Supplementary Information

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*[1,2,13]

Patient cohorts

Primary discovery set. All patients provided written consent for this study and were enrolled at 4 sites- the Massachusetts General Hospital (MGH; cutaneous melanoma (CM), breast cancer (BC) and Li-Fraumeni syndrome patients), the A. Sygros Hospital in Athens, Greece (CM patients) and the Massachusetts Eye and Ear Infirmary (MEEI; ocular melanoma (OM) patients) in Boston, MA, the Memorial Sloan Kettering Cancer Center in New York, NY (MSKCC; BC and colon cancer (CC) patients) - in accordance with protocols approved at these institutions.

All probands were considered "genetically enriched" based on the following criteria.

1. **MGH**:
    a. **CM**: a histologically-proven CM *AND* at least one 1st degree affected relative *OR* $\geq$2 affected relatives on one side of the family regardless of degree of relationship (proband CM + relative with CM, "Familial CM/CM"; proband CM + relative with OM, "Familial CM/OM") *OR* $\geq$3 primary melanomas regardless of family history ("MPM CM-CM").
    b. **BC**: a histologically-proven BC AND at least one 1st degree affected relative (Familial BC) OR age at clinical diagnosis less than 40 years old (Mean=34.29, SD=3.49).
    c. **Li-Fraumeni syndrome**: Patients under age of 45 with multiple primary cancers without family history OR patients under age 45 with familial history of multiple cancers ($\geq$2). Patients were screened for *TP53* variation and found negative for disease-relevant DNA alterations. Cancer syndromes observed in the cohort included breast cancer (36.4%), adenoid cystic carcinoma (2.72%), basal cell carcinoma (3.66%), central nervous system lymphoma (3.64%), colon cancer (4.55%), ductal carcinoma in situ (0.91%), Ewing sarcoma (0.91%), lung cancer (1.82%), Lynch syndrome (6.36%), melanoma (3.64%), neuroblastoma (2.73%), non-Hodgkin lymphoma (1.82%), osteosarcoma (5.45%), ovarian cancer (1.82%), prostate cancer (1.82%), sarcoma (1.82%), squamous cell carcinoma (2.73%),

thyroid cancer (3.64%), thymoma (1.82%) and other cancer types (11.74%). Mean age of the first cancer onset – 29.7 (SD=11.3).

2. **The A. Sygros Hospital**: a histologically-proven CM *AND* ≥1 affected relative on one side of the family ("Familial CM/CM") *OR* ≥2 primary melanoma ("MPM CM-CM").
3. **MEEI**: a histologically or clinically diagnosed OM *AND* ≥1 relative affected with either CM or OM (proband OM + relative with OM, "Familial OM/OM"; proband OM + relative with CM, "Familial OM/CM") *OR* a second CM ("MPM OM-CM").
4. **MSKCC**:
    a. **BC:** a histologically-proven BC AND at least one 1st degree affected relative (Familial BC) OR age at clinical diagnosis less than 35 years old (Mean=31.82, SD=3.11).
    b. **CC:** 54.5% patients were diagnosed with hereditary nonpolyposis colorectal cancer (HNPCC), satisfied Amsterdam criteria and had age at diagnosis less than 45 years. 45.5% patients did not satisfy Amsterdam criteria but had age at clinical diagnosis less than 45 years old. Age distribution: Mean=36.05, SD=6.51.

Genetic screening

At the time of enrollment patients were analyzed with respect to NCCN guidelines for corresponding cancer type management to determine whether they match criteria for genetic screening. Individuals who satisfied the criteria were screened and if found positive, removed from downstream analysis. Breast cancer cohort were analyzed for germline variation in *BRCA1, BRCA2, PALB2, ATM, CHEK2, TP53*, colon cancer cohort was analyzed for germline variation in *MLH1, MSH2, MSH6, APC, MYH* and *TP53.* Li-Fraumeni cohort was analyzed for *TP53* germline variation. Cutaneous and ocular melanoma cohorts were screened for *CDKN2A* and *BAP1* variants, respectively, with Sanger sequencing, though several positively identified subjects were kept in the study (5 CDKN2A and 4 BAP1 loss-of-function variants were identified, respectively) as positive

controls. DNA variants identified by exome sequencing in this cohort were confirmed with Sanger sequencing data obtained at the time of screening.

Unselected cases cohort

We used TCGA exome sequencing data from the phenotypes corresponding to those represented in selected cases (TCGA phenotype codes: SKCM, BRCA, COAD, UM). Data used in this manuscript is available from dbGAP under accession number phs000178.v1.p1.

Sequencing data processing

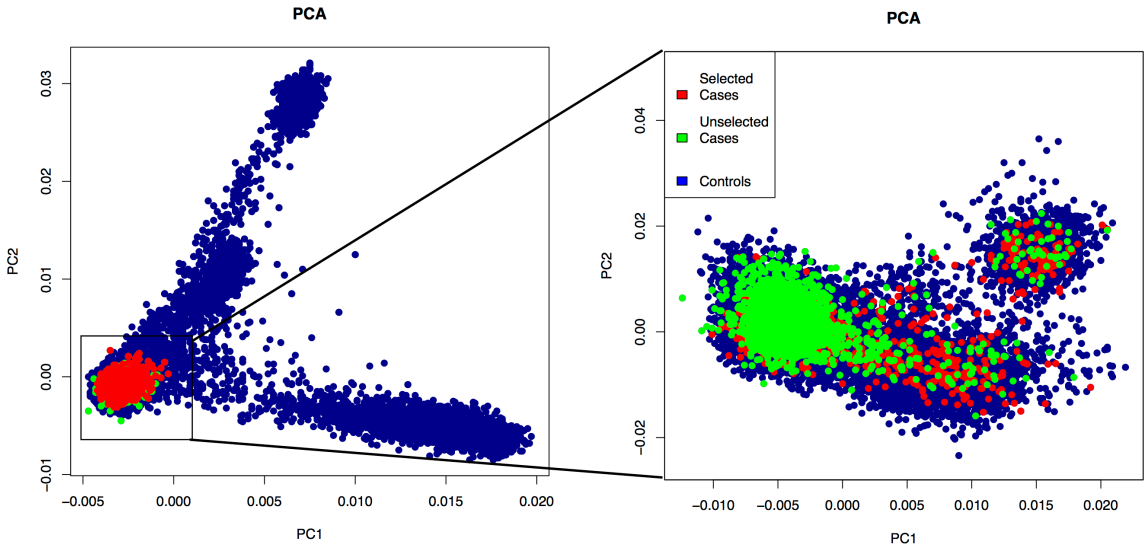
PCA. Downstream association studies require careful matching of cases and controls. Prinicpal component analysis was performed on the set of autosomal LD-prunned common variants (Purcell 5K) using Eigenstrat. Clustering of samples was performed using k-means clustering algorithm. Identification of ancestries within clusters was performed using 1000 Genomes samples as a training set.

Relatedness. Only European samples were subjected to downstream analysis. Within European PCA cluster we performed relatedness analysis on using PLINK and eliminated samples with PI_HAT>0.2. From the set of related samples the one with the greater average sequencing coverage was kept.

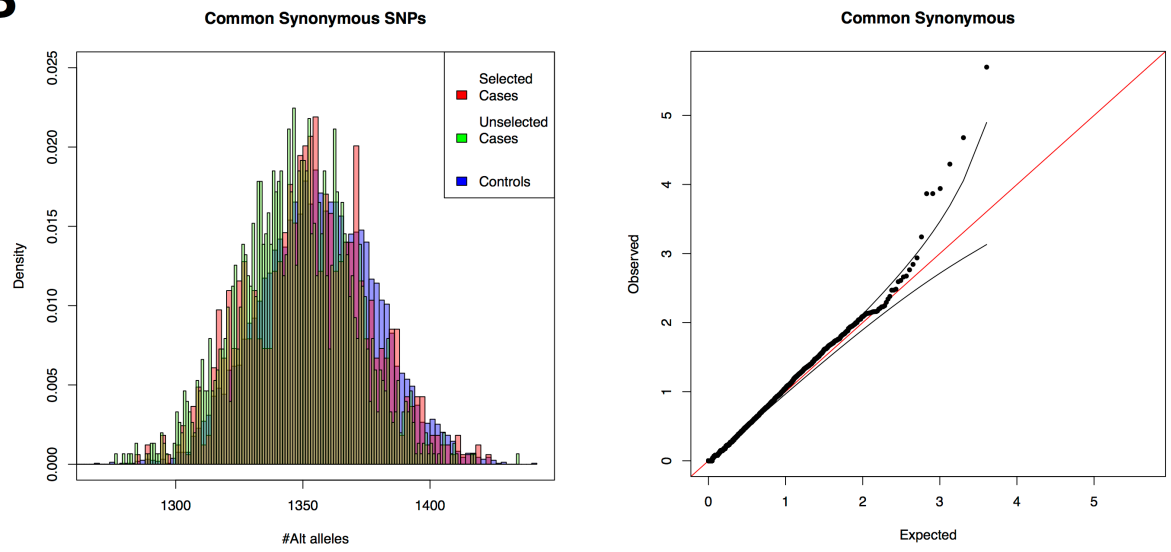
Case-control matching. Fisher's exact test was performed on the common synonymous variation to ensure the absence of systematic bias between cases and controls. Under the null hypothesis – common synonymous variation is benign and should be equally present in cases and controls. Null QQ-plot with per SNP p-values would thus prove the accuracy of matching.

Sup. Fig. 1

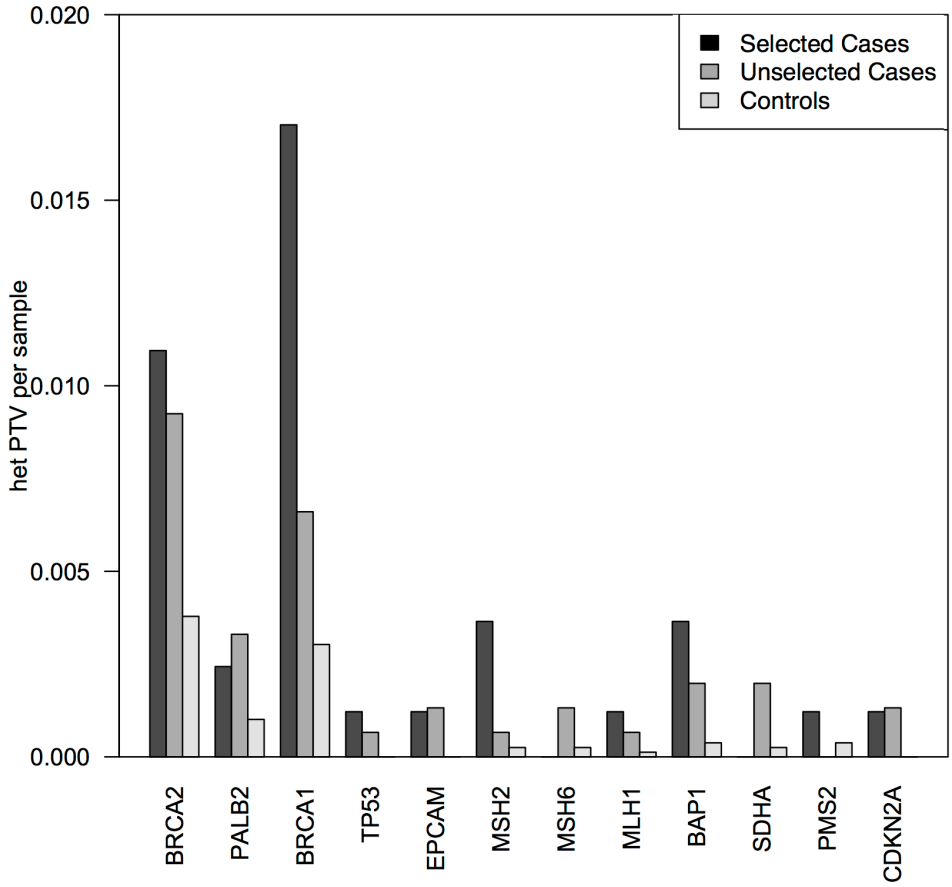
A



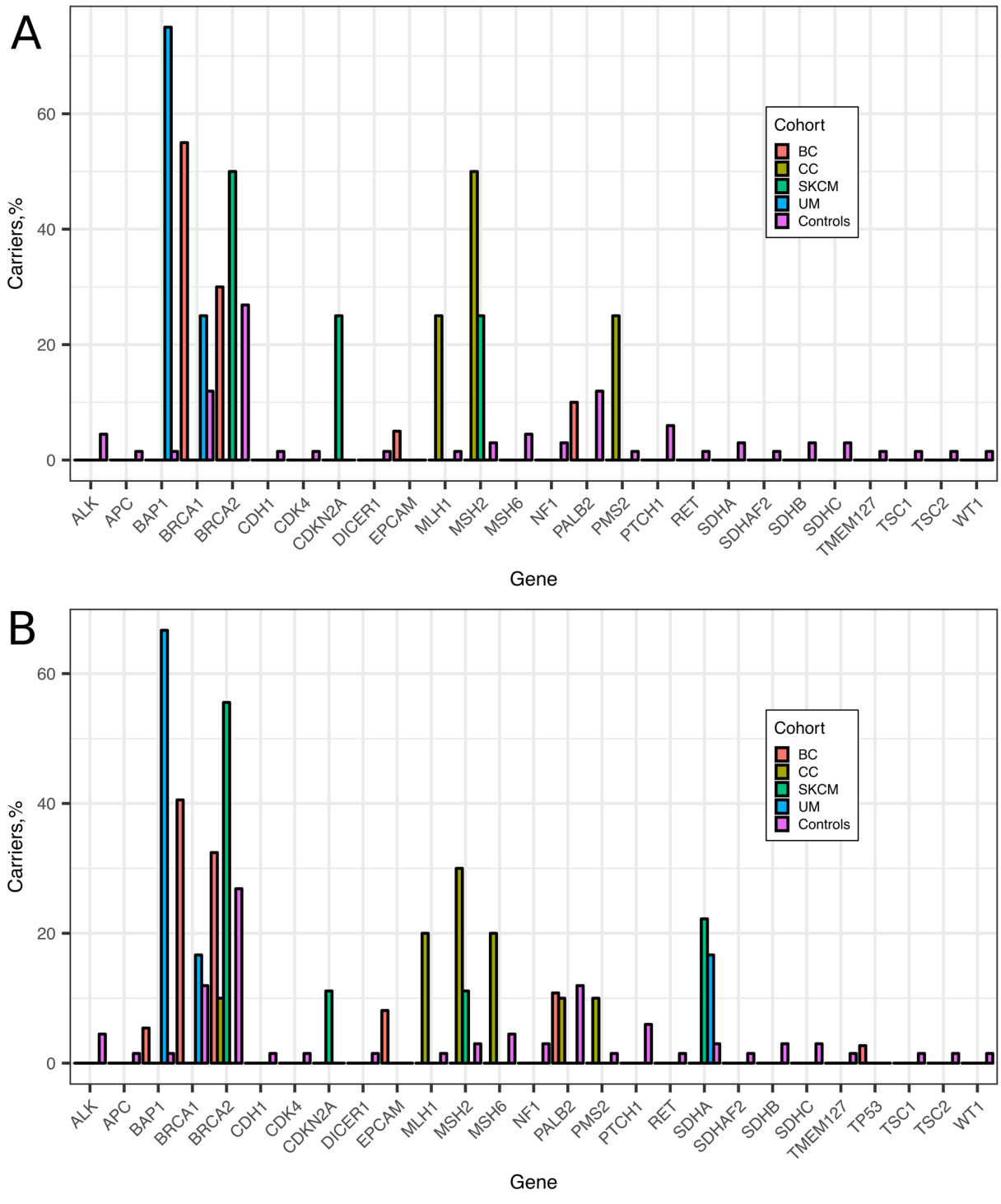
B



Supplementary Figure 1. (A) Principle component analysis on common autosomal variants; (B) Distribution of common synonymous variants burden and QQ-plot of Fisher's exact test, confirming case-control matching.

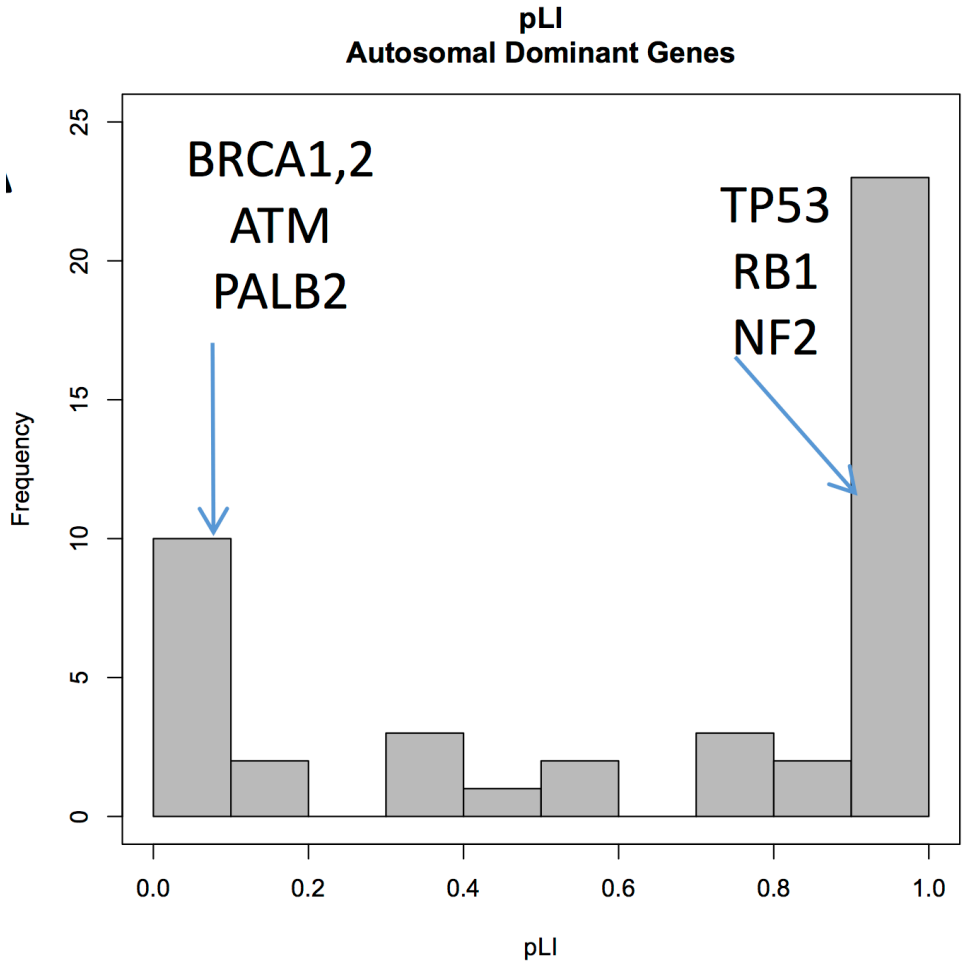


Supplementary Figure 2. Top genes contributing protein-truncating variants to the association signal from autosomal dominant disorders linked gene list.



Supplementary Figure 3. Carriers distribution across genes with at least 1 PTV. **(A)** Selected cases and controls; **(B)** Selected+unselected cases and controls.

BC – breast cancer, CC – colon cancer, SKCM – skin cutaneous melanoma, UM – uveal melanoma



Supplementary Figure 4. pLI distribution in autosomal dominant genes.

Supplementary Table 1. Dataset Summary

Genetically Selected Cases			
Phenotype	**Total Sequenced**	**Total Passed QC**	**Male/Female passed QC**
Breast Cancer	355	355	1/354
Colon Cancer	75	75	27/48
Cutaneous Melanoma	292	274	129/145
Ocular Melanoma	101	99	46/53
Li-Fraumeni	45	43	7/36
Total	868	846	210/636
Unselected Cases (TCGA)			
Breast Cancer	1060	820	9/811
Colon Cancer	250	250	135/115
Cutaneous Melanoma	397	379	233/146
Ocular Melanoma	47	47	27/20
Total	1754	1496	404/1092
Controls			
Controls	24612	7924	5689/2235
Samples used only for joint DNA variant calling			
Not used for analysis	10738	-	-

Supplementary Table 2. Known cancer genes tested.

Autosomal Dominant Genes: ALK, APC, BAP1, BMPR1A, BRCA1, BRCA2, CDC73, CDH1, CDK4, CDKN1C, CDKN2A, CEBPA, DICER1, EPCAM, FH, GATA2, MAX, MEN1, MLH1, MSH2, MSH6, NF1, NF2, PALB2, PAX5, PHOX2B, PMS2, PRKAR1A, PTCH1, PTEN, RB1, RET, RUNX1, SDHA, SDHAF2, SDHB, SDHC, SDHD, SMAD4, SMARCA4, SMARCB1, STK11, SUFU, TMEM127, TP53, TSC1, TSC2, VHL, WT1.
Autosomal Recessive: ATM, BLM, BRIP1, DDB2, ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, FANCA, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, MUTYH, NBN, NHP2, NOP10, RAD51C, RECQL4, SH2B3, TERT, WRN, XPA, XPC.
Tumor Suppressor: ARHGEF12, ARID1A, ASXL1, AXIN2, BARD1, BCL10, BCR, BUB1B, CAMTA1, CASP8, CBFA2T3, CDH11, CDKN1A, CDKN2C, CHEK2, CREBBP, CYLD, ETV6, EXT1, EXT2, FBXW7, FHIT, FLCN, FOXO1, FOXO3, FOXO4, FOXP1, GPC3, HOXB13, IDH1, IGF2R, KLF6, LRIG3, MAF, MAFB, MAP2K4, MN1, MUC1, NCOA4, NDRG1, NOTCH1, NR4A3, PBRM1, PHF6, PML, PMS1, PRDM1, SMO, SOCS1, SYK, TCF7L2, TET2, TFE3, TNFAIP3, TRIM24, WIF1, ZBTB16.

Supplementary Table 3. Gene list analyses. Variants with MAC<=10.

(A) Autosomal Dominant Genes					
Protein-Truncating Variants					
Selected Cases (N=846)	Controls (N=7924)	P (OR; OR CI)	Unselected Cases (N=1496)	Controls (N=7924)	P (OR; OR CI)
32	85	**3.16×10^{-8} (3.62; 2.32-5.54)**	40	85	**5.95×10^{-6} (2.53; 1.69-3.74)**
Damaging Missense Variants					
Selected Cases (N=846)	Controls (N=7924)	P (OR; OR CI)	Unselected Cases (N=1496)	Controls (N=7924)	P (OR; OR CI)
167	1615	0.69 (0.96; 0.8-1.15)	307	1615	0.92 (1.00; 0.88-1.16)
(B) Autosomal Recessive Genes					
Protein-Truncating Variants (Homozygotes or Double Hets)					
Selected Cases (N=846)	Controls (N=7924)	P (OR; OR CI)	Unselected Cases (N=1496)	Controls (N=7924)	P (OR; OR CI)
0	0	1.0 (-;-)	0	0	1.0 (-;-)
Damaging Missense Variants (Homozygotes or Double Hets)					
Selected Cases (N=846)	Controls (N=7924)	P (OR; OR CI)	Unselected Cases (N=1496)	Controls (N=7924)	P (OR; OR CI)
1	0	0.1 (-;-)	1	0	0.16 (-;-)
(C) Tumor-Suppressor Genes					
Protein-Truncating Variants					
Selected Cases (N=846)	Controls (N=7924)	P (OR; OR CI)	Unselected Cases (N=1496)	Controls (N=7924)	P (OR; OR CI)
2	43	0.32 (0.43; 0.05-1.67)	8	43	1 (0.99; 0.4-2.13)
Damaging Missense Variants					
Selected Cases (N=846)	Controls (N=7924)	P (OR; OR CI)	Unselected Cases (N=1496)	Controls (N=7924)	P (OR; OR CI)
83	661	0.15 (1.2; 0.93-1.52)	109	661	0.18 (0.86; 0.7-1.07)