DNASTAR
SOFTWARE FOR LIFE SCIENTISTS

A flexible, high-throughput software pipeline for identifying candidate variants and their effects on protein structure, starting from NGS data or VCF files

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

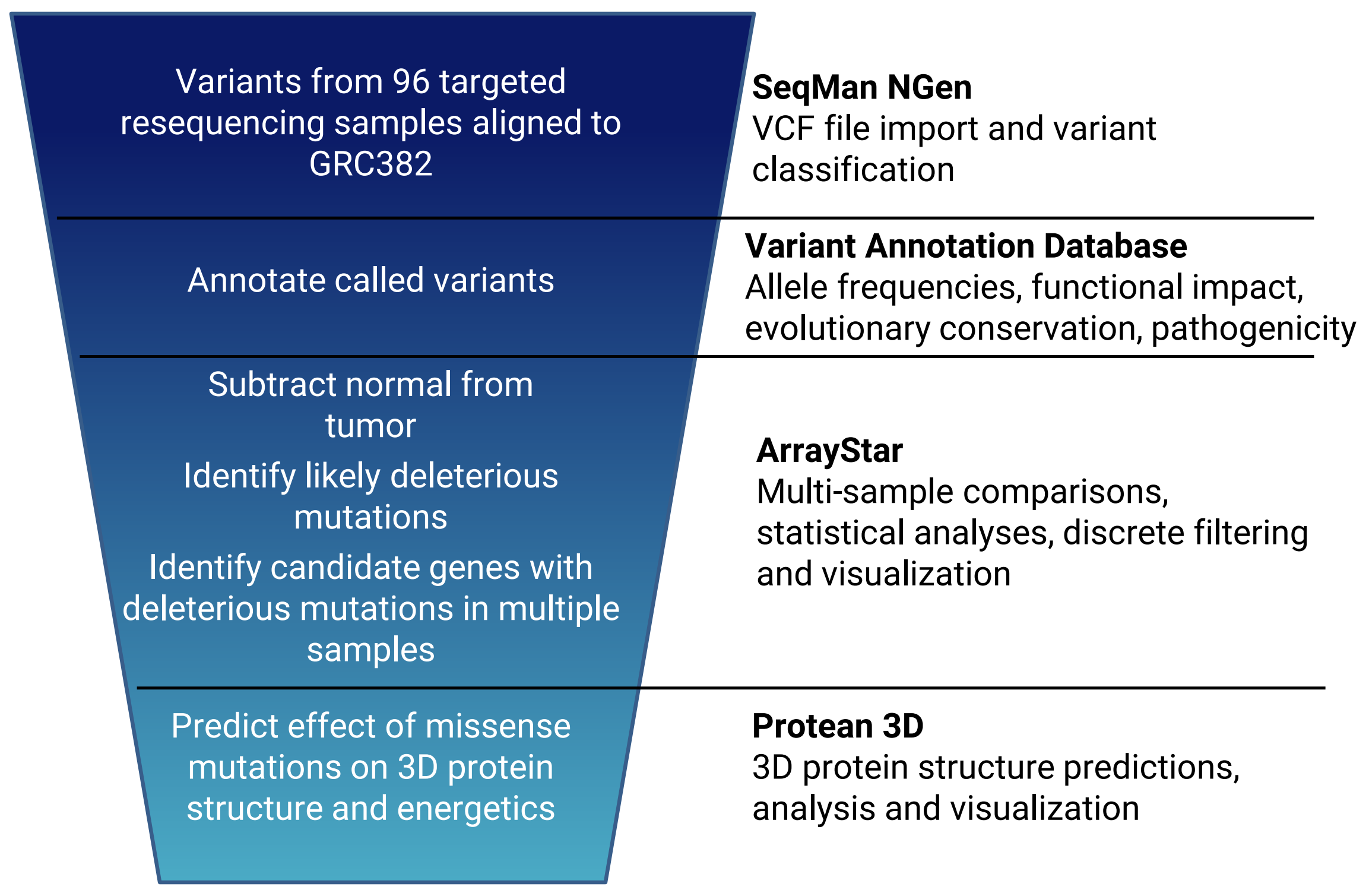
Advances in DNA sequencing have made identifying genetic variation from any individual routine, allowing researchers to analyze variation across populations or cohorts for candidate mutations causing diseases or traits of interest. To address the significant bioinformatic, data management and computing resource challenges inherent to this analysis, we have developed an integrated software solution for each step in the pipeline.

First, variant call profiles for each sample are either determined directly from raw NGS sequencing data or extracted from existing VCF files. For NGS data processing, reads from each sample are aligned to a reference genome using a fast, non-memory bound assembler. Gapped alignments are analyzed in-stream using a Bayesian modeled probabilistic variant caller to produce annotated single nucleotide variant (SNV) and small indel calls. To leverage existing variant information in VCF format, we have also developed a VCF annotation tool which maps the variants onto a corresponding annotated reference sequence and "decorates" each variant with information such as the affected gene(s) and impact on protein encoding regions and/or splice sites.

Next, annotated variant profiles from each sample, supplemented with their allele frequencies as well as predictions about functional impact and pathogenicity, are automatically combined into a single project for analysis. Various filtering and statistical methods then enable candidate genes and/or variants of interest to be readily identified. Further, for genes with known 3D protein structures, the effect of candidate missense mutations on the structure can be automatically predicted and analyzed using our molecular structure visualization module.

As a demonstration of the pipeline, we present results from our reanalysis of 96 targeted resequencing samples from a Chinese cohort with lung squamous cell carcinomas (LSCC, Li et al., Sci Rep. 2015. 5:14237). We will show how the software can be used to easily identify unique mutations in numerous samples across the cohort which all lead to nonsense and frameshift mutations in the TP53 tumor suppressor gene. We will also show how filtering on the functional impact and pathogenicity predictions can be used to identify likely deleterious TP53 missense mutations in other members of the cohort and how in one of those cases, the predicted protein structure change indicates that the DNA binding activity of the protein is likely directly affected.

Project Overview


Variants from 96 targeted resequencing samples aligned to GRC382
Annotate called variants
Subtract normal from tumor
Identify likely deleterious mutations
Identify candidate genes with deleterious mutations in multiple samples
Predict effect of missense mutations on 3D protein structure and energetics

SeqMan NGen
VCF file import and variant classification

Variant Annotation Database
Allele frequencies, functional impact, evolutionary conservation, pathogenicity

ArrayStar
Multi-sample comparisons, statistical analyses, discrete filtering and visualization

Protean 3D
3D protein structure predictions, analysis and visualization

Figure 1. Overview of DNASTAR Genomics Suite software pipeline for importing VCF variant data, classifying and annotating variants, filtering data, and modeling variants on protein structure, including steps taken to analyze LSCC samples.

Variant Annotation and Filtering

The DNASTAR Lasergene package can be used to analyze variants from raw sequencing data, or from VCF files produced by other assembly pipelines. Variants are annotated via a two-step process. The first annotation step classifies variants by their effect on coding regions, relative to the imported reference genome. The second annotation step includes import of the DNASTAR Variant Annotation Database (VAD), which combines data from a variety of SNP level annotation databases. The robust downstream filtering options provide researchers efficient means to identify candidate genes based on several criteria including mutation effects, as well as clinical and functional importance.

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.

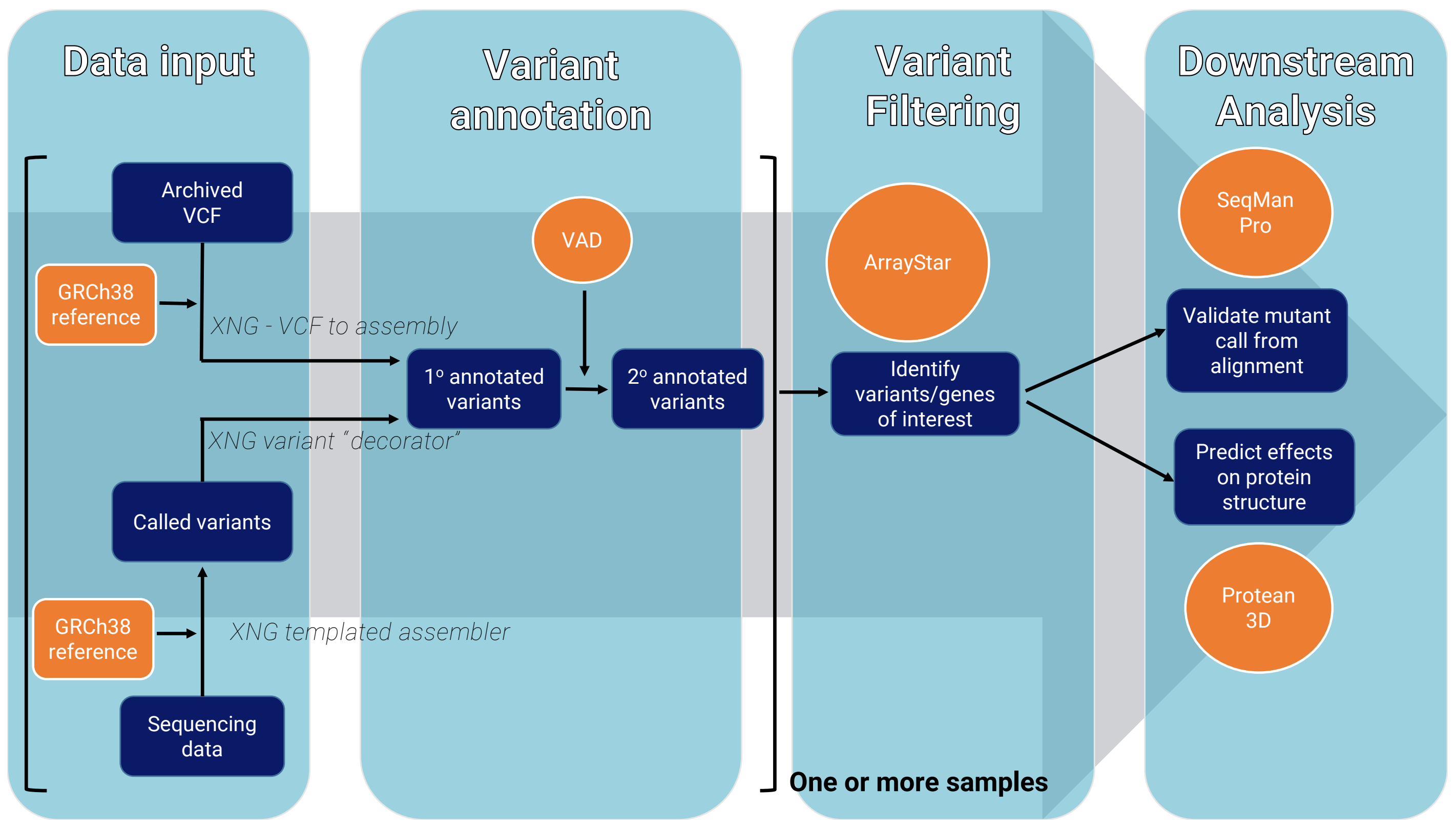


Figure 2. Steps for annotating and analyzing multi-sample variant data in Lasergene. DNASTAR software tools and supplemental data are shown in orange.

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	

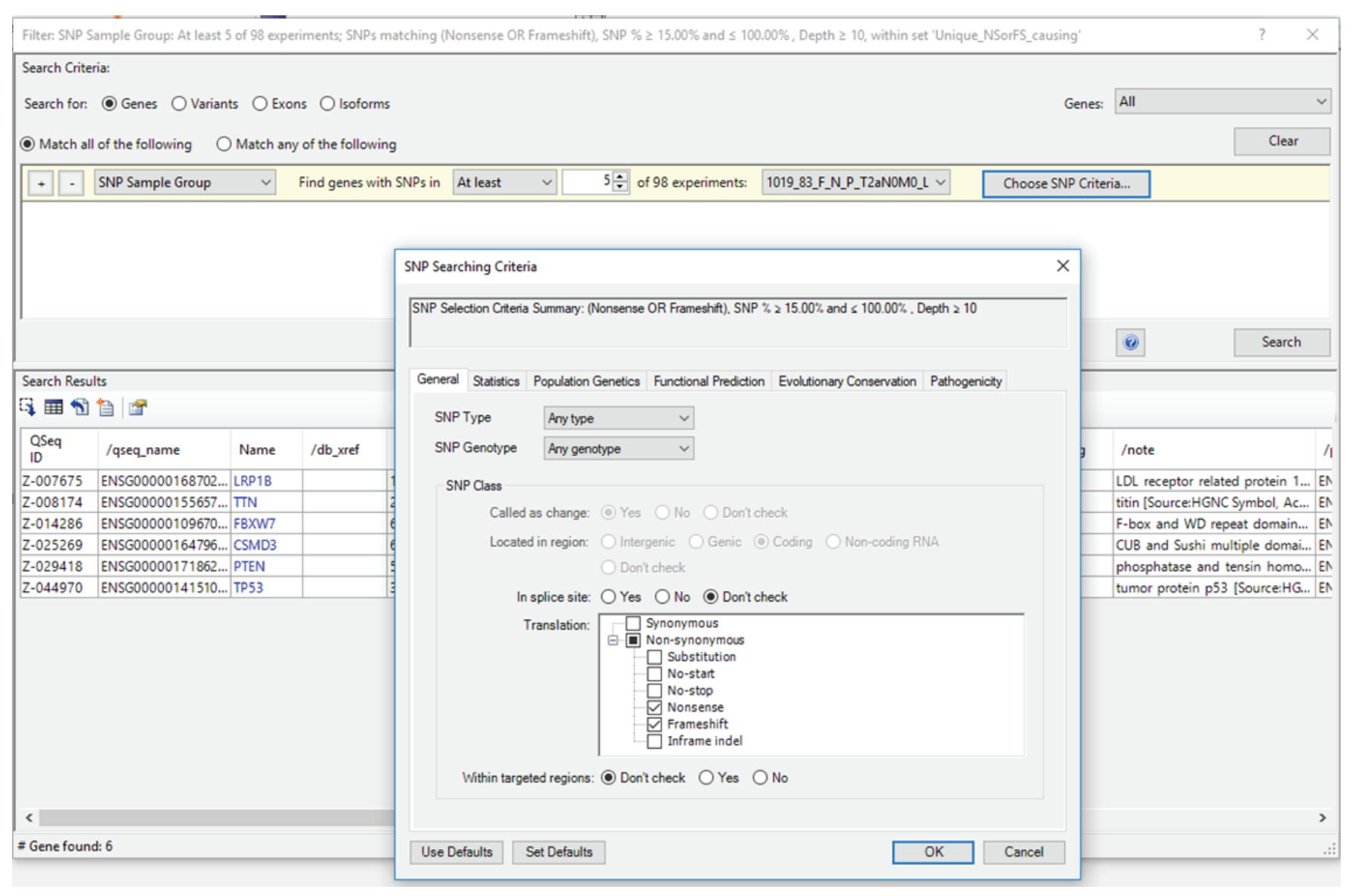
Table 2. Example of discrete filtering of missense variants in the LSCC samples

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


Filter: SNP Sample Group: At least 5 of 98 experiments; SNPs matching (Nonsense OR Frameshift), SNP % ≥ 15.00% and ≤ 100.00% , Depth ≥ 10, within set 'Unique_NSorFS_causing'

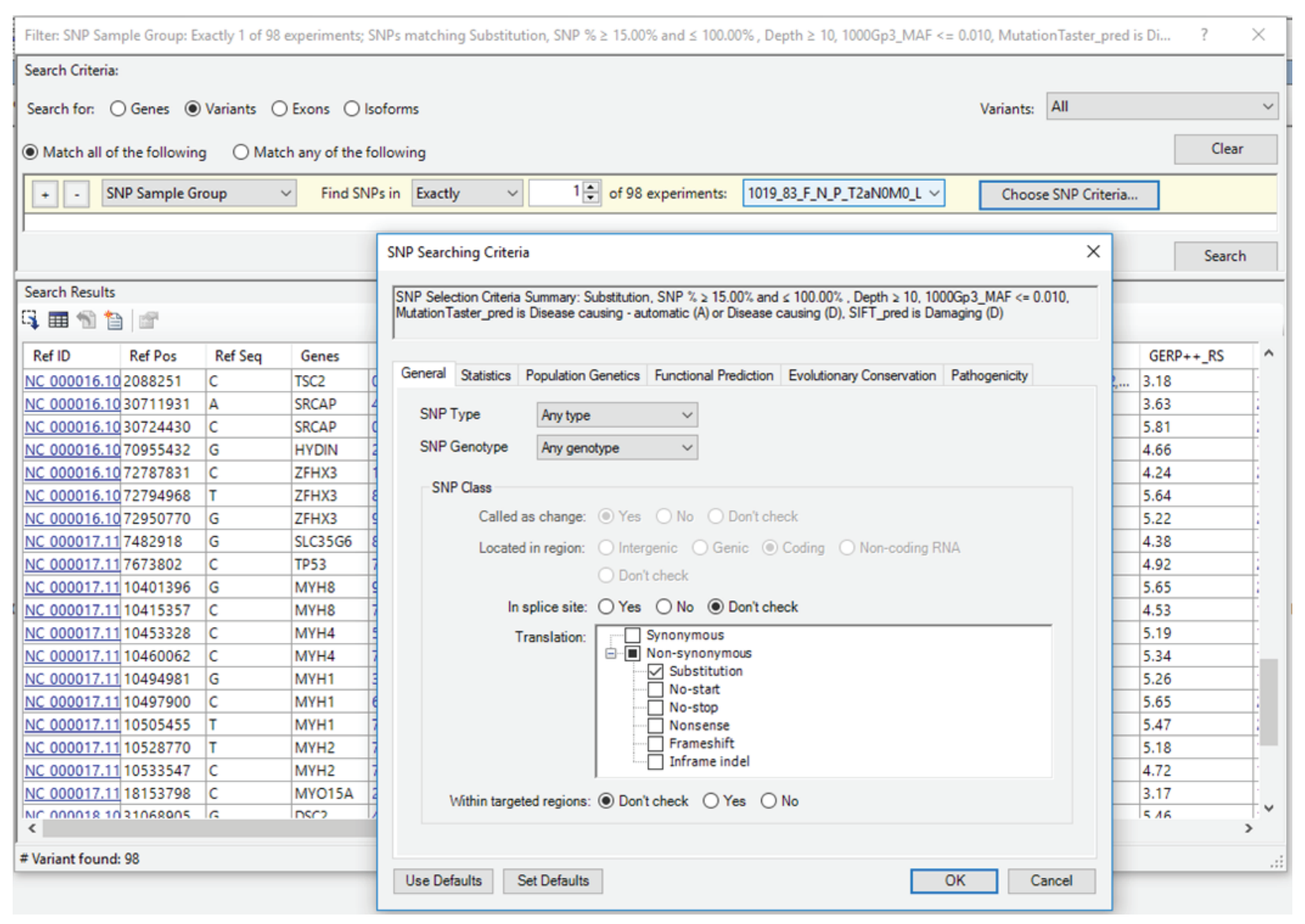
Search Criteria:
Search for: ● Genes ○ Variants ○ Exons ○ Isoforms
Genes: All
● Match all of the following ○ Match any of the following
SNP Sample Group — Find genes with SNPs in: At least 5 of 98 experiments: 1019_83_F_N_P_T2aN0M0_L

SNP Searching Criteria
SNP Selection Criteria Summary: (Nonsense OR Frameshift), SNP % ≥ 15.00% and ≤ 100.00% , Depth ≥ 10
SNP Type: Any type
SNP Genotype: Any genotype
SNP Class
Called as change: ● Yes ○ No ○ Don't check
Located in region: ○ Intergenic ○ Genic ● Coding ○ Non-coding RNA ○ Don't check
In splice site: ○ Yes ○ No ● Don't check
Translation: □ Synonymous ■ Non-synonymous □ Substitution □ No-start □ No-stop ☑ Nonsense ☑ Frameshift □ Inframe indel
Within targeted regions: ● Don't check ○ Yes ○ No

QSeq ID	/qseq_name	Name	/note
Z-007675	ENSG00000168702...	LRP1B	LDL receptor related protein 1...
Z-008174	ENSG00000155657...	TTN	titin [Source:HGNC Symbol, Ac...
Z-014286	ENSG00000109670...	FBXW7	F-box and WD repeat domain...
Z-025269	ENSG00000164796...	CSMD3	CUB and Sushi multiple domai...
Z-029418	ENSG00000171862...	PTEN	phosphatase and tensin homo...
Z-044970	ENSG00000141510...	TP53	tumor protein p53 [Source:HG...

Gene found: 6

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


Filter: SNP Sample Group: Exactly 1 of 98 experiments; SNPs matching Substitution, SNP % ≥ 15.00% and ≤ 100.00% , Depth ≥ 10, 1000Gp3_MAF <= 0.010, MutationTaster_pred is Di...

Search Criteria:
Search for: ○ Genes ● Variants ○ Exons ○ Isoforms
Variants: All
● Match all of the following ○ Match any of the following
SNP Sample Group — Find SNPs in: Exactly 1 of 98 experiments: 1019_83_F_N_P_T2aN0M0_L

SNP Searching Criteria
SNP Selection Criteria Summary: Substitution, SNP % ≥ 15.00% and ≤ 100.00% , Depth ≥ 10, 1000Gp3_MAF <= 0.010, MutationTaster_pred is Disease causing - automatic (A) or Disease causing (D), SIFT_pred is Damaging (D)
SNP Type: Any type
SNP Genotype: Any genotype
SNP Class
Called as change: ● Yes ○ No ○ Don't check
Located in region: ○ Intergenic ○ Genic ● Coding ○ Non-coding RNA ○ Don't check
In splice site: ○ Yes ○ No ● Don't check
Translation: □ Synonymous ■ Non-synonymous ☑ Substitution □ No-start □ No-stop □ Nonsense □ Frameshift □ Inframe indel
Within targeted regions: ● Don't check ○ Yes ○ No

Ref ID	Ref Pos	Ref Seq	Genes	GERP++_RS
NC_000016.10	2088251	C	TSC2	3.18
NC_000016.10	30711931	A	SRCAP	3.63
NC_000016.10	30724430	C	SRCAP	5.81
NC_000016.10	70955432	G	HYDIN	4.66
NC_000016.10	72787831	C	ZFHX3	4.24
NC_000016.10	72794968	T	ZFHX3	5.64
NC_000016.10	72950770	G	ZFHX3	5.22
NC_000017.11	7482918	G	SLC35G6	4.38
NC_000017.11	7673802	C	TP53	4.92
NC_000017.11	10401396	G	MYH8	5.65
NC_000017.11	10415357	C	MYH8	4.53
NC_000017.11	10453328	C	MYH4	5.19
NC_000017.11	10460062	C	MYH4	5.34
NC_000017.11	10494981	G	MYH1	5.26
NC_000017.11	10497900	C	MYH1	5.65
NC_000017.11	10505455	T	MYH1	5.47
NC_000017.11	10528770	T	MYH2	5.18
NC_000017.11	10533547	C	MYH2	4.72
NC_000017.11	18153798	C	MYO15A	3.17

Variant found: 98

Figure 3. ArrayStar advanced filtering results showing uniqe missense mutations that are predicted to be deleterious and occur in exactly one sample.

Structure Visualization and Mutation Modeling

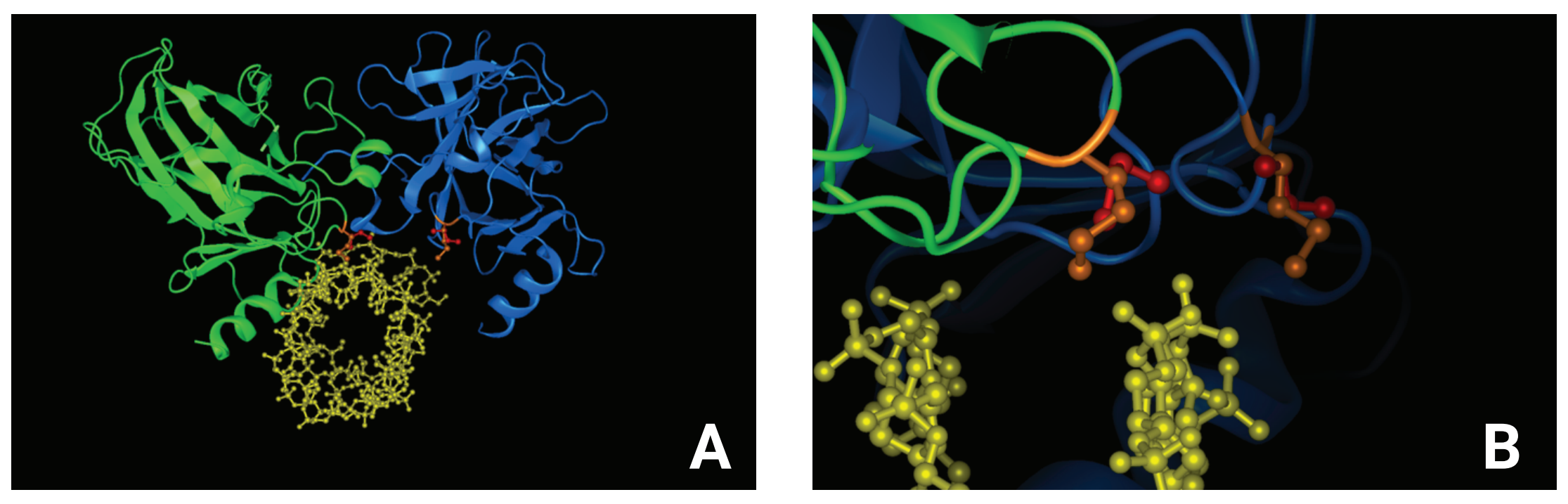


Figure 4. Protean 3D structure prediction of the M237I mutation in TP53 indicates it could affect DNA binding. (A) Crystal structure of a TP53 dimer (green and blue subunits) complexed with DNA (backbone shown in yellow ball and stick rendering). Methionine at position 237 (gold ball and stick) in each 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 affect binding and would in turn likely be involved in pathogenicity.

Discussion

The ability to detect and annotate variants from NGS data or VCF files, and integrate the results with multiple variant and gene annotation databases, allows researchers to quickly identify important or interesting mutations for larges number of samples. Fast and intuitive filtering tools allow users to filter on a variety of criteria and combine filtering results in unique ways. Analysis is not limited by the starting data format as long as a suitable reference genome is available.

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.