




In the format provided by the authors and unedited.

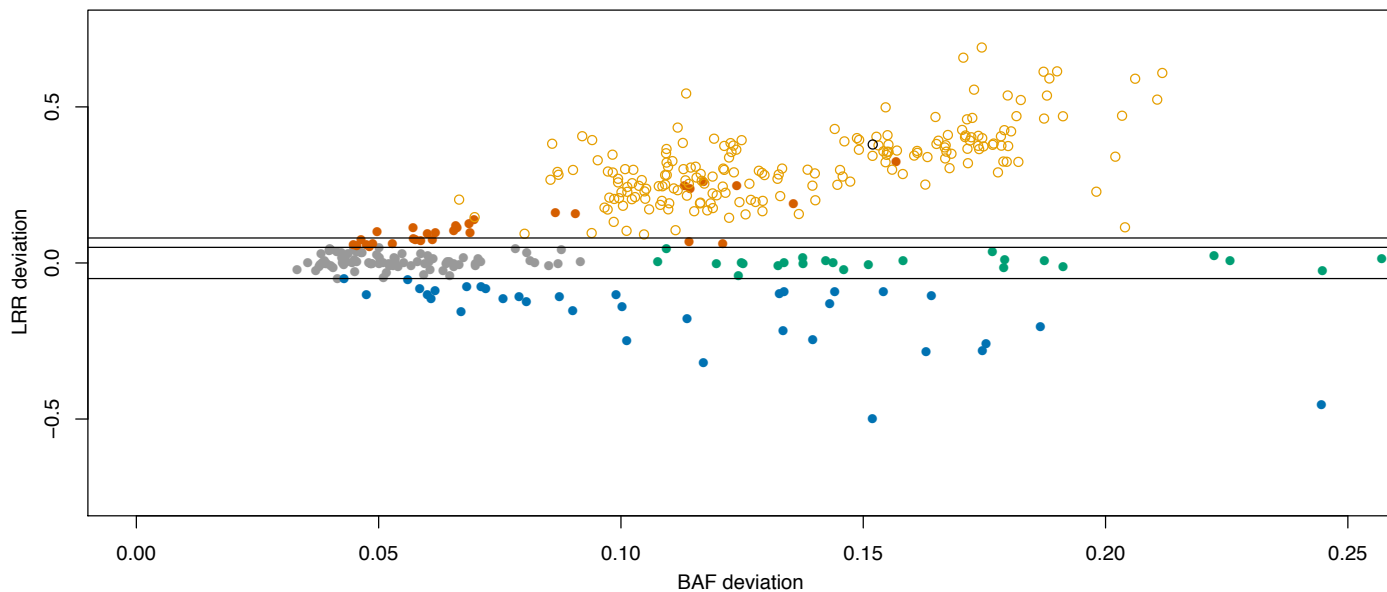
Large-scale analysis of acquired chromosomal alterations in non-tumor samples from patients with cancer

Y. A. Jakubek ^{1*}, K. Chang ¹, S. Sivakumar¹, Y. Yu¹, M. R. Giordano¹, J. Fowler ¹, C. D. Huff¹, H. Kadara², E. Vilar³ and P. Scheet¹

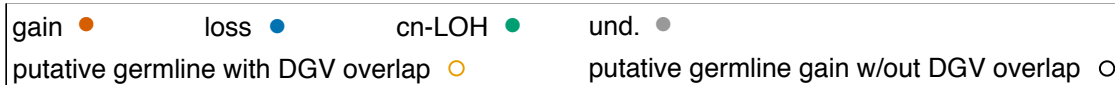
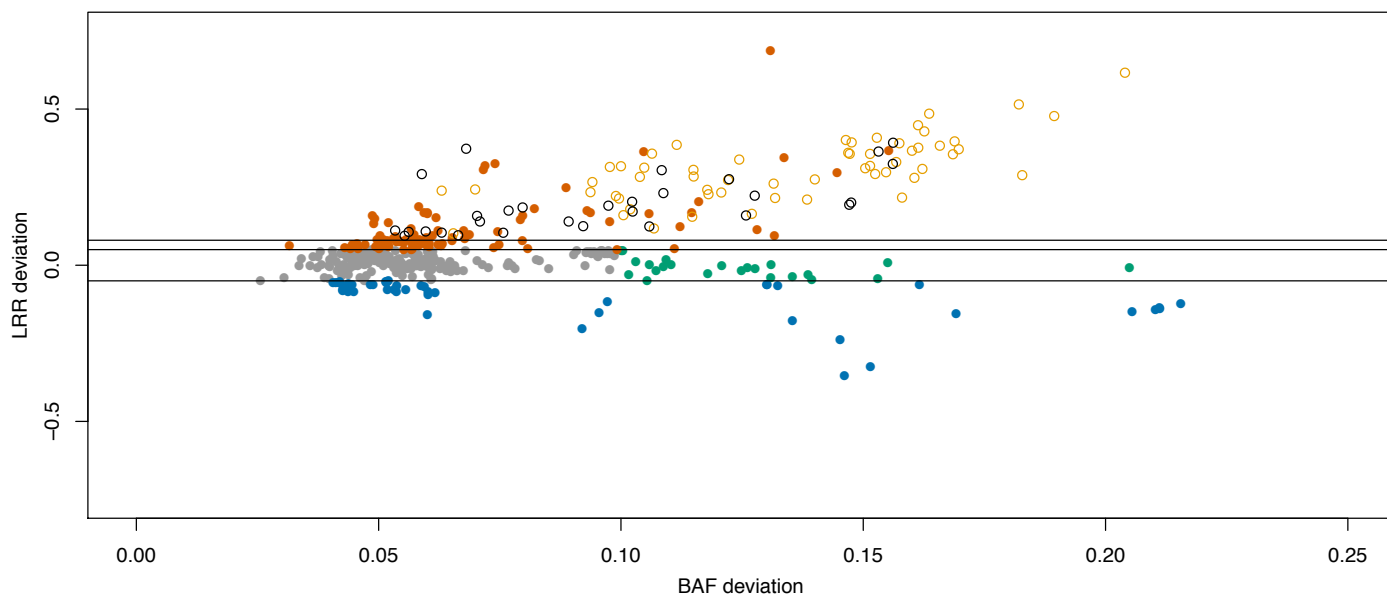
¹Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ²Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ³Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. *e-mail: yaj2@cornell.edu

Supplementary Figure 1: BAF and LRR deviation for genomic regions exhibiting allelic imbalance. Horizontal lines at LRR classification threshold of -0.05, 0.05, and 0.08. Open circles denote putative germline gains defined as allelic imbalance regions with LRR deviation > 0.08 and size < 5 Mb. Orange denotes putative germline gains that overlap with the database of genomic variants (DGV). For further information regarding these filters see Supplementary Note 1 and Methods.

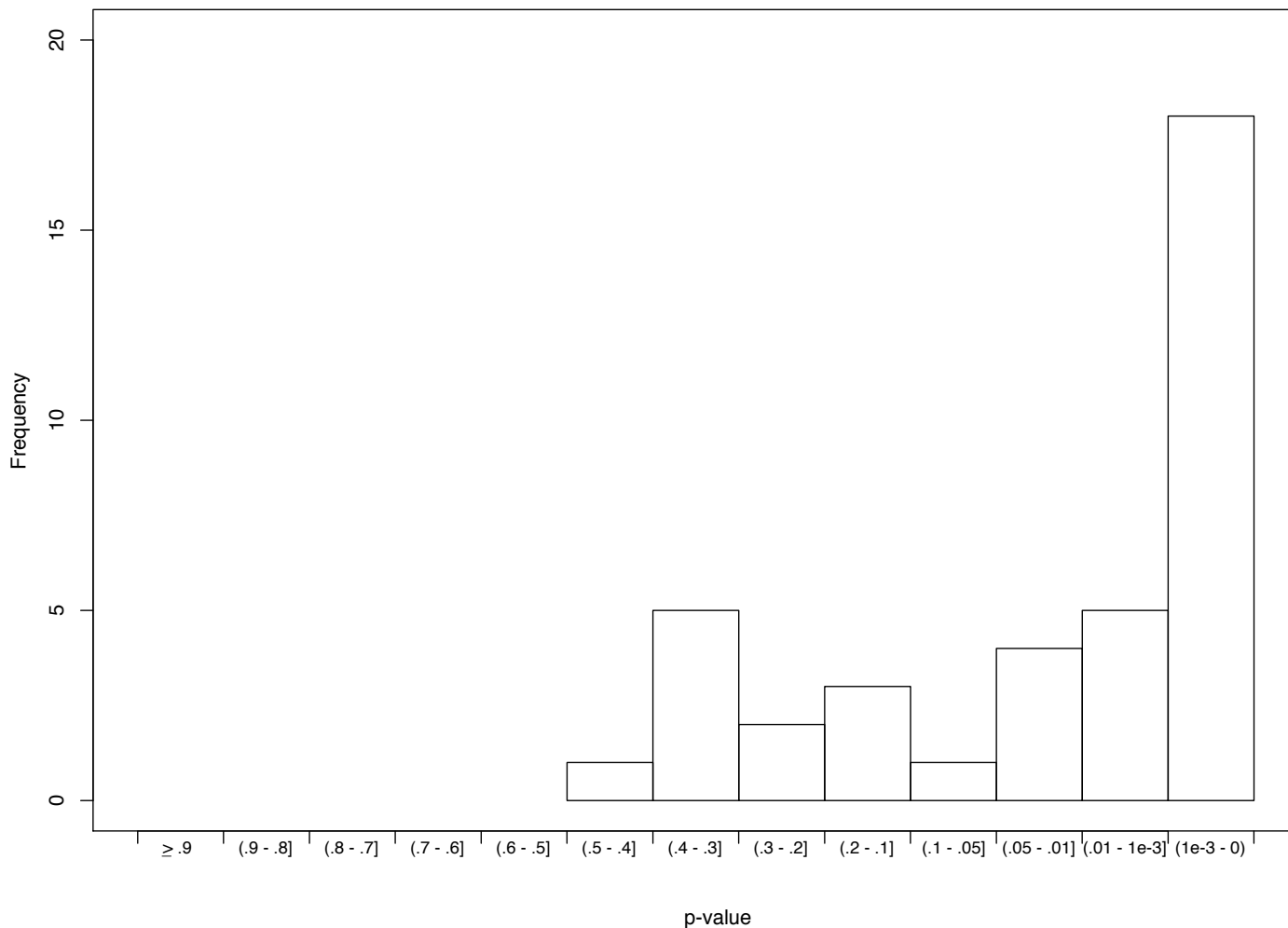
Blood



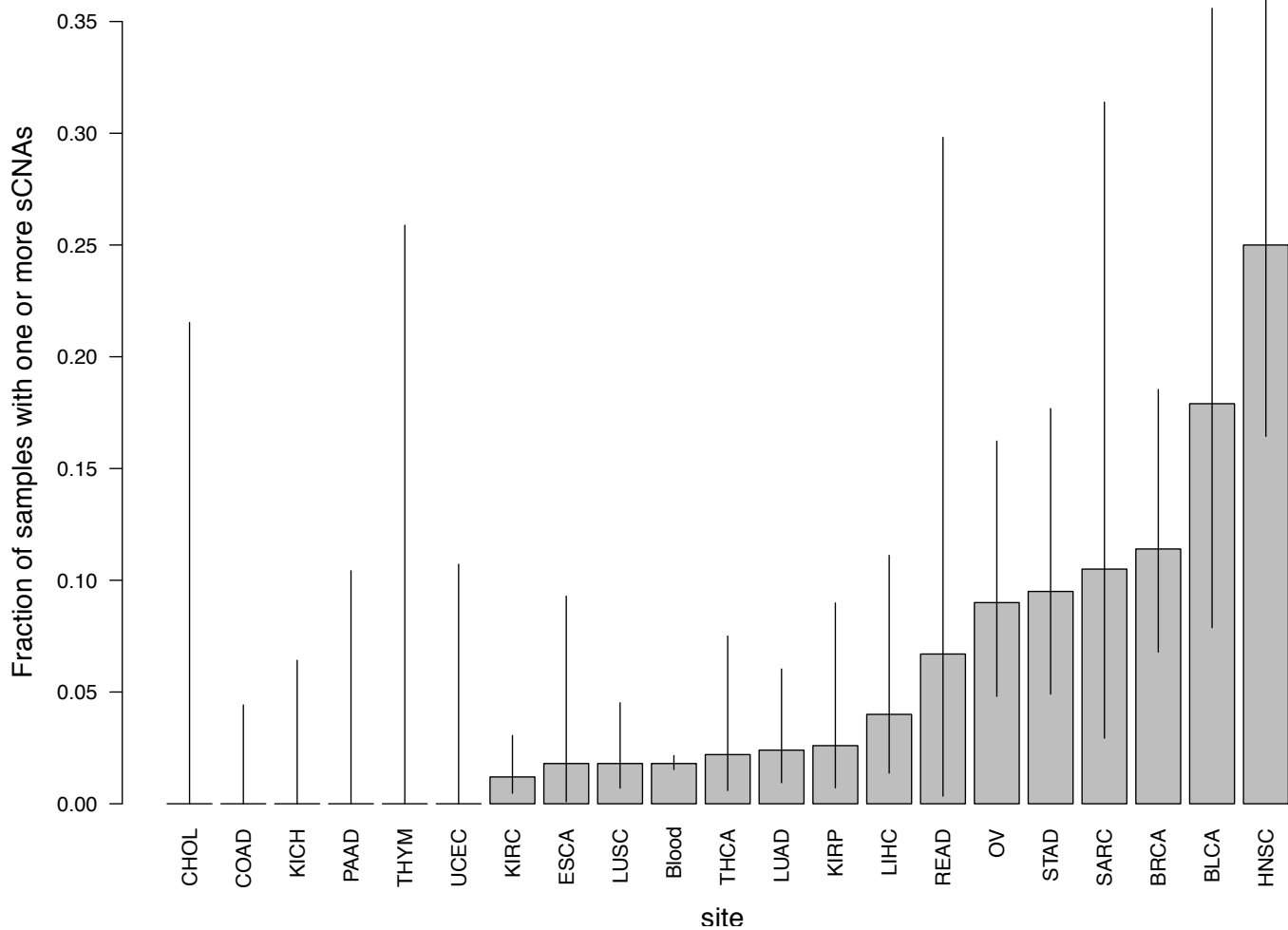
NAT



Supplementary Figure 2: Replication of sCNA calls from HNSC NAT tissues. For each genomic segment with a hapLOH call (from array) we tested for the presence of allelic imbalance in exome sequencing data, using a matched sample (with no call from array) as a control. This analysis is summarized with a histogram of p-values for each hapLOH call tested (n = 39 genomic segments, one-sided binomial test, not adjusted for multiple comparisons). 100% of the genomic segments with a hapLOH call had higher phase concordance (indicative of allelic imbalance) compared to the phase concordance of the matched genomic segment in the control sample, corroborating the allelic imbalance call.

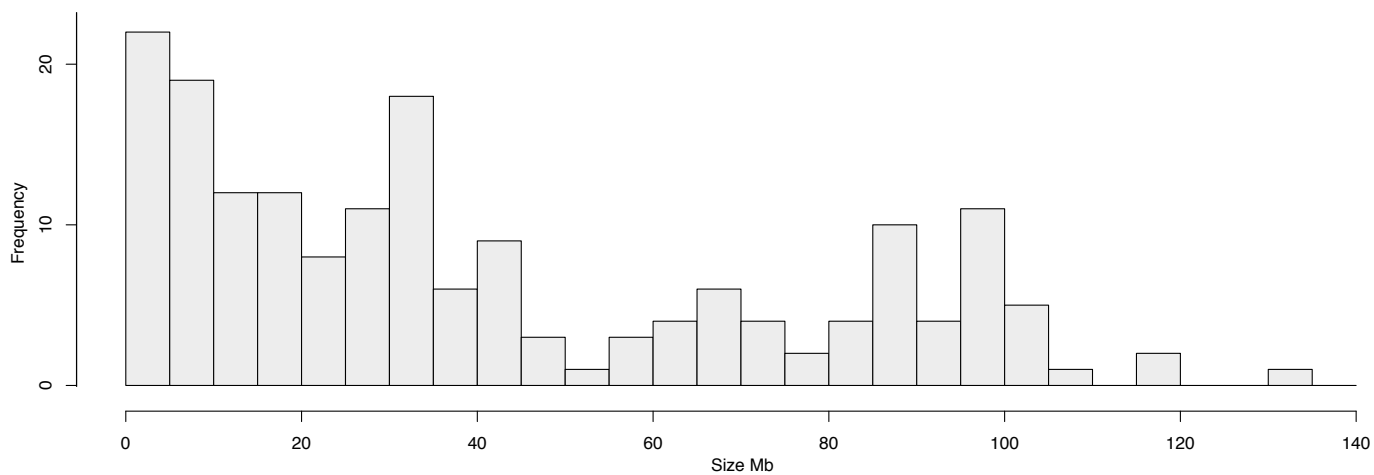


Supplementary Figure 3: sCNA rates. The fraction of mosaic NAT samples (number of samples with one or more sCNAs / number of samples surveyed) for different TCGA sites and blood. Lines indicate 95% confidence intervals (see Methods). For blood n = 7,149 samples surveyed. The number (n) of NAT samples surveyed for each cancer site is listed in Supplementary Table 3. Only TCGA sites with > 10 NAT samples surveyed are included.

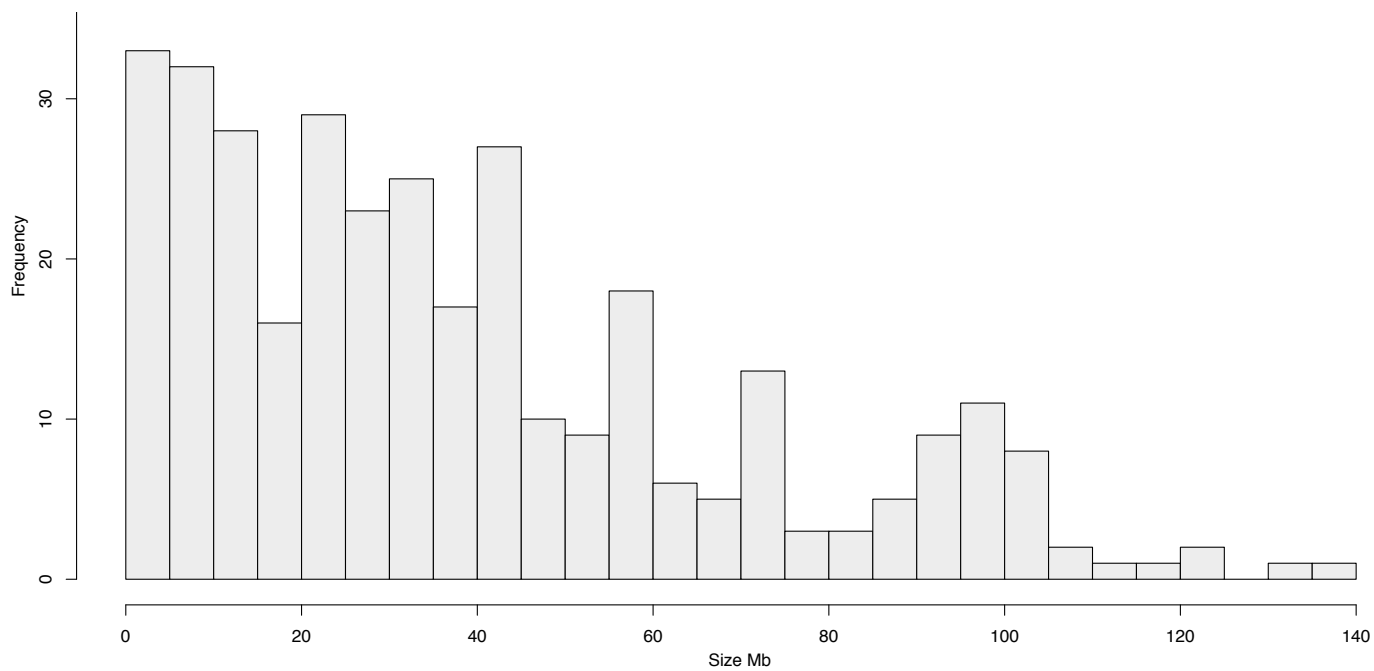


Supplementary Figure 4: Size distribution of sCNAs. Size of sCNAs in Mb.

Blood

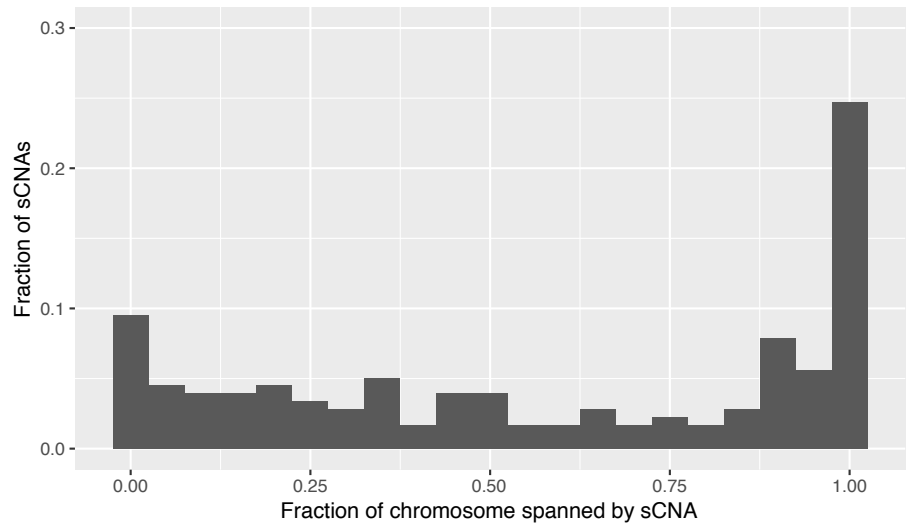


NAT

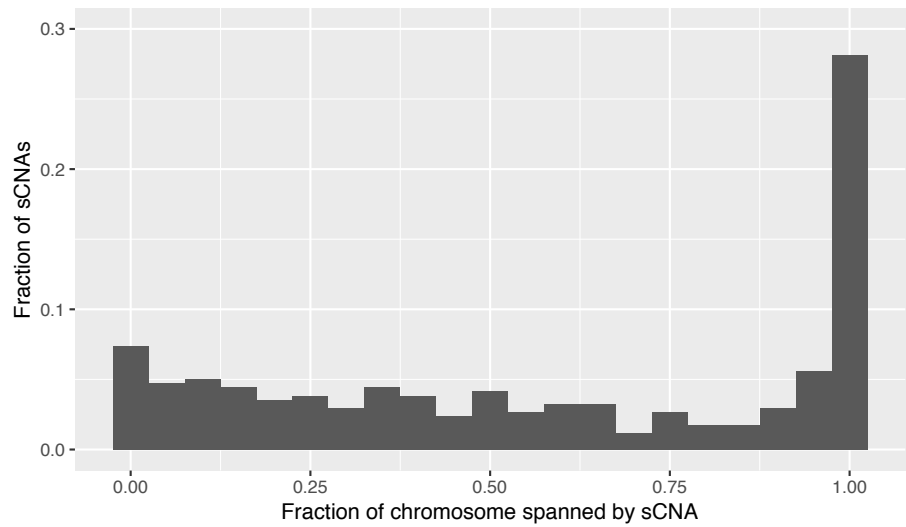


Supplementary Figure 5: Size distribution of sCNAs (fraction chromosome arm). Size distribution of sCNAs (as fraction of chromosome arm) for sCNAs detected in blood, NAT, and tumor samples from patients with detectable sCNAs in the NAT sample.

Blood



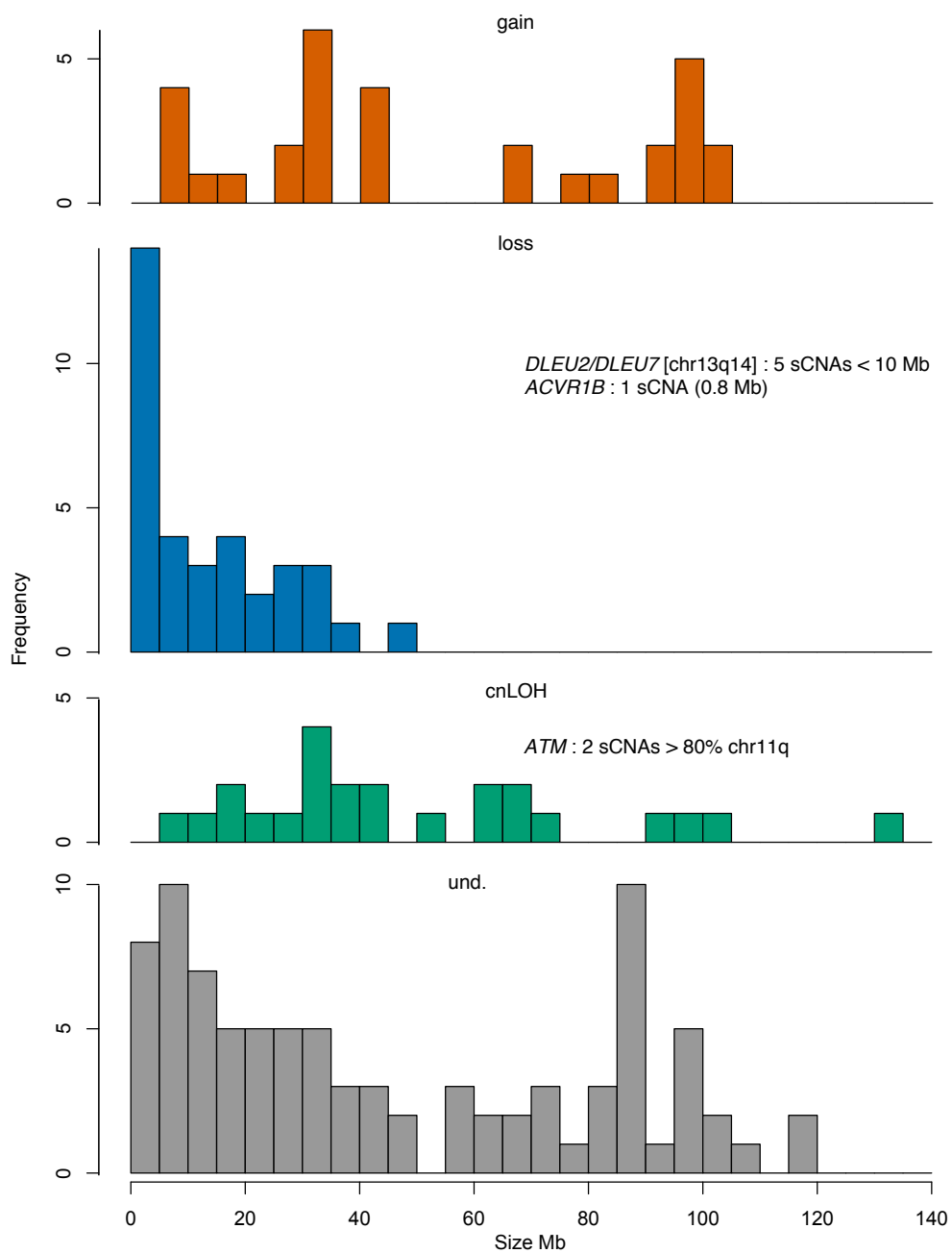
NAT



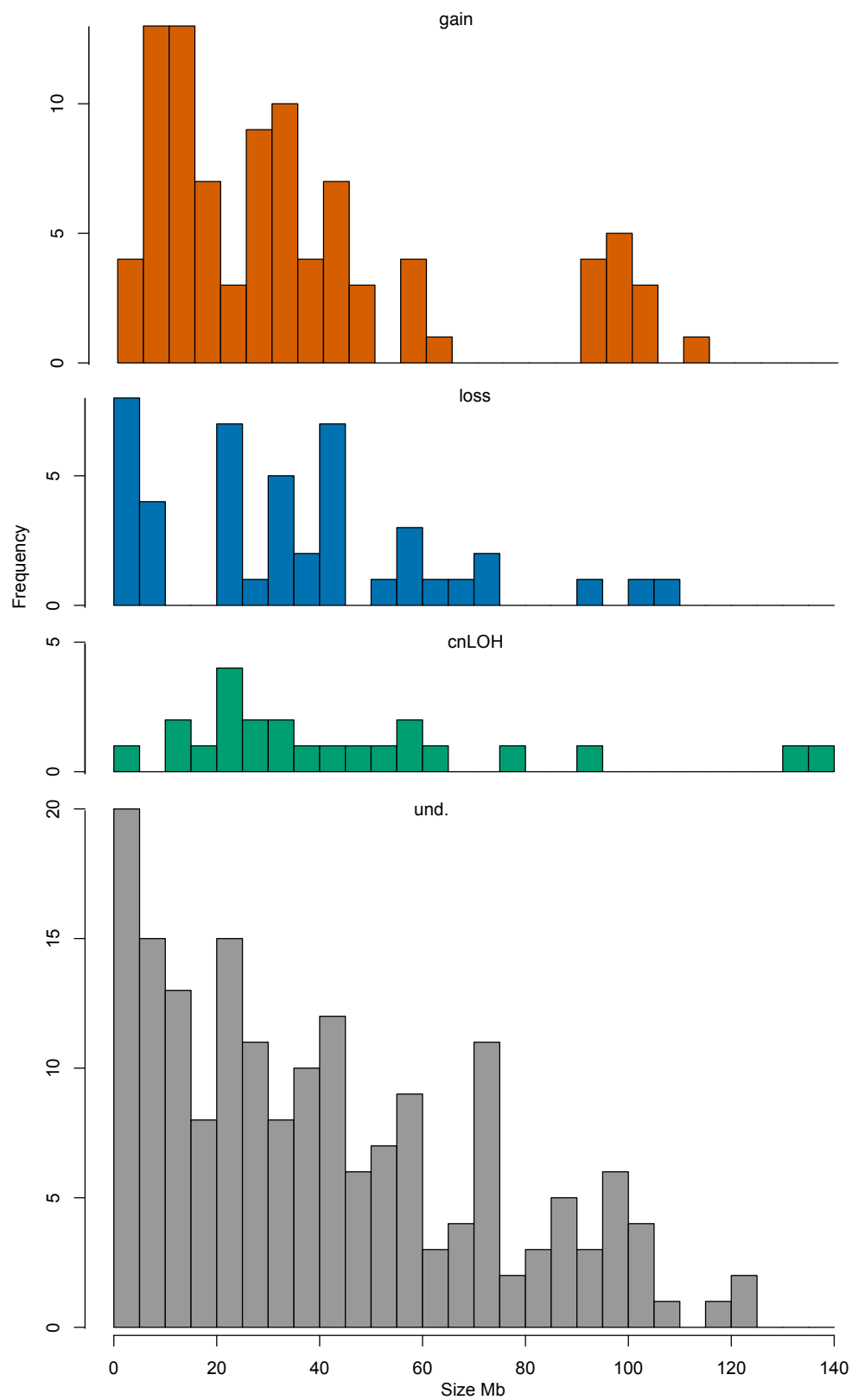
Tumor



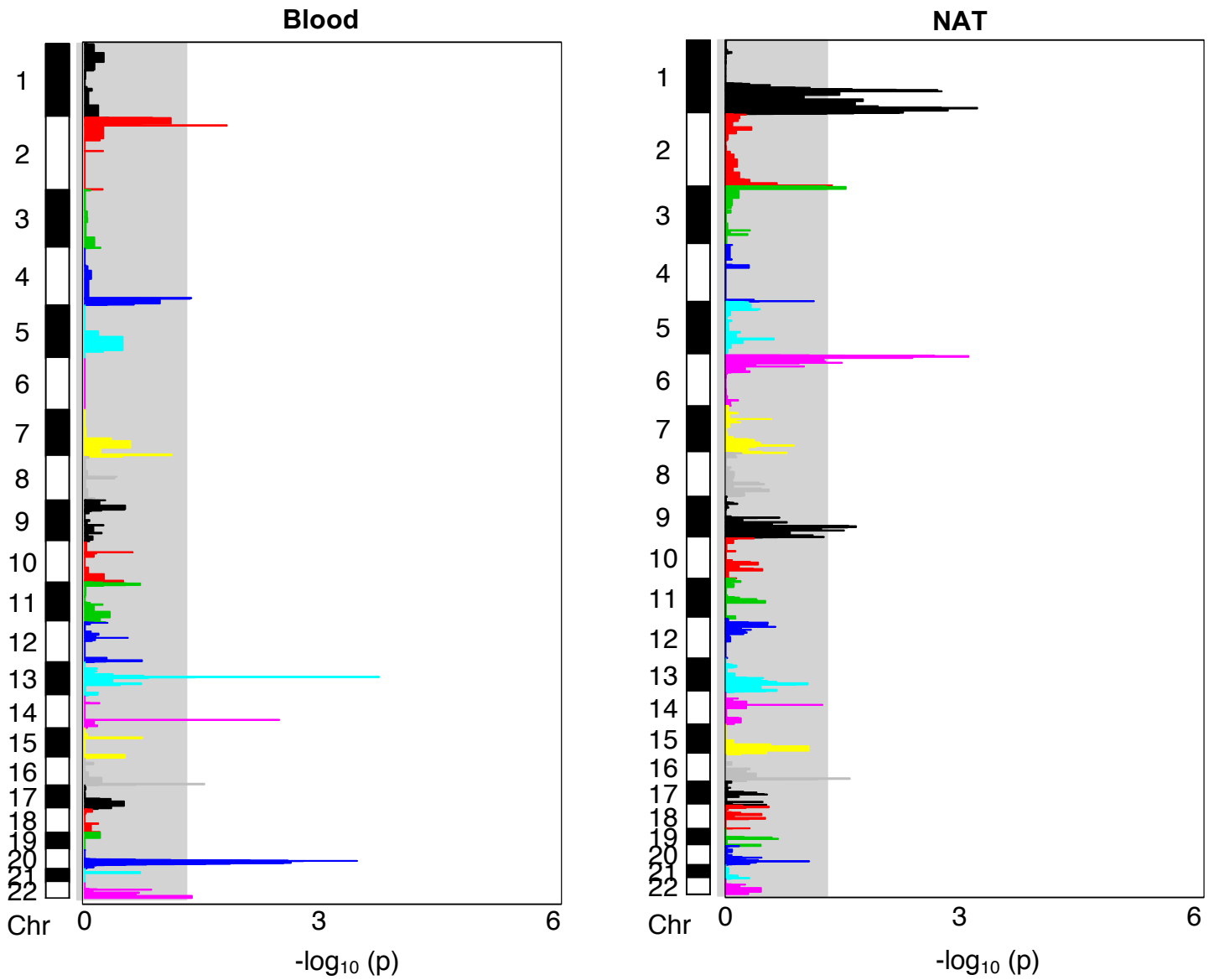
Supplementary Figure 6: Size distribution of sCNAs in blood by sCNA type. Observed gene losses and gains that have been previously reported in sCNA studies of blood are listed. Size of sCNAs in Mb.



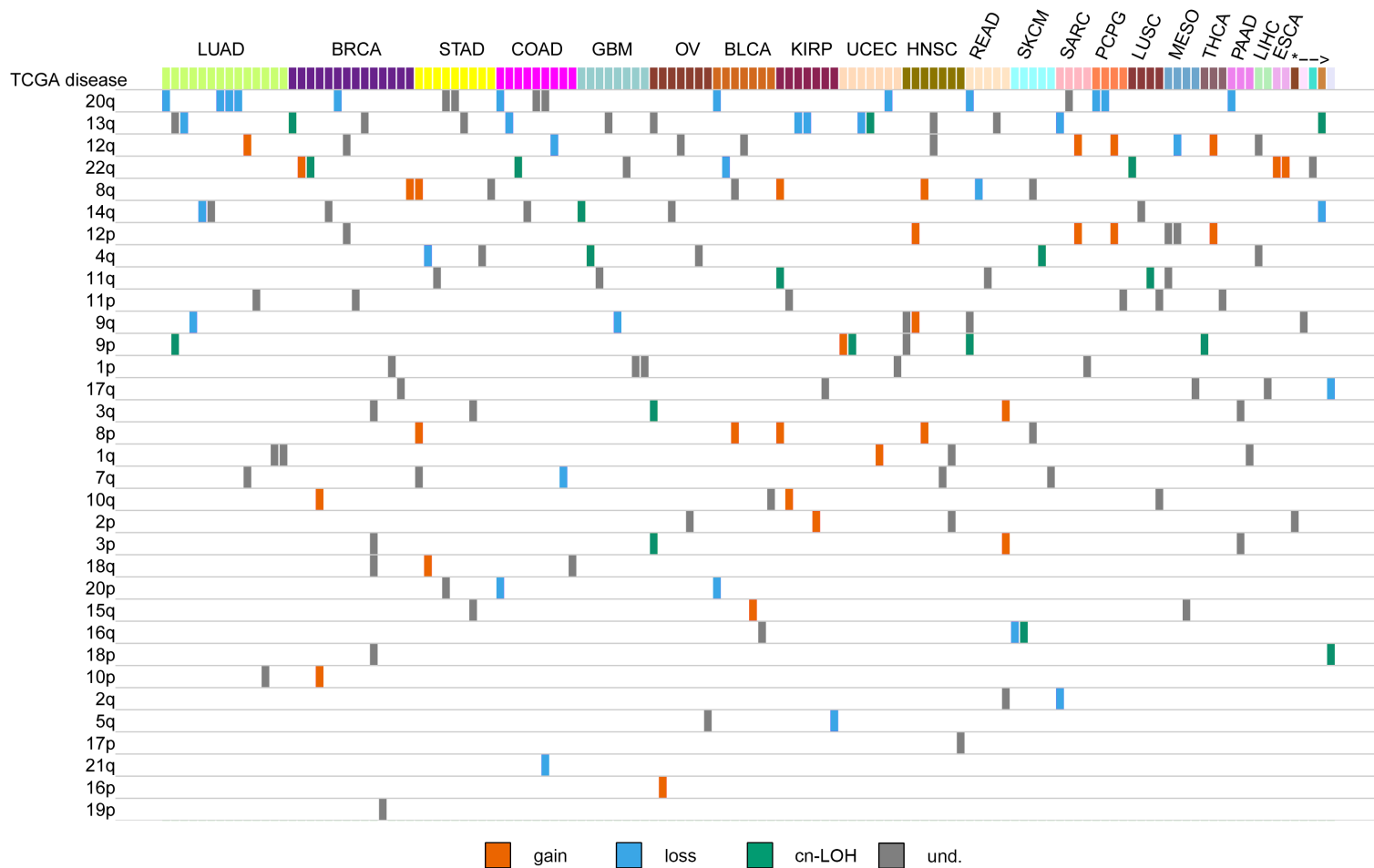
Supplementary Figure 7: Size distribution of sCNAs in NAT tissues by sCNA type. Size of sCNAs in Mb.



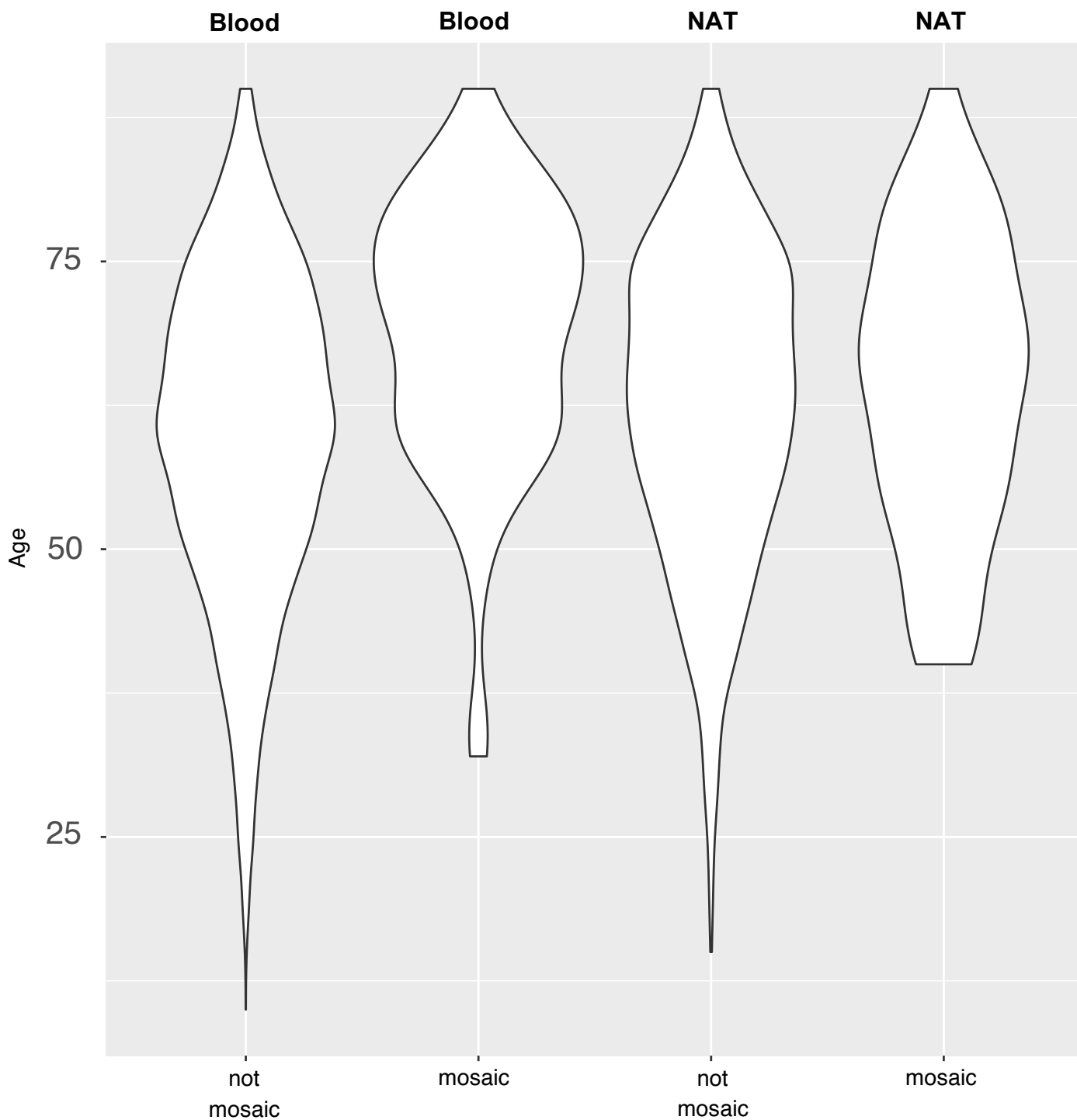
Supplementary Figure 8: Enrichment of sCNAs at genomic loci. Manhattan plot of p-values across the genome. We used a genomic random interval model to test for significance of overlap between sCNAs and genomic loci (see Reference 30). Separate tests were conducted for blood (n = 178 sCNAs modeled simultaneously) and for NAT tissues (n = 338 sCNAs modeled simultaneously).



Supplementary Figure 9: Arm-level sCNAs in blood. Each column represents one mosaic blood sample. Chromosome arms are ordered by sCNA frequency. *studies with one mosaic blood sample (UCS, KIRC, CHOL, CESC, ACC).

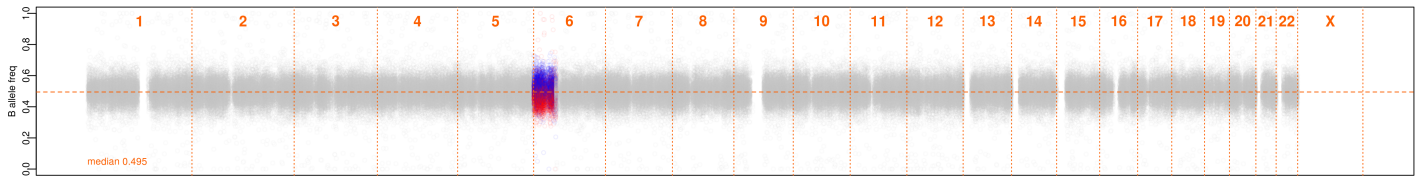


Supplementary Figure 10: Age and presence of sCNAs (mosaicism). Data shown is for patients with available age information. Mosaic is defined as a patient with one or more sCNAs. Blood mosaic (n = 129; 32, 61, 71, 78, 90), no detectable mosaicism in blood (n = 6,965; 10, 51, 61, 70, 90); NAT tissue mosaic (n = 78; 40, 55.25, 66, 74, 90) : no detectable mosaicism in NAT tissue (n = 1,630; 15, 54, 64, 73, 90) (n = number patients; min, 25th percentile, median, 75th percentile, max). Age is reported in years.

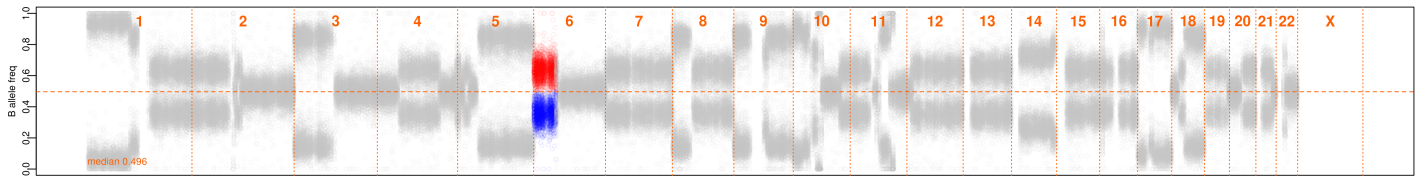


Supplementary Figure 11: Mirrored allelic imbalance in NAT-tumor pairs. Genomic regions exhibiting mirrored allelic imbalance in NAT-tumor sample pairs. BAF values for heterozygous markers that show an upwards shift in the reference sample are red and those with a downward shift are blue. The same coloring scheme is used for the non-reference sample.

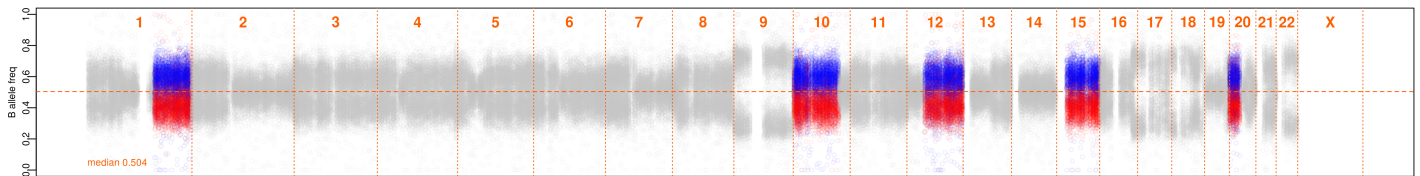
NAT - TCGA-CV-5973



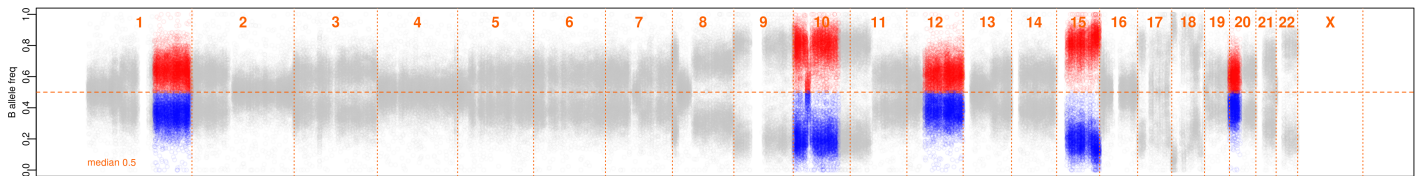
Tumor - TCGA-CV-5973



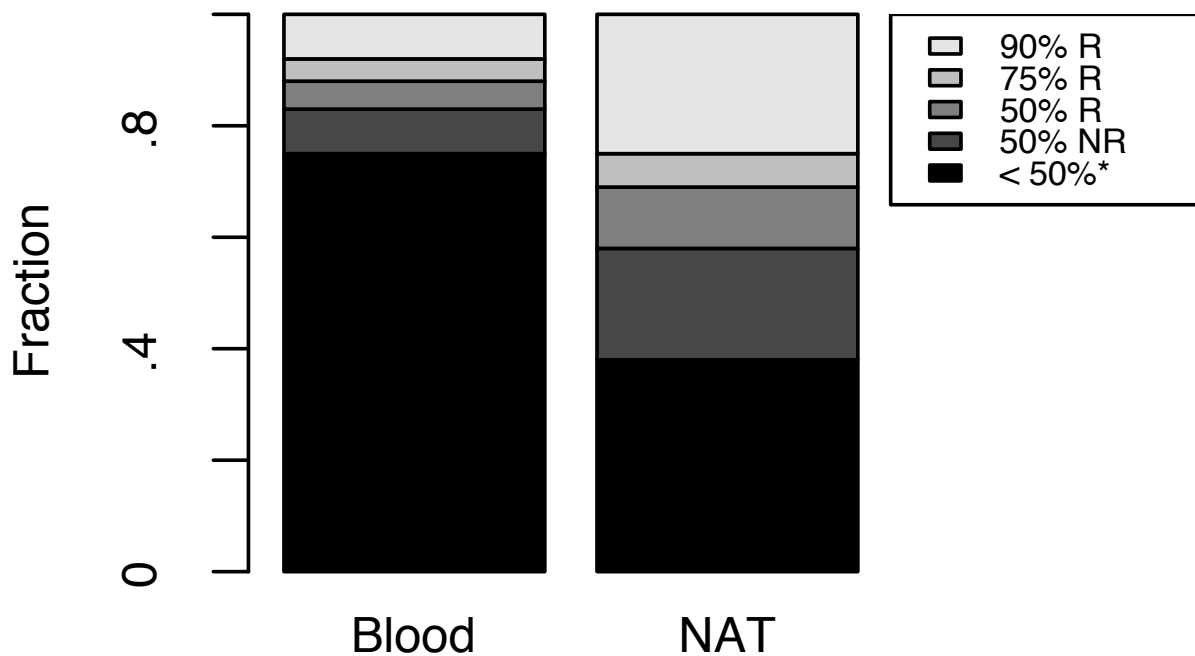
NAT - TCGA-V5-A7RE



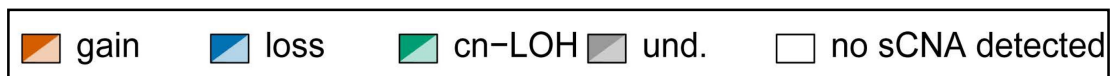
Tumor - TCGA-V5-A7RE



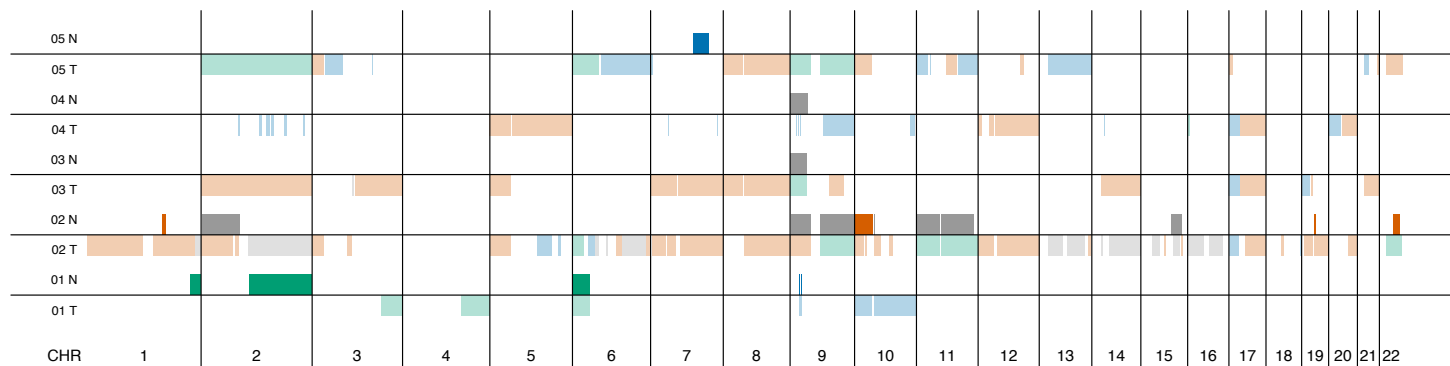
Supplementary Figure 12: Concordance for blood-tumor and NAT-tumor sCNAs. The fraction of sCNAs detected in blood and NAT tissues (concordant) that have 90%, 75%, 50% reciprocal (R) overlap with a tumor sCNA, those with 50% not reciprocal overlap (NR) and those with less than 50% overlap or conflicting sCNAs*.



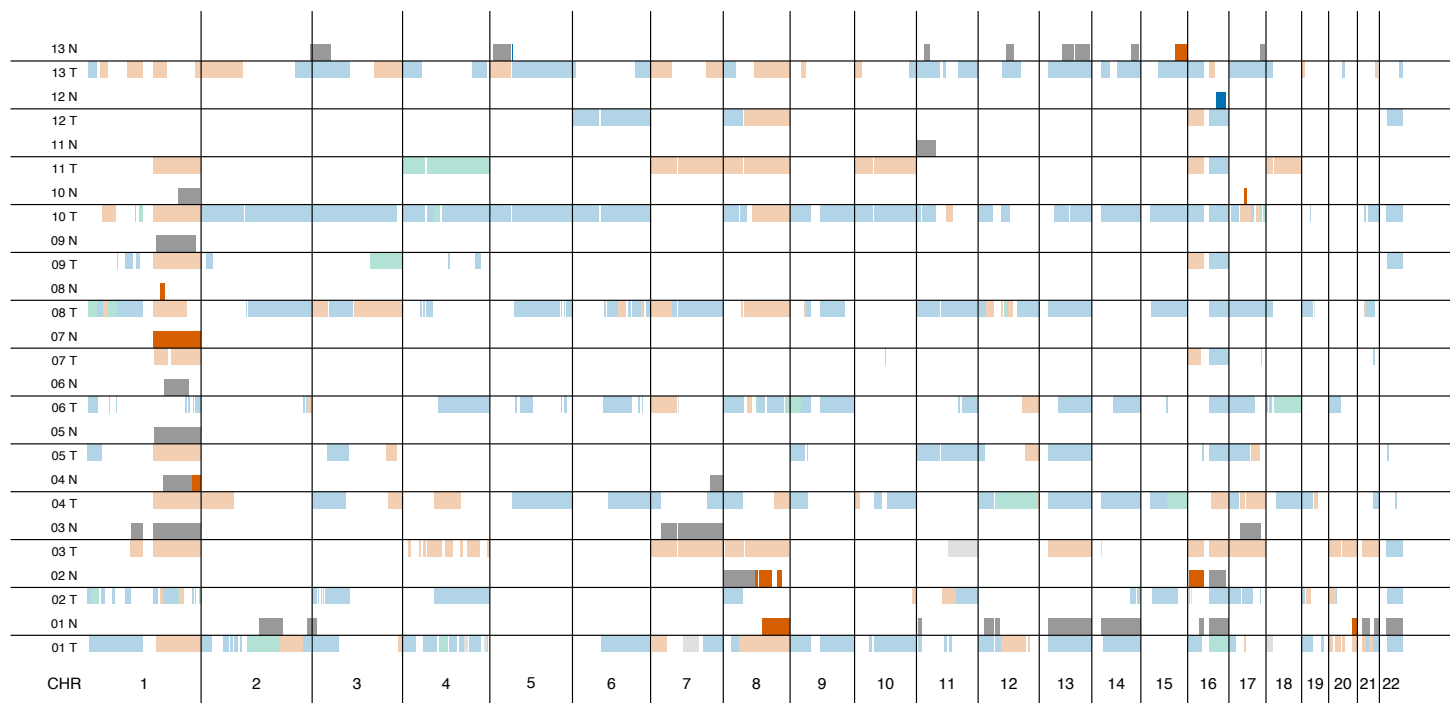
Supplementary Figure 13: NAT tissue sCNAs and tumor sCNAs across cancer sites. Pairs of NAT tissue (N) and tumor (T) sCNA profiles from the same patient. Black boxes indicate mirrored sCNAs (NAT and tumor with at least 50% overlap and opposite allelic imbalance shifts).



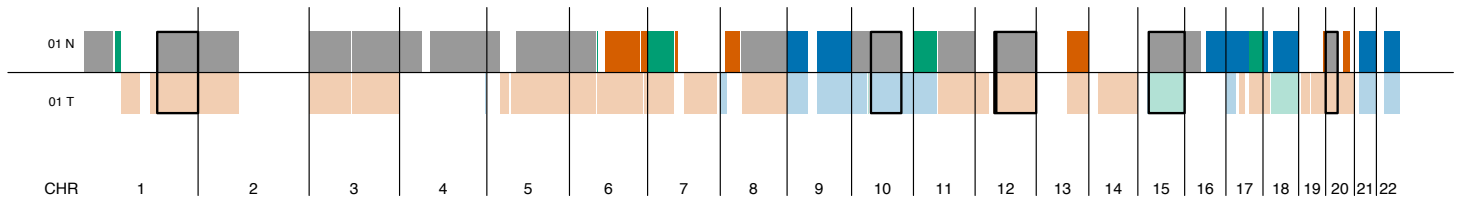
BLCA



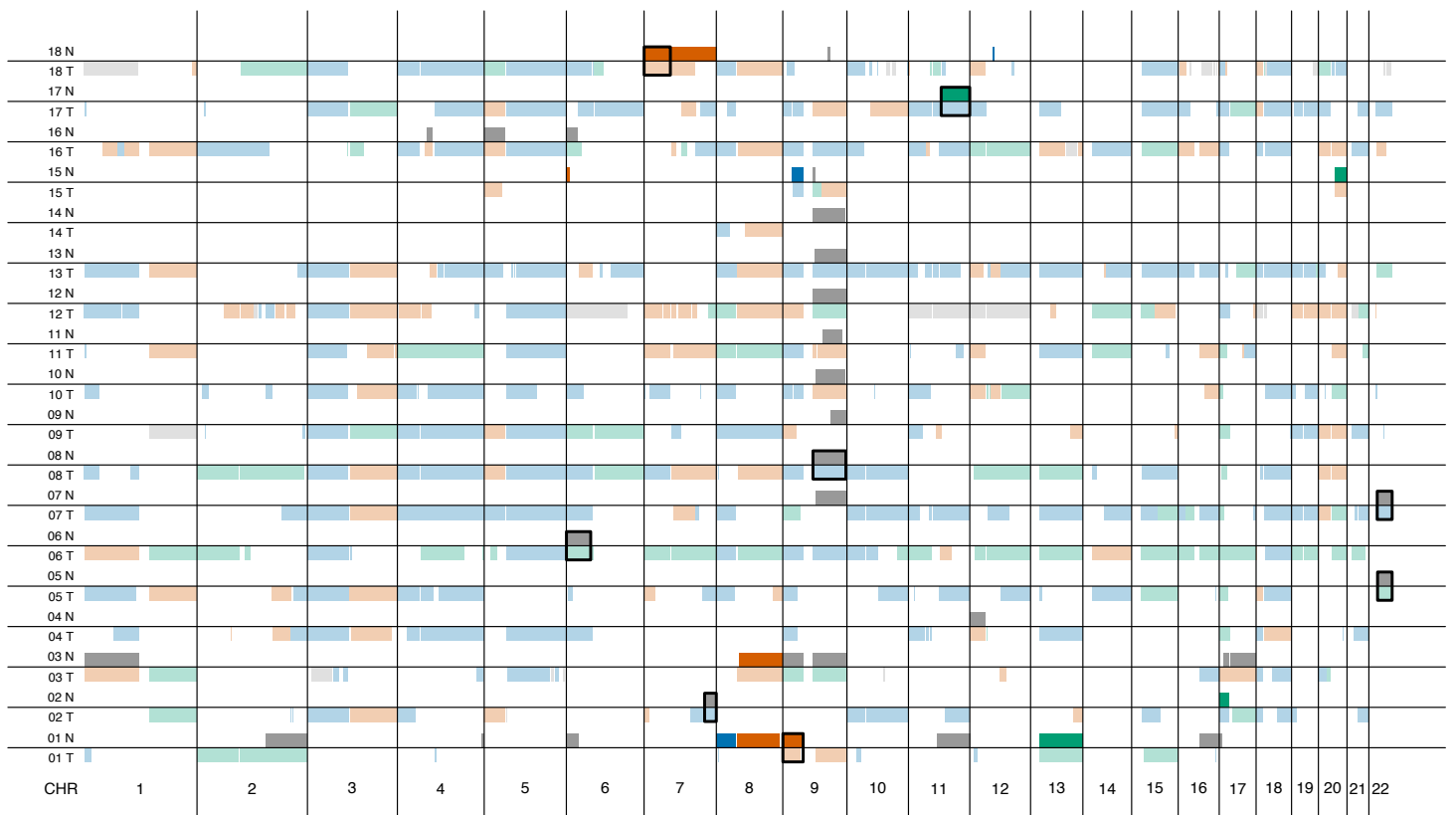
BRCA



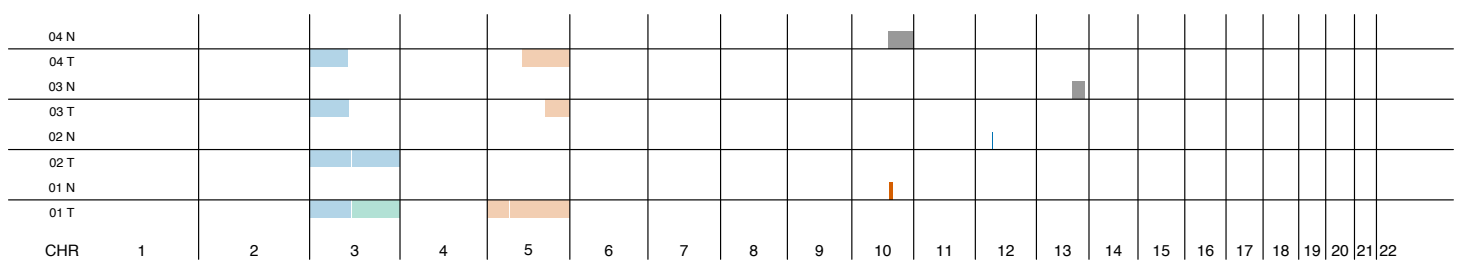
ESCA



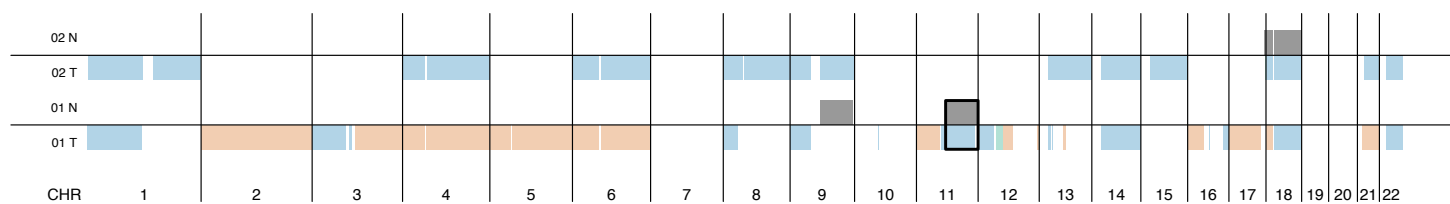
HNSC



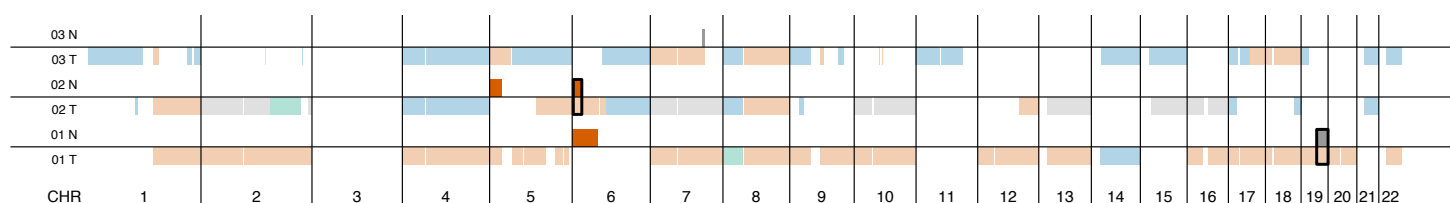
KIRC



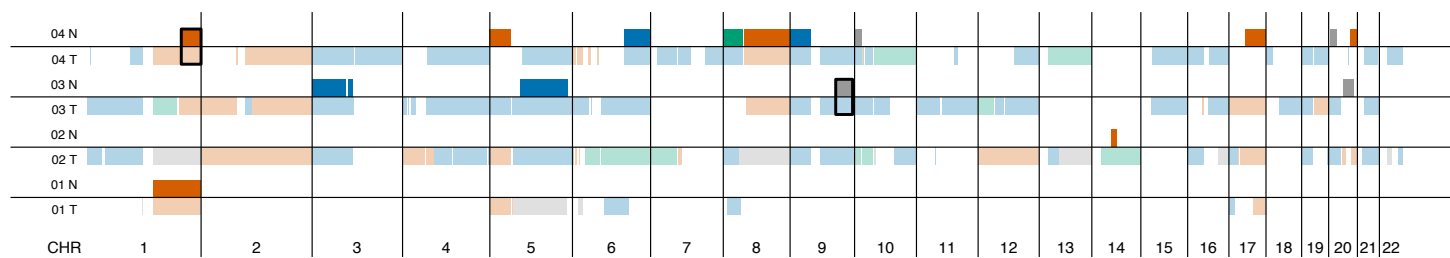
KIRP



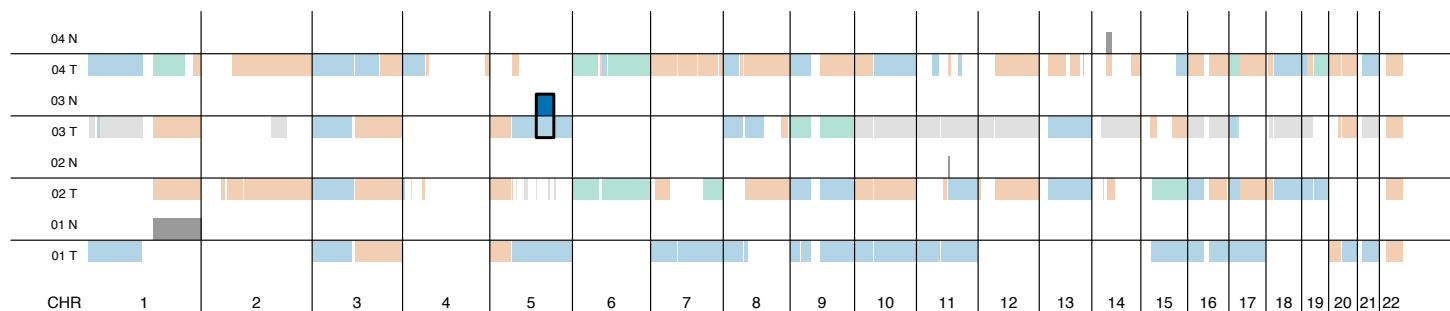
LIHC



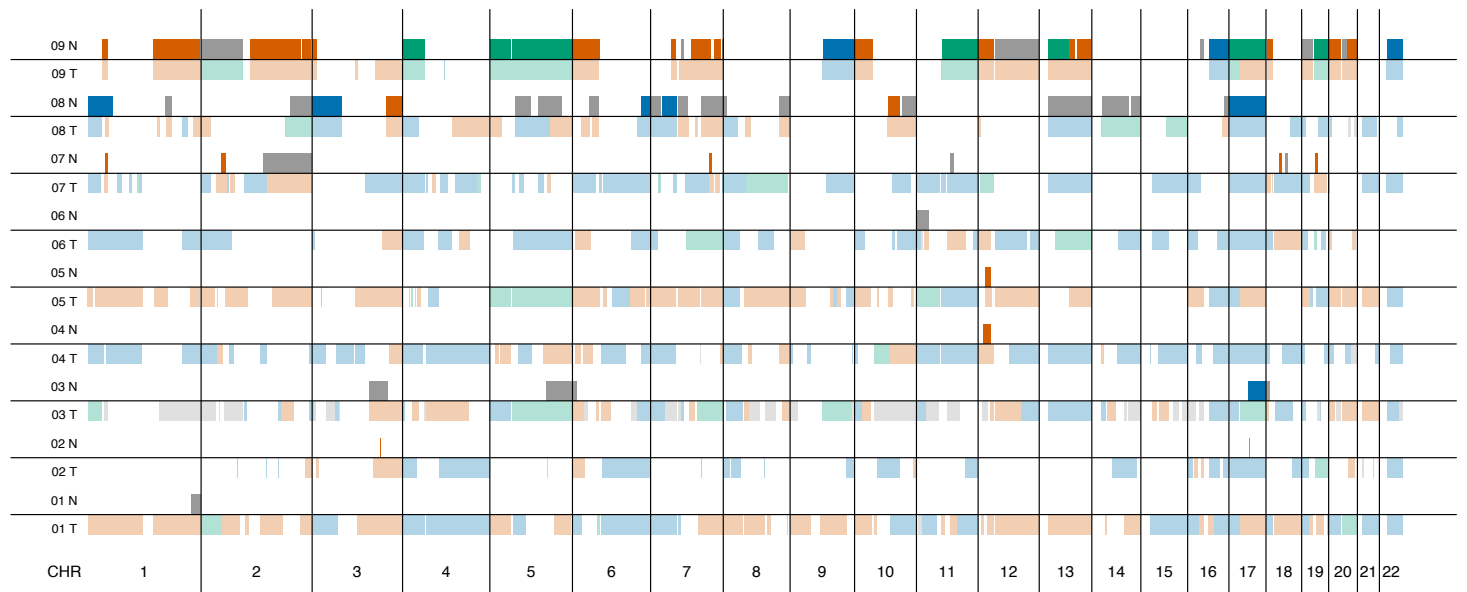
LUAD



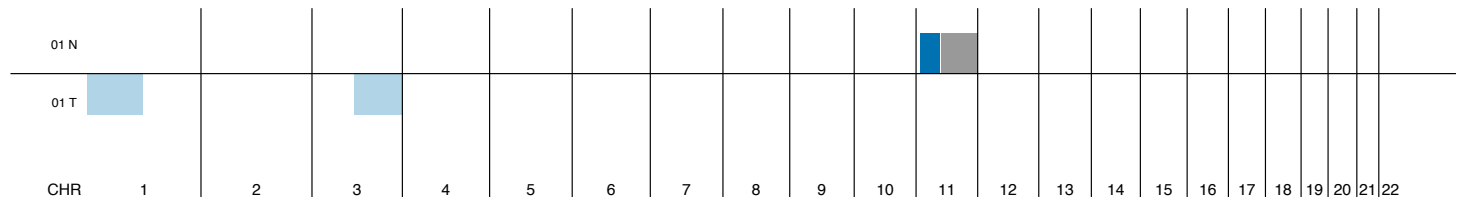
LUSC



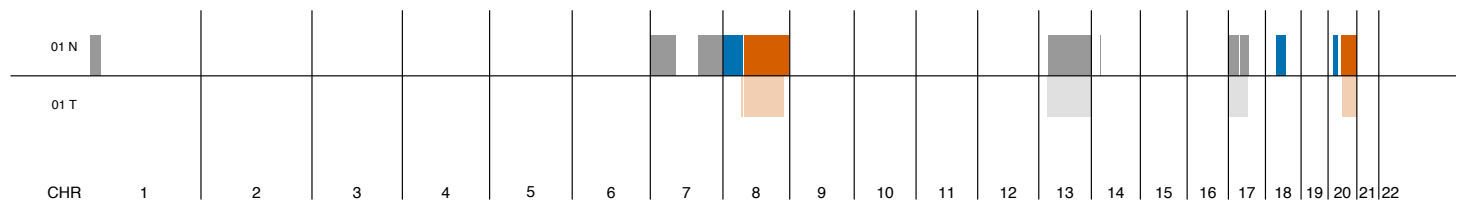
OV



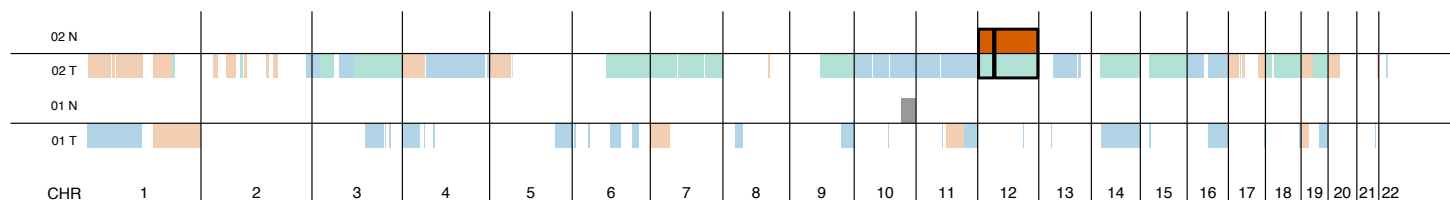
PCPG



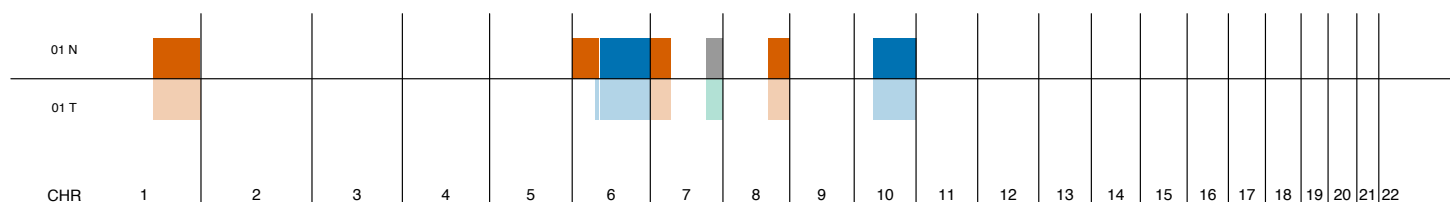
READ



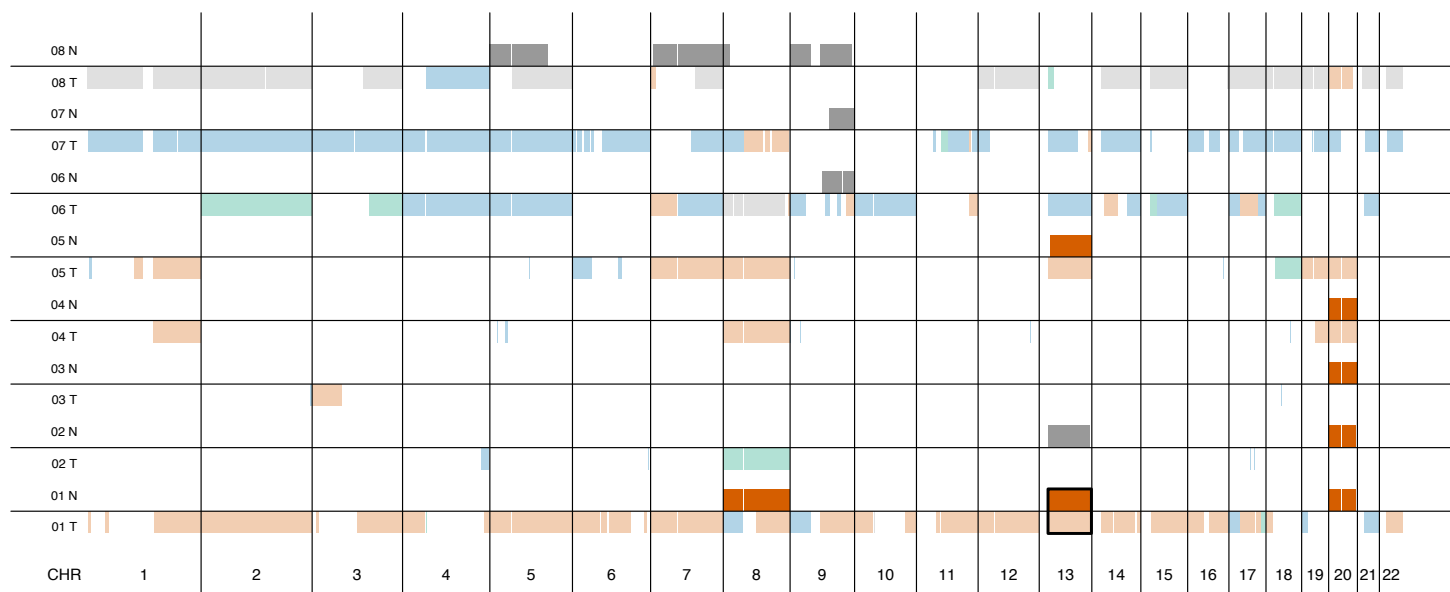
SARC



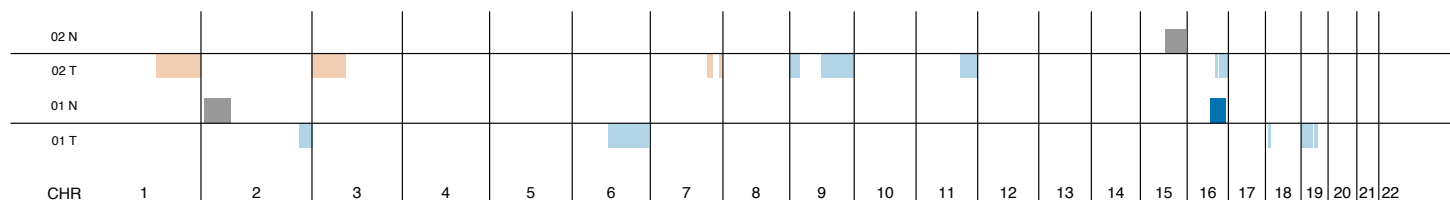
SKCM



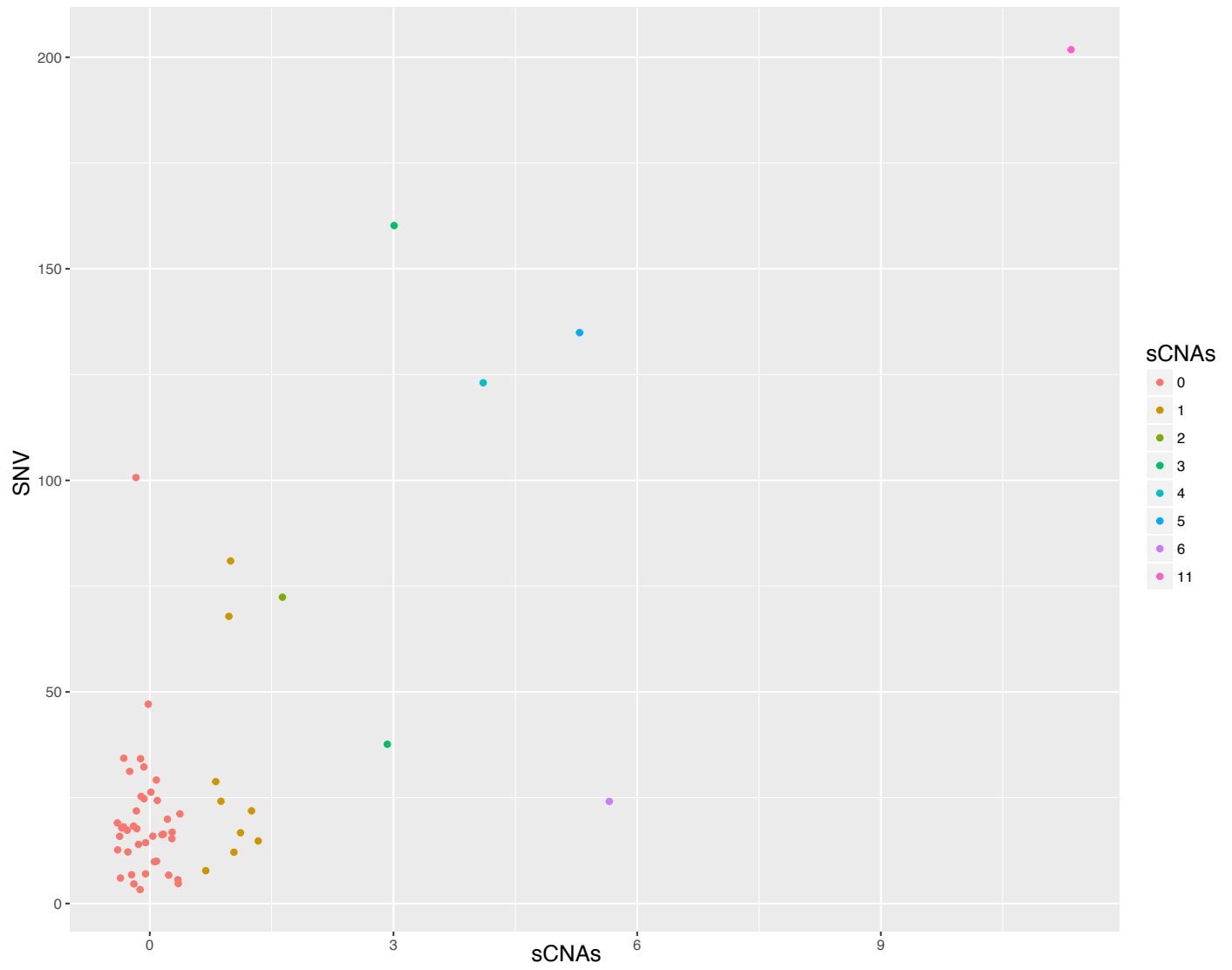
STAD



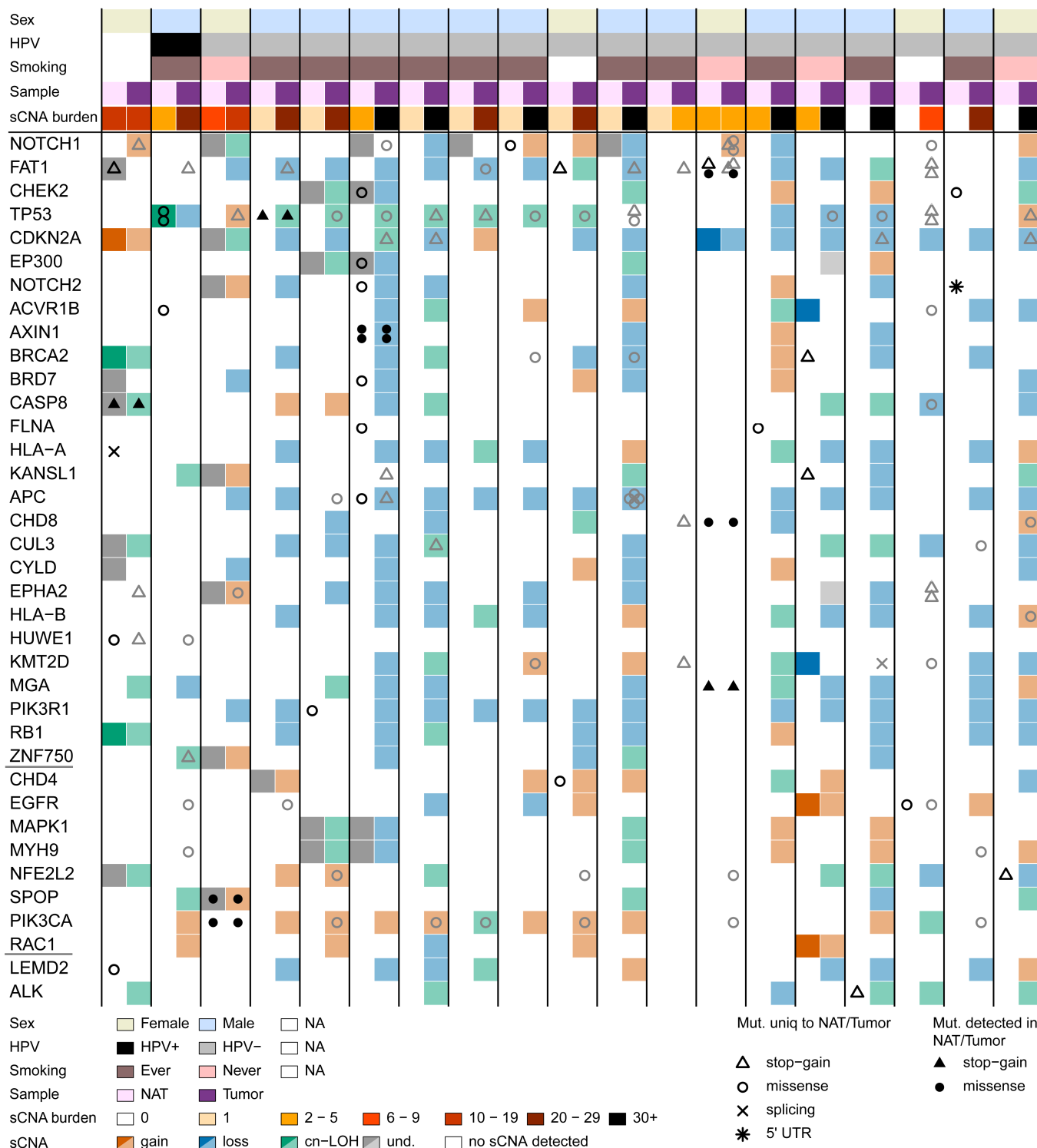
THCA



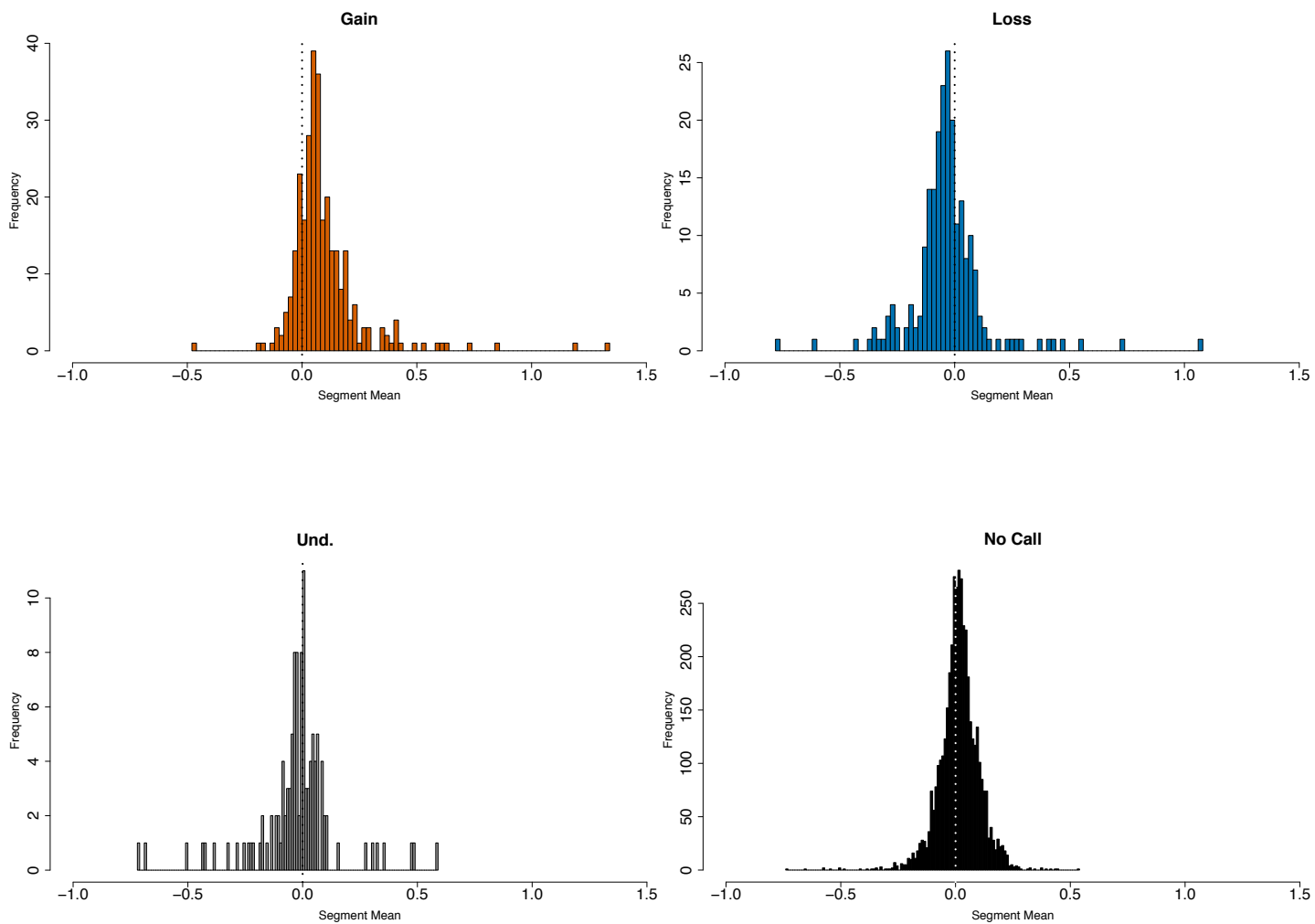
Supplementary Figure 14: Association between sCNA burden and SNV burden in HNSC NAT tissues.



Supplementary Figure 15: Somatic mutations (SNVs and sCNAs) of NAT-tumor HNSC samples. Genes appear in three groups from top to bottom 1) tumor suppressors [*NOTCH1* – *ZNF750*] 2) oncogenes [*CHD4* – *RAC1*] and 3) other [*LEMD2*, *ALK*]. Included are HNSC NAT tissues with putatively damaging SNVs in pan-cancer genes and those with SNVs/sCNAs in HNSC driver genes. One NAT tissue (TCGA-CV-7424) with a *NOTCH1* missense mutation and overlapping sCNA is not shown, because exome sequencing data was not available for the tumor.



Supplementary Figure 16: LRR segmentation of array data from mosaic NAT tissues. Distribution of segment means (LRR) for segments not spanning hapLOH calls (no-call) and those spanning hapLOH calls classified as gain, loss and undetermined.



Supplementary Notes

1- Exclusion of potentially germline gains

We filter out calls from hapLOH as potentially germline (and thus not in our analyses of sCNAs) based on the following criteria: (50% reciprocal overlap DGV and LRR > 0.05) or (LRR deviations > 0.08 and size < 5Mb); those that are excluded based on these criteria are deemed putative germline gains. Since these criteria are intentionally stringent to keep our somatic call set at a low false positive rate, undoubtedly some of the excluded allelic imbalance events will be somatic (Supplementary Figure 1). For blood, we note that one putative germline gain did not overlap with the DGV gold standard set of duplications. For NAT tissues, 29 putative germline gains did not overlap with the DGV gold standard set of duplications. The majority of these calls (55%) come from three NAT samples (TCGA-57-1586, TCGA-V5-A7RE, TCGA-57-1583). These samples had 44, 43 and 8 sCNAs that were not classified as putative germline gains by our criteria above. The other NAT tissue putative germline gains with no DGV overlap come from 11 NAT samples with one or more sCNAs that were not classified as germline; therefore, if we were to re-classify putative germline gains in NAT samples with no DGV overlap as sCNAs, this would not change mosaicism rates.

2- Association of mosaicism with clinical features

We note that analyses of survival should be viewed as preliminary due to issues arising from confounding of overall survival with age and limited/non-random missing data in the TCGA data set.

Previous studies report an association between mosaicism in blood and sex; however, we did not detect this association ($p = 0.41$ logistic regression adjusting for age). Similarly in NAT tissues, we did not detect an association between mosaicism and sex ($p = 0.48$ logistic regression adjusting for age and cancer site); however, our sample size is modest relative to those that report an association [24, 26].

The rate of mosaic NAT tissues from patients with late stage (III/IV) tumors (6%) was two times higher than for those with early stage (I/II) tumors ($p = 0.009$); however, this association appears to be driven by a positive association between the rate of mosaic NAT tissues and the proportion of late stage cases at each cancer site. When we accounted for differential mosaicism rates among malignancies, we did not detect an association

between mosaicism and stage for NAT tissues ($p = 0.31$, Supplementary Table 15). We tested for association between mosaicism and stage in cancer studies with 5 or more mosaic NAT samples (BLCA, BRCA, HNSC, STAD, OV). No association was detected in these studies (prior to multiple testing correction, Supplementary Table 15).

Data for survival analyses were obtained from the Genomic Data Commons Portal (<https://portal.gdc.cancer.gov>). The presence of sCNAs was treated as a binary phenotype. The comparison of survival between patients with mosaic NAT tissues and those without was performed using the survival analysis function (DAVE tools) of the Genomic Data Commons Portal (<https://portal.gdc.cancer.gov>). Patients with detectable sCNAs in NAT tissues had worse overall survival ($p = 1.6 \times 10^{-3}$). In order to determine if the association between survival and mosaicism in NAT tissues is driven by worse survival rates in malignancies that have higher mosaicism rates, we examined survival at sites with 5 or more mosaic NAT samples. The survival analysis at each cancer site was performed using the Mantel-Haenszel test (`survdif()`, Survival R package). In OV, the presence of mosaicism in the autosomes was marginally associated with shorter overall survival ($p = 0.052$; adj. $p = 0.26$, Bonferroni correction for 5 independent tests). No association was detected for BLCA ($p=0.809$; adj. $p = 1$), BRCA ($p = 0.277$; adj. $p = 1$), HNSC ($p = 0.168$; adj. $p = 0.84$), or STAD ($p=0.755$; adj. $p = 1$). Since OV NAT tissues are derived exclusively from females, we assessed whether the presence of sCNAs in the autosomes or X predicted overall survival and found that mosaicism was associated with shorter overall survival ($p = 0.009$, adj. $p = 0.05$, Bonferroni correction for 5 independent tests).

In blood, we observed a marginally significant association of mosaicism (autosomes) with shorter overall survival ($p = 0.089$). This result appears to be driven by age, as when we analyze survival in those 61 years and older (median age of blood donors) we do not observe an association ($p = 0.35$; survival data available for 99 donors with sCNAs present in blood and 3,497 donors with no detectable sCNAs).

We tested for association between mosaicism and smoking (never vs. ever), as well as HPV status (positive/negative), for HNSC NAT tissues. Both smoking and viral infections have been linked to HNSC. We did not detect an association between the presence of sCNAs in HNSC NAT tissues with smoking history ($p = 0.70$;

ever vs. never smoker) or HPV status ($p = 0.36$) (chi-square test with simulated p-value ($n = 1e6$), unadjusted p values).

3- Comparison of NAT-tumor sCNA profiles at each cancer site

When we contrast NAT-tumor sCNA profiles from cancer sites with 5 or more pairs, we observe that 60-89% of mosaic NAT samples in HNSC, BRCA, BLCA, and OV have one or more concordant sCNAs with the matched tumor (50% overlap threshold); the rate for STAD was 38% (Supplementary Table 17). When we compare STAD sCNA profiles, we see instances where the NAT tissue harbors a chromosome 20 gain, but the matched tumor does not (Supplementary Figure 13). The comparative analysis of STAD sCNAs also revealed a case where a mosaic 13q gain was present in the NAT tissue and tumor from the same STAD case, but directional allelic imbalance analysis revealed that the event had arisen in independent clones (Supplementary Figure 13). HNSC NAT-tumor sCNA analyses revealed independent gains of 7p and 9p and cases where 9q sCNAs in NAT tissues were not detected in the paired tumor. Of the 18 mosaic HNSC NAT-tumor pairs, 8 had mirrored sCNAs (Supplementary Figure 13). We also observed mirrored sCNAs in a Barrett's mucosa with mild dysplasia and the matched tumor (Supplementary Table 20). No mirrored sCNAs were detected in NAT-tumor pairs from BLCA, BRCA or OV. The majority of sCNAs in OV (92%) and BRCA (84%) NAT tissues were concordant (at least 50% overlap, non-conflicting event types, no evidence for mirrored allelic imbalance). Of the 8 BRCA patients with 1q sCNAs in NAT tissues, only one did not have a matching event in the adjacent tumor. We detected sCNAs in the tissues adjacent to a ductal carcinoma in situ (DCIS), referred to as stage 0 breast cancer, and the majority (8/10) of sCNAs in the adjacent tissue were concordant (Supplementary Table 19).

4- Comparison of NAT-tumor sCNA profiles from high-burden and low-burden NAT tissues

We sought to determine if a higher burden of autosomal sCNAs in NAT tissues was associated with a higher degree of concordant sCNAs. First, we stratified NAT samples into two groups: low-burden (1 or 2 sCNAs) and high-burden (3+) and assessed NAT-tumor sCNA concordance (50% overlap threshold). We observe that 93% of high-burden NAT samples have at least one concordant sCNA in contrast to 45% of low-burden NAT samples ($p = 1e-4$ chi-square, $n = 27$ high-burden, $n = 51$ low-burden samples); 67% of sCNAs in high-burden NAT

samples were concordant compared to 41% in low-burden NAT samples ($p = 1e-4$ chi-square). The sCNA concordance for low-burden NAT tissues was higher than for blood ($p = 0.03$ chi-squared test of independence).

5- Mosaicism in NAT tissues and mutational burden/MSI of tumors

In order to determine whether sCNAs in NAT tissues are associated with the mutational burden in the adjacent tumor, we tested for a correlation between NAT tissue and tumor mutations. For these analyses we used tumor mutation data, somatic point mutations and chromosome arm level copy number calls, downloaded from the Broad Firehose (<https://gdac.broadinstitute.org>). There was no correlation between the number of sCNAs in NAT samples and the number of single nucleotide variants (SNVs) ($r = -0.004$, $p = 0.86$ Pearson's correlation), or the somatic arm-level copy number changes in the adjacent tumor (tCNAs) ($r = 0.02$ Pearson's correlation $p = 0.45$). We then summarized data by cancer site, and tested for an association between the average number of sCNAs in NAT tissues (total number of sCNAs at site X / total NAT samples from site X) and the burden of mutations in tumors (median number of SNVs, as well as median number tCNAs in tumors from site X with available NAT samples) using a weighted correlation. The average number of NAT tissue sCNAs was positively correlated with median number of tCNAs ($r = 0.33$; $p = 0.10$, marginal association), but was not correlated with SNVs ($r = 0.09$ $p = 0.64$). The weighted correlation analyses were performed using a two-sided Pearson's product-moment correlation with weights equal to the number of NAT samples at each cancer site (`wtd.cor()`, `weights` R package).

We tested for an association between sCNAs in NAT tissues and the presence of microsatellite instability in the adjacent tumor. To do so we obtained MSI status available for a subset of TCGA tumors (Hause, R.J., et al., Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med*, 2016. 22(11): p. 1342-1350). We did not observe an association between MSI status in the tumor presence/absence and presence of sCNAs in NAT tissues ($p = 1$, chi-squared test with simulated p-value ($n = 1e6$)). These results are not conclusive due to modest sample size and the potential for non-random missing data.

6- Differences in sCNA rates between TCGA lung NAT tissues and previous studies

The percentages of NAT samples with detectable sCNAs were 2.4% and 1.8% for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) respectively, which are lower than those previously reported by our group (> 10%) in normal-appearing airways of non-small cell lung cancer patients [12]. However, the rate is similar to the rate reported for the uninvolved lung tissue, which served as a control in our previous report [12].

7- Specificity and Sensitivity

To assess specificity, we generated a “null” data set and sought to assess how often we detect an allelic imbalance event. First, we randomly resampled B-allele frequency (BAF) values among heterozygous sites (without replacement). Second, we applied hapLOH with the shuffled BAF values as input. There was 1 false-positive call (less than 1Mb in size) among all 1708 samples (all NAT samples that passed QC). Combined with our observation of 338 calls in the original data set, the estimated false positive rate is approximately 0.003.

To assess sensitivity, we applied BAF value shifts to capture the data perturbations reflected by chromosomal alterations of different types (gain, loss, cn-LOH), sizes (1, 10, 20, 50 Mb), and mutant cell fractions (5%, 7%, 10%, 15%, 20%, and 30%) in 300 blood samples with zero detectable sCNAs. For each sample, we simulated events by shifting BAF values according to the theoretical values generated from the formulas used to estimate mutant cell fraction (see Methods). These simulated sCNAs were randomly assigned to genomic loci, with a requirement that they fall within a chromosome arm. We simulated phasing errors by generating haplotype data with switch errors, mimicking a haplotype reconstruction accuracy of 98%, typical for modern software packages and large reference sets. We then used the simulated BAF and phase data as input. In addition, for all 300 samples, we simulated whole chromosome arm alterations (1q gain, 9p loss, and 9q cn-LOH) at mutant cell fractions of 5%, 7%, and 10%.

Sensitivity results are found in Supplementary Table 24. In summary, at 20% mutant cell fraction, hapLOH detected > 99% of chromosomal alterations of size 10Mb or larger. At 15% mutant cell fraction, hapLOH detected > 99% of alterations of size 20Mb or larger. At 10% mutant cell fraction, > 89% of alterations size 50

Mb or larger were detected. At 5% mutant cell fraction hapLOH detected large (50 Mb or larger) cn-LOH events with a sensitivity of 95%.

8- Ploidy and LRR segmentation

We used DNACopy (v 1.56.0; Venkatraman, E.S. and A.B. Olshen, A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 2007. 23(6): p. 657-63.) for segmentation of LRR values from NAT samples with autosomal sCNAs. We classify the resulting DNACopy segments (those with more than 50 markers) as gain, loss, und., or no-call based on the hapLOH output. DNACopy segments are considered as having overlap with a hapLOH call if greater than 80% of the segment spans a hapLOH call, or when the DNACopy segment spans more than 50% of a hapLOH call. Segments spanned by cn-LOH hapLOH calls are included in the no-call set, as they are not expected to result in LRR shifts. For segments classified as gains we observe an upwards shift in segment mean LRR relative to no-call and for losses we observe a downwards shift (Supplementary Figure 16).

We use ABSOLUTE (v1.0.6; Carter, S.L., et al., *Absolute quantification of somatic DNA alterations in human cancer*. *Nat Biotechnol*, 2012. 30(5): p. 413-21.) for ploidy estimates of NAT samples with autosomal sCNAs, using DNACopy segmentation data as input. We observe that the majority of samples with hapLOH calls displayed a statistically estimated ploidy of near 2 (Supplementary Table 25). None of the samples were deemed to have genome doubling.