A genome-wide case-only test for the detection of digenic inheritance in human exomes

Gaspard Kerner[a,b], Matthieu Bouaziz[a,b], Aurélie Cobat[a,b], Benedetta Bigio[a,b,c], Andrew T. Timberlake[d,e,f,g], Jacinta Bustamante[a,b,c,h], Richard P. Lifton[d,e,i,j], Jean-Laurent Casanova[a,b,c,e,k,1], and Laurent Abel[a,b,c,1]

[a]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM UMR 1163, Necker Hospital for Sick Children, 75015 Paris, France; [b]University of Paris, Imagine Institute, 75015 Paris, France; [c]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065; [d]Department of Genetics, Yale University School of Medicine, New Haven, CT 06510; [e]HHMI, New York, NY 10065; [f]Section of Plastic and Reconstructive Surgery, Department of Surgery, Yale University School of Medicine, New Haven, CT 06510; [g]Hansjörg Wyss Department of Plastic Surgery, New York University Langone Medical Center, New York, NY 10016; [h]Study Center for Primary Immunodeficiencies, Necker Hospital for Sick Children, 75015 Paris, France; [i]Yale Center for Genome Analysis, Yale School of Medicine, New Haven, CT 06510; [j]Laboratory of Human Genetics and Genomics, The Rockefeller University, New York, NY 10065; and [k]Pediatric Hematology-Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France

Contributed by Jean-Laurent Casanova, June 3, 2020 (sent for review January 8, 2020; reviewed by Michael Boehnke and David FitzPatrick)

Whole-exome sequencing (WES) has facilitated the discovery of genetic lesions underlying monogenic disorders. Incomplete penetrance and variable expressivity suggest a contribution of additional genetic lesions to clinical manifestations and outcome. Some monogenic disorders may therefore actually be digenic. However, only a few digenic disorders have been reported, all discovered by candidate gene approaches applied to at least one locus. We propose here a two-locus genome-wide test for detecting digenic inheritance in WES data. This approach uses the gene as the unit of analysis and tests all pairs of genes to detect pairwise gene × gene interactions underlying disease. It is a case-only method, which has several advantages over classic case-control tests, in particular by avoiding recruitment of controls. Our simulation studies based on real WES data identified two major sources of type I error inflation in this case-only test: linkage disequilibrium and population stratification. Both were corrected by specific procedures. Moreover, our case-only approach is more powerful than the corresponding case-control test for detecting digenic interactions in various population stratification scenarios. Finally, we confirmed the potential of our unbiased, genome-wide approach by successfully identifying a previously reported digenic lesion in patients with craniosynostosis. Our case-only test is a powerful and timely tool for detecting digenic inheritance in WES data from patients.

digenic inheritance | next-generation sequencing | genome-wide | case-only | craniosynostosis

Next-generation sequencing (NGS) is now widely used and is gradually being optimized for the detection of rare and common genetic variants underlying human diseases (1–3). These advances, including whole-exome sequencing (WES) in particular, have led to major new findings in the field of human genetics, particularly for monogenic disorders (4–12). The growing number of reports of incomplete penetrance or variable expressivity of monogenic disorders suggests that additional genetic contributions, other than the mono- or biallelic causal lesions, may contribute to clinical manifestations and outcome (13, 14). Digenic inheritance is the simplest genetic model of this type with alleles at two different loci being necessary and sufficient to determine disease status (15, 16). The recently established Digenic Diseases DAtabase (DIDA) contains detailed information about digenic inheritance for 258 reported digenic combinations, corresponding to 54 conditions, since 1994 (17). Well-known examples relate to genetic modifier variants influencing the expression of the clinical phenotype caused by a primary disease-causing mutation. Cystic fibrosis (CF) is a classic example of a monogenic disease for which several genetic modifier variants have been identified. An elegant WES-based study showed that two low-frequency (minor allele frequency [MAF] <5%) missense variants of *DCTN4* were associated with the severity of pulmonary *Pseudomonas aeruginosa* infections in CF patients (18). One remarkable example of digenic inheritance explaining incomplete penetrance was recently provided for craniosynostosis. Timberlake et al. (19) found a highly significant enrichment in rare damaging *SMAD6* mutations in patients with craniosynostosis ($n = 191$). However, variants were also carried by 13 asymptomatic family members. The authors showed that a common variant close to *BMP2*, a *SMAD6*-related gene, accounted for almost all of the observed incomplete penetrance.

Only 1% of the 5,442 traits listed in Online Mendelian Inheritance in Man (OMIM) (20) as single-gene disorders are also known to display digenic inheritance and are listed in DIDA. Interestingly, all of the lesions known to be caused by defects with digenic inheritance were discovered in candidate gene studies, rather than through unbiased genome-wide statistical tests. In some cases, as for the CF example cited above, the defects were identified by single-gene analyses of patients with known disease-causing variants at the primary causal locus (18). The craniosynostosis example is unique in that its discovery

Significance

Despite a growing number of reports of rare disorders not fully explained by monogenic lesions, digenic inheritance has been reported for only 54 diseases to date. The very few existing methods for detecting gene × gene interactions from next-generation sequencing data were generally examined in rare-variant association studies with limited simulation analyses for short genomic regions, under a case-control design. We describe a case-only approach designed specifically to search for digenic inheritance, which avoids recruitment of controls. We show, through both extensive simulation studies on real whole-exome sequencing datasets and application to a real example of craniosynostosis, that our method is robust and powerful for the genome-wide identification of digenic lesions.

Author contributions: G.K., J.-L.C., and L.A. designed research; G.K., M.B., and A.C. performed research; G.K. and J.B. contributed new reagents/analytic tools; G.K., M.B., A.C., B.B., A.T.T., and J.B. analyzed data; and G.K., R.P.L., J.-L.C., and L.A. wrote the paper.

Reviewers: M.B., University of Michigan; and D.F., University of Edinburgh.

The authors declare no competing interest.

Data deposition: R scripts and additional information to perform digenic analyses are available at GitHub, https://github.com/GaspardKerner/Digenic.

[1]To whom correspondence may be addressed. Email: jean-laurent.casanova@rockefeller.edu or laurent.abel@inserm.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1920650117/-/DCSupplemental.

First published July 27, 2020.

involved a combination of genome-wide single-gene analysis with prior knowledge of a common variant from genome-wide association study (GWAS) data (19, 21). For genetically heterogeneous diseases, such as Alport syndrome, for which there are three known disease-causing genes, long-QT syndrome and Bardet-Biedl syndrome, each with more than a dozen disease-causing genes, the proven digenic combinations display various modes of dominance and involve the known disease-causing genes (22, 23). However, other genetic modifier genes may be hidden among genes with an unknown functional impact on disease, or even genes with no detectable main effect. Similarly, many heritable conditions masked in apparently sporadic cases, for which the genetic etiology remains unknown, may be due to digenic inheritance.

There is, therefore, a need for two-locus genome-wide methods for the detection of digenic inheritance in NGS data, in particular WES data, which are widely used by geneticists (24–26). Very few methods have been developed for detecting gene × gene interactions in the general context of rare variant association studies; all techniques to date are based on case-control designs (27–29). Here, we propose a case-only approach to specific searches for digenic inheritance. This design avoids the need for control recruitment and the potential associated bias. Furthermore, case-only approaches have been shown to be more powerful than classic case-control tests when common variants are tested for interaction, particularly in the context of GWAS (30–34). Our approach is based on the aggregation of rare variants within a gene as the unit of analysis, addressing the lack of power inherent to studies of rare variants. Assessment of interactions at gene level also greatly decreases the number of tests required relative to testing at allele level, potentially also decreasing the amount of computer time required.

Materials and Methods

The Variant Aggregation Model. A strategy commonly used for low-frequency variants from NGS data involves tests based on the aggregation of variants within a genomic region. Several types of tests are used for this purpose: burden tests, adaptive burden tests, variance-component tests, and combinations of these three classes (35). Here, we propose a method based on the classic collapsing of variants within the unit of a gene. This approach optimizes statistical power under a hypothesis of genetic homogeneity, while making it possible to assess actual gene × gene interactions with a number of tests corresponding to the number of possible two-way combinations of genes. In this study, the aggregation of variants within a gene is based on the methodology of a class of burden tests known as the "cohort allelic sums test" (36). Formally, for each gene j and a given subset of variants S_j observed within this gene, if n is the number of individuals studied, we consider the following vector$(g_{j1}, \ldots, g_{jn})$ denoted G_j. For each $i = 1, \ldots, n$, g_{ji} is then defined as follows:

$$g_{ji} = \begin{cases} 1 & if individual i carries at least one variant in subset S_j \\ 0 & otherwise \end{cases}.$$

The term "carries" depends here on the biological inheritance model. For example, in a dominant model, $g_{ji} = 1$ if individual i harbors at least one copy of at least one variant allele from the set of variants studied S_j within gene j. In addition, the choice of S_j may be based on different features at the variant level, such as the MAF or functional impact prediction, as described below.

Table 1. Contingency table of carriers of rare variants for a given pair of genes *k* and *j* for affected and unaffected individuals

	Gene *k*	
Gene *j*	Carriers	Noncarriers
Carriers	$i_{kj,11}$	$i_{kj,01}$
Noncarriers	$i_{kj,10}$	$i_{kj,00}$

$i = \{n, m\}$. When $i = n$, cells account for the number of affected individuals; when $i = m$, cells account for the number of unaffected individuals.

The Case-Control Design for Interaction. Using this notation, data for genes k and j in a case-control dataset, with a binary disease status D, can be summarized into two 2 × 2 contingency tables, one for affected individuals (cases, D = 1) and one for unaffected individuals (controls, D = 0), as in Table 1. Based on these tables, let $N_{kj} = (n_{kj,00}, n_{kj,10}, n_{kj,01}, n_{kj,11})$ be a vector of the observed numbers of carriers for gene k and gene j among cases, such that, for example, $n_{kj,11} = \sum_{i \text{ in cases}} (g_{ki} \times g_{ji})$. Similarly, we define $M_{kj} = (m_{kj,00}, m_{kj,10}, m_{kj,01}, m_{kj,11})$ as a vector of the observed numbers of carriers for gene k and gene j among controls. The odds ratios for cases and controls, respectively, for genes k and j are defined as follows:

$$OR^a_{kj} = \frac{n_{kj,11} \times n_{kj,00}}{n_{kj,10} \times n_{kj,01}}, OR^u_{kj} = \frac{m_{kj,11} \times m_{kj,00}}{m_{kj,10} \times m_{kj,01}}.$$

Classic statistical analyses of interaction are based on the comparison of $OR^a_{kj}$and OR^u_{kj}. More specifically, the following classic case-control logistic regression model is often used to test for interaction:

$$\text{logit}P(D = 1) = \beta_0 + \beta_k G_k + \beta_j G_j + \beta_I G_k \times G_j, \quad [1]$$

where it can be shown that the interaction coefficient $\beta_I$equals $\log\left(\frac{OR^a_{kj}}{OR^u_{kj}}\right)$. This model also takes main effects into account, by considering coefficient terms for each gene (β_k and β_j). In addition, specific covariates, such as principal components (PCs), can easily be introduced into the model. Including a matrix of covariates X and a vector C of coefficients, the full logistic regression model takes the following form:

$$\text{logit}P(D = 1) = \beta_0 + \beta_j G_j + \beta_k G_k + \beta_I G_j \times G_k + CX. \quad [2]$$

Subsequently, the null hypothesis of no interaction $\beta_I = 0$ can be tested in a likelihood ratio test (LRT) with one degree of freedom, in the presence or absence of main genetic effects and/or covariate effects.

The Case-Only Model. Interactions can also be assessed by focusing exclusively on cases, such that all of the information is provided by the 2 × 2 contingency table for affected individuals (Table 1). In this situation, the standard full logistic regression model to test for interaction between genes *Gk* and *Gj* is now written as

$$\text{logit}P(G_k = 1) = \gamma_0 + \gamma_I G_j + CX, \quad [3]$$

where γ_I is equal to $\log(OR^a_{kj})$, X is a matrix of covariates, and C a vector of coefficients. As before, an LRT with one degree of freedom can be used to test the null hypothesis$\gamma_I = 0$.

Under the assumption that vectors *Gk* and *Gj* are not correlated, implying, in particular, that variants of the two genes are not in linkage disequilibrium (LD), a deviation from 1 of OR^a_{kj} indicates interaction. In addition, if the disease is rare, OR^u_{kj} is close to 1, and, consequently, β_I is approximatelyγ_I. The advantages of this test over case-control tests have been extensively studied theoretically (30, 34), in particular the gain of power. This gain stands from the nature of the estimators of the interaction coefficients of both designs. These estimators depend either on the ratio $\frac{OR^a_{kj}}{OR^u_{kj}}$ for the case-control or only onOR^a_{kj} for the case-only test. The asymptotic variances of the estimators are the sum of the reciprocal counts of Table 1, either for both affected and unaffected subjects (case-control design) or for affected individuals only (case-only) (30). Hence, the variance of the estimator of the case-control interaction coefficient has a larger variance, leading to a less powerful test. In addition, the advantages include the absence of a need to recruit controls, which, in the context of more prevalent diseases, would also avoid the problem of the misclassification of individuals with the unaffected phenotype. The only known limitation of this test is that it assumes independence in the general population of the variants tested. In agreement with previous analyses of the case-only test in GWAS data or theoretical contexts, our type I error analyses using WES data revealed similar sources of violation of this assumption.

Samples. For the simulation study we worked on real exome data, using samples from the 1000 Genome project (1000G) populations, and a subset of our in-house exome database, the Human Genetics of Infectious Diseases (HGID) database. Six populations from the 1000G database were used: four European populations—the Iberian population in Spain (IBS, n = 107),

Table 2. Empirical type I errors at a nominal value of $\alpha = 10^{-3}$ for the case-only and case-control tests in the absence of population stratification

Design	Model				
	Pg_0*	Pg_2†	$Pg_2 + 3PC$‡	$Pg_2 + C_{25}$§	$Pg_2 + C_{35}$¶
Case-only (IBS + TSI)	*0.00147* [0.0009–0.00110]	*0.00121* [0.0009–0.00110]	*0.00133* [0.0009–0.00110]	0.00109 [0.0009–0.00113]	0.00108 [0.0009–0.00114]
Case-control (IBS + TSI + GBR + FIN)	*0.00128* [0.0009–0.00110]	*0.00128* [0.0009–0.00110]	*0.00130* [0.0009–0.00110]	0.00107 [0.0009–0.00113]	0.00103 [0.0009–0.00114]

Boundaries of the 95% confidence intervals are shown in brackets. Type I error values lying outside the 95% confidence interval's boundaries are in italic.
*All pairs of genes with >15% of carriers of variants with MAF <5%.
†Pairs of genes as Pg_0 but with genes apart by at least 2 Mb.
‡Pairs of genes as Pg_2 with adjustment on the first three PCs.
§Pairs of genes as Pg_2 with >25% of carriers of variants with MAF <10%.
¶Pairs of genes as Pg_2 with >35% of carriers of variants with MAF <15%.

Toscani in Italy (TSI, $n = 107$), British in England and Scotland (GBR, $n = 91$) and Finnish in Finland (FIN, $n = 99$) — and two Asian populations of Chinese origin —Southern Han Chinese (CHS, $n = 105$) and Chinese Dai in Xishuangbanna, China (CDX, $n = 93$). From the HGID database, which includes data for > 4,000 individuals of various ethnic origins, including patients suffering from severe infectious diseases, we selected 1,331 individuals of European origin, as defined by PC analysis (PCA) on WES data, as previously described (37). Based on a refined PCA on these 1,331 individuals, together with the 404 European 1000G individuals, we identified three distinct subpopulations (*SI Appendix*, Fig. S1): "Northern Europeans" (N), "Middle Europeans" (M), and "Southern Europeans" (S). For the real data analysis we used the craniosynostosis WES dataset reported in ref. 19 (*SI Appendix*).

Results

Simulation Study. We first investigated the properties of our case-only test through simulations on real exome data from the 1000G populations and a subset of our in-house exome HGID database. We performed analyses under the null hypothesis of no digenic interactions and no main genetic effects, for which we assessed type I errors. We also worked under the alternative hypothesis of a digenic interaction, for which we assessed statistical power under genetic effects of different magnitudes. In these analyses, we compared the case-only approach to the corresponding case-control approach, for various population stratification (PS) scenarios.

Type I Error Analyses.

Case-only design. We first performed our case-only test on an ethnically homogeneous population based on the 214 IBS + TSI 1000G Southern European samples. After the application of quality-control filters (*SI Appendix*), 1,588 genes for which at least 15% of individuals carried rare variants were included in the analysis, resulting in 1,260,067 interaction tests. We observed a moderate inflation of type I error to 0.00147 for $\alpha = 10^{-3}$ (Table 2) and 0.0535 for $\alpha = 0.05$ (*SI Appendix*, Table S1). We therefore assessed the possible effect of LD (38), by restricting our analysis to pairs of genes physically separated by a minimal distance δ (measured in megabases [Mb]). Empirical type I errors decreased with increases in δ from 0.1 to 2 Mb (*SI Appendix*, Table S2), and a type I error of 0.00121 was obtained at a nominal value α of 10^{-3} when $\delta = 2$ Mb (Table 2). The distributions of P values for tests of pairs of genes with $\delta < 2$ Mb was strikingly inflated (*SI Appendix*, Fig. S2). Type I errors did not improve for $\delta > 2$ Mb (*SI Appendix*, Table S2). Globally, these results show that LD accounted for the lowest P values in the case-only test. The refined investigation of statistically significant pairs of genes located close together (680 with $P < 0.05$ among 4,082 pairs with $\delta < 2$ Mb in the IBS + TSI cohort) would require a case-control design. Even so, after simple LD correction based on removing the pairs of genes with $\delta < 2$ Mb, type I errors remained slightly above the corresponding upper limit of the confidence interval. No further improvement was obtained by adjusting our tests for the first three PCs, consistent with the fact that the IBS and the TSI populations are very close.

Case-control design. We conducted an analogous investigation with a case-control design on an enlarged European population consisting of the 404 IBS + TSI + GBR + FIN 1000G samples, in order to have ~200 cases and ~200 controls. We first applied it in a population-balanced scenario (Table 2), in which 1,563 genes were retained after the application of quality-control filters (*SI Appendix*). No inflation due to LD (as expected in a case-control design) or PS (as expected for a balanced scenario) was observed. Nevertheless, a slight inflation of type I error similar to that observed for the case-only test at $\alpha = 10^{-3}$ (Table 2) and $\alpha = 0.05$ (*SI Appendix*, Table S1) was found. We hypothesized that this inflation might be at least partly due to the small sample sizes in the contingency cells of Table 1. We tested this

Table 3. Empirical type I errors at a nominal value of $\alpha = 10^{-3}$ for the case-only and case-control tests in the presence of population stratification

Design	PC adjustment	
	No adjustment	3PC
Case-only* (IBS + CHS)	*0.01432* [0.0009–0.00113]	*0.00135* [0.0009–0.00113]
Case-control, balanced (IBS + TSI + CHS + CDX)	*0.00132* [0.0009–0.00113]	*0.00136* [0.0009–0.00113]
Case-control, unbalanced (IBS + TSI + CHS + CDX)	*0.00257* [0.0009–0.00113]	*0.00126* [0.0009–0.00113]

Boundaries of the 95% confidence intervals are shown in brackets. Type I error values lying outside the 95% confidence interval's boundaries are in italic.
*Using pairs of genes with genes apart by at least 2 Mb.

Table 4. Description of the schemes used in *Power Analyses*

	Schemes			
	A			2GR
	Genome-wide	2G	2GS	Two genes
Genes tested	All genes	Both common and nonstratified by population	Both common and stratified by population	One common and one rare nonstratified by population
OR_j	{1,2}	{1,2}	{1,2}	{1,2}
OR_k	{1,2}	{1,2}	{1,2}	{1,2}
OR_I	{1, . . . , 5}	{1, . . . , 5}	{1, . . . , 5}	{1, . . . , 10}

$OR_j$and OR_k are the odds ratios for the main effect of the first and the second gene of each pair, respectively.OR_I is the odds ratio for the interaction term of Eq. **1**.

hypothesis by repeating the analyses for both the case-only and the case-control tests with more common variants and a larger number of carriers at the gene level (i.e., variants with a MAF <10% and genes with carriage rates of at least 25%, and variants with a MAF <15% and genes with carriage rates of at least 35%; Table 2). The type I error was clearly lower and improved as the frequency of variants increased. For both tests, empirical type I errors were within the boundaries of the confidence interval for $\alpha = 10^{-3}$

Sample size investigation. We investigated the impact of contingency cell sample sizes and the number of tests on the case-only approach, by extending the previous scenario to two new settings with less stringent MAF thresholds. We first conducted a case-only test for all genes carried by at least 5%, rather than 15%, of individuals in the IBS + TSI population. This strategy increased the number of genes retained to 5,563, and, after the removal of genes in LD, we tested a total of 15,465,141 pairs of genes and generated the QQ-plot for *SI Appendix*, Fig. S3. The type I error was moderately inflated (0.057) for $\alpha = 0.05$, but there was a slightly conservative trend for lower α values (from 10^{-3} to 10^{-6}), which remained either close to the lower limit or within the corresponding confidence interval, particularly for the lowest nominal values (*SI Appendix*, Table S3). We also simulated the data for one gene considered "rare" (at least 1% carriers, total of 11,470 genes) and another considered "common" (at least 15% carriers, total of 1,588 genes). Under this scenario, 16,951,106 pairs of genes were tested, and the QQ-plot for *SI Appendix*, Fig. S4 was generated. The results were similar to those for the situation previously analyzed (5% vs. 5%), with a conservative trend for low nominal type I error values (*SI Appendix*, Table S3). Finally, we investigated the properties of the test for even lower α values (down to 10^{-7}) with a specific study design (*SI Appendix*). We found that the case-only test remained slightly conservative but within the confidence interval boundaries for $\alpha = 10^{-6}$ and $\alpha = 10^{-7}$ (*SI Appendix*, Table S4). Overall, these results suggest that the case-only test is robust for investigating a large range of carrier frequencies at the genome-wide level and for low nominal αvalues, provided that LD is considered.

Population stratification. We then investigated the effect of PS, again focusing only on genes for which at least 15% of the individuals in the study population were carriers and which were

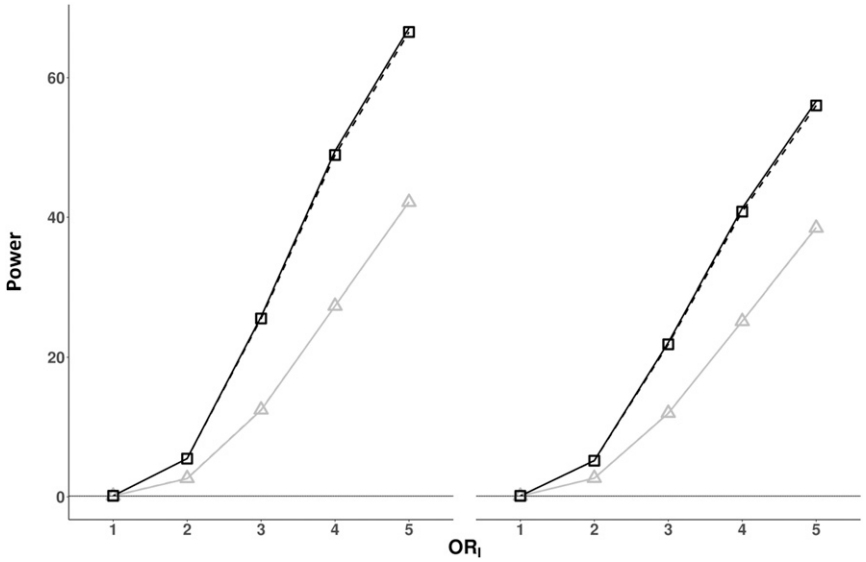

Fig. 1. Power of the case-only and case-control tests for the analysis of all pairs of genes (scheme A). Power values are presented as a percentage for a type I error of 10^{-3}, as a function of the odds ratio for interaction (OR_I), for the case-only (dark curves) and case-control (light curves) tests with (dotted lines with symbols) or without (solid lines without symbols) adjustment for the first three PCs. (*Left*) Obtained when no main gene effects are present. (*Right*) Results with a main effect of the second gene (OR = 2). Note that the results with and without adjustment are very similar and the strong superimposition of the corresponding curves.

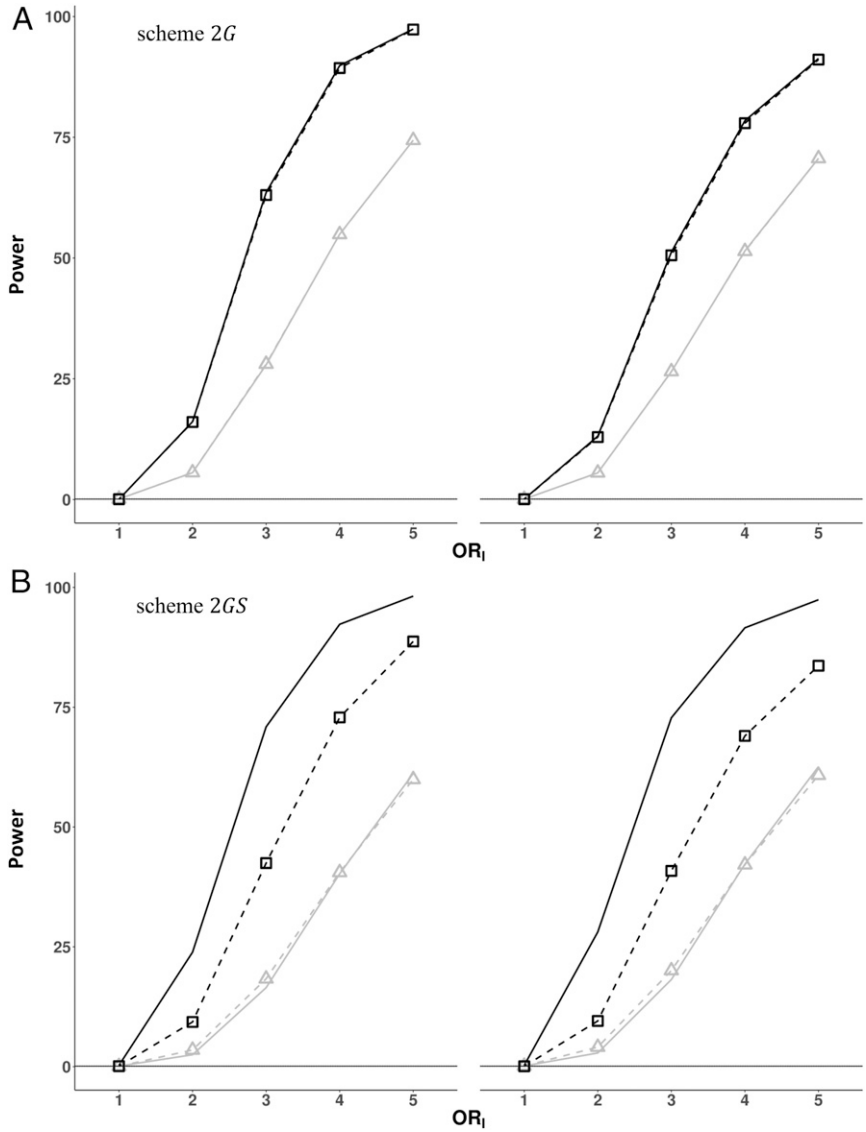

Fig. 2. Power of the case-only and case-control tests for the analysis of two specific pairs of genes in the absence (scheme 2G) or presence (scheme 2GS) of population stratification. Power values are presented as in Fig. 1. Results are shown for the analysis of (*A*) the two nonstratified genes *PKHD1L1* and *AHNAK* (scheme 2G, *Top*) and (*B*) the two stratified genes *ARPP21* and *MACF1* (scheme 2GS, *Bottom*). (*Left*) Obtained when no main gene effects are present. (*Right*) Results with a main effect (OR = 2) of the second gene, that is, *AHNAK* and *MACF1*, respectively.

separated by at least 2 Mb. For the case-only test, we used the 212 IBS + CHS samples, and we assessed 1,248 genes, in 776,879 tests (*SI Appendix*). Type I errors were highly inflated (0.0143 for $\alpha = 10^{-3}$ and 0.1264 for $\alpha = 0.05$) (Table 3 and *SI Appendix*, Table S5). The application of PS correction (adjustment for the first three PCs) brought empirical type I errors back down to levels very similar to those previously observed (0.0013 for $\alpha = 10^{-3}$ and 0.0550 for $\alpha = 0.05$). For the case-control test, we used the 412 IBS + TSI + CHS + CDX samples under an unbalanced population scenario, with 1,173 genes (*SI Appendix*). Inflated type I errors were also observed, and adjustment for PCs resulted in values similar to those for a situation without PS (Table 3 and *SI Appendix*, Table S5). Thus, provided that the search space was limited to pairs of genes far enough apart to avoid LD and adjustment for PCs was applied when required, our case-only test yielded reasonable type I error rates similar to those for the analogous case-control approach.

Power Analyses.

Average power scenario. Power studies were conducted on an enlarged European population consisting of 1,735 individuals from the four European 1000G populations (IBS, TSI, GBR, and FIN) and 1,331 individuals from the in-house HGID database (*SI Appendix*). We first estimated an "average" power by testing all possible pairs of genes (scheme A, Table 4), each with at least 15% carriers and separated by at least 2 Mb. In total, 370,530 tests were performed in 10 replicates (*SI Appendix*). Fig. 1 displays the results obtained for scenarios including one or no main genetic effect, corresponding to the most pertinent situations in which to search for a gene × gene interaction. The resulting curves adjusted or not with the first three PCs were superimposed, indicating that this analysis, in a European population, was not affected by PS. In all situations, power was always greater for the case-only test than for the case-control test. For example, a power of 65% at $\alpha = 10^{-3}$ was obtained when the odds ratio for interaction (OR_I) = 5 and no main effects were

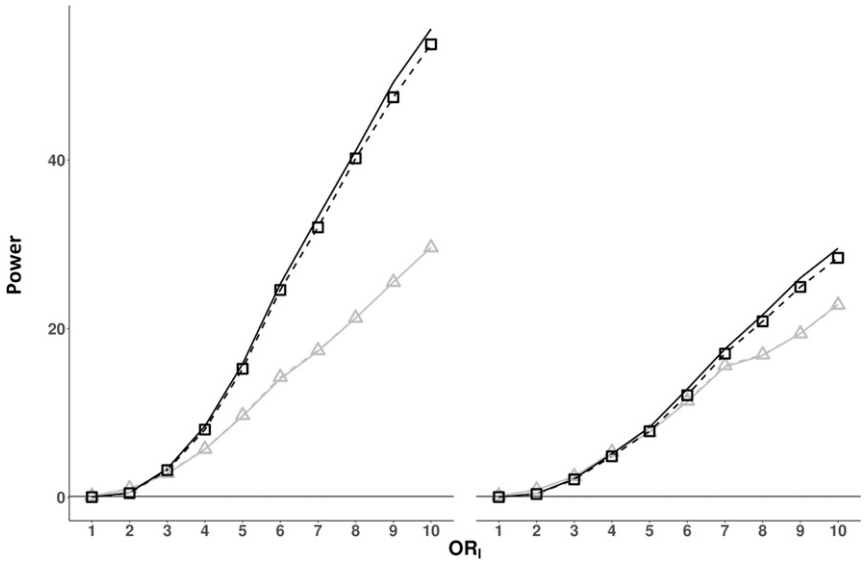

Fig. 3. Power of the case-only and case-control tests for analyzing a pair of genes with different proportions of variant carriers (scheme 2GR). Power curves are presented as in Fig. 1. Results are shown for the analysis of one "common" (*AHNAK*) and one "rare" gene (*MPC1*) (scheme 2GR). (*Left*) Obtained when no main effects are present. (*Right*) Results with a main effect (OR = 2) of the second gene, that is, MPC1.

considered, whereas a power of only 40% was obtained for the corresponding case-control test in the same conditions. Similar trends were observed when one main effect was present (Fig. 1 and *SI Appendix*, Fig. S5). We also assessed power in the case of genes with smaller numbers of carriers (>1% for one gene and >15% for the other) and for lower alpha values (10^{-5} and 10^{-6}), in the context of no main effects and for a larger range of OR_I (*SI Appendix*, Fig. S6). We found that the absolute difference in power between the two tests favored the case-only test and increased when alpha decreased or OR_I increased.

Two-gene power scenarios. We then focused on two specific pairs of genes, without (*AHNAK* and *PKHD1L1*; scheme 2G, Table 4) and with (*ARPP21* and *MACF1*; scheme 2GS, Table 4) PS (*SI Appendix*). In the analysis of scheme 2G, the case-only test performed better, overall, in terms of power (Fig. 2 and *SI Appendix*, Fig. S7, *Top*). In the absence of main effects, with $OR_I = 3$ and $\alpha = 10^{-3}$, a power value of 62% was obtained for the case-only test, vs. only 27% for the case-control test. For scheme 2GS, the power curves adjusted or not with the first three PCs of the case-only tests were clearly different, indicating an effect of PS (Fig. 2 and *SI Appendix*, Fig. S7, *Bottom*). We therefore used only the adjusted (for the first three PCs) case-only test for comparison. As expected, the case-control test was not affected by PS (0.0009 for $\alpha = 10^{-3}$) and had type I error values similar to those for the adjusted case-only test (0.0011 for$\alpha = 10^{-3}$). The adjusted case-only test clearly outperformed the case-control test, by reaching a power of 90% when $OR_I = 5$ without main effects, for example, whereas the corresponding power for the case-control test was only 60%. Finally, we also considered another specific pair of genes, including one "common" (26% carriers) and one "rare" (5% carriers) gene (scheme 2GR, Table 4). The case-only test was again more powerful than the corresponding case-control test (Fig. 3 and *SI Appendix*, Fig. S8), particularly in the absence of main effects, giving an absolute difference in power of almost 30% when $OR_I = 10$. Situations with a lower cumulative frequency of rare variants and a stronger OR might fit a Mendelian-like disorder hypothesis better and are of particular interest concerning the application of this approach to real data presented below.

Theoretical power of the case-only test. We used Eqs. **1** and **3** to derive theoretical power curves for the case-only test according to different parameters and for three nominal α values of 10^{-4}, 10^{-6}, and 10^{-8}. We first assessed power as a function of the genetic interaction effect, OR_I, for two genes each having a rate of 5% carriage of the variant and various sample sizes. For example, an $OR_I \sim 16$ gave a power of 80% at $\alpha = 10^{-4}$ with 200 cases or at $\alpha = 10^{-8}$ with 400 cases (Fig. 4*A*). Furthermore, as suggested by our previous results, for a fixed sample size of 5,000 controls, we show that, in the same conditions, the case-control test is less powerful than the case-only test (Fig. 4*B*). Results for a fixed OR_I of 10, for one gene with 5% carriers and a second gene with 1 to 20%, are shown in *SI Appendix*, Fig. S9. As expected, power strongly increased with the proportion of carriers, reaching 50% for a proportion of 15% with 300 cases, or for a proportion of 8% with 400 cases. Finally, we investigated the sample size required to reach a power of 80% as a function of OR_I for genes with 15% or 5% carriers (*SI Appendix*, Fig. S10). We observed that ~200 cases would be required to reach a power of 80% at $\alpha = 10^{-8}$ and $OR_I = 10$ in the case of genes with 15% carriers, or less than 400 cases at $\alpha = 10^{-6}$ and $OR_I = 15$, in the case of genes with 5% carriers.

Real Data Analysis: Craniosynostosis.

Background. We first applied our test to the dataset that led to the discovery of the first case of digenic inheritance of nonsyndromic midline craniosynostosis (MIM: 617439) (19). The original study showed a strong enrichment in rare heterozygous *SMAD6* mutations predicted to be damaging among cases (13 carriers among the 191 probands). Incomplete penetrance was observed in relatives of the carriers. The role of the common variant *rs1884302* (MAF = 0.33 in European populations), located close to the *BMP2* gene and previously associated with craniosynostosis through GWAS (21), was therefore investigated, and this variant was found to account for almost all of the observed phenotypic variation. Eleven of the 13 *SMAD6* probands were also carriers of *rs1884302*, whereas none of the healthy *SMAD6* carriers carried this variant. We used these data to determine whether our unbiased case-only test could detect this digenic association in the context of a genome-wide search (i.e., without prior knowledge of the role of the *SMAD6* and *BMP2* variants).

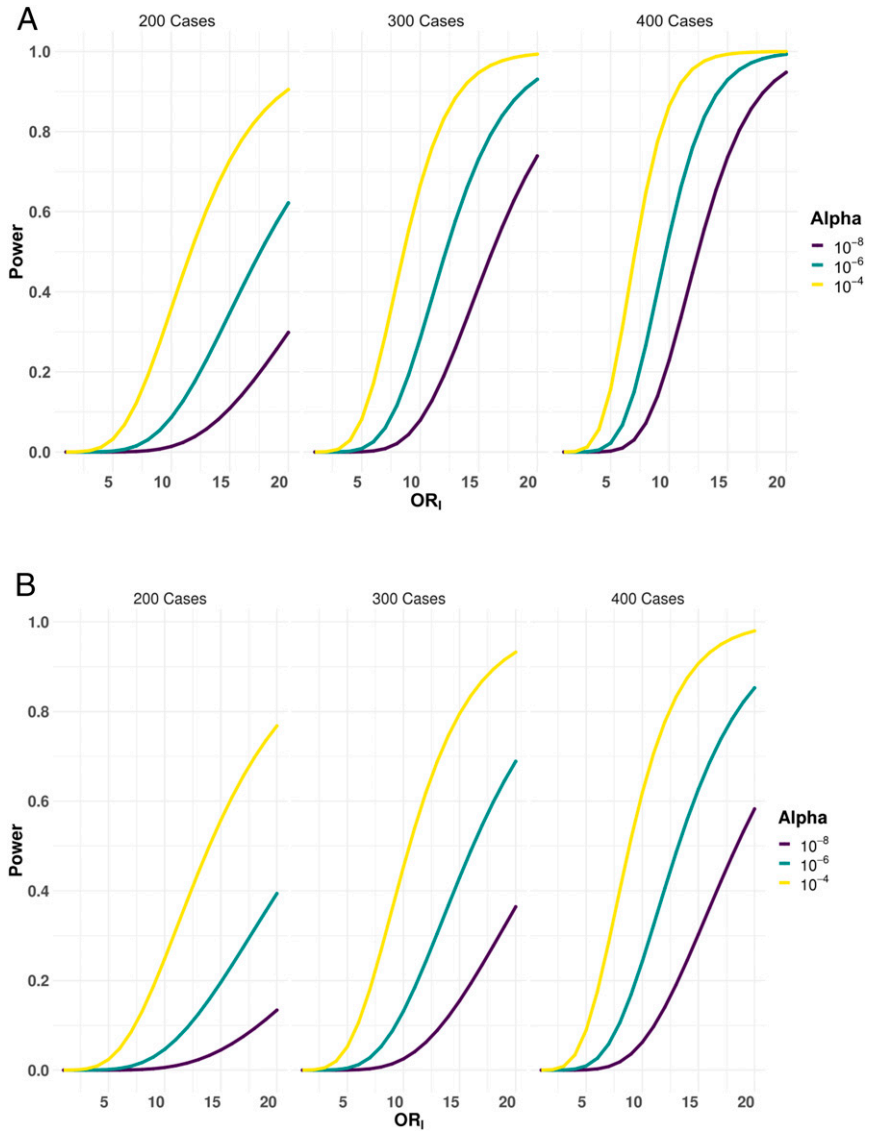

Fig. 4. Theoretical power according to sample size and nominal type I error for the (*A*) case-only and the (*B*) case-control tests. Power values were calculated for (*A*) the case-only test and (*B*) the case-control test, for type I errors of 10^{-4} (yellow curve), 10^{-6} (blue), and 10^{-8} (purple), and odds ratios of interactions ranging from 1 to 20. From left to right, panels represent scenarios with 200, 300, and 400 cases, respectively. In *B*, a fixed number of 5,000 controls was used for all scenarios. Prevalence was set to 10^{-4} and the percentage of carriers for both genes to 5%.

Genome-wide search. In total, 285,216 tests (83 genes and 8,102 variants) were conducted on the WES data for 191 patients after the application of quality control and other filters to the variants and genes (*SI Appendix*). The resulting QQ-plot shows no deviation from the expected distribution, with only one significant result over the expected *P* value line (Fig. 5). This result ($P = 1.58 \times 10^{-6}$, OR = 30.95) corresponds to the digenic combination of *SMAD6* and *rs1884302* and is one order of magnitude higher than the second result ($P = 1.04 \times 10^{-5}$), which is close to the expected line. The 2×2 contingency table for the top result is shown in *SI Appendix*, Table S6, and corresponds to the distribution found in the original paper (19). Thus, the two-locus genome-wide analysis focusing on genes harboring rare variants together with the potential contribution of a common modifier variant was able to detect the previously reported digenic inheritance for craniosynostosis (19). This analysis provides proof of concept that our statistical test can detect digenic inheritance without the need for biological assumptions concerning the disease studied, even when the disease is very rare.

Discussion

There is increasing evidence to suggest that digenic inheritance plays an important role in the genetic architecture of many conditions. The three previously reported approaches searching for gene × gene interactions in the general context of rare variant association studies are based on case-control designs (27–29). Moreover, these tests were assessed in limited simulation studies involving short genomic sequences of fewer than 500 variants ($n = 1$) or only 20 variants ($n = 2$) and were not based on WES-based simulated data. None was reported to have detected two genetic lesions at the genome-wide level. Indeed, all previously successful digenic inheritance studies relied on candidate gene approaches to overcome the lack of appropriate statistical resources to search for digenic inheritance at the genome-wide level (17). Digenic inheritance studies and statistical interaction approaches have thus been following separate paths. We show here, through both extensive simulation studies on real WES datasets and application to the example of craniosynostosis, that our

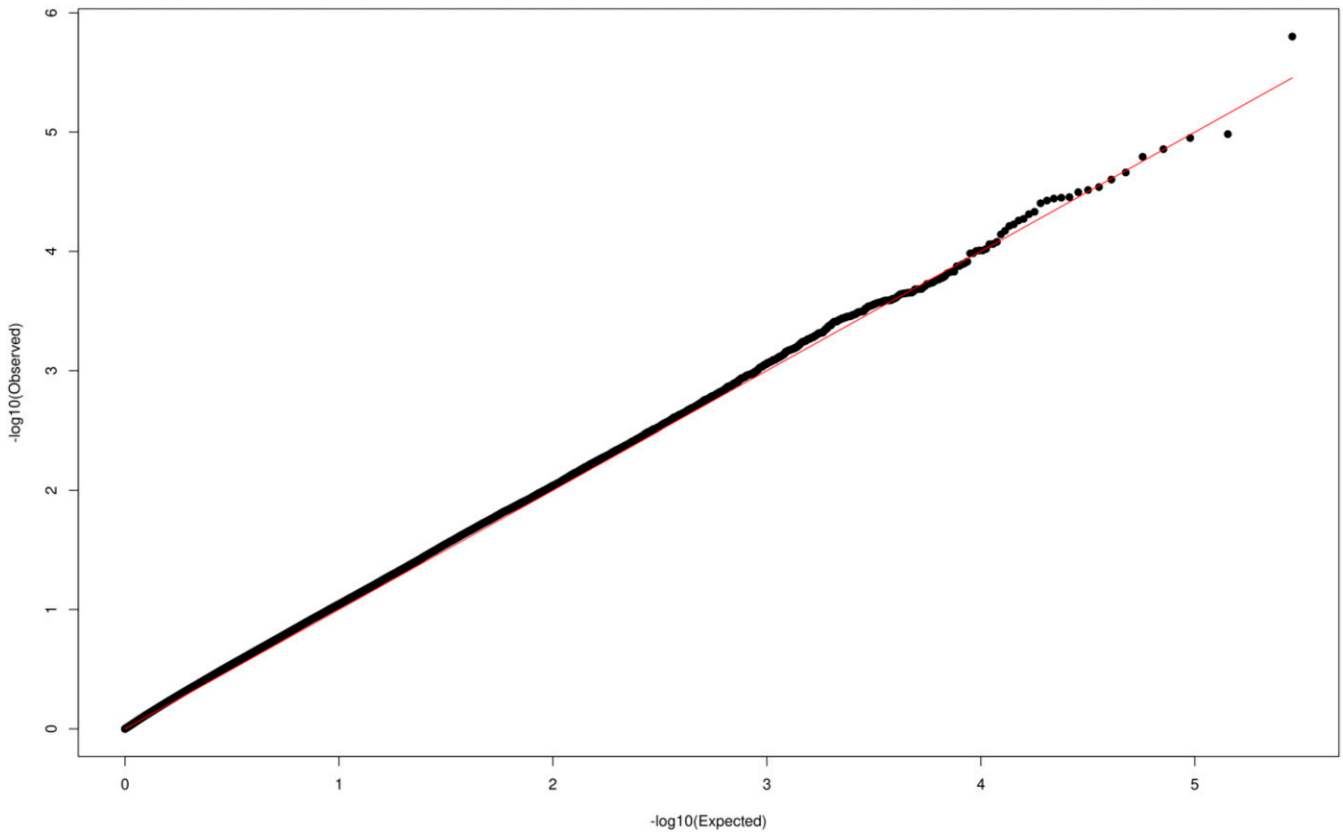

Fig. 5. QQ-plot for the genome-wide case-only test conducted on the 191 craniosynostosis probands. QQ-plot for a genome-wide analysis under a dominant mode of inheritance, adjusted for the first three PCs, and considering pairs of genes and variants at least 2 Mb apart with >5% carriers of rare variants a world-wide frequency >10% for the variant (n = 285,216 pairs).

method is robust and powerful for the identification of digenic lesions at the genome-wide level. Our unbiased genetic confirmation of the reported digenic lesions in the craniosynostosis dataset composed only of exome data from cases, a common feature of real datasets for rare disorders, justifies the choice of a case-only test based on the aggregation of rare variants. Further strong support for this approach is provided by the higher overall power for the case-only approach than for the corresponding case-control test, as shown here, for the same cases. Moreover, our theoretical analysis of the case-only test showed that sample sizes of between 200 and 400 cases were sufficient to achieve substantial power in the case of a digenic effect leading to an $OR_I < 10$ when the proportion of carriers is >15% for the two genes. Even for a lower proportion of carriers (e.g., > 5% for both genes), sample sizes of between 200 and 400 cases result in a powerful case-only test for an $OR_I > 10$, corresponding, for example, to the studied example of craniosynostosis ($OR_I \sim 30$).

The proposed methodology is simple to apply and flexible. It requires only the definition of a set of variants for testing, with filters based on features including MAF, variant annotations, and genetic models, defined before the analysis. It can, of course, be used at the gene level for the two loci studied. It can also directly assess the role of common variants as potential modifiers of a known monogenic defect. This assessment is achieved by simply replacing the gene by the variant as the unit of analysis, as illustrated in the craniosynostosis example. This technique proved useful, but it can increase the number of tests, leading to classical multiple testing issues. However, our analyses, demonstrating the robustness of our method for low nominal type I error values, are reassuring for the use of a low P value significance threshold, as determined by the widely used Bonferroni correction method. Our result also provides proof of concept that incomplete penetrance in disorders considered to be monogenic can be explained by a unique digenic combination. The frequency of carriers considered in our simulation studies may appear to be too high, but two important points must be considered when studying a rare disorder. First, these thresholds correspond to a cumulative frequency of the variants potentially contributing to the disease. The frequency of each individual allele may be much lower. Second, enrichment in the true disease-causing alleles would be expected in patients. For example, in the craniosynostosis dataset, the cumulative frequency of carriers of rare damaging *SMAD6* mutations is 6.8% (13 of 191), whereas the maximum cumulative frequency of carriers of these variants in gnomAD is 0.01%. The proposed case-only test thus already appears to be a powerful and timely tool for detecting digenic inheritance based on NGS data at the genome-wide level in disorders that are not explained or only partly explained by a monogenic lesion.

Data Availability. R scripts and additional information to perform digenic analyses are available at GitHub, https://github.com/GaspardKerner/Digenic.

ACKNOWLEDGMENTS. We thank the patients and their families, whose cooperation was essential for collection of the data used in this study. We thank all members of the Laboratory of Human Genetics of Infectious Diseases for helpful discussions and Céline Desvallées, Tatiana Kochetkov, Dominick Papandrea, Cécile Patissier, Mark Woollett, Dana Liu, and Yelena Nemirovskaya for their assistance. The datasets used for analyses described in this manuscript were partially obtained from dbGaP at https://www.ncbi.nlm.nih.gov/gap/ through dbGaP accession number phs000744.

The Laboratory of Human Genetics of Infectious Diseases is supported in part by institutional grants from Institut National de la Santé et de la Recherche Médicale, Paris Descartes University, St. Giles Foundation, The Rockefeller University Center for Clinical and Translational Science grant from the National Center for Research Resources and the National Center for Advancing Translational Sciences of the NIH (8UL1TR001866), the TBPATHGEN project (ANR-14-CE14-0007-01), grants from the French National Research Agency (ANR) under the "Investments for the future" program (ANR-10-IAHU-01), GENMSMD (ANR-16-CE17-0005-01 to J.B.) and MYCOPARADOX (ANR-16-CE12-0023-01), the Yale Center for Mendelian Genomics (UM1HG006504), funded by the National Human Genome Research Institute, and the Genome Sequencing Program Coordinating Center (U24 HG008956). Gaspard Kerner was supported by Institut Imagine (Imagine Thesis Award 2019-2020).

1. G. Andreoletti *et al*., Exome analysis of rare and common variants within the NOD signaling pathway. *Sci. Rep.* **7**, 46454 (2017).
2. M. J. Bamshad *et al*., Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
3. C. T. Johansen *et al*., Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684–687 (2010).
4. J. C. Cohen, H. H. Hobbs, Genetics. Simple genetics for a complex disease. *Science* **340**, 689–690 (2013).
5. S. Boisson-Dupuis *et al*., Tuberculosis and impaired IL-23–dependent IFN-γ immunity in humans homozygous for a common TYK2 missense variant. *Sci. Immunol.* **3**, eaau8714 (2018).
6. P. K. Brastianos *et al*., Exome sequencing identifies BRAF mutations in papillary craniopharyngiomas. *Nat. Genet.* **46**, 161–165 (2014).
7. M. Choi *et al*., Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19096–19101 (2009).
8. J. Kaiser, Human genetics. Affordable "exomes" fill gaps in a catalog of rare diseases. *Science* **330**, 903 (2010).
9. I. Meyts *et al*., Exome and genome sequencing for inborn errors of immunity. *J. Allergy Clin. Immunol.* **138**, 957–969 (2016).
10. S. B. Ng *et al*., Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
11. D. B. Zastrow *et al*.; Undiagnosed Diseases Network, Exome sequencing identifies de novo pathogenic variants in *FBN1* and *TRPS1* in a patient with a complex connective tissue phenotype. *Cold Spring Harb. Mol. Case Stud.* **3**, a001388 (2017).
12. V. G. Sankaran *et al*., Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *J. Clin. Invest.* **122**, 2439–2443 (2012).
13. C. J. Bell *et al*., Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra4 (2011).
14. D. N. Cooper, M. Krawczak, C. Polychronakos, C. Tyler-Smith, H. Kehrer-Sawatzki, Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
15. C. Deltas, Digenic inheritance and genetic modifiers. *Clin. Genet.* **93**, 429–438 (2018).
16. A. A. Schäffer, Digenic inheritance in medical genetics. *J. Med. Genet.* **50**, 641–652 (2013).
17. A. M. Gazzo *et al*., DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Res.* **44**, D900–D907 (2016).
18. M. J. Emond *et al*.; National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project; Lung GO, Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Nat. Genet.* **44**, 886–889 (2012).
19. A. T. Timberlake *et al*., Two locus inheritance of non-syndromic midline craniosynostosis via rare *SMAD6* and common *BMP2* alleles. *eLife* **5**, e20125 (2016).
20. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Online Mendelian Inheritance in Man. https://omim.org/. Accessed 9 August 2019.
21. C. M. Justice *et al*., A genome-wide association study identifies susceptibility loci for nonsyndromic sagittal craniosynostosis near BMP2 and within BBS9. *Nat. Genet.* **44**, 1360–1364 (2012).
22. M. A. Mencarelli *et al*., Evidence of digenic inheritance in Alport syndrome. *J. Med. Genet.* **52**, 163–174 (2015).
23. W. Peter *et al*., Compound mutations: A common cause of severe long-QT syndrome. *Circulation* **109**, 1834–1841 (2004).
24. M. Muona *et al*., A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nat. Genet.* **47**, 39–46 (2015).
25. S. H. Lelieveld *et al*., Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
26. J. F. McRae *et al*.; Deciphering Developmental Disorders Study, Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
27. R. Fan, S.-H. Lo, A robust model-free approach for rare variants association studies incorporating gene-gene and gene-environmental interactions. *PLoS One* **8**, e83057 (2013).
28. M. Kwon, S. Leem, J. Yoon, T. Park, GxGrare: Gene-gene interaction analysis method for rare variants from high-throughput sequencing data. *BMC Syst. Biol.* **12**, 19 (2018).
29. J. Zhao, Y. Zhu, M. Xiong, Genome-wide gene-gene interaction analysis for next-generation sequencing. *Eur. J. Hum. Genet.* **24**, 421–428 (2016).
30. P. S. Albert, D. Ratnasinghe, J. Tangrea, S. Wacholder, Limitations of the case-only design for identifying gene-environment interactions. *Am. J. Epidemiol.* **154**, 687–693 (2001).
31. W. J. Gauderman, Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.* **155**, 478–484 (2002).
32. P. Kraft, Y.-C. Yen, D. O. Stram, J. Morrison, W. J. Gauderman, Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–119 (2007).
33. B. L. Pierce, H. Ahsan, Case-only genome-wide interaction study of disease risk, prognosis and treatment. *Genet. Epidemiol.* **34**, 7–15 (2010).
34. Q. Yang, M. J. Khoury, F. Sun, W. D. Flanders, Case-only design to measure gene-gene interaction. *Epidemiology* **10**, 167–170 (1999).
35. S. Lee, G. R. Abecasis, M. Boehnke, X. Lin, Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
36. S. Morgenthaler, W. G. Thilly, A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* **615**, 28–56, 10.1016/j.mrfmmm.2006.09.003 (2007).
37. A. Belkadi *et al*., Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proc. Natl. Acad. Sci. U.S.A.*, https://doi.org/10.1073/pnas.1606460113 (2016).
38. P. Yadav, S. Freitag-Wolf, W. Lieb, M. Krawczak, The role of linkage disequilibrium in case-only studies of gene-environment interactions. *Hum. Genet.* **134**, 89–96 (2015).