and tumor, respectively, using 100 bp paired-end reads. Data, analyzed using the pre-release version of the Illumina Cancer Sequencing Workflow showed an average coverage of 40x for the normal and 90x for the tumor DNA with about 96% of loci covered with 10 or more reads. A subset of the output files were uploaded to BaseSpace for public sharing in accordance with the terms of a licensing agreement with UT Southwestern, the owners of the cell line[1].

Cancer Sequencing Workflow Informatics Pipeline

Analysis was performed using and early access version of the Cancer Sequencing Workflow, which includes alignment using the Isaac aligner[2], somatic variant calling using Strelka[3], and annotation. Small somatic variants are reported with RefSeq annotations, COSMIC annotations, functional consequence predictions, and regulatory motifs

[1] HCC cell lines were invented by Drs. Adi F. Gazdar and John D. Minna at the University of Texas Southwestern Medical Center. Rights in and to the HCC cell lines, progeny, and unmodified derivates thereof belong to the Board of Regents of The University of Texas System. Illumina, Inc. has obtained permission from the Board of Regents of The University of Texas System through the University of Texas Southwestern Medical Center to use the HCC cell lines and publish the data and results herein displayed.

[2] Raczy C, Petrovski R, Saunders C, Chorny I, Kruglyak S, Margulies E, Chuang HY, Kallberg M, Kumar SA, Liao A, Little KM, Stromberg M, Tanner S. Isaac: Ultra-fast whole genome secondary analysis on Illumina sequencing platforms. 2013

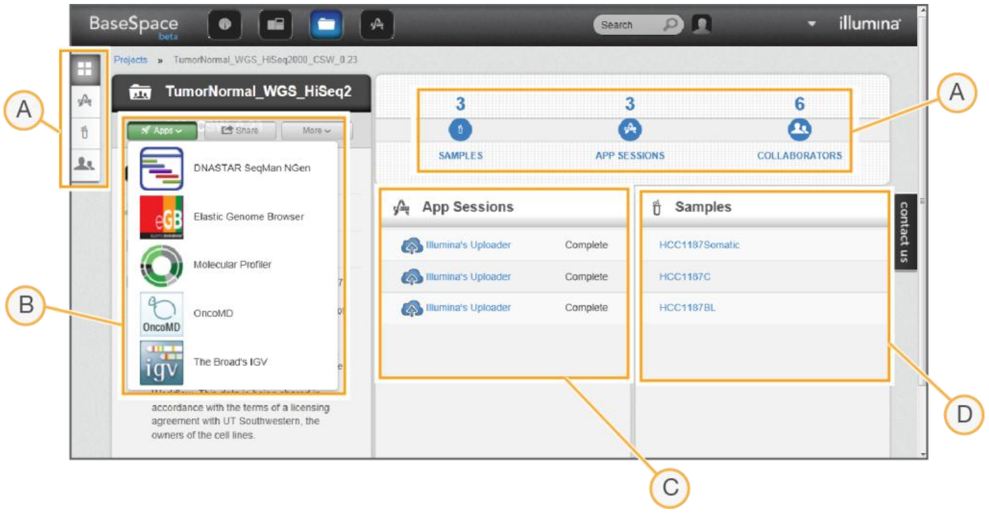
[3] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012 Jul 15;28(14):1811-7

Navigating a Tumor-Normal NGS Data Set in BaseSpace

BaseSpace provides a powerful infrastructure for data sharing, storage, and analysis. This section introduces BaseSpace features and describes how to navigate a typical tumor-normal whole genome sequencing project.

The tumor-normal sequencing project[4] opens to the Project Overview Session. Use the quick launch buttons to navigate to samples with data, app sessions results, and collaborator information.

Figure 1. Project Overview Screen



Illumina's Uploader Complete
Illumina's Uploader Complete
Illumina's Uploader Complete

HCC1187Somatic
HCC1187C
HCC1187BL

accordance with the terms of a licensing agreement with UT Southwestern, the owners of the cell lines.

A. Quick Launch Buttons—Links to samples with data, app sessions results, and collaborator information.

B. BaseSpace Apps—Links to applications used to analyze and view data.

C. Apps Sessions Pane—Links to results of the app sessions (BAM files, VCF files, sequencing and summary reports).

D. Samples Pane—Lists samples with links to data in FASTQ file format.

The tumor-normal sequencing project includes three data sets in compressed FASTQ[5] file format. These data sets are the result of whole-genome sequencing of the breast ductal carcinoma cell line, HCC1187C, and a normal lymphoblastoid cell line established from the same individual, HCC1187BL. The third data set, HCC1187Somatic, is generated post-single sample analysis and is the result of an algorithmic subtraction between the normal and tumor data. Tumor/Normal data subtraction, data alignment, and variant calling were performed using an early access version of the Cancer Sequencing Workflow.

[4] A project is a set of samples and corresponding app results that are managed by the data owner, but can also be shared by the owner with their collaborators. For full access to the tumor-normal sequencing project, register for a BaseSpace account or log in using your MyIllumina credentials.

[5] FASTQ files are text-based files that contain sequence information and quality scores per base, as well as information about the instrument, flow cell ID, and position of the read on the flow cell.

Table 1. Samples in the Tumor-Normal Sequencing Project

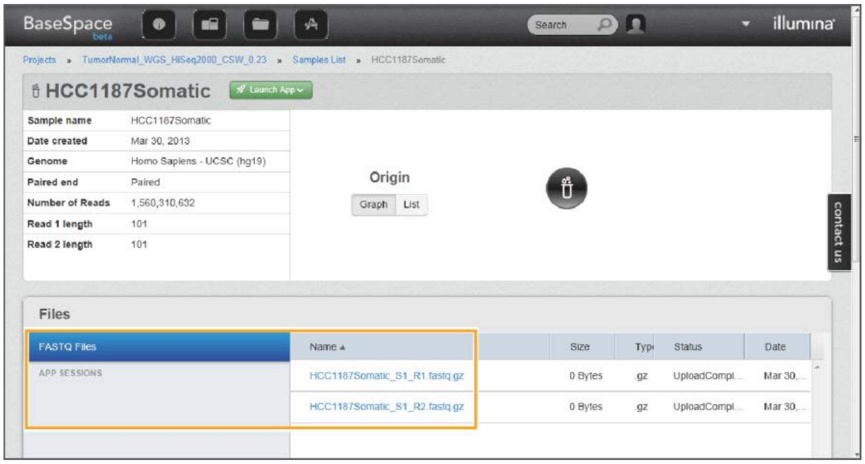
Sample ID	Ref Genome	Read Length	Description
HCC1187Somatic	H. sapiens UCSC hg19	2 x 100 bp	Data set resulting from subtraction of sequence shared between tumor and normal cell line
HCC1187C	H. sapiens UCSC hg19	2 x 100 bp	Breast ductal carcinoma cell line (tumor)
HCC11878BL	H. sapiens UCSC hg19	2 x 100 bp	EBV-transformed lymphoblastoid cell line from same individual (normal)

Sample Details and Access to FASTQ Files

Each sample name listed in the Samples section of the Project Overview screen is a link to detailed information about that sample, such as read length, total number of reads, and whether the data are from single-read or paired-end sequencing.

Links to FASTQ files for the sample are provided in the lower pane of the sample details screen. In this project, the FASTQ files are empty and uploaded as placeholders only.

Figure 2. Sample Information and Link to FASTQ Files



Projects » TumorNormal_WGS_HiSeq2000_CSW_0.23 » Samples List » HCC1187Somatic

HCC1187Somatic

Sample name	HCC1187Somatic
Date created	Mar 30, 2013
Genome	Homo Sapiens - UCSC (hg19)
Paired end	Paired
Number of Reads	1,560,310,632
Read 1 length	101
Read 2 length	101

Files

FASTQ Files

APP SESSIONS

Name ▲	Size	Typ	Status	Date
HCC1187Somatic_S1_R1.fastq.gz	0 Bytes	.gz	UploadCompl...	Mar 30,...
HCC1187Somatic_S1_R2.fastq.gz	0 Bytes	.gz	UploadCompl...	Mar 30,...

The Samples section contains sample details and links to FASTQ files.

App Session and Access to BAM Files, VCF Files, and Reports

The App Session pane within the Project Overview provides access to results of various data processes, such as data uploads and data analysis performed using one or more of the BaseSpace Apps.

The data sets for the various samples can be accessed through the Illumina Uploader hyperlinks within the App Session pane. The set of files for the subtracted data, HCC1187Somatic, consists of all the different VCF files that catalog single nucleotide variants (SNVs), insertions and deletions (somatic Indels), structural variants (SVs) and copy number alterations (CNAs). Also available in this directory is a Somatic Summary Report[6] from the analysis of the subtracted data (HCC1187Somatic).

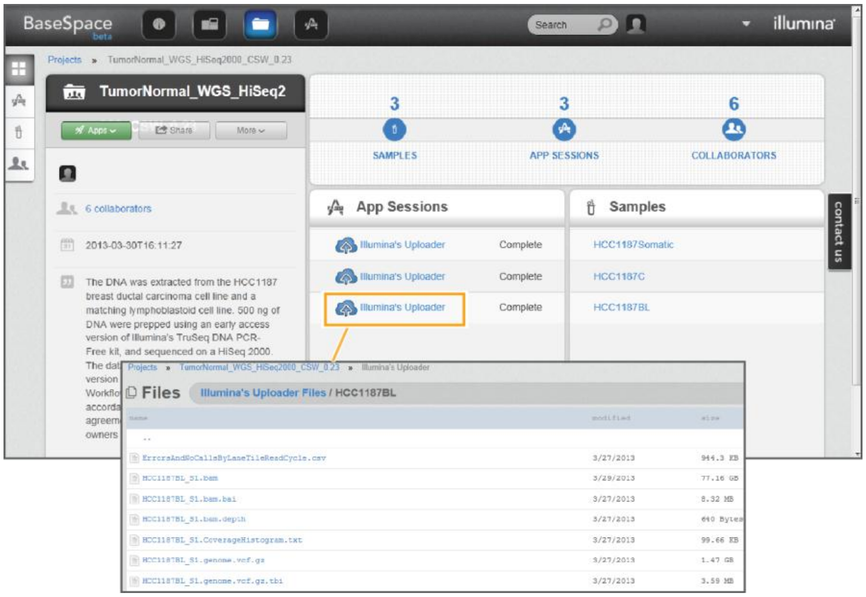
For the single-genome tumor or normal sample data, HCC1187C and HCC1187BL, respectively, the set of files consists of alignment files in BAM[7] file format, variant calls in VCF and genome VCF[8] file formats, and files that show run and sequencing data metrics.

[6] Somatic Summary Report is an output of the Cancer Sequencing Workflow that contains information about the sample, the data generated, and the variants that have been cataloged within that sample data set.

[7] BAM files (*.bam) are compact, indexable tab-delimited text files containing sequence alignment data in binary format. BAM is the recommended input file format for Integrative Genomics Viewer (IGV).

[8] VCF is a text file that contains meta-information, header and data lines that catalog and describe a specific position in the genome where a variant is detected in high detail. Information includes chromosome position, reference and variant allele, genotype quality, depth of coverage per allele, depth of coverage, among others. A genome VCF is a more fully annotated VCF that includes information on variant allele frequencies, predicted functional effects of the variants, and known associations with disease among others. See https://sites.google.com/site/gvcftools/home/about-gvcf.

Figure 3. Accessing BAM, VCF, and Summary Reports



2013-03-30T16:11:27

The DNA was extracted from the HCC1187 breast ductal carcinoma cell line and a matching lymphoblastoid cell line. 500 ng of DNA were prepped using an early access version of Illumina's TruSeq DNA PCR-Free kit, and sequenced on a HiSeq 2000. The dat version Workflo accorda agreem owners

App Sessions	
Illumina's Uploader	Complete
Illumina's Uploader	Complete
Illumina's Uploader	Complete

Samples
HCC1187Somatic
HCC1187C
HCC1187BL

Projects » TumorNormal_WGS_HiSeq2000_CSW_0.23 » Illumina's Uploader

Files Illumina's Uploader Files / HCC1187BL

name	modified	size
..		
ErrorsAndNoCallsByLaneTileReadCycle.csv	3/27/2013	944.3 KB
HCC1187BL_S1.bam	3/29/2013	77.16 GB
HCC1187BL_S1.bam.bai	3/27/2013	8.32 MB
HCC1187BL_S1.bam.depth	3/27/2013	640 Bytes
HCC1187BL_S1.CoverageHistogram.txt	3/27/2013	99.66 KB
HCC1187BL_S1.genome.vcf.gz	3/27/2013	1.47 GB
HCC1187BL_S1.genome.vcf.gz.tbi	3/27/2013	3.59 MB

The Illumina Uploader provides access to BAM files, VCF files, and summary reports.

Somatic Summary Report: A Summary of Somatic Variants

After analysis of the tumor and normal data using the Cancer Sequencing Workflow, the results are reported in the *Somatic Summary Report*. The report documents all identified somatic variants in the data set, in both graphical and tabular format, and can be downloaded from the Somatic folder in the Apps Sessions in BaseSpace.

An estimate of the purity[9] and ploidy[10] is provided. In this case, the ploidy of the sample has been estimated at 2.4 indicating that some chromosomal loci have been duplicated, and the purity estimate was reported to be 0.8, which suggests that for every 10 reads, 8 are derived from the tumor and 2 are derived from the contaminating normal or other molecularly distinct subclones.

Purity/Ploidy Estimate

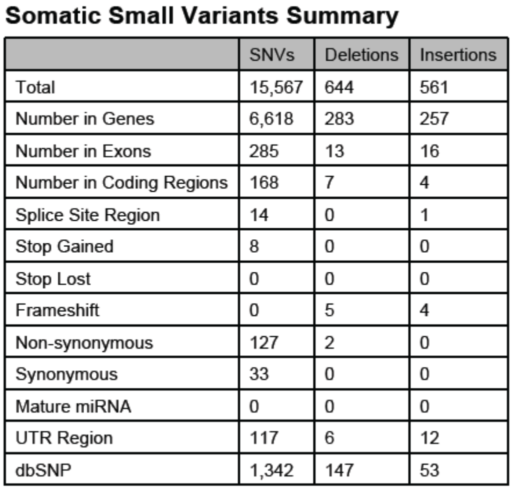
Sample	Purity	Ploidy
Cancer	0.8	2.4

When presented in tabular format, the variants are cataloged as small variants, representing single nucleotide changes or smaller indels (< 50 bases), or structural variants that include large insertions, duplications and deletions, copy number aberrations, and genomic rearrangements.

The **Somatic Small Variant Summary** table describes the small variant profile for the somatic data set relative to its DNA location (e.g., in exons, splice junctions, UTRs, etc.), the potential effect on the protein level (nonsynonymous, frameshift, etc.) as well as how many variants have previously been cataloged in dbSNP.

[9] Purity is an estimate of the homogeneity somatic data set with a purity of 1 indicating that all reads have been derived from the tumor sample. With increasing contaminating reads from the normal, the purity estimate decreases. HCC is a cell-line sample, so a purity < 1 is less likely due to the contamination of normal reads. The cell line might contain heterogeneous subclones.

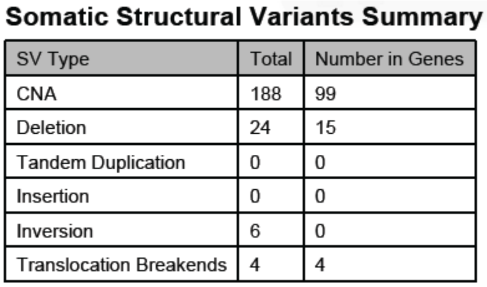
[10] Ploidy is the approximate number of sets of chromosomes in the sample as estimated from the data**.** A ploidy of greater than 2 indicates regions of chromosomal duplication and a ploidy of less than 2 indicates regions of chromosomal loss.



Somatic Small Variants Summary

	SNVs	Deletions	Insertions
Total	15,567	644	561
Number in Genes	6,618	283	257
Number in Exons	285	13	16
Number in Coding Regions	168	7	4
Splice Site Region	14	0	1
Stop Gained	8	0	0
Stop Lost	0	0	0
Frameshift	0	5	4
Non-synonymous	127	2	0
Synonymous	33	0	0
Mature miRNA	0	0	0
UTR Region	117	6	12
dbSNP	1,342	147	53

The **Somatic Structural Variants Summary** table reports the total number of identified events for each large variant type and the total number of each of these variants that are found within genes.

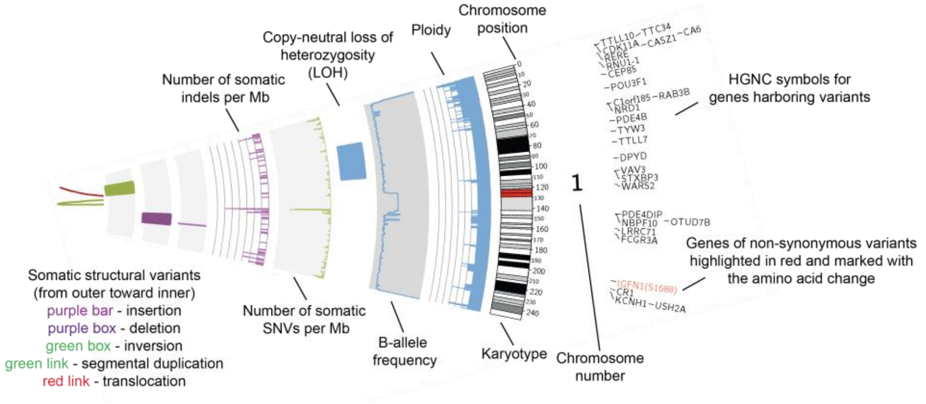


Somatic Structural Variants Summary

SV Type	Total	Number in Genes
CNA	188	99
Deletion	24	15
Tandem Duplication	0	0
Insertion	0	0
Inversion	6	0
Translocation Breakends	4	4

In addition to reporting the variants in a tabular format, the report also catalogs variants in a **Circos plot**, which is a graphical representation that provides a snapshot of the somatic mutations in a circular format. The Circos plot shows relationships between mutations, especially with translocations that, in most cases, involve two completely different chromosomes. The following figure shows the various tracks within a typical Circos plot from tumor-normal sequencing.

Figure 4. Various Tracks in a Typical Circos Plot



- **Structural variants** are cataloged in the inner tracks, closest to the center. Briefly, translocations are represented by red lines in the center of the circle, the ends of which show the two loci involved in the event. Green lines in the center represent segmental duplications while green boxes show inversions. Insertions and deletions are represented by purple bars and boxes, respectively.

- Immediately adjacent to structural variant tracks are descriptive tracks, such as the **somatic indels or SNV per Mb** tracks that show highly variant regions.
- The **copy-neutral loss of heterozygosity (CN-LOH)** track immediately follows and shows events that lead to loss of certain regions without a net effect on copy number. These events are also known as uniparental disomy and in tumor cells may be biologically equivalent to a second hit according to the Knudson hypothesis leading to the development of cancer.
- The **B-allele frequency** track provides information on the proportion of the total allele signal that can be explained by a single allele[11]. This track enables detection of low-level mosaic gains and losses.
- The **ploidy** track shows the fluctuation in copy numbers, followed by the track for karyotype and chromosomal position.
- The outermost track identifies the **genes within which variants are found**. Genes that have nonsynonymous variants are highlighted in red with the corresponding amino acid change noted.

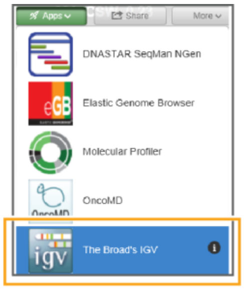
The Circos plot and summary tables, specifically generated from this tumor-normal sequencing data set, highlight the dynamism of the cancer genome by showing the numerous variants that were identified within a typical cancer sequencing data set. Among the variants immediately apparent are the translocations between chromosomes 1 and 8, and 7 and 10 as shown by the red lines in the center of the circle. Another observation is the frequent occurrence of copy-neutral LOH and changes in the ploidy, which notes various duplication and deletion events within the cancer genome suggesting genomic instability. And finally, the high number of nonsynonymous mutations within this data set suggests that the normal function of the various proteins within which these changes are found, as well as the signals conveyed by the pathways of which they are a part, might have been altered. Together, this plot highlights how the accumulation of mutations leads to a drastically altered cancer genome whose normal checkpoints have been dysregulated.

Visualizing Data Using Integrative Genomics Viewer

The Integrative Genomics Viewer (IGV)[12] is a fully-featured genome browser developed at the Broad Institute by Robinson and colleagues[13]. A web-downloadable version of this tool is available in BaseSpace to enable navigation through NGS data sets using alignment files and variant data in BAM and VCF formats, respectively, and to facilitate downstream analysis.

To start the BaseSpace IGV session, select the IGV app icon from the App drop-down list in Project Overview. If loading BaseSpace IGV for the first time, accept the terms and conditions for use of this application. Java is required for this application.

Figure 5. Accessing BaseSpace IGV



Access BaseSpace IGV from the Apps menu.

[11] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. 2011 Nature Rev Genet. 12: 363-376.

[12] For more information about IGV, see http://www.broadinstitute.org/igv/.

[13] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G,Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24-6.

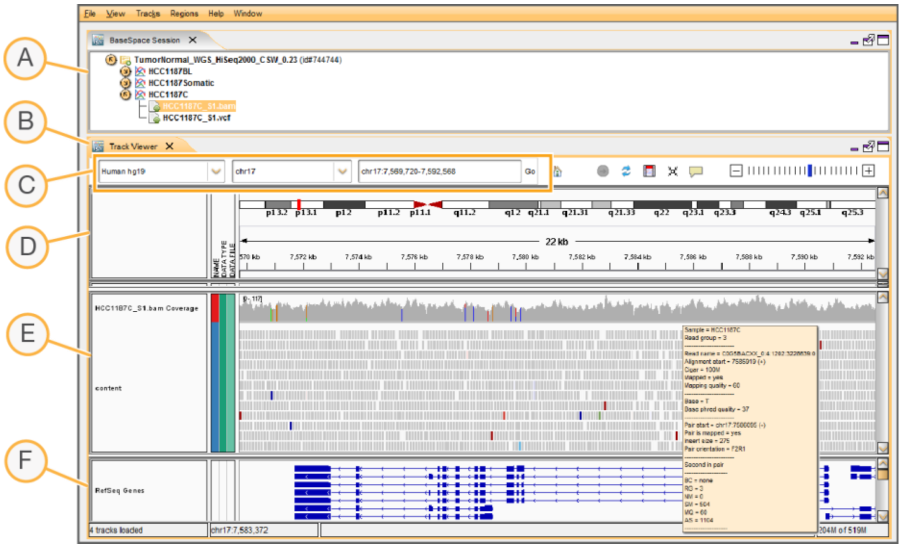
IGV User Interface

Data visualization and inspection of variant calls are facilitated by allowing multiple tracks, such as the aligned data (BAMs) from the sequenced samples and the different variant tracks (VCFs), to be compared against the reference genome as well as against each other. BaseSpace IGV is pre-loaded with the reference genome builds, and RefSeq genes track for human and a number of other model organisms.

The IGV user interface consists of the BaseSpace Session tab and the Track Viewer:

- The **BaseSpace Session** tab shows the directory structure of files as they are organized in BaseSpace. Files within each sample directory are accessible by clicking on the folder.
- The **Track Viewer** tab allows data browsing and inspection of specific genomic positions within the selected data sets. The Track Viewer consists of navigation controls that allow entry of specific locus information, a display window that allows visualization of the reads and attributes that are assigned to the data set, and the feature track viewer that displays features within the selected genomic locus.

Figure 6. BaseSpace IGV User Interface



TumorNormal_WGS_HiSeq2000_CSW_0.23 (id#744744)
HCC1187BL
HCC1187Somatic
HCC1187C
HCC1187C_S1.bam
HCC1187C_S1.vcf

Sample = HCC1187C
Read group = 3

Read name = C0G5BACXX_0:4:1202:3228639:0
Alignment start = 7586019 (+)
Cigar = 100M
Mapped = yes
Mapping quality = 60

Base = T
Base phred quality = 37

Pair start = chr17:7586095 (-)
Pair is mapped = yes
Insert size = 275
Pair orientation = F2R1

Second in pair

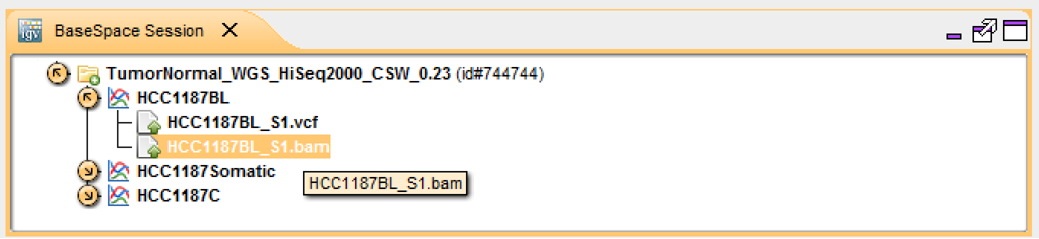
BC = none
RG = 3
NM = 0
SM = 504
MQ = 60
AS = 1104

A. BaseSpace Session Tab—Shows files as they appear in the BaseSpace directory.
B. Track Viewer Tab—Enables data browsing and inspection of genomic positions.
C. Navigation Fields
D. Chromosome Ideogram
E. Track Pane
F. Feature Track Pane

Working in BaseSpace IGV

To begin visualizing data in IGV, load files from the directories within the BaseSpace Session tab by double-clicking the file name. When loaded, the data appear on the Track Viewer tab. For this project, the BAM files for tumor and normal samples are loaded (HCC1187C_S1.bam and HCC1187BL_S1.bam). The variant data (HCC1187Somatic) is also loaded in the form of the various annotated VCF files (SNV, CAN, SV, LOH, and InDels).

Figure 7. Expanded Directories in the BaseSpace Session Tab



BaseSpace Session ✕

TumorNormal_WGS_HiSeq2000_CSW_0.23 (id#744744)
HCC1187BL
HCC1187BL_S1.vcf
HCC1187BL_S1.bam
HCC1187Somatic
HCC1187C

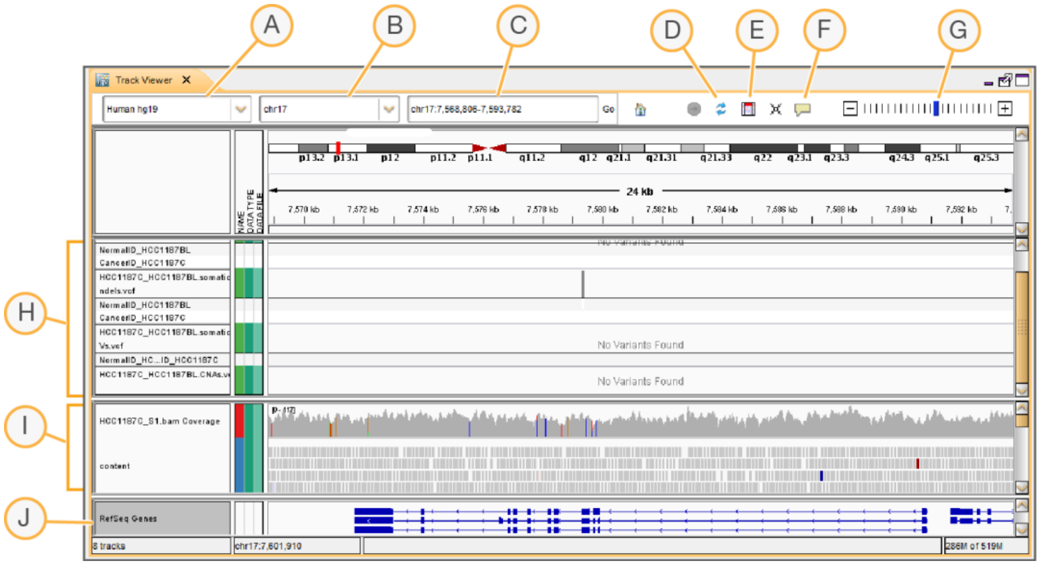
HCC1187BL_S1.bam

Click a file name to load the file and view data in the Track Viewer.

When the relevant sample tracks are loaded onto the Track Viewer, the BaseSpace Session tab can be closed by clicking the "X" on the tab or by clearing the BaseSpace Session checkbox under the Window menu at the top of the screen. Closing the BaseSpace Session tab provides more usable space for viewing the data.

By default, the topmost panel of the Track Viewer shows the chromosome ideogram. The panels below it display data as it is loaded. VCF files appear in the upper panels, single-genome BAM files appear in the lower panel. The lower-most panel shows the RefSeq Genes track, which allows the visualization of the known gene structure within a given chromosomal locus. For further expansion of the viewing area for any given track, the panels can be expanded or contracted by dragging the perimeter of the panel.

Figure 8. Features of the Track Viewer Tab



A. Reference Genome Field
B. Chromosome Field
C. Coordinate Field
D. Refresh Button
E. Link to Regions of Interest
F. Pop Up Enable/Disable
G. Zoom
H. Variant Tracks (VCFs)
I. Single-Genome Track (BAMs)
J. RefSeq Genes Track

With the tracks loaded, the reference genome to which the data has previously been aligned is selected, which in this case is hg19. Select the specific chromosome from the drop-down list to inspect entire chromosomal regions. Alternatively, specify a specific locus by typing either a gene name or chromosomal coordinates. These selections determine the chromosomal regions to be displayed.

Visualization of Variants within the Tumor-Normal WGS Project

The following two examples illustrate how to view variants in IGV:

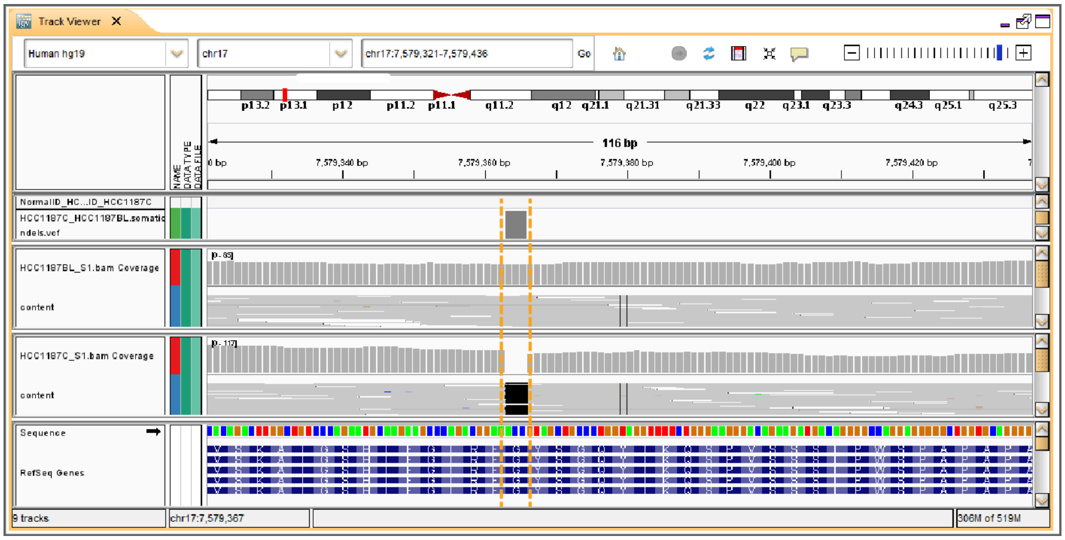
- *Example 1: A 3-nucleotide deletion in TP53 (chr17:7,579,321-7,579,436)*

  Within the TP53 locus, a small 3-nt homozygous in-frame deletion is identified (AACC > A), the functional consequence of which is the elimination of a glycine residue on the protein (p.G108del). This variant has previously been reported (COSMIC ID: 13119) by a separate study that also aimed to identify tumor-specific variants within the HCC1187 breast ductal carcinoma cell line. Visual inspection of this specific locus using IGV shows that there is no read coverage of this locus in the tumor data. In contrast, there are about 47 reads that cover this locus in the normal data.

  1. To view this data, enter the coordinate **chr17:7,579,321-7,579,436** in the Coordinates field.
  2. Click **Go** to navigate to the locus.
  3. Click the **Refresh** button to update the displayed tracks.

  Based on the chromosome ideogram, this view shows a window of 116 bases and only one variant track has an annotation within this locus: the somatic structural variant track (HCC1187C_HCC1187BL.somaticSVs.vcf).
  Based on the RefSeq Gene track, we can clearly see that this small deletion sits right on top of a codon, in this case, codon 108 that encodes a glycine residue.

Figure 9. Nucleotide Deletion in TP53



Human hg19

chr17

chr17:7,579,321-7,579,436

NormalID_HC...ID_HCC1187C

HCC1187C_HCC1187BL.somaticIndels.vcf

HCC1187BL_S1.bam Coverage

content

HCC1187C_S1.bam Coverage

content

Sequence

RefSeq Genes

9 tracks

chr17:7,579,367

306M of 519M

A three nucleotide deletion in an exon of TP53 leads to the loss of a residue of the protein.

- *Example 2: A translocation event between chromosomes 1 and 8*

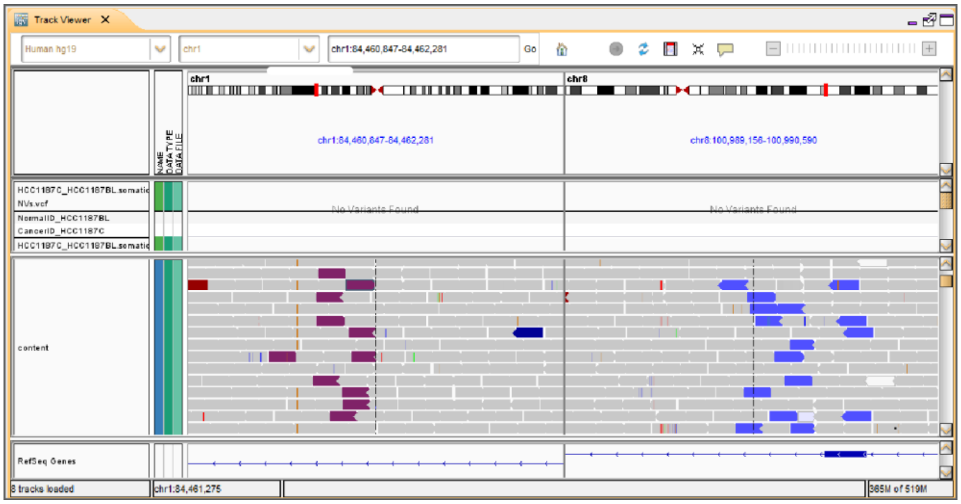
  In this data set, a translocation event was identified involving loci in chromosome 1 (chr1:84,460,847-84,462,281) and chromosome 8 (chr8:100,986,731-100,993,474). This variant has previously been reported (COSMIC ID: 17168) by an independent study, and is categorized as an inter-chromosomal mutation of unknown type with the breakpoint mapped as being from chromosome1: 84461564 to chromosome8:100990098 based on human genome build GRCh37.

  1. To further inspect the translocation, enter either one of the following coordinates into the navigation field highlighted below.

chr1:84,460,847-84,462,281
chr8:100,986,731-100,993,474

2. Click **Go**. Colored arrows appear in the panel showing the single-genome tumor BAM file, and indicate reads whose pair maps to a different chromosomal locus, indicating a rearrangement or translocation event.
3. Right-click on a colored arrow, and select **View mate region in split**. A side-by-side view appears that shows the two loci involved in the translocation event.
4. For a more global view of all reads within the specified loci, right-click on a read and select Squished, instead of Expanded.

Figure 10. Translocation Event Between Chromosomes 1 and 8



This technical note explains how to view some of the changes identified in the tumor-normal sequencing data set. Several other genomic changes are present and can be explored independently using the Broad's IGV.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com
For research use only

Pub. No. 770-2013-011 Current as of 05 April 2013