DNASTAR®
Software for life scientists

DNASTAR Lasergene

High throughput software pipeline for NGS-based association studies: from assembly to candidate variants, and their effects on 3D protein structures

Thomas Leary, Schuyler Baldwin, Dan Nash, Matt Larson, Matt Keyser, Steve Darnell and Tim Durfee

Affiliations

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 together with the enormous amount of data pose 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 will present results from the reanalysis of 96 targeted resequencing samples from a Chinese cohort with lung squamous cell carcinomas (LSCC, Li et al., Sci Rep. 2015. 5:14237). Results will be compared to the published observations highlighting new candidate genes/ variants that may have been initially missed. We will also show examples of how predicted protein structure changes caused by missense mutations can aid in prioritizing variants of interest.

Assembly and Variant Detection

Methods + Materials:
Software: Lasergene SeqManNGen 15.2 pre-release
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.

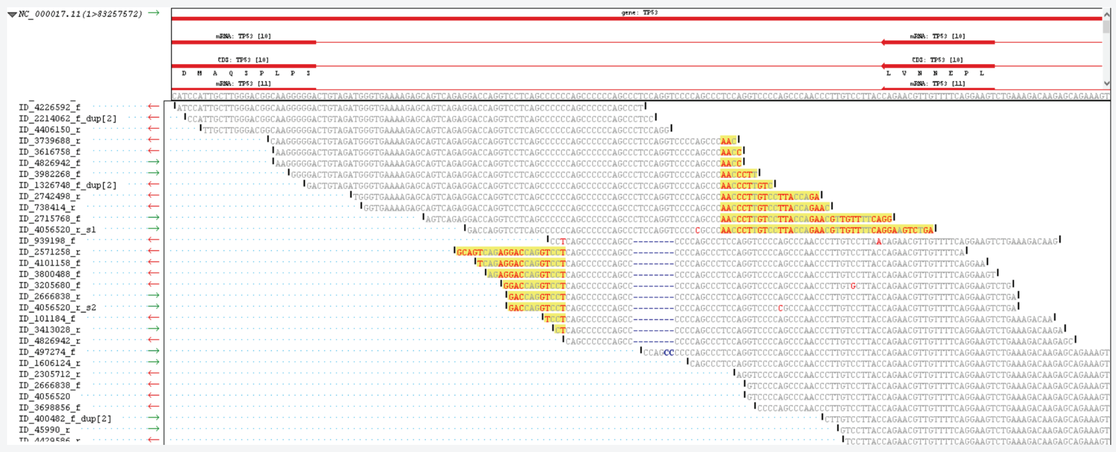


***Figure 1**. SeqMan Pro assembly view of oncogene TP53. Also, noted structrual variation due to hard edge overhang.*

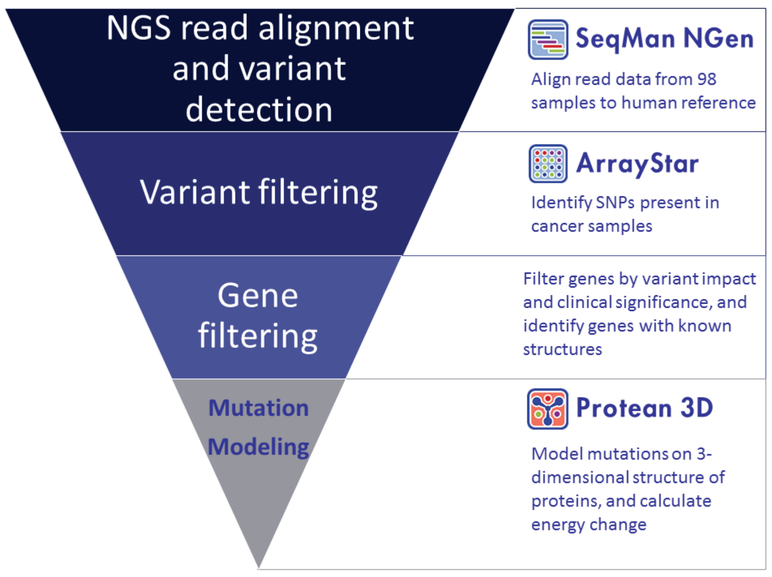
Variant and Gene Filtering

***Figure 2.** Gene table of variants found with somatic filtration. Minimum depth 10, Normal 0-3% and Lung Cancer Tumor 10-100% LRT Deleterious- Non Synonymous found in 10 out of 98 Samples.*

```
;;BEGIN Paired Somatic Batch Filter SCRIPT
;;98 Targeted Exome Paired Tumor Analysis

filter_clearCriteria

filter_setOptions
    dataType: SNP
    searchSpace: All
    comboMode: And

filter_addSNPSearchCriteria
    experiment: $project1
    classification: "Non-synonymous"
    SNPpct: 0.1
    SNPpctMax: 1
    depth: 10


filter_addSNPSearchCriteria
    experiment: $project2
    classification: "Non-synonymous"
    SNPpct: 0
    SNPpctMax: .03
    depth: 10

filter_runSearch

filter_saveResults
    name: "${project1}"
    force: true
```

Arraystar custom filtration is possible using built in scripting syntax. Batch script filter used to batch the paired analysis of tumor vs normal tissue.

Exome Assembly and Analysis Pipeline



Structure Visualization and Modeling

Variants detected from paired Tumor vs Normal tissue are shown in Protean 3D with respective PDB structure and dFire Energy Potential. Side chain is energy minimized and shown in conjunction with original structure.

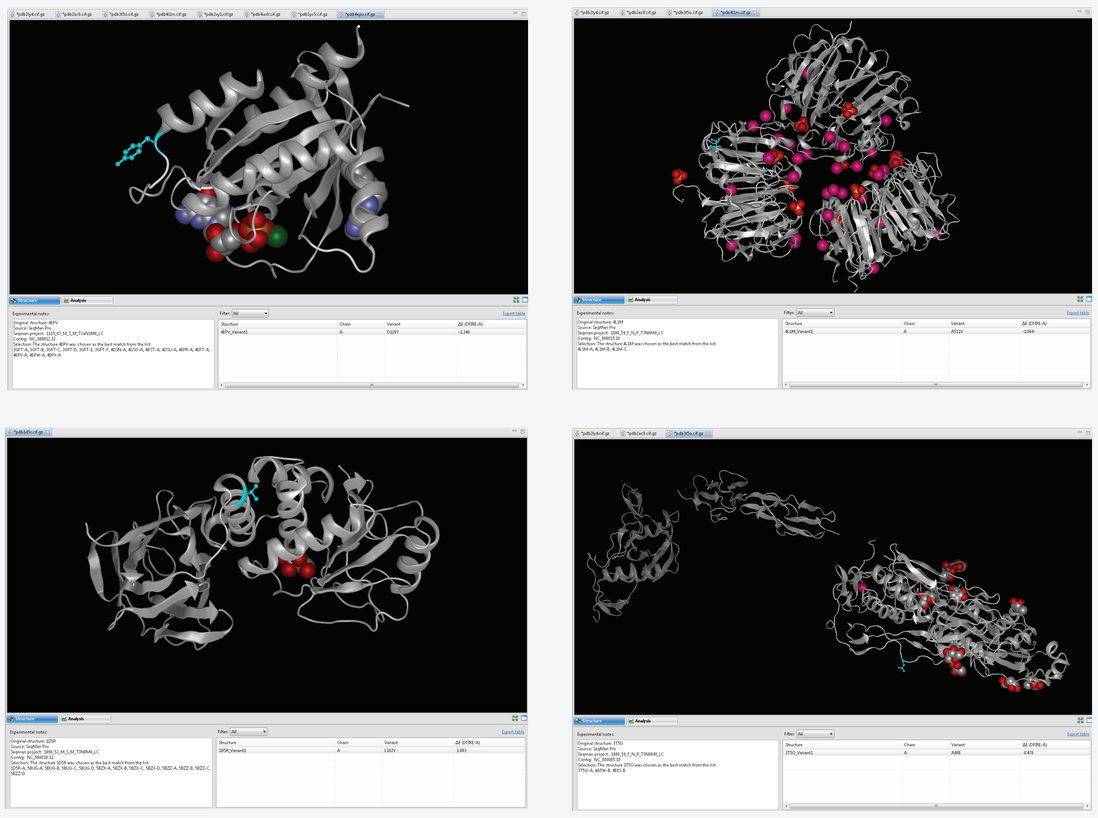


***Figure 3.** PDB structures for select genes found with somatic mutations in 40 out of 98 samples.*

Results and Discussion

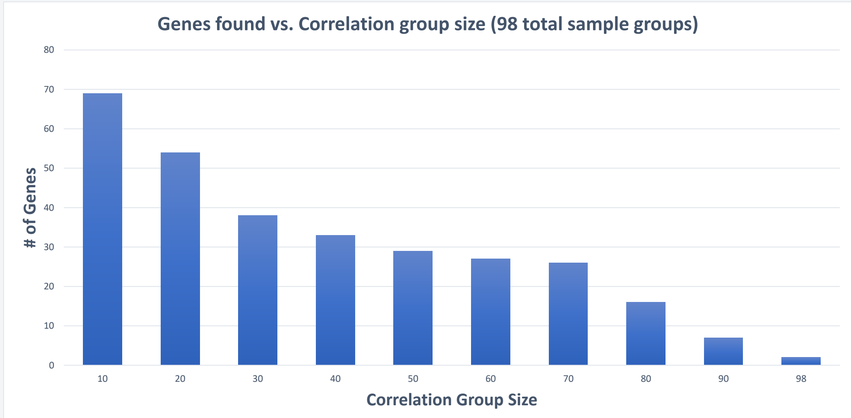


***Figure 4.** Filtered genes grouped by number of samples containing mutations of interest.*

Script based batch filtering was used to perform pairwise comparisons for each sample, and identify SNPs with frequency of 0-3% in the normal tissues and 10-100% in tumor samples. Logic filters were then appled to identify genes based on "or" logic filters; LRT - "Deleterious"; SIFT is "Damaging"; MutationTaster is "Disease Causing" ; 1000 GP3_EAS_MAF < 0.05. Correlation groups are shown in the chart above, demonstrating number of genes that meet these criteria in sub-groups, ranging from 10-98 of the samples.

By combining the read assembly, variant filtering, and protein visualization into a single pipeline, researchers can quickly go from large data sets to a few genes of interest.

www.dnastar.com

References:
Chinese cohort with lung squamous cell carcinomas (LSCC, Li et al., Sci Rep. 2015. 5:14237)
SRP030634 NCBI SRA archives