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

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 to be studied. Here, we describe the Genome of the Netherlands ("GoNL") Project, in which we sequenced whole genomes of 250 parent-offspring families from the Netherlands and constructed a phased haplotype map of 20.4 million single nucleotide variants and 1.2 million short insertions and deletions (indels). The intermediate coverage (~13x) and the trio design enabled extensive characterization of structural variation, including mid-size events (30-500 bp) previously poorly catalogued, and of *de novo* point mutations and larger structural changes. We demonstrate that the quality of the haplotypes significantly boosts imputation accuracy in independent samples, especially for lower frequency alleles (<5%). Population genetic analyses demonstrate fine-scale structure across the country and support multiple migration events in the past (consistent with historical sea level changes and floods in northern/westerns parts of the country) but relatively little movement in recent times. The GoNL Project illustrates how whole-genome sequencing can provide detailed characterization of genetic variation—inherited and *de novo*—within a relatively homogeneous population, and may guide the design of future population studies.

Introduction

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 single-nucleotide polymorphisms (SNPs) in various populations.[2-5] Fueled by the commercial development of microarrays for efficient SNP genotyping and collaborative projects in large samples, genome-wide association studies (GWAS) have provided a systematic approach to test genetic variants for a role in disease. To date, GWAS have reproducibly identified thousands of genomic 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, in particular with regard to the relative 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, not only single nucleotide variants (SNVs), but also structural DNA variations and *de novo* events.[10-12] The characterization of rare variation poses a major challenge. Since rare alleles have emerged, on average, relatively recently,[13] they show much greater geographic clustering[14] and are more susceptible to population stratification compared to common variation.[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—common and rare—for SNVs, short insertions and deletions (indels) and larger deletions in 769 individuals of Dutch ancestry selected from five 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 provinces of the Netherlands without ascertaining on phenotype or disease. (The twelfth province, Flevoland, was excluded from sampling as it was established by land reclamation in the 20th century.) By whole-genome sequencing these 250 families at ~13x coverage, our aim was to build a resource of 1,000 haploid genomes as representative for a small (41,543 km^2) but densely populated country (> 17 million inhabitants) in northwestern Europe (**Supplementary Note**).

Here, we provide the first detailed analysis of the GoNL data set after data 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) and 27,500 larger deletions (> 20 bp). Of the SNVs, 6.2 million are common (MAF > 5%), 4.0 million are low-frequency (MAF 0.5–5%), and 10.2 million are rare (MAF < 0.5%). Based on 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 of 1% frequency should be detectable (**Supplementary Note**). Indeed, we discovered 18.9M SNVs and 1.1M indels in the accessible, autosomal genome (**Supplementary Table 1**). Indels and large deletions were based on conservative consensus calls from several complementary methods that use information about mapped reads, read depth, read pair, split reads and *de novo* assembly (**Supplementary Note**). We used MVNcall for trio-aware phasing and linkage

disequilibrium-based imputation,[19] starting from the genotype likelihoods of SNVs and indels, which yielded a fully phased panel of 998 unique haplotypes after QC. 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 in all GoNL samples). The average coverage of 13.3x coupled with the family-based design allowed us to construct a high-quality whole-genome data set for further analysis, including characterization of structural variation, detection of *de novo* events, imputation, and demographic inference.

Novel variation in GoNL

To determine the number of novel variants, we investigated the overlap between GoNL and existing databases. We rediscovered almost all sites (98.2%) in the European sample (CEU) of HapMap Phase 2,[4] and 71.1% of sites in the European subset of the 1000 Genomes Project Phase 1 (1KG-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 SNVs observed in GoNL at 0.5% frequency or below (but excluding singletons) were observed in 1KG-EUR, 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 of dbSNP (build 137), although the majority (75.6%) of these novel variants are singletons. Considering that 16.5% of 2.0 million singletons in 1KG-EUR 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 within the Netherlands and 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 size spectrum shows

that larger structural events are less frequent than smaller indels, presumably reflecting their relative deleterious nature. We recognized specific peaks in the size spectrum that correspond to microsatellite instability (MSI) around 4 bp, short interspersed elements (SINEs) at 300 bp, and long interspersed elements (LINEs) at 6 kb. In general, we detected a large number of novel deletions throughout the size spectrum. In comparison to 1KG Phase 1, 54.4% of the indels (< 20 bp) and 93.3% of the larger deletions (> 20 bp) are novel (**Supplementary Note**). Taking advantage of the medium coverage data and using especially tuned methods (PINDEL, CLEVER), our analysis fills an important gap for the discovery of mid-size deletions (30–500 bp) where essentially all (98.4%) of the observed variants are novel. The novelty rate for deletions larger than 500 bp is still substantial (66.3%). We note that most of the deletions reported here are biased to be common because stringent filtering required variants to be present in at least three families and transmitted to at least one offspring. This strategy allowed us to generate a specific call set with an overall validation rate of 96.5% (**Supplementary Table 2**). A more sensitive and complete data set including duplications, inversions, mobile element insertions, and translocations is currently being assembled and validated.

Functional variation

Predicting the biological consequences of functional variants within a single genome is still an unresolved challenge with important implications for using next-generation sequencing in a clinical setting. To characterize the burden of loss-of-function (LoF) variants in detail, we classified all such variants in GoNL according to LoF definitions described by MacArthur et al.[21] and applied various filters to remove potential annotation errors (**Supplementary Note**). Amongst rare variants, we observed an excess of probably or possibly damaging missense variants (according to PolyPhen-2[22]), nonsense variants, and frameshift indels, consistent with a model in which such functional

variants are subject to purifying selection (**Supplementary Fig. 2**).[23,24] Of the larger deletions, we counted 66 LoF deletions that eliminate the first exon of a gene or more than half of its coding sequence,[21] and found evidence they were depleted relative to all deletions ($p = 0.005$ for 20–100 bp; $p = 2.6 \times 10^{-9}$ for > 100 bp). This effect was amplified when considering only genes listed in the OMIM compendium ($p = 2.4 \times 10^{-27}$), illustrating a strong selective force against large structural changes in key genes.

The overall patterns and per-individual distributions of LoF SNVs (premature stops or variants interrupting a splice site) and missense variants are consistent with those found in 1KG (**Table 1, Supplementary Fig. 3**). On average, an individual carries 60 LoF SNVs, 69 LoF frameshift indels, and 15 LoF large deletions. The bulk of these LoF mutations per individual are common (>5% frequency in the population), suggesting that these variants are not subject to strong selective pressure and, though they are protein-truncating mutations, are likely phenotypically benign. This emphasizes the need for caution in assigning pathogenicity to variants purely on the basis of their predicted impact on protein structure.

In contrast, when we considered only rare LoF variants, those more likely to be pathogenic, we found that the average individual in GoNL carries 4 nonsense variants, 2 variants interrupting a splice site, and 2 frameshift indels. By comparing these numbers to synonymous variants (which provides a baseline expectation under neutrality), we estimate that each individual carries an excess of 4–5 rare LoF SNVs (**Supplementary Note**), which are sufficiently deleterious that they will never reach high frequency in the population at large.

We also investigated the number of compound heterozygotes of rare LoF SNVs, short indels and large deletions. Across all genes in the genome, we observed three compound heterozygous events mapping to three genes in three separate individuals (average 0.01 compound heterozygous events per individual). The phenotypic impact of compound heterozygotes of rare LoF mutations demands further characterization and should be considered explicitly in rare variant studies.

Whereas compound heterozygotes of rare LoF variants are sparse, we expected compound heterozygotes of common LoF variants to be more prevalent assuming these variants are less likely to be deleterious. Indeed, we found that the average number of common-LoF compound heterozygotes per individual increased to 2.89 (range: 0–7). Interestingly, while overall there are 1,917 common-LoF compound heterozygous events across all GoNL samples, these are confined to only 11 genes (**Supplementary Table 3a**). All but one of these genes appear to have extreme Residual Variation Intolerance Scores (all >84th percentile across 16,956 genes), which is unlikely to occur by chance ($p = 1.41 \times 10^{-5}$, **Supplementary Fig. 4a**).[25] This suggests that these genes are more tolerant of disruptive mutations.

Because disease mutation databases are often employed to identify potential variants of interest, we annotated variants in GoNL listed as disease-causing ("DM") in the Human Gene Mutation Database (HGMD).[26] We observed that an individual in GoNL carries on average 20 such "DM" variants (range: 9–33) (**Table 1**). Since all GoNL samples were derived from population-based cohorts, the impact of these alleles is unclear. One possibility is that the presence of modifier alleles gives rise to incomplete penetrance or variable expressivity of "DM" variants depending on the particular background of the carrier.[27] An alternative explanation is that, in fact, HGMD contains a

considerable number of false-positive disease-causing mutations that need further scrutiny.[28] Of the 1,093 "DM" mutations occurring in GoNL, 32% have a frequency >1%, which is much higher than the prevalence of many of the rare diseases described in HGMD. Following 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, Supplementary Table 4**). In fact, one of these variants (chr14:94847262, an alpha 1 antitrypsin deficiency variant) was recently implicated as a pathogenic incidental finding in a set of 1,000 exomes after undergoing stringent filtering including in-depth literature reviews of the original findings.[29] The prevalence of alpha 1 antitrypsin deficiency (OMIM: 613490), an autosomal recessive disease, is estimated to be 0.02-0.06%, yet two unrelated GoNL individuals are homozygous carriers of this variant (prevalence = 0.4%, ~10x higher than the disease prevalence). Further, the typical age of onset of alpha 1 antitrypsin deficiency is between 20 and 50 years old, whereas the two homozygous carriers in GoNL were ages 60 and 63 at ascertainment. These results highlight the potential pitfalls of employing such databases in disease studies and the difficulty of interpreting personal genomes.

De novo mutations

A distinct advantage of the family-based study design was 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).[30] From an initial 4.5 million Mendelian violations in the original Unified Genotyper calls made in the 258 independent offspring, we prioritized 58,569 putative *de novo* mutation (DNM) candidates with PBT (**Supplementary Note**). After removing polymorphic sites, we counted a total of 29,162 autosomal DNM candidates, still many more than the expected average of 16,306, assuming 63.2 mutations per offspring.[31] To reduce false positives, we

evaluated to what extent sequencing features could help increase the DNM prediction accuracy. We selected 1,060 candidate sites across a broad confidence range and validated 569 sites of these as true DNMs using various technologies. We also classified another 1,139 candidates as false positives on the basis of Complete Genomics genotype data in 19 parents and 1 unrelated child. We trained a random forest classifier using 70% of the combined validation results on various features (**Supplementary Note**), and obtained a model with an estimated classification accuracy of 92.2% based on the remaining 30% of the data (**Fig. 2a**). This illustrates that the joint assessment of raw trio data and sequencing context can greatly boost prediction accuracy. After applying this classifier to our initial candidates, we obtained 11,258 high-confidence DNMs (with a range of 18–74 DNMs per individual) that we used for downstream analyses. Due to coverage fluctuating regionally, we expect a substantial fraction of genuine DNMs to be missed. We 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 the father's age at conception and the number of DNMs in the offspring (**Fig. 2b**), providing a third, independent estimate based on a larger sample size.[31,32] Accounting for a Poisson distributed background mutation rate and correcting for coverage, we estimate that each additional year of father's age is associated with a 2.5% increase of the mean number of DNMs. While parents' ages are highly correlated ($r^2 = 0.66$), comparing regression models based on the father's and mother's age at conception suggests that the observed age-related increase in DNMs is a predominantly paternal effect (**Supplementary Note**). Interpolating from the paternal model, we expect on average 75.4% of the DNMs in the GoNL offspring to derive from the father (assuming a linear increase in DNMs from puberty). Using read-pair information we were able to assign

parental origin to 1,321 DNMs, and found that indeed 74.5% were paternal in origin. When considering only mutations for which parental origin could be determined, the correlation with father's age remained significant (r^2 = 0.56, p = 0.012) but not with mother's age (p = 0.94) (**Supplementary Fig. 5**). The striking consistency of these results highlights the relative impact of paternal and maternal mutations.

Within a single family, we also tested if we could discover *de novo* indels and large deletions. Using strict filtering criteria for Mendelian violations followed by PCR-based Sanger sequencing, we confirmed 6 intergenic indels (4 insertions of 1 bp, 1 deletion of 1 bp and 1 deletion of 2 bp), as well as a large 113 kb deletion located in an intron of the *SUMF1* gene at chr3p26 (which seems unlikely to have a significant impact on gene function). These results illustrate that our predictions of indels and structural changes are a valuable source not only for commonly segregating alleles, but also for *de novo* events. Further work is needed to assess the frequency of such *de novo* deletions or other types of genomic rearrangements 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 Complete Genomics sequence data collected in 81 individuals of Dutch ancestry, independent from the GoNL samples, which we will refer to as the NTL data set. In these NTL samples, we masked all genotypes at SNVs not present on the Illumina Human-1M array, imputed these "non-genotyped" SNVs from the 1M-genotyped SNVs, and then compared the imputed to the known sequenced genotypes (**Supplementary Note**). The aggregate mean r^2 was 0.99 for common SNVs, 0.86 for low-frequency SNVs and 0.63 for rare SNVs, indicating that the overall quality was good (**Fig. 3**). We repeated

this evaluation based on the SNV content of other microarrays and obtained similar imputation performance for common SNVs, although there were significant differences for lower frequency alleles (**Supplementary Fig. 5**). To directly measure the impact of trio-based phasing we constructed a panel based on the unrelated parents alone, and re-evaluated the imputation quality in the NTL samples. The imputation accuracy dropped to a mean r^2 of 0.47 for rare variants (0.85 for low-frequency SNVs and 0.98 for common SNVs), indicating that the trio-based phasing contributed significantly to the imputation quality, especially for rare variants.

In comparison to 1KG as a reference panel, we observed better imputation accuracy with the GoNL panel for SNVs up to 10% frequency despite the larger sample size of 1KG (**Fig. 3**). To investigate the basis for the improved imputation accuracy with the GoNL panel, we constructed three reference panels based on 1KG-CEU (Northern Europeans from Utah), 1KG-TSI (Tuscans from Italy), and GoNL, all with a fixed sample size of 85 individuals. With each of these reference panels, we imputed in independent CEU, TSI and NTL samples with Complete Genomics data, and then evaluated their performance (**Supplementary Note**). Of the three panels, GoNL gave the highest imputation accuracy (especially for rare variants) not only for the Dutch NTL samples but also for the CEU samples, ruling out that the improved performance of the GoNL panel was simply due to shared Dutch ancestry of GoNL and NTL samples (**Supplementary Fig. 6**). Differentiation between northern and southern European populations may explain why the 1KG-CEU and GoNL panels gave roughly equivalent performance for TSI (but certainly worse than 1KG-TSI). Overall, these results suggest that the GoNL trios have enabled accurate reconstruction of long-range haplotypes with an advantageous effect on the imputation of rare alleles.

To assess the potential value of much larger reference panels, we combined the 1KG and GoNL panels with IMPUTE2,[33] and evaluated again the imputation accuracy in the NTL samples. Here we obtained an additional gain in imputation accuracy over the GoNL panel alone, reaching a mean r^2 of 0.70 for rare SNVs and 0.88 for low-frequency SNVs (**Fig. 3**). Thus, increasing the sample size of the reference panel will likely continue to improve imputation performance (especially for lower frequency alleles), which motivates a community-wide effort to create a unified reference panel across multiple ethnicities and populations.

Population structure and demographic inference

Although it is well understood that extensive migration and gene flow occurred amongst European populations,[34-36] we focused on creating a unified picture of Dutch demography in recent millennia. Because of unbiased ascertainment and 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 and the 14 populations represented in 1KG Phase I and found that F_{ST} patterns were consistent with continental clustering in principal component analysis (PCA) and with previous estimates (**Supplementary Table 5, Supplementary Fig. 7)**.[37] Among the European populations, the Dutch samples clustered best with 1KG samples from Great Britain and Northern Europe (F_{ST} = 0.0008 and 0.0006, respectively) and least with the Finnish (F_{ST} = 0.0068). To investigate more recent population connections, we focused on so-called f_2 variants that appear exactly twice (in two heterozygote carriers) in the joint data of GoNL and 1KG (**Supplementary Note**). As was observed in 1KG, within-

population f_2 sharing accounts for the majority (50.8%) of all f_2 alleles (**Supplementary Fig. 8**), but f_2 sharing revealed 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 of European descent (71.6%) or from the Americas (21.0%, due to substantial European admixture) than with an individual coming from Africa (6.2%) or East Asia (1.3%). These results underscore the high degree of geographic clustering of recent mutations within neighboring populations. Analysis of maternally inherited mitochondrial DNA (using 492 GoNL parent individuals) (**Supplementary Information**) revealed that the major haplogroups are H (39.4%), U (25.2%), J (10.4%) and T (10.8%), and the minor haplogroups are HV0 (4.9%), N1 (3.5%), W (3.3%), X (2.4%) and HV1 (0.2%), a composition that is in good agreement with previous observations in other European populations.[38]

Within the Netherlands (**Fig. 4a**), PCA revealed subtle substructure along a North–South gradient on the first two principal components (**Fig. 4b** and **Supplementary Note**), consistent with previous findings.[39,40] Because PCA has limitations in terms of demographic inference (in particular for migration patterns),[41] we also performed an independent analysis of identity-by-descent (IBD) sharing that revealed subtle signals of migration (**Supplementary Note**).[42] From the length distributions of the IBD segments,[43] we inferred demographic models and estimated effective population sizes of the Dutch provinces at different time scales, reflecting demographic changes throughout history (**Supplementary Note**).

Analysis of IBD segments of 1–2 cM (**Supplementary Fig. 9a**), corresponding to an estimated time-to-most-recent-common-ancestor ≈ 4,000 years, revealed rather homogeneous effective population sizes across the 11 provinces, consistent with common

genetic origins. Additionally, we observed a smooth south-to-north gradient of decreasing ancestral population size, accompanied by increased homozygosity in the northern provinces (correlation between average IBD sharing within provinces and latitude: $r = 0.952$, $p = 6 \times 10^{-6}$; **Supplementary Fig. 10**, **Supplementary Note**). Traditionally, this observation has been explained by a serial founder effect characterized by migration from the south to the north.[39]

Interestingly, GoNL samples, regardless of place of birth, tend to share more IBD segments with other individuals from the north of the country than with individuals from the same geographic region. In fact, although within-province IBD sharing is strong (diagonal values of the heat map), the excess sharing with the northern provinces (off-diagonal values for FR, GR, DR, OV, and NH) is evident (correlation between average province-province IBD sharing and average province latitude: $r = 0.943$, $p = 5 \times 10^{-7}$; **Fig. 4c, Supplementary Table 6a**). This pattern indicates that a simple south-to-north serial founder model is not sufficient to explain the observed IBD sharing. Grouping the provinces into three clusters (North, Center and South), we reconstructed possible coalescent time distributions within and across the clusters. Based on the reconstructed coalescent rates, the average individual from the Center or the South of the country is expected to co-inherit more IBD segments with the average northern individual in the past 4,000 years than with other individuals from the same geographic region (**Supplementary Fig. S11**). While different founder effect patterns emerge from these simulations, they all show support for a substantial amount of regional migration within the Netherlands. Assuming ancient serial migrations towards the North are causing 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 B.C.E. and 50 C.E. (**Supplementary Fig. 12**). More than half of the current Dutch territory is below sea level and a series of abandonments and resettlements of different regions were likely prompted by ocean level shifts and flooding that changed once-habitable land into dunes and marshes or buried regions under water entirely. We emphasize that other more complex demographic models may 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 13th century) increased land stability, which resulted in other forces influencing demography. An analysis of f_2 variants revealed non-random sharing within and across provinces. Though the proportion of within-province f_2 sharing comprises only 12% of all f_2 alleles, consistent with a homogeneous population, the proportion of within-province f_2 alleles is significantly larger than expected under the null hypothesis of uniform allele sharing across all provinces (**Fig. 4d**). This geographic localization of rare variants is suggestive of limited migration in recent centuries, which is consistent with about half of Dutch citizens still living in the same province in which their great-grandparents were born. Notably, Noord-Brabant and Overijssel show significantly stronger within-province f_2 sharing in comparison to the other provinces ($p = 1.2 \times 10^{-151}$, $p < 10^{-200}$, respectively), which is in agreement with small effective population sizes in these two provinces inferred from long (7–15 cM) IBD segment sharing (**Fig 4c**, **Supplementary Table 6b**, **Supplementary Fig. 9b**). Further, we found that within-region sharing in the northern and southern regions was substantially stronger when compared to the central regions ($p < 10^{-200}$, both comparisons). Altogether, these results suggest increased migration in the central region (as compared 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 the cosmopolitan, low-coverage effort by 1KG. 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 discovery of all types of DNA variation, but also enable population genetic analyses that can shed more light on the shared ancestry of the human species.

In spite of the intermediate coverage, we were able to reliably call *de novo* point mutations and confirm the relationship between paternal age and mutation load. Although 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 DNM discovery should be broadly applicable for disease studies where DNMs are suspected to play a role.[12] DNM represents an important class of DNA sequence variation that can further elucidate fundamental processes of mutagenesis, even if its absolute contribution may be limited in terms of explaining heritability (depending on the disease).[44] In cancer, for example, accounting for the genome-wide heterogeneity of mutation rates may be necessary to accurately pinpoint driver mutations against a background of random mutation.[45] Our results suggest 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.

As long as the cost of genotyping continues to be competitive with whole-genome sequencing, imputation on the basis of linkage disequilibrium 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[46,47] should therefore be considered a top priority. Through more complete interrogation of genetic variation, studies of large, well-phenotyped samples will continue to increase opportunities for development of diagnostic tools, prevention measures and therapeutics for human disease.

Methods Summary

All details concerning sample collection, data generation, processing and analysis can be found in the Supplementary Note.

References

1 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
2 Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-1079, doi:10.1126/science.1105436 (2005).
3 A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).
4 Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:10.1038/nature06258 (2007).
5 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).
6 Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* **14**, 549-558, doi:10.1038/nrg3523 (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, doi:10.1016/j.ajhg.2011.11.029 (2012).
8 McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210-217, doi:10.1016/j.cell.2010.03.032 (2010).
9 Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-145, doi:10.1038/nrg3118 (2011).
10 Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* **14**, 460-470, doi:10.1038/nrg3455 (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, doi:10.1038/nrg3373 (2013).
12 Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat Rev Genet* **13**, 565-575, doi:10.1038/nrg3241 (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, doi:10.1038/nature11690 (2013).
14 Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* **108**, 11983-11988, doi:10.1073/pnas.1019276108 (2011).
15 Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44**, 243-246, doi:10.1038/ng.1074 (2012).
16 Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet*, doi:10.1038/ejhg.2013.118 (2013).
17 Brandsma, M. *et al.* How to kickstart a national biobanking infrastructure – experiences and prospects of BBMRI-NL. *Norsk Epidemiologi* **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, doi:10.1038/ng.806 (2011).

19 Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84-91, doi:10.1093/bioinformatics/bts632 (2013).
20 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
21 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).
22 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).
23 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 (2012).
24 Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* **44**, 623-630, doi:10.1038/ng.2303 (2012).
25 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, doi:10.1371/journal.pgen.1003709 (2013).
26 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13, doi:10.1186/gm13 (2009).
27 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, doi:10.1007/s00439-013-1331-2 (2013).
28 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, doi:10.1002/humu.22375 (2013).
29 Dorschner, M. O. *et al.* Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes. *Am J Hum Genet* **93**, 631-640, doi:10.1016/j.ajhg.2013.08.006 (2013).
30 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
31 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475, doi:10.1038/nature11396 (2012).
32 Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-1442, doi:10.1016/j.cell.2012.11.019 (2012).
33 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, doi:10.1371/journal.pgen.1000529 (2009).
34 Lao, O. *et al.* Correlation between genetic and geographic structure in Europe. *Current biology : CB* **18**, 1241-1248, doi:10.1016/j.cub.2008.07.049 (2008).
35 Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98-101, doi:10.1038/nature07331 (2008).
36 Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol* **11**, e1001555, doi:10.1371/journal.pbio.1001555 (2013).
37 Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res* **23**, 1514-1521, doi:10.1101/gr.154831.113 (2013).
38 Zheng, H. X., Yan, S., Qin, Z. D. & Jin, L. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Scientific reports* **2**, 745, doi:10.1038/srep00745 (2012).

39 Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet*, doi:10.1038/ejhg.2013.48 (2013).
40 Lao, O. *et al.* Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history. *Investigative genetics* **4**, 9, doi:10.1186/2041-2223-4-9 (2013).
41 Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* **40**, 646-649, doi:10.1038/ng.139 (2008).
42 Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19**, 318-326, doi:10.1101/gr.081398.108 (2009).
43 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, doi:10.1016/j.ajhg.2012.08.030 (2012).
44 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, doi:10.1038/ng.2555 (2013).
45 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).
46 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, doi:10.1038/ng.2334 (2012).
47 Jia, X. *et al.* Imputing amino Acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683, doi:10.1371/journal.pone.0064683 (2013).

Acknowledgements

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; Mark Depristo, Eric Banks, Ryan Poplin, Guillermo del Angel from the Broad Institute for expert advice on setting up our alignment and calling pipeline; Kiran Garimella for the initial implementation of PhaseByTransmission; Ger Strikwerda, Wietze Albers, Robin Teeninga, Hans Gankema, and Haije Wind of the Groningen Center for Information Technology (http://www.rug.nl/cit) for support of the compute cluster and Target storage; Edwin Valentyn and Reese Williams of the Target project (http://www.rug.nl/target) for hosting project data on IBM GPFS storage; Tom Visser and Irene Nooren of BiG Grid (http://biggrid.nl) and SURFsara for providing backup storage, additional compute capacity and expert advice; the MOLGENIS team for software development support (http://www.molgenis.org); Hélène Lauvenberg for handling data access requests; Konrad Zych for the GoNL logo design; Lude Franke, Harm-Jan Westra, Javier Gutierrez-Achury for useful discussions, and Soumya Raychaudhuri and Ben Neale for their critical reading of the manuscript. Target is supported by Samenwerkingsverband Noord Nederland, European Fund for Regional Development, 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). AA is funded by the Centre for Medical Systems Biology-2 and DIB by the European Research Council

(ERC 230374). AS and PIWdB are recipients of a VIDI Award (NWO projects 016.138.318 and 016.126.354, respectively).

Author Contributions

PIWdB and MAS co-led the analysis group. FvD, PBTN, PD, LCF, AK, MD, HB, KJvdV, MAS formed the operational data stewardship and processing center. PBTN, FvD and MAS designed and implemented the compute cluster. MD, HB, AK and MAS designed and implemented the MOLGENIS compute platform to scale up analysis pipelines for alignment, variant calling, and imputation. FvD, LCF performed alignment with help from IJN, JB, BDCvS. LCF and FvD called SNVs. LCF, SLP, AM, EMvL, LCK, MSohail, AA, MV performed quality control. VG, KY, LCF, TM, AS, REH, SAM, WPK, FH, JYHK, EWL, AA, VK, HM, MHM, JB formed the structural variation subgroup. LCF developed the PhaseByTransmission module in the GATK and performed with PP *de novo* mutation analyses. AM performed haplotype phasing and imputation benchmarks. JHV and LHvdB provided Complete Genomics validation data. WPK and IR performed variant validation. CW and MP generated Immunochip data on all GoNL samples. SLP, CCE, AM, PFP, IP, AA, NA, MSohail, SRS performed population genetic analyses. MvO, MV, ML, JFJL, MStoneking, PdK, MKaiser performed mtDNA analysis. PD, AM, AK, EMvL, LCK, KE, MCMG, JvS, MKattenberg, JJH, DvE formed the imputation subgroup. PBTN, KJvdV and MAS were responsible for GoNL website and associated services (http://www.nlgenome.nl). CW conceived the GoNL Project. PIWdB drafted the initial manuscript. CW, DIB, GJBvO, LCK, AA, MAS, PES, SRS, JYHK, IP, JHV, PdK, WPK, TM, AS, VG, JTdD, MKaiser provided critical feedback on the manuscript. All authors have seen and approved the final manuscript.

Author Information

Raw sequence reads, aligned reads, variant calls, inferred genotypes, and phased haplotypes will be made available through the Database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) and the European Genome-phenome Archive (EGA access number EGAS00001000644; http://www.ebi.ac.uk/ega/). Data access requests can also be submitted through the GoNL website (http://www.nlgenome.nl/). Correspondence should be addressed to Paul de Bakker (pdebakker@umcutrecht.nl) or Cisca Wijmenga (c.wijmenga@umcg.nl).

The Genome of the Netherlands Consortium

Laurent C. Francioli[1,36], Androniki Menelaou[1,36], Sara L. Pulit[1,36], Freerk van Dijk[2,3,36], 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], Mingkun Li[12], Jeroen F.J. Laros[11], 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[13], Jan Bot[14], Isaäc J. Nijman[1], Ivo Renkens[1], Tobias Marschall[15], Alexander Schönhuth[15], Jayne Y. Hehir-Kwa[16], Robert E. Handsaker[17,18], Paz Polak[19], Mashaal Sohail[19], Dana Vuzman[19], Fereydoun Hormozdiari[20], David van Enckevort[21], Hailiang Mei[21], Vyacheslav Koval[22], Matthijs H. Moed[6], K. Joeri van der Velde[2,3], Fernando Rivadeneira[22], Karol Estrada[22], Carolina Medina-Gomez[22], Aaron Isaacs[9], Steven A. McCarroll[17,18], Margreet Brandsma[23], Marian Beekman[6], Anton J.M. de Craen[6], H. Eka D. Suchiman[6], Albert Hofman[9], Ben Oostra[24], André G. Uitterlinden[22], Gonneke Willemsen[8], LifeLines Cohort Study, Mathieu Platteel[2], Jan H. Veldink[25], Leonard H. van den Berg[25], Steven J. Pitts[26], Shobha Potluri[26], Purnima Sundar[26], David R. Cox[26,35], Shamil R. Sunyaev[17,19], Johan T. den Dunnen[11], Mark Stoneking[12], Peter de Knijff[27], Manfred Kayser[10], Qibin Li[28], Yingrui Li[28], Yuanping Du[28], Ruoyan Chen[28], Hongzhi Cao[28], Ning Li[29], Sujie Cao[29], Jun Wang[28,30,31], Jasper A. Bovenberg[32], Itsik Pe'er[4,33], P. Eline Slagboom[6], Cornelia M. van Duijn[9], Dorret I. Boomsma[8], Gertjan B. van Ommen[23], Paul I.W. de Bakker[1,17,19,34,37], Morris A. Swertz[2,3,37], Cisca Wijmenga[2,3,37]

Affiliations

[1] Department of Medical Genetics, 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, NY, USA

[5] The Genome Institute, Washington University, St. Louis, MO, 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 of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[8] Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands

[9] Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands

[10] Department of Forensic Molecular Biology, Erasmus Medical Center, Rotterdam, The Netherlands

[11] Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

[12] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[13] Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Medical Center, Amsterdam, The Netherlands

[14] SURFsara, Science Park, Amsterdam, The Netherlands

[15] Centrum voor Wiskunde en Informatica, Life Sciences Group, Amsterdam, The Netherlands

[16] Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

[17] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

[18] Department of Genetics, Harvard Medical School, Boston, MA, USA

[19] Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[20] Department of Genome Sciences, University of Washington, Seattle, WA, USA

[21] Netherlands Bioinformatics Centre, Nijmegen, The Netherlands

[22] Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands

[23] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

[24] Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands

[25] Department of Neurology, Brain Center Rudolph Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

[26] Rinat-Pfizer Inc, South San Francisco, CA, USA

[27] Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

[28] BGI-Shenzhen, Shenzhen, China

[29] BGI-Europe, Copenhagen, Denmark

[30] Department of Biology, University of Copenhagen, Copenhagen, Denmark

[31] The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark

[32] Legal Pathways Institute for Health and Bio Law, Aerdenhout, The Netherlands

[33] Department of Systems Biology, Columbia University, New York, NY, USA

[34] Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands

[35] Deceased

[36] These authors contributed equally

[37] These authors contributed equally

Figure legends

Figure 1 | Discovery of SNVs and structural variation. **a**, Venn diagram of all SNVs discovered in GoNL relative to dbSNP (build 137), 1KG Phase 1 and HapMap-CEU. The majority of the 7.6 million novel sites are rare (MAF < 0.5%), including 5.8 million singletons. **b**, Size spectrum of structural variation discovered in GoNL. Our detection strategy employed multiple approaches and provided a significant boost in novel SVs in the midsize range (30–500 bp). Peaks corresponding to long interspersed elements (LINEs), short interspersed elements (SINEs) and microsatellite instability (MSIs) are highlighted. The total number of variants called in GoNL are shown in orange, whereas SNVs found in dbSNP (build 137) and short indels and large deletions found in 1KG Phase 1 are shown in blue. For large deletions (> 20 bp), we required at least 80% reciprocal overlap between variants to be considered as similar.

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 line, evaluation subset only, 657 sites). The highlighted circle is the cutoff we used for our analyses with 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 = 0.59, $p < 2.2 \times 10^{-16}$), with the least-squares fit plotted in orange.

Figure 3 | Imputation accuracy. The aggregate r^2 between imputed and gold-standard genotype dosages 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 truth. The GoNL panel consistently outperforms the 1KG panels, especially at lower allele frequencies. A combined GoNL+1KG panel provides the best performance.

Figure 4 | Population genetic analyses in the Dutch population. a, Map of the Netherlands with its 12 provinces. We selected 769 individuals from five BBMRI-NL biobanks across all provinces except Flevoland. **b**, Principal component analysis. Individuals are projected onto the two dominant principal components, revealing 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), 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. Long IBD segment patterns are consistent with restricted geographic movement in recent times. **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 amongst northern or southern provinces is statistically significant compared to central provinces ($p < 10^{-200}$).

Table 1 | Individual variant load of coding mutations

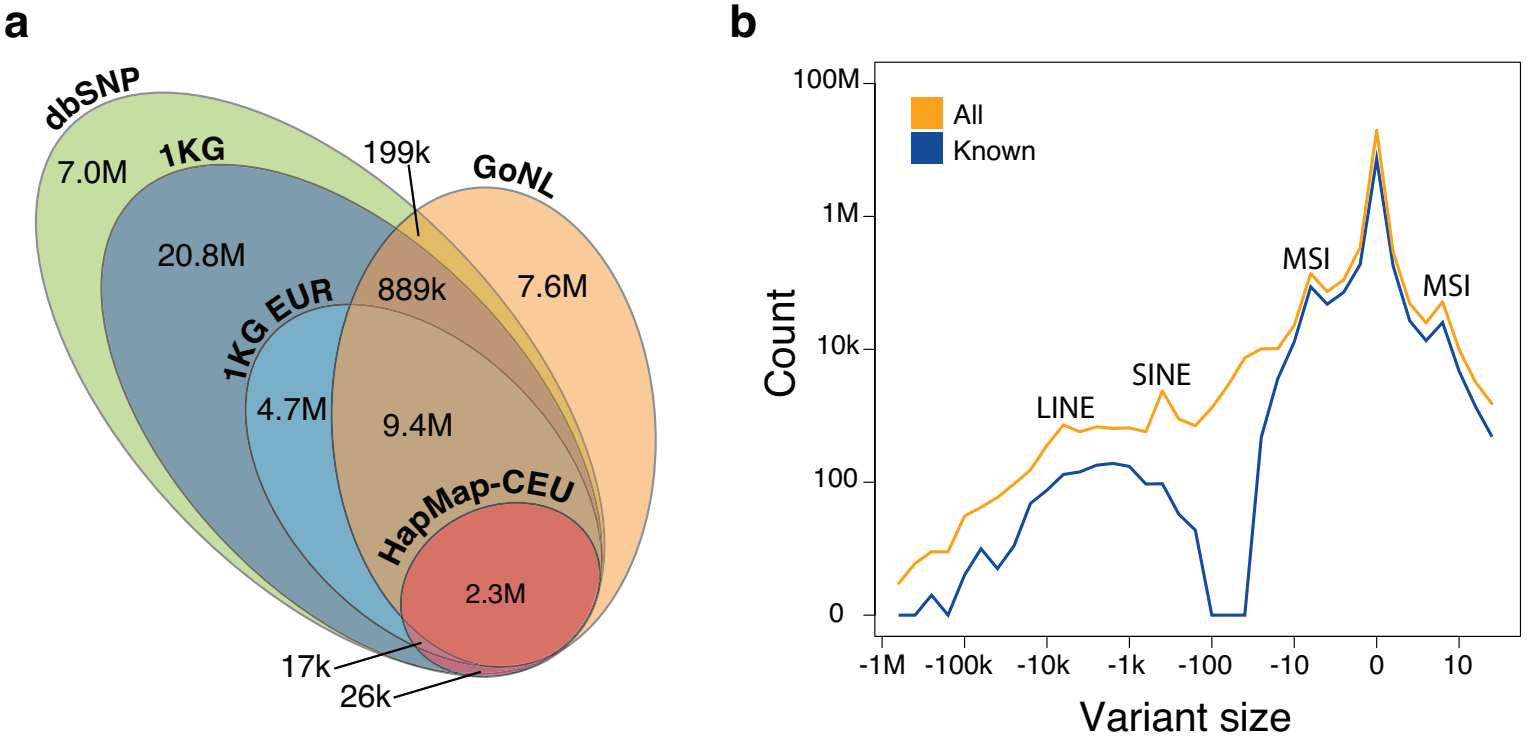
	Non-reference allele frequency		
	Rare (< 0.5%)	**Low-frequency (0.5–5%)**	**Common (> 5%)**
Variant type	Mean [SD]	Mean [SD]	Mean [SD]
All SNVs[1]	28,142 [3009.2]	130,190 [2448.1]	2.90M [10,080.9]
Novel[1,2]	17,751 [1,176.3]	4,354 [346.8]	620 [31.7]
Total conserved	1,892 [187.7]	7,593 [154.5]	106,824 [443.9]
Functional variation			
Synonymous	18 [4.9]	73 [8.9]	990 [19.0]
Nonsynonymous	101 [11.9]	238 [15.6]	2089 [31.8]
Probably damaging	32 [5.8]	58 [7.9]	394 [12.2]
Stop gain[1]	4 [1.9]	5 [2.2]	38 [4.3]
Splice site donor[1]	1 [0.9]	1 [0.9]	4 [1.5]
Splice site acceptor[1]	1 [0.7]	0.5 [0.6]	7 [1.4]
Total LoF[1]	5 [2.2]	6 [2.4]	49 [4.7]
Disease-associated variation			
OMIM	0 [0.6]	2 [1.6]	57 [4.9]
HGMD[3]	2 [1.2]	8 [2.7]	11 [2.3]
Indels (< 20 bp)			
Indel frameshift[1]	2 [1.4]	6 [2.6]	61 [4.8]
Indel non-frameshift[1]	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	6.7M bases		

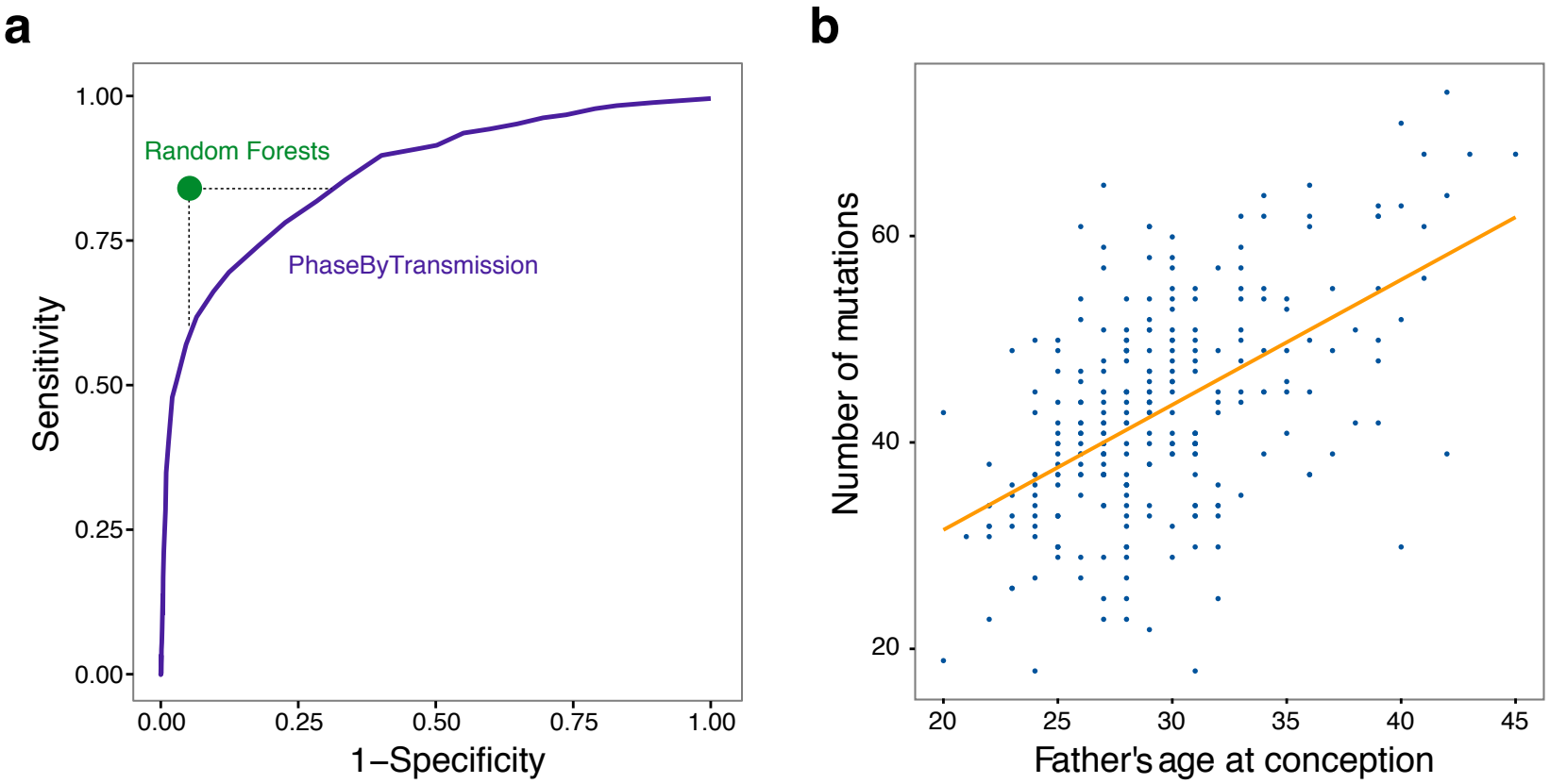
Only SNV sites at which ancestral state can be assigned with high confidence and that are highly conserved (GERP > 2.0) are reported. Frequency stratifications based on the unrelated samples only. OMIM, Online Mendelian Inheritance in Man.
[1]No conservation filter applied
[2]Not observed in dbSNP build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

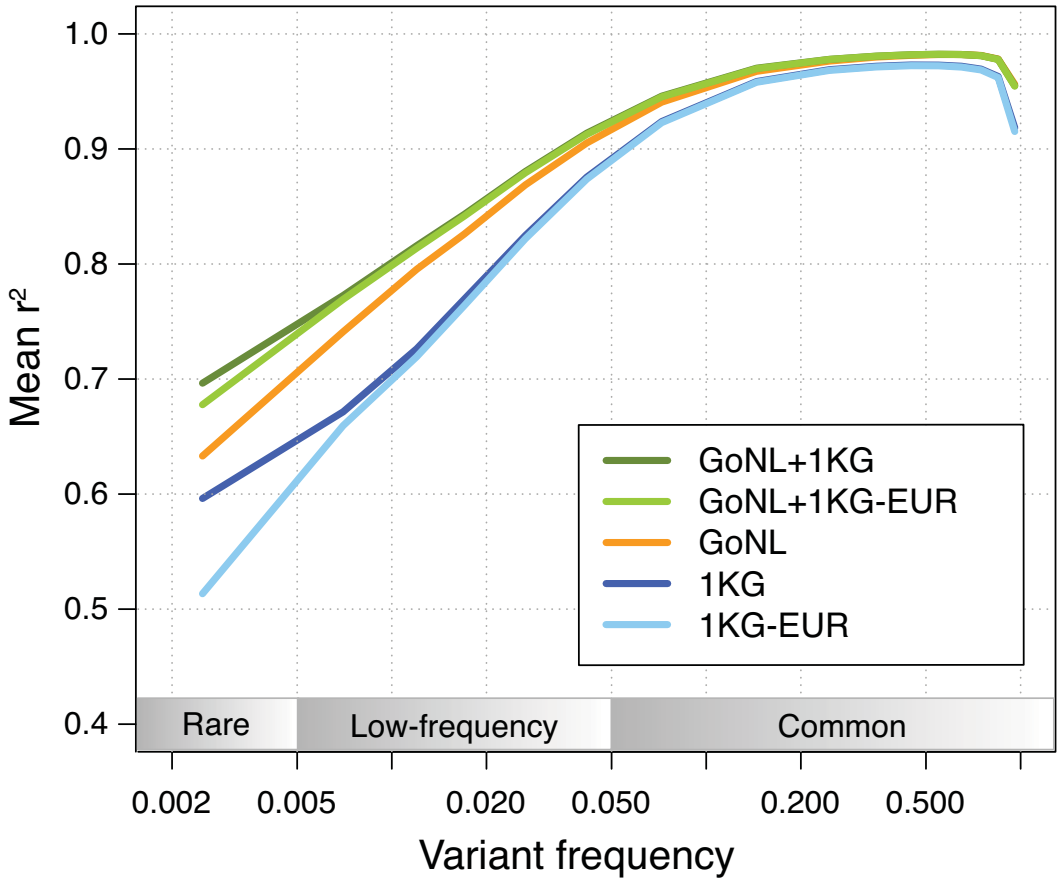
Table 2 | HGMD disease-causing mutations in the GoNL samples

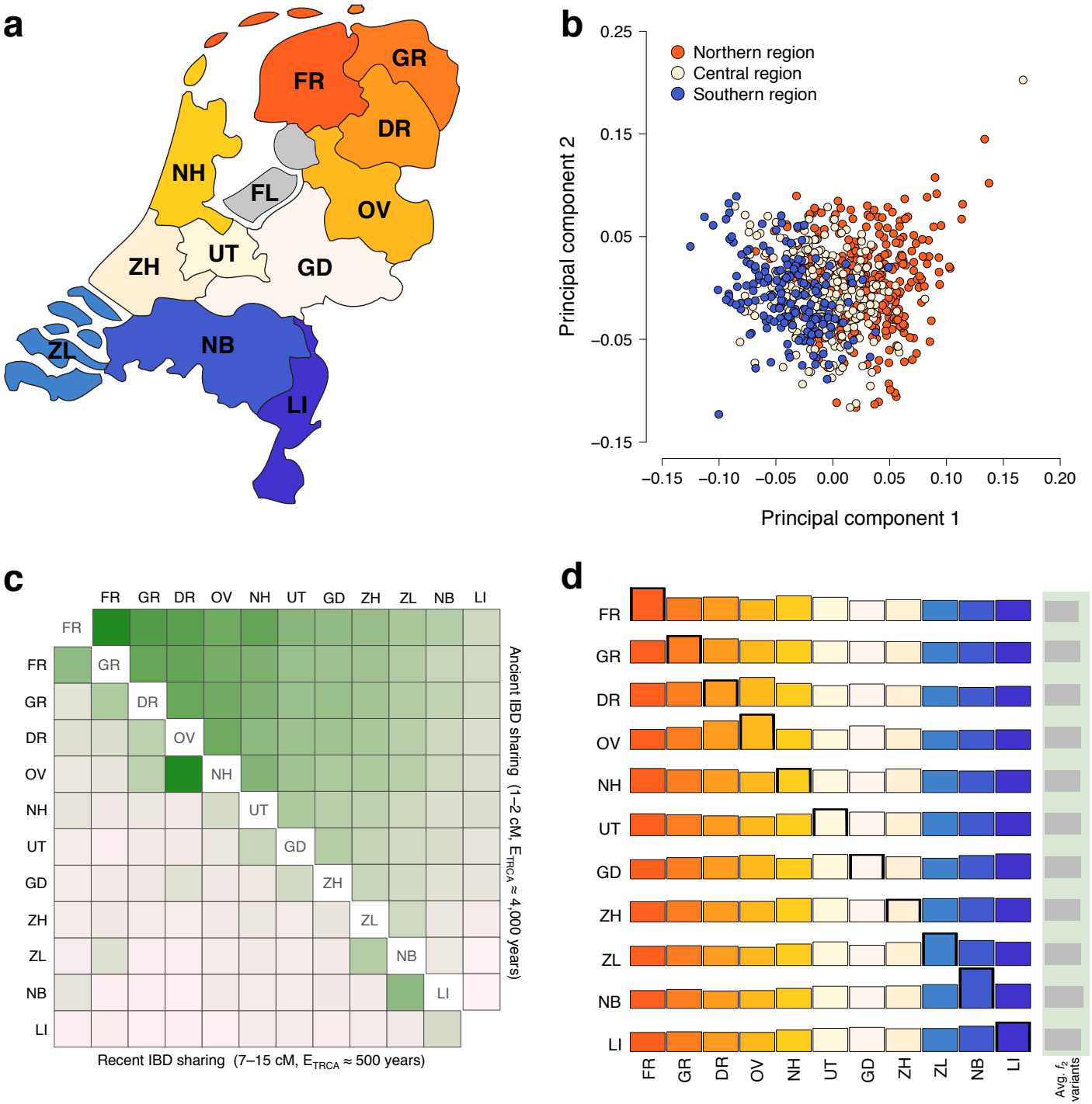
Chr	Pos	Gene	Mutation allele	Reference allele	Disease in HGMD	Disease prevalence	Inheritance pattern	Affected individuals[3]	Phenotypic manifestations[1]	OMIM ID(s)	Mutation allele frequency GoNL[4]	Mutation allele frequency 1KG – CEU
4	6302519	WFS1	A	G	Wolfram syndrome	0.0002%[1]	AR	257	Hyperglycemia, vision and hearing loss	604928, 222300	0.728	0.759
13	52515354	ATP7B	G	A	Wilson disease	0.003%[1]	AR	167	Liver disease, neuropsychiatric problems	277900	0.574	0.582
16	3304463	MEFV	T	C	Familial Mediterranean fever	0.10% in Mediterranean populations; rarer elsewhere[1]	AR	36	Recurrent fevers, inflammation of the abdomen, chest, joints	249100, 134610	0.277	0.224
11	6415463	SMPD1	A	G	Niemann-Pick disease	0.0004%[1]	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	61463522	COL9A3	A	C	Pseudo-achondroplasia	0.003%[1]	AD	177	Short stature, joint pain	177170	0.197	0.200
10	13340236	PHYH	A	G	Refsum disease	Unknown, current estimate 0.0001%[1]	AR	18	Anosmia, progressive blindness, deafness, hand/feet bone abnormalities, arrhythmia	266500	0.188	0.153
15	52643564	MYO5A	A	G	Griscelli syndrome	<0.0001%[2]	AR	10	Albinism (all types), intellectual disability (type 1), recurrent infection (type 2)	214450, 607624, 609227	0.159	0.141
19	36339247	NPHS1	T	C	Congenital nephrotic syndrome (Finnish type)	0.01% in Finland; rarer elsewhere[2]	AR	2	Proteinuria, rapid progression to renal failure	256300	0.082	0.082 (0.110)[5]
14	94847262	SERPINA1	A	T	Alpha 1 antitrypsin deficiency	0.02-0.06%[1]	AR	2	Lung disease, liver disease	613490	0.039	0.053

Acronyms are: HGMD (Human Gene Mutation Database); AR (autosomal recessive); AD (autosomal dominant); OMIM (Online Mendelian Inheritance in Man). [1]National Institutes of Health, Genetics Home Reference – USA. [2]National Institute of Health and Medical Research – France. [3]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). [4]Calculated from unrelated individuals. [5]Frequency in 1KG Phase I samples from Finland


a
dbSNP
7.0M
1KG
20.8M
199k
GoNL
7.6M
889k
1KG EUR
4.7M
9.4M
HapMap-CEU
2.3M
17k
26k
b
100M
1M
10k
100
0
Count
All
Known
LINE
SINE
MSI
MSI
-1M
-100k
-10k
-1k
-100
-10
0
10
Variant size


a
1.00
0.75
0.50
0.25
0.00
Sensitivity
Random Forests
PhaseByTransmission
0.00
0.25
0.50
0.75
1.00
1−Specificity
b
Number of mutations
60
40
20
20
25
30
35
40
45
Father's age at conception


1.0
0.9
0.8
0.7
0.6
0.5
0.4
Mean r²
GoNL+1KG
GoNL+1KG-EUR
GoNL
1KG
1KG-EUR
Rare
Low-frequency
Common
0.002
0.005
0.020
0.050
0.200
0.500
Variant frequency


a
FR
GR
DR
NH
FL
OV
ZH
UT
GD
ZL
NB
LI
b
Northern region
Central region
Southern region
Principal component 2
Principal component 1
0.25
0.15
0.05
−0.05
−0.15
−0.15 −0.10 −0.05 0.00 0.05 0.10 0.15 0.20
c
FR GR DR OV NH UT GD ZH ZL NB LI
Ancient IBD sharing (1–2 cM, E_TRCA ≈ 4,000 years)
Recent IBD sharing (7–15 cM, E_TRCA ≈ 500 years)
d
FR GR DR OV NH UT GD ZH ZL NB LI
Avg. f2 variants