DNASTAR®
Software for life scientists

# A high throughput software pipeline for NGS-based association studies: From assembly to candidate variants and their effect on 3D protein structures

Katie Maxfield, Matthew Keyser, Steven J. Darnell, Matt Larson, Erik Edlund, Richard Nelson, Schuyler Baldwin, Dan Nash and Tim Durfee
DNASTAR, Inc., Madison, Wisconsin, USA

## Abstract

Advances in DNA sequencing have made identifying genetic variation from any individual routine, and allowed variation across populations or cohorts to be analyzed for candidate mutations causing diseases or traits of interest. However, the computational complexity of the analysis, combined with an enormous amount of data, poses daunting challenges in terms of bioinformatics, data management and computing resources that are beyond all but the most well equipped laboratories.

We have developed a seamlessly integrated software pipeline to address this problem. NGS sequencing reads from each sample are aligned to a reference genome using SeqMan NGen, a fast non-memory bound assembler for data sets of any size. A Bayesian modeled probabilistic variant caller analyzes the gapped alignments in-stream to produce high accuracy single nucleotide variant (SNV) and small indel calls for each sample.

Assemblies can be done on standard desktop computers or on the DNASTAR Cloud which allows the samples to be processed in parallel. Variant profiles from each sample are automatically combined into a single project file for analysis in ArrayStar where various searching and statistical methods are available for identifying candidate genes and/or variants of interest. ArrayStar also allows gene and variant level annotations to be added, aiding in the identification and prioritization of candidates. Further, for genes with known 3D protein structures, the effect of candidate missense mutations can automatically be predicted through integration our molecular structure visualization and analysis module, Protean 3D.

As a demonstration of the pipeline, we present results from the reanalysis of 96 targeted resequencing samples from a Chinese cohort with lung squamous cell carcinomas (LSCC)[1]. Specifically, we will show how the software can be used to rapidly identify likely candidate genes and variants involved in tumor development as exemplified by the frequent occurrence of nonsense/frameshift and deleterious missense mutations in the key tumor suppressor gene, TP53. Further, we use Protean 3D to predict the effect of one deleterious missense mutation, M237I, on the 3-dimensional structure of TP53 and show how the predicted structural change implies DNA-binding may be impacted.

## Genomics Suite Pipeline

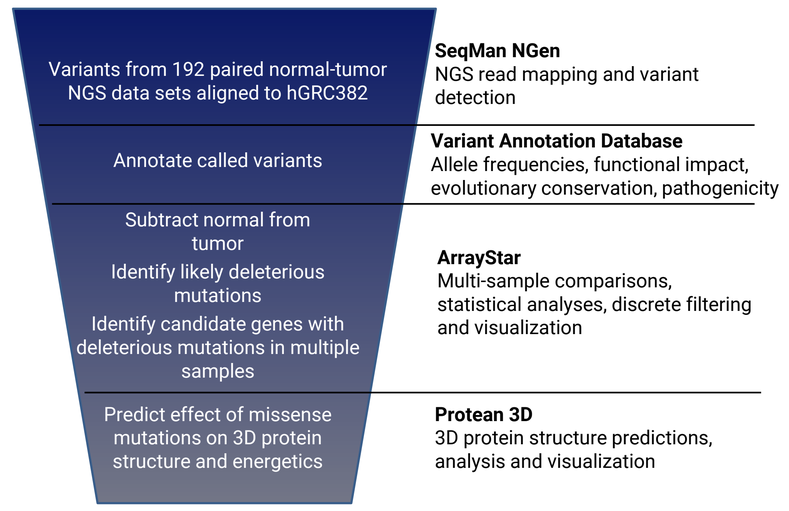


***Figure 1. Overview of DNASTAR Genomics Suite software pipeline for assembling NGS read data, calling variants, filtering data, and modeling variants on protein structure, including steps taken to analyze LSCC samples.***

## Assembly and Variant Detection

The DNASTAR Genomics Suite pipeline provides rapid and accurate assembly of NGS sequence data for the purpose of variant detection and analysis. Combined analysis tools, along with extensive variant/gene database annotation import, provide researchers efficient means to identify candidate genes based on several criteria including mutation effects, as well as clinical and functional importance.

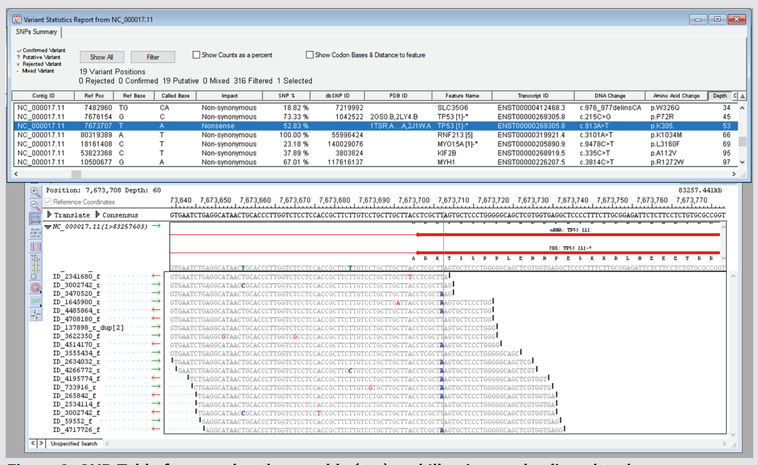

***Figure 2. SNP Table for completed assembly (top) and Illumina reads aligned to human genome reference sequence (bottom), with nonsense mutation in TP53 highlighted in blue.***

By applying progressive filters to variant data, we identified a small subset of genes with variants that are predicted to inactivate the gene in multiple tumor samples.

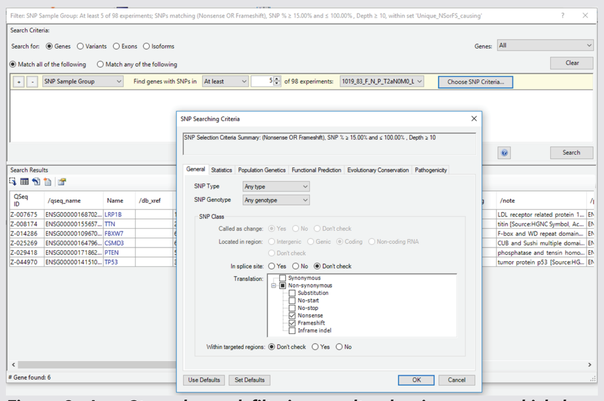

***Figure 3. ArrayStar advanced filtering results showing genes which have unique nonsense and frameshift mutations in multiple tumor samples.***

**Table 1. Genes with unique nonsense and/or frameshift mutations in multiple samples**

| Minimum number of samples | Number of Genes | Genes |
|---|---|---|
| **1** | 101 | |
| **5** | 6 | CSMD3, FBXW7, LRP1B, PTEN, TP53, TTN |
| **10** | 2 | TP53, TTN |
| **15** | 1 | TP53 |
| **24** | 1 | TP53 |
| **25** | 0 | |

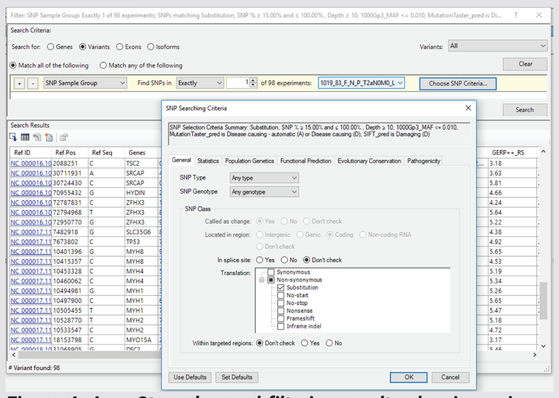

***Figure 4. ArrayStar advanced filtering results showing unique missense mutations that are predicted to be deleterious and occur in exactly one sample.***

**Table 2. Example of discrete filtering of missense variants in the LSCC samples**

| Step | Criteria | Number of variants | Number of genes |
|---|---|---|---|
| **1** | **Missense, minimum variant % = 15, minumum depth = 10** | 3689 | 472 |
| **2** | **" + unique to 1 LSCC sample** | 2134 | 433 |
| **3** | **" + Mutation Taster = Disease causing** | 1132 | 303 |
| **4** | **" + SIFT = Damaging** | 676 | 246 |
| **5** | **" + MAF >= 0.01** | 98 | 86 |

**Table 3. Genes with unique deleterious missense mutations in multiple samples[a]**

| Minimum number of samples | Number of Genes[b] | Genes |
|---|---|---|
| **1** | 208 | |
| **5** | 23 | |
| **10** | 7 | CDH10, COL11A1, MDN1, MYH4, NFE2L2, RYR1, TP53 |
| **12** | 2 | MDN1, TP53 |
| **15** | 1 | TP53 |

***[a] Starting with 246 gene set from Table 2, Step 4***
***[b] Filtered to genes with lengths < 250kb***

Methods + Materials:
**Software**: SeqManNGen 15.2
**Hardware**: 12 Core Intel / 64GB RAM / 1 TB Main drive/ 10 TB scratch drive/ liquid cooled cpu
**Samples Set**: 192 Targeted Exome Lung Cancer vs Normal paired data SRP030634[1]
**Assembly Time**: 84h, Variant Quantification Time: 40h, Total 124h, 37min/ Exome
**Scripting**: Arraystar / SeqMan NGen custom batch script used for pairwise comparision filter.

## Structure Visualization and Mutation Modeling

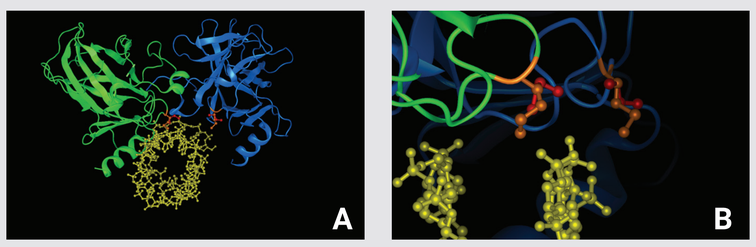


***Fig. 5. Protean 3D structure prediction of the M237I mutation in TP53 indicates it could affect DNA binding. (A) Crystal structure of a TP53 dimer (green and blue subunits) complexed with DNA (backbone shown in yellow ball and stick rendering). Methionine at position 237 (gold ball and stick) in each subunit was changed to isoleucine and the orientation of the side change (red ball and stick) predicted by the software. (B) Zoomed in image of the methionine (gold) and isoleucine (red) side chains at position 237 of the wild type and mutant, respectively, with the DNA backbone shown in ball and stick rendering (yellow).***

DNASTAR's integrated structure visualization and analysis tools allow users to model mutations of interest on the PDB protein structure. Built-in energy calculations allow users to make guided hypotheses about the effect of mutations on the protein structure and function.

Using the ArrayStar advanced filtering options, we interrogated the TP53 missense mutations to identify those which were predicted to be deleterious by all three functional impact predictors (SIFT, MAF and Mutation Taster) and were predicted to be pathogenic or likely pathogenic in ClinVar. This search yielded three variants, including M2371. Starting with the PDB structure 1TUP[3] which contains TP53 complexed to DNA, we used Protean 3D to mutate the methionine to isoleucine in both the A and B subunits. The structure shows that M237 is in close proximity to the DNA backbone. I237 is predicted to be rotated away from the DNA, a change that may affects binding and would in turn likely be involved in pathogenicity.

## Discussion

The ability to combine a large number of NGS assemblies with variant data, and integrate the results with multiple variant and gene annotation databases, allows researchers to quickly identify important or interesting mutations. Fast and intuitive filtering tools allow users to filter on a variety of criteria and combine filtering results in unique ways.

In addition, by combining structural bioinformatics with sequencing technologies, DNASTAR's integrated workflows can guide genomic and molecular biology researchers to create structure-based hypotheses and to investigate possibilities not evident by sequence data alone.

**www.dnastar.com**

References:

1. http://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP030634
2. Human genome reference with GRCh38 and ENSEMBL annotations
3. https://www.rcsb.org/structure/1tup

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R44HG006578 and by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R44GM110814. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.