Mosaic Mutations in Blood DNA Sequence Are Associated with Solid Tumor Cancers

Mykyta Artomov [1,2,3], Manuel A. Rivas[1,2], Giulio Genovese [1], Mark J. Daly[1,2]

[1] Broad Institute, Cambridge, MA, USA, 02139

[2] Analytic and Translational Genetics Unit, MGH, Boston, MA, USA, 02114

[3] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA, 02138

Please send correspondences to MJD (mjdaly@atgu.mgh.harvard.edu)

The authors disclose no potential conflict of interest

Abstract

Recent findings in understanding of the causal role of blood mosaic protein-truncating DNA variants in leukemia rose a question about generalizability of such observation for other cancer types. We used exome sequencing to compare 22 different cancer phenotypes from TCGA data (~8,000 samples) with more than 6,000 controls using a case-control study design and demonstrate that mosaic protein truncating variants in these genes are also associated with solid-tumor cancers. We analyzed tumor DNA samples from TCGA and observed that mosaic variants driving the association with cancer are absent from the tumors.

By analysis of the different cancer phenotypes we observe gene-specificity for mosaic mutations. PPM1D in previous reports has been linked to breast and ovarian cancer, which our analysis confirms as a specific gene for ovarian cancer. Also glioblastoma, melanoma and lung cancers show gene specific burden of the mosaic protein truncating mutations.

Taken together, these results represent an important observation of solid-tumor cancers being linked to the somatic blood DNA changes.

Introduction

Several recent studies[1,2,3] have reported associations of mosaic protein truncating variants (PTV) in *PPM1D, TET2, ASXL1* and *DNMT3A* with blood cancers. Intriguingly, such mosaic mutations in *PPM1D* have also been convincingly associated with breast and ovarian cancer – however, since these mutations are somatic, rather than germline, a role in causation has not been clear. We sought to more fully explore the relationship of these somatic mutations, clearly causally linked to blood cancers, in solid tumor cancer using a large assembly of germline and somatic exome DNA sequences of 7,979 cancer cases from TCGA[4] and performed a large-scale case-control study with 6,177 population controls with no cancer phenotype reported.

Results

Using data available from dbGAP, we performed a large-scale joint variant calling of germline DNA samples from the blood of cancer cases and controls – primarily from an assembly of TCGA samples (cases) compared with unselected population controls (with no known cancer status) from several studies (NHLBI-ESP, 1000 Genomes, ATVB, T2D, Ottawa Heart) appropriately consented for broad use as controls. Importantly, all cases and controls in this analysis have age at DNA sampling available (Supp. Table 1).

Observations of the mosaic mutations could be affected by several parameters – age, depth of coverage, variant calling probability. To make the case-control comparison robust we first identified what adjustments to the model of association are needed. We observed 348 PTVs (stop gain, essential splice site, frameshift mutations) in the four established somatic leukemia genes. Detection of somatic mutations with low non-reference allele balance depends importantly on depth. In order to insure no different sensitivity in our cases and controls we first compared depth of coverage in these genes in our cancer germline (average 33X coverage) and control (average 29X coverage) data. We further looked specifically at cases and controls with called PTVs. For germline heterozygous sites, the expected allele balance is 0.5 so we applied a binomial test to determine low allele balance genotypes based on the depth of coverage and number of alternative reads. Those with p<0.001 (i.e., heterozygotes with significantly less than 50% non-reference allele) and more than 20x coverage were determined to be mosaic and kept for further analysis (Supp. Fig 1,2). To further investigate any statistical bias due to a coverage of cases and controls - we tested whether there is a statistical difference in coverage and ref/alt reads counts between cancer cases and controls that carry at least one PTV in the 4 candidate genes with generalized linear model testing. The cancer status of the sample appears to be a non-significant (p=0.279, p=0.898 if adjusted for age) parameter, confirming that called PTVs are adequately covered in both cases and controls and protein-truncating mosaic events have equal chances to be detected in both cohorts. We finally evaluated the probability of calling a protein truncating DNA variant in cases and controls with respect to coverage (Sup. Fig. 3), since there is slightly higher sensitivity for the detection of DNA variants in cases we adjusted further analysis for the coverage differences. From these analyses, we conclude that all minor technical differences in sensitivity to find mosaic variants in cases and controls were accounted for – a pre-requisite for subsequent analyses.

We then assessed association between mosaic PTV and cancer status by generating a data set consisting of 7,979 cancer cases and 6,177 controls (See Methods). We applied a binomial generalized linear model considering age, coverage depth and mosaic PTV carrier status and found significant evidence of association with cancer status (P=0.00108, OR=1.26; OR CI=1.1-1.47). Since it was previously shown that *PPM1D* PTVs are associated with breast and ovarian cancers, we removed breast and ovarian cancer samples and repeated analysis, which confirmed the association to cancer status ($P=5.67x10^{-4}$, OR=1.3; OR CI=1.12-1.52) – suggesting the reported observations regarding *PPM1D* and breast and ovarian cancers are more general. As our set of controls was on average roughly 10 years younger than our cancer cohort and age has been shown to be a strong predictor of the existence of these somatic mosaic events, the inclusion of age in the above model is critical. We further evaluated the effect of the age on the probability of finding a mosaic event (Sup. Fig. 4)[5] and the generalized linear modeling above was adjusted for age differences between cases and controls. We also adjusted our model for minor coverage differences between cases and controls.

It is known that specifically PTVs in the last exon of *PPM1D* are enriched in cases of breast and ovarian cancer[1]. We observed the same enrichment in our dataset - 18 mosaic PTVs in *PPM1D*, 17 of which appeared in the last exon of the gene. Thus we tested other candidate genes for distribution of the PTVs. Variants in *TET2* also show strong exon specificity – 44 out of 50 total PTVs are found in the 3rd exon of the gene. Out of 40 ASXL1 PTVs 35 appear in the last exon and *DNMT3A* PTVs do not show any exon specificity (Supp Fig. 5). Previous reports of leukemia studies observe accumulation of the mosaic missense mutations in the last exons of *DNMT3A*. We found the same to be true in blood for solid tumor cancer cases as well ($p=4.78x10^{-4}$, Supp Fig. 6).

As it was previously demonstrated, mosaic PTVs in the list of candidate genes have a high association with the development of leukemia. Since mutations in these genes were demonstrated to precede and predict the development of leukemia, and because these genes were previously established to be leukemia genes, – strong evidence for a causative role in leukemias has been established[2.3]~~.~~ We next sought a key piece of evidence for evaluating the role of these mutations in solid tumors. Specifically we evaluated the quantity of these mosaic PTVs between tumor and germline DNA across all cancer samples included in this study with a detectable event in blood DNA and observed that mosaic PTVs in the candidate genes present in blood DNA were largely absent in the tumor DNA samples from the same individual (Fig. 1). This strongly indicates that these events in the blood did not represent residual evidence from driver mutations involved in tumor development (in which case we would have expected higher, or perhaps 100% of the mutated allele to be found). As before, we compared coverage in tumor and germline DNA samples, which shows, consistent with the design of TCGA, that tumors have similar or better coverage indicating that the deficit of these mosaic events in tumors is not sensitivity based (Supp. Fig 7). This observation is consistent with the findings of mosaic *PPM1D* variants in breast/ovarian cancers[1].

We considered whether presence of mosaic PTVs showed any evidence of cancer specificity. Under the null model, we would expect mosaic events to be found in all candidate genes at the same rate in each of the 20 cancer phenotypes. We tested for deviation from this null model and applied a multiple hypothesis testing correction procedure (p=.05/20) to reject the null that the rate is the same across all the phenotypes in all the genes. We first tested if any of the cancer phenotypes shows unusual burden of the mosaic PTVs by randomly drawing sample set (controlling for similarity of age distributions between the random and target sets) from the total cancer samples cohort and estimating how likely is observed amount of mosaic PTVs in each cancer phenotype. (Fig. 2a). Glioblastoma, melanoma and especially lung cancers show association by carrying increased burden of the mosaic PTVs compared to other cancers. We then looked on the distribution of mosaic PTVs across the candidate genes in each cancer phenotype (Fig. 2b, Sup. Table 2). Performing similar random permutations analysis. It appears that several cancer types show a trend for accumulation of the mosaic mutations in specific genes. Intriguingly, ovarian cancer is specifically associated with *PPM1D* mutations, which is supported by previous report[1]. We also observe associations of head and neck squamous cell carcinoma with *PPM1D*, colorectal adenocarcinoma and glioblastoma with *TET2*. Interestingly, cutaneous melanoma is associated with *ASXL1* mosaic mutations as *ASXL1* has protein-interaction with *BAP1*, a well-established risk factor for melanoma[6]. Lung cancer shows burden of the mosaic mutations that is distributed across several genes, suggesting no specificity in accumulation of the mosaic mutations. This perhaps could be related to the observation that smokers have higher rate of mosaic PTVs.

We used the previously reported set of samples from Swedish national patient registers[2] to estimate the frequency of mosaic PTVs and associated solid-tumor cancer development in a population unselected for cancer.

We removed from analysis all samples that had an evidence of leukemia or lymphoma developed before or after the DNA collection as well as those samples that have mosaic missense mutations in *DNMT3A* to estimate the contribution of the PTVs only. The final dataset for this analysis consisted of 10,966 (92 mosaic PTV carriers and 10,870 non-carriers) samples. There were 15 individuals with pre-DNA collection record of the solid-tumor cancer in the cohort of mosaic PTV carriers and 1,104 samples with record of solid-tumor cancer among non-carriers. We tried using different thresholds for age of the samples to estimate significance of enrichment. However, due to a small incidence of the mosaic mutations in the population unselected for cancer, this test was inconclusive (Supp. Table 3a).

We added mosaic missense *DNMT3A* mutations carriers to the mosaic samples cohort and repeated population analysis (Supp. Table 3b). This resulted in a total of 179 mosaic samples. There were 31 individuals (~17%) with pre-DNA collection record of the solid-tumor cancer in the cohort of mosaic PTV carriers (1,104 (~10%)cancer records in 10,870 non-mosaic samples). Once corrected for age this enrichment appears to be insignificant, thus for samples unselected for cancer a much larger cohort is needed to reach a significant conclusion.

Smoking status within the cohort unselected for cancer was tested for association with the burden of mosaic mutations. We observe slightly higher chances to carry a mosaic PTV among smokers and former smokers (p=0.615; OR=1.15; OR CI=0.676-1.91). Among the smokers or former smokers samples both mosaic and non-mosaic carriers have similar chances to have a history of solid-tumor cancer (p=0.2526; OR=0.53; OR CI=0.285-1.26) (Supp. Table 4).

Alternative possible driver of previously reported leukemia association with mosaic PTVs is a clinical intervention, specifically – radiation treatment, well known to be leading to increased risk of leukemia. Testing of the model for association of available clinical data, such as neoadjuvant treatment history (p=0.116), radiation therapy (p=0.348), pathologic tumor stage (p=0.354) and primary therapy outcome, showing whether patient had remission/response, progressive or stable disease as a result of clinical interventions (p=0.213) adjusted for age and cancer phenotype of the patients shows no association with mosaic PTV carrier status (Supp. Tables 5,6,7). However, this analysis appears somewhat limited by the clinical records available from TCGA portal.

Discussion

Our study investigates association of the mosaic protein-truncating variants in 4 previously associated with blood cancer risk genes on the solid-tumor cancer phenotypes.

Previously observed strong association of the mosaic PTVs with increased risk of leukemia in our observations is extended to the solid-tumor cancers. There are several possible reasons for observation of mosaic events. First, immune system changes in response to early pre-clinical stage of cancer. Our additional screening of early onset cancer cases (breast and ovarian cohort with cancer onset before 35, N=374) shows no enrichment in mosaic PTVs suggesting that this hypothesis is likely irrelevant and age of the samples plays important role (or serving as a trigger) for emergence of clonal expansion. Second, is the causal relationship. However, this is contradictory with absence of PTVs in tumors. At the same time observed differences in gene specificity to certain cancer phenotypes and increased burden of PTVs in cases suggests that there could be fundamental biological reasons for association signal. Third, is the result of clinical intervention. Previously reported association of mosaic PTVs in the last exon of *PPM1D* to breast and ovarian cancers suggests that such mutations result in enhanced p53 suppression in response to ionizing radiation.

Though finding a fundamental association of blood mosaic PTVs with solid-tumor cancers, analyses of different covariates in population cohorts do not explain wait is the underlying nature of such association due to a limited statistical power to detect deviation from null. Analysis of Swedish population unselected for cancer gives an estimate of how often such mosaic mutations could be seen in general population. Given a frequency of 0.82% to carry a PTV in one of candidate genes, a large-scale population study with a long-term post-DNA collection follow up would be needed to confidently answer the question whether blood mosaic PTVs are precursors of solid-tumor cancers.

References

1. Ruark E, Snape K, Humburg P, et al. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. Nature. 2013;493(7432):406-410. doi:10.1038/nature11725
2. Genovese G, Kähler AK, Handsaker RE, et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. The New England journal of medicine. 2014;371(26):2477-2487. doi:10.1056/NEJMoa1409405.
3. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med. 2014;371(26):2488-2498
4. The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/
5. Xie M, Lu C, Wang J, et al. Age-related cancer mutations associated with clonal hematopoietic expansion. Nature medicine. 2014;20(12):1472-1478. doi:10.1038/nm.3733.
6. Michele Carbone, Haining Yang, Harvey I. Pass, Thomas Krausz, Joseph R. Testa and Giovanni Gaudino. NATURE REVIEWS, VOLUME 13; 2013; 153-159.
7. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010 *GENOME RESEARCH 20:1297-303*
8. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M, 2011 *NATURE GENETICS 43:491-498*
9. Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M, 2013 *CURRENT PROTOCOLS IN BIOINFORMATICS 43:11.10.1-11.10.33*
10. Paul Flicek, M. Ridwan Amode, *et al.* Nucleic Acids Research 2014 42 Database issue:D749-D755 doi: 10.1093/nar/gkt1196

11. Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190
12. Alkes L Price[1, 2], Nick J Patterson[2], Robert M Plenge[2, 3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich, *Nature Genetics* 38, 904 - 909 (2006)
13. https://atgu.mgh.harvard.edu/plinkseq/
14. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Methods

Dataset

Genotypes dataset was created by joint variant calling of cancer cases and non-cancer controls using HaplotypeCaller (GATK-3.0)[7,8,9] with Broad Institute calling pipeline. For functional annotation of variants we used Variant Effect Predictor by Ensembl[10].

PCA was performed to keep for analysis only samples of European ancestry to eliminate possible population effects. PCA was performed with EIGENSTRAT[11,12].

Resulting genotype file was used to create a PLINK/SEQ[13] project for further manipulations.

Clinical data

For testing relevance of the mosaic PTVs to medical treatment/outcome clinical data was downloaded from TCGA web-site https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm.

Generalized linear model and statistical tests

For further statistical tests we used R-3.0[14].

Figure captions

Fig. 1. Blood vs Tumor allele balance for each sample with mosaic PTV in (A) PPM1D, (B) TET2, (C) DNMT3A, (D) ASXL1. Observed in blood mosaic mutations are strongly depleted from the tumor somatic DNA (Wilcoxon test $P<10^{-16}$).

Fig. 2. Solid-tumor cancer phenotypes show gene specificity with respect to mosaic PTVs. (A) Empirical enrichment of the different cancer phenotypes with mosaic PTVs (B) Per gene significance of mosaic PTV burden in each cancer phenotype. Experiment-wise significance level is set with Bonferroni correction for multiple phenotypes tested. Ovarian cancer shows previously reported specific association to PPM1D mosaic PTVs.

Supplementary Fig. 1. Allele balance for all PTVs with >20X coverage in 4 candidate genes in TCGA cancer samples.

Supplementary Fig. 2. Allele balance for all PTVs with >20X coverage in 4 candidate genes in control samples.

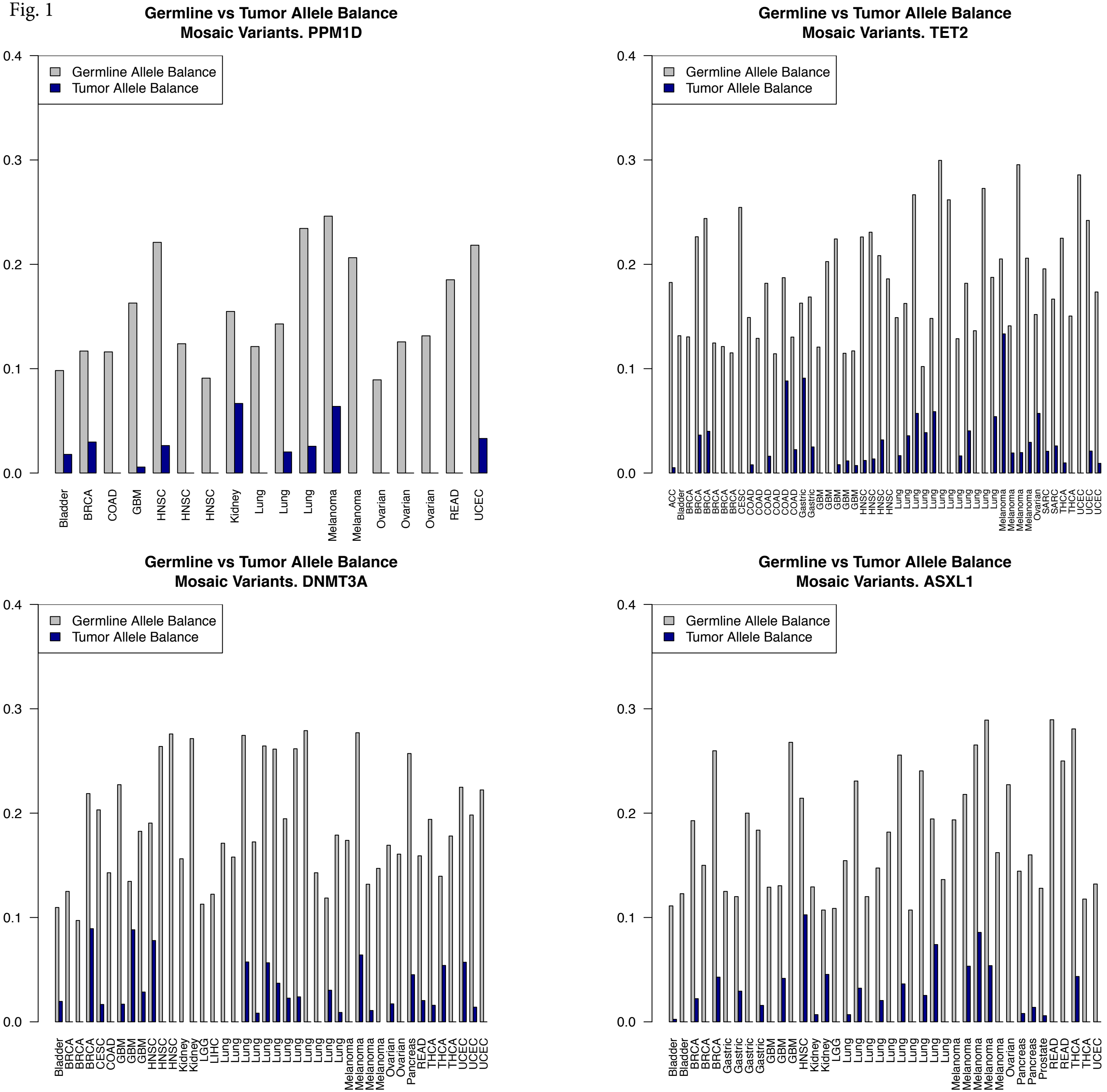
Supplementary Fig. 3. Probability of observation a protein-truncating variant in cancer and control samples with respect to the coverage. Controls show better or equal chances of PTV detection.

Supplementary Fig. 4. Mosaic PTV emergence in blood is strongly correlated with age, however probability of finding such mutations in cancer cases is much greater than in samples with no known cancer history.

Supplementary Fig. 5. Similarly to Genovese et. al. we observe enrichment of the last exons of DNMT3A with mosaic missense variants in the blood of both cases and controls.

Supplementary Fig. 6. Comparison of coverage between the TCGA blood and TCGA somatic DNA samples. On average tumor DNA has equal or better coverage, than blood DNA.

Fig. 1



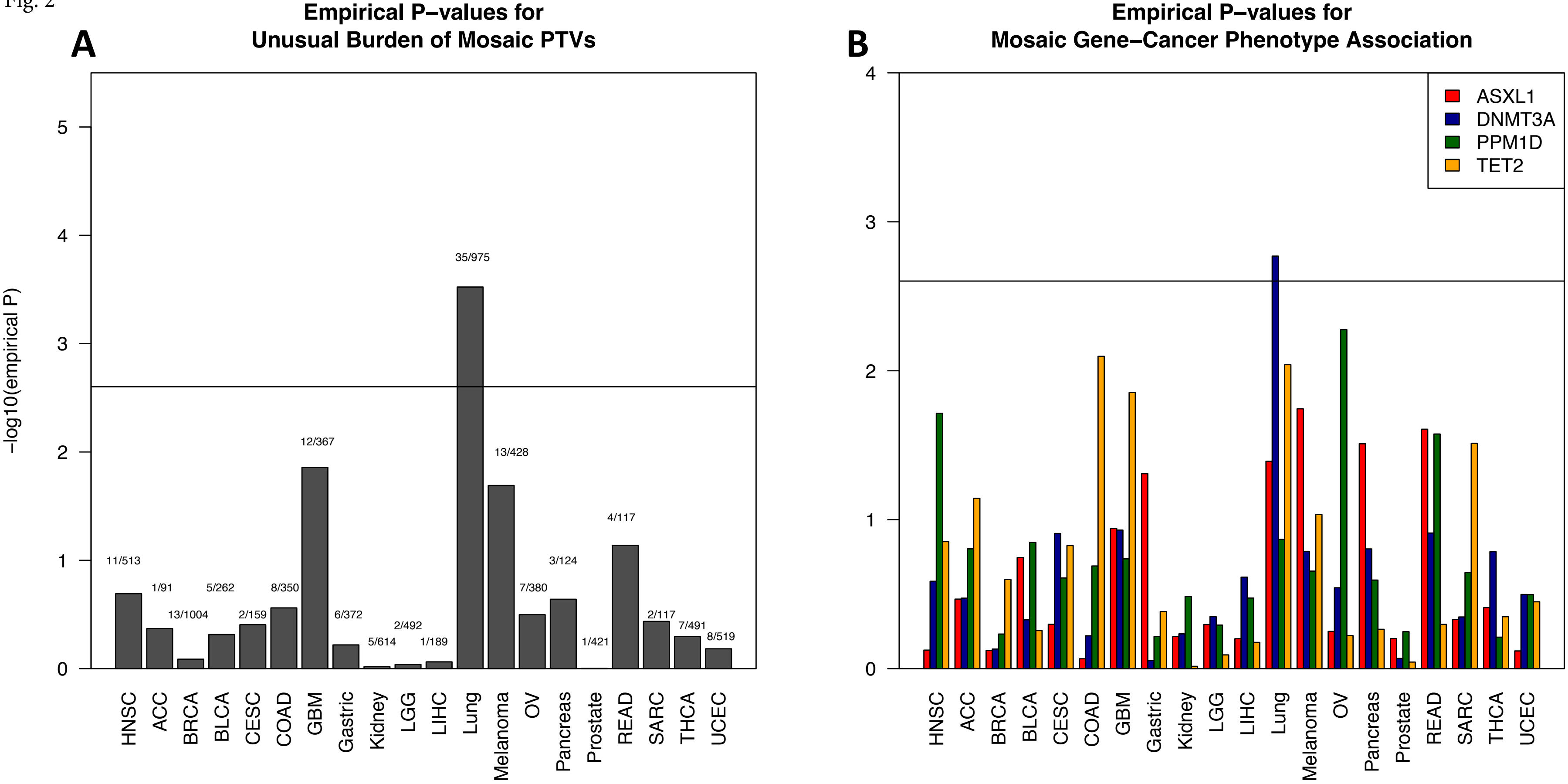
Germline vs Tumor Allele Balance
Mosaic Variants. PPM1D

Germline vs Tumor Allele Balance
Mosaic Variants. TET2

Germline vs Tumor Allele Balance
Mosaic Variants. DNMT3A

Germline vs Tumor Allele Balance
Mosaic Variants. ASXL1

Fig. 2


A
Empirical P−values for
Unusual Burden of Mosaic PTVs
−log10(empirical P)
0
1
2
3
4
5
11/513
1/91
13/1004
5/262
2/159
8/350
12/367
6/372
5/614
2/492
1/189
35/975
13/428
7/380
3/124
1/421
4/117
2/117
7/491
8/519
HNSC
ACC
BRCA
BLCA
CESC
COAD
GBM
Gastric
Kidney
LGG
LIHC
Lung
Melanoma
OV
Pancreas
Prostate
READ
SARC
THCA
UCEC
B
Empirical P−values for
Mosaic Gene−Cancer Phenotype Association
0
1
2
3
4
🟥 ASXL1
🟦 DNMT3A
🟩 PPM1D
🟧 TET2
HNSC
ACC
BRCA
BLCA
CESC
COAD
GBM
Gastric
Kidney
LGG
LIHC
Lung
Melanoma
OV
Pancreas
Prostate
READ
SARC
THCA
UCEC