Published in final edited form as:

Nat Methods. 2019 May ; 16(5): 401–404. doi:10.1038/s41592-019-0388-9.

Integrated Transcriptomic-Genomic profiling using Texomer reveals novel biology from cancer tissues

Fang Wang¹, Shaojun Zhang², Tae-Beom Kim¹, Yu-yu Lin¹, Ramiz Iqbal¹, Zixing Wang¹, Kanishka Sircar³, Jose A. Karam⁴, Michael C. Wendl⁵, Funda Meric-Bernstam⁶, John N. Weinstein^{1,7}, Li Ding⁵, Gordon B. Mills⁷, Ken Chen^{1,*}

¹Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX

²Department of Genome Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX

³Department of Pathology and Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX

⁴Department of Urology, The University of Texas MD Anderson Cancer Center, Houston, TX

⁵Department of Medicine, Department of Genetics, Washington University School of Medicine, St. Louis, MO

⁶Department of Investigational Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, TX

⁷Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX

Abstract

DNA/RNA integration bears the promise to further improve the power of genomic testing, yet novel analytical approaches are required to translate the increased data dimensionality, heterogeneity and complexity to patient benefits. We developed a statistical approach called Texomer (<https://github.com/KChen-lab/Texomer>) that performs allele-specific, tumor-deconvoluted transcriptome-exome integration of autologous bulk whole exome and transcriptome

*Corresponding author: Ken Chen (kchen3@mdanderson.org).

Author contributions

F. W. and K. C. designed the experiments. F. W. and S. Z. conceived the statistical model. F. W. developed the software. T. B. K., Z. W., M. C. W. and L. D. contributed to evaluation; Y. L. and R. I. software release; K. S. and J. A. K. pathology review; F. M., J. N. W. and G. B. M. biological interpretation.

K. C. and F. W. wrote the manuscript with input from all authors. K. C. supervised the work.

Data availability and Accession Code Availability Statements

We downloaded the bulk WES and WTS data of BRCA (N = 833) and SKCM samples (N = 465) from The Cancer Genome Atlas (TCGA, dbGAP Accession ID: phs000178.v9.p8). We downloaded the single cell RNA-seq data as well as matched bulk WES and WTS data from 11 breast cancer samples from NCBI under accession ID: GSE75688, SRP067248. We downloaded the WES and WTS data of breast cancer cell line HCC1143 and matched normal cell line HCC1143BL from the cancer cell line encyclopedia (CCLE) project of Genomic Data Commons (GDC) Data Portal.

Texomer is available in GitHub at <https://github.com/KChen-lab/Texomer>.

Competing Financial Interests Statement

None.

sequencing data. Texomer resulted in significantly improved accuracy in sample categorization and functional variant prioritization.

Molecular profiling of tissue (e.g., tumor) samples using bulk DNA sequencing are of limited power and precision¹. It generates long lists of variants of unknown significance (VUS)² and is limited in characterizing intra-tissue heterogeneity³. Multi-omics profiling bears the promise to further improve the power and precision^{4,5}. However, novel, systematic approaches are required to translate the increased data dimensionality, heterogeneity and complexity to patient benefits.

We developed a statistical approach called Texomer (Fig. 1a) that performs allele-specific, tumor-deconvoluted⁶ transcriptome-exome integration of autologous bulk whole exome (WES) and whole transcriptome sequencing (WTS) data, and outputs transformed tumor-specific DNA (TT-DNA) copy number and RNA (TT-RNA) expression profiles, differential allelic cis-regulatory effect (DACRE) scores, as well as tumor purity and intratumor heterogeneity estimations (Fig. 1b, Online Methods). Evaluation using simulated data and multiple real datasets indicated that Texomer achieved desirable technical accuracy and outperformed existing tools (Supplementary Note 1).

We applied Texomer to categorize tumor molecular subtypes using the WES and WTS data from autologous bulk breast invasive carcinoma (BRCA) samples in the cancer genome atlas (TCGA, Supplementary Note 2)⁷. We found that sample categorization based on original bulk WES and bulk WTS read counts at single nucleotide variant (SNV) sites had limited accuracy, resulting in clusters of samples of heterogeneous profiles (Fig. 1c and 1d); whereas categorization based on Texomer-transformed profiles achieved evidently improved accuracy, resulting in more clusters of samples of homogeneous profiles and distinct biological properties (Fig. 1e and 1f, Supplementary Note 2).

We further performed variance component analysis^{10–12} to quantify the relationship between the DNA and the RNA data (Supplementary Note 3). Around 10% of variance in the bulk RNA data (WTS read counts from the SNV sites) can be explained by the bulk DNA data (WES read counts from the SNV sites). After performing Texomer transformation, the amount of variance in the RNA data (TT-RNA expression profile) that can be explained by the DNA data (TT-DNA copy number profile) increased significantly to 23%. In contrast, only 2% of variance in the bulk WTS read counts can be explained by the TT-DNA copy number profile (Fig. 1g). These results indicated that the TT-DNA and TT-RNA profiles have more accurately matched tissue-origins and reflect more accurate genotype-phenotype association in the tumors than do the bulk data. Even more striking differences were observed in the allele-specific data, between the variance component of the allele-specific bulk RNA and DNA read counts, and that of the allele-specific TT-DNA copy number and RNA expression levels (Fig. 1h). Validation experiments using isogenic cell-line and *in silico* simulation data confirmed that Texomer can much more accurately integrate tumor DNA and RNA than other approaches (Supplementary Note 3).

The improved molecular characterization power achieved by Texomer from joint bulk WES/WTS profiling also manifested in significantly improved accuracy for functional

variant prediction, a critical mission for functional genomics and genomic medicine. By identifying SNVs that are selectively expressed, i.e., having allele-specific TT-RNA expression levels unexpectedly higher or lower than allele-specific TT-DNA copy number levels, we were able to identify from TCGA BRCA and skin cutaneous melanoma (SKCM) data putative functional variants and genes (Fig. 2a and b), which appeared enriched of known cancer targets contributing to predisposition and/or tumorigenesis (Supplementary Note 4). By further formulating the differential extent of selective expression between a variant and a wild-type allele as a differential allelic cis-regulatory effect (DACRE) score, we enabled systematic, exome-wide DNA/RNA-joint characterization of variant functions (Methods). The functional variants identified based on DACRE scores appeared associated significantly with extreme DNA methylation levels and known enhancer elements (Supplementary Note 4). We further compared *in silico* prediction scores with experimental scores obtained using an *in vitro* cell-line viability assay¹³. We confirmed that DACRE score has independent, additive values with respect to function impact scores computed by widely used DNA-based functional variant predictors (Supplementary Note 4).

We further assessed the potential utility of Texomer in clinical sequencing settings. We first predicted functional status of each of the somatic missense SNVs in TCGA BRCA and SKCM samples, based on the functional impact scores calculated by a set of 8 widely used functional variant predictors¹⁴. We further refined the predictions by filtering out somatic missense SNVs with negative DACRE scores. We found that in each of the cases, the filtered results had significantly higher (often doubled) precision (Fig. 2c and d), defined as the fraction of the variants known in the OncoKB database¹⁵. In contrast, filtering based on bulk WTS counts resulted in only marginal benefits (Supplementary Note 5).

Thus, our study revealed the analytical challenges involved in integrating autologous bulk WES and WTS data and presented a statistically robust solution Texomer to realize the increased power and precision. With the increasing expectation on precision medicine, many more patient samples will undergo both WES and WTS. Integrative approaches such as Texomer will be critically needed to deliver the promises. The source code of Texomer is available at <https://github.com/KChen-lab/Texomer>.

Online Methods

Overview of the methods

Given a set of autologous bulk whole exome (WES) and whole transcriptome sequencing (WTS) data, Texomer performs allele-specific, tumor-deconvoluted transformation of read counts observed at the germline single nucleotide polymorphisms (SNPs) and somatic single nucleotide variants (SNVs) sites. It outputs transformed tumor DNA (TT-DNA) allele-specific copy number (ASCN) and RNA (TT-RNA) allele-specific expression levels (ASELs), differential allelic cis-regulatory effect (DACRE) scores, as well as quantification of tumor purity and intratumor heterogeneity. The basic method consists of 6 steps (Fig. 1b).

Initial estimation of tumor purity and allele-specific copy numbers

Initial segmentation, tumor purity, and ASCNs are obtained using ASCAT¹⁶, TITAN¹⁷, sequenza¹⁸, and FACETS¹⁹, respectively from allelic read counts covering heterozygous germline SNPs.

The ploidy ratio between the tumor and the normal WES and the variant B allele frequency (BAF) at the i -th germline SNP site observed from the total (N_i) and allelic (y_i) read counts, can be estimated from tumor purity in the DNA sample (α_D), and the integer total (TCN_i) and the allele-specific copy numbers ($ASCN_i$) in the tumor sample, as described in equations (1) and (2), respectively,

$$Ploidy_ratio_i = \frac{N_i^{Tumor}}{N_i^{Normal}} \approx \frac{2(1 - \alpha_D) + \alpha_D \cdot TCN_i}{2}, \quad (1)$$

$$BAF_i = \frac{y_i^{Tumor}}{N_i^{Tumor}} \approx \frac{(1 - \alpha_D) + \alpha_D \cdot ASCN_i}{2(1 - \alpha_D) + \alpha_D \cdot TCN_i}. \quad (2)$$

TCN_i equals the summation of the $ASCN_i$ of the variant and the wildtype alleles. The BAF of the s -th somatic SNV is calculated using a different equation (3) because somatic SNVs exist only in the tumor cells:

$$BAF_s = \frac{y_s^{Tumor}}{N_s^{Tumor}} \approx \frac{\alpha_D \cdot SMCN_s}{2(1 - \alpha_D) + \alpha_D \cdot TCN_s}, \quad (3)$$

where $SMCN_s$ is the somatic variant allelic copy number corresponding to the s -th somatic SNV in the tumor cells.

Given α_D and $TCNs$, $SMCNs$ can be calculated using equation (3). Because copy numbers are by definition non-negative integer, a well-estimated $SMCN$ should be close to a non-negative integer. We define a metric to sum up the difference between the estimated $SMCNs$ and their nearest non-negative integers:

$$\Delta = \sum_{s=1}^S |SMCN_s - \text{int}(SMCN_s)|, \quad (4)$$

where indices over all somatic SNVs and $\text{int}(\cdot)$ rounds a $SMCN$ to the nearest non-negative integer. A separate value is estimated from the ASCAT, sequenza, TITAN and FACETS results, respectively. The result with the minimum value is selected as the best result, based on which further iterative optimizations are performed.

Since the assumption that true copy number equals to $\text{int}(SMCN)$ may not always hold true, particularly when $SMCNs$ become large, we investigated the distribution of $SMCNs$ in the breast invasive carcinomas (BRCA, N = 832) and skin cutaneous melanoma (SKCM, N = 465) samples from The Cancer Genome Atlas (TCGA, dbGAP Accession ID: phs000178.v9.p8). We found that the majority (98%) of genomic regions in BRCA and

SKCM data have a relatively moderate copy number (the nearest non-negative integer of $SMCN < 5$), implying that asymptotically our method will be robust in analyzing real data.

Optimizing tumor purity and ASCN based on somatic SNVs

After the initial estimation results are obtained from one of ASCAT, FACET, TITAN and sequenza, Texomer iteratively improves the results by including the additional somatic SNVs. In doing so, it first updates α_D using read counts at the somatic SNV sites (equation (3)) and then $ASCNs$ and $TCNs$ using read counts at germline SNP sites (equations (1) and (2)).

In order to quantify the extent of convergence, Texomer constructs a null distribution of based on 1,000 sets of randomly generated allelic counts across all the SNV sites. At each SNV site, an allelic read count is randomly sampled from a Binomial distribution $B(N, BAF)$, parameterized by the total read count N and the BAF observed from the real WES data. Corresponding $SMCNs$ are calculated from the simulated read counts. An empirical p -value is estimated for each observed in the real data using equation (5):

$$p_D = \frac{(\sum_{r \in R} index_r^A) + 1}{\|R\| + 1} index_o^A = \begin{cases} 1, A_r \leq A_o \\ 0, Otherwise \end{cases}, \quad (5)$$

where R represents the set of random samples, the $\|\cdot\|$ function measures the set size, the subscripts o and r designate the real and the random samples, respectively, and p_D calculates the statistical significance that the cumulative difference between the $SMCNs$ and the nearest integers is not caused by random fluctuations. In addition, because somatic SNVs are not expected to have 0 $SMCNs$, we estimate an empirical p -value to characterize the significance of observing a fraction of somatic SNVs ($0 \leq Z_0 \leq 1$) with 0 integerized $SMCN$:

$$p_Z = \frac{(\sum_{r \in R} index_r^Z) + 1}{\|R\| + 1} index_o^Z = \begin{cases} 1, Z_r \leq Z_o \\ 0, Otherwise \end{cases}, \quad (6)$$

where Z_r is the fraction of somatic SNVs of 0 integerized $SMCN$ in each of the R random samples.

Finally, we define an empirical score the proportion of somatic SNVs having 0 integerized $SMCN$. We update $\tau = p_D + p_Z$ to constrain the deviation of $SMCNs$ from integers and α_D and $ASCNs$ until τ converges.

Quantification of intra-tumor heterogeneity

In a clonal tumor, every cell should contain the same set of somatic SNVs and germline SNPs. $SMCNs$ derived from somatic SNVs should be restricted to discrete ranges of $ASCNs$ defined by germline SNPs. Peaks in the distribution of $SMCNs$ should overlap the peaks in the distribution of $ASCNs$. Deviation of the distribution of $SMCNs$ from that of $ASCNs$ reflects the presence of subclones. Thus, intrasample heterogeneity (ITH) level can be quantified by equation (7) (Supplementary Fig. 1a):

$$ITH = \int_A [p(x) - q(x)] dx \quad (7)$$

$$A = \{x | p(x) \geq q(x)\},$$

where $p(x)$ is the probability density of SMCNs and $q(x)$ the probability density of ASCNs. Interval A denotes the region of integration containing all the x values where $p(x) \geq q(x)$. A sample with low ITH score has similar SMCN and ASCN distributions and a relatively small area of difference between SMCN and ASCN distributions (Supplementary Fig. 1b and c), whereas a sample with a high ITH score has distinctive distributions and a relatively large area of difference (Supplementary Fig. 1d and e)

Jointly estimating tumor purity in the WTS data

The allelic WTS read counts y at a given variant v can be explained by two components:

$$p(y) = \pi \cdot p(y|I) + (1 - \pi) \cdot p(y|II), \quad (8)$$

where $p(y|I)$ represents the probability that allelic WTS read count of the variant is concordant with its ASCN, $p(y|II)$ is the probability of discordant with its ASCN which was parameterized by the allele-specific expression level (ASEL) and π is the weight between the two components. We can model $p(y|j)$ ($j = I, II$) as Beta-Binomial distributions:

$$p(y|j) = \frac{\binom{N}{y} \cdot B(N - y + \theta \cdot (1 - f_j), y + \theta \cdot f_j)}{B(\theta \cdot (1 - f_j), \theta \cdot f_j)},$$

where $B(\cdot, \cdot)$ is the Beta function, N the total number of reads, and θ an over-dispersion parameter. f_j represents the major allele frequency that have different expected values in the two components: I) $ASEL = ASCN$ (and $TEL = TCN$); II) $ASEL \neq ASCN$ (and $TEL \neq TCN$):

$$f_j = \begin{cases} \frac{1 - \alpha_R + \alpha_R \cdot ASCN}{2(1 - \alpha_R) + \alpha_R \cdot TCN}, j = I \\ \frac{1 - \alpha_R + \alpha_R \cdot ASEL}{2(1 - \alpha_R) + \alpha_R \cdot TEL}, j = II \end{cases} \quad (10)$$

where α_R is the tumor purity in the RNA data, $ASCN$ and TCN are allele-specific and total copy numbers estimated previously from the WES data, and $ASEL$ and TEL are allele-specific and total RNA expression levels. TEL equals to the summation of the $ASEL$ s of the two alleles.

For the m -th (copy number) segment that contains multiple germline SNPs, the likelihood of observing the variant allelic counts $Y = \{y_v | v \in V^m\}$ can be defined as:

$$L(Y | \alpha_R^m, ASEL^m, TEL^m, \theta^m, \pi^m) = \prod_{v \in V^m} [\pi^m \cdot p(y_v | I) + (1 - \pi^m) \cdot p(y_v | II)], \quad (11)$$

where V^m represents the set of variants located in segment m and $\Theta^m = (\alpha_R^m, ASEL^m, TEL^m, \theta^m, \pi^m)$ is unknown parameter corresponding to the m -th segment. $ASEL^m$ and TEL^m reflect allele-specific and total RNA expression levels of the segment.

We can evaluate the parameters through maximizing the likelihood (ML):

$$\begin{aligned}\hat{\Theta}^m &= \arg \max_{\Theta^m} L(Y) = \arg \max_{\Theta^m} \prod_{v \in V_m} [\pi \cdot p(y_v | I) + (1 - \pi) \cdot p(y_v | II)] \\ &= \arg \max_{\Theta^m} \sum_{v \in V_m} \log[\pi \cdot p(y_v | I) + (1 - \pi) \cdot p(y_v | II)],\end{aligned}\quad (12)$$

using the *bbmle* package in R (which supports a variety of customized likelihood functions) with the initial value $\pi = 0.5$, $\alpha_R = \alpha_D$, $ASEL = ASCN$, $TEL = TCN$ and $\theta = 9$. If iterations based on the initial values do not converge, we perform a grid search of α_R and π from 0.1 to 0.9 with a step size of 0.1 and re-perform the iterations until an ML solution is found. If the above ML estimation converges on different values of π , we select the one corresponding to the smallest difference between the estimated DNA and RNA purity values. This is grounded by the assumption that the DNA and the RNA data derived from autologous tissues should have similar purity values. Based on the estimated α_R from all of the segments weighted by the length of segments, the α_R value at the peak of its density distribution is output as the overall tumor purity at the RNA level.

Deconvoluting tumor allele-specific RNA expression level from bulk WTS data

The TCN and TEL of the i -th variant are related through

$$\frac{N_i^{DNA}}{N_i^{RNA}} \approx k \cdot ratio_i^E, \quad (13)$$

and

$$ratio_i^E = \frac{2(1 - \alpha_D) + \alpha_D \cdot TCN_i}{2(1 - \alpha_R) + \alpha_R \cdot TEL_i} \quad (14)$$

where N^{DNA} and N^{RNA} are the total read count spanning the variant site in the DNA-seq and RNA-seq data, respectively; k is a constant reflecting the sequencing depth difference between the entire DNA-seq and RNA-seq datasets, estimated from the read counts at all germline SNP sites; $ratio^E$ reflects a normalized level of difference in the tumor between the TEL and the TCN of each variant, calculated from the tumor purity values in the DNA and the RNA data, respectively. $ratio^E$ reflects relative transcriptional efficiency amongst different variants relative to their copy numbers. In this definition, the majority of variants are expected to have $ratio^E$ close to 1.

With calculated TEL of each variant through equation (14), we can further estimate $ASEL$ using

$$BAF = \begin{cases} \frac{(1 - \alpha_R) + \alpha_R \cdot ASEL}{2 \cdot (1 - \alpha_R) + \alpha_R \cdot TEL}, & \text{for germline SNPs} \\ \frac{\alpha_R \cdot ASEL}{2 \cdot (1 - \alpha_R) + \alpha_R \cdot TEL}, & \text{for somatic SNVs} \end{cases}, \quad (15)$$

where BAF is the B-allele frequency observed in the RNA-seq data over each variant.

Identification of selectively expressed variants

After the ASCNs and the ASELs are obtained, we can compute the posterior probability if the variant allele is selectively expressed, i.e. having RNA expression levels unexpected from DNA copy number level (II):

$$P(II|v) = \frac{(1 - \hat{\pi}^m) \cdot p(v|II)}{\hat{\pi}^m \cdot p(v|I) + (1 - \hat{\pi}^m) \cdot p(v|II)}, \quad (16)$$

$\hat{\pi}^m$ is from the ML-estimation corresponding to segment m that contains the variant. A variant is called selectively expressed if $P(II|v) > 0.5$.

DACRE score for measuring the functionality of a variant

The expression level of a gene can be decomposed into three components regulated respectively by trans-effects (T), copy number (C), cis-effects (M) and other remaining factors (ε)²⁰:

$$TEL = w_T \cdot T + w_C \cdot C + M + \varepsilon. \quad (17)$$

A total expression level (TEL) can be further decomposed into a variant ($ASEL_{variant}$) and a wildtype ($ASEL_{wildtype}$) allele-specific expression levels:

$$ASEL_{variant} = w_T \cdot T + w_{ASCN_{variant}} \cdot ASCN_{variant} + M + \varepsilon, \quad (18)$$

and

$$ASEL_{wildtype} = w_T \cdot T + w_{ASCN_{wildtype}} \cdot ASCN_{wildtype} + \varepsilon, \quad (19)$$

where w denotes a weight for each factor, assuming that where w the variant and the wildtype alleles are subject to similar regulatory effects except for the variant-allele-specific cis-regulation.

Based on these definitions, the differential allelic cis-regulatory effect (DACRE) associated with a variant allele (M) can be estimated from

$$DACRE = M = (ASEL_{variant} - ASEL_{wildtype}) - (w_{ASCN_{variant}} \cdot ASCN_{variant} - w_{ASCN_{wildtype}} \cdot ASCN_{wildtype}), \quad (20)$$

and w_{ASCN} measures the probability of a specific ASCN, that decreased away from $ASCN=1$:

$$w_{ASCN} = \lambda \cdot e^{-\lambda \cdot |ASCN-1|}, \quad (21)$$

where λ is constant (default 1). The weighting over ASCNs is motivated by the observed non-linear relationship between the paired ASEL and ASCN in real TCGA data (Supplementary Fig.2), particularly for somatic mutant alleles with high ASCNs. This may reflect increased transcriptional regulatory complexity of alleles that are duplicated multiple times and are placed into different regulatory context. We found that an exponential function can well approximate this observed non-linear effect, similar to what has been used in a previous study²¹.

DACRE nullifies trans-regulatory effects (which equally affect two alleles) and alleviates confounding effects introduced by tissue type, cell-type/states and/or environmental factors^{22,23} (Supplementary Fig.3). It is different from differential allelic expressions (DAE) and allelic expression imbalance (AEI) in that DACRE focuses specifically on *cis*-acting, local, potentially non-linear effects that perturb transcriptional regulation. It does so by removing dosage and, presumably, additive effects of copy number alterations, whereas DAE does not.

Life Science Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary Linked to this article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by the National Institute of Health [R01CA172652 to K. C., U01CA217842 to G. B. M., U24CA211006 to L. D., U24CA210950 to R. A.], the Cancer Prevention and Research Institute of Texas [RP180248 to K. C.], the MD Anderson Cancer Center Sheikh Khalifa Ben Zayed Al Nahyan Institute of Personalized Cancer Therapy and the National Cancer Institute Cancer Center Support Grant [P30 CA016672 to P. P.]. We also thank Y. Chen, T. Hart, B. Lim, G. Lozano, S. Xiong, L. Wang, X. Song for insightful discussions and X. Zheng for data curation.

References

1. Yohe S & Thyagarajan B Review of Clinical Next-Generation Sequencing. Arch Pathol Lab Med 141, 1544–1557 (2017). [PubMed: 28782984]
2. Richards S et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405–24 (2015). [PubMed: 25741868]
3. McGranahan N & Swanton C Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell 168, 613–628 (2017). [PubMed: 28187284]
4. Huang S, Chaudhary K & Garmire LX More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet 8, 84 (2017). [PubMed: 28670325]

5. Hasin Y, Seldin M & Lusis A Multi-omics approaches to disease. *Genome Biol* 18, 83 (2017). [PubMed: 28476144]
6. Yadav VK & De S An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in Bioinformatics* 16, 232–241 (2015). [PubMed: 24562872]
7. Hutter C & Zenklusen JC The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* 173, 283–285 (2018). [PubMed: 29625045]
8. van der Maaten L & Hinton G Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605 (2008).
9. Ester M, K. H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96) AAAI Press, 1996, pp. 226–231 (1996).
10. Hopper John L., V. PM Variance Component Analysis in Encyclopedia of biostatistics (ed. Armitage P, T. C) (Wiley Interscience, Hoboken, NJ, 2005).
11. Chambers JM Linear models in Chapter 4 of Statistical Models in S (ed. Hastie, C. JMa.J. T) (Wadsworth & Brooks/Cole, 1992).
12. Searle SR, Casella G, McCulloch CE Variance Components, (Wiley New York, 1992).
13. Dogruluk T et al. Identification of Variant-Specific Functions of PIK3CA by Rapid Phenotyping of Rare Mutations. *Cancer Res* 75, 5341–54 (2015). [PubMed: 26627007]
14. Tang H & Thomas PD Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics* 203, 635–47 (2016). [PubMed: 27270698]
15. Chakravarty D et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* 2017(2017).
16. Van Loo P et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* 107, 16910–5 (2010). [PubMed: 20837533]
17. Ha G et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* 24, 1881–93 (2014). [PubMed: 25060187]
18. Favero F et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 26, 64–70 (2015). [PubMed: 25319062]
19. Shen R & Seshan VE FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* 44, e131 (2016). [PubMed: 27270079]
20. Gamazon ER et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 47, 1091–8 (2015). [PubMed: 26258848]
21. Carter SL et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 30, 413–21 (2012). [PubMed: 22544022]
22. Carithers LJ & Moore HM The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* 13, 307–308 (2015). [PubMed: 26484569]
23. Regev A et al. The Human Cell Atlas. *bioRxiv* (2017).

Figure 1. Texomer improved DNA/RNA-joint TCGA BRCA sample categorization.

(a) Texomer jointly deconvolutes tumor variant (V) and wildtype (W) allele-specific DNA and RNA profiles from autologous bulk WES and WTS data and outputs pathological metrics (such as tumor purity and ITH) and functional (DACRE) scores for individual variants (color-filled circles) through subtractive comparison of transformed RNA and DNA profiles (Y axes). (b) Illustration of Texomer deconvolution steps. Texomer iteratively estimates ASCN (grey and green boxes), tumor purity (α_D), ITH scores from the bulk WES data (steps 1–4), and tumor purity (α_R) and ASELs from the bulk tumor WTS data (step 5). It then probabilistically classifies variants into 3 categories (step 6): ASCN-concordant ($ASEL \approx ASCN$), ASCN-discordant high ($ASEL > ASCN$) or low ($ASEL < ASCN$). Shown in (c) and (d) are the t-SNE 8 plot of BRCA samples based respectively on the bulk WES read counts (c) and the bulk WTS read counts (d) from 832 BRCA samples. Clustering was performed using DBSCAN 9 and clusters are labeled with integer IDs. Samples in (c) and (d) are co-colored based on their cluster IDs in (c). Shown in (e) and (f) are the t-SNE plot of tumor samples based respectively on Texomer TT-DNA copy number (e) and TT-RNA expression (f) profiles from the same sample size. Clustering was similarly performed using DBSCAN and samples in (e) and (f) are co-colored based on cluster IDs in (e). Shown in (g) and (h) are the variance of the bulk WTS read counts explained by the bulk WES read counts, variance of the bulk WTS read counts explained by the TT-DNA copy number profile, and variance of the TT-RNA expression profile explained by the TT-DNA copy number profile, respectively on the total (g) and the allele-specific (h) data. P value

determined by one-tailed t test ($n = 832$). Error bars correspond to the 95% confidence interval of the average proportion of variance explained across 832 samples.

Figure 2. Application of Texomer for functional variant characterization.

Plotted are the frequencies of selectively expressed germline variants (SEG, Y-axis) and selectively expressed somatic variants (SESV, X-axis) in TCGA 832 BRCA (a) and 465 SKCM (b) samples. Gene names are labeled for the top 10 highest frequent variants. Somatic mutations identified from WES data of 832 BRCA (c) and 465 SKCM (d) samples were annotated by a set of 8 widely-used functional variant annotators, then further filtered by Texomer DACRE scores (>0). Vertical bars contrast the averaged precision of functional mutation prediction before and after performing Texomer filtering. P value determined by one-tailed t test. Error bars correspond to the 95% confidence interval of the averaged precision of functional mutation prediction.