nature genetics

Whole-genome sequence variation, population structure and demographic history of the Dutch population

The Genome of the Netherlands Consortium*

Whole-genome sequencing enables complete characterization of genetic variation, but geographic clustering of rare alleles demands many diverse populations be studied. Here we describe the Genome of the Netherlands (GoNL) Project, in which we sequenced the whole genomes of 250 Dutch parent-offspring families and constructed a haplotype map of 20.4 million single-nucleotide variants and 1.2 million insertions and deletions. The intermediate coverage (~13×) and trio design enabled extensive characterization of structural variation, including midsize events (30–500 bp) previously poorly catalogued and *de novo* mutations. We demonstrate that the quality of the haplotypes boosts imputation accuracy in independent samples, especially for lower frequency alleles. Population genetic analyses demonstrate fine-scale structure across the country and support multiple ancient migrations, consistent with historical changes in sea level and flooding. The GoNL Project illustrates how single-population whole-genome sequencing can provide detailed characterization of genetic variation and may guide the design of future population studies.

Although the human genome reference sequence provides a common scaffold for the annotation of genes, regulatory elements and other functional units, it does not contain information about how individuals differ in their DNA sequences[1]. Initial efforts to map such variation across the human genome have successfully catalogued millions of common SNPs in various populations[2–5]. Fueled by the commercial development of microarrays for efficient SNP genotyping, genome-wide association studies (GWAS) have provided a systematic approach to test genetic variants for a role in disease. Thus far, GWAS have reproducibly identified thousands of loci, providing insight into underlying pathways of disease, in some cases with translational and clinical impact[6,7]. The importance of these discoveries notwithstanding, many questions remain about the allelic architecture of complex traits, especially with regard to the contributions of common versus rare variation[7–9].

To elucidate the genetic basis of disease, comprehensive sequencing-based approaches are required to interrogate all types of genetic variation, including single-nucleotide variants (SNVs), structural variations and *de novo* events[10–12]. The characterization of rare variation poses a major challenge. Because rare alleles have emerged, on average, relatively recently[13], they show greater geographic clustering[14] than common variants[15]. It is therefore imperative to study large samples across multiple populations, even within continental groups, to build a relatively complete catalog of rare variation in the human genome.

We initiated the Genome of the Netherlands (GoNL) Project to characterize DNA sequence variation for SNVs, short insertions and deletions (indels) and larger deletions in 769 individuals of Dutch ancestry selected from 5 biobanks under the auspices of the Dutch hub of the Biobanking and Biomolecular Research Infrastructure (BBMRI-NL)[16,17]. Specifically, we sampled 231 trios, 11 quartets with monozygotic twins and 8 quartets with dizygotic twins from 11 of the 12 Dutch provinces without ascertaining on the basis of phenotype or disease. By whole-genome sequencing these 250 families at ~13× coverage, our aim was to build a resource of 1,000 haploid genomes as representative of a small (41,543-km^2), densely populated (>17 million inhabitants) country in northwestern Europe (**Supplementary Note**).

Here we provide the first detailed analysis of the GoNL data after processing and quality control (**Supplementary Fig. 1** and **Supplementary Note**). To maximize sensitivity, we analyzed all samples jointly[18] and discovered 20.4 million biallelic SNVs, 1.2 million biallelic indels (<20 bp in length) and 27,500 larger deletions (>20 bp in length). Of the SNVs, 6.2 million are common (minor allele frequency (MAF) > 5%), 4.0 million are low frequency (MAF = 0.5–5%), and 10.2 million are rare (MAF < 0.5%). On the basis of coverage and mapping metrics, we estimate that 94.1% of the genome could be called reliably (the 'accessible' genome), within which 99.2% of SNVs with a frequency of 1% could be detected (**Supplementary Table 1** and **Supplementary Note**). The identification of indels and large deletions was based on conservative consensus calls from several complementary methods (**Supplementary Note**). We used MVNcall for trio-aware phasing and linkage disequilibrium–based imputation[19], starting from the genotype likelihoods of SNVs and indels, yielding a phased panel of 998 unique haplotypes. The non-reference genotype concordance for SNVs was 99.4% (compared to genotypes from Complete Genomics sequencing data in 20 overlapping samples) and 99.5% (compared to Illumina Immunochip genotypes collected for all samples). The average coverage of 13.3× coupled with the family-based design allowed us to construct a high-quality whole-genome data set for further analysis,

*A full list of authors and affiliations appear at the end of the paper.

Received 10 October 2013; accepted 6 June 2014; published online 29 June 2014; doi:10.1038/ng.3021

including characterization of structural variation, detection of *de novo* events, imputation and demographic inference.

RESULTS

Novel variation in GoNL

To determine the number of novel variants, we investigated the overlap between GoNL and existing databases. We detected the majority of sites (98.2%) present in the European sample (CEU) of HapMap Phase 2 (ref. 4) and 71.1% of sites in the European subset of the 1000 Genomes Project Phase 1 (EUR)[20], consistent with the expectation that commonly segregating alleles across European populations should also be detected in GoNL (**Fig. 1a**). Conversely, only 39.0% of rare SNVs observed in GoNL (excluding singletons) were observed in the 1000 Genomes Project EUR panel, highlighting the value of studying individual populations in greater depth. The contribution of 7.6 million novel SNVs in GoNL represents a 14.6% increase in dbSNP (Build 137), although the majority of variants (75.6%) were singletons. Considering that 16.5% of the 2.0 million singletons in the 1000 Genomes Project EUR panel were also observed in GoNL, we expect that a substantial number of the novel GoNL singletons will be encountered again as we continue to sequence larger samples across Europe.

Structural variation could be called confidently across a broad size range, from large deletions to short insertions (**Fig. 1b**). The overall shape of the distribution for size frequency shows that larger structural events are less frequent than smaller indels, presumably reflecting the relatively deleterious nature of the larger variants. We observed specific peaks in the spectrum of size frequency that corresponded to microsatellite instability (MSI) at ±4 bp, short interspersed elements (SINEs) at 300 bp and long interspersed elements (LINEs) at 6 kb. In comparison to 1000 Genomes Project data, 54.4% of the indels (≤20 bp) and 93.3% of the larger deletions (>20 bp) were novel (**Supplementary Note**). Our analysis thus fills an important gap in the discovery of midsize deletions (30–500 bp), where 98.4% of the observed variants were novel. The novelty rate for larger deletions (>500 bp) was still substantial (66.3%). We note that most of the deletions reported here were biased to be common because of stringent filtering (**Supplementary Note**), which allowed us to generate a call set with an overall validation rate of 96.5% (**Supplementary Table 2**). A more complete data set including duplications, inversions, mobile element insertions and translocations is currently being assembled.

Functional variation

Predicting the biological consequences of variants within a single genome is an ongoing challenge with important implications for using sequencing in a clinical setting. To characterize the burden of loss-of-function variants in detail, we classified all such variants in GoNL[21] (**Supplementary Note**). Among rare variants, we observed an excess of nonsense variants and frameshift indels, consistent with a model in which such functional variants are subject to purifying selection (**Supplementary Fig. 2**)[22,23]. We counted 66 larger loss-of-function deletions (removing the first exon of a gene or more than half of its coding sequence)[21], which showed a relative depletion in numbers when compared to all deletions ($P = 0.005$ for deletions of 20–100 bp; $P = 2.6 \times 10^{-9}$ for deletions of >100 bp). This effect was amplified when considering only genes listed in the Online Mendelian Inheritance in Man (OMIM) compendium ($P = 2.4 \times 10^{-27}$), illustrating strong selection against large structural changes in key genes.

The overall patterns and per-individual distributions of loss-of-function SNVs (variants introducing premature stop codons or variants interrupting splice sites) and missense variants were consistent with those found in the 1000 Genomes Project (**Table 1** and **Supplementary Fig. 3**). On average, an individual carried 60 loss-of-function SNVs, 69 loss-of-function indels and 15 loss-of-function large deletions. The bulk of these mutations for each individual were common, suggesting that these variants are not subject to strong selective pressure and, although they are protein-truncating mutations, are likely phenotypically benign. This observation emphasizes the need for caution in assigning pathogenicity to variants purely on the basis of their predicted impact on protein structure.

In contrast, when considering rare loss-of-function variants, which are more likely to be pathogenic, we found that the average individual in GoNL carried four nonsense variants, two variants interrupting a splice site and two frameshift indels. Comparing these numbers to those for synonymous variants (providing a baseline expectation under neutrality), we estimate that each individual carried an excess of 4–5 rare loss-of-function SNVs sufficiently deleterious that they would never reach high frequency in the population (**Supplementary Note**).

We also investigated the number of rare loss-of-function compound heterozygous events for SNVs, short indels and large deletions. Across all samples, we observed 3 such events mapping to 3 genes in 3 individuals (average of 0.01 events per individual). Given the rarity of such variants, the phenotypic impact of compound heterozygosity for rare loss-of-function mutations should be considered explicitly in disease studies.

Whereas compound heterozygosity for rare loss-of-function variants was sparse, we expected compound heterozygosity for common loss-of-function variants to be more prevalent, as these variants are less likely to be deleterious. Indeed, we found that the average number of common loss-of-function compound heterozygous variants per individual was 2.89 (range of 0–7). Interestingly, although there were 1,917 common loss-of-function compound heterozygous events across all samples, they were confined to only 11 genes (**Supplementary Table 3**). All but one of these genes have extreme residual variation intolerance scores[24] (RVIS; all but 1 gene above the 84th percentile across 16,956 genes), which is unlikely to occur by chance ($P = 1.41 \times 10^{-5}$;

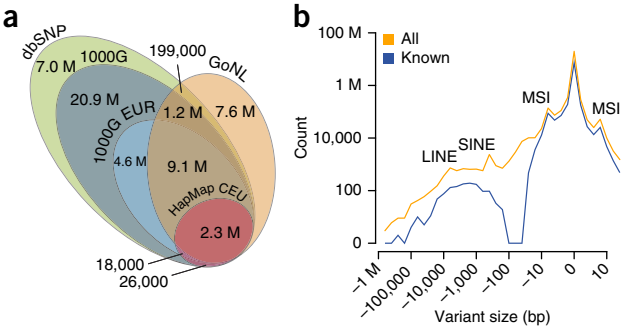

Figure 1 Discovery of SNVs and structural variation. (**a**) Venn diagram of all SNVs discovered in GoNL relative to dbSNP (Build 137) and the 1000 Genomes Project (1000G) Phase 1 and HapMap CEU panels. The majority of the 7.6 million novel sites are rare (MAF < 0.5%), including 5.8 million singletons. M, million. (**b**) Size frequency spectrum of all variation discovered in GoNL, where a negative size indicates a deletion. Our detection strategy employed multiple approaches and provided a substantial boost in the identification of novel structural variants in the midsize range (30–500 bp). Peaks corresponding to long interspersed elements (LINEs), short interspersed elements (SINEs) and microsatellite instability (MSI) are highlighted. The total number of variants called in GoNL is shown in orange, whereas SNVs found in dbSNP (Build 137) and short indels and large deletions found in 1000 Genomes Project Phase 1 data are shown in blue. For large deletions (>20 bp), we required at least 80% reciprocal overlap between variants for them to be considered similar.

Table 1 Individual variant load of coding mutations

	Non-reference allele frequency		
	Rare (<0.5%)	Low frequency (0.5–5%)	Common (>5%)
Variant type	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)
All SNVs[a]	28,142 (3009.2)	130,190 (2448.1)	2.90 M (10,080.9)
Novel[a,b]	17,751 (1,176.3)	4,354 (346.8)	620 (31.7)
Total conserved	1,892 (187.7)	7,593 (154.5)	106,824 (443.9)
Functional variation			
Synonymous	18 (4.9)	73 (8.9)	990 (19.0)
Nonsynonymous	101 (11.9)	238 (15.6)	2,089 (31.8)
Probably damaging	32 (5.8)	58 (7.9)	394 (12.2)
Stop gain[a]	4 (1.9)	5 (2.2)	38 (4.3)
Splice-site donor[a]	1 (0.9)	1 (0.9)	4 (1.5)
Splice-site acceptor[a]	1 (0.7)	0.5 (0.6)	7 (1.4)
Total loss of function[a]	5 (2.2)	6 (2.4)	49 (4.7)
Disease-associated variation			
OMIM	0 (0.6)	2 (1.6)	57 (4.9)
HGMD[c]	2 (1.2)	8 (2.7)	11 (2.3)
Indels (<20 bp)			
Indel frameshift[a]	2 (1.4)	6 (2.6)	61 (4.8)
Indel non-frameshift[a]	1 (1.1)	6 (2.6)	99 (5.9)
Deletions (>20 bp)			
Loss of function	0 (0.2)	1 (1.0)	14 (3.3)
Total bases deleted		6.7 million bases	

Only SNV sites at which ancestral state can be assigned with high confidence and that are highly conserved (GERP > 2.0) are reported. Frequency stratifications are based on the unrelated samples only. OMIM, Online Mendelian Inheritance in Man; M, million. [a]No conservation filter applied. [b]Not observed in dbSNP Build 137 (which includes all SNVs reported in the 1000 Genomes Project Phase I data release). [c]Frequency stratification and variant counts based on the reported mutation allele.

Supplementary Fig. 4) and suggests that these genes are more tolerant of disruptive mutations.

Because databases of disease-relevant mutation are often employed to identify potential variants of interest, we annotated variants in GoNL that were listed as disease causing (DM) in the Human Gene Mutation Database (HGMD)[25]. We observed that a sample carried, on average, 20 DM variants (range of 9–33) (**Table 1**). Because all samples were derived from population-based cohorts, the impact of these alleles is unclear. One possibility is that the presence of modifier alleles induces incomplete penetrance or variable expressivity of DM variants, depending on the carrier's genetic background[26]. An alternative explanation is that HGMD contains a considerable number of false positive disease-causing mutations[27]. Of the 1,093 DM mutations occurring in GoNL, 32% had a frequency of >1%, higher than the prevalence of many of the diseases described in HGMD. Given the inheritance patterns of the diseases conferred by these variants, many individuals in GoNL should have been affected by diseases with profound physical or even lethal manifestations (**Table 2** and **Supplementary Table 4**). In fact, one of these variants (chr. 14: 94,847,262; a variant for α1 antitrypsin deficiency) was recently implicated as a pathogenic incidental finding in a set of 1,000 exomes[28]. The prevalence of α1 antitrypsin deficiency (MIM 613490), an autosomal recessive disease, is estimated to be 0.02–0.06%, yet two unrelated GoNL individuals were homozygous carriers of this variant (prevalence = 0.4%, ~10× higher than the disease prevalence). Further, the typical age of onset of α1 antitrypsin deficiency is 20–50 years, whereas the two homozygous carriers in GoNL were ages 60 and 63 years at ascertainment. These results highlight the potential pitfalls of employing such databases in disease studies and the challenge of interpreting personal genomes.

De novo mutations

A distinct advantage of the family-based study design is the ability to call *de novo* events in genomic regions with sufficient coverage in a trio. To this end, we developed the PhaseByTransmission (PBT) module in the Genome Analysis Toolkit (GATK)[29]. From an initial 4.5 million mendelian violations in the original calls made in the 258 independent offspring, we prioritized 29,162 autosomal *de novo* mutation candidates at non-polymorphic sites with PhaseByTransmission (**Supplementary Note**). Given that the average number of *de novo* mutations per offspring was still higher than expected (~63.2 mutations per offspring[30]), we evaluated to what extent sequencing features could help increase the accuracy of *de novo* mutation prediction and reduce the number of false positives. We validated 592 candidate sites as true *de novo* mutations (**Supplementary Note**) and classified another 1,674 candidates as false positives (on the basis of validation experiments and Complete Genomics genotype data). We trained a random forests classifier on various sequencing features using 70% of the validation results (**Supplementary Note**) and obtained a model with an estimated classification accuracy of 92.2% using the remaining 30% of the data (**Fig. 2a**). This analysis illustrates how joint assessment of raw trio data and sequencing context can greatly boost prediction accuracy. We applied the classifier to our initial candidates and identified 11,020 high-confidence *de novo* mutations (18–74 per offspring) for downstream analyses. Owing to fluctuations in regional coverage, we expect a substantial fraction of genuine *de novo* mutations to have been missed. We also note that early embryonic somatic mutations would be indistinguishable from germline mutations.

We observed a significant positive correlation ($r^2 = 0.47$, $P < 2.2 \times 10^{-16}$) between father's age at conception and number of *de novo* mutations in the offspring (**Fig. 2b**), providing a third, independent estimate based on a larger sample size[30,31]. Assuming mutations are Poisson distributed among samples and adjusting for coverage, we estimated that each additional year of father's age caused a 2.5% increase in the mean number of *de novo* mutations in the offspring. Although parents' ages were highly correlated ($r^2 = 0.66$), comparing models based on father's and mother's age at conception suggested that the age-related increase in the frequency of *de novo* mutations was a predominantly paternal effect (**Supplementary Note**). Interpolating from the paternal model, we expected, on average, 75.4% of the *de novo* mutations in the GoNL offspring to originate from the father (assuming a linear increase in the number of *de novo* mutations from puberty). Using read-pair information, we were able to assign parental origin to 2,613 *de novo* mutations, and we found that, indeed, 76.0% were paternal. When considering only mutations for which parental origin could be determined, the correlation with father's age remained significant ($r^2 = 0.11$, $P = 2.0 \times 10^{-6}$; **Supplementary Fig. 5**), but the correlation was not significant for mother's age ($P = 0.94$), highlighting the relative impact of paternal and maternal mutations.

Within a single family, we attempted to identify *de novo* indels and large deletions. Using strict filtering criteria for mendelian violations followed by PCR-based Sanger sequencing (**Supplementary Note**), we confirmed six intergenic *de novo* indels (1–2 bp) and a large 113-kb *de novo* deletion located in an intron of the *SUMF1* gene (which seems unlikely to have substantial impact on gene function). These results show that our predictions of indels and structural changes are a valuable source for both commonly segregating alleles

Table 2 HGMD disease-causing mutations in GoNL samples

Chr.	Position	Gene	Mutation allele	Reference allele	Disease in HGMD	Disease prevalence	Inheritance pattern	Affected individuals[c]	Phenotypic manifestations[a]	OMIM ID	Mutation allele frequency in GoNL[d]	Mutation allele frequency 1000G CEU
4	6,302,519	*WFS1*	A	G	Wolfram syndrome	0.0002%[a]	AR	257	Hyperglycemia, vision and hearing loss	604928, 222300	0.728	0.759
13	52,515,354	*ATP7B*	G	A	Wilson disease	0.003%[a]	AR	167	Liver disease, neuropsychiatric problems	277900	0.574	0.582
16	3,304,463	*MEFV*	T	C	Familial Mediterranean fever	0.10% in Mediterranean populations; rarer elsewhere[a]	AR	36	Recurrent fevers, inflammation of the abdomen, chest, joints	249100, 134610	0.277	0.224
11	6,415,463	*SMPD1*	A	G	Niemann-Pick disease	0.0004%[a]	AR	37	Nervous system deterioration, failure to thrive, fatal in infancy or early childhood (type A)	257200, 607616, 257220, 607625	0.230	0.230
20	61,463,522	*COL9A3*	A	C	Pseudoachondroplasia	0.003%[a]	AD	177	Short stature, joint pain	177170	0.197	0.200
10	13,340,236	*PHYH*	A	G	Refsum disease	Unknown, current estimate 0.0001%[a]	AR	18	Anosmia, progressive blindness, deafness, hand/feet bone abnormalities, arrhythmia	266500	0.188	0.153
15	52,643,564	*MYO5A*	A	G	Griscelli syndrome	<0.0001%[b]	AR	10	Albinism (all types), intellectual disability (type 1), recurrent infection (type 2)	214450, 607624, 609227	0.159	0.141
19	36,339,247	*NPHS1*	T	C	Congenital nephrotic syndrome (Finnish type)	0.01% in Finland; rarer elsewhere[b]	AR	2	Proteinuria, rapid progression to renal failure	256300	0.082	0.082 (0.110)[e]
14	94,847,262	*SERPINA1*	A	T	α1 antitrypsin deficiency	0.02–0.06%[a]	AR	2	Lung disease, liver disease	613490	0.039	0.053

HGMD, Human Gene Mutation Database; AR, autosomal recessive; AD, autosomal dominant; OMIM, Online Mendelian Inheritance in Man; chr., chromosome.
[a]National Institutes of Health, Genetics Home Reference (United States). [b]National Institute of Health and Medical Research (France). [c]Unrelated individuals in GoNL carrying two copies of the mutation allele (for autosomal recessive diseases) or at least one copy of the mutation allele (for autosomal dominant diseases). [d]Calculated from unrelated individuals. [e]Frequency in 1000 Genomes Project (1000G) Phase I samples from Finland.

and *de novo* events. Further work is needed to assess the frequency of such *de novo* events in the general population.

Imputation

One of the goals of the GoNL Project was to provide a community resource for downstream imputation into GWAS samples. To evaluate the performance of the GoNL panel, we used independent Complete Genomics sequence data collected for 81 individuals of Dutch ancestry (the NTL data set). In these NTL samples, we masked all genotypes at SNVs not present on the Illumina Human-1M array, imputed these masked SNVs from the remaining SNVs and then compared the imputed and known genotypes (**Supplementary Note**). The aggregate mean r^2 value was 0.99 for common SNVs, 0.86 for low-frequency SNVs and 0.63 for rare SNVs, indicating good overall imputation quality (**Fig. 3**). We repeated this evaluation on the basis of the SNV content of other microarrays and obtained similar imputation performance for common SNVs, but we observed notable differences in imputation quality for lower frequency alleles (**Supplementary Fig. 6**). To directly measure the impact of trio-based phasing, we constructed a panel using data from the unrelated parents alone and

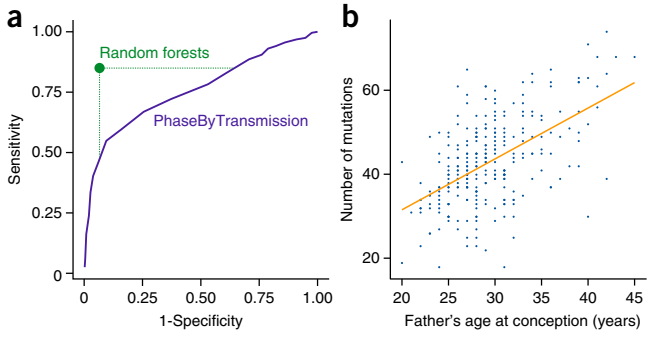

Figure 2 *De novo* mutation detection. (**a**) Receiver operating characteristics (ROC) curve to predict *de novo* mutations using PhaseByTransmission only (purple line; 2,199 sites) or using PhaseByTransmission followed by random forests classification trained on 70% of the validation data (green dot; evaluation subset only, 657 sites). The random forests classifier had an estimated 84.5% sensitivity and 94.6% specificity. (**b**) The number of *de novo* mutations in each of the 258 independent offspring is plotted (in blue) as a function of paternal age at conception. Linear regression of mutational load on paternal age is significant (Pearson's correlation = 0.47, $P < 2.2 \times 10^{-16}$), with the least-squares fit plotted in orange.

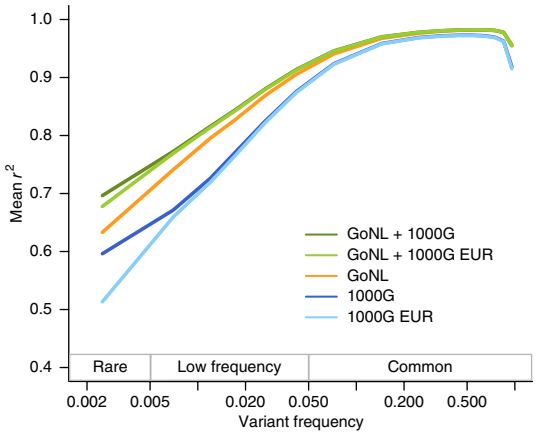

Figure 3 Imputation accuracy. The aggregate r^2 value between imputed and gold-standard genotype dosage is plotted as a function of allele frequency. We used genotypes from 81 Dutch samples (independent from GoNL) all sequenced with Complete Genomics as the gold standard. The GoNL panel consistently outperforms the 1000 Genomes Project panels, especially at lower allele frequencies. A combined GoNL and 1000 Genomes Project panel provided the best performance.

reevaluated imputation quality in the NTL samples. The imputation accuracy dropped to a mean r^2 value of 0.47 for rare variants (0.85 and 0.98 for low-frequency and common SNVs, respectively), indicating that trio-based phasing contributed substantially to the imputation quality of rare variants.

In comparison to imputation using 1000 Genomes Project data as the reference panel, we observed better imputation accuracy with the GoNL panel for SNVs with a frequency of up to 10%, despite the larger sample size of the 1000 Genomes Project (**Fig. 3**). To investigate the basis for the improved imputation accuracy with the GoNL panel, we constructed 3 reference panels using the 1000 Genomes Project CEU (Northern and Western Europeans from Utah) and TSI (Tuscans from Italy) panels and the GoNL panel, all with 85 individuals. Using each of these reference panels, we imputed into independent CEU, TSI and NTL samples with Complete Genomics data (**Supplementary Note**). Of the three panels, GoNL gave the highest imputation accuracy (especially for rare variants), not only for the NTL samples but also for the CEU samples, indicating that the improved performance of the GoNL panel was not simply due to shared ancestry of the GoNL and NTL samples (**Supplementary Fig. 7**). Differentiation between northern and southern European populations might explain why the 1000 Genomes Project CEU panel and the GoNL panel showed roughly equivalent performance for the TSI panel (with performance certainly worse than when the 1000 Genomes Project TSI panel was used as the reference). Overall, these results suggest that the GoNL design enabled accurate reconstruction of long-range haplotypes with marked improvement in imputation of rare alleles.

To assess the potential value of larger reference panels, we combined the 1000 Genomes Project and GoNL panels with IMPUTE2 (ref. 32), and we evaluated imputation accuracy in the NTL samples. Here we obtained an additional gain in accuracy over the GoNL panel alone, achieving a mean r^2 value of 0.70 for rare SNVs and 0.88 for low-frequency SNVs (**Fig. 3**). Increasing the sample size of the reference panel will likely continue to improve imputation performance (especially for lower frequency alleles), motivating a community-wide effort to create a unified reference panel across diverse populations.

Population structure and demographic inference

Although it is well understood that extensive migration and gene flow occurred among European populations[33–35], we focused on creating a unified picture of Dutch demography in recent millennia. Because of unbiased ascertainment and the inclusion of rare variation, whole-genome sequence data can potentially offer greater resolution for demographic inference than SNP array data.

First, we explored global relationships, analyzing both common and rare variants to elucidate ancient and recent population differentiation. We calculated Hudson's F_{ST} between the Dutch population and the 14 populations present in the 1000 Genomes Project, finding that F_{ST} patterns were consistent with continental clustering in principal-component analysis (PCA) and with previous estimates (**Supplementary Fig. 8** and **Supplementary Table 5**)[36]. To investigate more recent population connections, we focused on so-called f_2 variants, mutations appearing exactly twice (in two heterozygote carriers) in the joint data from GoNL and the 1000 Genomes Project (**Supplementary Note**). As was observed in the 1000 Genomes Project, within-population sharing accounted for the majority (50.8%) of all f_2 alleles (**Supplementary Fig. 9**), but f_2 sharing identified cross-population connections as well. For example, a Dutch sample sharing an f_2 variant with a non-Dutch individual was far more likely to share that variant with another individual from Europe (71.6%) or the Americas (21.0%; because of European admixture) than with an individual from Africa (6.2%) or East Asia (1.3%). These results underscore the high degree of geographic clustering of recent mutations. In analysis of mitochondrial DNA, the frequencies of the major haplogroups (H, 39.4%; U, 25.2%; J, 10.4%; T, 10.8%) and minor haplogroups were in agreement with previous observations in other European populations (**Supplementary Note**)[37].

Within the Netherlands (**Fig. 4a**), PCA showed subtle substructure along a north-south gradient (**Fig. 4b** and **Supplementary Note**), consistent with previous findings[38,39]. Because PCA cannot elucidate demography (particularly migration patterns)[40], we also performed an independent analysis of identity-by-descent (IBD) sharing that identified subtle signals of migration (**Supplementary Note**)[41]. From the length distributions of the IBD segments[42], we inferred demographic models and estimated the effective population sizes of the Dutch provinces at different time scales, reflecting historical changes in demography (**Supplementary Note**).

Analysis of IBD segments of 1–2 cM in length, corresponding to an estimated time to the most recent common ancestor of ~4,200 years, showed homogeneous effective population sizes across the 11 provinces, consistent with common genetic origins (**Supplementary Fig. 10**). Additionally, we observed a smooth south-to-north gradient of decreasing ancestral population size and increased homozygosity in the northern provinces (average within-province IBD sharing and latitude correlation, $r = 0.923$, $P = 5 \times 10^{-5}$; **Supplementary Figs. 10** and **11**, and **Supplementary Note**). Traditionally, such observations have been explained by a serial founder effect characterized by migration from the south to the north[38].

Interestingly, GoNL samples, regardless of place of birth, tended to share more IBD segments with other individuals from the north than with individuals from the same region. Although within-province IBD sharing was strong, excess sharing with individuals from the the northern provinces was evident (average between-province IBD sharing and average province latitude correlation, $r = 0.934$, $P < 1 \times 10^{-5}$; **Fig. 4c** and **Supplementary Table 6**). This pattern indicates that a simple south-to-north serial founder model might not be sufficient to explain the observed IBD sharing (**Supplementary Fig. 12** and **Supplementary Note**). Instead, different demographic

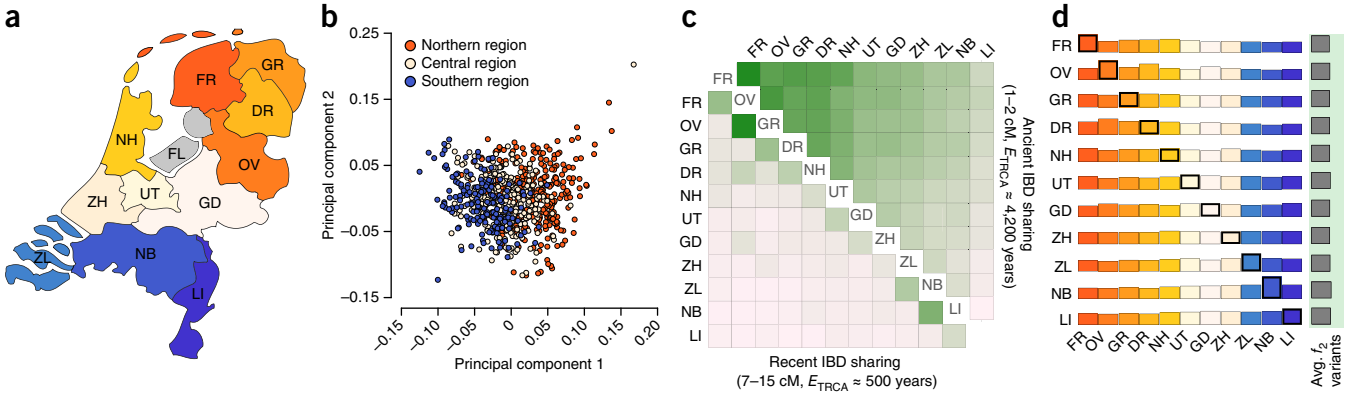

Figure 4 Population genetic analyses in the Dutch population. (**a**) Map of the Netherlands with its 12 provinces. We selected 769 individuals from 5 BBMRI-NL biobanks across all provinces except Flevoland. FR, Friesland; GR, Groningen; DR, Drenthe; OV, Overijssel; FL, Flevoland; NH, Noord-Holland; ZH, Zuid-Holland; UT, Utrecht; GD, Gelderland; ZL, Zeeland; NB, Noord-Brabant; LI, Limburg. (**b**) PCA. Individuals are projected onto the two dominant principal components, showing subtle substructure along a north-south axis within the Netherlands. (**c**) Heat map of IBD segment sharing within and across provinces. The upper half represents ancient IBD sharing (1–2 cM), and the bottom half represents recent IBD sharing (7–15 cM). Strikingly, all GoNL individuals, regardless of current residence, share more short IBD segments with individuals from the northern provinces than with other individuals from their own province. Patterns for long IBD segments are consistent with restricted geographic movement in recent times. E_{TRCA}, estimated time to recent common ancestor. (**d**) Sharing of rare doubleton (f_2) variants within and across provinces. The level of within-province sharing of f_2 variants exceeds that of across-province sharing, reflecting strong geographically localized clustering of these recent variants. The degree of f_2 sharing among northern or southern provinces is statistically significant compared to the sharing among central provinces ($P < 1 \times 10^{-200}$). Avg., average.

scenarios remain plausible, but all support a model of substantial regional migration. Assuming that ancient serial migrations toward the north of the country caused the observed gradient of increasing homozygosity, a possible explanation for these results is that additional migratory events out of the north took place after initial settlements. These subsequent migratory events are consistent with the dynamic nature of the Netherlands, particularly in the northern coastal regions, between 5000 BCE and 50 CE (**Supplementary Fig. 13**). A series of abandonments and resettlements were likely prompted by shifts in ocean level and flooding that changed once-habitable land into dunes and marshes or buried regions entirely underwater. We emphasize that other, more complex demographic models might yield similar patterns of IBD sharing; additional analyses are required to assess alternative scenarios.

In recent centuries, the advent of water-defense technologies (beginning in the thirteenth century) increased land stability, allowing for other forces to influence demography. An analysis of f_2 variants identified non-random sharing within and across provinces. Although the proportion of within-province f_2 sharing comprised only 12% of all f_2 alleles, consistent with homogenous ancestry, this is significantly greater than expected under the null hypothesis of uniform allele sharing ($P < 1 \times 10^{-200}$; **Fig. 4d**). This geographic localization of rare variants is suggestive of limited migration in recent centuries, consistent with current demographic studies. Notably, the Noord-Brabant and Overijssel provinces showed significantly stronger within-province f_2 sharing than the other provinces ($P = 1.2 \times 10^{-151}$ and $P < 1 \times 10^{-200}$, respectively), consistent with smaller effective population sizes in these two provinces inferred from sharing of long IBD segments (**Fig. 4c**, **Supplementary Fig. 9** and **Supplementary Table 6**). Further, we found that within-region sharing in the northern and southern regions was substantially stronger when compared to such sharing in the central region ($P < 1 \times 10^{-200}$, both comparisons). Taken together, these results suggest increased migration in the central region (in comparison to the northern and southern regions), consistent with recent urbanization in the wealthier central provinces.

DISCUSSION

The results presented here reflect the enormous wealth of knowledge that can be gleaned from whole-genome sequencing data and illustrate how intermediate-coverage sequencing within a single country complements cosmopolitan, low-coverage efforts[20]. The observed proportion of novel variation (in particular, for structural variation) underlines the added value of in-depth population studies such as GoNL. Combining sequencing data sets within and across populations will not only maximize sensitivity and resolution for the discovery of all types of DNA variation but will also enable population genetic analyses that can shed more light on local and global shared ancestry.

In spite of the intermediate coverage in GoNL, we were able to reliably call *de novo* point mutations and confirm the relationship between paternal age at conception and mutation load in offspring. We showed that we could also identify larger *de novo* events; these calls will have to be validated empirically and their properties studied across the entire cohort. The methods we developed for the discovery of *de novo* mutations should be broadly applicable for studies of diseases in which *de novo* mutations are suspected to have a role[12]. *De novo* mutation represents an important class of DNA sequence variation that can further elucidate fundamental processes of mutagenesis[43]. Our results indicate that trio-based sequencing of large samples at intermediate coverage may be a cost-effective way to ascertain genome-wide variation in mutation rates and establish a 'null expectation' for the general population against which mutations in cases can be compared. Similarly, population-based sequencing efforts are instrumental in defining guidelines for investigating the causality of variants that may have functional and phenotypic impact[44].

As long as the cost of genotyping continues to be competitive with whole-genome sequencing, imputation will remain important. The consolidation of available whole-genome data sets into a single cosmopolitan panel, including low-frequency, structural and other complex types of variation[45,46], should therefore be considered a top priority. Through more complete interrogation of genetic variation, studies of large, well-phenotyped samples will continue to increase

the number of opportunities for the development of diagnostic tools, prevention measures and therapeutics for human disease.

URLs. The Genome of the Netherlands Project, http://www.nlgenome.nl/; European Genome-phenome Archive (EGA), http://www.ebi.ac.uk/ega/; The Groningen Center for Information Technology, http://www.rug.nl/cit/; Target, http://www.rug.nl/target/; BiG Grid, http://www.biggrid.nl/; SURFsara, http://www.surfsara.nl/; MOLGENIS, http://www.molgenis.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data, variant calls, inferred genotypes and phased haplotypes have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under accession EGAS00001000644.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We wish to dedicate this work to the memory of David R. Cox, an enthusiastic supporter of human genetic research in the Netherlands for many years. The GoNL Project is funded by the BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007). We acknowledge additional financial support from eBioGrid, CTMM/TraIT, the Ubbo Emmius Fund, the Netherlands Bioinformatics Center (NBIC) and EU-BioSHARE. We thank the individual participants of the biobanks; M. Depristo, E. Banks, R. Poplin and G. del Angel from the Broad Institute for expert advice on setting up our alignment and calling pipeline; K. Garimella for the initial implementation of PhaseByTransmission; G. Strikwerda, W. Albers, R. Teeninga, H. Gankema and H. Wind of the Groningen Center for Information Technology (see URLs) for support of the compute cluster and Target storage; E. Valentyn and R. Williams of Target (see URLs) for hosting project data on IBM GPFS storage; T. Visser and I. Nooren of BiG Grid (see URLs) and SURFsara for providing backup storage, additional computing capacity and expert advice; the team from MOLGENIS (see URLs) for software development support; H. Lauvenberg for handling data access requests; K. Zych for design of the GoNL logo; L. Franke, H.-J. Westra and J. Gutierrez-Achury for useful discussions; and S. Raychaudhuri and B. Neale for their critical reading of the manuscript. Target is supported by Samenwerkingsverband Noord Nederland, the European Fund for Regional Development, the Dutch Ministry of Economic Affairs, Pieken in de Delta and the provinces of Groningen and Drenthe. Target operates under the auspices of Sensor Universe. BiG Grid and the Life Science Grid are financially supported by the Netherlands Organization for Scientific Research (NWO). A.A. is funded by the Center for Medical Systems Biology-2, and D.I.B. is funded by the European Research Council (ERC 230374). A.S. and P.I.W.d.B. are recipients of VIDI awards (NWO projects 016.138.318 and 016.126.354, respectively).

AUTHOR CONTRIBUTIONS

P.I.W.d.B., D.I.B., J.A.B., C.M.v.D., G.-J.B.v.O., P.E.S., M.A.S. and C.W. (chair) formed the steering committee of the GoNL Project. Biobanks are managed and organized by A.H., A.G.U., C.M.v.D., B.O., F.R., A.I. (for the Rotterdam and Erasmus Rucphen Family studies), D.I.B., G.W. (for the Netherlands Twin Register), P.E.S., M.B., A.J.M.d.C., H.E.D.S. (for the Leiden Longevity Study) and the members of the LifeLines Cohort Study. P.I.W.d.B. and M.A.S. jointly led the analysis group. Sequencing data were generated at BGI (Shenzhen, China) by Q.L., Y.L., Y.D., R.C., H.C., N.L., S.C. and J.W. Additional Complete Genomics sequencing data were generated by S.J.P., S.P., P.S. and D.R.C. through a partnership with Pfizer. F.v.D., P.B.T.N., P.D., L.C.F., A.K., M.D., H.B., K.J.v.d.V. and M.A.S. formed the operational data stewardship and processing center. P.B.T.N., F.v.D. and M.A.S. designed and implemented the compute cluster. M.D., H.B., A.K. and M.A.S. designed and implemented the MOLGENIS computing platform to scale up analysis pipelines for alignment, variant calling and imputation. F.v.D. and L.C.F. performed alignment with help from I.J.N., J.B. and B.D.C.v.S. L.C.F. and F.v.D. called SNVs. L.C.F., S.L.P., A.M., E.M.v.L., L.C.K., M. Sohail, A.A. and M.V. performed quality control. V.G., K.Y., L.C.F., T.M., A.S., R.E.H., S.A.M., W.P.K., F.H., J.Y.H.-K., E.-W.L., A.A., V.K., H.M., M.H.M. and J.B. formed the structural variation subgroup. L.C.F. developed the PhaseByTransmission module in GATK and performed *de novo* mutation analyses with P.P. A.M. performed haplotype phasing and imputation benchmarks. J.H.V. and L.H.v.d.B. provided Complete Genomics data for imputation benchmarking. W.P.K. and I.R. performed variant validation. C.W. and M.P. generated Immunochip data on all GoNL samples. S.L.P., C.C.E., A.M., P.F.P., I.P., A.A., N.A., M. Sohail, D.V. and S.R.S. performed population genetic analyses. M.v.O., M.V., M.L., J.F.J.L., M. Stoneking, P.d.K. and M. Kayser performed mitochondrial DNA analysis. P.D., A.M., A.K., E.M.v.L., L.C.K., K.E., C.M.-G., J.v.S., M. Kattenberg, J.J.H. and D.v.E. formed the imputation subgroup. P.B.T.N., K.J.v.d.V. and M.A.S. were responsible for the GoNL website and associated services (see URLs). C.W. conceived the GoNL Project. P.I.W.d.B. wrote the initial manuscript with critical input from L.C.F., A.M., S.L.P., P.F.P. and C.C.E. C.W., D.I.B., G.-J.B.v.O., L.C.K., A.A., M.A.S., P.E.S., S.R.S., J.Y.H.-K., I.P., J.H.V., P.d.K., W.P.K., T.M., A.S., V.G., J.T.d.D. and M. Kayser provided critical feedback on the manuscript. All authors have seen and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
3. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
4. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
5. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
6. Manolio, T.A. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* **14**, 549–558 (2013).
7. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
8. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
9. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2011).
10. Goldstein, D.B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).
11. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
12. Veltman, J.A. & Brunner, H.G. *De novo* mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
13. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
14. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* **108**, 11983–11988 (2011).
15. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
16. Boomsma, D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22**, 221–227 (2014).
17. Brandsma, M. *et al.* How to kickstart a national biobanking infrastructure—experiences and prospects of BBMRI-NL. *Nor. Epidemiol.* **21**, 143–148 (2012).
18. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
19. Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84–91 (2013).
20. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
21. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
22. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
23. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44**, 623–630 (2012).
24. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
25. Stenson, P.D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
26. Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
27. Cassa, C.A., Tong, M.Y. & Jordan, D.M. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum. Mutat.* **34**, 1216–1220 (2013).

28. Dorschner, M.O. *et al.* Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* **93**, 631–640 (2013).
29. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
30. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
31. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell* **151**, 1431–1442 (2012).
32. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
33. Lao, O. *et al.* Correlation between genetic and geographic structure in Europe. *Curr. Biol.* **18**, 1241–1248 (2008).
34. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
35. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol.* **11**, e1001555 (2013).
36. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A.L. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
37. Zheng, H.-X., Yan, S., Qin, Z.-D. & Jin, L. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Sci. Rep.* **2**, 745 (2012).
38. Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21**, 1277–1285 (2013).
39. Lao, O. *et al.* Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history. *Investig. Genet.* **4**, 9 (2013).
40. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**, 646–649 (2008).
41. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
42. Palamara, P.F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
43. Gratten, J., Visscher, P.M., Mowry, B.J. & Wray, N.R. Interpreting the role of *de novo* protein-coding mutations in neuropsychiatric disease. *Nat. Genet.* **45**, 234–238 (2013).
44. MacArthur, D.G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
45. Boettger, L.M., Handsaker, R.E., Zody, M.C. & McCarroll, S.A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**, 881–885 (2012).
46. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS ONE* **8**, e64683 (2013).

The Genome of the Netherlands Consortium:

Laurent C Francioli[1,40], Androniki Menelaou[1,40], Sara L Pulit[1,40], Freerk van Dijk[2,3,40], Pier Francesco Palamara[4], Clara C Elbers[1], Pieter B T Neerincx[2,3], Kai Ye[5,6], Victor Guryev[7], Wigard P Kloosterman[1], Patrick Deelen[2,3], Abdel Abdellaoui[8], Elisabeth M van Leeuwen[9], Mannis van Oven[10], Martijn Vermaat[11,12], Mingkun Li[13], Jeroen F J Laros[11,12], Lennart C Karssen[9], Alexandros Kanterakis[2,3], Najaf Amin[9], Jouke Jan Hottenga[8], Eric-Wubbo Lameijer[6], Mathijs Kattenberg[8], Martijn Dijkstra[2,3], Heorhiy Byelas[2,3], Jessica van Setten[1], Barbera D C van Schaik[14], Jan Bot[15], Isaäc J Nijman[1], Ivo Renkens[1], Tobias Marschall[16], Alexander Schönhuth[16], Jayne Y Hehir-Kwa[17,18], Robert E Handsaker[19,20], Paz Polak[19,21], Mashaal Sohail[19,21], Dana Vuzman[19,21], Fereydoun Hormozdiari[22], David van Enckevort[12], Hailiang Mei[12], Vyacheslav Koval[23], Matthijs H Moed[6], K Joeri van der Velde[2,3], Fernando Rivadeneira[9,23], Karol Estrada[19,23,24], Carolina Medina-Gomez[23], Aaron Isaacs[9], Steven A McCarroll[19,20], Marian Beekman[6], Anton J M de Craen[6], H Eka D Suchiman[6], Albert Hofman[9], Ben Oostra[25], André G Uitterlinden[9,23], Gonneke Willemsen[8], LifeLines Cohort Study[26], Mathieu Platteel[2], Jan H Veldink[27], Leonard H van den Berg[27], Steven J Pitts[28], Shobha Potluri[28], Purnima Sundar[28], David R Cox[28,39], Shamil R Sunyaev[19,21], Johan T den Dunnen[11,29], Mark Stoneking[13], Peter de Knijff[30], Manfred Kayser[10], Qibin Li[31], Yingrui Li[31], Yuanping Du[31], Ruoyan Chen[31], Hongzhi Cao[31], Ning Li[32], Sujie Cao[32], Jun Wang[31,33,34], Jasper A Bovenberg[35], Itsik Pe'er[4,36], P Eline Slagboom[6], Cornelia M van Duijn[9], Dorret I Boomsma[8], Gert-Jan B van Ommen[37], Paul I W de Bakker[1,38,41], Morris A Swertz[2,3,41] & Cisca Wijmenga[2,3,41]

[1]Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands. [2]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [3]Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [4]Department of Computer Science, Columbia University, New York, New York, USA. [5]The Genome Institute, Washington University, St. Louis, Missouri, USA. [6]Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands. [7]European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [8]Department of Biological Psychology, VU University Amsterdam, Amsterdam, the Netherlands. [9]Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands. [10]Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands. [11]Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [12]Netherlands Bioinformatics Center, Nijmegen, the Netherlands. [13]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [14]Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, the Netherlands. [15]SURFsara, Science Park, Amsterdam, the Netherlands. [16]Centrum Wiskunde & Informatica, Life Sciences Group, Amsterdam, the Netherlands. [17]Department of Human Genetics, Radboud University Nijmegen Medical Center, Nijmegen, the Netherlands. [18]Center for Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen Medical Center, Nijmegen, the Netherlands. [19]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [20]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [21]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [22]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [23]Department of Internal Medicine, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands. [24]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA. [25]Department of Clinical Genetics, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands. [26]A full list of members appears in the **Supplementary Note**. [27]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands. [28]Rinat-Pfizer, Inc., South San Francisco, California, USA. [29]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. [30]Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [31]BGI-Shenzhen, Shenzhen, China. [32]BGI-Europe, Copenhagen, Denmark. [33]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [34]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [35]Legal Pathways Institute for Health and Bio Law, Aerdenhout, the Netherlands. [36]Department of Systems Biology, Columbia University, New York, New York, USA. [37]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [38]Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands. [39]Deceased. [40]These authors contributed equally to this work. [41]These authors jointly directed this work. Correspondence should be addressed to P.I.W.d.B. (pdebakker@umcutrecht.nl) or C.W. (c.wijmenga@umcg.nl).

ONLINE METHODS

Sample collection. Five Dutch biobanks (LifeLines Cohort Study, Leiden Longevity Study, Netherlands Twin Registry, Rotterdam Study, Rucphen Study) contributed samples of Dutch ancestry (with parents born in the Netherlands). A total of 250 parent-offspring families (231 trios and 19 quartets, of which 11 had monozygotic twins and 8 had dizygotic twins) comprising 769 individuals were selected without phenotype ascertainment across 11 of the 12 provinces of the Netherlands (**Supplementary Fig. 4** and **Supplementary Table 7**). Limited phenotype data were available: age, sex, height, body mass index (BMI) and levels of total cholesterol, high-density lipoprotein (HDL)-cholesterol, low-density lipoprotein (LDL)-cholesterol and triglycerides (**Supplementary Fig. 14**). All participants provided written informed consent, and each biobank was approved by their respective institutional review board (IRB).

Data generation and processing. Samples were sequenced on an Illumina HiSeq 2000 instrument (91-bp paired-end reads, 500-bp insert size), and reads were aligned on the UCSC human reference genome build 37 using Burrows-Wheeler Aligner (BWA) 0.5.9-r16 (refs. 47,48). Samples were also genotyped on the Illumina Immunochip as well as on at least one other genotyping chip (**Supplementary Table 8**).

Single-nucleotide variant calling. SNV calling was performed on all samples jointly using GATK UnifiedGenotyper v1.6. Calls were filtered using GATK VariantQualityScoreRecalibration (**Supplementary Fig. 15** and **Supplementary Note**), and quality metrics were evaluated (**Supplementary Fig. 16**, **Supplementary Table 9** and **Supplementary Note**). We defined the accessible genome using the same methodology as the 1000 Genomes Project[20] (**Supplementary Fig. 17** and **Supplementary Note**). We assessed the robustness of our pipeline with respect to its stochastic components, its parameters and the version of the tools by reprocessing one trio under different conditions (**Supplementary Tables 10** and **11**, and **Supplementary Note**).

Indels and structural variants. To create reliable indel and structural variant call sets, we used a combination of ten algorithms (GATK UnifiedGenotyper, Pindel[49], 1-2-3SV (see URLs), Breakdancer[50], DWAC (see URLs), CNVnator[51], FACADE[52], MATE-CLEVER[53], GenomeSTRiP[54] and SOAPdenovo[55]). These algorithms are based on six approaches: (i) gapped reads, (ii) split reads, (iii) read pairs, (iv) read depth, (v) combined approaches and (vi) *de novo* assembly (**Supplementary Table 12**). Each of the ten tools was run, and calls for each were filtered separately (**Supplementary Note**). Variants were divided into three groups according to their size (1–20 bp, 20–100 bp and >100 bp), and different merging and filtering criteria were applied to obtain the final set (**Supplementary Note**).

Mitochondrial DNA. Unmapped reads were remapped to an appended version of the revised Cambridge Reference Sequence (rCRS)[56,57]. Consensus sequences were called using GATK and used for phylogenetic analyses. All sequences were assigned to haplogroups according to the human mitochondrial DNA phylogeny[58]. Analysis of molecular variance (AMOVA) based on provincial haplogroup frequencies was performed using Arlequin[59] v3.5.1.2 (**Supplementary Fig. 18** and **Supplementary Note**).

Validation of *de novo* variants. A total of 1,133 *de novo* mutations in 54 families were assayed using 3 sequencing technologies (**Supplementary Note**). Variants called using Sanger sequencing were analyzed manually using Phred[60,61]. MiSeq and IonTorrent data were aligned to the reference genome using the BWA and TMAP[62] aligners, respectively, and genotyped with the GATK UnifiedGenotyper. Putative *de novo* indels and structural variants in one trio were selected for validation. *De novo* indel candidates were sequenced using a MiSeq instrument, reads were aligned to the reference genome using BWA, and candidates were genotyped with the GATK HaplotypeCaller. *De novo* structural variant candidates were sequenced with Sanger sequencing, and traces were analyzed by NCBI BLAST[63].

Validation of polymorphisms. We randomly selected 433 deletions and 407 insertions for validation in one family, considering novelty (compared to the 1000 Genomes Project), allele frequency (rare, <0.5%; low frequency, 0.5–5%; common, >5%) and size (short, ≤10 bp; long, >10 bp). Candidates were sequenced using a MiSeq instrument, and reads were aligned to the reference genome, with additional non-reference allele contigs for all candidates larger than (1-kb padding). Indels of <6 bp in length were genotyped using the GATK HaplotypeCaller, and larger indels were genotyped using read counts mapping to the reference and non-reference allele contigs (**Supplementary Note**).

A random set of 96 medium-length (20- to 100-bp) and 48 large (>100-bp) deletions was assayed in one sample by Sanger sequencing (**Supplementary Table 2** and **Supplementary Note**). Sanger sequencing data were called using Phred[60,61] and aligned to the reference genome with NCBI BLAST[63]. Medium-length deletions were also sequenced on a MiSeq instrument, and reads were aligned to the reference genome with additional non-reference allele contigs. Genotyping was based on read counts mapping to the reference and non-reference allele contigs.

Variant annotation. We functionally annotated SNVs with the Variant Annotation Tool[64] (VAT) and SnpEff[65] (**Supplementary Note**), keeping only concordant annotations in coding regions and VAT annotations in noncoding regions (SnpEff provides annotations only for coding regions). Nonsynonymous SNVs were annotated with Polymorphism Phenotyping v2 (PolyPhen-2)[66]. SNVs in OMIM and 'disease-causing' SNVs in HGMD were annotated. SNVs were also annotated with Genomic Evolutionary Rate Profiling[67] (GERP) scores and ancestral and derived allele status. Indels were annotated using indelMapper in VAT with GENCODE v16 annotations (**Supplementary Note**). Structural variants were annotated on the basis of RefSeq[68] annotations, and loss-of-function annotations were defined using MacArthur *et al.*[21] (**Supplementary Note**). Overall and per-sample variant counts stratified by novelty and functional impact were computed for all variants (**Table 1**, **Supplementary Table 13** and **Supplementary Note**).

Loss-of-function analyses. We computed excess loss-of-function mutations per genome on the basis of the expected loss-of-function/synonymous mutation ratio for common SNPs (**Supplementary Note**)[20]. To identify purifying selection of loss-of-function variation, we tabulated counts of loss-of-function versus non-loss-of-function variation, stratified by frequency. We considered loss-of-function SNVs versus synonymous SNVs, loss-of-function indels versus non-frameshift indels and loss-of-function structural variants versus structural variants only removing intronic regions (**Supplementary Fig. 2**). Compound heterozygote loss-of-function events were extracted and stratified by frequency (**Supplementary Table 3**). Residual variation intolerance scores (RVIS)[24] were extracted for genes with loss-of-function variants in GoNL (**Supplementary Table 3**).

***De novo* mutation analyses.** *De novo* mutations were called using GATK PhaseByTransmission and filtered using a random forest machine learning algorithm (**Supplementary Note**). We fit both a linear model and a log-linear model (assuming a Poisson distribution of the residuals) to the number of *de novo* mutations in the offspring given the father's age at conception, conditioning on the depth of coverage of each trio. We also tested the effect of the father's age on the number of *de novo* mutations in the offspring, conditioning on mother's age at conception. Using read-phase information, parent of origin was determined using GATK ReadBackedPhasing, and analyses were repeated on the paternal and maternal *de novo* mutations separately (**Supplementary Fig. 5** and **Supplementary Note**).

Integrated phased panel construction. SNV genotype likelihoods (PLs) from the GATK UnifiedGenotyper were used as input for BEAGLE[69], treating all samples as unrelated to produce a first set of haplotypes. A subset of SNPs (based on the Omni2.5M array) was extracted from the BEAGLE output to construct a phased scaffold using SHAPEIT2 (ref. 70) with trio information. This scaffold was used by MVNcall[19] to phase the remaining SNVs. For chromosome X, we truncated the male PLs in non-pseudoautosomal regions, yielding a negligible heterozygous genotype likelihood.

SNP discovery power and genotype concordance. SNP discovery power was estimated by comparing called sites with SNPs genotyped on the Immunochip (**Supplementary Fig. 19** and **Supplementary Note**). Genotype concordance

doi:10.1038/ng.3021

was assessed by comparing called genotypes against (i) Immunochip genotypes in all samples and (ii) genotypes called by Complete Genomics[71] in 20 overlapping samples.

Imputation in Dutch samples. To evaluate imputation accuracy in Dutch samples, we used a set of 81 Dutch samples sequenced at ~40× coverage using Complete Genomics data from the population-based amyotrophic lateral sclerosis (ALS) study in the Netherlands (PAN)[72]. We used Complete Genomics genotypes at sites overlapping with Illumina Human-1M, 1000 Genomes Project and GoNL (702,253 SNPs) in these samples to impute the other Complete Genomics genotypes using IMPUTE2 (refs. 32,73) with 5 imputation panels: (i) GoNL, (ii) GoNL and 1000 Genomes Project, (iii) GoNL and 1000 Genomes Project EUR, (iv) 1000 Genomes Project and (v) 1000 Genomes Project EUR. Imputation accuracy was measured using the aggregate Pearson's r^2 correlation between the Complete Genomics genotypes and the imputed dosages (**Fig. 3** and **Supplementary Note**).

Imputation with country-specific reference panels. We created 3 country-specific imputation panels of equal size ($n = 85$) using the TSI and CEU samples from the 1000 Genomes Project and GoNL samples. Using the different panels, we imputed into non-overlapping samples sequenced with Complete Genomics ($n = 1$ for TSI; $n = 3$ for CEU; $n = 5$ for NTL) (**Supplementary Fig. 7** and **Supplementary Note**).

Comparison of imputation accuracy using different chips. Using NTL samples, we evaluated GoNL imputation accuracy for five different chips (Illumina-Human1M, HumanExomeCore, HumanOmniExpress, Affymetrix 500k and Affymetrix 6.0) by masking and imputing all variants not on the chip (**Supplementary Fig. 6** and **Supplementary Note**).

Principal-components analysis. We performed 3 sets of PCA using EIGENSTRAT[74] (i) across GoNL and all 14 populations of the 1000 Genomes Project, (ii) GoNL and the 1000 Genomes Project EUR panel, and (iii) GoNL only (**Supplementary Fig. 8**). We computed PCA in scenario (i) for SNPs included on the Omni2.5M chip and with frequency of >5% in each individual population. We removed SNPs with missingness of >0.1% and pruned by linkage disequilibrium the remaining SNPs ($r^2 < 0.3$). We computed PCA in scenario (ii) following the same procedure except that we extracted sites included on the Omni1M chip. PCA in scenario (iii) (**Fig. 4b**) was computed using SNPs in phased haplotypes with a frequency of >10%, no missingness and pruned for linkage disequilibrium ($r^2 < 0.3$). Principal components were calculated in unrelated individuals only, and offspring were projected onto these principal components. We checked for principal-component significance (Tracy-Widom) and for spousal correlation along the top ten principal components (**Supplementary Fig. 20**, **Supplementary Table 14** and **Supplementary Note**).

IBD-based demographic inference. We used phased SNPs (MAF >1%, phasing posterior = 1.0) and kept regions >45 cM in length (**Supplementary Table 15**) with IBD sharing within 5 s.d. of the mean. IBD sharing was inferred using GERMLINE and FastIBD[41,75]. Ancestral population sizes were inferred using the average fraction f of the genome spanned by segments of 1–2 cM in length for a pair of individuals. Recent effective size was inferred using segments >7 cM in length through the estimator $\hat{N} = 50\left(1 - f + \sqrt{1-f}\right)/(uf)$, where μ represents the minimum length in cM[76]. The average time to a common ancestor was estimated using DoRIS[76]. Populations were grouped into northern, central and southern using hierarchical clustering[77] on the basis of sharing of IBD segments >1 cM in length. Demographic models involving a single or multiple populations with migration were analyzed using DoRIS[76,78].

Runs of homozygosity. We used PLINK[79] to find runs of homozygous genotypes (SNPs with MAF of >5% and run length of >500 kb with at most one heterozygote genotype) using sliding windows of 5 Mb in unrelated GoNL and 1000 Genomes Project samples (**Supplementary Fig. 21**). We performed analysis of variance (ANOVA) to compare means between Dutch regions (northern, central and southern) on the basis of 1,000 bootstrap samples (**Supplementary Note**).

Population differentiation (F_{ST}). We computed Hudson's F_{ST} and Weir and Cockerham's (WC) F_{ST} between GoNL and each 1000 Genomes Project population. For the WC estimate, we calculated F_{ST} from allele frequency data using correction for small sample size[80]. We calculated Hudson's F_{ST} estimate on two different sets of SNPs, using (i) the 1000 Genomes Project YRI (Yoruba in Ibadan, Nigeria) population as an 'outgroup' population and (ii) sites polymorphic in the YRI population and in both populations for which the F_{ST} was calculated (**Supplementary Table 5** and **Supplementary Note**). We also computed Hudson's F_{ST} between (i) the 11 Dutch provinces in GoNL and (ii) between the 3 regions (northern, central and southern) used in the IBD and f_2 analyses (**Supplementary Table 16** and **Supplementary Note**).

Rare variant f_2 sharing analysis. We performed an interpopulation f_2 analysis by merging 1000 Genomes Project data and 88 random samples from GoNL (matching 1000 Genomes Project European population size) evenly distributed among the 11 provinces (**Supplementary Fig. 9** and **Supplementary Note**). We performed f_2 analysis using 330 GoNL samples selected evenly across the 11 provinces (**Fig. 4**). Provinces were then grouped into three regions (northern, central and southern), and proportions of f_2 sharing between provinces and regions were tested using a chi-squared test (**Supplementary Note**).

Singleton analysis. A filtered set of singletons was extracted from the SNPs. To account for sequencing biases, we computed the residuals of the following generalized linear regression (GLM) model: singletons per individual ~sequencing batch + biobank + depth of coverage + transmitted singletons. To investigate for possible geographic differences in genic singletons, we computed the Pearson's correlation between the principal components and (i) the singleton counts and (ii) the residuals of the GLM (**Supplementary Fig. 22**).

Impact of indels and structural variants. We used 1,000 permutations to compute differences in the distribution of indels and structural variants with respect to intergenic, intronic, exonic, OMIM and loss-of-function annotations (**Supplementary Table 17**).

47. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
50. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
51. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
52. Coe, B.P., Chari, R., MacAulay, C. & Lam, W.L. FACADE: a fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data. *Nucleic Acids Res.* **38**, e157 (2010).
53. Marschall, T., Hajirasouliha, I. & Schönhuth, A. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* **29**, 3143–3150 (2013).
54. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
55. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
56. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
57. Andrews, R.M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147 (1999).
58. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–E394 (2009).
59. Excoffier, L. & Lischer, H.E.L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
60. Ewing, B., Hillier, L., Wendl, M. & Green, P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
61. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
62. Wijaya, E., Frith, M.C., Suzuki, Y. & Horton, P. Recount: expectation maximization based error correction tool for next generation sequencing data. *Genome Inform.* **23**, 189–201 (2009).
63. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

64. Habegger, L. *et al.* VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28**, 2267–2269 (2012).
65. Reumers, J. *et al.* SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res.* **33**, D527–D532 (2005).
66. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit 7.20 (2013).
67. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
68. Pruitt, K.D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
69. Browning, B.L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.* **85**, 847–861 (2009).
70. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
71. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
72. Huisman, M.H.B. *et al.* Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J. Neurol. Neurosurg. Psychiatry* **82**, 1165–1170 (2011).
73. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457–470 (2011).
74. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
75. Browning, B.L. & Browning, S.R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88**, 173–182 (2011).
76. Palamara, P.F. & Pe'er, I. Inference of historical migration rates via haplotype sharing. *Bioinformatics* **29**, i180–i188 (2013).
77. Ward, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
78. Palamara, P.F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
79. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
80. Cockerham, C.C. & Weir, B.S. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40**, 157–164 (1984).

doi:10.1038/ng.3021