





Genome analysis

Chromosomal imbalances detected via RNA-sequencing in 28 cancers

Zuhal Ozcan ^{1,2}, Francis A. San Lucas³, Justin W. Wong ¹, Kyle Chang¹,
Konrad H. Stopsack^{4,5}, Jerry Fowler ¹, Yasminka A. Jakubek^{1,†} and
Paul Scheet ^{1,2,*†}

¹Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, ²The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA, ³Department of Hematopathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, ⁴Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA and ⁵Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Christina Kendzierski

Received on August 3, 2021; revised on November 5, 2021; editorial decision on December 12, 2021

Abstract

Motivation: RNA-sequencing (RNA-seq) of tumor tissue is typically only used to measure gene expression. Here, we present a statistical approach that leverages existing RNA-seq data to also detect somatic copy number alterations (SCNAs), a pervasive phenomenon in human cancers, without a need to sequence the corresponding DNA.

Results: We present an analysis of 4942 participant samples from 28 cancers in The Cancer Genome Atlas (TCGA), demonstrating robust detection of SCNAs from RNA-seq. Using genotype imputation and haplotype information, our RNA-based method had a median sensitivity of 85% to detect SCNAs defined by DNA analysis, at high specificity (~95%). As an example of translational potential, we successfully replicated SCNA features associated with breast cancer subtypes. Our results credential haplotype-based inference based on RNA-seq to detect SCNAs in clinical and population-based settings.

Availability and implementation: The analyses presented use the data publicly available from TCGA Research Network (<http://cancergenome.nih.gov/>). See Methods for details regarding data downloads. hapLOHseq software is freely available under The MIT license and can be downloaded from <http://scheet.org/software.html>.

Contact: pscheet@alum.wustl.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Cancer arises as a result of a gradual acquisition of molecular alterations (Hanahan and Weinberg, 2011). Genomic instability, a hallmark of cancer (Negrini *et al.*, 2010), leads to DNA alterations, such as somatic copy number alterations (SCNAs), which may span large genomic regions or entire chromosome arms. They can play a key role in the path to tumorigenesis by leading to loss of tumor suppressor genes and/or generating additional copies of oncogenes (Knudson, 1971). SCNAs have been associated with clinical features or outcomes and serve as prognostic indicators (Hieronymus *et al.*, 2018; Liang *et al.*, 2016; Nibourel *et al.*, 2017; Ried *et al.*, 2012; Shukla *et al.*, 2020; Taylor *et al.*, 2018; Wang *et al.*, 2016; Watkins *et al.*, 2020). Hence, detection and genome-wide characterization of

SCNAs is a key component for genomic studies of tumor initiation and progression, and of SCNA-associated clinical features and outcomes.

Typically, SCNAs are almost exclusively inferred directly from DNA, measured by technologies such as array comparative genomic hybridization, single nucleotide polymorphism (SNP) DNA microarray or next-generation sequencing (NGS) (Alkan *et al.*, 2009; Amarasinghe *et al.*, 2014; Bouska *et al.*, 2014; Callagy *et al.*, 2005; Weiss *et al.*, 2004). Investigations of RNA, either by microarray or NGS [RNA-sequencing (RNA-seq)], often complement DNA analyses through quantification of gene expression, and identification of novel transcripts and gene fusions (Alexandrov *et al.*, 2013; Peng *et al.*, 2015) or point mutations (Coudray *et al.*, 2018; Griffith *et al.*, 2015; Kridel *et al.*, 2012; Shah *et al.*, 2009; Yizhak *et al.*, 2019) to

further our understanding of disease. Yet, in many settings, particularly where tumor material or funding is limited, data exist from RNA-seq only. However, the extension of RNA-seq data into SCNA calling has not been as well developed. Inferring SCNAs from RNA-seq data is inherently difficult, since both regulation of expression and underlying DNA copy number will alter the observable quantities of mRNA. In addition, due to the non-uniform coverage of the genome from RNA-seq, it is challenging to differentiate between dynamically varying gene expression and SCNAs.

Recently the relative void of methods to detect SCNAs from RNA has been partially addressed. Most of these methods are exclusively tailored to single cell RNA-seq, such as HoneyBADGER (Fan *et al.*, 2018), CopyKAT (Gao *et al.*, 2021) and inferCNV (Fan *et al.*, 2018; Gao *et al.*, 2021; Tickle *et al.*, 2019), while some can be applied to bulk RNA, such as CaSpER (Serin Harmanci *et al.*, 2020) and SuperFreq (Flensburg *et al.*, 2021). CaSpER integrates genome-wide total gene expression and allelic signals to detect and visualize SCNAs; SuperFreq also uses both read counts and BAF dispersions for SCNA inference, requiring referent samples to be available for normalization. Another approach for detection of SCNAs from bulk RNA profiling integrated coverage data and tumor-specific SCNA frequency patterns from public, external, data to identify chromosome-arm level aneuploidy, which was in turn assessed for association with prostate cancer outcomes (Stopsack *et al.*, 2019). Yet, these methods do not utilize haplotype information (the genetic makeup of a single chromosome that is passed on from a parent), which has been shown to increase power for SCNA detection in studies with SNP microarray data (Baugher *et al.*, 2013; Loh *et al.*, 2018; Sivakumar *et al.*, 2021; Vattathil and Scheet, 2013).

We sought to facilitate inference of SCNAs from RNA by applying an approach that utilizes haplotypes for SCNA detection from bulk RNA-seq, opening avenues for joint analysis of aneuploidy and expression from population-scale data. Consideration of haplotype structure implicitly models the signal at multiple genomic loci (or SNP markers) *jointly*, which not only offers an opportunity for increased power, but also requires the patterns to sustain beyond individual transcripts, which may be modulated by factors beyond SCNAs. Our approach enables robust detection of megabase-scale SCNAs that represent gain, loss or copy neutral loss of heterozygosity (cn-LOH) events. The strength of our approach derives from modeling the allelic imbalance (AI) at genomic regions affected by SCNAs. AI refers to a deviation from the expected 1:1 ratio of 'A' and 'B' alleles at germline heterozygous (genotype 'AB') loci. Alterations such as deletion (genotype: A- or B-, ratio: 1:0 or 0:1), duplication (genotype: AAB or ABB, ratio: 2:1 or 1:2) and cn-LOH (AA or BB, ratio: 2:0 or 0:2) are representative examples of AI.

In this study, we demonstrate effective somatic chromosomal copy number alteration identification from RNA-seq, comparing results to those derived from a high-density SNP DNA microarray as a benchmark and so-called 'gold standard' for SCNA detection. We consider scenarios where data are available from RNA-seq only, as well as a complementary scenario where germline DNA data is available from another source such as routinely collected blood. We apply several novel techniques including using RNA-seq for inference of acquired AI and the incorporation of genotypes via an imputation step using publicly available large-scale genotype reference data, which improves our performance considerably by enhancing the quality of estimated genotypes and haplotypes. Our results demonstrate that comprehensive and robust inference of megabase-scale SCNAs is possible from bulk RNA-seq.

2 Materials and methods

2.1 Dataset

RNA-seq BAM files aligned against the human genome build hg38 (GRCh38) and the level 1 raw CEL files from Affymetrix Genome-Wide Human SNP Array 6.0 profiling of 4942 (primary solid) tumor samples across 28 cancer sites in The Cancer Genome Atlas (TCGA) were obtained from the Genomic Data Commons data portal along with BRCA clinical information. The level 1 raw CEL files

of the matched-normal (blood) samples across these sites were also downloaded to perform genotype imputation. In addition, for a subset of 7 cancer sites (BRCA, COAD, GBM, LUAD, LUSC, PAAD and PRAD), WES BAM files of 888 (primary solid) tumor samples aligned against the GRCh38 were obtained for comparisons.

2.2 Processing of the tumor RNA-seq array data

Our method for the detection of SCNAs relies on the allele-specific signals at germline heterozygous sites. For the purpose of deriving germline genotypes, the sample can come from the tumor itself or from a matched-normal. We explored the utility of using two different sources of data for obtaining germline genotype calls: (i) tumor RNA-seq and (ii) imputed genotypes derived from SNP array data from a matched-normal, specifically blood for the samples to which we had access.

2.2.1 Genotyping and phasing

2.2.1.1 Approach 1: genotypes from tumor RNA-seq. For this approach, using tumor RNA-seq, the genotypes were called at sites already known to be polymorphic from large-scale surveys of genetic variation. The Haplotype Reference Consortium (HRC; for individuals of European ancestry) was used as a reference and genotypes were called at these reference sites from the RNA data with the UnifiedGenotyper from Genome Analysis Toolkit (McKenna *et al.*, 2010) (GATK; version 3.6). Subsequently, the genotypes were phased using the MaCH software (Li *et al.*, 2010) to reconstruct haplotypes using the set of individual-level genotypes as an internal reference. Singleton SNPs—heterozygous markers that were observed only in one sample at a particular SNP locus within a cancer site—were removed.

2.2.1.2 Approach 2: genotypes imputed. The accuracy of haplotype reconstruction increases with larger reference/internal sample size. Therefore, haplotype reconstruction accuracy is limited particularly for smaller cancer sites when using an internal reference as done in approach 1. For approach 2, we leveraged the available blood genotype data from SNP DNA microarrays, representing genotypes from the matched-normal samples of the TCGA resource (blood). After calling genotypes using the SNP array data of the matched-normal samples with the Birdsuite software (Korn *et al.*, 2008), the genotypes were prepared for imputation. To assure the quality of the genotypes submitted for imputation, several quality control steps were performed. To filter out low-quality SNPs, we removed the SNPs that failed Hardy-Weinberg equilibrium test (P -value $< 1 \times 10^{-6}$), those with missing rate $> 5\%$ and excluded monomorphic sites. In addition, samples with greater than 5% missing genotype rate were removed from downstream analyses. Individuals of European ancestry were identified using principal component analysis [EIGENSTRAT (Price *et al.*, 2006)] using the genotyped SNPs at 1KG sites with. The cleaned, unphased genotypes from individuals of European ancestry were submitted to the Michigan Imputation Server (Das *et al.*, 2016) (MIS), using hg19 (GRCh37) genome build, the HRC panel (Version r1.1 2016) as the reference, 0.1 R^2 cutoff and EUR for population. Consequently, the imputed genotypes and estimated haplotypes for 4942 TCGA samples of European ancestry from 28 cancer sites were downloaded from the MIS and markers with an $R^2 < 0.3$ were removed to remove poorly imputed markers.

In both approaches, the centromeric, human leukocyte antigen (HLA), VDJ and Database of Genomic Variants (DGV) regions were masked for exclusion of putative germline copy number changes. We examined the effects of not performing any masking step. This resulted in slightly higher sensitivities and slightly lower specificities. For example, in COAD, sensitivity is 85% (versus 79% masked), specificity is 96% (versus 97% unmasked), in LUAD, sensitivity is 86% (versus 84% masked), specificity is 87% (versus 90% masked), and in LUSC, sensitivity is 90% (versus 88% masked), specificity is 85% (versus 88% masked). As in our approach, we are not modeling explicitly the repeat-rich nature of these regions, we determined it was a better practice in general to exclude them up front.

Table 1. Gene-level performance assessment

	BRCA		COAD		GBM		LUAD		LUSC		PAAD		PRAD	
	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)	Sens (%)	Spec (%)
hapLOHseq (RNA-seq)	71	94	47	99	79	93	74	91	73	89	69	97	45	97
hapLOHseq (RNA-seq + imputed genotypes)	84	92	79	97	89	92	84	90	88	88	80	95	66	94
hapLOHseq (WES)	93	89	94	92	94	96	89	92	91	90	94	89	76	97

Note: We evaluated the method at the gene level by comparing SCNA status of genes between the RNA-seq-derived SCNAs and the gold standard (array-based analysis) for seven cohorts in the TCGA. Sens, sensitivity; the proportion of genes covered by an SCNA in the gold standard that were also identified by the listed approach. Spec, specificity; the proportion of genes that are not covered by an SCNA event in the gold standard that were also not inferred to be covered by an SCNA by the listed approach.

2.2.2 Detection of SCNAs

As noted earlier, we investigate two approaches that differ on how the germline haplotypes are obtained. While the first approach uses tumor RNA to statistically estimate haplotypes, in approach 2, the phased germline genotypes are obtained through genotype imputation performed on a matched-normal from SNP array. hapLOHseq (San Lucas *et al.*, 2016) algorithm identifies the haplotype in excess and using the germline haplotypes, quantifies phase concordance via switch accuracy to determine genomic regions that harbor SCNAs. hapLOHseq software (version 0.1.2) was applied using the default parameters in both approaches, except `-end_param_event` (=0.9) and `-event_prevalence` (=0.05). SCNA calls with <10 markers or those smaller than 2 Mb were excluded. Furthermore, SCNA calls that contain large (>10 Mb) genomic regions without any heterozygous sites were split up into multiple regions.

2.3 Processing of the tumor DNA WES data

To identify SCNAs from WES tumor data, the genotypes were called at the HRC sites with UnifiedGenotyper first. Second, the genotypes were phased with MaCH software using the set of individual-level genotypes as an internal reference. Third, hapLOHseq software was used with the default parameters to detect the SCNAs after masking the centromeric, HLA, VDJ and DGV regions and removing singleton SNPs. SCNAs with <10 markers or those smaller than 2 Mb were excluded and the calls that contain large (>10 Mb) regions without any heterozygous sites were split up into multiple regions.

2.4 Processing of the tumor SNP array data

To detect SCNAs from SNP array tumor data, first the genotypes were called using the Birdsuite software, second the genotypes were phased using MaCH software, and third hapLOH software was used with the default parameters to identify regions that harbor SCNAs. Prior to the third step, the markers were mapped from genome build hg19 to hg38 and the centromeric, HLA, VDJ and DGV regions were masked. The SCNA calls detected from the SNP array constitute a gold standard for assessing the performance.

To ensure the consistency in the way the samples were processed, SyQADA (Fowler *et al.*, 2019) was used to automate the pipelines across the three platforms.

2.5 Performance assessment

We sought to assess the performance of our method for detection of SCNAs in tumors. To do so, we compared our set of SCNA calls to a gold standard set of SCNA calls from matched-tumor DNA samples processed using arrays, a gold standard set of calls. We contrasted the SCNA call sets at gene-, chromosome arm- and genome levels. At the gene level, we report sensitivity and specificity. Sensitivity represents the method's power to detect true SCNAs that are identified by the gold standard, while specificity represents the method's ability to correctly identify genes that do not fall within an SCNA region in the gold standard. Therefore, sensitivity (TPR) was calculated as $TP/(TP + FN)$ and specificity ($1 - FPR$) was calculated

as $TN/(TN + FP)$ where TP is true positive, FN is false negative, TN is true negative and FP is false positive. For each sample, sensitivity and specificity were calculated individually, then median sensitivity and specificity for samples in each TCGA cohort were reported as cohort-level summary statistics. When assessing the method's performance at the chromosome arm level, for each sample, we assessed presence or absence of a chromosome arm-level event, defining an arm-level event as present when at least 50% of the chromosome arm is affected by SCNAs. At the genome level, we calculated genomic burden for each sample, which reflects the percentage of a sample's genome that exhibits SCNAs. For each cancer site independently, we calculated each sample's genomic burden based on RNA-seq-derived SCNA calls and compared with the gold standard derived genomic burden.

2.6 Comparison to other methods

We followed the recommended workflow for SuperFreq of first applying VarScan2 (Koboldt *et al.*, 2012) for variant identification, followed by SuperFreq itself which is in R. RNA-seq from two adjacent-to-tumor breast samples in the BRCA resource were supplied to SuperFreq for normalization. Results from the *TP53* analysis of COAD samples were taken directly from their curated data in their GitLab page. Results from CaSpER were obtained directly from what they had curated previously, using their stored and available R data frames. We then summarized results by gene through direct tabulation or applying BEDTools intersect.

3 Results

To detect somatic (acquired) copy number alterations (SCNAs) using RNA, we applied a haplotype-based approach. In brief, hapLOHseq detects regions of the genome where the signal at heterozygous sites reflects one of the estimated haplotypes for that individual. A deviation from the expected 1:1 ratio of maternal to paternal DNA indicates a relative over-representation of one of the parental chromosomes, signaling the presence of an SCNA. This approach has been applied successfully to bulk DNA analyses of various tissues (Jakubek *et al.*, 2020; Loh *et al.*, 2018; Vattathil and Scheet, 2016).

Here, we explore the potential of this method for detection of large-scale SCNAs from NGS data of bulk RNA (RNA-seq). To do so, we obtained RNA-seq from seven large cancer sets in TCGA: Breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD) and prostate adenocarcinoma (PRAD). To identify genomic regions that harbor SCNAs, we applied hapLOHseq. To assess the accuracy and potential of this approach, we compare SCNAs inferred from RNA-seq to high-confidence SCNA calls detected from DNA SNP microarray data, which have been documented previously (Sivakumar *et al.*, 2021) with an estimated false-positive rate <3% (Vattathil and Scheet, 2016). For these purposes, since the RNA and DNA were derived from the same tissue (or tumor), we

Table 2. Gene-level performance summaries across 28 cancer sites

Tumor site (abbreviation) (sample size)	Sensitivity (%)	Specificity (%)
Adrenocortical carcinoma (ACC) ($n = 58$)	89	93
Bladder urothelial carcinoma (BLCA) ($n = 261$)	83	90
Breast invasive carcinoma (BRCA) ($n = 641$)	84	92
Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) ($n = 150$)	88	94
Cholangiocarcinoma (CHOL) ($n = 26$)	85	95
Colon adenocarcinoma (COAD) ($n = 256$)	79	97
Esophageal carcinoma (ESCA) ($n = 54$)	91	79
Glioblastoma multiforme (GBM) ($n = 99$)	89	92
Head and neck squamous cell carcinoma (HNSC) ($n = 346$)	86	94
Kidney chromophobe (KICH) ($n = 6$)	94	95
Kidney renal clear cell carcinoma (KIRC) ($n = 56$)	90	94
Kidney renal papillary cell carcinoma (KIRP) ($n = 143$)	92	95
Brain lower grade glioma (LGG) ($n = 388$)	79	96
Liver hepatocellular carcinoma (LIHC) ($n = 115$)	82	95
Lung adenocarcinoma (LUAD) ($n = 312$)	84	90
Lung squamous cell carcinoma (LUSC) ($n = 221$)	88	88
Mesothelioma (MESO) ($n = 76$)	87	95
Ovarian serous cystadenocarcinoma (OV) ($n = 249$)	87	87
Pancreatic adenocarcinoma (PAAD) ($n = 124$)	80	95
Pheochromocytoma and paraganglioma (PCPG) ($n = 133$)	89	96
Prostate adenocarcinoma (PRAD) ($n = 317$)	66	94
Rectum adenocarcinoma (READ) ($n = 133$)	80	96
Skin cutaneous melanoma (SKCM) ($n = 87$)	86	94
Stomach adenocarcinoma (STAD) ($n = 190$)	86	87
Testicular germ cell tumors (TGCT) ($n = 116$)	88	93
Thyroid carcinoma (THCA) ($n = 276$)	85	94
Uterine carcinosarcoma (UCS) ($n = 34$)	85	90
Uveal melanoma (UVM) ($n = 75$)	87	98

Note: For each cancer site, the study abbreviation, number of samples analyzed in the cohort and median gene level sensitivity and specificity are shown.

treat the SNP microarray results as a gold standard. For a subset of samples, we also applied hapLOHseq to whole exome sequencing of DNA (WES) to help interpret the results and assess where deficiencies may be attributed to technology, bioinformatic approaches or inherent limitations of inference from specific nucleic acids.

3.1 SCNA detection from RNA-seq

We compared and contrasted the RNA-seq derived SCNAs with the gold standard SCNAs with units of analysis as gene, chromosome arm and entire genome (i.e. burden) (Supplementary Figs S1 and S2), focusing here on gene-level summaries. Table 1 [first row: ‘hapLOHseq (RNA-seq)’] shows the sensitivity and specificity for SCNA detection solely using tumor RNA for seven cancer sites. Sensitivity is the method’s power to detect true SCNAs (defined as those identified by the gold standard) and specificity measures the method’s ability to correctly identify genomic regions where there is no SCNA detected by the gold standard. We obtained a generally high concordance between the events called from RNA-seq with those called from SNP arrays at a gene level, and this held for the other units of analysis as well. Our SCNA detection rate (sensitivity) from RNA-seq was highest for the GBM data at 79%, whereas the sensitivity for PRAD was markedly lower at 45%. We observed high specificities for all seven cancer sites, ranging between 89% and 99%.

To assess which factors drive the lower than average performance in the PRAD cohort, we further investigated the samples with the highest sensitivities. We observed that the samples with the highest quartile of sensitivity had the highest median number of heterozygous sites and the highest haplotype accuracies (Supplementary Fig. S3). After grouping the samples into quartiles based on their sensitivity, we statistically compared the groups with the Kruskal–Wallis test and noted that the groups were significantly different

when compared both by the number of heterozygous sites (P -value $< 1e-8$) and phase accuracy (P -value = 0.008).

To help understand potential limitations of our approach, we compared our results to those obtained from the application of hapLOHseq to TCGA DNA WES [Table 1; third row: ‘hapLOHseq (WES)’]. Not surprisingly, since it assays the DNA directly, WES consistently achieved higher sensitivities at similar specificities across sites, including PRAD with 76% sensitivity in comparison to 45% from RNA-seq. Although we observed a lower performance for RNA-seq than for WES, our results demonstrate that there exists sufficient information in the RNA-seq for generally accurate inference of SCNAs. Encouraged by this, we explored further the power of RNA-derived SCNA profiling approaches.

3.2 SCNA detection from RNA-seq and imputation-based haplotype inference

We hypothesized that the standard genotyping (variant calling) pipelines for NGS, along with modest reference sizes for haplotype phasing, were holding back the potential of our approach. To address this, we imputed germline genotypes and haplotypes from large-scale reference data using the optimized workflows in the Michigan Imputation Server (MIS). We leveraged the genotype data from a matched-blood sample, available for most participants in TCGA. While this sample does not provide any direct information about SCNAs of the tumor, it does provide more accurate identification of heterozygous sites and estimated haplotypes, central to our approach, but without the need to extract DNA from the tumor. We then applied hapLOHseq for detection of tumor SCNAs using RNA-seq signal combined with the more accurate haplotype information, as detailed in the methods.

The inclusion of imputed genotype calls and high-quality haplotypes provided a substantial improvement in overall SCNA detection [Table 1; second row: ‘hapLOHseq’ (RNA-seq + imputed

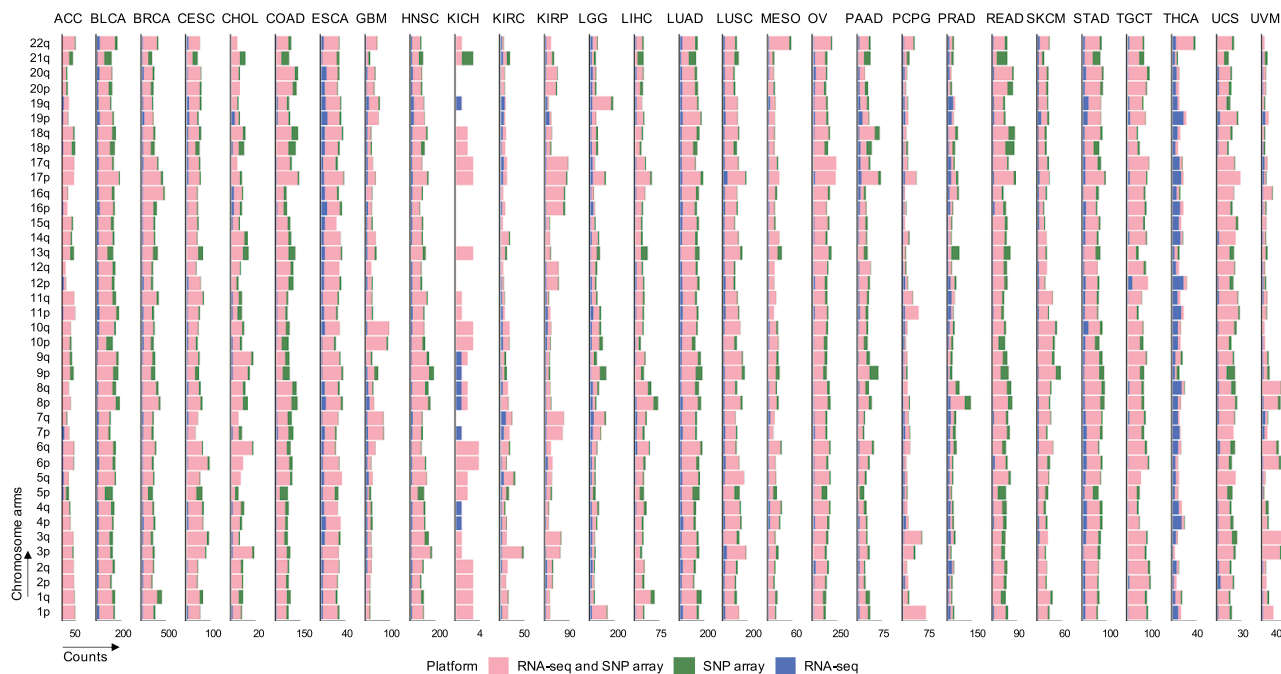


Fig. 1. Chromosome arm-level concordance assessment summaries across 28 cancer sites. We identified chromosome arms that were spanned by SCNAs ($\geq 50\%$) and for each arm we evaluated the concordance between RNA-seq and gold standard. The distribution of the non-acrocentric autosomal chromosome arms ($n = 39$) across the cancer sites are shown. For each site, a stacked bar plot of the number of samples with concordance-specific chromosome arm-level SCNAs are shown for all 39 chromosome arms

genotypes)]. After the imputation approach, sensitivities improved across all cancer sites, ranging from 10% to 32% in absolute increase, i.e. BRCA (13%), COAD (32%), GBM (10%), LUAD (10%), LUSC (15%), PAAD (11%) and PRAD (21%), while high specificities remained similar. In comparison to WES, the imputation approach has lower sensitivity with similar specificity. For instance, for the BRCA cohort, the imputation approach's sensitivity is 9% lower while specificity is 3% higher. These results indicated the potential for inference of SCNAs purely from RNA-seq, so long as there exist sufficiently informative germline genotypes. To comprehensively characterize the potential for SCNA inference from RNA-seq across a range of tissue types, we applied this approach to the remaining cancer sites from the TCGA for which data existed. Study abbreviations for all 28 TCGA cohorts investigated in this study are shown in [Table 2](#).

Across all cancer sites, the median SCNA size was 31.28 Mb, approximately equal to the median SCNA size (31.31 Mb) detected by the gold standard. The median number of SCNA events detected per sample was 20 (gold standard: 19). The highest frequency of SCNA calls per sample were observed in esophageal carcinoma (ESCA), ovarian serous cystadenocarcinoma (OV), LUSC, uterine carcinosarcoma (UCS) and bladder urothelial carcinoma (BLCA) with each with a median of 30 SCNAs or more per sample, consistent with the gold standard ([Supplementary Table S1](#)). In contrast, thyroid carcinoma (THCA) and uveal melanoma (UVM) had the fewest number of SCNAs per sample with each site having < 10 SCNAs per sample. These two sites were also ranked as having the lowest median number of SCNAs by the gold standard.

[Table 2](#) contains per cancer site gene-level summaries of sensitivity and specificity for 28 cancer sites, comprising 4942 samples in the TCGA, after applying the imputation workflow. At the gene level, our imputation-based approach achieved an 85% median sensitivity and 94% median specificity (all genes, all samples). Across the sites, median sensitivities ranged from 66% to 94%, with median specificities between 79% and 98%. With the exception of ESCA, specificity was always greater than or equal to the sensitivity for a given site. At 79%, 79% and 66%, COAD, brain lower grade glioma (LGG) and PRAD (respectively) were the only cancer sites

with sensitivity below 80%. Interestingly, kidney cancers (kidney chromophobe: KICH; kidney renal clear cell carcinoma: KIRC; and kidney renal papillary cell carcinoma: KIRP) were the three cohorts that we observed the best performances for with 94%, 90% and 92% sensitivity and 95%, 94% and 95% specificity.

We evaluated the method at the chromosome arm level as well ([Fig. 1](#)). For each chromosome arm, we assessed the concordance with the gold standard calls and the results indicate that the majority of the true arm-level SCNAs were inferred correctly across all cancer sites. However, several, such as 5p, 9p, 13q and 21q were missed with RNA consistently across the cohorts. We also note that the majority of the chromosome arm-level SCNAs inferred from RNA-seq in the THCA cohort were not present in the gold standard set.

To evaluate the SCNA patterns at a whole genome level, we calculated 'genomic burden' – the proportion of a sample's genome exhibiting SCNAs. Marginally, we observed a median 0.28 genomic burden across all samples across all cancer sites, compared with 0.31 from DNA microarrays. Further, we investigated the patterns of genomic burden per cancer site. The highest genomic burden was observed in ESCA (0.63), followed by OV (0.53) and TGCT (0.53). THCA was the lowest (0.05), followed by PRAD (0.09). Median genomic burden for all cancer sites along with the corresponding array-derived genomic burden are shown in [Figure 2](#). Next, we assessed the correlation of the RNA-derived genomic burden with the gold standard-derived genomic burden at a sample level ([Supplementary Fig. S5](#)), with most of the sites exhibiting correlations larger than 0.9.

SCNA calls discovered by RNA-seq that were not detected in the DNA (gold standard) are putative false positives in our analysis. However, some appear to be false negatives in the gold standard set. To attempt to discriminate between these, we examined in more detail the COAD data, where these putative calls made up 6% of all SCNAs discovered in RNA-seq. We explicitly tested the genomic regions of the putative false positive SCNAs in the corresponding SNP array data with a specific binomial test leveraging phase concordance. Among the 6%, we found that approximately one-fourth (23%) of the calls were validated in the SNP array data (at a P -value < 0.05),

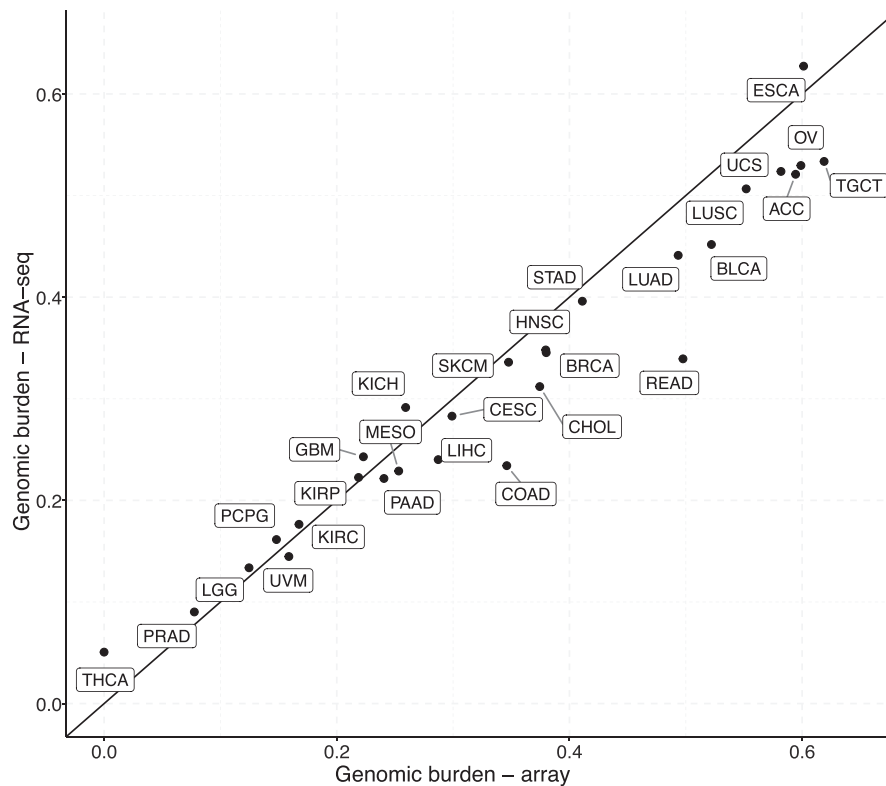


Fig. 2. Concordance assessment at genome level ‘genomic burden’ across 28 cancer sites. Genomic burden is defined as the fraction of the genome that is affected by SCNAs. A scatter plot demonstrating the concordance between RNA-seq- and gold standard-derived genomic burden (median) for each cancer site is shown

Table 3. hapLOHseq and CaSpER comparison

Method	BRCA (<i>n</i> = 77)		GBM (<i>n</i> = 98)	
	Sens (%)	Spec (%)	Sens (%)	Spec (%)
hapLOHseq (RNA-seq)	74	94	81	93
hapLOHseq (RNA-seq + imputed genotypes)	85	91	89	92
CaSpER	43	82	59	95

Note: hapLOHseq and CaSpER performance evaluation. Rows 1–3 show performance results at the gene level obtained by comparing each method to the gold standard (Sivakumar et al., 2021).

indicative that the reported false positive rates are modestly overestimated.

We investigated the method’s performance for different SCNA categories (Supplementary Table S2). We observed that ‘undetermined’ events—subtle events in bulk samples with low mutant cell fraction that cannot be classified as gain, loss or cn-LOH—in the gold standard were the SCNA category that was most frequently missed by RNA-based detection. The difference was most striking in the PAAD and PRAD sites. Exclusion of these undetermined (low mutant cell fraction) events results naturally in higher overall sensitivity rates. Perhaps not surprisingly, the cn-LOH category consistently had the best performance, presumably due to greater BAF perturbation for this SCNA type.

3.3 Comparison to other methods for bulk RNA-seq

While most methods for SCNA inference from RNA-seq have been designed for single cell data, we were able to conduct a detailed comparison with one state-of-the-art method for bulk RNA. Table 3 contains a summary of results from our method,

hapLOHseq and CaSpER, for BRCA and GBM, sites analyzed in the original paper for CaSpER. We compared both methods to the gold standard. Compared with the gold standard calls at the gene level, for both cancers, hapLOHseq offered a superior performance, with a substantial increase in sensitivities with absolute gains of 30% and 42%. Against CaSpER’s own benchmark, the methods appeared more similar, with an edge to hapLOHseq in detection but at some cost in specificity (Supplementary Table S3).

We were also able to successfully run SuperFreq on a subset of the BRCA samples. From analysis of 12 samples where we had results from both SuperFreq and hapLOHseq, the sensitivity for hapLOHseq was 85% versus 77% for SuperFreq (specificities were 95% and 98%, respectively). The authors of SuperFreq demonstrated high sensitivities analyzing *TP53* alterations in high mutant cell fraction settings for COAD. We were able to detect SCNAs in *TP53* in this set with hapLOHseq at an equivalent rate but without the need for additional RNA samples for normalization.

3.4 Translational/prognostic use

Finally, to demonstrate a translational potential, we examined the portability of conclusions from others’ analyses of DNA to ours from RNA-seq. In an example in breast cancer, we recapitulated the distinct genomic burden distributions across different subtypes previously observed in TCGA BRCA data (Cancer Genome Atlas Network, 2012), demonstrating that those of a basal subtype are characterized by high genomic burden in comparison to the others (Fig. 3A). We observed that the samples in the basal subtype that have more than 40% of their genome altered comprise 84% of all basal samples, consistent with the previous report. Furthermore, analyzing the RNA-seq, we were also powered to observe that chromosome arm 5q is more frequently altered in the basal subtype, whereas chromosome arms 1q and 16q are more frequently altered in the luminal subtypes, consistent with the previous findings from analyzing DNA directly (Cancer Genome Atlas Network, 2012) (Fig. 3B). Finally, we specifically investigated the concordance

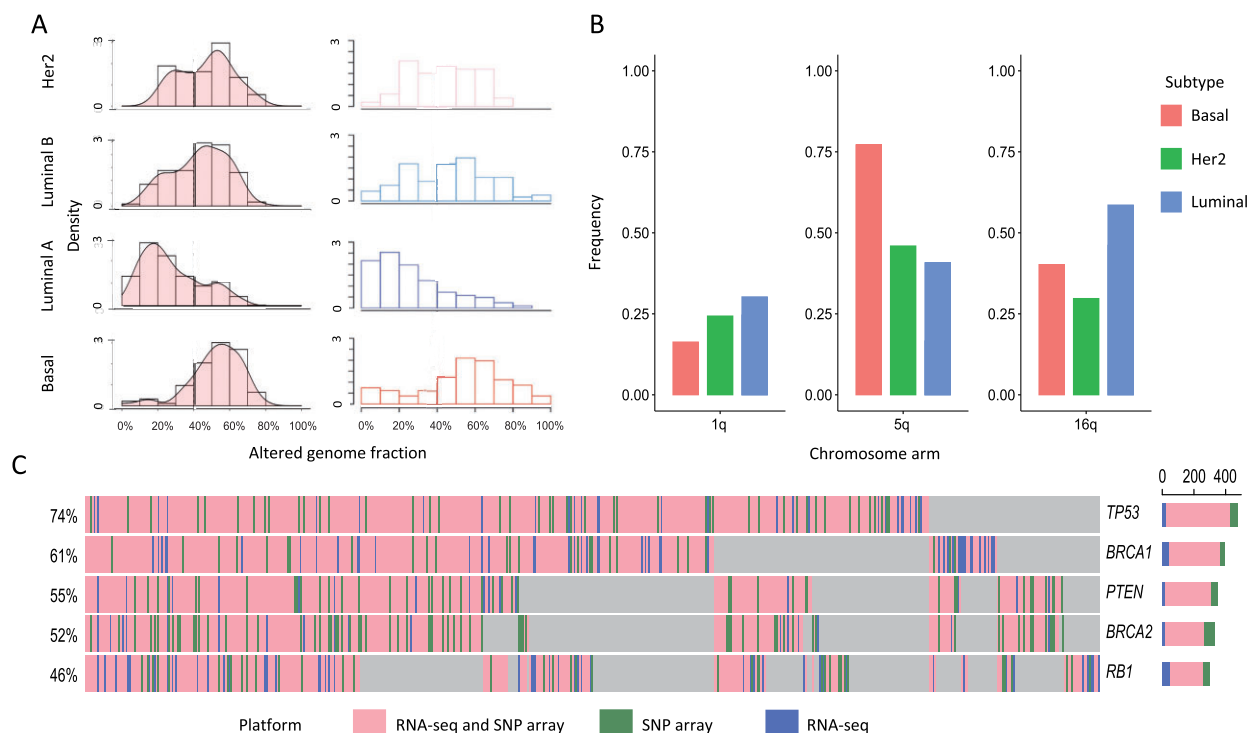


Fig. 3. Clinical efficacy of hapLOHseq results demonstrated using TCGA BRCA cohort. (A) Recapitulating the genomic burden distribution across different subtypes: left: from the supplementary material of the TCGA BRCA paper (Cancer Genome Atlas Network, 2012), right: hapLOHseq results; histogram of sample genomic burden across the cohort grouped by subtypes. (B) Frequency of chromosome arm level alterations in 1q, 5q and 16q as a fraction of number of samples across different subtypes. (C) Concordance assessment for the five genes that are frequently affected by SCNA events. Rows represent the genes and columns represent the samples in the cohort

between hapLOHseq and gold standard results for five genes that are frequently affected by CNA events for the BRCA cohort, i.e. *BRCA1/2*, *PTEN*, *RB1* and *TP53*. We showed that hapLOHseq obtained promising results that have a potential clinical use with 93% sensitivity and 85% specificity for *BRCA1*, 78% sensitivity and 95% specificity for *BRCA2*, 85% sensitivity and 88% specificity for *PTEN*, 87% sensitivity and 94% specificity for *RB1* and 90% sensitivity and 86% specificity for *TP53*. Sample-level concordance assessment for each of the genes is shown in Figure 3C.

4 Discussion

In this study, we detect and characterize the genomic landscapes of SCNAs from tumor bulk RNA-seq using a haplotype-aware statistical method. We proposed two approaches that differ in the way germline genotypes are obtained for subsequent analysis of ‘B allele’ frequencies (BAFs). While the first approach solely uses RNA-seq from tumor to estimate the haplotypes, the second approach leverages available or potentially collectible SNP array (or equivalent) data from a matched-normal sample to achieve higher accuracies in genotyping and haplotype reconstruction through a popular imputation pipeline.

In an analysis of 28 cancer sites from TCGA, our method achieved high sensitivity for SCNA detection with the imputation approach (85% versus 68%), retaining high specificity as well (.95%). Summaries of SCNA genomic burden were sufficiently high as to potentially obviate the need for analyzing DNA. In sites with lower sensitivities, such as PRAD, analyses of DNA exome sequencing reflected difficulties as well, indicating challenges for such sites more specific to targeted sequencing data.

Our imputation approach addresses difficulties associated with genotype calling from RNA-seq, e.g. due to non-uniform coverage, which in turn provides highly accurate and phased genotypes. Indeed, we explored these factors as direct contributors to improved

performance in PRAD (Supplementary Fig. S3). Our germline heterozygote identification could improve other methods for bulk RNA analysis, as well, such as CaSpER and SuperFreq. In different statistical implementations, each of these combine information from not only total read counts but also BAF dispersion at heterozygotes. Whereas SuperFreq relies on external data for normalization (e.g. paired normal RNA samples), CaSpER’s approach works on a sample-by-sample basis, as does hapLOHseq. We conducted a detailed comparison to CaSpER, observing higher sensitivities with our haplotype-based approach. Ultimately, getting the absolute performance characteristics will depend on improved gold standard datasets. We note that these methods use information orthogonal to that leveraged by hapLOHseq and thus may offer improvements when applied in combination, or integrated for joint analyses, an area of future study.

Blood, buccal or adjacent normal samples can serve as representative of the germline. The first two are non-surgical and more easily collected, whereas the third may be available for some specimens. Array-based genotyping of these samples presents an economical approach for improved tumor SCNA characterization. This is feasible for existing clinical cohorts with banked patient blood samples or biobanks with existing genotype data. In our exhibition, we focused on individuals of European ancestry to assess the performance of our method in a ‘best case’ scenario given current resources. However, efforts such as TOPMed (Taliun et al., 2021) will generate high-density genotype panels for individuals of non-European ancestry. Our approach further highlights the need for genetic panels of high diversity in biomedical research.

Our approach, as we have demonstrated here, does not attempt to detect balanced duplications, i.e. those where maternal and paternal chromosome segments are present in equal ratios. This may be overcome through integration of coverage data with our existing approach, which would be feasible for large balanced duplications as shown previously (Serin Harmanci et al., 2020; Stopsack et al., 2019). While any method for RNA-seq will have natural limitations

in detecting alterations that do not span expressed genes, over larger regions limiting factors will be mitigated or averaged toward genome levels. Further, molecular alterations caused by focal SCNAs of key cancer drivers may be detectable through traditional RNA analyses of altered gene expression, including at the pathway level, and/or identification of specific transcripts.

In summary, we show that the proposed haplotype-based approach for RNA-derived SCNA calls is robust for detection of megabase-scale somatic mutations. Overall, detection rates were generally higher than 85% at specificities high enough for *de novo* discoveries and assessments of genomic associations with clinical phenotypes and malignancies. Indeed, we successfully recapitulated SCNA features associated with clinical subtypes of breast cancer. Our findings show that our method can be used to increase the utility of bulk RNA-seq by allowing for a more comprehensive molecular profiling of tumors in settings where DNA analysis is impractical due to limited tissue sample availability or financial constraints and enables secondary analyses of existing data from high-value clinical cohorts.

Acknowledgement

The authors acknowledge the High Performance Research Computing Center at the University of Texas, MD Anderson Cancer Center.

Funding

This work was supported by National Institutes of Health [R01HG005855 and P30CA016672]; and the following award from the Cancer Prevention Research Institute of Texas: RP160668.

Conflict of Interest: none declared.

References

- Alexandrov, L.B. *et al.*; ICGC PedBrain. (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
- Alkan, C. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.*, **41**, 1061–1067.
- Amarasinghe, K.C. *et al.* (2014) Inferring copy number and genotype in tumour exome data. *BMC Genomics*, **15**, 732.
- Baughner, J.D. *et al.* (2013) Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method. *BMC Genomics*, **14**, 367.
- Bouska, A. *et al.* (2014) Genome-wide copy-number analyses reveal genomic abnormalities involved in transformation of follicular lymphoma. *Blood*, **123**, 1681–1690.
- Callagy, G. *et al.* (2005) Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays. *J. Pathol.*, **205**, 388–396.
- Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
- Coudray, A. *et al.* (2018) Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ*, **6**, e5362.
- Das, S. *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
- Fan, J. *et al.* (2018) Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.*, **28**, 1217–1227.
- Flensburg, C. *et al.* (2021) Detecting copy number alterations in RNA-Seq using SuperFreq. *Bioinformatics*, **37**, 4023–4032.
- Fowler, J. *et al.* (2019) System for quality-assured data analysis: flexible, reproducible scientific workflows. *Genet. Epidemiol.*, **43**, 227–237.
- Gao, R. *et al.* (2021) Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.*, **39**, 599–608.
- Griffith, M. *et al.* (2015) Optimizing cancer genome sequencing and analysis. *Cell Syst.*, **1**, 210–223.
- Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
- Hieronymus, H. *et al.* (2018) Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. *Elife*, **7**, e37294.
- Jakubek, Y.A. *et al.* (2020) Large-scale analysis of acquired chromosomal alterations in non-tumor samples from patients with cancer. *Nat. Biotechnol.*, **38**, 90–96.
- Knudson, A.G. Jr (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA*, **68**, 820–823.
- Koboldt, D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
- Korn, J.M. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.
- Kridel, R. *et al.* (2012) Whole transcriptome sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma. *Blood*, **119**, 1963–1971.
- Li, Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.
- Liang, L. *et al.* (2016) Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene*, **35**, 1475–1482.
- Loh, P.-R. *et al.* (2018) Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*, **559**, 350–355.
- McKenna, A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
- Negrini, S. *et al.* (2010) Genomic instability—an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.*, **11**, 220–228.
- Nibourel, O. *et al.* (2017) Copy-number analysis identified new prognostic marker in acute myeloid leukemia. *Leukemia*, **31**, 555–564.
- Peng, L. *et al.* (2015) Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Sci. Rep.*, **5**, 13413.
- Price, A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
- Ried, T. *et al.* (2012) The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochim. Biophys. Acta*, **1819**, 784–793.
- San Lucas, F.A. *et al.* (2016) Rapid and powerful detection of subtle allelic imbalance from exome sequencing data with hapLOHseq. *Bioinformatics*, **32**, 3015–3017.
- Serin Harmanci, A. *et al.* (2020) CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat. Commun.*, **11**, 89.
- Shah, S.P. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.
- Shukla, A. *et al.* (2020) Chromosome arm aneuploidies shape tumour evolution and drug response. *Nat. Commun.*, **11**, 449.
- Sivakumar, S. *et al.* (2021) Pan cancer patterns of allelic imbalance from chromosomal alterations in 33 tumor types. *Genetics*, **217**, 1–12.
- Stopsack, K.H. *et al.* (2019) Aneuploidy drives lethal progression in prostate cancer. *Proc. Natl. Acad. Sci. USA*, **116**, 11390–11395.
- Taliun, D. *et al.*; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, **590**, 290–299.
- Taylor, A.M., Cancer Genome Atlas Research Network. *et al.* (2018) Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell*, **33**, 676–689.e3.
- Tickle, T. *et al.* (2019) *inferCNV of the Trinity CTAT Project*. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Boston.
- Vattathil, S. and Scheet, P. (2013) Haplotype-based profiling of subtle allelic imbalance with SNP arrays. *Genome Res.*, **23**, 152–158.
- Vattathil, S. and Scheet, P. (2016) Extensive hidden genomic mosaicism revealed in normal tissue. *Am. J. Hum. Genet.*, **98**, 571–578.
- Wang, H. *et al.* (2016) Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. *Oncogene*, **35**, 2011–2019.
- Watkins, T.B.K. *et al.* (2020) Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature*, **587**, 126–132.
- Weiss, M.M. *et al.* (2004) Genomic alterations in primary gastric adenocarcinomas correlate with clinicopathological characteristics and survival. *Cell. Oncol.*, **26**, 307–317.
- Yizhak, K. *et al.* (2019) RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, **364**, eaaw0726.