PLOS GENETICS

RESEARCH ARTICLE

Contribution of Common Genetic Variants to Familial Aggregation of Disease and Implications for Sequencing Studies

Andrew Schlafly[1,2], Ruth M. Pfeiffer[3], Eduardo Nagore[4], Susana Puig[5], Donato Calista[6], Paola Ghiorzo[7], Chiara Menin[8], Maria Concetta Fargnoli[9], Ketty Peris[10,11], Lei Song[3], Tongwu Zhang[1], Jianxin Shi[3], Maria Teresa Landi[1☯]*, Joshua Neil Sampson[3☯]*

1 Integrative Tumor Epidemiology Branch: Division of Cancer Epidemiology and GeneticsNational Cancer Institute, Rockville, Maryland, United States of America, **2** Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **3** Biostatistics Branch: Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America, **4** Department of Dermatology, Instituto Valenciano de Oncología, València, Spain, **5** Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Spain, **6** Department of Dermatology, Maurizio Bufalini Hospital, Cesena, Italy, **7** Genetics of Rare Cancers, Department of Internal Medicine (DiMI), University of Genoa and Ospedale Policlinico San Martino Genoa, Genoa, Italy, **8** Immunology and Molecular Oncology Unit, Veneto Institute of Oncology IOV—IRCCS, Padua, Italy, **9** Department of Dermatology, Department of Biotechnological and Applied Clinical Sciences, University of L'Aquila, L'Aquila, Italy, **10** Institute of Dermatology, Catholic University, Rome, Italy, **11** Fondazione Policlinico Universitario A. Gemelli, IRCCS, Rome, Italy

☯ These authors contributed equally to this work.
* landim@mail.nih.gov (MTL); joshua.sampson@nih.gov (JNS)

Check for updates

OPEN ACCESS

Citation: Schlafly A, Pfeiffer RM, Nagore E, Puig S, Calista D, Ghiorzo P, et al. (2019) Contribution of Common Genetic Variants to Familial Aggregation of Disease and Implications for Sequencing Studies. PLoS Genet 15(11): e1008490. https://doi.org/10.1371/journal.pgen.1008490

Editor: Heather J. Cordell, Newcastle University, UNITED KINGDOM

Received: June 6, 2019

Accepted: October 23, 2019

Published: November 15, 2019

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Data Availability Statement: The coefficients for the PRS score and the list of the 29 variants are provided in the supplementary files.

Funding: The study at the Melanoma Unit, Hospital Clinic, Barcelona, was supported, in part, by grants from Fondo de Investigaciones Sanitarias, Spain (P.I. 12/00840, PI15/00956 and PI15/00716); by the CIBER de Enfermedades Raras of the Instituto de Salud Carlos III, Spain, co-funded by 'Fondo Europeo de Desarrollo Regional (FEDER). Unión Europea. Una manera de hacer Europa'; by AGAUR

Abstract

Despite genetics being accepted as the primary cause of familial aggregation for most diseases, it is still unclear whether afflicted families are likely to share a single highly penetrant rare variant, many minimally penetrant common variants, or a combination of the two types of variants. We therefore use recent estimates of SNP heritability and the liability threshold model to estimate the proportion of afflicted families likely to carry a rare, causal variant. We then show that Polygenic Risk Scores (PRS) may be useful for identifying families likely to carry such a rare variant and therefore for prioritizing families to include in sequencing studies with that aim. Specifically, we introduce a new statistic that estimates the proportion of individuals carrying causal rare variants based on the family structure, disease pattern, and PRS of genotyped individuals. Finally, we consider data from the MelaNostrum consortium and show that, despite an estimated PRS heritability of only 0.05 for melanoma, families carrying putative causal variants had a statistically significantly lower PRS, supporting the idea that PRS prioritization may be a useful future tool. However, it will be important to evaluate whether the presence of rare mendelian variants are generally associated with the proposed test statistic or lower PRS in future and larger studies.

2014_SGR_603 and 2017_SGR_1134 of the Catalan Government, Spain; by a grant from 'Fundació La Marató de TV3, 201331-30', Catalonia, Spain; by the European Commission under the 6th Framework Programme [contract no.: LSHC-CT-2006-018702 (GenoMEL)]; by CERCA Programme/Generalitat de Catalunya and by a Research Grant from 'Fundación Científica de la Asociación Española Contra el Cáncer' GCB15152978SOEN, Spain; and by a grant from the European Academy of Dermatology and Venereology (PPRC-2017/19). Part of the work was developed at the building Centro Esther Koplowitz, Barcelona. The studies from Genoa are partially supported by grants from the Italian Ministry of Health, RF-2016-02362288 and 5x1000 RC2019 to Ospedale Policlinico san Martino. The sponsors had no role in the design or conduct of the study; in the collection, analysis and interpretation of data; nor in the preparation, review or approval of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Multiple members in a family can be diagnosed with the same disease. In such families, genetics may be a significant factor in disease risk. However, it remains unclear whether such familial aggregation of disease is likely due to a single highly penetrant rare variant (HPRV), many minimally penetrant common variants, or a combination of the two types of variants. We therefore use recent estimates of SNP heritability and the liability threshold model to estimate the proportion of afflicted families likely to carry a rare, causal variant. We then show that Polygenic Risk Scores (PRS) may be useful for identifying families likely to carry such a rare variant and introduce a related statistic that can be used to select families for sequencing studies trying to identify HPRV.

Introduction

Genetics is the primary cause for familial aggregation of disease [1]. In afflicted families, members often share either a single highly penetrant rare genetic variant (HPRV) [2], a large number of minimally penetrant common variants [3], or both[4]. In this article we have two objectives. First, we show that a non-negligible fraction of familial aggregation may be attributable to common variants. Second, given this fact, we show that sequencing studies should try to select families with members having low Polygenic Risk Scores (PRS) (i.e. few risk alleles at common variants) when the study's objective is to identify new highly penetrant rare variants. Importantly, the genotyping needed for calculating PRS is usually inexpensive relative to full sequencing.

The importance of common variation in the familial aggregation of disease is demonstrated by the large estimates of SNP heritability [5] and the elevated PRS in members of afflicted families. High PRS has already been reported for families with migraines [6], dyslipidemia [7, 8], Crohn's disease [9], Alzheimer's [10], schizophrenia [11], and breast cancer with [12] and without [13–15] BRCA mutations. The increased number of common risk variants can either be solely responsible for the familial aggregation of disease or can exaggerate [16] the penetrance of an HPRV. In the first half of this paper, we suggest that for many diseases with high SNP heritability and a prevalence above 1%, a non-negligible proportion of afflicted families may not harbor an HPRV.

As only a fraction of afflicted families will have an HPRV, sequencing studies intended to identify new HPRV should preferentially select those families most likely to carry such a variant. For our discussion, we consider a sequencing study that compares the proportion of individuals in the selected families with a specific HPRV with the proportion of individuals in a large, biobank-sized, set of controls carrying that HPRV. Therefore, the power for this illustrative study, as well for many other types of sequencing studies, will be determined by the proportion of tested family members with the HPRV. As initially discussed in Jostins [17], such preferential selection can be achieved by selecting families with the lowest PRS. Therefore, in the second half of this paper, we define a PRS-based statistic for prioritizing families and explore the potential for preferential selection to increase the statistical power of the illustrative study. We acknowledge that prioritization may only be useful in the near future, while the genotyping needed to obtain the PRS is actionably less expensive than sequencing or we continue to work with families that have already been genotyped. Furthermore, we use results from a recent Whole-Exome-Sequencing (WES) Study of Melanoma Families from the MelaNostrum consortium to provide an example where families with putative HPRV appear to have lower PRS than other afflicted families.

Our article proceeds as follows. In the results section, we describe HPRV prevalence in afflicted families, demonstrate the potential for PRS-guided selection to increase study power, and discuss the relationship between PRS and the presence of HPRV in the MelaNostrum consortium. In the discussion section, we review the key points, relate our approach to the literature, and suggest next steps. In the methods section, we summarize notation, define our PRS statistic, provide full details on the scenarios used for evaluating familial aggregation, and describe the MelaNostrum consortium. Finally, it is important to note that this article, with its theory grounded in a simple liability threshold model, is intended to provide only rough approximations of HPRV prevalence and study power. However, despite the dependence on this model, we believe that our conclusions, at least qualitatively if not quantitively, are valid.

Results

We start by describing how the specific features of a disease affect the proportion of affected family members expected to carry an HPRV and the ability of our proposed PRS-based statistic to identify those families with members carrying an HPRV. Specifically, we consider these relationships in affected sibships (e.g. bottom level of Fig 1A) and multi-generational families (e.g. bottom two levels of Fig 1B). We explore the relationships using the simulations described in the Methods section and, for reference, we note that the notation used to describe the evaluated features are fully listed in the Methods: Notation section.

The first result is that the proportion of affected family members carrying an HPRV depends on the genetic architecture of the disease. The proportion (E_p) carrying an HPRV increases with increasing HPRV MAF (p_G), penetrance in individuals carrying a single HPRV (π_1), and disease frequency in the family; and the proportion decreases with increasing disease prevalence (π^*) and polygenic heritability (σ^2_P). We note, at this time, the polygenic heritability is usually an order of magnitude higher than the PRS heritability (σ^2_S). Figs 2 and S2 show examples for affected sibships and Figs 3 and S3 show examples for multi-generational families. For example, consider diseases that have a prevalence of $\pi^* = 0.02$ and a heritability of

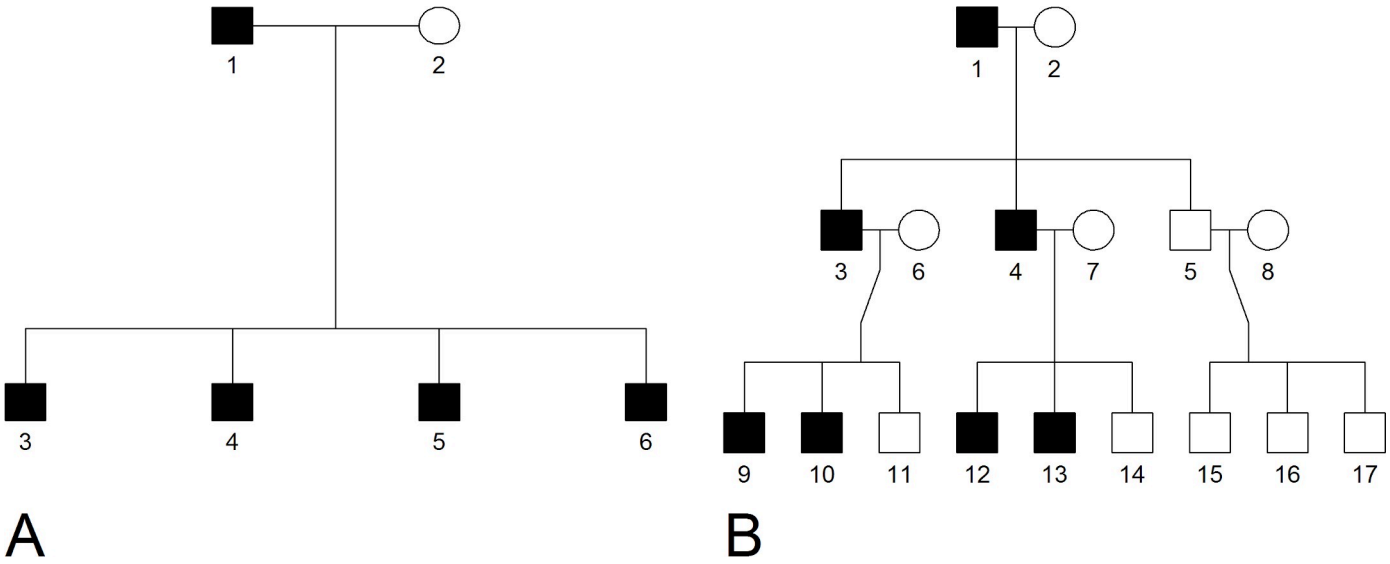

Fig 1. Examples family structure. (A) This example of the two-generation family structure has a founding couple with $n_1^D = 4$ affected children and $n_1^T = 4$ total children. (B) This example of the three-generation family structure has a founding couple with $n_1^D = 2$ affected children, $n_1^T = 3$ total children, $n_2^D = 4$ affected grandchildren, and $n_2^T = 9$ total grandchildren.

https://doi.org/10.1371/journal.pgen.1008490.g001

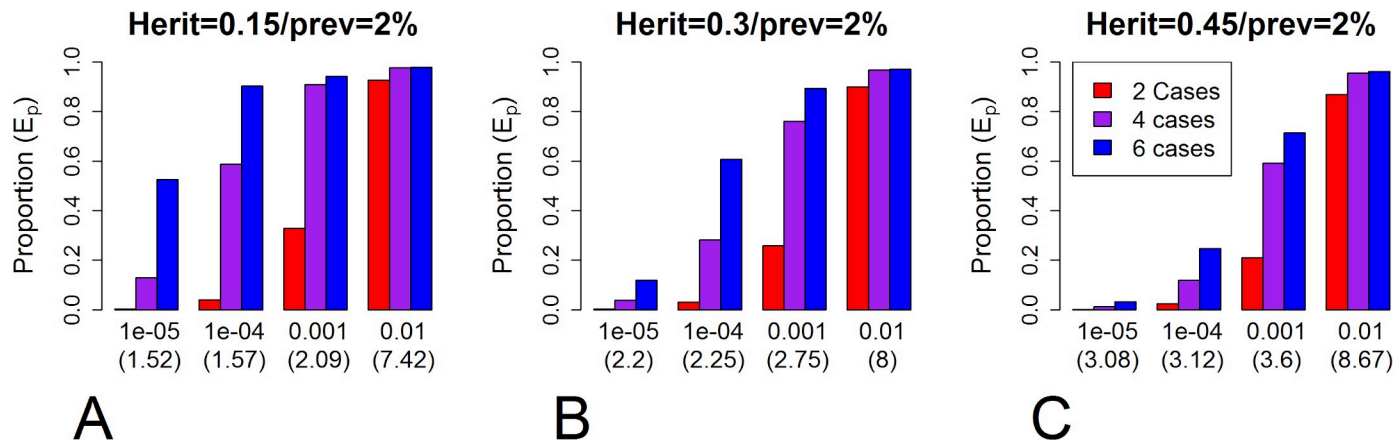

Fig 2. Proportion carrying HPRV in affected sibships. The proportion (E_p) of individuals carrying a highly penetrant rare variant (HPRV) in affected sibships for a disease with a 2% prevalence ($100 \times \pi^*$) in the population. Panels A, B, and C represent different values of polygenic heritability (A: $\sigma_P^2 = 0.15$, B: $\sigma_P^2 = 0.30$, C: $\sigma_P^2 = 0.45$). Within a panel, the four sets of bars represent different MAF (p_G = 0.00001, 0.0001, 0.001, or 0.01; note, the number in parenthesis is the resulting Sibling Relative Risk). Within a set of bars, the colors represent the different number of children (red: $n_1^T = n_1^D = 2$, purple: $n_1^T = n_1^D = 4$, blue: $n_1^T = n_1^D = 6$). See Supporting Information: S2 Fig for diseases with other prevalences.

https://doi.org/10.1371/journal.pgen.1008490.g002

$\sigma_P^2 = 0.3$ (e.g. [5] ischemic stroke, Barret's esophagus, schizophrenia). In sibships where all four siblings develop disease, approximately 5%, 30%, and 80% of the affected siblings will have an HPRV if the total MAF of HPRVs is 0.00001, 0.0001, and 0.001 (Fig 2B) in our evaluated scenarios. Note, the true MAF is unknown. In multi-generation families ($n_1^T = 3$, $n_2^T = 9$; with n_g^T being the number of biologically related individuals in the gth generation) where 6 of the 12 relevant individuals have disease, approximately 10%, 60%, and 90% of cases carry an HPRV when the MAF is 0.00001, 0.001, and 0.001 (Fig 3B). Note, percentages decrease dramatically when the prevalence increases to 5% (S2 and S3 Figs).

The second result is that families with larger values of our newly proposed test statistic, T_i, will be more likely to include individuals with an HPRV. Although the formal definition is postponed until the Methods, T_i reflects the expected number of HPRV per affected individual

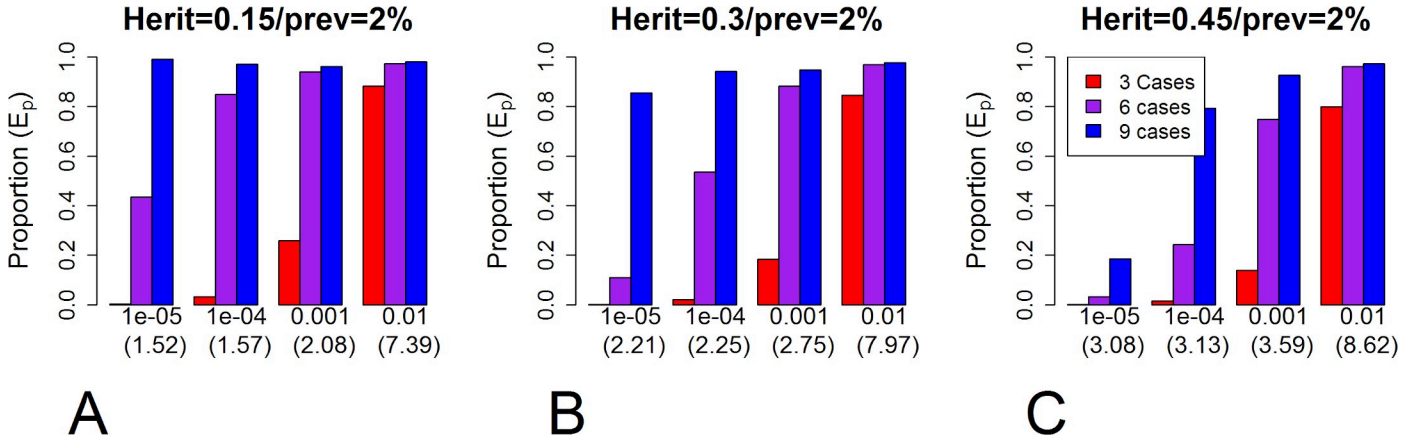

Fig 3. Proportion carrying HPRV in multi-generational families. The proportion (E_p) of individuals carrying a highly penetrant rare variant (HPRV) in affected multi-generational families for a disease with a 2% prevalence ($100 \times \pi^*$) in the population. Panels A, B, and C represent different values of polygenic heritability (A: $\sigma_P^2 = 0.15$, B: $\sigma_P^2 = 0.30$, C: $\sigma_P^2 = 0.45$). Within a panel, the four sets of bars represent different MAF (p_G = 0.00001, 0.0001, 0.001, or 0.01; note, the number in parenthesis is the resulting Sibling Relative Risk). Within a set of bars, the colors represent the different numbers of total affected individuals (red: $n_1^D + n_2^D = 3$, purple: $n_1^D + n_2^D = 6$, blue: $n_1^D + n_2^D = 9$). See Supporting Information: S3 Fig for diseases with other prevalences.

https://doi.org/10.1371/journal.pgen.1008490.g003

in a given family based on the pattern of disease and polygenic risk scores (PRS). We note that although T_i can be used to rank families, the ability of that ranking to differentiate families with and without an HPRV strongly depends on both the genetic architecture and predictive accuracy of the PRS. As an example, we describe HPRV enrichment in diseases that have a total polygenic heritability of $\sigma_P^2 = 0.3$, a prevalence of $\pi^* = 0.02$, and an HPRV MAF of $p_G = 0.0001$ (Fig 4). Here, enrichment, M, is defined as a ratio, where we divide the average number of HPRV in afflicted families at a given value of T_i by the average number of HPRV in all afflicted families. For an affected sibship with four cases, families with T_i at the top 20th percentile have a 2.5-fold enrichment in HPRV, as compared to the average affected sibship, when the PRS heritability is high ($\sigma_S^2 = \sigma_P^2 = 0.3$) and a 1.5-fold enrichment when the PRS heritability is more modest ($\sigma_S^2 = 0.05$). Here, σ_S^2 denotes the polygenic heritability captured by the PRS. For the multi-generational family with six cases, families with T_i at the top 20th percentile have a 1.8-fold enrichment and a 1.4-enrichment in HPRV when the PRS heritability is 0.3 and 0.05 respectively. We note that relatively high enrichment for HPRV even when the σ_S^2 is small is consistent with the behavior of the liability threshold model (see S1 Fig) which ensures individuals with low PRS are unlikely to have the disease.

The next result is that studies which preferentially select families with higher values of our proposed statistic can have increased power to detect associated HPRV. Details of the assumed study design and statistical test are presented in the Methods Section Briefly, we first assume that we select a group of families with a specific family structure for the sequencing study. Then, we compare the proportion of afflicted family members carrying a specific HPRV to the proportion of a large set of unaffected controls carrying that HPRV. Again, we consider the ideal example with $\sigma_P^2 = 0.3$, $\pi^* = 0.02$, and $p_G = 0.0001$ (Fig 4). Note, for the discussed families, these parameters are ideal because about 50% of the families carry an HPRV and therefore intelligent selection can be beneficial. For an affected sibship with four cases or the multi-generational family with 6 cases, a study that selects families with T_i in the top 20th percentile has notably higher power to detect a statistically significant association than a study that randomly selects the same number of families, even when the PRS heritability is modest ($\sigma_S^2 = 0.05$). Note, if using all families results in a power of approximately 0.5, it is expected that a study using only a randomly selected 20% of the families yields a power near 0. However, a study that enriches the HPRV in that 20% of subjects by a factor of 1.5–2.0x (i.e. selects families

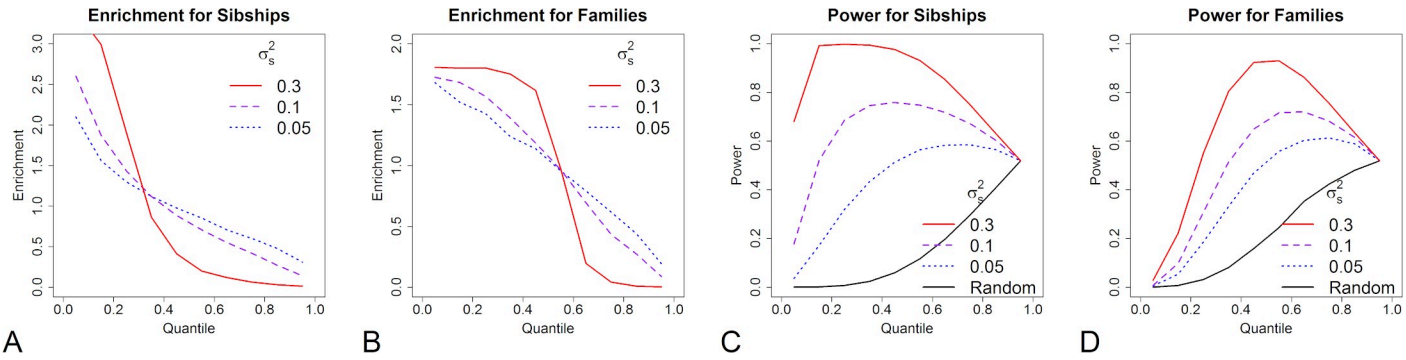

Fig 4. Enrichment and power to detect HPRV by the PRS statistic. Panels A and B show the enrichment (M) of HPRV as a function of the PRS statistic for a sibship with four affected individuals and a multi-generational family with six affected individuals. The X-axis is the quantile of the statistic (i.e. 0.1 represents a family at the top 10th percentile). Panels C and D show the power (β) to detect an association with the HPRV when a sibship has four affected individuals and when a multi-generational family has six affected individuals. The X-axis is the quantile of the statistic (i.e. 0.1. represents a study where we select the 10% of affected families with the highest PRS statistic). In all four panels, the color indicates the PRS heritability (red, purple, blue indicates $\sigma_S^2 = 0.3$, 0.1, and 0.05 respectively), the MAF is fixed at $p_G = 0.0001$, the polygenic heritability is fixed at $\sigma_P^2 = 0.3$, and the disease prevalence is fixed at $100\pi^* = 2\%$. See Supporting Information: S4 Fig for other diseases with other prevalences and HPRV MAFs.

https://doi.org/10.1371/journal.pgen.1008490.g004

based on T_i) results in approximately 10–20% power to detect a HPRV. Moreover, when that enrichment approaches 3x (e.g. $\sigma_S^2 = \sigma_P^2 = 0.3$ in affected sibships), the power using only the enriched subset can exceed that using all subjects. We note that for other examples, especially those where nearly all affected individuals carry the HPRV, selection by the PRS statistic can, at best, only modestly improve power.

Finally, we evaluated the MelaNostrum families. In this study, there are 229 families from Italy and 175 families from Spain with at least one melanoma case genotyped and sequenced. The median number of individuals per family is 11 and the corresponding interquartile range (IQR) is 6–19 individuals per family. The median number of cases per family is 3 with 34% and 15% of families having at least 4 and 6 cases per family, respectively. The total number of genotyped individuals is 990, the number of genotyped cases is 711, and the number of sequenced cases is 606. We identified 29 families where at least one affected individual had a rare potentially deleterious variant, as defined in the Methods, from *ACD* (3 families), *BAP1* (2), *CDKN2A* (8), *POT1* (11), *TERF2IP* (3), and *TERT* (2). We found that the mean family PRS (i.e. the average PRS over all affected and genotyped relatives) in these 29 families was significantly lower than the mean PRS in the 368 families without a known variant, 0.09 vs 0.55 (p-value = 0.01 from t-test adjusted for country). However, we did not find the family structure or disease pattern to be related to HPRV status. Even simple metrics, such as the number or proportion of family members with disease, did not significantly differ by HPRV status. Therefore, not surprisingly, the test statistic, T_i was similar in the two sets of families with the mean (Standard Error, SE) in families with and without an HPRV being 0.51 (0.05) and 0.44(0.02) respectively. Finally, we note that in this example, the PRS provided only modest contributions to T_i given the limited number of genotyped individuals in each family (Fig 5) and that the spearman correlation coefficient between the test statistics, T_i and T_i^*, that consider and do not consider PRS was 0.92.

Discussion

Genetics is the primary cause of the familial aggregation of diseases. Either sharing a highly penetrant rare variant (HPRV), risk alleles at multiple common variants, or a combination of both can lead to numerous cases of a disease within a family. In this paper, we aimed to describe the proportion of afflicted families likely to carry an HPRV, evaluate the ability for polygenic risk scores (PRS) to identify those families carrying an HPRV, and to offer a test statistic for prioritizing families for sequencing studies.

The first message is that for diseases with a prevalence $> 1\%$, a proportion of afflicted families will not carry an HPRV. Unfortunately, as the number or penetrance of HPRVs for any disease is unknown, we cannot yet predict the true proportion of afflicted families who will carry HPRV. The purpose of this paper is only to show that, at least in some scenarios, there may be a significant proportion of families who do not carry an HPRV. In the future, as we continue to collect genetic information about afflicted families and new HPRVs, we might be able to postulate which of the simulated scenarios most likely reflects truth. Currently, we are limited to combining observed sibling relative risks, estimates of polygenic heritability, and disease specific knowledge of genetics to determine the realistic boundaries of possible models. We note that for rare diseases (e.g. disease prevalence 0.5%), our simulations suggest that it is difficult to obtain significant familial enrichment (e.g. affected sibships with four siblings) without the presence of an HPRV.

The second message is PRS can be used to identify those families most likely to be harboring an HPRV. Although this was expected, the surprising finding was that families with a low average PRS tended to be significantly enriched for HPRV even when the heritability explained

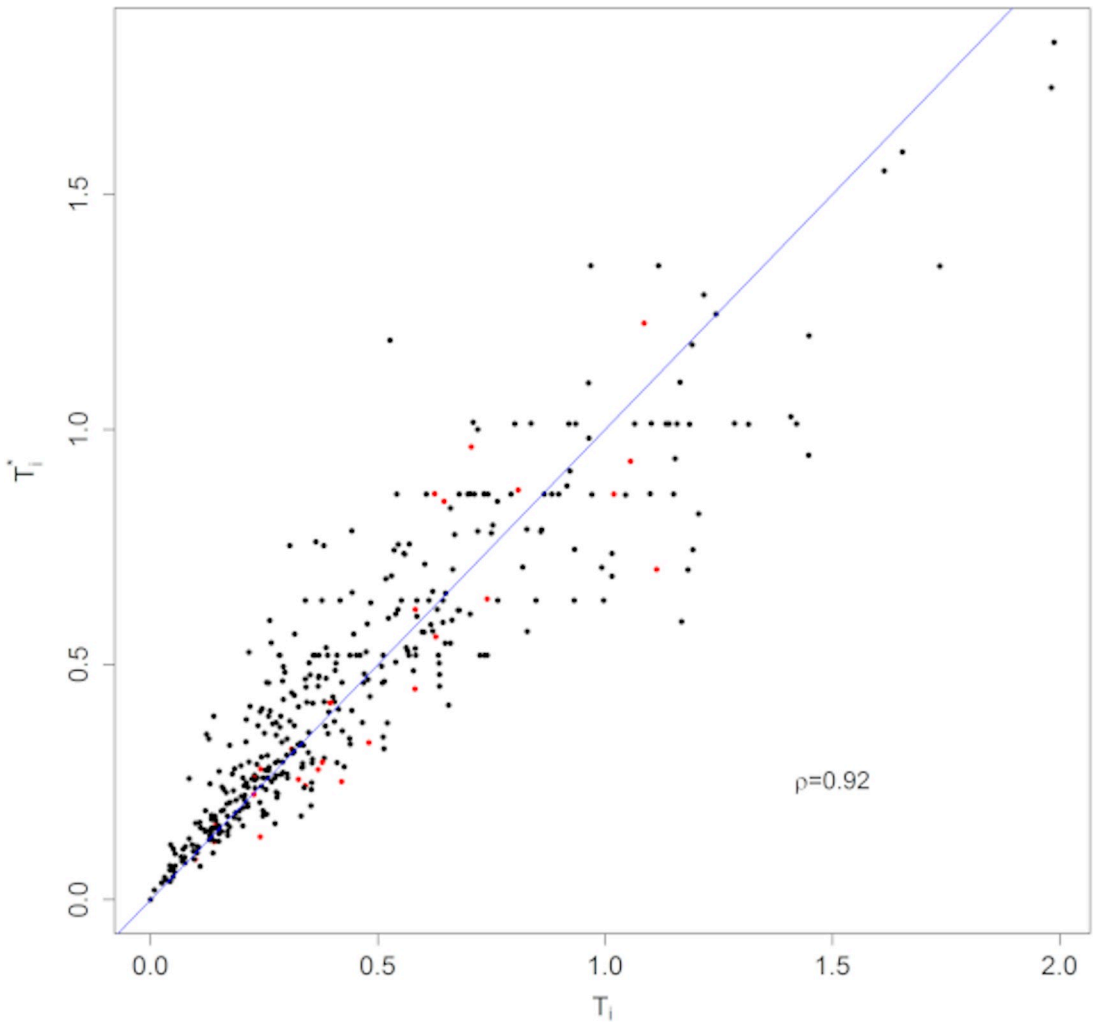

Fig 5. This figure compares the test statistic, T_i, that accounts for PRS and the test statistic, T_i^*, that does not account for PRS in the 404 families with Melanoma. Each dot represents one family and red dots indicate families with an HPRV. The spearmen correlation (ρ) between T_i and T_i^* is 0.92.

https://doi.org/10.1371/journal.pgen.1008490.g005

by the PRS was relatively low ($\sigma_S^2 = 0.05$), suggesting that current versions of PRS could be potentially useful in selecting families for sequencing studies for non-rare diseases. This is important, since the cost per sample of SNP genotyping is much lower than that for whole exome sequencing. However, before promoting its usage, it would be important to show, in multiple studies, that the PRS are lower in families carrying HPRV. MelaNostrum would be one example, albeit with limited sample size, to suggest that families with an HPRV have lower PRS, despite melanoma being a relatively rare disease in these Mediterranean populations (age standardized incidence rate per 100,000 persons = 6.9 and 11.4 in Spain and Italy, respectively [18]). However, the newly developed statistic, which in theory should be more strongly correlated with the presence of an HPRV, did not differ across the two sets of families with and without an HPRV. We expect that our statistic will perform better in future studies, but it is worth considering why the statistic likely failed to differentiate families with and without an HPRV in MelaNostrum. First, we note that our statistic combines two types of information, the disease pattern in the family and the PRS. The first type, or disease pattern, is a time-tested

means for identifying families carrying an HPRV; families with a large number of afflicted individuals tend to carry HPRV. The second type, or PRS, is the comparatively novel information used for identifying such families. However, in the MelaNostrum example, it is the disease patterns that do not appear to have utility. One partial explanation is that, in this study, where each family had only an average of three cases, the disease patterns might not have varied enough as to be able to differentiate carriers of HPRV. Otherwise, we are left with conjectures, such that families may have misreported disease incidence and there are many HPRV beyond those in predefined genes. If a low mean PRS, as opposed to a high T, continues to be a better predictor of the presence of an HPRV in future studies, the simpler statistic might be a better way to select families. Note, the superiority of the PRS statistic might be possible, if not likely, if identifying diseased individuals in extended families is prone to error.

We showed some extreme examples where simple case-control studies had higher statistical power when only using the subset of cases with the highest values of T_i, as opposed to all cases. In such scenarios, we could clearly gain power by using all cases if we modified the statistical test. For example, we might compare the proportion of controls carrying an HPRV to a weighted average of the genotypes in the affected family members, where the weights are lower for individuals with higher PRS. As sequencing costs continue to decrease and studies no longer need to select a limited number of individuals, weighting will become a more valuable use of PRS. Therefore, future investigations should evaluate the potential gains of weighting.

We are not the first to suggest using PRS to select individuals for sequencing studies and specifically built our approach upon the elegant ideas proposed by Jostins [17]. We extend prior ideas by introducing a new statistic that attempts to estimate the number of HPRVs, accounting for ungenotyped individuals by modeling the PRS as a multivariate normal distribution instead of imputing individual genotypes, and by considering the influence of the total polygenic heritability. We expect the performance of both statistics, as well as the simpler average PRS in a family, to be evaluated with actual data as it becomes available.

We have pointed out two main limitations of our study. First, both our predictions about the presence of an HPRV and the form of our test statistic rely on the liability threshold model. Although this model has been a staple of genetics for over 80 years [19], the model has never been evaluated for its accuracy in estimating the predictive capacity of polygenic risk scores. Second, without knowledge of either the number or penetrance of these rare causal variants, we cannot offer precise predictions about the true proportion of families carrying an HPRV or the utility of PRS in selecting families for sequencing. Finally, it is worth highlighting a general limitation of PRS. For most diseases, PRS were developed in European populations and their predictive accuracy will be lower in other populations. Therefore, the benefits of our proposed method for selecting afflicted families may be greatly diminished when the PRS was, in fact, developed in a European population but the study's families have non-European ancestry. The performance of the method would be further comprised if the imputation quality of the imputed SNPs used in the PRS was poor because the families' ancestries were not properly included on the imputation panel. Despite these limitations, we have proposed a new test statistic for identifying families with an HPRV and believe that the performance of the statistic may be worth exploring in future studies.

Methods

We divide our Methods into four sections, describing the notation, the PRS-statistic, the scenarios for inquiry, and the WES study of Melanoma.

Notation

We assume that there is a single highly penetrant rare variant (HPRV), but note that our framework holds for multiple HPRVs with appropriate modifications to the definitions (e.g. "p_G" would be the combined frequency of all HPRVs).

n_F is the total number of families available for sequencing ($i \in \{1,\ldots n_F\}$ indexes families).

N_i is the number of individuals in family i ($k \in \{1,\ldots N_i\}$ will index family members).

$\Omega_i \subset \{1, \ldots N_i\}$ is the subset of diseased individuals to be sequenced.

$D_{ik} \in \{0,1\}$ indicates the disease status; $D_i = \{D_{i1}, \ldots, D_{iN_i}\}$.

$G_{ik} \in \{0,1,2\}$ indicates the genotype at the rare variant; $G_i = \{G_{i1}, \ldots, G_{iN_i}\}$.

S_{ik} is the Polygenic Risk Score (PRS); $S_{i\cdot} = \{S_{i1}, \ldots, S_{iN_i}\}$.

p_G is the Minor Allele Frequency (MAF) of the variant defining G_{ik}.

$\sigma_S^2 = \mathrm{var}(S_{ik})$ is the polygenic risk score heritability.

$\pi_\cdot = \{\pi_0,\pi_1,\pi_2\}$ are the prevalences/penetrances of disease in individuals with G_{ik} = 0, 1, and 2.

$\pi^* = \pi_0(1-p_G)^2 + \pi_1 2p_G(1-p_G) + \pi_2 p_G^2$ is the overall prevalence of disease in the population.

σ_P^2 is the total polygenic heritability.

We are assuming that the liability threshold model, described in the next paragraph, is true and that the underlying disease liability (L_{ik}) can be decomposed into a polygenic effect from identified sources (S_{ik}), a polygenic effect from unidentified sources (S_{ik}^U), and a non-genetic effect (ϵ_{ik}) with $L_{ik} = S_{ik} + S_{ik}^U + \epsilon_{ik}$ and the polygenic heritability defined by $\sigma_P^2 = \mathrm{var}(S_{ik} + S_{ik}^U)$.

Liability threshold model

The disease liability and its three components follow multivariate normal distributions. Letting I be the $N_i \times N_i$ identity matrix and Σ_i be the kinship matrix for family i,

$$L_{i\cdot} = S_{i\cdot} + S_{i\cdot}^U + \epsilon_{i\cdot}$$

$$S_{i\cdot} \sim N(0, \sigma_S^2 \Sigma_i)$$

$$S_{i\cdot}^U \sim N(0, \sigma_U^2 \Sigma_i)$$

$$\epsilon_{i\cdot} \sim N(0, \sigma_\epsilon^2 I)$$

where $S_{i\cdot}$, $S_{i\cdot}^U$, and ϵ_i are independent of each other and $\sigma_S^2 + \sigma_U^2 + \sigma_\epsilon^2 = 1$.

The disease thresholds ($c_{i\cdot}$) depend on Genotypes ($G_{i\cdot}$) and population-level penetrances (π_0, π_1,π_2). In other words, the disease threshold for a given person depends on whether that individual has a rare variant (i.e. in the presence of a rare variant, liability need not be as large for the disease to occur). Letting Φ be the cumulative distribution for the standard normal distribution, define $c_g = \Phi^{-1}(1-\pi_g)$ and $c_{ik} = c_{G_{ik}}$. Then, the liability threshold model states that

the disease status for an individual is determined by

$$D_{ik} = \begin{cases} 1 \; if \; L_{ik} \geq c_{ik} \\ 0 \; if \; L_{ik} < c_{ik} \end{cases}$$

PRS statistic

We will define our PRS statistic so that families with a higher value are, by some measure, more likely to carry a risk allele at the rare variant under investigation.

For defining the statistic, let $X_i \equiv X_i(D_{i\cdot}, G_{i\cdot}) = \sum_{k \in \Omega_i} G_{ik}$ be the total number of rare alleles among the n_i sequenced cases in a family and $V_i = var_0(X_i)$ be the variance of Xi under the null hypothesis, when the variant does not affect disease. Then, we define the PRS-statistic for the i^{th} family to be

$$T_i = \frac{E[X_i | D_{i\cdot}, S_{i\cdot}]}{\sqrt{n_i V_i}}$$

and show how to calculate T_i in the Supporting Information: S1 Doc, with the assumptions that the liability threshold model is true and all parameters are known. We show how to calculate T_i when only a subset of individuals have PRS information in the Supporting Information: S1 Doc. For comparison, we also consider the test statistic, T_i^*, that does not consider PRS, $T_i^* = E[X_i | D_{i\cdot}] / \sqrt{n_i V_i}$.

We offer some intuition behind our test statistic T_i, by rewriting the statistic as

$$T_i \propto \frac{E[X_i \; | D_{i\cdot}, S_{i\cdot}]}{n_i} \times \frac{\sqrt{n_i p_G (1 - p_G)}}{\sqrt{V_i}}$$

The first term on the right side is proportional to the expected number of HPRV per individual. The second term is the adjustment that accounts for the correlation of the genotypes among family members. In other words, we are not only interested in the average number of HPRV per individual, but we are interested in subsequent test statistic that accounts for the variance of that average.

Scenarios

We describe the metrics and scenarios that we use to evaluate the characteristics of familial aggregation and study power. We estimate the metrics by simulating populations (and studies) according to the liability threshold model. We start by identifying the pedigree of interest, which is either the sibship or multi-generational family. Based on the genetic parameters (i.e. MAF, $\sigma_S^2, \sigma_p^2, p_G$), we then simulate $>$ 50,000,000 pedigrees (i.e. $G_{i\cdot}$ $D_{i\cdot}$ $L_{i\cdot}$, $S_{i\cdot}$). Finally, we identify at least 10,000 families with the specified disease pattern and calculate their test statistics. From this group of families, we calculate all desired metrics.

We start by defining the three key metrics: E_p, $\beta(q_T)$ and $M(q_T)$. Our first objective is to estimate the proportion, E_p, of individuals in afflicted families that have at least one rare-variant

$$E_p = \mathrm{E}[1(G_{ik} > 0) | D_{ik} = 1]$$

and show that E_p is not always close to 1. Our second objective is to estimate the power, $\beta(q_T)$, of a hypothetical study that selects the $q_T \times n_F$ families with the highest values of T_i. Note, n_F is the total number of available families and, for our exposition, will be defined so that the study

power is 0.6 when $q_T = 1$. This hypothetical study then sequences all diseased individuals from the selected families, calculates an estimate, $\hat{p}_G(q_T)$, of the MAF in those individuals, and tests for an association using the statistic $Z(q_T) = (\hat{p}^*_G(q_T) - p^*_G)/\sqrt{var(\hat{p}^*_G(q_T))}$, where we assume the overall MAF accounts for 1000 rare variants, $p^*_G = p_G/1000$ and $\hat{p}^*_G = \hat{p}_G/1000$. Then the power is defined by

$$\beta(q_T) = \Pr(Z(q_T) > 5.32)$$

where the threshold is chosen to correspond to a p-value of 5×10^{-8}. As part of this second objective, we will also discuss the enrichment $M(q_T)$ of rare-variants in families at the top q^{th}_T percentile of T_i among afflicted families

$$M(q_T) = \frac{\mathrm{E}[1(G_{ik} > 0)|D_{ik} = 1, T_i = F^{-1}_T(1 - q_T)]}{\mathrm{E}[1(G_{ik} > 0)|D_{ik} = 1]}$$

where F_T is the distribution function of T_i.

We next define the scenarios for evaluating the three metrics. We consider two simple family structures (Fig 1). The two-generation family includes one founding couple with n^D_1 of their n^T_1 children having disease. The three-generation family has an additional n^D_2 of the n^T_2 grandchildren having disease, with the grandchildren evenly split among the children. Note, we only consider the individuals in the second and third generations (i.e. assume we cannot sequence the founding generation). We simulate either a large population of two-generation families with $n^D_1 = n^T_1 \in \{2, 4, 6\}$ or three generation families with $n^T_1 = 3, n^T_2 = 9$, and $n^D_1 + n^D_2 \in \{3, 6, 9\}$. We simulate these families assuming $\sigma^2_P \in \{0.15, 0.30, 0.45\}$, $\sigma^2_S \in \{0.1, 0.3\}$, $p_g \in \{0.00001, 0.0001, 0.001, 0.01\}$, and $\pi^* \in \{0.005, 0.02, 0.05\}$. We acknowledge that an allele with MAF = 0.01 would not be considered rare, but this MAF could represent the combined frequency of 1000's of rare variants. For all scenarios, $\pi_1 = 0.5$, and $\pi_2 = 0.5$.

MelaNostrum consortium

We considered melanoma-prone families participating in the MelaNostrum consortium. As part of this consortium, 6961 affected individuals from Italy, Spain, and Greece and 5553 controls were successfully genotyped using the Illumina OmniExpress Array. Moreover, a select subcohort of cases from 404 melanoma-prone families successfully received Whole Exome Sequencing (WES) or had been previously identified as carrying a *CDKN2A* mutation. Full details of genotyping, sequencing, and family selection have been described elsewhere [20, 21] and a brief summary is also provided in the Supporting Information: S1 Doc. For the PRS, we used the SNPs previously identified in the MelaNostrum consortium [22]. Of the 204 identified SNPs, 193 were carried by the subgroup of cases from the melanoma-prone families. We multiplied the genotype, coded as 0/1/2, by the beta coefficient for these SNPs and calculated the PRS by subtracting off the mean of the country-specific controls and dividing by the standard deviation in those same controls. For reference, we approximated the PRS heritability of melanoma in the Italy and Spain cohorts to be respectively 0.03 and 0.05 by performing linear regression with disease status and PRS as the dependent and independent variables, obtaining the coefficient $\hat{\beta}_{PRS}$ and estimating σ^2_S by $1 - var(R_i)/var(D_i)$ where $R_i = D_i - \hat{\beta}_{PRS}S_i$. We then identified the potential Highly Penetrant Rare Variants (HPRV) in one of 7 genes (*CDKN2A*, *CDK4*, *BAP1*, *TERT*, *POT1*, *ACD*, *TERF2IP*) previously associated with melanoma [23]. After creating a list of all variants passing QC filters, we removed variants meeting any of the following criteria: 1) Minor Allele Frequency (MAF) > 0.001 in ExAC Non-Finish Europeans,

MAF > 0.001 in both ExAC and the 1000 genome database, or MAF > 0.001 in our in-house Eagle controls (same population). 2) Listed as low/modifier consequences mutations ("3'UTR", "5'UTR", "3'Flank", "Targeted_Region", "Silent", "Intron", "RNA", "IGR", "Splice_Region", "5'Flank", "lincRNA"). 3) In WES samples, a note of "CScoreFilter" under FILTER column in VCF files or a genotype called as "0/0". We note that 33 variants were missense mutations, while the remaining three included two frameshift deletions (*CDKN2A*, *ACD*) and one non-sense mutation (*POT1*).

Supporting information

S1 Doc. This word document contains an expanded methods sections detailing genotyping and sequencing procedures, a discussion about the properties of the liability-threshold model, and results from additional simulations.
(DOCX)

S1 Tab. This .txt file lists the SNPs and their coefficients used for calculating the polygenic risk score.
(TXT)

S2 Tab. This .txt files lists the rare potentially deleterious variants identified in the Mela-Nostrum families.
(TXT)

S1 Fig. The relative risk of disease as a function of the PRS score. Note that the liability threshold model has interesting implications for polygenic risk scores (PRS). In brief, individuals in the population with a low PRS have an extremely low risk of disease. If the disease has a prevalence of 0.01, then the relative risk of disease (e.g. probability of disease divided by 0.01) in individuals at the bottom 10^{th} percentile are approximately 0.01, 0.2, and 0.4 for PRS heritabilities of 0.05, 0.1, and 0.3. Moreover, the total heritability is irrelevant to these calculations. These results emphasize that even when the polygenic risk scores account for minimal heritability, individuals in the lowest percentiles are at a greatly reduced risk of disease.
(TIF)

S2 Fig. Proportion carrying HPRV in affected sibships. The proportion (E_p) of individuals carrying a highly penetrant rare variant (HPRV) in affected sibships for a disease with a 0.5%, 2% or 5% prevalence ($100 \times \pi^*$) in the population. Columns represent different values of polygenic heritability and rows represent different prevalences. Within a panel, the four sets of bars represent different MAF (p_G = 0.00001, 0.0001, 0.001, or 0.01; note, the number in parenthesis is the resulting Sibling Relative Risk). Within a set of bars, the colors represent different the number of children (red: $n_1^T = n_1^D = 2$, purple: $n_1^T = n_1^D = 4$, blue: $n_1^T = n_1^D = 6$).
(TIF)

S3 Fig. Proportion carrying HPRV in multi-generational families. The proportion (E_p) of individuals carrying a highly penetrant rare variant (HPRV) in affected multi-generational families for a disease with a 0.5%, 2%, or 5% prevalence ($100 \times \pi^*$) in the population. Columns represent different values of polygenic heritability and rows represent different prevalences. Within a panel, the four sets of bars represent different MAF (p_G = 0.00001, 0.0001, 0.001, or 0.01; note, the number in parenthesis is the resulting Sibling Relative Risk). Within a set of bars, the colors represent different the numbers of total affected individuals (red: $n_1^D + n_2^D = 3$, purple: $n_1^D + n_2^D = 6$, blue: $n_1^D + n_2^D = 9$).
(TIF)

S4 Fig. Enrichment and power to detect HPRV by the PRS statistic. The first two columns show the enrichment (M) of HPRV as a function of the PRS statistic when a sibship has four affected individuals and when a multi-generational family has six affected individuals. The X-axis is the quantile of the statistic (i.e. 0.1 represents a family at the top 10^{th} percentile). The last two columns show the power (β) to detect an association with the HPRV when a sibship has four affected individuals and when a multi-generational family has six affected individuals. The X-axis the quantile of the statistic (i.e. 0.1. represents a study where we select the 10% of affected families with the highest PRS statistic). In all four panels, the color indicates the PRS heritability (red, purple, blue indicates $\sigma^2_S = 0.3$, 0.1, and 0.05 respectively) and the polygenic heritability is fixed at $\sigma^2_P = 0.3$. The MAF and disease prevalence are listed at the top of each graph.
(TIF)

Acknowledgments

We acknowledge the contributions of the participants, their families and the many clinicians, geneticists, genetic counsellors and allied health professionals involved in their management. This work was performed in participation with members of the Barcelona study center: Paula Aguilera, Llúcia Alós, Celia Badenas, Alicia Barreiro, Neus Calbet, Cristina Carrera, Carlos Conill, Mireia Domínguez, Daniel Gabriel, Pablo Iglesias, Josep Malvehy, M. Eugenia Moliner, Javiera Pérez, Ramon Pigem, Miriam Potrony, Joan Anton Puig Butille, Ramon Rull, Marcelo Sánchez, Gemma Tell-Martí, Sergi Vidal-Sicart, and Oriol Yelamos. Finally, this work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov).

Author Contributions

Conceptualization: Andrew Schlafly, Jianxin Shi, Maria Teresa Landi, Joshua Neil Sampson.

Data curation: Lei Song, Tongwu Zhang.

Formal analysis: Andrew Schlafly, Ruth M. Pfeiffer, Maria Teresa Landi, Joshua Neil Sampson.

Investigation: Eduardo Nagore, Susana Puig, Donato Calista, Paola Ghiorzo, Chiara Menin, Maria Concetta Fargnoli, Ketty Peris, Maria Teresa Landi.

Methodology: Andrew Schlafly, Ruth M. Pfeiffer, Lei Song, Tongwu Zhang, Jianxin Shi, Maria Teresa Landi, Joshua Neil Sampson.

Software: Lei Song, Tongwu Zhang.

Supervision: Maria Teresa Landi, Joshua Neil Sampson.

Writing – original draft: Andrew Schlafly, Maria Teresa Landi, Joshua Neil Sampson.

Writing – review & editing: Andrew Schlafly, Eduardo Nagore, Susana Puig, Donato Calista, Paola Ghiorzo, Chiara Menin, Maria Concetta Fargnoli, Ketty Peris, Jianxin Shi, Maria Teresa Landi, Joshua Neil Sampson.

References

1. Khoury MJ, Beaty TH, Liang KY. Can Familial Aggregation of Disease Be Explained by Familial Aggregation of Environmental Risk-Factors. Am J Epidemiol. 1988; 127(3):674–83. PubMed PMID: WOS: A1988M162600023. https://doi.org/10.1093/oxfordjournals.aje.a114842 PMID: 3341366
2. Online Mendelian Inheritance in Man, OMIM McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) [12/12/2018]. Available from: https://omim.org/.

3. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018; 50(9):1219-+. PubMed PMID: WOS:000443151300011. https://doi.org/10.1038/s41588-018-0183-z PMID: 30104762
4. Potrony M, Puig-Butille JA, Aguilera P, Badenas C, Tell-Marti G, Carrera C, et al. Prevalence of MITF p. E318K in Patients With Melanoma Independent of the Presence of CDKN2A Causative Mutations. Jama Dermatol. 2016; 152(4):405–12. PubMed PMID: WOS:000373916600012. https://doi.org/10.1001/jamadermatol.2015.4356 PMID: 26650189
5. Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ, Consortium U. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017; 49(7):986-+. PubMed PMID: WOS:000404253300006. https://doi.org/10.1038/ng.3865 PMID: 28530675
6. Gormley P, Kurki MI, Hiekkala ME, Veerapen K, Happola P, Mitchell AA, et al. Common Variant Burden Contributes to the Familial Aggregation of Migraine in 1,589 Families (vol 98, pg 743, 2018). Neuron. 2018; 99(5):1098-. PubMed PMID: WOS:000443712200015. https://doi.org/10.1016/j.neuron.2018.08.029 PMID: 30189203
7. Jarauta E, Perez-Ruiz MR, Perez-Calahorra S, Mateo-Gallego R, Cenarro A, Cofan M, et al. Lipid phenotype and heritage pattern in families with genetic hypercholesterolemia not related to LDLR, APOB, PCSK9, or APOE. J Clin Lipidol. 2016; 10(6):1397–405. PubMed PMID: WOS:000390829400015. https://doi.org/10.1016/j.jacl.2016.09.011 PMID: 27919357
8. Ripatti P, Ramo JT, Soderlund S, Surakka I, Matikainen N, Pirinen M, et al. The Contribution of GWAS Loci in Familial Dyslipidemias. Plos Genet. 2016; 12(5). PubMed PMID: WOS:000377197100067.
9. Levine AP, Pontikos N, Schiff ER, Jostins L, Speed D, Lovat LB, et al. Genetic Complexity of Crohn's Disease in Two Large Ashkenazi Jewish Families. Gastroenterology. 2016; 151(4):698–709. PubMed PMID: WOS:000389548500027. https://doi.org/10.1053/j.gastro.2016.06.040 PMID: 27373512
10. Tosto G, Bird TD, Tsuang D, Bennett DA, Boeve BF, Cruchaga C, et al. Polygenic risk scores in familial Alzheimer disease. Neurology. 2017; 88(12):1180–6. PubMed PMID: WOS:000397383000018. https://doi.org/10.1212/WNL.0000000000003734 PMID: 28213371
11. Boies S, Merette C, Paccalet T, Maziade M, Bureau A. Polygenic risk scores distinguish patients from non-affected adult relatives and from normal controls in schizophrenia and bipolar disorder multi-affected kindreds. Am J Med Genet B. 2018; 177(3):329–36. PubMed PMID: WOS:000427234000005.
12. Kuchenbaecker KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, et al. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. Jnci-J Natl Cancer I. 2017; 109(7). PubMed PMID: WOS:000405496200004.
13. Li HY, Feng BJ, Miron A, Chen XQ, Beesley J, Bimeh E, et al. Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. Genet Med. 2017; 19(1):30–5. PubMed PMID: WOS:000391911100005. https://doi.org/10.1038/gim.2016.43 PMID: 27171545
14. Muranen TA, Mavaddat N, Khan S, Fagerholm R, Pelttari L, Lee A, et al. Polygenic risk score is associated with increased disease risk in 52 Finnish breast cancer families. Breast Cancer Res Tr. 2016; 158 (3):463–9. PubMed PMID: WOS:000380711700006.
15. Sawyer S, Mitchell G, McKinley J, Chenevix-Trench G, Beesley J, Chen XQ, et al. A Role for Common Genomic Variants in the Assessment of Familial Breast Cancer. J Clin Oncol. 2012; 30(35):4330–6. PubMed PMID: WOS:000312195900014. https://doi.org/10.1200/JCO.2012.41.7469 PMID: 23109704
16. Begg CB. On the use of familial aggregation in population-based case probands for calculating penetrance. J Natl Cancer I. 2002; 94(16):1221–6. PubMed PMID: WOS:000177474200011.
17. Jostins L, Levine AP, Barrett JC. Using Genetic Prediction from Known Complex Disease Loci to Guide the Design of Next-Generation Sequencing Experiments. Plos One. 2013; 8(10). PubMed PMID: WOS:000326029300022.
18. Ward WF, JM. Cutaneous Melanoma: Etiology and Therapy. Brisbane, AU: Codon Publications; 2017.
19. Wright S. The results of crosses between inbred strains of guinea pigs, differing in number of digits. Genetics. 1934; 19(6):0537–51. PubMed PMID: WOS:000201599900003.
20. Shi JX, Yang XHR, Ballew B, Rotunno M, Calista D, Fargnoli MC, et al. Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. Nat Genet. 2014; 46(5):482–6. PubMed PMID: WOS:000335422900016. https://doi.org/10.1038/ng.2941 PMID: 24686846
21. Landi MT, Goldstein AM, Tsang S, Munroe D, Modi W, Ter-Minassian M, et al. Genetic susceptibility in familial melanoma from northeastern Italy. J Med Genet. 2004; 41(7):557–66. Epub 2004/07/06. https://doi.org/10.1136/jmg.2003.016907 PMID: 15235029; PubMed Central PMCID: PMC1735833.
22. Gu FY, Chen TH, Pfeiffer RM, Fargnoli MC, Calista D, Ghiorzo P, et al. Combining common genetic variants and non-genetic risk factors to predict risk of cutaneous melanoma. Hum Mol Genet. 2018; 27

(23):4145–56. PubMed PMID: WOS:000452536200012. https://doi.org/10.1093/hmg/ddy282 PMID: 30060076

23. Potrony M, Badenas C, Aguilera P, Puig-Butille JA, Carrera C, Malvehy J, et al. Update in genetic susceptibility in melanoma. Ann Transl Med. 2015; 3(15):210. Epub 2015/10/22. https://doi.org/10.3978/j.issn.2305-5839.2015.08.11 PMID: 26488006; PubMed Central PMCID: PMC4583600.