

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 most well equipped laboratories.

g genetic 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.

Genomics Suite Pipeline

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

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.

Variants from 192 paired normal-tumor NGS data sets aligned to hGRC382	SeqMan NGen NGS read mapping and variant detection
Annotate called variants	Variant Annotation Database Allele frequencies, functional impact, evolutionary conservation, pathogenicity
Subtract normal from tumor Identify likely deleterious mutations Identify candidate genes with deleterious mutations in multiple samples	ArrayStar Multi-sample comparisons, statistical analyses, discrete filtering and visualization
Predict effect of missense mutations on 3D protein structure and energetics	Protean 3D 3D protein structure predictions, analysis and visualization

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.

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.

Search Crite	ria:					
Search for:	● Genes ○ Varian	ts 🔿 Exo	ns 🔘 Isoform	s Ge	enes: All	
Match all	l of the following) Match an	of the followir	a		Clear
	SNP Sample Group	~	Find genes wit	SNPs in At least 5 of 98 experiments: 1019 83 E.N.P.T2aN0M0 L Choose SNP	Criteria	
	Sill Sample Group		rind genes int		citteria	
				SNP Searching Criteria	1	
				SNP Selection Criteria Summary: (Nonsense OR Frameshift), SNP % ≥ 15.00% and ≤ 100.00% , Depth ≥ 10		
					0	Search
			_			
Search Resu	lts			General Statistics Population Genetics Functional Prediction Evolutionary Conservation Pathogenicity		
1 🎟 🔊				SNP Type V		
QSeq ID	/qseq_name	Name	/db_xref	SNP Genotype V	g /note	
Z-007675	ENSG00000168702	LRP1B		SNP Class	LDL recept	tor related protein 1
Z-008174	ENSG00000155657	TTN		Called as change: @ Yes: O No. O Don't check	titin [Source	ce:HGNC Symbol, Ac
Z-014286	ENSG00000109670	FBXW7			F-box and	WD repeat domain
Z-025269	ENSG00000164796	CSMD3		Located in region: O Intergenic O Genic O Coding O Non-coding RNA	CUB and S	Sushi multiple domai
Z-029418	ENSG00000171862	PTEN		O Don't check	phosphata	ase and tensin homo
Z-044970	ENSG00000141510	TP53		In splice site: O Yes O No O Don't check	tumor pro	tein p53 [Source:HG
				Indistation		
				Substitution		
				No-start		
				No-stop		
				Frameshift		
				Inframe indel		
				Within targeted regions: Don't check Yes No		
				Within targeted regions. O Don't check () Tes () No		
<						



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 SRP030634¹ **Assembly Time**: 84h, Variant Quantification Time: 40h, Total 124h, 37min/ Exome **Scripting**: Arraystar / SeqMan NGen custom batch script used for pairwise comparision filter.

Community of the second	Report nonn	110_00017										
s Summary												
onfirmed Variant tative Variant	Show All	I	Filter	Show Counts as a perce	nt 🗌	Show Codon Bases	s & Distance to feature					
xed Variant	19 ∨ariant 0 Rejecteo	Positions 0 Confin	med 19 Put	ative O Mixed 316 Filter	ed 1 Selected							
Contig ID	Ref Pos	Ref Base	Called Ba	re Impact	SNP %	dbSNP ID	PDB ID	Feature Name	Transcript ID	DNA Change	Amino Acid Change	Depth
000017.11	7482960	TG	CA	Non-synonymous	18.82 %	7219992		SLC35G6	ENST00000412468.3	c.976 977delinsCA	p.W326Q	34
000017.11	7676154	G	C	Non-synonymous	73.33 %	1042522	2GS0.B.2LY4.B	TP53 [1]-*	ENST00000269305.8	c.215C>G	p.P72R	45
000017.11	7673707	T	A	Nonsense	52.83 %		1TSR.A., A 2J1WA	TP53 [1]-*	ENST00000269305.8	c.913A>T	p.K305.	53
000017.11	80319389	A	Т	Non-synonymous	100.00 %	55996424		RNF213 [5]	ENST00000319921.4	c.3101A>T	p.K1034M	66
000017.11	18161408	С	т	Non-synonymous	23.18 %	140029076		MY015A [1]-*	ENST00000205890.9	c.9478C>T	p.L3160F	69
000017.11	53823368	c	Т	Non-synonymous	37.89 %	3803824		KIF2B	ENST00000268919.5	c.335C>T	p.A112V	95
000017.11	10500677	G	A	Non-synonymous	67.01 %	117616137		MYH1	ENST00000226207.5	c.3814C>T	p.R1272W	97
		-										
Position	on: 7,673,7	'08 Depth	: 60								83257.	441kb
200		-	70.0	10 7 673 650 7 673 66	0 7.673.670	7.673.680 7.6	673 690 7 673 700	7.673.710 7.6	73 720 7 673 730 7 67	3,740 7,673,750 7,6	673,760 7,673,77	70
	nce Coordinater	0	(3.6	•0 r.0rJ.0J0 r.0rJ.06								
🛛 🗹 Refere	nce Coordinates	S	/3,64	····························				فتتلبينا تبيتا	. lil.ilil.i	أسطينا أسطين		
Reference 🖌 🕞 Reference	nce Coordinates slate 🕨 Cor	s nsensus	73,6 GTGA	ATCTGAGGCATAACTGCACCO	CTTGGTCTCCTC	CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGAT	TCTCTTCCTCTGTGC	GCCGGT
Reference ► Trans ► NC_00	nce Coordinates slate Cor 00017.11(1>	s nsensus -83257603	(3,6) GTGA	ATCTGAGGCATAACTGCACCI	CTTGGTCTCCTC	CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGAT	TCTCTTCCTCTGTGC	GCCGGT
Reference NC_00 NC_	nce Coordinates slate P Cor 00017.11(1>	s nsensus -83257603	(3,6- GTGA) →	ATCTGAGGCATAACTGCACC	CTTGGTCTCCTC	CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGAT		GCCGGT
	nce Coordinates slate 🕨 Cor 00017.11(1>	s nsensus 83257603	(3,6) GTGA) →	ATCTGAGGCATAACTGCACC	CTTGGTCTCCTC	CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGAT		GCCGGT
Refere	nce Coordinates slate Cor 00017.11(1>	s nsensus 83257603	(73,6- GTGA) →	ATCTGAGGCATAACTGCACC	CTTGGTCTCCTC	CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	TIAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT NSRA: TI CD3: TP5 P. L. E. N. H. P. E.	CCCCTTTCTTGCGGAGAT		GCCGGT
	nce Coordinates slate D Cor 00017.11(1>	s nsensus 83257603	73,6- GTGA) →	ATCTGAGGCATAACTGCACC					GGGGCAGCTCGTGGTGAGGCT mRNA: TI CDS: TPS P L E H H P E CDS: CDS	CCCCTTTCTTGCGGGGGT		B
	nce Coordinates	s nsensus 83257603	73,6- GTGA) → GTGA	ATCTGAGGCATAACTGCACC	CTTGGTCTCCTC	CACCGCTTCTTGT		x T S C P CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT RRNA: TI CDS: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGAT 23 [1]-* 6 X X B L B CCCCTTTCTTGCGGAGAT	E E E T R	B
	nce Coordinate: slate ▶ Cor 00017.11(1> 341680_f 002742 r	s nsensus 83257603	(TGA) GTGA (GTGA) (GTGA) (GTGA) (GTGA) (GTGA) (GTGA)	ATCTGAGGCATAACTGCACC	CTTGGTCTCCTC	CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	X T 3 6 P CTTIAGTGCTCCCTG CTTIAI CTTIAI	GGGGCAGCTCGTGGTGAGGCT wERA: TI CDS: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGAT	E E E T R	B
	nce Coordinate slate Cor 00017.11(1> 0002742_r 1002742_r	s nsensus 83257603	(73,6) (75,	ATCTGAGGCATAACTGCACCI ATCTGAGGCATAACTGCACCI ATCTGAGGCATAACTGCACCI ATCTGAGGCATAACTGCACCI	CTTGGTCTCCTC(CTTGGTCTCCTC(CTTGGTCTCCTC(CTTGGTCTCCTC(CTTGGTCTCCTC)		CCTGCTTGCTTACCTCG	X T S 6 P CTTIAGTGCTCCCTG CTTIAGTGCTCCCTG CTTIAGI CTTIAGI CTTIAGI	GGGGCAGCTCGTGGTGAGGCT mE20A: T CDS: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGAT	E E E T R	B B B CCCGGT
✓ Reference ♦ Transition ♦ NC_00 ID_23 (Ø) ID_30 ID_16	nce Coordinate: slate Cor 00017.11(1> 341680_f 102742_r 470520_f 45900_r	s 15ensus 83257603	(73,6) (75,	ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC		A R CCTGCTTGCTTGCTTGCTTGC CCTGCTTGCTTGCTTGC	X T S C P CTTIAGTGCTCCCTG CTTIAGTGCTCCCTG CTTIAGI CTTIAGI CTTIAGI	GGGGCAGCTCGTGGTGAGGCT «S284: TI CD3: TP9 P L E H H P E GGGGCAGCTCGTGGTGAGGCT cl	CCCCTTTCTTGCGGAGAT	E E E T B	B
	nce Coordinate: slate Cor 00017.11(1> 841680_f 002742_r 170520_f 545900_r 185864 r	s 15ensus 83257603	(73,6) (75,	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A B CCTGCTTGCTTACCTCG A B CCTGCTTGCTTGCTTGC CCTGCTTGCTTGCTTGC CCTGCTTGCT	X T S C P CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAG CTTAGI CTTAGI CTTAGI CTTAGI CTTAGI CTTAGI CTTAGI CTTAGI CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT MRNA: TI CDS: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI cl	CCCCTTTCTTGCGGAGAT	E E E T R	B
	nce Coordinate: slate Cor 20017.11(1> 341680_f 102742_r 470520_f 545900_r 485864_r 708180 f	s 13ensus 83257603	(73,6) (75,	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	X T S G P CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT RRNA: TI CDS: TP9 P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI GI GI	CCCCTTTCTTGCGGAGAT	E E E T R	B
	nce Coordinate: slate	s 13en3u3 83257603	(73,6) (75,	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A B CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG	X T G P CTTIAGTGCTCCCTG CTTIAG CTTIAG CTTIAG CTTIAG CTTIAG CTTIAG CTTIAGTGCTCCCTG CTTIAGTGCTCCCTG CTTIAGTGCTCCCTG CTTIAGTGCTCCCTG CTTIAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT mRNA: TI CDS: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI GI GI GI	CCCCTTTCTTGCGGAGAT 23 [1] 3 [1]-* 6 X X B L N CCCCTTTCTTGCGGAGAT	E E E T R	B GCCGGT
✓ Referee ♦ Transi ♦ Transi ♥ NC_00 ID_23 ID_30 ID_30 ID_44 ID_44 ID_44 ID_47 ID_47 ID_47 ID_36	nce Coordinate: slate	s 13en3u3 83257603 [2]	(73,6) (75,	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	TTGGTCTCCTCI TTGGTCTCCTCI TTGGTCTCCTCI TTGGTCTCCTCI TTGGTCTCCTCI TTGGTCTCCTCI TTGGTCTCCTCI TTGGTCTCCTCI		CCTGCTTGCTTACCTCG	X T S C P CTTAGTGCTCCCTG X T S C P CTTAGTGCTCCCTG CTTAG CTTAGI CTTAGI CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT wRRA: TI CD3: TP5 P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI GI GI GI GI GI GI	CCCCTTTCTTGCGGAGATT 2 [1] 3 [1]-* 5 X R L R CCCCTTTCTTGCGGAGATT	E E E T R	B GCCGGT R GCCGGT
 ✓ Reference ▶ Transi ▼ NC_00 ID_233 ID_303 ID_304 ID_447 ID_447 ID_447 ID_446 ID_446 ID_447 	nce Coordinate: slate Cor 00017.11(1> 341680_f 102742_r 470520_f 545900_r 185864_r 708180_f 37898_r_dup 522350_f 514170 r	s 13ensus 83257603 [2]	(73,6) (75,	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCCCCTCC CTTGGTCCCCTCC CTTGGTCCCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A B CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG	x T 3 6 P CTTAGTGCTCCCTG TTAGTGCTCCCTG CTTAG CTTAGI CTTAGI CTTAGI CTTAGI CTTAGI CTTAGI CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGGAGAT 3 [1] 5 X X R L R CCCCTTTCTTGCGGAGAT	E E E T B	B CCCGGT
 ✓ Reference ▶ Transi ▼ NC_00 ID_233 ID_303 ID_304 ID_444 ID_474 ID_474 ID_484 ID_435 ID_355 	nce Coordinate: slate ▶ Cor 20017.11(1> 341680_f 302742_r 470520_f 545900_r 485864_r 708180_f 37898_r_dup 522350_f 514170_r 555434 f	s 13en3u3 83257603 [2]	(73,6) (75,	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A B CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG	Image: Contract of the second secon	GGGGCAGCTCGTGGTGAGGCT MRNA: TI CDS: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI GI GI GGI GGI GGI GGI GGGCAGCI	CCCCTTTCTTGCGGAGAT	E E E T R TCTCTTCCTCTGTGC	B B B CCCGGT
✓ Reference ✓ Reference ✓ Transist ✓ NC_00 ✓ ID_233 ✓ ID_303 ✓ ID_304 ✓ ID_443 ✓ ID_445 ✓ ID_356 ✓ ID_266	nce Coordinate: slate Cor 20017.11(1> 341680_f 3002742_r 370520_f 545900_r 185864_r 708180_f 37898_r_dup 522350_f 514170_r 5555434_f 334032 r	s 13en3u3 83257603 [2]	(73,6) (75,	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	Image: Contract of the second secon	GGGGCAGCTCGTGGTGAGGCT "RNA: TI CDS: TPS P L E H H P E GGGGGCAGCTCGTGGTGAGGCT GI GI GI GGI GGI GGI GGGI GGGGCAGCI GGGGCAGCTCGI	CCCCTTTCTTGCGGAGAT 752 [1] 76 X X R L N CCCCTTTCTTGCGGAGAT	E E E T R	B
	nce Coordinate: slate Cor 200017.11(1> 341680_f 0002742_r 470520_f 545900_r 485864_r 708180_f 37898_r_dup 522350_f 514170_r 555434_f 534032_r 266772_r	s 13en3u3 83257603 [2]	$\begin{array}{c} 73, 6\\ \hline \\ $	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	CCTGCTTGCTTACCTCG	X T S C P CTTAGTGCTCCCTG X T S C P CTTAGTGCTCCCTG CTTAGI CTTAGI CTTAGI CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT wBBA: TI CD3: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI GI GI GGI GGI GGI GGGCAGCT GGGGCAGCI GGGGCAGCTCGI GGGGCAGCTCGI GGGGCAGCTCGI	CCCCTTTCTTGCGGAGATT 2 [1]-* 5 X X R L N CCCCTTTCTTGCGGAGATT	E E E T R	B CCCGGT
 ▶ Tran: ▶ Tran: ▼ NC_00 □ D_23 □ D_23 □ D_44 	nce Coordinate: slate Cor 200017.11(1> 000017.11(1> 00017.11(1> 00017.11(1> 00017.11(1> 00017.11(1> 0	s 13en3u3 83257603 [2]	$\begin{array}{c} 736\\ \hline \\ \hline$	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A R CCTGCTTGCTTGCTTGCCTGC CCTGCTTGCTTGCCTGC CCTGCTTGCT	X T S C P CTTAGTGCTCCCTG TTAGTGCTCCCTG CTTAI CTTAGI CTTAGI CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGATT 9 [1] 9 [1]-* 6 K K R L R CCCCTTTCTTGCGGAGATT	E E E T B	B GCCGGT
 ✓ Reference ▶ Transi ▶ NC_00 ID_23 ID_303 ID_304 ID_304 ID_447 ID_447 ID_447 ID_454 ID_34 ID_454 ID_447 ID_441 ID_441	nce Coordinate: slate Cor 00017.11(1> 341680_f 102742_r 470520_f 485864_r 708180_f 37898_r_dup 522350_f 514170_r 555434_f 334032_r 266772_r 195774_f 3916 r	s 13en3u3 83257603 [2]	$\begin{array}{c} 73,6\\ 73,6\\ 74,7\\$	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A B CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG CCTGCTTGCTTACCTCG	Image: Contract of the second secon	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGGAGAT 3 [1] 5 X X R L R CCCCTTTCTTGCGGAGAT	E E E T B	B
	nce Coordinate: slate Cor 20017.11(1> 341680_f 102742_r 170520_f 545900_r 185864_r 708180_f 37896_r_dup 522350_f 514170_r 555434_f 334032_r 266772_r 195774_f 33916_r 55842_f	s 1sensus 83257603 [2]	$\begin{array}{c} 736\\ \hline 736\\ \hline 7764\\ \hline \\ \hline$	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCC ATCTGAGCATAACTGCACCCC ATCTGAGCATAACTGCACCCC ATCTGAGCATAACTGCACCCCC ATCTGAGCATAACTGCACCCCC ATCTGAGCATAACTGCACCCCCCCCCCCCCCCCCCCCCC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A B CCTGCTTGCTTACCTCG CCTGCTTGCTTCCTCG CCTGCTTGCTTCCTCG CCTGCTTGCT	Image: Contract of the second secon	GGGGCAGCTCGTGGTGAGGCT MRNA: TI CDS: TPS P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI GI GI GGI GGI GGG GGG GGGCAGCTCGT GGGGCAGCTCGTI GGGGCAGCTCGTGGTI GGGGCAGCTCGTGGTI GGGGCAGCTCGTGGTI GGGGCAGCTCGTGGTG	CCCCTTTCTTGCGGAGAT	E E E T R TCTCTTCCTCTGTGC	B CCCGGT
	nce Coordinate: slate Cor 200017.11(1> 341680_f 3002742_r 370520_f 545900_r 185864_r 708180_f 37898_r_dup 522350_f 514170_r 555434_f 354032_r 266772_r 195774_f 33916_r 55842_f 534114_f	s 13en3u3 83257603 [2]	$\begin{array}{c} 73,6\\$	ATCTGAGGCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGCACCACTGCACCC ATCTGCACCACCACCACCACCACCC ATCTGCACCACCACCACCC ATCTGCACCCACCACCACCCC ATCTGCACCCACCACCACCCCCACCCC ATCTGCACCCCCACCCCCCCCCCCCCCCCCCCCCCCCCC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	CCTGCTTGCTTACCTCG CCTGCTTGCTTTCCTCG CCTGCTTGCTTTCCTCG CCTGCTTGCT	X T S C P CTTAGTGCTCCCTG X T S C P CTTAGTGCTCCCTG CTTAG CTTAGI CTTAGI CTTAG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT wBBA: TI CD3: TP5 P L E H H P E GGGGCAGCTCGTGGTGAGGCT GI GI GGI GGI GGI GGGCAGCTCGI GGGGCAGCI GGGGCAGCTCGI GGGGCAGCTCGI GGGGCAGCTCGTGGTGI GGGGCAGCTCGTGGTGI GGGGCAGCTCGTGGTGI GGGGCAGCTCGTGGTGAI	CCCCTTTCTTGCGGAGATT 253 [1] 2 [1]-* 5 X X B L N CCCCTTTCTTGCGGAGATT	E E E T R	B
	nce Coordinate: slate Cor 200017.11(1> 000017.11(1> 00017.11(1> 00017.11(1> 00017.11(1> 00017.11(1> 0	s 13en3u3 83257603 [2]	$\begin{array}{c} 736\\ \hline \\ 776\\ \hline \\ \hline$	ATCTGAGGCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATAACTGCACCC ATCTGAGCCATACTGCACCC ATCTGCACCACCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A R CCTGCTTGCTTACCTCG	X T S C P CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAG CTTAGI CTTAGI CTTAGI CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG CTTAGTGCTCCCTG	GGGGCAGCTCGTGGTGAGGCT GGGGCAGCTCGTGGTGAGGCT GI GI GI GI GI GGI GGI GGI GG	CCCCTTTCTTGCGGAGATT 9 [1] 9 [1]-* 6 X X R L R CCCCTTTCTTGCGGAGATT	E E E T R	B
▶ Tran: ▶ Tran: ▶ Tran: ▶ NC_00 □ <	nce Coordinate: slate Cor 200017.11(1> 341680_f 341680_f 302742_r 470520_f 485864_r 708180_f 37898_r_dup 522350_f 314170_r 555434_f 334062_r 266772_r 195774_f 33916_r 55842_f 534114_f 102742_f 5552_f	s 13en3u3 83257603 [2]	$\begin{array}{c} 73,6\\ 73,6\\ 74,7\\$	ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC CATCTGAGGCATAACTGCACC ATCTGAGGCATAACTGCACC C TCGAGGCATAACTGCACC C TGGAGGCATAACTGCACC C C GAGGCATAACTGCACC C C CGAGGCATAACTGCACC C C CGAGGCATAACTGCACC C C C CGAGGCATAACTGCACC C C C C C C C C C C C C C C C C C	CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC CTTGGTCTCCTCC	CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT CACCGCTTCTTGT	A B CCTGCTTGCTTACCTCG	Image: Contract of the second secon	GGGGCAGCTCGTGGTGAGGCT	CCCCTTTCTTGCGGAGATT 959 [1] 9 [1]-* 6 X X R L R CCCCTTTCTTGCGGAGATT	E E E T B	B

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.

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				

NC 000017.11 18153798 C MYO15A 2	Within targeted regions: O Don't check O Yes O No	3.17
<pre> # Variant found: 98 </pre>		>
	Use Defaults Set Defaults OK	Cancel

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

able 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	m Table 2, Step 4				
^b Filtered to genes with lengths	< 250kb				

Structure Visualization and Mutation Modeling



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 1TUP³ 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.

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).



References:

http://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP030634
 Human genome reference with GRCh38 and ENSEMBL annotations
 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 R44HG006578 and by the National Institutes of Health.