BioSkryb GENOMICS

Technical Note

The BaseJumper™ Bioinformatics Platform

Cells explored. Answers revealed.

Authors

Victor Weigman, PhD
Senior Director
Bioinformatics
BioSkryb Genomics

Viren Amin, PhD
Data Engineer
Bioinformatics
BioSkryb Genomics

Corey Culler
Director
Information Technologies
BioSkryb Genomics

Enabling Rapid, Multi-omic Analyses and Interactive Visualization from Anywhere

Translating the molecular alterations, especially those from a single cell, can be a powerful tool to understand underlying mechanisms of disease and therapies. Along with managing the data volume from sequencing technologies, the complexity of these changes requires a multi-dimensional approach to first compute and identify the variance found within the cells of a study and then display these effects concomitantly. The BaseJumper platform is designed to enable the researcher with powerful computing and visualization tools that put them in control of driving their own analysis and interpretation.

Figure 1. The BioSkryb BaseJumper cloud-based bioinformatic platform. The platform was designed to operate on standard laboratory computers with internet access. Researchers' sequencing data can be directly uploaded securely and processed based on pre-loaded analysis pipelines through a standard web portal. BaseJumper automates many of the functions typically requiring bioinformatics staff time, deep computational power, access to comprehensive annotation and iteration of visualization. Convenient dashboards organize projects, link analysis pipelines and provide reports that can be downloaded and shared.

Accessing and Uploading Data

Leveraging any computer typically found in a laboratory (Figure 1), new researchers to the BaseJumper Platform can request their own account at the initial landing page. This initial step will immediately allow access to a demonstration workspace of exemplar BioSkryb datasets and projects. Through the information provided, BaseJumper administrators will link the account to an existing "Organization" space, or create a new one. This critical step ensures sensitive genomic information is only shared to authorized individuals. During this setup, Organizations are able to assign administrators to their space and allow access-granting to additional subgroups, team "Workspaces," and assign roles to other researchers.

Getting sequencing data into BaseJumper enables researchers to leverage existing datasets held in the Illumina BaseSpace™ Sequence Hub and from other cloud or local repositories. Accessing the latter is enabled through Globus[1], where the researchers' existing data endpoints can be transferred directly to BaseJumper's data delivery portal.

During the project creation process, a researcher identifies the sequencing runs or file directories that comprise the experiment. Mechanisms for performing sample filters on file names can rapidly search through large data volumes to identify the correct files. During this process, the researcher applies metadata, such as chemistry and library preparation workflows. The metadata collection is a valuable step which can later be leveraged to determine workflow-specific biases. Following this, data paths for all experimental samples will be aggregated, copied and stamped with date and size within the project and can be used for later workflows. An email is created and sent to the researcher to denote initiation and completion of projects.

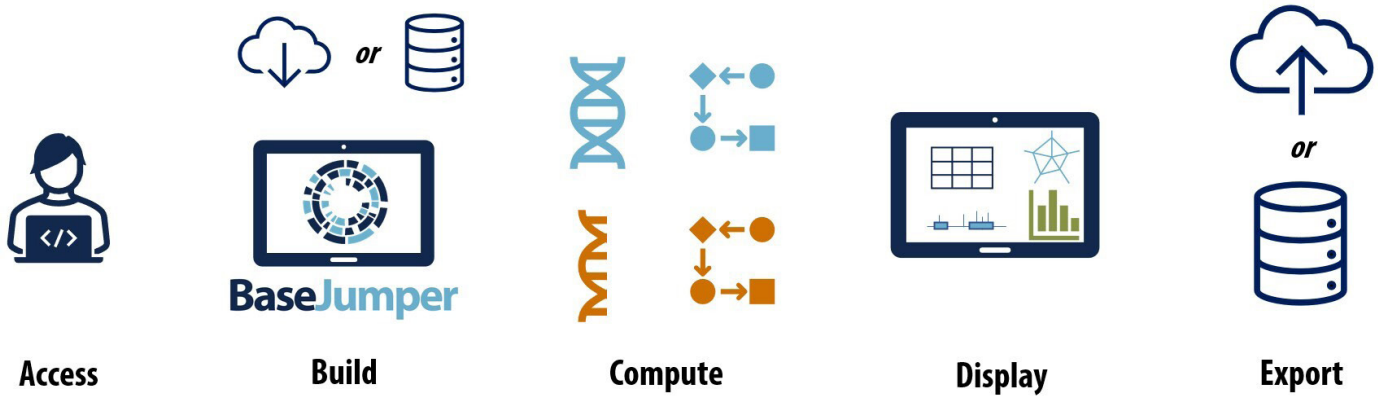

Figure 2. The ABCs of the BaseJumper Workflow. Researchers create accounts as a part of an Organization and are assigned into one or more of that Organization's workspaces. Once in a workspace, they can create projects, pulling data from Illumina BaseSpace and/or their own share. Bioinformatics pipelines can then be queued to immediately launch after data transfer to BaseJumper. Researchers control which modules and parameters are passed to pipelines set for DNA or RNA workflows. Output from these pipelines can then be fed into dynamic visualization applications and exported into individual's specific data endpoints.

Analysis Workflow

Project creation is just the first and most fundamental part of leveraging the BaseJumper platform. While enabling rich, dynamic visualizations is at the core of BaseJumper's mission, the project is the extension of the experiment and careful setup is the foundation of meaningful downstream interpretation. In Figure 2, each of the steps performed by the BaseJumper platform are outlined and make up the core features researchers are able to perform:

- **Access** - The platform puts control of genomic information in the hands of each Organization. Researchers request or are assigned roles specific to their Organization's access policies.
- **Build** - Create a project from previous sequencing data. This can be done through existing access of BaseSpace or from other data endpoints, accessible through Globus.
- **Compute** - Researchers can determine whether one or multiple bioinformatics pipelines can be executed (or a subset of samples within a project).
- **Display** - Projects feature a visualization section with several application portals to display and manipulate pipeline generated results.
- **Export** - Data generated by BaseJumper can be exported either by direct download of files (or graphics) or at a project-level into a researchers' data endpoint.

The BaseJumper platform is designed in such a way to be *complementary* to your organization's existing bioinformatics team. You can choose one or more of these steps that best fits your internal workflows. Projects and visualization results can not only be shared internally but to external collaborators as needed and authorized by researcher.

To facilitate understanding the status of each step, a messaging system within BaseJumper leverages the email within a researcher's profile to notify when each step begins and completes, enhancing the efficiency of the researcher's time.

Analysis Pipelines

After projects are created and samples' sequencing files (FASTQs) have been identified, the researcher can select from analysis pipelines suitable to highlight the genomic and transcriptomic variability found in the samples within the projects. Pipeline modules and parameters can be configured to allow flexibility to match prior analyses.

Currently, BaseJumper provides these pipeline capabilities:

- **QC** - Metrics such as library complexity, error rates, chromosomal coverage, library anomalies and read numbers are returned. This can be run to do rapid (down-sampled) determination of the quality of your sequencing results prior to longer, more expensive analyses.
- **DNA-based** - The genomic pipeline performs analyses to report variation in: single nucleotide variants (SNVs), small insertions or deletions (Indels), and copy number variants (CNVs). Regions of interest can be provided to frame these results against exome or other targeted panel capture technologies.
- **RNA-based** - The transcriptomic pipeline makes the most of isoform and gene-level counting and normalization suitable for BioSkryb's full transcript length results or end-tagging data. Results include gene-level counts, isoform-level counting, count normalization and cell identification.

Annotation of results is vitally important to provide the context of results to known features. While the RNA-based workflows provide counts specific to individual splice forms and common HGNC[2] gene symbols, identification of cellular information is provided to predict:

- Cell cycle state (G1, S, G2M)
- Primary cell / progenitor (Human Primary Cell Atlas - HPCA[3])
- Tissue of origin (GTeX[4])
- Associated cancer tissue (TCGA, multiple cancers)

Figure 3. The BaseJumper Visualization Portal.
Visualization of project-level datasets is accessed through different embedded views. Within each embedded application there are manipulatable representations of genomic and transcriptomic data. A key component of each application is the selection and storing of gene and sample sets which can be used to label and group datasets in linked applications to enable data overlays. Each graphic can be downloaded directly to a user's computer for sharing and the configuration file saved to reuse on a separate project - enabling consistency in interpretation.

For variants identified through the DNA-based pipelines there are more complex annotations for DNA-based variation:

- Polymorphic annotations - gnomAD[5]
- Pathogenic designations (ClinVar[6], COSMIC[7])
- Prevalent Cancer Mutations (frequency in TCGA)
- Effect on Protein[8]

Predicting the impact of a variant on a protein function reports values from: Sorting Intolerant From Tolerant (SIFT)[9], Likelihood ratio test (LRT)[10], FATHM[11], Protein Variation Effect Analyzer (PROVEAN)[12] and MetaSVM/MetaLR[13].

Visualization

Complex and detailed visualization is at the heart of elucidating discovery for the researcher, across applications spaces. The BaseJumper™ bioinformatics platform was designed to provide a streamlined solution to immediately visualize and interpret pipeline data. This can be done from data generated within the BaseJumper pipeline system or through another solution.

Within a project is a portal of applications that can produce a dynamic view (Figure 3) of pipeline results. Within the application, the researcher is free to select subsets of the data to perform a more detailed interpretation. This can be done across the scale of sample and gene-level to individually queried mutations or transcripts.

Visualizations currently served by BaseJumper include:

- **Genome Views** - The genome variability within the samples of a project can be freely navigated. In addition, specific loci or markers can be viewed specifically
- **Copy Number** - Levels of aneuploidy across the dataset are reported to identify common areas of gain or loss. Selecting an area provides a list of the genes that occupy those regions
- **Variant Filtering** - The genetic variation of samples across the project can be simultaneously viewed in tabular format, along with meaningful annotation. The researcher can then chose a myriad of filtering options for meaningful or actionable changes and observe the extent of heterogeneity across the project
- **Sample Similarity** - Correlation of samples that share similar genomic or transcript variation can be identified to select for subsets
- **Variability of Expression** - Heatmaps at the gene and transcript level can be annotated for provided gene/pathway membership along with sample-related annotation (such as cellular ID) identified through other visualization applications
- **Differential Expression** - The same sample characteristics can be used to perform supervised analysis across groups and new gene lists associated to biological groups identified

Each of the applications' visualizations can be exported and downloaded directly to the researcher's computer.

Exporting Results

Researchers can decide to export results at *any stage* of the analysis workflow. After a pipeline has been run, there is a file explorer available to download individual graphical reports, metadata or result files. Each graphical application within the visualization portal has a direct image download to capture screenshots and add to any presentation. Bulk project-level data can be flagged for transfer and passed directly to users' data endpoints (through Globus).

Summary

BaseJumper is a scalable and secure platform that allows the interpretation of large datasets and fuels discovery of new biomarkers within single cell sequencing data. Modular workflows enable the researcher to control which parts of the platform can be leveraged to complement and enhance internal capabilities. Visualization applications can take datasets across genomic and transcriptomic results and can be manipulated based on expert interpretation of the researcher to maximize biological leverage. Compiling all of this within a browser maximizes accessibility of data so you can perform analyses anywhere inspiration takes you.

References

1. Foster, I., "Globus Online: Accelerating and Democratizing Science through Cloud-Based Services," Internet Computing, IEEE , vol. 15, no. 3, pp. 70,73, May-June 2011
2. Tweedie S, Braschi B, Gray KA, Jones TEM, Seal RL, Yates B, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res.
3. D. Aran, A.P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage Nat Immunol, 20 (2019), pp. 163-172
4. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nature genetics, 45(6), 580–585.
5. Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).
6. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res . 2018 Jan 4. PubMed PMID: 29165669 .
7. John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, Simon A Forbes, COSMIC: the Catalogue Of Somatic Mutations In Cancer, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D941–D947, Robinson, J.T., et al., Integrative genomics viewer. Nat Biotechnol, 2011. 29(1): p. 24-6.
8. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.", Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). 2012 Apr-Jun;6(2):80-92.
9. Sim, N.L., et al., SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res, 2012. 40(Web Server issue): p. W452-7.
10. Qian, M. and Y. Shao, A likelihood ratio test for genome-wide association under genetic heterogeneity. Ann Hum Genet, 2013. 77(2): p. 174-82.
11. Shihab, H.A., et al., An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics, 2015. 31(10): p. 1536-43.
12. Shihab, H.A., et al., Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics, 2013. 29(12): p. 1504-10.
13. Shihab, H.A., et al., Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat, 2013. 34(1): p. 57-65.
14. Shihab, H.A., et al., Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Hum Genomics, 2014. 8: p. 11.
15. Choi, Y., et al., Predicting the functional effect of amino acid substitutions and indels. PLoS One, 2012. 7(10): p. e46688.
16. Dong, C., et al., Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet, 2015. 24(8): p. 2125-37.
17. Rebhan, M., et al., GeneCards: integrating information about genes, proteins and diseases. Trends Genet, 1997. 13(4): p. 163.
18. Howe, K.L., et al., Ensembl 2021. Nucleic Acids Res, 2021. 49(D1): p. D884-D891.

For more information or technical assistance:
info@bioskryb.com

Published by:

BioSkryb GENOMICS

2810 Meridian Parkway, Suite 110
Durham, NC 27713
www.bioskryb.com

TAS.018, 05/2022