bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Case-control analysis identifies shared properties of rare germline variation in cancer predisposing genes

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

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

Correspondence:

Mark J. Daly, 185 Cambridge St, CPZN-6823A, Boston, MA, 02114, USA;

mjdaly@atgu.mgh.harvard.edu; phone: (617)-726-5936

Kenneth Offit, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY, 10065;

offitk@mskcc.org; phone: (212)-639-2000

Authors declare no conflict of interests

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ABSTRACT

Traditionally, genetic studies in cancer are focused on somatic mutations found in tumors and absent from the normal tissue. Identification of shared attributes in germline variation could aid discrimination of high-risk from likely benign mutations and narrow the search space for new cancer predisposing genes. Extraordinary progress made in analysis of common variation with GWAS methodology does not provide sufficient resolution to understand rare variation. To fulfil missing classification for rare germline variation we assembled datasets of whole exome sequences from >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 and analyzed germline variation in known cancer predisposing genes to identify common properties of disease associated mutations and new candidate cancer susceptibility genes. Lists of all cancer predisposing genes were divided into subclasses according to the mode of inheritance of the related cancer syndrome or contribution to known major cancer pathways. Out of all subclasses only genes linked to dominant syndromes presented significant rare germline variants enrichment in cases. Separate analysis of protein-truncating and missense variation in this subclass of genes confirmed significant prevalence of protein-truncating variants in cases only in loss-of-function tolerant genes (pLI<0.1), while ultra-rare missense mutations were significantly overrepresented in cases only in constrained genes (pLI>0.9). Taken together, our findings provide insights into the distribution and types of mutations underlying inherited cancer predisposition.

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 and helped in mitigating the risk using appropriate clinical management[1]. Common in cancer genetics approach involves studying kindred with multiple samples and searching for DNA variation segregated between affected and non-affected members of the family. However, segregating mutations could be uniquely observed in a given kindred and do not provide compelling information about their capacity to explain cancer cases outside of kindred of interest. Multiple cohort-based studies of inherited variation in cancer with GWAS methods reached great success in identifying low to moderate risk common mutations[2]. Understanding rare coding variation on population scale requires massive genome/exome sequencing data both for cases and controls. Only recently sufficient statistical power was gained to discover new cancer susceptibility genes using case-control analysis of rare germline variation[3]. However, systematic description of rare inherited variation architecture in cancer cases in comparison to control subjects has not been reported yet.

In order to identify shared properties of rare germline variation in cancer 866 patients diagnosed with early onset and/or familial history either of breast cancer (MIM: [114480]), colon cancer (MIM: [114500]), cutaneous melanoma (MIM: [155600]), ocular melanoma (MIM: [155720]), Li-Fraumeni syndrome (MIM: [151623]) with family history and/or early onset of disease (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

materials, Supplementary Table 1) were recruited at MGH (Boston, USA), MEEI (Boston, USA), MSKCC (New York, USA), Andreas Sygros Hospital (Athens, Greece). Written informed consent was obtained from all individuals. Patients in this cohort were subjected to initial genetic screening (Supplementary Methods) and further identified as “selected cases”. We additionally included 1754 cancer samples with matching phenotypes from The Cancer Genome Atlas with no ascertainment for family history and age of onset. This cohort was identified as “unselected cases”. Control set of 24,612 samples was assembled from dbGAP whole exome studies of non-cancer phenotypes (Supplementary Table 1). Whole exome DNA sequences from all samples were aligned on a reference genome and processed through GATK Haplotype Caller[4–6] variant discovery pipeline as a single batch. Genotypes in assembled dataset were then subjected to quality check on per variant and per individual level (Supplementary Materials). To ensure close ancestral matching, we performed principal component analysis (PCA; Supplementary Figure 1A). For reduction of heterogeneity due to diverse population admixture, only the largest cluster of samples 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), resulting in total 846 selected cases, 1496 unselected cases and 7924 matched controls included in the final dataset used for analysis. Further, examination of common synonymous variants (MAF>5%) revealed a null-distribution of the Fisher’s exact test (two-sided)

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

statistic between cases and controls with genomic inflation factor λ=1.012 (Supplementary Figure 1B).

Results from *Zhang et al.*[7] provided good reference to known cancer susceptibility genes (germline and somatic) clustering based on dominant/recessive nature of linked cancer syndromes and contribution to common cancer pathways (Supplementary Table 2). We examined cumulative burden of rare (minor allele count of less or equal to 10) variants in cases compared to controls within each set of genes. Only genes linked to dominant cancer disorders exhibited significant burden in both selected and unselected cases compared to controls. Isolated analysis of damaging missense and protein-truncating variants (PTVs) established the main role of the latter in observed association signal (Supplementary Table 3A-D). We also observed enrichment in DNA repair pathway list only in selected cases, to further investigate this signal we noted, that *BRCA1* (MIM: [113705]) and *TP53* (MIM: [191170]) are present in both this and autosomal dominant disorders gene lists. 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). Interestingly, significant abundance of risk alleles was observed both in selected (Two-sided Fisher's test p=5.6x10^{-8}; OR=3.53; OR CI=2.26-5.39) and unselected cases (Two-sided Fisher's test p=1.28x10^{-5}; OR=2.49; OR CI=1.66-3.69), however, burden of risk alleles in selected group was greater than in unselected group (Figure 1A). Genes linked to breast cancer disorders carry substantial number of PTVs in controls, while

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

genes linked to more severe phenotypes, like Li-Fraumeni syndrome (*TP53*), uveal and cutaneous melanomas show no or very low count of PTV carriers in control cohort (Supplementary Figure 2). Downstream examination of allele frequency spectrum for mutations driving this association signal affirmed significance of singleton burden (Two-sided Fisher's test p=1.49x10^{-8}, p=2.36x10^{-5}; OR=5.25, OR=3.23; OR CI=2.99-9.01, OR CI=1.88-5.46; selected and unselected cases, respectively) while variants with minor allele count 2-10 did not show significant enrichment (Two-sided Fisher's test p=0.1, p=0.07 selected and unselected cases, respectively). Considering overrepresentation of PTVs in cases it was feasible to test genes linked to dominant cancer disorders for loss-of-function intolerance. We used probability of loss-of-function intolerance (pLI) from ExAC database[8] to separate genes into loss-of-function tolerant (pLI<0.1) and intolerant (pLI>0.9) groups (Supplementary Figure 3). Given that our case cohort does not have pediatric cancer patients we expectedly observed significant burden of singleton PTVs only in tolerant genes (Two-sided Fisher's test p=1.5x10^{-8}, OR=3.66, OR CI=2.03-6.36; p=3.0x10^{-4}, OR=2.74, OR CI=1.57-4.65, selected and unselected cases respectively, Figure 1B-C), as expected for adult onset disorder. While constrained genes are depleted in protein-truncating variants in cancer, we sought to test whether missense mutations are uniformly distributed between constrained and tolerant genes linked to dominant cancer syndromes. We did not observe any enrichment in damaging missense variants among cases using minor allele count of 1 (MAF~1x10^{-4}) and 1-10 (MAF~1x10^{-4} - 1x10^{-3}) as a frequency cutoff (Figure 1 D-F). To examine this further, we used

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

non-TCGA subset of ExAC[8] database to keep for analysis only variants with MAF<2.3x10^{-5} (not present in ExAC and singletons in ExAC non-Finnish Europeans). Ultra-rare missense variant analysis revealed significant burden in selected cases driven by loss-of-function intolerant gene contribution (Two-sided Fisher's test p=0.045, p=0.025; OR=1.26, OR=1.44; OR CI=1.00-1.58, OR CI=1.03-1.97; all and constrained autosomal dominant disorder genes, respectively; Figure 1 G-I). Previous analysis of cutaneous melanoma cohort used in this study identified *EBF3* (MIM: [607407]) as a new germline predisposition gene demonstrating tumor suppressor functional activity[3]. Interestingly, this gene has pLI=1 and carried ultra-rare missense variation in conserved protein domains, consistently with our observations above.

It is worth noting, however, that selected cases dataset was assembled by initial genetic screening of probands that satisfy NCCN genetic testing criteria[9]. If tested positive, they were not subsequently included in this study. Thus, genetically enriched cases have had more genetic screening and 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.

Further, we subjected selected breast cancer cohort to independent analysis, as the largest cohort in our dataset. We kept only female samples for analysis, resulting in 354 breast cancer cases and 2171 controls subjected to downstream analysis. First, we analyzed burden of singleton PTVs in loss-of-function tolerant

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

genes. *BRCA1* (p=1.34x10^{-8}, 8/355 cases and 5/7924 controls) and *BRCA2* (p=6x10^{-4}, 5/355 cases and 12/7924 controls) were two top genes in this analysis. Significance threshold was estimated using Bonferroni correction for number of genes tested – 0.05/7861 (3.34x10^{-5}). Analysis of missense singletons in constraint genes did not return any significant genes likely due to power limitations.

Overall, observed germline variation in both selected and unselected cases in established cancer susceptibility genes is linked to dominant cancer disorders, majorly represented by PTVs and has ultra-low frequency in population (singletons). While we observed ultra-rare missense mutations enrichment in cases, proportion of cases explained by this type of variation is likely very small, though potentially providing additional genes contributing to cancer pathways. Understanding power limitations of our study and potential effects of imbalance between cancer cohort sizes, yet our results provide a reference point for allele frequencies and variation type for future search of new genes contributing to inherited cancer susceptibility through rare DNA variation. We expect that overall majority of cancer cases would be explained by sporadic somatic mutations and inherited polygenic risk (mostly driven by common DNA variation), however, analysis of enriched kindred is bound to ultra-rare variation which our results could significantly aid.

WEB RESOURCES

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

OMIM, http://www.omim.org

ExAC database, http://exac.broadinstitute.org

Non-TCGA ExAC,

ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/subsets

All case and control genotypes are publically available through dbGAP database.

bioRxiv preprint doi: https://doi.org/10.1101/162479; this version posted February 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

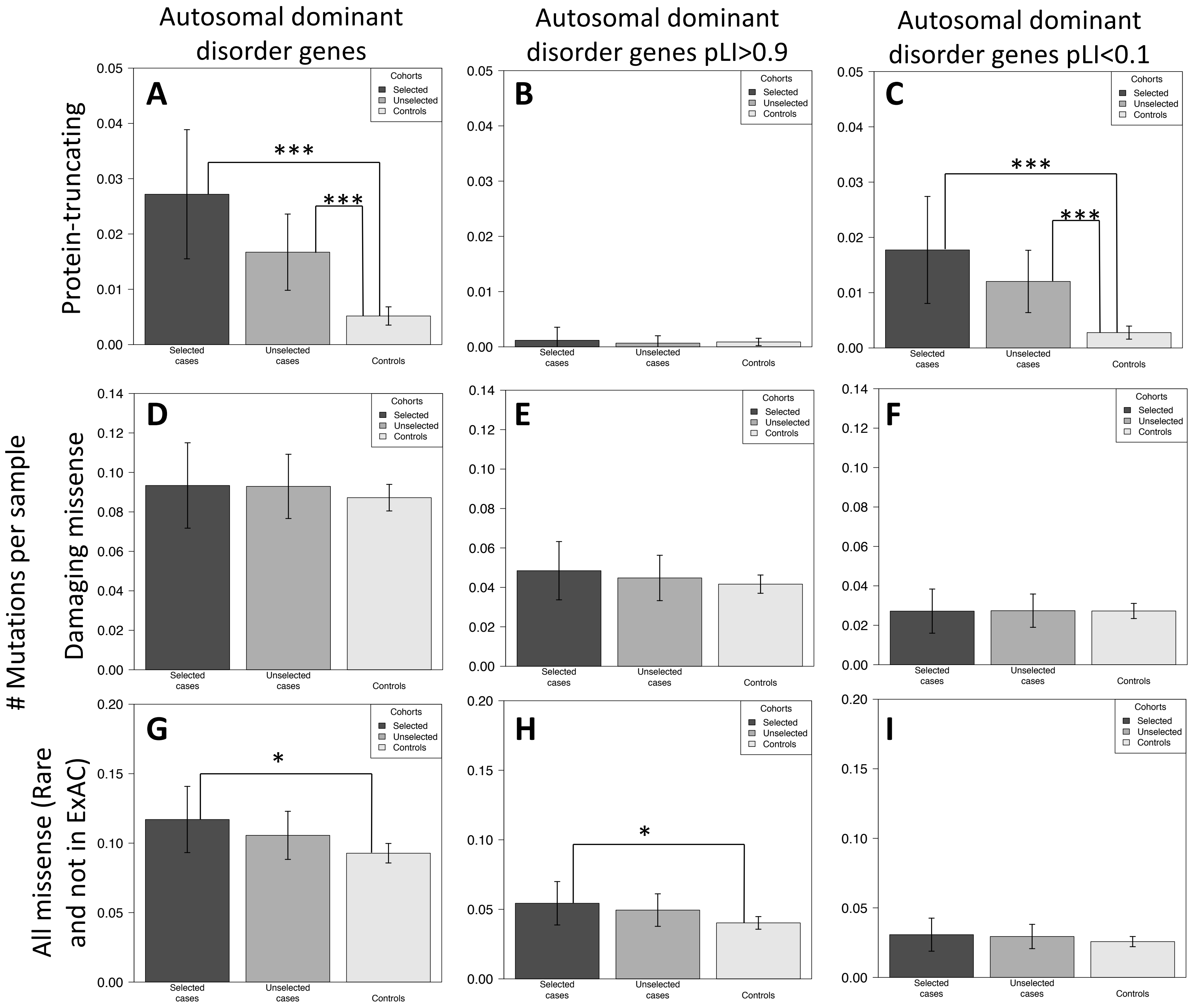
REFERENCES

1. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505,** 302–8 (2014).
2. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer* **17,** 692–704 (2017).
3. Artomov, M. *et al.* Rare Variant, Gene-Based Association Study of Hereditary Melanoma Using Whole-Exome Sequencing. *JNCI J. Natl. Cancer Inst.* **109,** (2017).
4. 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).
5. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
6. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).
7. Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N. Engl. J. Med.* **373,** 2336–2346 (2015).
8. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
9. NCCN. NCCN Guidelines. (2017).

FIGURE CAPTIONS

Figure 1. Mutational landscape overview. Mean mutation count per sample for protein truncating variants (MAC=1; MAF~1x10^{-4}) (**A-C**); Damaging missense mutations (MAC=1; MAF~1x10^{-4}) (**D-F**); Missense mutations that are not or rare (MAC=1; MAF=<2.3x10^{-5}) in non-TCGA ExAC (**G-I**); estimated across all genes linked to autosomal dominant disorders (**A, D, G**); autosomal dominant disorders linked genes with pLI>0.9 (**B, E, H**); autosomal dominant disorders linked genes with pLI<0.1 (**C, F, I**); * - p<0.05; *** - p<0.001.

Fig. 1



Autosomal dominant disorder genes
Autosomal dominant disorder genes pLI>0.9
Autosomal dominant disorder genes pLI<0.1
# Mutations per sample
Protein-truncating
Damaging missense
All missense (Rare and not in ExAC)
A
B
C
D
E
F
G
H
I
Cohorts
■ Selected
■ Unselected
□ Controls
Selected cases
Unselected cases
Controls
0.00
0.01
0.02
0.03
0.04
0.05
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.05
0.10
0.15
0.20
***
***
*