DNASTAR
SOFTWARE FOR LIFE SCIENTISTS

Automating Workflows in DNASTAR's Lasergene Genomics Suite for High-Throughput Applications

Thomas Leary III, Matthew Keyser, Thomas Lynch, PhD, Schuyler Baldwin, Richard Nelson, PhD, Timothy Durfee, PhD, Andrew Thomas, John Schroeder, Frederick Blattner, PhD

Affiliations: DNASTAR, Inc., Madison, Wisconsin, USA

Abstract

In the era of increasingly large and complex data sets, it is becoming more important to automate tasks for thorough and complete data analysis. DNASTAR's Lasergene Genomics Suite has proven itself as a powerful and accurate tool for genomic assembly and analysis and it is used globally by researchers who need an easy to use and feature-rich genomics toolset for a wide range of workflows. The ease of project setup and the Graphical User Interface (GUI) have made Lasergene the preferred software for thousands of molecular biologists. Beneath the easy to use GUI is a powerful scripting language that gives advanced users the capability to automate a wide range of workflows. Harnessing the scripting capability, which is available throughout the Lasergene Genomics Suite, can support the creation of high-throughput pipelines that can significantly advance genomic assembly and analysis. Nearly any manual process can be integrated into an automated assembly pipeline using the scripting commands found in the Lasergene Genomics Suite applications. Analysis can also be automated using the advanced scripting language so researchers can focus on genomic regions, variations, and other areas of greatest interest to their work. The scripting capability in Lasergene Genomics Suite is easily learned and most researchers are able to begin automating tasks after minimal training. We will highlight several workflows that are conducive to automation using the DNASTAR scripting language, including several examples shared by current Lasergene users.

Note: DNASTAR-developed handler systems are used to connect the processes and initiate logic control of existing and validated manual methods. The handler systems can initiate assembly runs and also dynamically create the scripts required for the NGS process pipeline. These systems were also developed to work with existing pipeline components.

Automated Gene Analysis

Situation: A 96-well sample sequenced with Illumina technology needed to be batch processed to expand throughput of the pilot consensus calling procedure. Critical to the process was the assembly consensus calling for each sample individually.

Approach: 96 samples were assembled in each run, producing individual assemblies and SNP reports, including feature, coverage and Wiggle files. Automated analysis was performed, ending with visualization of files in GenVision Pro.

Benefit: A process that previously took one week was able to be accomplished in two hours and repeated as needed.

```
#!/usr/bin/xng
setDefaultDirectory defaultDirectory: "."
runscript table: "Tar10_Arabidopsis_SRP105913.table.txt" script: "Tar10_Arabid
;command to run 16 sample templated RNASEQ assembly from batch table script

set $projectDir: "D:\ResultDisk\"
set $baseProjectDir: "D:\ResultDisk\Tar10_Arabidopsis_SRP105913"
set $baseProjectName: "Tar10_Arabidopsis_SRP105913"
;set variables for assembly parameters

if  setValue: "${assembly.success}" then {
    ;calculate RPKM and save quantification data

    tell command: qng {
        set $$QNGprojectDir: "${baseProjectDir}"
        set $$QNGprojectName: "${baseProjectName}"
        setupData workflow: RNASeq
        runscript
            inline: "addExperiments file: $$QNGprojectDir/$$project.assembly "
            table: "$QNGprojectName.table.txt"
        qseq_setLayoutOptions
            AutoSetNormalizationMethod: false
            NormalizationMethod: RPKM
        qseq_setTemplateGeneHandling
            featureFilter: "CDS,rRNA,tRNA,ncRNA,tmRNA,misc_RNA"
            excludePseudogenes: true
        processData

        exportToNGen
        saveProject file: "$QNGprojectDir/$QNGprojectName"
```

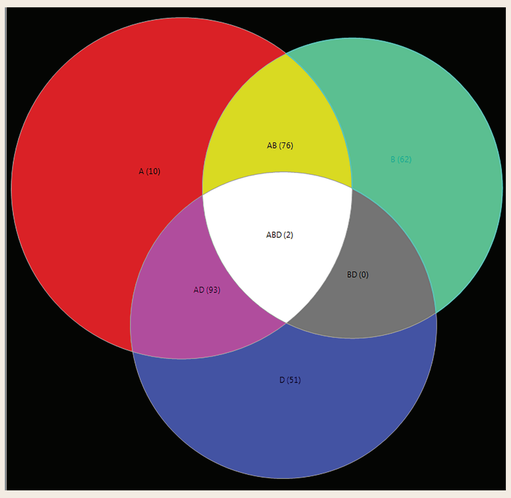
GenVision Pro view of the differential expression of 8 RNA-Seq replicate sets from the SRP105913 study displaying the SPY gene after templated assembly (reference was Tair10 Arabidopsis) and gene expression quantification in the DNASTAR RNA-Seq pipeline. Inset: Lasergene Genomics Suite scripting code to run the 16 sample SRP105913 study.[1]

Clinical Research Process Automation

Situation: A clinical researcher was interested in identifying mutations in thousands of samples that correlate with lung disease.

Approach: High-throughput sample analysis of human exome data was needed for a standard operating protocol that included a cross compared reference with a user supplied VCF file (Validated Gene Panel). The Lasergene workflow scripting language was utilized to provide an integrated solution to assembly, including applying specific filtering criteria to identify somatic mutations and create a list of putative SNPs for each sample. Comparison across samples was performed based on the filtered results. Visualization of highly filtered results was performed in ArrayStar.

Benefit: A reproducible process was created and able to be repeated across thousands of samples per year, very efficiently.



The Ashkenazi Jewish Trio, a candidate NIST reference exome data set, assembled with the exome pipeline and shown in ArrayStar represented as a three sample filtered Venn Diagram.[2]

Ref ID	Ref Pos	Gene Name	Daughter - Called Seq	Father - Called Seq	Mother - Called Seq	dbSNP ID	Daughter - 1000Gp3_CEU_MAF	Father - 1000Gp3_CEU_MAF	Mother - 1000Gp3_CEU_MAF	Daughter - SIFT_pred	Father - SIFT_pred	Mother - SIFT_pred
NC_000001	881627	NOC2L	G>A	G>A	G>A	2272757	0.6970	0.6970	0.6970			
NC_000001	883899	NOC2L	T>G|T	T	T>G|T	72631890	0.0051	------	0.0051	Tolerated (T)		Tolerated (T)
NC_000001	887436	NOC2L	G>C|G	G>C|G	G	202229466	------	------	------	Tolerated (T)	Tolerated (T)	
NC_000001	887449	NOC2L	G>A|G	G	G	144009138	------	------	------	Damaging (D)		
NC_000001	887801	NOC2L	A>G	A>G	A>G	3828047	0.9343	0.9343	0.9343			
NC_000001	888639	NOC2L	T>C	T>C	T>C	3748596	0.9343	0.9343	0.9343			
NC_000001	888659	NOC2L	T>C	T>C	T>C	3748597	0.9343	0.9343	0.9343	Tolerated (T)	Tolerated (T)	Tolerated (T)
NC_000001	889158	NOC2L	GA>CC	GA>CC	GA>CC	13303056	------	------	------			
NC_000001	889423	NOC2L	A>G|A	A>G|A	A>G|A		------	------	------			
NC_000001	892502	NOC2L	G	G>T|G	G>T|G	751975897	------	------	------		Damaging (D)	Damaging (D)
NC_000001	892510	NOC2L	C	C>T|C	C>T|C	746921496	------	------	------		Tolerated (T)	Tolerated (T)
NC_000001	897325	KLHL17	G>C	G>C	G>C	4970441	0.9343	0.9343	0.9343			

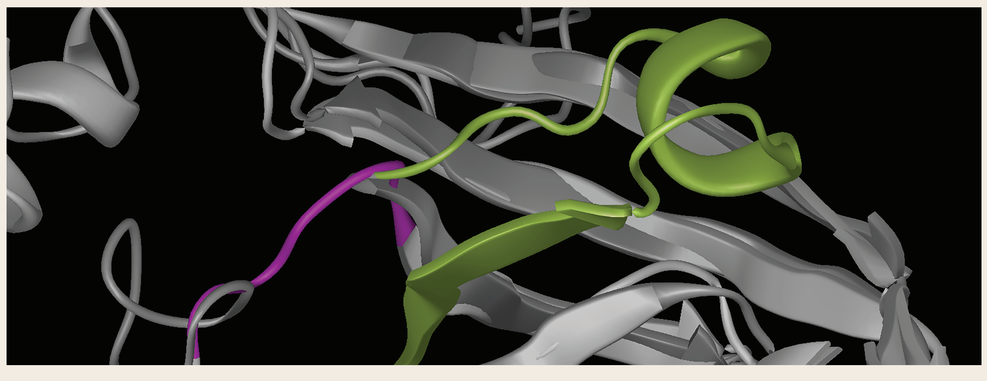
ArrayStar SNP table showing a sample of Ashkinazi Jewish Trio (Candidate NIST RM8392) SNPs with the following filter criteria: Non-synonymous AND within targeted regions, P not ref > 90%, Depth >10, 1000Gp_3_CEU_MAF <= 0.05, SIFT_pred is Damaging. SNP quantification data was integrated with DNASTAR's Variant Annotation Database (VAD) before filtering.[2]

Structural Biologist Automated Genetic Mutation Analysis

Situation: A structural biologist was interested in interrogating the impact of numerous genetic variations on a wide range of proteins within a genome.

Approach: The genome was assembled from Illumina data and SNPs were called. SeqNinja, a molecular biology automation application within Lasergene, was used to translate genetic sequences to amino acid sequences and to launch NovaFold to predict protein structures. Selected proteins of interest were aligned in Protean 3D for comparing the impact of protein sequence variation on structure.

Benefit: An automated pipeline was created to identify sequence variations and the structures for selected proteins were predicted and compared using DNASTAR tools.



Human Tax1 Binding Protein 3 (kiwi) model structurally aligned using TM-Align to Human Tax1 Binding Protein 3 splice variant 4 (magenta) model. Both structural models were generated using NovaFold. Splice variant (Δ 54-79) eliminates β-sheet 5 (aa 55-60) and a-helix 2 (aa 65-69) observed in wild-type structural model.

Summary

The Lasergene genomics package is built upon a powerful genomics pipeline that allows for an integrated and expandable scripting-based genomics NGS platform suitable for high-throughput automation.

LASERGENE®
Genomics Suite

1. Severing, E. Splicing-related genes are alternatively spliced upon changes in ambient temperatures in plants; SRP105913; Accession PRJNA328771; 2017-05-02; Wageningen University, Netherlands.
2. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data 3:160025 doi: 10.1038/sdata.2016.25 (2016).

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Numbers R44GM110814 and 5R44GM100520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

www.dnastar.com