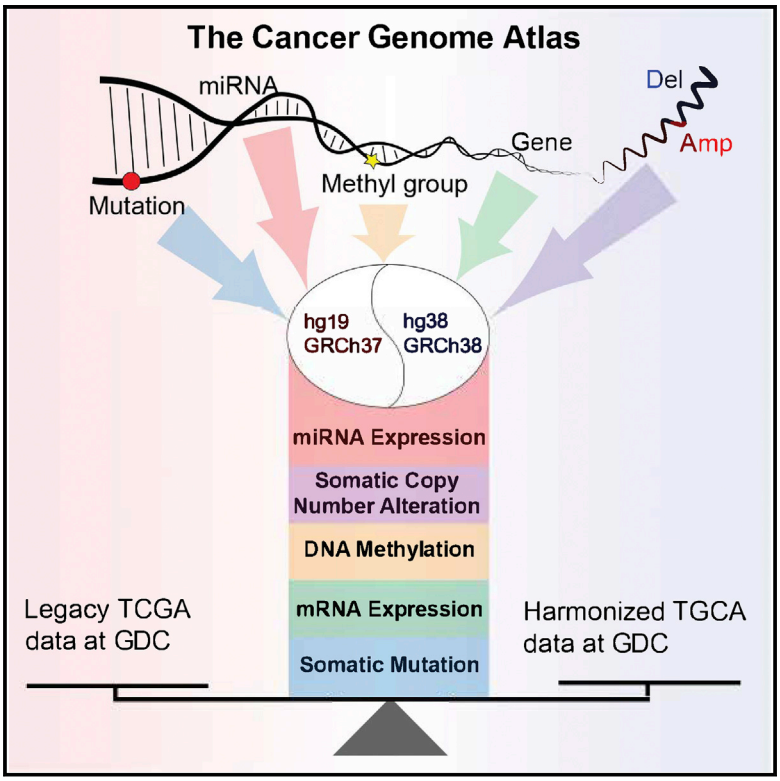
Article

Cell Systems

Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data

Graphical Abstract



Highlights

- A systematic analysis on how the reference genome affects various TCGA data types
- The GRCh37 (hg19) and GRCh38 (hg38) TCGA data versions are highly concordant
- Generate the gene lists showing significant differences between the two versions
- Provide detailed information about TCGA software, pipelines, and annotations

Authors

Galen F. Gao, Joel S. Parker, Sheila M. Reynolds, ..., The Genomic Data Analysis Network, Han Liang, Michael S. Noble

Correspondence

hliang1@mdanderson.org (H.L.), mnoble@cogenimmune.com (M.S.N.)

In Brief

Gao et al. performed a systematic analysis of the effects of synchronizing the large-scale, widely used, multi-omic dataset of The Cancer Genome Atlas to the current human reference genome. For each of the five molecular data platforms assessed, they demonstrated a very high concordance between the 'legacy' GRCh37 (hg19) TCGA data and its GRCh38 (hg38) version as 'harmonized' by the Genomic Data Commons.

Gao et al., 2019, Cell Systems 9, 24–34
July 24, 2019
https://doi.org/10.1016/j.cels.2019.06.006

CellPress

OPEN ACCESS
CellPress

Cell Systems
Article

Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data

Galen F. Gao,[1,2,23] Joel S. Parker,[3,23] Sheila M. Reynolds,[4,23] Tiago C. Silva,[5,6,23] Liang-Bo Wang,[7,8,9,23] Wanding Zhou,[10,23] Rehan Akbani,[11] Matthew Bailey,[7,8,9] Saianand Balu,[12] Benjamin P. Berman,[5,13] Denise Brooks,[14] Hu Chen,[11,15] Andrew D. Cherniack,[1,16] John A. Demchok,[17] Li Ding,[7,8,9] Ina Felau,[17] Sharon Gaheen,[18] Daniela S. Gerhard,[17] David I. Heiman,[1] Kyle M. Hernandez,[19,20] Katherine A. Hoadley,[3] Reyka Jayasinghe,[7] Anab Kemal,[17] Theo A. Knijnenburg,[4] Peter W. Laird,[10] Michael K.A. Mensah,[17] Andrew J. Mungall,[14] A. Gordon Robertson,[14] Hui Shen,[10] Roy Tarnuzzer,[17] Zhining Wang,[17] Matthew Wyczalkowski,[7,8,9] Liming Yang,[17] Jean C. Zenklusen,[17] Zhenyu Zhang,[21] The Genomic Data Analysis Network, Han Liang,[11,15,22,24,*] and Michael S. Noble[1,*]

[1]Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA
[2]The University of Texas Southwestern Medical School, Dallas, TX 75390, USA
[3]Department of Genetics, Lineberger Comprehensive Cancer Center, the University of North Carolin at Chapel Hill, Chapel Hill, NC 27599, USA
[4]Institute for Systems Biology, Seattle, WA 98109, USA
[5]Center for Bioinformatics and Functional Genomics, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
[6]Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP 14.040-905, Brazil
[7]Department of Medicine, Washington University in St Louis, Saint Louis, MO 63108, USA
[8]McDonnell Genome Institute, Washington University in St Louis, Saint Louis, MO 63108, USA
[9]Siteman Cancer Center, Washington University in St Louis, Saint Louis, MO 63108, USA
[10]Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI 49503, USA
[11]Department of Bioinformatics and Computational Biology, the University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[12]Lineberger Comprehensive Cancer Center, Bioinformatics Core, the University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[13]Faculty of Medicine, Department of Developmental Biology and Cancer Research, the Hebrew University of Jerusalem, Jerusalem 91120, Israel
[14]Canada's Michael Smith Genome Sciences Centre, Vancouver, BC V5Z 4S6, Canada
[15]Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX 77030, USA
[16]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[17]National Cancer Institute, Bethesda, MD 20892, USA
[18]Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD 21702, USA
[19]Department of Pediatrics, the University of Chicago, Chicago, IL 60637, USA
[20]Center for Research Informatics, the University of Chicago, Chicago, IL 60637, USA
[21]Center for Translational Data Science, the University of Chicago, Chicago, IL 60615, USA
[22]Department of Systems Biology, the University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[23]These authors contributed equally
[24]Lead Contact
*Correspondence: hliang1@mdanderson.org (H.L.), mnoble@cogenimmune.com (M.S.N.)
https://doi.org/10.1016/j.cels.2019.06.006

SUMMARY

We present a systematic analysis of the effects of synchronizing a large-scale, deeply characterized, multi-omic dataset to the current human reference genome, using updated software, pipelines, and annotations. For each of 5 molecular data platforms in The Cancer Genome Atlas (TCGA)—mRNA and miRNA expression, single nucleotide variants, DNA methylation and copy number alterations—comprehensive sample, gene, and probe-level studies were performed, towards quantifying the degree of similarity between the 'legacy' GRCh37 (hg19) TCGA data and its GRCh38 (hg38) version as 'harmonized' by the Genomic Data Commons. We offer gene lists to elucidate differences that remained after controlling for confounders, and strategies to mitigate their impact on biological interpretation. Our results demonstrate that the hg19 and hg38 TCGA datasets are very highly concordant, promote informed use of either legacy or harmonized omics data, and provide a rubric that encourages similar comparisons as new data emerge and reference data evolve.

INTRODUCTION

Over the course of a decade The Cancer Genome Atlas (TCGA) helped usher in the era of extreme-scale team science, yielding numerous biological insights and many widely cited papers (Hutter and Zenklusen, 2018). Underlying this progress in understanding the molecular bases of cancer is one of the broadest, deepest, and most integratively characterized biological datasets ever assembled: on the order of 2 petabytes of primary and secondary data, in the form of 84,000 data aliquots from

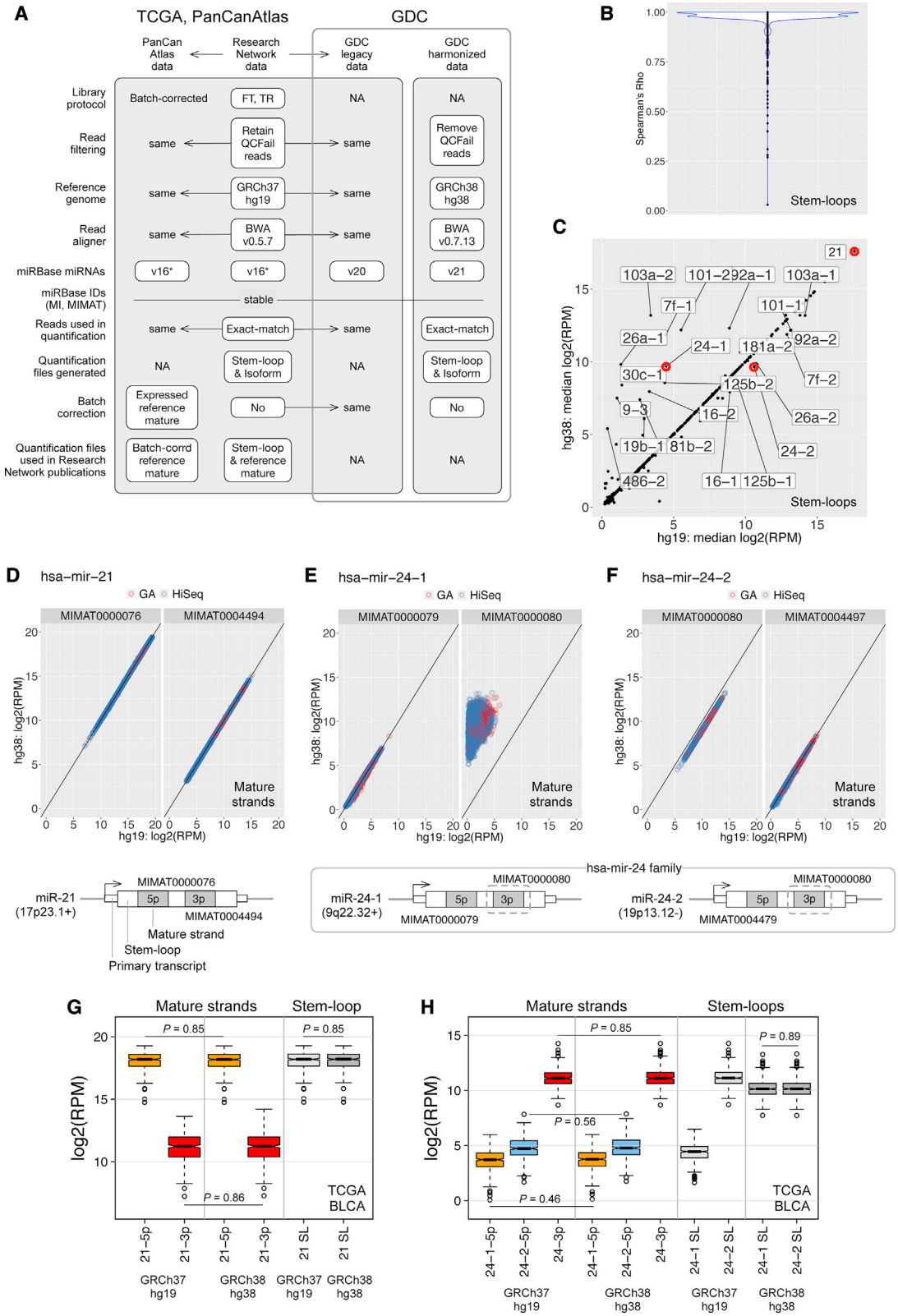


Figure 1. miRNA-Seq Data Processing and Data Comparison in TCGA Legacy and the GDC

(A) Overview of processing steps (rows) and data sets (columns). GDC legacy data and PanCancer Atlas data were derived from the TCGA quantitation-level data. GDC harmonized data were regenerated from TCGA sequence data, using an updated version of the TCGA sequence data processing pipeline. QC comparisons

(legend continued on next page)

some 11,300 patients across 33 disease studies. Most TCGA samples were originally aligned against the Genome Reference Consortium build GRCh37 (hg19), with a small fraction (from the pilot phase of TCGA) having been aligned against NCBI Build 36.1 (hg18). Since TCGA was initiated, however, the research community has undergone tremendous evolution, not only in the characterization machinery, due to the enormous drop in sequencing costs, but also in the surrounding ecosystem of reference data, sequence alignment methods, variant calling tools, RNA quantification methods, quality controls used to help distinguish signal from noise, and analysis software. For this reason, the Genomic Data Commons (GDC, https://gdc.cancer.gov/) was conceived by the National Cancer Institute (NCI) as more than just a massive warehouse of digitized samples: instead, by harmonizing those samples to a uniform reference alignment and gene annotation, then characterizing samples with established tools in consistent workflows and providing updates at regular intervals, the GDC also helps navigate an orderly course through this sea of constant change. The GDC thus offers promise as a force-multiplier for researchers, who can now spend more time exploring their biological questions and less on resolving inconsistencies in data and software versions.

In this paper, we examine the results of the first major harmonization effort undertaken at the GDC: in which the corpus of legacy TCGA data was either aligned or lifted over to the GRCh38 build (hg38) with a GDC workflow assembled from updated versions of bioinformatic tools and reference files used by sequencing and characterization centers in TCGA. While the mechanics of evaluation varied for each data platform, owing largely to natural differences between them and/or how their hg19 counterparts were harmonized to hg38 (e.g., realignment of single nucleotide variants [SNVs] versus liftover of SNP6 copy number arrays), in each case ''the aim was to categorize observed differences in analytic results as a function of their sources and control for such to discern potential impact upon biological interpretation.'' The sources of variation are given in a figure for each platform and include, among others: (1) genome reference; (2) gene annotation: e.g. UCSC genes, GENCODE, miRBase; (3) upstream methods used in alignment, variant calling and quantification, including: BWA (Li and Durbin, 2009), STAR (Dobin et al., 2013), RSEM (Li and Dewey, 2011), FPKM-HTSeq (Anders et al., 2015), and MuTect (Cibulskis et al., 2013); (4) downstream methods used in clustering, correlation, or significance analysis, such as GISTIC (Mermel et al., 2011); (5) parameterizations such as: thresholds for filtering, p-values or q-values; and (6) auxiliary data, such as: GISTIC marker files and CNV lists, or panels of normals used to remove suspect SNVs. In the interest of reproducibility, the supplement describes the software codes and parameterizations used to carry out these studies, and for each platform includes manifests of the input files upon which our analyses were executed. In the remainder of the text we use the terms ''legacy data'' and ''harmonized data'' interchangeably with ''hg19 data'' and ''hg38 data,'' respectively.

RESULTS

miRNA Expression

TCGA miRNA-seq data were generated with a process in which gel-based size selection enriched for library constructs containing ~21-nt inserts, i.e., for 5p and 3p mature strands (Chu et al., 2016). During TCGA, sequencing instruments changed from the Illumina Genome Analyzer II (GAII) to the HiSeq, and sequencing chemistry kits evolved. Each TCGA project used one of two alternative miRNA-seq library construction protocols; the RNA was either the flow-through following poly(A) mRNA purification, or total RNA. For TCGA, sequencing 'QCFail' reads were retained, BWA v0.5.7 aligned reads to the reference human genome (Li and Durbin, 2009), and miRNA expression quantification considered only exact-match aligned reads. All miRNA-seq expression data were initially generated using miRBase v.16 annotations for stem-loops and mature strands; the sequence data were later reprocessed with miRBase's most mature hg19 annotations, v20 (Griffiths-Jones et al., 2006). For GDC harmonized data the reference genome changed to hg38 and the miRBase annotations to v21, QCFail reads were removed prior to alignment (as of GDC data release v11.0), and reads were aligned with BWA v0.7.15.

Figure 1A summarizes differences and similarities between legacy (GRCh37/hg19) and harmonized (GRCh38/hg38) normalized expression data (reads per million mapped reads, RPM). For context, it includes data for TCGA Research Network publications and PanCancer Atlas publications (https://gdc.cancer.gov/node/977). In Figure 1, panels B and C summarize the stem-loop comparisons made in the current study, and panels D through H show examples. All study data are available from the GDC and for interactive querying using SQL from BigQuery tables hosted by the ISB-CGC (Reynolds et al., 2017). For additional details see Supplemental Methods. Our analysis shows that the hg19 and hg38 versions of TCGA miRNA expression quantifications are highly concordant: the $\log_2$(RPM) values for 1,137 (83%) out of the 1,367 miRNA mature strands detected in at least 1,000 samples have a

were done between legacy TCGA hg19 and GDC hg38 harmonized data. Library construction protocols: FT: the flow-through from poly(A) mRNA purification, and TR: Total RNA. Asterisks indicate that while the source data were generated using v16 miRBase annotations, names reported for stem-loops and 5p/3p mature strands, in TCGA publications, may be from a more recent miRBase version; in contrast to names, miRBase MI and MIMAT identifiers are stable.

(B–F) Results of QC comparisons for GDC miRNA-seq data. (B) Distribution of rank correlation coefficients for hg19 versus hg38 reads-per-million normalized abundance (RPMs) for stem-loops across all cancer types and miRNAs. (C) Comparison of hg19 versus hg38 median RPMs for stem-loops. The red circles highlight hsa-mir-21 and two hsa-mir-24 family members, see (D–F). (D–F) RPM comparisons for mature strands: (D) hsa-mir-21, (E) hsa-mir-24-1, and (F) hsa-mir-24-2. Dots represent samples, and are colored to indicate the sequencing instrument (GAII or HiSeq). Schematics below the RPM scatterplots show miRNA stem-loops and cytoband locations for hsa-mir-21, and for hsa-mir-24 family's hsa-mir-24-1 and -2. Dashed lines highlight the 3p mature strand, MIMAT000080, whose reference sequence is identical in each family.

(G and H) Distributions of RPMs for legacy (GRCh37/hg19) and harmonized (GRCh38/hg38) mature strands and stem-loops, for primary tumors from the TCGA muscle-invasive bladder cancer (BLCA) cohort (n = 409): (G) hsa-mir-21, (H) hsa-mir-24-1 and -2. p values are from Wilcoxon test. 'SL': stem-loop. See also Table S1.

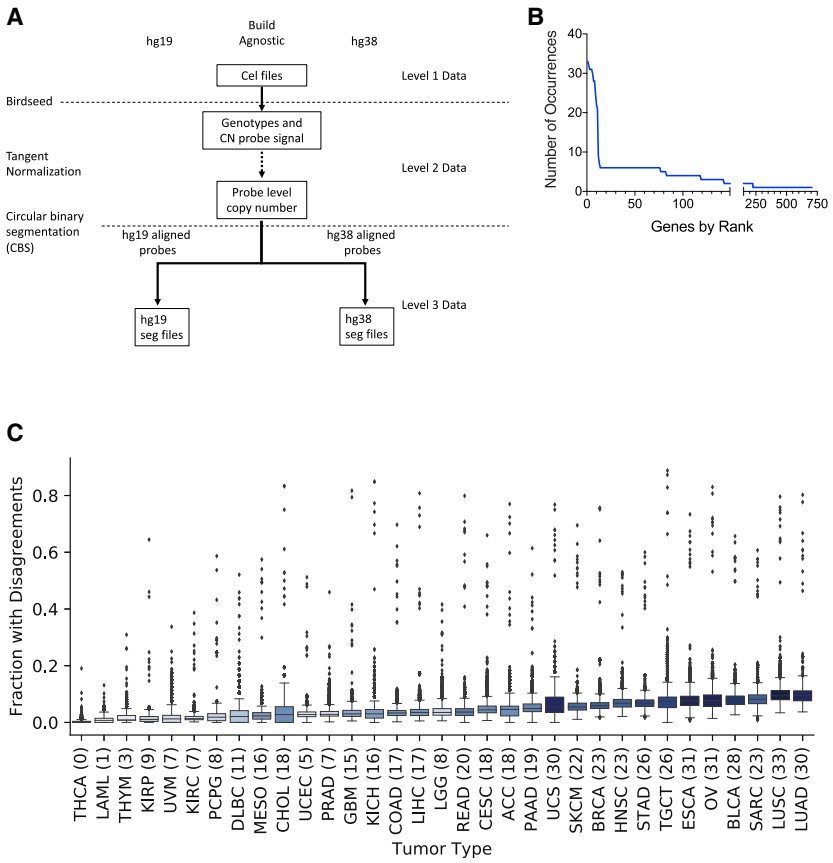


Figure 2. Somatic Copy Number Processing and Data Comparison in TCGA Legacy and the GDC

(A) Affy SNP array copy number pipeline: other than lifting probe loci over to hg38, the pipeline was identical for hg19 and hg38. Probesets used to create Level 3 hg19 and hg38 data were not identical in that 14,811 (0.8%) probes that could not be uniquely mapped in hg38 were not used in segmentation. (B) Genes sorted by number of cancer types in which each gene is "deviant" (as defined earlier). We observe a very small subset of genes that are deviant in more than a few cancer disease types.

(C) Distribution of SCNA disagreements for 20,616 genes between the hg19-aligned run and the hg38-aligned run for 33 TCGA tumor types ordered by increasing median fraction. Boxes are colored by median number of recurrent SCNAs (listed in parentheses) in each tumor type as determined by GISTIC2.0 with the hg19 reference build. See also Figure S1.

correlation coefficient greater than 0.98. We also found that the $\log_2$(RPM) quantification values were highly comparable, despite the removal of QCFail reads: the median absolute difference between hg19 and hg38 $\log_2$(RPM) expression values was less than 0.05 for 1,290 (94%) out of the same 1,367 miRNA mature strands. For 124 (~5%) of the 2,577 miRBase v20 5p or 3p mature strands, reference sequences are identical across members of a miRNA family (e.g., hsa-let-7-5p has the same reference sequence in hsa-let-7a-1 [9q22.32], hsa-let-7a-2 [11q24.1] and hsa-let-7a-3 [22q13.31]). The short (~21-nt) miRNA sequence reads for such mature strands will exact-match align to two or more genomic miRNA locations (Chu et al., 2016). Because the newer version of BWA (v0.7.15) distributed reads more uniformly between such multi-mapping locations than the older version (v0.57), legacy-to-harmonized differences were larger for stem-loops from miRNA families (Figures 1C–1H).

Given the data-generating process and our comparison of the legacy and harmonized datasets, analyses that use overall mature strand expression, e.g., differential expression, or miR-gene targeting, should be insensitive to choosing legacy or harmonized data, and to whether a strand can be expressed from locations across a miRNA family. Consistent with this, mature strand RPM expression values for primary tumors from the muscle-invasive bladder cancer (BLCA) cohort were similar in GDC legacy and harmonized data (Figures 1G and 1H). In contrast, other analyses that depend on specific genomic locations, e.g., stem-loop expression, or where DNA methylation or copy number alterations influence the expression of mature strands and stem-loops, may be sensitive to which dataset is used. Because the data-generating process is not able to quantify the relative contributions of a miRNA family's members to a mature strand's reads, in the harmonized data shared mature strands and stem-loops for miRNA family members are assigned similar RPMs (Figure 1H), and all location-based analyses should consider this when they involve members of a miRNA family.

Somatic Copy Number Alterations

The copy number data studied here were generated from Affymetrix (Santa Clara, California, United States of America) Genome-Wide Human SNP6.0 arrays through probe intensity normalization, tangent normalization, and circular binary segmentation (Olshen et al., 2004), using identical pipelines for both hg19 and hg38 (Figure 2A). Tumor samples were profiled for each disease cohort, with corresponding normal samples when available (blood or adjacent tissue), and germline-subtracted as described by TCGA. We examined both individual genes as well as driver events within peaks of significantly recurring focal alterations [as determined by GISTIC2.0 (Mermel et al., 2011)]. Pre-computed GISTIC analyses of the hg19-aligned data are available at firebrowse.org, while both the legacy hg19-aligned and the harmonized hg38-aligned copy number segmentation profiles are available for download from the GDC and FireBrowse.

To compare relative mean copy numbers, i.e., copy numbers uncorrected for tumor purity and ploidy, we began by identifying the set of genes annotated with the same HUGO name in both hg19 and hg38. For each of these 20,616 genes, we computed the average difference in relative copy number between the hg19 and hg38 runs, over all samples in each TCGA cohort (Data S2.3, excerpted in Figure S1A). We then filtered for deviant genes, those with average differences exceeding 4 standard

deviations from 0 in each disease type. Finally, we examined which of these genes were recurrently deviant across multiple cancer types (Data S2.4, Figure 2B). As described in Supplemental Methods, this analysis identified that from over 20,000 genes only 11 (0.04%) differ in copy number between the two builds in more than 10 TCGA tumor types (*NAA38, SNORD29, SNORA44, MIR1255B1, SNORD79, SNORD44, SNORD22, SNORD31, FAM230C, MIR1827,* and *SNORD30*); and with the notable exception of NAA38, none of these genes encodes proteins. Deviant genes generally appeared with greater frequency in cancer types having large numbers of copy number alterations (LUSC, LUAD, ESCA, SARC, BLCA, and OV), whereas fewer disagreements tended to occur within cancer types harboring fewer copy number alterations (THCA, LAML, KIRP, UVM, THYM, and KIRC) (Figure 2C).

Finally, we compared GISTIC 2.0 analysis utilizing hg19 and hg38 data. To do this, we determined whether driver genes that were previously reported in TCGA marker papers were located within GISTIC peaks found in analyses of the entire TCGA cohort using either genome build. Almost all (93%; 482/521) of these driver alterations in both marker papers and hg19 GISTIC analyses were matched in the analysis of the hg38 data, meaning they were either found in both hg19 and hg38 or not present in either (Data S2.6). For 23 of 39 driver alterations that were found in analyses utilizing only one of the builds, peaks were found within 1.2 mb of the driver gene identified in the analysis utilizing the other build. Significant changes in the location of some SNP probes may account for differences in the boundary of some GISTIC peaks regions. In addition, GISTIC analysis was performed utilizing standard analysis for all tumor types, and so optimizing GISTIC parameters for each disease cohort separately may produce an even greater concordance between analysis using hg19 and hg38 data.

DNA Methylation

The DNA methylation data in this study were derived from one of two array-based platforms: the Infinium Human Methylation 27k (HM27) or Infinium Human Methylation 450k (HM450). As depicted in Figure 3A, the hg38 data were generated by (1) remapping the array features (probes) from the hg19 to the hg38 reference genome sequence, then (2) re-annotating the associations between features and genes using a newer gene annotation database [GENCODE v22 (Harrow et al., 2012)]. Importantly, the processing of raw data to Level 3 methylation (a.k.a. "beta") values was not altered, and thus methylation beta values for individual probes were identical between the hg19 and hg38 versions. The major consequence of the probe remapping was the invalidating of a relatively small number of probes that no longer had a uniquely identifiable location in the hg38 genome (2.0% of probes for the HM27 array and 1.1% of probes for the HM450 array).

In contrast to the relatively small number of changes introduced by genome remapping, the gene reannotation step introduced a large number of changes to probe-gene mappings. While the majority of probes with a gene association in hg19 remained associated to the same gene in hg38 (64% of HM27 and 67% of HM450 probes), a large number of probes (28% of HM27 and 25% of HM450 probes) became associated with one or more new GENCODE v22 genes (Figure S2). Among them, many were non-coding genes, which were largely absent from the earlier annotations used for the hg19 data [RefSeq Gene v. 2010 (Pruitt et al., 2007)]. These included a large number of new linkages to antisense and other lncRNA genes (Figure S2), which have been shown to play important roles in cancer biology (Chiu et al., 2018; Huarte, 2015; Wang et al., 2018) . We quantified the additional biological value of the new associations by performing a global analysis to identify gene expression changes associated with promoter epigenetic regulation (Figure 3B). We identified many more associations using the hg38 annotations, and these occurred across both protein coding ("protein_coding") and the various non-coding (e.g., "antisense" and "lincRNA") gene categories (Figure 3C). We found about 75% more protein coding associations in each cancer type in the hg38 version than in the hg19 version (Figure 3C, left), whereas antisense and lncRNA annotations were almost entirely new (Figure 3C, right). Some of the new protein coding associations involved alternative promoters of known cancer genes that were not represented in the hg19 version, such as epigenetic regulation of the *PAX8* gene in a subset of CHOL tumors (Figures 3D and 3E). While *PAX8* activity has been associated with cancers originating from the thyroid, Müllerian, and renal tracts (Ghannam-Shahbari et al., 2018; Laury et al., 2011), neither this new isoform nor upregulation in cholangiocarcinoma have been previously described. In Supplemental Methods, we describe additional resources available at the GDC that can aid users in analyzing TCGA methylation data: (1) improved HM27/HM450 probe annotation (Zhou et al., 2017) and data generation pipelines (Zhou et al., 2018b), which are planned to become the default processing version in a forthcoming GDC data release; and Whole-Genome Bisulfite Sequence data for 47 TCGA samples (Zhou et al., 2018a), which can be used to investigate whole-genome methylation patterns.

mRNA Expression

TCGA RNA-seq data were generated in a process in which polyA+ RNA were selected and sequenced. Similar to miRNA seq data, during TCGA, both sequencing instruments (from the Illumina Genome Analyzer GA to HiSeq) and sequencing chemistry kits evolved. As described in Supplemental Methods, the bioinformatic workflow for generating hg38 mRNA-Seq data at the GDC differs substantially from that used to generate hg19 mRNA-seq data in TCGA. These differences—in alignment, quantification, normalization, and references (Figure 4A)—are expected to introduce bias between the hg19 and hg38 abundance estimates. To characterize that bias and evaluate concordance with prior results, we performed a large-scale comparison of alignment and gene expression estimates. A total of 2,302 samples were used from the breast, head and neck, and lung squamous cohorts (BRCA = 1205, HNSC = 546, LUSC = 551), across which 19,744 protein-coding genes were studied—all those which unambiguously mapped in both the hg19 and hg38 data. Expression values for hg19 were derived from upper-quartile normalized count estimates, while FPKM estimates were used to derive hg38 expression values.

Genome annotation and alignment biases resulted in increased reporting of rRNA alignments (legacy median = 0.12% of bases; current median = 1.0%; Figure S3), but did not change total mRNA alignments (legacy median = 75.4% of bases; current median = 75.8%). Correlation was performed to assess the concordance of workflow results for each sample.

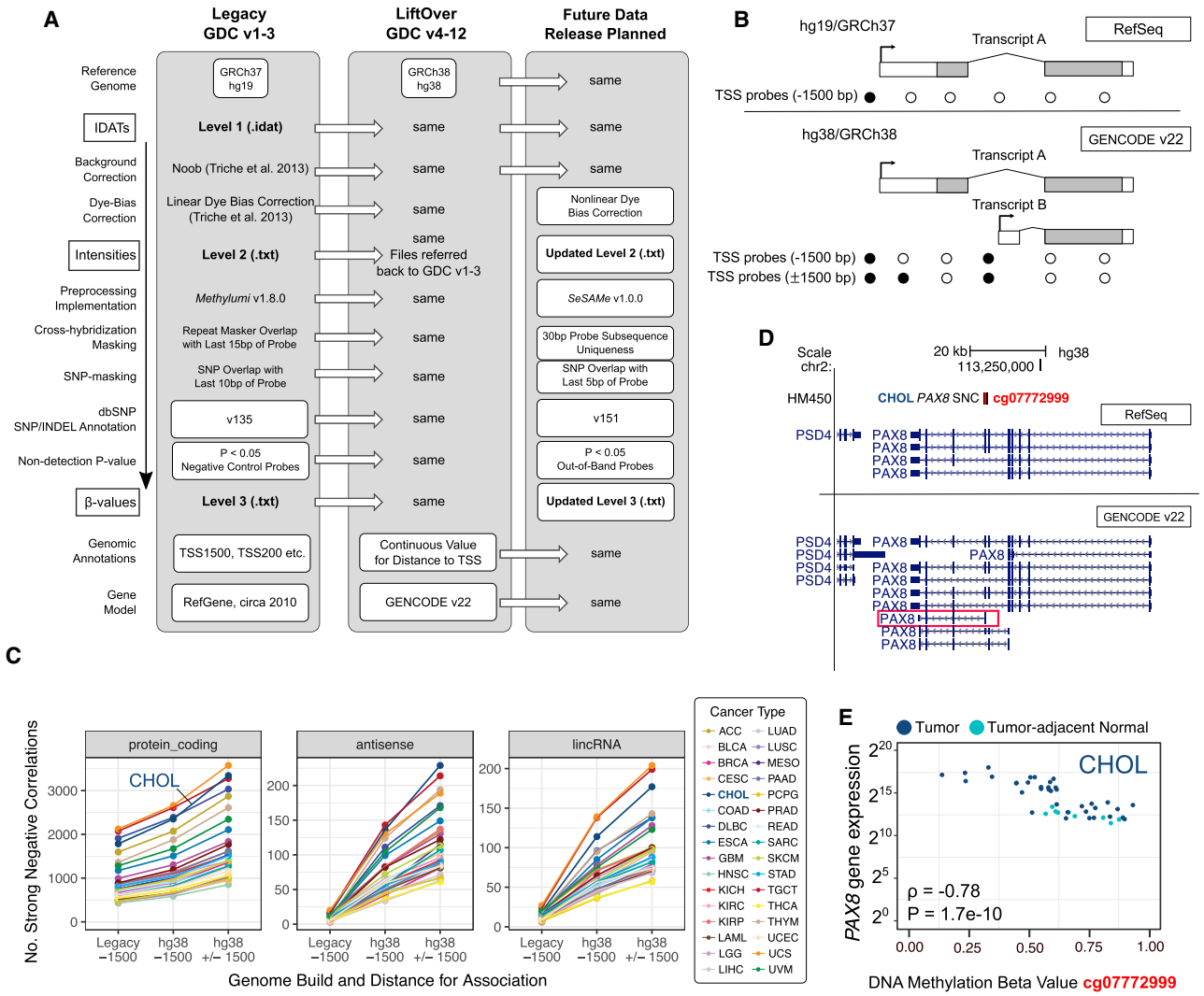


Figure 3. DNA Methylation Processing and Data Comparison in TCGA Legacy and the GDC

(A) Summary of HM27/HM450 processing differences between legacy (hg19, GDCv1-3) and current (hg38, GDCv4-12) versions, and an upcoming version available for manual download in the GDC Community Tools repository (see Supplemental Information for details).

(B) Associating array features with genes in the hg19 and hg38 pipelines: hg19 used the RefSeq version 40 annotations from the Illumina HM450 manifest, and only associated probes within 1,500 bp upstream of a transcript start site ("TSS -1500"); hg38 used GENCODE 22 annotations, and includes distance from the nearest TSS, which can be used to associate probes both upstream or downstream from a TSS ("TSS +/−1,500"). GENCODE 22 often includes additional alternative promoters for the same gene.

(C) Number of Strong Negative Correlations (SNCs) between DNA methylation beta value and RNA expression, using different associations: "Legacy -1500" used hg19 associations, "hg38-1500" used hg38 annotations but only upstream associations, and "hg38 +/−1500" used the same annotations but both upstream and downstream associations. The number of SNCs increased for all transcript types (only three shown here).

(D) Example of a new alternative promoter for *PAX8* present in hg38 annotations but not hg19, which also coincided with an SNC identified in the hg38 but not hg19 version.

(E) Methylation versus expression for this SNC (cg07772999-PAX8) across all TCGA-CHOL samples—about 50% of tumors are demethylated at this alternative promoter and overexpress *PAX8*. See also Figure S2, Tables S2–S5.

Spearman's rank correlation was used because the expression estimates are reported in different scales (estimated count by RSEM versus FPKM). Results indicate that gene rank order is generally preserved (mean Spearman's rho = 0.943; range = 0.893–0.959) (Figure 4B).

Bias in differential expression estimates was evaluated by comparing subtypes within the BRCA, HNSC, and LUSC cohorts. Subtypes of each tumor type were assigned by sample based on the published PAM50 subtypes for BRCA, and the transcriptome subtypes for HNSC and LUSC. Differential gene expression was estimated as the log ratio of counts between two subtypes of the same disease. The log ratios were calculated independently for both workflows, and concordance was estimated by the adjusted r^2 of the two workflows for the same pair of subtypes. The transcriptome-wide effects from the BRCA luminal versus basal demonstrate excellent concordance

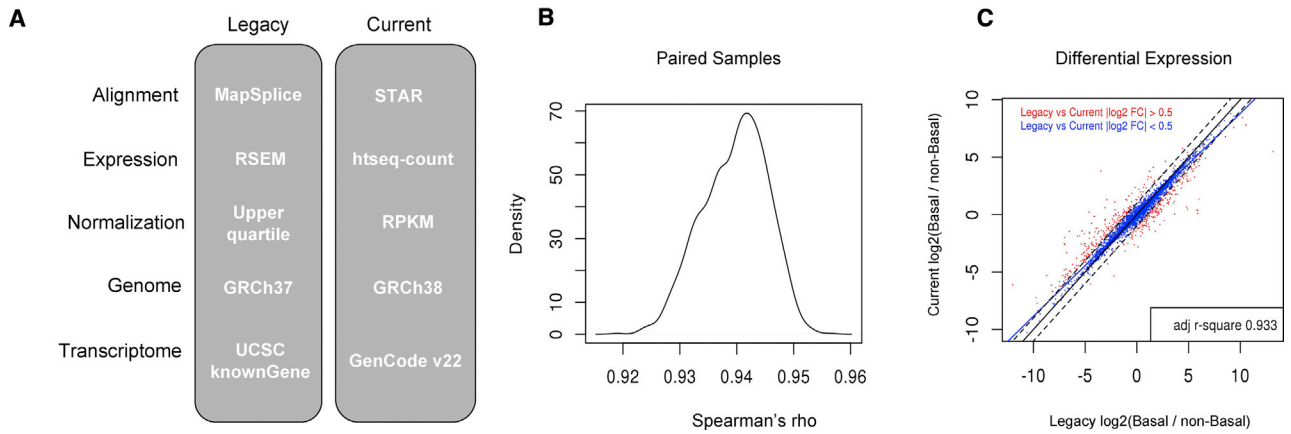


Figure 4. mRNA-Seq Processing and Data Comparison in TCGA Legacy and the GDC

(A) Outline of bioinformatic pipeline steps for TCGA Legacy (hg19) and current GDC (hg38) data. All aspects of sample processing differ including computational methods, the reference genome, and reference transcriptome.

(B) The distribution of sample rank correlation coefficients between matched samples of the two data versions from the BRCA cohort (n = 1205). Correlation estimates arise from comparing gene level counts of the Legacy RSEM output to gene level counts from the Current htseq-count workflow.

(C) Comparison of log ratios between Legacy and Current for the BRCA basal versus non-basal comparison. Each point represents the log ratio of subtypes (basal/non-basal) from the Legacy (x axis) or Current (y axis) workflow. Genes exhibiting >1.5-fold change in either direction are highlighted in red. Log ratio estimates were derived from upper quartile normalized gene level estimates for the Legacy workflow and FPKM transformed gene level estimates from the Current workflow. Log (base 2) ratios between subtypes demonstrate large changes across many genes, while changes between workflows are far fewer in both number and magnitude. See also Figure S3.

(adjusted r^2 of 0.933) (Figure 4C) between the two workflows. Further, the relative change between conditions is preserved across all subtype comparisons attempted (mean R-square = 0.91; range 0.862–0.947; Figure S3). The bias in absolute counts prevents direct comparison of abundance between these workflows, but we find relative abundance results to be highly concordant when restricting to a single workflow.

From our analyses in BRCA, HNSC, and LUSC, we identified 319 genes with a mean absolute difference greater than 1, representing at least a 2-fold change in differential gene expression between the legacy and harmonized pipelines. Many of these genes were from gene families with similar sequence homology such as olfactory genes (Zozulya et al., 2001), keratin-associated proteins (Shibuya et al., 2004), and antigens of the GAGE, PAGE, and XAGE families (Brinkmann et al., 1999), among others. Many mapped to regions that were previously not as well annotated in hg19, including genes near centromeres (~20%) and those near the ends of chromosomes (~30%) (Church et al., 2011; Genovese et al., 2013). Therefore, genes with a significant expression change were often homologs, which is a direct result of the differences in quantification approaches used in the two workflows.

Somatic Mutations

TCGA somatic mutation data were generated by whole-exome sequencing in which exome capture was performed using the Agilent (Santa Clara, California, United States of America) SureSelect Human All Exon kit. During TCGA, multiple sequencing platforms (Illumina and SOLiD) were employed. As described in Figure 5A, the analysis pipeline for calling hg38 somatic mutations at the GDC differs substantially from that for calling hg19 somatic mutations in TCGA legacy version [i.e., multi-Center Mutation Calling in Multiple Cancers; MC3 (Ellrott et al., 2018)]. Although both pipelines used a multiple-caller strategy, there are still some differences—in the processing of alignment files, versions of mutation callers, mutation filters, and gene annotations (Figure 5A)—between the MC3 (hg19) and GDC (hg38) mutations. To characterize those differences and evaluate concordance with prior results, we compared the public somatic mutation calling of SNVs on multiple TCGA cohorts, using 2,069 samples from the breast, leukemia, colorectal, and ovarian cancer cohorts (BRCA, LAML, COAD, and OV). We also investigated the 'protected' somatic mutation calls of the two groups, which represented the pre-filtered calls. A protected call was excluded from the public call set if it was considered low quality or potentially germline by the filters. GDC protected calls collected all the raw somatic mutation calls detected by all the callers, while MC3 counterpart excluded some low-confidence calls from the raw somatic calls.

The overlap between GDC hg38 and MC3 hg19 mutation calls was calculated by matching their genomic locations and tumor alleles. Across 1,902 shared tumor samples, the mutation overlap between GDC and MC3 contained a total of 488,138 public somatic SNV calls from 21,535 genes (Figure 5B). The two groups shared 386,350 SNVs (79%), leaving 71,967 GDC-unique calls and 29,821 MC3-unique calls. We thought the protected somatic calls of one group should represent the universe of all confident somatic mutations; however, there were 45,773 GDC-unique calls and 7,419 MC3-unique calls that the other group could not recover. Those unrecoverable unique calls (21%) implied that the raw mutation calling from the two groups have different characteristics, so those calls were only reported by the mutation callers in one group (Figures S4A and S4B).

We then analyzed the recoverable unique calls to investigate how different filtering strategies affected the generation of the public calls from the protected calls. The GDC reported 26,194 recoverable unique calls (the dark red region in Figure 5B). The

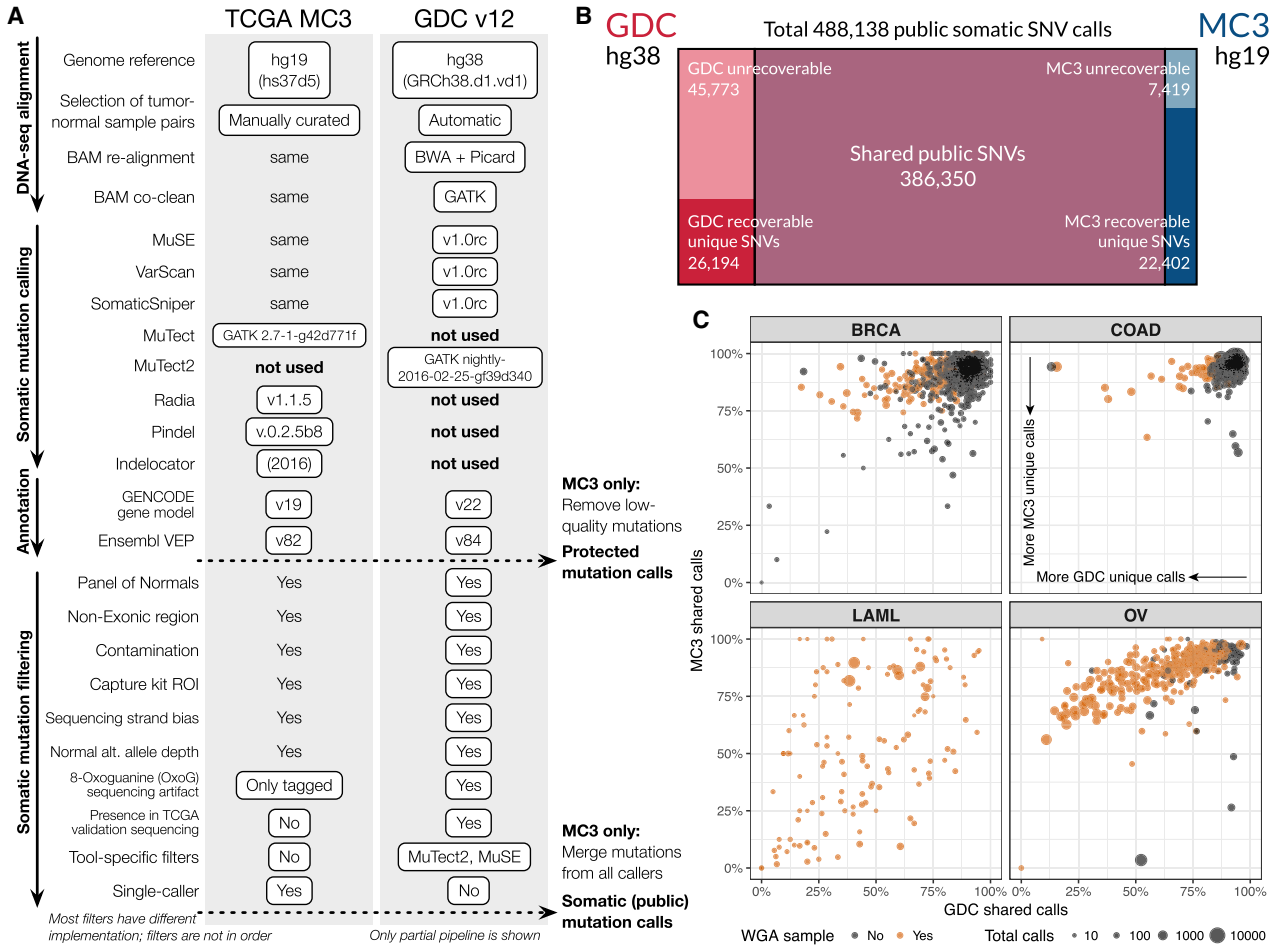


Figure 5. Somatic Mutation Processing and Data Comparison in TCGA Legacy and the GDC

(A) Outline of pipeline steps for TCGA MC3 (hg19) and current (v12) GDC release (hg38).

(B) Overlapping somatic mutation calls between GDC and MC3. Red and blue shaded regions represent the public somatic SNV calls unique in GDC and MC3, respectively. The lighter red and blue shaded regions represent the unrecoverable calls that were available in the public call of one group but were not found in neither public nor protected calls of the other group.

(C) The overlap of somatic mutation call per sample in four different cancer types. The X and Y axes represent the proportion of shared calls over the total calls from GDC and MC3, respectively. Each dot represents a sample, and the dot size indicates the numbers of somatic SNVs called. A sample has more GDC-unique or MC3-unique calls is closer to the origin. The color indicates whether WGA sequencing was employed. See also Figure S4.

stringent one-caller mutation removal in MC3 contributed to the majority of the recoverable GDC-unique calls (59.0%). Different definitions of non-exonic mutations were the second major source of the recoverable unique calls (36.2%). The gene annotation changes can also alter the exonic definition (at least 2.1% of all recoverable GDC-unique calls), such as genes *CCDC168* and *EFCAB8* (Figures S4C and S4D). The change can be systematically detected by comparing the transcript ID versions between two annotations. Different panel-of-normal (PoN) samples chosen by the two groups was another decisive factor for the identification of somatic mutations (4.4%). Usage of validation sequencing could also alter the somatic status of a variant call. The GDC labels a mutation call 'public' when it is also found in the validation sequencing, regardless of the filter status, whereas MC3 calling did not utilize validation status.

MC3 reported 22,402 recoverable unique (the dark blue region in Figure 5B). By following the GDC documentation for somatic MAF file generation, we were able to identify the specific filtering stage where each mutation call was 'protected'. A majority of the recoverable MC3-unique calls (96.8%) were protected in GDC since they were marked by multiple caller-specific filters in filtering stage 3 of the GDC pipeline, whereas those filters were not used to exclude mutations in MC3. We identified a few filters frequently associated to those recoverable MC3-unique calls, including 't_lod_fstar' filter for MuTect2 calls of low quality (41.8%), 'bSeq' filter for mutation calls with strand-biased read support (30.3%), 'oxog' filter for 8-oxoguanine (OxoG) artifacts (16.2%), all 'Tier*' filters for MUSE calls with poor evidence (45.2%), and 'clustered_events' filter for MuTect2 calls located at a reassembled haplotype with too many mutations (13.5%).

To understand whether the overlap between GDC hg38 and MC3 hg19 somatic mutation calls varies across different samples and cancer types, we calculated the proportion of shared calls over all GDC and MC3 somatic calls for every sample (Figure 5C). Overall,

we observed that different cancer types exhibited very different levels of concordance. COAD samples had the largest fraction of concordant calls, where 243 out of the total 376 samples (88%) had a higher shared call percentage than the median in both GDC and MC3. In contrast, LAML samples exhibited the worst concordance between the two groups, which may be due to this cohort having incomplete demultiplexing, and using Whole Genome Amplification (WGA) in library construction (Bodini et al., 2015).

From our analysis, we found that 79% of all the public somatic mutation calls from two groups were concordant. When we excluded LAML samples, 80% of all the public somatic calls were concordant, and the average of the fraction of concordant calls per sample improved from 72.5% to 75.9%. We were able to explain the remaining non-concordant public mutation calls by the three major sources of the differences (Data S5.3): for unrecoverable unique calls, mutation caller; for recoverable unique calls, filtering strategy, and gene annotation version.

DISCUSSION

The publication of the first human reference genome unleashed a torrent of cancer research, in which the field witnessed a steady transition from a largely qualitative and wet lab-based practice to one that is far more quantified and digital. While TCGA has played important catalytic and leadership roles in this transformation, the resulting increase in volume and complexity of data are pushing the limits of our capacity to store, process, and make sense of it. At the same time, characterization and software technologies, analytic methods, and reference data continue to rapidly evolve. The cumulative effect of these forces—size, complexity, and accelerated change—mean researchers have to confront substantial technical and logistical challenges before they can begin to ask basic biological questions.

This study helps address those challenges, and is important to the research community in several ways: (1) scientifically, because confidence in and usability of global resources like TCGA must remain high even as the data they offer and fields they serve experience enormous growth and change. Findings published from such resources, which purport to describe fundamental biology, should remain evident no matter how the primary data are transformed after original collection; and any findings refuted after such transformations should be revised or discarded. By demonstrating significant concordance between the legacy hg19 and GDC-harmonized hg38 versions of TCGA data, our study girds the corpus of TCGA-related research and suggests it will continue to play a valuable role into the foreseeable future. (2) Technologically, because the GDC will play an important role in the usability of many large data sets beyond TCGA—e.g., TARGET and others which are currently generating data—and in offering these data to the world the GDC has updated or introduced new data models, ontologies, back-end infrastructure as well as front-end portals and APIs. When combined with the fact that many of the scientific algorithms and knowledge bases used to generate legacy TCGA data have either evolved or been superseded, this means that the number of moving parts—and therefore the sources of potential variation—in the harmonized hg38 data served by the GDC, is high. The work reported here isolates confounding factors in the technology stack for each platform, offers sets of outliers for each platform, and indicates that TCGA results should be relatively insensitive to changes between legacy hg19 and harmonized hg38 data processing. (3) Efficiency and cost-effectiveness, because the scope of the vetting effort reported here would be impractical for many labs or academic departments to conduct themselves prior to confidently utilizing GDC data in their work. Our study provides a framework that may guide similar comparative analyses in the future, online resources for follow-up exploration, and tools that can be hardened, generalized, and deployed to form the basis of future QC efforts in other large-scale projects.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Analyses of miRNA Expression
  - Analyses of Somatic Copy Number Alterations
  - Analyses of DNA Methylation
  - Analyses of mRNA Expression
  - Analyses of Somatic Mutations
- DATA AND CODE AVAILABILITY
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.06.006.

ACKNOWLEDGMENTS

We thank the U.S. National Cancer Institute for funding through grants 1U24CA210999-01, 1U24CA210974-01, 1U24CA211006-01, 1U24CA210949-01, 1U24CA210978-01, 1U24CA210952-01, 1U24CA210989-01, 1U24CA210957-01, 1U24CA210990-01, 1U24CA211000-01, 1U24CA210950-01, 1U24CA210969-01, and 1U24CA210988-01. We are grateful for advice and dialogue from numerous colleagues at our respective institutions, TCGA and GDAN collaborators, the technical support team from GDC; and especially the NCI Office of Cancer Genomics at NCI, for steadfast organizational support.

AUTHOR CONTRIBUTIONS

H.L. and M.S.N. working group leaders; D.I.H. manuscript coordinator. Analyses of miRNA Expression: S.R., G.R., and T.K. section leaders; D.B., A.J.M., R.A., S.S., and M.S.N. contributors. Analyses of Somatic Copy Number Alterations: A.C. and G.G. section leaders; K.H. and M.S.N contributors. Analyses of DNA Methylation: B.P.B, W.Z., and T.C.S. section leaders; H.S., P.W.L., T.K., A.C., and M.S.N contributors. Analyses of mRNA Expression: J.P. and S.B. section leaders; K.H., S.C., D.H., and H.L. contributors. Analyses of Somatic Mutations: L.B.W. section leader; L.D., B.B., R.J., M.B., M.W., and H.L. contributors. All the authors contributed to the project administration and helpful discussion.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 19, 2019
Revised: March 18, 2019
Accepted: June 13, 2019
Published: July 24, 2019

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288–295.

Bodini, M., Ronchini, C., Giacò, L., Russo, A., Melloni, G.E., Luzi, L., Sardella, D., Volorio, S., Hasan, S.K., Ottone, T., et al. (2015). The hidden genomic landscape of acute myeloid leukemia: subclonal structure revealed by undetected mutations. Blood *125*, 600–605.

Bowen, N.J., Logani, S., Dickerson, E.B., Kapa, L.B., Akhtar, M., Benigno, B.B., and McDonald, J.F. (2007). Emerging roles for PAX8 in ovarian cancer and endosalpingeal development. Gynecol. Oncol *104*, 331–337.

Brinkmann, U., Vasmatzis, G., Lee, B., and Pastan, I. (1999). Novel genes in the PAGE and GAGE family of tumor antigens found by homology walking in the dbEST database. Cancer Res. *59*, 1445–1448.

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Chiu, H.S., Somvanshi, S., Patel, E., Chen, T.W., Singh, V.P., Zorman, B., Patil, S.L., Pan, Y., Chatterjee, S.S., et al.; Cancer genome Atlas research Network (2018). Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. Cell *23*, 297–312.

Chu, A., Robertson, G., Brooks, D., Mungall, A.J., Birol, I., Coope, R., Ma, Y., Jones, S., and Marra, M.A. (2016). Large-scale profiling of microRNAs for the Cancer Genome Atlas. Nucleic Acids Res. *44*, e3.

Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.C., Agarwala, R., McLaren, W.M., Ritchie, G.R., et al. (2011). Modernizing reference genome assemblies. PLoS Biol. *9*, e1001091.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T., Malta, T.M., Pagnotta, S.M., Castiglioni, I., Ceccarelli, M., Bontempi, G., and Noushmehr, H. (2015). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv1507.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR. Curr. Protoc. Bioinformatics *51*, 14.1–19.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. *281*, 271–281.

Fan, Y., Xi, L., Hughes, D.S., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A., and Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. *17*, 178.

Fortin, J.P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M., and Hansen, K.D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. *15*, 503.

Genovese, G., Handsaker, R.E., Li, H., Kenny, E.E., and McCarroll, S.A. (2013). Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. Am. J. Hum. Genet. *93*, 411–421.

Ghannam-Shahbari, D., Jacob, E., Kakun, R.R., Wasserman, T., Korsensky, L., Sternfeld, O., Kagan, J., Bublik, D.R., Aviel-Ronen, S., Levanon, K., et al. (2018). PAX8 activates a p53-p21-dependent pro-proliferative effect in high grade serous ovarian carcinoma. Oncogene *37*, 2213–2224.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets, and gene nomenclature. Nucleic Acids Res. *34*, D140–D144.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res. *22*, 1760–1774.

Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat. Med. *21*, 1253–1261.

Hutter, C., and Zenklusen, J.C. (2018). The cancer genome Atlas: creating lasting value beyond its data. Cell *173*, 283–285.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

Köster, J., and Rahmann, S. (2012). Snakemake - A scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*, 311–317.

Laury, A.R., Perets, R., Piao, H., Krane, J.F., Barletta, J.A., French, C., Chirieac, L.R., Lis, R., Loda, M., Hornick, J.L., et al. (2011). A comprehensive analysis of PAX8 expression in human epithelial tumors. Am. J. Surg. Pathol. *35*, 816–826.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Liu, J., and Siegmund, K.D. (2016). An evaluation of processing methods for HumanMethylation450 BeadChip data. BMC Genomics *17*, 469.

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). SWAN: subset-quantile within array normalization for Illumina infinium HumanMethylation450 BeadChips. Genome Biol. *13*, R44.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Nonaka, D., Chiriboga, L., and Soslow, R.A. (2008). Expression of pax8 as a useful marker in distinguishing ovarian carcinomas from mammary carcinomas. Am. J. Surg. Pathol. *32*, 1566–1571.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557–572.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35*, D61–D65.

Radenbaugh, A.J., Ma, S., Ewing, A., Stuart, J.M., Collisson, E.A., Zhu, J., and Haussler, D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS One *9*, e111516.

Reynolds, S.M., Miller, M., Lee, P., Leinonen, K., Paquette, S.M., Rodebaugh, Z., Hahn, A., Gibbs, D.L., Slagel, J., Longabaugh, W.J., et al. (2017). The ISB cancer genomics cloud: a flexible cloud-based platform for cancer genomics research. Cancer Res. *77*, e7–e10.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). DbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Shibuya, K., Obayashi, I., Asakawa, S., Minoshima, S., Kudoh, J., and Shimizu, N. (2004). A cluster of 21 keratin-associated protein genes within introns of another gene on human chromosome 21q22.3. Genomics *83*, 679–693.

Silva, T.C., Coetzee, S.G., Gull, N., Yao, L., Hazelett, D.J., Noushmehr, H., Lin, D., and Berman, B.P. (2018). ELMER v.2: An R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty902.

Smit, A.F. (1996). The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. *6*, 743–748.

Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics *29*, 189–196.

Triche, T.J., Jr., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., and Siegmund, K.D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. *41*, e90.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Wang, Z., Yang, B., Zhang, M., Guo, W., Wu, Z., Wang, Y., Jia, L., Li, S., Cancer genome Atlas research Network, Xie, W., and Yang, D. (2018). lncRNA epigenetic landscape analysis identifies EPIC1 as an oncogenic lncRNA that interacts with MYC and promotes cell-cycle progression in cancer. Cancer Cell *33*, 706–720.

Wickham, H. (2014). Tidy data. J. Stat. Soft. *59*, 23.

Ye, K., Wang, J., Jayasinghe, R., Lameijer, E.W., McMichael, J.F., Ning, J., McLellan, M.D., Xie, M., Cao, S., Yellapantula, V., et al. (2016). Systematic discovery of complex insertions and deletions in human cancers. Nat. Med. *22*, 97–104.

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2013). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics (Oxford, England), btt730.

Zhou, W., Dinh, H.Q., Ramjan, Z., Weisenberger, D.J., Nicolet, C.M., Shen, H., Laird, P.W., and Berman, B.P. (2018a). DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat. Genet. *50*, 591–602.

Zhou, W., Laird, P.W., and Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. *45*, e22.

Zhou, W., Triche, T.J., Jr., Laird, P.W., and Shen, H. (2018b). SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. *46*, e123.

Zozulya, S., Echeverri, F., and Nguyen, T. (2001). The human olfactory receptor repertoire. Genome Biol. *2*, RESEARCH0018.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
TCGA hg38 miRNA-seq data	NCI Genomic Data Commons (GDC)	https://portal.gdc.cancer.gov miRNA and Isoform Expression Quantification
TCGA hg19 miRNA-seq data	NCI GDC Legacy Archive	https://portal.gdc.cancer.gov/legacy-archive miRNA gene and isoform quantification
ISB-CGC TCGA data in BigQuery	Reynolds et al. (2017)	www.isb-cgc.org
TCGA hg38 somatic copy number alteration data	NCI GDC	https://portal.gdc.cancer.gov somatic copy number alteration
TCGA hg19 somatic copy number alteration data	NCI GDC Legacy Archive	https://portal.gdc.cancer.gov/legacy-archive somatic copy number alteration
TCGA hg38 DNA methylation data (v. 12.0)	NCI GDC	https://portal.gdc.cancer.gov/ - Data release 12.0
TCGA hg19 DNA methylation legacy gene mapping	Illumina	https://support.illumina.com/downloads/infinium_human methylation450_product_files.html
TCGA hg38 mRNA-seq data	NCI GDC	https://portal.gdc.cancer.gov mRNA gene and isoform quantification
TCGA hg19 mRNA-seq data	NCI GDC Legacy Archive	https://portal.gdc.cancer.gov/legacy-archive mRNA gene and isoform quantification
TCGA hg38 somatic mutation data	NCI GDC	https://portal.gdc.cancer.gov/ - Data release 12.0
TCGA hg19 somatic mutation data	Ellrott et al. (2018)	https://gdc.cancer.gov/about-data/publications/mc3-2017
SuppData-1.1	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-1.2	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-2.1	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-2.2	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-2.3	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-2.4	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-2.5	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-2.6	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-3.1	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-3.2	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-3.3	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-3.4	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-3.5	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-4.1	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-4.2	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-5.1	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-5.2	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
SuppData-5.3	this study	https://gdc.cancer.gov/about-data/publications/HG38QC
Software and Algorithms		
GISTIC2.0	Mermel et al. (2011)	https://software.broadinstitute.org/software/cprg/?q=node/31
R 3.5.1 (DNA methylation, mRNA expression, and somatic mutation)	R Development Core Team	https://www.R-project.org
TCGAbiolinks v2.4.3 (DNA methylation)	Colaprico et al. (2016)	http://bioconductor.org/packages/TCGAbiolinks/
ELMER v.2.6.1 (DNA methylation)	Silva et al. (2018)	http://bioconductor.org/packages/ELMER/
Infinium probe-gene links R markdown (DNA methylation)	this paper	Supp. Data file methylation_R_code.tar.gz or https://github.com/zwdzwd/GDC_DNA_methylation_QC
Methylation-expression associations R markdown (DNA methylation)	this paper	Supp. Data file methylation_R_code.tar.gz or https://github.com/tiagochst/GDC_DNA_methylation_QC

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
miRNA-seq QC analysis code (Jupyter notebooks and R script)	this paper	https://github.com/GDAC-miRNA/TCGA-hg19-hg38-QC
Snakemake v5.4.0 (somatic mutation)	Köster et al. (2012)	https://snakemake.readthedocs.io/
CrossMap v0.2.8 (somatic mutation)	Zhao et al. (2013)	http://crossmap.sourceforge.net
SQLite v3.22.0 (somatic mutation)	SQLite Development Team	https://sqlite.org/
Somatic mutation QC analysis code (R notebooks and Snakemake workflow)	this paper	https://github.com/ding-lab/gdc_qc_analysis

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new unique reagents. Further information and requests for resources and materials should be directed to and will be fulfilled by the Lead Contact, Han Liang (HLiang1@mdanderson.org).

METHOD DETAILS

Analyses of miRNA Expression

During the long-running TCGA project (2008-2018), the production and analysis of miRNA data transitioned from array-based assays (used for only two tumor types: GBM and OV) to sequencing-based assays. In this study we focused only on the sequencing data produced by the BCGSC. TCGA miRNA-Seq data is available for 10,250 cases, and 11,022 samples. The data were primarily derived from solid tumor samples (n=9729), but some data were derived from normal or adjacent tissue samples (n=675), metastatic samples (n=379) or other sample types (n=239). Two primary data sources will be referenced: first and foremost, the GDC repositories, but in addition we also make use of certain Google BigQuery tables made available by the ISB-CGC (ISB Cancer Genomics Cloud (Reynolds et al., 2017) an NCI Cloud Resource), which allows the research community to use SQL to explore and compare the relevant data and metadata.

As of this writing (and based on GDC data release 14.0, December 18, 2018), the data available from the GDC Data Portal (https://portal.gdc.cancer.gov) are divided into two main archives: an "active" archive, and a "legacy" archive. The legacy archive contains data which the GDC inherited from two previous data repositories: the TCGA DCC and CGHub. This legacy data is primarily based on human genome reference hg19/GRCh37, with miRNA annotations initially from miRBase v16, prior to a transition to miRBase v20 in early 2016. The active archive contains data which has been re-processed at the GDC, using updated references, including hg38/GRCh38 and miRBase v21. Note that there are two separate "entry points" for these two data sets: the main GDC Data Portal (https://portal.gdc.cancer.gov) for the hg38 data, and the Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive) portal for the older data.

Legacy Archive

The legacy archive repository can be queried for TCGA miRNA-Seq data interactively by selecting TCGA in the Cancer Program section and miRNA-Seq as the Experimental Strategy. The result of this query is a list of 65,427 files:

- data Category counts: 19,081 raw sequencing data + 46,346 gene expression
- data Type counts: 19,081 aligned reads + 23,173 each miRNA isoform quantification and miRNA gene quantification
- Data Format counts: 19,081 BAM + 46,346 TXT
- Platform counts: 10,157 Illumina GA + 55,270 Illumina HiSeq
- Access Level counts: 19,081 controlled-access + 46,346 open-access

Considering BAM files alone (i.e. by selecting only BAM files in the Data Format section), these data are derived from 10,250 cases (5357 female, 4853 male, and 40 cases with no clinical information available), and 11,022 samples. Most samples have more than one BAM file in the legacy archive for a variety of reasons, including multiple aliquots per sample, or updates to reference sources and/or to pipeline parameters: 4262 samples have only one BAM file, 5510 samples have two, 1203 samples have three, and 45 samples have four, and 2 samples have five. When new BAM files were produced, older versions remained available because they may have been referenced in publications. In addition, updated quantification files were released for many samples after the transition to miRBase v20.

Active Archive

The active archive repository can also be queried for TCGA miRNA-Seq data interactively, by selecting the TCGA Program in the Cases filter tab, and selecting miRNA-Seq as the Experimental Strategy in the Files filter tab. As of GDC data release 14.0, this yields a total of 33,246 files, with 3 files typically given per sample (two TSV files for each BAM file):

- Data Category counts: 11,082 raw sequencing data + 22,164 transcriptome profiling
- Data Type counts: 11,082 each; aligned reads, isoform expression quantification, and miRNA expression quantification

- Workflow Type counts: 11,082 BWA-aln + 22,164 BCGSC miRNA profiling
- Data Format counts: 11,082 BAM + 22,164 TSV
- Platform counts: all data are derived from the Illumina platform (note the loss of information about the specific platform used)
- Access Level counts: 11,082 controlled-access + 22,164 open-access

These data are derived from 10,250 cases (5357 female, 4853 male, and 40 cases with no clinical information available), and 11,020 samples (there are 15 cases with three samples each, 740 cases with two samples each, with the remaining 9495 cases having one sample each; and there are 62 samples with two aliquots each).

hg19 to hg38 File Mapping

Each hg38 miRNA-Seq BAM file was created by running the GDC Alignment Workflow (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/miRNA_Pipeline/#alignment-workflow) which, in the case of miRNA data, runs BWA-aln2 on one of the hg19 miRNA-Seq BAM files. Although the detailed information specifying which input file was used to create each hg38 file is not currently available from the GDC APIs, the mapping between hg19 and hg38 BAM files is available in a BigQuery table (tcga-qc:miRNA.legacy2active_BAMid_map).

Differences in Case, Sample, and File Counts

Based on the descriptions and counts above, there are some minor discrepancies between the hg19 and hg38 datasets with regards to sample, aliquot, and file counts (Table S1, Data S1.1 and S1.2). The GDC has assigned UUIDs to each distinct entity referenced or contained in its archives: files, cases, samples, *etc*. Each data file is associated with a single aliquot, which is uniquely identified by a UUID and by a "TCGA barcode" of length 24.

The number of cases with miRNA-Seq data are identical between the two archives. There are two samples which were discarded during the harmonization process (TCGA-AB-2888-03A and TCGA-AB-2990-03A) because for each sample a second 'B' vial existed and was retained (TCGA-AB-2888-03B and TCGA-AB-2990-03B). The difference in the number of aliquots is larger because one objective of the harmonization process was to retain a single aliquot (and associated data) for each sample. Finally, the difference in miRNA BAM file counts is large (7999) because, as described above, most samples have more than one BAM file in the legacy archive. It is particularly important that miRNA BAMs derived from the older BCGSC adapter-trimming algorithm not be used. The results of examining 19846 miRNA-Seq BAMs in the legacy archive to verify trim-length are available in the BigQuery table named trim_length_test (tcga-qc:miRNA.reads_trim_length_test). These tests indicated that 13115 out of 19846 BAMs had the correct trim-length (and are marked as "KEEP" in the table), derived from the newer adapter-trimming algorithm. We advise users to take careful note of this when using hg19 miRNA-Seq BAM files from the legacy archive.

miRNA-Seq Pipeline

The miRNA-Seq pipeline implemented by the GDC (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/miRNA_Pipeline/) is largely identical to the pipeline used to generate data over the course of TCGA[3]. The main differences between the hg19 and hg38 versions of the pipeline involve the versions of some software components, and the reference data files used (Figure 1A). An additional change involves reads that had been flagged as 'QCFail' in sequencing. These had been retained in the original hg19 TCGA BAMs, but were removed in the GDC hg38 BAMs (as of GDC data release 11.0). This new filtering influenced miRNA expression measures as follows. In TCGA, legacy and harmonized GDC miRNA-seq data, only reads with exact-match alignments contributed to expression quantification, i.e. to read counts and normalized RPMs for miRNA stem-loops and isoforms. While QCFail filtering reduced miRNA read counts, relatively few QCFail reads will have exact-match miRNA alignments, and RPM is a normalized abundance metric. Given these factors, we expected and found that RPM values were relatively insensitive to removing QCFail reads.

We note that while the overall TCGA data set was influenced by batch effects related to platforms and/or protocols (Figure 1A), and batch-corrected TCGA miRNA-Seq data have recently been published, comparing batch-corrected to uncorrected miRNA-seq data is outside of the scope of the work reported here. Platforms and protocols used to process each aliquot are also available in BigQuery (tcga-qc:miRNA.protocol_platform_info).

Data Online in Cloud Resource

The ISB-CGC platform (Reynolds et al., 2017) provides BigQuery tables containing all available open-access TCGA molecular data, as well as clinical, biospecimen, and other metadata for all of the TCGA cases, samples, and data files, and a variety of other genomic reference sources. These tables, which can be queried using Standard SQL, provide an accessible environment for comparing hg19 and hg38 miRNA quantification data. There are four tables of miRNA expression counts: one table for each of the two outputs of the miRNA-Seq pipeline, and one pair for each genome-reference:

- isb-cgc:TCGA_hg19_data_v0.miRNAseq_Isoform_Expression
- isb-cgc:TCGA_hg19_data_v0.miRNAseq_Expression
- isb-cgc:TCGA_hg38_data_v0.miRNAseq_Isoform_Expression
- isb-cgc:TCGA_hg38_data_v0.miRNAseq_Expression

Each table is in "tidy" format (Wickham, 2014) and contains several columns of metadata including the case, sample, and aliquot barcodes, and the GDC file UUID. For the two isoform (isomiR) tables, the key data columns ("fields") are: aliquot_barcode, mirna_id,

mirna_accession, mirna_transcript, chromosome, start_pos, end_pos, strand, read_count, reads_per_million_miRNA_mapped, and cross_mapped. These and related data from this study can be further explored through the interactive BigQuery webUI (https://cloud.google.com/bigquery/bigquery-web-ui).

Comparing hg19 and hg38 miRNA Expression Levels

Our analysis shows that the hg19 and hg38 versions of TCGA miRNA expression quantifications are highly concordant: the $\log_2$(RPM) values for 1137 (83%) out of the 1367 miRNA mature strands detected in at least 1000 samples have a correlation coefficient greater than 0.98. We also found that the $\log_2$(RPM) quantification values were highly comparable, despite the removal of QCFail reads: the median absolute difference between hg19 and hg38 $\log_2$(RPM) expression values was less than 0.05 for 1290 (94%) out of the same 1367 miRNA mature strands. For miR families that have identical mature sequences in multiple genomic loci, we noted that expression levels from hg19 BAM files were systematically higher for one member of the family, while expression from hg38 BAM files were more evenly distributed between members, resulting in lower than expected correlation between the stem-loop and isoform quantification values (Figure 1). Further investigation identified the cause of these differences to be associated with the BWA versions. Specifically, the alignment algorithm in BWA v0.5.7 was used to produce the hg19 BAM files, whereas BWA v0.7 was used by the BCGSC and GDC for newer hg38 alignments. When aligning reads which map identically to multiple loci in the genome, as in the case of miR families, the former arbitrarily maps more reads to one location while the latter distributes reads evenly between the locations, resulting in differences in normalized read counts used for isoform quantification.

Discordant miRs

Despite generally high correlations between the hg19 and hg38 miRNA expression quantifications across the entire TCGA dataset, there are cases where the correlation for a particular miR, within a specific tumor type, may be significantly higher or lower than when considering data across all tumor types. To help readers more easily identify these outlier miRs, we computed Pearson and Spearman correlations for each high-confidence, mature-strand miRNA, grouped by TCGA project (i.e. tumor-type), and made these available in a BigQuery table (tcga-qc:miRNA.mirna_corrs_by_project). In general, such "discordant miRs" have very low expression levels in one or more tumor types.

Analyses of Somatic Copy Number Alterations

As described in the main text, copy number alterations were identified from TCGA Level 3 segmented copy number profiles, generated from the same pipeline for both hg19 and hg38. In TCGA Level 1 and Level 2 genotype and copy number data are at the probe level, and thus agnostic to genome build. Therefore Levels 1 and 2 data at the GDC are identical for hg19 and hg38; only Level 3 data and above show differences. Individual Affymetrix SNP6 probes were mapped to genomic locations in either hg19 or hg38, and copy number values from the final Level 2 data were used to generate copy number segments. The probesets used for hg19 and hg38 were not identical, however, in that 14811 (0.8%) probes that could not be uniquely mapped to hg38 were not used in segmentation. A total of 10,705 tumor samples across 33 TCGA disease types were used here, with colon and rectal adenocarcinomas combined into a single COADREAD cohort (file manifests for these data are given as Data S2.1 and S2.2). Two evaluations of concordance were performed: between gene-level copy number calls as well as gene membership within significant focal alterations as called by GISTIC2.0. For hg19 samples significance analysis of focal amplifications and deletions was taken from the pre-computed GISTIC results in firebrowse.org, while custom analysis was performed in FireCloud for hg38 samples, by running GISTIC2.0.23 (Mermel et al., 2011) with the parameters:

Broad Length Cutoff = 0.7
Cap Values = 1.5
Confidence Level = 0.99
Amplification Threshold = 0.1
Deletion Threshold = 0.1
Do Gene GISTIC = 1

Gene Collapse Method = ‘‘extreme’’

Join Segment Size = 4
Maximum Sample Segments = 2000

Q-value threshold: 0.25

Remove X-Chromosome = 0

Gene Level Copy Number Calls

The average difference in relative copy number between the hg19 and hg38 samples in each TCGA cohort is given in Data S2.3, with an excerpt from one cancer cohort plotted in Figure S1A. The list of ‘‘deviant’’ genes is given in Data S2.4 and summarized in Figure 2B. In addition to raw relative copy number values, GISTIC2.0 also assigns to each gene of every sample a thresholded copy number level that reflects the magnitude of its deletion or amplification. These are integer values ranging from -2 to 2, where 0 means no amplification or deletion of magnitude greater than the threshold parameters described above. Amplifications are represented

by positive numbers: 1 means amplification above the amplification threshold; 2 means amplification larger than the arm level amplifications observed in the sample. Deletions are represented by negative numbers: -1 means deletion beyond the threshold; -2 means deletions greater than the minimum arm-level deletion observed in the sample. To compare these thresholded calls for each of the 20,616 genes, we simply counted the number of samples with a disagreement in thresholded calls between the hg19-aligned and hg38-aligned GISTIC2.0 runs for each gene; i.e. all samples off the diagonal of the 5×5 confusion matrix that compares thresholded calls for each gene between the hg19-aligned and the hg38-aligned GISTIC2.0 runs (examples given in Figures S1B and S1C). 624655 of 680328 (tumor type, gene) pairs retained the same thresholded copy number call in over 90% of all samples in both hg19-aligned and hg38-aligned GISTIC2.0 runs. However, the same set of 11 recurrently deviant genes mentioned previously again showed a high proportion of samples in each TCGA disease type with differing thresholded copy number calls: each with >35% disagreement when averaged over all cancer types.

Significantly Altered Focal Peaks

To assess concordance of focal amplification and deletion peaks identified as significant, we tallied the genes called within focal peaks in either the hg19-aligned or hg38-aligned analyses and examined the concordance between these two gene lists (Data S2.5). Out of 131,894 unique gene-tumor type pairs called in significant peaks in hg19 and 132,390 in hg38, only 84,540 were found in the intersection between the two runs (Figures S1D and S1E). As an example, among focal amplifications in READ, only 375 genes were called in both runs' peaks compared to 240 and 1396 genes found only in the hg19-aligned and hg38-aligned GISTIC2.0 amplification peaks, respectively. Similarly, among focal deletions in COAD, only 1519 genes were called in both runs' peaks compared to 938 and 809 genes found only in the hg19-aligned and hg38-aligned GISTIC2.0 deletion peaks, respectively. Overall, the majority of genes present in focal peaks, especially those near the peak boundaries, are likely passengers as opposed to drivers, and so this abundance of passengers may explain much of the relatively low conservation of genes in these peaks.

To test this hypothesis, we reanalyzed cervical endometrial squamous carcinomas (CESC) as a representative tumor type using GISTIC2.0, this time lowering the confidence level parameter from 99% to 25%, thereby narrowing the called peaks to their regions of highest likelihood density. The genes encompassed by these narrow (25% confidence) peaks show better overlap with those in their corresponding wide (99% confidence) peaks. Of 119 genes in narrow amplification peaks aligned to hg19, only 3 were not found in the hg38 wide amplification peaks, and all of the 35 genes in narrow amplification hg38 peaks were found in the hg19 wide amplification peaks.

Finally, we performed a more biologically grounded assessment of conservation of significant copy number driver events between GISTIC2.0 runs. Using the body of published TCGA tissue-specific marker papers, we compiled a comprehensive list of 521 key copy-number-driven oncogenes and tumor suppressors in each of 26 tumor types. Five out of the 33 total TCGA tumor types were excluded as they lack many common, significant copy number driver events (DLBC, KICH, MESO, THYM, and UVM), while COAD and READ were analyzed as a combined COADREAD cohort. 482 out of 521 putative disease driver events mentioned in the marker papers were either explicitly found in both hg19 and hg38 runs or were absent from both runs (Data S2.6). Forty drivers were absent from both hg19 and hg38 GISTIC2.0 runs and likely stem from multiple causes: the marker papers often used smaller sets of TCGA samples to discover these drivers, utilized manual inspection and rescue of drivers proximal to identified copy number peaks but not explicitly called within these peaks, and used earlier versions of GISTIC running on hg18-aligned copy number data. Only 20 drivers were explicitly found in the hg19-aligned GISTIC2.0 run but were absent from the hg38-aligned GISTIC2.0 run. Conversely, only 19 drivers were explicitly found in the hg38-aligned GISTIC2.0 run but not in the hg19-aligned GISTIC2.0 run.

Analyses of DNA Methylation

DNA methylation data in TCGA is based on Illumina's Infinium DNA methylation BeadChips (Cancer Genome Atlas Research Network, 2011). In total, data from 11,172 primary tumor samples from 33 cancer types were generated as part of TCGA, initially using the HM27 platform but then switching to HM450 during the project. 194 acute myeloid leukemia (LAML) samples were assayed on both platforms. In addition, 1,109 matched adjacent normal samples were profiled, of which five were later determined to reflect occult tumors. The data were generated in 281 batches. Each batch contains a quality control sample from a lymphoblastoid cell line (Coriell's GMO6990), expanded either at the Nationwide Children's Hospital (NCH, those having the sample code TCGA-07-0227) or at the International Genomics Consortium (IGC, those having the sample code TCGA-AV-A03D). In brief, the HM27 platform contains ca. 27 thousand individual CpG features, all within 1,500 bp of an annotated gene promoter; while the HM450 platform contains ca. 480 thousand features, adding many from a number of other selected genomic contexts. Over 90% of the CpG features present on the HM27 array are also present in the HM450 array (Zhou et al., 2017).

DNA Methylation Pipeline for Array-Based Data

The files generated by the array scanner contain raw fluorescence intensities, one for each of the two channels (red/green). These "Raw intensity" (.idat) files are labeled Level 1 data in the TCGA data type hierarchy, and are identical between GDC data release versions (although individual samples may be added or removed). IDAT files are processed to control for dye bias and signal background, to generate Level 2 "Normalized intensity" files, in a tab-delimited (.txt) format. Finally, red and green intensities are compared to compute a single *beta* value for each probe, defined as the percentage of methylated DNA molecules. These results are stored as Level 3 "Methylation beta value" files, which are tab-delimited (.txt) and contain one beta value per feature. Some CpG features present on the array are omitted ("masked") from the Level 3 data files, because they are considered unreliable due to likely cross-hybridization and/or the presence of polymorphic sequences. Additional details about this "experiment-independent" probe masking can be found below in our study (Zhou et al., 2017). In addition, "experiment-specific" probe masking is

performed, based on weak signal or high background due to array quality, experimental failure, or genomic deletions in the sample. These probes are present in Level 3 data, but given a value of "N/A". More details about experiment-dependent masking can also be found Zhou et al., 2018b.

All processing steps are summarized in Figure 3A, along with the specific methods used for the GDC hg19 Legacy and hg38 versions. UCSC RefSeq (Gene, 2010) (Pruitt et al., 2007) and dbSNP version 13 (Sherry et al., 2001) were used to create the original gene associations and SNP overlaps in Illumina's manifest file, which was used in the hg19 Legacy version. We used Repeatmasker (Smit, 1996) to perform repeat-based masking, and the *Methylumi* pipeline to perform signal processing steps, using the standard dye bias correction and "Noob" background correction described in the *Methylumi* publication (Triche et al., 2013). Note that while additional normalization methods have been proposed and used by other groups (including normalization that explicitly quantile normalize beta value distribution of Type-I and Type-II probes (Maksimovic et al., 2012; Teschendorff et al., 2013) or ones that made use of internal control measurements (Fortin et al., 2014), these methods may introduce artifacts due to internal assumption on the data distribution (Liu and Siegmund, 2016). Hence we employed only minimal reliable background subtraction and did not use any additional normalization.

Overview of DNA Methylation Data at the GDC

For each TCGA sample, there are two Level 1 idat files ("Raw intensities", one for each color channel) and two tab-separated files: the Level 2 file containing normalized probe intensities and other low-level information ("Normalized intensities"), and the Level 3 file containing final methylation "beta" values, genomic coordinates, and gene context ("Methylation beta value"). The hg19-derived (Legacy) data is reflected in GDC versions 1.0 through 3.0, while hg38-derived data are reflected in versions 4.0 onward (Table S2). The number of samples processed for each platform is shown in Table S3, and reflects that: (i) a small number of samples were run on both the HM27 and HM450 platforms (as noted above); and (ii) multiple tissues may be collected per patient/case (such as primary and/or metastatic tumor, and normal), thus yielding multiple aliquots for the same patient.

GDC data release versions 4.0 through 12.0 were generated with a newer pipeline, which not only introduced hg38 but also improved probe mapping and annotation (detailed below). These later releases contain only Methylation beta values (Level 3), to reflect the fact that Levels 1 and 2 data (respectively the raw IDAT and Methylation Intensity files) were not modified by the pipeline and thus remained identical. Users wishing to access the Levels 1 or 2 methylation data should thus retrieve them from the Legacy archive at the GDC. Note that 59 tumor samples with Level 1 & 2 data in the legacy archive do not have harmonized hg38 data in the GDC portal (Table S3): 3 are from the HM27 platform, 2 are from HM450, with the remaining 54 being dual-platform LAML cases. Three adjacent normal samples were also excluded (two HM27 and one HM450). Sample manifests for both hg19 and hg38 versions are available as Data S3.1–S3.4.

Differences between Level 3 hg19 and hg38 DNA Methylation Data at the GDC

As of GDC version 4.0, probe sequences were re-mapped to hg38 (GRCh38) coordinates, resulting in a small number of additional probes that were quality-masked (i.e. flagged) due to poor mapping to the new genome assembly: specifically, in GDC data release 14.0 we flagged 5,120 (HM450) and 544 (HM27) probes by removing genomic coordinates from the Level 3 data file. In addition, gene annotations were substantially updated, resulting in a number of new probe-gene associations. The updated processing steps are outlined in Figure 3 and described below.

Differences in Probe Remapping and Masking

While the GDC hg38 pipeline for methylation is called "Liftover", all probe sequences were re-mapped to hg38 as described in (Zhou et al., 2017). In addition to the 88,058 probes masked with N/A values in the hg19 Legacy version, another 5,120 HM450 and 544 HM27 probes were invalidated because they either had a mapping quality of <10, or were Type-I probes for which the methylated and unmethylated probes mapped to different locations in the genome (Zhou et al., 2017). These probes and beta values were retained in the hg38 level 3 data files, but genomic coordinates were removed (changed to NA or chr=*, start=-1, end=-1) to indicate poor mapping. The vast majority of probes remained validly mapped without mismatches in hg38 (98.0% of HM27 and 98.9% of HM450, Table S4). 86% of probes were mapped with the highest quality (=60) in both genome builds (Table S4, left). 23 HM450 probes have been moved from the primary reference assembly to "decoy" (unmapped) sequences in the new hg38 version (Table S4, right).

Differences in Genomic Annotations

In the hg19 data files a single Gene_Symbol column was present, which contained a semicolon-delimited list of associated gene symbols or an empty value if the probe did not overlap any gene annotation. The annotations were from Illumina's manifest file for the HM450 platform (and based on RefSeq Gene v. 2010), and an association was only included if the probe overlapped the body of an annotated gene, i.e. the interval from the Transcription Start Site (TSS) to the Transcription Termination Site (TTS) (Bibikova et al., 2011). In contrast, GENCODE v22 (Harrow et al., 2012) was used for gene annotation of the hg38 data, and as shown in Table S5, the format and content of hg38 data files is considerably different: (i) the probe overlapping criterion was changed so that Gene_Symbol would reflect an association between a probe and transcript if it was located within the range from 1500-bp upstream of the TSS to the TTS, (ii) the Transcript_ID, Gene_Type and CGI_coordinate columns were added to describe the associated transcript, its functional type, and overlapping CpG Islands; (iii) the relative distance of the interrogated CpG to the TSS (Position_to_TSS column) is provided to allow more flexible thresholding. Because the gene symbol and type are listed once for each transcript designated by the transcript ID, the same symbol/type may appear repeatedly. Likewise, multiple genes can be associated if they overlap the same interrogated CpG.

Most probes (64% HM27 and 67% HM450 probes) were completely concordant between hg19 and hg38, meaning that the gene(s) annotated were completely identical between the two releases (Figures S2A and S2B). 25% (HM450) and 28% (HM27) of probes retained the hg19 association(s) but were associated with additional genes in hg38. A substantial fraction of these augmented associations was due to the expanded non-coding gene annotation in GENCODE v22, with the three most frequently augmented gene categories being protein coding, antisense, and lncRNA (Figures S2A and S2B). Gene name changes also contributed to these differences: for roughly 25k probes (23,508 HM450 and 1,467 HM27), the probe was associated with the same gene in both hg19 and hg38 releases, but the gene symbol was updated in hg38. While this might be considered a trivial change, it could affect analyses that rely on symbol matching, and we thus recommend always using ENSEMBL identifiers included in the new hg38 version. However, remapping to canonical gene names reduces the level of discordant gene associations to 2.1% for HM450 and 1.2% for HM27 (Figures S2C and S2D).

To better understand how these gene annotation changes could affect discovery in a real-world scenario, we performed a search for genes with expression that was significantly correlated with methylation of the gene's promoter. This approach was commonly used in TCGA to infer epigenetic silencing of genes such as BRCA1 in ovarian cancer (Cancer Genome Atlas Research Network, 2011). Following the methodology used in (Wang et al., 2018), we searched for strongly negatively correlated probe-gene pairs (SNCs) in all TCGA cancer types by calculating a Spearman correlation for each HM450 probe within 1,500bp of each annotated TSS, using RNA-seq expression values obtained from GDC data release 12. We then calculated an FDR based on all pairwise comparisons, and considered any pair with FDR<=0.05 and Rho<=-0.5 to be an SNC.

Because the newer gene associations contained more genes and alternative transcripts, and associated CpG probes both upstream and downstream of the TSS (illustrated in Figure 3B), we were able to identify substantially more significant associations in the hg38 methylation data than in its hg19 predecessor (Figure 3C). For protein-coding genes, the number of SNCs was almost 2-fold higher, whereas non-coding RNAs were almost completely restricted to hg38 since most were not included in the early RefSeq version used for hg19. An example of a new protein coding association is PAX8, an important cancer-associated gene in ovarian cancer and other cancer types (Bowen et al., 2007; Nonaka et al., 2008), for which a novel SNC was identified for an alternative promoter that was annotated only in the hg38 data (Figure 3D). De-methylation of the associated promoter CpG is associated with increased expression in a subset of TCGA-CHOL tumors (Figure 3E).

Additional GDC Resources for DNA Methylation

While the analysis here uses data from the most recent (v.12) GDC release of hg38, our group has already begun to implement improved methods that may be incorporated into future GDC pipelines. We have made some of these methods and tools available on the GDC Community Tools webpage (https://gdc.cancer.gov/access-data/gdc-community-tools), including the following. *Improved DNA Methylation Array Probe Annotation* provides a set of probe annotation sets for the different Infinium methylation arrays that provide improved experiment-independent quality masking methodology described in our previous study (Zhou et al., 2017). *SeSAMe (SEnsible Step-wise Analysis of Methylation data)* provides a Bioconductor R package that can be used for improved signal processing and experiment-specific probe masking of Infinium methylation array data (Zhou et al., 2018b). Whole-Genome Bisulfite Sequence (WGBS) data for 47 TCGA samples (Zhou et al., 2018a) used to validate these new methods is available at GDC only in the Legacy hg19 archive (Data S3.5). Users wishing to obtain hg38 alignments for this dataset can find them at the supplemental data website: http://zwdzwd.github.io/pmd.html.

Analyses of mRNA Expression

TCGA samples that were assayed by RNA were largely tumor tissue samples, but some were derived from normal or adjacent tissue. As described in the main text, we obtained the RNA-seq data available from the GDC Data Portal (based on GDC data release 10), which is divided into two main archives: the ''current'' archive, and a "legacy" archive. The legacy archive contains data which the GDC inherited from two previous data repositories: the TCGA DCC and CGHub. This is primarily hg19 data. The current archive contains data which has been re-processed at the GDC, using updated references, including hg38 and Gencode v22. (Note that there are two separate "entry points" for these two data sets: the main GDC Data Portal for the hg38 data, and the Legacy Archive portal for the older data) A complete set of files used during this evaluation is provided in Data S4.1.

Legacy Archive

The legacy archive repository can be queried for TCGA mRNA-Seq data interactively. The results indicate:

- Data Category counts:
  - 82,256 Gene Expression
  - 25,968 Raw sequencing data
  - 3,817 simple nucleotide variation
  - 3,147 structural rearrangement
- Data Type counts:
  - 11,373 Unaligned reads
  - 14,533 Aligned reads
  - 22,566 Isoform expression quantification
  - 27,460 Gene Expression quantification
- Data Format counts:

- 82,318 TXT
- 14,533 BAM
- 10,113 FASTQ
- 5,395 VCF
- 1,569 FA
- 1,260 TAR

- Platform counts:
  - 104,253 Illumina HiSeq
  - 10,935 Illumina GA
- Access Level counts:
  - 82,318 open
  - 32,870 controlled

The legacy data were submitted from two different source sites with separate processing platforms. This explains the small subsets for some data types such as simple nucleotide variation, and data formats such as VCF, FA, and TAR.

Current Archive

The current archive repository can be queried for TCGA mRNA-Seq data interactively. The results indicate:

- Data Category counts:
  - 34,713 Transcriptome profiling
  - 11,604 Raw Sequencing Data
- Data Type counts:
  - 34,713 gene expression quantification
  - 11,604 Aligned Reads
- Workflow Type counts:
  - 11,604 STAR 2-Pass. HTSeq-Counts
  - 11,604 STAR 2-Pass. HTSeq-Counts
  - 11,604 HTSeq-FPKM
  - 11,604 HTSeq-FPKM-UQ
- Data Format counts:
  - 11,604 BAM
  - 34,713 TXT
- Platform counts:
  - 11,096 Illumina
- Access Level counts:
  - 34,713 open
  - 11,096 controlled

Workflows

Methods for accurate alignment of RNA-seq and estimation of transcript abundance were under rapid development during TCGA. Initial methods utilized transcriptome alignment, but improved methods quickly followed that permitted genome alignment and translation back to the transcriptome space. These tools have merged and matured resulting in a simplified alignment workflow that provides accurate alignments for use by an arbitrary quantification algorithm. The procedure for quantification was also simplified to facilitate interrogation and interpretation.

The legacy workflow aligned fastqs to the hg19 genome using MapSplice (Wang et al., 2010), translated the genome coordinates to the transcriptome based on adaptation of UCSC knownGene, and performed quantification of this transcriptome with RSEM. The resulting count estimates were normalized to fixed upper quartile values (500 for isoform estimates and 1000 for gene estimates) and formatted for dissemination. The complete set of commands and references to replicate a bam file or abundance estimate are provided here: https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/.

The current workflow begins with legacy bam files which are reformatted as fastqs using biobambam. These are then aligned to the hg38 genome using the STAR 2-pass approach (Dobin and Gingeras, 2015). The Gencode v22 transcriptome definition is then quantified using htseq-count procedure within samtools. Raw counts, FPKM, and upper quartile normalized FPKM estimates are provided: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/.

Comparative Analysis

The data used for this comparison were RNA-seq legacy Genomic Data commons data based on hg19 annotation (“legacy”; Legacy Data) and current GDC (current data) data based on hg38 annotation. Data were extracted for all samples in the BRCA, HNSC, and LUSC cohorts. The legacy data annotation for each sample used Entrez geneid and gene symbol which was mapped to ENSEMBL id then merged with the current workflow ENSEMBL id. A total of 2302 samples (BRCA=1205, HNSC=546, LUSC=551) and 19,744 genes could be unambiguously mapped between the legacy and current data. Subtype labels for each sample were retrieved

from the supplemental information of the primary publications and merged by TCGA barcode. Legacy expression data were represented by the upper-quartile normalized count estimates, and FPKM estimates were used to represent results from the current cohort. For paired samples, we calculated the Spearman correlation of count values for all mRNA between the legacy and current quantification values. Differential expression among subtypes was assessed using the log ratio of mean normalized count between groups. We used the r-squared estimate to represent the difference between log ratios of group comparisons by legacy and current dataset. A summary of gene wise absolute differences across workflows is provided in Data S4.2.

We first assessed the differences at the level of sequence alignment and abundance estimation. The results of sequencing alignment indicate a higher proportion of bases aligned in the current workflow relative to legacy (Figure S3A). In contrast, little difference was observed in alignments to mRNA (Figure S3B). Instead, the increase in total alignment is explained by a substantial increase in rRNA annotated alignments (Figure S3C).

A very high concordance between the legacy and current quantification values was found based on computing the Spearman correlation of count values for all mRNAs. A mean Spearman's rho of 0.943 is observed with range of 0.893–0.959 across all cohorts. Similarly, the median abundance estimate by gene was observed to be highly concordant (adjusted r^2 of 0.866) despite the difference in measurement scale (upper quartile normalized counts versus FPKM) (Figure S3D). Similar results are observed for the LUSC and HNSC cohorts, or when utilizing the upper quartile normalized form of the Current workflow.

We expected that treating the data as relative abundance, as when comparing experimental groups, the scale bias would be removed. The published subtypes of BRCA, HNSC, and LUSC cohorts were used as experimental groups, and differential expression was assessed using the log ratio of mean normalized count between groups. Indeed, when comparing BRCA luminal to basal and then comparing differential abundance we find excellent concordance (adjusted r^2 of 0.933). Only 3% (517 of 18556) of genes are observed to exceed a 0.5-fold change between the estimates.

These same experiments were repeated within the LUSC and HNSC cohorts. In all cases high correlation of sample measurements are observed when comparing each sample across the two protocols, and more annotated genes are detected in the current data. Most importantly, as evident by a mean R-square of 0.91 with range 0.862-0.947, the relative change between conditions is preserved across every comparison that was attempted (Figures S3E and S3H).

While the overall evidence for differential expression between groups generally agrees, a relatively small number of genes consistently demonstrated discordant results. These genes, characterized by more than 0.5-fold change in relative expression ($|\text{current } \log_2(A/B) - \text{legacy } \log_2(A/B)| > 1$), constitute 3.9% of estimates on average (range 3.1% - 5.1%). A complete list of genes and their concordance across these experiments is provided in the table of discordance measures (Data S4.2).

The transcriptome definitions were compared to assess if the concordant and discordant genes showed differences in annotation. We observed a large magnitude of difference in the legacy transcriptome (primarily built from the UCSC knownGene definitions in 2010) defined in hg19 versus the current Gencode v22 definitions defined for hg38. A total of 3573 Gencode transcripts could be perfectly matched to their definitions in the legacy annotation. The remainder were split into 131,457 Gencode transcripts with imperfect matches, and 12,308 Gencode transcripts that could not be matched to any legacy transcript. The discrepant annotations were not associated with discordant genes. Further, when considered with previous results, we find that the large magnitude of change at the level of transcript annotation does not greatly affect gene level abundance estimates.

Users are advised that alignments or expression should not be compared directly between legacy and current workflows. Comparisons should be restricted to samples originating from the same workflow. Under this recommendation, we find relative abundance results to be generally concordant across both workflows.

Analyses of Somatic Mutations

Throughout the decade that spanned the TCGA project, somatic mutation calling has been constantly improved and all the accumulated knowledge was recently applied to a harmonized set of mutation calls across all TCGA samples: the Multi-Center Mutation Calling in Multiple Cancers (MC3) project, as a part of the TCGA Pan-Cancer Atlas effort. The unified MC3 somatic mutations were called using standardized protocols and annotated with various filters to detect potential sequencing artifacts and label low quality variant calls. Two centers, DNAnexus and FIREHOSE, generated somatic mutation calls for each pipeline (7 tools) on each sample (>10,000 tumor-normal pairs), from 33 cancer types (Ellrott et al., 2018). The tool developers supplied the identification of tool-specific, sample-specific, and mutation-specific filters. The filter flags were present in a comma separated column which were carried through to the mutation annotation file (MAF). These mutation calls supply the basis for many PanCancer Atlas analysis working groups.

Despite the uniformity, the MC3 pipeline was designed for the human genome reference hg19. On the other hand, GDC has developed a different somatic mutation calling pipeline for a newer human genome reference version, hg38. The GDC pipeline was designed to be fully automatic. Therefore, whenever an algorithm of variant calling or filtering gets improved, GDC can generate an updated set of uniform variant calls across all samples by updating the data release version.

Data Source

GDC MAFs in hg38 were obtained through the GDC Data Portal. GDC version 12.0 was used, which was released in June 2018. The query for the file retrieval was: cases.project.project_id in ["TCGA-LAML", "TCGA-BRCA", "TCGA-COAD", "TCGA-OV"] and files.data_format = "MAF". The query returned both somatic and protected MAFs, totaling 32 files and covering 2,069 tumor and paired normal samples. Note that protected somatic mutation calls did not reflect the full pool of germline variants. Protected MAFs contained the raw mutation calls from all somatic mutation callers. Each protected MAF underwent various filters to remove calls

of low quality or potential germline variants reported in the germline variant databases like dbSNP (Sherry et al., 2001) and Exome Aggregation Consortium (ExAC) (Lek et al., 2016). Germline variant calling used a different algorithm, thus not all the germline variants would be captured in the protect MAFs, which is beyond the scope of our study.

TCGA MC3 MAFs were obtained from Synapse (somatic: syn7214402; protected: syn5917256). The genomic coordinates of MC3 variant calls were lifted over from hg19 to hg38 using CrossMap v0.2.7 and chain files from University of California, Santa Cruz (UCSC). Variant calls that could not be lifted over were dropped and excluded from our analysis. Among the four cancer types, 1,902 samples that had somatic mutation calls from both MC3 and GDC were used for the rest of the comparison.

Mutational Calling Pipelines

The GDC variant calling pipeline of release 12.0 was described as follows: the pipeline started with genome re-alignment to GRCh38.d1.vd1 by extracting sequencing reads from a sample's BAM files using Biobambam. The re-aligned BAM files were merged and cleaned up using Picard and GATK. Four tools were used for the variant calling: MuSE, MuTect2, VarScan2, and SomaticSniper. MuTect employed a "Panel of Normals" (PoN) to reduce the rate of germline variant call and, more often, recurrent sequencing artifacts. The PoN was selected from the TCGA blood normal genomes curated to be cancer-free. VarScan and MuTect are also able to generate indel calls, which were collected together with the single nucleotide variant (SNV) calls. Variant calls were annotated using Variant Effect Predictor (VEP) version 84, which was based on GENCODE version 22. Various filters were added to both MAF and Variant Call Format (VCF) files during the annotation. Please refer to the following GDC documentation for more details:

- File Format: MAF: https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/
- File Format: VCF: https://docs.gdc.cancer.gov/Data/File_Formats/VCF_Format/
- Bioinformatics Pipeline: DNA-Seq Workflow—Somatic Variant Calling Workflow: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/#somatic-variant-calling-workflow

Complete details of the MC3 variant calling pipeline were provided in the MC3 publication (Ellrott et al., 2018), with the key aspects as follows. Non hg19 aligned samples were excluded from the pipeline. Implementation of these tools were performed by DNAnexus including SomaticSniper (Larson et al., 2012), MuSE (Fan et al., 2016), Pindel (Ye et al., 2016), Radia (Radenbaugh et al., 2014), and Varscan (Koboldt et al., 2012) and by Broad Institute including MuTect (Cibulskis et al., 2013) and Indelocator. Filtering for each tool was provided by the tool developers. Filtered VCFs were merged and converted to MAF. During the conversion, annotations were provided by many databases such as Ensembl version 75, GENCODE v19 using Variant Effect Predictor (VEP) version 82. A Panel of Normals (filter flag: broad_PoN_v2) was also used. Sample level annotations were also added to the MAF, including additional filters: gapfillter, contest, badseq, nonpreferredpair, and wga. The MAF file was then split to 2 MAFs. A flagged MAF available to the public, and a protected MAF harboring all mutations that were merged after flagging potential artifacts.

Comparative Analysis

The variant call overlap between GDC and MC3 was done by matching their genomic location and tumor allele 2. To identify the potential causes for the unique calls in the two groups, we investigate the overlap result in the following subsets: unique calls in each group, sample-wise overlap, and recoverable unique calls in each group. To understand whether the overlap between GDC and MC3 calls varies across different samples and cancer types, we calculated the proportion of shared calls over total GDC and MC3 calls for every sample.

DATA AND CODE AVAILABILITY

Supplemental Data provide key information for reproducing the results reported here, such as the index files for the NCI GDC raw data files used in this study. They are available at https://gdc.cancer.gov/about-data/publications/HG38QC.

ADDITIONAL RESOURCES

GDC Community Tools Webpage: https://gdc.cancer.gov/access-data/gdc-community-tools.