



## NIH PUBLIC ACCESS

## Author Manuscript

*Nat Biotechnol.* Author manuscript; available in PMC 2012 May 07.

Published in final edited form as:

*Nat Biotechnol.* 2011 January ; 29(1): 24–26. doi:10.1038/nbt.1754.

## Integrative Genomics Viewer

**James T. Robinson<sup>1</sup>, Helga Thorvaldsdóttir<sup>1</sup>, Wendy Winckler<sup>1</sup>, Mitchell Guttman<sup>1,2</sup>, Eric S. Lander<sup>1,2,3</sup>, Gad Getz<sup>1</sup>, and Jill P. Mesirov<sup>1</sup>**<sup>1</sup>Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA<sup>2</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA<sup>3</sup>Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA

### To the Editor

Rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data, including data from exome and whole genome sequencing, epigenetic surveys, expression profiling of coding and non-coding RNAs, SNP and copy number profiling, and functional assays. Analysis of these large, diverse datasets holds the promise of a more comprehensive understanding of the genome and its relation to human disease. Experienced and knowledgeable human review is an essential component of this process, complementing computational approaches. This calls for efficient and intuitive visualization tools able to scale to very large datasets and to flexibly integrate multiple data types, including clinical data. However, the sheer volume and scope of data poses a significant challenge to the development of such tools.

To address this challenge we developed the Integrative Genomics Viewer (IGV), a lightweight visualization tool that enables intuitive real-time exploration of diverse, large-scale genomic datasets on standard desktop computers. It supports flexible integration of a wide range of genomic data types including aligned sequence reads, mutations, copy number, RNAi screens, gene expression, methylation, and genomic annotations (Figure S1). The IGV makes use of efficient, multi-resolution file formats to enable real-time exploration of arbitrarily large datasets over all resolution scales, while consuming minimal resources on the client computer (see Supplementary Text). Navigation through a dataset is similar to Google Maps, allowing the user to zoom and pan seamlessly across the genome at any level of detail from whole-genome to base pair (Figure S2). Datasets can be loaded from local or remote sources, including cloud-based resources, enabling investigators to view their own genomic datasets alongside publicly available data from, for example, The Cancer Genome Atlas (TCGA)<sup>1</sup>, 1000 Genomes ([www.1000genomes.org/](http://www.1000genomes.org/)), and ENCODE<sup>2</sup> ([www.genome.gov/10005107](http://www.genome.gov/10005107)) projects. In addition, IGV allows collaborators to load and share data locally or remotely over the Web.

IGV supports concurrent visualization of diverse data types across hundreds, and up to thousands of samples, and correlation of these integrated datasets with clinical and phenotypic variables. A researcher can define arbitrary sample annotations and associate them with data tracks using a simple tab-delimited file format (see Supplementary Text). These might include, for example, sample identifier (used to link different types of data for

---

Corresponding authors: Jill P. Mesirov [mesirov@broad.mit.edu](mailto:mesirov@broad.mit.edu) and James T. Robinson, [jrobinso@broadinstitute.org](mailto:jrobinso@broadinstitute.org).

**Author Contributions:** JTR and HT designed and developed the software; WW, MG, ESL, GG, and JPM contributed to the design of the interface and data views; JPM and GG oversaw the project; and JTR, HT, WW, GG, and JPM wrote the manuscript.

the same patient or tissue sample), phenotype, outcome, cluster membership, or any other clinical or experimental label. Annotations are displayed as a heatmap but more importantly are used for grouping, sorting, filtering, and overlaying diverse data types to yield a comprehensive picture of the integrated dataset. This is illustrated in Figure 1, a view of copy number, expression, mutation, and clinical data from 202 glioblastoma samples from the TCGA project in a 3 kb region around the *EGFR* locus<sup>1,3</sup>. The investigator first grouped samples by tumor subtype, then by data type (copy number and expression), and finally sorted them by median copy number over the *EGFR* locus. A shared sample identifier links the copy number and expression tracks, maintaining their relative sort order within the subtypes. Mutation data is overlaid on corresponding copy number and expression tracks, based on shared participant identifier annotations. Several trends in the data stand out, such as a strong correlation between copy number and expression and an overrepresentation of *EGFR* amplified samples in the Classical subtype.

IGV's scalable architecture makes it well suited for genome-wide exploration of next-generation sequencing (NGS) datasets, including both basic aligned read data as well as derived results, such as read coverage. NGS datasets can approach terabytes in size, so careful management of data is necessary to conserve compute resources and to prevent information overload. IGV varies the displayed level of detail according to resolution scale. At very wide views, such as the whole genome, IGV represents NGS data by a simple coverage plot. Coverage data is often useful for assessing overall quality and diagnosing technical issues in sequencing runs (Figure S3), as well as analysis of ChIP-Seq<sup>4</sup> and RNA-Seq<sup>5</sup> experiments (Figures S4 and S5).

As the user zooms below the ~50 kb range, individual aligned reads become visible (Figure 2) and putative SNPs are highlighted as allele counts in the coverage plot. Alignment details for each read are available in popup windows (Figures S6 and S7). Zooming further, individual base mismatches become visible, highlighted by color and intensity according to base call and quality. At this level, the investigator may sort reads by base, quality, strand, sample and other attributes to assess the evidence of a variant. This type of visual inspection can be an efficient and powerful tool for variant call validation, eliminating many false positives and aiding in confirmation of true findings (Figures S6 and S7).

Many sequencing protocols produce reads from both ends ("paired ends") of genomic fragments of known size distribution. IGV uses this information to color-code paired ends if their insert sizes are larger than expected, fall on different chromosomes, or have unexpected pair orientations. Such pairs, when consistent across multiple reads, can be indicative of a genomic rearrangement. When coloring aberrant paired ends, each chromosome is assigned a unique color, so that intra- (same color) and inter- (different color) chromosomal events are readily distinguished (Figures 2 and S8). We note that misalignments, particularly in repeat regions, can also yield unexpected insert sizes, and can be diagnosed with the IGV (Figure S9).

There are a number of stand-alone, desktop genome browsers available today<sup>6</sup> including Artemis<sup>7</sup>, EagleView<sup>8</sup>, MapView<sup>9</sup>, Tablet<sup>10</sup>, Savant<sup>11</sup>, Apollo<sup>12</sup>, and the Integrated Genome Browser<sup>13</sup>. Many of them have features that overlap with IGV, particularly for NGS sequence alignment and genome annotation viewing. The Integrated Genome Browser also supports viewing array-based data. See Supplementary Table 1 and Supplementary Text for more detail. IGV focuses on the emerging integrative nature of genomic studies, placing equal emphasis on array-based platforms, such as expression and copy-number arrays, next-generation sequencing, as well as clinical and other sample metadata. Indeed, an important and unique feature of IGV is the ability to view all these different data types together and to use the sample metadata to dynamically group, sort, and filter datasets (Figure 1 above).

Another important characteristic of IGV is fast data loading and real-time pan and zoom – at all scales of genome resolution and all dataset sizes, including datasets comprising hundreds of samples. Finally, we have placed great emphasis on the ease of installation and use of IGV, with the goal of making both the viewing and sharing of their data accessible to non-informatics end users.

IGV is open source software and freely available at <http://www.broadinstitute.org/igv/>, including full documentation on use of the software.

## Supplementary Material

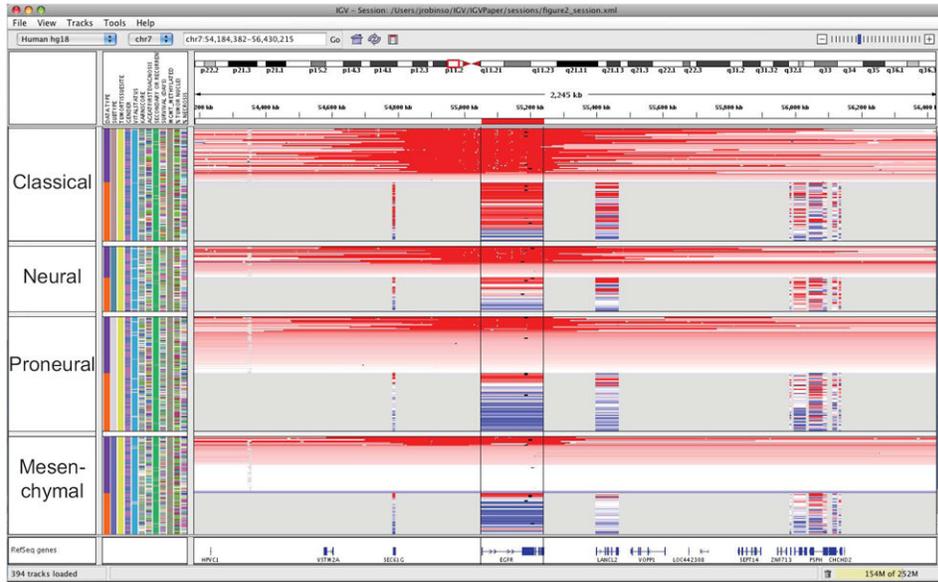
Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

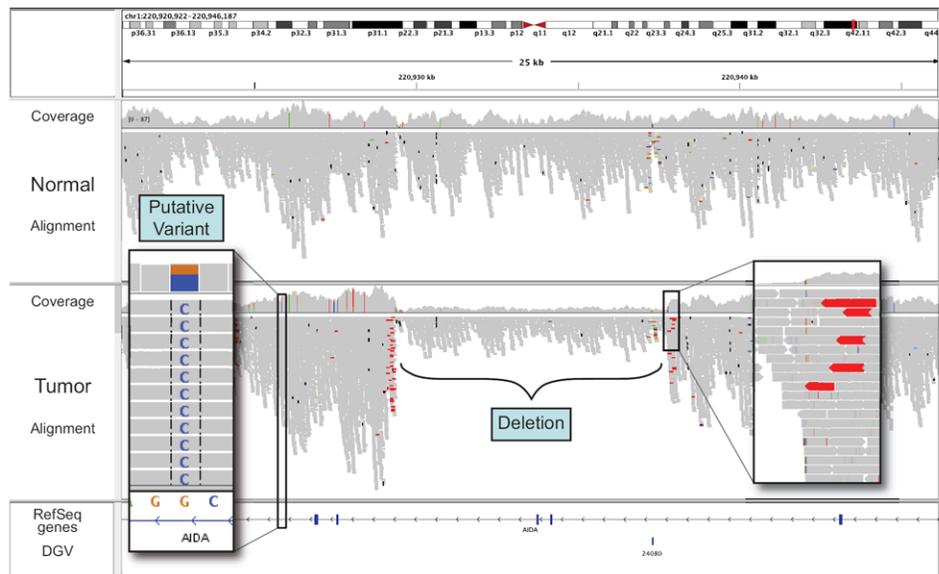
Thanks to M.-D. Nazaire, IGV software developer, J. Boehm, M. DePristo, M. Garber, M. Reich, and B. Weir who worked closely with the development team with suggestions for features and usability, and J. Bistline for expert assistance with figures. M. Garber provided source code for multiple alignment tracks. This work was supported by NIGMS (R01GM074024), NCI (R21CA135827), and NHGRI (U54HG003067).

## References

1. Cancer Genome Atlas Research Network. *Nature*. 2008; 455:1061–1068. [PubMed: 18772890]
2. The ENCODE Project Consortium. *Science*. 2004; 306:636–640. [PubMed: 15499007]
3. Verhaak RG, et al. *Cancer Cell*. 2010; 17:98–110. [PubMed: 20129251]
4. Guttman M, et al. *Nature*. 2009; 458:223–227. [PubMed: 19182780]
5. Berger MF, et al. *Genome Res*. 2010; 20 Epub ahead of print.
6. Nielsen C, Cantor M, Dubchak I, Gordon D, Wang T. *Nature Methods*. 2010; 7:S5–S15. [PubMed: 20195257]
7. Rutherford K, et al. *Bioinformatics*. 2000; 16:944–945. [PubMed: 11120685]
8. Huang W, Marth G. *Genome Res*. 2008; 18:1538–1543. [PubMed: 18550804]
9. Bao H, et al. *Bioinformatics*. 2009; 25:1554–1555. [PubMed: 19369497]
10. Milne I, et al. *Bioinformatics*. 2010; 26:401–402. [PubMed: 19965881]
11. Fiume M, Williams V, Brook A, Brudno M. *Bioinformatics*. 2010; 26:1938–1944. [PubMed: 20562449]
12. Lewis SE, et al. *Genome Biol*. 2002; 3 RESEARCH0082.1–0082.14.
13. Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE. *Bioinformatics*. 2009; 25:2730–2731. [PubMed: 19654113]



**Figure 1. Copy number, expression, and mutation data grouped by tumor subtype**  
 This figure illustrates an integrated, multi-modal view of 202 glioblastoma multiforme TCGA samples. Copy number data is segmented values from Affymetrix SNP 6.0 arrays. Expression data is limited to genes represented on all TCGA employed platforms and displayed across the entire gene locus. Red shading indicates relative up-regulation of a gene and the degree of copy gain of a region; blue shading indicates relative down-regulation and copy loss. Small black squares indicate the position of point missense mutations. Samples are grouped by tumor subtype (2<sup>nd</sup> annotation column) and data type (1<sup>st</sup> sample annotation column), and sorted by copy number of the EGFR locus. Linking via sample attributes insures that the order of sample tracks is consistent across data types within their respective tumor subtypes.



**Figure 2. View of aligned reads at 20kb resolution**

Coverage plot and alignments from paired-end reads for a matched tumor/normal pair. Sequencing was performed on an Illumina GA2 platform and aligned with Maq. Alignments are represented as gray polygons with reads mismatching the reference indicated by color. Loci with a large percentage of mismatches relative to the reference are flagged in the coverage plot as color-coded bars. Alignments with unexpected inferred insert sizes are indicated by color. There is evidence for an approximately 10kb deletion (removing 2 exons of AIDA) in the tumor sample not present in the normal.