EJHG Open

**European Journal of Human Genetics (2017) 25,** 1253–1260
Official journal of The European Society of Human Genetics
www.nature.com/ejhg

ARTICLE

# SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population

Adam Ameur*[1,2], Johan Dahlberg[2,3], Pall Olason[4,5], Francesco Vezzi[2,6], Robert Karlsson[7], Marcel Martin[5,6], Johan Viklund[4,5], Andreas Kusalananda Kähäri[4,5], Pär Lundin[6], Huiwen Che[1], Jessada Thutkawkorapin[8], Jesper Eisfeldt[8], Samuel Lampa[5,9], Mats Dahlberg[5,6], Jonas Hagberg[5,6], Niclas Jareborg[5,6], Ulrika Liljedahl[2,3], Inger Jonasson[1,2], Åsa Johansson[1], Lars Feuk[1], Joakim Lundeberg[2,10], Ann-Christine Syvänen[2,3], Sverker Lundin[10], Daniel Nilsson[8], Björn Nystedt[4,5], Patrik KE Magnusson[7] and Ulf Gyllensten*[1,2]

**Here we describe the SweGen data set, a comprehensive map of genetic variation in the Swedish population. These data represent a basic resource for clinical genetics laboratories as well as for sequencing-based association studies by providing information on genetic variant frequencies in a cohort that is well matched to national patient cohorts. To select samples for this study, we first examined the genetic structure of the Swedish population using high-density SNP-array data from a nation-wide cohort of over 10 000 Swedish-born individuals included in the Swedish Twin Registry. A total of 1000 individuals, reflecting a cross-section of the population and capturing the main genetic structure, were selected for whole-genome sequencing. Analysis pipelines were developed for automated alignment, variant calling and quality control of the sequencing data. This resulted in a genome-wide collection of aggregated variant frequencies in the Swedish population that we have made available to the scientific community through the website https://swefreq.nbis.se. A total of 29.2 million single-nucleotide variants and 3.8 million indels were detected in the 1000 samples, with 9.9 million of these variants not present in current databases. Each sample contributed with an average of 7199 individual-specific variants. In addition, an average of 8645 larger structural variants (SVs) were detected per individual, and we demonstrate that the population frequencies of these SVs can be used for efficient filtering analyses. Finally, our results show that the genetic diversity within Sweden is substantial compared with the diversity among continental European populations, underscoring the relevance of establishing a local reference data set.**

*European Journal of Human Genetics* (2017) **25,** 1253–1260; doi:10.1038/ejhg.2017.130; published online 23 August 2017

## INTRODUCTION

Human whole-genome sequencing (WGS) is now being performed at an unprecedented scale because of technology developments, making large-scale sequencing projects affordable. The majority of this sequencing involves patient samples with a specific phenotype of possible genetic nature. Access to population-based reference control data sets, based on high-quality whole-genome sequences, is important for identification of candidate disease causing genetic variants in clinical sequencing, for improved prediction of genetic effects in research studies as well as for better imputation of samples typed on SNP arrays. International efforts to determine the genetic variability pattern in population-based samples, such as the 1000 Genomes Project,[1] have contributed important information, but because of the small sample size for many populations, this provides merely an overview of the global pattern of variability. European populations differ in their genetic structure, and there is a need to assess the genetic variability at a more detailed level in individual populations. Such efforts have been initiated for example in The Netherlands, Denmark and Iceland.[2–4] Large-scale sequencing is also being performed in the United Kingdom but focusing on patient samples.[5] Similar WGS projects have also been initiated in many other parts of the world.[6–9]

The genetic structure of the Swedish population was recently studied using genome-wide SNP data from 5174 Swedes with extensive geographical coverage.[10] The results showed that there are pronounced regional differences, in particular between the northern and the remaining counties. In addition, the number of extended homozygous segments showed large differences between southern and northern Sweden, as well as between southern regions. Population sequencing efforts in Sweden have shown that a large number of the genetic variants in human populations have not yet been identified. For example, in sequencing 200 kb from 500 individuals from each of five local European populations, including northern Sweden, we have previously reported that 17% of single-nucleotide variants (SNVs) and 61% of small insertion/deletions were not present in the public databases.[11] Most novel genetic variants showed a very limited geographic distribution, with 62% of the novel SNVs and 59% of novel insertion/deletion variants detected in only one of the local populations.

[1]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden; [2]National Genomics Infrastructure, Science for Life Laboratory, Sweden; [3]Science for Life Laboratory, Department of Medical Sciences, Molecular Medicine, Uppsala University, Uppsala, Sweden; [4]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden; [5]National Bioinformatics Infrastructure, Science for Life Laboratory, Sweden; [6]Science for Life Laboratory, Department of Biochemistry and Biophysics (DBB), Stockholm University, Stockholm, Sweden; [7]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [8]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; [9]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden; [10]Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Royal Institute of Technology, Stockholm, Sweden
*Correspondence: Dr A Ameur or Professor U Gyllensten, Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, PO Box 815, Uppsala 75108, Sweden.
Tel: +46 18 425 02 79 or +46 18 471 48 46; E-mail: adam.ameur@igp.uu.se or ulf.gyllensten@igp.uu.se

Received 24 March 2017; revised 9 June 2017; accepted 18 July 2017; published online 23 August 2017

Thus, results from previous studies demonstrate that the genetic differentiation between local communities in Sweden is substantial when viewed on a European scale. The large number of population-specific low-frequency SNVs and indels underscores the need to establish a database of genetic variants using whole human genome sequences for the Swedish population. It is well known from array-based genome-wide association studies (GWASs) that genetic stratification within a population may introduce a bias, particularly for rare variants.[12] Careful handling of influences from genetic ancestry is therefore necessary to avoid false positives. The problem is likely to be more severe for WGS studies because of the multitude of low-frequency variants. Comparisons between unmatched cases and controls will inflate levels of false positive and false negative findings.[13] Therefore, in order to make full use of WGS in association studies or in clinical diagnostics, there is a strong need to establish sufficiently large control data sets for local populations. Finally, studies of human evolution and migrations have been brought to another level of understanding using WGS of both extant and archaic humans.[14–16] For these types of studies, population-based cross-sectional sampling provides the optimal strategy for estimating unbiased frequencies of individual variants as well as haplotypes.

Through the SweGen project described here, we make available a population-based genetic variant data set for the Swedish population. The undertakings of this project were to identify a control cohort that reflects the genetic structure of the Swedish population, to employ WGS on these samples using Illumina Inc. (San Diego, CA, USA) sequencing, to construct robust analyses pipelines that are capable of handling large-scale human WGS data and to make the resulting variant frequencies available to researchers and clinicians. The genetic variant data set based on these 1000 individuals will enable scientifically sound whole-genome sequence-based association studies for national patient cohorts by providing data on well-matched national controls selected on the basis of the genetic structure of the population. It will also be an important resource in studies of familial heritable disorders and clinical sequencing. Large cross-sectional population studies are needed to better understand the penetrance and disease trait expressivity of pathogenic variants, as well as for interpretation and prioritization of candidate disease-causing genetic variation.[17] However, in today's globalized world, many populations contain a high degree of heterogeneity because of immigration and admixture. This comprises a big challenge for the creation of population-based genetic resources such as the SweGen data set. The issue could be addressed by WGS efforts targeting certain ethnic minority groups, or by efficient sharing of data from WGS initiatives in different parts of the world.

## METHODS

### DNA samples and ethical consent

The Swedish Registry (STR) was established in the 1960s and, at present, holds information on 85 000 twin pairs, both monozygotic and dizygotic. DNA from blood or saliva has been collected and extracted from 55 000 of these individuals. A total of 10 000 participants (one per twin pair for monozygous twins) in the TwinGene project were subjected to high-density SNP array typing. TwinGene is a nation-wide and population-based study of Swedish-born twins agreeing to participate. The TwinGene sample collection represents the Swedish geographic population density distribution. In our initial principal component analysis (PCA), all 10 000 individuals were included, and from this analysis, 942 unrelated individuals were selected for WGS, mirroring the density distribution. All participants gave their written informed consent and the TwinGene study was approved by the regional ethics committee (Regionala Etikprövningsnämnden, Stockholm, dnr 2007-644-31, dnr 2014/521-32).

The Northern Sweden Population Health Study (NSPHS) is a health survey in the county of Norrbotten, Sweden, to study the medical consequences of lifestyle and genetics. Blood samples were collected (serum and plasma) and stored at − 70 °C on site. DNA was extracted using organic extraction. Based on PCA, 58 samples were selected. The NSPHS study was approved by the local ethics committee at the University of Uppsala (Regionala Etikprövningsnämnden, Uppsala, 2005:325 and 2016-03-09). All participants gave their written informed consent to the study including the examination of environmental and genetic causes of disease in compliance with the Declaration of Helsinki.

### Library preparation and Illumina sequencing

Library preparation and sequencing was performed by the National Genomics Infrastructure (NGI) in Sweden, at the Genomics Production site in Stockholm (NGI-S) and the SNP&SEQ Technology Platform in Uppsala (NGI-U). The 1000 samples were divided between NGI-S (509 samples) and NGI-U (491 samples) and library preparation and sequencing were performed independently by each facility. DNA samples were fragmented with Covaris E220 (Covaris Inc., Woburn, MA, USA) to 350 bp insert sizes and sequencing libraries were prepared from 1.1 $\mu$g/1 $\mu$g DNA (NGI-S and NGI-U, respectively) using the TruSeq DNA PCR free sample preparation kit (Illumina Inc.) according to the manufacturer's instructions (guide 15036187). The protocols were automated using an Agilent NGS workstation (Agilent Technologies, Santa Clara, CA, USA) and a Biomek FXp (Beckman Coulter, Brea, CA, USA) at NGI-S and NGI-U, respectively. Paired-end sequencing with 150 bp read length was performed on Illumina HiSeq X (HiSeq Control Software 3.3.39/RTA 2.7.1) with v2.5 sequencing chemistry. A sequencing library for the phage PhiX was included as 1% spike-in in the sequencing run.

### Alignment and variant calling

Raw reads were aligned to the GRCh37 version of the human reference using BWA-MEM 0.7.12.[18] The resulting alignments were sorted and indexed using samtools 0.1.19,[19] and alignment quality control statistics were gathered using qualimap v2.2.[20] Alignments for the same sample from different flow cells and lanes were merged using Picard MergeSamFiles 1.120 (broadinstitute.github.io/picard). The raw alignments were then processed according to the GATK best practice,[21] using version 3.3 of the GATK software suite. Alignments were realigned around indels using GATK RealignerTargetCreator and IndelRealigner, duplicate marked using Picard MarkDuplicates 1.120 and base quality scores were recalibrated using GATK BaseRecalibrator, resulting in a final BAM file for each sample. Finally, gVCF files were created for each sample using the GATK HaplotypeCaller 3.3. Reference files from the GATK 2.8 resource bundle were used throughout. All these steps were coordinated using Piper v1.4.0.[22] For details on commands and parameters used see Supplementary Methods 1.1.

### Joint genotyping and generation of variant frequency file

Joint genotyping was executed on the whole cohort of 1000 samples as recommended by GATK guidelines. This produced a single VCF file containing all variants identified in the 1000 sequenced individuals. Because of the large number of samples, five batches of 200 samples each were merged into five separate gVCF files using GATK CombineGVCFs. The five individual gVCF files were used as input for GATK GenotypeGVCF. Subsequently, GATK VQSR steps were executed to recalibrate SNVs and indels (ie, GATK VariantRecalibrator and GATK ApplyRecalibration). In order to reduce the overall time, the first two steps (GATK CombineGVCFs and GATK GenotypeGVCF) were run on 60 Mbp non-overlapping segments of the Human Genome. For more details, see Supplementary Methods.

### Structural variation analyses

Manta 1.0.0[23] was used to call structural variants using the sample-specific, preprocessed BAM files as input. The output VCF files were then converted to BED, treating heterozygous and non-reference homozygous variants identically. Statistics reported here were computed with the help of these files. The 1000 BED files were merged into a single one with summarized variant frequencies, where SVs were considered identical when their type (DEL/DUP/INS/INV) and position (chromosome, start, end) are identical. The resulting file is available for download (see Data availability). To evaluate the efficiency of the SweGen

SV data for filtering, the SV variant frequency BED file was recomputed excluding one individual, and the Manta VCF file from that individual was filtered with all events in the recomputed BED file above a given frequency threshold (0.1, 0.5, 1 or 5%), using reciprocal overlaps (100, 95, 75 or 50%) between events as implemented in BEDTools.[24] This computation was permuted for all 1000 individuals in the cohort. For insertions, BEDTools only considers the exact insertion site as a definition of event, and hence the reciprocal overlap between events cannot be defined for insertions. SV calling and SV filtering tools are available from the WGS-structvar workflow (https://github.com/NBISweden/wgs-structvar).

### Principal component analyses

PCA is a commonly used method for population stratification and identification of outliers. We used PCA for both selecting the 1000 individuals to be included in the project based on data from SNP arrays (Figure 1a), and evaluating the final WGS data in the context of genomic data from other populations (Figure 5).

The PCA on SNP array data in Figure 1a is based on 19 978 autosomal SNPs, selected to fulfill the following: (1) must be present (by name) on SNP chips Affymetrix 5.0 (Affymetrix Inc., Santa Clara, CA, USA), Affymetrix 6.0, Illumina OmniExpress and in the 1000 genomes phase 1 EUR data set; (2) per SNP missingness ≤ 2% in the combined sample; (3) minor allele frequency ≥ 5% in the combined sample; (4) pairwise LD ($R^2$) ≤ 0.05 with other SNPs within a 1000-SNP window; and (5) not in a region known to harbor strong long-range LD[25] (see Supplementary Table S1). SNP weights were defined by running the PCA in European 1000 Genomes samples, and then projecting STR and NSPHS into the resulting space defined by the analysis.

The PCA on WGS data in Figure 5 was performed on the 942 STR samples together with data from the 1000 Genomes Project. In order to work on a representative yet manageable number of variants, we extracted 713 027 SNPs contained in the Infinium OmniExpress-24 v1.2 chip from the jointly called VCF file. After filtering we retained 648 379 SNVs shared between the SweGen project and 1000 Genomes. Genotypes for the same set of SNVs were extracted from the 1000 Genomes phase 3 data set.[26] To run PCA, VCFtools[27] was employed to convert VCF into tped format (input format for PLINK 1.9[28]). Subsequently, PLINK was used to compute the principal components.

## RESULTS

### Strategy for selecting individuals

The first aim was to identify a set of 1000 individuals to be used for the construction of the SweGen reference cohort. As this data set will be used for a wide range of research projects as well as for routine clinical investigations, we decided to construct a population-based cross-sectional cohort that reflects the genetic structure of the Swedish population. As a first step, we performed an inventory of available cohorts that fulfilled three criteria: (1) population-based cross sectional cohorts (ie, not focusing on specific outcomes), (2) SNP array data available to be able to study the genetic structure, and (3) DNA or blood samples available. A number of alternatives were identified and we decided to use samples from two of these: the STR[29] supplemented by samples from the NSPHS.[30] Both of these are population-based collections. STR is a national registry of Swedish-born twins and the vast majority of samples were taken from this resource. NSPHS was used to contribute a limited number of samples from the very northern part of the country. This cross-sectional selection is likely to reflect the genetic variation that distinguishes the Swedish population from continental Europe. One caveat is that already established and well-characterized national sample collections such as STR and NSPHS do not reflect the genetic background of the most recent migrants to Sweden. Therefore, the SweGen control cohort is likely to represent the genetic structure of Swedish individuals who been present in Sweden for at least one generation.

From the STR, 10 000 samples had previously been genotyped using high-density SNP arrays, and from NSPHS, similar data were available for 1033 individuals. Figure 1a shows a plot with the first and second principal components of these two Swedish cohorts alongside the continental European populations included in the 1000 Genomes Project.[26] The results indicate that the combined allele frequencies of known polymorphisms form a distinct pattern in the Swedish samples, thereby underscoring the need to generate a population-specific reference cohort. We identified a set of 1000 samples for the SweGen data set, 942 selected from STR and 58 selected from NSPHS, and

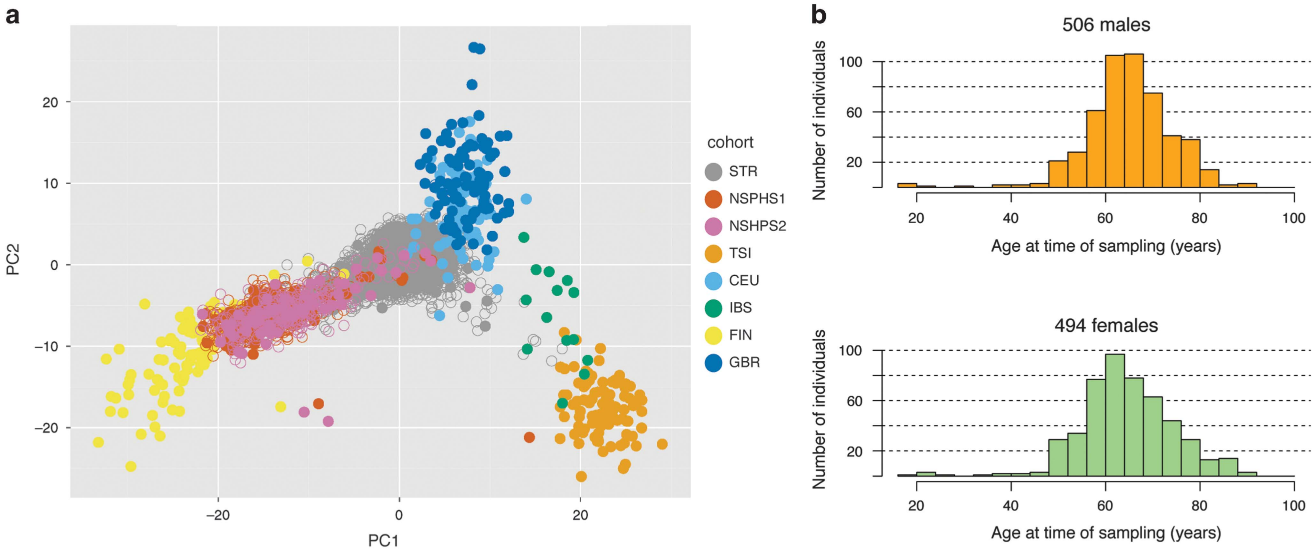

**Figure 1** Selection of 1000 individuals based on genetic variation within Sweden. (**a**) PCA of SNP array data from the Swedish Twin Registry (STR) and the Northern Sweden Population Health Study (NSPHS1 and NSPHS2, collected in two different phases) compared with data from European 1000 Genomes populations (CEU: Utah Residents with Northern and Western Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian Population in Spain, TSI: Toscani in Italia). A total of 19 978 SNP positions were used to generate this plot (see Methods). (**b**) Age and gender distribution for the 1000 individuals in the SweGen data set. The median age at sampling is 65.4 years for males, 64.9 years for females and 65.2 in the combined data set.

confirmed that they captured the diversity seen in the country (ie, the pattern seen for the 10 000 STR individuals and the 1033 NSPHS individuals). The 1000 samples are composed of 506 males and 494 females, and their average age was 65.2 years at the time of sampling (Figure 1b).

**Results of pilot experiments**

The 1000 Swedish individuals were sequenced on Illumina HiSeq X instruments at two different sites: NGI Uppsala (NGI-U) and NGI Stockholm (NGI-S) (see Methods). Before sequencing of the entire data set, two small pilot studies were conducted to validate the quality of the WGS data. In the first pilot experiment, performed at NGI-S, two independent libraries prepared from the same sample were analyzed. Sequencing and analysis of the two replicated libraries resulted in ~3.86 million SNV positions at which a non-reference genotype was called for either of the library preparations. Of these, 90 446, or 2.3%, had discordant genotypes, resulting in a genotyping concordance of 97.7%. The discordant genotypes are likely to be caused by heterozygous SNPs that are incorrectly called as homozygous (and with different alleles) in the two replicates.

In a second pilot study, eight samples were sequenced in parallel at NGI-U and NGI-S to study potential biases between the two sites. Our results revealed that out of an average of 8 056 903 genotyped variant sites for each of the eight sample pairs, 7 942 822 genotype calls on average overlap between the two replicates before any filtering. The genotype concordance between the samples was 97.5% on average. As very similar concordance results were obtained from the first pilot study, we conclude that the site effects between NGI-U and NGI-S are negligible. After variant quality score recalibration (VQSR), the number of overlapping genotypes, passing filters, was on average 7 448 632 sites per sample pair (93.8% of the number of sites pre-VQSR) and the genotype concordance was on average 98.6%. The genotype concordance of variants pre- and post-VQSR filtering shows the effectiveness of this approach, as more than half of the discordant sites (104 981/196 822 = 53.4%) were removed, while filtering only 6.2% of the total number of sites. To assess whether there was any systematic lack of coverage in coding regions for the eight pilot samples, we listed all exons that have zero coverage for at least 50% of their bases in more than half of the samples. A total of 27 exons from 9 gene loci were found to be less than 50% covered (see Supplementary Table S2). Only one single exon had less than half of its positions covered in all eight samples.

**Analysis and quality control of the WGS data**

Of the 1000 samples, 509 were sequenced at NGI-S and 491 at NGI-U. For the management, alignment and SNV analysis of the WGS data, we developed a robust and efficient pipeline that was used for all samples (see Figure 2). Briefly, this pipeline aligned the reads to the GRCh37 version of the human reference genome, processed the alignments according the GATK best practice recommendations[21] and carried out variant calling using the GATK HaplotypeCaller.[31] The workflow was executed using Piper.[22]

A series of analyses were performed to assess the data quality. To ensure that no sample mix-ups occurred during the sequencing procedure, all samples were independently genotyped at a limited number of SNP positions and checked for concordance with the WGS data. We also performed a kinship analysis to ensure that no sample was represented in duplicate in the final data set. Further, we performed a PCA on all 1000 samples to search for unexpected outliers or obvious biases between samples sequenced at the two different sites (Supplementary Figure S1). After having performed the basic quality control steps, we generated a file containing the SNV and indel frequencies observed in the 1000-sample data set. The resulting frequency file is available for the research community both as a flat

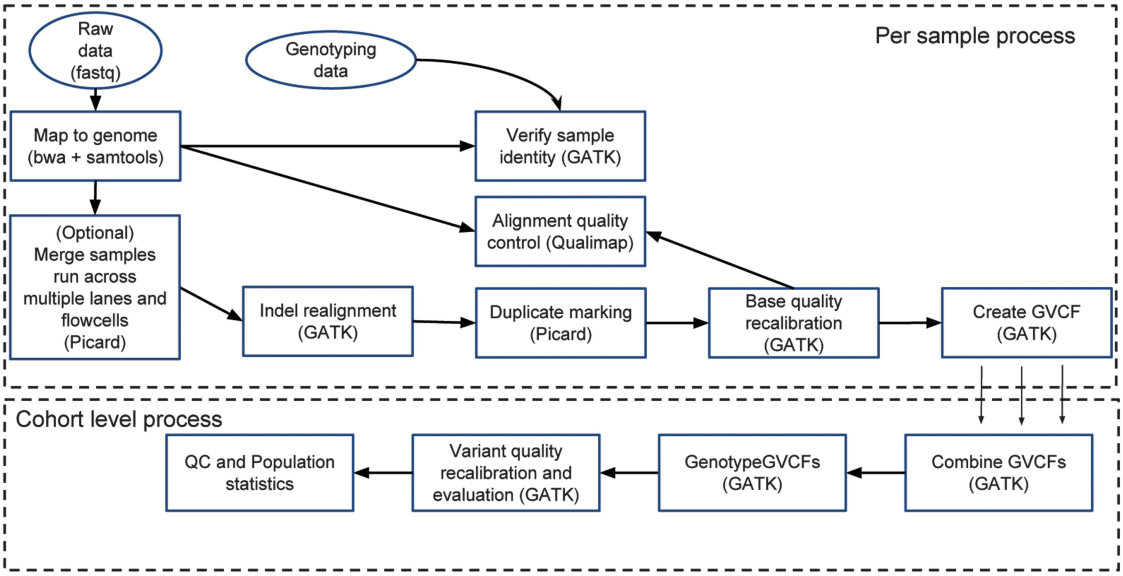
Per sample process
Raw data (fastq)
Genotyping data
Map to genome (bwa + samtools)
Verify sample identity (GATK)
Alignment quality control (Qualimap)
(Optional) Merge samples run across multiple lanes and flowcells (Picard)
Indel realignment (GATK)
Duplicate marking (Picard)
Base quality recalibration (GATK)
Create GVCF (GATK)
Cohort level process
QC and Population statistics
Variant quality recalibration and evaluation (GATK)
GenotypeGVCFs (GATK)
Combine GVCFs (GATK)

**Figure 2** Overview of workflow for alignment and SNV and indel detection. The process has two phases: first each sample is processed individually and then the entire cohort is processed together. The first phase begins by aligning the raw reads to the reference genome using bwa, converting the resulting alignments to bam format and sorting and indexing them using samtools. Preliminary sample identity is verified by checking concordance with genotyping data and alignment quality is assessed using Qualimap. Once all alignments from a sample have been merged, they are processed according to the GATK Best practice workflow, with indel realignment, duplicate marking and base quality score recalibration, before using the GATK Haplotypecaller to create genomic VCF files (GVCF). The second phase is carried out on a cohort level. This is followed by variant quality recalibration. Finally, quality control metrics and population statistics are computed for the final call set.

text file and through a local installation of the ExAC[32] genome browser at the website https://swefreq.nbis.se. The data have also been submitted to dbSNP[33] (see Data availability section).

### Overview of SNVs and indels in the SweGen data set

An overview of the WGS data for all 1000 samples in the SweGen data set is shown in Table 1. The average autosomal coverage after duplicate removal was 36.7×, with values ranging between 20.2× and 97.6× for individual genomes. A total of 29.2 million SNV sites and 3.8 million small insertion/deletion (indel) sites were detected in the 1000 samples. The vast majority of indel sites were <50 bp in length, with a few of them ranging up to 200 bp (Supplementary Figure S2). In version 147 of dbSNP,[33] a database that includes information from phase 3 of the 1000 Genomes Project and many other sources, 8.9 million SNVs and 1.0 million indels were not present. The minor allele frequency (MAF) distribution shows that the vast majority of variants in the SweGen data set are rare and have a MAF of at most 2% (Figure 3a). Virtually all of the novel genetic variants have a MAF between 0.05 and 1% (Figure 3b). Of the novel variants, 7 198 518 were detected in only one of the samples, thereby resulting in an average contribution of 7199 novel variants per individual. As a comparison, 7215 novel variants were contributed by each individual among the European samples in a recent study of 10 000 whole human genomes.[8] For the novel genetic variants, 23 396 SNVs and 3239 indels were found to have a direct effect on the amino acid sequence of some protein (ie, non-synonymous, stop–gain, stop–loss, frame shift or splicing). Our results thus indicate the presence of a substantial number of genetic variants in the Swedish population that are not represented in current databases, some of which could have a direct effect on a protein.

### Analysis of SVs in the SweGen data set

A total of 8 636 141 structural variation (SV) events were reported in the 1000 samples using the software Manta[23] (see Table 1). On average, 2417 insertions (INS), 5245 deletions (DEL), 537 duplications (DUP) and 436 inversions (INV) were detected per individual (see Figure 4a). Because of difficulties to reliably identify SV events from short read sequencing data, we expect a relatively large fraction of the SV calls to be false positives. However, preliminary analyses suggest that using the same SV calling method, both true and false positive SV calls appear to be well reproducible across samples with a similar genetic background, and that an aggregated SV frequency database can therefore provide an important resource to aid the identification of rare SV events in clinical settings.[34] Such filtering will remove variants common in the population, and at the same time significantly reduce the number of false positive calls and background noise arising, for example, from alignment artifacts. To assess the power of our SV frequency table for this type of analysis, we performed a computational test where the SVs from each of the 1000 individuals were filtered so that all events occurring at a frequency of at least 1% were removed (Figure 4b), using four different overlap criteria for comparing SV coordinates, ranging from a stringent 100% reciprocal overlap to a more relaxed 50% reciprocal overlap. The filtering resulted in an average remaining number of 78 insertions, and an average of 185–355 deletions, 45–105 duplications and 28–89 inversions per individual, depending on the overlap criteria used (Figure 4b). The filtering thus gives a more manageable number of all types of SV events to be further scored and evaluated that might be of importance for example in a clinical setting. More relaxed or stringent filtering can be applied by increasing or reducing the 1% frequency level (Supplementary Figure S3). A robust bioinformatics workflow to perform the SV calls and to utilize the SweGen SV frequency table for filtering is provided at https://github.com/NBISweden/wgs-structvar.

**Table 1 Overview of the WGS data for the 1000 Swedish samples**

| | *SweGen data set* |
|---|---|
| No. of samples | 1000 |
| Avg coverage (min /max) | 36.7 (20.2/97.5) |
| Total no. SNP sites (not in dbSNP147) | 29 162 141 (8 856 354) |
| Total no. of indel sites (not in dbSNP147) | 3 825 043 (1 001 047) |
| Ts/tv ratio | 2.01 |
| Avg homozygous SNPs | 1 486 648 |
| (min/max) | (1 332 739/1 578 505) |
| Avg heterozygous SNPs | 2 366 095 |
| (min/max) | (2 190 824/2 603 755) |
| Avg singleton SNPs | 10.975 |
| (min/max) | (1030/31 087) |
| Total number of SVs | 8 636 141 |
| (INS/DEL/DUP/INV) | (2 417 420/5 245 403/537 422/435 896) |

Abbreviations: DEL, deletions; DUP, duplications; INV: inversions; INS, insertions.

### Relating WGS data for Swedish individuals to other populations

In order to study the genetic variation within the Swedish WGS samples in the context of other populations, we identified nearly 650k positions from the Illumina OmniExpress chip where SNVs in the SweGen data set overlapped with SNVs from phase 3 of the 1000 Genomes Project.[26] These positions were used as a basis for PCA. We restricted the PCA to the 942 samples from the STR data set, as those individuals are more representative of a cross-section of the Swedish population as compared with the complete data set that includes the 58 samples from the northern NSPHS population. In addition, the NSPHS individuals have a higher degree of relatedness that might influence the PCA. Our results show that the STR cohort is genetically close to the other European 1000 Genomes populations, but with a distinct bias toward the Asian (EAS and SAS) populations (Figure 5a). We also performed a PCA restricted only to European populations (Figure 5b) that clearly shows that the Swedish genomes contain a substantial amount of genetic variation as compared with the other European 1000 Genomes populations. The analysis shows a long tail that is intermixed with the Finnish samples, mainly represented by genome sequences from northern parts of Sweden. Our results thus support previous findings of a high degree of genetic diversity between different parts of the country.[10]

## DISCUSSION

We have generated a comprehensive data resource that describes the landscape of genetic variability in the Swedish population. The data set is constructed from a population-based selection of individuals, taking into account the main aspects of the genetic structure of the Swedish population. The 942 individuals from STR represent a cross-section of the population, and the remaining 58 individuals from NSHPS increase the coverage of genetic variation in the northern part of the country. No metadata, including possible disease state, has been used in the selection of the 1000 individuals. This means that individuals affected by diseases may have been included. However, the median age of the participants is relatively high, on average 65.2 years at the time

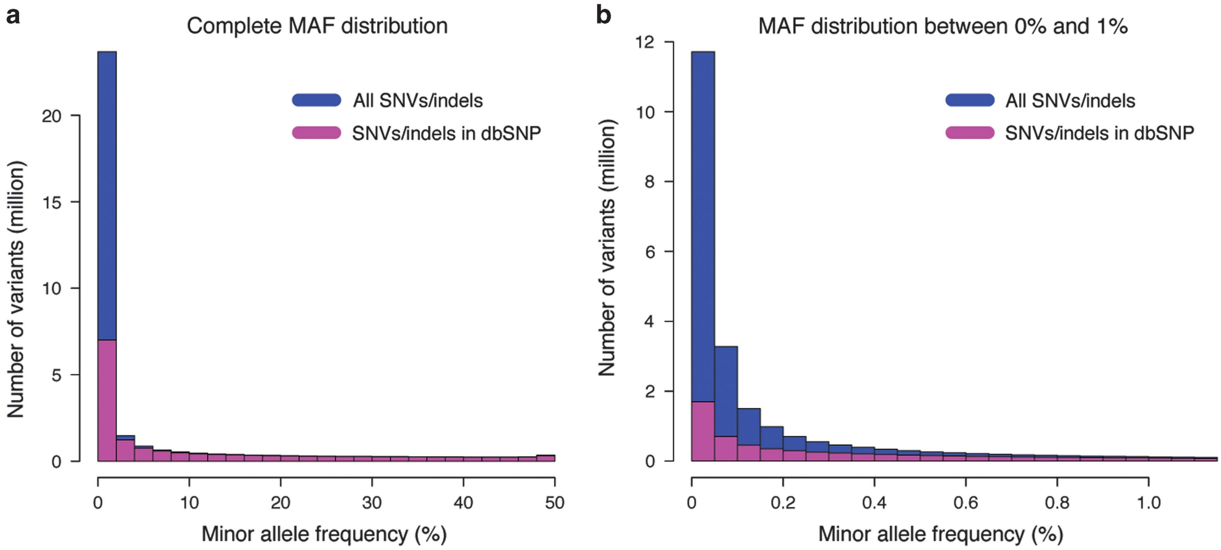

**Figure 3** Minor allele frequency (MAF) distribution in the SweGen data set. (**a**) MAF distribution for all SNVs and indel variants in the data set. The known variants (colored in pink) are those that are found in version 147 of dbSNP. All other variants (colored in blue) are novel. (**b**) MAF distribution for variants occurring in at most 1% of the SweGen individuals.

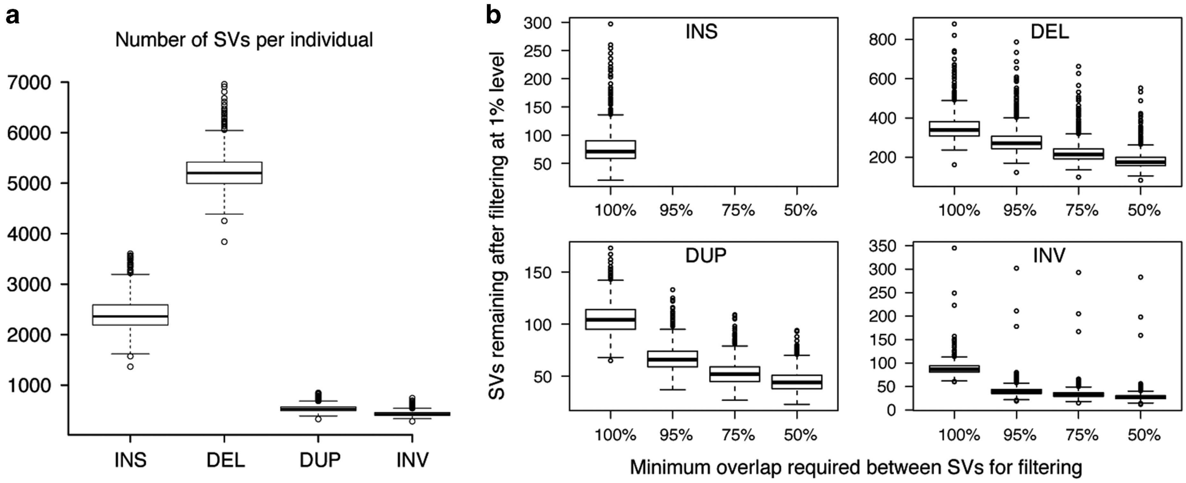

**Figure 4** Analysis of structural variation in the SweGen data set. (**a**) Structural variations (SVs) were detected by the Manta software and the box plots show distributions of the number of insertions (INS), deletions (DEL), duplications (DUP) and inversions (INV) detected in each of the 1000 SweGen samples. The average numbers are the following: 2417 INS, 5245 DEL, 537 DUP and 436 INV. (**b**) Number of structural variants remaining in a WGS sample after filtering all events occurring at a frequency of at least 1% in the SweGen data set. For each of the 1000 genomes, INS, DEL, DUP and DEL calls were filtered against the SweGen SV frequencies to produce a box plot distribution for the number of SVs remaining after filtering. For each of the SV types, four different analyses were performed requiring a reciprocal overlap of 100, 95, 75 and 50% between SVs in order to be filtered. As partial overlaps are not defined for INS (see Methods), only the 100% data are shown for these events.

of sampling. The SweGen resource will therefore be of immediate use for variant interpretation in most highly penetrant rare disorders, including some with late onset. For more complex disorders, the SweGen data set can be used as a starting point, although more detailed information about the samples might be required. Access to phenotypic data can be requested from the Swedish Twin Registry. Each application will be subjected to an individual review process before any data access is granted.

A limitation of the SweGen data set is that it lacks representation of the genetic constitution of individuals who have arrived recently to Sweden. Approximately 10% of the present Swedish population has an origin in a country outside Europe, and have only recently settled in Sweden. This fraction has been further affected by the refugee situation in the past years, with many individuals seeking permanent residence in Sweden. In order to appropriately also include the genetic diversity of these groups, specific additions to the control cohort may be needed. However, because of the broad ethnic diversity of migrant groups recently arriving in Europe, there is also a need for similar genome initiatives to be carried out in other parts of the world in order to appropriately span the genetic structure of all individual populations. Ideally, reference data sets from different geographic regions would be combined into a single publicly available resource, as this would increase overall numbers and facilitate data sharing. However, before this can happen, the human WGS community needs

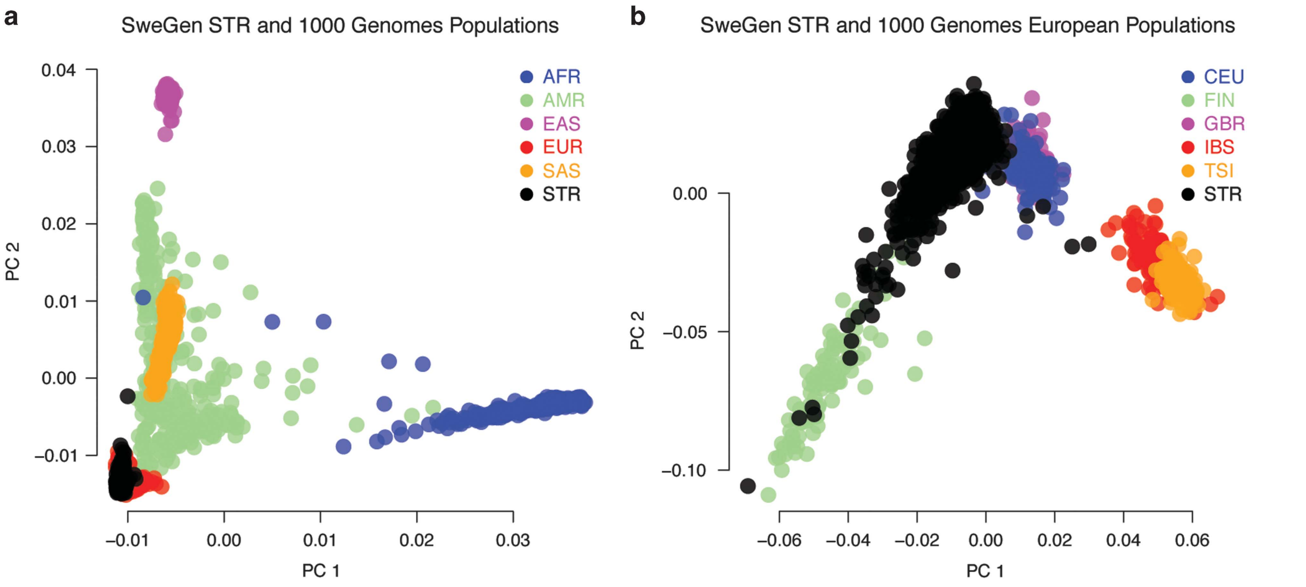

**Figure 5** Genetic variation in Sweden in relation to 1000 Genomes populations. (**a**) Results of PCA of SweGen WGS data, comparing the 942 Swedish STR samples with 1000 Genomes populations (AFR = African, AMR = Ad Mixed American, EAS = East Asian, EUR = European, SAS = South Asian). (**b**) Results of PCA of SweGen WGS data, comparing the 942 Swedish STR samples with the European 1000 Genomes populations (CEU: Utah Residents with Northern and Western Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian Population in Spain, TSI: Toscani in Italia). A total of 648 379 SNP positions were used to generate these two PCA plots (see Methods).

to resolve a number of important issues related to personal integrity regulations, standardization of laboratory/analysis procedures and data security. In spite of these limitations, the SweGen data set represents a valuable annotation of the genetic variability pattern in the majority of the Swedish population.

In this first release, the data are presented as a joint frequency table, annotating the frequency of SNVs, small insertion/deletions (indels) as well as larger SVs at each position across the genome. This is aimed at users in need of identifying and filtering out non-pathogenic population-specific low-frequency variants from their patient sequences. However, these frequency data can also be used for WGS-GWAS, although these analyses are more powerful when based on individual genotypes. Our intention is to make the individual genotype information available to individual projects granted permission to access the data. This will enable more powerful studies of the underlying genetic structure of the population, including haplotype-based analyses and imputation.

Our SV results are limited by the challenges of detecting structural events from short read sequencing data. This issue has been highlighted by several recent studies employing long-read technologies for sequencing of whole human genomes.[35–37] In these studies, only a fraction of the structural variants detected by long-read sequencing could be found by analysis of Illumina WGS data obtained from the same individuals. In the future, we might have reason to revisit the SweGen samples and apply novel technologies to get a more comprehensive view of these events in the Swedish population. Nevertheless, the SV data made available here can be seen a first step toward creating a catalog of structural events in this population and estimating their frequencies. Our preliminary results indicate that already this first release may have a clinical value for WGS-based diagnostics of genomics aberrations by removing false positive events and SVs occurring at high frequency in the population (Figure 4). We are hopeful that the future will bring more efficient technologies to detect SVs at a population scale, as well as standardization and efficient sharing of structural variation data between human WGS projects around the world.

The size of the control cohort is a compromise between the statistical power to identify genetic variants and the cost for WGS. The control database of 1000 individuals has sufficient size to include most of the genetic variants that occur in the population at a frequency of ≥ 1%, although there are many rare variants that may be missed. Given that the cost of WGS continues to plummet, the data set could be extended to increase our ability to detect a higher fraction of the low-frequency genetic variants. In terms of the power in association studies, sample sizes on at least the order of $10^3$ are needed to detect effects of low frequency variants.[38]

Our analyses indicate that the SweGen data set contains a large number of genetic variants that are not represented in the previously studied European populations, such as those included in the 1000 Genomes Project. By contributing information on this previously unreported genetic variation, as well as building a reference data set for the population, we believe that the SweGen data set will provide a valuable national resource for genetics research and clinical diagnostics. In fact, information from our website (https://swefreq.nbis.se) has already started to be used within the health-care system. Furthermore, it will serve as an international resource, for example, in the United States a substantial part of the present population was founded by immigrants from the Nordic countries, or as a geographical data point for larger studies of human population genetics.

## Data availability

The SweGen variant frequency data are available from the site https://swefreq.nbis.se (doi = 10.17044/NBIS/G000003). Individual positions can be browsed using a local installation of the ExAC browser.[22] Flat files containing SNV, indel and SV frequency data for the whole genome are available upon registration and agreement to terms and conditions for data download. The SNV and indel data are also deposited in dbSNP under batch ID 1062836 and will be included in dbSNP build 151 (scheduled for release during 2017). Access to phenotypic information can be requested from the Swedish Twin Registry (http://ki.se/en/research/the-swedish-twin-registry).

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

This work was funded by Science for Life Laboratory (SciLifeLab) as a National Project, supported by the Knut and Alice Wallenberg Foundation (2014.0272), and the Swedish Research Council (PI:UG). Illumina sequencing was performed by the National Genomics Infrastructure (NGI), which is hosted by SciLifeLab in Stockholm and Uppsala. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC sens2016003. PO, MM and BN were financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. LF was supported by ERC Starting Grant Agreement 282330.

1 1000 Genomes Project Consortium, Abecasis GR, Auton A *et al*: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
2 Besenbacher S, Liu S, Izarzugaza JM *et al*: Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* 2015; **6**: 5969.
3 Boomsma DI, Wijmenga C, Slagboom EP *et al*: The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* 2014; **22**: 221–227.
4 Gudbjartsson DF, Helgason H, Gudjonsson SA *et al*: Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 2015; **47**: 435–444.
5 UK10K Consortium, Walter K, Min JL *et al*: The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–90.
6 Fakhro KA, Staudt MR, Ramstetter MD *et al*: The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum Genome Var* 2016; **3**: 16016.
7 Nagasaki M, Yasuda J, Katsuoka F *et al*: Rare variant discovery by deep whole-genome sequencing of 1070 Japanese individuals. *Nat Commun* 2015; **6**: 8018.
8 Telenti A, Pierce LC, Biggs WH *et al*: Deep sequencing of 10 000 human genomes. *Proc Natl Acad Sci USA* 2016; **113**: 11901–11906.
9 Wong LP, Ong RT, Poh WT *et al*: Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet* 2013; **92**: 52–66.
10 Humphreys K, Grankvist A, Leu M *et al*: The genetic structure of the Swedish population. *PLoS ONE* 2011; **6**: e22547.
11 Zaboli G, Ameur A, Igl W *et al*: Sequencing of high-complexity DNA pools for identification of nucleotide and structural variants in regions associated with complex traits. *Eur J Hum Genet* 2012; **20**: 77–83.
12 Mathieson I, McVean G: Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012; **44**: 243–246.
13 Jiang Y, Epstein MP, Conneely KN: Assessing the impact of population stratification on association studies of rare variation. *Hum Hered* 2013; **76**: 28–35.
14 Allentoft ME, Sikora M, Sjogren KG *et al*: Population genomics of Bronze Age Eurasia. *Nature* 2015; **522**: 167–172.
15 Mathieson I, Lazaridis I, Rohland N *et al*: Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 2015; **528**: 499–503.
16 Pagani L, Schiffels S, Gurdasani D *et al*: Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am J Hum Genet* 2015; **96**: 986–991.
17 Whiffin N, Minikel E, Walsh R *et al*: Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* 2017; e-pub ahead of print 18 May 2017; doi:10.1038/gim.2017.26.
18 Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.
19 Li H, Handsaker B, Wysoker A *et al*: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.
20 Okonechnikov K, Conesa A, Garcia-Alcalde F: Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 2016; **32**: 292–294.
21 Van der Auwera GA, Carneiro MO, Hartl C *et al*: From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013; **43**: 11 10 11–11 10 33.
22 Dahlberg J, Smeds P: NationalGenomicsInfrastructure/piper: v1.4.0. Zenodo http://doi.org/10.5281/zenodo.154586, 2016.
23 Chen X, Schulz-Trieglaff O, Shaw R *et al*: Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016; **32**: 1220–1222.
24 Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**: 841–842.
25 Abraham G, Inouye M: Fast principal component analysis of large-scale genome-wide data. *PLoS ONE* 2014; **9**: e93766.
26 1000 Genomes Project Consortium, Auton A, Brooks LD *et al*: A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.
27 Danecek P, Auton A, Abecasis G *et al*: The variant call format and VCFtools. *Bioinformatics* 2011; **27**: 2156–2158.
28 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
29 Magnusson PK, Almqvist C, Rahman I *et al*: The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin Res Hum Genet* 2013; **16**: 317–329.
30 Johansson A, Marroni F, Hayward C *et al*: Common variants in the JAZF1 gene associated with height identified by linkage and genome-wide association analysis. *Hum Mol Genet* 2009; **18**: 373–380.
31 McKenna A, Hanna M, Banks E *et al*: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.
32 Lek M, Karczewski KJ, Minikel EV *et al*: Analysis of protein-coding genetic variation in 60, 706 humans. *Nature* 2016; **536**: 285–291.
33 Sherry ST, Ward MH, Kholodov M *et al*: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**: 308–311.
34 Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A: TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Research* 2017; **6**: 664.
35 Pendleton M, Sebra R, Pang AW *et al*: Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* 2015; **12**: 780–786.
36 Seo JS, Rhie A, Kim J *et al*: De novo assembly and phasing of a Korean human genome. *Nature* 2016; **538**: 243–247.
37 Shi L, Guo Y, Dong C *et al*: Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* 2016; **7**: 12065.
38 Bansal V, Libiger O, Torkamani A, Schork NJ: Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010; **11**: 773–785.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)