Supplementary Materials

Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny

Eugene Urrutia[1], Hao Chen[2], Zilu Zhou[3], Nancy R Zhang[4], Yuchao Jiang[1,5,*]

[1] Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA, [2] Department of Statistics, University of California, Davis, Davis, CA 95616, USA, [3] Genomics and Computational Biology Graduate Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, [4] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA, [5] Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA.

* To whom correspondence should be addressed. Email: yuchaoj@email.unc.edu.

Contents

Availability

MARATHON's GitHub page:
https://github.com/yuchaojiang/MARATHON.

MARATHON's R notebook:
https://rawgit.com/yuchaojiang/MARATHON/master/notebook/MARATHON.html.

MARATHON'S R script:
https://github.com/yuchaojiang/MARATHON/blob/master/notebook/MARATHON.Rmd.

MARATHON depends on CODEX, CODEX2, iCNV, FALCON, FALCON-X, and Canopy, which are all publicly available R packages available at CRAN, Bioconductor, and/or GitHub.

CODEX: http://bioconductor.org/packages/CODEX/
Published in Nucleic Acids Research (Jiang, et al., 2015)

CODEX2: https://github.com/yuchaojiang/CODEX2/
Available on bioRxiv (Jiang, et al., 2017)
https://www.biorxiv.org/content/early/2017/11/13/211698

iCNV: https://github.com/zhouzilu/iCNV
Available on bioRxiv (Zhou, et al., 2017)
https://www.biorxiv.org/content/early/2017/11/30/172700

FALCON: https://CRAN.R-project.org/package=falcon
Published in PLOS Computational Biology (Chen, et al., 2015)

FALCON-X: https://CRAN.R-project.org/package=falconx
Published in Annals of Applied Statistics (Chen, et al., 2017)

Canopy: https://CRAN.R-project.org/package=Canopy
Published in Proceedings of the National Academy of Sciences (Jiang, et al., 2016)

Supplementary Results

Tumor phylogeny analysis (Pipeline 5: FALCON → Canopy)

We first demonstrate a cancer phylogenetic study, where we adopt a WGS dataset of normal, primary tumor, and relapse genome of a neuroblastoma patient PASGAP (Eleveld, et al., 2015). Since this is WGS, and not WES data, we can simply apply FALCON to profile ASCN. If this were WES data, one can substantially improve accuracy by first applying CODEX2 to perform normalization, followed by ASCN analysis using FALCON-X.

On this data set, the ASCN estimated by FALCON are shown to be consistent with the output from Sequenza (Favero, et al., 2015) (Supplementary Figure S1-S2). Sequenza simultaneously estimates purity and ploidy of the tumor, and outputs integer absolute copy number of each allele assuming all copy number changes are clonal events. The caveat is that purity and ploidy are often not identifiable from the data (Supplementary Figure S3) and estimation of integer-valued ASCN requires the assumption of clonal copy number alternations, which is usually not known prior to knowing the phylogenetic history. FALCON has less stringent assumptions and returns fractional ASCNs with standard deviations estimated by a bootstrap-based method (Supplementary Table S1). We focus on chr4p loss, chr6p gain, chr14 LOH, and chr20p loss, of which the primary and relapse have distinct profiles, to reconstruct the tumor phylogeny.

Somatic SNVs are called by GATK and MuTect and further annotated by ANNOVAR. Stringent quality control procedures are adopted to remove possible germline mutations, low-quality indels, variants with missing genotypes, variants within segmental duplication regions, and variants with low coverage, etc. We also focus on variants that are annotated as deleterious by at least one scoring metric, resulting in 32 high-confidence SNVs – 7 are unique to the primary, 21 are unique to the relapse, and 4 are shared (Supplementary Table S2, Supplementary Figure S4). Notably, a relapse-specific SNV in the gene *CORIN* lies within the heterozygous deletion in chr4 and has a variant allele frequency of 68.8% in the relapse.

The SNVs, as well as the ASCNs, are used as input for Canopy, which returns a most likely tree with four leaves, including one leaf representing the normal cells (Supplementary Figure S4). The purity estimates are largely concordant with those returned by Sequenza, which selects posterior modes with ploidy equal to two (Supplementary Figure S3). The heterozygous deletion in chr4p occurs on the first tree branch, followed by the point mutation in gene *CORIN* on the remaining allele, together with the other relapse-specific SNVs. These SNVs exist in clone 2 and 3, which have cellular fractions of 41.5% and 30.2%, respectively. Chr6p duplication separates clone 3 from clone 2, which are merged if the tree is only allowed to have three leaves.

Tumor allele-specific copy number inference (Pipeline 4: CODEX2 → FALCON-X)

Using Falcon-X, we also analyze WES sequencing data from 39 breast and ovarian tumors with matched normal blood DNA. All samples have germline *BRCA*1/2 mutations (gBRCA1/2). An in-depth study of these samples is described in Maxwell et al. (Maxwell, et al., 2017), where the goal is to delineate molecular mechanisms of tumorigenesis in gBRCA1/2 carriers and to identify potentially druggable alterations in these tumors. Whole exome sequencing on these samples was performed using the Agilent All-Exon Kit. Tumors were sequenced by Illumina Hi-Seq 2000 to an average depth of 141X and blood DNA to an average mean depth of 155X. The sequenced reads were aligned to the hg19 genome assembly using the Burrows-Wheeler Aligner (BWA) for short-read alignment. The aligned data was analyzed as described in Figure 1 using GATK (DePristo, et

al., 2011), CODEX2 (Jiang, et al., 2017), and FALCON-X (Chen, et al., 2017). Specifically, inherited heterozygous sites were called in the matched normal samples using GATK, the position-specific total coverage biases were estimated by CODEX2, and allele-specific copy number was finally estimated through the FALCON-X model and algorithm.

Allele-specific copy number estimates can be validated through procedures such as digital droplet PCR or targeted sequencing, both of which are laborious procedures that are usually only applied to a small number of events. It is too costly to apply such validation techniques on the genome scale, and so, to assess the quality of FALCON-X estimates, we compare our analysis of the 39 breast cancer samples to an existing genotyping-array-based analysis of 47 gBRCA1/2 breast tumors from The Cancer Genome Atlas Project (TCGA) (Cancer Genome Atlas, 2012). Since analysis methods for genotyping arrays are now more mature than those for high-throughput sequencing data, and since TCGA applied rigorous quality control to their data sets, we expect that high-level trends observed in the TCGA samples should be reproduced in our breast cancer cohort. Although no two cancer patients have the same chromosome copy number profile, it has been shown that breast cancer patients with gBRCA1/2 mutations, and similarly gBRCA1/2 ovarian cancer patients, often share recurrent gain and loss regions. We adopt that most of these recurrent CNAs have been seen in the TCGA cohort and we expect to observe similar recurrent gains and losses between the TCGA gBRCA1/2 breast cancer samples and our Basser gBRCA1/2 samples. Supplementary Figure S6 shows the frequency of detected gain and loss at each genome position for the TCGA gBRCA1/2 breast cancers as well as for the Basser gBRCA1/2 samples analyzed by Falcon-X and by Falcon. For each plot, blue bars in the "positive" direction show the proportion of the samples with a detected gain at the given position, and red bars in the "negative" direction show this proportion for losses. Since copy number changes are scattered somewhat randomly in the genomes of all gBRCA1/2 tumors due to genome instability, almost all positions are marked as gained or lost in at least some of the patients. Yet, the Falcon-X results clearly indicate that there are genome regions that are more frequently altered than others, such as loss of 8p and 17p and gain of 3q, 8q and 17q. This agrees with the recurrent regions reported in the literature on gBRCA1/2 breast tumors. Note that the recurrent regions found by Falcon-X are more similar to those found by TCGA, as compared to the Falcon results. Falcon analysis detects much more copy number events, as seen by the elevated occurrence of both gains and losses at all genome positions across the cohort. Against this uniformly elevated background of detections, Falcon results do not show marked evidence for recurrence at the known positions reported in the literature, which are found by Falcon-X. We believe many of the Falcon detections are false positives caused by the biases inherent in WES data.

Supplementary Figure S6 does not explicitly show the frequency of copy-neutral loss-of-heterozygosity (LOH) events, where one of the parental alleles have been lost and replaced by a duplication of the allele from the other parent. Supplementary Figure S7, which plots the frequency of copy-neutral LOH events along the genome, shows that copy-neutral LOH events are frequent in the Basser gBRCA1/2 cancer data. These events would not have been detected if we only estimate total copy number. Using Falcon-X, we identified copy-neutral LOH that helped us better understand the initiation mechanism of BRCA1/2 tumors. These events are described and analyzed in Maxwell et al. (Maxwell, et al., 2017).

Total copy number analysis in tumor (Pipeline 2: CODEX2)

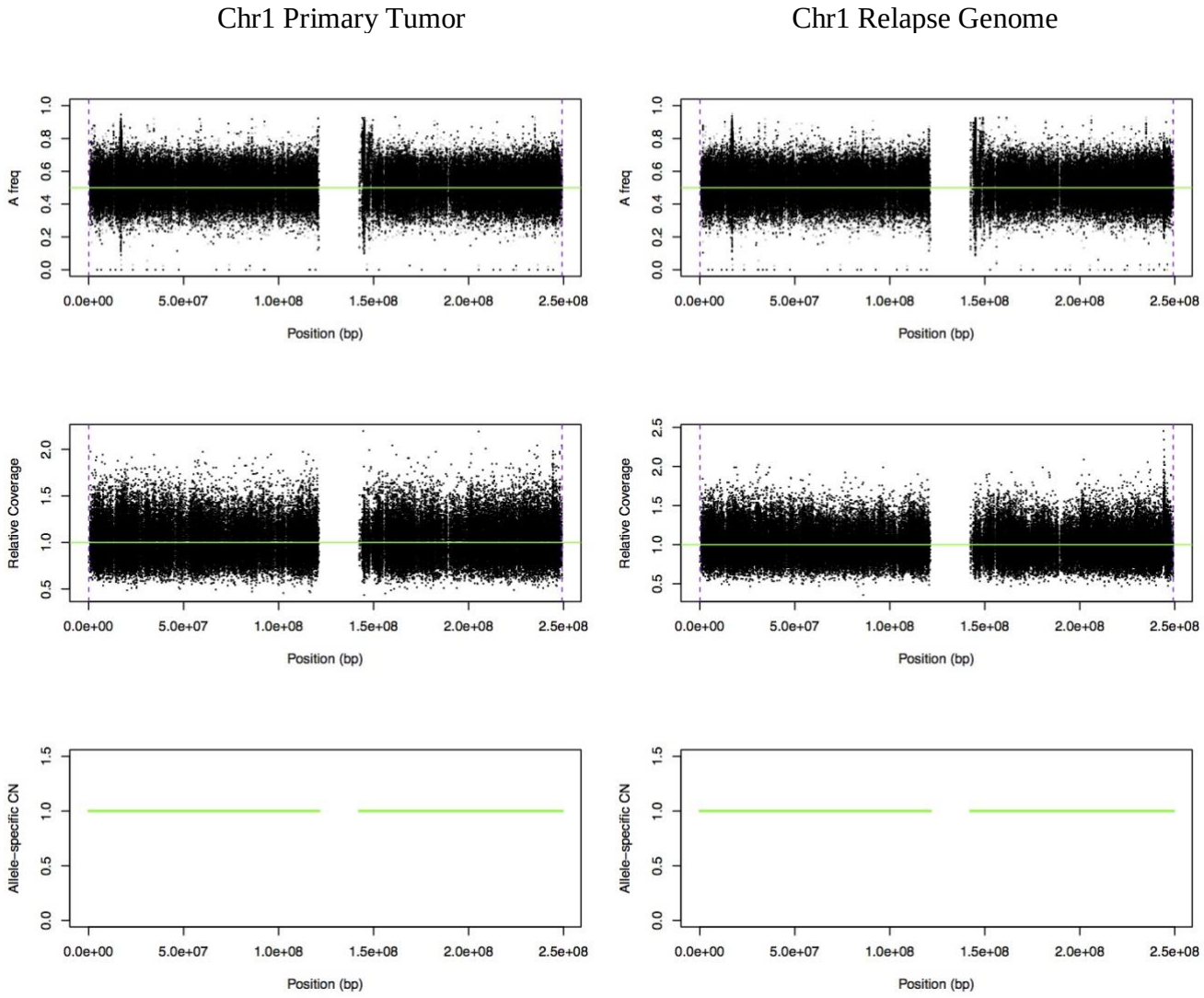
We further applied CODEX2 to a cohort study of melanoma from Garman et al. (Garman, et al., 2017) including 334 cases (untreated human melanoma cell lines, patient-derived xenografts, and

tumors) and 16 controls. Samples were sequenced on a custom capture panel of 108 genes previously implicated in tumorigenesis of melanoma. For almost all tumor suppressor genes, the entire gene region (exons and introns) were sequenced to facilitate CNV calling; for oncogenes, only exons were sequenced. For cases where the full gene is captured and sequenced, we separated the gene region into consecutive windows of 500 base pairs. This resulted in a panel of 1398 targets across 350 samples.

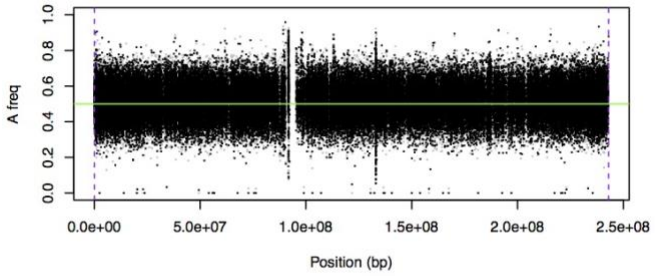
We applied CODEX2 to this data set and compare to CODEX. The number of Poisson latent factors in the background model is determined by the Bayesian information criterion (BIC) for both programs. The use of negative controls in estimating the background model allowed CODEX2 to be more robust to model tuning. For CODEX2, the number of latent factors had minimal effect on normalization and more generally on CNV detection, as only the normal samples were used to estimate the bias coefficient for each exon (Supplementary Table S3). In comparison, for CODEX, the number of CNV events decreased as the number of latent factors increased (Supplementary Table S3). Since the 108 genes are sparsely scattered across the genome, segmentation is performed within each gene separately. Furthermore, due to clonal heterogeneity and normal cell contamination, copy numbers may not be integers, and are assumed to be continuous and fractional to represent attenuated mean estimates of the genome mixtures. We categorize a CNA event to be high gain, gain, diploid, one-copy deletion, and two-copy deletion, if the profiled copy number is above 3.3, between 2.3 and 3.3, between 1.7 and 2.3, between 0.7 and 1.7, and below 0.7, respectively. Supplementary Figure S8 shows the heatmaps of the segmentation results by CODEX and CODEX2. Each row corresponds to a sample, with the first 16 samples towards the bottom corresponding to the normal controls; each column corresponds to a target in the gene panel. In melanoma, somatic deletions of tumor suppressors (e.g., *PTEN*) and duplications of oncogenes (e.g., *BRAF*) are known to have high incidence rates (Cancer Genome Atlas, 2015). From visual inspection of the heatmaps in Supplementary Figure S8, we see that CODEX2 successfully retains these expected recurrent deletions and duplications, while CODEX, which does not make use of the negative control samples in fitting the background model, misinterprets the recurrent signals as a background latent factor.

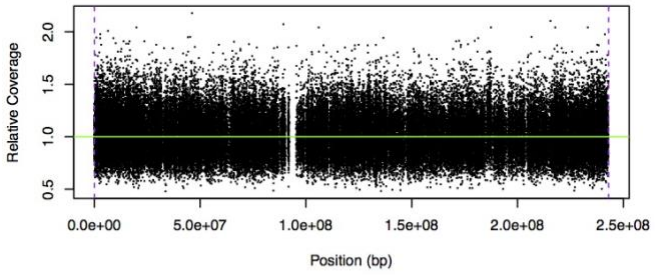
To rigorously evaluate CODEX2's accuracy on this data, we compared the frequencies of the profiled gains and losses, that is, the proportion of samples in which a call is made, with frequencies from an independent melanoma patient cohort in TCGA (Cancer Genome Atlas, 2015). Specifically, for each gene target, we plotted in Supplementary Figure S9 the proportion of samples carrying a deletion (or duplication) in TCGA, versus this corresponding proportion in our current data set. CODEX2 achieves much higher concordance with TCGA results, with overall correlation across genes reaching 0.842 for deletions and 0.853 for gains, as compared to CODEX (correlation = 0.52 for deletions and 0.049 for gains). CODEX2 detects in these cell line samples a higher frequency of *BRAF* amplification and *CDKN2A* loss, as compared to the frozen-tissue derived TCGA results, which is not surprising due to the relative *in vitro* growth advantage of cells carrying these mutations. Based on the results by CODEX2, Garman et al. (Garman, et al., 2017) further separated the cohort based on cancer subtypes and clinicopathological characteristics (responses to targeted and/or immunotherapy) and investigated the differences in mutational profiles between groups. The accurate profiling of CNVs in this data set enables unbiased downstream analysis.

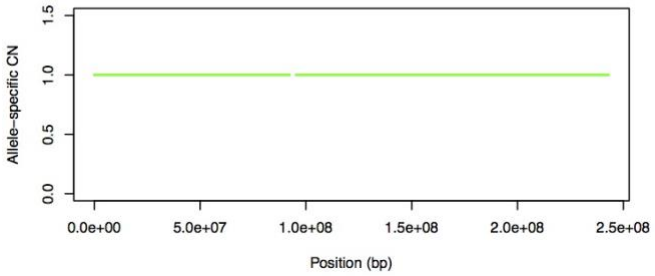
Supplementary Figure S1. Genome-wide segmentation results returned by FALCON (Chen, et al., 2015). Fractional ASCNs are used as input for Canopy (Jiang, et al., 2016) to infer tumor phylogeny. For deconvolution, we focus on chr4p loss, chr6p gain, chr14 LOH, and chr20p loss, which show distinct profiles between the primary and relapse.



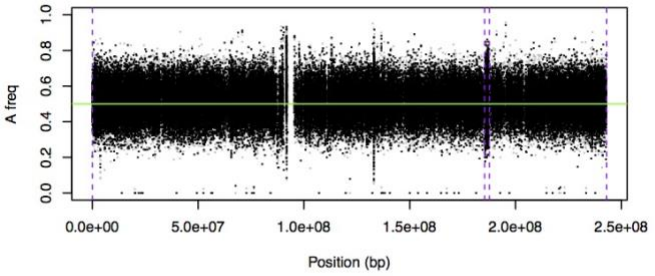
Chr2 Primary Tumor

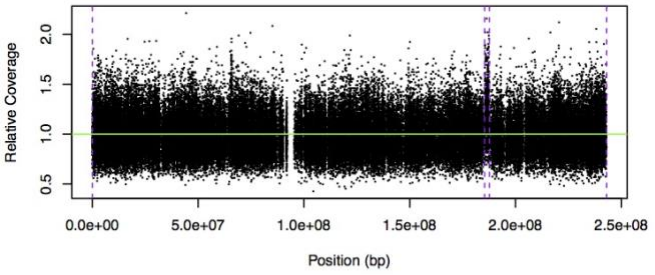


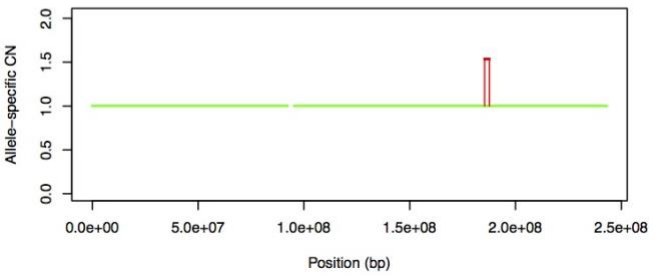




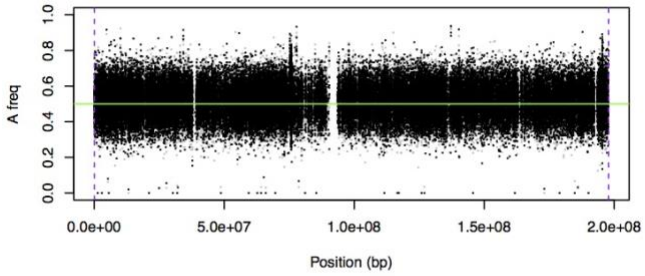
Chr2 Relapse Genome

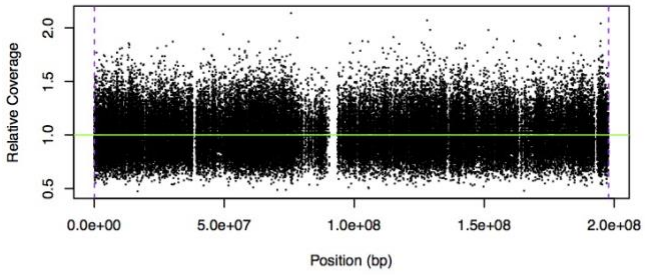


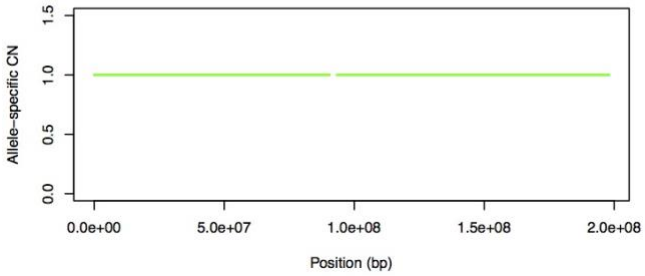




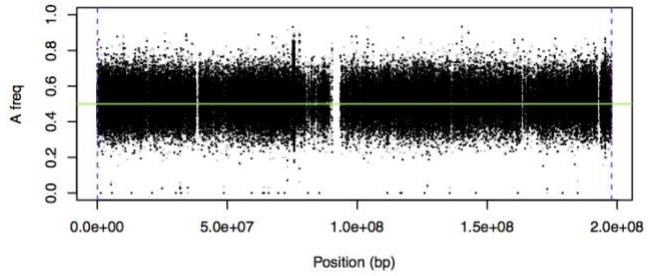
Chr3 Primary Tumor

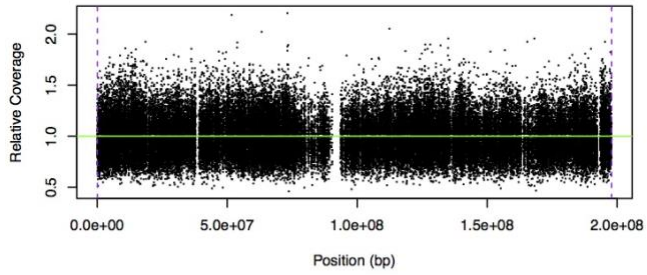


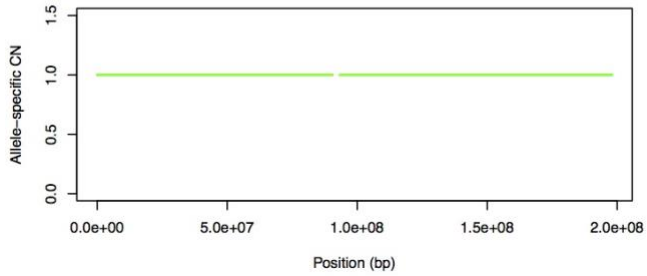




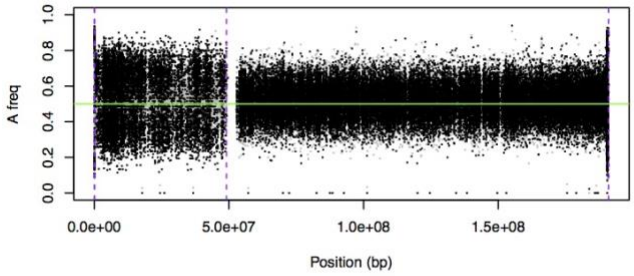
Chr3 Relapse Genome

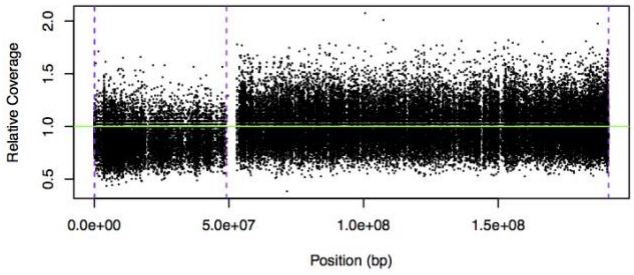


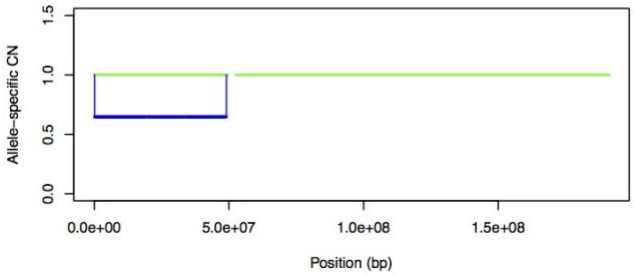




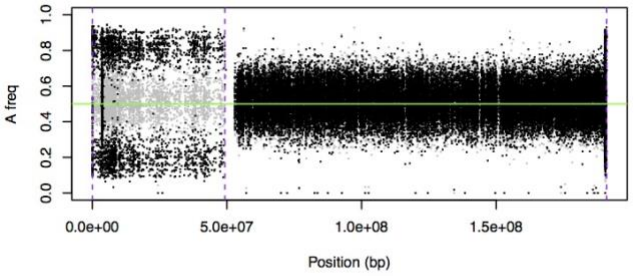
Chr4 Primary Tumor

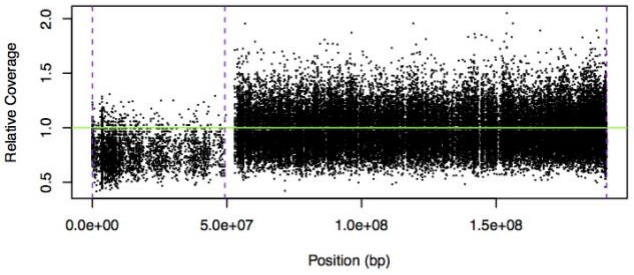


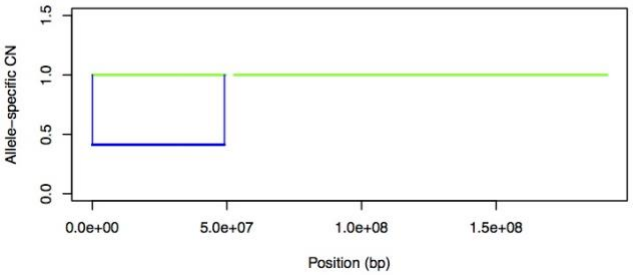




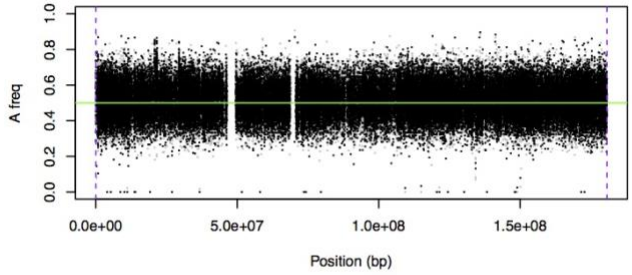
Chr4 Relapse Genome

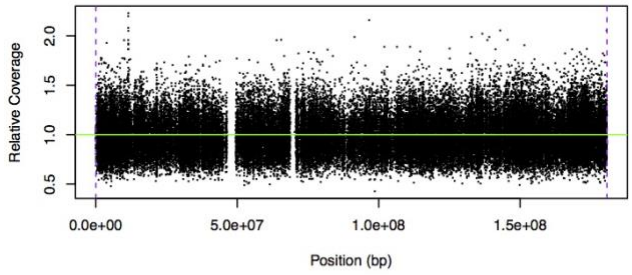


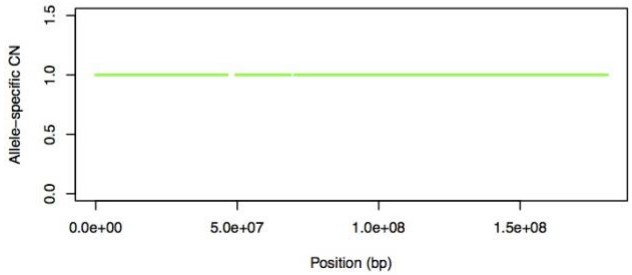




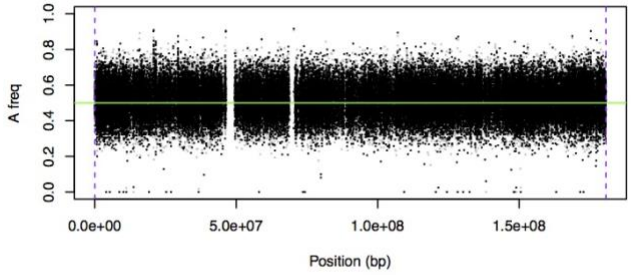
Chr5 Primary Tumor

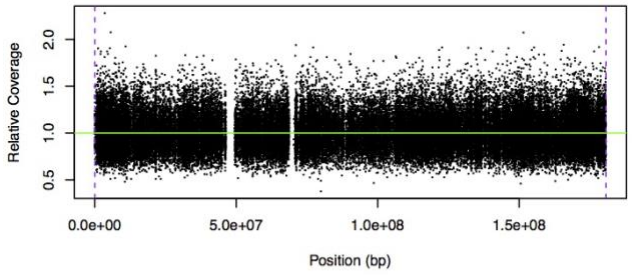


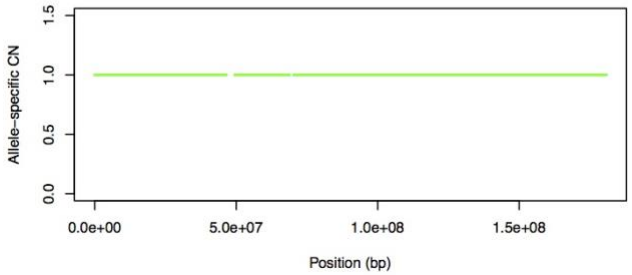




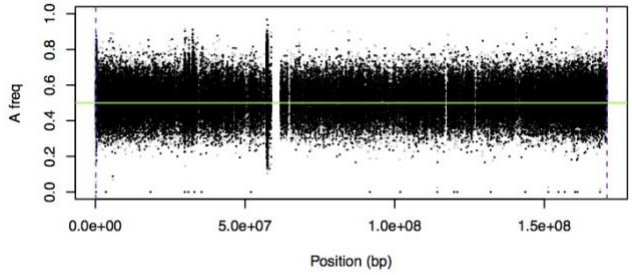
Chr5 Relapse Genome

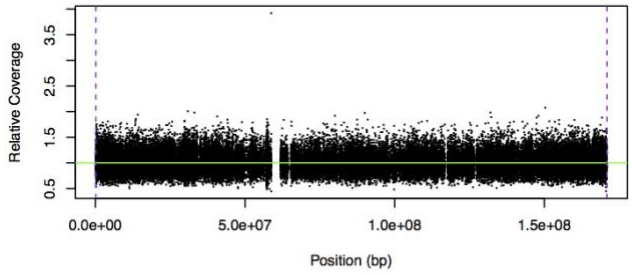


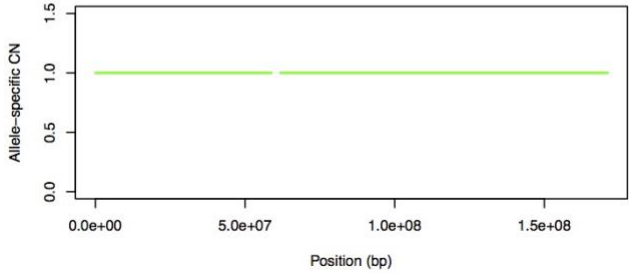




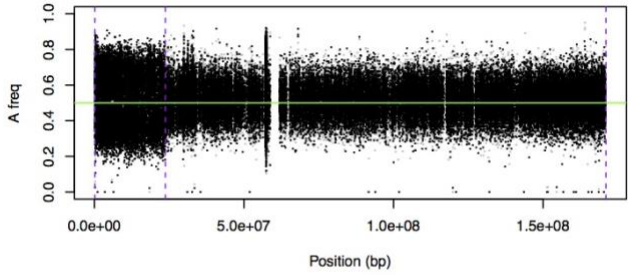
Chr6 Primary Tumor

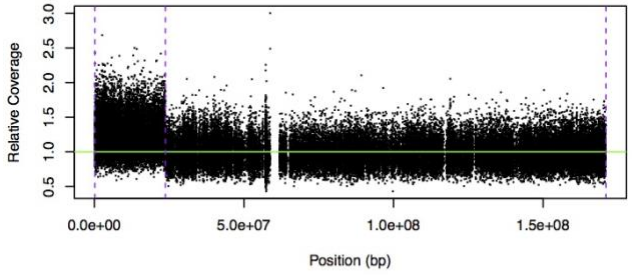


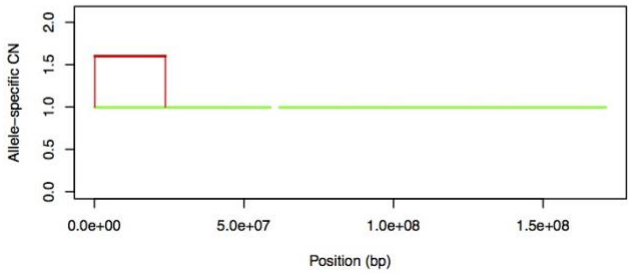




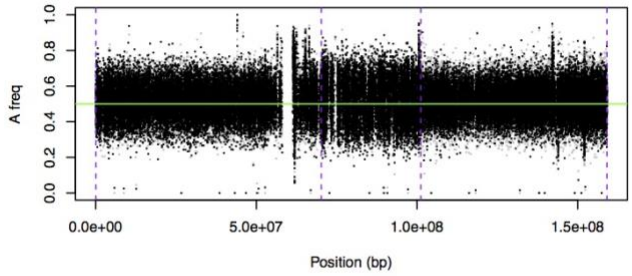
Chr6 Relapse Genome

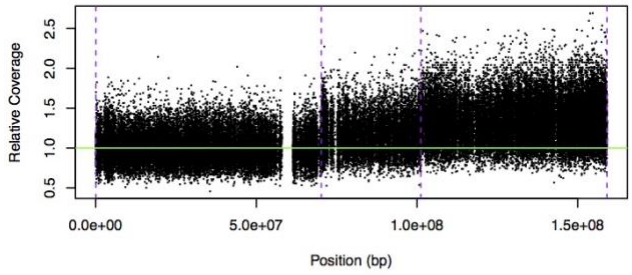


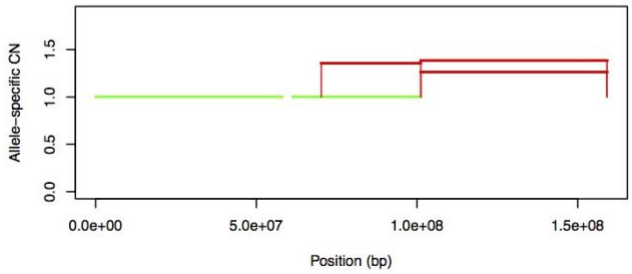




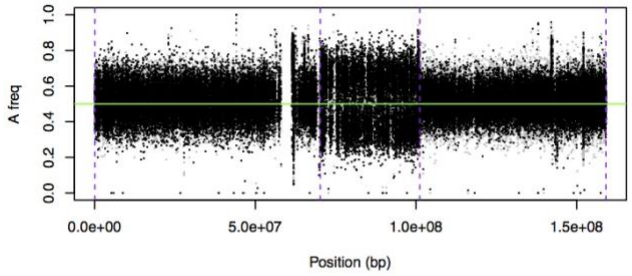
Chr7 Primary Tumor

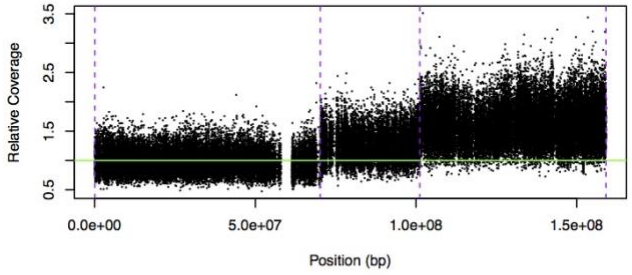


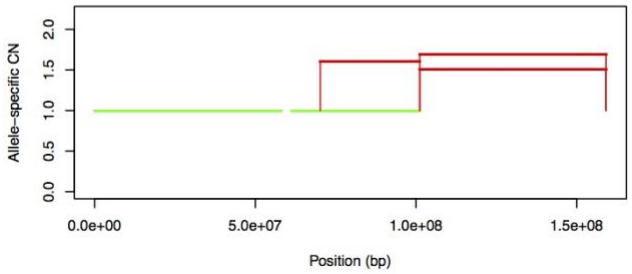




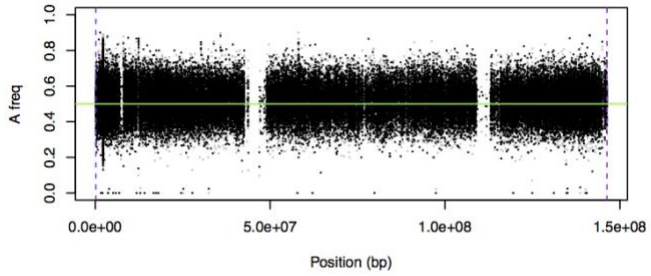
Chr7 Relapse Genome

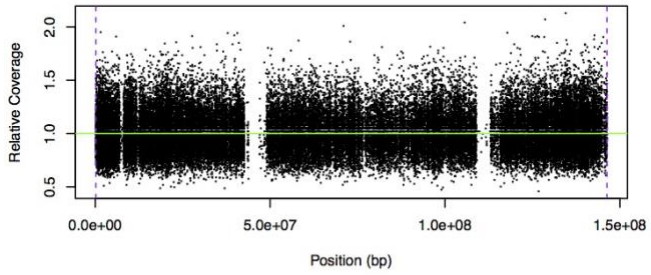


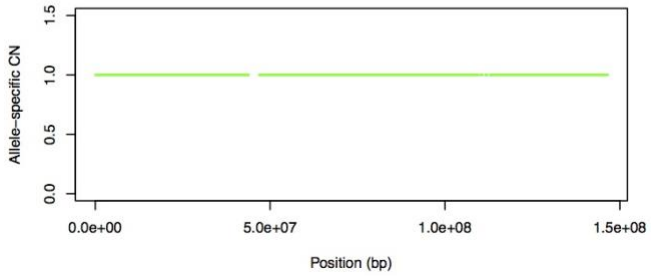




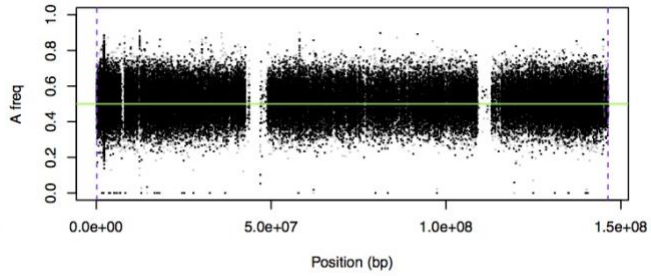
Chr8 Primary Tumor

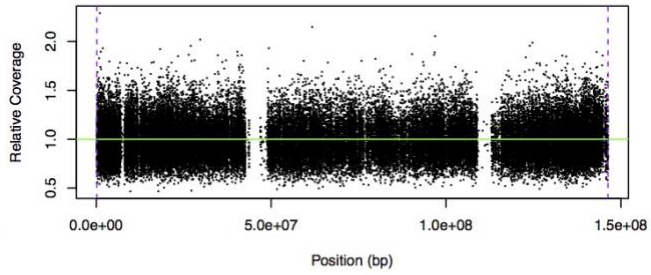


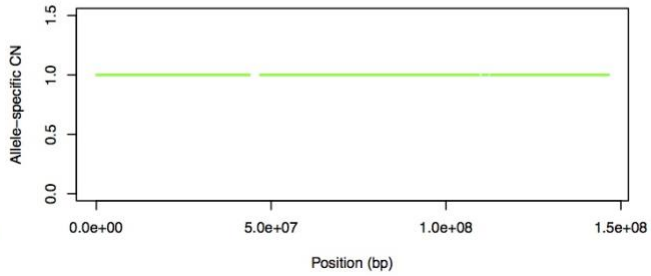




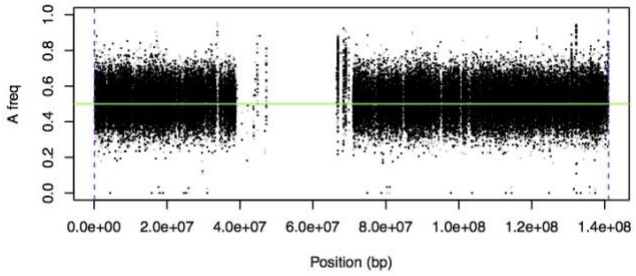
Chr8 Relapse Genome

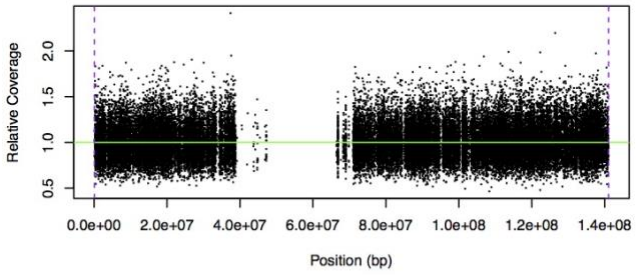


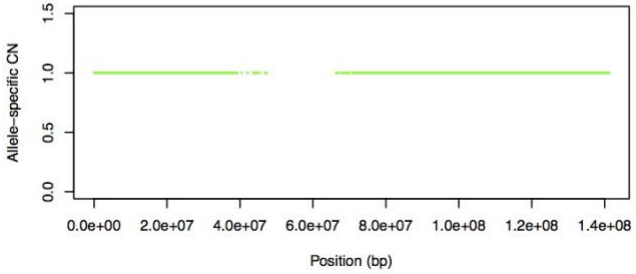




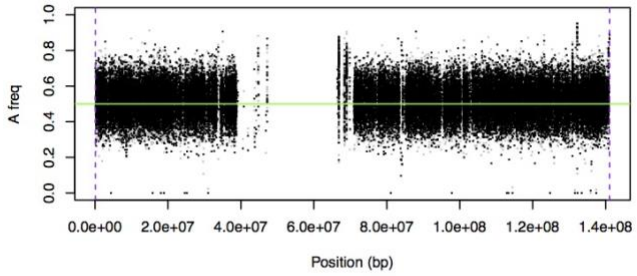
Chr9 Primary Tumor

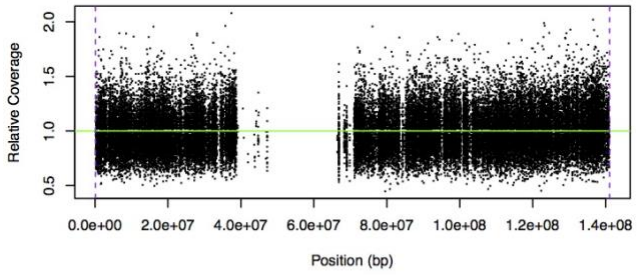


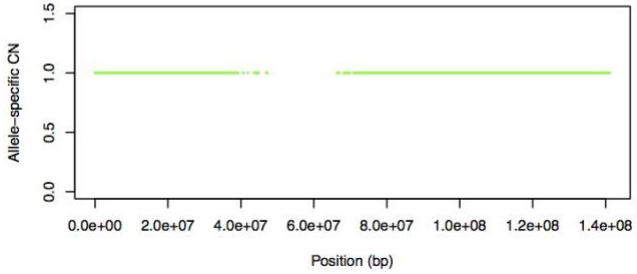




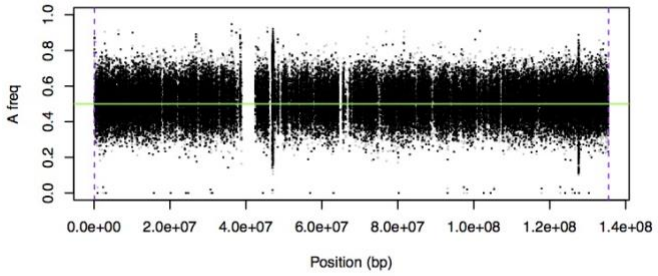
Chr9 Relapse Genome

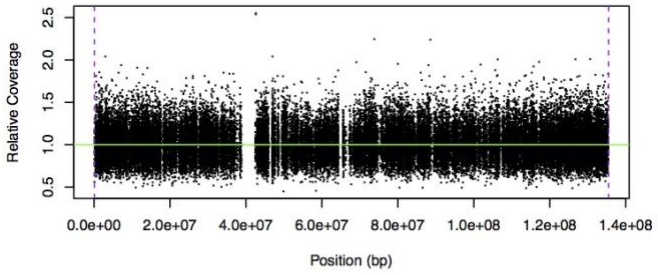


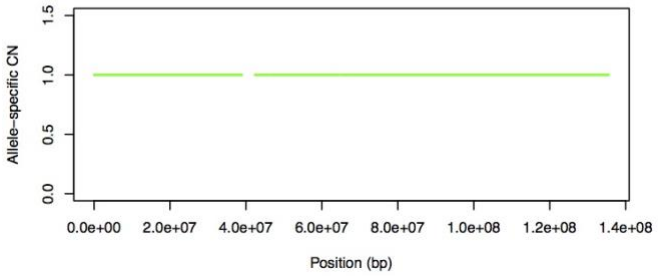




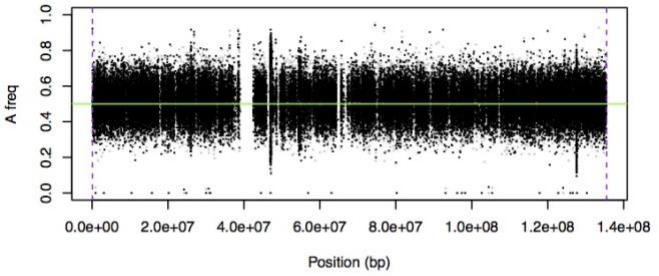
Chr10 Primary Tumor

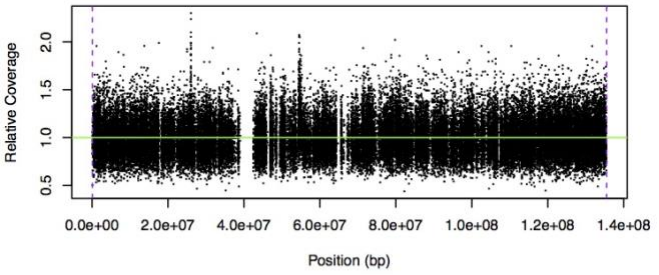


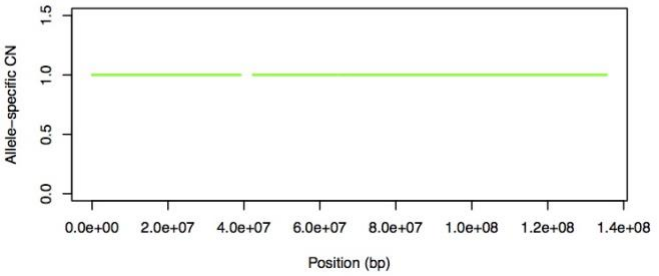




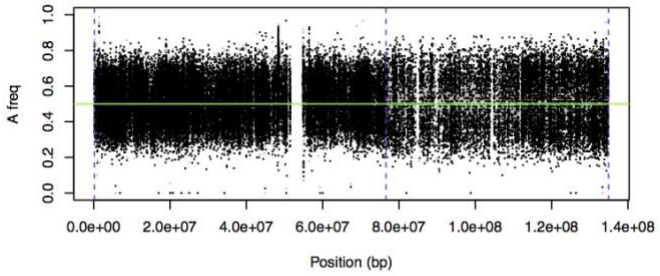
Chr10 Relapse Genome

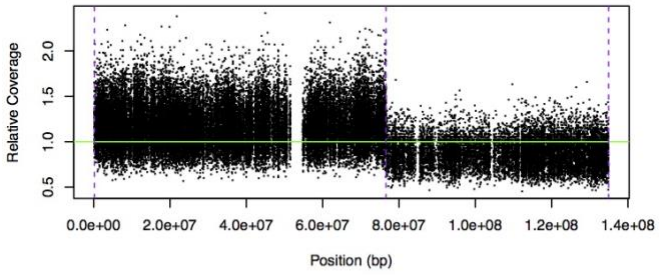


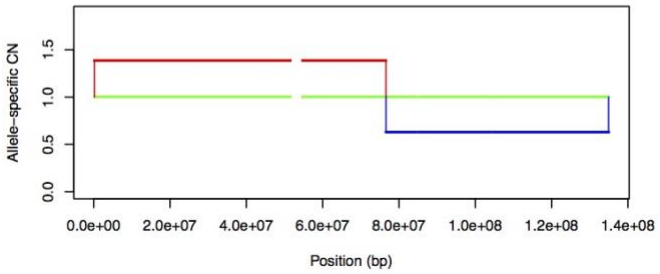




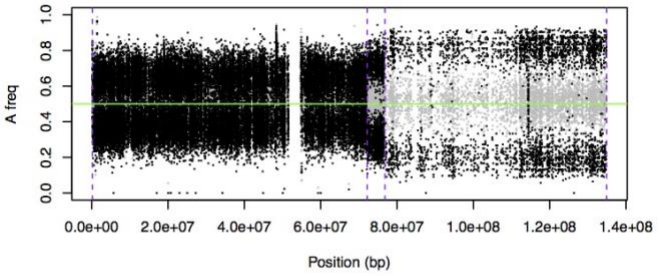
Chr11 Primary Tumor

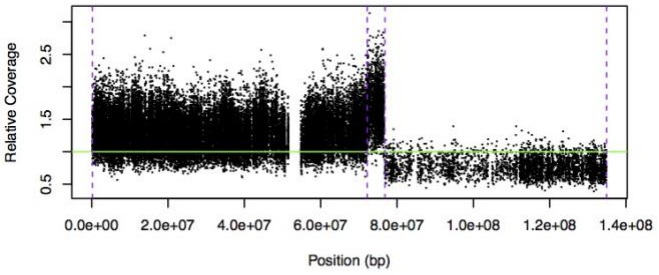


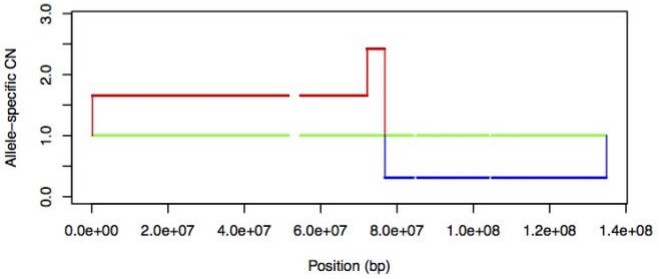




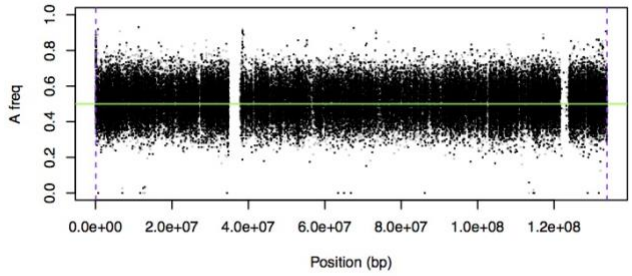
Chr11 Relapse Genome

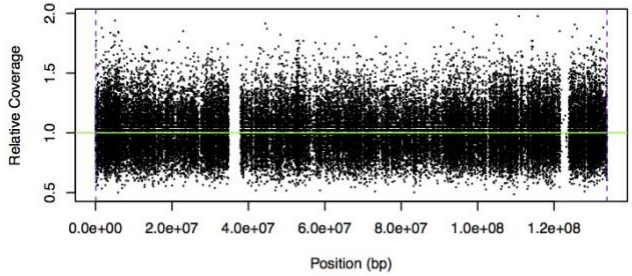


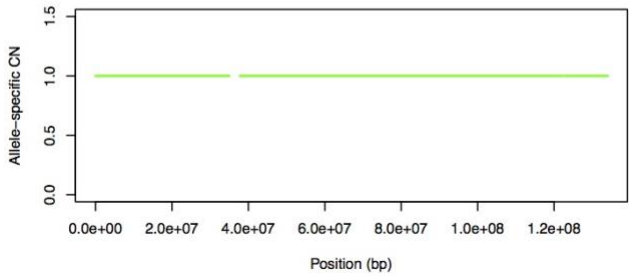




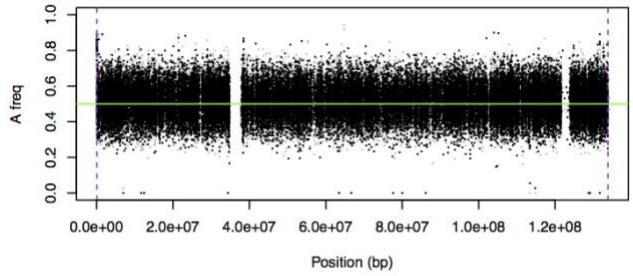
Chr12 Primary Tumor

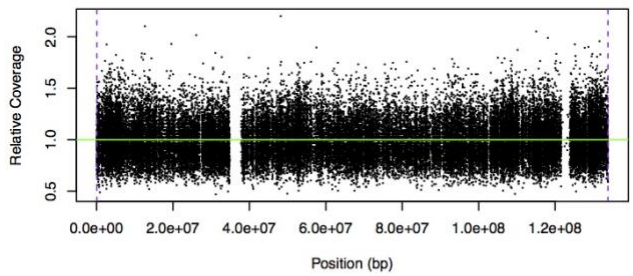


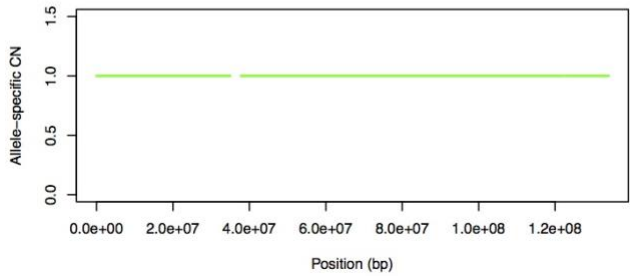




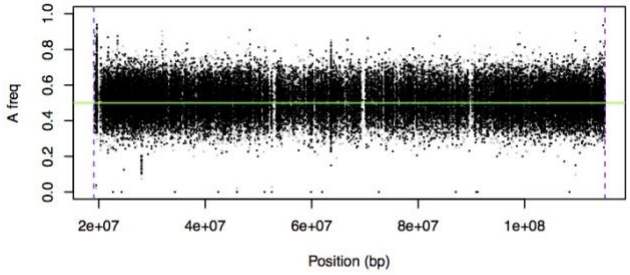
Chr12 Relapse Genome

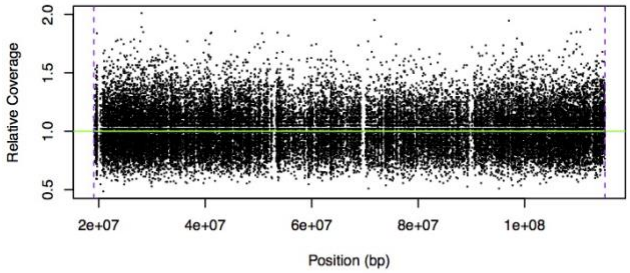


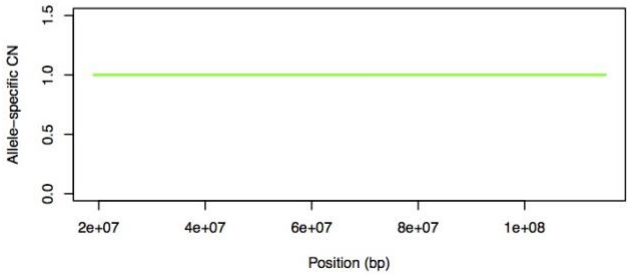




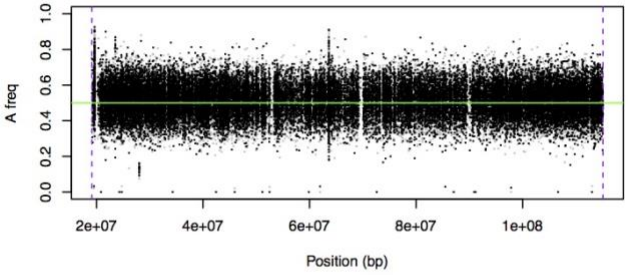
Chr13 Primary Tumor

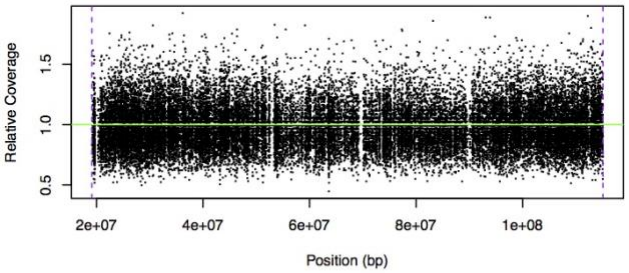


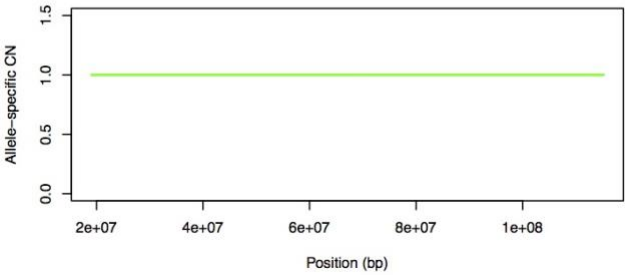




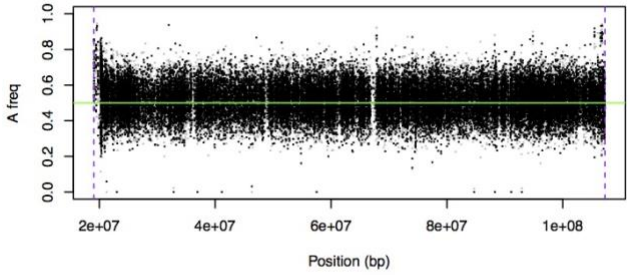
Chr13 Relapse Genome

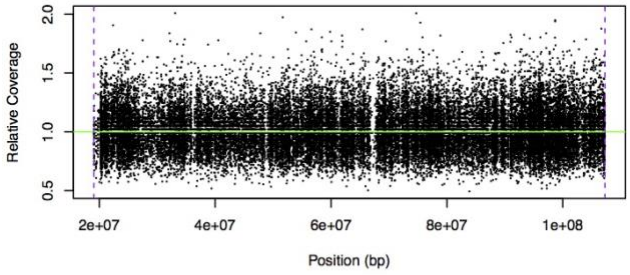


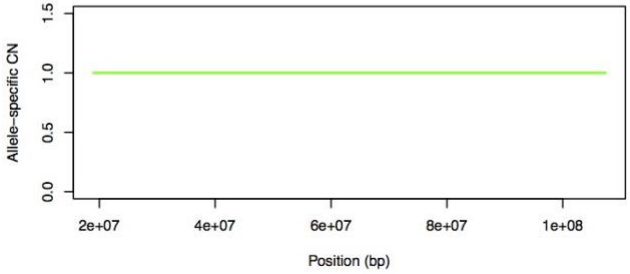




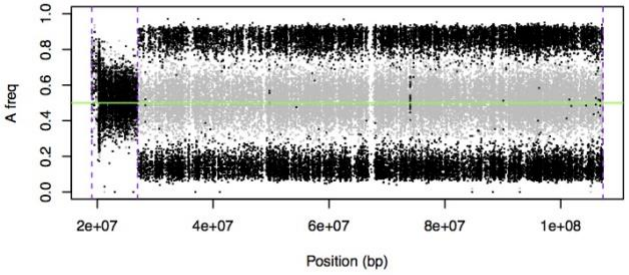
Chr14 Primary Tumor

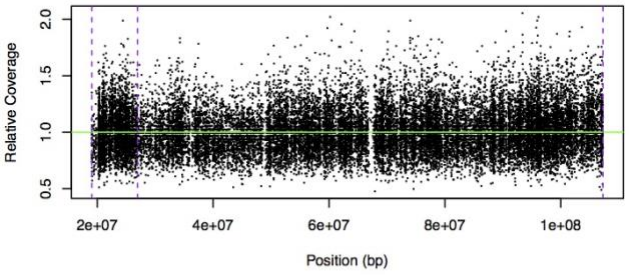


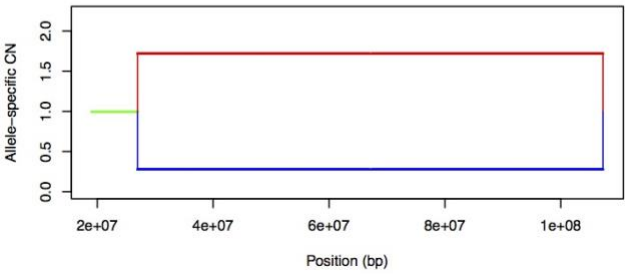




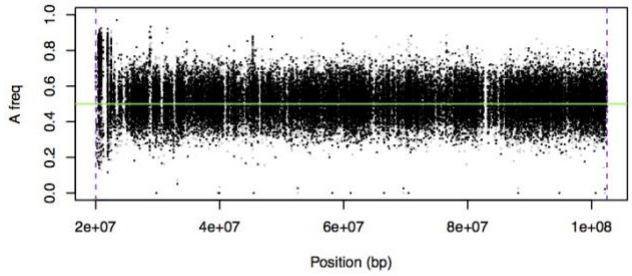
Chr14 Relapse Genome

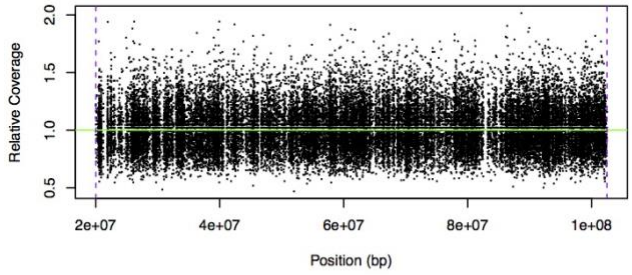


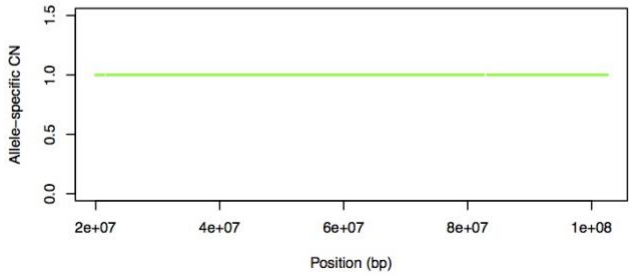




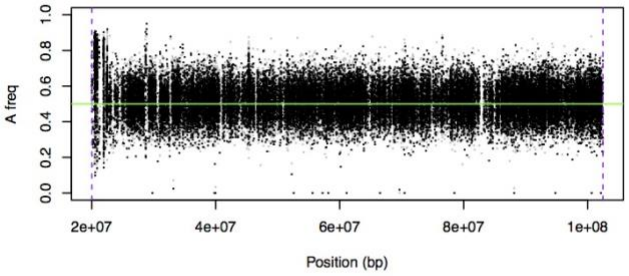
Chr15 Primary Tumor

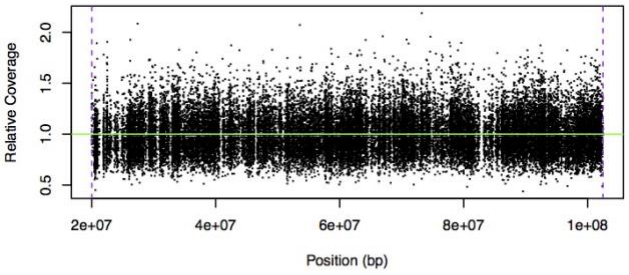


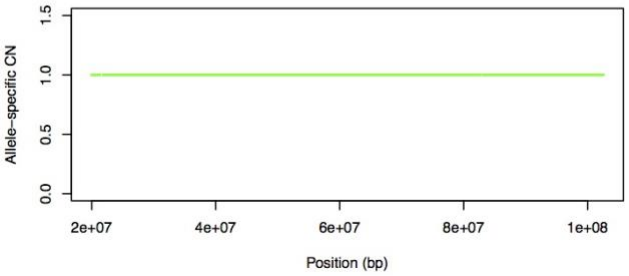




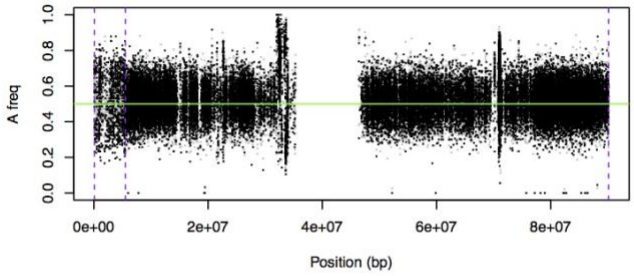
Chr15 Relapse Genome

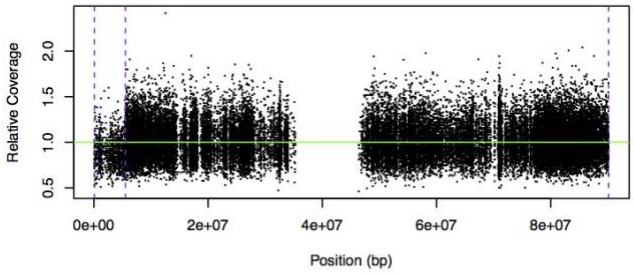


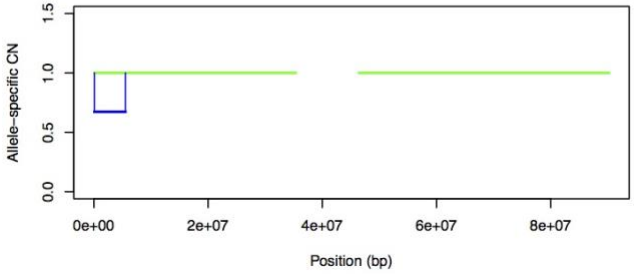




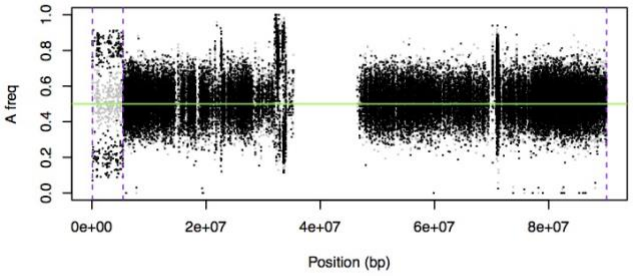
Chr16 Primary Tumor

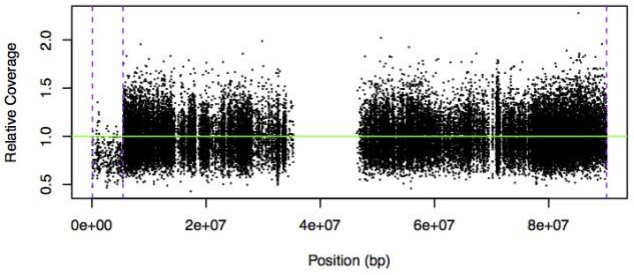


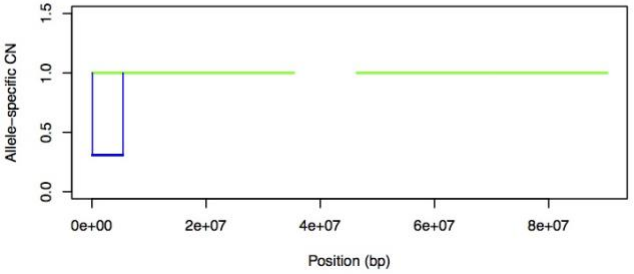




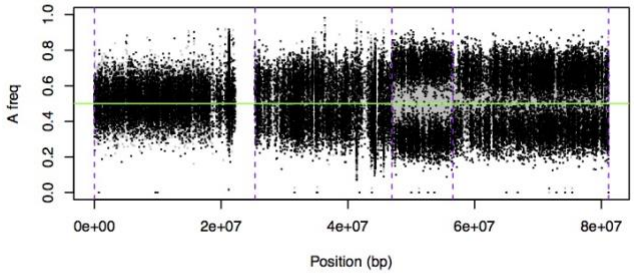
Chr16 Relapse Genome

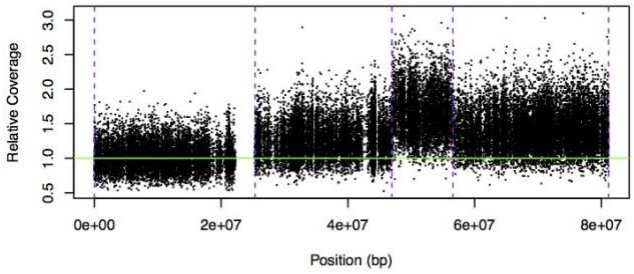


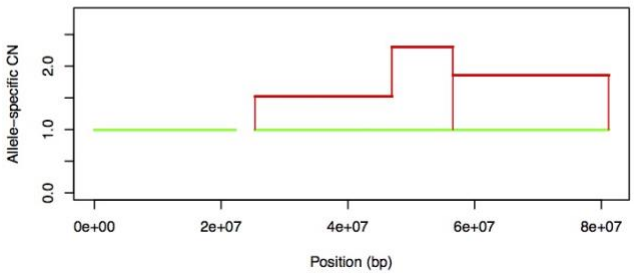




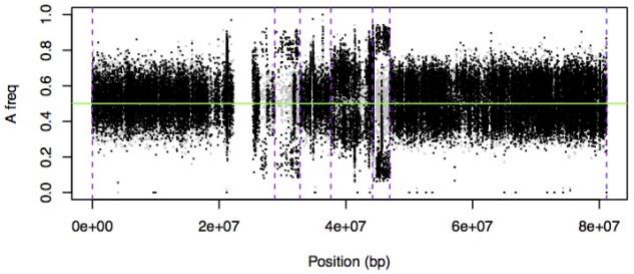
Chr17 Primary Tumor

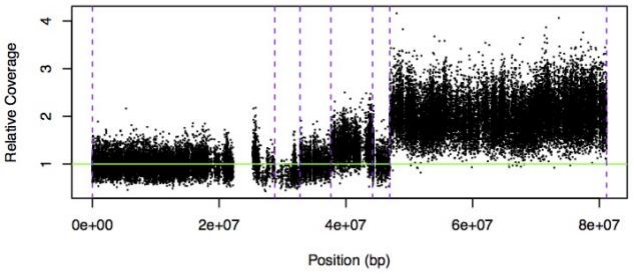


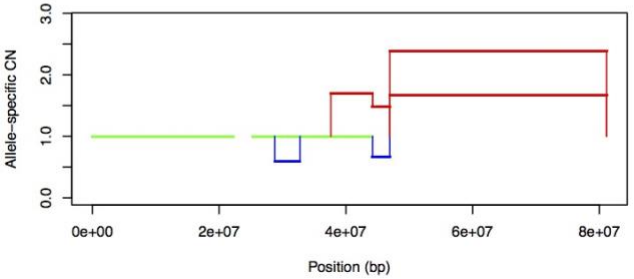




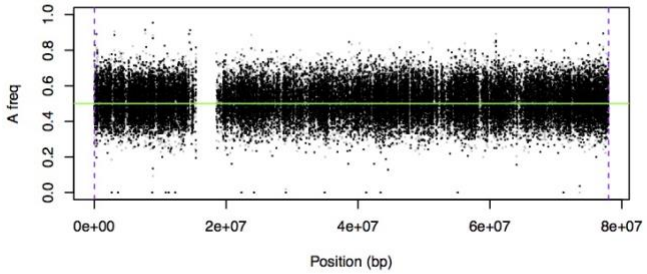
Chr17 Relapse Genome

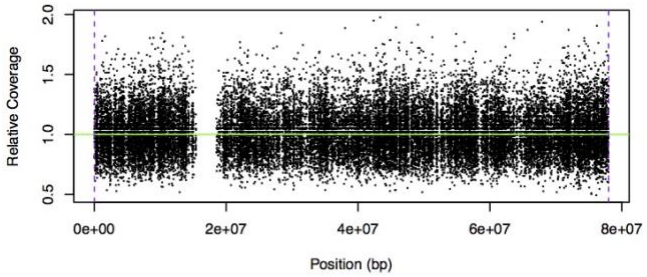


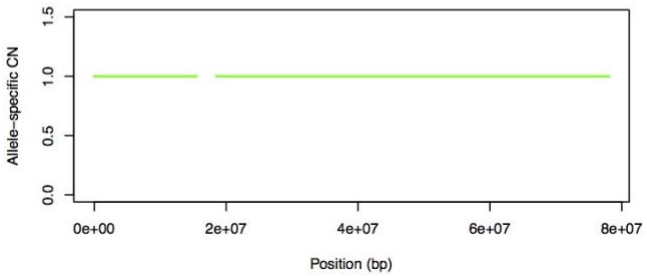




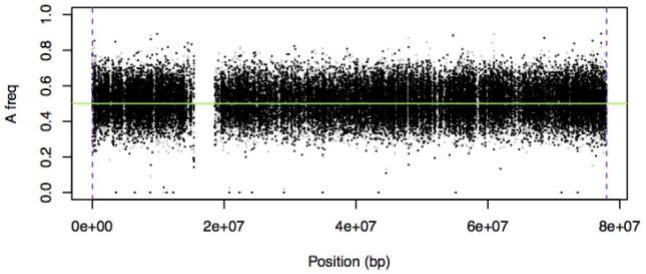
Chr18 Primary Tumor

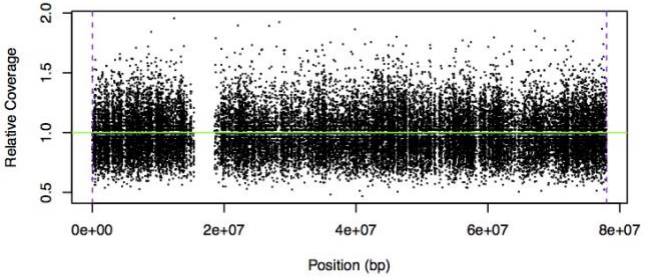


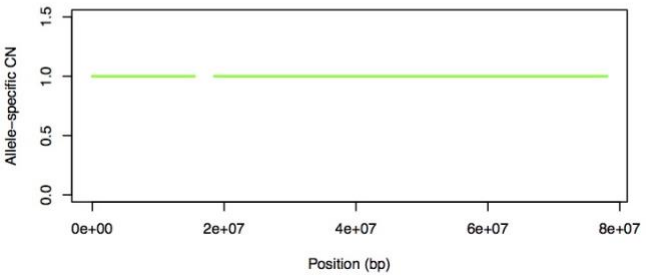




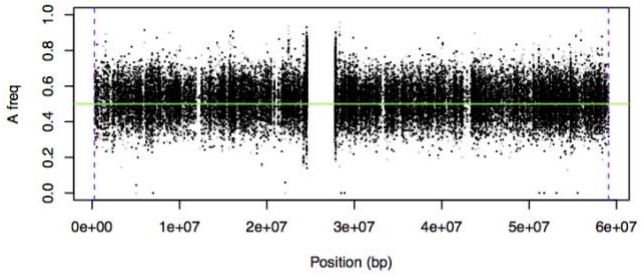
Chr18 Relapse Genome

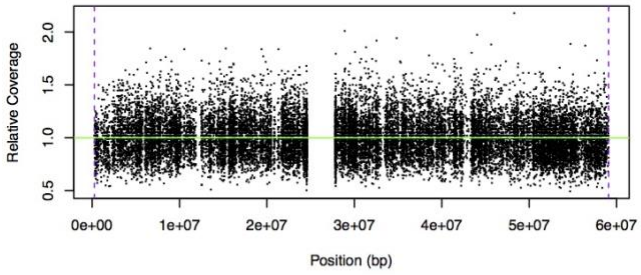


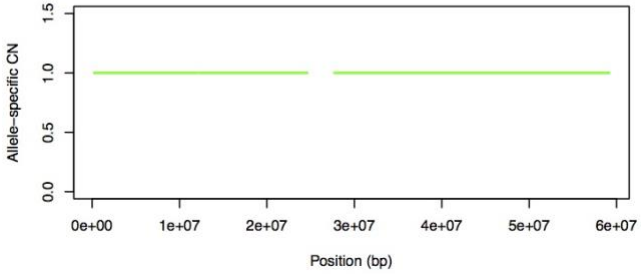




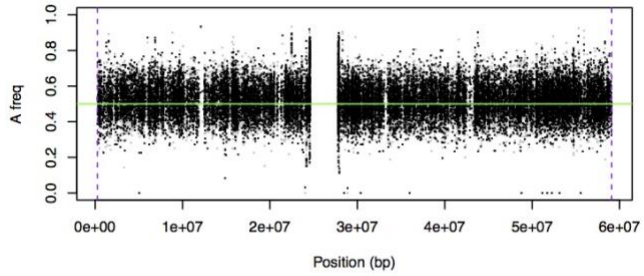
Chr19 Primary Tumor

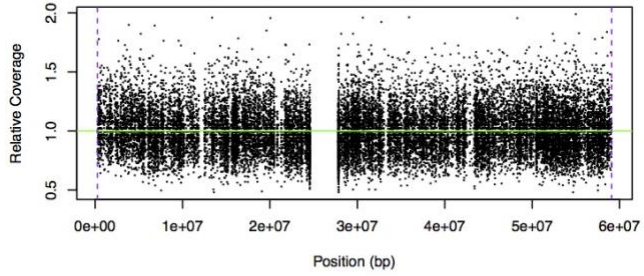


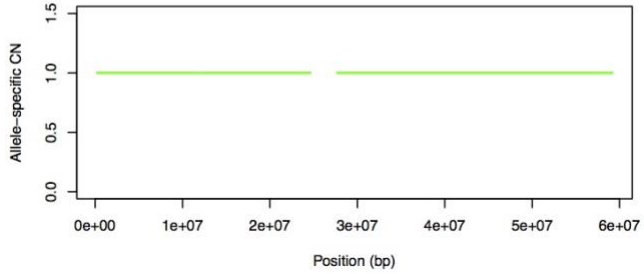




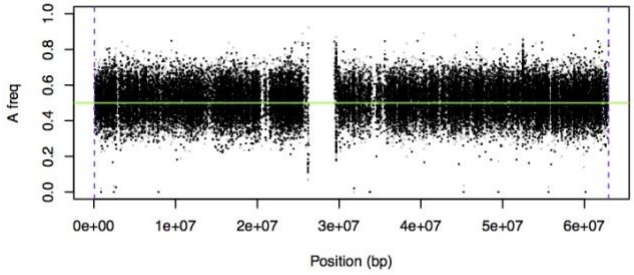
Chr19 Relapse Genome

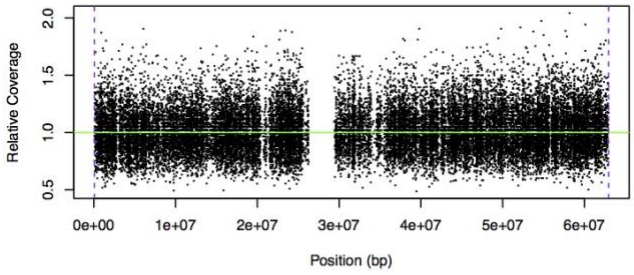


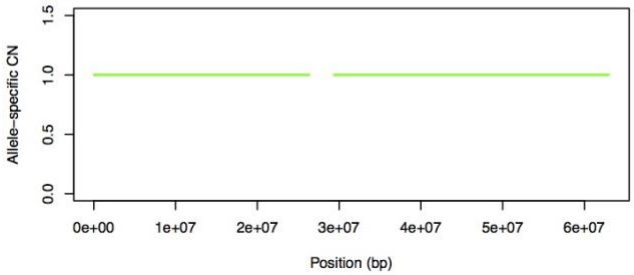




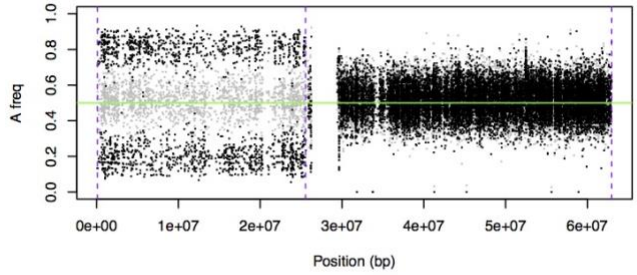
Chr20 Primary Tumor

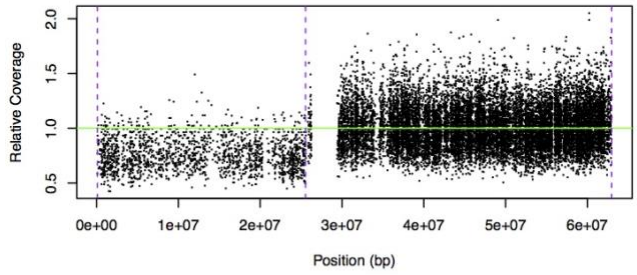


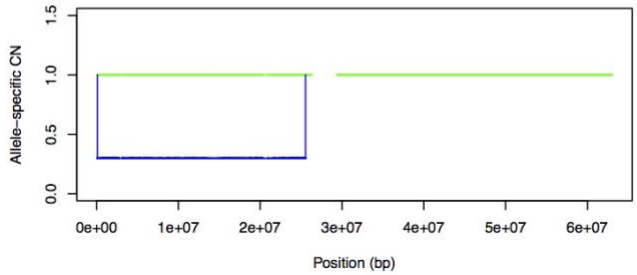




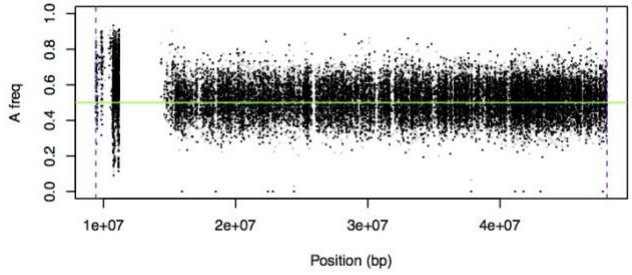
Chr20 Relapse Genome

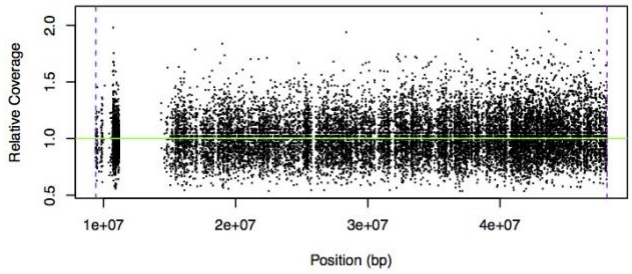


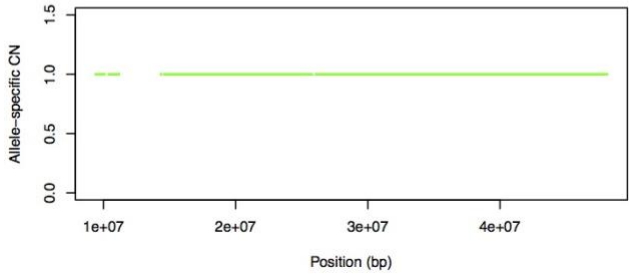




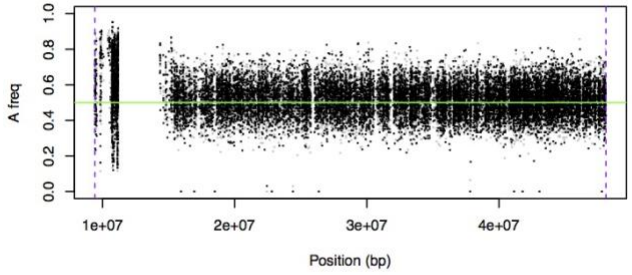
Chr21 Primary Tumor

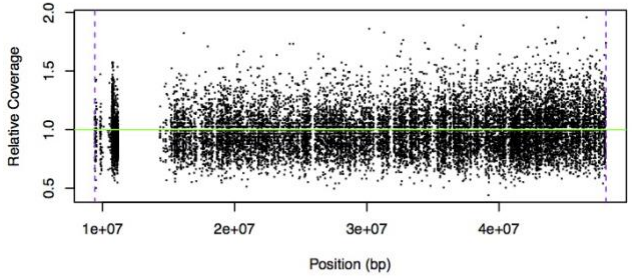


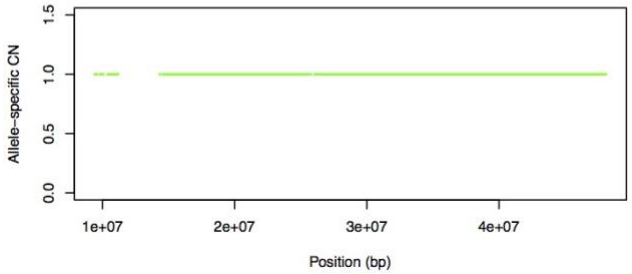




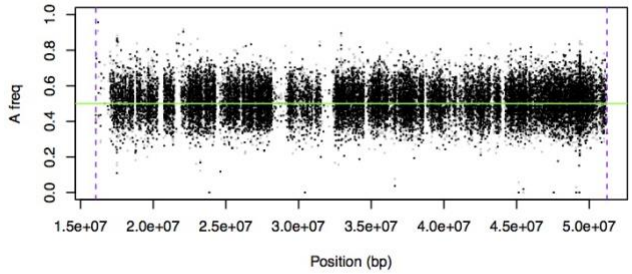
Chr21 Relapse Genome

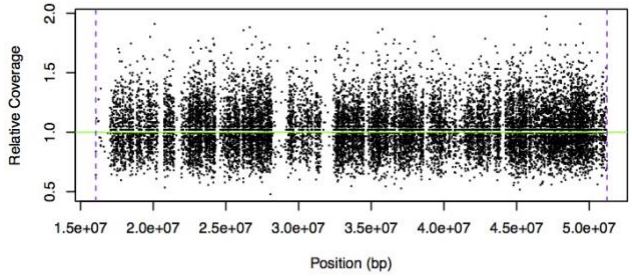


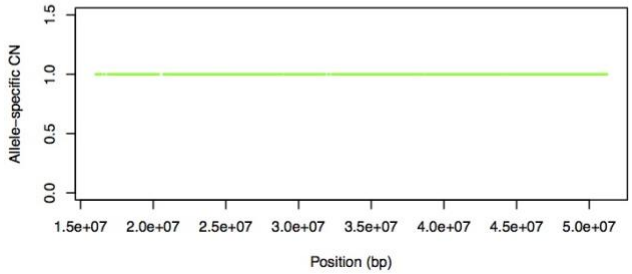




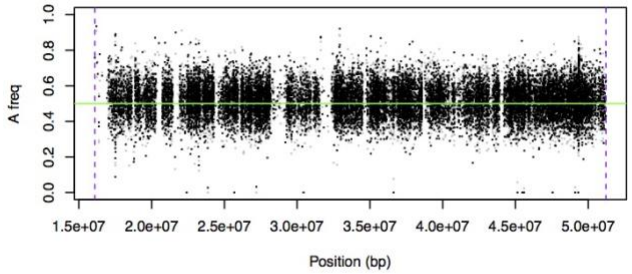
Chr22 Primary Tumor

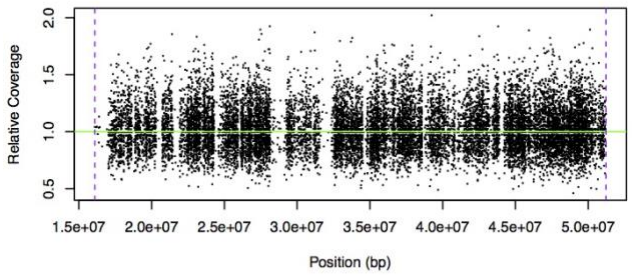


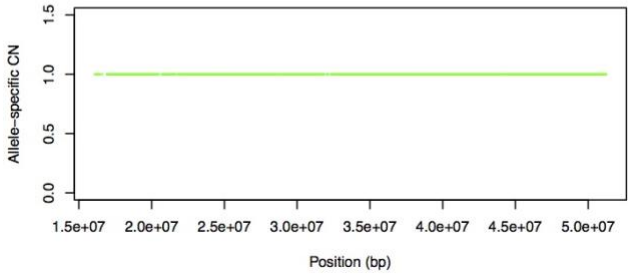




Chr22 Relapse Genome

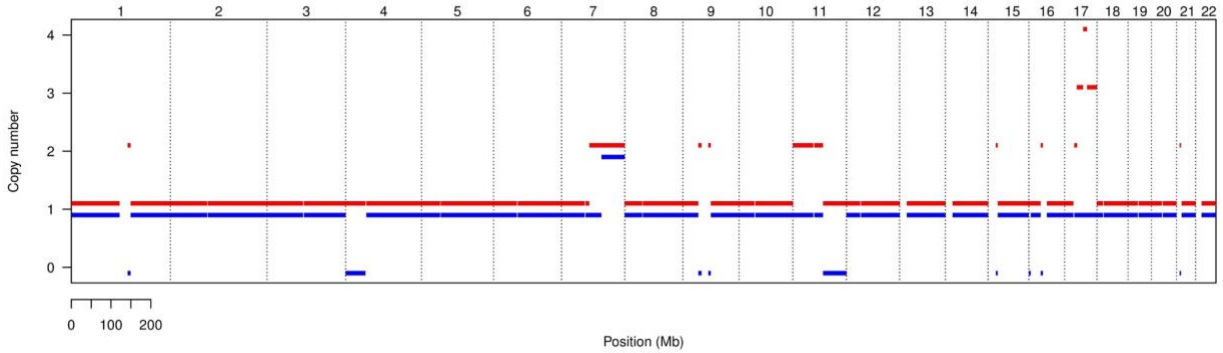




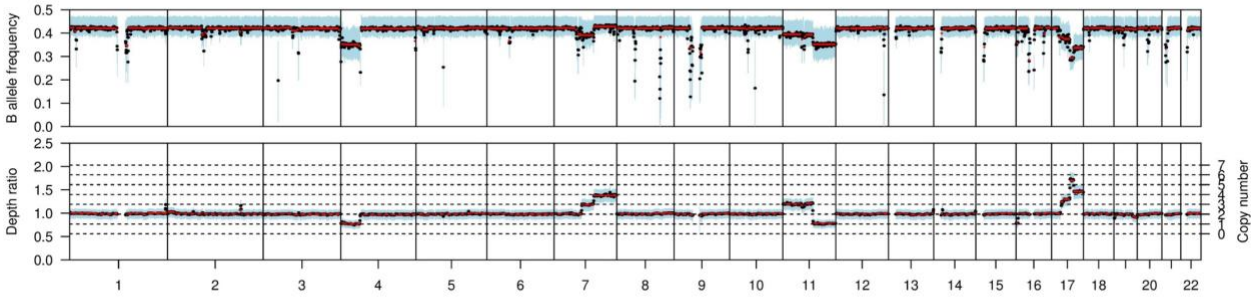


Supplementary Figure S2. Genome-wide ASCN profiles and segmentation results by Sequenza (Favero, et al., 2015), which assumes clonal copy number change events. Integer-value ASCNs are returned with adjustment of tumor purity and ploidy.

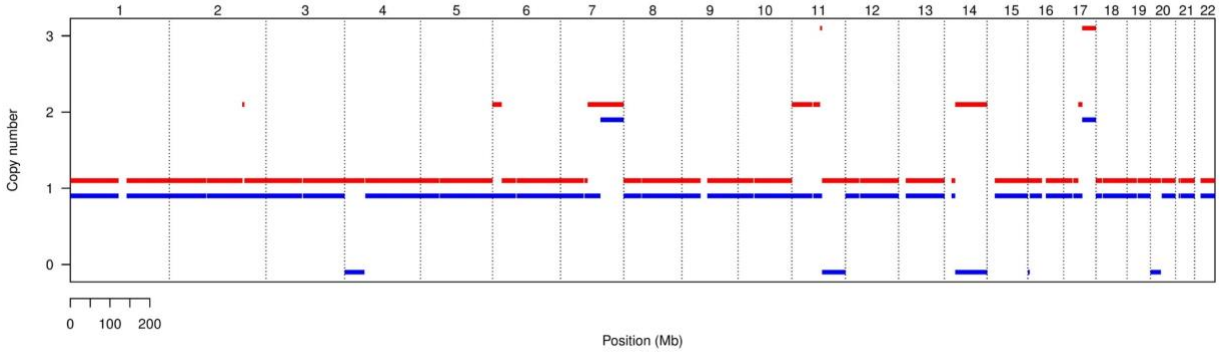
(A) Integer-value ASCN of primary returned by Sequenza.



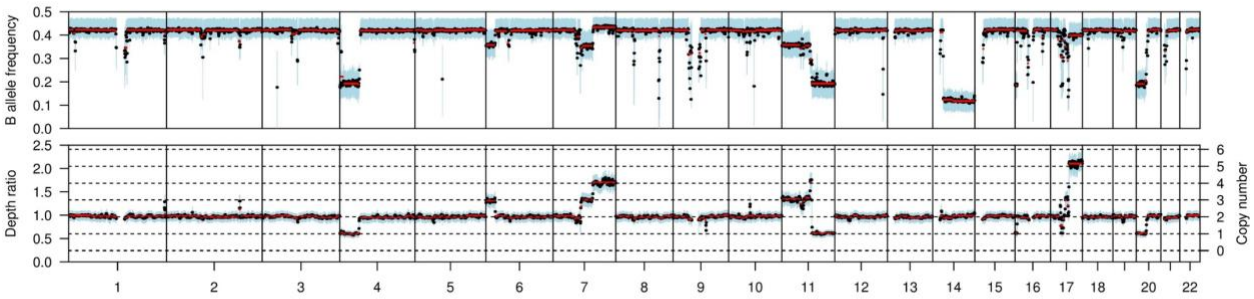
(B) B allele frequency and depth ratio of primary returned by Sequenza.



(C) Integer-value ASCN of relapse returned by Sequenza.

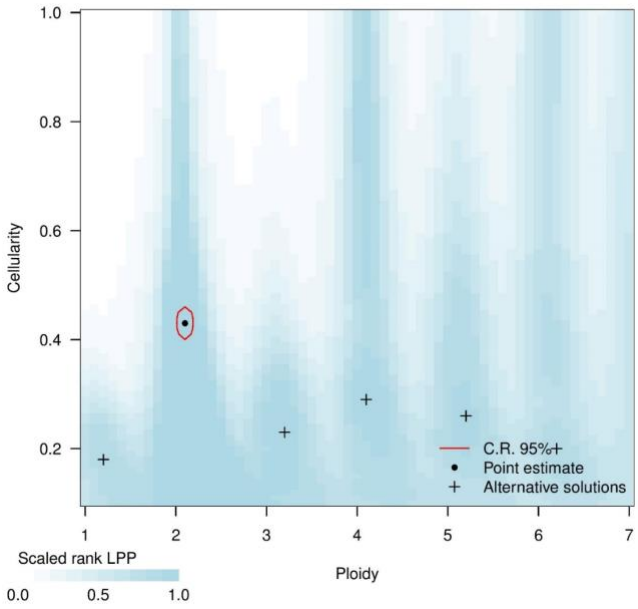


(D) B allele frequency and depth ratio of relapse returned by Sequenza.

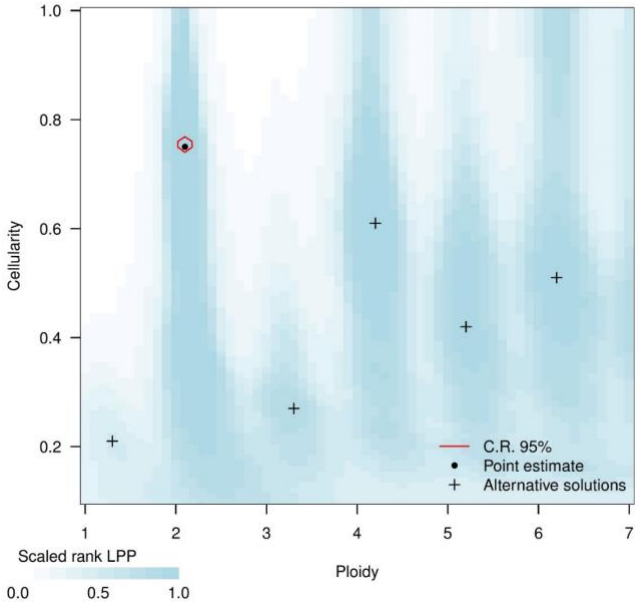


Supplementary Figure S3. Sequenza's results from the inference over the defined range of cellularity/purity and ploidy. Color intensity indicates the log posterior probability of corresponding cellularity/ploidy values. Multiple modes exist in the posterior where the purity and ploidy are not always identifiable.

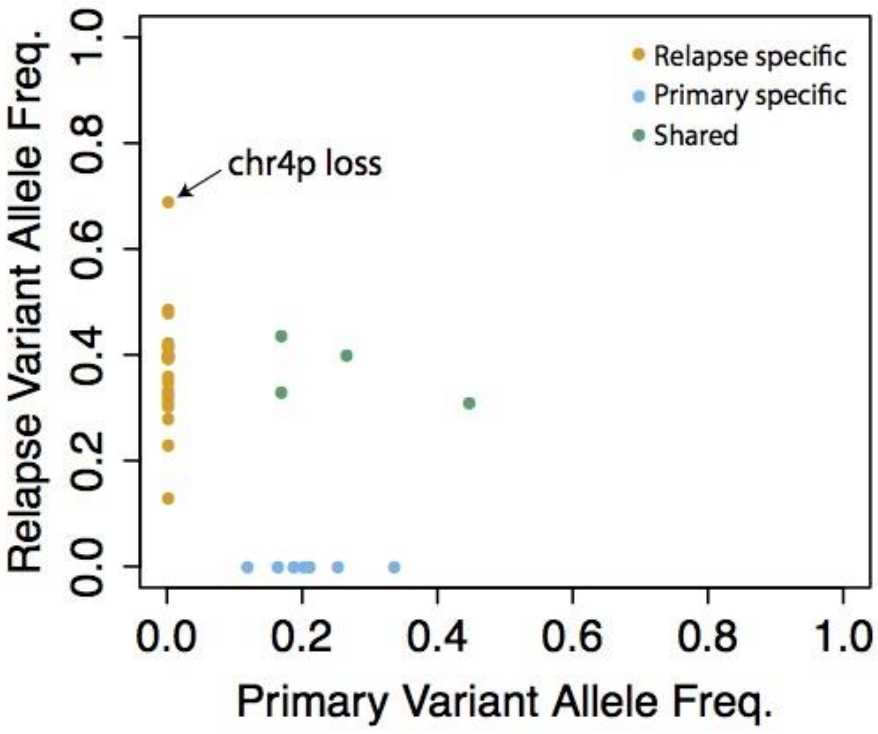
(A) Primary tumor.



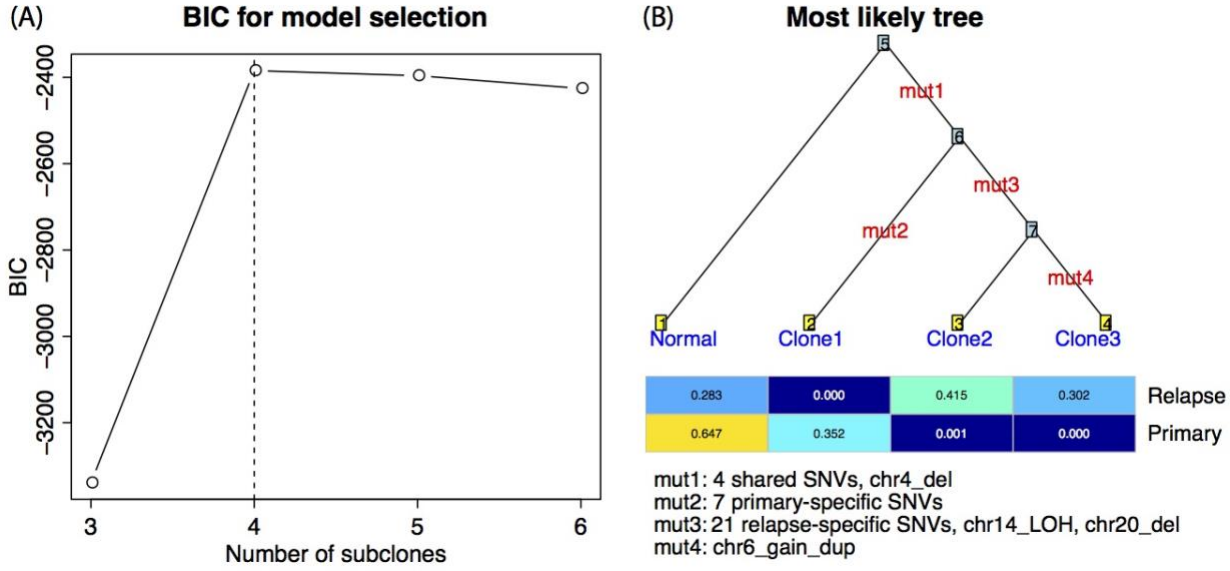
(B) Relapse genome.



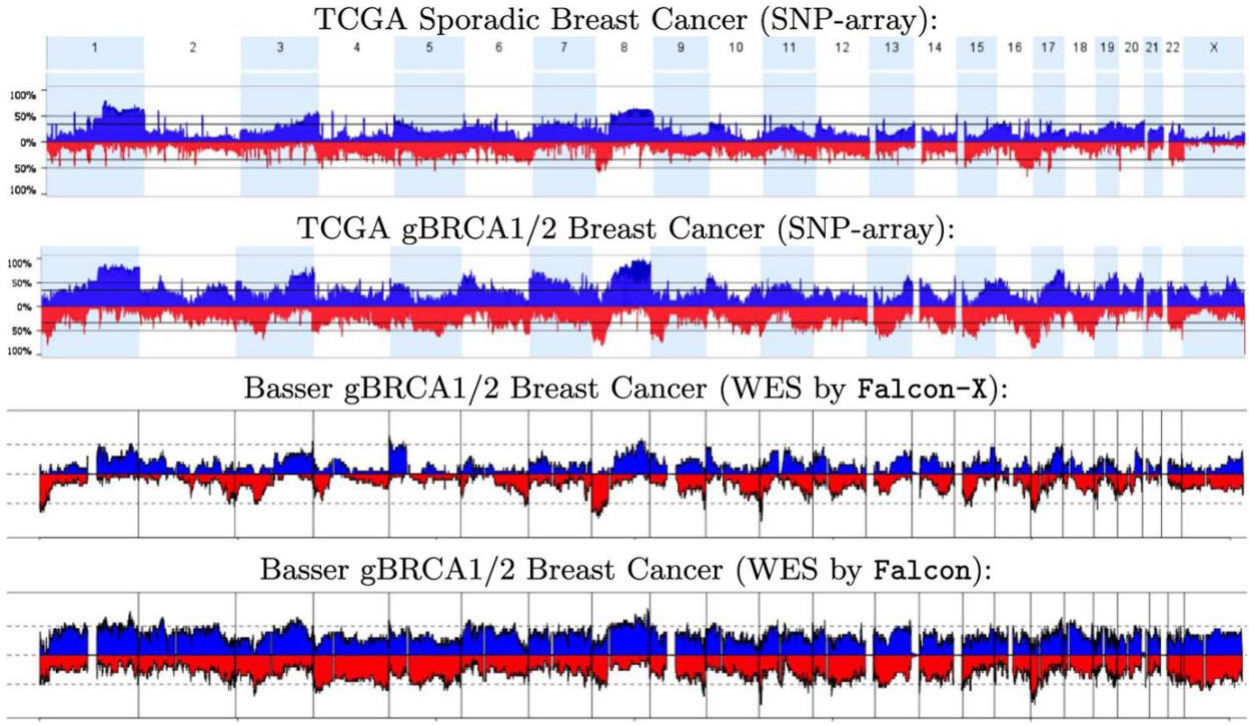
Supplementary Figure S4. Variant allele frequencies (VAFs) of SNVs in primary tumor and relapse genome. 32 high-confidence SNVs are categorized as deleterious, of which 7 are unique to the primary (blue), 21 are unique to the relapse (orange), and 4 are shared between the two bulk samples (green). All variants have VAFs less than 50%, except for one in gene *CORIN*, which lies in a LOH region in chr14 and is thus enriched in the relapse with a VAF of 68.8%.



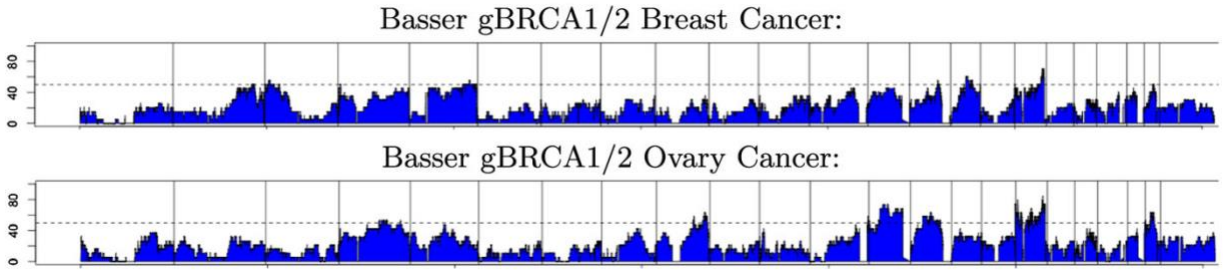
Supplementary Figure S5. Tumor phylogeny reconstructed by Canopy (Jiang, et al., 2016). (A) BIC as a model selection metric to determine the number of subclones including one for the normal cells. (B) Most likely tree with 4 subclones returned by Canopy. There is only one posterior tree configuration in the tree space. Quantities of tree elements are estimated from the sampling posterior with confidence assessment.



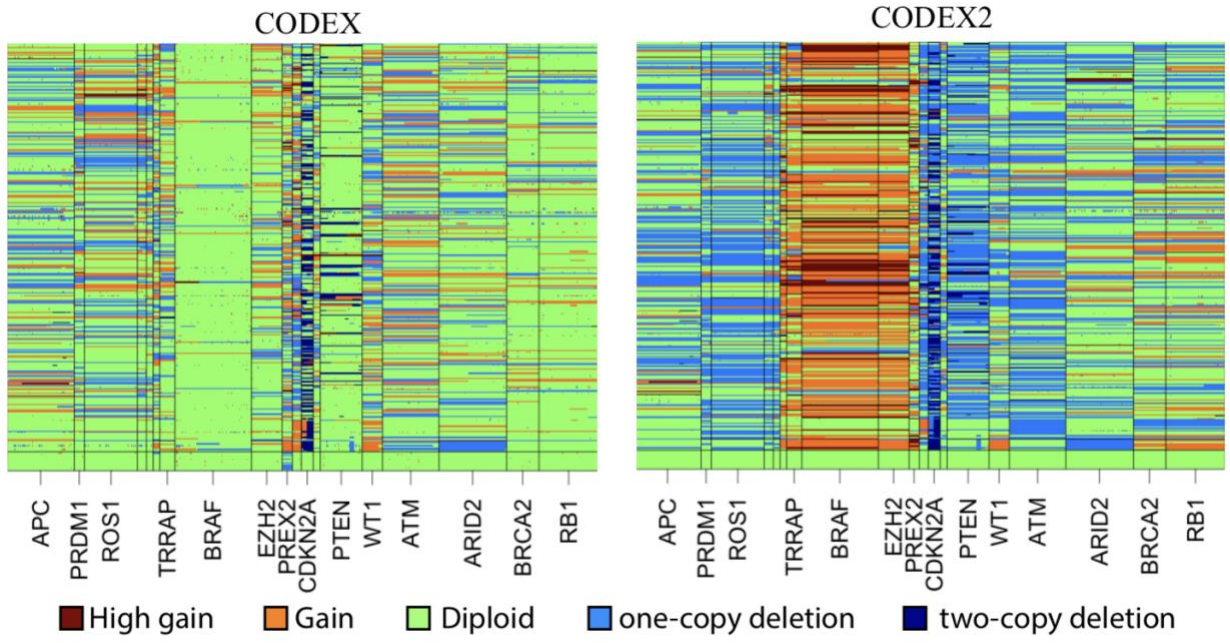
Supplementary Figure S6. Frequency of detected occurrence of gains (in blue, above the axis) and losses (in red, below the axis) of total copy number in three breast cancer cohorts: TCGA sporadic breast cancers, TCGA gBRCA1/2 breast cancers, and our Basser gBRCA1/2 breast cancers. The TCGA cohorts, shown in the top two plots, were profiled by the genotyping array. The Basser samples were profiled by WES and analyzed by Falcon-X, shown in the third plot from the top, and by Falcon, shown in the bottom plot. The horizontal axis shows genome location, and is aligned between the four plots. The vertical axis shows the proportion of samples where a call is made. Chromosome boundaries are marked by vertical lines or color shading.



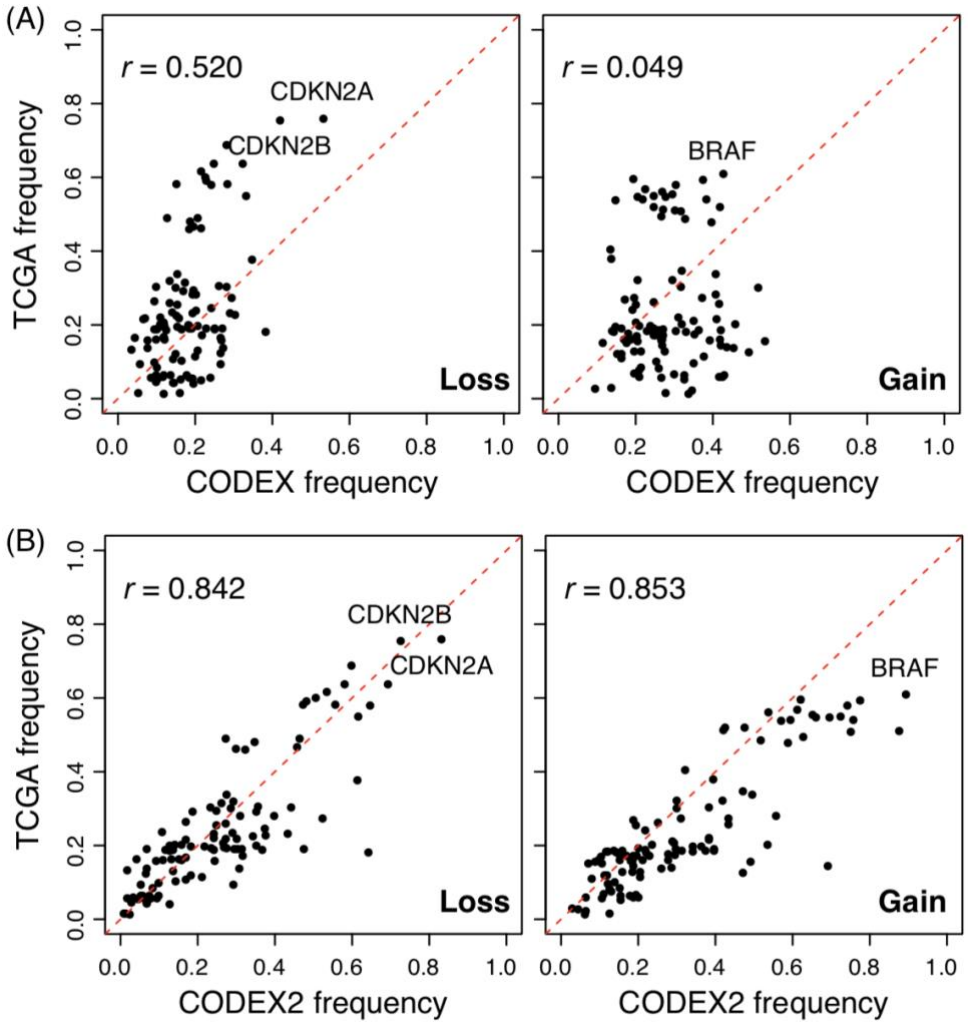
Supplementary Figure S7. Frequency of occurrence of Copy-neutral loss of heterozygosity (LOH) found by Falcon-X in the Basser gBRCA1/2 breast cancer cohort and the Basser gBRCA1/2 ovarian cancer cohort. The horizontal axis shows genome location aligned between the two plots, and vertical axis shows percentage of samples where LOH is detected. Vertical lines denote chromosome boundaries.



Supplementary Figure S8. Heatmap of CODEX and CODEX2 normalization/segmentation results for the melanoma cohort. Each column is one target in the gene panel; each row is one sample, with the first 16 towards the bottom in the heatmap being normal. Profiled CNVs are categorized as high gain, gain, diploid (null), one-copy loss, and two-copy loss based on the estimated copy numbers. Only part of the oncogenes and tumor suppressors with greater than 30 targets are shown.



Supplementary Figure S9. Assessment of profiled CNVs in the melanoma cohort with comparison to TCGA. CNVs are separated by states: losses on the left panels and gains on the right panel. Each dot corresponds to one gene in the targeted sequencing panel. CNV frequencies detected by (A) CODEX and (B) CODEX2 from the melanoma cohort is compared to the TCGA cohort with CODEX2 having drastically higher correlation.



Supplementary Table S1. ASCN profiles of primary tumor and relapse genome of neuroblastoma patient PASGAP (Eleveld, et al., 2015) returned by FALCON (Chen, et al., 2015).

(A) Primary tumor ASCN profile.

chr	st_snp	end_snp	st_bp	end_bp	Minor_copy	Major_copy	Minor.sd	Major.sd
4	141	10571	68603	49067896	0.647	1	0.0019	0.0029
7	26581	37821	70209300	101172171	1	1.359	0.002	0.003
7	37821	67110	101172171	159118443	1	1.381	0.0014	0.0019
11	1	35551	181188	76579669	1	1.383	0.0012	0.0018
11	35551	47396	76579669	134944606	0.63	1	0.0016	0.0024
16	1	971	84926	5518787	0.672	1	0.006	0.0091
17	10411	20871	25342455	46941578	1	1.52	0.0021	0.003
17	20871	26081	46941578	56568680	1	2.308	0.0032	0.0063
17	26081	39614	56568680	81149937	1	1.856	0.0019	0.0032

(B) Relapse genome ASCN profile.

chr	st_snp	end_snp	st_bp	end_bp	Minor_copy	Major_copy	Minor.sd	Major.sd
2	67261	68021	185395311	187588855	1	1.532	0.0095	0.0145
4	1	3371	17207	49223101	0.412	1	0.0044	0.0063
6	1	14741	149609	23724469	1	1.604	0.0022	0.0035
7	26411	38481	70200455	101161506	1	1.603	0.0019	0.003
7	38481	68550	101161506	159118443	1.511	1.696	0.0016	0.0021
11	1	35641	181188	72167491	1	1.658	0.0012	0.0019
11	35641	37571	72167491	76821335	1	2.418	0.0056	0.0115
11	37571	41141	76821335	134929444	0.313	1	0.0019	0.004
14	3451	24617	26963559	107287663	0.277	1.719	9.00E-04	0.0032
16	1	311	86709	5449790	0.307	1	0.0078	0.0176
17	11061	11631	28778189	32751138	0.592	1	0.0085	0.0079
17	13551	17021	37626663	44210988	1	1.698	0.0036	0.0052
17	17021	18151	44210988	46924980	0.668	1.479	0.0067	0.0091
17	18151	37154	46924980	81149937	1.673	2.391	0.0019	0.0028
20	1	1731	98930	25543198	0.298	1	0.0032	0.0076

Supplementary Table S2. Somatic single nucleotide variants profiled by the GATK UnifiedGenotyper. Stringent quality control procedures are carried out to remove possible germline mutations, low-quality indels, variants with missing genotypes, and variants with low depth of coverage. Annotation is carried out using ANNOVAR. SNVs that are deleterious by at least one scoring metric are used as input for Canopy to infer tumor phylogeny.

chr	pos	ref	alt	refGene	ExonicFunc	P_ref	P_alt	R_ref	R_alt	P_VAF	R_VAF
1	1118426	A	G	TTLL10	nonsynonymous SNV	20	5	0	0	0.2	0
1	86578276	C	A	COL24A1	nonsynonymous SNV	0	0	18	8	0	0.308
1	155932794	C	T	ARHGEF2	nonsynonymous SNV	10	5	0	0	0.333	0
1	201177341	C	G	IGFN1	nonsynonymous SNV	0	0	38	25	0	0.397
2	160672032	G	T	LY75;LY75-CD302	nonsynonymous SNV	0	0	12	8	0	0.4
2	186661228	T	C	FSIP2	nonsynonymous SNV	35	7	37	18	0.167	0.327
3	57447290	G	C	DNAH12	unknown	10	2	13	10	0.167	0.435
3	167159941	G	T	SERPINI2	nonsynonymous SNV	0	0	11	8	0	0.421
4	47644012	G	T	CORIN	nonsynonymous SNV	0	0	5	11	0	0.688
4	68919697	G	T	TMPRSS11F	nonsynonymous SNV	0	0	27	13	0	0.325
4	119145719	C	G	NDST3	nonsynonymous SNV	0	0	17	5	0	0.227
4	166964497	C	A	TLL1	nonsynonymous SNV	0	0	23	10	0	0.303
5	74807166	A	G	COL4A3BP	nonsynonymous SNV	0	0	13	5	0	0.278
5	148407267	C	A	SH3TC2	nonsynonymous SNV	0	0	36	18	0	0.333
6	31696700	A	C	DDAH2	nonsynonymous SNV	18	6	0	0	0.25	0
6	108492724	G	A	NR2E1	nonsynonymous SNV	0	0	14	9	0	0.391
7	44579412	G	T	NPC1L1	nonsynonymous SNV	0	0	27	15	0	0.357
8	143356161	G	A	TSNARE1	nonsynonymous SNV	19	5	0	0	0.208	0
9	95608857	G	C	ZNF484	nonsynonymous SNV	0	0	17	12	0	0.414
9	97535426	G	T	C9orf3	nonsynonymous SNV	0	0	15	8	0	0.348
10	6008290	G	T	IL15RA	nonsynonymous SNV	0	0	18	17	0	0.486
10	25145905	C	T	PRTFDC1	nonsynonymous SNV	0	0	38	25	0	0.397
12	51450158	T	C	LETMD1	nonsynonymous SNV	0	0	34	5	0	0.128
12	83359438	T	C	TMTC2	nonsynonymous SNV	26	5	0	0	0.161	0
12	113405742	G	T	OAS3	nonsynonymous SNV	38	5	0	0	0.116	0
15	45781095	C	A	SLC30A4	stopgain	14	5	12	8	0.263	0.4
15	91422159	A	T	FURIN	nonsynonymous SNV	0	0	12	8	0	0.4
17	2290292	A	C	MNT	nonsynonymous SNV	22	5	0	0	0.185	0
18	8784236	G	T	MTCL1	nonsynonymous SNV	10	8	18	8	0.444	0.308
18	28914139	A	T	DSG1	nonsynonymous SNV	0	0	17	8	0	0.32
19	19466859	G	T	MAU2	unknown	0	0	21	15	0	0.417
22	43010848	G	C	POLDIP3	nonsynonymous SNV	0	0	13	12	0	0.48

Supplementary Table S3. Performance of CODEX and CODEX2 with different number of latent factors. CODEX and CODEX2 are applied to the melanoma targeted sequencing data set with the number of latent factors K ranging from 0 to 10. Correlations of the profiled losses and gains (r_{loss} and r_{gain} respectively) by CODEX and CODEX2 with those reported by TCGA, as well as the number of BRAF gains and PTEN losses out of 334 tumor samples are used as measurements. CNV profiles by CODEX2 are consistent since only the negative control samples (16 normal samples) are used to estimate the target-specific bias and artifacts. For CODEX, true CNV signals are attenuated with a large K resulting in less CNV events and lower correlations.

	CODEX2				**CODEX**			
K	r_{loss}	r_{gain}	*BRAF* gains	*PTEN* losses	r_{loss}	r_{gain}	*BRAF* gains	*PTEN* losses
0	0.787	0.859	288	256	0.571	0.112	170	146
1	0.840	0.841	297	230	0.589	-0.039	163	138
2	0.837	0.839	296	228	0.580	-0.003	161	133
3	0.841	0.830	296	227	0.567	0.013	160	117
4	0.837	0.839	292	229	0.581	0.034	158	113
5	0.842	0.848	302	229	0.556	0.046	158	110
6	0.845	0.845	301	229	0.546	0.033	155	108
7	0.842	0.853	298	231	0.520	0.049	142	107
8	0.838	0.859	295	237	0.453	0.027	130	97
9	0.831	0.858	299	234	0.435	0.042	128	95
10	0.835	0.859	297	237	0.442	0.039	125	98

References

Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490(7418):61-70.

Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* 2015;161(7):1681-1696.

Chen, H., *et al.* Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res* 2015;43(4):e23.

Chen, H., *et al.* Allele-specific copy number estimation by whole exome sequencing. *The Annals of Applied Statistics* 2017;11(2):1169-1192.

DePristo, M.A., *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43(5):491-498.

Eleveld, T.F., *et al.* Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nat Genet* 2015;47(8):864-871.

Favero, F., *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015;26(1):64-70.

Garman, B., *et al.* Genetic and Genomic Characterization of 462 Melanoma Patient-Derived Xenografts, Tumor Biopsies, and Cell Lines. *Cell Rep* 2017;21(7):1936-1952.

Jiang, Y., *et al.* CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *bioRxiv* 2017:211698.

Jiang, Y., *et al.* CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res* 2015;43(6):e39.

Jiang, Y., *et al.* Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc Natl Acad Sci U S A* 2016;113(37):E5528-5537.

Maxwell, K.N., *et al.* BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. *Nat Commun* 2017;8(1):319.

Zhou, Z., *et al.* Integrative DNA copy number detection and genotyping from sequencing and array-based platforms. *bioRxiv* 2017:172700.