Assessing the Gene Regulatory Landscape in 1,188 Human Tumors

Calabrese C[1,*,$], Lehmann K[2,3,4,*,$], Urban L[1,5,*,$], Liu F[7,$], Erkek S[5], Fonseca NA[1], Kahles A[2,3], Kilpinen H[10,11], Markowski J[6], PCAWG Group 3, Waszak SM[5], Korbel JO[5], Zhang Z[7], Brazma A[1,#], Rätsch G[2,3,4,8,9,#], Schwarz RF[1,6,#], Stegle O[1,5,#]

[*] These authors contributed equally and are listed in alphabetical order.
[$] First authors.
[#] Last authors in alphabetical order.

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
[2] Department of Computer Science, ETH Zürich, Switzerland
[3] Memorial Sloan Kettering Cancer Center, New York, USA
[4] University Hospital Zurich, Switzerland
[5] European Molecular Biology Laboratory, Heidelberg, Germany
[6] Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany
[7] Beijing Advanced Innovation Center for Genomics and College of Life Sciences, Peking University, China
[8] Department of Biology, ETH Zürich, Switzerland
[9] Weill Cornell Medical College, New York, USA
[10] UCL Great Ormond Street Institute of Child Health, University College London, UK
[11] Wellcome Trust Sanger Institute, Hinxton, UK

Abstract

Cancer is characterised by somatic genetic variation, but the effect of the majority of non-coding somatic variants and the interface with the germline genome are still unknown. We analysed the whole genome and RNA-Seq data from 1,188 human cancer patients as provided by the Pan-cancer Analysis of Whole Genomes (PCAWG) project to map *cis* expression quantitative trait loci of somatic and germline variation and to uncover the causes of allele-specific expression patterns in human cancers. The availability of the first large-scale dataset with both whole genome and gene expression data enabled us to uncover the effects of the non-coding variation on cancer. In addition to confirming known regulatory effects, we identified novel associations between somatic variation and expression dysregulation, in particular in distal regulatory elements. Finally, we uncovered links between somatic mutational signatures and gene expression changes, including *TERT* and *LMO2*, and we explained the inherited risk factors in APOBEC-related mutational processes. This work represents the first large-scale assessment of the effects of both germline and somatic genetic variation on gene expression in cancer and creates a valuable resource cataloguing these effects.

Introduction

Cancer is characterised by extensive somatic genetic alterations. These variations can result in important cellular phenotypes with relevance for disease, including uncontrolled proliferation, immune evasion and metastasis (Weir, Zhao, and Meyerson 2004; Knudson 2002). Somatic mutagenesis is in part explained by environmental and intrinsic risk factors, with growing evidence for the relevance of the germline background of the patient (Pleasance et al. 2010; Jia, Pao, and Zhao 2014; Saini et al. 2016; Nik-Zainal et al. 2014). However, the interplay and the functional relevance of these different genetic factors are not sufficiently known.

One strategy to shed light on this question is association analysis with molecular readouts such as gene expression levels. Previous efforts using exome- and transcriptome sequencing data from The Cancer Genome Atlas (TCGA, Cancer Genome Atlas Research Network, Weinstein, et al. 2013) and the International Consortium of Cancer Genomes (ICGC, J. Zhang et al. 2011) have identified associations between somatic variants in coding regions and gene expression. Although these studies have helped to identify and characterise regulatory drivers, the role of the much larger number of non-coding somatic variants is not fully understood (Cancer Genome Atlas Research Network, Kandoth, et al. 2013; Kanchi et al. 2014). Recent studies have begun to address this by identifying genomic loci that are recurrently altered by somatic mutations, including their linkages to gene expression levels (Weinhold et al. 2014; Fredriksson et al. 2014; Smith et al. 2015; Ding et al. 2015). Additionally, substantial work has focused on the effects of variation in promoters of established cancer-genes, including *TERT* and *BCL2 (*Weinhold et al. 2014; Fredriksson et al. 2014; Smith et al. 2015; Ding et al. 2015). However, thus far, a comprehensive analysis of associations between non-coding somatic variation and gene expression is missing.

We carried out joint genetic analyses that integrate coding and non-coding somatic variation with germline variants to investigate regulatory effects on gene expression levels in 27 cancer types. Building on 1,188 consistently processed genomes (whole genome sequencing, WGS) and transcriptomes (RNA-Seq) from the Pan-cancer Analyses of Whole Genomes (Campbell et al. 2017; PCAWG Consortium 2017) project, we have derived a detailed regulatory map that integrates different dimensions of germline and somatic variation, including single nucleotide variants (SNVs), somatic copy-number alterations (SCNAs) and signatures that capture differences in the prevalence of mutational processes across patients. Our approach combines complementary strategies, including allele-specific expression (ASE) analyses, somatic and germline expression quantitative trait loci (eQTL) mapping and the analysis of gene expression associations with global somatic signatures (**Fig. 1**).

Collectively, our analyses provide a comprehensive picture of the regulatory landscape in human cancers, thereby identifying previously underappreciated associations between somatic regulation in distal regulatory elements and gene expression, as well as *de novo* functional annotations of mutational signatures. We report several associations that involve cancer-testis (CT) genes with known immunogenic properties (R.-F. Wang and Wang 2017;

Scanlan et al. 2002), which exhibit high expression in sperm and some cancers but are repressed in healthy tissues (Simpson et al. 2005). Our results point at non-coding somatic dysregulation as a functional driver of carcinogenesis beyond mutations of the coding sequence and major contributor to inter- and intra-tumor heterogeneity.

Results

We investigated 1,188 cases of 27 different tumor types obtained as part of the PCAWG working group (Campbell et al. 2017; Yung et al. 2017; PCAWG Consortium 2017), with WGS and RNA-Seq data (Yung et al. 2017; Whalley et al. 2017; Fonseca et al. 2017; PCAWG Group 1 2017; PCAWG Group 8 2017; PCAWG Group 3 2017, **Methods**). Across this panel, we quantified gene expression levels for 18,898 protein-coding genes (FPKM ≥ 0.1, in at least 1% of the patients, **Methods**), and allele-specific expression (**Methods**, **Fig. 1**).

Germline Regulatory Variants in Cancer

We considered common germline variants (MAF≥1%) proximal to individual genes (±100kb around the gene) to map eQTL across the entire cohort (**Fig. S2 A**). This pan-cancer analysis identified 3,509 genes with an eQTL (FDR ≤ 5%, hereafter denoted eGenes; **Methods, Table S2**), enriched in transcription start site (TSS) proximal regions as expected from eQTL studies in normal tissues (Bryois et al. 2014**, Fig. S2 B**). Analogous tissue-specific eQTL analyses in seven cancer types that were represented with at least patients identified between 106 (Breast-AdenoCA) and 472 eGenes (Kidney-RCC) (**Fig. S3 A, Table S2, Methods**).

To identify regulatory variants that are cancer-specific, we compared our eQTL set to eQTL maps from normal tissues obtained from the GTEx project (GTEx Consortium et al. 2017), adapting the strategy devised in Kilpinen et al. (2017). For each eQTL lead variant, we assessed the marginal replication in GTEx tissues (P<0.01, Bonferroni adjusted for 42 somatic tissues excluding cell lines, using proxy variants r^2>0.8 for missing variants, **Methods**). For 87.5% of the eQTL that could be assessed in at least one GTEx tissue (2,982 of altogether 3,408 eQTL, **Fig. 2D**), this identified a replicating eQTL in at least one GTEx tissue, whereas 426 eQTL did not replicate in GTEx tissues, indicating cancer-specific regulation of gene expression (**Table S3**). One such example is *SLAMF9*, a member of the CD2 subfamily, with known roles in immune response and cancer (X. A. Zhang et al. 2003, **Fig. S4 A**). Similarly, we identified cancer-specific eQTL for genes with known roles in cancer such as *SLX1A*, an important regulator of genome stability that is involved in DNA repair and recombination and is associated with Fanconi Anemia (Saito et al. 2012; Medves et al. 2016, **Fig. S4 B**). The majority of these cancer-specific regulatory variants could not be explained by differences in gene expression level between cancer and normal tissues (328/426 exhibit at most a 2-fold increase in median gene expression compared to GTEx, e.g. *SLX1A*, **Fig. S3 B**), with a remaining set of 98 genes showing clear evidence of upregulation or ectopic expression (e.g. *SLAMF9*). This latter group contained

immunoglobulin genes and nine CT genes including *RAD21L* (**Fig. 2, Fig. S3 B**). We also identified instances of eQTL that replicated in GTEx tissues but not in their corresponding normal tissues. One such example is *TEKT5*, which is expressed in our cohort but otherwise specific to testis in GTEx normal tissues, pointing to upregulation of selected genes in cancer (**Fig. S3 C-E**).

Allele-specific Expression Captures Cancer-specific Dysregulation

To robustly identify genetic elements that contribute to somatic dysregulation of gene expression, we considered ASE, a locally controlled readout that enables assessing differential regulation between haplotypes in the same patient (Korir and Seoighe 2014). We quantified ASE at heterozygous germline variants, following Castel et al. (2015) (**Methods)**. To maximise detection power, we aggregated ASE counts across heterozygous sites within genes, leveraging phased germline variant maps derived using a combination of statistical, copy-number (CN) and read-based variant phasing (**Methods**). This allowed quantification of ASE for between 588 and 7,728 genes per patient (median of 4,112 genes with at least 15 ASE reads in 1,120 patients, **Fig. S5**, **Methods**).

We tested for allelic expression imbalance (AEI) (FDR ≤ 5%, binomial test, **Methods**), finding substantial differences in the fraction of genes with AEI between cancer types (median percentages between 14.2% in Prost-AdenoCA and 46.8% in Lung-SCC among tumors; $P=2.2x10^{-13}$ Mann-Whitney-Wilcoxon Test, **Fig. 3A**), and between AEI in cancer and the corresponding normal tissue (**Fig. 3B**). Cancers with extensive chromosomal rearrangements, including lung, breast and ovarian cancers, were associated with most frequent AEI events (**Fig. 3A, Fig. S6**), which is consistent with previous reports that have implicated SCNAs in allelic dysregulation in cancers (Ha et al. 2012).

Motivated by this, we used logistic regression to model the determinants of AEI (**Methods**), accounting for the presence or absence of germline eQTL, local allele-specific SCNAs and the mutational burden of proximal somatic SNVs, weighted by their respective cancer cell fraction and stratified into functional categories (upstream, downstream, promoter, 5'UTR, intron, synonymous, missense, stop gain, 3'UTR, **Fig. 1**, **Methods**). In aggregate, SCNAs accounted for 86.14% of the explained AEI variability, followed by germline eQTL (9.03%) and somatic SNVs (4.83%) (**Fig. 3C**). While cumulatively, non-coding variants were more relevant than coding variants (**Fig. 3C**), somatic protein truncating variants ('stop-gained') were the most predictive individually (**Fig. 3D**), which confirms the importance of nonsense-mediated decay (NMD) in cancer gene regulation (Lindeboom, Supek, and Lehner 2016). SNVs within splice regions, 5' UTR and promoters were also strongly associated with AEI presence and we observed a global trend of decreased relevance of variants as a function of the distance from the TSS (**Fig. 3D**). We also considered a quantitative model on ASE ratios using phased variants as features, confirming downregulation of allelic expression by NMD (**Fig. S7 A-D**).

Using the trained model for AEI, we set out to characterise sets of genes with strong allelic dysregulation that can be attributed to different different genetic factors. We ranked genes according to average scores across the cohort, based on (i) the predicted AEI from the

germline component and (ii) the predicted AEI from somatic components (SCNAs and SNVs) without germline effects. For comparison, we also considered (iii) the empirical AEI frequency in the cohort; and (iv) the burden of loss-of-function (LoF) and gain-of-function (GoF) mutations derived from genetic data only (**Fig. 3E**). When assessing these rank lists using known cancer genes (COSMIC, Forbes et al. 2017), we found cancer genes to be enriched among genes with high somatic AEI score (P≤0.005, Gene Set Enrichment Analysis, Subramanian et al. 2005, **Fig. S8 C**), however we observed no enrichment among genes with recurrent AEI (P=0.99, **Fig. S6 A**). As expected, genes with AEI due to germline eQTL were depleted for cancer genes (negative enrichment, P≤0.001, **Fig. S8 B**). Finally, consistent with the traditional definition of cancer genes based on recurrent mutations (Forbes et al. 2017), genes with recurrent LoF/GoF mutations were most strikingly enriched for the COSMIC census (P≤0.001, **Fig. S8 D**). The top 10% of genes of the somatic AEI scores were enriched for Gene Ontology (GO) categories with relevance to cancer, including chemotaxis, cell motility, locomotion and cell migration, which notably were absent when considering LoF/GoF mutations (**Fig. 3F**, Ashburner et al. 2000; Gene Ontology Consortium 2015a). These results suggest that somatic AEI could be used to prioritise regulatory variants to identify genes with roles in cancers, which extends previous observations in a single cancer (Ongen et al. 2014) to a pan-cancer setting.

Due to the strong effect of SCNAs on AEI we specifically investigated genes that were primarily dysregulated by SNVs. Of the 4,007 genes in the upper quartile based on the prevalence of overall AEI across all tumors, 1,843 genes exhibited SNV-linked AEI. When we ranked these genes based on the predicted AEI from SNVs, the top 10 genes were *FBXO5*, *ASPM*, *PSCA*, *CDKN1A*, *KIF20B*, *TP53* and *CLDN4,* which have previously been linked to cancer (Z. Wang et al. 2014; W.-Y. Wang et al. 2013; Gu et al. 2000; Abbas and Dutta 2009; Liu, Gong, and Huang 2013; Shang et al. 2012; Olivier, Hollstein, and Hainaut 2010), but also *EXO1*, *SYNE1* and *STON1-GTF2A1L*, genes that have not been prominently linked to cancer. *SYNE1* controls nuclear polarity and spindle orientation, which is upstream of NOTCH signaling in squamous lineage development (Lasorella, Benezra, and Iavarone 2014; Garraway and Lander 2013). Notably, based on the CADD analysis (Kircher et al. 2014), three melanoma cases preferentially expressed deleterious missense mutations (all CADD scores >25) in *SYNE1*, likely leading to a relative decrease of gene expression in these tumors. Based on the GEPIA web server (Tang et al. 2017), we also found that low *SYNE1* expression was associated with worse overall survival in TCGA melanoma patients (Log rank P=0.002, Hazard ratio=0.57, **Fig. S9 A**), providing further support for its relevance in disease. *EXO1* is known to be involved in mismatch repair and recombination, and exhibited significant AEI for a deleterious missense (CADD score 34, SIFT score 0, Kircher et al. 2014; Kumar, Henikoff, and Ng 2009) and a nonsense mutation in a colorectal adenocarcinoma patient. Similarly, TCGA colorectal adenocarcinoma patients with lower expression of *EXO1* showed worse overall survival (Log rank P=0.022, HR=0.57, **Fig. S9 B**), implicating *EXO1* as a potential tumor suppressor in human colorectal cancer. Consistent with this finding, *EXO1* knockout mice exhibited defects in DNA damage response and increased tumorigenesis (Schaetzlein et al. 2013).

Motivated by the observed cancer-specific germline regulation of CT genes, we investigated AEI in these genes. Notably, CT genes were depleted when considering the full somatic score (SCNAs and SNVs, 25/476 CT genes in the top 10% of genes, 48 expected,

chi-square test, $P=6x10^{-4}$), but enriched in the AEI score based on SNVs (66/476 CT genes in the top 10% of genes, 48 expected, chi-square test P=0.006). One potential explanation is that CT genes with low or no expression in differentiated tissues have to undergo somatic re-activation by SNVs before any subsequent SCNA can have an effect. To elucidate this, we used mutation timing data (Gerstung et al. 2017; PCAWG Group 11 et al. 2018, **Methods**), stratifying SNVs into the categories *early* (SNV occurred before SCNA at the same locus) and *late* (SNV occurred after SCNA at the same locus). We found strong over-representation of *early* SNVs in 329 out of 7525 CT gene-patient pairs (216 expected, chi-square test, $P=3.96x10^{-14}$), suggesting that somatic re-activation of developmental genes through SNVs is selected for and precedes SCNA events at the same locus.

Somatic eQTL Mapping Reveals Widespread Associations with Non-coding Variants

We next explored the effect of *cis* somatic variation on gene expression by aggregating somatic SNVs using local burdens in intronic and exonic genomic intervals (Gencode annotations, Harrow et al. 2012) and in consecutive genomic intervals within 1Mb from the gene boundaries (2kb regions, 1kb overlap, **Methods**), hereafter denoted as flanking regions.

Initially, we used these somatic burdens in conjunction with *cis* germline variants and SCNAs to decompose variation in gene expression of individual genes into their underlying components (**Methods, Fig. 4A**). Consistent with the ASE analysis, this identified SCNAs as the major source of variation (27.3% on average across all genes, **Fig. 4A**), followed by flanking somatic and germline variants. Notably, *cis* germline effects, although exhibiting smaller effects on average, were the largest variance component for 11,905 genes, compared to 3,568 genes which variation was primarily explained by somatic factors. Consistent with ASE analysis, we observed that non-coding somatic variants had more explanatory power than variants in coding regions.

Next, we tested for associations between recurrently mutated intervals (burden frequency ≥ 1%) and expression levels of individual genes (18,708 protein coding genes; median 952 intervals per gene, **Fig. S10 C**), accounting for local and global differences in mutational burden across patients as well as tumor purity, cancer type, local SCNAs and other technical covariates (**Methods**). We assessed alternative strategies to estimate mutational burdens and found that weighted burden that accounts for variant clonality maximised detection power (**Fig. S11 A-D**). Genome-wide, this identified 649 somatic eQTL (FDR ≤ 5%; **Table S6**) in 567 genomic regions. Among these, 11 somatic eQTL were explained by the mutational burden in exons or introns, including genes with known roles in the pathogenesis of specific cancers such as *CDK12* in ovarian cancer (Bajrami et al. 2013; Ekumi et al. 2015)*, PI4KA* in hepatocellular carcinoma (Ilboudo et al. 2014), *IRF4* in leukemia (Havelange et al. 2011)*, AICDA* in skin melanoma (Nonaka et al. 2016)*, C11orf73* in clear cell renal cancer (Bhalla et al. 2017) and *BCL2* and *SGK1* in lymphoma (Weinhold et al. 2014; Hartmann et al. 2016**, Fig. S12 A-G**). For the majority of 444 eGenes (70%), we observed associations with flanking non-coding regions (272 intergenic and 172 intronic regions). Hereby, 43.6% of the 567 unique genomic regions were intergenic and did not overlap any

other annotated gene feature (**Fig. 4B)**. The associated elements were generally mutated in two or more cancer types (**Fig. S13, Table S6**). Unlike germline eQTL, the associations tended to be located in distal regions (≥20kb, 88%) with on average larger effect sizes ($|\beta|$=3.3) than associations in regions proximal to the TSS ($|\beta|$=1.4) (**Fig. S14**), which points to the relevance of somatic mutations at distal regulatory elements.

Motivated by this, we tested lead flanking regions for enrichments in cell type specific regulatory annotations in 127 cell types from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al. 2015) as well as transcription factor binding sites (TFBS) from ENCODE (8 cancer cells and one embryonic stem cell line, ENCODE Project Consortium 2012), comparing the overlap of true associations to distance and burden frequency matched random regions (**Methods**). This identified a significant enrichment (FDR ≤ 10%) of 13 out of 25 epigenetic annotations (**Fig. 4C**, **Table S7**), including poised promoters, weak and active enhancers and heterochromatin in more than two cell lines (**Fig. 4C**), but no significant enrichment of TFBS (**Tables S8**).

Poised or bivalent promoters are a hallmark of developmental genes and prepare stem cells for somatic differentiation (Lesch and Page 2014). Re-activation of poised promoters is one mechanism of upregulation of developmental genes in cancer, including CT genes (Bernhart et al. 2016). CT genes were marginally more frequent among genes with somatic eQTL than expected (45/982, P=0.06, Fisher's exact test), and we observed an enrichment for eQTL in bivalent promoters for the CT genes (P=0.04, Fisher's exact test). One such gene is *TEKT5,* an integral component of sperm, that has been found to be aberrantly expressed in a variety of cancers (Hanafusa et al. 2012). We observed a positive association between *TEKT5* expression and somatic mutational burden (prevalently observed in non-Hodgkin lymphoma patients) in a bivalent promoter site close to the 5' end of the gene (**Fig. 4D**). The prevalence of developmental genes among the somatic eGenes was also consistent with a global enrichment (FDR ≤ 10%) for GO categories related to cell differentiation and developmental processes (**Table S9**). Together with the ASE analysis, these results emphasise the relevance of non-coding genetic variation for changes in gene expression and the importance of re-activation of developmental genes through somatic mutations for cancer progression.

Associations between global somatic mutational signatures and gene expression

In addition to associations with *cis* elements proximal to genes, it is plausible that global differences in mutational states between patients are associated with cell phenotypes, including gene expression levels. Global variations in mutational patterns can be quantified using mutational signatures which tag mutational processes specific to their tissue-of-origin and environmental exposure (Alexandrov et al. 2013). However, the relationship between mutational signatures and gene expression levels is unknown.

We considered 28 mutational signature readouts (Forbes et al. 2017) derived using non-negative matrix factorization of context-specific mutation frequencies in the PCAWG cohort (PCAWG-7 beta 2 release, PCAWG Group 7 et al. 2018). We tested for associations

between patient-specific signature prevalence and expression levels of individual genes, accounting for total mutational burden and additional technical as well as biological confounders (18,831 genes, FPKM filtering of genes across 1,159 patients, **Methods**). Across all signatures, this identified 1,176 genes associated with at least one signature (FDR ≤ 10%, **Fig. S15, Methods**), a markedly different set of genes compared to genes associated with the total mutational burden that ignores the assignment to different signatures (**Table S10 F**). Lymphoma Signature 9 was associated with the largest set of genes, followed by the smoking-related Signature 4 (**Fig. 5A, Tables S10 A,E, Fig. S16 D**).

While aetiologies for some mutational signatures are well understood, others have not previously been characterised. To annotate the signatures, we considered 18 signatures with at least 20 associated genes and tested for enrichments using GO and Reactome Pathways (Fabregat et al. 2016; Milacic et al. 2012) categories. Of these, 11 signatures were enriched for at least one category (FDR ≤ 10%, **Table S10 D**), revealing several associations that were consistent with known aetiologies (**Fig. 5A**). For example, Signature 9 is known to be active in certain lymphomas and leukemias (Forbes et al. 2017), and was associated with 354 genes enriched for lymphocyte/leukocyte-related processes and immune response, including *TCL1A*, *LMO2* and *TERT* ($P=1.22x10^{-10}$, $6.84x10^{-10}$, and $1.96x10^{-09}$). Smoking Signature 4 (Forbes et al. 2017) was associated with 119 genes, enriched for biological oxidation associated processes (e.g. processing of benzo[a]pyrene) and including the gene *CYP24A1*, which is known to be down-regulated in tobacco-smoke exposed tissue (**Fig. 5B**, Woenckhaus et al. 2006). We further identified 70 genes associated with the APOBEC Signature 2 (Forbes et al. 2017), which were significantly enriched for DNA deaminase pathways.

Among those signatures with previously unknown aetiology, our results link Signature 8, which is primarily prevalent in medulloblastoma (Forbes et al. 2017), to a set of 25 genes enriched for ABCA-transporter pathways, which are targeted by drugs that are in clinical trials for medulloblastoma (Milacic et al. 2012; Ingram et al. 2013). Similarly, Signature 38, which is correlated with the canonical UV Signatures 7 (e.g. 7a: $r^2=0.375$, $P=5x10^{-40}$, **Fig. S16 C**), was linked to melanin processes (**Fig. 5A**). Melanin synthesis is known to subject melanocytes to oxidative stress by involving main oxidation reactions and superoxide anion/hydrogen peroxide generation (Kvam and Tyrrell 1999; Denat et al. 2014). Our data linked Signature 38 to the oxidative stress promoting gene Tyrosinase (*TYR,* $P=1x10^{-4}$, Jimbow et al. 2001). As Signature 38 is characterised by an excess of C>A mutations, a typical product of reactive oxygen species (ROS) mediated by activity of 8-hydroxy-2′-deoxyguanosine (DFG, 2010), it might represent DNA damage indirectly caused by UV after direct sun exposure due to oxidative damage (Premi and Brash 2016), with *TYR* as a possible damage mediator.

The cause-and-effect relationship of correlated somatic variations and gene expression changes are not clear *a priori*. To begin addressing the directionality of these linkages, we explored the utility of germline variation as an anchor to gain directed mechanistic insights. Building on our eQTL map, we queried germline eQTL lead variants of mutational signature-associated genes and tested for associations between these variants and the corresponding mutational signature. This eQTL-guided approach entails substantially fewer tests than genome-wide analyses that consider all germline variants (Waszak et al. 2017;

PCAWG Group 8 2017). Among 1,176 signature-linked genes, 197 also had a germline eQTL, but only eQTL variant rs12628403 was associated with the respective signature (FDR ≤ 10%, multiple testing over 197 association tests, **Table S10 E**). This germline variant is a known germline predisposition factor for Signature 2 prevalence (Middlebrooks et al. 2016), an eQTL for *APOBEC3A/B* and significantly associated with Signature 2 prevalence in our cohort ($P=5.13x10^{-7}$, **Fig. 5C**). Colocalisation analysis (Wallace 2013) confirmed that the variant rs12628403 is a plausible genetic determinant of both, expression levels of *APOBEC3A/B* and Signature 2 (**Table S10 E, Methods**). Finally, we carried out mediation analysis (Baron and Kenny 1986; Preacher and Hayes 2004) to formally test whether expression levels of *APOBEC3A/B* confer the genetic effect to the signature (**Methods**). We found that *APOBEC3B, POTEI* (both $P<10^{-10}$) and *APOBEC3A* ($P=0.004$) expression levels conferred the effect of rs12628403 to Signature 2. The proportion of the effect of the variant on the signature that is mediated by only *APOBEC3B* expression is a remarkable 87.11% (nonparametric bootstrapping, 1,000 simulations, **Fig. S17**, **Methods**).

Discussion

The regulatory landscape of cancer is highly heterogeneous, cancer type specific and influenced by the germline background. This study provides a comprehensive picture of how different germline and somatic variations alter gene expression levels. Our results show that coregulation of the same genes by multiple different types of variants is common in cancer (**Fig. S1**). Here, we have assessed the relative magnitudes of these effects (**Fig. 4A**). Previous studies have been limited by the lack of whole genome sequencing data, which is essential for identifying contributions of non-coding variants to gene expression variability. Indeed, our analysis which is based on data from the currently largest cohort of matched tumor WGS and RNA-Seq data of 1,188 patients, demonstrates that the impact of non-coding variation can be profound.

We have produced comprehensive across-tissue and tissue-specific germline eQTL maps and have identified associated genetic risk variants. By comparing these cancer eQTL maps to the GTEx catalogue, we have observed substantial overlap between cancer and normal tissues, but also a smaller number (12.14%) of potentially interesting cancer-specific eQTL. This estimate is conservative since we have assessed replication in any GTEx tissues and not only in the corresponding normal tissues. In selected cases we have observed that eQTL in cancer mimic regulatory effects in other non-corresponding tissues, in particular testis-specific genes (**Fig. 2A**).

In parallel to germline eQTL, we have considered the effect of somatic mutational burden in different genomic regions on gene expression changes by building a systematic map of somatic eQTL. Our analysis accounts for variation in clonality as well as local hypermutations, thereby identifying likely causal associations between somatic burden and gene expression. However, this approach has limitations and we cannot rule out that a fraction of the associations we have identified are due to technical or biological factors that

jointly affect gene expression and local mutation rates. We also note that an analysis of cancer type specific somatic eQTL is currently not feasible with the given sample size.

We have used ASE readouts for integrated modelling of genetic variation in *cis* and fine-grained characterisation of the genetic elements that have the largest regulatory effects. We have demonstrated the extent to which AEI follows allelic imbalance on the genomic level. While ASE is sensitive to heterozygous genetic variation, the considered phased somatic mutation set is based on read phasing, which has only been possible for around 20% of all SNVs. Further, ASE readouts can only be derived in cases with at least one heterozygous germline variant in the gene in question, reducing overall sample size and hindering gene-level associations.

Our analysis suggests somatic re-activation of CT genes through SNVs, supported by mutation timing analysis. CT genes were also enriched in cancer-specific germline and somatic eGenes linked to SNV burden in nearby bivalent promoters. Somatic eQTL were further enriched for tissue development and differentiation. Due to the low number of expressed CT genes these findings need to be evaluated on additional data. However, we have found CT gene implicated in three out of four independent analyses (germline eQTL, somatic eQTL and ASE), reinforcing the potential impact of these findings.

Mutational signatures capture global variations across individuals, for example due to environmental factors or exogenous damage, which are distinct from local somatic variations assessed using QTL mapping and ASE. We have explored the utility of associations between these mutational signatures and gene expression levels, thereby deriving *de novo* annotations of signatures with previously unknown roles. Finally, we have carried out proof of concept analyses to integrate germline factors, somatic variants and gene expression, thereby unpicking the molecular chain of events for the common APOBEC mutational process and its germline component. Due to the tissue specificity of mutational signatures, it will be important to conduct similar analyses for individual cancer types and in cohorts with larger sample sizes.

This is the first large-scale study assessing the effects of both germline and somatic genetic variation on gene expression from WGS data in a pan-cancer setting. The somatic and germline eQTL resources will be a valuable resource to address a wide range of downstream analyses, providing a comprehensive overview of gene expression determinants in cancer and insights into the underlying biology. The systematic assessment of regulatory non-coding genetic variation significantly improves our understanding of the aetiology and functional implications of intra- and inter-tumor heterogeneity and the selective forces applied to these heterogeneous genomes.

Acknowledgements. L.U., R.F.S. and O.S. received support from core funding of the European Molecular Biology Laboratory and the European Union's Horizon2020 research and innovation programme (grant agreement number N635290). K.L., A.K. and G.R. received core funding from Memorial Sloan Kettering Cancer Center (New York) and from ETH Zurich. R.F.S. and J.M. received support from the Helmholtz Foundation and the Max Delbrueck Center for Molecular Medicine. F.L. and Z.Z. received support from National Natural Science Foundation of China (grant agreement numbers 31530036 and 81573022). C.C, N.F. and A.B. received support from core funding of the European Molecular Biology Laboratory and from the EU FP7 Programme projects EurocanPlatform (grant agreement number 260791) and CAGEKID (grant agreement number 241669).

Contributions. The authors are listed according to their contributions. Senior contributors are listed in parentheses and in alphabetical order.

Data Preprocessing: Lehmann K, Calabrese C, Kahles A, Fonseca N, (Brazma A, Rätsch G, Stegle O)

Germline eQTL mapping/analysis: Lehmann K, Calabrese C, (Rätsch G, Stegle O)

Somatic eQTL mapping/analysis: Calabrese C, Lehmann K, Fonseca N, Urban L, (Brazma A, Rätsch G, Schwarz RF, Stegle O)

ASE preprocessing: Schwarz RF, Erkek S, Waszak S, (Schwarz RF, Korbel J, Stegle O, Zhang Z)

ASE analysis: Schwarz RF, Urban L, Liu F, Kilpinen H, Markowski J, (Schwarz RF, Stegle O, Zhang Z)

Signature analysis: Urban L, (Schwarz RF, Stegle O)

Germline/Signature interface: Urban L, Lehmann K, (Schwarz RF, Stegle O)

Figures

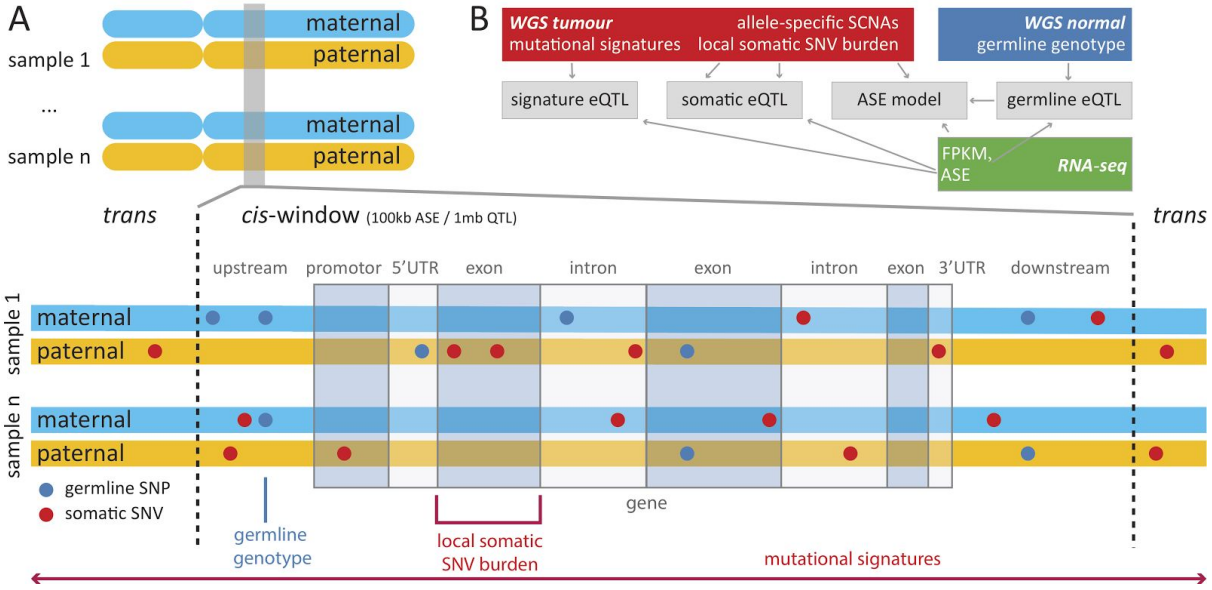

Fig. 1. Integrative analysis approach for assessing the regulatory landscape in human cancers and overview of different classes of genetic variation considered here. **A)** *Cis* genetic effects of individual germline variants (blue) on gene expression were assessed using eQTL mapping. Somatic variants (red) were aggregated in mutational burden categories and assessed using i) ASE, ii) gene level *cis* eQTL and iii) associations with mutational signatures in patients. **B)** Data sources and processing steps, including germline and somatic variant calling from WGS data and allele-specific expression quantification from RNA-seq.

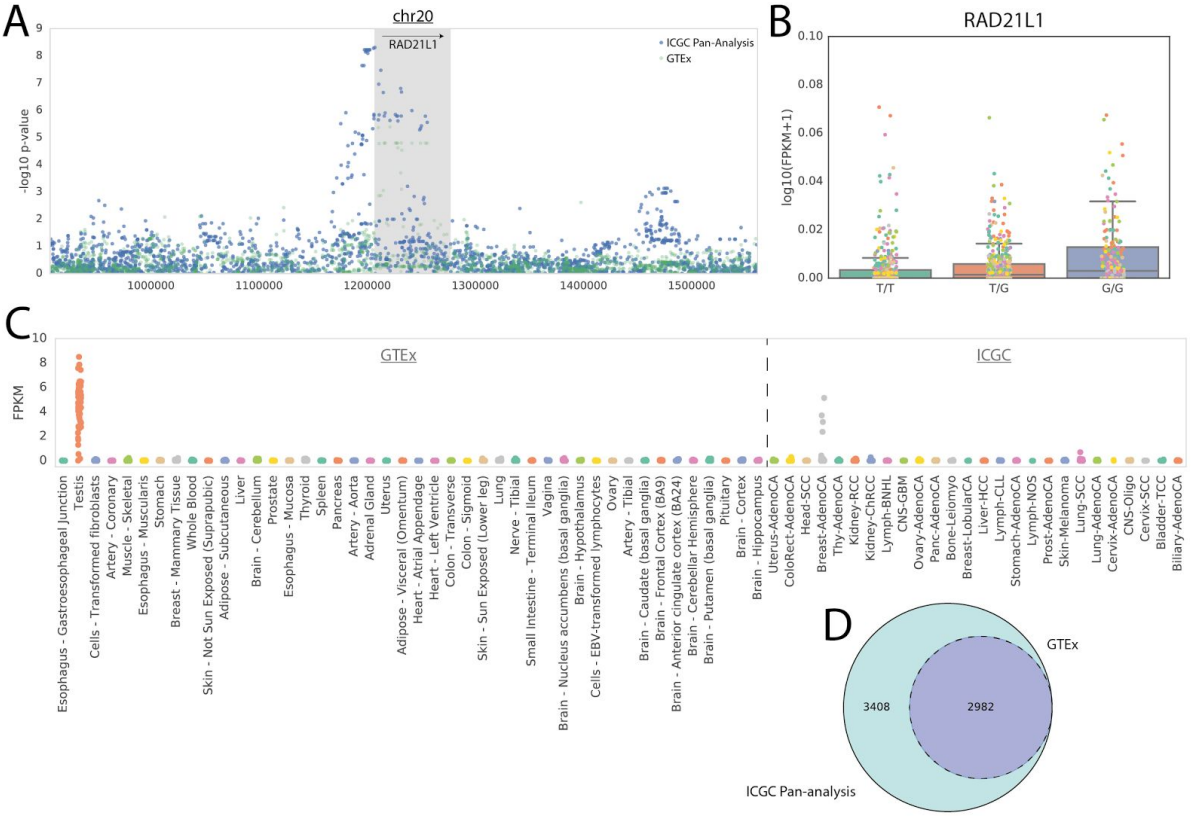

Fig. 2. Germline eQTL analysis. **A)** Manhattan plot for *RAD21L1*, showing associations in the ICGC cohort (blue) and in GTEx testis tissue (green). **B)** Boxplots of the effect of the eQTL lead variant for *RAD21L* in the ICGC cohort. Colors denote cancer types of individual patients. **C)** Gene expression distribution of *RAD21L* across GTEx tissues and ICGC cancer types. *RAD21L* expression is increased in testis tissue, breast adenocarcinoma patients ('Breast-AdenoCA') and leukemia cell lines. **D)** The Venn diagram shows the number of eQTL identified in the ICGC cohort and the fraction of QTL that replicate in the GTEx cohort. 426 eQTL were specific to the ICGC cohort.

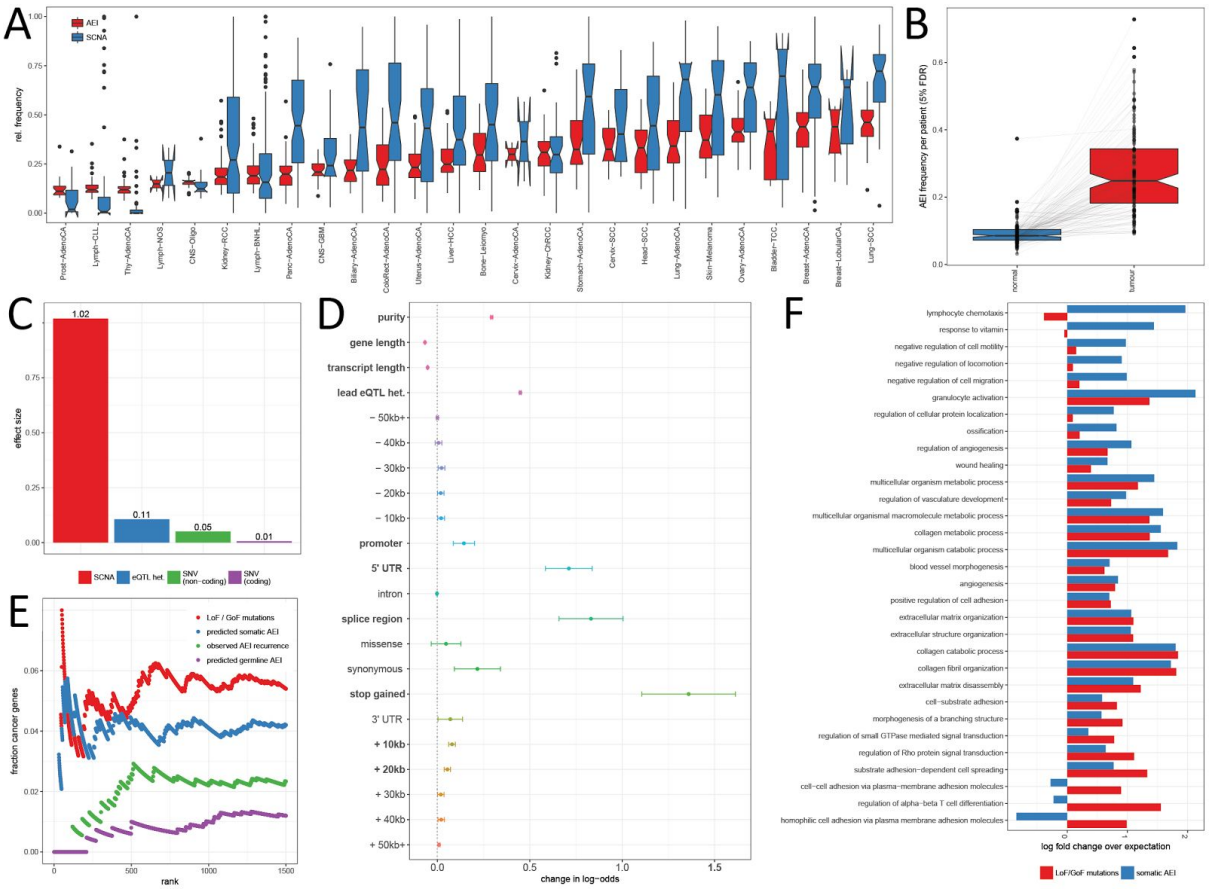

Fig. 3. Allele-specific expression analysis. **A)** Distribution of the proportion of genes with significant allele-specific imbalance (AEI, FDR ≤ 5%) (red) and SCNAs (blue) across patients for different cancer types. Cancer types with prevalent chromosomal instability (frequent SCNAs) exhibit most frequent AEI. **B)** Proportion of genes with AEI (FDR ≤ 5%) in tumor and matched normal RNAseq patients across the cohort. **C)** Effect sizes of SCNAs, germline eQTL, coding and non-coding variants on AEI status. **D)** Relevance of individual somatic mutation types, germline eQTL and other co-variates for AEI status. Significant covariates (FDR ≤ 5%) highlighted in bold. **E)** COSMIC cancer gene enrichment for the four models of gene dysregulation: (i) average observed AEI frequencies (green); (ii) predicted germline AEI (purple); (iii) predicted somatic AEI excluding germline (blue), and (iv) LoF and GoF mutations (red). **F)** Enrichment of top 10% of genes ranked according to LoF/GoF mutations (red) or predicted somatic AEI (blue), using GO ontologies (FDR ≤ 5%).

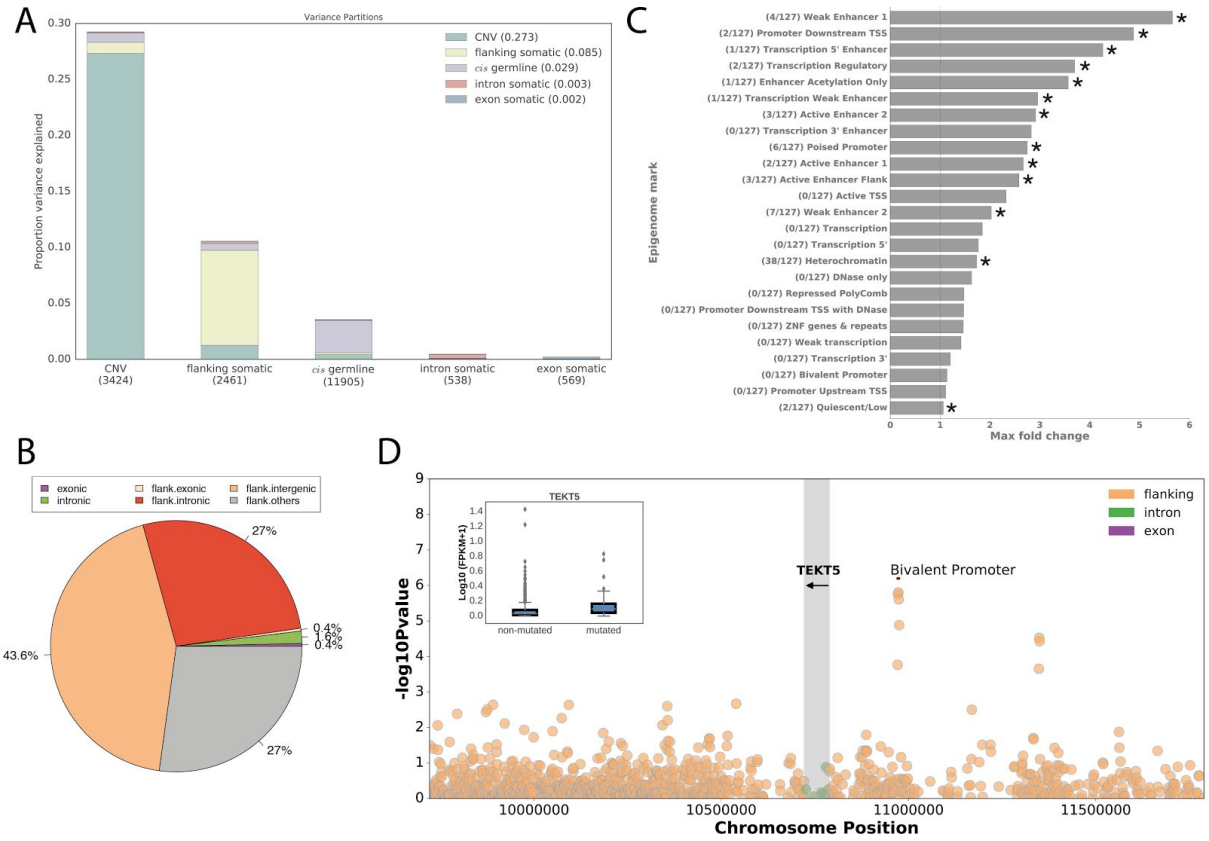

Fig. 4. Somatic burden eQTL analysis. **A)** Variance component analysis for gene expression levels (**Methods**). Shown is the average proportion of variance explained by different germline and somatic factors for different sets of genes, considering genes for which the largest variance component are i) copy number effects (CNV), ii) somatic variants in flanking regions, iii) *cis* germline effects and iv) somatic intron and exon mutations, respectively. The number of genes in each set is indicated in parentheses. **B)** Breakdown of 567 genomic regions that underlie the observed *cis* somatic eQTL by variant category (Intronic = eGene intron; Exonic = eGene exon; Flank. = 2kb flanking region within 1Mb distance to the eGene start; Flank.intergenic = flanking region in a genomic location without gene annotations; Flank.intronic = flanking region overlapping an intron of a nearby gene; Flank.others = flanking region partially overlapping exonic and intronic annotations of a nearby gene). **C)** Maximum fold enrichment of epigenetic marks from the Roadmap Epigenomics Project across 127 cell lines. The number of cell lines with significant enrichments is indicated in parentheses (FDR ≤ 10%); asterisks denote significant enrichments in at least one cell line. **D)** Manhattan plot showing nominal p-values of association for *TEKT5* (highlighted in gray), considering flanking, intronic and exonic intervals. The leading somatic burden is associated with increased *TEKT5* expression ($P=1.56x10^{-06}$; $\beta=0.221$) and overlaps an upstream bivalent promoter (red box; annotated in 81 Roadmap cell lines, including 8 ESC, 9 ES-derived and 5 iPSC cell lines). The inset boxplot shows a positive association between the mutation status and expression levels.

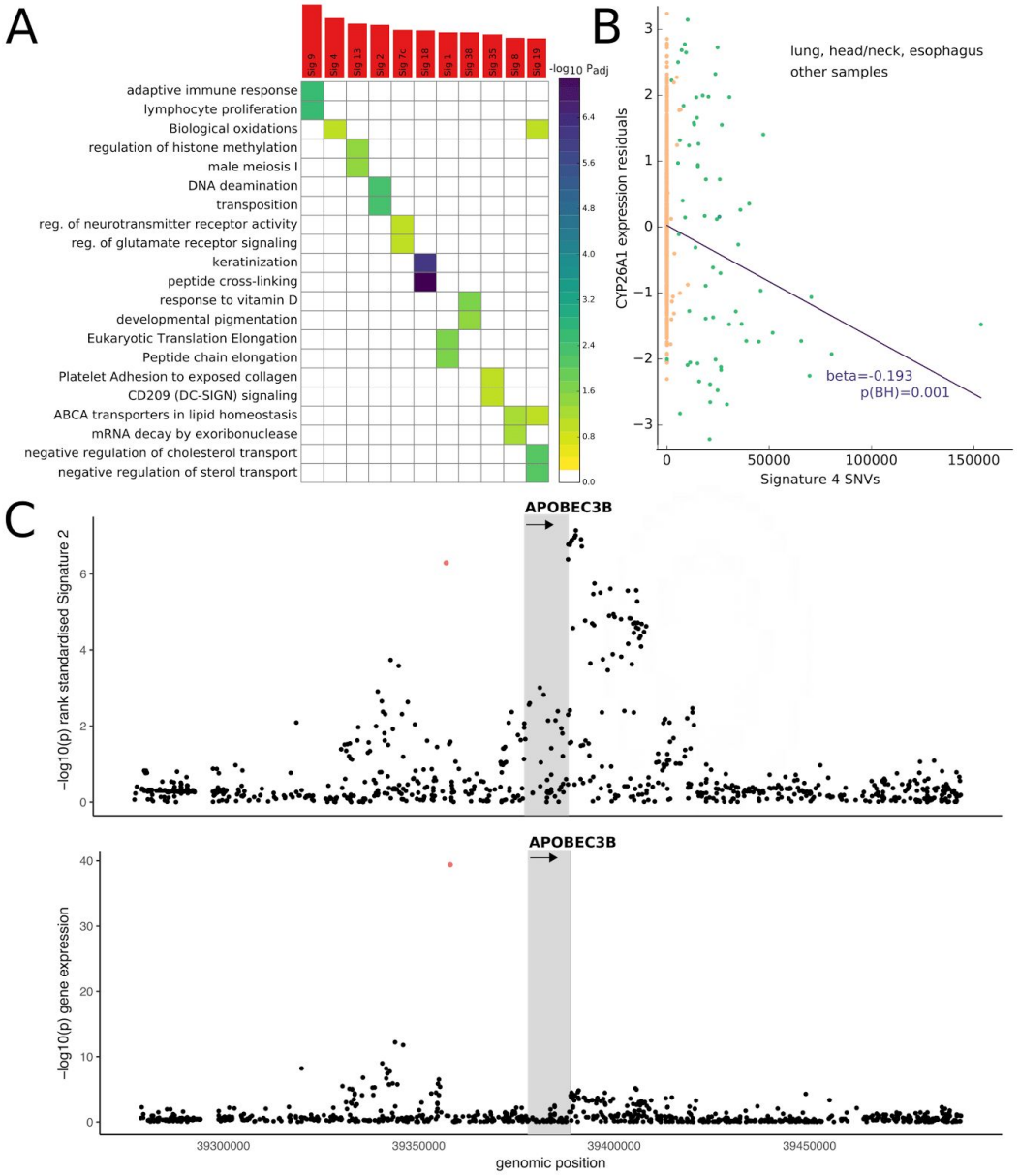

Fig. 5. Associations between mutational signatures and gene expression**. A)** Summary of significant associations. Top panel: Total number of associated genes per signature (FDR ≤ 10%). Bottom panel: Enriched GO categories or Reactome pathways for genes associated with each signature (FDR ≤ 10%, significance level encoded in color, $-\log_{10} P_{adj}$). **B)** Representative signature-gene association, depicting a negative association between *CYP26A1* expression and Signature 4. **C)** Manhattan plots of associations between *cis* germline variants proximal to *APOBEC3B* (plus or minus 100kb from the gene boundaries) and Signature 2 (top panel) or *APOBEC3B* gene expression level (bottom panel). The gray region denotes the gene body, the orange variant the lead eQTL variant rs12628403.

References

Abbas, Tarek, and Anindya Dutta. 2009. “p21 in Cancer: Intricate Networks and Multiple Activities.” *Nature Reviews. Cancer* 9 (6):400–414.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. “Signatures of Mutational Processes in Human Cancer.” *Nature* 500 (7463):415–21.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1):25–29.

Bajrami, I., J. R. Frankum, A. Konde, R. E. Miller, F. L. Rehman, R. Brough, J. Campbell, et al. 2013. “Genome-Wide Profiling of Genetic Synthetic Lethality Identifies CDK12 as a Novel Determinant of PARP1/2 Inhibitor Sensitivity.” *Cancer Research* 74 (1):287–97.

Baron, R. M., and D. A. Kenny. 1986. “The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.” *Journal of Personality and Social Psychology* 51 (6):1173–82.

Bernhart, Stephan H., Helene Kretzmer, Lesca M. Holdt, Frank Jühling, Ole Ammerpohl, Anke K. Bergmann, Bernd H. Northoff, et al. 2016. “Changes of Bivalent Chromatin Coincide with Increased Expression of Developmental Genes in Cancer.” *Scientific Reports* 6 (November):37393.

Bhalla, Sherry, Kumardeep Chaudhary, Ritesh Kumar, Manika Sehgal, Harpreet Kaur, Suresh Sharma, and Gajendra P. S. Raghava. 2017. “Gene Expression-Based Biomarkers for Discriminating Early and Late Stage of Clear Cell Renal Cancer.” *Scientific Reports* 7 (March):44997.

Bryois, Julien, Alfonso Buil, David M. Evans, John P. Kemp, Stephen B. Montgomery, Donald F. Conrad, Karen M. Ho, et al. 2014. “Cis and Trans Effects of Human Genomic Variants on Gene Expression.” *PLoS Genetics* 10 (7):e1004461.

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

Cancer Genome Atlas Research Network, Cyriac Kandoth, Nikolaus Schultz, Andrew D. Cherniack, Rehan Akbani, Yuexin Liu, Hui Shen, et al. 2013. “Integrated Genomic Characterization of Endometrial Carcinoma.” *Nature* 497 (7447):67–73.

Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. “The Cancer Genome Atlas Pan-Cancer Analysis Project.” *Nature Genetics* 45 (10):1113–20.

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

Denat, Laurence, Ana L. Kadekaro, Laurent Marrot, Sancy A. Leachman, and Zalfa A. Abdel-Malek. 2014. “Melanocytes as Instigators and Victims of Oxidative Stress.” *The Journal of Investigative Dermatology* 134 (6):1512–18.

Ding, Jiarui, Melissa K. McConechy, Hugo M. Horlings, Gavin Ha, Fong Chun Chan, Tyler Funnell, Sarah C. Mullaly, et al. 2015. “Systematic Analysis of Somatic Mutations Impacting Gene Expression in 12 Tumour Types.” *Nature Communications* 6 (October):8554.

Ekumi, Kingsley M., Hana Paculova, Tina Lenasi, Vendula Pospichalova, Christian A. Bösken, Jana Rybarikova, Vitezslav Bryja, Matthias Geyer, Dalibor Blazek, and Matjaz Barboric. 2015. “Ovarian Carcinoma CDK12 Mutations Misregulate Expression of DNA

Repair Genes via Deficient Formation and Function of the Cdk12/CycK Complex." *Nucleic Acids Research* 43 (5):2575–89.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414):57–74.

Fabregat, Antonio, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, et al. 2016. "The Reactome Pathway Knowledgebase." *Nucleic Acids Research* 44 (D1):D481–87.

Fonseca, Nuno A., Andre Kahles, Kjong-Van Lehmann, Claudia Calabrese, Aurelien Chateigner, Natalie R. Davidson, Deniz Demircioğlu, et al. 2017. "Pan-Cancer Study of Heterogeneous RNA Aberrations." *bioRxiv*. https://doi.org/10.1101/183889.

Forbes, Simon A., David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, et al. 2017. "COSMIC: Somatic Cancer Genetics at High-Resolution." *Nucleic Acids Research* 45 (D1):D777–83.

Fredriksson, Nils J., Lars Ny, Jonas A. Nilsson, and Erik Larsson. 2014. "Systematic Analysis of Noncoding Somatic Mutations and Gene Expression Alterations across 14 Tumor Types." *Nature Genetics* 46 (12):1258–63.

Garraway, Levi A., and Eric S. Lander. 2013. "Lessons from the Cancer Genome." *Cell* 153 (1):17–37.

Gene Ontology Consortium. 2015. "Gene Ontology Consortium: Going Forward." *Nucleic Acids Research* 43 (Database issue):D1049–56.

Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Thomas J. Mitchell, Yulia Rubanova, et al. 2017. "The Evolutionary History of 2,658 Cancers." *bioRxiv*. https://doi.org/10.1101/161562.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. "Genetic Effects on Gene Expression across Human Tissues." *Nature* 550 (7675):204–13.

Gu, Z., G. Thomas, J. Yamashiro, I. P. Shintaku, F. Dorey, A. Raitano, O. N. Witte, J. W. Said, M. Loda, and R. E. Reiter. 2000. "Prostate Stem Cell Antigen (PSCA) Expression Increases with High Gleason Score, Advanced Stage and Bone Metastasis in Prostate Cancer." *Oncogene* 19 (10):1288–96.

Ha, Gavin, Andrew Roth, Daniel Lai, Ali Bashashati, Jiarui Ding, Rodrigo Goya, Ryan Giuliany, et al. 2012. "Integrative Analysis of Genome-Wide Loss of Heterozygosity and Monoallelic Expression at Nucleotide Resolution Reveals Disrupted Pathways in Triple-Negative Breast Cancer." *Genome Research* 22 (10). Cold Spring Harbor Laboratory Press:1995–2007.

Hanafusa, Tadashi, Ali Eldib Ali Mohamed, Shohei Domae, Eiichi Nakayama, and Toshiro Ono. 2012. "Serological Identification of Tektin5 as a Cancer/testis Antigen and Its Immunogenicity." *BMC Cancer* 12 (November):520.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project." *Genome Research* 22 (9):1760–74.

Hartmann, S., B. Schuhmacher, T. Rausch, L. Fuller, C. Döring, M. Weniger, A. Lollies, et al. 2016. "Highly Recurrent Mutations of SGK1, DUSP2 and JUNB in Nodular Lymphocyte Predominant Hodgkin Lymphoma." *Leukemia* 30 (4):844–53.

Havelange, Violaine, Yuri Pekarsky, Tatsuya Nakamura, Alexey Palamarchuk, Hansjuerg Alder, Laura Rassenti, Thomas Kipps, and Carlo M. Croce. 2011. "IRF4 Mutations in Chronic Lymphocytic Leukemia." *Blood* 118 (10):2827–29.

Ilboudo, Adeodat, Jean-Charles Nault, Hélène Dubois-Pot-Schneider, Anne Corlu, Jessica Zucman-Rossi, Michel Samson, and Jacques Le Seyec. 2014. "Overexpression of Phosphatidylinositol 4-Kinase Type IIIα Is Associated with Undifferentiated Status and

Poor Prognosis of Human Hepatocellular Carcinoma." *BMC Cancer* 14 (January):7.
Ingram, Wendy J., Lisa M. Crowther, Erica B. Little, Ruth Freeman, Ivon Harliwong, Desi Veleva, Timothy E. Hassall, Marc Remke, Michael D. Taylor, and Andrew R. Hallahan. 2013. "ABC Transporter Activity Linked to Radiation Resistance and Molecular Subtype in Pediatric Medulloblastoma." *Experimental Hematology & Oncology* 2 (1):26.
Jia, Peilin, William Pao, and Zhongming Zhao. 2014. "Patterns and Processes of Somatic Mutations in Nine Major Cancers." *BMC Medical Genomics* 7 (February):11.
Jimbow, K., H. Chen, J. S. Park, and P. D. Thomas. 2001. "Increased Sensitivity of Melanocytes to Oxidative Stress and Abnormal Expression of Tyrosinase-Related Protein in Vitiligo." *The British Journal of Dermatology* 144 (1):55–65.
Kanchi, Krishna L., Kimberly J. Johnson, Charles Lu, Michael D. McLellan, Mark D. M. Leiserson, Michael C. Wendl, Qunyuan Zhang, et al. 2014. "Integrated Analysis of Germline and Somatic Variants in Ovarian Cancer." *Nature Communications* 5:3156.
Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, et al. 2017. "Common Genetic Variation Drives Molecular Heterogeneity in Human iPSCs." *Nature* 546 (7658):370–75.
Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3):310–15.
Knudson, Alfred G. 2002. "Cancer Genetics." *American Journal of Medical Genetics* 111 (1):96–102.
Korir, Paul K., and Cathal Seoighe. 2014. "Inference of Allele-Specific Expression from RNA-Seq Data." *Methods in Molecular Biology* 1112:49–69.
Kumar, Prateek, Steven Henikoff, and Pauline C. Ng. 2009. "Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm." *Nature Protocols* 4 (7):1073–81.
Kvam, E., and R. M. Tyrrell. 1999. "The Role of Melanin in the Induction of Oxidative DNA Base Damage by Ultraviolet A Irradiation of DNA or Melanoma Cells." *The Journal of Investigative Dermatology* 113 (2):209–13.
Lasorella, Anna, Robert Benezra, and Antonio Iavarone. 2014. "The ID Proteins: Master Regulators of Cancer Stem Cells and Tumour Aggressiveness." *Nature Reviews. Cancer* 14 (2):77–91.
Lesch, Bluma J., and David C. Page. 2014. "Poised Chromatin in the Mammalian Germ Line." *Development* 141 (19):3619–26.
Lindeboom, Rik G. H., Fran Supek, and Ben Lehner. 2016. "The Rules and Impact of Nonsense-Mediated mRNA Decay in Human Cancers." *Nature Genetics* 48 (10):1112–18.
Liu, Xinran, Hao Gong, and Kun Huang. 2013. "Oncogenic Role of Kinesin Proteins and Targeting Kinesin Therapy." *Cancer Science* 104 (6):651–56.
Medves, Sandrine, Morgan Auchter, Laetitia Chambeau, Sophie Gazzo, Delphine Poncet, Blandine Grangier, Aurélie Verney, et al. 2016. "A High Rate of Telomeric Sister Chromatid Exchange Occurs in Chronic Lymphocytic Leukaemia B-Cells." *British Journal of Haematology* 174 (1):57–70.
Middlebrooks, Candace D., A. Rouf Banday, Konichi Matsuda, Krizia-Ivana Udquim, Olusegun O. Onabajo, Ashley Paquin, Jonine D. Figueroa, et al. 2016. "Association of Germline Variants in the APOBEC3 Region with Cancer Risk and Enrichment with APOBEC-Signature Mutations in Tumors." *Nature Genetics* 48 (11):1330–38.
Milacic, Marija, Robin Haw, Karen Rothfels, Guanming Wu, David Croft, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. 2012. "Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome." *Cancers* 4 (4):1180–1211.
Nik-Zainal, Serena, David C. Wedge, Ludmil B. Alexandrov, Mia Petljak, Adam P. Butler,

Niccolo Bolli, Helen R. Davies, et al. 2014. “Association of a Germline Copy Number Polymorphism of APOBEC3A and APOBEC3B with Burden of Putative APOBEC-Dependent Mutations in Breast Cancer.” *Nature Genetics* 46 (5):487–91.

Nonaka, Taichiro, Yoshinobu Toda, Hiroshi Hiai, Munehiro Uemura, Motonobu Nakamura, Norio Yamamoto, Ryo Asato, et al. 2016. “Involvement of Activation-Induced Cytidine Deaminase in Skin Cancer Development.” *The Journal of Clinical Investigation* 126 (4):1367–82.

Olivier, Magali, Monica Hollstein, and Pierre Hainaut. 2010. “TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use.” *Cold Spring Harbor Perspectives in Biology* 2 (1):a001008.

Ongen, Halit, Claus L. Andersen, Jesper B. Bramsen, Bodil Oster, Mads H. Rasmussen, Pedro G. Ferreira, Juan Sandoval, et al. 2014. “Putative Cis-Regulatory Drivers in Colorectal Cancer.” *Nature* 512 (7512):87–90.

PCAWG Consortium et al. 2017. “PCAWG Marker Paper.” *In Preparation*.

PCAWG Group 1 et al. 2017. “PCAWG-1 Marker Paper.” *In Preparation*.

PCAWG Group 3 et al. 2017. “PCAWG-3 Marker Paper.” *In Preparation*.

PCAWG Group 7 et al. 2018. “PCAWG-7 Marker Paper.” *In Preparation*.

PCAWG Group 8 et al. 2017. “PCAWG-8 Marker Paper.” *In Preparation*.

PCAWG Group 11 et al. 2018. “PCAWG 11 Marker Paper.” *In Preparation*.

Pleasance, Erin D., R. Keira Cheetham, Philip J. Stephens, David J. McBride, Sean J. Humphray, Chris D. Greenman, Ignacio Varela, et al. 2010. “A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome.” *Nature* 463 (7278):191–96.

Preacher, Kristopher J., and Andrew F. Hayes. 2004. “SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models.” *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc* 36 (4):717–31.

Premi, Sanjay, and Douglas E. Brash. 2016. “Unanticipated Role of Melanin in Causing Carcinogenic Cyclobutane Pyrimidine Dimmers.” *Molecular & Cellular Oncology* 3 (1):e1033588.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518 (7539):317–30.

Saini, Natalie, Steven A. Roberts, Leszek J. Klimczak, Kin Chan, Sara A. Grimm, Shuangshuang Dai, David C. Fargo, et al. 2016. “The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts.” *PLoS Genetics* 12 (10):e1006385.

Saito, Takamune T., Firaz Mohideen, Katherine Meyer, J. Wade Harper, and Monica P. Colaiácovo. 2012. “SLX-1 Is Required for Maintaining Genomic Integrity and Promoting Meiotic Noncrossovers in the Caenorhabditis Elegans Germline.” *PLoS Genetics* 8 (8):e1002888.

Scanlan, Matthew J., Ali O. Gure, Achim A. Jungbluth, Lloyd J. Old, and Yao-Tseng Chen. 2002. “Cancer/testis Antigens: An Expanding Family of Targets for Cancer Immunotherapy.” *Immunological Reviews* 188 (October):22–32.

Schaetzlein, Sonja, Richard Chahwan, Elena Avdievich, Sergio Roa, Kaichun Wei, Robert L. Eoff, Rani S. Sellers, et al. 2013. “Mammalian Exo1 Encodes Both Structural and Catalytic Functions That Play Distinct Roles in Essential Biological Processes.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (27):E2470–79.

Shang, Xiying, Xinjian Lin, Edwin Alvarez, Gerald Manorek, and Stephen B. Howell. 2012. “Tight Junction Proteins Claudin-3 and Claudin-4 Control Tumor Growth and Metastases.” *Neoplasia* 14 (10):974–85.

Simpson, Andrew J. G., Otavia L. Caballero, Achim Jungbluth, Yao-Tseng Chen, and Lloyd J. Old. 2005. “Cancer/testis Antigens, Gametogenesis and Cancer.” *Nature Reviews. Cancer* 5 (8):615–25.

Smith, Kyle S., Vinod K. Yadav, Brent S. Pedersen, Rita Shaknovich, Mark W. Geraci, Katherine S. Pollard, and Subhajyoti De. 2015. “Signatures of Accelerated Somatic Evolution in Gene Promoters in Multiple Cancer Types.” *Nucleic Acids Research* 43 (11):5307–17.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43):15545–50.

Tang, Zefang, Chenwei Li, Boxi Kang, Ge Gao, Cheng Li, and Zemin Zhang. 2017. “GEPIA: A Web Server for Cancer and Normal Gene Expression Profiling and Interactive Analyses.” *Nucleic Acids Research* 45 (W1). Oxford University Press:W98–102.

Wallace, Chris. 2013. “Statistical Testing of Shared Genetic Control for Potentially Related Traits.” *Genetic Epidemiology* 37 (8):802–13.

Wang, Rong-Fu, and Helen Y. Wang. 2017. “Immune Targets and Neoantigens for Cancer Immunotherapy and Precision Medicine.” *Cell Research* 27 (1):11–37.

Wang, Wei-Yu, Chung-Chi Hsu, Ting-Yun Wang, Chi-Rong Li, Ya-Chin Hou, Jui-Mei Chu, Chung-Ta Lee, et al. 2013. “A Gene Expression Signature of Epithelial Tubulogenesis and a Role for ASPM in Pancreatic Tumor Progression.” *Gastroenterology* 145 (5):1110–20.

Wang, Zhiwei, Pengda Liu, Hiroyuki Inuzuka, and Wenyi Wei. 2014. “Roles of F-Box Proteins in Cancer.” *Nature Reviews. Cancer* 14 (4):233–47.

Waszak, Sebastian M., Grace Tiao, Bin Zhu, Tobias Rausch, Francesc Muyas, Bernardo Rodriguez-Martin, Raquel Rabionet, et al. 2017. “Germline Determinants of the Somatic Mutation Landscape in 2,642 Cancer Genomes.” *bioRxiv*. https://doi.org/10.1101/208330.

Weinhold, Nils, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. 2014. “Genome-Wide Analysis of Noncoding Regulatory Mutations in Cancer.” *Nature Genetics* 46 (11):1160–65.

Weir, Barbara, Xiaojun Zhao, and Matthew Meyerson. 2004. “Somatic Alterations in the Human Cancer Genome.” *Cancer Cell* 6 (5):433–38.

Whalley, Justin P., Ivo Buchhalter, Esther Rheinbay, Keiran M. Raine, Kortine Kleinheinz, Miranda D. Stobbe, Johannes Werner, et al. 2017. “Framework For Quality Assessment Of Whole Genome, Cancer Sequences.” https://doi.org/10.1101/140921.

Woenckhaus, M., L. Klein-Hitpass, U. Grepmeier, J. Merk, M. Pfeifer, Pj Wild, M. Bettstetter, et al. 2006. “Smoking and Cancer-Related Gene Expression in Bronchial Epithelium and Non-Small-Cell Lung Cancers.” *The Journal of Pathology* 210 (2):192–204.

Yung, Christina K., Brian D. O’Connor, Sergei Yakneen, Junjun Zhang, Kyle Ellrott, Kortine Kleinheinz, Naoki Miyoshi, et al. 2017. “Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments.” https://doi.org/10.1101/161638.

Zhang, Junjun, Joachim Baran, A. Cros, Jonathan M. Guberman, Syed Haider, Jack Hsu, Yong Liang, et al. 2011. “International Cancer Genome Consortium Data Portal--a One-Stop Shop for Cancer Genomics Data.” *Database: The Journal of Biological Databases and Curation* 2011 (September):bar026.

Zhang, Xin A., William S. Lane, Stephanie Charrin, Eric Rubinstein, and Lei Liu. 2003. “EWI2/PGRL Associates with the Metastasis Suppressor KAI1/CD82 and Inhibits the Migration of Prostate Cancer Cells.” *Cancer Research* 63 (10). American Association for

Cancer Research:2665–74.

DFG (Deutsche Forschungsgemeinschaft) (2010) The MAK-Collection Part IV: BAT Value Documentations, Vol. 5. *WILEY-VCH Verlag GmbH & Co*. KGaA, Weinheim. ISBN: 978-3-527-32614-3.