doi:10.1038/nature11329

Supplemental Text 1: Detailed materials and methods

Patient selection & consent

Informed consent to collect and perform genomic studies on tumor tissue and peripheral blood was provided by families of medulloblastoma patients treated at Children's Hospital Boston, The Hospital for Sick Children Toronto and member institutions contributing to the Children's Oncology Group tumor bank through the Cooperative Human Tissue Network. All tumors in this cohort were obtained at the initial surgical resection and recurrent tumors were excluded from our analysis. Hematoxylin and eosin stained slides of tumor samples in this study were pathologist reviewed to confirm the diagnosis of medulloblastoma, determine histological subtype when able, and assess tumor purity. DNA was isolated from snap-frozen tumor specimens and matched peripheral blood using the Gentra PureGene DNA extraction kit, as previously described[1].

Exome sequencing and analysis

The generation, sequencing, and analysis of 92 pairs of exome libraries at the Broad Institute was performed using a detailed, previously described, protocol[2] and methods described at http://www.broadinstitute.org/cancer/cga/. A summary of deviations from the published protocol is provided here. Exonic regions were captured by in-solution hybridization using RNA baits similar to those described[2] but supplemented with additional probes capturing additional genes listed in ReqSeq[3] in addition to the original Consensus Coding Sequence (CCDS)[4] set. In total, ~33 Mb of genomic sequence was captured, consisting of 193,094 exons from 18,863 genes annotated by the CCDS and RefSeq databases as coding for protein or micro-RNA (accessed November 2010). Sequencing-by-synthesis (SBS) of 76 bp paired-end reads was performed using Illumina HiSeq 2000 instruments by the Broad Sequencing Platform. Reads were aligned to the hg19 (GRCh37) build of the human reference genome sequence using BWA[5] and further processed by the Broad Picard pipeline including marking of duplicate reads, realignment around suspected indels, and recalibration of base quality scores[6]. Further analysis was coordinated by the Broad Cancer Program's Firehose software. Candidate somatic base substitutions were detected using muTect (previously referred to as muTector[2], Cibulskis et al, in preparation). Candidate somatic insertions and deletions were detected using Indelocator, as previously described[2] (Sivachenko et al, in preparation). Mutations were annotated using Oncotator[7] (Ramos et al, in preparation). To confirm sample identity, fingerprint genotyping of 24 SNPs was used to match tumor/normal pairs and copy number profiles derived from sequence data (CapSeg, see below) were compared with those previously derived from microarray data from each case, when available.

Significance analysis

Genes mutated at a statistically significant frequency were identified using version 2.0 of MutSig[8], a method that corrects for gene length and callable sequence in each

tumor/normal pair as well as sample-specific mutation rates, non-silent to silent mutation ratio, clustering of mutations within genes, and base conservation across species. Implementations of this algorithm have been described previously[9–11]. This method was applied to the entire cohort to identify genes frequently mutated in medulloblastoma (see main text, Table 1) as well as tumors from each subtype independently (Supplemental Table 3). Pathway and gene-set analysis achieved by grouping genes into sets annotated by the Gene Set Enrichment Analysis software (GSEA)[12] and testing each set for increased mutation frequency relative to the other sets. This approach, including the gene set annotations, is identical to our previous report[2].

Validation of candidate mutations

Analysis of whole exome sequence data from 92 medulloblastoma/normal pairs uncovered twelve genes with somatic mutations occurring at statistically significant frequencies. This list included known medulloblastoma genes *CTNNB1* (beta-catenin), *TP53*, and *SMARCA4* as well as novel candidates *DDX3X* and *CTDNEP1* (also known as *DULLARD*). To verify these mutations using an orthogonal technology and assess whether additional variants were missed during exome sequencing, we used a microfluidic PCR platform (Fluidigm Access Array)[13] to amplify 48 targets from 96 medulloblastomas and sequenced them using single molecule real-time sequencing technology (SMRT, Pacific Biosciences). In total, these targets cover:

1) All coding amplicons (no UTRs) from *CTNNB1*, *DDX3X*, and *CTDNEP1*
2) 12 of 35 exons from *SMARCA4* that encode two helicase domains with somatic mutations in our medulloblastomas
3) 6 of 11 exons from *TP53* that contained somatic or germline variants in 80 medulloblastoma exomes and, according to COSMIC, are the most frequently mutated exons in cancer.

PCR was performed using 50 ng in batches of 48 samples and 48 PCR primer pairs in parallel using the Access Array (Fluidigm) following the manufacturer's recommendations. PCR primers were tailed with molecular barcode sequences to facilitate subsequent assignment of sequencing reads to individual samples therefore enabling sample multiplexing within a sequencing run. Pacific Biosciences provided barcode sequences and pre-release barcode identification software.

In total, 96 medulloblastoma DNAs were processed in two batches. 48 of these were from whole genome amplified (WGA) material and 48 were using native genomic DNA. Of the 96, 87 were from our discovery set and 9 were new samples not part of the discovery cohort. Digital PCR was performed using 50 ng from each of 48 samples and 48 PCR primer pairs in parallel using a commercially-available microfluidic device as recently described[13] and following manufacturer's instructions (Fluidigm Access Array). PCR products were then pooled in sets of 48 samples and subjected to standard Pacific Biosciences library construction and sequencing following the manufacturer's recommendations (version 1.2.3 system specifications). A total of 20 sequencing runs yielded a total 389,215 post-filtered

reads, resulting in 4,367,852 filtered subreads. As a result of barcode identification, 2,834,170 subreads (64%) were assigned to samples.

Reads were aligned to the hg19 (GRCh37) build of the human reference genome sequence using BWA-SW[5], with settings tolerant of the high indel artifact rate present in these reads. Specifically, the command "bwa bwasw -b5 -q2 -r1 -z20 -t16" was used to:

-b5 increase the mismatch penalty to 5 [default 3]
-q2 decrease the gap open penalty to 2 [default 5]
-r1 decrease the gap extension penalty to 1 [default 2]
-z20 increase Z-best to 20, as PCR amplicons are highly targeted [default 1]
-t16 ran 16 threads on multi-core machine [default 1]

This alignment was followed by base-quality recalibration using the Genome Analysis Toolkit[14]. Candidate mutations were confirmed by manual review using the Integrative Genomics Viewer[15]. See Supplemental Figure 1 for screenshots at sites of mutations targeted for validation sequenced by PCR/SMRT and hybrid-capture/SBS.

19 of 19 mutations with sufficient coverage from this experiment were validated by this method. Notably, despite the high indel error rate of SMRT sequencing, a 2 bp deletion in *CTDNEP1* was readily differentiated from the otherwise random indel artifacts currently inherent to this technology. *DDX3X* p.R376C was not confirmed due to a PCR failure (i.e. 0 coverage of that amplicon). Coverage was low for all amplicons from this sample, possibly due to a bad WGA reaction (although other WGA samples had sufficient coverage for mutation validation). We also didn't confirm *CTNNB1* p.D32G or *DDX3X* p.D506Y as there wasn't enough genomic DNA available from these samples for the validation run.

Functional validation

Plasmids construction

The pOTB7-DDX3X wild type plasmid was obtained from Open Biosystems (Lafayette, CO). Point mutations were introduced into DDX3X using the QuikChange II Site-Directed Mutagenesis kit from Agilent Technologies (Santa Clara, CA) and then verified by direct sequencing of the whole gene. The wild type and mutated DDX3Xs were then subcloned into the lentiviral vector pCDH-CMV-MCS-EF1-copGFP (System Biosciences, Mountain View, CA) between EcoR1 and BamH1 sites to generate DDX3 lentiviral expression vectors. β-catenin WT and S33Y plasmids were a gift of Bert Vogelstein (Addgene plasmid #16518 and #16519) which were then subcloned into pCDH-CMV-MCS-EF1-RFP (System Biosciences) between Nhe and Not1 restriction sites.

Lentiviral production and stable cell line generation

pCDH-copGFP-DDX3 and pCDH-RFP-β-catenin constructs were transfected into 293T cells and pseudovirus particles were generated according to manufacturer's instructions (System Biosciences). Subsequently, D425 medulloblsatoma cells were infected at a MOI of 2 in the presence of 5μg/ml polybrene (Sigma, St Louis, MO).

Cell viability assays

Protocol was adapted from methods described in Weeraratne et al., 2010. Briefly, D425 stable cell lines (DDX3 wild-type or mutants, Beta-catenin wild-type or mutant, and both in combination) were seeded at 25000 cells per well in 96 well plates (Corning, Corning, NY) in 100µl of antibiotic free media. Each sample was plated in quintuplicate. Cell viability was determined 48 hours post-plating using the 3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium salt (MTS) assay with CellTiter 96 AQueous One Solution Reagent (Promega, Madison, WI) following the manufacturer's protocol. Finally, the optical density of each well was measured with a spectrophotometric microplate reader (Bio-Rad, Hercules, CA) at 490 nm.

TOPflash assay

TOPflash luciferase reporter containing the TCF/LEF consensus sequence was used to assay Wnt signaling in the presence of various combinations of wild type, mutant DDX3 and β-catenin plasmids. Briefly, 15 x 10^3 cells/well of 293T cells were transfected using Lipofectamine 2000 in 96 well plates (Invitrogen) with 200 ng TOPFlash or FOPFlash plasmids(Millipore, Billerica, MA) together with 50 ng *Renilla* plasmid (Promega). Firefly and *Renilla* luciferase activities were measured 48 hours after transfection by a Dual-luciferase Assay (Promega) using an LB9507 luminometer according to the manufacture's instructions. Firefly luciferase activity derived from TOPFlash or FOPFlash was normalized based on Renilla luciferase activity.

Model of DDX3 in complex with RNA and Mg-ATP

Superposition of human Ddx3 and Drosophila Vasa. The two recA-like domains that form the catalytic core from the crystal structure of Ddx3 (PDB 2I4I)[16] were superimposed individually onto the functionally equivalent domains from the structure of Vasa (PDB 2DB3)[17]. The superposition was produced using the program MatchMaker and the Needleman-Wunsch alignment algorithm within the Chimera suite of programs[18]. Each recA-like domain of Ddx3 supimposed onto Vasa with an r.m.s.d. of >1.1 Å. Molecular graphics images were produced using the UCSF Chimera package (University of California, San Francisco; http://www.cgl.ucsf.edu/chimera)[18].

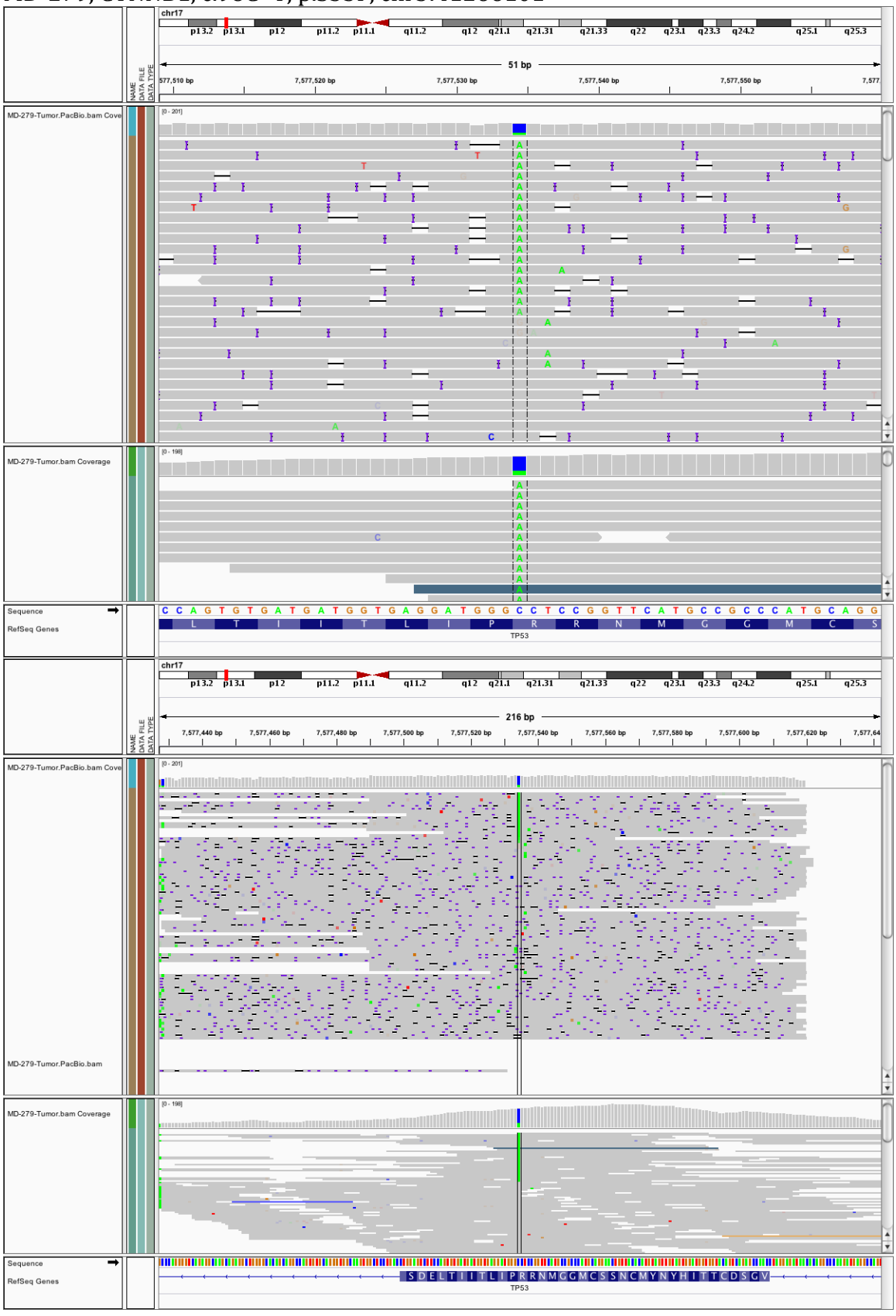
References

1. Cho, Y.-J. *et al.* Integrative Genomic Analysis of Medulloblastoma Identifies a Molecular Subgroup That Drives Poor Clinical Outcome. *Journal of Clinical Oncology* **29**, 1424 -1430 (2011).
2. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).
3. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, D32-36 (2009).
4. Pruitt, K.D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316-1323 (2009).
5. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
6. Picard webpage. at <http://picard.sourceforge.net/>
7. Ramos, Alex *et al.* Oncotator. at <http://www.broadinstitute.org/oncotator/>
8. Getz, G. *et al.* Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers."*Science* **317**, 1500 (2007).
9. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).
10. Wang, K. *et al.* Integrative genomics identifies LMO1 as a neuroblastoma oncogene. *Nature* **469**, 216-220 (2011).
11. Getz, G. *et al.* Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers."*Science* **317**, 1500 (2007).
12. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545 -15550 (2005).
13. Hollants, S., Redeker, E.J.W. & Matthijs, G. Microfluidic Amplification as a Tool for Massive Parallel Sequencing of Familial Hypercholesterolemia Genes. *Clinical Chemistry* (2012).doi:10.1373/clinchem.2011.173963
14. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
15. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat Biotech* **29**, 24-26 (2011).
16. Högbom, M. *et al.* Crystal structure of conserved domains 1 and 2 of the human DEAD-box helicase DDX3X in complex with the mononucleotide AMP. *J. Mol. Biol.* **372**, 150-159 (2007).
17. Sengoku, T., Nureki, O., Nakamura, A., Kobayashi, S. & Yokoyama, S. Structural basis for RNA unwinding by the DEAD-box protein Drosophila Vasa. *Cell* **125**, 287-300 (2006).
18. Pettersen, E.F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).

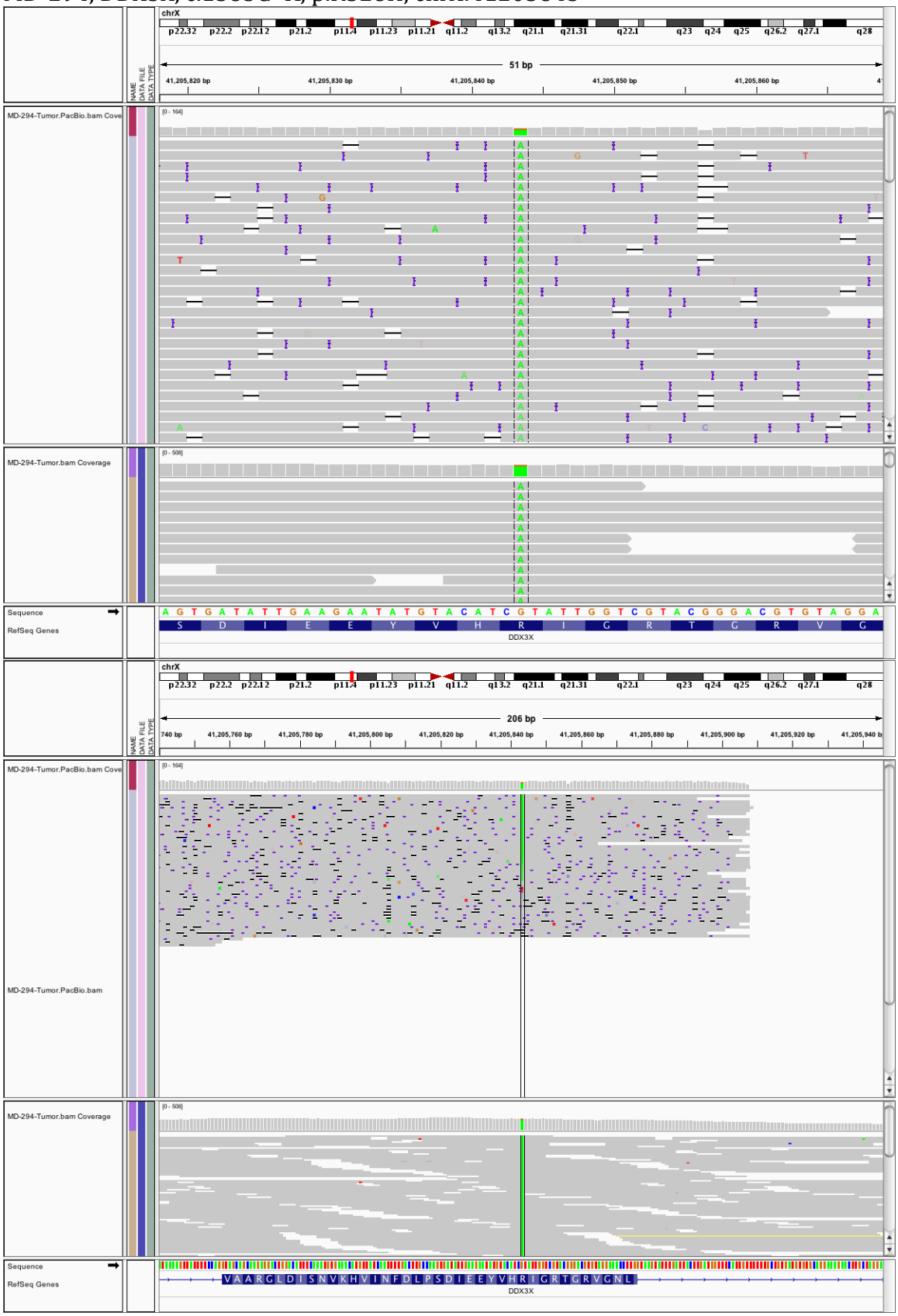
Supplemental Figure 1: Integrated Genomics Viewer screenshots at candidate mutation sites in discovery and validation sequence data sets

Following are representative screenshots taken using the Integrative Genomics Viewer (IGV) of sequence read alignments at sites of candidate somatic mutation sequenced by hybrid-capture followed by sequencing-by-synthesis (SBS, Illumina) and by digital PCR followed by single-molecule real-time sequencing (SMRT, Pacific Biosciences) (See Supplemental Text for methods). The top track contains the SMRT read alignments from PCR products used for validation. The bottom track contains the SBS read alignments generated from hybrid capture products used for discovery. For each mutation, two screenshots were taken. The top screenshot provides base level resolution of reads that support the mutation in a ~50 bp window. The bottom screenshot uses IGV's "squished" option to display all reads at the site and is often of a wider region (~50-200 bp).

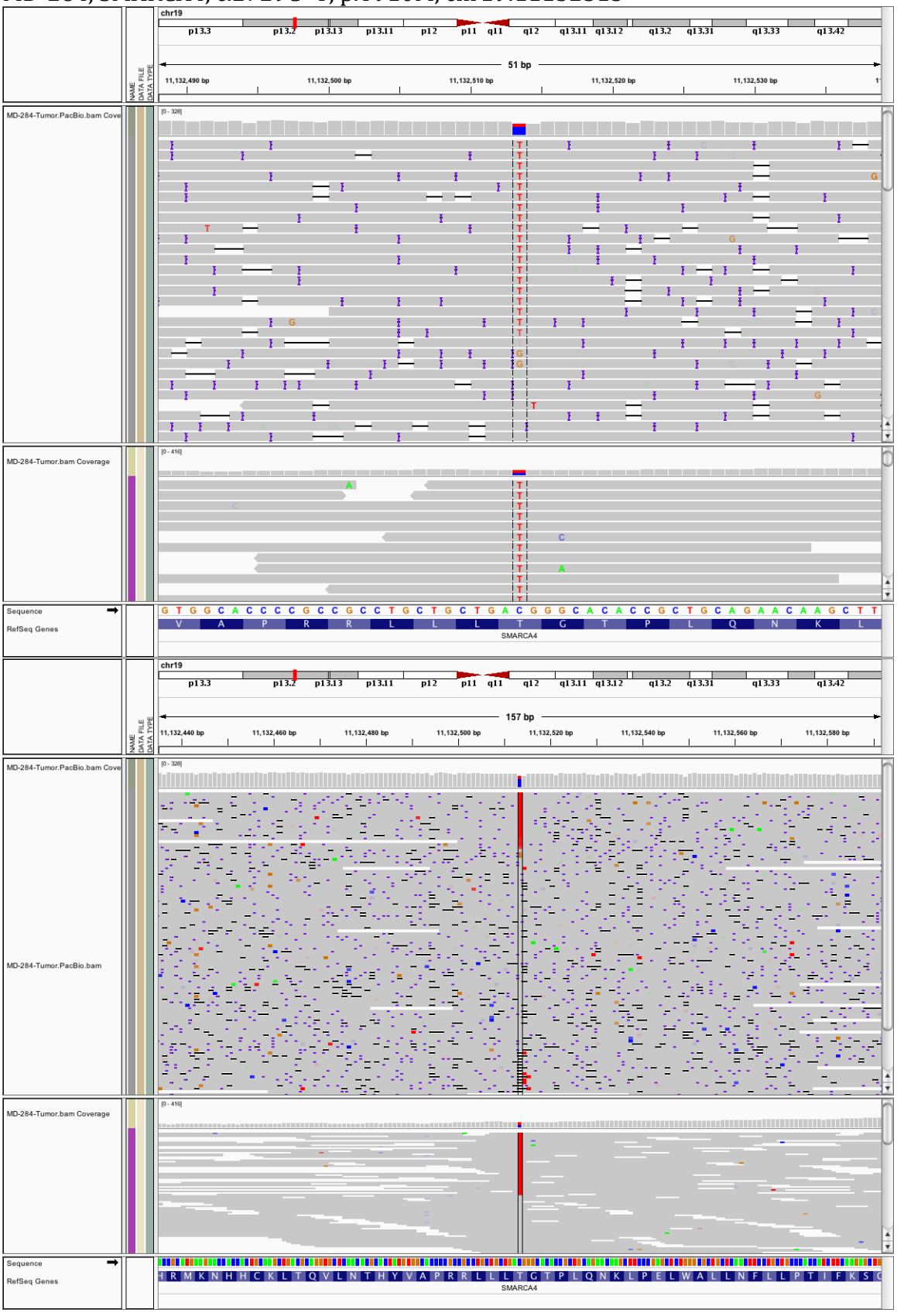
MD-279, CTNNB1, c.98C>T, p.S33F, chr3:41266101



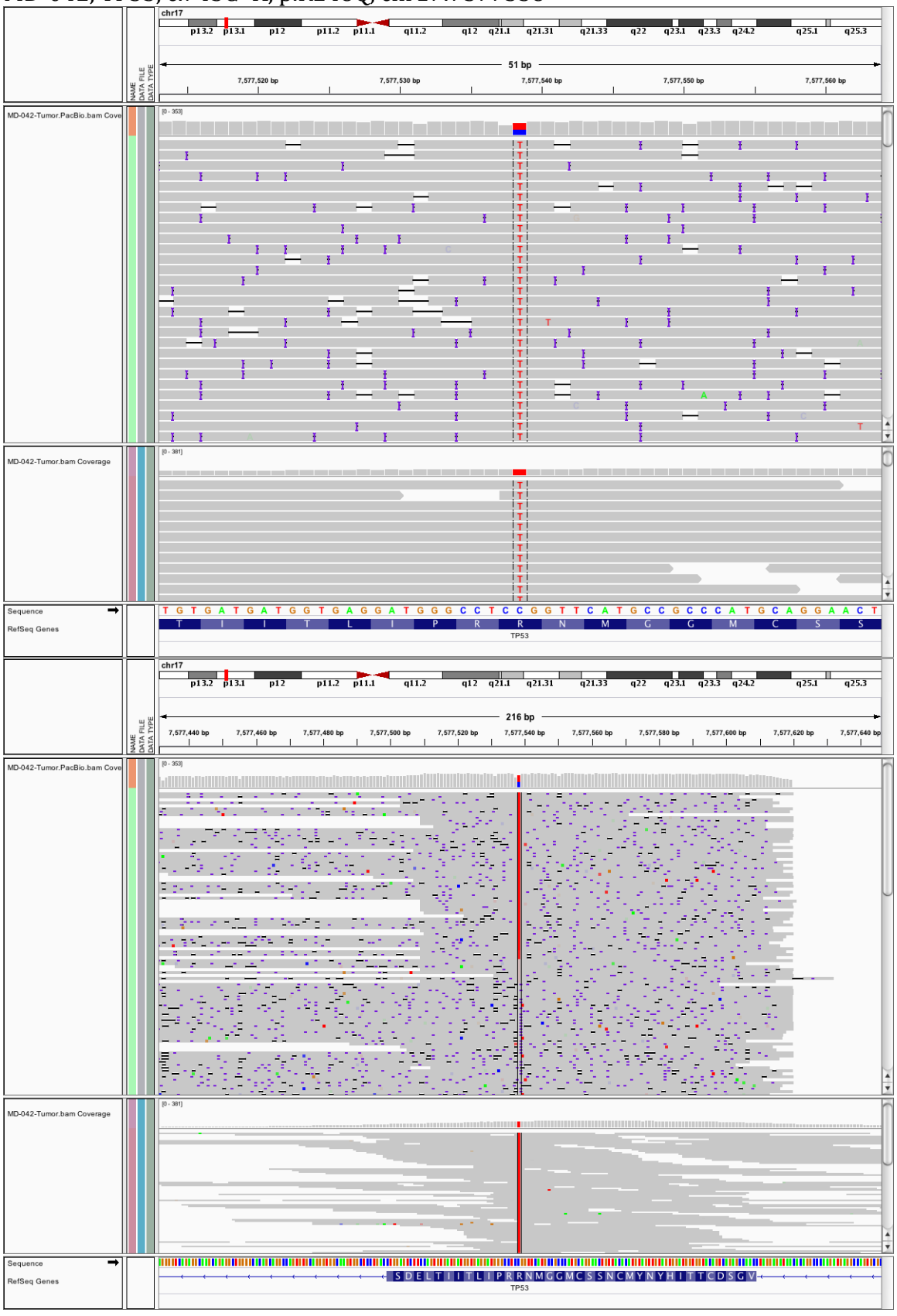
MD-294, DDX3X, c.1583G>A, p.R528H, chrX:41205843



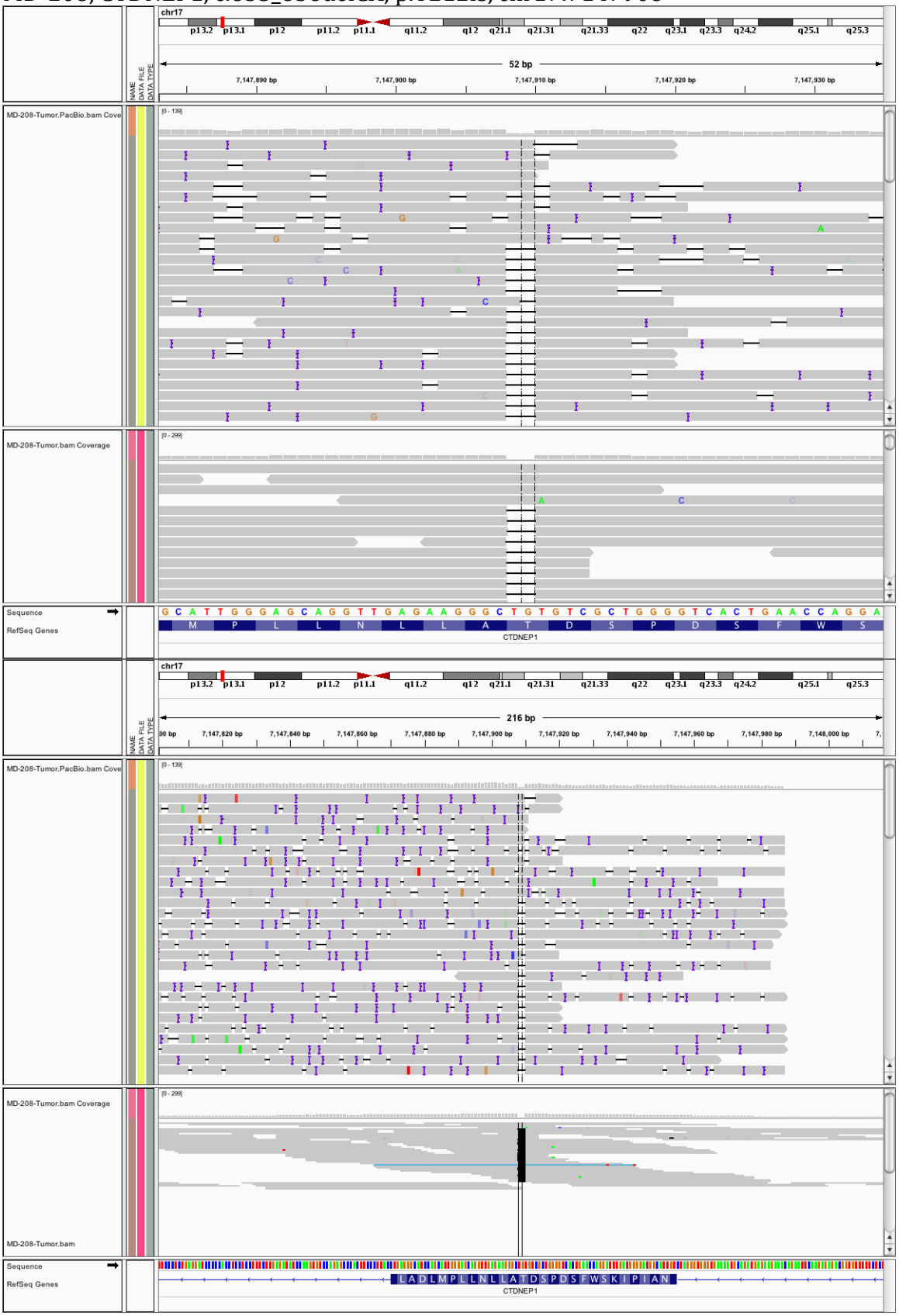
MD-284, SMARCA4, c.2729C>T, p.T910M, chr19:11132513



MD-042, TP53, c.743G>A, p.R248Q, chr17:7577538

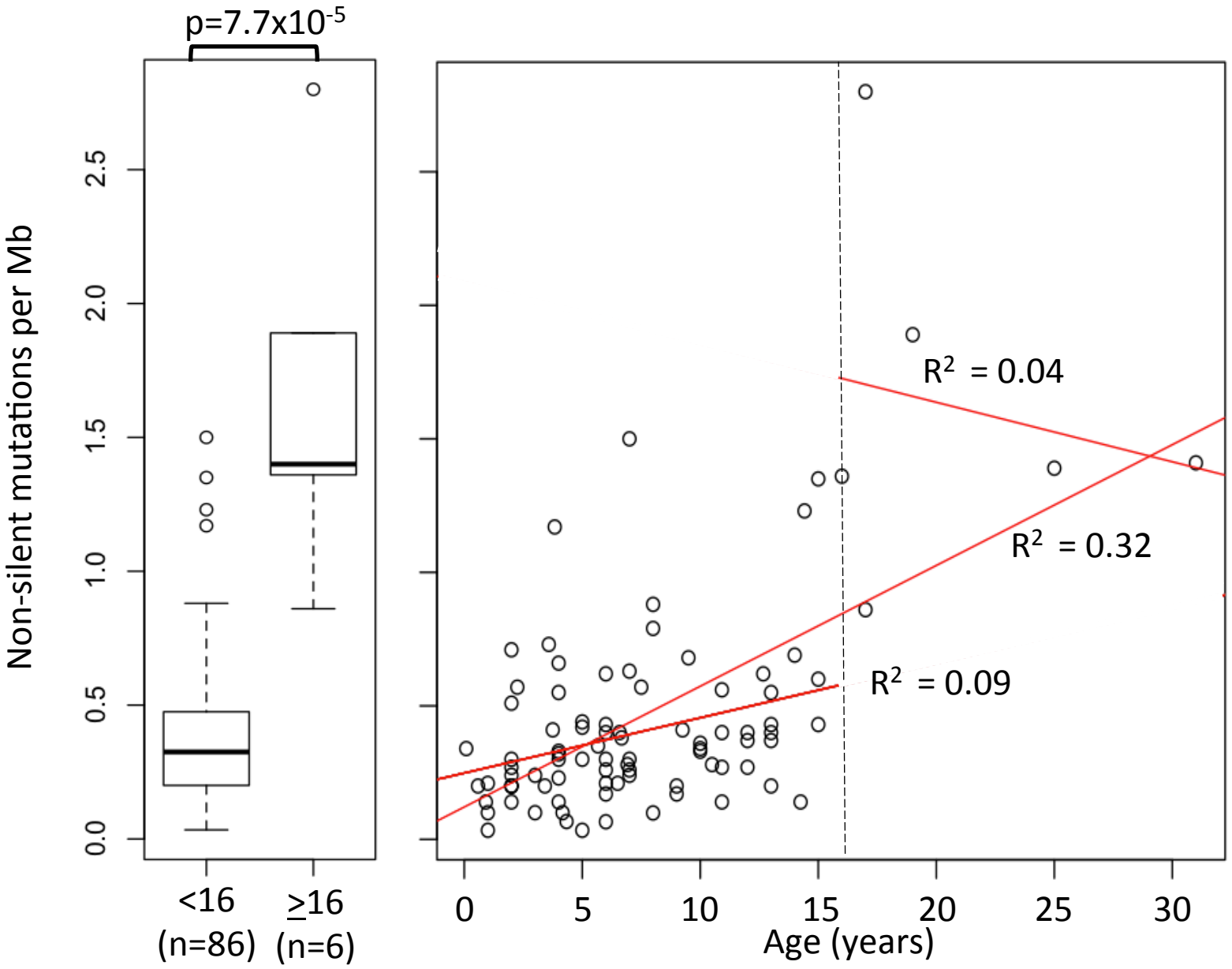


MD-208, CTDNEP1, c.635_636delCA, p.T212fs, chr17:7147908



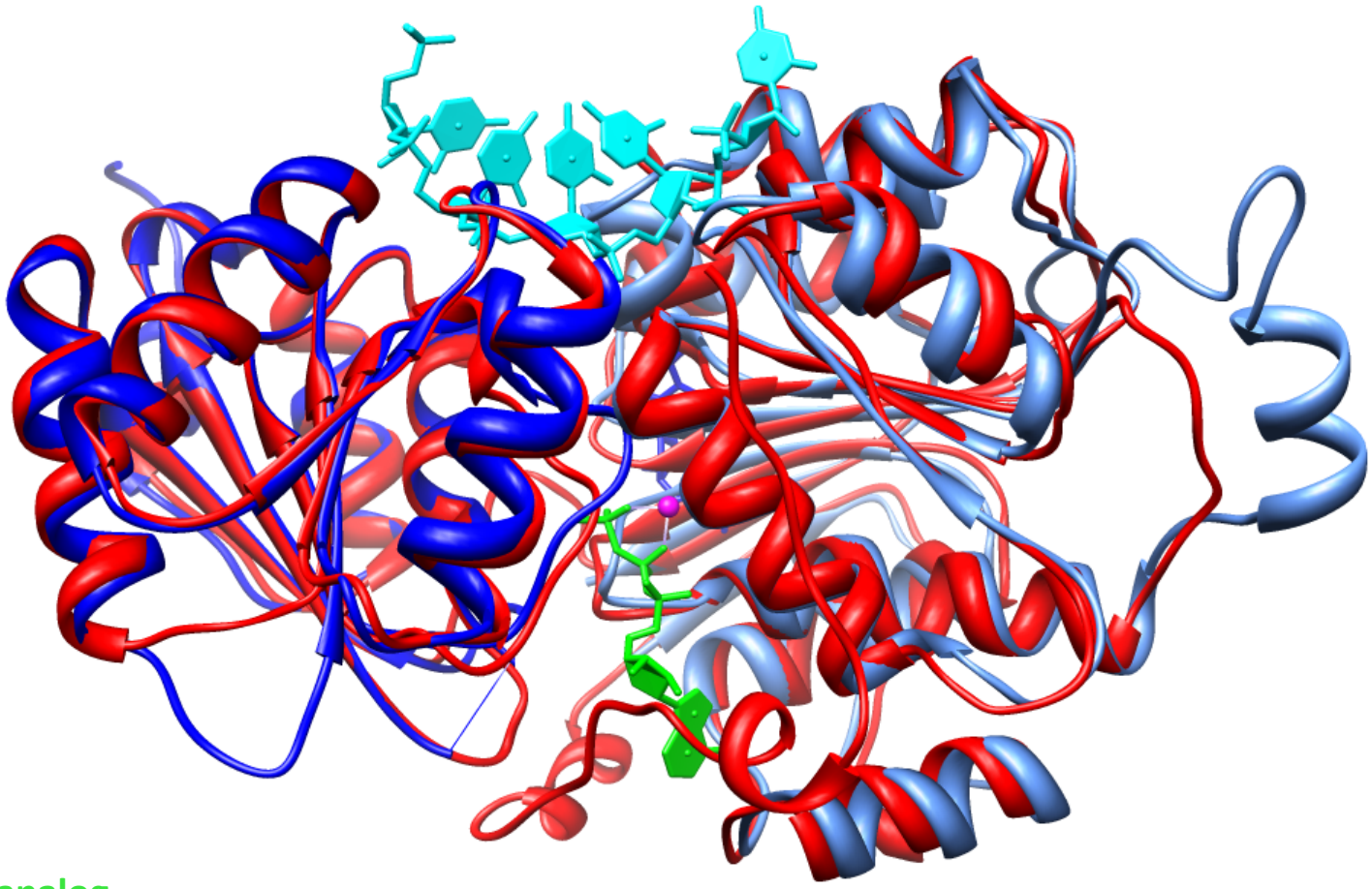
MD-274, CTDNEP1, c.360_splice, p.V120_splice, chr17:7150109

Supplemental figure 2 -- mutation rate versus age (divided into <16 years and >16 years in the left panel, or in dot plot of mutation rate per sample versus age in the right panel)



Supplemental figure 3.

Superposition of the two rec-A like domains of DDX3X (blue) onto Vasa (red)



Vasa
RNA
Mg-ATP analog
166-405
406-582

Supplementary Table 5. Possible impact of DDX3X mutations.

DExD-box helicases can be divided into two structural domains: a helicase ATP-binding domain (residues 211-403 in DDX3X) and a helicase C-terminal domain (residues 414-575 in DDX3X). Most DExD-box helicases contain nine consensus sequences that form the conserved core of the helicase.

Motif	Function(s) of morif	Location in DDX3X	Sequence in DDX3X
Q	RNA recognition	-	-
I	Walker A motif, crucial for ATPase and helicase activities	224-231	AQTGSGKT
Ia	RNA binding	274-279	PTRELA
Ib	RNA binding	323-327	TPGRL
II (DEAD-box)	Walker B motif, necessary for the ATPase activity	347-350	DEAD
III	May participate in linking ATPase and helicase activities	382-384	SAT
IV	RNA binding	401-407	IFLAVGT
V	RNA binding	444-455	TLVFVETKKGAD
VI	Interface between domains 1 and 2 - important for ATPase activity and RNA binding	527-534	HRIGRTGR

DDX3X mutations R276K, D354H, and R376C are in the ATP-binding domain, while mutations D506Y, R528H, R534H, and P568L are in the C-terminal domain.

Mutation	Proposed impact on DDX3X based on structural model
R276K	Part of motif Ia - an interaction with RNA may be disrupted.
D354H	Proximal to DEAD-box motif – ATP binding pocket may be disrupted.
R376C	R376 may form a hydrogen bond with the backbone amide of F396. Mutation may destroy this proposed interaction and disrupt structure.
D506Y	D506 is proximal to VI-motif residues near the interface of domains 1 and 2 - ATPase activity and/or RNA binding may be disrupted.
R528H	R528 is part of the VI-motif – ATPase activity and/or RNA binding may be disrupted.
R534H	R534 is part of the VI-motif – ATPase activity and/or RNA binding may be disrupted.
P568L	Loss of proline may result in extension of the helix into Serine-rich sequence.