scientific reports

Check for updates

OPEN

How local reference panels improve imputation in French populations

Anthony F. Herzig[1]✉, Lourdes Velo-Suárez[1,2], The FrEx Consortium[1*], The FranceGenRef Consortium[3*], Christian Dina[4], Richard Redon[4], Jean-François Deleuze[5,6] & Emmanuelle Génin[1,2]

Imputation servers offer the exclusive possibility to harness the largest public reference panels which have been shown to deliver very high precision in the imputation of European genomes. Many studies have nonetheless stressed the importance of 'study specific panels' (SSPs) as an alternative and have shown the benefits of combining public reference panels with SSPs. But such combined approaches are not attainable when using external imputation servers. To investigate how to confront this challenge, we imputed 550 French individuals using either the University of Michigan imputation server with the Haplotype Reference Consortium (HRC) panel or an in-house SSP of 850 whole-genome sequenced French individuals. With approximate geo-localization of both our target and SSP individuals we are able to pinpoint different scenarios where SSP-based imputation would be preferred over server-based imputation or vice-versa. This is achieved by showing to a high degree of resolution the importance of the proximity of the reference panel to target individuals; with a focus on the clear added value of SSPs for estimating haplotype phase and for the imputation of rare variants (minor allele-frequency below 0.01). Such benefits were most evident for individuals from the same geographical regions in France as the SSP individuals. Overall, only 42.3% of all 125,442 variants evaluated were better imputed with an SSP from France compared to an external reference panel, however this rises to 58.1% for individuals from geographic regions well covered by the SSP. By investigating haplotype sharing and population fine-structure in France, we show the importance of including SSP haplotypes for imputation but also that they should ideally be combined with large public panels. In the absence of the unattainable results from a combined panel of the HRC and our French SSP, we put forward a pragmatic solution where server-based and SSP-based imputation outcomes can be combined based on comparing posterior genotype probabilities. We show that such an approach can give a level of imputation accuracy in excess of what could be achieved with either strategy alone. The results presented provide detailed insights into the accuracy of imputation that should be expected from different strategies for European populations.

Population-based genotype imputation remains a widely used technique for enriching datasets of genotyped or low-coverage sequenced individuals. Advances in software capabilities have been rapid, enormous haplotype reference panels have been assembled, and dedicated computation servers have been created at the University of Michigan[1] and at the Sanger institute[2].

Numerous studies have compared the effectiveness of different imputation strategies. The important point of consensus being that imputation benefits from a reference panel that is both large and diverse[2–6]. Public reference panels widely used for imputation include the 1000 Genomes Project (1000G)[7], the HRC panel[2] and the TOPMED panel[8]. The size and variety of origin of the reference haplotypes in such panels aims to ensure accurate imputation of target-individuals from different populations. Many groups have published results underlying the importance of preferring 'local reference panels' or 'study specific panels' (SSPs)—the intuitive concept being that the best panel for imputation should contain reference haplotypes that closely resembles the target individuals. Furthermore as rarer genetic variants are often younger[9,10], they are expected to be geographically clustered and hence only successfully imputed with geographically relevant reference haplotypes. Increased imputation

[1]Univ Brest, Inserm, EFS, UMR 1078, GGB, Brest, France. [2]CHRU Brest, Brest, France. [3]LABEX GENMED, Centre National de Recherche en Génomique Humaine, Evry, France. [4]Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, Nantes, France. [5]Centre National de Recherche en Génomique Humaine (CNRGH), Université Paris-Saclay, CEA, Evry, France. [6]Fondation Jean Dausset - Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France. *A list of authors and their affiliations appears at the end of the paper. ✉email: anthony.herzig@inserm.fr

accuracy coming from SSPs has been shown in populations such as the Netherlands[11], Estonia[12], Norway[13], and Japan[14]. SSP imputation also improves the power of genome-wide association studies (GWAS) involving both common and rare variants[13,15–17]. The benefits of using SSPs have been shown to be particularly evident in the context of isolated populations[17–21].

SSPs may often be relatively small and so the best approach may often be to combine an SSP with a large cosmopolitan reference panel. Though combining public and study specific reference panels is computationally feasible, it remains problematic for other practical reasons. Panels such as the HRC[2] or TOPMED[8] are only fully available through online servers and hence it is not possible to merge their data with one's in-house sequencing data. Hence, most published results cited above involving a combination of panels have merged an SSP with the freely available (but smaller in comparison) 1000G. It should also be recognised that as leading imputation servers are located outside of the European Union, General Data Protection Regulations have added significant complications for the use of imputation servers[22]. In this study, we elucidate precisely what is to be gained or lost from pursuing the use of such servers compared to in-house imputation using SSPs.

Leading population-based imputation software invoke haplotype copying models based on the Li-Stephens model[23]. This model uses coalescent theory, capturing the idea that if two chromosomes (at a given position) are followed back in time, they will eventually coalesce, sharing a (most recent) common ancestor and this will translate into stretches of shared haplotypes between individuals. For two unrelated individuals, any given genomic region would likely contain many differences representing a very long coalescent time between the pair. But with a large enough sample of a population and in a given genomic region, each observed haplotype can be expected to have a shared lineage (and hence have a relatively recent common ancestor) with at least one other haplotype in the sample. Thus these two haplotypes would likely share a near identical haplotype (allowing for only a few very recent mutations) that would stretch far enough to contain multiple common genetic variants. Extending this idea across regions, a given chromosome from the sample can be described as a mosaic of small haplotype segments present in the pool of all other chromosomes in the sample. This concept is harnessed by imputation software; each target individual chromosome is modelled as a mosaic of reference panel haplotypes using genotyping information for the target individual on a set of common variants. Once a likely chain of copying haplotypes is estimated based on similarities for common genetic variants, missing genotypes can be inferred. Or more often, posterior probabilities of missing genotypes across many potential chains are estimated. Developments in imputation software have been driven by the need to make inference from larger and larger reference panels, but also to operate efficiently to find the best subsets of reference individuals for each chromosomal region. In particular, the PBWT[24] algorithm has allowed for very rapid sub-selections of reference panel individuals to serve as region-specific reference haplotype pools. PBWT can be employed as a phasing and imputation software on its own but the algorithm has also been incorporated into other software such as EAGLE2[25], IMPUTE5[5] and SHAPEIT4[26]. With the concepts of the Li-Stephens model in mind, it is intuitive that imputation will be successful if the reference panel contains relevant haplotypes which closely match the target individual but also enough diversity to enable good haplotype matching across the target's whole chromosome—i.e. there are no weak links in the chain. This can explain potentially counter-intuitive results such as the inclusion of the UK10K[27] imputation panel improving the imputation of Italian[28] and even Chinese[29] genomes.

Aside from choice of reference panel, an important consideration is the estimation of haplotypes—referred to herein simply as 'phasing'. The accuracy of phasing has also been widely evaluated, with a parallel rapid development of competing software. Population based phasing software use broadly the same haplotype copying models as imputation software, only that two chains of mosaics have to be found simultaneously rather than a single one. An important difference is that when phasing, inference is often made between individuals in the study. Conversely when imputing, each target individual has missing genotypes imputed from their pre-phased data using only the reference panel. Older software versions such as IMPUTE2[3] and MaCH[30] provide the possibility of phasing and imputing simultaneously. Avoiding pre-phasing has been shown to give small increases in imputation accuracy though this comes at a price of a huge increase in computation complexity[31]. Therefore, this approach is unlikely to be considered for imputation involving large target and/or reference panel sample sizes (such as those analysed here); and in particular is not possible on current imputation servers.

Imputation accuracy has not been investigated in the French population. The French population has considerable internal diversity[32,33] and does not have direct representations in panels such as 1000G, HRC, or TOPMED. Recently, 856 French individuals were whole-genome sequenced at 30–40×, this makes up the FranceGenRef panel (Labex GENMED http://www.genmed.fr/); an obvious candidate for an imputation panel for French genomes. However, as FranceGenRef is relatively small, it is unclear as to whether it will be competitive with a panel such as the HRC (38,821 individuals) for imputation. Furthermore, FranceGenRef does not include individuals from all corners of France and so may not be appropriate for imputing missing genotypes for all French genomes. In this study, we will evaluate potential approaches for both phasing and imputation of French data using either the Sanger and Michigan imputation servers or in-house phasing and imputation. We will also analyse the interplay of population structure within France and the impact that this can have on phasing and imputation accuracy.

Results

Evaluating imputation servers

Our study involves two French datasets: FrEx, a panel with exome data on 574 individuals recruited in six French cities and FranceGenRef (FGR) with whole genome sequence data on 856 individuals with ancestry in different French regions (Fig. 1). The constitutions of both datasets are described fully in the "Methods". To motivate the use of a French SSP for imputation of French genomes, an initial investigation of the performance of imputation servers for French individuals was performed. Our technique was to send sets of common variants extracted from

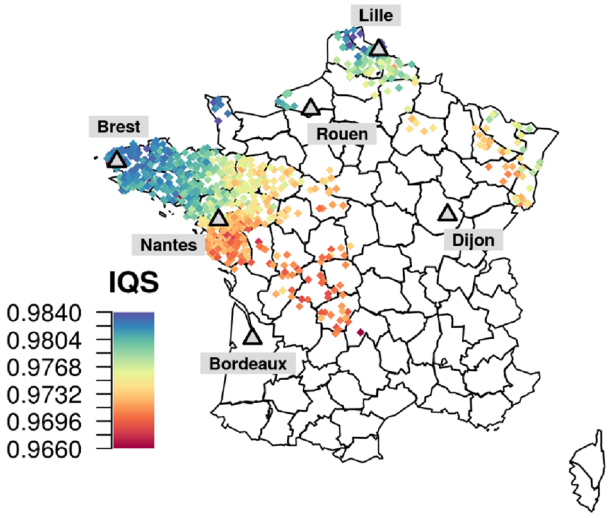

Figure 1. Geographical localities of the participants of FGR (diamonds) and FrEx (grey triangles) and individual Imputation Quality Scores. For individuals in FGR, positions were estimated as the mean latitude and longitude of the four birth-places of their four grand-parents. For FrEx, recruitment was centred around 6 cities in France (Brest, Nantes, Bordeaux. Rouen, Lille, Dijon). The individuals in FrEx are assumed to have origins close to their recruitment centre as information regarding the origins of each individual's recent ancestors were used to select participants. Individual IQS scores in FGR from imputation using the Michigan server and the HRC panel are represented by colour. Individual IQS scores range from 0.9661 to 0.9836 (scores closer to 1 represent greater imputation accuracy). IQS was measured on over 17 million variants, a difference in 0.01 between two individuals' IQS scores approximately represents a swing of 100,000 more or less correctly imputed genotypes.

FGR to two imputation servers (Michigan and Sanger) in order to be imputed with the HRC reference panel. We could then compare imputed genotypes to the sequence data in FGR. In order to assess imputation accuracy we calculated the imputation quality score (IQS)[34] per individual, assuming that the true genotypes were those from the sequence data. We established that using the Michigan server and the list of positions on the UK Biobank imputation array provided the most accurate imputation (Supplementary Materials, Supplementary Fig. 1). Furthermore, between the different imputation pipelines that we tested, there were always strong correlations between individual IQS statistics (Supplementary Fig. 2); the same individuals were always imputed the best (or worst) across our sample. This suggested that underlying characteristic of each individual were determining their individual IQS score (relative to the rest of the sample). A likely cause would be fine-scale population stratification within the sample. To show this, we plotted individual's IQS scores against their geographical location in France, and a striking pattern emerged (Fig. 1); individuals from the North and West of France were imputed with greater accuracy (top deciles of individual IQS scores). The HRC panel contains many individuals of Northern European and Britannic ancestry, this likely explains the higher imputation accuracy for individuals from the North and West of France. This suggests that the internal population structure of France may have a strong influence on the quality of imputation that can be achieved with certain reference panels. Furthermore, using only a panel such as the HRC for imputation in France could lead to an unwanted confounding between imputation accuracy and internal population stratification.

Testing different reference panels

To test the impact of using different imputation reference panels in France, we enlisted data from the French Exome Project (FrEX). Here, 550 individuals were analysed (see "Methods") who have whole-exome sequencing (WES) data and also genotype data from Illumina OmniExpressExome arrays. We took the array data for FrEx as a basis for imputation and used the WES data for calculating the accuracy of imputation (IQS scores). We sent the array data from FrEx to the Michigan imputation server to be imputed with the HRC panel using the phasing algorithm EAGLE2 and imputation software MINIMAC4. We were also able to effectively use the Michigan server to perform imputation of FrEx using the WGS data of our SSP. This was achieved through the *docker* provided by the Michigan server, allowing us to run their exact phasing and imputation pipeline using our WGS data from FGR as a reference panel whilst being required to send out our WGS panel overseas.

When using the Michigan imputation server, the HRC clearly outperformed FGR (Fig. 2, comparing far-left and far-right boxplots MICHIGAN:FGR:FGR against MICHIGAN:HRC:HRC, the notation of each strategy is Place:PRP:IRP where Place (MICHIGAN, SANGER or LOCAL) refers to where the imputation took place, PRP refers to the phasing reference panel, and IRP to the imputation reference panel). Results are split among the six French cities of FrEx. As the HRC panel contains many more individuals than FGR, far more variants can be imputed. The HRC was able to impute 12.6 million variants genome-wide with an RSQ score > 0.5 (the RSQ is the imputation quality score provided by MINIMAC4), compared to 5.2 million variants with an RSQ > 0.5

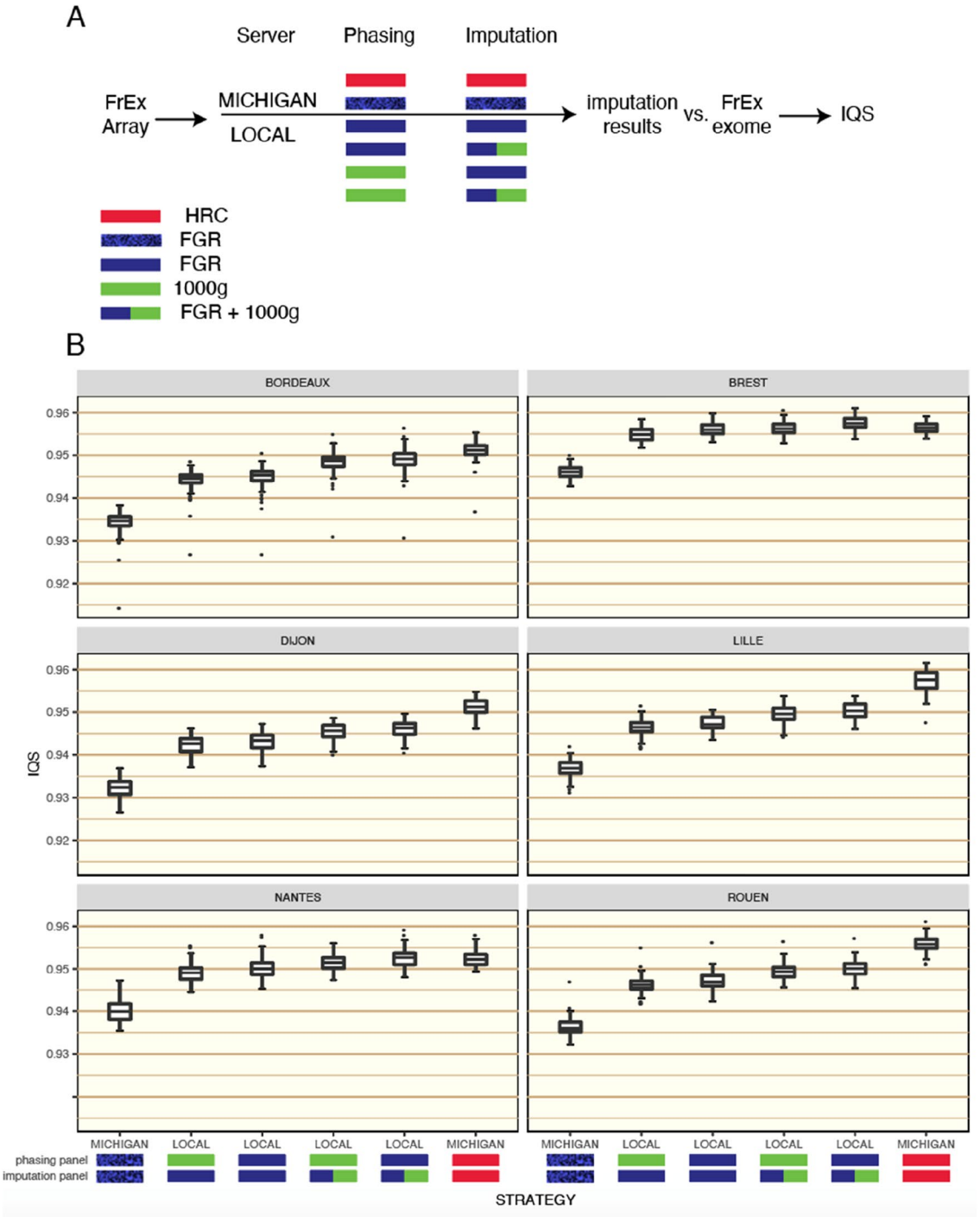

Figure 2. IQS scores for individuals in FrEx for different pipelines. Results are split between the 6 cities of FrEx. Section (**A**) depicts the different possible phasing and imputation strategies that were tested, running either on the Michigan server or locally (LOCAL) in our lab and with different combinations of phasing and imputation panels. Section (**B**) gives boxplots of individual level IQS scores for each strategy. Of the 550 individuals analysed, 89 are from Bordeaux, 96 from Brest, 87 from Dijon, 93 from Lille, 90 from Nantes, and 95 from Rouen.

by FGR. This is due to the fact that our SSP only contains variants with an observed Minor Allele Count (MAC) ≥ 5 in FGR. The superiority of the public panel over the SSP contrasts against many of the results presented in our literature review where SSPs were regularly shown to be the most effective imputation panels. Possible explanations include the small size of FGR and that previous studies had often compared SSP imputation against imputation using the 1000G. But also, this could be partly due to the use of the imputation servers. Examining the described pipeline of the Michigan server, we observed that the phasing step could be working at a disadvantage to the SSP strategy. EAGLE2 is able to very efficiently take advantage of the HRC as a huge phasing reference

panel (PRP). However, it has been shown to be less optimal for taking advantage of within-sample phasing[19], relying more on comparing each haplotype separately to the PRP, which could have had an impact during both the phasing of FrEx and FGR when using the Michigan imputation server.

We constructed our own phasing-imputation pipeline to assess the impact of using the imputation servers (description in "Methods"). To further investigate the impact of the pre-phasing step, we phased FrEx using SHAPEIT4 with either FGR or 1000G as a PRP. By using IMPUTE2 with the merge-ref-panel option, we were able to use a combined reference panel of the 1000G and FGR without having to restrict to a common list of variants. When the PRP and IRP were FGR, the imputation quality was improved substantially by using our in-house imputation pipeline compared to the results achieved with the Michigan server via the docker (LOCAL:FGR:FGR against MICHIGAN:FGR:FGR). In the cities of Brest and Nantes, SSP-based imputation became competitive with the server-based approach using the HRC panel. These two cities lie in the regions that are most well represented by FGR. In terms of per-variant accuracy, 42.3% of all 125,442 variants evaluated were better imputed (had a higher per-variant IQS score) with the LOCAL:FGR:FGR strategy compared to MICHIGAN:HRC:HRC. However, this rises to 58.1% if we considered only individuals from Brest and Nantes. For all six cities, a marginal improvement in imputation accuracy was gained by including the 1000G in the reference panel (LOCAL:FGR:FGR against LOCAL:FGR:FGR + 1000G).

Regarding the key comparison of imputation results between LOCAL:FGR:FGR and MICHIGAN:HRC:HRC, we found that the advantage brought by the HRC panel over FGR was particularly evident for rare variants (see Supplementary Fig. 3 where the mean IQS scores per-variant were compared in detail between these two strategies for different categories of MAF). Only in Brest and Nantes did the two pipelines perform similarly. However, this analysis does not tell the full story as IQS was calculated on only variants that can both be imputed by the HRC and our FGR. Indeed, this ignores the existence of potentially-population specific variants that cannot be imputed by panels such as 1000G or the HRC. In Supplementary Fig. 4, we show the proportions of variants observed in FrEx that are also observed in the other three datasets (FGR, 1000G, HRC) pertinent to this study. For the rarer variations in FrEx, large proportions of variants are not observed in all three of FGR, 1000G, and HRC including many that are only observed in FrEx and FGR. Over 50% of singletons and over 10% of doubletons in FrEx were not observed in any reference panel. Supplementary Fig. 4 highlights the importance of including an SSP in order to impute potentially population specific variants (as seen by the change in the proportion of variants only seen in FrEx and FGR). Supplementary Fig. 4 also shows how using an SSP in conjunction with large public panels will capture variants that are rare in the study population and so might be missed by a relatively small SSP but found in a large cosmopolitan panel.

Investigating haplotype sharing between target and reference individuals

We investigated the composition of the estimated haplotypes in FrEx to reveal where our imputation pipeline was most accurate. As imputation is haplotype based, the haplotype-estimation pre-phasing step clearly has an impact on the quality of imputation. We postulated that using an SSP as PRP could be particularly beneficial as haplotype estimation could be improved and imputation using the same SSP would thus be facilitated. This we demonstrated by comparing phasing-imputation run LOCAL:FGR:FGR against LOCAL:1000G:FGR. To approximate the accuracy of phasing, we applied the principle of phasing uncertainty. This involved repeatedly performing phasing using different random seeds and evaluating the stability of the final estimated haplotypes. Using such a method, we could have an approximation of the Switch Error Rate (SER) of the haplotypes constructed (see "Methods"). We refer to the two different phasing outputs involved in these two pipelines as FGR-PRP and 1000G-PRP.

A correlation is observed between individual SER and IQS (Supplementary Fig. 5—left panel). Improvement on both SER and IQS is observed with the use of FGR-PRP compared to 1000G-PRP and this especially true for individuals from the two cities of Brest and Nantes compared to the other four cities of FrEx (Supplementary Fig. 5—right panel). As Brest and Nantes are located in the French regions that are the best represented in FGR, FrEx individuals from those two cities likely exhibit greater haplotype-sharing with FGR than FrEx individuals from the other four cities.

Furthermore, we explicitly looked for regions of haplotype-sharing between FrEx individuals and FGR individuals by estimating IBD segments using RefinedIBD[35]. We clustered individuals in FGR based into 12 groups (see "Methods") using finestructure[36]. These coincided with different geographical regions of France (Fig. 3). We calculated the total IBD shared between each city in FrEx and each cluster of reference haplotypes in the two different phasing scenarios described above: FGR-PRP and 1000G-PRP. Far greater total shared IBD was estimated between Individual's from Brest and Nantes and FGR under FGR-PRP. What is more, the increase in detectable haplotype-sharing pertained largely to shared segments of length greater than 3cM between individuals of Brest and Nantes in FrEx and individuals in FGR from the regions of France close to Brest and Nantes. For the detection of shorter segments, the choice of PRP had less impact.

To demonstrate the interplay of the estimation of shared IBD segments between the target and reference panel and imputation accuracy, we examined the imputation of rare variants (determined by minor allele frequency less than 0.01 in FrEx) for the FrEx individuals from Brest. Given that such variants can be imputed by our FGR panel (otherwise they cannot be analysed for imputation accuracy) we can assume that we are focusing on variants with a frequency in France roughly between 0.001 and 0.01. Imputation accuracy was evaluated in two variant sets: those inside long IBD segments and those outside long IBD segments. Specifically, for each rare-variant, we tabulated the imputed dosages of heterozygous sites observed in the individuals of Brest inside and outside long IBD segments (> 3cM) shared with Clusters 1 and 2 of FGR that cover the Finistere department where Brest is located. Histograms of these imputed dosages (which should be equal to 1 if the imputation has been successful at a heterozygous site) are presented in Supplementary Fig. 6 for two imputation strategies: LOCAL:FGR:FGR and

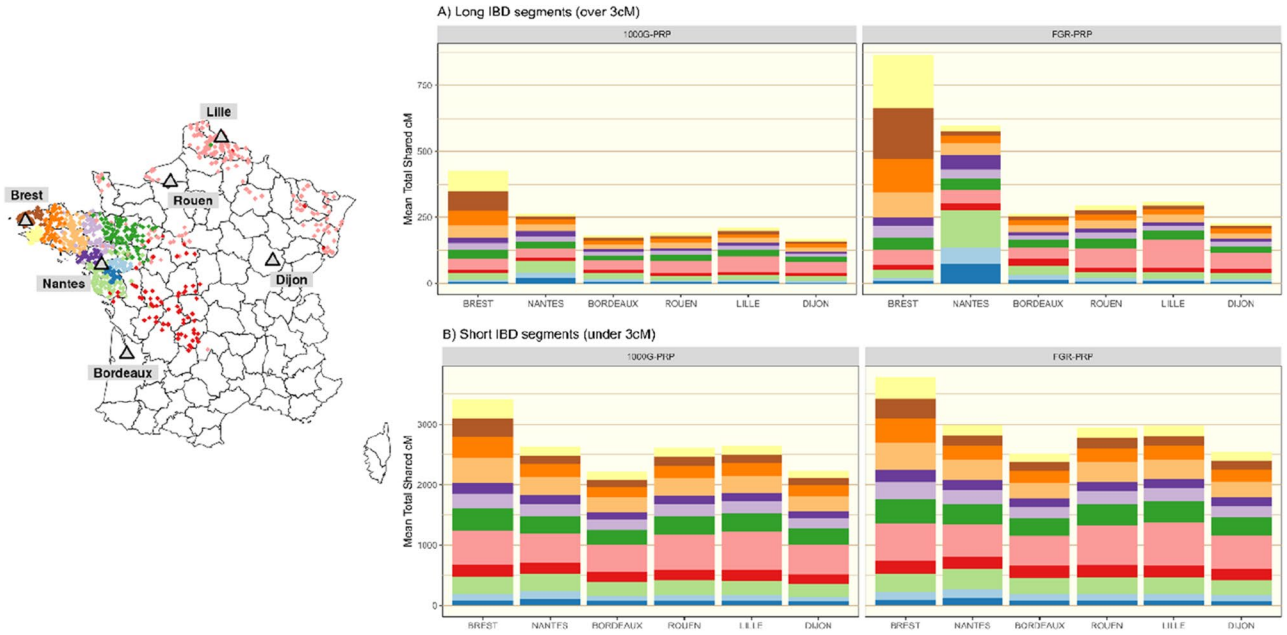

Figure 3. Left: A map of France with the 850 individuals of FGR coloured by the 12 haplotype-sharing clusters identified with finestructure. The 6 cities of FrEx are highlighted (grey triangles). Right: Haplotype sharing between individuals from the different FrEx cities and individuals from FGR. For each city in FrEx, the mean total length of shared IBD segments between each individual in FrEx and individuals in FGR in each of the 12 clusters of FGR detected with finestructure is shown. Colours correspond to those in the left panel. The PRP used in the phasing step was either 1000G (panels marked 1000G-PRP—left column) or FGR (panels marked FGR-PRP—right column) and IBD segments were split into long segments over 3 cM ((**A**)—top row) or small segments under 3 cM ((**B**)—bottom row).

MICHIGAN:HRC:HRC. There was a clear higher proportion of correctly imputed genotypes for variants within IBD segments compared to variants outside IBD segments under the LOCAL:FGR:FGR imputation strategy but not under the MICHIGANC:HRC:HRC strategy. This shows how imputation using a local reference panel can be expected to improve accuracy, and in particular for the rare local variants that are expected to lie on long IBD segments. This is because rare variants are expected to be younger than common variants and thus shared haplotypes around rare variants are longer.

Pragmatic imputation strategy for France

We have thus far demonstrated that by optimising certain stages of our phasing and imputation strategy we could achieve a comparable imputation quality using FGR compared to imputation servers that have access to the HRC panel. However, this was only the case for the individuals of Brest and Nantes in FrEx. Furthermore, it was also observed through analysing haplotype-sharing that even for Brest and Nantes the SSP imputation was not performing uniformly and that its accuracy would vary in different genomic regions. Ideally, a combined panel of FGR and HRC could be used. Without a straightforward path to this solution, one possibility is to attempt to combine imputed data from two different imputation runs in the spirit of the PedPop method put forward by Saad & Wijsman[37]. This simply involves merging two or more imputation outputs such that all variants imputed by any of the strategies are present. For variants that are imputed by multiple imputation strategies, a simple 'most confident vote' selection is used (see "Methods"). We combined imputation using the MICHIGAN:HRC:HRC with our own LOCAL:FGR:FGR imputation in such a manner (see "Methods") and the overall improvement to the imputation accuracy was substantial (Fig. 4). This hybrid imputation coming from this combination is denoted as HYB.

To illustrate how the HYB method was improving the imputation, for each individual, IQS scores were calculated for two sets of variants: One set where there was agreement between the two imputation strategies regarding the most likely genotype (Accord), and a second where there was disagreement (Discord). The average percentages of genotypes in agreement for each individual for the six cities of FrEx were: Bordeaux 98.5%, Brest 98.7%, Dijon 98.5%, Lille 98.6%, Nantes 98.6%, and Rouen 98.6%. Overall, agreement between HRC and FGR imputation corresponded with the correct genotype being assigned the highest probability 98.5% of the time. Hence, agreement between HRC and FGR was a reliable indication of accurate imputation. Choosing the set of imputation probabilities with the greatest top probability in the case where the two imputation runs are in Accord will therefore produce a dosage closer to the true genotype for the majority of cases. This gave a significant boost to the IQS statistics (Fig. 4, Supplementary Fig. 7) and would also lead to an increase in power for prospective association tests. In Supplementary Fig. 7, we also observe that this improvement afforded by the HYB strategy was present for both rare and common genetic variants.

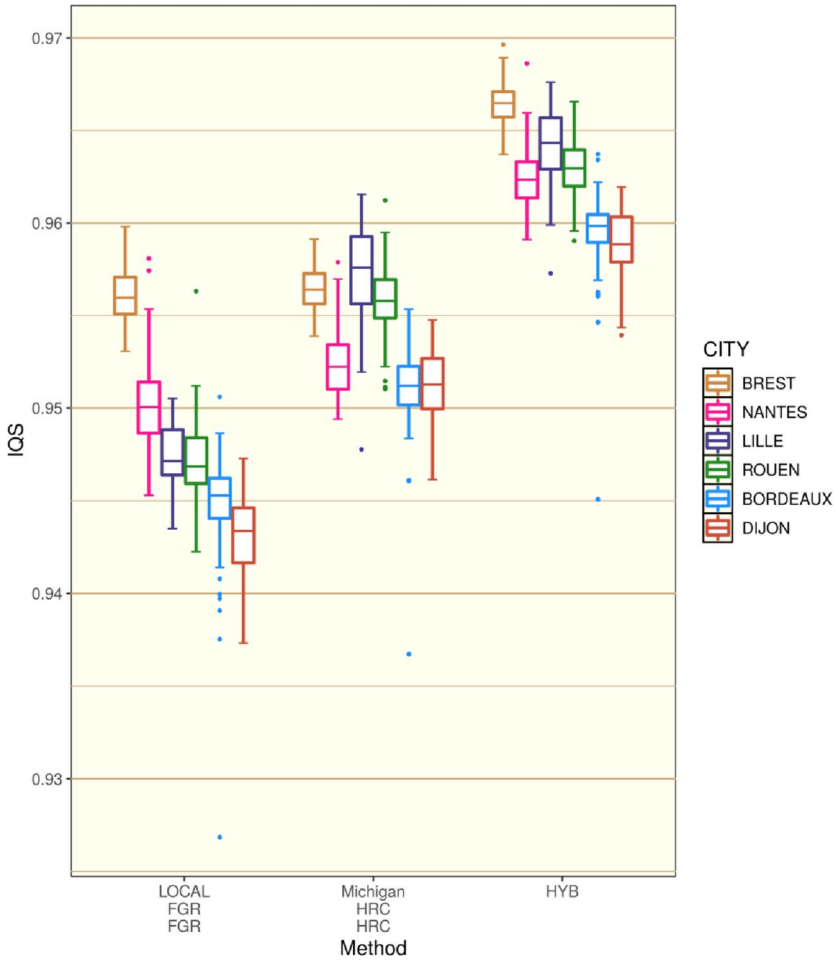

Figure 4. Individual IQS scores for the hybrid (HYB) imputation strategy split by city in FrEx against the previously calculated scores for strategies LOCAL:FGR:FGR and MICHIGAN:HRC:HRC described in Fig. 2.

The two imputation runs were not in agreement (Discord) for 1.5% of all genotypes analysed (125,442 exonic variants). This corresponds to approximately 1800 genotypes per individual. In this set, the percentage of variants where HRC imputation was correct was the following for the different cities: Bordeaux 64%, Brest 59%, Dijon 66%, Lille 69%, Nantes 61%, and Rouen 67%. The hybrid imputation strategy chose the correct genotype more often than not (Bordeaux 68%, Brest 68%, Dijon 67%, Lille 70%, Nantes 68%, and Rouen 69%). Hence, the HYB strategy coped well with disagreement between the two pipelines. In Brest and Nantes, HYB even provided an improvement in imputation in the Discord set of variants (Supplementary Fig. 7); both for rare and common variants. Therefore, a simple combination of in-house and server-based imputation could provide a pragmatic and most effective imputation strategy.

Finally, we calculated IQS scores per gene under the strategies LOCAL:FGR:FGR, MICHIGAN:HRC:HRC, and HYB to see if any particular regions of the genome had a noticeable differential in imputation quality between strategies (Supplementary Fig. 8a–c). Globally, genes that were imputed with lower scores by MICHIGAN:HRC:HRC were also imputed with lower scores by LOCAL:FGR:FGR. We could also again observe the overall trend of better imputation across all of FrEx given by the MICHIGAN:HRC:HRC; but also that many genes were imputed with a higher IQS by LOCAL:FGR:FGR; as well as the effectiveness of the HYB strategy. The differences in IQS per-gene are generally very small, with only two genes standing out: TCF7L2 and HLA-DQB1. Further investigation of the imputation of these genes are detailed in Supplementary Fig. 9a,b and Supplementary Table 1a,b. The difference for TCF7L2 appears incidental and likely linked to data quality in the reference panels, while HLA-DQB1 could hint at the added importance of SSPs for imputing complex regions such as the major histocompatibility complex.

Discussion

Many factors affect the accuracy of phasing and imputation, though most focus has been put on the size and composition of the haplotype reference panel. Furthermore, imputation has shifted from an operation performed in-house using publically available data and free academic software to an operation that is increasingly performed at distance using publically accessible imputation servers. The main motivations for using external imputation servers is their convenience and the access they provide to the largest (and hence most powerful for

imputation) public reference panels. However, certain compromises currently have to be made if one decides to use an imputation server. Importantly, there is not the possibility to combine an SSP with public reference panels, and there is less control of haplotype phasing. In this study we have shown that such considerations can make a significant difference for imputation quality. Hence, the optimisation of an imputation strategy goes far beyond simply choosing the largest available reference panel. If an SSP is to be used, we suggest that at this current time it is still preferable to perform phasing and imputation in-house rather than turning to online imputation servers.

We decided that, for this study, it was not necessary to test the very recently developed TOPMED server. This was for three reasons, firstly the TOPMED panel is aligned to genomic build 38 and it was beyond the scope of this study to re-call the FranceGenRef and FrEx datasets that are both aligned to build 37. Secondly, whilst the TOPMED reference panel is significantly larger than the HRC, it contains a comparable number of individuals of European ancestry. The TOPMED panel has been shown to greatly outperform the HRC for imputation of individuals of Latin American and African ancestry[8] but would not provide such a significant improvement in accuracy compared to the HRC for French individuals. Finally, as the TOPMED server only became available during the course of our study, its use had not been specified in our design and hence prohibited without making specific re-applications for data usage which given the first two reasons may not be well justified in this study.

It would have been possible to gain permission from the European Phenome-Genome Archive to download a part of the HRC panel (https://ega-archive.org/studies/EGAS00001001710). This option was used in one simulation study[19], where this subset of the HRC was combined with an SSP for evaluation of imputation in an isolated population. Using this subset of the HRC has recently been shown to be effective for imputation in conjunction with SSPs by Quick et al.[38] We decided not to pursue this avenue in this study for two reasons. Firstly, having to download this subset of the HRC and then perform imputation using IMPUTE2 and the merge-ref-panel option is very computationally heavy, requires a lot of storage space, and requires the submission of a specific request for the HRC subset. Hence, this is a strategy that may not be suitable for all researchers and so is not a realistic recommendation to make. Secondly, in this study we wished to focus on the pros and cons of the choice of using the relatively easy server-based approach against in-house imputation. Hence, to test the HRC we preferred to access it through the server; furthermore, this allowed us to test the HRC panel in its entirety.

Our results chime with previous results regarding the benefits of local reference haplotypes[11,12,14–16,18,20]. However, by including the complete HRC panel in our study, we showed certain limitations to SSP-based imputation. This was possible by investigating the fine genetic structure in France and its impact on the imputation of French genomes. Our SSP was successful in improving imputation beyond the possibilities of the HRC panel but only for target individuals that were from the regions that were densely covered in FGR (Brest and Nantes). In the other four cities, the HRC clearly afforded higher accuracy. We note that evaluating imputation accuracy per-individual rather than the more commonly used per-variant calculation was important in uncovering such patterns. Indeed, imputation in France using only the HRC led to a clear gradient of imputation quality in FGR; with individual IQS scores varying from 0.984 to 0.966. As IQS was measured on over 17 million variants in Fig. 1, a difference in 0.01 between two individuals' IQS scores approximately represents a swing of 100,000 more or less correctly imputed genotypes. This further motivates the use of local reference haplotypes to avoid the potential of introduced bias as panels such as the HRC will likely provide stronger imputation for individuals from the North and West of France.

On the other hand, it should also be noted that a certain limitation of this study is that we have imputed individuals (either in FrEx or in FGR) with a specific geographical based recruitment. This could have led to a slightly artificial advantage for the SSP over the HRC in this study. For other imputation studies in French populations, for example of patient data, we should expect diversity to be required in the pool of reference haplotypes and hence the importance of utilising large cosmopolitan panels; ideally alongside an SSP.

Furthermore, whilst the whole-exome sequencing data of FrEx and the whole-genome sequencing data of FGR were not generated at the same time period nor on the same sequencers, they were generated by the same centre and so if any batch effects are present between the three key datasets in this study (HRC, FGR, and FrEx) it is likely that the one dataset that differentiates most from the others would be the HRC. The correlations of our results with geographical data should clearly demonstrate that our conclusions are not largely driving by such effects but this could represent another subtle source of advantage for the SSP in this study. However, this is not unique to this study, the same could be said of all studies that have compared SSPs with public panels and in fact represents an additional argument for using SSPs as in this case one can have greater control on the formation and quality of one's reference panel.

We have also demonstrated that the benefits to SSP-based imputation coincided with the sharing of haplotypes between the target and reference individuals. This highlights the importance of optimising haplotype estimation, an area we have concentrated on in this study. Indeed, the imputation of rare-variants would likely benefit noticeably from greater accuracy in the phasing of the SSP. Further improvements to phasing performance could also be sought either through read-based phasing algorithms[39] or through consensus based phasing[40,41]. Another promising approach is to replace array based genotyping with low-pass sequencing[42,43]. In future work, it will also be of interest to explore the impact of including study-specific haplotypes in the reference panel for the imputation of certain complex regions. Indeed, we were able to have an indication that this could be a valuable avenue for further investigation by observing that the imputation of variants in HLA-DQB1 was much improved when imputing with FGR compared to the HRC.

The FGR panel used here contained 850 individuals. The largest prospective SSP for France, the POPGEN project of the French medical genomics initiative[44,45] will contain roughly 4,000 individuals. Joining the dots, the significant improvements to the estimated SER and IQS for the individuals of Brest and Nantes would suggest that imputation could be highly accurate for individuals from across France using this novel reference panel. Particularly as we observed that the HRC panel performed less well for individuals from towards the South of France, an area that will be better represented in POPGEN. However, there may still be room to incorporate imputed

variants from imputation servers due to the undeniable power of huge public reference panels for imputation; in particular, for rare variants. Rare variants that arrived recently in the population can be expected to have a high level of IBD-sharing[46] lying within long shared segments. Such variants should be expected to be well imputed using an SSP. This does not hold for older variants that are observed to be rare due to many generations of purifying natural selection[47]; for such variants, the breadth provided by large cosmopolitan panels may provide the best imputation. The POPGEN dataset will cover the whole of France but realistically may not provide a significantly denser coverage than what is given by FGR for the regions of Bretagne and Pays-de-la-Loire (the regions that surround Brest and Nantes, respectively). Combining panels allows for a greater number of overall variants to be imputed as panels will not have coinciding lists of observed variants; each will have a set of variants only observed in that panel. Without the current possibility to combine an SSP with the full HRC or TOPMED, we have put forward a simple pragmatic approach for combining in-house and server-based imputation and showed that this can give more complete and accurate imputation. Shortly after we completed this study, a formalised approach to combine imputation runs using different reference panels, known as meta-imputation, was developed[48]. This represents another possibility for researchers to combine inference from an SSP and an imputation server and would likely produce similar results of a similar trend to our simple combination approach. We would certainly recommend its use as it we feel it responds to a clear need for a tool to combine separate imputation runs using different panels in the case where it is not possible to otherwise combine the panels. It was beyond the remit of this study to go back and reproduce imputation results using the specialised functionalities of the Michigan imputation server required for meta-imputation.

Conclusion

This study serves to demonstrate exactly why and to what extend SSPs can improve imputation in a European population using the example of France and by observing various correlations between fine-scale population structure and imputation precision. Without an SSP available, external imputation servers allowing access to huge cosmopolitan reference panels will be an appropriate choice, though as demonstrated here this comes with certain inconveniences and we were also able to show how careful control of an in-house imputation pipeline can improve results. Specifically, pre-phasing the data with the SSP conferred a significant advantage for the imputation of rare variants. As general advice to those wishing to perform imputation, it appears that one should search for a way to have as large a reference panel as possible which includes some haplotypes that are as specific as possible to one's target data; down to a very fine scale of population structure. When it is not possible to combine imputation panels, we have set out a simple pragmatic approach to approximate such a joint imputation.

Methods

The two sequencing datasets used in this study, FGR and FrEx were prepared using VCFprocessor[49] using in-house settings (described in Supplementary Material). FGR includes 856 individuals selected using strict criteria on ancestral places of birth. Specifically, individual were only sequenced if their four grand-parents were known to have all been born within no more than 30 km of each other. By taking the barycentre of the co-ordinates of all 4 grand-parents, we approximated the ancestral location of each individual in FGR. The recruitment of France-GenRef (http://www.genmed.fr/index.php/en/recherche/projets/france-genref) is described in full elsewhere[33]. Individuals included in the FREX project (https://www.france-genomique.org/databases/frex-the-french-exome-project-database/?lang=en) are 574 healthy individuals sampled in 6 different regions of France around 6 cities (Bordeaux, Nantes, Brest, Rouen, Lille and Dijon)[50]. All individuals signed informed consent for genetic studies at the time they were enrolled and had their blood collected. Neither phenotypic nor clinical data were collected. Declaration and ethical approval for the present study was accorded by the Ministry of Research; specifically, from the local Committee of Protection of Persons (CPP in Nantes), the Advisory Committee on Information Processing for Health Research, and the National Commission on Informatics and Liberty. The CPP in Nantes represents a research ethics committee, who gave approval to the present study. All methods applied in the course of the present study are in accordance with the relevant guidelines and regulations.

The FrEx data analysed here comprises 557 (out of 574) individuals who are those who have both genotyping (Illumina OmniExpressExome arrays) and WES data. A total of 824,279 variants are found in the WES data after QC. Our SSP was built with 850 individuals with WGS data from FGR. Six individuals were removed from the reference panel for having an autosomal missingness above 2% as well as removing one individual from each pair of relatives in order to facilitate population structure analyses of the reference panel. Seven individuals were removed from the target panel (FrEx) as certain individuals were also present in FGR or had a high autosomal missingness in the whole-exome sequencing data. Finally, 550 individuals from FrEx were imputed using 850 individuals from FGR. We kept only variants with a minor allele count above 5 for the creation of an imputation panel. For both datasets, we have approximated geographical locations for each individual.

Imputation quality was measured using IQS[34] calculated per-individual across various sets of genetic variants. This imputation score measures the concordance between the truth set and the posterior imputation probabilities whilst taking into account the expected level of concordance by chance. When splitting results by minor allele frequencies (MAFs), we used the naive MAF estimates from FrEx and results are shown either for rare variants ($MAF < 0.01$) or non-rare variants ($MAF \geq 0.01$). For the analyses pertaining to Fig. 1, IQS was calculated across 17,192,131 variants observed in FGR and imputed by the HRC. All other analyses focus on FrEx and hence IQS was calculated on a set of 125,442 exonic variants across the 22 autosomal chromosomes that could be imputed with the constructed imputation panel of FGR (i.e. variants that passed the quality control measures in FGR and thus had a minor allele count superior to 5). Hence the raw IQS values are not directly comparable between Fig. 1 and all other Figures. To describe an imputation strategy, we use the following notation: Place:PRP:IRP

where Place refers to the location of the imputation (either Michigan imputation server, the Sanger server, or in-house at LOCAL), PRP refers to the phasing reference panel, and IRP refers to the imputation reference panel.

In order to use FGR as a reference panel for our-in house LOCAL pipeline, it was phased using SHAPEIT4[26] and the 'sequencing' option to optimise the algorithm for WGS data. Furthermore, in an effort to improve the phasing performance of SHAPEIT4, we specified the following iteration programme: '8b,1p,1b,1p,1b,1p,1b,1p,15m'. Conversely, when using the Michigan imputation server with the FGR panel for the Michigan:FGR:FGR strategy, the phasing of FGR was performed using the Michigan server and the phasing-only functionality.

FrEx was phased with SHAPEIT4 and imputed using IMPUTE2. The choice of IMPUTE2 may seem questionable given the availability of more recent version such as IMPUTE5[5] as well as competing software such as MINIMAC4 or BEAGLE5[51]. IMPUTE2 was chosen purely due to the availability of the merge-ref-panel option, allowing for a combined panel of the 1000G and FGR to be used. The importance of this option is demonstrated by the observation that 0.54% of all variants in the SSP we constructed from FGR are not present in the 1000 Genomes Project. Without this cross-imputation option, these SSP-specific variants would be lost. Given that software cited above rely on similar methodology and have similar performances[5] (with more recent versions admittedly bringing incremental improvements), we felt that this was a suitably choice for putting forward an imputation strategy involving a SSP. The improvements that have been made to imputation software beyond IMPUTE2 are concerned with the ability to leverage vast reference panels such as the HRC or TOPMED. Our imputation strategy LOCAL:FGR:FGR + 1000G involves a combined reference panel of only 6708 haplotypes and so it is reasonable to employ IMPUTE2 in this scenario. However, using IMPUTE2 with a combination of the HRC and our SSP would encounter excessive runtime.

To approximate Switch Error Rate (SER) without knowing the true phase in FrEx, we simply ran SHAPEIT4 21 times using 21 different random seeds. Across the 21 repetitions, and for each pair of adjacent heterozygous genotypes, we assumed that the phase configuration assigned by the majority of random seeds was the correct phase; this allowed us to estimate SER in each seed before finally calculating an average SER across all 21 replicates.

IBD segments in FGR were estimated using RefinedIBD[35]. The resultant matrix of IBD sharing between individuals was then treated as matrix of 'chunk lengths' and supplied to finestructure[36] to establish 12 groups of individuals likely having similar genetic backgrounds. As described in Bycroft et al.[52], using a chunk-length matrix necessitated the estimation of the 'c-factor' parameter from within the sample, for which we followed the instructions given in the supplementary material of Bycroft et al.[52]. The choice of 12 groups was made by inspection and in order to give a set of easily interpretable groups. Up until 12 groups, each cluster identified corresponded to over 10 individuals and to specific geographical region. Beyond 12, groups become small and lacked easily interpretable links to geographical regions. We note that finestructure was unable to distinguish the individuals in FranceGenRef from the North and the East of France. We attribute this to the fact that we don't have a sufficient sample size in these regions and that, as observed by the Eigen decomposition of the IBD sharing matrix (see Supplementary Fig. 10), the most evident sources of variation in the data come from the proximity of individuals to the source populations of the Brittany region and the Pays-de-la-Loire region. As FranceGenRef does not represent a fair sampling of the French population, it is not surprising that the finestructure analysis largely reflects only the variation in the West of France; where we have by far the most individuals. However, the clusters presented here are still relevant for the West of France and are instructive in showing the potential for extensive fine-structure in the French population.

To combine the imputation pipelines for the HYB imputation. We simply compared the maximal probabilities for each pair of genotype from the pipelines MICHIGAN:HRC:HRC and LOCAL:FGR:FGR. For example, if the posterior imputation probabilities for a genotype of a given individual are $I_A = (0.95, 0.05, 0.00)$ & $I_B = (0.85, 0.15, 0.00)$ from imputation strategies A and B, respectively, then only the posterior probabilities I_A will be retained as they are the most certain. The concept that the more certain a set of genotype probabilities the more accurate the imputation is well known and underpins the calculation of most imputation quality metrics[53]. Inspection suggested that when the two maximal probabilities were very close, little could be gained by selecting the trio with the highest probability. Furthermore, due to the differences in imputation software (MINIMAC4 against IMPUTE2), we often saw that the maximal probability of LOCAL:FGR:FGR (denoted as P_{max}^{FGR}) was larger than the its counterpart P_{max}^{HRC} but only by an order of 10^{-2}. We found that an effective combination method was to select the imputation trio of posterior probabilities from LOCAL:FGR:FGR if and only if $P_{max}^{FGR} > P_{max}^{HRC} + 0.05$, hence giving priority to HRC when the P_{max}^{FGR} and P_{max}^{HRC} were very close. This rule was used to form the HYB imputation presented in the Results section. Variants were split into groups denoted as Accord and Discord, based on whether P_{max}^{FGR} and P_{max}^{HRC} indicated the same genotype or not.

Data availability

Data from the FranceGenRef panel will be submitted to the French Centralized Data Center of the France Medicine Genomic Plan that is under construction. Enquiries for the use of this data can be addressed to GENMED LABEX (http://www.genmed.fr/index.php/en/contact). Summary information for the FrEx dataset is available at http://lysine.univ-brest.fr/FrExAC/, those wishing to access the data on a collaborative basis should contact Emmanuelle Génin (emmanuelle.genin@inserm.fr).

Received: 28 July 2023; Accepted: 13 December 2023
Published online: 03 January 2024

References

1. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

2. The Haplotype Reference Consortium *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet. 48*, 1279–1283 (2016).
3. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 Genes Genomes Genetics (Bethesda)* **1**, 457–470 (2011).
4. Zhang, P., Zhan, X., Rosenberg, N. A. & Zöllner, S. Genotype imputation reference panel selection using maximal phylogenetic diversity. *Genetics* **195**, 319–330 (2013).
5. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the positional burrows wheeler transform. *bioRxiv* https://doi.org/10.1101/797944 (2020).
6. Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* **84**, 235–250 (2009).
7. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
8. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLOS Genet.* **15**, e1008500 (2019).
9. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLOS Genet.* **10**, e1004528 (2014).
10. Kimura, M. & Ohta, T. The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199 (1973).
11. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **22**, 1321–1326 (2014).
12. Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).
13. Zhou, W. *et al.* Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genet. Epidemiol.* **41**, 744–755 (2017).
14. Yasuda, J. *et al.* Regional genetic differences among Japanese populations and performance of genotype imputation using whole-genome reference panel of the Tohoku Medical Megabank Project. *BMC Genom.* **19**, 551 (2018).
15. Cocca, M. *et al.* A bird's-eye view of Italian genomic variation through whole-genome sequencing. *Eur. J. Hum. Genet.* **28**, 435–444 (2020).
16. Kals, M. *et al.* Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses. *bioRxiv* https://doi.org/10.1101/579201 (2019).
17. Joshi, P. K. *et al.* Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies. *PLOS ONE* **8**, e68604 (2013).
18. Pistis, G. *et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: Implications for cost-effective study designs. *Eur. J. Hum. Genet.* **23**, 975–983 (2015).
19. Herzig, A. F. *et al.* Strategies for phasing and imputation in a population isolate. *Genet. Epidemiol.* **42**, 201–213 (2018).
20. Surakka, I. *et al.* Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome Res.* **20**, 1344–1351 (2010).
21. Zeggini, E. Next-generation association studies for complex traits. *Nat. Genet.* **43**, 287–288 (2011).
22. Molnár-Gábor, F. *et al.* Bridging the European data sharing divide in genomic science. *J. Med. Internet Res.* **24**, e37236 (2022).
23. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
24. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
25. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
26. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
27. The UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82 (2015).
28. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
29. Chou, W.-C. *et al.* A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Sci. Rep.* **6**, 39313 (2016).
30. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
31. Roshyara, N. R., Horn, K., Kirsten, H., Ahnert, P. & Scholz, M. Comparing performance of modern genotype imputation methods in different ethnicities. *Sci. Rep.* **6**, 34386 (2016).
32. Saint Pierre, A. *et al.* The genetic history of France. *Eur. J. Hum. Genet.* **28**, 853–865 (2020).
33. Alves, I. *et al.* Genetic population structure across Brittany and the downstream Loire basin provides new insights on the demographic history of Western Europe. *bioRxiv* https://doi.org/10.1101/2022.02.03.478491 (2022).
34. Lin, P. *et al.* A new statistic to evaluate imputation reliability. *PLOS ONE* **5**, e9697 (2010).
35. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
36. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLOS Genet.* **8**, e1002453 (2012).
37. Saad, M. & Wijsman, E. M. Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees. *Genet. Epidemiol.* **38**, 579–590 (2014).
38. Quick, C. *et al.* Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. *Genet. Epidemiol.* **44**, 537–549 (2020).
39. Bansal, V. Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes. *Bioinformatics* **35**, i242–i248 (2019).
40. Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K. & Goudey, B. Exploring effective approaches for haplotype block phasing. *BMC Bioinform.* **20**, 540–540 (2019).
41. Al Bkhetan, Z., Chana, G., Ramamohanarao, K., Verspoor, K. & Goudey, B. Evaluation of consensus strategies for haplotype phasing. *bioRxiv* https://doi.org/10.1101/2020.07.13.175786 (2020).
42. Wasik, K. *et al.* Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genom.* **22**, 197 (2021).
43. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
44. Lethimonnier, F. & Levy, Y. Genomic medicine France 2025. *Ann. Oncol.* **29**, 783–784 (2018).
45. Lévy, Y. Genomic medicine 2025: France in the race for precision medicine. *Lancet* **388**, 2872 (2016).
46. Albrechtsen, A., Moltke, I. & Nielsen, R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* **186**, 295–308 (2010).
47. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases?. *Am. J. Hum. Genet.* **69**, 124–137 (2001).
48. Yu, K. *et al.* Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet.* **109**, 1007–1015 (2022).

49. Ludwig, T. E., Marenne, G. & Génin, E. VCFProcessor. http://lysine.univ-brest.fr/vcfprocessor/index.html. Accessed 08/10/2020. (2020).
50. Génin, E. *et al.* The French Exome (FREX) Project: A Population-based Panel of Exomes to Help Filter Out Common Local Variants. *The 2017 Annual Meeting of the International Genetic Epidemiology Society* **41(7)**, 691–691 (2017).
51. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
52. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat. Commun.* **10**, 551 (2019).
53. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

Acknowledgements
We would like to thank all of the members of the FranceGenRef cohort and the FrEx cohort for their participation in this study.

Author contributions
A.F.H. and E.G. designed the study. A.F.H. wrote the manuscript and carried out all major analyses. L.V.S. assisted in the analyses, notably in relation to the use of the external imputation servers. All analyses in the study were discussed and decided upon by A.F.H., L.V.S., and E.G. C.D., R.R., J.F.D., and E.G. contributed to data production for both the FrEx and FranceGenRef datasets. All authors participated in the final redaction of the manuscript.

Funding
This work was supported by LABEX GENMED funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (grant number ANR-10-LABX-0013), and by the French regional council of Pays-de-le-Loire (VaCaRMe project). This work was also supported by the POPGEN project as part of the Plan Médecine Génomique 2025 (FMG2025/POPGEN) and by Inserm cross-cutting project GOLD. Funding for exome sequencing in the FrEx project was obtained from France Genomique 2013 call for sequencing. This study also received financial support the Agence Nationale de la Recherche in France (ANR; FROGH, ANR-16-599-CE12-0033).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-49931-3.

Correspondence and requests for materials should be addressed to A.F.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The FrEx Consortium

Principal Investigators

Emmanuelle Génin[1,2], Dominique Campion[7], Jean-François Dartigues[8], Jean-François Deleuze[5,6] & Jean-Charles Lambert[9] & Richard Redon[4]

Collaborators

Bioinformatics

Thomas Ludwig[1,2], Benjamin Grenier-Boley[9], Sébastien Letort[1,2], Pierre Lindenbaum[4], Vincent Meyer[5] & Olivier Quenez[7]

Statistical genetics

Christian Dina[4], Céline Bellenguez[9], Camille Charbonnier-Le Clézio[7] & Joanna Giemza[4]

[9]Inserm UMR 1167, Institut Pasteur, Lille, France.

Data collection

Stéphanie Chatel[4], Claude Férec[1], Hervé Le Marec[4], Luc Letenneur[8], Gaël Nicolas[7] & Karen Rouault[1]

[7]Univ Rouen, Inserm UMR 1079, Rouen, France. [8]Univ Bordeaux, Inserm UMR 1219, Bordeaux, France.

Sequencing

Delphine Bacq[5], Anne Boland[5] & Doris Lechner[5]

The FranceGenRef Consortium

Principal Investigators

Jean-François Deleuze[5,6] & Emmanuelle Génin[1,2]Richard Redon[4]

Collaborators

Data collection

Chantal Adjou[10], Stéphanie Chatel[4], Claude Férec[1], Marcel Goldberg[11], Philippe-Antoine Halbout[10], Hervé Le Marec[4], David L'Helgouach[10], Karen Rouault[1], Jean-Jacques Schott[4], Anne Vogelsperger[10] & Marie Zins[11]

[10]Etablissement Français du Sang, La Plaine, Saint-Denis, France. [11]UMS 11, Inserm, Université de Versailles Saint-Quentin-en-Yvelines, Versailles, France.

Sample preparation/sequencing

Delphine Bacq[5], Hélène Blanché[6], Anne Boland[5] & Robert Olaso[5]

Bioinformatics

Pierre Lindenbaum[4], Thomas Ludwig[1,2], Vincent Meyer[5], Florian Sandron[5], Damien Delafoy[5] & Lourdes Velo-Suárez[1,2]

Statistical Genetics

Isabel Alves[4], Ozvan Bocher[1], Christian Dina[4], Anthony F. Herzig[1], Matilde Karakachoff[4], Gaëlle Marenne[1], Aude Saint Pierre[1] & Véronique Geoffroy[1]