Pan-cancer study of heterogeneous RNA aberrations

Nuno A. Fonseca[1,38*], Andre Kahles[2,3,38*], Kjong-Van Lehmann[2,3,38*], Claudia Calabrese[1*], Aurelien Chateigner[4*], Natalie R. Davidson[2,3,5*], Deniz Demircioğlu[6,7*], Yao He[8*], Fabien C. Lamaze[4*], Siliang Li[9,10*], Dongbing Liu[9,10*], Fenglin Liu[8*], Marc D. Perry[4,11*], Hong Su[9,10*], Linda Xiang[4*], Junjun Zhang[4*], Samirkumar B. Amin[12$], Peter Bailey[13], Brian Craft[14], Milana Frenkel-Morgenstern[15], Mary Goldman[14], Liliana Greger [1], Katherine A. Hoadley[16], Yong Hou[9,10], Ekta Khurana[5], Jan O. Korbel[17], Chang Li[9,10], Xiaobo Li[9,10], Xinyue Li[9], Xingmin Liu[9,10], Youyong Lu[18], Maximillian G. Marin[14], Matthew Meyerson[19,20,21], Tannistha Nandi[7], Morten M. Nielsen[22], Akinyemi I. Ojesina[23,24], B.F. Francis Ouellette[25], Qiang Pan-Hammarström[9,26], Chandra Sekhar Pedamallu[27], Jakob S. Pedersen[22], Yuichi Shiraishi[28], Reiner Siebert[29], Cameron M. Soulette[14], Stefan G. Stark[2], Patrick Tan[7,30], Bin Tean Teh[31], Alfonso Valencia[32], Jian Wang[9], Rui Xing[18], Heng Xiong[9,10], Sergei Yakneen[33], Chen Ye[9,10], Christina Yung[4], Fan Zhang[8], Xiuqing Zhang[9], Liangtao Zheng[8], Jingchun Zhu[14], Shida Zhu[9,10], Philip Awadalla[34#], Chad J. Creighton[35#], Jonathan Göke[7#], Roland F. Schwarz[36#], Oliver Stegle[1#], Kui Wu[9,10#], Huangming Yang[9#], Zemin Zhang[8#], Alvis Brazma[1,39#], Gunnar Rätsch[2,3,5,37,39#], Angela N. Brooks[14,19,21,39#], PCAWG Network

[1]European Molecular Biology Laboratory, Hinxton, CB10 1SD, UK , [2]ETH Zurich, Zurich, 08092, Switzerland, [3]Memorial Sloan Kettering Cancer Center, New York, 10065, USA, [4]Ontario Institute for Cancer Research, Toronto, M5G 0A3, Canada, [5]Weill Cornell Medical College, New York, 10065, USA, [6]National University of Singapore, Singapore, 117417, Singapore, [7]Genome Institute of Singapore, Singapore, 138672, Singapore, [8]Peking University, Beijing, 100871, China, [9]BGI-Shenzhen, Shenzhen, 518083, China, [10]China National GeneBank-Shenzhen, Shenzhen, 518083, China, [11]University of California, San Francisco, Toronto, M5G 0A3, Canada, [12]The UT MD Anderson Cancer Center, Houston, 06032, USA, [13]University of Glasgow, Glasgow, G61 1BD, UK , [14]University of California, Santa Cruz, Santa Cruz, 95064, USA, [15]Bar-Ilan University, Safed, 13195, Israel, [16]The University of North Carolina at Chapel Hill, Chapel Hill, 27599, USA, [17]European Molecular Biology Laboratory, Heidelberg, 69117, Germany, [18]Peking University Cancer Hospital & Institute, Beijing, 100142, China, [19]Dana-Farber Cancer Institute, Boston, 02215, USA, [20]Harvard Medical School, Boston, 02215, USA, [21]Broad Institute, Cambridge, 02142, USA, [22]Aarhus University, Aarhus, DK-8200, Denmark, [23]University of Alabama at Birmingham, Birmingham, 35294, USA, [24]HudsonAlpha Institute for Biotechnology, Birmingham, 35294, USA, [25]University of Toronto, Toronto, M5G 0A3, Canada, [26]Karolinska Institutet, Stockholm, 14186, Sweden, [27]Broad Institute, Boston, 02215, USA, [28]The University of Tokyo, Minato-ku, 108-8639, Japan, [29]Ulm University, Ulm, 89081, Germany, [30]Duke-NUS Graduate Medical School, Singapore, 169857, Singapore, [31]National Cancer Centre Singapore, Singapore, 169610, Singapore, [32]Barcelona Supercomputing Center, Barcelona, 08034, Spain, [33]European Molecular Biology Laboratory, Heidelberg, 69120, Germany, [34]University of Toronto, Toronto, M8G 0A8, Canada, [35]Baylor College of Medicine, Houston, 77030, USA, [36]Max Delbruck Center for Molecular Medicine, Berlin, 13125, Germany, [37]University Hospital Zurich, Zurich, 08091, Switzerland, [$]Present address: The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

*Co-first author

Senior authors

[38]Key-contributors

[39]Working-group leaders: Alvis Brazma (brazma@ebi.ac.uk), Gunnar Rätsch (raetsch@inf.ethz.ch), Angela N. Brooks (anbrooks@ucsc.edu)

Abstract

Pan-cancer studies have transformed our understanding of recurrent somatic mutations that contribute to cancer pathogenesis; however, there has yet to be a full investigation of the multiple mechanisms in which genes can be somatically altered, particularly at the transcriptome level. We present the most comprehensive catalogue of cancer-associated gene alterations through extensive characterization of tumor transcriptomes from 1,188 donors of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project with matched whole-genome sequence data. We processed the RNA-seq data with a unified analysis pipeline that included both sequence alignment and extensive quality control. Subsequently, we identified gene alterations through gene expression, alternative splicing, alternative transcription starts, allele-specific expression, RNA-edited sites, and gene fusions, and by comparing with RNA-Seq from a panel of normal tissue samples from the Genotype-Tissue Expression (GTEx) project. Our data represent an extensive pan-cancer catalog of RNA-level aberrations for each gene and will be the basis for further analyses within PCAWG. NOTE TO READERS: This is a draft of a marker paper from the PCAWG Transcriptome Working Group and is intended to describe *technical* aspects of RNA-Seq analysis associated with the PCAWG project. The full marker paper is currently in preparation.

Introduction

The International Cancer Genome Consortium's Pan-Cancer Analysis of Whole Genomes (ICGC PCAWG) project was formed to obtain a more complete view of cancer genome alterations that have been missed by previous pan-cancer whole-exome studies. To increase power to detect commonly mutated genes and genomic regions, as well as identify common mutational patterns across cancers, the consortium pooled together all available whole-genome sequencing data from ICGC, including a large portion of data coming from The Cancer Genome Atlas (TCGA). This resulted in 2,583 donors with whole-genome sequencing data (Campbell et al. 2017), thereby generating a unique resource to study both coding and noncoding genetic alterations for 27 cancer types.

The vast majority of somatic cancer genome alterations impacts the non-coding genome (Khurana et al. 2016). Somatic non-coding alterations have been shown to alter gene regulatory information found in enhancers, promoters, and introns to give rise to RNA-level changes. (Khurana et al. 2016; Rheinbay et al. 2017). The identification of such alterations requires not only profiling of somatic mutations in the tumor genome, it also requires profiling of the matched transcriptome. Genes can be altered at the RNA-level through multiple mechanisms such as regulation of gene expression, alternative splicing, alternative promoter usage, formation of gene fusions, allele-specific expression, and RNA editing, warranting a comprehensive analysis of cancer transcriptomes. Here we describe the largest currently available set of matched whole genome and transcriptome sequencing for 1,188 samples from 27 different cancer types. We analyzed heterogeneous transcriptome alterations generating a key resource of the PCAWG project.

Pan-cancer unified analysis of RNA-level alterations from 1,188 donors

For our comprehensive transcriptome analysis, we processed all tumor samples with RNA-Seq data from the PCAWG consortium provided by 30 projects. To reduce batch effects due to the use of different computational pipelines, we reanalyzed 2,217 RNA-Seq libraries with metadata information using a unified RNA-Seq analysis pipeline.

Core components of the unified analysis were spliced alignment of the RNA-Seq data followed by gene expression quantification (Figure 1a). To maximize the utility of the unified alignments with multiple downstream analysis tools, we aligned the data using STAR (Dobin et al. 2013) and TopHat2 (Kim et al. 2013) (Figure 1a).

Gene expression, in the form of read counts normalized by gene length, was quantified with HTSeq (Anders, Pyl, and Huber 2015) separately for both alignment methods. Quantification from both alignment approaches were found to be highly correlated (Extended Data Figure 1a); therefore, we generated consensus gene expression measurements by averaging read counts for each gene. After upper-quartile normalization (UQ) to create a more uniform distribution of values across the full dataset (Bullard et al. 2010; Dillies et al. 2013) (Extended Data Figure 1b),

we produced final gene expression measurements as fragments per kilobase of million mapped reads with upper quartile normalization (FPKM-UQ). As expected, for high to moderately expressed genes, FPKM-UQ quantification across TCGA samples were highly correlated with TCGA-reported gene expression using RSEM (B. Li and Dewey 2011), with an overall median correlation of 0.95 (Extended Data Figure 2). Transcript isoform-specific expression was computed using Kallisto (Bray et al. 2016).

Quality control (QC) analysis of all datasets was performed using metrics of the raw read data (FASTQ), the aligned data, and gene expression measurements. In total we defined 6 QC criteria to evaluate whether a single sample should be marked for exclusion from downstream analysis (see Methods, Extended Data Figure 3). After QC filtering and merging 722 libraries identified as technical replicates, we obtained a total number of 1,359 RNA-seq sample aliquots from 1,188 donors. 755 (64%) of the tumor transcriptomes were derived from The Cancer Genome Atlas. Using these uniformly processed samples, we studied additional forms of RNA alteration through fusions, alternative promoter usage, alternative splicing, allele-specific expression, and RNA editing (Figure 1b).

The number of samples per histotype varied from 2 to 154 (Figure 2a) with approximately a 1:1 ratio of male and female donors (Figure 2b). Only 13 of the 27 histotypes had matched, adjacent normal tissue data, giving a total of 150 normal tissues samples (Figure 2a,c). To increase the panel of normal tissue expression for comparison with tumor expression, we processed RNA-Seq from 3,274 samples in the Genotype-Tissue Expression (GTEx) Consortium (version phs000424.v4.p1) using the same computational pipeline used for the PCAWG dataset. Generally, we observed that samples derived from a similar tissue of origin had similar expression profiles. More specifically, tumor samples from donors derived from different projects were similar and also tissue derived from GTEx versus matched normal tissue were similar (Figure 2d, Extended Data Figure 4), suggesting that this is a useful resource to identify cancer-specific RNA aberrations. Other forms of RNA alterations were examined across tumor samples yielding 3,297 RNA fusions, 70,937 promoter events, 588,324 alternative splicing events, 4,109 genes with allele-specific expression in at least 10% of samples, and 1,646 RNA editing events (Table 1).

Here, we describe the largest resource of matched cancer transcriptome and whole-genome profiling, providing numerous opportunities to identify novel cancer-associated genes, to investigate the impact of coding and non-coding mutations on the transcriptome, and to link genetic alterations with molecular transformations that contribute to cancer development.

Methods

RNA-Seq alignment and quality control analysis

Normal and tumour ICGC RNA-seq data, included in the PCAWG cohort (Campbell et al. 2017), was aligned to the human reference genome (GRCh37.p13) using two read aligners: STAR (Dobin et al. 2013) (version 2.4.0i, 2-pass), performed at MSKCC and ETH Zürich, and TopHat2

(Kim et al. 2013) (version 2.0.12), performed at the European Bioinformatics Institute. Both tools used Gencode (release 19) as the reference gene annotation. For the STAR 2-pass alignment, an initial alignment run was performed on each sample to generate a list of splice junctions derived from the RNA-seq data. These junctions were then used to build an augmented index of the reference genome per sample. In a second pass, the augmented index was used for a more sensitive alignment. Alignment parameters have been fixed to the values reported in (https://github.com/akahles/icgc_rnaseq_align). The TopHat2 alignment strategies also followed the 2-pass alignment principle, but was performed in a single alignment step with the respective parameter set. For the TopHat2 alignments the irap analysis suite (Fonseca et al. 2014) was used. The full set of parameters is available along with the alignment code in (https://hub.docker.com/r/nunofonseca/irap_pcawg/). For both aligners, the resulting files in BAM format were sorted by alignment position, indexed, and are available for download in the GDC portal (https://portal.gdc.cancer.gov/) and in the Bionimbus Data Cloud (https://bionimbus-pdc.opensciencedatacloud.org/pcawg/). The individual accession numbers and download links can be found in the ICGC data release table: http://pancancer.info/data_releases/may2016/release_may2016.v1.4.tsv.

Quality control of all datasets was performed at three main levels: i) assessment of initial raw data using FastQC (Bioinformatics 2011) (version 0.11.3); ii) assessment of aligned data (percentage of mapped and unmapped reads for both alignment approaches); iii) quantification (by correlating the expression values produced by the STAR and TopHat2 based expression pipelines) (Extended Data Figure 1). In total we defined 6 QC criteria to assess the quality of the samples. We marked a sample as a candidate for exclusion if: (1) three out five main FastQC measures (base wise quality, k-mer overrepresentation, guanine-cytosine content, content of N bases and sequence quality) did not pass; (2) more than 50% of reads were unmapped or less than 1M reads could be mapped in total using the STAR pipeline; (3) more than 50% of reads were unmapped or less than 1M reads could be mapped in total using the TopHat2 pipeline; (4) we measured a degradation score (Cancer Genome Atlas Research Network 2015) greater than 10; (5) the fragment count in the aligned sample (averaged over STAR and TopHat2) was less than 5M; (6) the correlation between the expression counts of both pipelines was less than 0.95. If a sample did not pass one of these six criteria it was marked as problematic and were placed on a graylist. If more than two criteria was not passed, we excluded the sample.

A subset of 722 libraries from the projects ESAD-UK, OV-AU, PACA-AU, and STAD-US were identified as technical replicates generated from the same sample aliquot. These libraries were integrated post-alignment for both the STAR and the TopHat2 pipelines using samtools (H. Li et al. 2009) into combined alignment files. Further analysis was based on these files. Read counts of the individual libraries were integrated to a sample-level count by adding the read counts of the technical replicates.

Initially, a total of 2,217 RNA-seq libraries were fully processed by the pipeline. QC filtering and integration of technical replicates (722 libraries) gave a final number of 1,359 fully processed RNA-seq sample aliquots from 1,188 donors.

GTEx data analysis

For a panel of normal RNA-Seq data from a variety of tissues, data from 3,274 samples from GTEx (phs000424.v4.p1) were used and analyzed with the same pipeline as PCAWG data for quantifying gene expression. A list of GTEx identifiers are provided at https://www.synapse.org/#!Synapse:syn7596611.

Quantification and normalization of transcript and gene expression

STAR and TopHat2 alignments were used as input for HTSeq (Anders, Pyl, and Huber 2015) (version 0.6.1p1) to produce gene expression counts. Gencode v19 was used as the gene annotation reference. Quantification on a per transcript level was performed with Kallisto (Bray et al. 2016) (version 0.42.1). This implementation is available as a Docker container at https://hub.docker.com/r/nunofonseca/irap_pcawg. The implementation of the STAR and TopHat2 quantification is available as Docker containers in:

- https://github.com/akahles/icgc_rnaseq_align
- https://hub.docker.com/r/nunofonseca/irap_pcawg/

, respectively. Consensus expression quantification was performed by taking the average expression based on STAR and TopHat2 alignments. Gene counts were normalized by adjusting the counts to fragments per kilobase of million mapped (FPKM) (Mortazavi et al. 2008) as well as fragments per kilobase of million mapped with upper quartile normalization (FPKM-UQ) where the total read counts in the FPKM definition has been replaced by the upper quartile of the read count distribution multiplied by the total number of protein-coding genes.

The FPKM and FPKM-UQ calculations were:

- FPKM = $(C \times 10^9) / (N \times L)$ where N = total fragment count to protein coding genes, L = length of gene, C = fragment count
- FPKM-UQ = $(C \times 10^9) / (UQ \times L \times G)$ where UQ= upper quartile of fragment counts to protein coding genes on autosomes unequal to zero, G = number of protein coding genes on autosomes

t-SNE analysis

The t-SNE plot in Figure 2D was produced using the RTsne package (Krijthe 2015) (with a perplexity value of 18) based on the Pearson correlation of the aggregated expression (log+1) of the 1000 most variable genes. FPKM expression values per gene were aggregated (median) by tissue (GTEx) and study (PCAWG). Coefficient of variation for each gene was also computed per tissue (GTEx) and study (PCAWG) to determine the 1000 most variable genes.

Estimation of alternative promoter activity

We estimated promoter activities using RNA-seq data and Gencode (release 19) annotations for 70,937 promoter in 20,738 genes. We grouped transcripts with overlapping first exons under the assumption that they are regulated by the same promoter (Frith et al. 2008). We quantified the expression of each transcript from the RNA-Seq data using Kallisto (Bray et al. 2016) and calculated the sum of expression of the transcripts initiated at each promoter to obtain an

estimate of promoter activity. To obtain the relative activity for each promoter, we normalized each promoter's activity by the overall gene's expression. More details can be found in (Demircioğlu et al. 2017).

Allele-specific expression

Allele-specific expression was called using the GATK ASEReadCounter (Castel et al. 2015) toolkit (version 3.4.46) using the TopHat2 alignments. Pre- and post-processing of ASE data was done following (Castel et al. 2015). In particular, ASE sites were restricted to uniquely mappable regions using the ENCODE 50-mer mappability tracks and statistically tested for sequencing errors and misannotation of genotypes. Details of ASE pre- and post-processing can be found in (Brazma et al. *in preparation*).

Site-level ASE data was aggregated to gene-level ASE data through phased genotypes. Phasing was achieved using a combination of statistical phasing based on IMPUTE2 (Howie, Donnelly, and Marchini 2009) and Battenberg copy-number phasing (Nik-Zainal et al. 2012). Details of the phasing procedure can be found in (Brazma et al. *in preparation*). Site- and gene-level ASE calls were tested for allelic imbalance using binomial tests, taking reference bias, the expected allelic imbalance based on the copy-number state of the expressed region and the purity of the sample into account. For mathematical details on the expected allelic imbalance, please refer to (Brazma et al. *in preparation*). Site- and gene-level significance values were independently corrected for multiple testing using Benjamini-Hochberg and a FDR of 5% across the whole cohort.

Identification of RNA fusions

Gene fusions between any two genes were identified based on two different gene fusions detection pipelines: FusionMap (version 2015-03-31) pipeline (Ge et al. 2011) and FusionCatcher (version 0.99.6a)/STAR-Fusion (version 0.8.0) pipeline (Nicorici et al. 2014). The detailed procedure is described in (Fonseca et al. 2017).

Briefly, FusionMap was applied to all unaligned reads from the PCAWG aligned TopHat2 RNA-seq BAM files for each aliquot to detect gene fusions. In the FusionCatcher/STAR-Fusion pipeline, for each aliquot with paired-end RNA-seq reads FusionCatcher was applied to the raw reads, with the genome reference. To reduce the number of false positive fusions, the two sets of fusions were filtered to exclude fusions based on the number of supporting junction reads, sequence homology, occurrence in normal samples (from the GTEx and the PCAWG cohort). To get a high-confident consensus fusion call set from these two pipelines, a fusion to be included in the final set of fusions had to: i) be detected by both fusion detection tools in at least one sample; and/or ii) be detected by one of the methods and have a matched SV in at least one sample. The consensus WGS-based somatic structural variants (version 1.6) were obtained from the PCAWG repository in Synapse (https://www.synapse.org/#!Synapse:syn7596712). Finally, 3,540 fusion events were included as the consensus fusion call set, from these 2,268 were detected by both FusionCatcher/STAR-

Fusion and FusionMap (from these 1821 had SV support) and 1,112 were detected by only one method and had SV support. All fusions are available in Synapse (https://www.synapse.org/#!Synapse:syn10003873).

Identification of alternative splicing

We used the alignments based on the STAR pipeline to collect and quantify alternative splicing events with SplAdder (Kahles et al. 2016). The software has been run with its default parameters with confidence level 3. We generated individual splicing graphs for each RNA-Seq sample for both tumor as well as matched normal samples (when available). All graphs were then integrated into a merged graph to comprehensively reflect all splice junctions observed in all samples together. Based on this combined graph, SplAdder was used to extract alternative splicing events of the following types: alternative 3' splice site, alternative 5' splice site, exon skipping, intron retention, mutually exclusive exons, coordinated exon skip (see Figure in (Kahles et al. 2016)). Each identified event was then quantified in all samples by counting split alignments for each splice junction in any previously identified event and the average read coverage of each exonic segment involved in the event was determined. We then computed a percent spliced in (PSI) value for each event that were then used for further analysis. We further generated different subsets of events, filtered at different levels of confidence, where confidence is defined by the SplAdder confidence level (generally 2), the number of aligned reads supporting each event, the number of samples that were found to support the event by SplAdder, and the number of samples that passed the minimum aligned read threshold.

Identification of RNA editing events

We used an RNA editing events calling pipeline, which is an improved version of a previously published one (Han et al. 2015). Firstly, we summarized the base calls of pre-processed aligned RNA-reads to the human reference in pileup format. Secondly, the initially identified editing sites were then filtered by the following quality-aware steps: (1) The depth of candidate editing site, base quality, mapping quality and the frequency of variation were taken into account to do a basic filter: the candidate variant sites should be with base-quality≥20, mapping-quality≥50, mapped reads≥4, variant-supporting reads ≥3, and mismatch frequencies (variant-supporting-reads/mapped-reads)≥0.1 (2) Statistical tests based on the binomial distribution B(n, p) were used to distinguish true variants from sequencing errors on every mismatch site (Q. Li et al. 2014), where p denotes the background mismatch rate of each transcriptome sequencing, and n denotes sequencing depth on this site. (3) Discard the sites present in combined DNA SNP datasets (dbSNP v.138, 1000 Genome SNP Phase 3, human populations of Dutch (Liao, Smyth, and Shi 2013), and BGI in-house data. (Combined datasets deposited at: ftp://ftp.genomics.org.cn/pub/icgc-pcawg3) (4) Estimate strand bias and filter out variants with strand bias based on two-tailed Fisher's exact test. (5) Estimate and filter out variants with position bias, such as sites only found at the 3'-end or at 5'-end of a read. (6) Discard the variation site in simple repeat region or homopolymer region or <5 bp from splicing site. (7) To reduce false positives introduced by misalignment of reads to highly similar regions of the reference genome, we performed a realignment filtering. Specifically, we extracted variant-

supporting reads on candidate variant sites and realign them against a combination reference (hg19 genome + Ensembl transcript reference v75) by bwa0.5.9-r16. We retain a candidate variant site if at least 90 % of its variant-supporting reads are realigned to this site. Finally, all high confident RNA editing sites were annotated by ANNOVAR (Wang, Li, and Hakonarson 2010). (8) To remove the possibility of an RNA-editing variant being a somatic variant, the variant sites are positionally filtered against PCAWG WGS somatic variant calls. The calls were obtained from the PCAWG repository in Synapse (https://www.synapse.org/#!Synapse:syn7364923). (9) The final two steps of filtering are designed to enrich the number of functional RNA-editing sites. Firstly, we only keep events that occur more than two times in at least one cancer type. Secondly, we only keep events that occur in exonic regions with a predicted function of either missense, nonsense, or stop-loss.

Table 1: Synapse identifiers for RNA alteration data

RNA data	Synapse ID
Extended RNA-Seq metadata	syn7416381
Gene expression (PCAWG)	syn3104297
Gene expression (GTEx)	syn8105922
Transcript-level expression (PCAWG and GTEx)	syn5974793
RNA fusions	syn3107127
Alternative promoters	syn3354819
Alternative splicing	syn3107126
Allele-specific expression	syn5816006
RNA editing	syn7183004

The table gives Synapse identifiers to RNA analyses reported in this study. To find a particular dataset, the url is:https://www.synapse.org/#!Synapse:<Synapse ID> For example, for gene expression data, the full url is: https://www.synapse.org/#!Synapse:syn3104297

Figure Legends

Figure 1: Unified RNA-Seq analysis to identify RNA-level alterations. a. Workflow of RNA-Seq alignment and quantification of gene expression. **b.** Computational methods used to detect additional types of RNA alterations including RNA fusions, alternative promoters, alternative splicing, allele-specific expression, and RNA editing.

Figure 2: Pan-cancer expression profiling of 1,188 PCAWG donors. a. Tumor and normal RNA-Seq data from 27 histotypes. Total number of samples are given to the right of bars. grey bars, normal samples. **b.** Number of female versus male donors. **c.** Total number of tumor and normal samples from the PCAWG study. A subset of tumors were metastatic. **d.** t-SNE analysis of median gene expression aggregated within each project and each GTEx tissue.

References

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq--a Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics* 31 (2): 166–69.

Bioinformatics, Babraham. 2011. “FastQC: A Quality Control Tool for High Throughput Sequence Data.” *Cambridge, UK: Babraham Institute*.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Bullard, James H., Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. 2010. “Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments.” *BMC Bioinformatics* 11 (February): 94.

Campbell, Peter J., Gad Getz, Joshua M. Stuart, Jan O. Korbel, Lincoln D. Stein, and -ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net. 2017. “Pan-Cancer Analysis of Whole Genomes.” *bioRxiv*. doi:10.1101/162784.

Cancer Genome Atlas Research Network. 2015. “The Molecular Taxonomy of Primary Prostate Cancer.” *Cell* 163 (4): 1011–25.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. “Tools and Best Practices for Data Processing in Allelic Expression Analysis.” *Genome Biology* 16 (September): 195.

Demircioğlu, Deniz, Martin Kindermans, Tannistha Nandi, Engin Cukuroglu, Claudia Calabrese, Nuno A. Fonseca, Andre Kahles, et al. 2017. “A Pan Cancer Analysis of Promoter Activity Highlights the Regulatory Role of Alternative Transcription Start Sites and Their Association with Noncoding Mutations.” *bioRxiv*. doi:10.1101/176487.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.” *Briefings in Bioinformatics* 14 (6): 671–83.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Fonseca, Nuno A., Yao He, Liliana Greger, Alvis Brazma, Zemin Zhang, and -. Pcawg-. 2017. “Comprehensive Genome and Transcriptome Analysis Reveals Genetic Basis for Gene Fusions in Cancer.” doi:10.1101/148684.

Fonseca, Nuno A., Robert Petryszak, John Marioni, and Alvis Brazma. 2014. “iRAP - an Integrated RNA-Seq Analysis Pipeline.” doi:10.1101/005991.

Frith, Martin C., Eivind Valen, Anders Krogh, Yoshihide Hayashizaki, Piero Carninci, and Albin Sandelin. 2008. “A Code for Transcription Initiation in Mammalian Genomes.” *Genome Research* 18 (1): 1–12.

Ge, Huanying, Kejun Liu, Todd Juan, Fang Fang, Matthew Newman, and Wolfgang Hoeck. 2011. “FusionMap: Detecting Fusion Genes from next-Generation Sequencing Data at Base-Pair Resolution.” *Bioinformatics* 27 (14): 1922–28.

Han, Leng, Lixia Diao, Shuangxing Yu, Xiaoyan Xu, Jie Li, Rui Zhang, Yang Yang, et al. 2015. “The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers.” *Cancer Cell* 28 (4): 515–28.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. “A Flexible and Accurate Genotype Imputation Method for the next Generation of Genome-Wide Association Studies.” *PLoS Genetics* 5 (6): e1000529.

Kahles, André, Cheng Soon Ong, Yi Zhong, and Gunnar Rätsch. 2016. “SplAdder: Identification, Quantification and Testing of Alternative Splicing Events from RNA-Seq Data.” *Bioinformatics* 32 (12): 1840–47.

Khurana, Ekta, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A. Rubin, and Mark Gerstein. 2016. “Role of Non-Coding Sequence Variants in Cancer.” *Nature Reviews. Genetics* 17 (2): 93–108.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.” *Genome Biology* 14 (4): R36.

Krijthe, Jesse H. 2015. “Rtsne: T-Distributed Stochastic Neighbor Embedding Using Barnes-Hut Implementation.” https://github.com/jkrijthe/Rtsne.

Liao, Y., G. K. Smyth, and W. Shi. 2013. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30.

Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12 (August): 323.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Li, Qiye, Zongji Wang, Jinmin Lian, Morten Schiøtt, Lijun Jin, Pei Zhang, Yanyan Zhang, et al. 2014. “Caste-Specific RNA Editomes in the Leaf-Cutting Ant Acromyrmex Echinatior.” *Nature Communications* 5: 4943.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28.

Nicorici, Daniel, Mihaela Satalan, Henrik Edgren, Sara Kangaspeska, Astrid Murumagi, Olli Kallioniemi, Sami Virtanen, and Olavi Kilkku. 2014. “FusionCatcher - a Tool for Finding Somatic Fusion Genes in Paired-End RNA-Sequencing Data.” doi:10.1101/011650.

Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, et al. 2012. “The Life History of 21 Breast Cancers.” *Cell* 149 (5): 994–1007.

Rheinbay, Esther, Prasanna Parasuraman, Jonna Grimsby, Grace Tiao, Jesse M. Engreitz, Jaegil Kim, Michael S. Lawrence, et al. 2017. “Recurrent and Functional Regulatory Mutations in Breast Cancer.” *Nature* 547 (7661): 55–60.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research* 38 (16): e164.

Figure 1

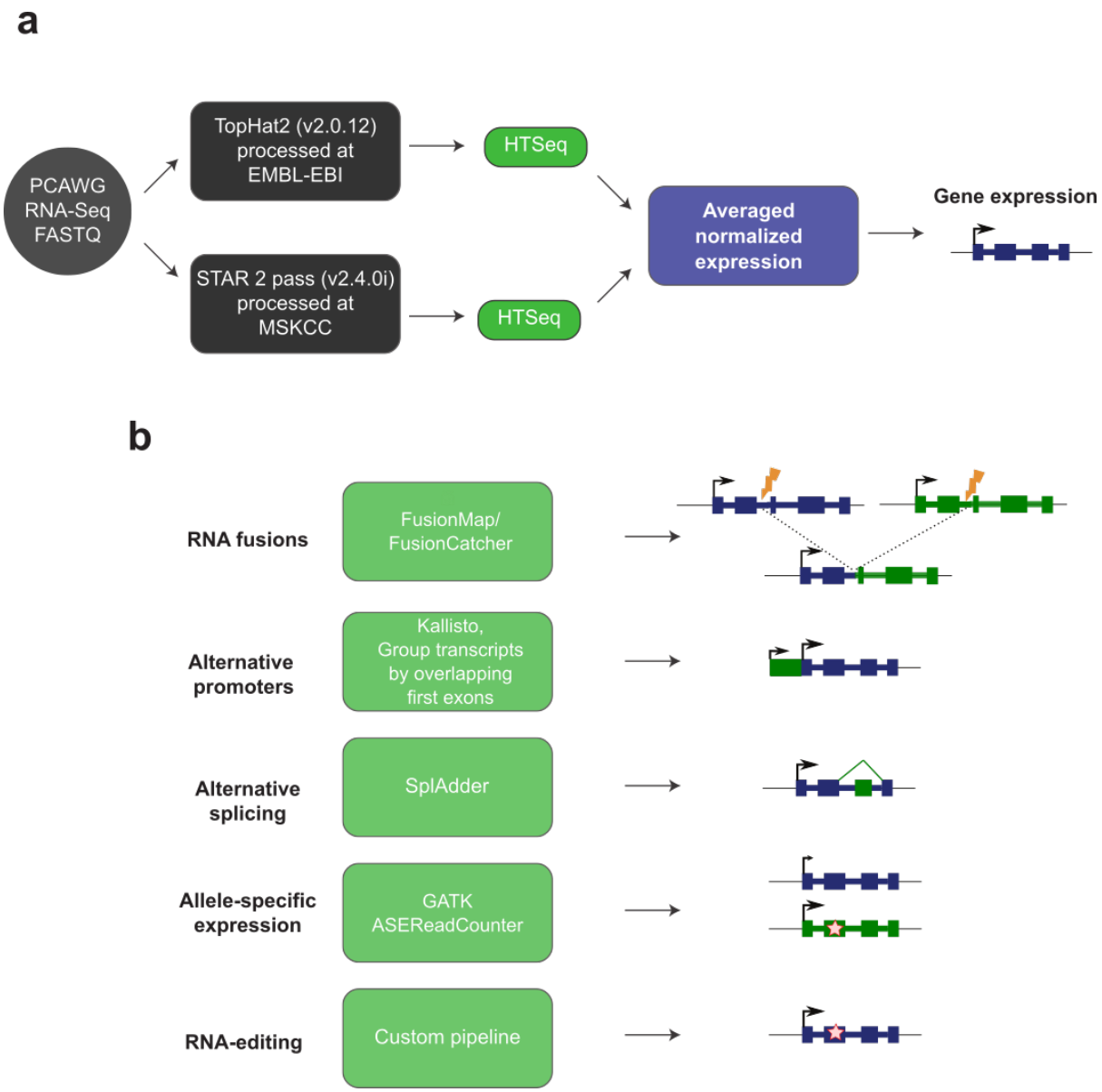
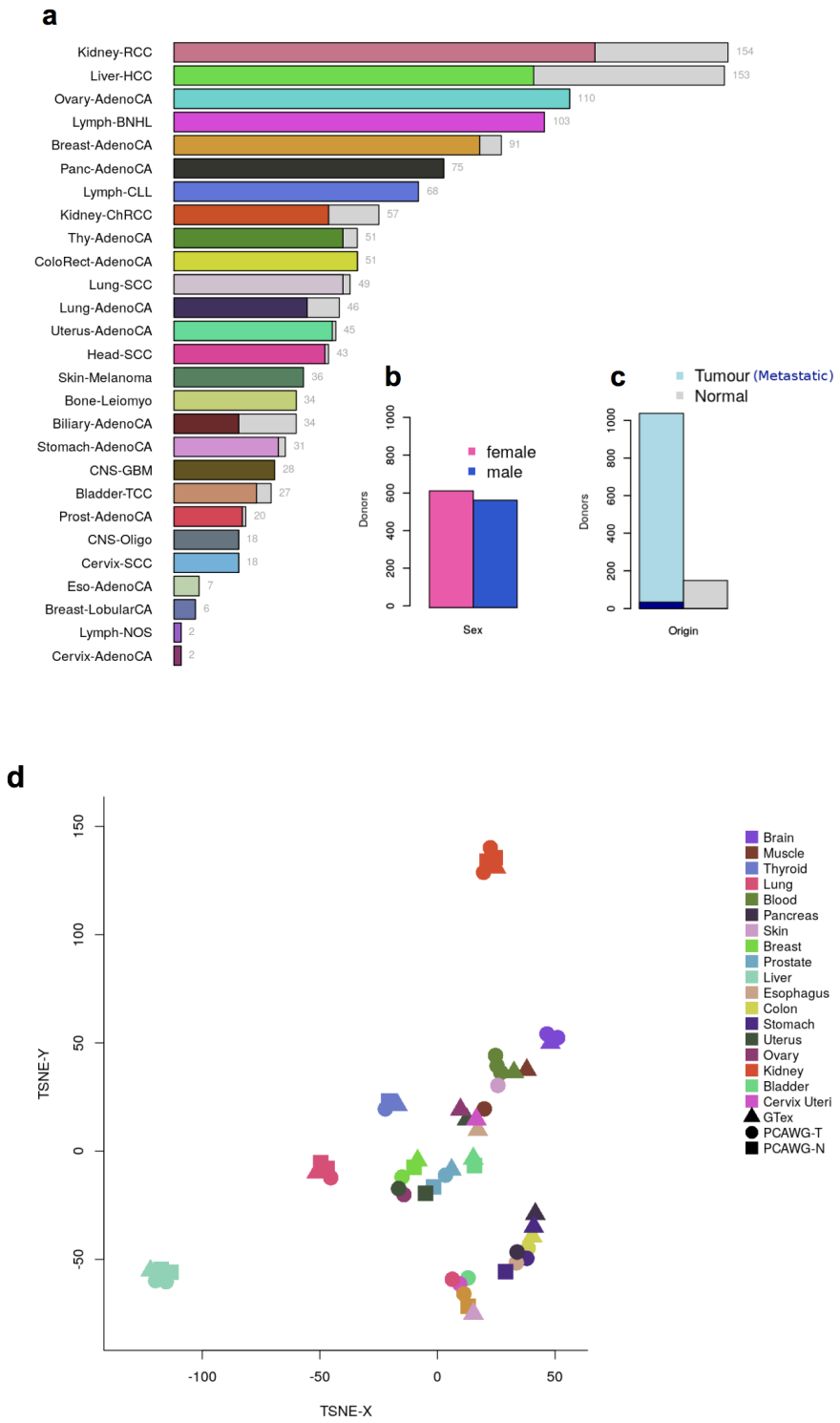
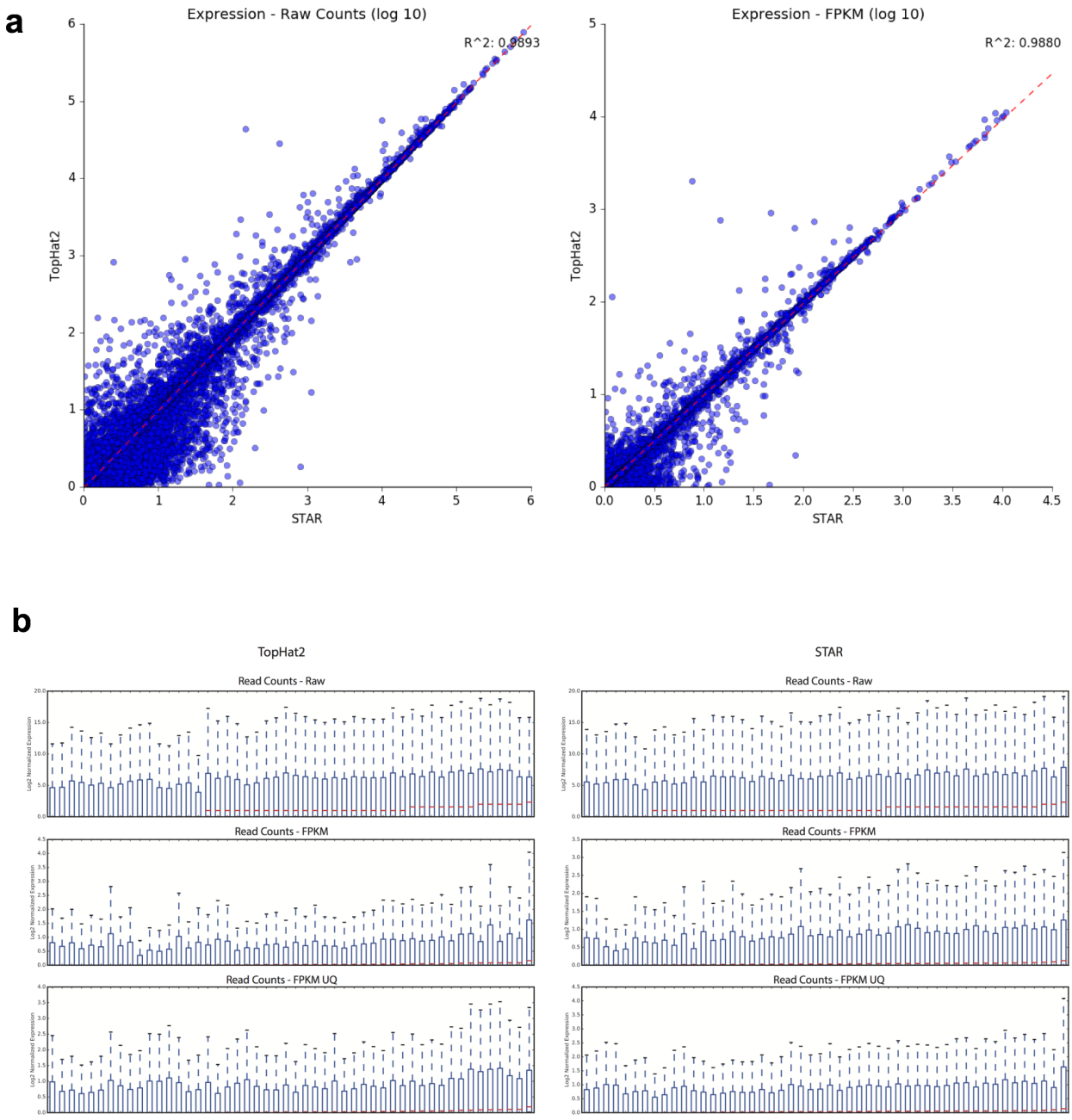

a
PCAWG RNA-Seq FASTQ
TopHat2 (v2.0.12) processed at EMBL-EBI
STAR 2 pass (v2.4.0i) processed at MSKCC
HTSeq
HTSeq
Averaged normalized expression
Gene expression
b
RNA fusions
FusionMap/ FusionCatcher
Alternative promoters
Kallisto, Group transcripts by overlapping first exons
Alternative splicing
SplAdder
Allele-specific expression
GATK ASEReadCounter
RNA-editing
Custom pipeline

Figure 2

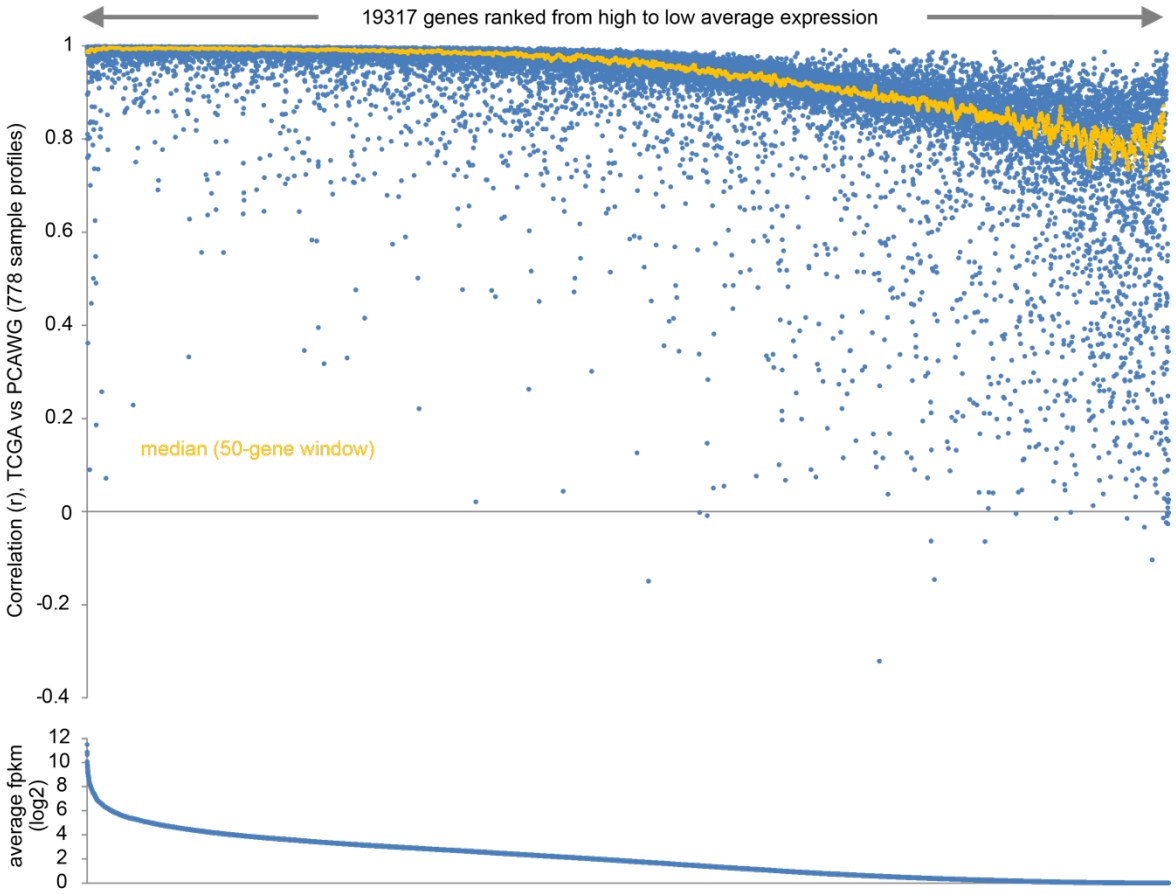

a
Kidney-RCC 154
Liver-HCC 153
Ovary-AdenoCA 110
Lymph-BNHL 103
Breast-AdenoCA 91
Panc-AdenoCA 75
Lymph-CLL 68
Kidney-ChRCC 57
Thy-AdenoCA 51
ColoRect-AdenoCA 51
Lung-SCC 49
Lung-AdenoCA 46
Uterus-AdenoCA 45
Head-SCC 43
Skin-Melanoma 36
Bone-Leiomyo 34
Biliary-AdenoCA 34
Stomach-AdenoCA 31
CNS-GBM 28
Bladder-TCC 27
Prost-AdenoCA 20
CNS-Oligo 18
Cervix-SCC 18
Eso-AdenoCA 7
Breast-LobularCA 6
Lymph-NOS 2
Cervix-AdenoCA 2
b
Donors
0 200 400 600 800 1000
female
male
Sex
c
Tumour (Metastatic)
Normal
Donors
0 200 400 600 800 1000
Origin
d
TSNE-Y
150 100 50 0 -50
TSNE-X
-100 -50 0 50
Brain
Muscle
Thyroid
Lung
Blood
Pancreas
Skin
Breast
Prostate
Liver
Esophagus
Colon
Stomach
Uterus
Ovary
Kidney
Bladder
Cervix Uteri
GTex
PCAWG-T
PCAWG-N

Extended Data Figure 1



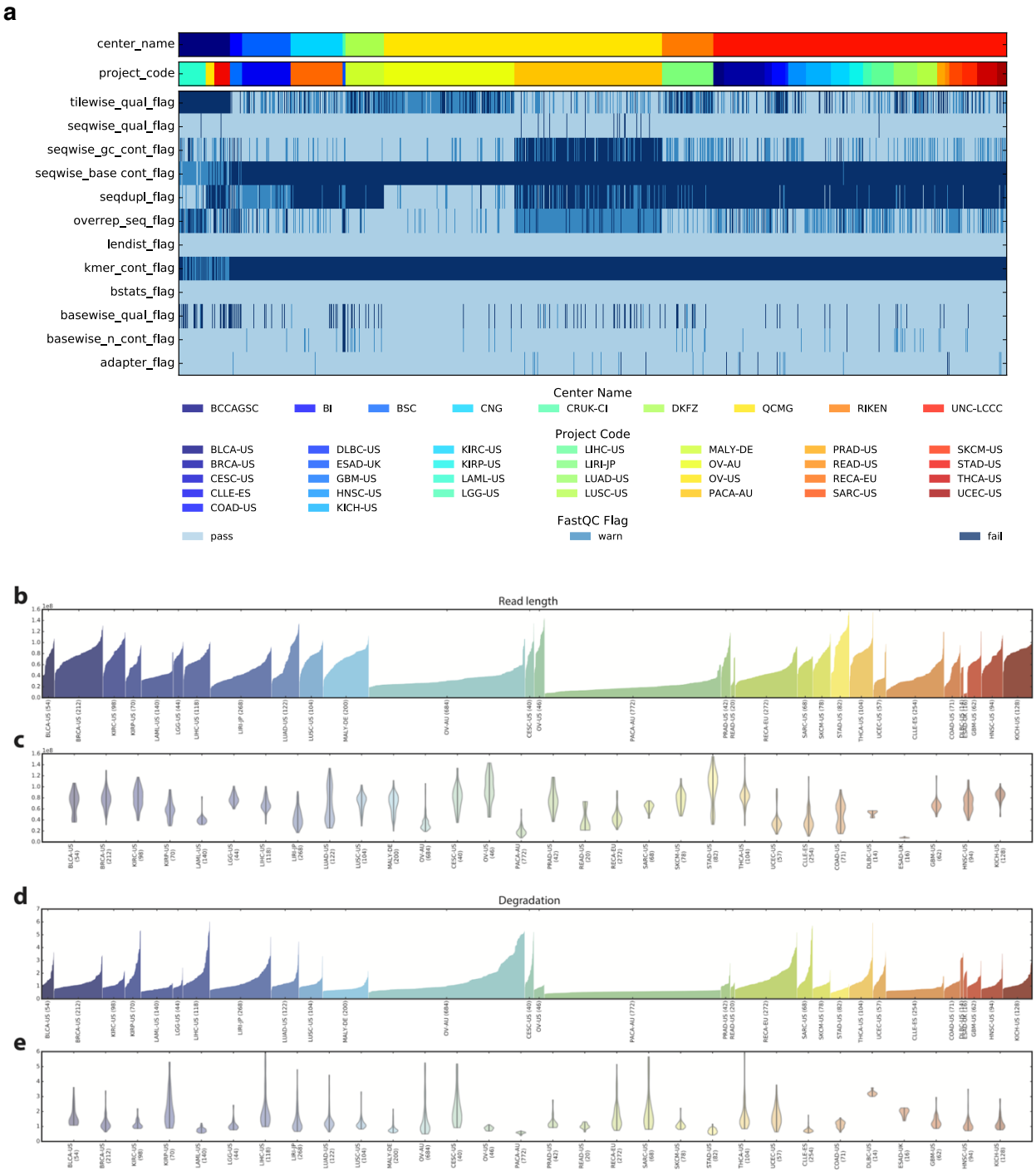
Extended Data Figure 1 | Consensus gene expression quantification and upper-quartile normalization. a. Correlation of STAR vs TopHat2 HTSeq counts. b. Boxplots of raw counts, FPKM, and upper-quartile normalized FPKM values (FPKM-UQ) for representative samples.

Extended Data Figure 2



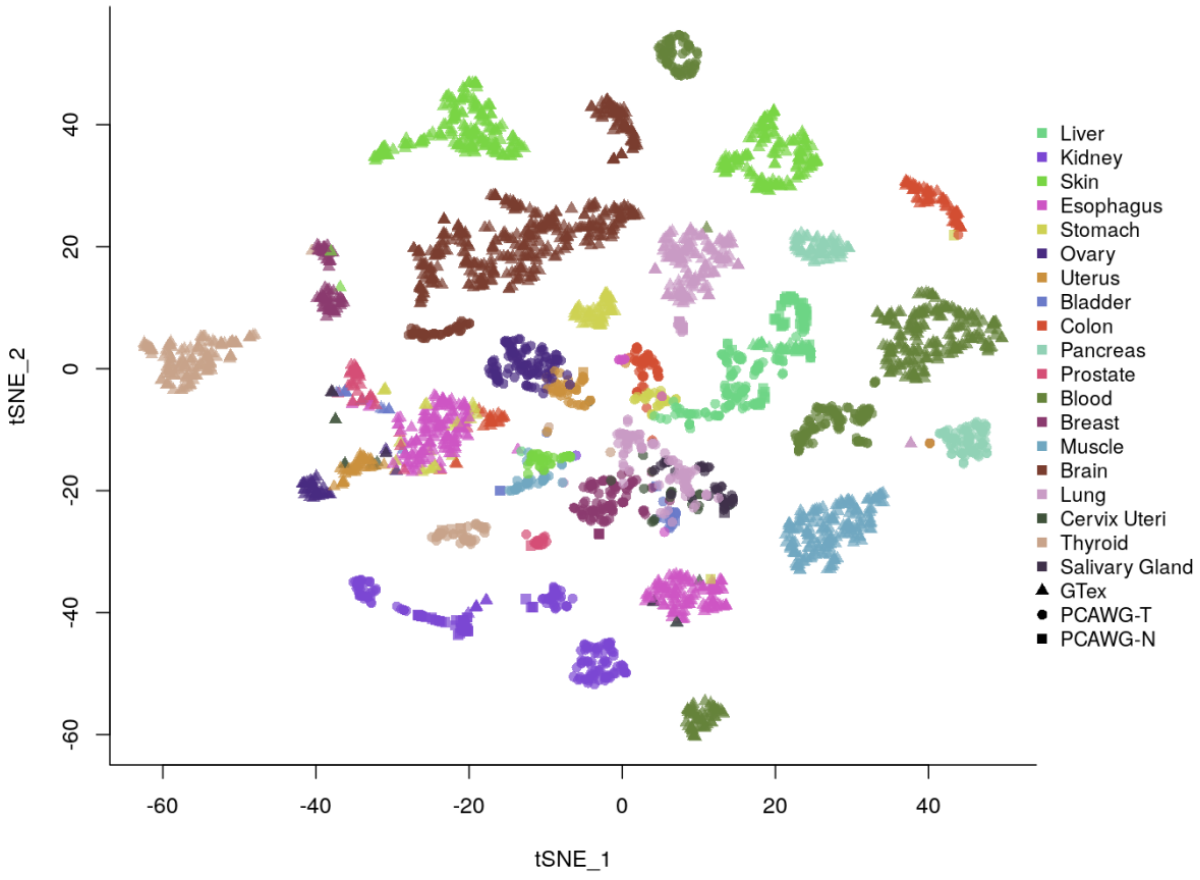
Extended Data Figure 2 | Comparison of TCGA RSEM and FPKM-UQ quantification

Extended Data Figure 3



Extended Data Figure 3 | Quality control analysis of RNA-Seq libraries. a. FastQC flags for each analyzed RNA-Seq library; b. and c. read length distribution; d. and e. degradation measurement based on 3'->5' expression bias. Number of RNA-Seq libraries for each project are indicated.

Extended Data Figure 4



Extended Data Figure 4 | t-SNE analysis of gene expression with all samples shown

Extended Data Table 1

Extended Data Table 1| RNA-editing base substitution frequencies

	Frequency	
Base Pair Substitution	All Regions	Exonic Regions
A>C	110	24
A>G	279615	264
A>T	121	22
C>A	226	2
C>G	98	1
C>T	3456	526
G>A	3013	470
G>C	69	5
G>T	126	6
T>A	140	15
T>C	271442	309
T>G	53	2