2014

Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma

Charles Lu
Washington University School of Medicine in St. Louis
Li Ding
Washington University School of Medicine in St. Louis
Elaine R. Mardis
Washington University School of Medicine in St. Louis
Richard K. Wilson
Washington University School of Medicine in St. Louis
et al

OPEN ACCESS
CellPress

Cell Reports
Report

Recurrent Somatic Structural Variations Contribute to Tumorigenesis in Pediatric Osteosarcoma

Xiang Chen,[1,12] Armita Bahrami,[2,12] Alberto Pappo,[3] John Easton,[1] James Dalton,[2] Erin Hedlund,[1] David Ellison,[2] Sheila Shurtleff,[2] Gang Wu,[1] Lei Wei,[1] Matthew Parker,[1] Michael Rusch,[1] Panduka Nagahawatte,[1] Jianrong Wu,[4] Shenghua Mao,[4] Kristy Boggs,[1] Heather Mulder,[1] Donald Yergeau,[1] Charles Lu,[6] Li Ding,[6] Michael Edmonson,[1] Chunxu Qu,[1] Jianmin Wang,[1] Yongjin Li,[1] Fariba Navid,[3] Najat C. Daw,[5] Elaine R. Mardis,[6,7,8] Richard K. Wilson,[6,7,9] James R. Downing,[3] Jinghui Zhang,[1,*] and Michael A. Dyer,[10,11,*] for the St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project

[1]Department of Computational Biology
[2]Department of Pathology
[3]Department of Oncology
[4]Department of Biostatistics
St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[5]University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[6]The Genome Institute
[7]Department of Genetics
[8]Department of Medicine
[9]Siteman Cancer Center
Washington University School of Medicine in St. Louis, St. Louis, MO 63108, USA
[10]Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[11]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[12]These authors contributed equally to this work
*Correspondence: jinghui.zhang@stjude.org (J.Z.), michael.dyer@stjude.org (M.A.D.)
http://dx.doi.org/10.1016/j.celrep.2014.03.003

SUMMARY

Pediatric osteosarcoma is characterized by multiple somatic chromosomal lesions, including structural variations (SVs) and copy number alterations (CNAs). To define the landscape of somatic mutations in pediatric osteosarcoma, we performed whole-genome sequencing of DNA from 20 osteosarcoma tumor samples and matched normal tissue in a discovery cohort, as well as 14 samples in a validation cohort. Single-nucleotide variations (SNVs) exhibited a pattern of localized hypermutation called kataegis in 50% of the tumors. We identified p53 pathway lesions in all tumors in the discovery cohort, nine of which were translocations in the first intron of the *TP53* gene. Beyond *TP53*, the *RB1*, *ATRX*, and *DLG2* genes showed recurrent somatic alterations in 29%–53% of the tumors. These data highlight the power of whole-genome sequencing for identifying recurrent somatic alterations in cancer genomes that may be missed using other methods.

INTRODUCTION

Osteosarcoma is the most common malignant bone tumor in children and adolescents, with approximately 400 new cases each year in the United States (Ottaviani and Jaffe, 2009). Although most cases are sporadic, the risk of osteosarcoma is increased in patients with various genetic diseases, including hereditary retinoblastoma, Li Fraumeni syndrome, and germline mutations of *RecQL4* (Hicks et al., 2007; Kleinerman et al., 2005; McIntyre et al., 1994). Current multimodal therapies that incorporate surgical excision and combination chemotherapy (i.e., doxorubicin, methotrexate, and cisplatin) cure approximately 70% of patients (Meyers et al., 2005). However, clinical outcomes and therapeutic strategies have remained virtually unchanged over the past 20 years (Smith et al., 2010).

In this study, we characterized the genomic landscape of osteosarcoma by performing whole-genome sequencing (WGS) on 34 osteosarcoma tumor and matched nontumor tissue samples from 32 patients. Our results demonstrate that pediatric osteosarcomas have one of the highest rates of SVs of any pediatric cancer sequenced to date (Downing et al., 2012), but relatively few recurrent single-nucleotide variations (SNVs). However, when SVs and SNVs were combined, inactivating mutations were identified in several cancer pathways. Taken together, our results provide insights into the molecular pathology of pediatric osteosarcoma and demonstrate that comprehensive WGS is required to elucidate the complete genetic landscape of osteosarcoma.

RESULTS

WGS of Primary and Metastatic Osteosarcomas

Using a paired-end sequencing approach, we generated 10,265 Gb of sequence data for DNA in 20 osteosarcomas and matched

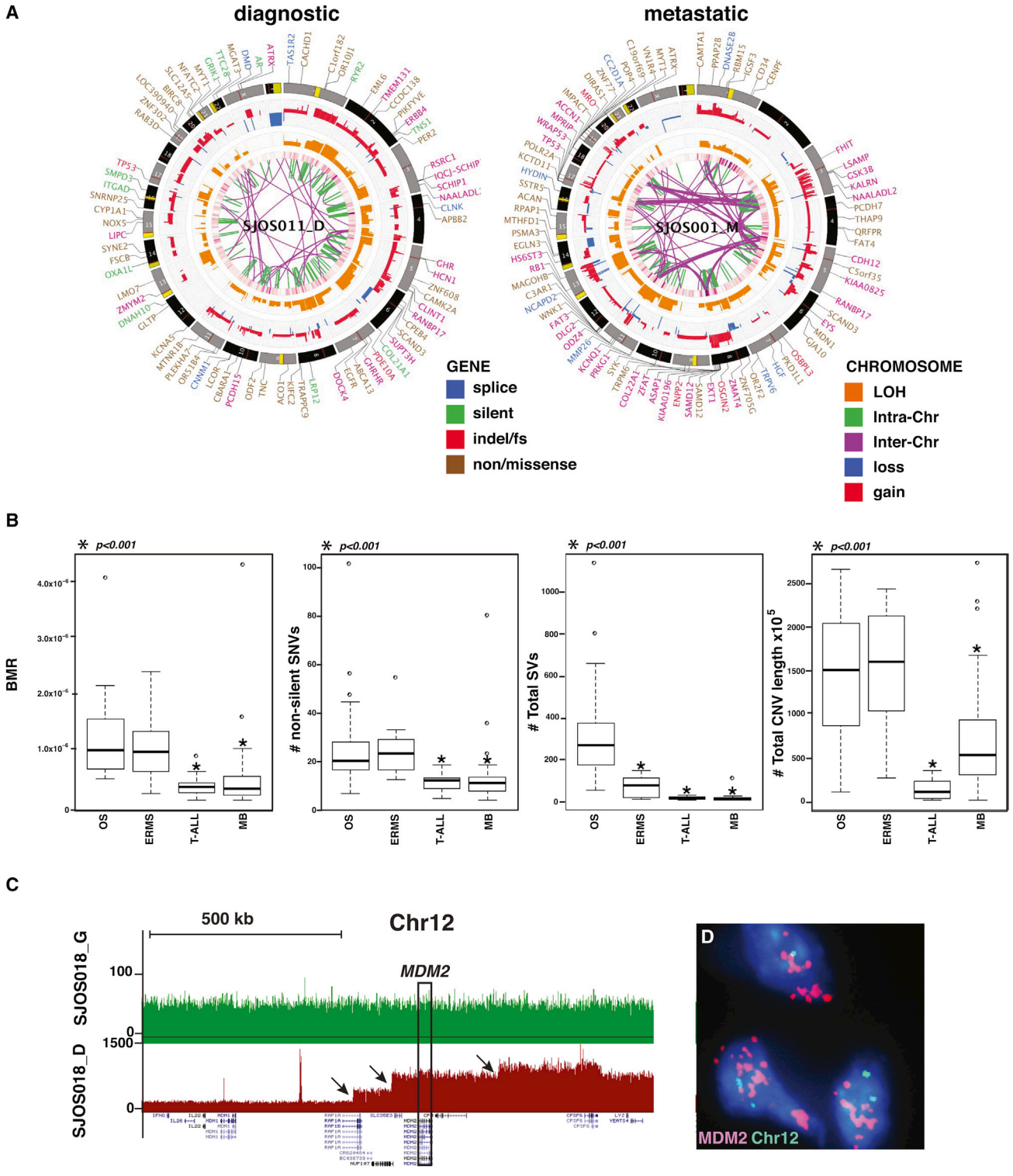


Figure 1. WGS of Osteosarcoma

(A) Representative CIRCOS plots of validated mutations and chromosomal lesions in diagnostic and metastatic osteosarcoma tumors from different patients. LOH (orange), gain (red), and loss (blue) are shown. Intrachromosomal (green lines) and interchromosomal (purple lines) translocations are indicated. Sequence mutations in RefSeq genes included silent SNVs (green), nonsense and missense SNVs (brown), splice-site mutations (dark blue), and insertion/deletion mutations (red). An additional track was added to the innermost ring of the plot showing the density of SNVs to highlight regions adjacent to SVs characteristic of kataegis.

(legend continued on next page)

normal DNA from 19 osteosarcoma patients in a discovery cohort, and 14 tumor specimens and matched normal DNA from 13 patients in a validation cohort (Table S1); 9,671 Gb (94%) were successfully mapped to the reference genome (Table S2). In the discovery cohort, the samples included 17 pre-treatment diagnostic samples (16 primary and one metastatic), one recurrent metastatic sample (SJOS001), and two tumor specimens (SJOS010_D and SJOS010_M) from the same patient with metachronous osteosarcoma (Table S1).

The average genome coverage was 44× and the average exon coverage was 39×; 99% of SNPs detected across all 39 genomes showed concordance with their corresponding SNP array genotype calls (Table S2). Validation was carried out using custom liquid capture for all SNVs, SVs, and insertions or deletions (indels) identified in the original sequence data. Combining the discovery and validation cohorts, we identified 50,426 validated somatic sequence mutations and 10,806 SVs (Table S3). These included 856 nonsilent tier 1 mutations in genes, 4,651 tier 2 mutations in evolutionarily conserved regions of the genome, and 43,782 tier 3 mutations in nonrepetitive regions of the genome that are not part of tier 1 or tier 2 (Table S3). The average number of sequence mutations was 1,483.1 per case (range 610–5,178), with 25.2 mutations per case (range 5–103) resulting in amino acid changes (Table S3). The estimated mean mutation rate was 1.15×10^{-6} per base (range 4.90×10^{-7}–3.99×10^{-6}). Among the validated SVs, 377 were predicted to produce an in-frame fusion protein (Table S3). Good-quality RNA sequencing (RNA-seq) data were available for five tumors with 64 predicted fusion SVs. Among them, 15 SVs (23%) were expressed (Table S3).

Primary and metastatic osteosarcomas had high rates of validated SVs (Figures 1A and S1). The number of SVs and CNVs, background mutation rate, and number of nonsilent tier 1 mutations were significantly higher in osteosarcoma compared with medulloblastoma and T-ALL (Robinson et al., 2012; Zhang et al., 2012; Figure 1B). However, only the number of SVs was significantly higher in osteosarcoma compared with another pediatric solid tumor with high rates of somatic alterations (embryonal rhabdomyosarcoma) (Chen et al., 2013; Figure 1B). The global patterns revealed by the WGS analysis of osteosarcoma suggest that the majority of SVs and CNVs were generated by sequential accumulation of SVs (Figures 1C and 1D), but chromothripsis (Stephens et al., 2011) was detected at specific genomic regions in four samples (chr14 in SJOS002_D, chr17q in SJOS003_D, chr6q in SJOS005_D, and chr13 in SJOS010_M; Supplemental Experimental Procedures). We used a modified version of GISTIC analysis to identify regions of the osteosarcoma genome with recurrent copy number alterations in the discovery cohort. The *TP53*, *RB1*, *MYC*, and *PTEN* pathways, as well as *ATRX*, *LSAMP-AS3*, *CCNE1*, and a genomic region on chromosome 16 containing *COPS3*, *PMP22*, *MAPK7*, *NCOR1*, and *UBB*, were recurrently mutated (Figure S1C). Among SNVs with sufficient coverage in both SJOS010 samples (20×), we validated 673 SNVs in both samples, 1,686 in diagnostic-only samples, and 1,408 in metastasis-only samples, indicating that these two tumors shared a limited amount of common mutations and were divergent early in the progression.

Applying the GRIN method (Pounds et al., 2013) on functional mutations (including SNVs and indels) and SVs, we identified *TP53* (false discovery rate [FDR] = 3.6E-51, mutated in 28/34 samples) *RB1* (FDR = 1.1E-5, mutated in 10/34 samples), *ATRX* (FDR = 2.4E-4, mutated in 10/34 samples), and *DLG2* (FDR = 0.044, mutated in 18/34 samples) as significantly mutated genes. All genes except *DLG2* were mutated by point mutations (nine for *TP53*, three for *RB1*, and five for *ATRX*) and SVs in multiple tumors (18 for *TP53*, seven for *RB1*, and five for *ATRX*). *DLG2* was exclusively mutated by SVs.

Osteosarcoma Tumor Purity and Tumor Heterogeneity

Using the purity-adjusted mutant allele fraction (MAF) derived from deep sequencing of all SNVs by capture enrichment and Illumina sequencing, we analyzed intratumor heterogeneity. Eleven tumors (SJOS001_M, SJOS004, SJOS005, SJOS008, SJOS012, SJOS013, SJOS015, SJOS001103_D1, SJOS001105_D1, and SJOS001123_D1, and SJOS001125_D1) were excluded from quantitative heterogeneity analysis due to an insufficient number of SNVs in copy-neutral regions. Statistical modeling demonstrated that 61% (14/23) of osteosarcomas in this group had evidence of multiple clones, including metastatic samples SJOS010_M, SJOS001107_M1 and SJOS001107_M2 (Figure S2).

Kataegis in Osteosarcoma

To determine whether there was any relationship between the SVs and location, distribution, or type of SNV in the osteosarcoma genomes, we plotted the validated SVs and SNVs for each sample and analyzed the intermutation distance (Figure S2). Hypermutable regions with the five hallmarks of kataegis (Nik-Zainal et al., 2012) were identified in 17 of the osteosarcoma tumors (Figure 2A). These five hallmarks of kataegis are (1) enriched C->T and C->G substitutions at TpCpX trinucleotides (Figures 2B and 2C), (2) the same class of nucleotide mutation occurring for contiguous stretches before switching to a different class (Figure 2D), (3) mutations within short stretches of the genome occurring on the same parental chromosome (Figure S2), (4) clustering of heavily mutated short stretches of the genome at multiple scales (Figure 2E), and (5) association of the hypermutated region with SV breakpoints (Figure 2E). The regions of the genome with kataegis were not recurrent in our cohort and were not associated with recurrently mutated genes

(B) Boxplots of validated basal mutation rate (BMR) and numbers of nonsilent SNVs, total SVs, and total CNVs in embryonal rhabdomyosarcoma (ERMS) and alveolar rhabdomyosarcoma (ARMS) tumors in the discovery cohort. * represents statistical significance of $p < 0.001$ as compared with the osteosarcoma genomes.

(C) Representative plot of sequence reads on chromosome 12 for the matched germline (green) and tumor (red) sample. Several distinct regions of copy number change are identified (arrows) spanning the *MDM2* gene consistent with sequential chromosomal lesions.

(D) MDM2 FISH of SJOSO18 showing amplification (red) relative to the probe for chromosome 12 (green). T-ALL, T cell acute lymphoblastic leukemia; MB, medulloblastoma.

See also Figures S1–S3 and Tables S1–S3.

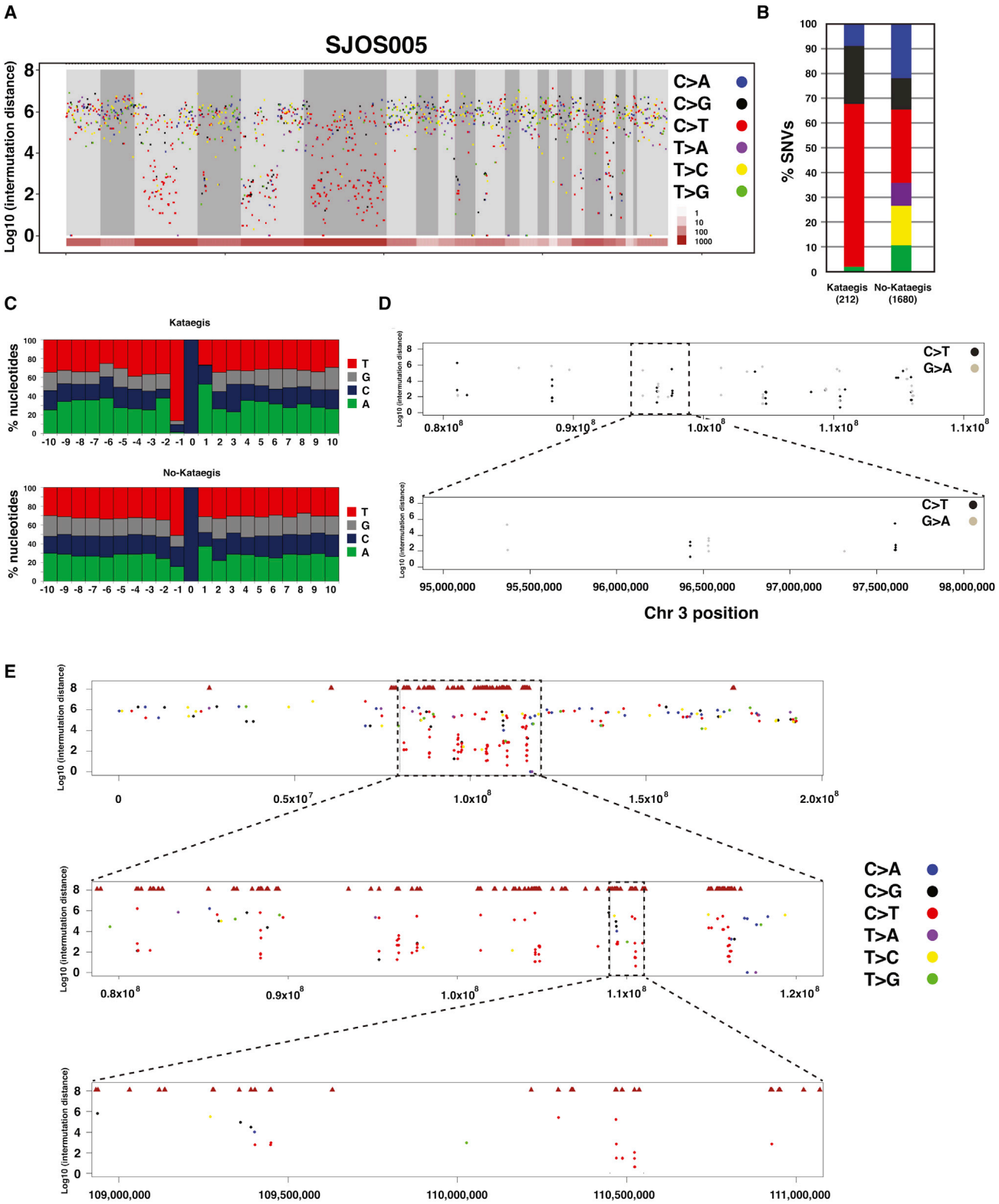


Figure 2. Kataegis in Osteosarcoma

(A) Rainfall plot showing the Log_{10} of the intermutation distance versus genomic position for a representative osteosarcoma sample (SJOS005) with evidence of kataegis. The chromosomes are demarcated by gray shading and the number of SVs in each chromosome is shown in brown at the bottom. The validated SNVs are plotted and color-coded by the type of mutation.

(legend continued on next page)

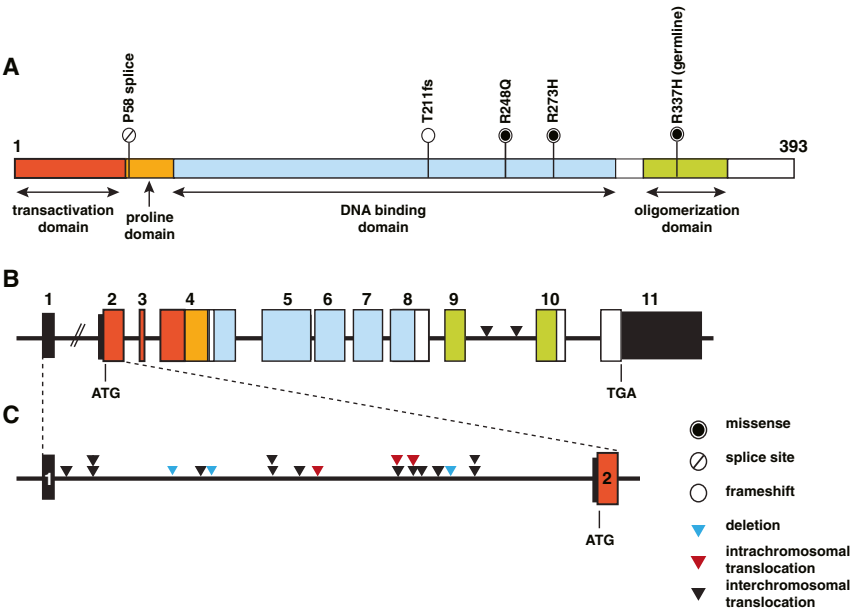


Figure 3. Validated Mutations in *TP53*

(A) Structure of the *TP53* gene showing the transactivation, proline, DNA binding, and oligomerization domains with splice-site, frameshift, and missense mutations in tumors of the 19 patients in the discovery cohort.

(B) Structure of the genomic locus of the *TP53* gene showing the exon boundaries color-coded in accordance with the protein domains shown in (A). Sites of interchromosomal translocations are indicated by black arrowheads between exons 9 and 10. The sizes of the introns and exons are scaled proportionally except for intron 1, which is much larger than the other introns in human *TP53*.

(C) A magnified view of intron 1 of *TP53* showing the deletions (blue arrowheads), intrachromosomal translocations (red arrowheads), and interchromosomal translocations (black arrowheads).

See also Figure S3 and Table S4.

in osteosarcoma (Figure S2). Tier 1 SNVs in kataegis regions were not significantly associated with the expression status ($p = 0.16$ by Fisher's exact test).

Chronology of Kataegis, SVs, and Aneuploidy in SJOS005

SJOS005 had the highest proportion (11%) of kataegis SNVs in our cohort. The large number of kataegis SNVs (n = 212) coupled with the accurate measurement of the MAFs of all SNVs derived from deep sequencing allowed us to analyze the chronology of kataegis in relation to other mutational events in this tumor. First, we examined MAFs of SNVs in kataegis microclusters containing five or more consecutive kataegis SNVs within 10 kb. The MAF variance was relatively small (6.7% of overall variance) within a microcluster, although there was a wide range of MAFs across microclusters (range 0.142–0.839, median 0.364; Figure S2). This pattern, along with the observation that SNVs in a microcluster occurred on the same parental chromosome, supports the hypothesis that SNVs in a kataegis microcluster originated from a single event. MAF analysis of SVs flanking "kataegis" clusters (range 0.132–0.866, median 0.396) also showed a significant positive correlation ($p = 4.56E-5$) with those of "kataegis" SNVs, and there was no significant difference between them ($p = 0.143$ by Wilcoxon signed rank test), indicating that the neighboring SVs likely arise simultaneously with kataegis SNVs (Figure S2).

SJOS005, like most osteosarcomas, exhibits aneuploidy with copy number gains spanning more than 50% of the genome. Kataegis SNVs were significantly enriched in regions of the genome with four or more copies compared with nonkataegis SNVs ($p < 2.2e-16$ by Fisher's exact test; Figure S2). However, the MAF distribution of kataegis SNVs showed a large fraction of SNVs with multiple copies of the mutant allele in amplified regions, whereas only a single copy of mutant alleles was found in the majority of the nonkataegis SNVs (Figure S2). Taken together, these data suggest that kataegis likely occurs before global aneuploidy, and nonkataegis SNVs occur primarily after the aneuploidy.

SVs in *TP53*

The p53 pathway was mutated in all 20 tumor samples from the 19 patients in our discovery cohort. The majority (95%, 19/20) had either sequence mutations or SVs in the *TP53* gene, and one (SJOS018) had an *MDM2* amplification (see Figures 1C and 1D; Table S3). Surprisingly, 55% of the tumors (11/20) had SVs in the *TP53* gene, and the majority of those were translocations with breakpoints that were confined to the first intron of the gene (90%, 19/21 SV breakpoints; Figures 3A–3C; Table S4). Indeed, some tumors had rearrangements in both alleles of *TP53*, resulting from two or more independent translocations (Table S4). One patient's tumor (SJOS006) had a germline SNV (R337H), one (SJOS012) had a somatic splice-site mutation, and two (SJOS004 and SJOS010) had somatic missense SNVs (Figures 3A–3C; Table S4). The remaining four patients had tumors that harbored indels in the *TP53* gene. Loss of heterozygosity (LOH) at the *TP53* locus

(B) The proportion of each type of validated SNV in osteosarcomas with evidence of kataegis versus those without kataegis.

(C) The distribution of each nucleotide sequence 5′ to the C mutations in tumors with kataegis and those without kataegis.

(D) A rainfall plot in a representative regions of chromosome 3 in SJOS005 with kataegis showing the strand of the hypermutation based on the C>T or G>A sequence clusters.

(E) A macrocluster of hypermutation with evidence of kataegis on chromosome 3 of SJOS005, with two sequential magnifications (boxes) showing the existence of microclusters within a single macrocluster.

See also Figure S2.

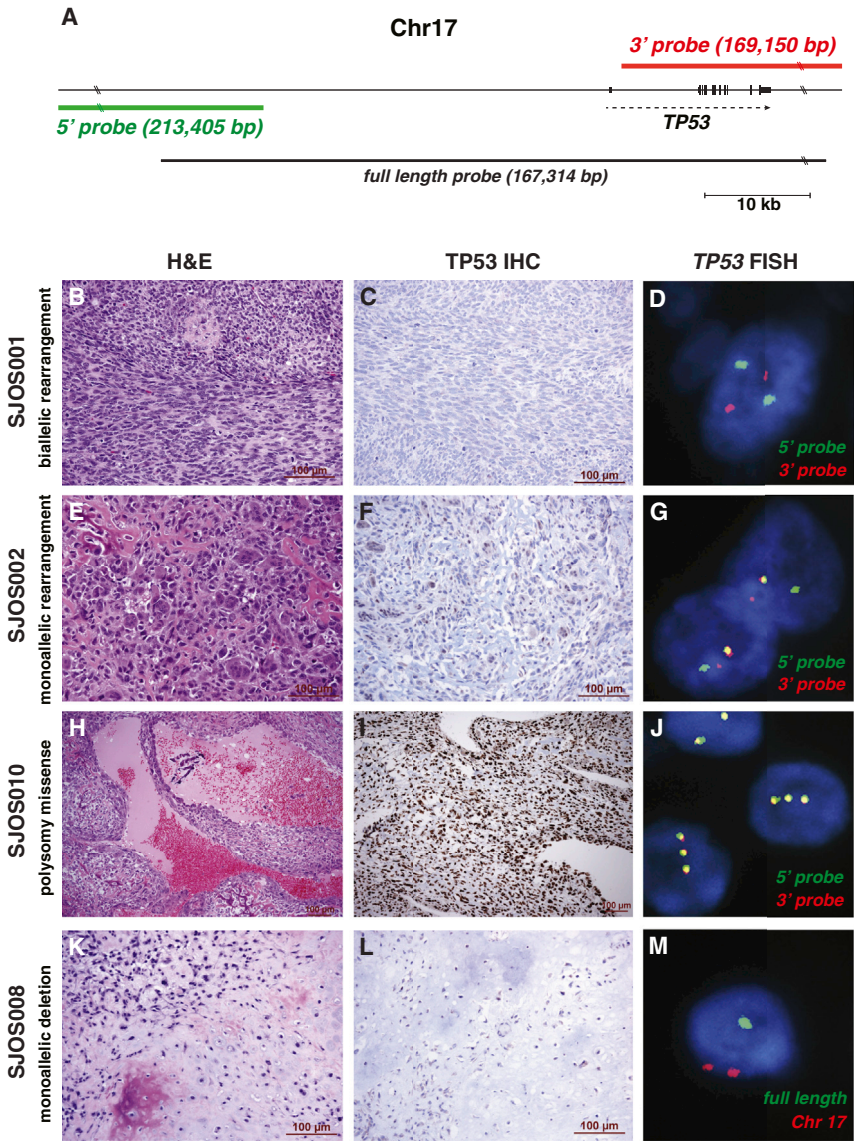


Figure 4. FISH and Immunohistochemistry for TP53 in Osteosarcoma

(A) Genomic location of the 5′ (green) and 3′ (red) break-apart FISH probes showing their position relative to the *TP53* gene. The full-length probe used to identify deletions at the *TP53* locus is shown in black.

(B–D) Images of hematoxylin and eosin (H&E), TP53 immunohistochemistry (IHC), and break-apart FISH for SJOS001 with biallelic rearrangement of the *TP53* gene.

(E–G) Images of H&E, TP53 IHC, and break-apart FISH for SJOS002 with monoallelic rearrangement of the *TP53* gene.

(H–J) Images of H&E, TP53 IHC, and break-apart FISH for SJOS010 with polysomy and a missense mutation leading to elevated accumulation of nuclear TP53 protein.

(K–M) Images of H&E, TP53 IHC, and FISH using the full-length probe (green) for SJOS008 with monoallelic deletion of *TP53*.

See also Table S5.

was evident in 40% (8/20) of the osteosarcoma tumors. In total, 15 tumors had biallelic inactivation of *TP53*, four had monoallelic inactivation of *TP53*, and one had *MDM2* amplification (Figure 1C; Table S4).

To further validate the translocations in the *TP53* gene identified by WGS, we developed a break-apart fluorescence in situ hybridization (FISH) assay with separate probes spanning the 5′ and 3′ regions of the gene (Figure 4A). We also developed a FISH assay with a probe spanning the entire *TP53* gene (Figure 4A) to assess ploidy and determine whether the gene was deleted. To complement the FISH analysis, we performed p53 immunostaining to verify that the tumors with missense mutations had accumulated high levels of nuclear p53 protein. We successfully performed FISH in 18 of 20 tumors and p53 immunostaining on all 20 tumors (Table S4). Overall, there was perfect concordance between the WGS data and the FISH data (Figures 4B–4M; Table S4).

In an additional cohort of patient tumor samples, we found that 50% (16/32) had *TP53* rearrangements, 22% (7/32) had missense mutations, 16% (5/32) had nonsense mutations, 6% (2/32) had a *TP53* deletion, and 3% (1/32) had an *MDM2* amplification (Table S5). Three patients with tumor showed no evidence of a p53 pathway mutation.

We did not find any significant difference in CNV ($p = 0.20$ by Wilcoxon rank sum test), SV ($p = 0.85$), SNV ($p = 0.43$), nonsilent tier 1 mutations ($p = 0.66$), or background mutation rate ($p = 0.43$) in the osteosarcoma samples with mutant p53 versus those with inactivating (nonsense, deletion and truncation) mutations in *TP53*. Survival analysis, including event-free survival and overall survival, did not show a significant difference in outcome for the patients whose tumors carried *TP53*-missense mutations (ten patients) versus those with *TP53*-truncating mutations (34 patients), with log rank test p values of 0.88 and 0.64, respectively.

RB1, *ATRX*, and *DLG2* Are Recurrently Mutated in Osteosarcoma

ATRX is part of a multiprotein complex that regulates chromatin remodeling, nucleosome assembly, and telomere maintenance. It was recently shown that *ATRX* mutations in neuroblastoma are associated with age at diagnosis (Cheung et al., 2012). Most neuroblastomas with *ATRX* mutations show evidence of alternative lengthening of telomeres (ALT), as measured by WGS, telomere FISH, and telomere quantitative PCR (qPCR) (Cheung et al., 2012). In our osteosarcoma discovery cohort, we identified five tumors (SJOS001, SJOS002, SJOS007, SJOS001112-M2, and SJOS001117-D1) with point mutations in *ATRX*, and five

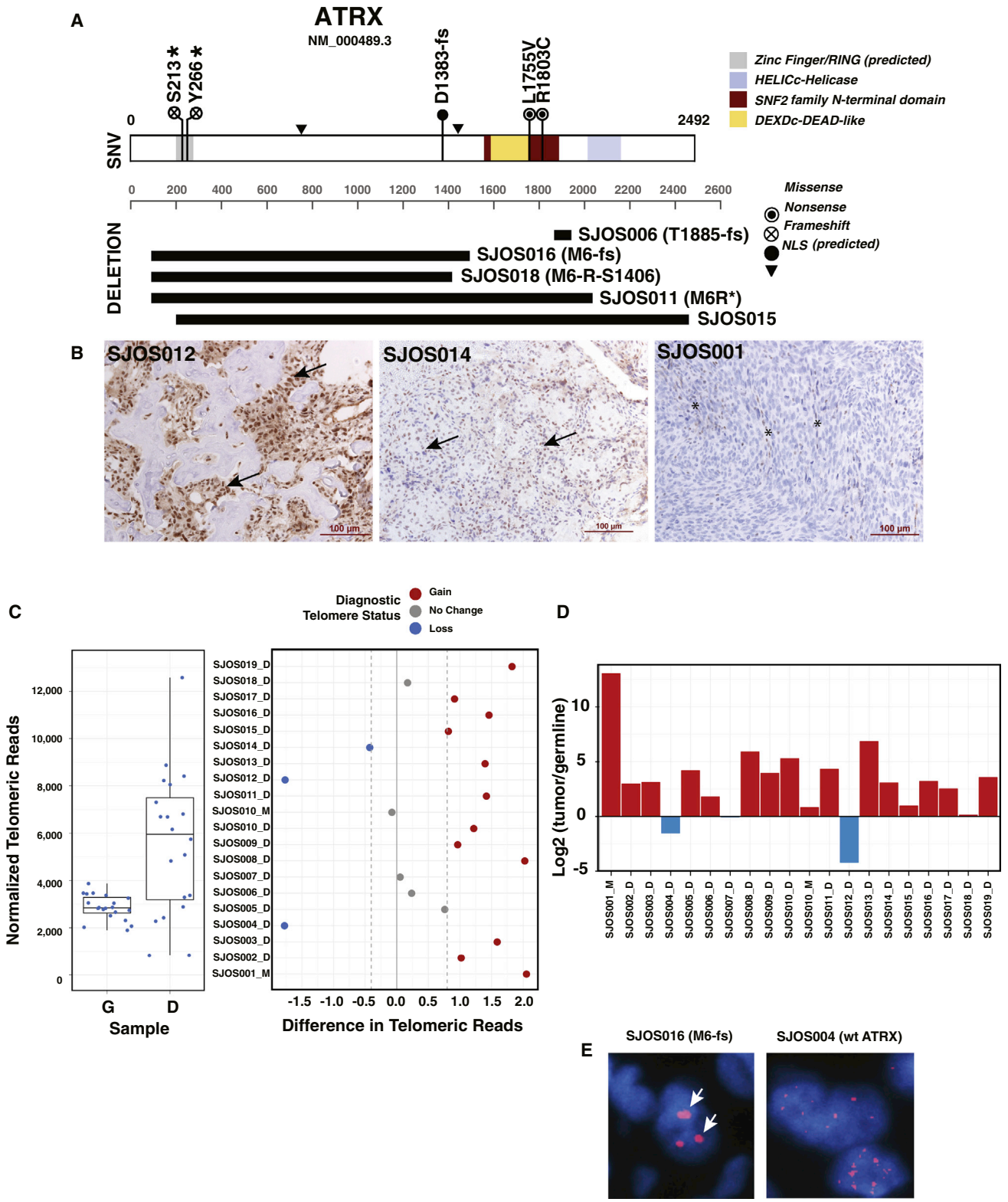


Figure 5. *ATRX* Mutations Correlate with ALT in Osteosarcoma

(A) Diagram of the five SNVs, four deletions, and one interchromosomal SV found in the *ATRX* genes of the osteosarcoma cohort. Three of the samples with ATRX SVs (SJOS006, SJOS018, and SJOS011) had matching RNA-seq data. SJ006 has a short deletion at exon 23 and the RNA-seq data confirmed a readthrough

(legend continued on next page)

with focal deletions or SVs affecting the coding region of the gene (Figure 5A; Table S6). There was no significant gender bias in *ATRX* mutations (p = 0.25 by Fisher's exact test) even though it is located on the X chromosome. By immunohistochemistry, 31% (6/19) of the tumors in the discovery cohort were ATRX negative (Figure 5B; Table S6). The sample with a missense mutation (SJOS007-R1803C) and one with an SV (SJOS018) were heterogeneous for ATRX protein expression. Analysis of telomere sequence reads from the WGS data and qPCR of telomeres showed that the majority of osteosarcomas had longer telomeres (Figures 5C and 5D), and ALT was found in 85% (12/14) of the samples using telomere FISH (Table S6).

Beyond *TP53* and *ATRX*, there were significant recurrent mutations in *RB1* (10/34, FDR *q* = 1.1E-5) and *DLG2* (18/34, FDR *q* = 0.044). *DLG2* encodes a multi-PDZ domain protein that is involved in epithelial polarity during cell division and has been implicated in cancer cell invasion. In *Drosophila*, DLG is a tumor suppressor, but a clear tumor-suppressor function has not yet been confirmed for *DLG2* in human cancer.

SVs in Cancer Genes

SVs contributed 91% (9,605/10,523) of all functional genetic lesions in our osteosarcoma cohort. In total, 122 cancer genes had at least one SV breakpoint (Table S7) and all but one tumor (SJOS001118_D1) had at least one breakpoint (range 1–40) in a cancer gene. SV breakpoint enrichment in the cancer genes was highly significant even when we excluded *TP53* from the list (p = 2.5E-6). Twelve of the 34 tumors (35%) achieved significant enrichment of SV breakpoints in cancer genes individually. In addition, some tumors have "fold-back intrachromosomal translocations" (Campbell et al., 2010) to inactivate tumor-suppressor genes (Figure S3). These results further support the hypothesis that genomic instability leads to lesions in various cancer genes.

DISCUSSION

WGS of osteosarcoma demonstrated that the rate of SNVs was similar to that in other pediatric solid tumors, and only a few recurrent SNVs were detected. Approximately half of the osteosarcomas in our discovery cohort had a pattern of hypermutation associated with SVs, called kataegis (Nik-Zainal et al., 2012). The regions of the genome with kataegis were not recurrent, and none of the most recurrently mutated genes were found in regions of kataegis. Chromosomal lesions, rather than SNVs, were the major mechanism of recurrent mutations, and many of the most significant chromosomal lesions were found in known cancer genes, including *TP53*, *RB1*, and *ATRX.*

Genomic Stability and Osteosarcoma Initiation and Progression

The most frequent mutation in osteosarcoma is in *TP53*. By our estimates, both alleles are mutated in as many as 80% of tumors, and at least one allele was mutated in >90% of tumors. These data suggest that p53 mutations are a major oncogenic driver in osteosarcoma. Although this finding is not novel, what is surprising is the mechanism of inactivation. Most *TP53* mutations are SVs in intron 1, which suggests that either the *TP53* locus is particularly susceptible to SVs or SVs occur at a high rate in the osteosarcoma tumor-initiating cell. Aside from osteosarcomas and prostate cancers (Baca et al., 2013; Berger et al., 2011), there is no evidence of *TP53* SVs in any other cancer, so the locus is probably not uniquely susceptible to chromosomal rearrangements. These data raise an intriguing possibility: genomic instability characterized by high rates of CNVs and SVs may precede *TP53* inactivation, and may be the underlying mechanism that initiates and promotes osteosarcoma.

Kataegis in Osteosarcoma

In a recent WGS study, Nik-Zainal et al. (2012) described a distinct hypermutation phenomenon in breast cancer that they termed kataegis. Here, we found SNV clusters with the same five characteristics of kataegis in 50% of the osteosarcomas analyzed by WGS. Interestingly, genomic regions encoding *TP53* and *ATRX*, the two most frequently mutated genes in osteosarcoma, did not exhibit this pattern of local hypermutation. Furthermore, there was no association between kataegis and *TP53* mutation type (i.e., SNV, indel, or SV).

TP53-Mutant or -Null Osteosarcomas

Previous studies have estimated that 20%–70% of osteosarcomas carry mutations in the p53 pathway (Lonardo et al., 1997; Wunder et al., 2005), but our data suggest that the proportion is much higher. For example, Wunder et al. (2005) sequenced exons 4–10 of the *TP53* gene in 196 osteosarcoma samples and found that 19.4% (38/196) had *TP53* SNVs. The investigators concluded that the remaining 80.6% (158/196) had wild-type *TP53* (Wunder et al., 2005). They went on to show that event-free survival was indistinguishable between the two groups (wild-type and mutant *TP53*) (Wunder et al., 2005). SVs in the first intron of *TP53* were not analyzed in that study, even though such lesions had previously been reported in osteosarcoma (Miller et al., 1990). Our data suggest that the majority of the tumors identified as *TP53* wild-type in the study

event that would result in a T1885 frameshift. For SJOS011, the RNA-seq and WGS data supported a junction connecting exon 1 to exon 28, creating a nonsense mutation (M6R*). For SJOS018, the RNA-seq and WGS data supported a deletion connecting exon 1 to exon 13, thereby creating an in-frame fusion protein (M6RS1406). The WGS for SJOS016 predicts a deletion that connects exon 1 to exon 16, producing a frameshift (M6fs).

(B) Representative IHC for ATRX showing nuclear ATRX in a sample with intense staining and wild-type ATRX (SJOS012), a sample with fainter nuclear localized ATRX (SJOS014), and a sample with a nonsense mutation (SJOS001). Arrows indicate representative nuclei stained positive for ATRX. Asterisks indicate ATRX immunopositive vascular endothelial cells among the tumor cells that are negative for ATRX IHC.

(C and D) Relative telomere length in the osteosarcomas compared with that in the matched germline DNA, as analyzed by WGS and qPCR.

(E) Representative telomere FISH showing characteristics of ALT (arrow) in an osteosarcoma with an ATRX deletion.

See also Tables S6 and S7.

by Wunder et al. (2005) actually had inactivating SVs in *TP53*. Therefore, it may be useful to revisit the association of *TP53* pathway inactivation with osteosarcoma outcome in a large cohort of patient samples.

EXPERIMENTAL PROCEDURES

Full details regarding sample acquisition, molecular and biochemical procedures, informatics, and WGS are provided in the Supplemental Information. All tumors in this study were obtained from St. Jude Children's Research Hospital (SJCRH) patients. The SJCRH IRB approved experiments involving human subjects and informed consent was obtained from all subjects.

ACCESSION NUMBERS

The European Bioinformatics Institute accession number for the sequencing data reported in this paper is EGAS00001000263.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.03.003.

ACKNOWLEDGMENTS

This work was supported, in part, by Cancer Center Support (CA21765) from the NCI, grants to M.A.D from the NIH (EY014867, EY018599, and CA168875), and the American Lebanese Syrian Associated Charities (ALSAC). M.A.D. is an HHMI Investigator. The whole-genome sequencing was supported as part of the St. Jude Children's Research Hospital -Washington University Pediatric Cancer Genome Project.

Received: May 10, 2013
Revised: November 22, 2013
Accepted: March 3, 2014
Published: April 3, 2014

REFERENCES

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666–677.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214–220.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109–1113.

Chen, X., Steward, E., Shelat, A., Qu, C., Bahrami, A., Hatley, M., Wu, G., Bradley, C., McEvoy, J., Pappo, A., et al.; St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project (2013). Targeting oxidative stress in embryonal rhabdomyosarcoma. Cancer Cell *24*, 710–724.

Cheung, N.K., Zhang, J., Lu, C., Parker, M., Bahrami, A., Tickoo, S.K., Heguy, A., Pappo, A.S., Federico, S., Dalton, J., et al. (2012). Association of age at diagnosis and genetic mutations in patients with neuroblastoma. JAMA *307*, 1062–1071.

Downing, J.R., Wilson, R.K., Zhang, J., Mardis, E.R., Pui, C.H., Ding, L., Ley, T.J., and Evans, W.E. (2012). The Pediatric Cancer Genome Project. Nat. Genet. *44*, 619–622.

Hicks, M.J., Roth, J.R., Kozinetz, C.A., and Wang, L.L. (2007). Clinicopathologic features of osteosarcoma in patients with Rothmund-Thomson syndrome. J. Clin. Oncol. *25*, 370–375.

Kleinerman, R.A., Tucker, M.A., Tarone, R.E., Abramson, D.H., Seddon, J.M., Stovall, M., Li, F.P., and Fraumeni, J.F., Jr. (2005).). Risk of new cancers after radiotherapy in long-term survivors of retinoblastoma: an extended follow-up. J. Clin. Oncol. *23*, 2272–2279.

Lonardo, F., Ueda, T., Huvos, A.G., Healey, J., and Ladanyi, M. (1997). p53 and MDM2 alterations in osteosarcomas: correlation with clinicopathologic features and proliferative rate. Cancer *79*, 1541–1547.

McIntyre, J.F., Smith-Sorensen, B., Friend, S.H., Kassell, J., Borresen, A.L., Yan, Y.X., Russo, C., Sato, J., Barbier, N., Miser, J., et al. (1994). Germline mutations of the p53 tumor suppressor gene in children with osteosarcoma. J. Clin. Oncol. *12*, 925–930.

Meyers, P.A., Schwartz, C.L., Krailo, M., Kleinerman, E.S., Betcher, D., Bernstein, M.L., Conrad, E., Ferguson, W., Gebhardt, M., Goorin, A.M., et al. (2005). Osteosarcoma: a randomized, prospective trial of the addition of ifosfamide and/or muramyl tripeptide to cisplatin, doxorubicin, and high-dose methotrexate. J. Clin. Oncol. *23*, 2004–2011.

Miller, C.W., Aslo, A., Tsay, C., Slamon, D., Ishizaki, K., Toguchida, J., Yamamuro, T., Lampkin, B., and Koeffler, H.P. (1990). Frequency and structure of p53 rearrangements in human osteosarcoma. Cancer Res. *50*, 7950–7954.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979–993.

Ottaviani, G., and Jaffe, N. (2009). The epidemiology of osteosarcoma. Cancer Treat. Res. *152*, 3–13.

Pounds, S., Cheng, C., Li, S., Liu, Z., Zhang, J., and Mullighan, C. (2013). A genomic random interval model for statistical analysis of genomic lesion data. Bioinformatics *29*, 2088–2095.

Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature *488*, 43–48.

Smith, M.A., Seibel, N.L., Altekruse, S.F., Ries, L.A., Melbert, D.L., O'Leary, M., Smith, F.O., and Reaman, G.H. (2010). Outcomes for children and adolescents with cancer: challenges for the twenty-first century. J. Clin. Oncol. *28*, 2625–2634.

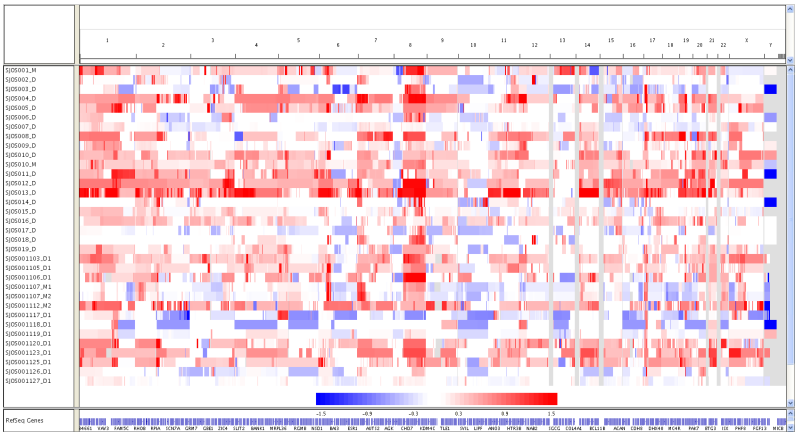
Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.

Wunder, J.S., Gokgoz, N., Parkes, R., Bull, S.B., Eskandarian, S., Davis, A.M., Beauchamp, C.P., Conrad, E.U., Grimer, R.J., Healey, J.H., et al. (2005). TP53 mutations and outcome in osteosarcoma: a prospective, multicenter study. J. Clin. Oncol. *23*, 1483–1490.

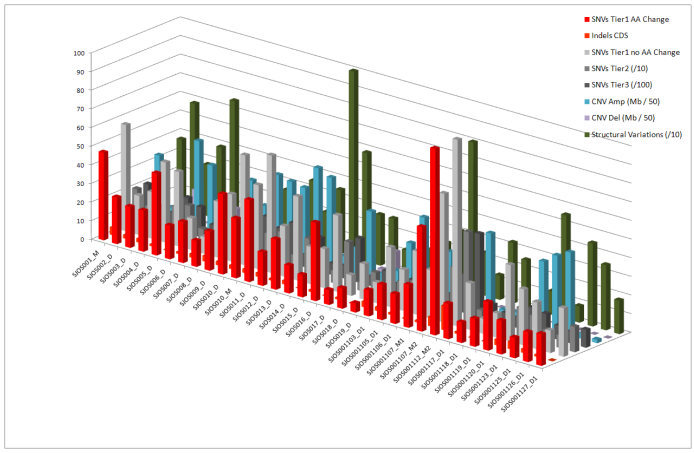
Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163.

SUPPLEMENTAL DATA

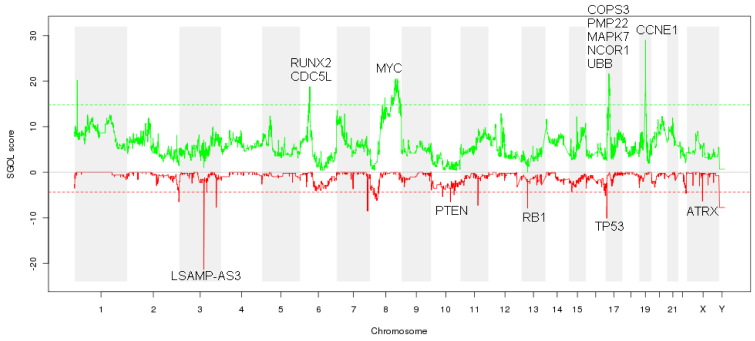
A



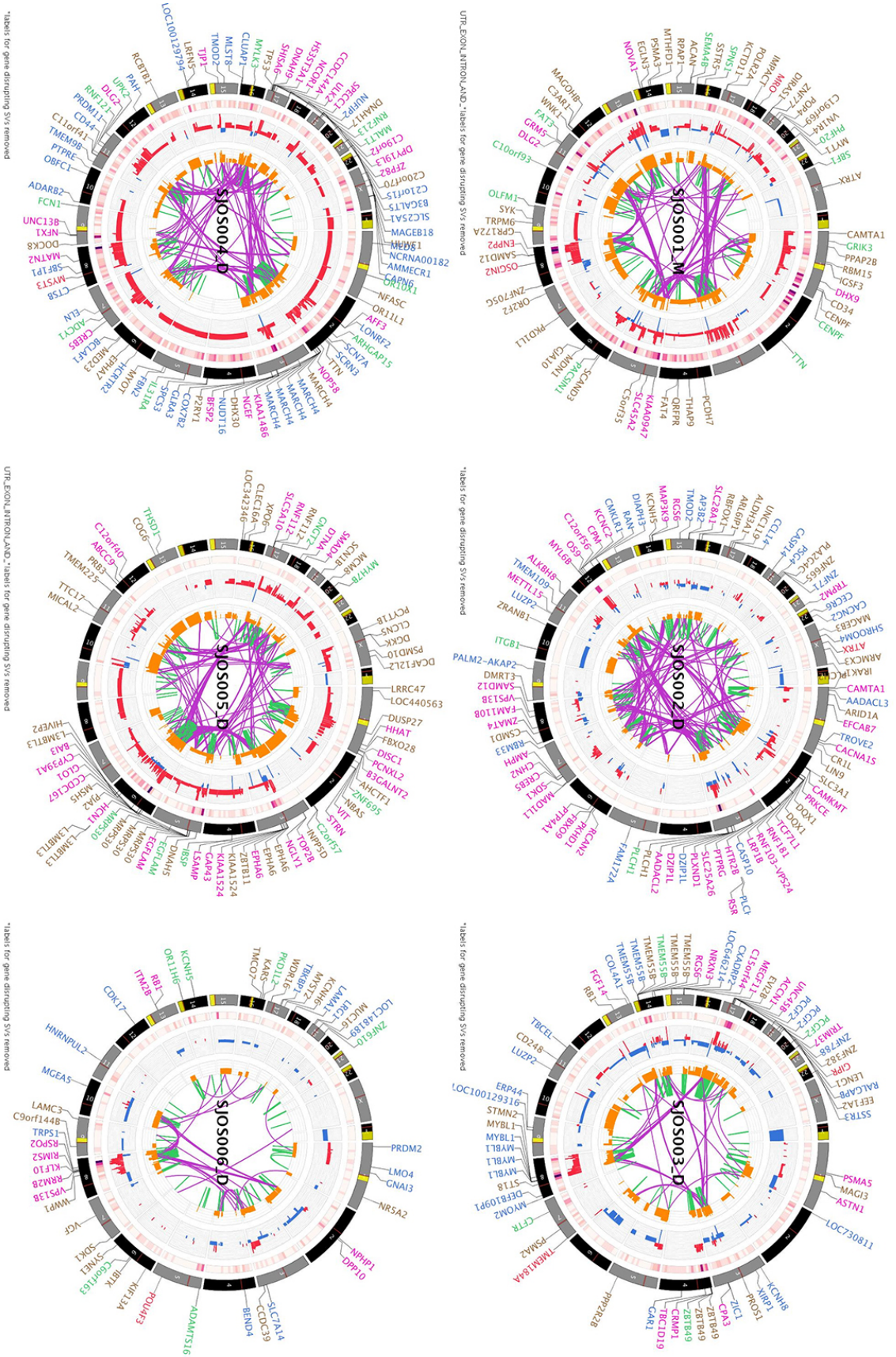
B

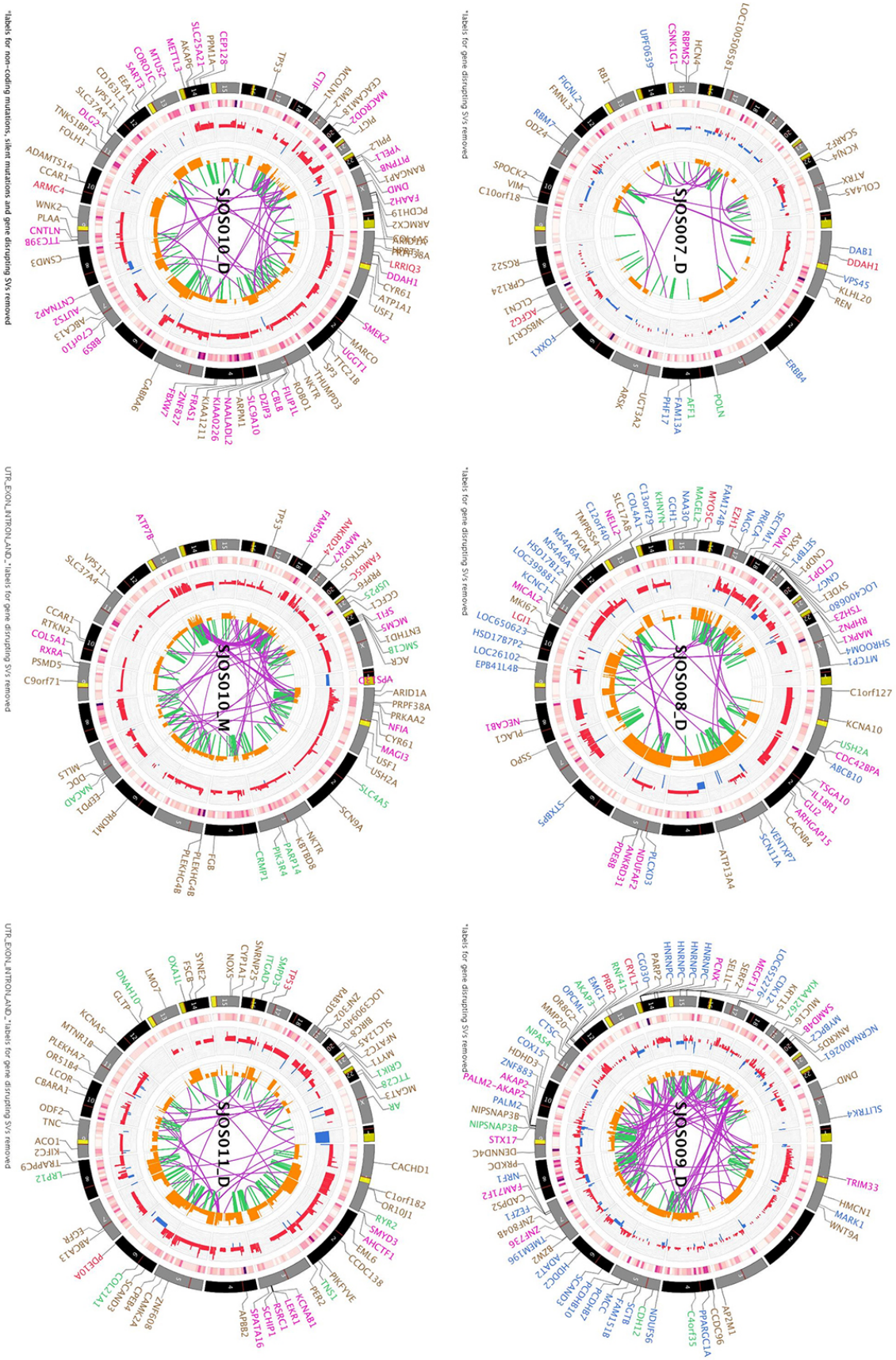


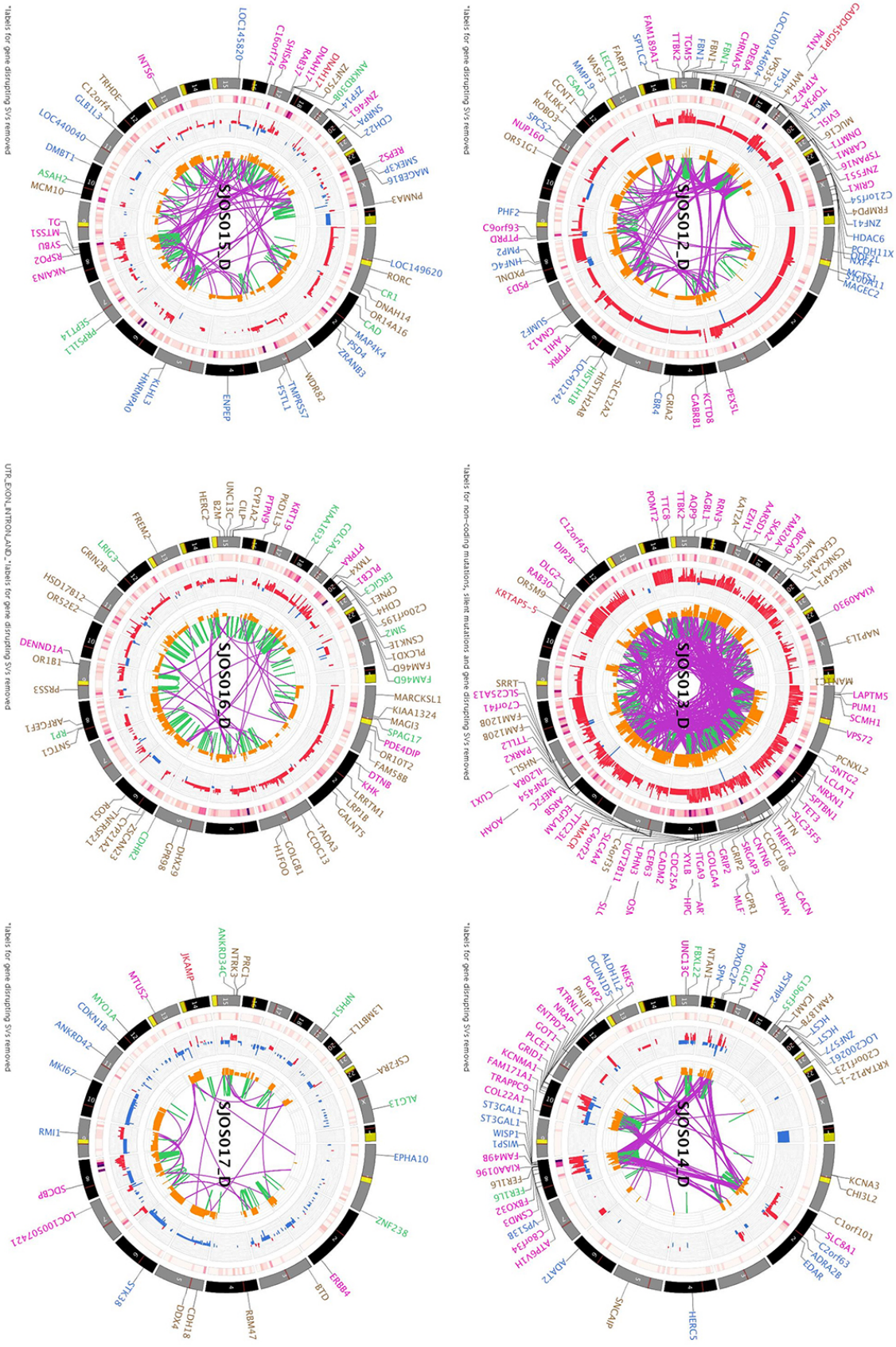
C

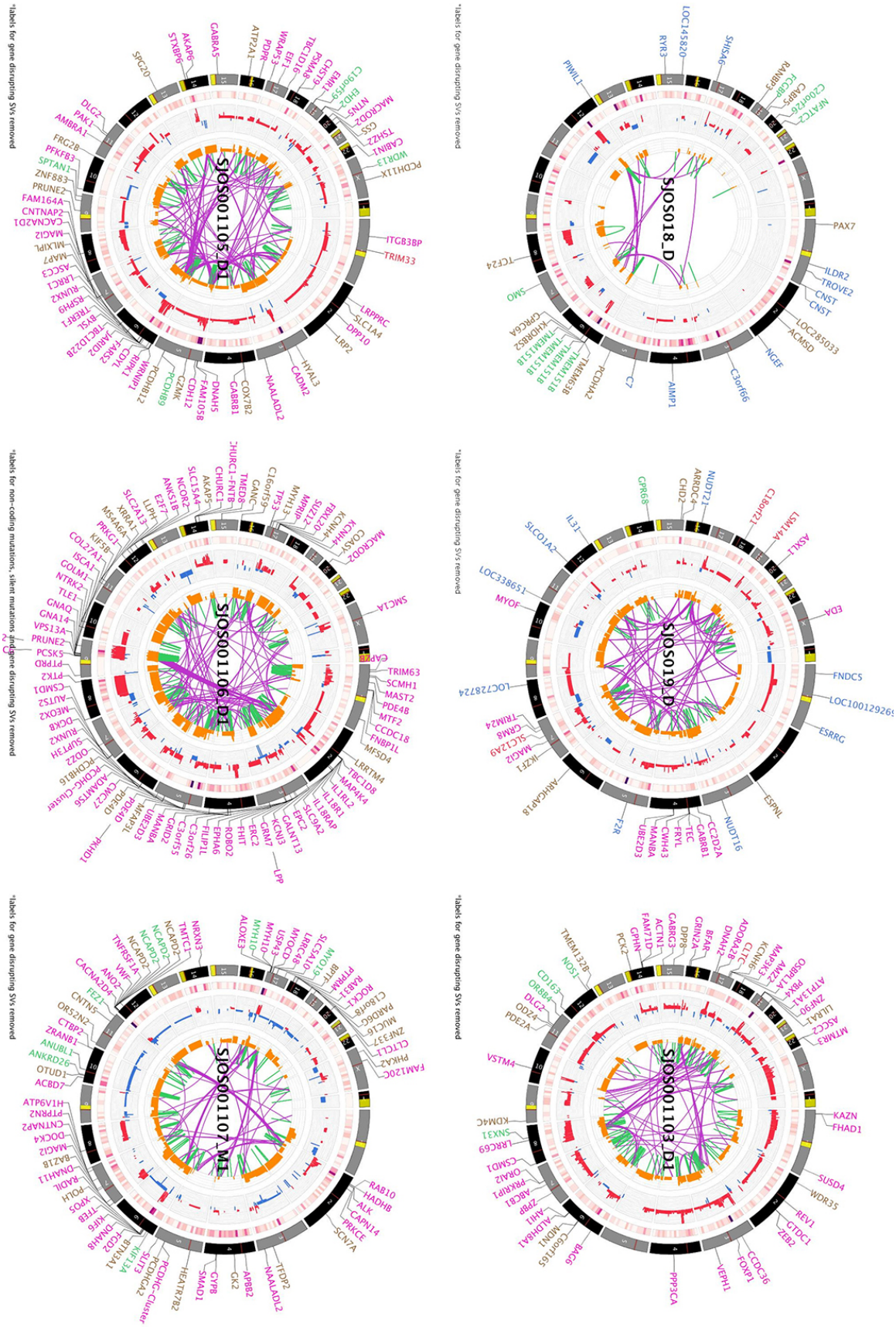


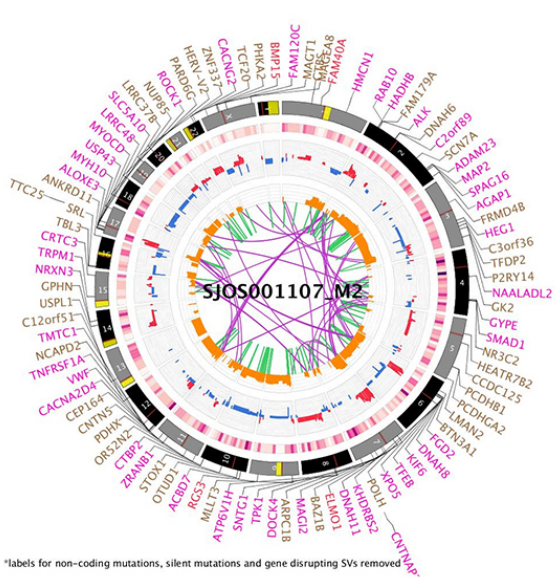
D



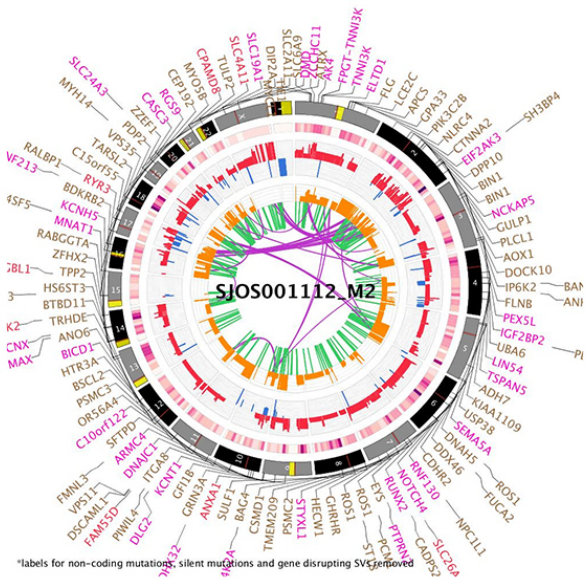

*labels for non-coding mutations, silent mutations and gene disrupting SVs removed
SJOS010_D
*labels for gene disrupting SVs removed
SJOS007_D
UTR_EXON_INTRON_LAND, *labels for gene disrupting SVs removed
SJOS010_M
*labels for gene disrupting SVs removed
SJOS008_D
UTR_EXON_INTRON_LAND, *labels for gene disrupting SVs removed
SJOS011_D
*labels for gene disrupting SVs removed
SJOS009_D


*labels for gene disrupting SVs removed
SJOS012_D
*labels for gene disrupting SVs removed
SJOS015_D
*labels for non-coding mutations, silent mutations and gene disrupting SVs removed
SJOS013_D
UTR_EXON_INTRON_LAND; *labels for gene disrupting SVs removed
SJOS016_D
*labels for gene disrupting SVs removed
SJOS014_D
*labels for gene disrupting SVs removed
SJOS017_D

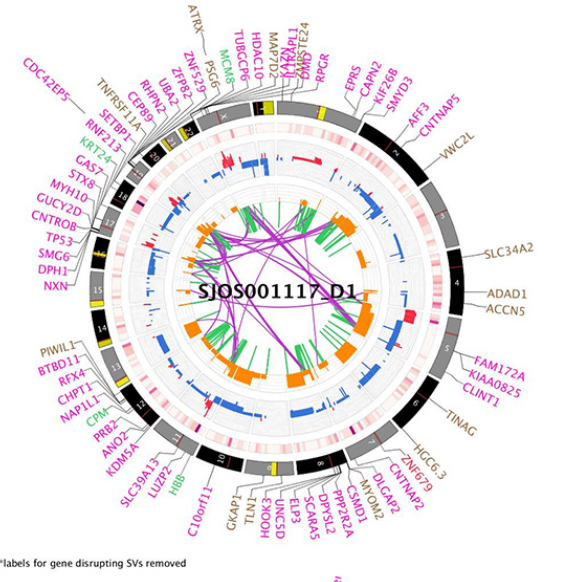

*labels for gene disrupting SVs removed
SIOS00105_D1
ITGB3BP
TRIM33
LRPPRC
SLC1A4
DPP10
LRP2
HYAL3
CADM2
NAALADL2
COX7B2
GABRR1
DNAH5
FAM105B
CDH12
CZCM2
PCDHB2
PCDHB9
WNK1
UPK1
CHYL
FAR52
FAR2
TBC1D22B
BYSL
TREX1
RSPH9
RUNX2
LRRC1
ASCC3
MAP7
MLXIPL
MAGI2
CACNA2D1
CNTNAP2
FAM164A
PRUNE2
ZNF883
SPTAN1
PFKFB3
FRG2B
AMBRA1
PAK1
DLG2
SPG20
STX8P6
AKAP6
GABRA5
ATP2A1
PDE8A
WNK1
LIF1
MAVS
TBC1D16
PRKAB
CHST9
EMC1
FHDC1
C19orf57
NTN5
MACROD2
CSS
TSHZ2
CABIN1
WDR13
PCDH11X
*labels for gene disrupting SVs removed
SIOS00118_D
LOC145820
RYR3
SHISA9
NFATC2
C20orf26
CABP5
FCGBP
RANBP3
PIWIL1
PAX7
ILDR2
TROVE2
CNST
LOC285033
ACMSD
NCEF
C3orf66
AIMP1
C7
PCDHA2
TMEM63B
TMEM151B
TMEM151B
TMEM151B
TMEM151B
CHDH8B52
CPRCRA
SMO
TCF24
*labels for non-coding mutations, silent mutations and gene disrupting SVs removed
SIOS00110_D1
SMCHD1
SMCA1A
CAPN3
TRIM63
SCMH1
MAST2
PDE4B
MTF2
CCDC18
FNBP1L
MFSD4
LRRTM4
TBC1D8
MAP4K4
IL1R2
IL18R1
IL18RAP
SLC9A2
EPC2
GALNT13
KCNJ3
CRM7
ERC2
FHIT
ROBO2
EPHA6
FHLIP1L
C3orf26
CRBD5
MANBA
MRAP2L
MRAP2
PDE4D
PDE4D
CWC27
ADAMTS6
PCDHG-Cluster
PCDHB15
OOZ2
SUPT3H
RUNX2
OCA2
NEOX2
AUTS2
CSMD1
PTK2
PTPRD
PCSK5
PRUNE2
VPS13A
GNA14
GNAQ
TLE1
NTRK2
COLM1
ISCA1
COL27A1
KIF5B
PRKG1
M6A4GA
XRRA1
SLC2A13
ANKS1B
LLPH
E2F7
SLC15A4
NCOR2
AKAP5
CHURC1
CHURC1-FNTB
TMED8
CANC
C16orf59
TAT53
MPHH13
MPHIP
SUZ12
FBXL20
KCNH4
KCNH4
COASY
MACROD2
*labels for gene disrupting SVs removed
SIOS00119_D
NUDT21
CPSF6
CHD2
ARRDC4
CPR68
IL31
SLCO1A2
LOC338651
MYOF
EDA
ASXL1
LSM14A
C18orf21
FNDC5
LOC100129269
ESRRG
NUDT16
ESPNL
CCDC28A
CABRB1
FRYL
TEC
CWH43
MANBA
UBE2D3
PZR
ANKCAP18
MKZF1
MAGI2
SLC12A9
CRM8
TRIM24
LOC728724
*labels for gene disrupting SVs removed
SIOS00103_D1
MTMR3
ASCC2
LLPAN1
ZNF90
ATP13A1
PRX4
OSBPL1A
MAP3K3
AMZ2
KCNH6
CLTC
ADORA2B
DNAH2
GRIN2A
BFAR
GABRG3
DPP8
ACTN1
FAM71D
GPHN
PCK2
TMEM132B
NOS1
CD163
ORS8S4
DLG2
OD24
PDE2A
VSTM4
KDM4C
SNX31
LRRC69
CSMD1
ORAI2
PRKRIP1
ABCB1
ZPBP
AHI1
ALDH6A1
MON1
C6orf165
BACH
PPP3CA
VEPH1
FOXP1
CCDC36
REV1
CTDC1
ZER2
WDR35
SUSD4
KAZN
FHAD1
*labels for non-coding mutations, silent mutations and gene disrupting SVs removed
SIOS00107_M1
RAB10
HADHB
ALK
CAPN14
PRKCE
SCN7A
TFDP2
NAALADL2
APBB2
GK2
CYP8
SMAD1
HEATR7B2
PCDHC-Cluster
PCDHGA2
SUFU
NIF3A
KCQ2
RITA1
DNAH8
KIF6
TFB
WIPO5
POLH
RADIL
DNAH11
MAGI2
BAZ1B
DOCK4
CNTNAP2
PTPRN2
ATP6V1H
ACBD7
OTUD1
ANKRD26
ANUBL1
ZRANB1
CTBP2
ORS2N2
CNTN5
FEZ1
CACNA2D4
ANO4
WIF
TNFRSF1A
NCAPD2
NCAPD2
NCAPD2
NCAPD2
TMTC1
NRXN3
ALOXE3
MYH10
MYH10
LRP4
ISM1
LRRC4B
MYO10
SLC4A10
MYO19
BPGT
PPP4R4
RAB31
ROCK1
C18orf8
PARD6G
MICU1
ZNF337
CTDP1
PHKA2
FAM120C



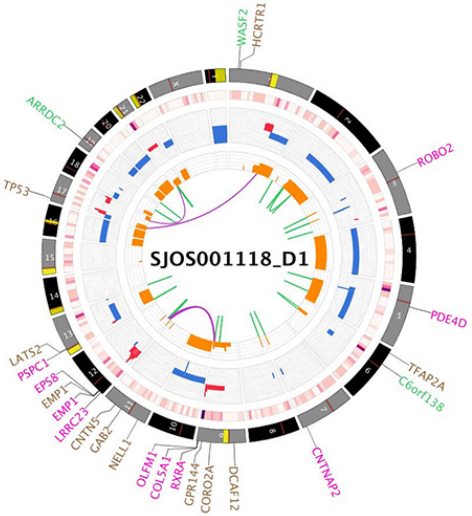
*labels for non-coding mutations, silent mutations and gene disrupting SVs removed



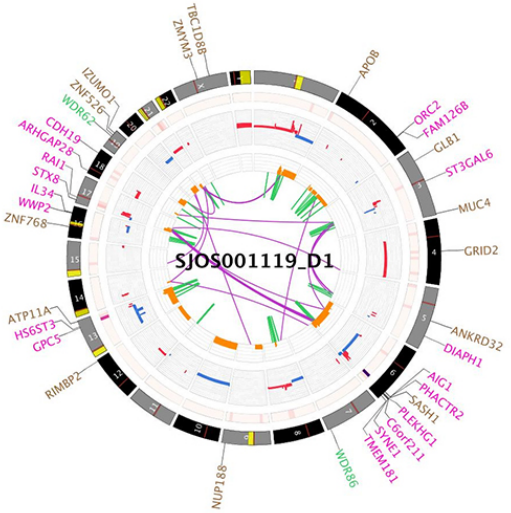
*labels for non-coding mutations, silent mutations and gene disrupting SVs removed



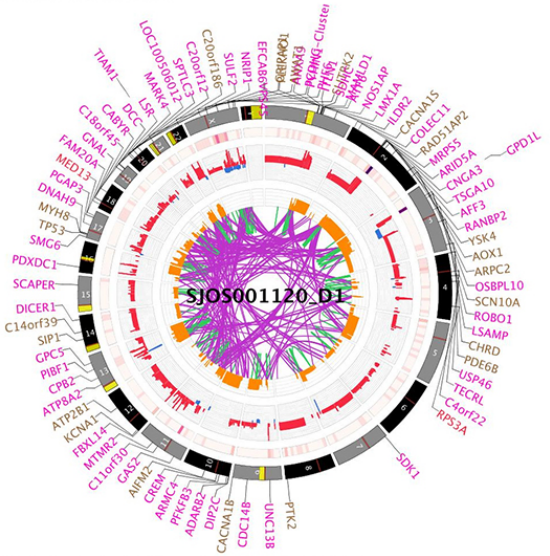
*labels for gene disrupting SVs removed



*labels for gene disrupting SVs removed



*labels for gene disrupting SVs removed



*labels for non-coding mutations, silent mutations and gene disrupting SVs removed

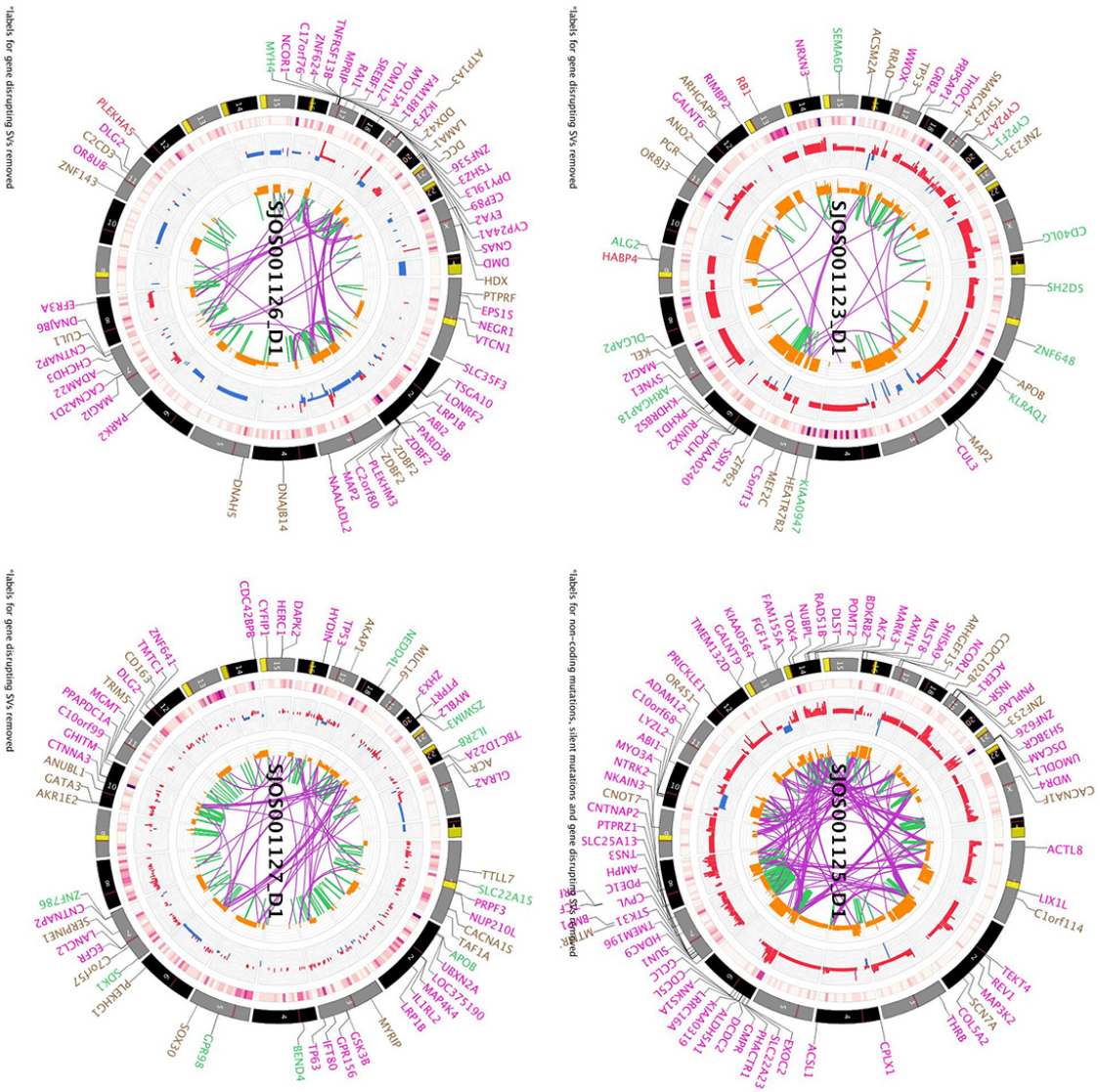

*labels for gene disrupting SVs removed
SIGS00112126_D1
MYH4
C17orf76
NCOR1
ZNF624
TNRSF13B
MPHOSPH8
MAP1A
SGEF
TOMM22
MYO15A
FAM189A1
IKZF3
ODF2L
LAMA1
DCC
ZNF536
TSHZ3
DPY19L3
CEP89
EYA2
CYP24A1
CNAS
DMD
HDX
PTPRF
EPS15
NEGR1
VTCN1
SLC35F3
TSCA10
LONRF2
LRP1B
ABI2
PARD3B
ZDBF2
C2orf80
PLEKHM3
MAP2
NAALADL2
DNAJB14
DNAH5
PAK2
MAGI2
CACNA2D1
ADAM22
CNTNAP2
CUL1
DNAJB6
EFR3A
ZNF143
OR8U8
C2CD3
DLG2
PLEKHA5
ATP1A3
*labels for gene disrupting SVs removed
SIGS00112123_D1
NRXN3
SEMA6D
ACSM2A
RRAD
WWOX
TP53
GAB2
PRPSAP1
SMARCA4
THOC1
TSHZ3
CYP24A1
ZNF233
CD40LG
SH2D5
ZNF648
APOB
KLRAQ1
MAP2
CUL3
KIAA0947
HEATR7B2
MEF2C
C5orf13
ZFP62
SSH1
POLH
KIAA0240
PLXNC1
PKHD1
PKHD8S2
ARHGAP18
SYNE1
MAGI2
KEL
DLGAP2
HABP4
ALG2
OR8J3
PGR
ANO2
GALNT6
ARHGAP6
RIMBP2
RB1
*labels for gene disrupting SVs removed
SIGS00112127_D1
DAPK2
HERC1
CYFIP1
CDC42BPB
HYDIN
TJP3
AP4
AK4
NEDD4L
MUC16
ZHX3
PTPRT
MYBL2
ZSWIM3
IL2RB
TBC1D22A
ACK
GLRA2
TTLL7
SLC22A15
PRPF3
NUP210L
CACNA1S
TAF1A
APOB
UBXN2A
LOC375190
MAP4K4
IL1RL2
LRP1B
MYRIP
GSK3B
CPR156
FT80
TP63
BEND4
SOX30
CPR98
PLEKHG1
SOX1
C7orf57
EGFR
LANCL2
SERPINE1
CNTNAP2
ZNF786
ZNF641
TMTC1
CD163
DLG2
TRIM5
MCMT
PPAPDC1A
C10orf99
GHITM
CTNNA3
ANUBL1
GATA3
AKR1E2
*labels for non-coding mutations, silent mutations and gene disrupting SVs removed
SIGS00112125_D1
CACNA1F
WDR4
UMODL1
DSCAM
SH3BGR
ZNF826
ZNF259
PNPLA6
NSFL
ACER1
CCDC102B
NCOR1
ARHGEF15
SHISA9
MLST8
TNIK
MARK3
AK7
BDKRB2
POMT2
DLST
RAD51B
NUBP1
TOX4
FAM155A
FGF14
KIAA0564
GALNT9
TMEM132D
PRICKLE1
OR4S1
ADAM12
C10orf68
LYZL2
ABI1
MYO3A
NTRK2
NKAIN3
CNOT7
CNTNAP2
PTPRZ1
SLC25A13
TNS3
AMPH
PDE1C
CPVL
STX3
TMEM196
HOXC9
SUN1
CCLC
CCSL
ANKS1A
TARBP1
KIAA0319
ALDH5A1
DCDC2
PHACTR1
SLC22A23
EXOC2
ACSL1
CPLX1
THRB
COL5A2
SGCNZA
MAP3K2
REV1
TEKT4
C1orf114
LIX1L
ACTL8
*labels for gene disrupting SVs removed

Figure S1 related to Figure 1. Copy number analysis of osteosarcoma discovery cohort. A) Copy number analysis for all 34 osteosarcomas in the discovery cohort with red indicating gain and blue indicating loss. **B)** Histogram of the type of mutations (SNV, CNV, SV, indel) across the osteosarcoma tumors in the discovery cohort. **C)** GISTIC analysis of the copy number changes in the osteosarcoma discovery cohort with green indicating gains and red indicating loss. The dashed lines represent cutoff for statistical significance and the individual chromosomes are labeled along the bottom of the plot. **D)** CIRCOS plots of all 34 tumors sequenced by WGS in this study.

Table S1 related to Figure 1. Clinical features of discovery and validation cohorts.

Provided as a separate file.

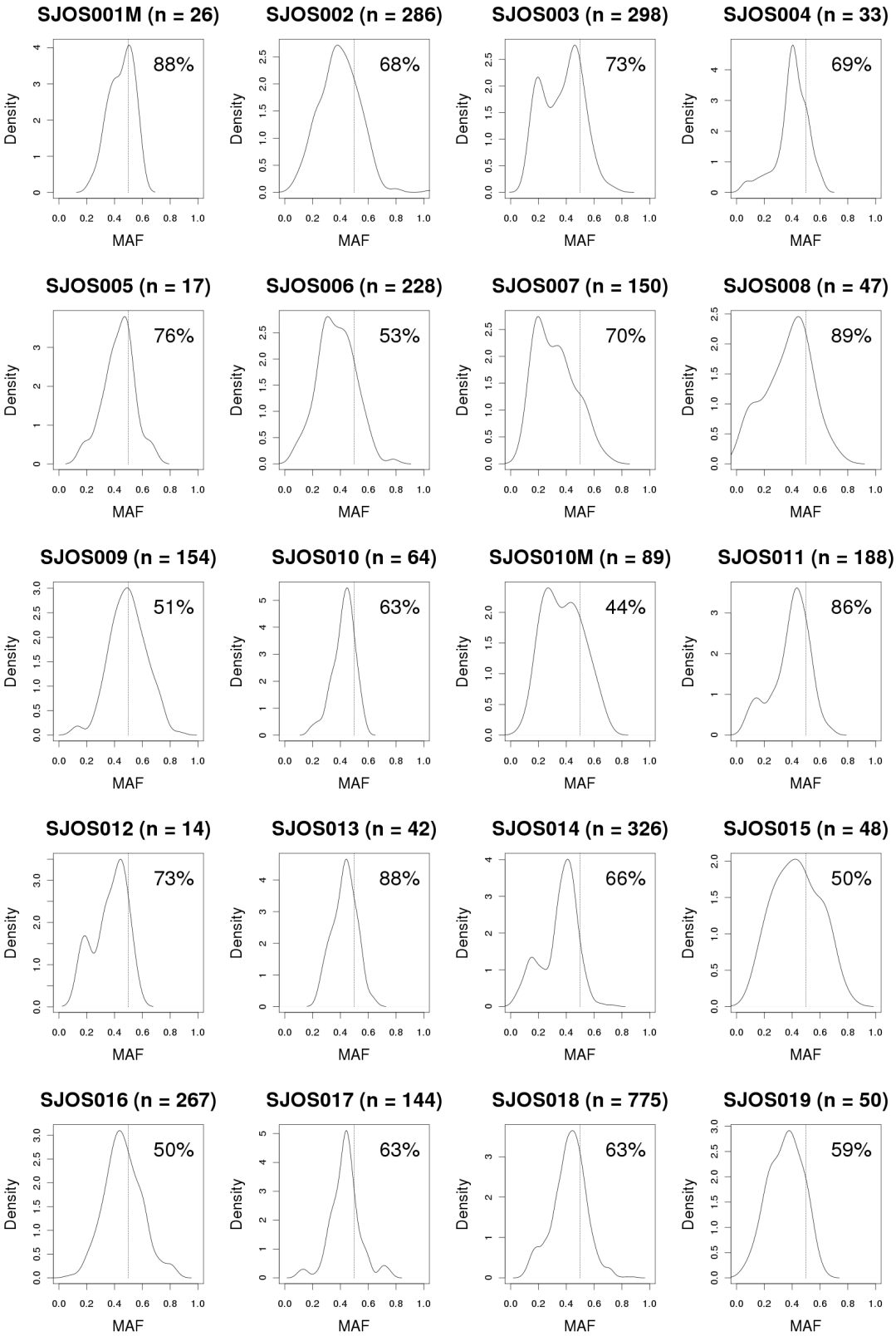
Table S2 related to Figure 1. Sequence coverage data.

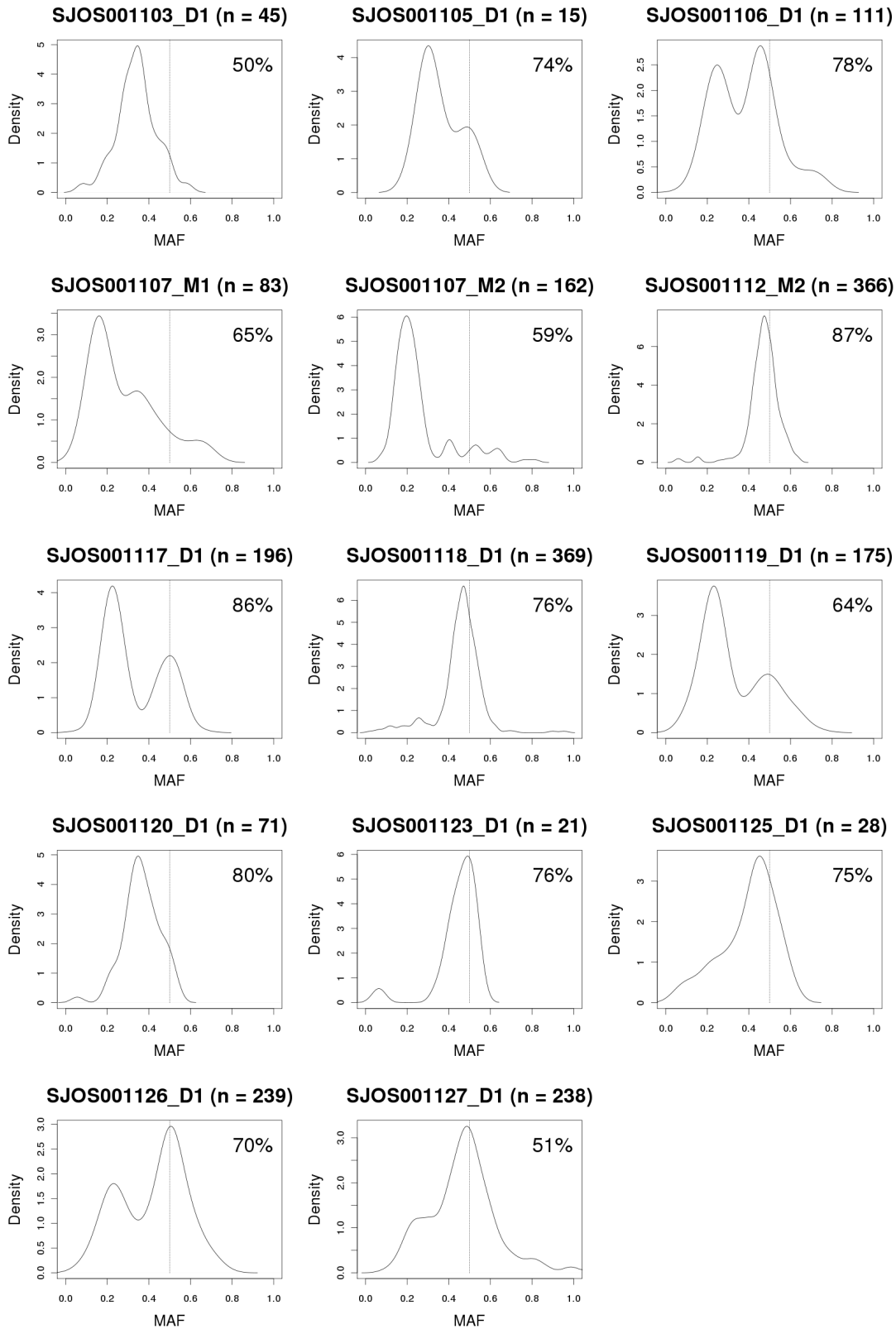
Provided as a separate file.

Table S3 related to Figure 1. Validated mutations.

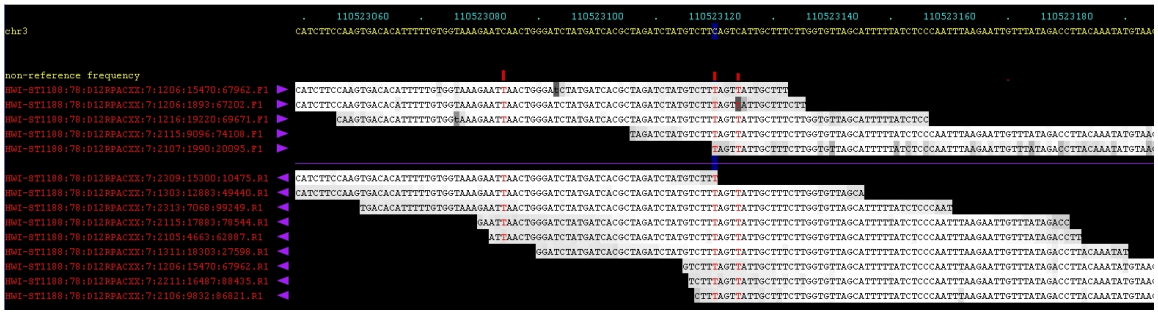
Provided as a separate file.

A

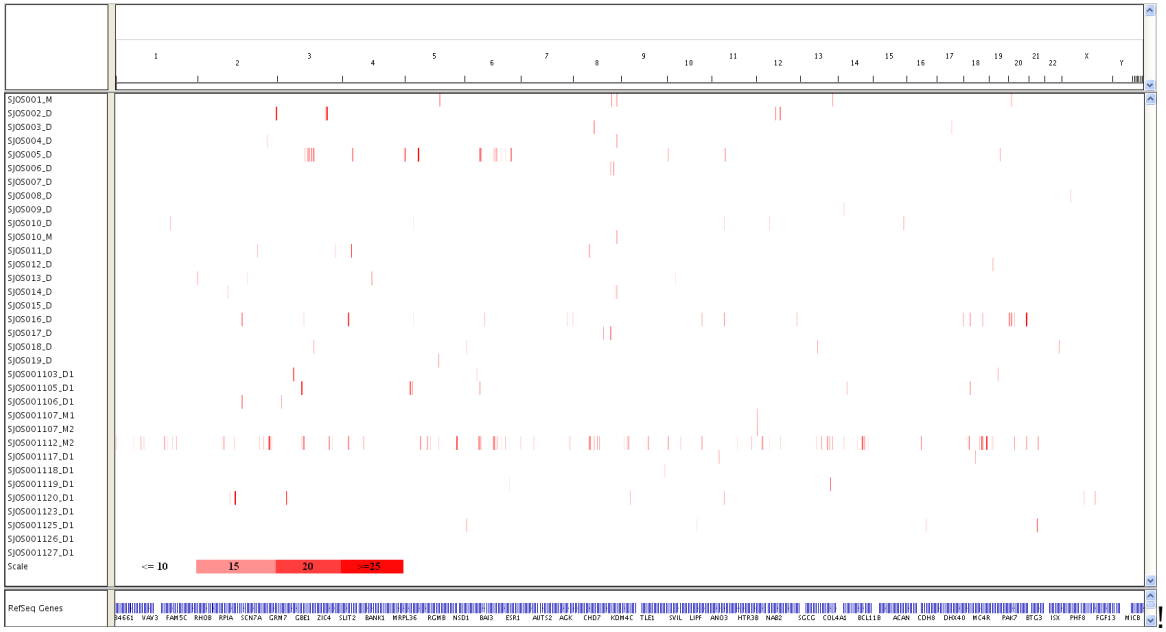

SJOS001M (n = 26)
88%
SJOS002 (n = 286)
68%
SJOS003 (n = 298)
73%
SJOS004 (n = 33)
69%
SJOS005 (n = 17)
76%
SJOS006 (n = 228)
53%
SJOS007 (n = 150)
70%
SJOS008 (n = 47)
89%
SJOS009 (n = 154)
51%
SJOS010 (n = 64)
63%
SJOS010M (n = 89)
44%
SJOS011 (n = 188)
86%
SJOS012 (n = 14)
73%
SJOS013 (n = 42)
88%
SJOS014 (n = 326)
66%
SJOS015 (n = 48)
50%
SJOS016 (n = 267)
50%
SJOS017 (n = 144)
63%
SJOS018 (n = 775)
63%
SJOS019 (n = 50)
59%
Density
MAF


SJOS001103_D1 (n = 45)
50%
SJOS001105_D1 (n = 15)
74%
SJOS001106_D1 (n = 111)
78%
SJOS001107_M1 (n = 83)
65%
SJOS001107_M2 (n = 162)
59%
SJOS001112_M2 (n = 366)
87%
SJOS001117_D1 (n = 196)
86%
SJOS001118_D1 (n = 369)
76%
SJOS001119_D1 (n = 175)
64%
SJOS001120_D1 (n = 71)
80%
SJOS001123_D1 (n = 21)
76%
SJOS001125_D1 (n = 28)
75%
SJOS001126_D1 (n = 239)
70%
SJOS001127_D1 (n = 238)
51%
Density
MAF

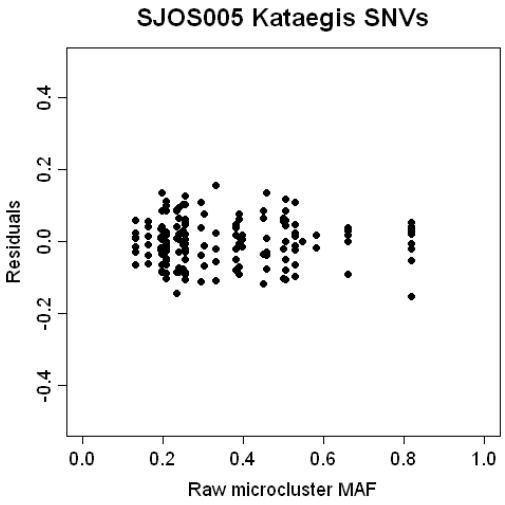
B


chr3
110523060
110523080
110523100
110523120
110523140
110523160
110523180
non-reference frequency

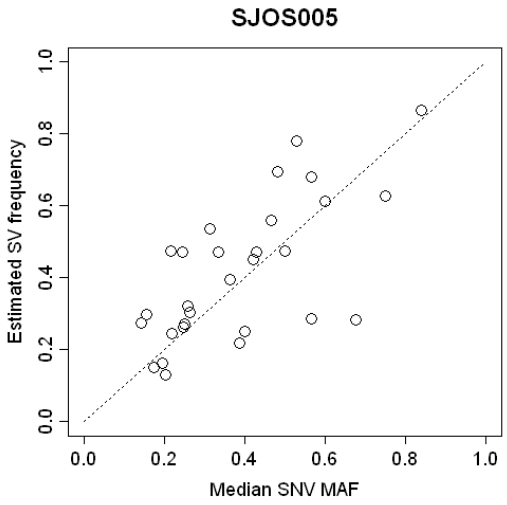
C

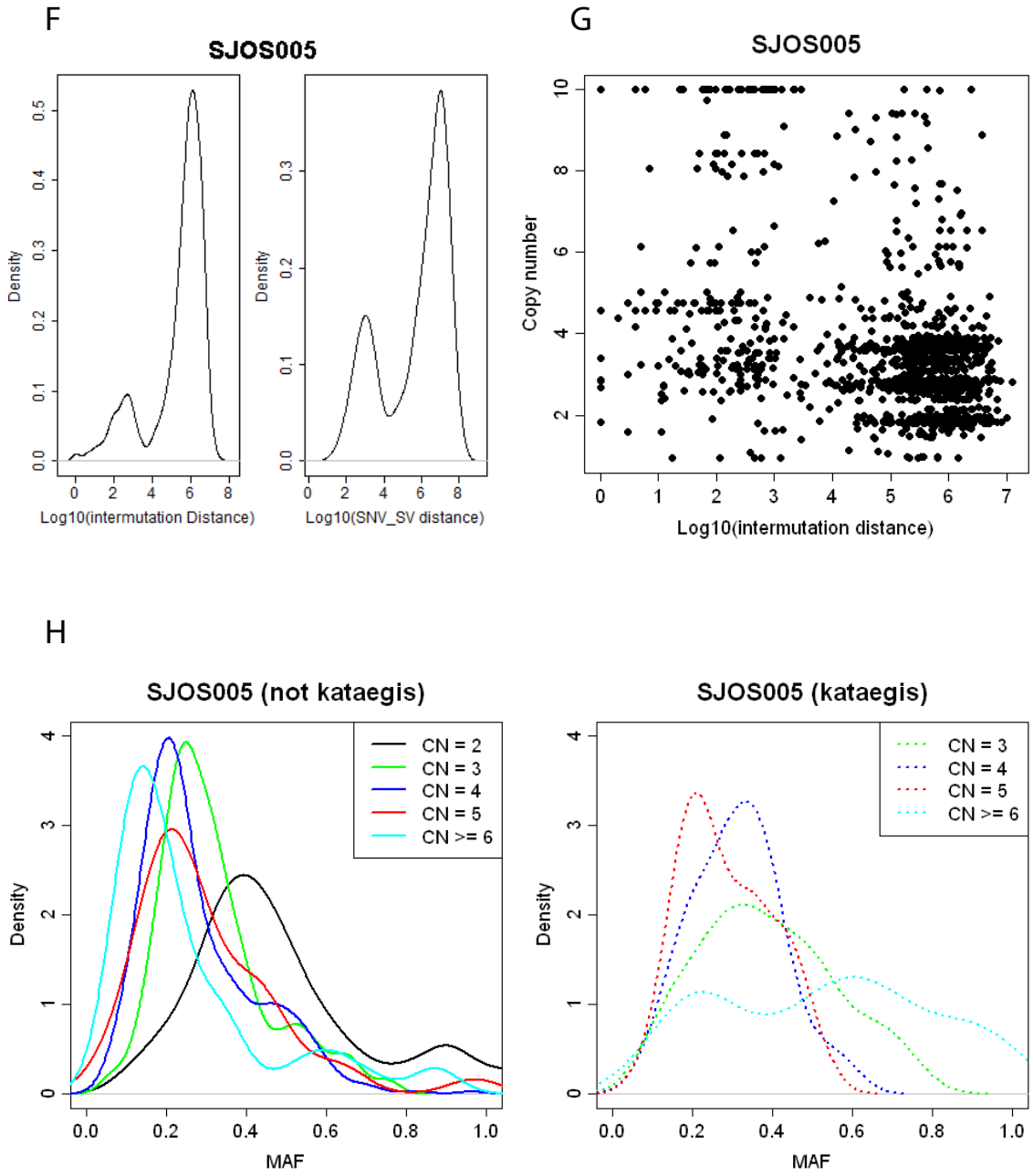

Scale
<= 10
15
20
>=25
RefSeq Genes

D

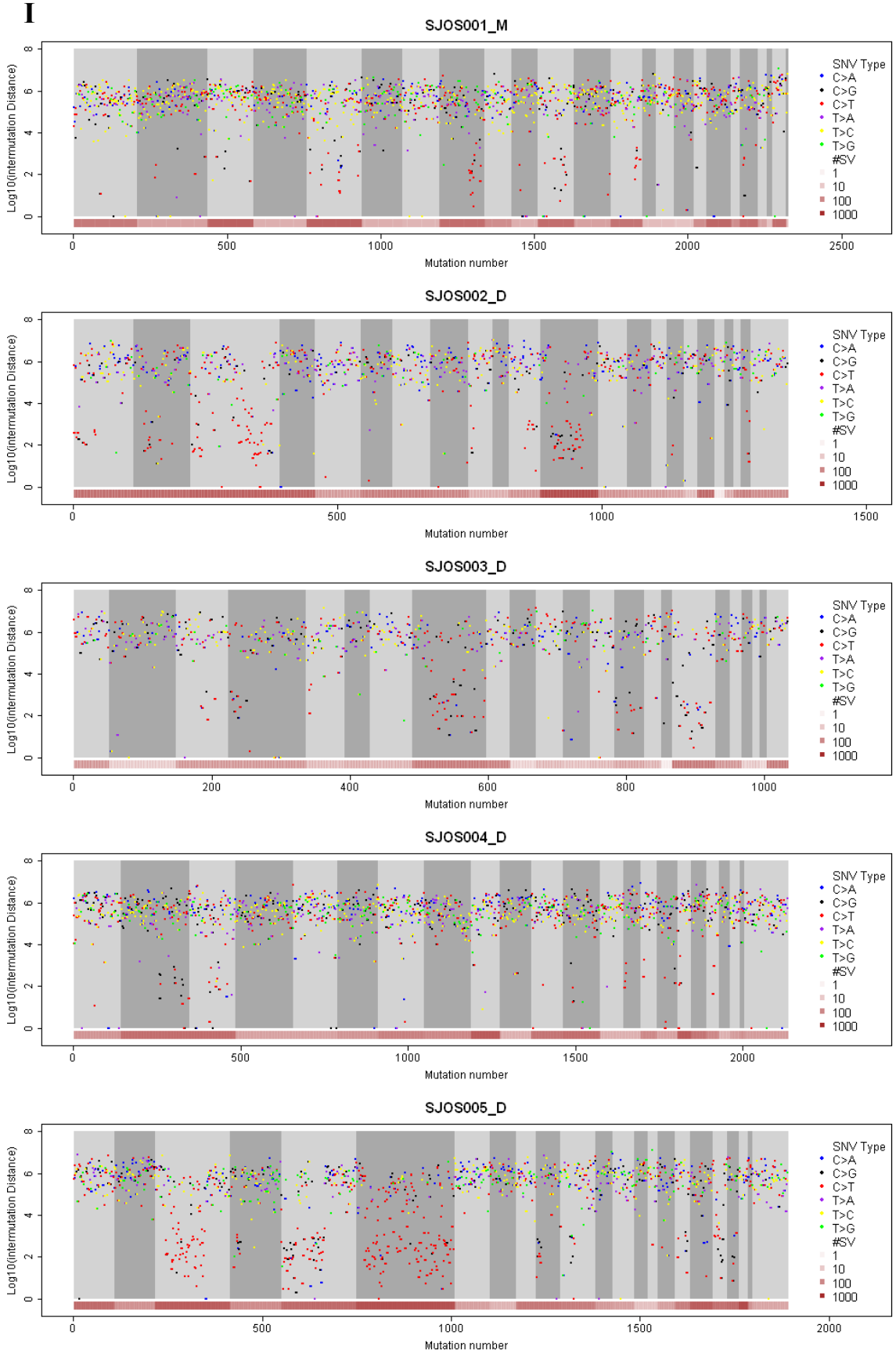

SJOS005 Kataegis SNVs
Residuals
Raw microcluster MAF

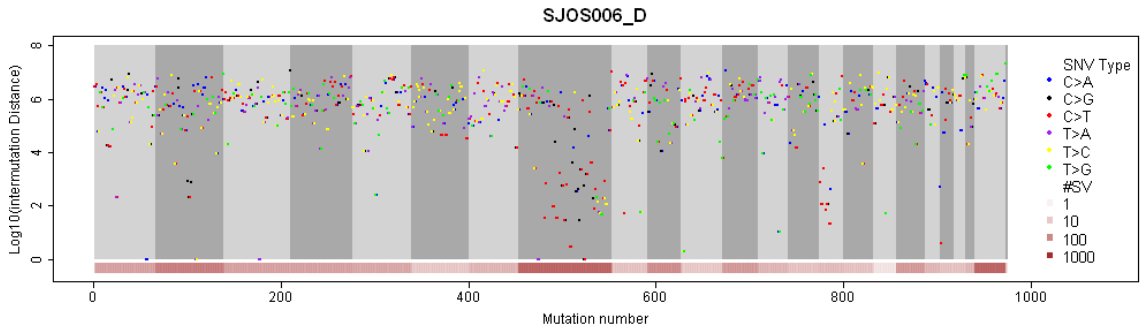
E

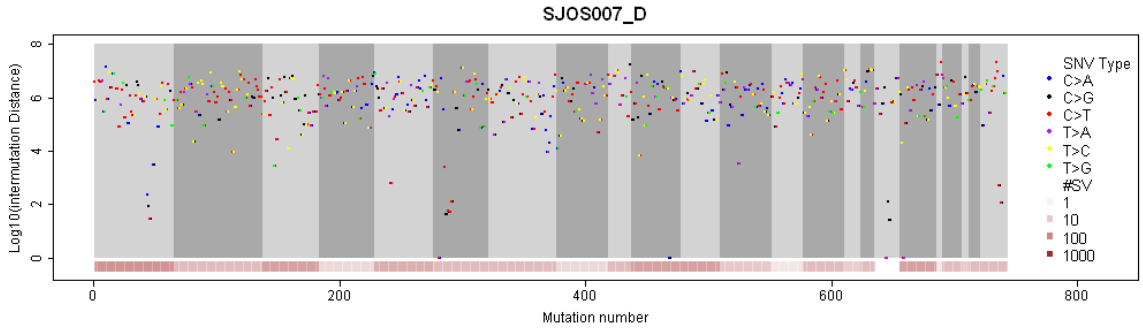

SJOS005
Estimated SV frequency
Median SNV MAF

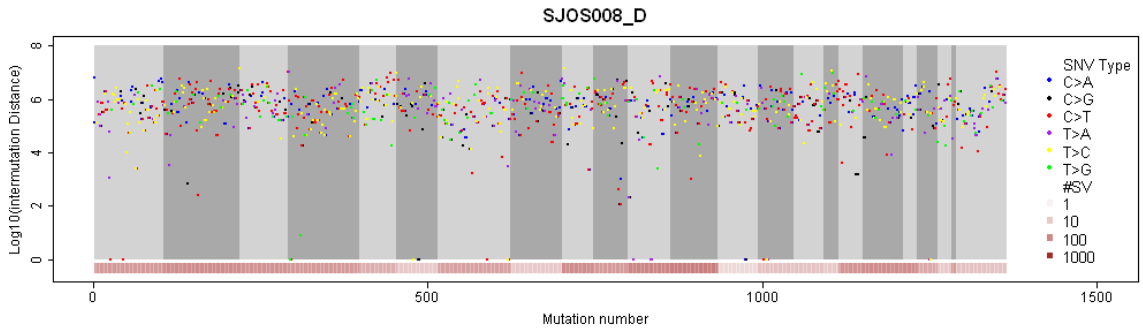

F
SJOS005
Density
Log10(intermutation Distance)
Log10(SNV_SV distance)
G
SJOS005
Copy number
Log10(intermutation distance)
H
SJOS005 (not kataegis)
CN = 2
CN = 3
CN = 4
CN = 5
CN >= 6
Density
MAF
SJOS005 (kataegis)
CN = 3
CN = 4
CN = 5
CN >= 6
Density
MAF

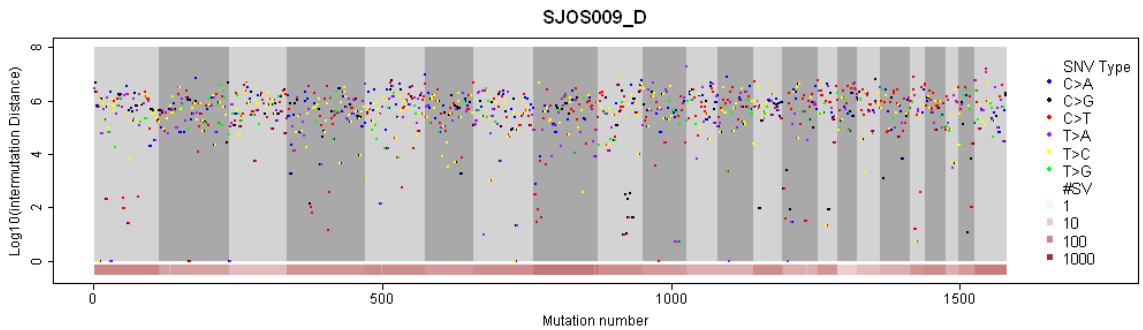
I

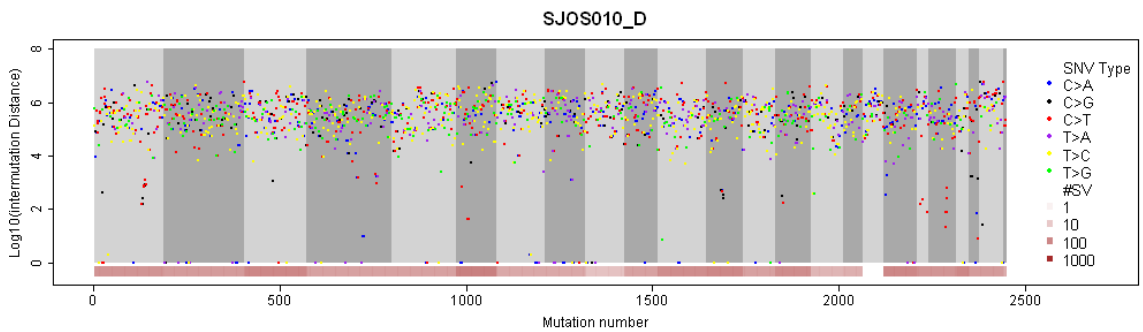

SJOS001_M
SJOS002_D
SJOS003_D
SJOS004_D
SJOS005_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

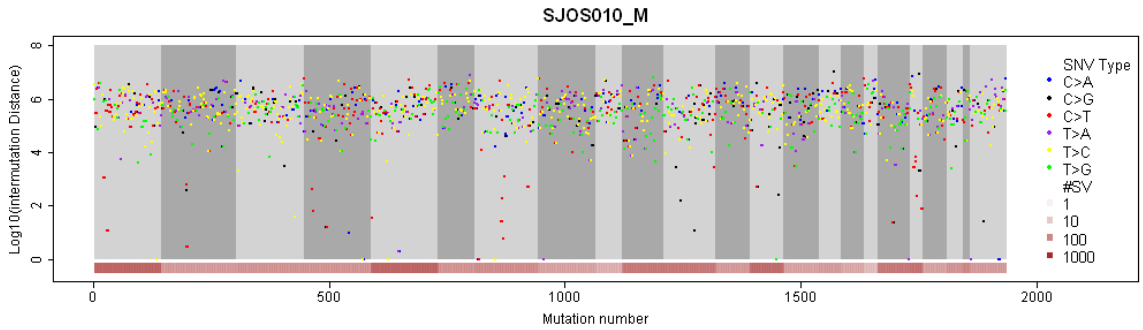

SJOS006_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

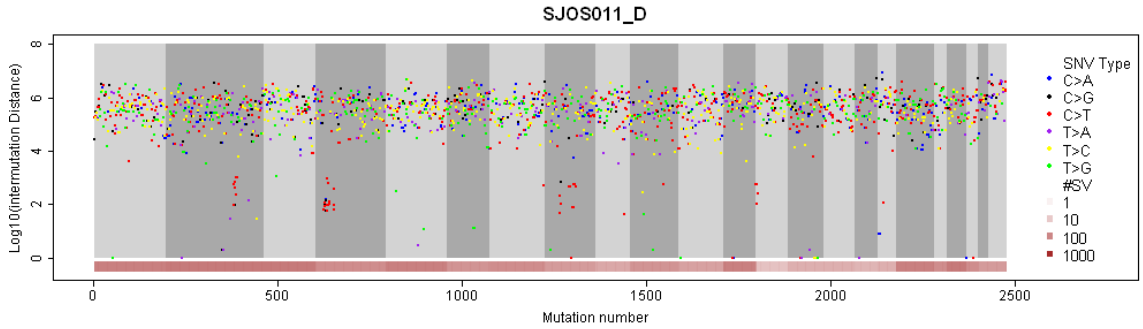

SJOS007_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

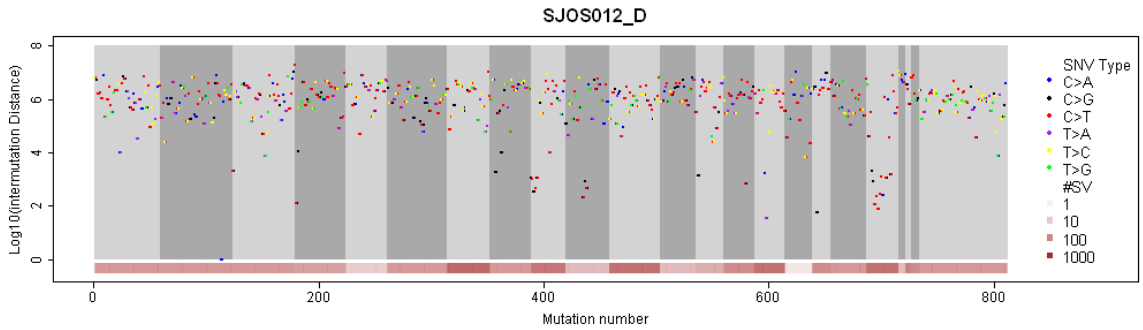

SJOS008_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

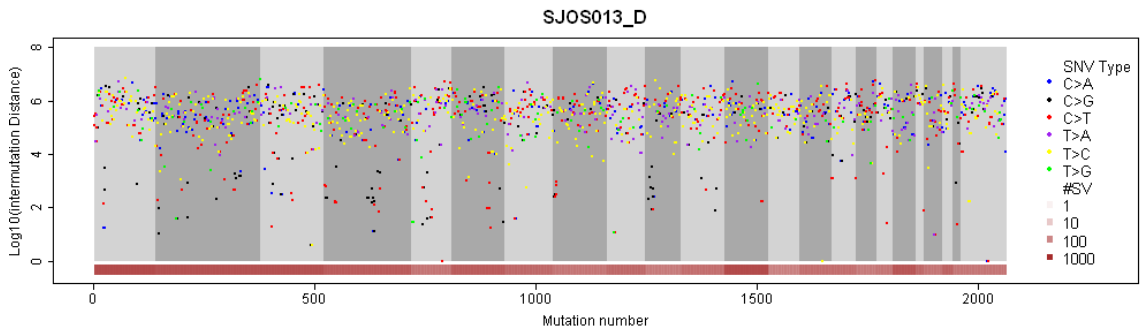

SJOS009_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

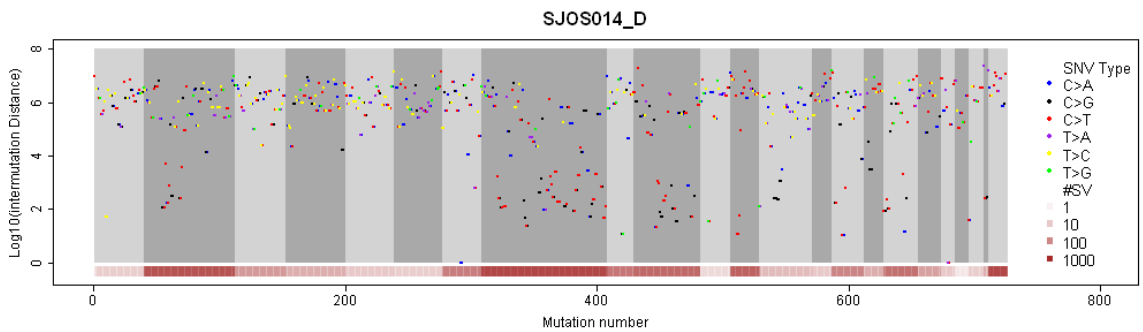

SJOS010_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

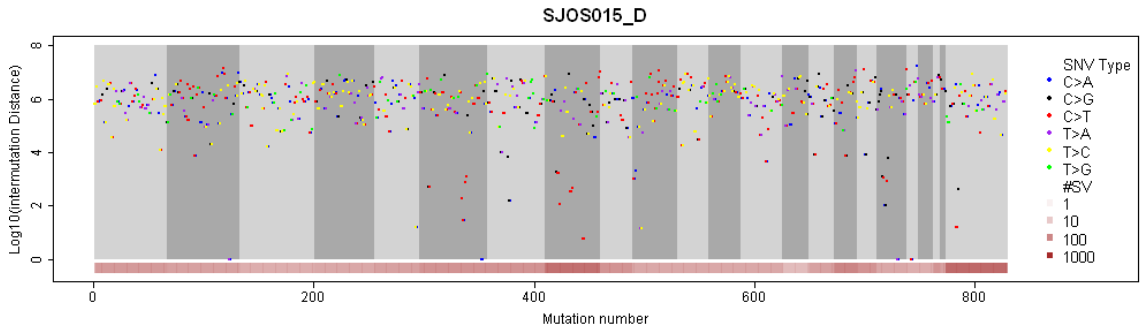

SJOS010_M
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

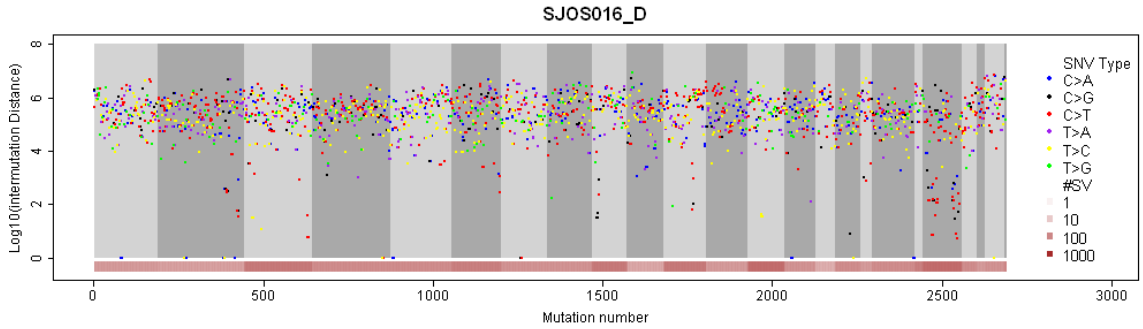

SJOS011_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

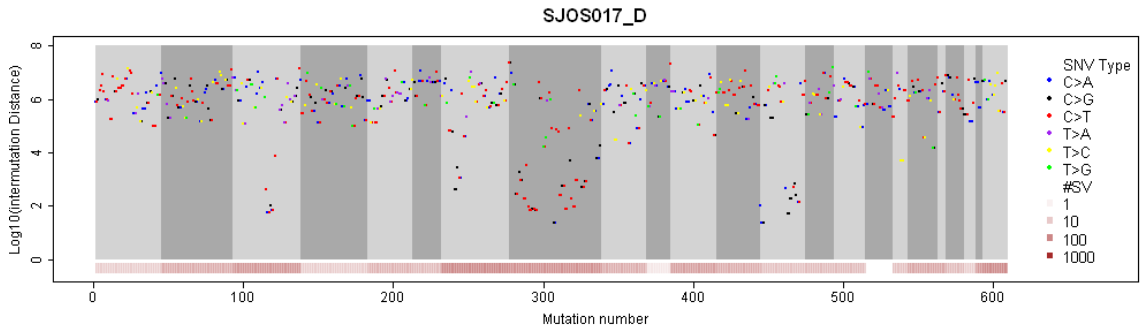

SJOS012_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

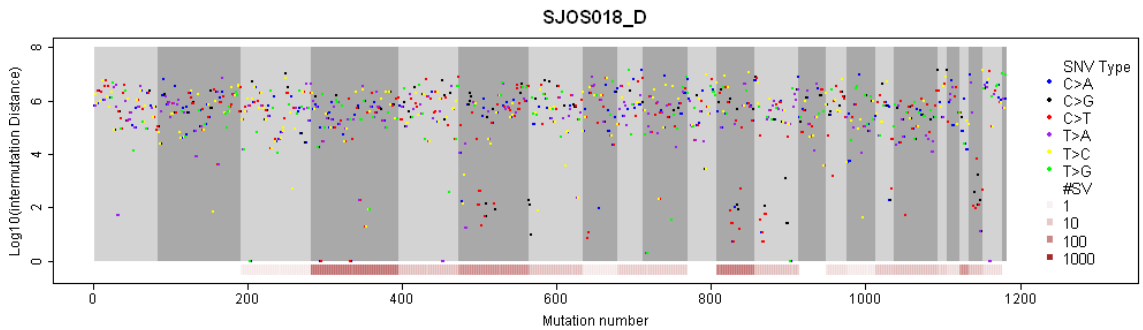

SJOS013_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

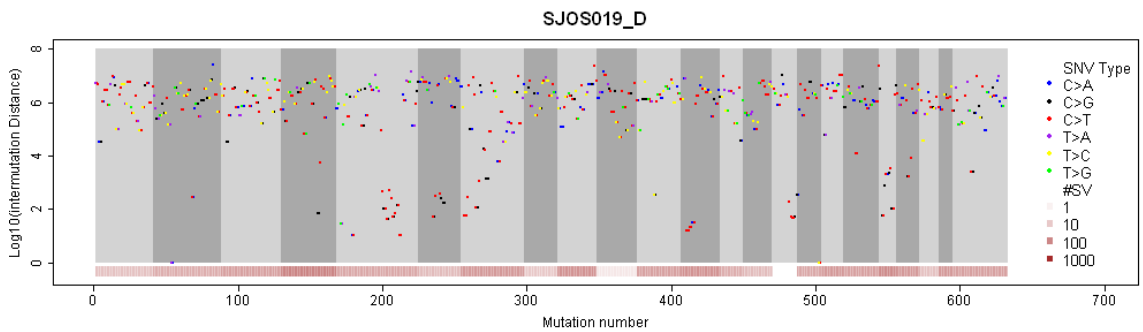

SJOS014_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

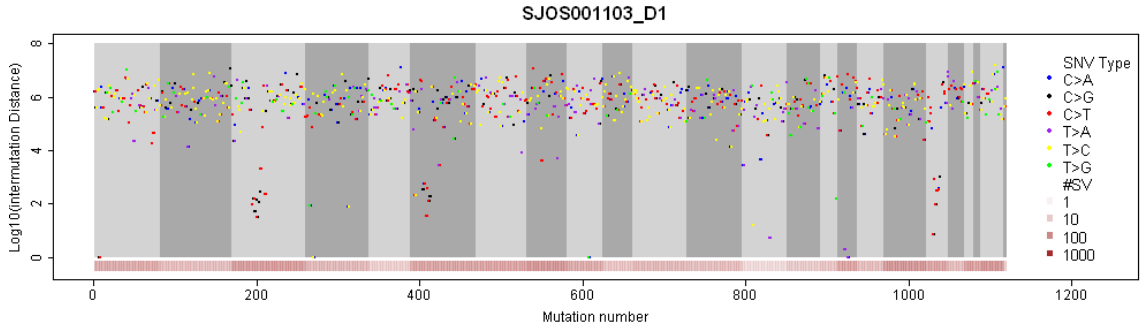

SJOS015_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

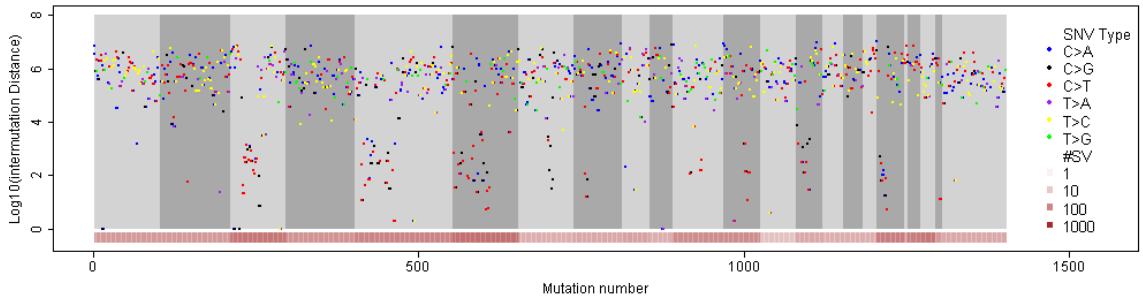

SJOS016_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

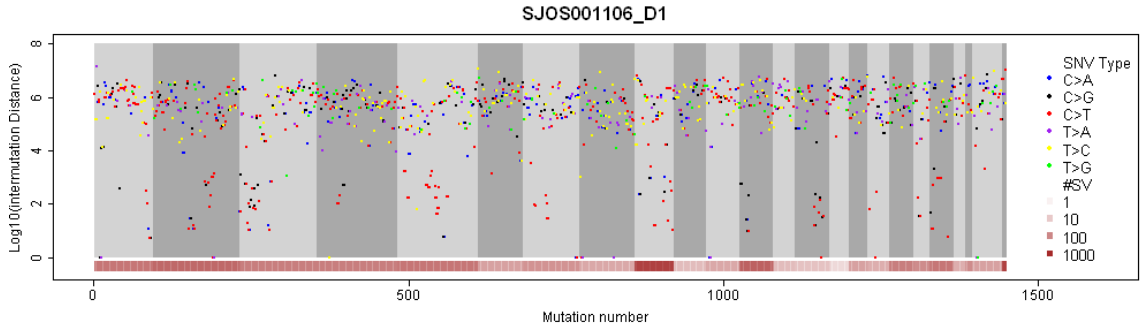

SJOS017_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

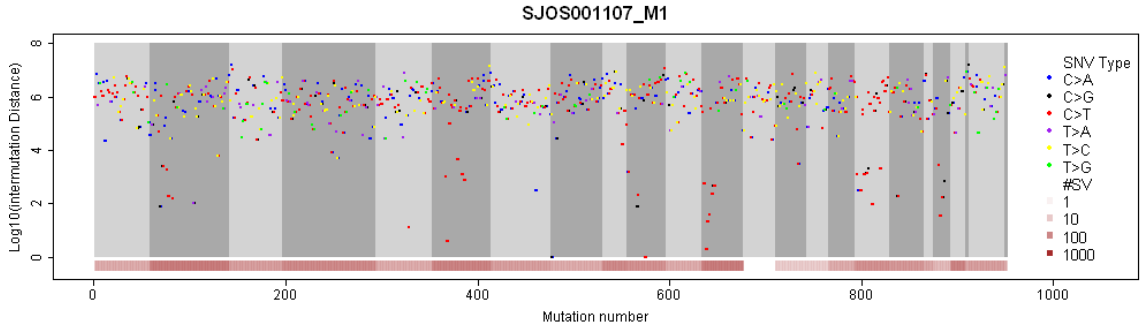

SJOS018_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

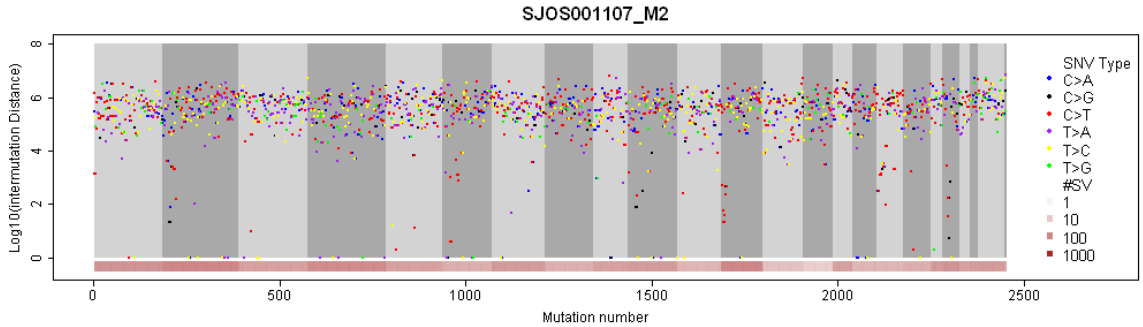

SJOS019_D
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

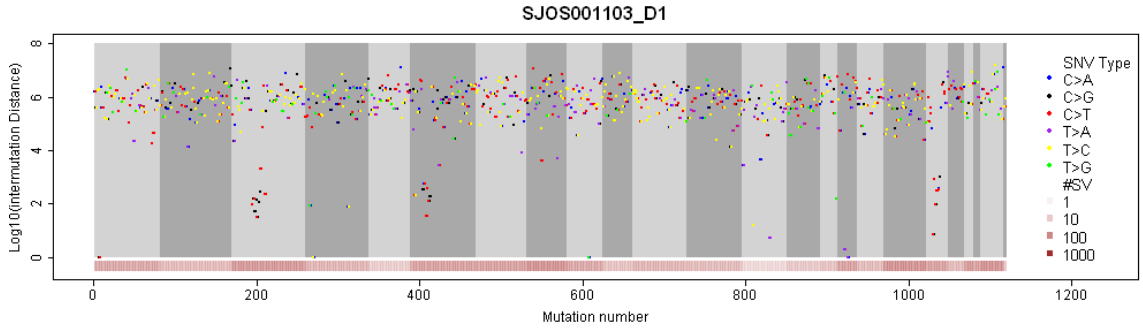

SJOS001103_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

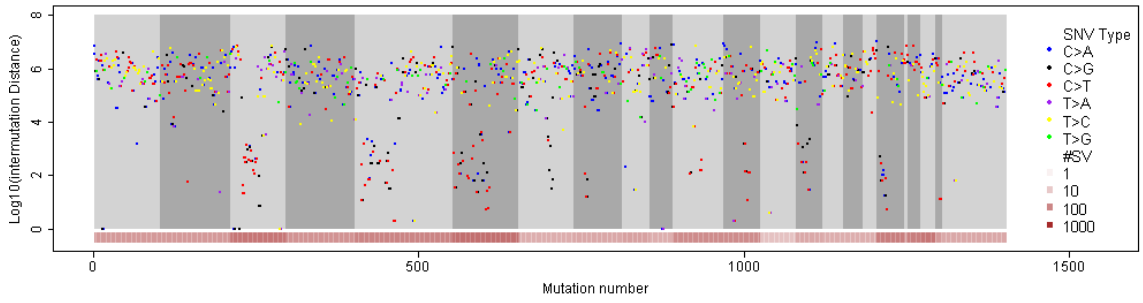

SJOS001105_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

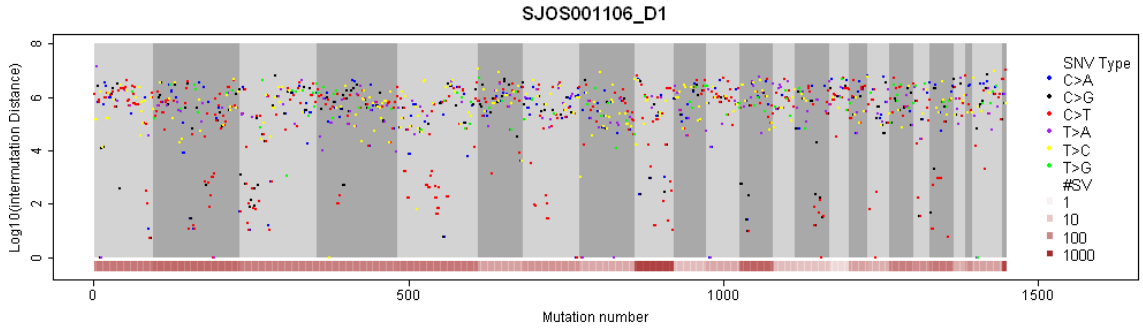

SJOS001106_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

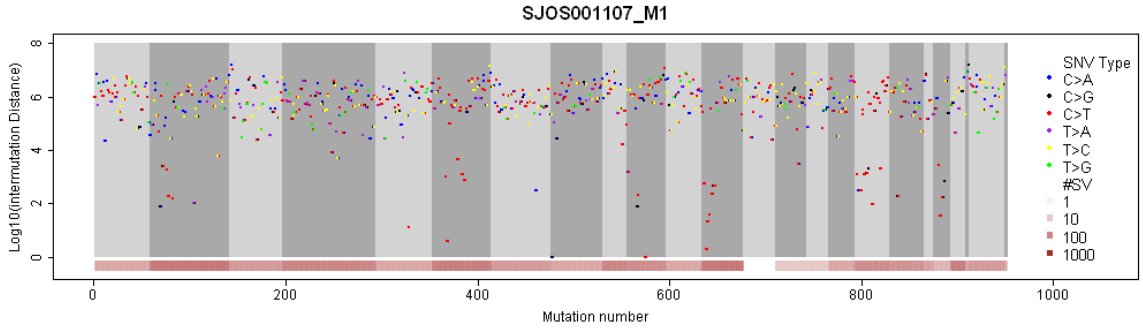

SJOS001107_M1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

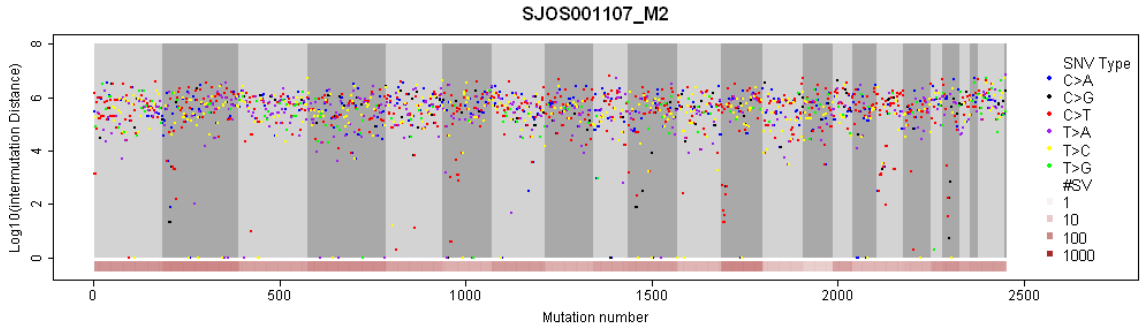

SJOS001107_M2
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

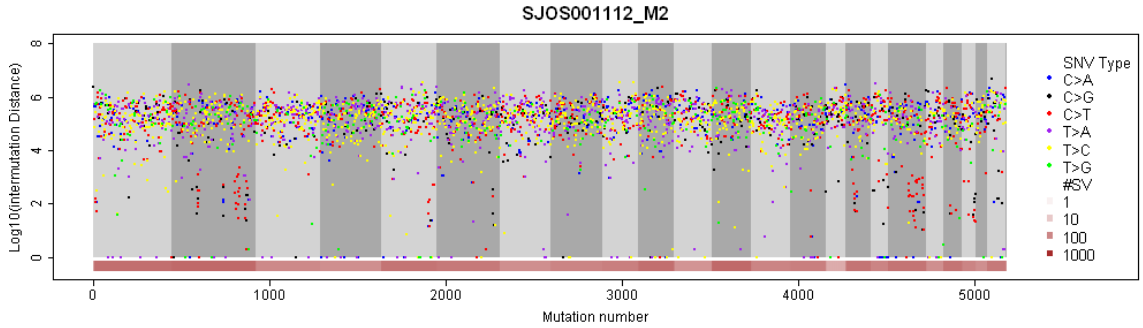

SJOS001103_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

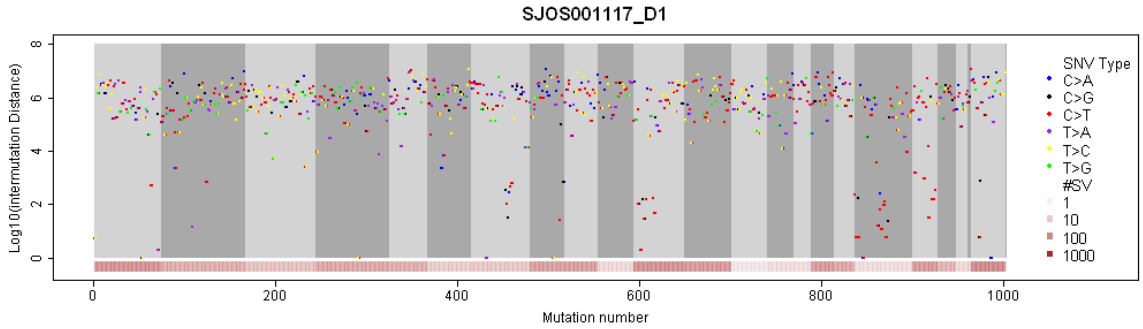

SJOS001105_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

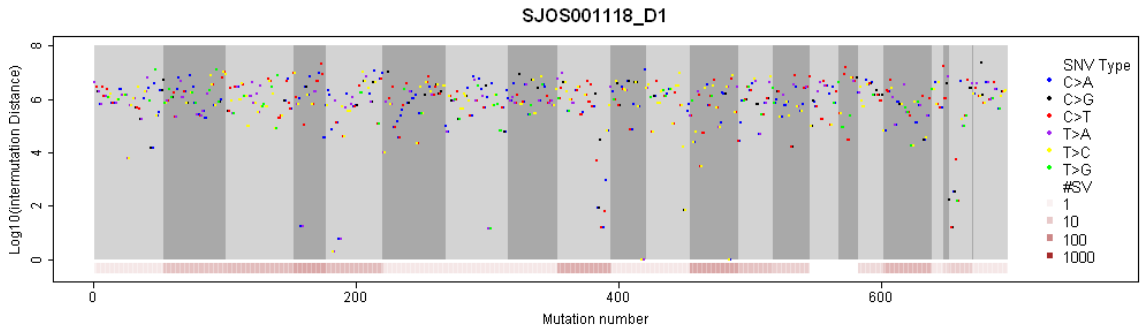

SJOS001106_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

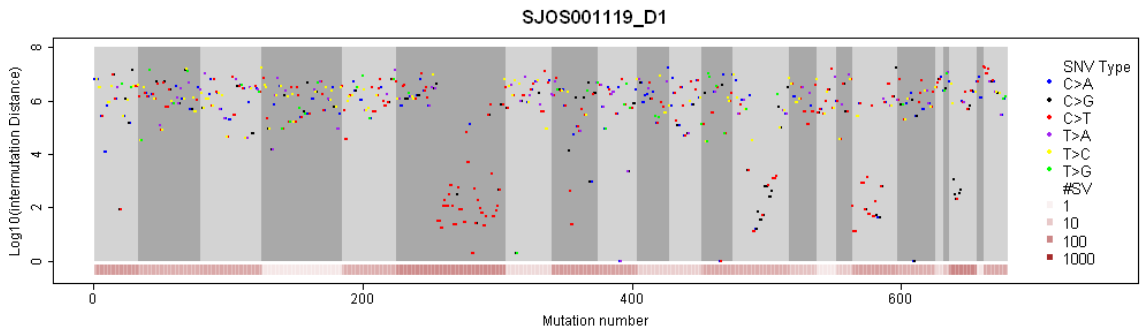

SJOS001107_M1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

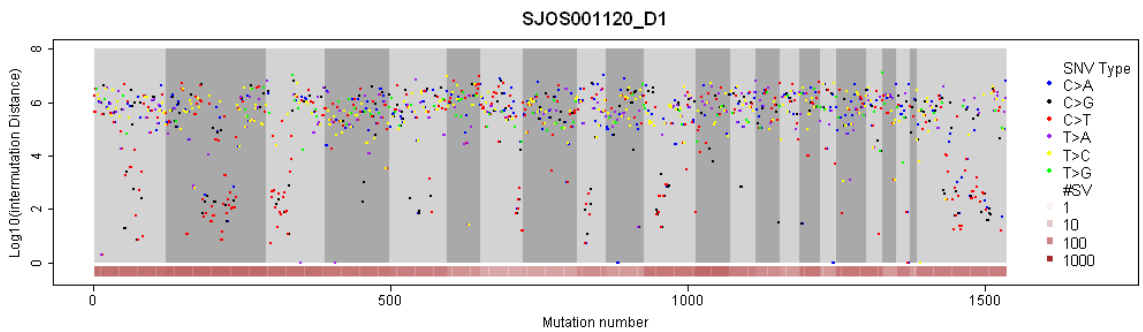

SJOS001107_M2
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

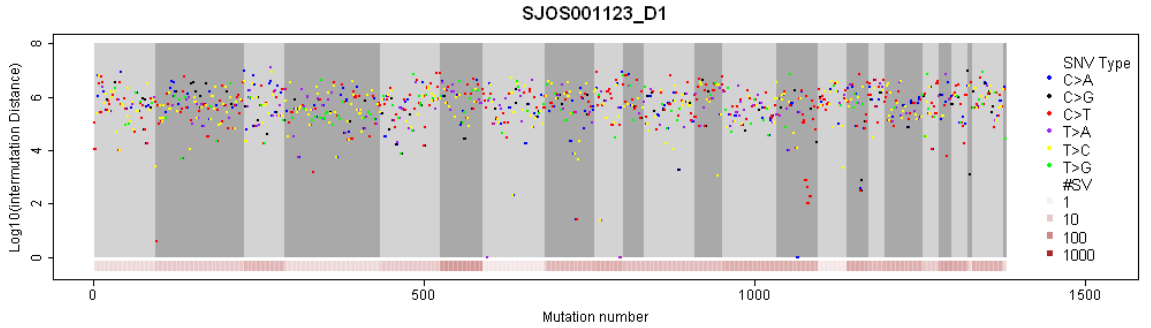

SJOS001112_M2
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

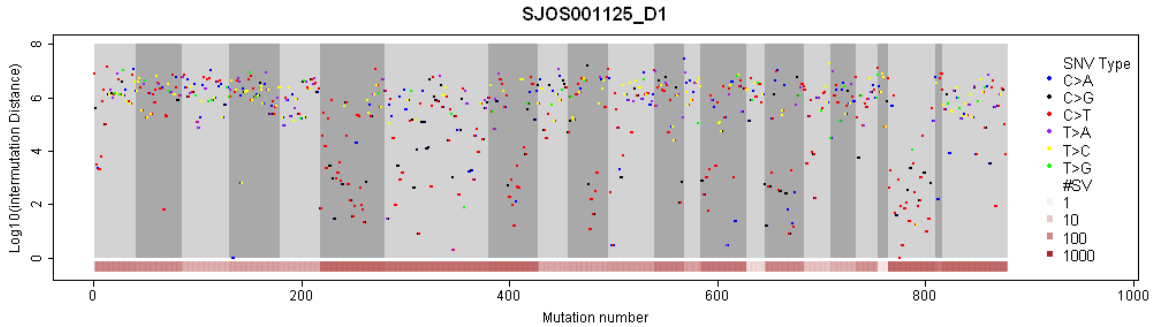

SJOS001117_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

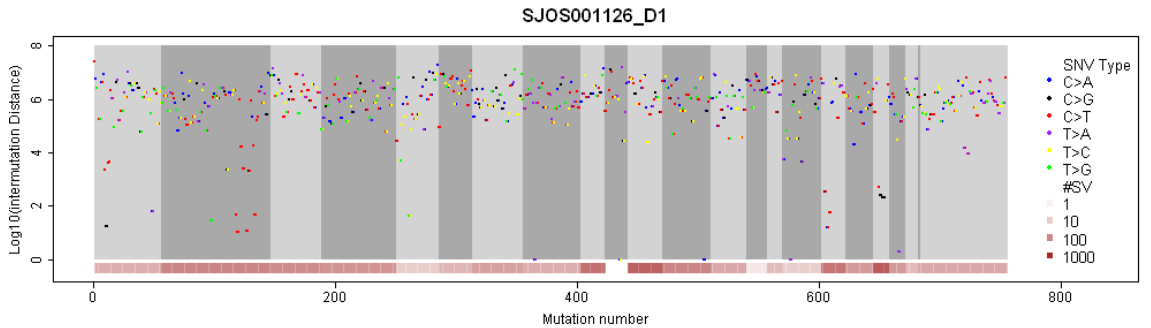

SJOS001118_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000


SJOS001119_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000


SJOS001120_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000


SJOS001123_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000


SJOS001125_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000


SJOS001126_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

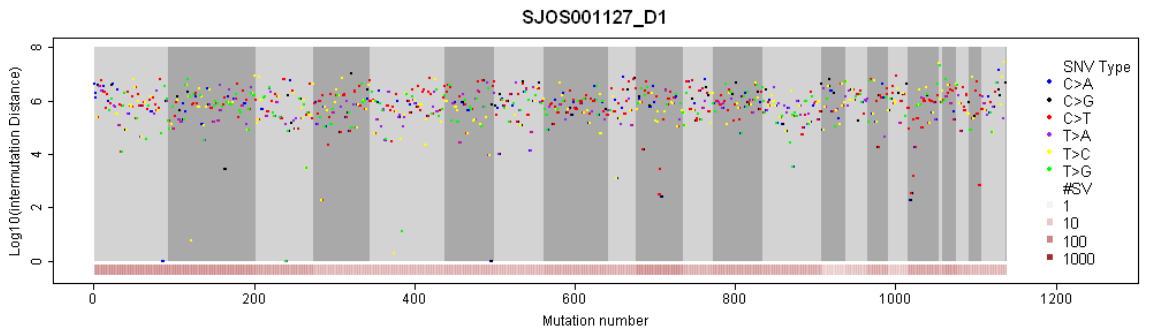
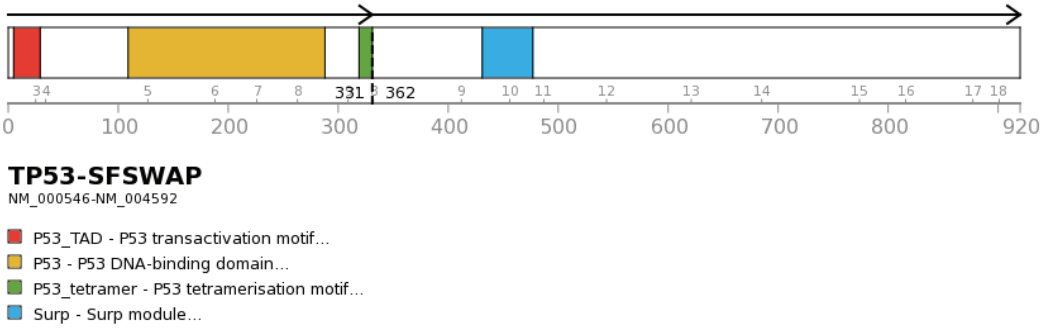

SJOS001127_D1
Log10(intermutation Distance)
Mutation number
SNV Type
C>A
C>G
C>T
T>A
T>C
T>G
#SV
1
10
100
1000

Figure S2 related to Figure 2. Analysis of tumor heterogeneity and kataegis. A) Tumor purity adjusted mutant allele fraction (MAF) for samples analyzed by whole genome sequencing. The actual tumor purity that was used to adjust the MAF is shown as a percentage for each tumor. The number of qualifying SNVs (n) used to plot the density plot is shown for each tumor. **B)** Screenshot of WGS data with individual sequence reads showing kataegis hypermutation on the same DNA strand. **C)** Genomewide distribution of mutation hotspots in osteosarcoma shown as the ratio between observed mutation rate in the window over the genomewide mutation rate within a 3.2 Mb window across the genome for the 34 samples analyzed by WGS. **D)** Kataegis SNVs showed different MAFs among different microclusters although SNVs within a microcluster shared similar MAFs. **E)** Comparison of MAFs of SVs and SNVs in kataegis regions. **F)** The distribution of intermutation distance and distance from a SNV to nearest SV breakpoint in SJOS005, showing majority of mutations in the genome have a intermutation distance around 1 Mb while a small portion of SNVs have an intermutation distance smaller than 10 kb. **G)** Distribution of copy number variation and SNV intermutation distance (log10) which shows higher proportion of closely spaced SNVs (i.e. kategis SNVs) occur in amplified regions. **H**)Distribution of mutant allele fraction of non-kataegis SNVs (left panel) and kataegis SNVs (right panel) in regions of different copy number state in SJOS005. SNVs acquired before amplification show multiple mutant allele fraction (MAF) peaks, depending on whether the mutant allele or the reference allele is amplified while mutations acquired after amplification showed a single MAF corresponding to mutation on a single copy. **I)** Rainfall plot of all 34 genomes in the WGS cohort.

A


34
5
6
7
8
331
362
9
10
11
12
13
14
15
16
17
18
0
100
200
300
400
500
600
700
800
920
TP53-SFSWAP
NM_000546-NM_004592
P53_TAD - P53 transactivation motif...
P53 - P53 DNA-binding domain...
P53_tetramer - P53 tetramerisation motif...
Surp - Surp module...

B

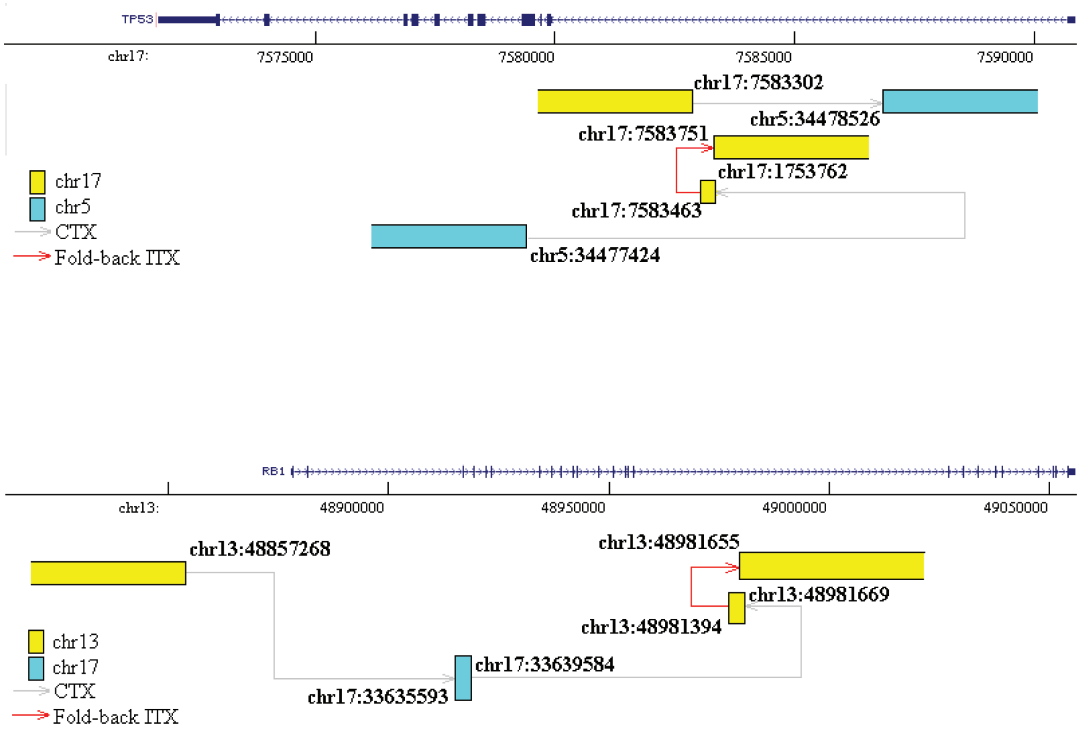

TP53
chr17:
7575000
7580000
7585000
7590000
chr17:7583302
chr5:34478526
chr17:7583751
chr17:1753762
chr17:7583463
chr5:34477424
chr17
chr5
CTX
Fold-back ITX
RB1
chr13:
48900000
48950000
49000000
49050000
chr13:48857268
chr13:48981655
chr13:48981669
chr13:48981394
chr17:33639584
chr17:33635593
chr13
chr17
CTX
Fold-back ITX

Figure S3 related to Figure 3. ***TP53* translocations in osteosarcoma**. **A)**Diagram of the predicted fusion gene generated by the interchromosomal translocation between the *TP53* gene and the *SFSWAP* gene in SJOS007_D. **B)** Diagram of the fold-back translocation in the *TP53* gene in SJOS001_M (upper panel). Diagram of the fold-back translocation in the *RB1* gene in SJOS015_D (lower panel). The translocations are indicated by the red arrows.

Table S4 related to Figure 3. Analysis of p53 mutations in osteosarcoma.

Provided as a separate file.

Table S5 related to Figure 4. Analysis of p53 analysis and clinical features.

Provided as a separate file.

Table S6 related to Figure 5. ATRX analysis in osteosarcoma.

Provided as a separate file.

Table S7 related to Figure 5. Cancer gene mutations in osteosarcoma.

Provided as a separate file.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Tumor Samples

Twenty high-grade intermedullary osteosarcoma samples with matched normal tissue from 19 patients were subjected to the whole genome sequence (WGS) analysis. The samples included 16 untreated primary and 4 metastatic tumors, the latter obtained from 3 patients, of whom two had metastatic disease at presentation and one had tumor recurrences. The tumors occurred in 11 males and 8 females ranging in age from 8-22 years of age (median age, 14 years). H&E slides of the tumors were retrieved from the Pathology Archives at St. Jude Children's Research Hospital for review and the tumors were classified in the following histologic subtypes: osteoblastic (10 tumors), mixed pattern (3 tumors), telangiectatic (3 tumors), fibroblastic (2 tumor), chondroblastic (1 tumor), and small cell (1 tumor). Clinical features of the validation cohort are provided in Suppl. Table 1. For the p53 analysis cohort, 38 samples from 31 additional patients were analyzed for alterations in the p53 pathway. The clinicopathologic data and the results of the molecular genetic assays on validation cohort are provided in Supplemental Table 4.

TP53 Immunostaining

The corresponding formalin-fixed, paraffin-embedded (FFPE) tissue blocks for each specimen were cut at 4 micron thickness. Immunohistochemical staining was performed using an antibody directed against p53 protein (DO-7, DAKO, 1:50) and processed with standard heat-induced epitope retrieval (Ventana CC1) and the Ventana IVIEW detection systems. p53 nuclear staining was scored using a previously published scoring system(Papai et al., 1997) with a minor modification as follows: samples with no staining

cells were scored as negative; samples with <5% p53 immunopositive cells were scored as rare; samples with 5-25% immunopositive cells were scored as 1+; samples with 26-50% immunopositive cells were scored as 2+; samples with >50% immunopositive cells were scored as 3+.

TP53 and MDM2 Fluorescence In Situ Hybridization Studies

The fluorescence in situ hybridization (FISH) probe sets were designed using bacterial artificial chromosome (BAC) clones according to the UCSC Genome Bioinformatics database (http://genome.ucsc.edu). The *TP53* dual-color break-apart FISH assay was developed by using RP11-1081A10 BAC clone, flanking the 3′ end of the gene– labeled with Rhodamine (red fluorochrome)– and the RP11-709J3 BAC clone, flanking the 5′ end of the gene– labeled with AlexaFluor-488 (green fluorochrome). In addition, a *TP53* FISH probe mixture was constructed to enumerate the gene copy using the RP11-89D11 BAC clone, spanning the entire sequence of *TP53* that was labeled with green fluorochrome and the control RP11-64J19 that was labeled with red fluorochrome. FISH for *MDM2* was set up using the RP11-611O2 BAC clone spanning the entire gene (red fluorechrome) and a probe targeting *ATF1* at 12q13.1 as the control probe (green fluorochrome. DNA was isolated from BAC clones (BACPAC Resources, Oakland, CA) according to a modified Qiagen (Valencia, CA) extraction protocol. The probes were labeled by nick translation using a modification of the manufacturer's protocol (Life Technologies, Inc., Carlsbad, CA). FISH analysis was performed on 4 micron-thick FFPE tissue sections using the previously published methods(Bahrami et al., 2012).

Hybridization signals were evaluated in 200 interphase nuclei of each sample. FISH images were captured and processed as previously described(Bahrami et al., 2012).

Telomeres were analyzed in the discovery cohort using 3 different methods. The whole genome sequencing (WGS) data was analyzed for telomere length (described below) for all 20 tumors and matched normal tissue in the discovery cohort. Quantitative PCR (described below) was performed to validate the results from WGS analysis for all 10 tumors in the discovery cohort with *ATRX* mutations and an additional 4 samples with wild type *ATRX* to serve as controls. Matched normal DNA was used as internal control for each patient's sample. Telomere FISH (see below) was performed on all tissue samples in the discovery cohort that had available FFPE material (non-decalcified tumor tissue). All of the samples that were analyzed by telomere FISH were also analyzed for ATRX protein expression by immunohistochemistry (described below).

Whole-Genome Sequencing, RNA-Seq and Exome Seq

Using a paired-end sequencing approach, we sequenced DNA from 20 tumors and their matching germline control DNA with an average of 30× haploid coverage per genome. Single nucleotide variations (SNVs) and insertions/deletions (indels) were identified independently algorithms by Washington University Genome Sequencing Center (WUGSC) and St. Jude Children's Research Hospital (SJCRH) using different approaches. The results generated were then compared and a final candidate SNV and indel list was developed for experimental validation.

At WUGSC, SNVs were found by Somatic Sniper that defines high quality somatic predictions as those sites with a somatic score greater than 40 and an average

mapping quality greater than 40. The predicted SNVs are compared to the most current version of dbSNP(Sherry et al., 2001) (build 129-130). For SNVs, we require both positional and allele match. In addition we also compared the predicted SNVs to SNPs found in CEU and YRI trios as described(Ding et al.). All predicted SNVs were filtered through a SNV false-positive filter developed at the Genome Institute that is based on a standard set of criteria including mapping quality score, average supporting read length, average position of the variant in the read, strand bias and the presence of homopolymer. Indels were called using modified SAMtools(Li et al., 2009) indel-calling algorithm as described(Ding et al.), Pindel(Ye et al., 2009) and GATK(Zerbino and Birney, 2008).

At SJCRH, putative sequence variants including SNVs and indels were initially detected by running the variation detection module of Bambino(Edmonson et al.) using the following three parameters: (1) a high quality threshold for pooled tumor and matching normal bam files (min-quality=20, min-flanking-quality=20, min-alt-allele-count=3, min-minor-frequency=0, broad-min-quality=10, mmf-max-hq-mismatches=4, mmf-min-quality=15, mmf-max-any-mismatches=6; (2) a low quality threshold for pooled tumor and matching normal bam files (min-quality=10, min-flanking-quality=10, min-alt-allele-count=2, min-minor-frequency=0, broad-min-quality=10); and (3) a high tolerance for the number of mismatches for normal bam file alone (min-quality=20, min-flanking-quality=15, min-alt-allele-count=2, min-minor-frequency=0, mmf-max-hq-mismatches=15, mmf-min-quality=15, mmf-max-any-mismatches=20). In addition to Bambino, putative indels were also found by a *de novo* assembly process which construct contigs using unmapped reads and re-map them to the reference genome followed by a Smith-Waterman alignment to detect indels. In this process, unmapped reads include (1)

unmapped reads whose mate are mapped to the genome; (2) reads with indels in CIGAR (Compact Idiosyncratic Gapped Alignment Report) string; (3) reads with at least 4 high-quality (quality value >=20) mismatches; and (4) reads with high-quality (quality value at least 20) soft-clipped bases in the CIGAR string. All putative sequence variants were further assessed to determine their accuracy and somatic original using the processes described below. Velvet(Zerbino and Birney, 2008), BLAT(Kent, 2002) and SIM(Huang et al., 1990) were the three programs used for assembly, mapping, and Smith-Waterman alignment, respectively.

A putative somatic sequence mutation determined by SJCRH process was collected based on the following criteria: (1) the variant site is absent in the normal-only analysis; (2) Fisher's exact test *P* value indicates that the number of reads harboring non-reference allele is significantly higher in tumor; (3) the non-reference allele frequency in normal is <=5%; and (4) mutant alleles present in both orientations. Higher *P* value and absence of non-reference allele in normal is required for a variant to be considered somatic if it matches dbSNP build 130 or is located in an unmappable region (determined by recurrence of 75mers across the reference genome) or is inside a polynucleotide repeat. Substitution variants are classified into four categories based on combination of their *P* value and sequence quality scores: High quality, high *P* value; high quality, low *P* value; low quality, high *P* value; low quality, low *P* value. *P* value refers to the *P* value of Fisher's exact test comparing the distribution of the alternative allele in tumor and normal. High *P* value, $P<0.05$; low *P* value, $0.05<P<0.10$. A final review process re-maps and re-aligns the reads harboring the non-reference allele to the reference genome to filter potential false positive calls introduced by mapping in repetitive regions and

alignment artifacts. For putative somatic indels, the review process re-aligns all reads in tumor and normal at the indel site to a mutant allele template sequence constructed by substituting the wild-type allele with the indel. Presence of reads in normal that cover the mutant allele is considered a germline variant. Structural variations including the 5 deletions in *ATRX* were detected using the CREST algorithm (Wang et al., 2011) and CONSERTING algorithm. The data have been deposited in EBI with accession number: EGAS00001000263.

Paired-end reads from mRNA-seq were aligned to the following 4 database files using BWA (0.5.5) aligner (4): (i) human NCBI Build 37 reference sequence, (ii) RefSeq, (iii) a sequence file that represents all possible combinations of non-sequential pairs in RefSeq exons, and (iv) AceView flat file downloaded from UCSC and representing transcripts constructed from human EST. The final BAM file was constructed by selecting the best alignment in the four databases. SV detection was carried out using CREST (1) and deFuse (5) as well as a novel algorithm that searched for the predicted junction breakpoints from detected SVs in matching WGS samples.

For exome sequencing, OS DNA libraries were prepared from 1 ug of WGA material from matched samples using the Illumnia TruSeq DNA library prep kit following the recommended manufacturer's protocol. Libraries were analyzed on an Agilent Bioanalyzer to inspect quality of each library construction. Germline and diagnostic library samples were independently pooled and applied for exome capture using the Illumina TruSeq Exome Enrichment kit as described by the manufacturer. Captured libraries were then clustered on the Illumina c-bot and were sequenced on an

Illumina HiSeq 2000 platform with 100 base pair end multiplexed reads at an equivalent of 3 samples per lane.

We used cghMCR (an R implementation of a modified version of GISTIC analysis(Aguirre et al., 2004)) to find common regions of copy number alterations. To identify genes of significant DNA copy number alterations, we defined the genes with Segments Of Gain Or Loss (SGOL) scores above the 3 standard deviations of the mean SGOL scores of all 'gains' scores as significantly amplified genes. The genes with SGOL scores below the 3 standard deviations of the mean SGOL scores of all 'losses' scores were sselected as significantly deleted genes.

Sequence Validation

For enrichment of the regions containing putative alterations, genomic coordinates of the putative WGS targets were used to order Nimbelgen Seqcap EZ solution bait sets (Roche). The library construction and target enrichment was performed per manufacturer's instructions using repli-G (Qiagen) whole genome amplified DNA. Enriched targets were sequenced on the Illumina platform using paired end 100 cycle sequencing. The resulting data was converted to FASTQ files using CASAVA 1.8.2 (Illumina) and mapped with BWA prior to pipeline analysis.

Statistical Methods

Kaplan-Meier method was used to estimate the overall survival and event-free survival curves. Log-rank test was performed to test the significant difference of survival curves

between *TP53* missense mutation group and the *TP53* truncating mutation group in SAS version 9.2.

We used the MuSiC software(Dees et al., 2012) to identify significantly mutated genes with point mutations. For significantly mutated genes with SVs, the "background" base-level mutation rate for SVs in each tumor under the null hypothesis that SV breakpoints were distributed randomly within the genome and the number of tumors mutated by SVs for a specific gene follows the Poisson binomial distribution under the null hypothesis. The significance level was estimated from the Le Cam's theorem.

Telomere Analysis

Telomere length was predicted in silico by counting the number of next-generation sequencing reads containing the telomeric-repeat sequence TTAGGG (Castle et al., 2010). The resulting number of reads was normalized to the average genomic coverage, and the difference in diagnostic and germline telomeric sizes was calculated. Telomere length was validated in vitro in NBs expressing an *ATRX* aberration as described previously (Cawthon, 2002; O'Callaghan et al., 2008). Briefly 15-20ng of diagnostic and germline WGA amplified DNA was subject to qPCR using two sets of primers in separate reactions, one to amplify telomeric sequence and one to amplify a common gene; *36B4 (RPLP0)*. Ct values obtained were compared to those of two standard curves, a telomeric standard curve performed on known quantities of a telomeric 84mer and one using an oligomer of *36B4 (RPLP0)*. All reactions were performed in triplicate with both tumor and germline DNA and both assays on the same plate. All reactions were carried out using Brilliant III Ultra-Fast SYBR Green master mix (Agilent) on a Stratagene

Mx3000 thermal cycler using a melting temperature of 60ºC. This allowed us to determine the telomere length in Kb per diploid genome. The forward primer for telomere analysis was:

5'- CGGTTTGTTTGGGTTTGGGTTTGGGTTTGGGTTTGGGTT-3'

The reverse primer for telomere analysis was:

5'-GGCTTGCCTTACCCTTACCCTTACCCTTACCCTTACCC-3'

The forward primer for the internal control *36B4 (RPL0)* gene was:

5'- CAGCAAGTGGGAAGGTGTAATCC-3'

The reverse primer for the internal control *36B4 (RPL0)* gene was:

5'- CCCATTCTATCATCAACGGGTACAA-3'

The standard used to generate the standard curve for telomeres was:

5'-$(TTAGGG)_{14}$-3'

The standard used to generate the standard curve for the internal control *36B4 (RPL0)* was:

5'- CAGCAAGTGGGAAGGTGTAATCCGTCTCCACAGACAAGGCCAGGACTCG TTTGTACCCGTTGATGATAGAATGGG-3'

ATRX Immunohistochemistry

Formalin-fixed, paraffin-embedded tissues were cut into 4-μm-thick sections and immunostained with a polyclonal antibody against ATRX (1:600; Sigma-Aldrich) by using heat-induced epitope retrieval and Leica Polymer Refine Detection Kit (Leica Microsystems) on a Leica Bond system after 15-minute antibody incubation.

Telomere FISH

Interphase FISH was performed on 4-μm-thick, formalin-fixed, paraffin-embedded tissue sections. The Cy3-labeled TelG probe (PNAbio) was co-denatured with the target cells on a hotplate at 90 °C for 12 minutes. The slides were incubated for 48 hours at 37 °C and then washed in 4 M Urea/2× SSC at 45 °C for 5 minutes. Nuclei were counterstained with DAPI (200 ng/mL) (Vector Labs).

Tumor Purity Estimations

For germline heterogeneous SNPs, loss of heterozygosity (LOH) measures the absolute difference between the mutant allele fraction in tumor and that in germline sample (0.5). LOH is the result of copy number alterations and/or copy neutral-LOH in tumor cells. Compared to copy number gains (a single copy gain in 100% tumor results in a LOH value of 0.167), regions with copy number loss showed stronger LOH (a single copy loss in 100% tumor result in a LOH value of 0.5). Consequently, we used LOH signals in copy neutral or heterozygous copy number loss regions (CNA value between [-1, 0]) to estimate tumor purity for all WGS samples. Briefly, a single copy loss in x% tumor cells resulted in an estimated CNA value of $-\frac{x}{100}$ and a LOH value of $\frac{x}{400-2x}$. Assuming the remaining LOH signal came from CN-LOH (CN-LOH in x% tumor cell resulted in a LOH value of $\frac{x}{200}$), the tumor content in a region could be estimated as the sum of the fraction with copy number loss and the fraction with CN-LOH by: $-CNA + 2 * \left(LOH - \frac{-CNA}{4-2CNA}\right)$. Using tumor content estimates from various regions within the genome, we performed an unsupervised clustering analysis using the *mclust* package (version 3.4.8) in R (version 2.11.1). The tumor purity of the sample was defined as the highest cluster center value among all clusters.

Purity Adjusted Mutant Allele Fraction (MAF) Estimation

MAF for validated SNVs was estimated as $\frac{\#Mutant\ reads}{(\#Total\ reads)\times(tumor\ purity)}$ using deep sequencing data. The frequency of SV was determined by a process of re-mapping all reads at breakpoints to both SV and non-SV templates using a BWA Smith-Waterman based approach. To do this, we use the assembled consensus sequence from CREST result as SV template. From comparison, a pool of non-SV templates were constructed by including: 1, directly pull out the flanking sequences of 100 bp of each side of the breakpoint from reference genome (GRCh37-Lite); 2, assemble non-SV reads around the breakpoint from the bam file, where non-SV reads were defined by: any non-duplicate, non-softclipped, reads that contains at least 10 bases mapped on each side of a breakpoint, and requiring for at least one of two sides of that breakpoint, all bases are mapped in the read within a 10 bp continuous window immediately next to the breakpoint. We then extracted all reads at both SV break points, together with any unmapped or partially mapped (soft-clipped) reads within 4 kb of the breakpoints, and perform a pair-wise mapping and comparison for the SV and each of the non-SV templates to determine the status of individual read. Reads only covering the breakpoint in the SV template, but not in the non-SV template, are considered as SV supporting reads, and verse versa. If there are any reads covering breakpoint in both SV and normal template, we calculated a local alignment score within a 10-bp window of the breakpoint from SV and normal templates, and chose the template with higher score. In the end, the statuses of every read from all pair-wise comparisons were summarized to generate a consensus status. Any reads with conflicting statuses, i.e., called as SV in one run and

non-SV in another run will be considered as "unknown". The SV mutant allele frequency was calculate by the ratio of number of SV reads to the total number of SV, non-SV, and unknown reads.

Tumor Heterogeneity Estimation

We used all validated autosomal SNVs satisfying the following criteria in heterogeneity analysis:

1) In copy neutral region (Log2ration between (-0.1, 0.1) in CNV analysis).
2) Not in regions with LOH (LOH value < 0.12 + min (0.08 purity*0.1).
3) With MAF > 0.05 or mutant allele count > 2.

We drew the kernel density estimate plot for MAFs of the qualifying SNVs using the *density* function in the *stat* package in R. For samples with at least 50 qualifying SNVs, we also estimated the number of significant peaks and the relative MAF component for each peak (peaks with less than 5 SNVs, peaks with less than 1% SNVs, and peaks with excessive variance were ignored). A sample with heterogeneity shows density peaks at a MAF smaller than 0.5 (the expected MAF assuming heterogeneous SNVs).

Kataegis Analysis

Kataegis analysis was performed on all validated Tier1-3 SNVs and SV breakpoints for each sample. The intermutation distance for a SNV was calculated as the distance to its nearest neighbor. For each SNV, its distance to the nearest validated SV breakpoint was also calculated. We defined microclusters of kataegis as clusters that contain at least 5 consecutive SNVs with inter-variant distance less than 10 kb. Mutant allele frequency

(MAF) was estimated for SNVs with at least 20X coverage in tumor BAMs based on deep sequencing of custom capture validation.

We derived copy number of SNVs from the CONSERTING analysis. For each SNV, the CNV segment covering the SNV was identified and the corresponding CN was estimated after tumor purity adjustment and rounded to the nearest integer.

Statistical evaluation of chromothripsis in OS tumors analyzed by WGS

Chromothripsis was described as localized chromosome shattering and repair occurs in a single event. The initial criterion is oscillation between restricted CNV states (Stephens *et al*. 2011(Stephens et al., 2011)), which were found in 4 OS tumors in this study. Most recently, Korbel and Campbell (Korbel and Campbell, 2013) proposed four potential criteria for assessing chromothripsis: 1) clustering of breakpoints; 2) randomness of DNA fragment joins; 3) randomness of DNA fragment order; and 4) ability to walk the derivative chromosome. Since randomness of DNA fragment order (Criterion 3) was not entirely valid even in Korbel and Campbell's own analysis, we decided not to evaluate this feature. For the 4 tumors in Supplementary Table 5, we performed Bartlett's goodness-of-fit test for exponential distribution to assess whether the distribution of SV breakpoints in each tumor departs from the null hypothesis of random distribution. A significant departure from random distribution supports clustering of SV breakpoints. To evaluate whether there is any bias in the DNA fragment joints categorized by the SV types (*i.e*. deletion, tandem duplication, head-to-head re-arrangements and tail-to-tail re-

arrangements), we applied goodness-of-fit test separately for inter- and intra-chromosomal events with a minimum of 5 SVs. A significant *p* value suggests biased fragment joins, which would *not* support chromothripsis. When both inter- and intra-chromosomal data are available, we reported the lower *p* value to represent a more conservative assessment of the random distribution for DNA fragment joints.

The significant chromothripsis regions were chromosome 14 in SJOS002_D (*p=2.09E-09)*, chromosome 17 in SJOS003_D (*p=9.65E-05)*, chromosome 6 in SJOS005_D (*p=1.75E-90)* and chromosome 13 in SJOS010_M (*p=2.21E-35)*.

REFERENCES

Aguirre, A. J., Brennan, C., Bailey, G., Sinha, R., Feng, B., Leo, C., Zhang, Y., Zhang, J., Gans, J. D., Bardeesy, N., *et al.* (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. Proceedings of the National Academy of Sciences of the United States of America *101*, 9067-9072.

Bahrami, A., Dalton, J. D., Krane, J. F., and Fletcher, C. D. (2012). A subset of cutaneous and soft tissue mixed tumors are genetically linked to their salivary gland counterpart. Genes, chromosomes & cancer *51*, 140-148.

Castle, J. C., Biery, M., Bouzek, H., Xie, T., Chen, R., Misura, K., Jackson, S., Armour, C. D., Johnson, J. M., Rohl, C. A., and Raymond, C. K. (2010). DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. BMC genomics *11*, 244.

Cawthon, R. M. (2002). Telomere measurement by quantitative PCR. Nucleic acids research *30*, e47.

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., *et al.* (2012). MuSiC: identifying mutational significance in cancer genomes. Genome research *22*, 1589-1598.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999-1005.

Edmonson, M. N., Zhang, J., Yan, C., Finney, R. P., Meerzaman, D. M., and Buetow, K. H. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics *27*, 865-866.

Huang, X. Q., Hardison, R. C., and Miller, W. (1990). A space-efficient algorithm for local similarities. Comput Appl Biosci *6*, 373-381.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Korbel, J. O., and Campbell, P. J. (2013). Criteria for inference of chromothripsis in cancer genomes. Cell *152*, 1226-1236.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
O'Callaghan, N., Dhillon, V., Thomas, P., and Fenech, M. (2008). A quantitative real-time PCR method for absolute telomere length. BioTechniques *44*, 807-809.
Papai, Z., Feja, C. N., Hanna, E. N., Sztan, M., Olah, E., and Szendroi, M. (1997). P53 Overexpression as an Indicator of Overall Survival and Response to Treatment in Osteosarcomas. Pathology oncology research : POR *3*, 15-19.
Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308-311.
Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A*., et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27-40.
Wang, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., Rusch, M. C., Chen, K., Harris, C. C., Ding, L*., et al.* (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nature methods *8*, 652-654.
Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865-2871.
Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821-829.