A Strategy for Large-Scale Systematic Pan-Cancer Germline Rare Variation Analysis

Mykyta Artomov*[1,2,3], Joseph Vijai*[4], Grace Tiao[2], Tinu Thomas[4], Kasmintan Schrader[4], Robert Klein[4], Adam Kiezun[2], Namrata Gupta[2], Lauren Margolin[2], Alexander J. Stratigos[5], Ivana Kim[6], Kristen Shannon[7], Leif W. Ellisen[7,8], Daniel Haber[7,8,9], Gad Getz[2], Hensin Tsao[10,11], Steven Lipkin[12], David Altshuler[2], Kenneth Offit[4,13]** and Mark J. Daly[1,2]**

[1]Analytic and Translational Genetics Unit, MGH, Boston, MA, USA
[2]Broad Institute, Cambridge, MA, USA
[3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA
[4] Clinical Genetics Research Laboratory, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[5]1st Department of Dermatology-Venereology, National and Kapodistrian University of Athens School of Medicine, Andreas Sygros Hospital, Athens, Greece
[6]Retina Service, Massachusetts Eye and Ear Infirmary, Boston, MA, USA
[7] Massachusetts General Hospital Cancer Center, Boston, MA, USA
[8] Harvard Medical School, Boston, MA, USA
[9] Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA
[10] Department of Dermatology, Wellman Center for Photomedicine, MGH, Boston, MA, USA
[11] Melanoma Genetics Program, MGH Cancer Center, MGH, Boston, MA, USA
[12] Department of Medicine, Weill-Cornell Medicine, New York, NY, USA
[13] Cancer Biology and Genetics Program, Sloan Kettering Institute, New York, NY, USA

* - M.A. and V.J. contributed equally

** - K.O. and M.J.D. co-corresponding authors

ABSTRACT

Traditionally, genetic studies in cancer are focused on somatic mutations found in tumors and absent from the normal tissue. However, this approach omits inherited component of the cancer risk. We assembled exome sequences from about 2,000 patients with different types of cancers: breast cancer, colon cancer and cutaneous and ocular melanomas matched to more than 7,000 non-cancer controls. Using this dataset, we described germline variation in the known cancer genes grouped by inheritance mode or inclusion in a known cancer pathway. According to our observations, protein-truncating singleton variants in loss-of-function tolerant genes following autosomal dominant inheritance mode are driving the association signal in both genetically enriched and unselected cancer cases. We also performed separate gene-based association analysis for individual phenotypes and proposed a list of new cancer risk gene candidates. Taken together, these results extend existing knowledge of germline variation contribution to cancer onset and provide a strategy for novel gene discovery.

INTRODUCTION

Analysis of inherited predisposition to cancer usually involves cohorts of early disease onset patients or large kindreds. Here we analyzed a large cohort of genetically enriched (early onset and/or familial) and unselected cases of breast cancer, colon cancer and cutaneous and ocular melanomas (in total about 2,000 cases matched to more than 7,000 non-cancer controls) to develop a search strategy for novel germline cancer risk genes. By first analyzing known cancer

predisposition genes, we demonstrate that protein truncating, rather than missense, mutations are the main driver of inherited solid tumor cancer predisposition and generally these occur in genes tolerant of loss-of-function mutations – distinct from the highly-constrained genes more often somatically mutated and found to be drivers in tumors. Interestingly, we find that unselected cancer cases have a significant burden of protein-truncating variants in known cancer risk genes, similar to that observed in genetically enriched (familial and early-onset, herein referred to as 'selected') patients. Using these observations to design our search for new cancer genes, we analyzed individual cancer cohorts with matched controls and constructed a ranked list of new potential candidate risk genes.

RESULTS

Cohort and overview

For this study, germline DNA from selected "genetically-enriched" cases (individuals with familial cancer and/or onset of the disorder at age of 35 or earlier) of breast cancer, colon cancer, cutaneous and ocular melanomas, and Li-Fraumeni syndrome (with primary breast cancer) was collected (Inclusion criteria is included in Supplementary Methods). We also included anonymous germline DNA sequences from specific cancer types from TCGA that were used as "unselected" cancer cases (not controlling for family history or age of onset). In total 845 "genetically-enriched" cases, 1496 "unselected" cases and 7924

controls passed quality-check and were included in subsequent analysis (Supplementary Table 1).

To ensure close ancestral matching, we performed principal component analysis (PCA; Sup. Fig. 1A) of the case and control cohorts. To reduce heterogeneity due to diverse population admixture, only the single largest cluster representing predominantly European ancestry was further analyzed. Within European-ancestry samples we performed relatedness analysis and removed all duplicates and first-degree relatives (PI_HAT > 0.2). Examination of common synonymous variants (MAF>5%) revealed a null-distribution of the Fisher's exact test statistic between cases and controls (details in Methods Section) (Sup. Fig. 1B).

Search strategy for new cancer risk genes

Discovery of over 100 germline predisposition genes in cancer have not only revolutionized identification of individuals and families at higher risk, but also provided novel mechanistic insights into the role of pathways in cancer development[1] and helped in mitigating the risk using appropriate clinical management. This is true, not only in adult cancers, but also in diverse pediatric cancers. To define an exome-wide strategy to search for new cancer predisposition genes, we began by analyzing rare genetic variation in known risk genes. Specifically, we examined the abundance of risk alleles in known genes grouped by reported mendelian models of inheritance and known tumor

suppressive activity or involvement in DNA repair pathways. We identified features common to genes in each group and compared genetic association observed in selected and unselected cancer cases (Sup. Table 2)[2]. Out of four lists that we tested – autosomal dominant, autosomal recessive, tumor suppressor genes and DNA repair pathway, only the autosomal dominant model genes shows true enrichment in cancer cases (Sup. Table 3 A-D, Sup. Fig. 2A). We also observed enrichment in DNA repair pathway list only in selected cases, to further investigate this signal we noted, that *BRCA1* and *TP53* are present in both this and autosomal dominant model list. We removed these two genes from both lists and repeated analysis. Autosomal dominant genes remain highly significant and DNA repair pathway genes association signal is lost (Sup. Table 3 E-F). We further considered autosomal dominant model to be the only significantly associated. Separate analysis of protein-truncating variants (nonsense, frameshift and essential splice site) and damaging missense (Supplementary Methods) was performed. Unselected cases ($p=6.41 \times 10^{-6}$; OR=2.45; OR CI=1.66-3.56) show similar significant enrichment to genetically enriched cases ($p=5.26 \times 10^{-10}$; OR=3.67; OR CI=2.47-5.37) with rare (minor allele count less or equal to 10) protein-truncating variants only, while we observed no enrichment in damaging missense variation (p=0.68 and p=0.37 for selected and unselected respectively) (Fig. 1). It is worth noting, however, that selected cases dataset was assembled by initial genetic screening of probands that satisfy NCCN genetic testing criteria. If tested positive, they were not subsequently included in this study, and it is possible that genetically enriched cases have had

more genetic screening in general and that some diagnosed cases were removed before being entered in this study sample – likely attenuating the strength of association to the group with known autosomal dominant cancer predisposition genes. We tried multiple analyses for the recessive model, including counting of samples with more than 1 heterozygous genotype in the same gene and expanding the set of included variants up to minor allele frequency less than 1%, however the counts were still very low and inconclusive, thus the recessive model was not further tested. We looked into the frequency spectrum for the variants with MAC<=10 and observed that this association signal is driven almost entirely by singletons (Sup. Fig. 2B).

We then asked whether there were any additional features characterizing which genes within the autosomal dominant list were driving the truncating variant association signal. Using a metric of genic tolerance to truncating variation (pLI) defined by the Exome Aggregation Consortium (ExAC), we separately estimated association in genes tolerant of loss-of-function mutations (pLI<0.1) and intolerant of such variation (pLI>0.9) (Sup. Fig. 3A). While this list contains genes that carry either heritable risk of cancer or high-risk somatic mutations (or both) we observe high enrichment of protein-truncating variants in highly tolerant genes ($p=1.5 \times 10^{-6}$ selected cases and $p=3 \times 10^{-4}$ unselected cases, Sup. Fig 3B, 3C), consistent with modest selection pressure due generally post-parental age of the disease onset.

Using these identified properties of the known cancer susceptibility genes, we can infer what features we should expect to observe in novel germline candidate

genes. We therefore targeted our search for mutations (primarily truncating) with autosomal dominant model of inheritance, in genes tolerant of loss-of-function mutations (as predicted by pLI score metric) and driven by a substantial burden of singletons (or independent variants) in both genetically enriched and unselected cases.

Case-control analysis

We applied this search strategy to our complete dataset. We found 4021 and 6254 singleton protein-truncating variants in selected and unselected cases respectively. We kept only genes with pLI<0.1 for further analysis. Because of earlier demonstrated significant contribution of inherited risk in unselected cases we joined both case cohorts for further analysis. Considering the burden of singleton truncating variants, among the top 5 genes identified with our methodology, 3 are known cancer risk genes – *BRCA1, BRCA2* and *ATM*. While this serves as a good proof-of-concept and suggests that the exome sequencing and analysis approach has some degree of both sensitivity and specificity, there are clearly no significant novel candidates discovered from this approach (Supplementary Tables 4-6).

Individual Cohorts Analysis

We then performed analysis of the individual phenotypic cohorts for each cancer. Additional 3526 controls were matched to the unselected cases and were used as a replication set (Sup. Fig. 4). In addition to our primary analysis focused on

burden of protein-truncating variants, two other previously reported models for rare variant association studies (RVAS) were used for analysis[3] which added additional variants to the truncating variants: addition of the missense mutations (c-alpha, VT tests) and ultra-rare variation analysis (variants filtered for MAF<10^{-5} in ExAC). Detailed analysis of the cutaneous and ocular melanomas cohorts is available in our earlier report[4].

We performed gene-based rare variant association testing for each gene using the earlier developed composite multi-test model[4,5]. For analysis of the breast cancer patients, we eliminated all male samples from the dataset, resulting in comparison of 354 genetically enriched cases with 2190 matched controls. Despite the screening of the previously known *BRCA1* risk mutations in the breast cancer cohort we still observe genome-wide significant rare loss-of-function variants burden in this gene (Sup. Tables 7-9). Genes with p-value less than $1x10^{-4}$ were taken into replication. *MKL2* was also included in the short-list of genes as it appears second only to *BRCA1* in the burden of protein-truncating variants (Supplementary Table 8). According to GTEx database[6] – *MKL2* is primarily expressed in adipose and mammary tissues. Two genes show evidence of replication – *BRCA1 and HSD17B1*. Interestingly, four genes out of our short-list of candidates are known to be associated with worse outcome of breast cancer, once mutated or amplified in tumors – *BRCA1, HSD17B1, PCDHB15 and MED28*[7–10].

Similarly, *ATM* appears as a top gene in RVAS of the colon cancer cohort. Being a known predisposing gene for this phenotype it does not reach significance threshold due to statistical power limitations (Table 2, Sup. Tables 10-12). Another known colorectal cancer susceptibility gene *MSH2* ($P=5.8x10^{-4}$) was also rediscovered in these analyses (Sup. Table 11). Some of the top candidate genes such as *OBP2A* and *TMEM14C* do not have expression specificity to colon tissues.

Discussion

Our study develops a systematic approach for search of the novel cancer risk genes through analysis of the rare variation in the known susceptibility genes. Moreover, we observe notable enrichment of the known inherited genetic risk factors in the unselected cancer cohort (TCGA). Despite common beliefs that sporadic cancer cases are mostly elucidated by replicative oncogenesis[11], aging and carcinogen exposure, genetic predisposition plays substantial role in the disease onset, comparable to genetically enriched cohort. List of the genes that we used as a training model includes well-established high-risk genes. Expectedly, we did not find potential candidate genes in our dataset with comparable effect size. At the same time, lower-risk genes require large cohorts providing enough statistical power. Individual cohort analysis provides alternative approach to the search for candidate genes. We identified three potential candidates – *HSD17B1, PCDHB15, MED28* with reported association to worse outcome for patients with somatic mutations in these genes. The *HSD17B1* gene

produces an enzyme that catalyzes the conversion of estrone to estradiol and estrogen exposure influences risks of breast and endometrial cancer[8]. Instead of looking at the true cancer driver genes, inclusion of the missense variation allows to screen regulatory genes potentially altering functionality of the driver genes. This effect was observed in the recent analysis of the cutaneous melanoma[4] and here we suggest a short list of potential candidates. Intriguingly, Lu et al. found that overexpression of full-length *MED28* in HEK293 human embryonic kidney cells or human breast cancer cell lines caused significantly increased *in vitro* cell proliferation[12]. Also, functional studies report importance of the *MED28* for breast cancer progression. *MED28* modulates cell growth through FOXO3a and NFkB in human breast cancer cells[10]. Not only is *MED28* involved in cellular migration and invasion but also in cell cycle progression in human breast cancer cells[10].

Analysis of the colon cancer cohort is strongly influenced by the small cohort size, although it still reveals a known predisposition gene – *ATM.* While proving the concept, a significantly larger genetically enriched cohort would be needed to facilitate discovery of the new candidates.

Methods

Patient cohorts. Details could be found in Supplementary Methods.

Exome sequencing, Variant processing and calling. Whole exome libraries were prepared using a modified version of Agilent's Exome Capture kit and protocol, automated on the Agilent Bravo and Hamilton Starlet, followed by

sequencing on the Illumina HiSeq 2000. We used an aggregated set of samples consented for joint variant calling resulting in 37,607 samples (Supplementary Table 1). All samples were sequenced using the same capture reagents at the Broad Institute and aligned on the reference genome with BWA-MEM algorithm (version 0.7.12-r1039)[13] and the best-practices GATK/Picard Pipeline, followed by joint variant calling with all samples processed as a single batch using GATK v 3.1-144 Haplotype Caller[14–16]. The resulting dataset had 7,094,027 distinct variants. Haplotype Caller, which was used for the ExAC database[17], was also used to detect indels. Selected mutations in *CDKN2A, BRCA1* and *BAP1* were confirmed with Sanger sequencing.

We performed principal component analysis (PCA) on common (MAF>5%) autosomal independent SNPs to filter out all non-European samples with Eigenstrat[18]. Relatedness analysis among Europeans was conducted with PLINK[19,20] as suggested in the PLINK best practices. We used VEP[21] for functional annotation of the DNA variants. Common and rare variants analyses were conducted using PLINK/SEQ[22], which allows indexing of the large datasets. A burden test was used for rare protein truncating variants. In addition, the VT[23] and C-alpha[24] tests were chosen as an adaptive burden test and variance-component test, respectively, to complement each other and to boost the power of rare missense and protein truncating variation association detection[25]. See details in supplementary methods.

Statistical Methods. Gene-based association was performed using 3 distinct, but related, analytical frameworks. In the first analysis, a burden test was applied to all rare (MAF<1%) protein truncating variants (PTV) since the functional impact is presumed to be severe and most directly inferred. Then, to expand on all rare variants (missense and PTV), a second analysis using both the C-alpha and variable threshold (VT) tests was employed. A third analysis applied the burden test to examine "ultra-rare" (MAF<.0001; ExAC database[17]) variants as these may represent the most highly penetrant alleles. Genome-wide significance was determined by Bonferroni correction (0.05 /17,337 genes tested, i.e. p<2.88E-06). Additional multi-test model application examples are available in multiple phenotype studies[4,5].

References

1. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505,** 302–8 (2014).
2. Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N. Engl. J. Med.* **373,** 2336–2346 (2015).
3. Yu, H. *et al.* A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. *J. Clin. Invest.* **126,** (2016).
4. Artomov, M. *et al.* Rare Variant, Gene-Based Association Study of Hereditary Melanoma Using Whole-Exome Sequencing. *JNCI J. Natl. Cancer Inst.* **109,** 94–98 (2017).
5. Yu, H. *et al.* A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. *J. Clin. Invest.* **126,** 1603–1603 (2016).
6. GTEx Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).
7. Zhang, C. *et al.* The Identification of Specific Methylation Patterns across Different Cancers. *PLoS One* **10,** e0120361 (2015).
8. Setiawan, V. W., Hankinson, S. E., Colditz, G. A., Hunter, D. J. & De Vivo, I. HSD17B1 gene polymorphisms and risk of endometrial and breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **13,** 213–9 (2004).
9. Gunnarsson, C. *et al.* Amplification of HSD17B1 and ERBB2 in primary breast cancer. *Oncogene* **22,** 34–40 (2003).
10. Li, C.-I., Hsieh, N.-T., Huang, C.-Y., Chang, H.-C. & Lee, M.-F. MED28 Modulates Cell Cycle Progression in Human Breast Cancer Cells. *FASEB*

J. **29,** (2015).

11. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science (80-.).* **355,** 1330–1334 (2017).
12. Lu, M. *et al.* The Novel Gene EG-1 Stimulates Cellular Proliferation. *Cancer Res.* **65,** 6159–6166 (2005).
13. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
14. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).
15. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).
16. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
17. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
18. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
19. Purcell, S. PLINK.
20. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and

Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

21. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).
22. Https://atgu.mgh.harvard.edu/plinkseq/. PLINK/SEQ.
23. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86,** 832–8 (2010).
24. Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genet.* **7,** e1001322 (2011).
25. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

Table 1. Early Onset Breast Cancer RVAS female cases and controls.

		Target Set			Replication Set		
Gene Name	Method	Cases (N=354)	Controls (N=2190)	P	Cases (N=504)	Controls (N=1870)	P
ENPP5	VT	11	16	3.00E-06	3	14	1
BRCA1	PTV Burden	11	7	5.30E-06	5	3	**0.01367**
PCDHB15	Exac Burden	6	1	2.18E-05	0	0	1
USP35	VT	46	124	4.20E-05	15	53	0.76
CCDC9	VT	10	18	4.84E-05	5	12	0.224
COL5A2	C-alpha	23	72	7.38E-05	6	34	0.391304
MED28	C-alpha	7	6	8.54E-05	0	6	1
HSD17B1	C-alpha	15	22	8.77E-05	3	5	**0.039**
MKL2	PTV Burden	5	2	8.85E-04	2	1	0.2

Table 2. Early Onset Colon Cancer RVAS.

		Target Set			Replication Set		
Gene Name	Method	Cases (N=75)	Controls (N=7654)	P	Cases (N=190)	Controls (N=3526)	P
ATM	PTV burden	4	12	1.66E-05	1	6	0.31
OBP2A	ExAC burden	5	43	4.81E-05	1	9	0.26
TMEM14C	C-alpha	2	12	9.83E-05	0	10	1

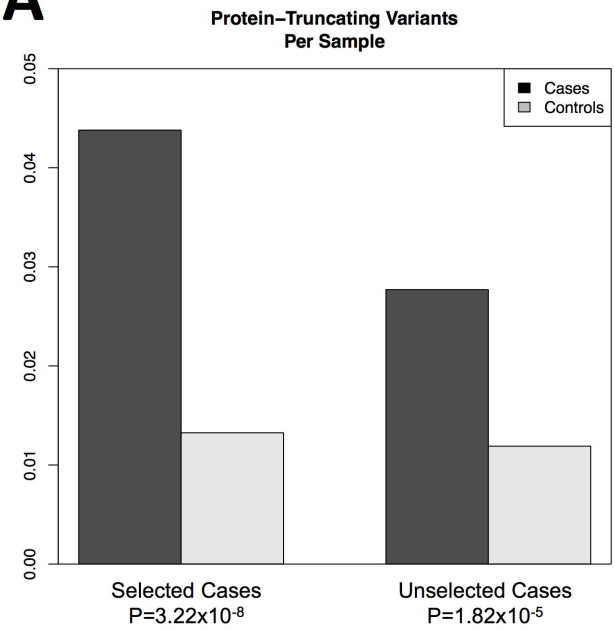
Figure Captions

Figure 1. Burden of rare (MAC<10) (A) protein-truncating variants per sample; (B) Damaging missense variants per sample; in autosomal dominant genes.

Supplementary figure captions could be found in supplementary information.

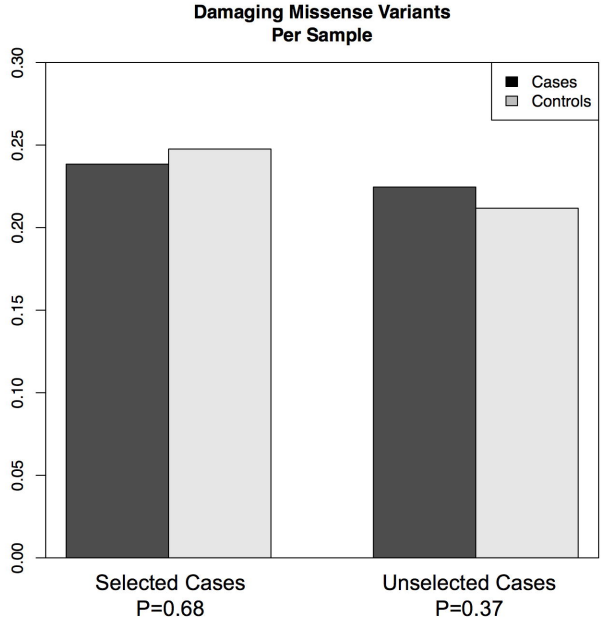
Fig. 1

A



Cohort	Alt Cases	Total Cases	Alt Controls	Total Controls	Fisher P
Selected Cases	40	822	105	7924	5.26×10^{-10}
Unselected Cases	44	1514	94	7924	6.41×10^{-6}

B



Cohort	Alt Cases	Total Cases	Alt Controls	Total Controls	Fisher P
Selected Cases	196	822	1962	7924	0.68
Unselected Cases	340	1514	1678	7924	0.37