Bioinformatics workflows for genomic analysis of tumors from Patient Derived Xenografts (PDX): challenges and guidelines

Xing Yi Woo[*,1], Anuj Srivastava[*,1], Joel H. Graber[*,3], Vinod Yadav[1], Vishal Kumar Sarsani[2], Al Simons[2], Glen Beane[2], Stephen Grubb[2], Guruprasad Ananda[1], Rangjiao Liu[1], Grace Stafford[2], Jeffrey H. Chuang[1], Susan D. Airhart[2], R. Krishna Murthy Karuturi[1], Joshy George[#,1], Carol J. Bult[#,2]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA

[2] The Jackson Laboratory, Bar Harbor, ME 04609, USA

[3] MDI Biological Laboratory, Bar Harbor, ME 04609, USA

[*] The authors contributed equally to this work.

[#] Corresponding authors: Carol J. Bult (Email: Carol.Bult@jax.org), Joshy George (Email: Joshy.George@jax.org)

Abstract

Bioinformatics workflows for analyzing genomic data obtained from xenografted tumor (e.g., human tumors engrafted in a mouse host) must address several challenges, including separating mouse and human sequence reads and accurate identification of somatic mutations and copy number aberrations when paired normal DNA from the patient is not available. We report here data analysis workflows that address these challenges and result in reliable identification of somatic mutations, copy number alterations, and transcriptomic profiles of tumors from patient derived xenograft models. We validated our analytical approaches using simulated data and by assessing concordance of the genomic properties of xenograft tumors with data from primary human tumors in The Cancer Genome Atlas (TCGA). The commands and parameters for the workflows are available at https://github.com/TheJacksonLaboratory/PDX-Analysis-Workflows.

Introduction

Patient-Derived Xenograft (PDX) models are *in vivo* preclinical models of human cancer for translational cancer research and personalized therapeutic selection [1-7]. Previous studies have demonstrated engrafted human tumors retain key genomic aberrations found in the original patient tumor [3, 8, 9] and that treatment responses of tumor-bearing mice typically reflect the responses observed in patients [6, 10]. PDXs have been used successfully as a platform for pre-clinical drug screens [6, 7, 10], to facilitate the development of potential biomarkers of drug response and resistance [6, 7, 11], and to select appropriate therapeutic regimens for individual patients [8].

The Jackson Laboratory (JAX) PDX Resource has over 400 PDX models from more than 20 different types of cancer. A schematic summarizing the processes used for model generation, quality control, and characterization process for the resource is shown in Figure 1. Genome characterization of PDX tumors includes the identification of somatic mutations, copy number alterations, and transcriptional profiles. Over 100 of the models have been assessed to date for responses to various therapeutic agents. The integration of results from dosing studies with genomic data for the models has been successfully applied to the identification of novel genomic biomarkers associated with treatment responses [12].

To generate accurate calls for mutations and copy number variants for human tumors engrafted in a mouse host, several challenges had to be addressed. First, because human stroma is replaced by mouse cells and tissues during tumor engraftment, sequence data generated for PDX tumors includes both mouse and human sequences. As the protein-coding regions of the mouse and human genomes are 85% identical on average, there is a high risk of introducing false positive variants in functional regions and erroneous gene expression [13-15]. Second, because the tumor material used to create models in the JAX PDX Resource consisted of material that remained following clinical pathology assessment (i.e. material was not collected specifically for xenograft model creation), paired normal samples were not available for the majority of tumor samples used to generate the PDXs. The absence of normal tissue complicates the

ability to distinguish germline variants from somatic alterations (point mutations, indels and copy number aberrations) in the tumor [16-19]. Third, false positive (FP) variants due to errors in sequencing and mapping require additional filtering steps in the computational workflow [20-22]. Finally, it has been reported previously that the immunodeficient host mice are susceptible to forming B-cell human lymphomas during engraftment due to Epstein-Barr virus (EBV)-associated lymphomagenesis [23-27]. Systematic screening of PDX tumor samples for EBV transformation is an important step in quality assurance for the integrity of PDX repositories.

Here, we describe bioinformatics analysis workflows and guidelines (https://github.com/TheJacksonLaboratory/PDX-Analysis-Workflows) that we developed for the for the analysis of genomic data generated from PDX tumors (http://www.tumor.informatics.jax.org/mtbwi/pdxSearch.do). These workflows incorporated established tools and public databases and were tailored to address the specific challenges mentioned above by tuning parameters and addition of filters. We demonstrate how our methods, using simulated and experimental data, improve the accuracy in the detection of somatic alterations in PDX models. We also developed a classifier based on expression data to systematically identify and filter out EBV transformed samples. Finally, to verify the effectiveness of our workflows, we show the overall concordance of the genomic and transcriptomic profiles of the PDX models in the JAX PDX resource with relevant tumor types from The Cancer Genome Atlas (TCGA).

Results

Workflow for calling somatic point mutations and indels in PDX tumors

A schematic of the variant calling workflow we implemented for human tumors engrafted in mice is shown in Figure 2A and 2B (see Methods).

Preprocessing and removal of mouse reads. Human and mouse DNA reads were classified by Xenome [13], which had shown reliable performance in separate studies [28], and only human reads were used for subsequent variant calling. The percentage of mouse reads within the PDX samples in the JAX resource has a

median value of 5.30% (range: 0.00163% - 65.1%) (Figure 3A). Using simulated CTP datasets, we verified that omitting the Xenome step to filter the mouse reads resulted in very low precision (Figure 3B), i.e. large number of FPs, in the absence of the quality hard filters (Supplementary Table S1). These FPs were due to mouse reads being aligned to the reference genome with mismatches and subsequently called as variants with low quality scores (QD).

While the default thresholds for GATK hard filtering parameters [29] removed a large proportion of the FPs, applying Xenome to filter for human reads yielded superior performance in terms of substantially higher precision, as well as improvement in recall. In addition, Xenome filtering maintained the correlation between the predicted versus actual allele frequencies, which would otherwise decrease with higher mouse contamination (Supplementary Table S2).

Filtering germline variants. To enhance filtering out germline variants from somatic mutations, we sequenced and analyzed 20 normal blood samples using the CTP targeted panel. As shown in Supplementary Figure S1A and S1B, 87% of the variants identified in normal blood had allele frequencies of 40% - 60% or >90% across all the samples, indicating the presence of heterozygous or homozygous common variants, respectively. Ninety-one percent of the variants identified in these 20 samples were annotated in the public germline databases. 4% of these variants were not found in public germline databases, but were recurrent in these normal samples or across the PDX tumors in our collection (Supplementary Figure S1C) and so were added to our list of putative germline variants. Only 5% of all of the variants in the 20 samples were private events. Based on these observations, the variants in each PDX tumor with an allele frequency of 40% - 60% or >90%, and present in either public germline database or our list of putative germline variants (Supplementary Table S3) were filtered out as germline variants (Supplementary Table S3). This was a more conservative approach given that these known germline variants in regions of copy number alterations where the ratio of both alleles were not balanced would not be filtered. Figure 3C shows that the germline filters effectively rectified the estimated somatic mutational load in the PDX tumors (Supplementary Table S5) by about

four-fold reduction (Supplementary Table S4), which was reasonable as a large proportion of the variants were expected to be germline.

Filtering false positives due to systematic errors. Putative somatic variants with no known effects in cancer that recur across large numbers of PDX samples are potentially FPs arising from sequence assembly based error in the reference the genome, sequencing errors or alignment errors in low mappability regions [30]. To detect these, we filtered out the variants at loci that were recurrently mutated in ≥25% of PDX tumors (Figure 2C). The distribution of tumor types for each of these recurrently mutated positions (n=52) was highly similar to the overall distribution of tumor types in the PDX resource (Supplementary Figure S2A) with Pearson correlation coefficient >0.9 (Supplementary Figure 2B). This implies that these mutations were systematic errors and were not selected for any tumor type, and thus, biologically irrelevant. Filtering these highly recurrent loci did not significantly reduce the predicted mutational load per tumor (Figure 3C and Supplementary Table S4).

Rescuing variants. The germline filters might filter out actual somatic events in each PDX sample, leading to false negatives. However, retaining all variants represented in cancer variant databases such as COSMIC would lead to excess FPs. For example, 46% of the variants in the normal samples are present in the COSMIC database (Supplementary Figure S1A). To address the balance of false positive and false negative mutation calls, we "rescued" variants that were initially filtered out based on curated annotations available in the JAX-Clinical Knowledgebase (CKB, https://ckb.jax.org/) [31]. The criteria for rescuing variants included those with 1) known or predicted gain or loss of protein function, 2) potential treatment approach for any cancer type and 3) drug sensitivity and resistance effects in clinical or preclinical studies (Supplementary Table S4). We also included an additional indel caller, Pindel [32], in the workflow in order to increase the sensitivity of indel prediction. As Pindel results contained a large number of FPs, we only included those that were present in the JAX-CKB by the same criteria. Overall, 127 unique variants from 52 genes (1.03% of the total and 2.21 % of the filtered unique variants detected by the CTP platform) were rescued from 381

PDX CTP samples. Nine of these mutations have been validated to be present in the PDX model (Figure 3D). Almost all were initially filtered as germline events, as many well-known actionable cancer mutations (e.g. BRAF V600E and KRAS G12C) are present in the dbSNP database and were filtered if they fall within the germline allele frequency. Two other variants that were not called by GATK initially but were detected by Pindel were rescued as they were annotated clinically relevant.

Optimized workflow achieves high performance in somatic mutation calling. Figure 3B shows that our full feature workflow on the simulated datasets achieved the highest precision in variant calling, with insignificant compromise on the recall (Supplementary Table S1). We observed that the allele frequencies of the true positive (TP) variants correlates well (Pearson correlation coefficient >0.99) with the input allele frequencies for all samples (Supplementary Figure S3 and Figure S4, and Supplementary Table S2). Although the estimated allele frequencies were lower than the true allele frequencies, this difference was marginal and could be attributed to the reads carrying the variants being classified as non-human reads by Xenome or not mapped to the genome. Moreover, all (20 out of 20) clinically relevant mutations experimentally validated or clinically reported in the corresponding patient tumors were detected in the PDX tumors (Figure 3D).

Gene expression analysis in PDXs

A schematic overview of the PDX gene expression workflow is provided in Figure 4A (see Methods).

Screening of EBV-associated lymphomas by RNA-Seq expression data. We observed that the EBV-associated lymphoma tumors that arise in PDX samples display a distinct and highly reproducible expression pattern, regardless of the platforms in which the expression was measured (RNA-Seq, Affymetrix Human Gene 1.0 ST arrays and Human Gene 133 Version 2 arrays). The PDX tumors identified as EBV-associated routinely showed higher correlation in expression profiles than distinct pairs of PDX models derived from common original tumor materials (Supplementary Figure S5). This expression profile was also

independent of the tissue of origin of the tumors from which the EBV-associated lymphomas were derived. Given the high similarity in expression profiles, we identified a gene signature based on the most differentially expressed genes between EBV-associated lymphomas and non-EBV-associated tumors (data not shown). Using gene set analysis, we observed that genes associated with B-lymphocytes and other immune processes were over-expressed, while cell-to-cell communication and adherence genes were suppressed (data not shown). We developed a classifier that scored each PDX sample based on the expression levels of the genes in the gene signature (Supplementary Table S6). This single score, when applied on RNA-Seq data, was able to effectively distinguish PDX tumors that were either EBV-transformed or originated from human lymphomas from non-lymphoma PDX tumors (Figure 4B). Overall, 8.5% (32 out of 376) of the non-lymphoma PDX samples with RNA-Seq data in the PDX resource progressed to EBV-associated lymphomas. These tumors were further confirmed to be CD45 positive by immunohistochemistry (IHC) staining, which is the primary tool at JAX to identify PDX tumors that are EBV-transformed.

Copy Number Variant (CNV) analysis in PDXs

A schematic overview of the PDX CNV workflow is provided in Figure 5A (see Methods).

***Effect of mouse DNA* on CNV calls.** We studied the effect of mouse contamination on array data by hybridizing DNA of the NSG mouse on the human SNP array, and observed that the signal intensity from mouse DNA is negligible (Supplementary Figure S6). Samples with higher mouse content are more likely to result in failure of the standard array quality control or the analysis workflow, due to lower amount of human DNA to give sufficient probe signal, thus enabling samples with substantial mouse contamination to be screened out.

Absence of matched normal to call somatic copy number aberrations. We compared the results of the single-tumor CNV analysis with the tumor-normal CNV analysis to access the reliability of the single-tumor CNV analysis results. For the limited number of PDX samples with paired normal samples, we observed

overall high similarity between the segmented copy number profiles analyzed with and without the paired-normal sample (Supplementary Figure S7). The gene-based log2(total CN/ploidy) showed good correlation between the single-tumor and tumor-normal CNV analysis (Pearson correlation >0.81, n=9), with 8 out of 9 PDX samples having a correlation of >0.93 (Supplementary Table S7), indicating that the single-tumor CNV analysis was sufficiently robust.

Establishing the appropriate baseline to call copy number gains and losses. We analyzed the effects of using different baselines for "normal state" to compute copy number gains and losses using a list of significantly amplified and deleted genes from TCGA (Supplementary Figure S8). When the overall cancer genome ploidy was used as the normal baseline, we observed a balance of a larger proportion of the significantly amplified being called copy number gain, and similarly a larger proportion of the significantly deleted genes being called copy number loss among the PDX samples (Supplementary Figure S9). However, more of both significantly amplified and deleted genes were being classified as amplified when copy number aberrations were calculated relative to the diploid state. While the average ploidy could be estimated differently across the samples for the same model, the copy number changes relative to ploidy remained consistent (Figure 5B and Supplementary Figure S7).

Effects copy number aberrations on expression changes. We observed that the estimated copy number gains and losses of known oncogenes (n=23) and tumor suppressor genes (n=40) [33], relative to the average ploidy per PDX sample, generally results in expression fold change (relative to the average expression at copy number normal state) in the same direction (Supplementary Table S8) [11, 34, 35]. Most of these genes show significant over-expression with copy number gain and significant under-expression with copy number loss across the PDX samples ($p<0.05$) (Figure 5C and Supplementary Figure S10). This shows that the baselines to call copy number gain and loss, and over and under-expression, were correctly established. This significant observation, however, did not hold when we did a global analysis across all genes instead of selected oncogenes and tumor suppressor genes. This was because many genes were not expressed in the

respective tissue types even though they were in regions affect by copy number alterations, and the expression of many genes, despite being non-altered regions, could be regulated by other mutations or epigenetic mechanisms in the tumors.

Comparison of genomic and transcriptomic profiles of PDX models and TCGA patient tumors

Due to the lack of paired-normal samples for the PDX models in the JAX PDX Resource, we were unable to experimentally validate the somatic calls predicted from the various workflows. To determine if the results of our genomic analysis workflows were similar to known somatic profiles of the same tumor type, we compared the overall genomic and transcriptomic profiles for selected tumor types between the JAX PDX resource and patient tumor cohorts in the TCGA.

Frequently mutated genes in primary patient tumors in TCGA detected in the PDX resource. The distribution of somatic coding non-silent mutational load of the CTP genes for each tumor type was comparable between PDX and TCGA (Figure 6A). Despite the much smaller sample size for each PDX tumor type, we still observed higher mutational load in colorectal cancer and melanoma. Nonetheless, the overall mutational load remained higher in PDX tumors, which could be possibly due to the fact that the PDX tumors were sequenced at a higher coverage (>900X) using the CTP targeted panel, and thus more variants were detected per base pair compared to exome sequencing (~100X) of TCGA tumors. Moreover, known germline variants with allele frequency outside the range of 40% - 60% and >90%, possibly due to errors in allele frequency estimation or copy number aberrations at the variant position, as well as private germline variants, were not filtered. The mutations in TCGA were curated with partial experimental validations, hence the mutation count and FP rate were expected to be lower. Given that there were more samples in the TCGA cohorts, we compared the genes that were mutated at 5% frequency with genes that were mutated in at least one sample within the same tumor type in the PDX resource. Almost all genes mutated at high frequencies in TCGA tumors were mutated in PDX tumors, with significant p-values ($p < 1\times10^{-4}$) by Fisher's exact test (Figure 6B, Supplementary

Table S9). This indicates that the key drivers by mutation within each cancer type were preserved in PDX tumors.

Expression signatures of primary patient tumors in TCGA recapitulated in the PDX resource. The top 1000 most varying genes by expression z-scores in 6 TCGA tumor types (Supplementary Table S10) were able to independently cluster both TCGA samples and the PDX samples by their tumor types (Figure 6C). We observed clusters of genes that were highly expressed in specific tumor types in TCGA were recapitulated in the PDX expression data (hypergeometric p-value < 1×10^{-8}), which demonstrated the replicability of TCGA expression signatures in the PDX resource. The frequencies of over- and under-expression for the top-varying genes for each tumor type displayed better correlation for the same tumor type for PDX versus TCGA compared to other tumor types (Figure 6D). The varying level of concordance between different tumor types in TCGA data was also maintained in the PDX versus TCGA comparison (Supplementary Figure S11). Alternatively, the differentially expressed genes of each tumor type versus all other tumors within the TCGA or PDX samples displayed significant overlaps ($p<1\times e^{-6}$), despite different sample sizes and different proportion of tumor types (Supplementary Table S11).

Copy number profiles of primary patient tumors in TCGA recapitulated in PDX resource. We showed that the frequency of genome-wide copy number aberrations for each tumor type in the PDX resource (Supplementary Table S12, Supplementary Figure S12) were similar to the primary tumors in TCGA (Supplementary Figure S13). Moreover, the PDX tumors had the highest correlation in gain and loss frequencies of significantly amplified and deleted genes for the same tumor type in TCGA compared to other tumor types (Figure 6E and Supplementary Figure S14A). The varying levels of correlation between different tumor types were preserved between the TCGA versus TCGA tumors and the TCGA versus PDX tumors (Figure 6E and Supplementary Figure S15B). Consistent with the earlier observations, there was a weaker concordance with TCGA data when amplification and deletion was called relative to the diploid state (Supplementary Figure S14B).

Discussion

The application of PDX models in pre-clinical research and personalized therapy requires that the engrafted human tumors are accurately characterized for tumor-specific mutations [3]. The development of bioinformatics workflows to call somatic mutations (SNVs, Indels), copy number aberrations and gene expression from PDX sequencing or array data requires balancing sensitivity and specificity [22, 30], especially when paired normal samples for engrafted tumors are not available. Using genomic and transcriptomic data from models in the JAX PDX Resource, we conducted a systematic analysis to address several key data analysis challenges and tailored our workflows to optimize the sensitivity and specificity of the results.

Our recommendations for the somatic mutation calling from PDX DNA sequencing data in the absence of paired-normal samples are as follows:

- Remove mouse reads with Xenome (or equivalent) to eliminate variants called from mouse reads mapping to the human reference genome
- Filter with germline variant databases to improve somatic mutation calling
- Filter highly recurrent mutations to remove false positives arising from sequencing or analysis related errors
- Rescue clinically relevant variants which were filtered in the upstream steps as they were likely to be present as important mutations in the tumor

Despite implementing multiple filters to remove putative germline and other FP mutations, the mutation rate remains higher in PDX tumor types when compared to TCGA. One possible reason for this difference that is not related to the informatics challenges described in this paper is that many of the human tumor samples used to generate PDX models arose from metastatic lesions and from patients with prior treatment whereas many of the tumor samples used for TCGA were early stage tumors. PDX tumors were thus expected to harbor more mutations due to longer tumor evolution [36, 37]. Also, previous studies have noted that PDX engraftment success is better for late stage tumors that are likely to have more aggressive phenotypes than early stage tumors [38, 39]. As such,

there is a likelihood for biased selection towards such tumor subtypes in the engrafted tumors that are known to harbor more mutations than tumors from early stages.

For evaluation of gene expression differences in individual tumors, matched normal tissue is ideal but not available for PDX models in the JAX PDX Resource. To compare gene expression among the engrafted tumors, we used expression z-scores across all tumor types as the best proxy for calling over- and under-expression. In a subset of PDX samples in which both expression and copy number data are available, we estimated the "normal" expression of each gene with the average expression for samples with normal copy number state, given that sufficient samples are available for the tumor type. While this approach neglects other mechanisms of gene regulation, we were able to better estimate the normal expression for some genes like MYC which tends to be frequently amplified and over-expressed across many tumor types. For copy number, we defined the "normal" state of each PDX tumor using the estimated ploidy to call relative gain and losses as this takes into account errors in ploidy estimation.

As one approach to assessing the results of our genomic characterization workflows, we compared the JAX PDX models with patient cohorts in TCGA at the genomic and transcriptomic level. Other than small differences in genomic mutations, the engrafted PDX tumors reflected the human tumors in copy number variations and gene expression. Using colorectal cancer as an example, we demonstrated that the integration of different data types showed that known perturbed pathways in cancer were altered in a consistent manner across PDX and TCGA tumors (Supplementary Figure S16), with similar combinations of alterations occuring at comparable frequencies. Taken together, we have created a set of workflows for the analysis of genomic and transcriptomic data from PDX tumors that have no paired normal sample to reliably identify true somatic mutations and expression changes.

Acknowledgements

The workflows reported in this publication was partially supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA034196. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The public portal for JAX PDX data is supported by R01CA089713.

Methods

Genomic and transcriptomic profiling of samples

DNA sequencing. Flash frozen tissues were pulverized using a Bessman Tissue Pulverizer (Spectrum Chemical) and homogenized in Nuclei Lysis Buffer (Promega) using a gentleMACS dissociator (Miltenyi Biotec Inc). DNA was isolated using the Wizard Genomic DNA Purification Kit (Promega) according to manufacturer's protocols. DNA quality and concentration were assessed using a Nanodrop 2000 spectrophotometer (Thermo Scientific), a Qubit dsDNA BR Assay Kit on a Qubit Fluorometer (Thermo Scientific), and the Genomic DNA ScreenTape on a 4200 TapeStation (Agilent Technologies). Libraries were prepared using the Hyper Prep Kit (KAPA Biosystems) and SureSelectXT Target Enrichment System with the JAX Cancer Treatment Profile (CTP) targeted panel (Agilent Technologies), according to the manufacturer's instructions. Briefly, the protocol entails shearing the DNA using the Covaris E220 Focused-ultrasonicator (Covaris), ligating Illumina specific adapters, and PCR amplification. Amplified DNA libraries are then hybridized to the CTP probes, amplified using indexed primers, and checked for quality and concentration using the High Sensitivity D5000 ScreenTape (Agilent Technologies) and Qubit dsDNA HS Assay Kit (Thermo Scientific). Libraries were pooled and sequenced 150 bp paired-end on the NextSeq 500 (Illumina) using NextSeq v2 reagents (Illumina).

RNA sequencing. Tissues preserved in RNAlater were homogenized in TRIzol (ThermoFisher Scientific) using a gentleMACS dissociator (Miltenyi Biotec Inc). Total RNA was isolated using the miRNeasy Mini kit (Qiagen) according to manufacturer's protocols, including the optional DNase digest step. RNA quality and concentration were assessed using the RNA 6000 Nano LabChip assay on the

2100 Bioanalyzer instrument and Nanodrop 2000 spectrophotometer (Thermo Scientific). Prior to 2016, non-stranded libraries were constructed using TruSeq RNA Library Prep Kit v2 (Illumina). Starting in 2016, stranded libraries were prepared by the Genome Technologies core facility at The Jackson Laboratory using the KAPA mRNA HyperPrep Kit (KAPA Biosystems), according to the manufacturer's instructions. Briefly, the protocol entails isolation of polyA containing mRNA using oligo-dT magnetic beads, RNA fragmentation, first and second strand cDNA synthesis, ligation of Illumina-specific adapters containing a unique barcode sequence for each library, and PCR amplification. Libraries were checked for quality and concentration using the DNA 1000 assay (Agilent Technologies) and quantitative PCR (KAPA Biosystems), according to the manufacturers' instructions. Libraries were pooled and sequenced 75 bp paired-end on the NextSeq 500 (Illumina) using NextSeq High Output Kit v2 reagents (Illumina), or 100 bp paired-end on the HiSeq2500 (Illumina) using TruSeq SBS v3 reagents (Illumina).

SNP array. DNA samples were sent to the Genotyping Core at the Hussman Institute for Human Genomics (University of Miami) for genotyping on the Genome-Wide Human SNP Array 6.0 (Affymetrix). Quality control on the CEL files was carried out using the standard Contrast QC metric from the Affymetrix Genome Wide SNP 6.0 array manual.

Somatic point mutation and indel calling workflow

Preprocessing and removal of mouse reads. DNA sequence data generated from PDX tumors underwent initial data processing as follows: (i) sequence reads with 70% of the bases having a quality score <30 (Q30) were discarded, (ii) bases with quality scores less than Q30 were trimmed from the 3' end of the read, (iii) sequence reads with <70% of bases remain after trimming were discarded, (iv) both reads from pair-end sequencing were discarded if either read was discarded. If <50% of the total reads remained following the preprocessing steps, the sample was removed from the analysis. Following the initial data processing step described above, mouse reads were identified and filtered out using Xenome

v1.0.0 [13]. Only read pairs with both reads classified as human were included in further analyses.

Sequence reads that passed all pre-processing steps were mapped to the reference human genome (build GRCh38.p5 with 262 alternate loci) using the BWA-MEM alignment tool with ALT-Aware mapping (Supplementary Figure S14) [40, 41]. Because low sequence coverage leads to poor sensitivity in variant calling, samples with less than 75% of the target region covered at least at ≥100X by human reads were excluded from further analysis.

Variant calling. The GATK best practices workflow (https://gatkforums.broadinstitute.org/gatk/categories/best-practices-workflows) using the UnifiedGenotyper, was used for variant discovery analysis [42-44], which is comprised of the following steps: (i) sorting the SAM/BAM file by coordinate, (ii) removing duplicates to mitigate biases introduced by library preparation steps such as PCR amplification by Picard (https://broadinstitute.github.io/picard/), and (iii) recalibrating the base quality scores as the variant calling algorithms rely heavily on the quality scores assigned to the individual base calls in each sequence read. Pindel [32] was also incorporated into the workflow to call indels that have been missed by the GATK UnifiedGenotyper.

Quality filtering of variants for targeted sequencing. High quality variants from both variant callers in the PDX samples were obtained based on GATK hard filtering (see below), and have a read depth (DP) of ≥140 and allele frequency (ALT_AF) of ≥5%. These DP and ALT_AF thresholds were optimized using a set of known and validated mutations and samples reported earlier for the JAX CTP targeted panel sequencing at high coverage (average 941X) [45]. The parameters for GATK hard filtering [29] were set as default as recommended by GATK best practices (https://software.broadinstitute.org/gatk/documentation/article.php?id=6925, https://software.broadinstitute.org/gatk/documentation/article.php?id=3225, https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set):

(i) for point mutations, QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0

(ii) for indels, QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0.

In addition, we verified that these default thresholds were able to detect all the known mutations in the CTP samples [45]. The average number of variants before and after quality filtering across the CTP samples is shown Supplementary Table S4.

Annotation of variants. Variants were annotated for their effect (gene, consequence, amino acid change, etc.) using SnpEff v4.3 [46] based on gene annotations from Ensembl (version GRCh38.84) and information from COSMIC version 80 [47], dbSNP build 144 [48]. The observed variant allele frequency in 1000 Genomes Project [49] and ExAC version 0.3 [45, 50] database were obtained using SnpSift tool by utilizing dbNSFP3.2a.txt database. We further annotated each variant with 1) known or predicted gain or loss of protein function, 2) potential treatment approach for any cancer type and 3) drug sensitivity and resistance effects in clinical or preclinical studies, based on curated clinical information from the JAX clinical knowledge base (CKB, https://ckb.jax.org/) [31]. The average number of variants annotated to be clinically relevant across the CTP samples is shown in Supplementary Table S4.

Filtering of germline variants. Since normal samples were unavailable for patients whose tumors were used to generate the PDX models, we generated a dataset of putative human germline variants using data from several public resources: (i) dbSNP, (ii) 1000 Genomes Project, (iii) ExAC database with MAF ≥1%, and (iv) a compendium of variants from 20 normal blood samples that were prepped and sequenced on the CTP panel using the same protocol as the PDX samples, with a frequency of 2/20 in normal samples or 1/20 in normal samples and 2/20 in PDX models. The number of variants in each of these databases are shown in Supplementary Table S3. The variants identified via GATK and Pindel in the PDX model tumors were annotated as germline and filtered out of the model's somatic mutation calls if they were present in our aggregated dataset of putative germline variants and had allele frequencies between 40% to 60% or more than 90%.

Filtering putative false positives. Variants not in our aggregated dataset of putative germline variants described above but occurred at a frequency of 25% or greater across all PDX models (n=236) were considered to be putative false positive (FP) mutations. The rationale for this data filtering step was based on our observation that the maximum recurrent frequency of somatic mutated base positions was 6% across a compendium of TCGA tumor samples (n=3576, 9 tumor types that were also represented in the PDX model). Thus, we would expect that any mutated loci recurring across PDX samples at significantly higher rates to likely be FP. Systematic technical errors in sequencing and/or mapping are possible explanations for the common recurrent non-somatic mutations identified PDX models.

Rescuing the false negative variants. An exception to the germline and false positives exclusion process was made for variants (from GATK or Pindel) that were annotated as clinically relevant in JAX CKB. We rescued any filtered variants that were curated into the proprietary JAX-Clinical Knowledgebase (CKB, https://ckb.jax.org/) [31] with 1) known or predicted gain or loss of protein function, 2) potential treatment approach for any cancer type and 3) drug sensitivity and resistance effects in clinical or preclinical studies.

Benchmarking of PDX somatic mutation workflow

To benchmark the PDX somatic mutation workflow, a simulated dataset (45 samples) was generated that included sequenced reads that includes sequencing errors of an Illumina HiSeq were generated *in-silico* for different samples with, 1) varying sequencing coverage, 2) spiked-in mutations to the reference human sequence representative of different tumor types, and 3) different proportions of spiked-in mouse reads (Supplementary Table S1).

Generation of simulated sequence reads. SeqMaker was used to generate simulated sequencing data based on human genome assembly GRCh38 with varying sequencing depth, read length, duplication rate, sequencing error and base quality range [51]. Reference sequences were extracted from target region of the CTP

panel. Sequence reads for 5 samples were simulated using predicted mutations from PDX models of different cancer types from the CTP dataset to represent different spectrum of mutations, with a range of allele frequency to mimic germline and somatic mutations. For each simulated sample, we generated three technical replicates at 500X, 1000X and 1500X coverage.

Addition of mouse reads. Mouse sequencing reads were added in different fractions to the human-specific simulated dataset to mimic mouse contamination observed in PDX models. The mouse reads were extracted from the sequencing data of mouse DNA isolated from fresh spleen tissue of NSG mice on the CTP. For each simulated human-specific sample, we added mouse reads in three proportions (10, 15 and 25% of the total coverage).

Calculate sensitivity and specificity of mutation results based on different workflow filters. To evaluate the effect of each filter used in our workflow, we modified the somatic mutation workflow by: (i) omitting Xenome to filter mouse reads, and (ii) mapping to the reference sequence using BWA-MEM. Each modified workflow was used to process each PDX simulated library and each set of results, with and without quality filters, was used to compute the lists of true positive, false positive, true negative and false negative variants. As such, we can calculate the range of sensitivities and specificities of the predicted variants for all the simulated PDX models. We compared the distributions of precision, recall and F1-score (2*(Recall*Precision)/(Recall+Precision)) for different variations of the variant calling workflow on the simulated datasets. Furthermore, we compared the predicted allele frequencies of the true positives of each sample with the input by correlation.

RNA-Seq expression workflow

Data processing and expression estimation. Prior to alignment to the human transcriptome, sequences from PDX tumors were processed for sequence quality. Only sequences with base qualities ≥30 over 70 percent of read length were used in downstream analyses. Quality trimmed reads were then analyzed using the default parameters of Xenome v1.0.0 (k=25) [13] to separate human, mouse,

and ambiguous sequences (i.e., sequences that cannot be reliably classified as mouse or human). Sequence reads that passed the quality and Xenome screening were aligned to a human transcriptome dataset (ENSEMBL version GRCh38.84) using Bowtie v2.2.0 [52, 53]. Only samples with at least 1 million human reads were retained for expression analysis. Gene expression estimates were determined using RSEM v1.2.19 [54] (*rsem-calculate-expression*) with default parameters. We further normalized the expression estimate (expected_count from RSEM) using upper quantile normalization of non-zero expected counts and scaling to 1000.

Classifier for EBV-associated PDX lymphomas

A gene signature for differentiating EBV-associated lymphomas was derived from the most highly differentially expressed genes between 20 EBV-associated lymphomas and 100 non-EBV tumors based on upper-quantile normalized RNA-Seq counts (RSEM). Gene set analysis on the resulting expression vector was performed with GSEA using the GenePattern webserver and default parameters (data not shown). 24 up-regulated and 24 downregulated genes from the set of differentially expressed genes were used to define the list of classifier genes (Supplementary Table S10). For each PDX sample, the upper-quantile normalized counts from RSEM of the classifier genes were transformed into z-scores using the mean and standard deviation computed across all PDX samples for each gene. Subsequently, a sign corresponding to the direction of regulation in the classifier table was multiplied to each z-score and the sum of these modified z-scores resulted in a single score for each PDX sample. A classifier score of >3.0 was used to identify a PDX tumor sample as a potential EBV-associated lymphoma.

Copy Number Variant (CNV) workflow

Assessing the effects of mouse DNA on SNP array. DNA of the NSG mouse was hybridized on the Affymetrix SNP 6.0 array, and the signal intensity was extracted from the CEL files using Affymetrix Power Tools (apt-cel-extract). The mouse content for each PDX sample was estimated by the mouse reads proportion computed by Xenome of the mutation calling pipeline for the CTP sequencing of the same PDX sample.

Single-tumor CNV analysis. PennCNV-Affy and Affymetrix Power Tools [55-57] were used to extract the B-allele frequency (BAF) and Log R Ratio (LRR) from the resulting CEL files of the Affymetrix Human SNP 6.0 array. Due to the absence of paired-normal samples, the allele-specific signal intensity for each PDX tumor were normalized relative to 300 randomly selected sex-matched Affymetrix Human SNP 6.0 array samples obtained from the International HapMap project [58]. The single tumor version of ASCAT 2.4.3 [59] was then used for GC correction, predictions of the heterozygous germline SNPs and estimation of ploidy, tumor content and copy number segments with allele-specific copy number.

Annotation of CNV segments. The resultant copy number segments were annotated with loss of heterozygosity (LOH) and $\log_2$ ratio of total copy number relative to diploid state (copy number 2) and predicted ploidy from ASCAT. A segment was defined as LOH when the major-allele copy number was ≥ 0.5 and the minor-allele copy number was ≤ 0.1. Gene-level copy number and LOH were estimated by intersecting the genome coordinates of copy number segments with genome coordinates of genes (Ensembl annotation version 84 for genome assembly GRCh38). In cases where a segment boundary was contained within a gene's coordinates, the most conservative (lowest) estimate of copy number was used and the gene was annotated with the number of overlapping segments.

Defining copy number gain and loss. The low-level copy number gain or loss of a gene was defined by the $\log_2$ ratio of the copy number relative to the average ploidy of the sample or diploid state with a threshold of ±0.4 respectively. We compiled a list of genes with focal copy number aberrations that were significantly amplified (n=273) or deleted (n=820) in the 8 tumor types (Supplementary Table S8) from the GISTIC 2.0 analysis from the TCGA FireBrowse website (http://firebrowse.org/). Using this set of genes, we compared the proportion of genes that would be classified as gain and loss when using different baselines (diploid state 2 or ASCAT predicted ploidy) for PDX models listed in Supplementary Table S12.

Comparison of copy number aberrations with gene expression. Using annotations from the Cancer Census resource [33] we analyzed the relationship between copy number aberrations and gene expression using a list of 23 oncogenes that are commonly amplified in cancers and a list of 40 tumor suppressor genes that are commonly deleted in cancers. These genes were classified into copy number states of high-level loss (log2(CN/ploidy) < -1), normal (-1 ≤ $\log_2$(CN/ploidy) ≤ +1) and high-level gain ($\log_2$(CN/ploidy) > +1). The expression fold change of each gene was calculated as the $\log_2$(TPM+1) relative to the mean expression across PDX samples with a stringent normal copy number state (-0.4 ≤ log2(CN/ploidy) ≤ 0.4). The significance of expression changes of each gene for the entire PDX resource with copy number gain or loss relative to the normal state was calculated using the Student's t-Test.

Comparison between PDX and TCGA data

Somatic mutations. We calculated the distribution of mutational load (number of non-silent, coding mutations in exonic regions per sample) of the CTP genes for 6 tumor types with at least 10 models in the PDX resource (colorectal cancer, lung adenocarcinoma, lung squamous cell carcinoma, melanoma, bladder carcinoma and triple-negative breast cancer , Supplementary Table S5). MAF files for somatic mutations based on whole-exome sequencing of the TCGA samples of 6 tumor types [60-64] were obtained from TCGA Data Portal and were used to compute the mutation frequency for CTP genes only. The Fisher's exact test was used to test the significance of overlap of mutated genes between the PDX resource and TCGA patient cohorts for each tumor type. The genes in each PDX resource were considered if they were mutated in at least one sample, while the genes in each TCGA tumor cohort were considered if they were mutated with at least 5% frequency, due to a much larger sample size.

RNA-Seq gene expression. 6 tumor types with at least 10 models in the PDX resource were selected for comparison with TCGA (colorectal cancer, lung adenocarcinoma, lung squamous cell carcinoma, melanoma, bladder carcinoma and triple-negative breast cancer, Supplementary Table S10). The scaled estimate

(TPM $\times$ 10^{-6}) from the RNA-Seq data of 6 tumor types in TCGA [60-65] were obtained from the TCGA FireBrowse website (http://firebrowse.org/). Non-expressed genes across all tumor types were removed ($\log_2$(TPM+1) < 2), and the top 1000 most varying genes based on their z-scores of $\log_2$(TPM+1) across all tumor types were selected to cluster the samples by hierarchical clustering. The frequencies of over-expression and under-expression of each gene is defined by the z-scores of $\log_2$(TPM+1) of $\pm$1. Correlation of the gene expression frequencies in each tumor type was computed using Pearson correlation. The differential gene expression of each tumor type compared to all other tumor types was computed using limma [66] based on $\log_2$(TPM+1) values. Up-regulated (adjusted p-value < 0.05, log (fold change of TPM+1) > 1 by limma) or down-regulated (adjusted p-value < 0.05, log (fold change of TPM+1) < -1 by limma) genes were obtained for the PDX resource and TCGA patient cohorts separately. The significance of overlap of each set of genes between PDX and TCGA RNA-Seq data was determined using hypergeometric p-value.

Copy number aberrations. 8 tumor types with at least 10 models in the PDX resource (colorectal cancer, lung adenocarcinoma, lung squamous cell carcinoma, melanoma, glioblastoma multiforme, bladder carcinoma, triple-negative breast cancer and ovarian carcinoma, Supplementary Table S12) selected to compare with corresponding primary tumors in the TCGA [60-65, 67-69]. For PDX samples, the low-level copy number gain or loss of a gene was defined by the $\log_2$ ratio of the copy number relative to the average ploidy of the sample (or copy number state 2) with a threshold of $\pm$0.4 respectively. The amplification or deletion calls of each gene for the TCGA samples were provided (loss=-1, normal=0, gain=1) by FireBrowse (http://firebrowse.org/). Using the list of genes with focal copy number aberrations that were significantly amplified (n=273) or deleted (n=820) in the 8 tumor types from the GISTIC 2.0 analysis from the TCGA FireBrowse website, we calculated the copy number gain and loss frequencies of these genes for each tumor type in the PDX resource and TCGA cohorts using the respective gain and loss calls.

References

1. Hidalgo, M., et al., *Patient-derived xenograft models: an emerging platform for translational cancer research.* Cancer Discov, 2014. **4**(9): p. 998-1013.
2. Whittle, J.R., et al., *Patient-derived xenograft models of breast cancer and their predictive power.* Breast Cancer Res, 2015. **17**: p. 17.
3. Byrne, A.T., et al., *Interrogating open issues in cancer precision medicine with patient-derived xenografts.* Nat Rev Cancer, 2017.
4. Day, C.P., G. Merlino, and T. Van Dyke, *Preclinical mouse cancer models: a maze of opportunities and challenges.* Cell, 2015. **163**(1): p. 39-53.
5. Tentler, J.J., et al., *Patient-derived tumour xenografts as models for oncology drug development.* Nat Rev Clin Oncol, 2012. **9**(6): p. 338-50.
6. Bruna, A., et al., *A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds.* Cell, 2016. **167**(1): p. 260-274 e22.
7. Krepler, C., et al., *A Comprehensive Patient-Derived Xenograft Collection Representing the Heterogeneity of Melanoma.* Cell Rep, 2017. **21**(7): p. 1953-1967.
8. Garralda, E., et al., *Integrated next-generation sequencing and avatar mouse models for personalized cancer treatment.* Clin Cancer Res, 2014. **20**(9): p. 2476-84.
9. Reyal, F., et al., *Molecular profiling of patient-derived breast cancer xenografts.* Breast Cancer Res, 2012. **14**(1): p. R11.
10. Gao, H., et al., *High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response.* Nat Med, 2015. **21**(11): p. 1318-25.
11. Dong, G., et al., *Integrative analysis of copy number and transcriptional expression profiles in esophageal cancer to identify a novel driver gene for therapy.* Sci Rep, 2017. **7**: p. 42060.
12. Menghi, F., et al., *The tandem duplicator phenotype as a distinct genomic configuration in cancer.* Proc Natl Acad Sci U S A, 2016. **113**(17): p. E2373-82.
13. Conway, T., et al., *Xenome--a tool for classifying reads from xenograft samples.* Bioinformatics, 2012. **28**(12): p. i172-8.
14. Tso, K.Y., et al., *Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts?* BMC Genomics, 2014. **15**: p. 1172.
15. Rossello, F.J., et al., *Next-generation sequence analysis of cancer xenograft models.* PLoS One, 2013. **8**(9): p. e74432.
16. Jones, S., et al., *Personalized genomic analyses for cancer mutation discovery and interpretation.* Sci Transl Med, 2015. **7**(283): p. 283ra53.
17. Hiltemann, S., et al., *Discriminating somatic and germline mutations in tumor DNA samples without matching normals.* Genome Res, 2015. **25**(9): p. 1382-90.
18. Sandmann, S., et al., *Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data.* Sci Rep, 2017. **7**: p. 43169.
19. Hsu, Y.C., et al., *Detection of Somatic Mutations in Exome Sequencing of Tumor-only Samples.* Sci Rep, 2017. **7**(1): p. 15959.

20. Reumers, J., et al., *Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing.* Nat Biotechnol, 2011. **30**(1): p. 61-8.
21. Hofmann, A.L., et al., *Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers.* BMC Bioinformatics, 2017. **18**(1): p. 8.
22. Hwang, S., et al., *Systematic comparison of variant calling pipelines using gold standard personal exome variants.* Sci Rep, 2015. **5**: p. 17875.
23. Choi, Y.Y., et al., *Establishment and characterisation of patient-derived xenografts as paraclinical models for gastric cancer.* Sci Rep, 2016. **6**: p. 22172.
24. Zhang, L., et al., *The extent of inflammatory infiltration in primary cancer tissues is associated with lymphomagenesis in immunodeficient mice.* Sci Rep, 2015. **5**: p. 9447.
25. Bondarenko, G., et al., *Patient-Derived Tumor Xenografts Are Susceptible to Formation of Human Lymphocytic Tumors.* Neoplasia, 2015. **17**(9): p. 735-741.
26. Butler, K.A., et al., *Prevention of Human Lymphoproliferative Tumor Formation in Ovarian Cancer Patient-Derived Xenografts.* Neoplasia, 2017. **19**(8): p. 628-636.
27. Dieter, S.M., et al., *Patient-derived xenografts of gastrointestinal cancers are susceptible to rapid and delayed B-lymphoproliferation.* Int J Cancer, 2017. **140**(6): p. 1356-1363.
28. Ahdesmaki, M.J., et al., *Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples.* F1000Res, 2016. **5**: p. 2741.
29. De Summa, S., et al., *GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data.* BMC Bioinformatics, 2017. **18**(Suppl 5): p. 119.
30. Li, H., *Toward better understanding of artifacts in variant calling from high-coverage samples.* Bioinformatics, 2014. **30**(20): p. 2843-51.
31. Patterson, S.E., et al., *The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies.* Hum Genomics, 2016. **10**: p. 4.
32. Ye, K., et al., *Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.* Bioinformatics, 2009. **25**(21): p. 2865-71.
33. Futreal, P.A., et al., *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-83.
34. Ohshima, K., et al., *Integrated analysis of gene expression and copy number identified potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors.* Sci Rep, 2017. **7**(1): p. 641.
35. Jabs, V., et al., *Integrative analysis of genome-wide gene copy number changes and gene expression in non-small cell lung cancer.* PLoS One, 2017. **12**(11): p. e0187246.
36. Yates, L.R., et al., *Genomic Evolution of Breast Cancer Metastasis and Relapse.* Cancer Cell, 2017. **32**(2): p. 169-184 e7.
37. Robinson, D.R., et al., *Integrative clinical genomics of metastatic cancer.* Nature, 2017. **548**(7667): p. 297-303.

38. Oh, B.Y., et al., *Correlation between tumor engraftment in patient-derived xenograft models and clinical outcomes in colorectal cancer patients.* Oncotarget, 2015. **6**(18): p. 16059-68.
39. Moon, H.G., et al., *Prognostic and functional importance of the engraftment-associated genes in the patient-derived xenograft models of triple-negative breast cancers.* Breast Cancer Res Treat, 2015. **154**(1): p. 13-22.
40. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
41. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
42. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
43. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
44. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11 10 1-33.
45. Ananda, G., et al., *Development and validation of the JAX Cancer Treatment Profile for detection of clinically actionable mutations in solid tumors.* Exp Mol Pathol, 2015. **98**(1): p. 106-12.
46. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
47. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.
48. Kitts, A., et al., *The Database of Short Genetic Variation (dbSNP)*, in *The NCBI Handbook [Internet]*. 2014: Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
49. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
50. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
51. Shifu, C., et al. *SeqMaker: A next generation sequencing simulator with variations, sequencing errors and amplification bias integrated*. in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2016.
52. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
53. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
54. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
55. Wang, K., et al., *PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.* Genome Res, 2007. **17**(11): p. 1665-74.

56. Diskin, S.J., et al., *Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms.* Nucleic Acids Res, 2008. **36**(19): p. e126.
57. Wang, K., et al., *Modeling genetic inheritance of copy number variations.* Nucleic Acids Res, 2008. **36**(21): p. e138.
58. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
59. Van Loo, P., et al., *Allele-specific copy number analysis of tumors.* Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16910-5.
60. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
61. Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
62. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.
63. Cancer Genome Atlas, N., *Genomic Classification of Cutaneous Melanoma.* Cell, 2015. **161**(7): p. 1681-96.
64. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
65. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of urothelial bladder carcinoma.* Nature, 2014. **507**(7492): p. 315-22.
66. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
67. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
68. Brennan, C.W., et al., *The somatic genomic landscape of glioblastoma.* Cell, 2013. **155**(2): p. 462-77.
69. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.

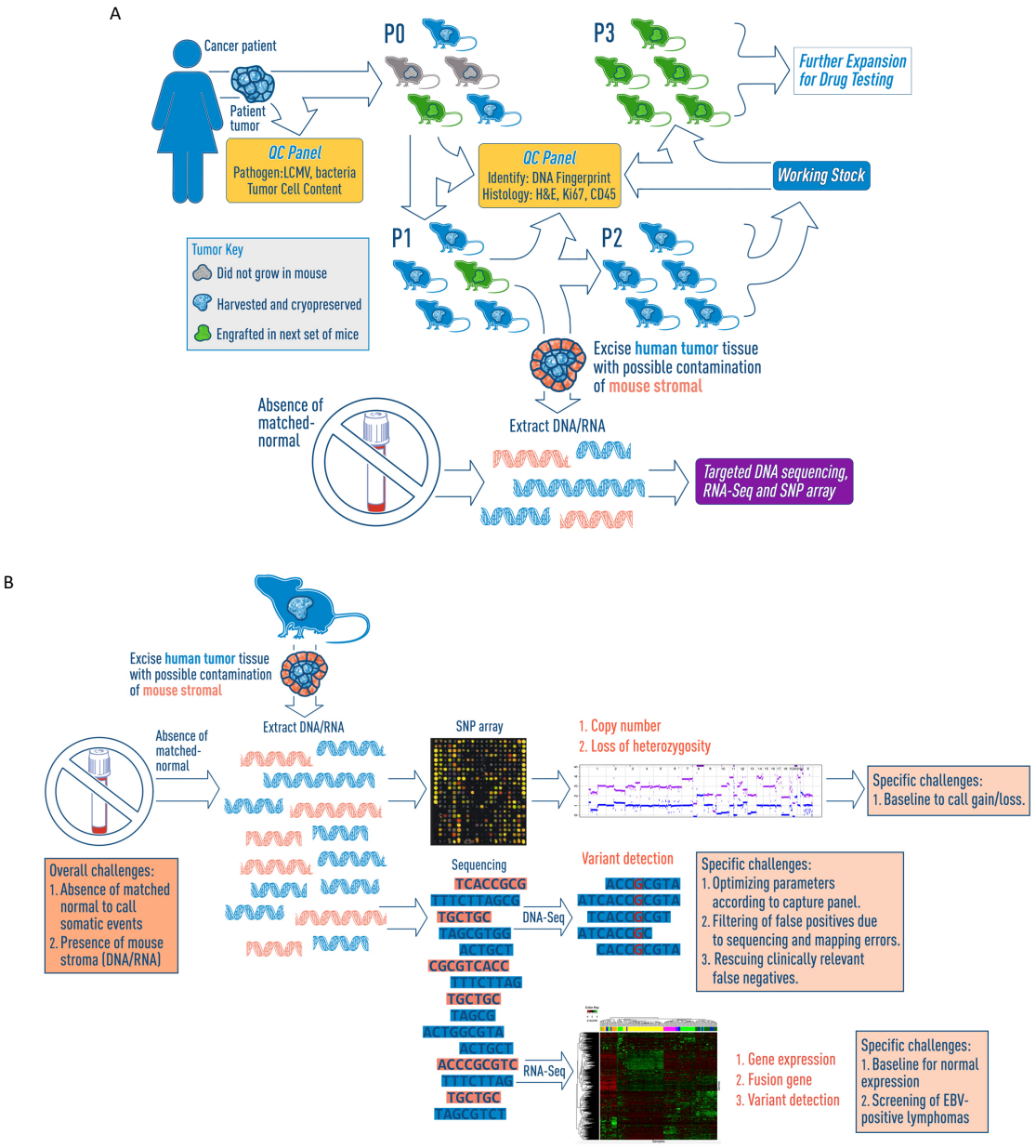


Figure 1

(A) The Jackson Laboratory (JAX) has generated, clinically annotated, and genomically characterized more than 450 patient-derived xenograft (PDX) cancer models from about 20 different types of cancer using the immunodeficient NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (aka, NSG^{TM}) mouse as the host strain (http://tumor.informatics.jax.org/mtbwi/index.do). This figure shows the workflow of PDX model generation from patient tumor, the process of engraftment and passaging that supplies to the JAX PDX resource, and the generation of genomic and transcriptomic data to profile the PDX models.

(B) The PDX models are profiled by: 1) DNA mutations from capture sequencing using the JAX Cancer Treatment Profile^{TM} (CTP, https://www.jax.org/clinical-genomics/clinical-offerings/jax-cancer-treatment-profile), the Illumina Truseq panel or whole-exome sequencing, 2) DNA copy-number variations using Affymetrix SNP 6.0 arrays, and 3) gene expression profiles from Affymetrix microarrays or RNA sequencing (Illumina HiSeq). The analysis of the genomic and transcriptomic data of PDX models poses several challenges which we have developed several strategies to circumvent these issues.

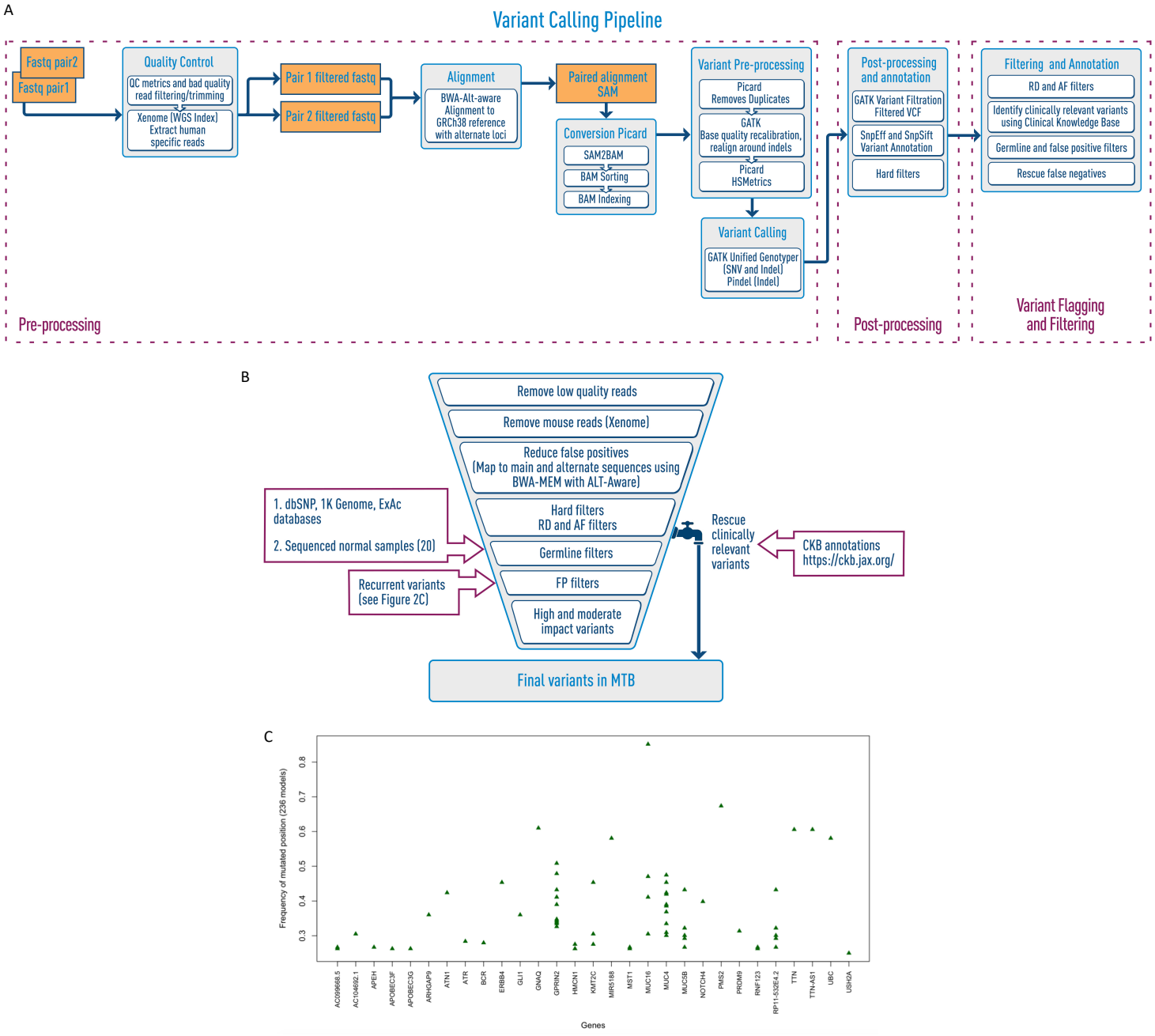
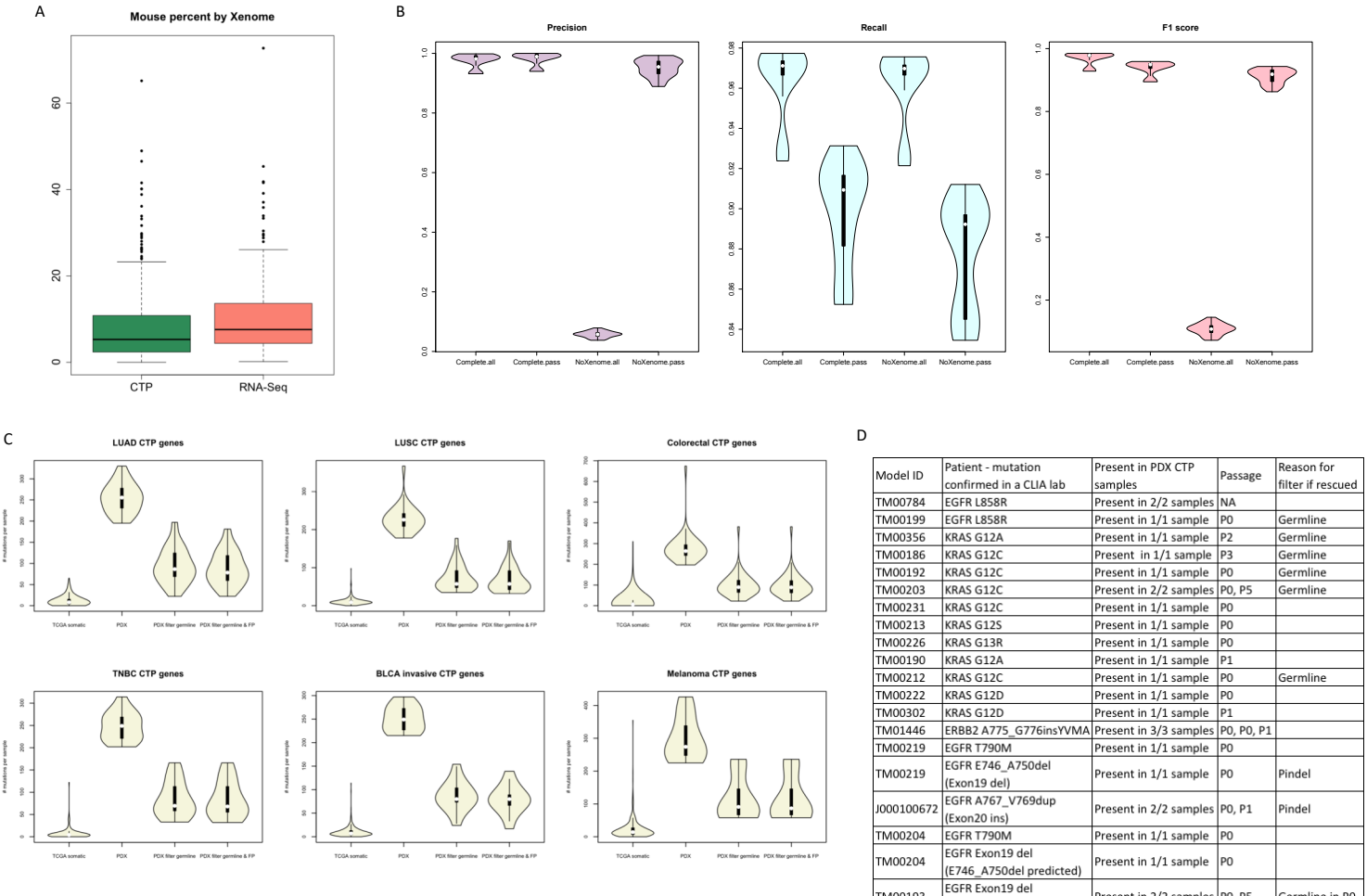


Figure 2

(A) This flow chart describes the variant calling pipeline for PDX DNA sequencing data.

(B) This figure shows the different filters used the variant calling pipeline for PDX DNA sequencing data applied to the CTP panel sequencing (see Methods for details). MTB is the Mouse Tumor Biology Database in JAX, PDX models in the JAX PDX resource can be searched in http://tumor.informatics.jax.org/mtbwi/pdxSearch.do. (RD: Read depth, AF: Allele-frequency, FP: False positives

(C) The recurrent frequencies of the mutated positions (after germline filtering) for various genes that were found to be recurrent in more than 25% of PDX samples. These were identified as additional false positive variants due to sequencing errors or mapping issues.



D

Model ID	Patient - mutation confirmed in a CLIA lab	Present in PDX CTP samples	Passage	Reason for filter if rescued
TM00784	EGFR L858R	Present in 2/2 samples	NA	
TM00199	EGFR L858R	Present in 1/1 sample	P0	Germline
TM00356	KRAS G12A	Present in 1/1 sample	P2	Germline
TM00186	KRAS G12C	Present in 1/1 sample	P3	Germline
TM00192	KRAS G12C	Present in 1/1 sample	P0	Germline
TM00203	KRAS G12C	Present in 2/2 samples	P0, P5	Germline
TM00231	KRAS G12C	Present in 1/1 sample	P0	
TM00213	KRAS G12S	Present in 1/1 sample	P0	
TM00226	KRAS G13R	Present in 1/1 sample	P0	
TM00190	KRAS G12A	Present in 1/1 sample	P1	
TM00212	KRAS G12C	Present in 1/1 sample	P0	Germline
TM00222	KRAS G12D	Present in 1/1 sample	P0	
TM00302	KRAS G12D	Present in 1/1 sample	P1	
TM01446	ERBB2 A775_G776insYVMA	Present in 3/3 samples	P0, P0, P1	
TM00219	EGFR T790M	Present in 1/1 sample	P0	
TM00219	EGFR E746_A750del (Exon19 del)	Present in 1/1 sample	P0	Pindel
J000100672	EGFR A767_V769dup (Exon20 ins)	Present in 2/2 samples	P0, P1	Pindel
TM00204	EGFR T790M	Present in 1/1 sample	P0	
TM00204	EGFR Exon19 del (E746_A750del predicted)	Present in 1/1 sample	P0	
TM00193	EGFR Exon19 del (E746_A750del predicted)	Present in 2/2 samples	P0, P5	Germline in P0

Figure 3

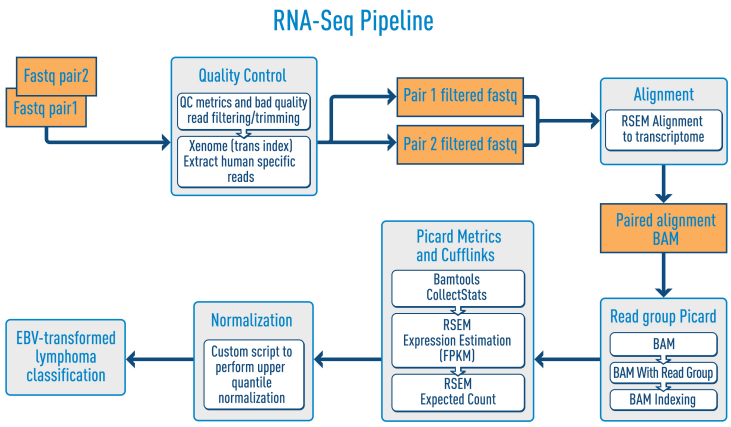
(A) Proportion of mouse reads detected by Xenome for CTP and RNA sequencing data of PDX models.

(B) This figure shows the benchmarking of the CTP variant calling pipeline using 45 simulated sequencing datasets different samples, sequencing coverages, and mouse DNA content (see Supplementary Table S2) using precision, recall and F1 score based on the input variants for each sample. Complete: variant calling pipeline with all steps included; NoXenome: variant calling pipeline with Xenome omitted; all: all variants called by the pipeline; pass: variants annotated as "PASS" in the pipeline which pass the hard filters, minimum read depth and minimum alternate allele frequency of the variant.

(C) Distribution of mutational load per sample of non-silent coding somatic mutations of CTP genes from exome sequencing TCGA samples and from CTP-panel sequencing of PDX models. TCGA somatic: TCGA somatic mutations reported in maf files; PDX: all variants annotated as "PASS" (pass the hard filters, minimum read depth and minimum alternate allele frequency of the variant); PDX filter germline: all variants annotated as "PASS" and filtered from putative germline variants; PDX filter germline & FP: all variants annotated as "PASS" and filtered from putative germline variants and false positives.

(D) Mutations in PDX models that were detected by CTP-panel sequencing and experimentally validated in the corresponding patient tumor. Some of these variants were rescued after initial filtering.

A



B

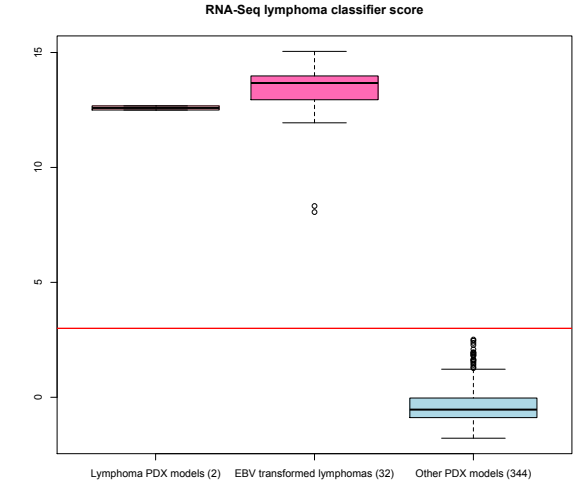


Figure 4

(A) This flow chart describes the RNA expression pipeline and fusion gene prediction for PDX RNA sequencing data.

(B) Distribution of lymphoma classification scores of PDX tumors.

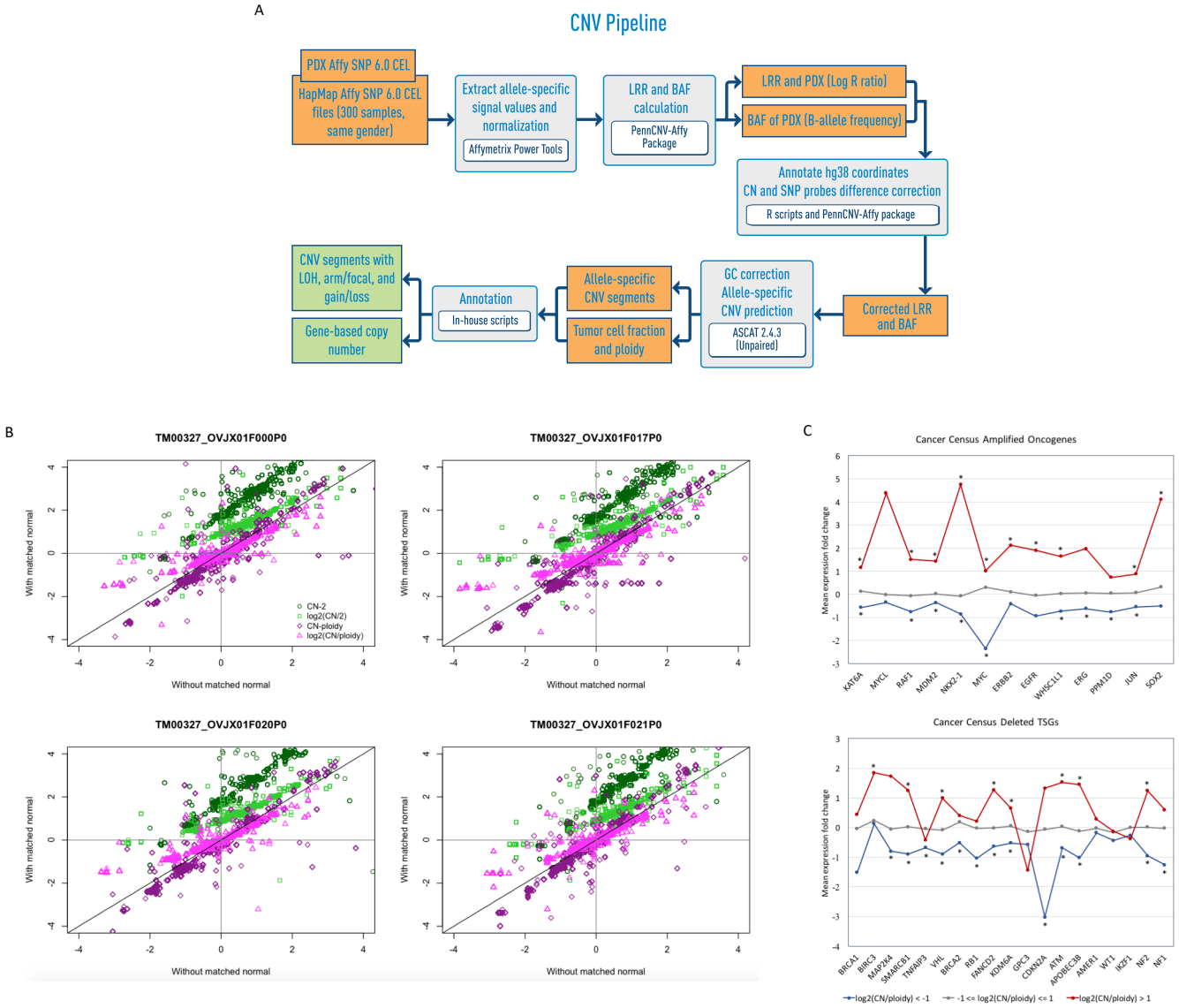


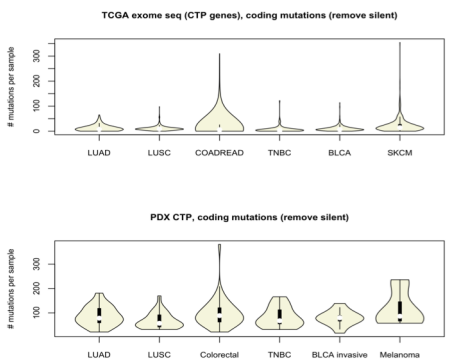
Figure 5

(A) This flow chart describes the CNV and LOH prediction pipeline for PDX SNP array data.

(B) Comparison of copy number relative to the estimated overall ploidy of the PDX sample or the diploid state between analyses with and without matched normal.

(C) Mean expression fold change of genes with copy number normal, gain and loss state for a selected list of known oncogenes that are amplified in cancers and known tumor suppressor genes that are deleted in cancers from the Cancer Census [34]. Overexpressed and under-expressed genes marked with * indicates significant differences in expression fold change with copy number gain or loss state respectively relative to the normal state across all PDX samples.

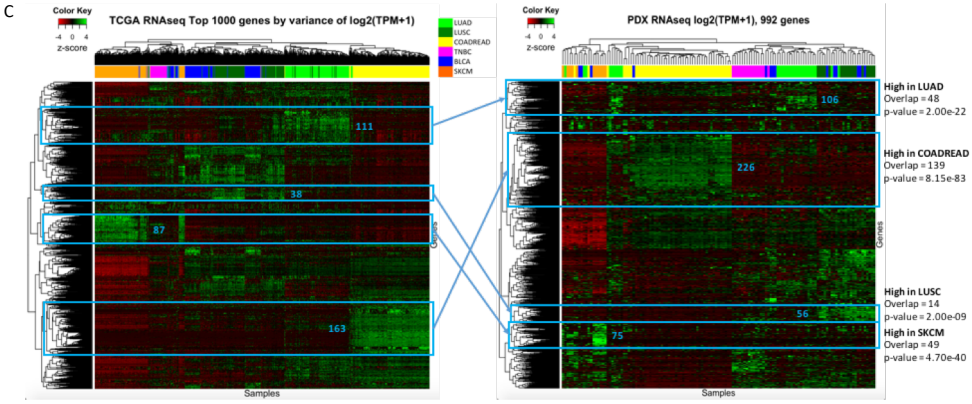
A



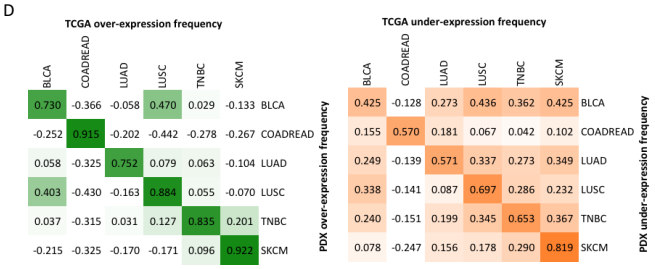
B

Tumor type	Fisher's exact test	
	p-value	Odds ratio
Lung adenocarcinoma	1.46E-05	18.5
Lung squamous cell carcinoma	1.42E-05	19.2
Colorectal cancer	3.61E-05	Inf
Triple negative breast cancer	8.88E-05	17.2
Bladder cancer, muscle invasive	1.25E-06	6.7
Melanoma	5.83E-14	14.3

C



D



E

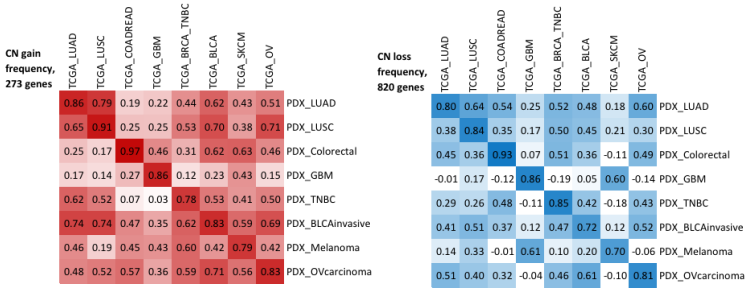


Figure 6

(A) Distribution of mutational load per sample of non-silent coding somatic mutations of CTP genes from exome sequencing of TCGA samples and from CTP-panel sequencing of PDX models (all filters included).

(B) Overlap of CTP genes that have non-silent coding somatic mutations with >5% mutation frequency in TCGA data with genes that have at least one non-silent coding somatic mutation in PDX CTP data (all filters and rescue of clinically relevant variants included) for each tumor type. Fisher's exact test is used to compute the significance of the overlap.

(C) Hierarchical clustering of z-score of expression (log_2(TPM+1)) of top 1000 most varying genes of TCGA RNA-Seq samples across different tumor types. The same set of genes (omitting non-expressed genes) is used to cluster the expression z-score by Hierarchical clustering of PDX RNA-Seq models across different tumor types. Gene sets were identified to be high expression in specific tumor types TCGA and PDX separately and were found to share significant overlap.

(D) Correlation frequency of genes that are over-expressed (z-score of log_2(TPM+1) > 1, green) or under-expressed (z-score of log_2(TPM+1) < -1, orange) across each tumor type between PDX models and TCGA samples.

(E) Correlation of frequency of copy number gain (red) or loss (blue) of selected genes frequently amplified or deleted in TCGA tumors predicted by GISTIC analysis for each tumor type between PDX and TCGA datasets.

Supplementary Figures and Tables

Bioinformatics workflows for genomic analysis of tumors from Patient Derived Xenografts (PDX): challenges and guidelines

Xing Yi Woo[*,1], Anuj Srivastava[*,1], Joel H. Graber[*,3], Vinod Yadav[1], Vishal Kumar Sarsani[2], Al Simons[2], Glen Beane[2], Stephen Grubb[2], Guruprasad Ananda[1], Rangjiao Liu[1], Grace Stafford[2], Jeffrey H. Chuang[1], Susan D. Airhart[2], R. Krishna Murthy Karuturi[1], Joshy George[#,1], Carol J. Bult[#,2]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA

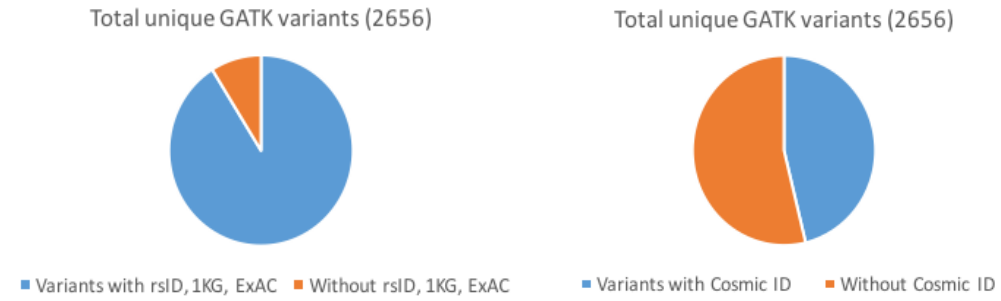
[2] The Jackson Laboratory, Bar Harbor, ME 04609, USA

[3] MDI Biological Laboratory, Bar Harbor, ME 04609, USA

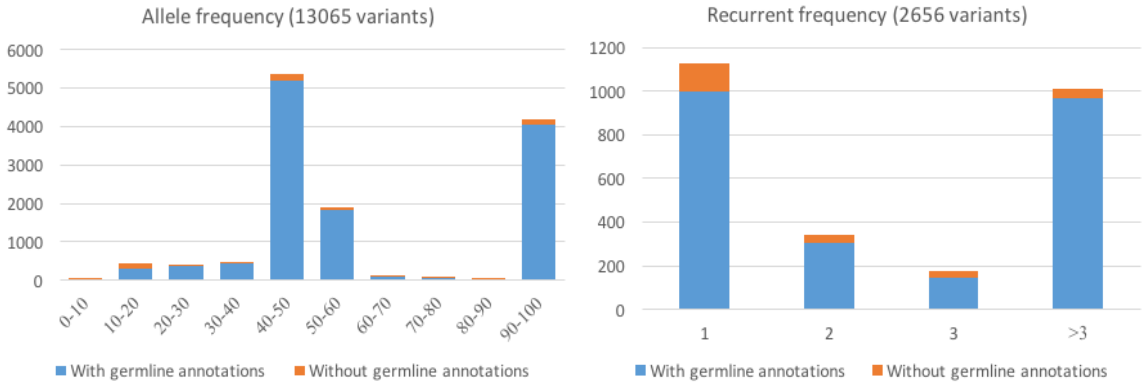
[*] The authors contributed equally to this work.

[#] Corresponding authors: Carol J. Bult (Email: Carol.Bult@jax.org), Joshy George (Email: Joshy.George@jax.org)

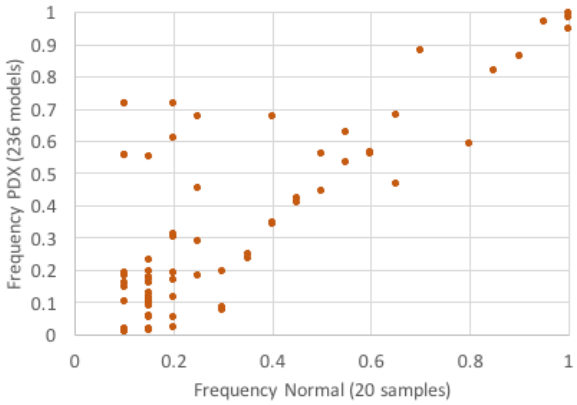
A



B



C



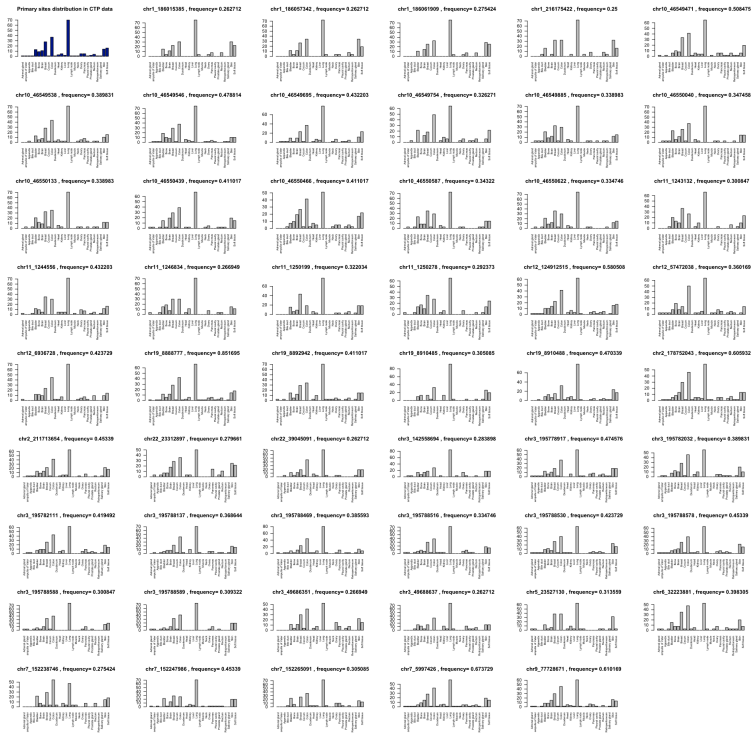
Supplementary Figure S1

(A) This figure shows the annotation of variants of the 20 normal samples in JAX using public databases (dbSNP Build 144, 1000 Genomes, ExAC version 0.3, and COSMIC version 80)

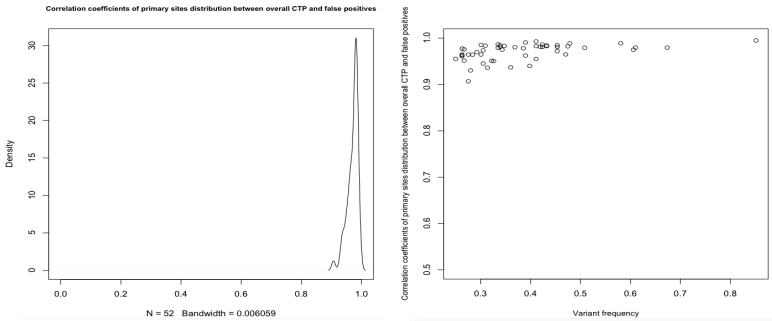
(B) The allele frequencies and recurrent frequencies of the variants.

(C) Recurrent frequency of variants (> 1 sample) found in 20 normal samples and the corresponding recurrent frequency across 236 PDX models of different tumor types.

A



B

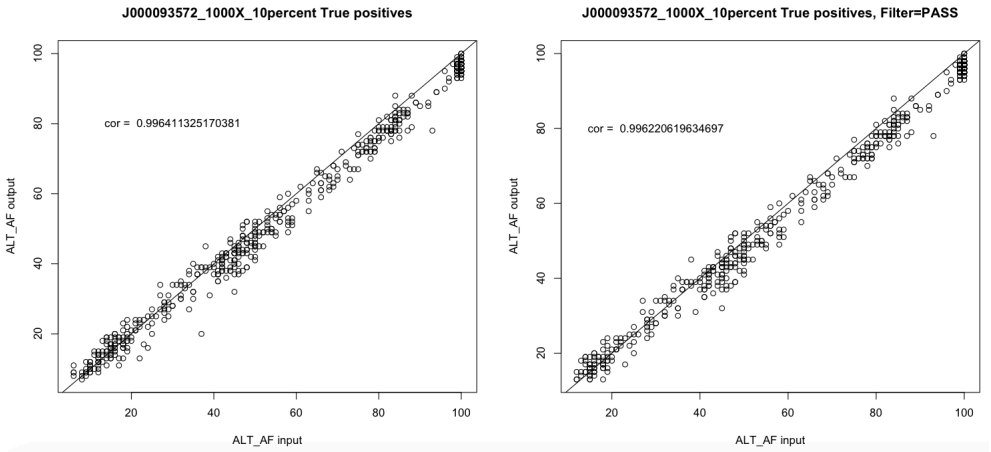


Supplementary Figure S2

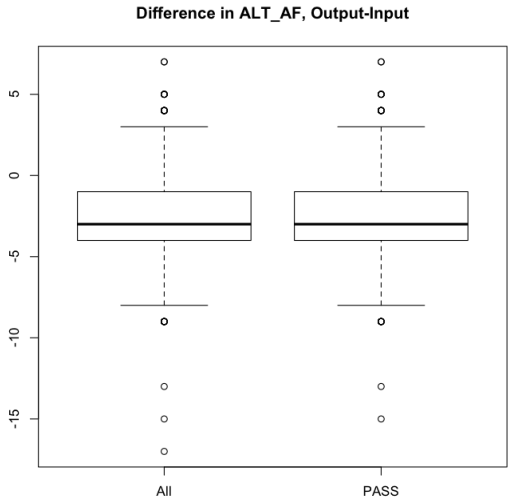
(A) Top left (dark blue): Distribution of tumor types across 236 PDX models; Others (grey): Frequency of tumor types for each frequently mutated position normalized by the recurrent frequency of the mutation.

(B) Correlations of tumor type frequency for the 236 PDX models (dark blue in B) with the tumor type frequency for each frequently mutated position (grey in B).

A



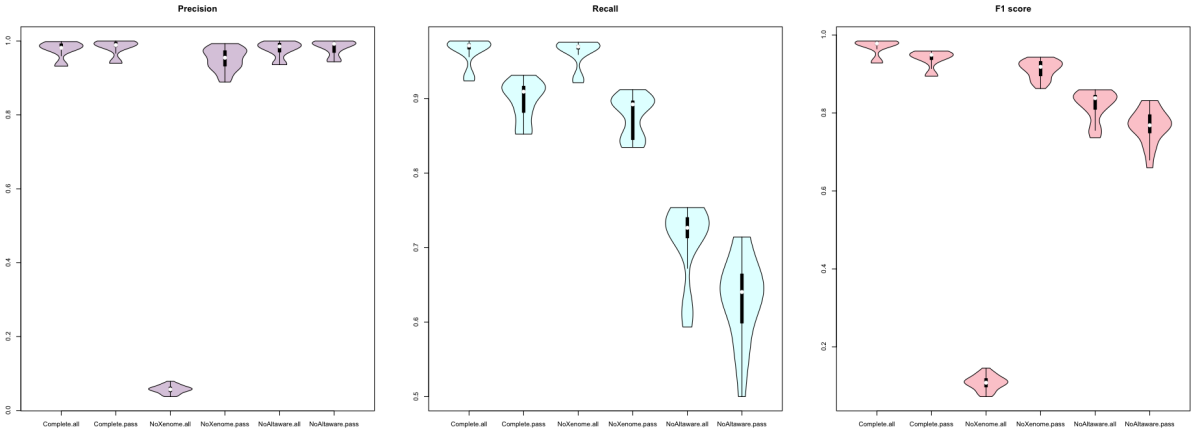
B



Supplementary Figure S3

(A) Correlation of alternate allele frequencies between input and true positive variants for one of the simulated samples for the complete feature pipeline. ALL: all variants called by the pipeline; PASS: variants annotated as “PASS” in the pipeline which pass the hard filters, minimum read depth and minimum alternate allele frequency of the variant. The correlation coefficient for all simulated samples are found in Supplementary Table S3.

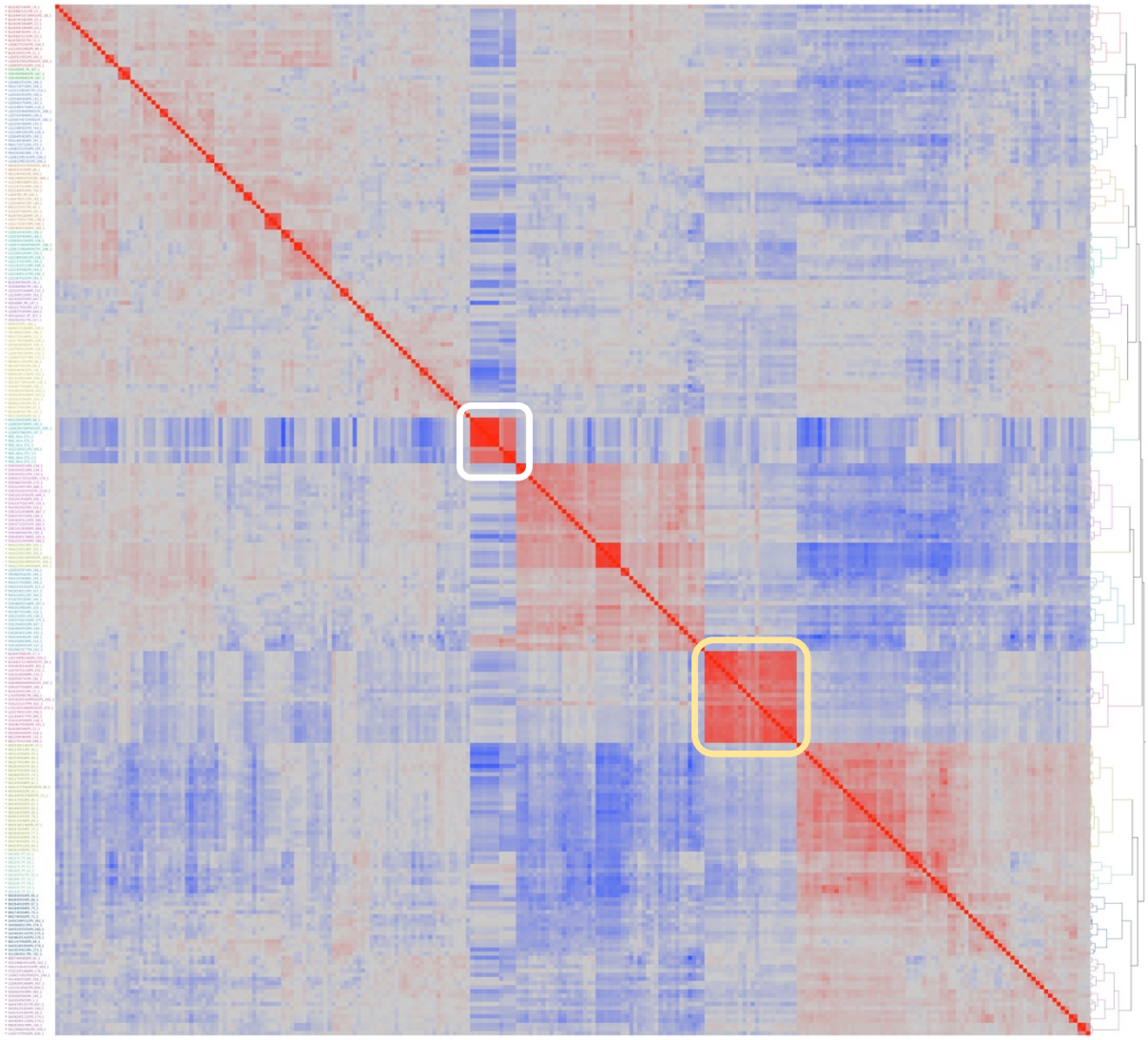
(B) Difference in alternate allele frequencies between input and true positive variants for one of the simulated samples (J000093572_1000X_10percent) for the complete feature pipeline. ALL: all variants called by the pipeline; PASS: variants annotated as “PASS” in the pipeline which pass the hard filters, minimum read depth and minimum alternate allele frequency of the variant.



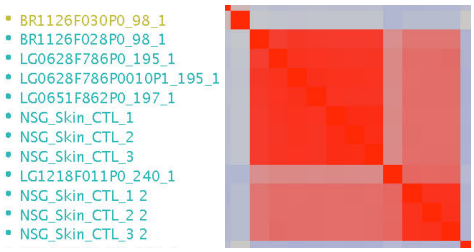
Supplementary Figure S4
This figure shows the benchmarking of the CTP variant calling pipeline using 45 simulated sequencing datasets different samples, sequencing coverages, and mouse DNA content (see Supplementary Table S2) using precision, recall and F1 score based on the input variants for each sample. Complete: variant calling pipeline with all steps included; NoXenome: variant calling pipeline with Xenome omitted; NoAltaware: variant calling pipeline using hg38 reference with alternate sequences but using standard BWA for mapping instead of BWA-ALT-Aware; all: all variants called by the pipeline; pass: variants annotated as "PASS" in the pipeline which pass the hard filters, minimum read depth and minimum alternate allele frequency of the variant.

Presence of alternate loci in the genome assembly. The GRCh38.p5 human genome assembly includes 262 regions of alternate loci to account for human chromosomal regions that exhibit sufficient variability to prevent adequate representation by a single sequence [29]. As such, we aligned the reads to both primary and alternate chromosomal reference sequences using BWA-MEM with ALT-aware. When alignment is performed using BWA-MEM only, the recall of the variants is much lower (~30%) than the standard pipeline with or without hard-filtering (Supplementary Figure S14 and Supplementary Table S2). This shows that using an alignment tool not catered for alternate loci mapping reduces the overall sensitivity of the variant calling due to lesser reads being correctly mapped. The correlation of allele frequencies also decreases and the reduction in median allele frequency increases up to 15% (Supplementary Table S3).

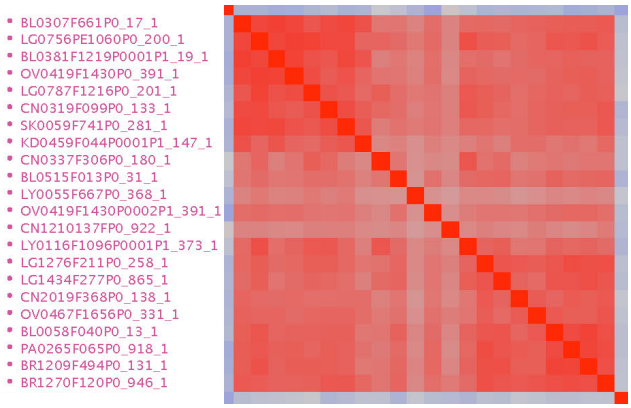
A



B



C



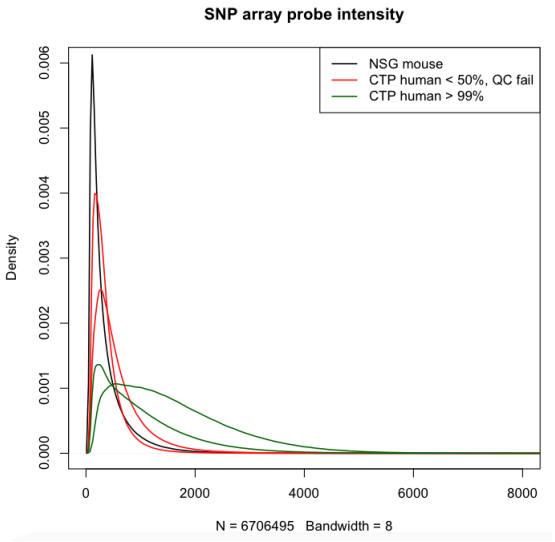
Supplementary Figure S5

(A) Hierarchical clustering of the pairwise correlations of microarray expression between pairs of models, with red representing perfect correlation (+1), and blue perfect anti-correlation (-1). The two largest red blocks (highlighted in white and yellow) show the mouse introgressed and EBV transformed models. The other blocks, which much lower average correlation, typically show related tumor types (e.g., the lower right block is all neurological tumors).

(B) A small fraction of tumors, highlighted in white in (A), that were heavily introgressed by mouse tissues were clustered with expression of NSG mouse (skin sample).

(C) EBV Lymphoma models, highlight in yellow in (A), show an extremely highly correlated expression pattern regardless of the original tissue or tumor type.

A



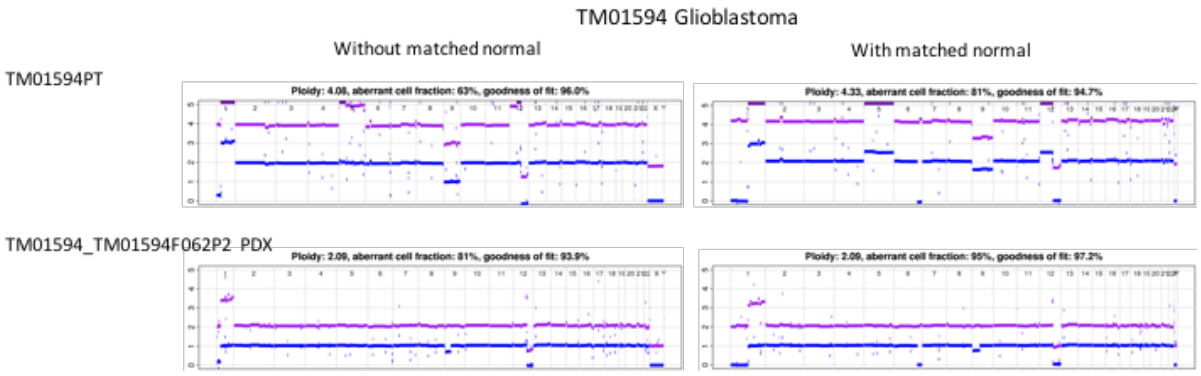
B

Supplementary Figure S6

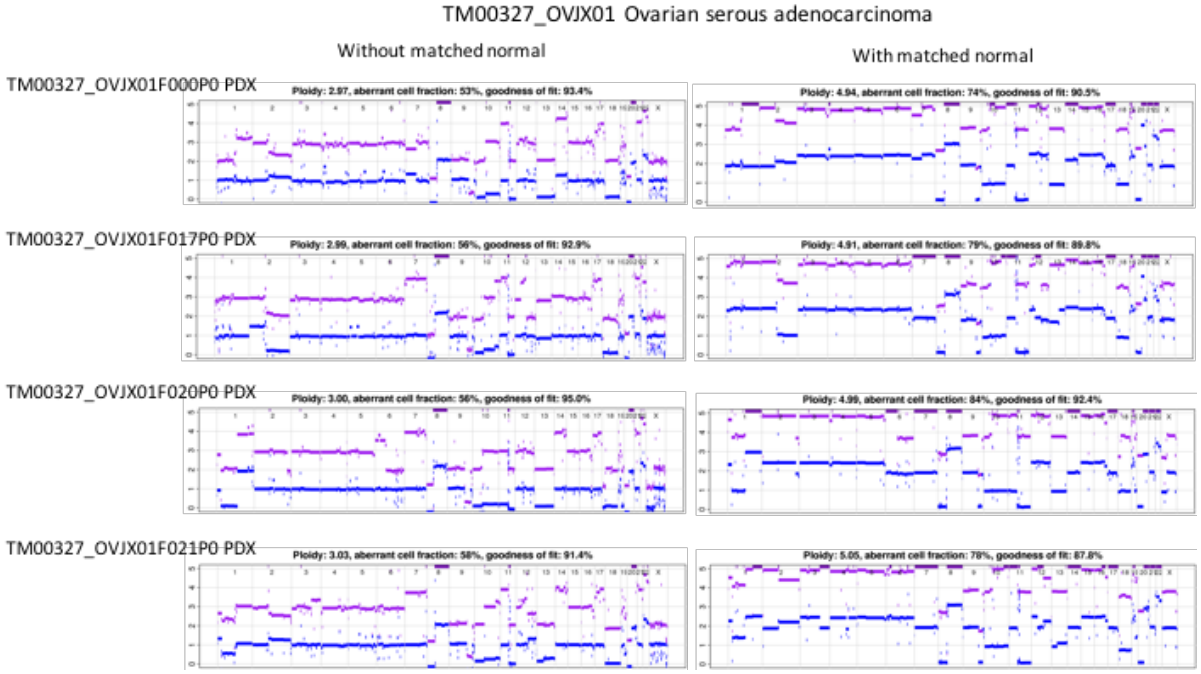
(A) Distribution of probe intensity of the SNP array for PDX samples with different mouse DNA content (using the percentage of human reads estimated from the CTP sequencing as a proxy for human DNA content on the SNP array): > 99% human DNA (green), < 50% human DNA with QC failure of SNP array CEL file (red), and 100% NSG mouse DNA (black).

(B) Human DNA content of PDX samples classified by successful CNV prediction (black squares), failure in QC of CEL files (orange circles), and failure in ASCAT analysis red triangles).

A

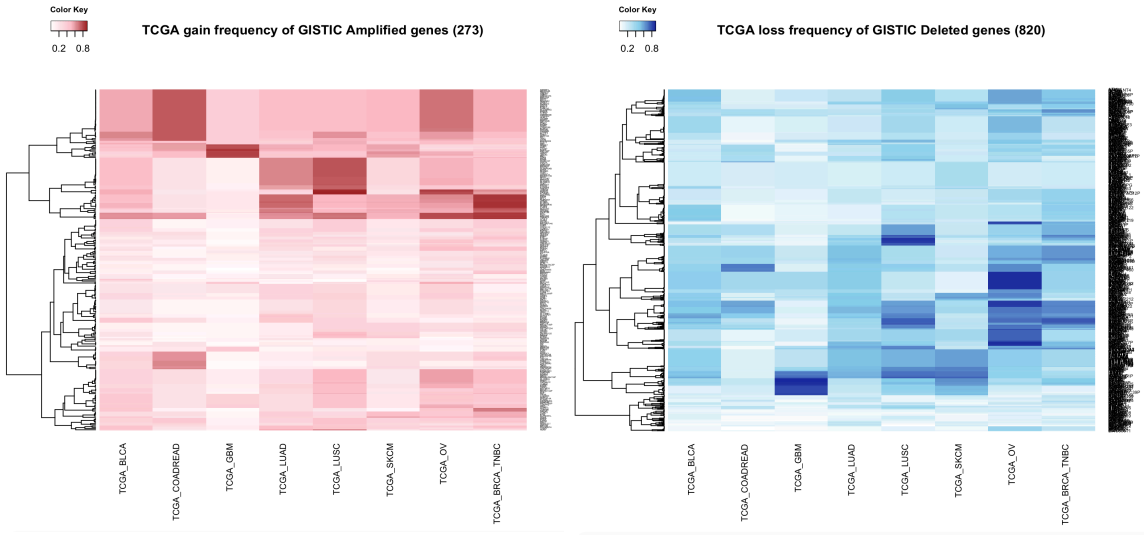


B



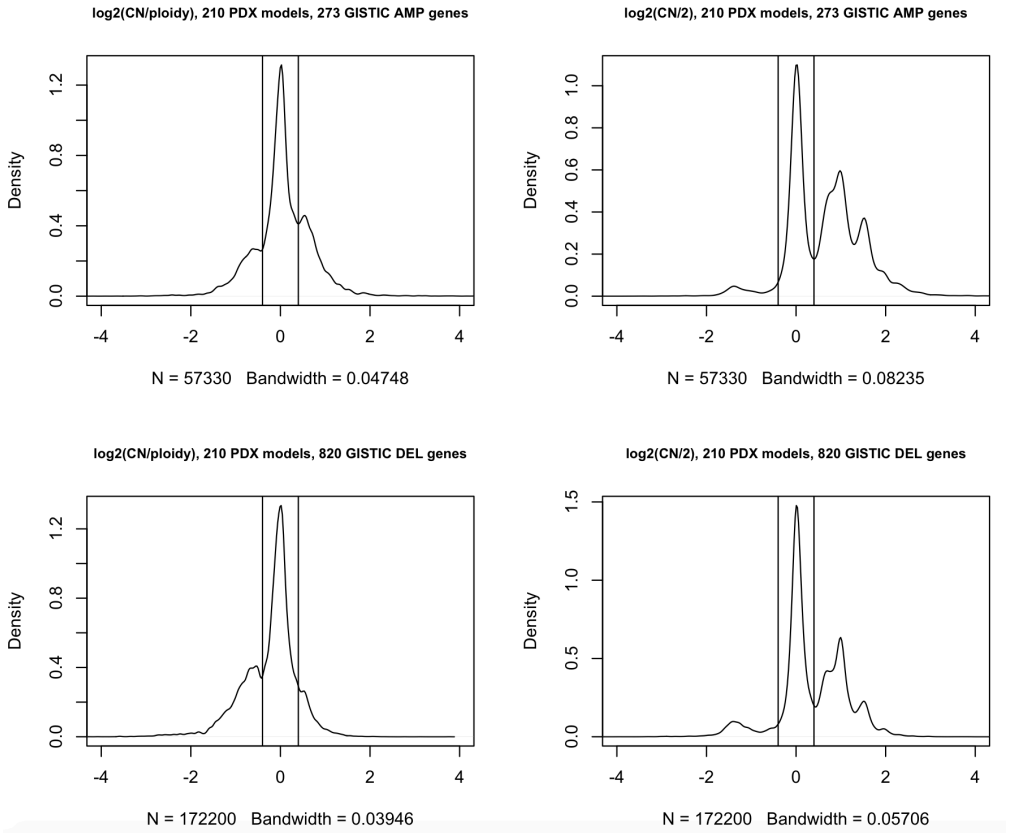
Supplementary Figure S7

(A) And (B): CNV profiles of PDX models with matched models and multiple samples from the corresponding patient tumor, multiple passages or multiple samples (mouse) of same passage.



Supplementary Figure S8

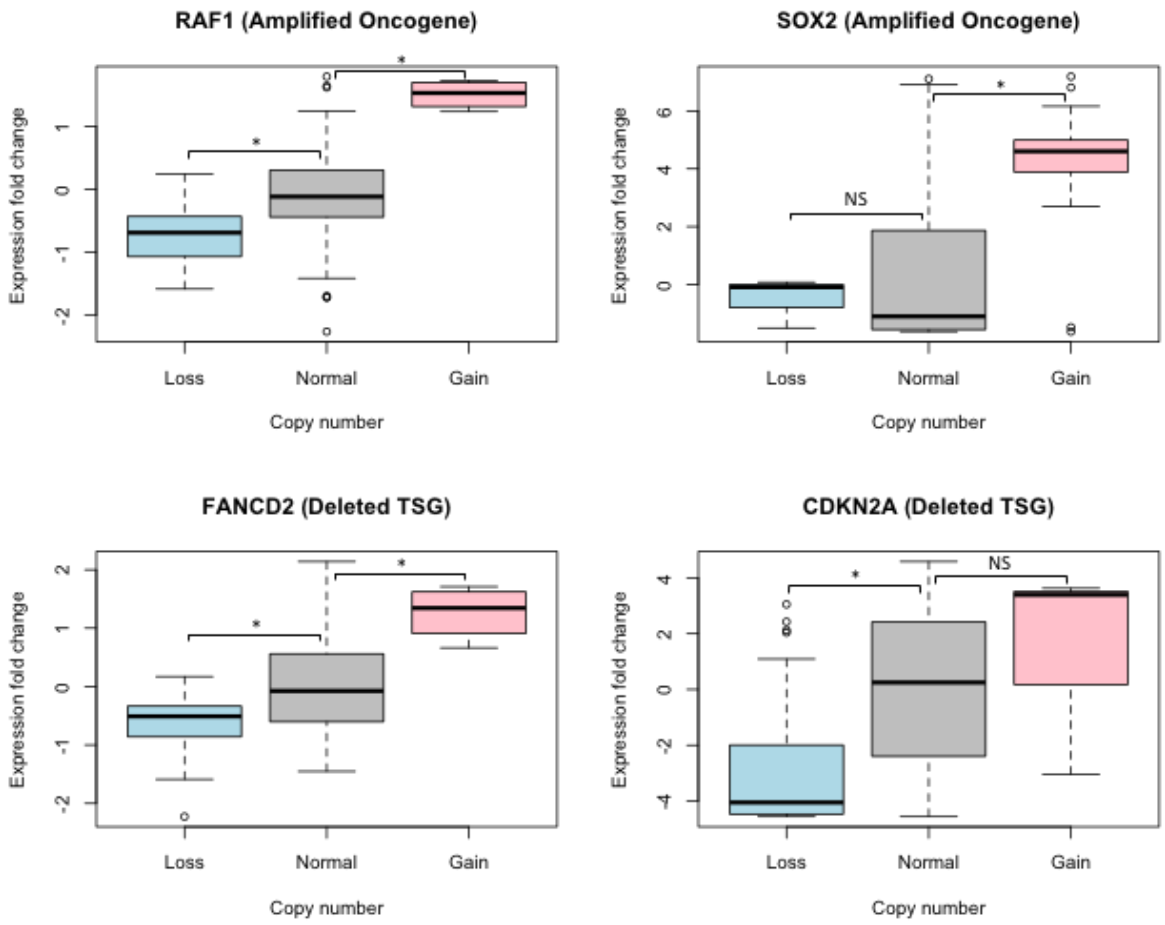
Frequency of copy number gain (red) or loss (blue) of selected genes frequently amplified or deleted in TCGA tumors predicted by GISTIC analysis for each tumor type in TCGA SNP array datasets.



Supplementary Figure S9

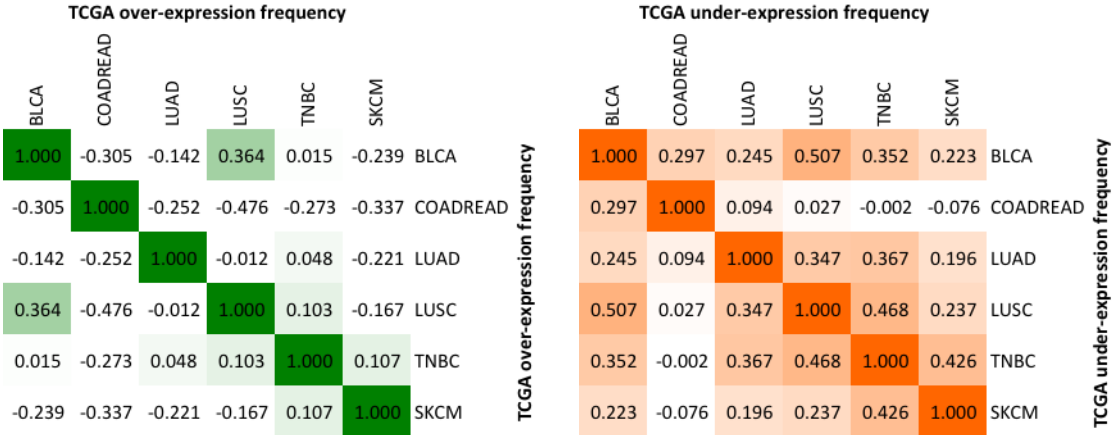
Distribution of $\log_2$ ratio of gene copy number relative to the estimated overall ploidy of each individual PDX sample or the diploid state for selected frequently amplified or deleted genes in TCGA tumors predicted by GISTIC analysis across all PDX samples. The threshold of low-level gain and loss is defined as $\log_2(CN/ploidy) > +0.4$ and $\log_2(CN/ploidy) < -0.4$ respectively.

A



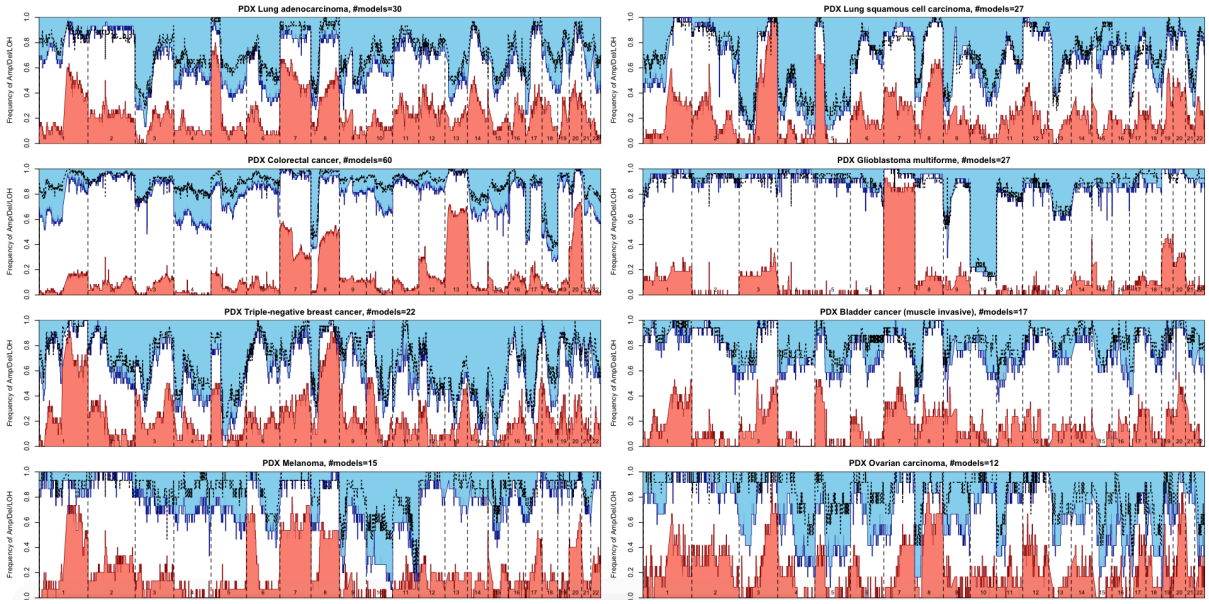
Supplementary Figure S10

Expression fold change of each gene across all PDX samples, defined by fold change of $\log_2$(TPM+1) relative to the mean expression of samples with a stringent normal copy number state (-0.4 < log2(CN/ploidy) < 0.4). Here, a higher-level copy number gain and loss is defined as $\log_2$(CN/ploidy) > +1 and $\log_2$(CN/ploidy) < -1 respectively. The normal copy number state is defined as -1 < $\log_2$(CN/ploidy) < +1. Significance in differences in expression by Student's t-test (*: p-value < 0.005, NS: non-significant).

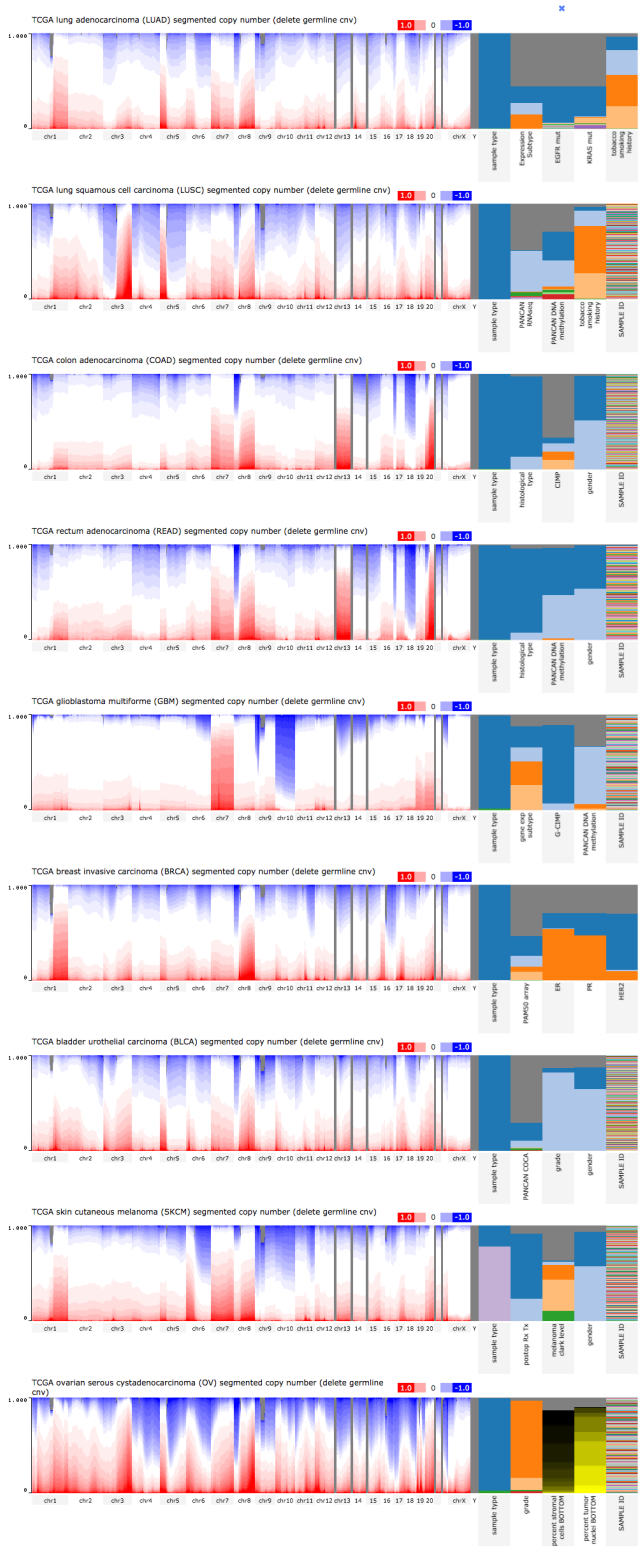


Supplementary Figure S11

Correlation frequency of genes that are over-expressed (z-score of $\log_2$(TPM+1) > 1, green) or under-expressed (z-score of $\log_2$(TPM+1) < -1, orange) between each tumor type in TCGA RNA-Seq samples.

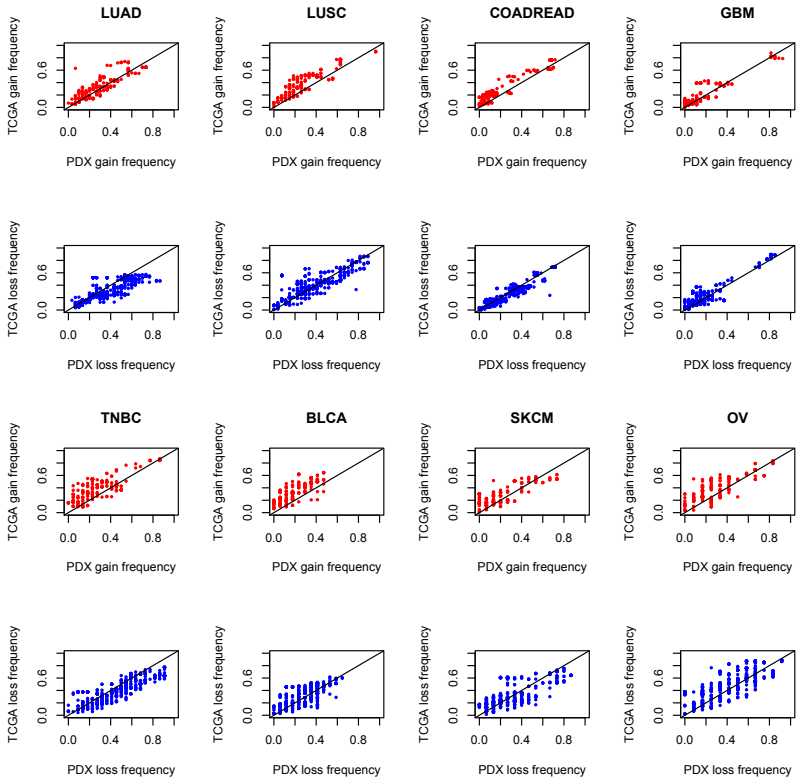


Supplementary Figure S12
Frequency of genome-wide copy number gain, loss and LOH across PDX models for 8 tumor types.

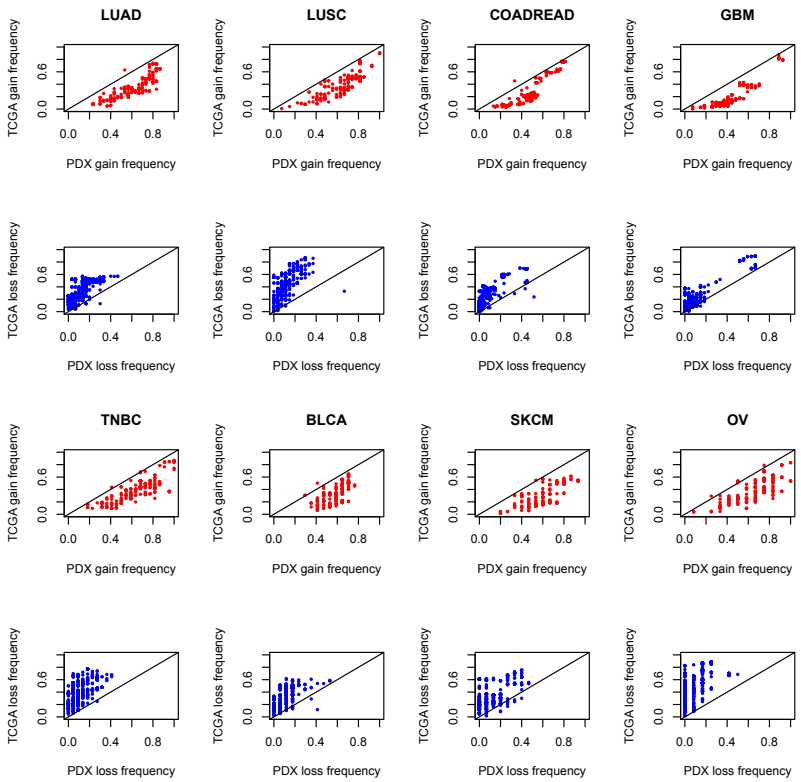


Supplementary Figure S13
Frequency of genome-wide copy number gain and loss across TCGA samples for 8 tumor types. (Compiled from UCSC Cancer Browser, https://genome-cancer.ucsc.edu/)

A



B

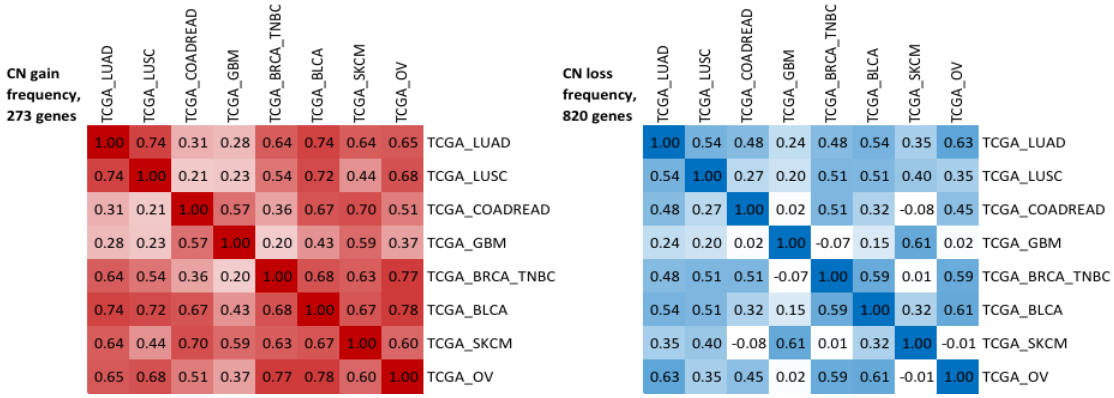


Supplementary Figure S14

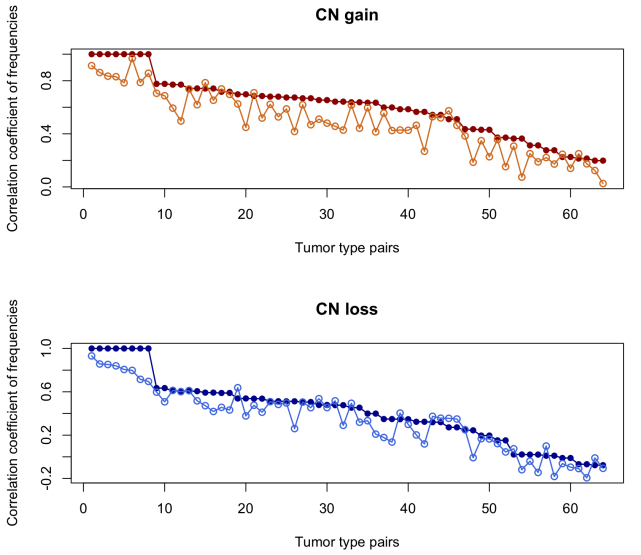
(A) Comparison of frequency of copy number gain (red) or loss (blue) of selected genes frequently amplified or deleted in TCGA tumors predicted by GISTIC analysis for each tumor type between PDX and TCGA datasets using predicted ploidy as a reference state.

(B) Comparison of frequency of copy number gain (red) or loss (blue) of genes frequently amplified or deleted in TCGA tumors predicted by GISTIC analysis for each tumor type between PDX and TCGA datasets using diploid as a reference state.

A



B



Supplementary Figure S15

(A) Correlation of frequency of copy number gain (red) or loss (blue) of selected genes frequently amplified or deleted in TCGA tumors predicted by GISTIC analysis between each tumor type in TCGA SNP array datasets.

(B) Ranked correlation coefficients based on Figure 4b and Supplementary Figure S6.

		Non-Silent Mutation		Amplification		Deletion		Over expression		Under expression	
		PDX	TCGA	PDX	TCGA	PDX	TCGA	PDX	TCGA	PDX	TCGA
WNT signalling	APC	0.610	0.553	0.017	0.075	0.373	0.308	0.104	0.029	0.333	0.227
	CTNNB1	0.024	0.050	0.000	0.112	0.220	0.133	0.063	0.093	0.188	0.043
	DKK1	NA	0.008	0.051	0.050	0.288	0.196	0.042	0.026	0.167	0.251
	DKK2	NA	0.015	0.017	0.032	0.441	0.339	0.000	0.177	0.000	0.128
	DKK3	NA	0.010	0.119	0.148	0.119	0.120	0.042	0.058	0.271	0.081
	DKK4	NA	0.010	0.305	0.419	0.119	0.127	0.333	0.254	0.000	0.000
	LRP5	NA	0.029	0.034	0.123	0.085	0.162	0.104	0.270	0.000	0.008
	FZD10	NA	0.027	0.085	0.182	0.136	0.133	0.250	0.142	0.000	0.000
	AMER1	0.098	0.097	0.458	0.167	0.153	0.154	0.125	0.204	0.167	0.046
	AXIN2	NA	0.049	0.136	0.235	0.119	0.127	0.667	0.773	0.021	0.002
	TCF7L2	0.098	0.061	0.034	0.044	0.305	0.255	0.271	0.435	0.000	0.013
	FBXW7	0.122	0.139	0.000	0.041	0.458	0.347	0.042	0.014	0.167	0.230
	ARID1A	0.146	0.094	0.000	0.015	0.458	0.394	0.063	0.021	0.271	0.225
	SOX9	0.171	0.121	0.119	0.216	0.119	0.149	0.250	0.637	0.021	0.000
TGF-β signalling	TGFBR1	NA	0.024	0.102	0.166	0.186	0.144	0.042	0.016	0.375	0.196
	TGFBR2	0.146	0.024	0.000	0.107	0.254	0.138	0.229	0.153	0.000	0.051
	ACVR1B	0.049	0.044	0.136	0.195	0.102	0.106	0.313	0.326	0.083	0.022
	ACVR2A	NA	0.045	0.051	0.195	0.051	0.049	0.125	0.043	0.125	0.109
	SMAD2	0.049	0.044	0.017	0.024	0.644	0.672	0.042	0.026	0.292	0.262
	SMAD3	NA	0.032	0.017	0.031	0.356	0.396	0.021	0.008	0.188	0.160
	SMAD4	0.122	0.086	0.017	0.013	0.729	0.705	0.042	0.043	0.313	0.209
PI3K signalling	IGF2	NA	0.007	0.153	0.222	0.119	0.120	0.167	0.153	0.000	0.026
	IRS2	NA	0.015	0.695	0.617	0.017	0.032	0.271	0.240	0.042	0.019
	PIK3CA	0.146	0.087	0.102	0.183	0.085	0.070	0.021	0.002	0.313	0.262
	PIK3R1	0.171	0.042	0.034	0.073	0.322	0.276	0.104	0.093	0.125	0.102
	PTEN	0.049	0.039	0.000	0.029	0.356	0.260	0.083	0.040	0.042	0.112
RTK-RAS signalling	ERBB2	0.195	0.037	0.136	0.260	0.102	0.114	0.042	0.042	0.063	0.008
	ERBB3	0.049	0.057	0.136	0.192	0.102	0.106	0.104	0.091	0.042	0.003
	KRAS	0.146	0.082	0.288	0.226	0.136	0.114	0.104	0.053	0.188	0.101
	NRAS	0.024	0.007	0.000	0.047	0.373	0.281	0.083	0.088	0.208	0.107
	BRAF	0.098	0.101	0.288	0.503	0.051	0.013	0.000	0.003	0.396	0.350
p53 signalling	TP53	0.366	0.113	0.000	0.029	0.576	0.597	0.083	0.259	0.125	0.165
	ATM	0.268	0.077	0.034	0.135	0.102	0.183	0.063	0.005	0.271	0.297

Supplementary Figure S16

Frequency of genes altered PDX and TCGA tumors for each genomic datatype for colorectal cancer. These genes are identified by commonly affected pathways in colorectal cancer reported in TCGA studies. For both PDX and TCGA cohorts of colorectal cancer, we observed high frequencies in the 1) mutation of APC and over-expression of AXIN2 in the WNT signaling pathway, 2) amplification of IRS2 in the PI3K signaling pathway, 3) copy number loss of SMAD2 and SMAD4 in the TGF-β signaling pathway, 4) under-expression of BRAF in the RTK-RAS signaling pathway, and 5) copy number loss of TP53 in the TP53 signaling pathway.

model	coverage	mouse %	input variants	Complete pipeline												Without Xenome pipeline												Without ALT-aware pipeline											
				ALL						PASS						ALL						PASS						ALL						PASS					
				output variants	TP	FP	FN	precision	recall	output variants	TP	FP	FN	precision	recall	output variants	TP	FP	FN	precision	recall	output variants	TP	FP	FN	precision	recall	output variants	TP	FP	FN	precision	recall	output variants	TP	FP	FN	precision	recall
J000093572 (lung squamous cell carcinoma)	1000X	10	658	638	636	2	22	0.997	0.967	580	580	0	78	1.000	0.881	12468	637	11831	21	0.051	0.968	591	561	30	97	0.949	0.853	495	494	1	164	0.998	0.751	409	409	0	249	1.000	0.622
		15	658	638	636	2	22	0.997	0.967	580	580	0	78	1.000	0.881	11114	635	10479	23	0.057	0.965	572	560	12	98	0.979	0.851	495	494	1	164	0.998	0.751	409	409	0	249	1.000	0.622
		25	658	638	636	2	22	0.997	0.967	580	580	0	78	1.000	0.881	10733	635	10098	23	0.059	0.965	563	558	5	100	0.991	0.848	495	494	1	164	0.998	0.751	410	410	0	248	1.000	0.623
	1500X	10	658	645	640	5	18	0.992	0.973	581	581	0	77	1.000	0.883	12772	638	12134	20	0.050	0.970	595	556	39	102	0.934	0.845	496	496	0	162	1.000	0.754	405	405	0	253	1.000	0.616
		15	658	645	640	5	18	0.992	0.973	581	581	0	77	1.000	0.883	12274	638	11636	20	0.052	0.970	572	556	16	102	0.972	0.845	496	496	0	162	1.000	0.754	405	405	0	253	1.000	0.616
		25	658	645	640	5	18	0.992	0.973	581	581	0	77	1.000	0.883	11526	637	10889	21	0.055	0.968	561	555	6	103	0.989	0.843	496	496	0	162	1.000	0.754	405	405	0	253	1.000	0.616
	500X	10	658	640	639	1	19	0.998	0.971	587	587	0	71	1.000	0.892	10972	638	10334	20	0.058	0.970	605	569	36	89	0.940	0.865	489	488	1	170	0.998	0.742	394	394	0	264	1.000	0.599
		15	658	640	639	1	19	0.998	0.971	587	587	0	71	1.000	0.892	10219	637	9582	21	0.062	0.968	581	569	12	89	0.979	0.865	489	488	1	170	0.998	0.742	394	394	0	264	1.000	0.599
		25	658	640	639	1	19	0.998	0.971	587	587	0	71	1.000	0.892	9217	636	8581	22	0.069	0.967	568	564	4	94	0.993	0.857	489	488	1	170	0.998	0.742	394	394	0	264	1.000	0.599
TM00055 (giant cell glioblastoma)	1000X	10	492	488	479	9	13	0.982	0.974	456	451	5	41	0.989	0.917	12445	478	11967	14	0.038	0.972	480	440	40	52	0.917	0.894	310	302	8	190	0.974	0.614	278	273	5	219	0.982	0.555
		15	492	488	479	9	13	0.982	0.974	456	451	5	41	0.989	0.917	11214	478	10736	14	0.043	0.972	464	443	21	49	0.955	0.900	310	302	8	190	0.974	0.614	279	274	5	218	0.982	0.557
		25	492	488	479	9	13	0.982	0.974	456	451	5	41	0.989	0.917	11136	478	10658	14	0.043	0.972	453	441	12	51	0.974	0.896	310	302	8	190	0.974	0.614	278	273	5	219	0.982	0.555
	1500X	10	492	489	480	9	12	0.982	0.976	457	450	7	42	0.985	0.915	12748	478	12270	14	0.037	0.972	493	441	52	51	0.895	0.896	312	306	6	186	0.981	0.622	281	276	5	216	0.982	0.561
		15	492	489	480	9	12	0.982	0.976	457	450	7	42	0.985	0.915	12058	478	11580	14	0.040	0.972	459	439	20	53	0.956	0.892	313	307	6	185	0.981	0.624	282	277	5	215	0.982	0.563
		25	492	489	480	9	12	0.982	0.976	457	450	7	42	0.985	0.915	11199	478	10721	14	0.043	0.972	454	440	14	52	0.969	0.894	313	307	6	185	0.981	0.624	282	277	5	215	0.982	0.563
	500X	10	492	487	479	8	13	0.984	0.974	454	449	5	43	0.989	0.913	11382	480	10902	12	0.042	0.976	489	438	51	54	0.896	0.890	301	292	9	200	0.970	0.593	254	246	8	246	0.969	0.500
		15	492	487	479	8	13	0.984	0.974	454	449	5	43	0.989	0.913	10486	480	10006	12	0.046	0.976	462	439	23	53	0.950	0.892	305	295	10	197	0.967	0.600	258	250	8	242	0.969	0.508
		25	492	487	479	8	13	0.984	0.974	454	449	5	43	0.989	0.913	9843	480	9363	12	0.049	0.976	449	437	12	55	0.973	0.888	304	294	10	198	0.967	0.598	257	249	8	243	0.969	0.506
TM00926 (colon adenocarcinoma)	1000X	10	840	837	782	55	58	0.934	0.931	764	720	44	120	0.942	0.857	12722	778	11944	62	0.061	0.926	789	708	81	132	0.897	0.843	645	606	39	234	0.940	0.721	583	553	30	287	0.949	0.658
		15	840	838	782	56	58	0.933	0.931	764	720	44	120	0.942	0.857	11327	776	10551	64	0.069	0.924	760	705	55	135	0.928	0.839	644	605	39	235	0.939	0.720	581	551	30	289	0.948	0.656
		25	840	837	782	55	58	0.934	0.931	764	720	44	120	0.942	0.857	10900	774	10126	66	0.071	0.921	752	702	50	138	0.934	0.836	645	606	39	234	0.940	0.721	583	553	30	287	0.949	0.658
	1500X	10	840	836	779	57	61	0.932	0.927	763	717	46	123	0.940	0.854	13220	777	12443	63	0.059	0.925	792	704	88	136	0.889	0.838	646	605	41	235	0.937	0.720	572	542	30	298	0.948	0.645
		15	840	836	779	57	61	0.932	0.927	763	717	46	123	0.940	0.854	12780	777	12003	63	0.061	0.925	759	701	58	139	0.924	0.835	648	607	41	233	0.937	0.723	575	545	30	295	0.948	0.649
		25	840	836	779	57	61	0.932	0.927	763	717	46	123	0.940	0.854	11967	777	11190	63	0.065	0.925	752	701	51	139	0.932	0.835	646	605	41	235	0.937	0.720	573	543	30	297	0.948	0.646
	500X	10	840	832	776	56	64	0.933	0.924	761	716	45	124	0.941	0.852	10739	779	9960	61	0.073	0.927	785	705	80	135	0.898	0.839	640	599	41	241	0.936	0.713	570	538	32	302	0.944	0.640
		15	840	832	777	55	63	0.934	0.925	761	716	45	124	0.941	0.852	10725	779	9946	61	0.073	0.927	762	704	58	136	0.924	0.838	639	598	41	242	0.936	0.712	569	537	32	303	0.944	0.639
		25	840	832	776	56	64	0.933	0.924	761	716	45	124	0.941	0.852	9825	775	9050	65	0.079	0.923	754	703	51	137	0.932	0.837	640	599	41	241	0.936	0.713	570	538	32	302	0.944	0.640
TM01117 (bladder urothelial papillary carcinoma)	1000X	10	662	653	647	6	15	0.991	0.977	604	602	2	60	0.997	0.909	12562	645	11917	17	0.051	0.974	620	589	31	73	0.950	0.890	485	481	4	181	0.992	0.727	445	441	4	221	0.991	0.666
		15	662	652	647	5	15	0.992	0.977	604	602	2	60	0.997	0.909	11293	645	10648	17	0.057	0.974	608	592	16	70	0.974	0.894	484	480	4	182	0.992	0.725	444	440	4	222	0.991	0.665
		25	662	652	647	5	15	0.992	0.977	604	602	2	60	0.997	0.909	10934	644	10290	18	0.059	0.973	599	592	7	70	0.988	0.894	485	481	4	181	0.992	0.727	445	441	4	221	0.991	0.666
	1500X	10	662	654	647	7	15	0.989	0.977	604	602	2	60	0.997	0.909	12689	644	12045	18	0.051	0.973	641	595	46	67	0.928	0.899	497	490	7	172	0.986	0.740	439	436	3	226	0.993	0.659
		15	662	654	647	7	15	0.989	0.977	605	603	2	59	0.997	0.911	12244	644	11600	18	0.053	0.973	611	594	17	68	0.972	0.897	497	490	7	172	0.986	0.740	440	437	3	225	0.993	0.660
		25	662	654	647	7	15	0.989	0.977	603	601	2	61	0.997	0.908	11574	644	10930	18	0.056	0.973	602	593	9	69	0.985	0.896	496	489	7	173	0.986	0.739	439	436	3	226	0.993	0.659
	500X	10	662	645	641	4	21	0.994	0.968	607	605	2	57	0.997	0.914	11330	641	10689	21	0.057	0.968	636	595	41	67	0.936	0.899	481	478	3	184	0.994	0.722	420	418	2	244	0.995	0.631
		15	662	645	641	4	21	0.994	0.968	607	605	2	57	0.997	0.914	10503	643	9860	19	0.061	0.971	613	596	17	66	0.972	0.900	480	477	3	185	0.994	0.721	420	418	2	244	0.995	0.631
		25	662	645	641	4	21	0.994	0.968	607	605	2	57	0.997	0.914	9225	642	8583	20	0.070	0.970	604	597	7	65	0.988	0.902	481	478	3	184	0.994	0.722	420	418	2	244	0.995	0.631
TM01443 (melanoma)	1000X	10	728	721	707	14	21	0.981	0.971	679	670	9	58	0.987	0.920	12415	709	11706	19	0.057	0.974	691	653	38	75	0.945	0.897	545	537	8	191	0.985	0.738	524	520	4	208	0.992	0.714
		15	728	721	707	14	21	0.981	0.971	680	671	9	57	0.987	0.922	11231	709	10522	19	0.063	0.974	680	657	23	71	0.966	0.902	545	537	8	191	0.985	0.738	524	520	4	208	0.992	0.714
		25	728	721	707	14	21	0.981	0.971	679	670	9	58	0.987	0.920	11171	707	10464	21	0.063	0.971	674	659	15	69	0.978	0.905	545	537	8	191	0.985	0.738	523	519	4	209	0.992	0.713
	1500X	10	728	723	707	16	21	0.978	0.971	679	671	8	57	0.988	0.922	12902	706	12196	22	0.055	0.970	703	652	51	76	0.927	0.896	547	539	8	189	0.985	0.740	521	519	2	209	0.996	0.713
		15	728	722	707	15	21	0.979	0.971	680	671	9	57	0.987	0.922	12475	706	11769	22	0.057	0.970	676	653	23	75	0.966	0.897	548	540	8	188	0.985	0.742	522	520	2	208	0.996	0.714
		25	728	722	707	15	21	0.979	0.971	680	671	9	57	0.987	0.922	11429	705	10724	23	0.062	0.968	668	652	16	76	0.976	0.896	547	539	8	189	0.985	0.740	521	519	2	209	0.996	0.713
	500X	10	728	723	708	15	20	0.979	0.973	686	678	8	50	0.988	0.931	11533	706	10827	22	0.061	0.970	710	664	46	64	0.935	0.912	539	531	8	197	0.985	0.729	493	490	3	238	0.994	0.673
		15	728	723	708	15	20	0.979	0.973	686	678	8	50	0.988	0.931	10618	706	9912	22	0.066	0.970	680	659	21	69	0.969	0.905	539	531	8	197	0.985	0.729	493	490	3	238	0.994	0.673
		25	728	723	708	15	20	0.979	0.973	685	677	8	51	0.988	0.930	9474	706	8768	22	0.075	0.970	674	660	14	68	0.979	0.907	539	531	8	197	0.985	0.729	493	490	3	238	0.994	0.673

Supplementary Table S1

This table summarizes the results from the benchmarking studies of the CTP variant calling pipeline using 45 simulated sequencing datasets different samples, sequencing coverages, and mouse DNA content.

Model	Coverage (X)	Mouse percent	ALT_AF Correlation						ALT_AF Difference Median					
			Output ALL	Output PASS	NoXenome ALL	NoXenome PASS	NoAltAware ALL	NoAltAware PASS	Output ALL	Output PASS	NoXenome ALL	NoXenome PASS	NoAltAware ALL	NoAltAware PASS
J000093572	500	10	0.995	0.995	0.985	0.984	0.982	0.981	-3	-3	-3	-3	-13	-15
		15	0.995	0.995	0.976	0.975	0.982	0.981	-3	-3	-3	-3	-13	-15
		25	0.995	0.995	0.957	0.952	0.982	0.981	-3	-3	-3	-3	-13	-15
	1000	10	0.996	0.996	0.986	0.985	0.983	0.984	-3	-3	-3	-3	-13.5	-14
		15	0.996	0.996	0.977	0.974	0.983	0.984	-3	-3	-3	-3	-14	-15
		25	0.996	0.996	0.960	0.955	0.983	0.984	-3	-3	-3	-3	-13	-15
	1500	10	0.997	0.997	0.986	0.985	0.976	0.974	-3	-3	-3	-3	-13	-15
		15	0.997	0.997	0.977	0.974	0.976	0.974	-3	-3	-3	-3	-13	-15
		25	0.997	0.997	0.961	0.955	0.976	0.974	-3	-3	-3	-3	-13	-15
TM00055	500	10	0.991	0.990	0.954	0.927	0.970	0.958	-3	-3	-4	-4	-12	-12
		15	0.990	0.990	0.931	0.878	0.970	0.958	-3	-3	-4	-4	-12	-12
		25	0.990	0.990	0.888	0.777	0.970	0.959	-3	-3	-4	-4	-12	-12
	1000	10	0.990	0.989	0.952	0.935	0.966	0.955	-3	-3	-4	-4	-12	-12
		15	0.991	0.989	0.929	0.894	0.963	0.950	-3	-3	-4	-4	-12	-12
		25	0.991	0.989	0.888	0.813	0.963	0.951	-3	-3	-4	-4	-12	-12
	1500	10	0.991	0.988	0.953	0.931	0.968	0.953	-3	-4	-4	-4	-12	-12
		15	0.991	0.989	0.930	0.882	0.969	0.954	-3	-4	-4	-4	-12	-12
		25	0.991	0.989	0.888	0.811	0.969	0.953	-3	-4	-4	-4	-12	-12
TM00926	500	10	0.990	0.990	0.974	0.977	0.976	0.980	-3	-3	-3	-3	-14	-15
		15	0.990	0.990	0.963	0.965	0.976	0.980	-3	-3	-3	-3	-14	-15
		25	0.990	0.990	0.939	0.943	0.976	0.980	-3	-3	-3	-3	-14	-15
	1000	10	0.992	0.992	0.977	0.980	0.977	0.982	-3	-3	-3	-3	-15	-15
		15	0.992	0.992	0.965	0.970	0.977	0.982	-3	-3	-3	-3	-15	-15
		25	0.992	0.992	0.943	0.947	0.977	0.982	-3	-3	-3	-3	-15	-15
	1500	10	0.993	0.993	0.977	0.981	0.974	0.976	-3	-3	-3	-3	-14	-15
		15	0.992	0.993	0.966	0.971	0.974	0.976	-3	-3	-3	-3	-14	-15
		25	0.992	0.993	0.945	0.950	0.974	0.977	-3	-3	-3	-3	-14	-15
TM01117	500	10	0.994	0.992	0.979	0.979	0.977	0.977	-3	-3	-3	-3	-13	-14
		15	0.994	0.993	0.969	0.968	0.977	0.977	-3	-3	-3	-3	-13	-14
		25	0.994	0.992	0.946	0.942	0.977	0.977	-3	-3	-3	-3	-13	-14
	1000	10	0.995	0.994	0.979	0.981	0.981	0.983	-3	-3	-3	-3	-14	-14
		15	0.995	0.994	0.969	0.969	0.981	0.983	-3	-3	-3	-3	-14	-14
		25	0.995	0.994	0.949	0.945	0.981	0.983	-3	-3	-3	-3	-14	-14
	1500	10	0.995	0.995	0.980	0.982	0.978	0.978	-3	-3	-3	-3	-13	-14
		15	0.995	0.994	0.971	0.971	0.979	0.979	-3	-3	-3	-3	-13	-13.5
		25	0.995	0.995	0.951	0.947	0.978	0.977	-3	-3	-3	-3	-13	-14
TM01443	500	10	0.984	0.988	0.962	0.970	0.971	0.972	-3	-3	-3	-3	-14	-15
		15	0.984	0.988	0.950	0.959	0.971	0.972	-3	-3	-3	-3	-14	-15
		25	0.984	0.988	0.921	0.934	0.971	0.972	-3	-3	-4	-4	-14	-15
	1000	10	0.981	0.990	0.961	0.973	0.970	0.972	-3	-3	-4	-3	-15	-15
		15	0.982	0.990	0.947	0.959	0.970	0.972	-3	-3	-4	-4	-15	-15
		25	0.982	0.990	0.925	0.937	0.970	0.972	-3	-3	-4	-4	-15	-15
	1500	10	0.984	0.991	0.962	0.974	0.968	0.968	-3	-3	-3	-3	-14	-15
		15	0.984	0.991	0.949	0.960	0.967	0.967	-3	-3	-3	-4	-15	-15
		25	0.984	0.991	0.924	0.939	0.968	0.967	-3	-3	-4	-4	-14	-15

Supplementary Table S2

This table shows the correlation and difference in median of alternate allele frequencies between input and true positive variants for all the simulated samples. ALL: all variants called by the pipeline; PASS: variants annotated as “PASS” in the pipeline which pass the hard filters, minimum read depth and minimum alternate allele frequency of the variant.

Database	Number of variants
dbSNP Build 144	143,257,868
1000 Genomes	637,385
ExAC >1% MAF	38,129
Recurrent in 20 normals on CTP panel	105
Cosmic v80	18,974,374

Supplementary Table S3
Number of variants in each germline databases.

CTP (383 samples)	SNP			INDEL		
Number of unique variants	Mean	Median	Standard deviation	Mean	Median	Standard deviation
Total	798	779	107	30	28	13
DP, AF and Hard filters	601	592	80	23	22	11
Germline filters	164	146	84	9.9	9	7.9
Recurrent variants filter and High/Moderate impact variants	67	57	42	3.1	2	3.2
Clincally relevant variants rescue	71	61	42	3.2	2	3.6
Clincally relevant variants	5.6	5	2.1	0.18	0	0.65

Supplementary Table S4

Number of unique variants in each CTP sample (n=383), represented by mean, median and standard deviation, called by GATK and after each filtering or rescue step. The last row shows the average number of variants annotated as clinically relevant.

PDX CTP	Number of models/samples	TCGA Whole Exome (Somatic)	Number of samples (patients)
Lung adenocarcinoma (PDX_LUAD)	36	Lung adenocarcinoma (TCGA_LUAD)	571 (569)
Lung squamous cell carcinoma (PDX_LUSC)	28	Lung squamous cell carcinoma (TCGA_LUSC)	494 (494)
Colorectal cancer (PDX_Colorectal)	41	Colorectal adenocarcinoma (TCGA_COADREAD)	595 (592)
Triple negative breast cancer (PDX_TNBC)	12	Breast invasive carcinoma, Triple-negative (TCGA_BRCA_TNBC)	132 (131)
Bladder cancer, Muscle invasive (PDX_BLCAinvasive)	12	Bladder urothelial carcinoma (TCGA_BLCA)	413 (412)
Melanoma (PDX_Melanoma)	12	Skin cutaneous melanoma (TCGA_SKCM)	472 (470)

Supplementary Table S5
Number of PDX and TCGA samples for 5 tumor types used for analysis of variant calling.

Ensembl ID	Gene Symbol	Up or Down regulation
ENSG00000002919	SNX11	1
ENSG00000007866	TEAD3	-1
ENSG00000009790	TRAF3IP3	1
ENSG00000028277	POU2F2	1
ENSG00000056558	TRAF1	1
ENSG00000061676	NCKAP1	-1
ENSG00000062370	ZFP112	-1
ENSG00000072818	ACAP1	1
ENSG00000076662	ICAM3	1
ENSG00000084070	SMAP2	1
ENSG00000086730	LAT2	1
ENSG00000088256	GNA11	-1
ENSG00000102096	PIM2	1
ENSG00000104067	TJP1	-1
ENSG00000104814	MAP4K1	1
ENSG00000110031	LPXN	1
ENSG00000110777	POU2AF1	1
ENSG00000116473	RAP1A	1
ENSG00000120256	LRP11	-1
ENSG00000122386	ZNF205	-1
ENSG00000126822	PLEKHG3	-1
ENSG00000130147	SH3BP4	-1
ENSG00000137693	YAP1	-1
ENSG00000138185	ENTPD1	1
ENSG00000142192	APP	-1
ENSG00000144677	CTDSPL	-1
ENSG00000147799	ARHGAP39	-1
ENSG00000150760	DOCK1	-1
ENSG00000152990	GPR125	-1
ENSG00000156052	GNAQ	-1
ENSG00000157985	AGAP1	-1
ENSG00000162627	SNX7	-1
ENSG00000163625	WDFY3	-1
ENSG00000167984	NLRC3	1
ENSG00000172578	KLHL6	1
ENSG00000173200	PARP15	1
ENSG00000180096	1-Sep	1
ENSG00000180891	CUEDC1	-1
ENSG00000187079	TEAD1	-1
ENSG00000187164	KIAA1598	-1
ENSG00000188822	CNR2	1
ENSG00000197702	PARVA	-1
ENSG00000197763	TXNRD3	-1
ENSG00000198833	UBE2J1	1
ENSG00000211895	IGHA1	1
ENSG00000213402	PTPRCAP	1
ENSG00000213999	MEF2B	1
ENSG00000249096	RP11-290F5.1	1

Supplementary Table S6

Classifier gene table to classify EBV-associated PDX lymphomas versus other tumors. Up-regulation: +1; Down-regulation: -1.

Model ID	Sample ID	Correlation coefficient (Pearson)
J000077712	J000077712PT	0.811
J000079689	J000095142	0.938
J000079689	J000079689PT	0.983
TM00327	OVJX01F000P0	0.974
TM00327	OVJX01F017P0	0.955
TM00327	OVJX01F020P0	0.969
TM00327	OVJX01F021P0	0.968
TM01594	TM01594F062P2	0.966
TM01594	TM01594FPT	0.978

Supplementary Table S7
Pearson correlation coefficient of gene-based log2(total CN/ploidy) between the single-tumor and tumor-normal CNV analysis.

Cancer Census	Gene symbol	Loss: log2(CN/ploidy) < -1		Normal: -1 <= log2(CN/ploidy) <= 1		Gain: log2(CN/ploidy) > 1		P-value	
		# Samples	Mean Exp FC	# Samples	Mean Exp FC	# Samples	Mean Exp FC	Loss vs Norm	Gain vs Norm
Amplified Oncogenes	KAT6A	13	-0.574	170	0.124	12	1.160	9.91E-04	5.40E-03
	MYCL	9	-0.352	182	-0.018	4	4.380	2.26E-01	5.65E-02
	RAF1	36	-0.755	155	-0.073	4	1.506	4.58E-10	7.03E-05
	MDM2	4	-0.362	182	0.022	9	1.431	7.08E-02	2.38E-03
	NKX2-1	18	-0.859	168	-0.083	9	4.744	1.59E-07	1.00E-03
	MYC	5	-2.363	143	0.295	47	1.009	2.05E-02	2.64E-04
	ERBB2	3	-0.407	186	0.108	6	2.126	1.69E-01	3.12E-03
	EGFR	4	-0.945	171	-0.056	20	1.906	8.21E-02	4.42E-04
	WHSC1L1	29	-0.732	156	0.026	10	1.648	3.65E-09	3.18E-04
	ERG	21	-0.627	170	0.051	4	1.977	4.52E-05	1.24E-01
	PPM1D	4	-0.775	182	0.037	9	0.729	6.15E-03	5.23E-02
	JUN	10	-0.545	182	0.061	3	0.876	2.32E-02	1.59E-02
	SOX2	3	-0.509	173	0.317	19	4.109	1.17E-01	2.98E-07
Deleted TSGs	BRCA1	3	-1.515	189	-0.032	3	0.445	1.40E-01	1.90E-01
	BIRC3	26	0.131	160	0.237	9	1.847	4.15E-01	4.10E-02
	MAP2K4	39	-0.798	153	-0.050	3	1.731	2.50E-10	8.16E-02
	SMARCB1	13	-0.892	178	0.022	4	1.252	1.96E-06	1.83E-02
	TNFAIP3	30	-0.679	162	-0.044	3	-0.422	1.19E-02	6.50E-01
	VHL	33	-0.896	156	-0.083	6	0.988	1.58E-05	1.71E-02
	BRCA2	21	-0.517	151	0.191	23	0.407	6.08E-05	1.99E-01
	RB1	26	-1.041	150	-0.022	19	0.217	4.11E-05	1.74E-01
	FANCD2	34	-0.638	155	-0.018	6	1.265	1.00E-08	1.29E-04
	KDM6A	27	-0.527	159	0.051	9	0.655	5.14E-05	2.00E-02
	GPC3	27	-0.572	156	-0.139	12	-1.442	1.50E-01	9.98E-01
	CDKN2A	68	-3.020	124	-0.064	3	1.326	7.86E-15	2.95E-01
	ATM	44	-0.682	148	0.036	3	1.527	1.80E-05	2.00E-14
	APOBEC3B	17	-1.016	172	-0.124	6	1.449	1.81E-02	1.70E-02
	AMER1	23	-0.173	165	-0.026	7	0.282	2.24E-01	1.53E-01
	WT1	14	-0.434	176	-0.162	5	-0.130	8.75E-02	4.75E-01
	IKZF1	10	-0.262	166	0.004	19	-0.381	1.54E-01	9.89E-01
	NF2	14	-0.950	178	0.014	3	1.239	7.53E-05	1.16E-02
	NF1	12	-1.263	179	-0.028	4	0.599	2.86E-04	1.02E-01

Mean Expression Fold Change = log2(TPM+1) - Mean[log2(TPM+1)]norm, where norm: -0.4 <= log2(CN/ploidy) <= 0.4

Supplementary Table S8

Mean expression fold change of genes with copy number normal, gain and loss state for genes found in the Cancer Census that is listed as oncogenes affected by amplification and tumor suppressor genes affected by deletions (refer to Supplementary Figure S16). The p-value, calculated by Student's t-test, measures if the difference in expression for each gene between the copy number loss models versus normal models, and between the copy number gain models versus normal models is significant.

COADREAD CTP (coding, non-silent)	TCGA mut >=5%	TCGA mut <5%	Total
PDX mut	54	244	298
PDX not mut	0	60	60
Total	54	304	358

LUAD CTP (coding, non-silent)	TCGA mut >=5%	TCGA mut <5%	Total
PDX mut	44	220	264
PDX not mut	1	93	94
Total	45	313	358

LUSC CTP (coding, non-silent)	TCGA mut >=5%	TCGA mut <5%	Total
PDX mut	29	197	226
PDX not mut	1	131	132
Total	30	328	358

BLCA CTP (coding, non-silent)	TCGA mut >=5%	TCGA mut <5%	Total
PDX mut	37	151	188
PDX not mut	6	164	170
Total	43	315	358

SKCM CTP (coding, non-silent)	TCGA mut >=5%	TCGA mut <5%	Total
PDX mut	84	145	229
PDX not mut	5	124	129
Total	89	269	358

BRCA TNBC CTP (coding, non-silent)	TCGA mut >=5%	TCGA mut <5%	Total
PDX mut	18	173	191
PDX not mut	1	166	167
Total	19	339	358

Supplementary Table S9

Contingency table for Fisher's Exact Test for CTP genes with coding, non-silent mutations for different tumor types in the PDX and TCGA cohort. For PDX, the number of CTP genes with and without coding, non-silent mutations in each tumor type cohort was counted. For TCGA data with more samples than PDX, the number of CTP genes with coding, non-silent mutations at ≥5% and <5% frequency in each tumor type cohort was counted.

PDX RNA-Seq	Number of models/samples	TCGA RNA-Seq	Number of samples (patients)
Lung adenocarcinoma (PDX_LUAD)	28	Lung adenocarcinoma (TCGA_LUAD)	517 (515)
Lung squamous cell carcinoma (PDX_LUSC)	17	Lung squamous cell carcinoma (TCGA_LUSC)	501 (501)
Colorectal cancer (PDX_Colorectal)	48	Colorectal adenocarcinoma (TCGA_COADREAD)	626 (623)
Triple negative breast cancer (PDX_TNBC)	15	Breast invasive carcinoma, Triple-negative (TCGA_BRCA_TNBC)	140 (139)
Bladder cancer, Muscle invasive (PDX_BLCAinvasive)	14	Bladder urothelial carcinoma (TCGA_BLCA)	408 (408)
Melanoma (PDX_Melanoma)	11	Skin cutaneous melanoma (TCGA_SKCM)	472 (469)

Supplementary Table S10
Number of PDX and TCGA samples for 6 tumor types used for analysis of RNA-Seq expression profiling.

	Up-regulated (Adjusted p-value < 0.05, logFC > 1)					Down-regulated (Adjusted p-value < 0.05, logFC < -1)				
	TCGA	PDX	Overlap	Jaccard index	Hypergeometric p-value	TCGA	PDX	Overlap	Jaccard index	Hypergeometric p-value
Lung adenocarcinoma (LUAD)	952	681	323	0.165	3.33E-241	296	267	19	0.033	1.01E-07
Lung squamous cell carcinoma (LUSC)	779	1230	454	0.184	0.00E+00	232	363	26	0.042	1.49E-12
Colorectal cancer (COADREAD)	981	796	290	0.140	2.32E-170	1149	1599	379	0.121	3.50E-128
Triple negative breast cancer (TNBC)	625	1011	264	0.139	1.02E-169	430	504	42	0.043	1.97E-12
Bladder cancer, Muscle invasive (BLCA)	394	293	89	0.115	1.15E-77	613	106	22	0.030	7.61E-12
Melanoma (SKCM)	856	1051	293	0.133	4.13E-154	1540	1402	512	0.148	7.87E-212

Total genes = 17,978

Supplementary Table S11

Number of genes that are up-regulated (adjusted p-value < 0.05, log (fold change of TPM+1) > 1 by limma) or down-regulated (adjusted p-value < 0.05, log (fold change of TPM+1) < -1 by limma) for each tumor types versus all other tumor types for PDX and TCGA RNA-Seq data respectively. This table shows the overlap of each set of genes between PDX and TCGA RNA-Seq data.

PDX CNV	Number of models/samples	TCGA CNV	Number of tumors/samples
Lung adenocarcinoma (PDX_LUAD)	30	Lung adenocarcinoma (TCGA_LUAD)	516
Lung squamous cell carcinoma (PDX_LUSC)	27	Lung squamous cell carcinoma (TCGA_LUSC)	501
Colorectal cancer (PDX_Colorectal)	60	Colorectal adenocarcinoma (TCGA_COADREAD)	616
Glioblastoma multiforme (PDX_GBM)	27	Glioblastoma multiforme (TCGA_GBM)	577
Triple negative breast cancer (PDX_TNBC)	22	Breast invasive carcinoma, Triple-negative (TCGA_BRCA_TNBC)	136
Bladder cancer, Muscle invasive (PDX_BLCAinvasive)	17	Bladder urothelial carcinoma (TCGA_BLCA)	408
Melanoma (PDX_Melanoma)	15	Skin cutaneous melanoma (TCGA_SKCM)	367
Ovarian carcinoma (PDX_OVcarcinoma)	12	Ovarian serous cystadenocarcinoma (TCGA_OV)	579

Supplementary Table S12
Number of PDX and TCGA samples for 8 tumor types used for analysis of copy number and LOH predicted from SNP array.