Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage

Aziz Belkadi[a,b,1], Vincent Pedergnana[a,b,c,1], Aurélie Cobat[a,b], Yuval Itan[d], Quentin B. Vincent[a,b], Avinash Abhyankar[e], Lei Shang[d], Jamila El Baghdadi[f], Aziz Bousfiha[g], the Exome/Array Consortium[2], Alexandre Alcais[a,b], Bertrand Boisson[a,b,d], Jean-Laurent Casanova[a,b,d,h,i,3], and Laurent Abel[a,b,d,3]

[a]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, 75015 Paris, France; [b]Imagine Institute, Paris Descartes University, 75015 Paris, France; [c]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom; [d]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York, NY 10065; [e]New York Genome Center, New York, NY 10013; [f]Genetics Unit, Military Hospital Mohammed V, Hay Riad, 10100 Rabat, Morocco; [g]Clinical Immunology Unit of the Pediatric Infectious Disease Service, Ibn Roshd Hospital, Hassan II University, Casablanca 20460, Morocco; [h]Howard Hughes Medical Institute, New York, NY 10065; and [i]Pediatric Haematology–Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France

Contributed by Jean-Laurent Casanova, April 27, 2016 (sent for review March 3, 2016; reviewed by Luis B. Barreiro and Jacques Fellay)

Principal component analysis (PCA), homozygosity rate estimations, and linkage studies in humans are classically conducted through genome-wide single-nucleotide variant arrays (GWSA). We compared whole-exome sequencing (WES) and GWSA for this purpose. We analyzed 110 subjects originating from different regions of the world, including North Africa and the Middle East, which are poorly covered by public databases and have high consanguinity rates. We tested and applied a number of quality control (QC) filters. Compared with GWSA, we found that WES provided an accurate prediction of population substructure using variants with a minor allele frequency > 2% (correlation = 0.89 with the PCA coordinates obtained by GWSA). WES also yielded highly reliable estimates of homozygosity rates using runs of homozygosity with a 1,000-kb window (correlation = 0.94 with the estimates provided by GWSA). Finally, homozygosity mapping analyses in 15 families including a single offspring with high homozygosity rates showed that WES provided 51% less genome-wide linkage information than GWSA overall but 97% more information for the coding regions. At the genome-wide scale, 76.3% of linked regions were found by both GWSA and WES, 17.7% were found by GWSA only, and 6.0% were found by WES only. For coding regions, the corresponding percentages were 83.5%, 7.4%, and 9.1%, respectively. With appropriate QC filters, WES can be used for PCA and adjustment for population substructure, estimating homozygosity rates in individuals, and powerful linkage analyses, particularly in coding regions.

exome sequencing | genotyping array | population structure | homozygosity mapping | linkage analysis

Whole-exome sequencing (WES) has become the leading strategy for uncovering germ-line exome variants in humans. A number of gene- and variant-level methods have been proposed for the analysis of WES data to select candidate variants in rare Mendelian disorders and more common traits (1–13). These analyses benefit from the use of additional information, such as familial linkage, homozygosity rate, and ethnic background, which are commonly used in the study of inherited diseases (14–17). Genome-wide single-nucleotide variant array (GWSAs) are the gold standard method for linkage analysis, because they provide maximal linkage information for the whole genome (18). GWSAs are also classically used to estimate homozygosity rate in patients, confirming or sometimes, revealing parental consanguinity through the inbreeding coefficient parameter F in particular (19, 20). Population stratification can be an issue in the analysis of population-based genetic data, including WES, particularly for association studies (21–24). Population structures have been widely determined by GWSA (25, 26) in European (27), African (28, 29), Asian (30), Jewish (31), Mexican (32), and other populations (33). These analyses are mostly based on principal component analysis (PCA) (34), which can also be used to confirm or reveal the ethnicity of an individual patient (or his or her parents).

Unlike WES, which provides thorough coverage for less than 2% of the human genome for both rare and common variants, GWSAs cover the whole genome for common variants but only patchily, with a mean interval between variants of about 2–4 kb. Obtaining both WES and GWSA data in patients, kindreds, or populations is DNA-, resource-, and time-consuming. Two studies comparing WES and GWSA in linkage analyses based on real data from three families (35) or both simulated and real data from two families (36) showed that the two sets of genetic data defined linkage peaks (35) and excluded genomic regions (36) in a consistent manner. A recent study estimating homozygosity rates with both GWSA and WES data in patients born to consanguineous families provided recommendations for the detection of homozygous regions by WES (37). Finally, a method for estimating individual ancestry from a PCA map generated from data for a reference set of individuals also showed added value for a combination of single-nucleotide variant (SNV) data from exome chips or targeted sequencing with genotyping and imputed data for accurate ancestry estimation, particularly for European populations (38). We performed both GWSA and WES on 110 subjects originating from various regions of the

Significance

We compared the information provided by whole-exome sequencing (WES) and genome-wide single-nucleotide variant arrays in terms of principal component analysis, homozygosity rate estimation, and linkage analysis using 110 subjects originating from different regions of the world. WES provided an accurate prediction of population substructure using high-quality variants with a minor allele frequency > 2% and reliable estimation of homozygosity rates using runs of homozygosity. Finally, homozygosity mapping in 15 consanguineous families showed that WES led to powerful linkage analyses, particularly in coding regions. Overall, our study shows that WES could be used for several analyses that are very helpful to optimize the search for disease-causing exome variants.

MEDICAL SCIENCES

Author contributions: A. Belkadi, V.P., A.C., Y.I., A. Alcais, J.-L.C., and L.A. designed research; A. Belkadi, V.P., A.C., Y.I., Q.B.V., A. Abhyankar, L.S., J.E.B., A. Bousfiha, E./A.C., and B.B. performed research; A. Belkadi, V.P., A.C., Y.I., Q.B.V., A. Abhyankar, and B.B. contributed new reagents/analytic tools; A. Belkadi, V.P., A.C., Y.I., Q.B.V., A. Abhyankar, L.S., A. Alcais, B.B., J.-L.C., and L.A. analyzed data; A. Belkadi, V.P., A.C., Y.I., A. Alcais, J.-L.C., and L.A. wrote the paper; and J.E.B., A. Bousfiha, and E./A.C. provided clinical data.

Reviewers: L.B.B., Centre Hospitalo-Universitaire (CHU) Sainte-Justine/University of Montreal; and J.F., Ecole Polytechnique Fédérale de Lausanne (EPFL).

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1] A. Belkadi and V.P. contributed equally to this work.

[2] A complete list of the Exome/Array Consortium can be found in *Supporting Information*.

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

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606460113/-/DCSupplemental.

world, including North Africa and the Middle East. Both of these regions are poorly covered by the HapMap Project and the 1000 Genomes Project and have high consanguinity rates. We compared the information provided by the two datasets for the estimation of homozygosity rate and linkage analysis by homozygosity mapping. We also defined the optimal criteria for selecting WES variants to optimize PCA and ancestry prediction for individuals of various ethnic origins.

Results

We performed genotyping with the Affymetrix GWSA 6.0 array and WES with the Agilent Sureselect All Exons V4 Kit on 110 unrelated individuals (58 male and 52 female subjects) originating from six regions of the world, including North Africa (27 subjects) and the Middle East (16 subjects) (Table 1). After the application of quality control (QC) filters (*Methods*), 810,914 high-quality (HQ) GWSA SNVs and 249,310 HQ WES SNVs were retained for our analyses (Fig. S1). In total, 10,598 of these SNVs, with a call rate (CR) of 100%, were common to both WES and GWSA. We checked the genotype matching rate of these common variants between WES and GWSA with the PLINK Identity by State matrix (39). The mean Identity by State genotype matching rate between WES and GWSA was 99.37% (SD = 1.02%), a value similar to that reported in previous studies (40).

We first conducted PCA using 375 unrelated individuals from five world regions as the reference (Table S1). Data for these individuals were present in both the HapMap Project (HapMap release 3) (41) and the most recent 1000 Genomes Project phase 3 (42) database available since May of 2013. We merged our GWSA data with the HapMap data, such that all of the 810,914 HQ SNVs of our sample were present in the HapMap dataset. The resulting merged database was then used for PCA. We found that 183,065 (73.4%) of the 249,310 HQ SNVs detected in our 110 WES samples were present in the SNVs included in the 1000 Genomes Project phase 3 (Fig. S1). The difference in the number of variants in our WES data and the 1000 Genomes Project reflects the enrichment of our WES data in rare variants. We first conducted PCA on the HapMap/GWSA data (Fig. 1). Consistent with their geographic origin, the North African individuals mapped between the European and African clusters, whereas the Middle Eastern individuals mapped between the European and Asian clusters (Fig. 1). Like subjects from the Middle East, Central and South Americans were located between the European and Asian clusters for the first two principal components (PCs). However, the South American and Middle Eastern subjects were separated by the third PC.

We then performed a more formal comparison of PCA between the GWSA/HapMap data used as a gold standard and the WES/1000 Genomes Project data using the R_W correlation coefficient weighted by the eigenvalues of the significant PCs (*Methods*). We considered different CRs (range = 95–100%) and different minor allele frequency (MAF) thresholds (range = 0–5%) for the WES SNVs, because higher CRs and MAFs would be expected to increase variant quality while decreasing the number of variants (range = 39,391–183,013) (Table S2). The R_W correlation coefficient was calculated for the 14 PCs significant at $P < 0.05$ (Table S3). Correlations were particularly strong ($R_W > 0.98$) for the four first PCs, which accounted for >85% of the scaled eigenvalues in both GWSA and WES (Table S3). Overall, we found strong correlations (range = 0.813–0.892) between the PCA coordinates obtained by GWSA and WES for our 110 subjects for all combinations of CR and MAF (Fig. 2). The exclusion of rare variants (MAF < 2%) from the PCA clearly decreased the number of variants but increased the strength of the correlation. The strongest correlations were observed with WES variants with an MAF > 2%, and for MAF values in this range, CR had very little influence. The panel of WES variants with an MAF > 3% and a CR > 98% provided the highest R_W value at 0.892, corresponding to 85,112 variants in total, whereas the corresponding value was 183,013 in the largest panel (Table S2). The results of PCA with this panel of 85,112 SNVs are shown in Fig. 1, in which the distribution of population structures is very similar to that derived from the GWSA/HapMap data.

Table 1. Origin of 110 individuals included in the analysis

World region	No. of individuals
Central and South America*	5
Middle East[†]	16
North Africa[‡]	27
Sub-Saharan Africa[§]	6
Western Europe[¶]	53
Mixed origin[#]	3

*Individuals from Colombia, Brazil, and Mexico.

[†] Individuals from Turkey, Pakistan, Kuwait, India, Iran, Qatar, and Afghanistan.

[‡] Individuals from Morocco, Algeria, Tunisia, and Egypt.

[§] Individuals from Mali, Senegal, Comoros Islands, and Madagascar.

[¶] Individuals from France, Italy, Spain, Portugal, and the United Kingdom.

[#] Individuals with parents from sub-Saharan Africa and Europe and from the Middle East and Europe.

Next, we considered the prediction of coordinates for a single individual from WES data and a sample of publicly available data. Based on our previous findings, we used variants with an MAF > 3% and a CR > 98% when WES data for this single individual were merged with the 1000 Genomes Project data. Predictions were made independently for each of our 110 individuals, and the R_W correlation coefficient for the whole sample was again very strong at 0.844 (0.841 when MAF was >2% and CR was >98%). Interestingly, this correlation was also very strong when we considered only ethnic groups not represented in the reference panels of the HapMap Project and the 1000 Genomes Project, such as 16 individuals from the Middle East ($R_W = 0.853$), 27 individuals from North Africa ($R_W = 0.829$), and 3 subjects of mixed origin ($R_W = 0.949$). All of these results indicate that WES data based on common variants are appropriate for use in population structure analyses and inferring the ethnic ancestry of an individual. Finally, we compared the performance of WES and GWSA in terms of local ancestry inference using Hapmix (43), which can consider two ancestral populations (*Methods*). The correlation between the proportions of ancestry obtained by GWSA or WES data in our 110 individuals was high, varying from 0.84 to 0.99 (Fig. S2) according to the two ancestral populations considered among the four HapMap/1000 Genomes populations European (CEU), Han Chinese (CHB), Yoruba Nigerian (YRI), and Mexican (MEX). An example for the analysis using CEU and YRI as ancestral populations is shown in Fig. S2. These high correlations are consistent with our PCA results, further indicating that WES data could be used to infer local ancestry.

We then estimated homozygosity rates by calculating the inbreeding coefficient F by two approaches: one based on the search for runs of homozygosity (ROHs) over a given length of the genome (20) and the other based on the use of Markov processes to model homozygous states throughout the genome by the FEstim method (19). We identified ROHs with PLINK (39), in which a sliding window of 1,000 kb is passed across the genome, with homozygosity determined at each window. We considered different numbers of SNVs within the sliding window (*Methods*). With GWSA data, the mean homozygosity of our sample, estimated by FEstim ($F_{\text{ESTIM-GWSA}}$), was 1.64% (SD = 3.44%; range = 0–15.50%) (Table S4). As expected, with the ROH approach, the mean homozygosity ($F_{\text{ROH-GWSA}}$) increased as the number of SNVs included in the window decreased from 1.34% (300 SNVs) to 2.27% (100 SNVs). The $F_{\text{ROH-GWSA}}$ values obtained with 200 SNVs ($F_{\text{ROH-GWSA200}} = 1.67\%$) and 250 SNVs ($F_{\text{ROH-GWSA250}} = 1.47\%$) were the closest to $F_{\text{ESTIM-GWSA}}$, which could be considered the reference estimate (44). They were also strongly correlated with the $F_{\text{ESTIM-GWSA}}$ estimates ($r = 0.973$ and $r = 0.975$, respectively). With WES data, the estimated F_{ESTIM} value was higher than that obtained with GWSA data at 2.53% (SD = 5.23%; range = 0–22.50%), and the coefficient of correlation with the $F_{\text{ESTIM-GWSA}}$ estimates was 0.889.

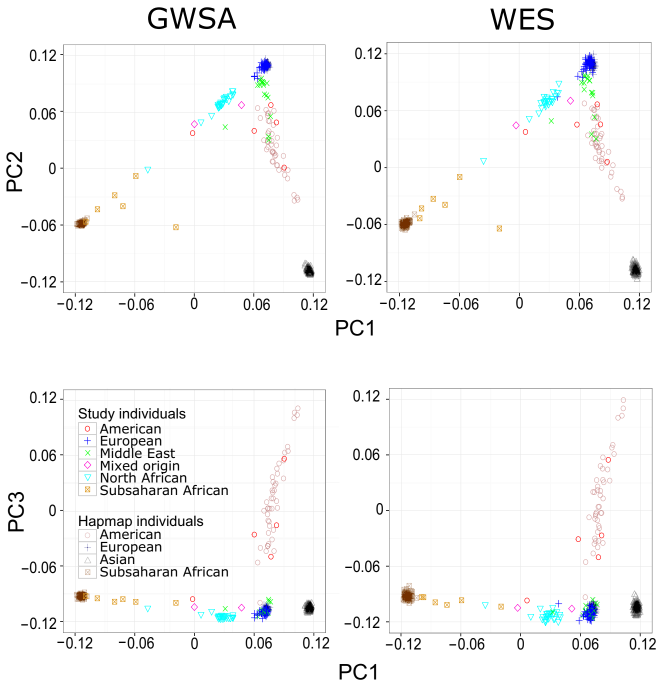

Fig. 1. PCA was performed with smartPCA software on both (*Left*) GWSA and (*Right*) WES data. The results for WES data are presented for SNVs with an MAF > 0.03 and a CR > 98%. (*Upper*) The first two PCs and (*Lower*) the first and third PCs are plotted. We included a total of 110 individuals (colored plots) and 375 HapMap individuals (black plots) in the analysis.

However, less than 20% of the submaps generated by the $F_{\text{ESTIM-WES}}$ approach could be used for analysis in 18 subjects who were from various geographic regions (12 from Europe, 3 from sub-Saharan Africa, 2 from North Africa, and 1 from the Middle East). We noted that these 18 individuals had significantly higher missing rates of WES variants than the 92 others (3% vs. 0.9%; P = 0.0003). In addition, we found that the mean chromosomal segment lengths homozygous by descent (HBD) were significantly lower ($P = 0.0004$) when using WES data (mean = 1.91 Mb) than when using GWSA data (mean = 2.64 Mb), indicating that WES data may lead more often to exclusion of submaps compared with GWSA because of smaller HBD segment lengths. Overall, these findings suggest that the $F_{\text{ESTIM-WES}}$ estimates may be less reliable, at least for these 18 individuals.

When the ROH approach was applied to WES data, the mean $F_{\text{ROH-WES}}$ estimates varied from 0.80% (100 SNVs) to 1.95% (20 SNVs) (Table S4). The $F_{\text{ROH-WES}}$ values obtained with 50 SNVs ($F_{\text{ROH-WES50}}$ = 1.73%) and 30 SNVs ($F_{\text{ROH-WES30}}$ = 1.87%) were the closest to $F_{\text{ESTIM-GWSA}}$ and also strongly correlated with $F_{\text{ESTIM-GWSA}}$ (r = 0.933 and r = 0.930, respectively) and $F_{\text{ROH-GWSA200}}$ (r = 0.952 and r = 0.951, respectively) (Table S5) data. With the optimal parameters proposed in a previous study (37), the mean $F_{\text{ROH-WES10}}$ was higher at 2.97%, with correlation coefficients of 0.931 with $F_{\text{ESTIM-GWSA}}$ and 0.985 with $F_{\text{ROH-WES30}}$. Thus, for both GWSA and WES data, the most appropriate number of SNVs within a 1,000-kb window for calling an ROH providing F_{ROH} estimates similar to $F_{\text{ESTIM-GWSA}}$ was close to the mean number of SNVs per 1,000 kb (corresponding to 0.37% of the autosomal genome) from the GWSA (~242 SNVs per 1,000 kb) and WES (~27 SNVs per 1,000 kb) data (*Methods*). These results indicate that WES can be used to obtain reliable homozygosity estimates by ROH methods if the number of SNVs within a window of 1,000 kb used corresponds to about 0.37% of the total number of available autosomal HQ WES SNVs (~30 SNVs in this analysis).

Based on the homozygosity results, we selected 15 individuals with $F_{\text{ROH-GWSA250}}$ and $F_{\text{ROH-WES30}}$ above 3% for linkage analysis by homozygosity mapping, because the offspring of first cousin marriages may have as little as 3% of their genome identical by descent (19); 11 of these 15 individuals were known to have been born to consanguineous parents. Information about consanguinity was not available for the other four subjects, although inbreeding was considered likely given their high rates of homozygosity. We also assumed that family structure was the same across families and that the patient was the only person genotyped/sequenced in each family. We performed homozygosity mapping with either GWSA or WES data (including all HQ variants with CRs > 98%). We first compared the linkage information content provided by the two methods, because this content provides some indication as to how closely the available markers approach the ideal situation of complete inheritance information concerning the segregation of the chromosomal region tested. Over the 22 autosomes, GWSA provided 51% more information, on average, than WES data. The ratio of the amount of information provided by WES to that provided by GWSA ranged from 0.41 on chromosome 21 to 0.90 on chromosome 19 (Fig. 3). This ratio was strongly correlated (Pearson's correlation coefficient = 0.72) with the proportion of coverage by the exome kit for each chromosome defined as the number of bases covered by the probes over the total length of each chromosome (Fig. 3). For example, chromosomes 19 and 22 contain a high proportion of coding sequences. They are, therefore, more densely covered by WES data than the other chromosomes, resulting in a higher information ratio. We then restricted our linkage analysis to the regions covered by the exome kit. These regions included a total of 10,674, and 73,565 autosomal SNVs in GWSA and WES data, respectively. In these regions, the amount of information provided was 1.97 higher, on average, with WES data than with GWSA data. Indeed, the WES/GWSA information ratio ranged from 1.35 on chromosome 14 to 3.71 on chromosome 21 (Fig. 3). Thus, for the regions covered by the exome kit, WES data clearly provided more information for linkage analysis than GWSA data.

Finally, we compared the linked regions larger than 1 Mb with a logarithm of the odds (LOD) score above 1 (the maximum expected LOD score in the family structure that we analyzed was 1.2), which we identified by conducting the analysis with three different sets of SNVs from (*i*) GWSA, (*ii*) WES, and (*iii*) the combination of GWSA and WES data (GWSA+WES). The third set of data with the largest number of SNVs was used as the reference for the linkage results (this combined set would be expected to provide the true linked regions). From these GWSA+WES results, we were able to estimate the proportion of linked regions identified by both GWSA and WES, those identified only by GWSA, and those identified only by WES. At the

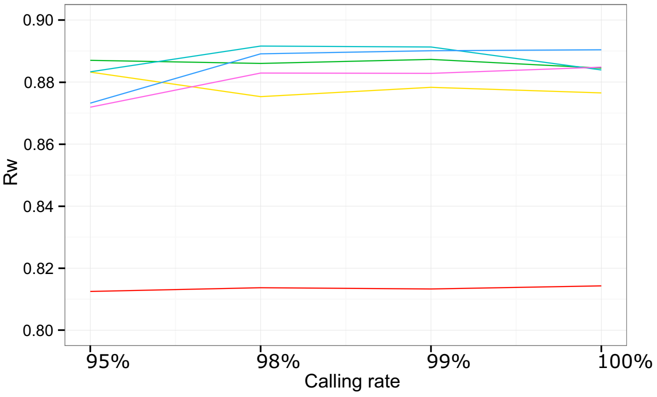

Fig. 2. Weighted correlation, R_W, between the PCA coordinates obtained with GWSA and WES for 110 individuals of our sample as a function of the CR (x axis) and the MAF thresholds used to filter WES SNVs: no MAF filter (red line), MAF > 1% (yellow line), MAF > 2% (green line), MAF > 3% (turquoise line), MAF > 4% (blue line), and MAF > 5% (purple line). R_W was calculated as described in *Methods*.

genome-wide scale, 76.3% of these regions were found by both GWSA and WES in 15 families, 17.7% were found by GWSA only, and 6.0% were found by WES only (Fig. S3). The WES/GWSA information ratio was higher in the regions found by WES only (mean WES/GWSA information ratio of 1.19) than in those found exclusively by GWSA (mean information ratio of 0.48). We conducted a similar analysis restricted to coding regions covered by the exome kit. We found that 83.5% of the regions were found by both GWSA and WES in 15 families, and slightly more regions were found by WES only (9.1%) than by GWSA only (7.4%) (Fig. S3). The linked regions found by GWSA only were regions in which WES supplied less information than GWSA (mean information ratio of 0.39) because of the small number of variants in the targeted sequenced segments. Overall, WES seems to provide reasonable linkage results at the genome-wide scale, with 82.3% of linked regions correctly detected vs. 94% for GWSA. In coding regions, WES is more informative overall and more powerful, detecting 92.6% of linked regions correctly, whereas 90.9% of these regions were correctly detected by GWSA.

Discussion

PCA is usually performed on common markers provided by GWSA. We conducted the first comprehensive PCA comparison of GWSA and WES, to our knowledge, by measuring a specific correlation. We found that performing PCA on WES data with HQ variants (with a CR > 95%) with an MAF > 2% provided a distribution of population structures very similar to that obtained from GWSA data for individuals of various ethnic origins. These criteria substantially decreased (>50%) the total number of WES variants used for the analyses, but they clearly provided an optimal tradeoff for HQ PCA. WES studies can be carried out on limited numbers of individuals, sometimes a single family or a single patient (45). We also showed that WES can accurately predict the ethnic origin of a single individual when using a sample of publicly available data including individuals belonging to ethnic groups that are not directly represented in the reference panel or who are born to parents of different origins. We also found reliable estimates when using WES data to infer local ancestry by means of the Hapmix approach (43). These results indicate that WES data are appropriate for use in population structure analyses and inferring the ethnic ancestry of an individual. The extent to which rare WES variants (MAF < 1%) could be used to refine population substructures remains to be investigated in depth, because it has been shown that rare variants could show stratification patterns that are different from those captured by common variants (24, 46). It will be particularly important to assess the influence of these stratification patterns on the association studies focusing on the role of rare variants (23, 46).

Genetic data from GWSA are used to estimate the homozygosity rate in patients to predict or confirm parental consanguinity in particular. We used the two most widely used approaches to estimate F from GWSA and WES data. We searched for ROHs and used Markov processes to assess homozygous states throughout the genome by the F_{ESTIM} approach. Using F_{ESTIM} on multiple sparse maps, as recommended (47), we obtained reliable homozygosity estimates with GWSA data. With WES data, we observed that ∼16% of individuals had a high proportion (>80%) of submaps that could not be used for the estimation of F_{ESTIM}. Although this aspect requires additional investigation, a first analysis indicated that WES data may be more sensitive to submap exclusions with the F_{ESTIM} approach because of smaller HBD segments than those obtained with GWSA data, in particular in subjects who have more missing data. Using ROH methods, we found that optimal F_{ROH} estimates for both GWSA and WES data (compared with $F_{\mathrm{ESTIM\text{-}GWSA}}$) were obtained by considering a number of SNVs for calling an ROH within a 1,000-kb window close to the mean number of SNVs per 1,000 kb available in the GWSA (∼250 SNVs in our study) or the WES (∼30 SNVs in our study) data. In this context, estimates of mean homozygosity from WES were very similar to those obtained with GWSA, and there was a strong correlation between the two estimates of F_{ROH} ($r = 0.95$ between $F_{\mathrm{ROH\text{-}GWSA250SNVs}}$ and $F_{\mathrm{ROH\text{-}WES30SNVs}}$). This result is consistent with the findings of a previous study (37), although the optimal configuration for detecting ROH from WES data in this previous study included fewer SNVs (10) within the 1,000-kb window. The detection of ROHs from WES data could also be improved by adding genotyped SNVs from other family members (17). In any case, reliable homozygosity estimates could be obtained from WES data only if ROHs were identified with PLINK, considering a number of SNVs within a 1,000-kb window corresponding to ∼0.37% of the total number of available HQ SNVs.

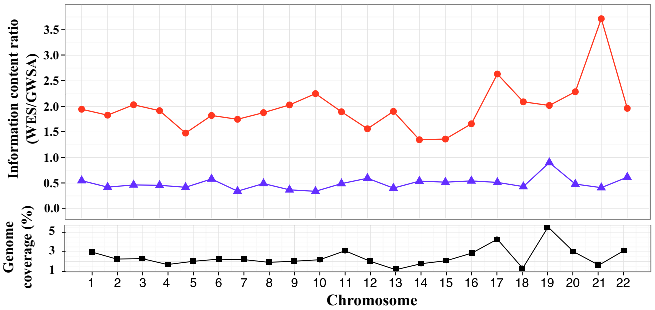

Fig. 3. (*Upper*) Ratio of information content (IC) for linkage analysis performed with GWSA and WES SNPs calculated with Merlin software. Each dot represents the IC ratio (IC for WES/IC for GWSA). The IC is the mean amount of information for all SNPs per chromosome computed over all of 15 families. Blue triangles indicate the ratio at the whole-genome level; red circles indicate the ratio for the analysis conducted with SNPs located in the regions covered by the SureSelect Exome Kit. (*Lower*) Black squares indicate the proportion of the whole genome covered by the probes of the SureSelect Exome Kit defined as the number of bases covered by the probes divided by the total length of each chromosome.

Many linkage studies have been conducted with WES data in the context of Mendelian disorders (1–3, 6, 15), but to our knowledge, only two have formally compared their results with those obtained with GWSA data. Using real genetic data from three families as an example, Smith et al. (35) showed that accurate genetic linkage mapping could be performed with WES SNVs. Gazal et al. (36) performed a linkage study of two families with both simulated and real data. They reported similar performances for linkage analyses conducted with GWSA or WES (36). As mentioned above, the recent study by Kancheva et al. (37) was based on the detection of ROHs in patients born to consanguineous families without a formal linkage analysis. Here, we extended the analysis to 15 individuals with high homozygosity rates (>3%) in the specific context of linkage analysis by homozygosity mapping, a frequent situation in which WES data may be available for the patient only.

We first analyzed the linkage information content provided by GWSA and WES across the genome. The linkage information obtained with WES was generally only about one-half that obtained with GWSA at the genome-wide level and highest for chromosomal regions with a high density of coding regions. Consistent with this result, we found that, at the genome-wide level, WES detected a smaller proportion of linked regions than GWSA, although this proportion remained substantial at 82.3% (vs. 94% with GWSA). GWSA, nevertheless, missed 6% of the linked regions, corresponding to regions in which the information content was higher for WES than for GWSA. In the regions covered by the exome kit, the information content obtained with WES was generally twice that obtained with GWSA, and WES detected slightly more linked regions than GWSA (92.6% vs. 90.9%). However, in some coding regions, the segments sequenced by WES included only a small number of SNVs, resulting in a low information content and accounting for the small proportion of linked coding regions (7.4%) detected only by GWSA. Clearly, with the decreasing cost of whole-genome sequencing (48), optimal approaches will, in the future, involve linkage analysis together with other analyses of whole-genome sequencing data (49). However, it is currently possible to use WES data for PCA after the application of the appropriate QC filters and adjustment for population substructure to estimate homozygosity rates by ROH and perform reliable linkage analyses, particularly for coding regions.

Methods

Study Subjects. The individuals used in the analysis were selected from samples ascertained by our laboratory and recruited with the collaboration of many clinicians. They presented a variety of severe infectious diseases and/or primary immunodeficiencies. Although these individuals do not form a random sample, they were ascertained through a number of distinct phenotypes and in different countries. Cohort-specific effects are, therefore, not expected to bias patterns of variation. Among these patients, we studied only 110 individuals who had both WES by Agilent Sureselect All Exons V4 (50 Mb) Single-Sample Capture and genotyping by the Affymetrix Genome-Wide SNV 6.0 Array. The retained 110 subjects studied (58 male and 52 female patients) originated from different regions of the world (Table 1). Written consent was obtained from all subjects included in this study, which was overseen by the Comité de Protection des Personnes (Institutional Review Board) Ile de France 2 (Institutional Review Board no. 00001072).

WES. WES was performed on an Illumina HiSeq 2000 by Agilent Sureselect All Exons V4 (50 Mb) Single-Sample Capture at the Rockefeller core facilities and the New York Genome Center. Sequencing was performed with 2 × 100 bp paired end reads, and we pooled five samples per lane. We used the Genome Analysis Software Kit (GATK) best practice pipeline to analyze our WES data (50). Reads were aligned with the human reference genome (hg19) with the Maximum Exact Matches algorithm in Burrows–Wheeler Aligner (51). Local realignment around indels was performed with the GATK (52). PCR duplicates were removed with Picard tools (broadinstitute.github.io/picard/). The GATK base quality score recalibrator was applied to correct sequencing artifacts. Individual genomic variant call files were generated with the GATK HaplotypeCaller, and joint genotyping was performed with the GATK Genotype genomic variant call files. The calling process targeted regions covered by the WES 50-Mb Kit, including 200 bp flanking each region.

All variants with a Phred-scaled SNV quality ≤ 30 were filtered out. We then used the GATK Variant Quality Score Recalibrator (50) on the combined variant call file for 110 samples. We retained 1,213,952 SNVs that passed the Variant Quality Score (VQS) Recalibrator filter (VQS log-odds > -0.682). We filtered out sample genotypes with a coverage $< 8\times$, a genotype quality < 20, or a ratio of reads for the less covered allele (reference or variant allele) over the total number of reads covering the position at which the variant was called in the heterozygous genotypes of $<20\%$ using an in-house script. Finally, we excluded from the analysis 704,954 variants, for which more than 10% of the genotypes were missing. A set of 249,310 HQ variants was retained for the analysis (Fig. S1).

GWSA. In total, 110 individuals were genotyped with the Affymetrix Genome-Wide SNV 6.0 Array. Genotype calling was achieved with Affymetrix Power Tools (www.affymetrix.com/estore) for all individuals. In total, 909,622 raw SNVs were detected. We applied QC criteria similar to those used in Hapmap release 3 (41) by removing SNVs with a CR $<$ 95% and a P value in Fisher's exact test for Hardy–Weinberg equilibrium on 53 European individuals of $<10^{-6}$. In total, 810,914 HQ SNVs passed this Hapmap filter and were retained for analysis.

PCA and Local Ancestry Inference. PCA was carried out with the smartPCA program (53). We initially included 375 unrelated individuals from five regions of the world (Table S1) present in both the 1000 Genomes Project and the Hapmap (Hapmap release 3) Project. We used the data from the Affymetrix 6.0 array and the 1000 Genomes Project for these 375 individuals as a reference for our PCA with GWSA and WES data, respectively. We further considered four different CRs for WES SNVs (95%, 98%, 99%, and 100%) and different MAF thresholds for WES variants (0.01, 0.02, 0.03, 0.04, and 0.05), because these parameters may affect the results of the PCA (54).

We compared PCAs on GWSA and WES data using our whole sample of 110 individuals by calculating the weighted correlation, R_W, between the coordinates of our individuals obtained with GWSA or WES data. These correlations were summed over the M significant PCs and weighted by the mean eigenvalues of the corresponding GWSA and WES components as follows:

$$R_W = \sum_{j=0}^{M} \frac{\left(P_{\mathrm{WES}_j} + P_{\mathrm{GWSA}_j}\right)}{2} \operatorname{cor}_{\mathrm{Pearson}}\left(\mathrm{WES}_j,\ \mathrm{GWSA}_j\right),$$

where P_{WES_j} and P_{GWSA_j} are the normalized eigenvalues of the PC j in the analysis of WES and GWSA data, respectively; WES_j and GWSA_j are the vectors of the coordinates for PC j in our 110 individuals obtained in PCA on WES and GWSA data, respectively; M is the number of significant PCs (P value $<$ 0.05) obtained with unsupervised Tracy–Widom statistics (Table S3); and the R_W correlation coefficient was calculated for each of 25 combinations of CRs and MAF shown in Table S2.

The local ancestry for 110 study individuals was inferred by Hapmix (43). Because Hapmix assumes two ancestral populations, we ran the software for six sets of two ancestral populations from four HapMap/1000 Genomes Projects: CEU and YRI, CEU and CHB, CEU and MEX, YRI and CHB, YRI and MEX, and CHB and MEX. Because the MEX population included only 44 independent individuals with both HapMap and 1000 Genomes data, we also used a set of 44 independent individuals for three other ancestral populations. The correlation between the proportions of ancestry estimated in our 110 individuals using the GWSA or the WES data was computed over the whole autosomal genome for each of six sets of ancestral populations.

Estimation of Homozygosity. Several approaches have been proposed for estimating the inbreeding coefficient F from genetic data (20). Chromosomal regions that are HBD can be identified by searching for ROHs over a given length, providing an estimate of F based on the proportion of the autosomal genome in ROHs (20). For these analyses, we used the HQ autosomal SNVs with an MAF $>$ 0.05 (654,155) identified by GWSA and 73,565 SNVs with a CR $>$ 98% and an MAF $>$ 0.05 identified by WES. We identified ROHs with PLINK (39), which has several advantages over other methods (37). We used the classical PLINK method with default parameters, in which a 1,000-kb window is moved across the genome, with homozygosity determined for each window. We varied the number of SNVs within the 1,000-kb window required to call an ROH using a smaller number for WES (20, 30, 50, and 100) than for GWSA (100, 200, 250, and 300) to account for the lower total autosomal SNV counts in WES than in GWSA data (37). The choice of these numbers was based on the fact that a window of 1,000 kb corresponds to ∼0.37% of the autosomal genome, giving mean numbers of available SNVs per 1,000 kb of ∼27 for WES data and ∼242 for GWSA data. We also considered the PLINK parameters reported to be optimal in a recent study (37) for the analysis of the WES data. These parameters included 10 SNVs within the 1,000-kb window. We obtained a genomic measurement of individual homozygosity (F_{ROH}) by determining the proportion of the autosomal genome present in ROHs (20).

Another approach for estimating F involves modeling the HBD states of the different markers of one individual along the genome as a Markov process using hidden Markov models as initially proposed in the F_{ESTIM} method (19). This method assumes that marker alleles are independent conditionally on HBD state, which is not true for dense SNVs (in array or exome data), for which linkage disequilibrium (LD) may occur. We used the FEstim_SUBS method to minimize LD between SNVs as recommended in a previous study (44) for the random extraction of sparse markers every 0.5 cM to create 1,000 submaps. This strategy does not require the estimation of LD scores for the data, and F is estimated by calculating the median value of the estimates obtained from the different maps. The FSuite program was used to calculate F_{ESTIM} for each individual from both GWSA and WES data (47).

Linkage Analysis. We performed linkage analysis assuming autosomal recessive inheritance with complete penetrance (homozygosity mapping) on individuals found to have a high rate of homozygosity. For each individual, we created the same family structure based on a unique consanguinity loop at the first cousin level. The main goal of our study was to compare the linkage information provided by WES with that provided by GWSA using the same familial structure and the same data for all families, consisting of nine individuals with a single genotyped subject assumed to be affected (the offspring of the youngest generation). We carried out parametric multipoint linkage analysis by homozygosity mapping (55) with Merlin software (56). A population disease allele frequency of 0.0001 was specified together with a fully penetrant recessive genetic model. LOD scores were calculated for every marker (from WES or GWSA data), and 1000 Genomes Project allele frequencies were used (42). Information content was also estimated for both WES and GWSA data, because this parameter provides an indication of how closely the available markers approach the ideal situation of complete inheritance information for the segregation of the chromosomal region considered.

ACKNOWLEDGMENTS. We thank all Exome/Array Consortium members for contributing to the collection of samples: Waleed Al-Herz, Cigdem Arikan, Peter Arkwright, Cigdem Aydogmus, Olivier Bernard, Lizbeth Blancas-Galicia, Stéphanie Boisson-Dupuis, Damien Bonnet, Omar Boudghene Stambouli, Lobna Boussafara, Jeannette Boutros, Jacinta Bustamante, Michael Ciancanelli, Theresa Cole, Antonio Condino-Neto, Mukesh Desai, Claire Fieschi, José Luis Franco, Philippe Ichai, Emmanuelle Jouanguy, Melike Keser-Emiroglu, Sara S. Kilic, Seyed Alireza Mahdaviani, Nizar Malhaoui, Davood Mansouri, Nima Parvaneh, Capucine Picard, Anne Puel, Didier Raoult, Nima Rezaei, Ozden Sanal, Silvia Sanchez Ramon, François Vandenesch, Guillaume Vogt, and Shen-Ying Zhang (*Supporting Information*). We also thank Lahouari Amar and Yelena Nemirovskaya for their invaluable help. The Laboratory of Human Genetics of Infectious Diseases is supported by European Research Council Grant ERC-2010-AdG-268777, the French National Research Agency under the "Investments for the Future" Program Grant ANR-10-IAHU-01, National Institute of Allergy and Infectious Diseases Grant 5U01AI088685, and grants from INSERM, Paris Descartes University, the St. Giles Foundation, and the Rockefeller University.

1. Ng SB, et al. (2010) Exome sequencing identifies the cause of a Mendelian disorder. *Nat Genet* 42(1):30–35.
2. Bolze A, et al. (2010) Whole-exome-sequencing-based discovery of human FADD deficiency. *Am J Hum Genet* 87(6):873–881.
3. Bamshad MJ, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12(11):745–755.
4. Kiezun A, et al. (2012) Exome sequencing and the genetic basis of complex traits. *Nat Genet* 44(6):623–630.
5. Tennessen JA, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.
6. Bolze A, et al. (2013) Ribosomal protein SA haploinsufficiency in humans with isolated congenital asplenia. *Science* 340(6135):976–978.
7. Chakravarti A, Clark AG, Mootha VK (2013) Distilling pathophysiology from complex disease genetics. *Cell* 155(1):21–26.
8. Itan Y, et al. (2013) The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci USA* 110(14):5558–5563.
9. Rausell A, et al. (2014) Analysis of stop-gain and frameshift variants in human innate immunity genes. *PLOS Comput Biol* 10(7):e1003757.
10. Cirulli ET, et al.; FALS Sequencing Consortium (2015) Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* 347(6229):1436–1441.
11. Itan Y, et al. (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci USA* 112(44):13615–13620.
12. Itan Y, Casanova J-L (2015) Can the impact of human genetic variations be predicted? *Proc Natl Acad Sci USA* 112(37):11426–11427.
13. Itan Y, et al. (2016) The mutation significance cutoff: Gene-level thresholds for variant predictions. *Nat Methods* 13(2):109–110.
14. Boisson B, et al. (2013) An ACT1 mutation selectively abolishes interleukin-17 responses in humans with chronic mucocutaneous candidiasis. *Immunity* 39(4):676–686.
15. Byun M, et al. (2013) Inherited human OX40 deficiency underlying classic Kaposi sarcoma of childhood. *J Exp Med* 210(9):1743–1759.
16. Carr IM, et al. (2013) Autozygosity mapping with exome sequence data. *Hum Mutat* 34(1):50–56.
17. Santoni FA, Makrythanasis P, Antonarakis SE (2015) CATCHing putative causative variants in consanguineous families. *BMC Bioinformatics* 16:310.
18. Goddard KAB, Wijsman EM (2002) Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. *Genet Epidemiol* 22(3):205–220.
19. Leutenegger A-L, et al. (2003) Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* 73(3):516–523.
20. McQuillan R, et al. (2008) Runs of homozygosity in European populations. *Am J Hum Genet* 83(3):359–372.
21. Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 60(3):227–237.
22. Moore CB, et al. (2013) Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genet* 9(12):e1003959.
23. Zawistowski M, et al. (2014) Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *Eur J Hum Genet* 22(9):1137–1144.
24. O'Connor TD, et al.; NHLBI GO Exome Sequencing Project; ESP Population Genetics and Statistical Analysis Working Group, Emily Turner (2015) Rare variation facilitates inferences of fine-scale population structure in humans. *Mol Biol Evol* 32(3):653–660.
25. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11):e1002967.
26. Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* 197(2):573–589.
27. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456(7218):98–101.
28. Bryc K, et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA* 107(2):786–791.
29. Pickrell JK, et al. (2014) Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci USA* 111(7):2632–2637.
30. Abdulla MA, et al.; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium (2009) Mapping human genetic diversity in Asia. *Science* 326(5959):1541–1545.
31. Behar DM, et al. (2010) The genome-wide structure of the Jewish people. *Nature* 466(7303):238–242.
32. Moreno-Estrada A, et al. (2014) Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344(6189):1280–1285.
33. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100–1104.
34. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
35. Smith KR, et al. (2011) Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol* 12(9):R85.
36. Gazal S, et al. (2016) Can whole-exome sequencing data be used for linkage analysis? *Eur J Hum Genet* 24(4):581–586.
37. Kancheva D, et al. (October 22, 2015) Novel mutations in genes causing hereditary spastic paraplegia and Charcot-Marie-Tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing. *Genet Med*, 10.1038/gim.2015.139.
38. Wang C, Zhan X, Liang L, Abecasis GR, Lin X (2015) Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am J Hum Genet* 96(6):926–937.
39. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
40. Szpiech ZA, et al. (2013) Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet* 93(1):90–102.
41. International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52–58.
42. 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
43. Price AL, et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5(6):e1000519.
44. Gazal S, et al. (2014) Inbreeding coefficient estimation with dense SNP data: Comparison of strategies and application to HapMap III. *Hum Hered* 77(1-4):49–62.
45. Casanova J-L, Conley ME, Seligman SJ, Abel L, Notarangelo LD (2014) Guidelines for genetic studies in single patients: Lessons from primary immunodeficiencies. *J Exp Med* 211(11):2137–2149.
46. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44(3):243–246.
47. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L (2014) FSuite: Exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics* 30(13):1940–1941.
48. Belkadi A, et al. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 112(17):5473–5478.
49. Ott J, Wang J, Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 16(5):275–284.
50. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
51. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589–595.
52. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
53. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
54. He H, et al. (2011) Effect of population stratification analysis on false-positive rates for common and rare variants. *BMC Proc* 5(Suppl 9):S116.
55. Lander ES, Botstein D (1987) Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. *Science* 236(4808):1567–1570.
56. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30(1):97–101.