genes

Article

De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data

Adam Ameur [1,*], Huiwen Che [1], Marcel Martin [2], Ignas Bunikis [1], Johan Dahlberg [3], Ida Höijer [1], Susana Häggqvist [1], Francesco Vezzi [2], Jessica Nordlund [3], Pall Olason [4], Lars Feuk [1] and Ulf Gyllensten [1]

[1] Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, 752 36 Uppsala, Sweden; chehuiw@hotmail.com (H.C.); ignas.bunikis@igp.uu.se (I.B.); ida.hoijer@igp.uu.se (I.H.); susana.haggqvist@igp.uu.se (S.H.); lars.feuk@igp.uu.se (L.F.); ulf.gyllensten@igp.uu.se (U.G.)

[2] Science for Life Laboratory, Department of Biochemistry and Biophysics (DBB), Stockholm University, 114 19 Stockholm, Sweden; marcel.martin@scilifelab.se (M.M.); francesco.vezzi@scilifelab.se (F.V.)

[3] Science for Life Laboratory, Department of Medical Sciences, Molecular Medicine, Uppsala University, 752 36 Uppsala, Sweden; johan.dahlberg@medsci.uu.se (J.D.); jessica.nordlund@medsci.uu.se (J.N.)

[4] Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, 752 36 Uppsala, Sweden; pallolason@gmail.com

* Correspondence: adam.ameur@igp.uu.se

Received: 28 August 2018; Accepted: 5 October 2018; Published: 9 October 2018

Abstract: The current human reference sequence (GRCh38) is a foundation for large-scale sequencing projects. However, recent studies have suggested that GRCh38 may be incomplete and give a suboptimal representation of specific population groups. Here, we performed a de novo assembly of two Swedish genomes that revealed over 10 Mb of sequences absent from the human GRCh38 reference in each individual. Around 6 Mb of these novel sequences (NS) are shared with a Chinese personal genome. The NS are highly repetitive, have an elevated GC-content, and are primarily located in centromeric or telomeric regions. Up to 1 Mb of NS can be assigned to chromosome Y, and large segments are also missing from GRCh38 at chromosomes 14, 17, and 21. Inclusion of NS into the GRCh38 reference radically improves the alignment and variant calling from short-read whole-genome sequencing data at several genomic loci. A re-analysis of a Swedish population-scale sequencing project yields > 75,000 putative novel single nucleotide variants (SNVs) and removes > 10,000 false positive SNV calls per individual, some of which are located in protein coding regions. Our results highlight that the GRCh38 reference is not yet complete and demonstrate that personal genome assemblies from local populations can improve the analysis of short-read whole-genome sequencing data.

Keywords: de novo assembly; SMRT sequencing; GRCh38; human reference genome; human whole-genome sequencing; population sequencing; Swedish population

1. Introduction

Due to advances in DNA sequencing technologies, whole genome sequencing (WGS) has become an established method to study human genetic variation at a population scale. Large human WGS projects have been initiated in several countries and geographic regions [1–6], in some cases comprising 10,000 individuals or more [7,8]. These genome projects will provide a wealth of information for future research on human genetics, evolution, and disease. Today, the vast majority of human WGS is performed using short-read Illumina sequencing technology, and requires an alignment of the sequence reads to a human reference sequence. The gold standard reference is the GRCh38 release from 2013, which is based on DNA from multiple donors and intended to represent a pan-human genome, rather than a single individual or population group [9]. However, the current GRCh38 reference might not be optimal in the context of population specific WGS projects, and more information could be gained from WGS data by instead using local references genomes, tailored to a specific country or population. For instance, the de novo assembly of 150 Danish individuals based on Illumina mate-pair sequencing have strengthened the hypothesis that regional reference genomes can increase the power of association studies and improve precision medicine [10]. Since Illumina's technology is limited by short read lengths and amplification biases [11], it is not a viable alternative for creating human de novo assemblies comparable to GRCh38 in terms of completeness and contiguity.

A number of sequencing technologies have emerged that are capable of reading very long DNA molecules without prior amplification. These methods can resolve complex regions of the human genome, such as GC-rich regions or repeats, which are difficult to determine with amplification-based and short-read approaches [12]. In particular, PacBio's single-molecule real-time (SMRT) sequencing technology has proven to be an excellent method for de novo genome assembly. In 2015, the first human de novo SMRT sequencing project was reported; the assembly of the CHM1 cell line derived from a haploid hydatidiform mole [13]. Since then, a handful of human genomes have been assembled using combinations of long-read, linked-read, and optical mapping technologies [14–17], including the AK1 cell line originating from a Korean individual [15] and the HX1 genome originating from a healthy male Han Chinese [14]. These personal genomes have been assembled to a high level of completeness. For example, the AK1 assembly has a contig N50 size of 17.9 megabases (Mb) and scaffold N50 size of 44.8 Mb, with eight chromosome arms resolved into single scaffolds [15]. The contigs assembled from SMRT sequence data, as opposed to most assemblies based on Illumina data, are completely gap-free and contain no ambiguous bases (represented by N's). Approximately 20,000 structural variations (SVs) are detected by SMRT sequencing of a human individual [14,15] and a majority of these SVs are missed by analyses of short-read Illumina WGS data [16]. The assemblies generated using SMRT sequencing have also indicated that a substantial amount of the sequence is missing from the GRCh38 version of the human reference. For example, 12.8 Mb of novel sequences (NS) were detected in the Chinese HX1 assembly [14], and a recent study of 17 individuals from five diverse populations sequenced using linked-read technology revealed 2.1 Mb of NS [18]. Also, an average of 0.7 Mb per individual not present in GRCh38 was found among the 10,000 samples in a population-scale Illumina WGS project [8], showing that NS can, to some extent, also be detected in short read data.

The human de novo assemblies available based on long-read data thus indicate that each personal genome contains a significant amount of dark matter of SV that is not detected by short-read WGS, and several million bases of a NS that cannot be matched to GRCh38. At present, it is unknown how many of these SVs and NS are common to all humans, and thus represent errors in the GRCh38 reference, and how many of them are polymorphic between individuals. To address this question, there is a need to assemble several personal genomes from different populations around the world to a high degree of completeness. Such a collection of de novo genomes would make it possible to improve on GRCh38, and eventually to construct complete new population-specific genomes. In this study, we performed de novo assembly of genomes from two individuals from the Swedish population, in order to investigate the missing pieces of GRCh38, and to evaluate the benefits using a local reference for single nucleotide variant (SNV) calling in population-based WGS data.

2. Materials and Methods

2.1. Samples

Swe1 and Swe2 were selected from the 1000 individuals included in the SweGen project [1]. Samples of whole blood from these two individuals were collected in 2006 and frozen without the separation of white and red blood cells at −70 °C on site, as part of the Northern Sweden Population Health Study

(NSPHS), which aims to study the medical consequences of lifestyle and genetics. Genomic DNA was extracted using organic extraction. 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.

2.2. PacBio Library Preparation and Sequencing

Four PacBio libraries were produced for each of the Swe1 and Swe2 samples using the SMRTbell™ Template Prep Kit 1.0 (PacBio, Menlo Park, CA, USA) according to the manufacturer's instructions. In brief, 10 µg of genomic DNA per library was sheared into 20 kb fragments using the Megaruptor system, followed by exo VII treatment, DNA damage repair, and end-repair before ligation hair-pin adaptors to generate SMRTbell™ libraries for circular consensus sequencing. Libraries were then subjected to exo treatment and PB AMPure bead wash procedures for clean-up before they were size selected with the BluePippin system with a cut-off value of 9500 bp. The libraries were sequenced on the PacBio RSII (PacBio, Menlo Park, CA, USA) instrument using C4 chemistry and P6 polymerase, and a 240 min movie time in a total of 225 SMRTcells™ per sample.

2.3. De Novo Assembly of SMRT Sequencing Reads

Raw data was imported into SMRT Analysis software 2.3.0 (PacBio) and filtered for subreads longer than 500 bp or with a polymerase read quality above 75. A de novo assembly of filtered subreads was generated using FALCON [19] assembler version 0.4.1 (configuration file is provided as Supplementary Information). In order to improve the accuracy of the assembly, two rounds of sequence polishing were performed using the Quiver consensus calling algorithm [19]. For subsequent analysis, primary contigs shorter than 20 kb were excluded from the Swe1 and Swe2 assemblies to reduce putative assembly errors. This is slightly more conservative compared to the Korean AK1 study [15], where a 10 kb cut-off of primary contigs was used.

2.4. Generation of BioNano Optical Maps and Hybrid Assembly

DNA extraction for optical maps was performed at BioNano Genomics (San Diego, CA, USA), starting from frozen blood from Swe1 and Swe2. Optical mapping was performed on the Irys system (BioNano Genomics) using the two labeling enzymes BssSI and BspQI for each individual. The resulting data was used for a two-step hybrid assembly of the PacBio contigs using the IrysView software.

2.5. The hg38 Reference Genome

The hg38 reference genome used in this study is identical to the original, full analysis set of GRCh38 (accession GCA_000001405.15) described in a study by Zheng-Bradley et al. (2017) [20]. This implies that hg38 consists of the primary GRCh38 sequences (autosomes and chromosome X and Y), mitochondria genome, un-localized scaffolds that belong to a chromosome without a definitive location and order, unplaced scaffolds that are in the assembly without a chromosome assignment, the Epstein-Barr virus (EBV) sequence (AJ507799.2), ALT contigs, and the decoy sequences (GCA_000786075.2).

2.6. Quality Control and Alignment of the Two Swedish De Novo Assemblies

Components of MUMmer3 [21] (NUCmer, delta-filter, and dnadiff) were used to assess the quality of the Swe1 and Swe2 de novo assemblies and to perform genome alignments. NUCmer (-maxmatch–l 150 –c 400) was used to align each of the assemblies to hg38. After the alignments, delta-filter (-q) was used to filter out repetitive alignments and to keep the best alignment for each assembled contig. Summary statistics for the filtered whole genome alignments were generated by dnadiff.

2.7. Detection of Structural Variation in PacBio Data

We utilized NGMLR (v0.2.3) (https://github.com/philres/ngmlr) and Sniffles [22] (v1.0.5) to detect SVs from PacBio long reads. Filtered subreads (min subread length 500 bp and polymerase read quality above 75) were first aligned to the hg38 using NGMLR with default parameters. Sniffles (-s 10 –l 50) was subsequently used to identify SVs ≥ 50 bp with at least 10 reads support. Only SV detected on chr1-22, X, and Y from Swe1, and on chr1-22 and X from Swe2, were kept for analysis.

2.8. Detection of Novel Sequences

To identify NS, we performed two rounds of sequence mapping using NUCmer [21]. The first round of mapping was the same as described above, when aligning contigs to the hg38 reference. Contigs and part of contigs that failed to align to the reference in the first step were then processed in a second round of mapping, where more relaxed settings were used in an attempt to have more sequences aligned by NUCmer (-maxmatch–l 100–c 200). Duplicated sequences were then removed to obtain a set of NS for each of the two individuals (Swe1 and Swe2). All NS in the final set have a sequence length of at least 100 bp, with a sequence identity to the hg38 reference that is less than 80%.

2.9. Repeat Analysis and BLAST Comparison of Novel Sequences

Repeats in NS were analyzed using RepeatMasker (-species human -s–x; http://www.repeatmasker.org). NS were searched against the nucleotide collection database using BLAST [23] (2.2.31+) (-max_target_seqs 1–task blastn–num_threads 16). In order to obtain matched sequences of a relatively high similarity, the BLAST results were post processed by setting an E-value threshold at 10^{-50} and by keeping only the top hit for each NS.

2.10. Anchoring Novel Sequences on Human Chromosomes

To determine the potential genomic position of the NS that may be anchored, we first used NUCmer (-maxmatch–l 100–c 200) and delta-filter (-q) to map the NS to the hybrid scaffolds that were generated by PacBio contigs and BioNano optical maps. After anchoring of the hybrid scaffolds, NUCmer (-maxmatch–l 150–c 400) and delta-filter (-q) were ran to identify the location of the alignments on hg38 chromosomes. NS that mapped to anchored hybrid scaffolds were further analyzed to identify unique or multiple location anchors. NS that were anchored to decoy sequences included in the hg38 reference were excluded from the final results.

2.11. Construction of an Extended Reference Based on Swedish Novel Sequences

Novel sequences detected in Swe1 and Swe2 were appended to hg38 to create an extended version of the human reference sequence (named hg38+NS). For NS overlapping between both Swedish individuals, only the Swe1 version of the NS was used. The resulting hg38+NS reference added 17.3 Mb of NS to hg38.

2.12. Re-Alignment of SweGen Illumina Data to hg38 and hg38+NS

In total, 200 of the SweGen samples [1] were processed with the Cancer Analysis Workflow (CAW) pipeline (https://github.com/SciLifeLab/CAW) in normal-only mode (no tumor samples), once with hg38 and again with hg38+NS as the reference. CAW implements a workflow based on GATK best practices. In summary, reads were aligned using BWA-MEM 0.7.15 with the ALT-aware option turned off. Duplicates were then marked with Picard’s MarkDuplicates 2.0.1. The tools RealignerTargetCreater, IndelRealigner, CreateRecalibrationTable, HaplotypeCaller, and GenotypeGVCFs from GATK 3.7.0 (https://software.broadinstitute.org/gatk/) were then used in that order to realign around indels, recalibrate base qualities, and call variants, respectively, resulting in a final CRAM and VCF file for each sample.

Also, a similar analysis was re-run for 150 of the 200 SweGen samples, but with the ALT aware option turned on in the BWA-MEM alignment. This ALT aware analysis resulted in a much higher number of lost and gained SNVs compared to the non-ALT aware alignment. We therefore decided to focus on the non-ALT aware analysis in this study, i.e., the analysis run on the 200 samples. By a non-ALT aware alignment, we get a conservative estimate of the number of lost and gained SNVs and do not exaggerate the effect of adding NS to the hg38 reference.

2.13. Analysis and Annotation of SNVs in SweGen Re-Alignments

To detect SNVs that were consistently gained or lost among the 200 SweGen samples when the NS were added to hg38, we employed a filtering strategy using custom scripts in Perl and R. ANNOVAR [24] was used to annotate gained and lost SNVs with information about human genetic variation from dbSNP [25] v147 and protein coding genes from the NCBI RefSeq database [26].

3. Results

3.1. De Novo Assembly of Two Swedish Individuals

To construct two high-quality genome references for the Swedish population, DNA was extracted from blood samples obtained from one male (Swe1) and one female (Swe2). The two individuals were unrelated and selected from the 1000 samples included in SweGen, which is a project where the genetic variation in a cross-section of the Swedish population was studied using Illumina WGS [1]. A principal component analysis (PCA) shows that Swe1 and Swe2 are relatively distant from each other in the context of the genetic variation within Sweden (Figure 1A). The long tail of SweGen samples that are intermixed with the Finnish genomes mainly represents genome sequences from the northern parts of the country, and thus the two genomes contain a large portion of the common genetic variation in the Swedish population.

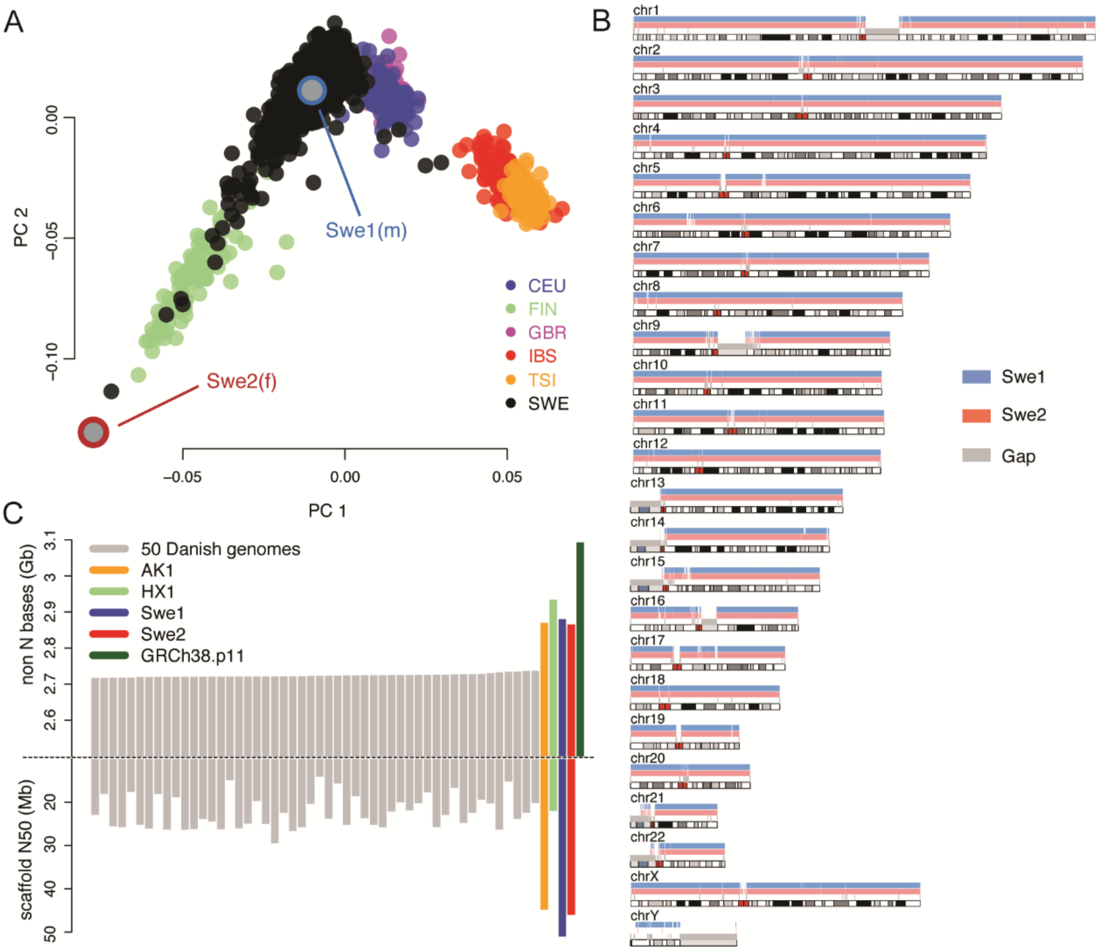

Figure 1. Selection of individuals and de novo assembly results. (**A**) Results of principal component analysis (PCA) of whole genome sequencing (WGS) data from the SweGen project [1], compared to the European 1000 Genomes data [27] (CEU: Utah Residents with Northern and Western Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian Population in Spain, TSI: Toscani in Italia). The black dots indicate 942 samples from the Swedish Twin Registry (STR), which were sequenced within the SweGen project and represent a cross-section of the Swedish population. Swe1 and Swe2 are the individuals selected for de novo sequencing. (**B**) Alignment of contigs for Swe1 (blue) and Swe2 (red) to the human GRCh38 reference. A total of 6812 contigs could be aligned for Swe1 and 6924 for Swe2. Only the male Swe1 sample has extensive coverage of the Y chromosome. (**C**) The bars show the total number of non-N bases (top) and scaffold N50 values (bottom) for Swe1, Swe2, and a selection of other human de novo assemblies. The grey bars represent the top 50 genomes with the highest number of non-N bases from an Illumina mate-pair assembly of 150 individuals [10]. The Korean (AK1) and Chinese (HX1) genomes were assembled by a combination of single-molecule real-time (SMRT) sequencing and optical mapping. Scaffold N50 is not shown for GRCh38 (in green) since it is much higher than for the personal genomes and difficult to fit into the same plot.

SMRT sequencing data was generated at an average coverage of 78.7× for Swe1 and 77.8× for Swe2 (Table S1). By de novo assembly [28], followed by two iterations of genome polishing, we were able to construct sequence assemblies of 2.996 Gb and 2.978 Gb for Swe1 and Swe2, consisting of 7166 and 7186 contigs, respectively (Table S2). Each of the assemblies contained about 3000 primary contigs and an additional 4000 alternative contigs originating from regions with high heterozygosity. The alternative contigs only cover a small fraction of the genome; about 115 Mb in each individual. N50 values for the primary contigs were 9.5 Mb for Swe1 and 8.5 Mb for Swe2. For both individuals, we also generated BioNano optical mapping data with two different labeling enzymes, at over 100× coverage per enzyme. A two-step hybrid scaffolding of the SMRT sequencing contigs together with the optical maps resulted in assemblies of size 3.1 Gb and scaffold N50 of 49.8 Mb (Swe1) and 45.4 Mb (Swe2) (Table S3). These numbers are similar to the 44.8 Mb scaffold N50 obtained for the first published Korean genome [15], and substantially larger than the median scaffold N50 of 21 Mb obtained for 150 Danish genomes [10]. It is worth noting that the DNA samples used for optical mapping of Swe1 and Swe2 were extracted from blood collected in 2006. Our results thus demonstrate that it is possible to obtain very high-quality genome assemblies starting from frozen blood that has been stored in the freezer for over a decade.

3.2. Evaluating the Quality of the De Novo Assemblies

To assess the quality of the two de novo assemblies, we aligned the contigs for Swe1 and Swe2 to the hg38 reference genome. Throughout this article, the abbreviation hg38 is used to denote a sequence that is identical to the full analysis set of GRCh38, which includes un-localized scaffolds and decoy sequences [20] (see Methods). For Swe1 and Swe2, respectively, 2.971 Gb (99.14%) and 2.956 Gb (99.24%) of the assembled sequence could be uniquely aligned to hg38 (see Figure 1B and Table S4). The slightly higher number of aligned bases for Swe1, who is a male, can be explained by sequences on the Y chromosome that are not present in the female Swe2 sample. The average identity between the contigs and hg38 was over 99.7% for both genomes. Intriguingly, a higher fraction of the Swe2 sequence data can be uniquely aligned to the Swe1 de novo assembly (99.55%) compared to hg38 (99.24%), thus suggesting that the hg38 reference does not contain all sequences present in these Swedish individuals. The corresponding analysis for Swe1 is not relevant in this context, since Swe1 is expected to contain a sequence on the Y chromosome not present in the female Swe2.

In order to discover a NS missing from hg38, it is essential that Swe1 and Swe2 were assembled to a high degree of completeness and with as few gaps as possible. To evaluate this, we compared our Swedish de novo assemblies to results obtained for the Korean AK1 [15], the Chinese HX1 [14], and 150 Danish genomes [10]. As seen in Figure 1C, the primary contigs of Swe1 and Swe2 contain a similar number of unambiguous (non-N) bases as the other SMRT sequencing assemblies (i.e., AK1 and HX1). Importantly, the assemblies obtained from SMRT sequencing contain over 100 Mb of additional sequence as compared to the Illumina mate-pair assemblies. The GRCh38 reference contains almost 3.1 Gb, which is significantly more compared to the ~2.9 Gb for Swe1 and Swe2. To a certain extent, these differences can be explained by the fact that GRCh38 is based on a combination of DNA sequences and haplotypes from several individuals [9], which could lead to an inflated genome size, and also that primary contigs shorter than 20 kb were excluded from the Swe1 and Swe2 assemblies. N50 scaffold values are highest for the Swe1, Swe2, and AK1 assemblies, which all used BioNano data from two labeling enzymes for hybrid scaffolding. A single enzyme was used for HX1 and this assembly has a scaffold N50 similar to those obtained for the Danish genomes.

3.3. Structural Variation in Swedish Genomes

Analysis of SV resulted in a total of 17,936 SVs for Swe1 and 17,687 SVs for Swe2 (Table S5). These numbers can be compared with the 20,175 and 18,210 SVs detected in the Chinese HX1 and in the Korean AK1 assembly, respectively. The SV length distribution shows an enrichment of ALU repeat elements at around 300 bp and of LINE elements at around 6100 bp, similar to what has been previously reported [15] (Figure S1).

3.4. Detection of Novel Sequences Not Present in the Human Reference

Even though most of the contigs in our assemblies were in good agreement with the human reference, we detected 25.6 Mb of sequences in Swe1 and 22.6 Mb in Swe2 that could not be aligned to hg38 (Table S4). To refine these sequences further, we performed a two-step re-alignment using more relaxed settings and removed duplicated sequences (see Methods). This resulted in 2859 NS in Swe1 of a total length of 13.8 Mb, and 2786 NS in Swe2 of a total length of 10.6 Mb (Table S6 and Data S1). The NS were required to be at least 100 bp in length and have at most 80% identity to hg38. They could either originate from contigs that could not be aligned to hg38, or from inserted elements in the aligned contigs. As seen in Figure 2A, most of the NS are relatively short (between 100 bp and 5 kb). However, 83% of NS bases originate from sequences that are over 5 kb in length.

Repeat masking using sensitive settings, revealed an abundance of repetitive elements in the NS (see Figure 2B and Table S7). For Swe1, 88.58% of the NS bases were found to be repetitive. A slightly lower repeat content, 83.60%, was detected in the NS from Swe2. Since the repeat content is around 50% among all Swe1 and Swe2 contigs, there is a high enrichment of repeats in the NS. Also, the GC level is slightly elevated, with values of 42.68% (Swe1 NS) and 43.45% (Swe2 NS), compared to 40.95% in all contigs (Table S8). SMRT sequencing is known to perform well in repetitive regions and high-GC regions, and therefore these results are not unexpected. Annotation of all the repeats showed that satellites and simple repeats make up 82% of the NS bases, but only 3% of all of the primary contig sequences (see Table S7). Interestingly, all other groups of repeats are underrepresented among the NS. Even though a high proportion of the NS are repetitive, there is also a substantial amount of non-repetitive sequence. For Swe1 and Swe2, 1.58 Mb and 1.73 Mb of NS remained after the repeat masking.

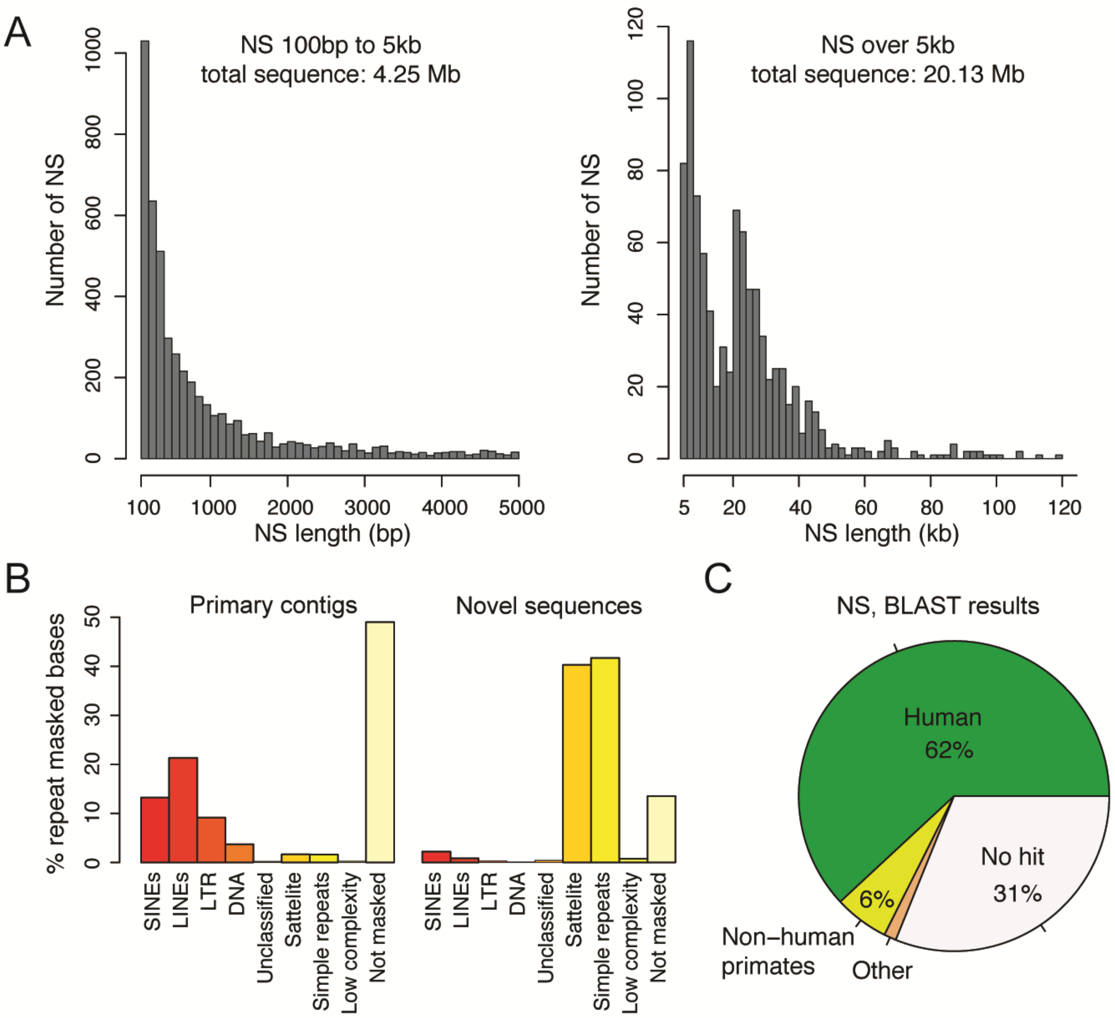

Figure 2. Characterization of novel sequences (NS) found in Swe1 and Swe2. (**A**) The histograms show the length distribution of all NS found in Swe1 and Swe2. Shorter NS are displayed in the left panel (100 bp to 5 kb), and longer NS are shown in the right panel (>5 kb). The longer NS comprise the majority of the NS in Swe1 and Swe2. (**B**) Results of repeat masking in primary contigs (left) and NS (right). Within the primary contigs, 51% of the bases are found to be repetitive using the repeat masker software, while 86% of the bases are repetitive within the NS. Satellite and simple repeats make up 82% of the bases in the NS. (**C**) Results of matching the 5645 NS in Swe1 and Swe2 to the NCBI database using BLAST [23]. Each piece of the pie chart represents the number of NS that were assigned to a particular species as the top hit. The No hit category (in white) contains NS where no E-value reached 10^{-50} or lower. A total of 72 of the

NS are in the Other category, which includes matches to a number of parasitic worms (both for Swe1 and Swe2) and a complete human papilloma virus 35 (HPV35) genome (only for Swe2).

3.5. Origin of the Novel Sequences

To further investigate the contents of the NS, we performed a BLAST search [29] against all sequences present in the NCBI database (see Figure 2C, Table S9 and Data S2). Thirty-one percent of the NS did not produce a BLAST hit, implying that they have not been previously reported. For the remaining NS, the majority matched to human entries in NCBI, thereby suggesting that a majority of our NS have been detected previously, but originate from regions or haplotypes that have not been included in the hg38 reference. We also detected 5% non-human primate sequences, which most likely originate from regions missing in hg38 that have been sequenced in another primate. Nearly 1% of the NS match to other, non-primate, species. Of note, several of the hits show high similarity to parasitic worms, including *Spirometra erinaceieuropaei*, *Enterobius vermicularis*, and *Dracunculus medinensis*. Since it is highly unlikely that the two Swedish individuals indeed have DNA from these parasites present in their blood, a more plausible explanation is that the worm genome assemblies contain a fraction of human sequence. The initial worm assemblies were based on short-read sequencing of samples extracted from human patients [30], and contigs not aligning to GRCh38 may have been mis-annotated as the worm sequence, thus explaining the overlap with our NS. Notably, the Swe2 sample also contained a complete human papilloma virus 35 (HPV35). This could either originate from an HPV35 infection in the blood of this individual, or from a contamination in the sample.

3.6. Comparing Novel Sequences between Swedish Individuals and the Chinese HX1

To investigate whether the NS are individual-specific, population-specific, or shared between different populations, we compared our results to those obtained for the Chinese HX1 genome [14]. The HX1 assembly was based on 103× genome-wide SMRT sequencing of DNA from a human blood sample, where 12.8 Mb of NS was found. Starting from the NS identified in Swe1 or Swe2, we determined whether the same sequences could be identified in the other Swedish assembly, or in HX1 (see Figure 3A and Table S10). For Swe1 and Swe2, 55% (7.65 Mb) and 52% (5.51 Mb) of the NS, respectively, could also be found in the other Swedish individual, as well as in HX1. The higher overlap obtained when starting the analysis from Swe1 is explained by certain repetitive elements that occur with a higher copy number in Swe1 compared to Swe2. Our results also show the presence of over 5 Mb of NS in the three-way overlap category, i.e., found in all three individuals. A smaller amount of NS (~1.5 Mb) was only common between the two Swedish individuals, while not found in the Chinese HX1, thus representing a possible population-specific sequence. We also identified a substantial amount of individual-specific sequences, 3.27 Mb for Swe1 and 3.22 Mb for Swe2. Interestingly, a much higher amount of NS were shared between Swe1 and HX1 (1.36 Mb) compared to Swe2 and HX1 (0.29 Mb). Since Swe1 and HX1 are both males, while Swe2 is a female, the ~1 Mb of additional NS shared between Swe1 and HX1 may at least partly be explained by segments of the Y chromosome that are missing from the hg38 reference.

3.7. Anchoring Novel Sequences on Human Chromosomes

We next aimed to anchor the NS onto human chromosomes using information provided by the PacBio long-read data and BioNano optical maps (see Methods section). Only a minority of sequences could be placed into the human genome using this approach, but this analysis still provided valuable insights about the genomic localization of the NS (see Figure 3B). For Swe1 and Swe2, 2.08 Mb and 1.97 Mb of NS, respectively, could be uniquely anchored to a chromosome, while 1.70 Mb and 1.55 Mb were anchored to multiple chromosomes (see Table S11). This again shows that many NS contain repetitive or transposable elements. For the uniquely anchored NS, we observed an accumulation at certain chromosomes. The highest amount of three-way overlap sequences is present on chromosome 21, while chromosomes 13, 14, and 22 also show enrichment (Figure 3C). The NS placed on these chromosomes are mainly localized to centromeric or telomeric regions, suggesting that placement of these sequences has previously been difficult to determine due to their repetitive content. Interestingly, we detected an accumulation of population-specific NS present in both Swedish individuals, but not in the Chinese HX1 on chromosome 17. A relatively large amount of sequences (121 kb) shared only between the two male individuals (Swe1 and HX1) could be anchored to the Y chromosome. Surprisingly, we also noted an accumulation of NS shared between Swe1 and HX1 on chromosome 17.

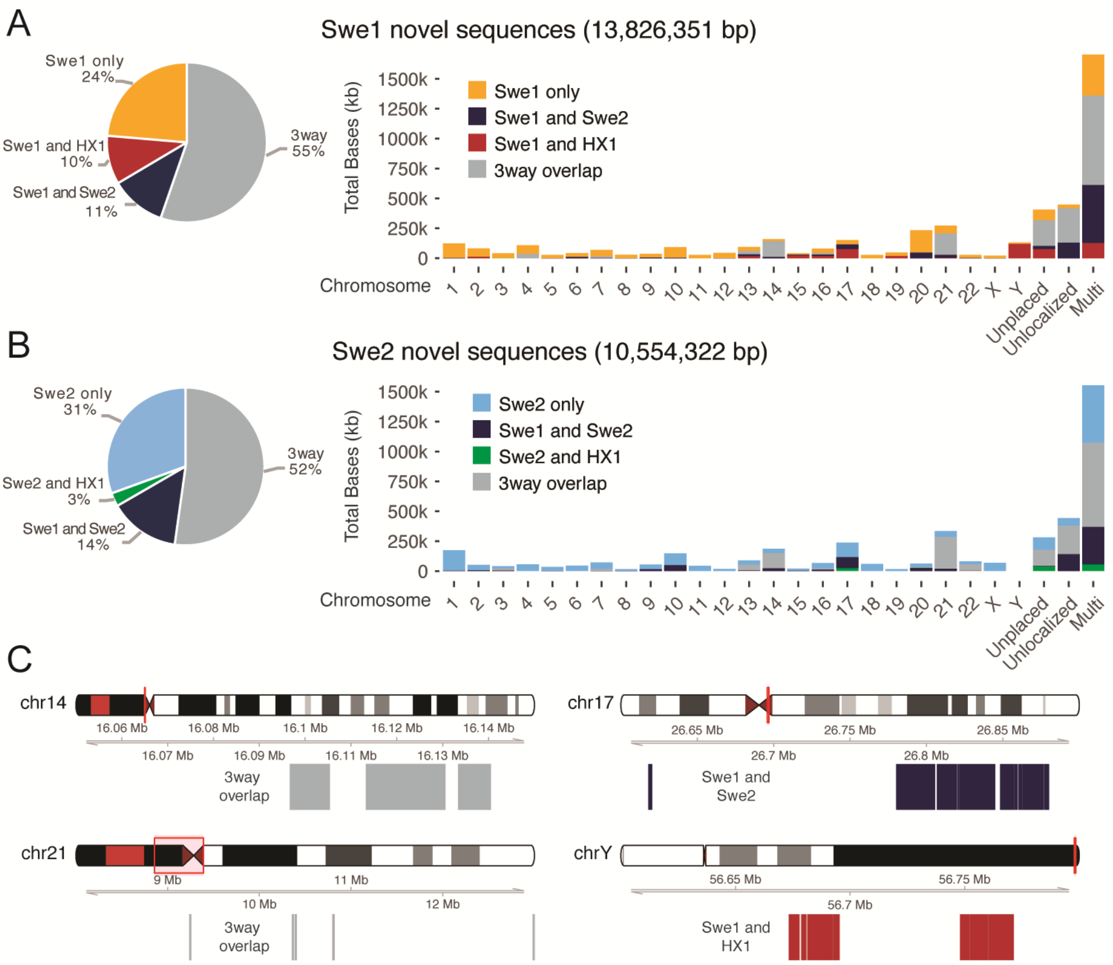

Figure 3. Anchoring of Swe1 and Swe2 NS to the hg38 reference. (**A**) The pie chart to the left shows the proportion of Swe1 NS (in total 13.8 Mb) that are also found in Swe2 or in the Chinese HX1. The category 3way (grey) represents NS that are found in all three individuals. The bars to the right show the amount of NS that can be anchored to the hg38 genome. The category unplaced represents sequences in hg38 that are not associated with any chromosome, and unlocalized corresponds to sequences that are associated with a specific chromosome but have not been assigned an orientation and position. The multi category furthest to the right represents NS that are mapping to multiple chromosomes. (**B**) Similar results for NS detected in Swe2. (**C**) Examples of chromosomal regions where a high amount of NS are detected. The two plots to the left show the localization of 3way overlap sequences (i.e., found in Swe1, Swe2, and HX1) near the centromeric regions of chr14 and chr21. The top right panel displays a region on chr17 where an excess of NS found only in Swe1 and Swe2 could be anchored. The bottom left panel shows NS detected only in the two males (Swe1 and HX1) that could be anchored to regions close to the telomere of chromosome Y.

3.8. Application of Novel Sequences for Population Scale WGS Analysis

Having identified several Mb of DNA not present in the human reference, we were interested to see whether these NS would improve the results of whole genome re-sequencing of the Swedish population. We therefore created a new reference consisting of hg38 combined with all the NS detected in Swe1 and Swe2 (named hg38+NS), after which we leveraged the Illumina WGS data from the SweGen dataset [1] and aligned the reads from 200 individuals both to hg38, as well as to hg38+NS. The aim of this analysis was to study whether the number of SNVs was altered as a result of appending NS to the reference, through an analysis procedure outlined in Figure 4A.

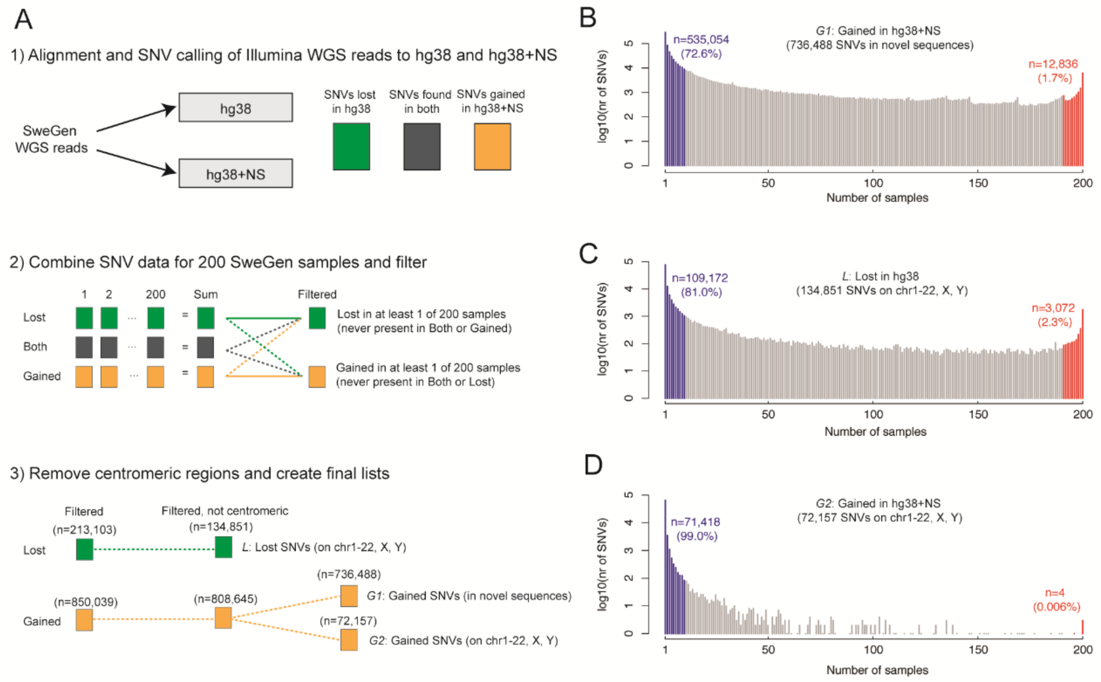

Figure 4. Re-analysis of Illumina WGS data using a Swedish human reference. (**A**) Overview of our method to evaluate the effect of NS on SNV calls from Swedish Illumina WGS data. In the first step, reads from 200 SweGen samples [1] were aligned both to hg38 and to an extended reference (hg38+NS), where 17.3 Mb of NS detected in Swe1 and Swe2 were appended to hg38. In step 2, single nucleotide variants (SNVs) for each of the samples were sorted into three groups: (i) SNVs found only in hg38, but not in hg38+NS (named Lost, in green); (ii) SNVs found both in hg38 and hg38+NS ('Both', in grey); and (iii) SNVs found only in hg38+NS, but not in hg38 (Gained, in orange). After such SNV tables were generated for all 200 individuals, a summary file was created for the Lost and Gained group. The Lost SNVs were not allowed to be detected in any of the Gained or Both files. A similar filtering was also performed for the Gained group. In step 3, we further filtered the SNV lists by removing all centromeric regions (from file centromeres_UCSC_hg38.txt). The resulting Gained SNVs were separated into two distinct groups, those present in hg38 chromosomes (chr1-22, X or Y) and those present in the NS. (**B**) Frequency distribution of the 736,488 SNVs that were gained in the NS. The x-axis shows the not peer-reviewed is the author/funder. It is made available under a CC-BY 4.0 International license. bioRxiv preprint first posted online on 18 February 2018; doi:http://dx.doi.org/10.1101/267062. The copyright holder for this preprint (which was 21 SweGen samples (out of 200) and the *y*-axis show the number of gained SNVs for each number of samples on a log10-scale. Most of the gained SNVs are detected only in a few samples. The blue and red areas show the number of SNVs that are gained in at most 5% and at least 95% of samples, respectively. (**C**) Frequency distribution of the SNVs that were lost in hg38 when adding NS to the hg38 reference. (**D**) Frequency distribution of the gained SNVs on chromosomes 1-22, X, or Y (i.e., not in NS) when adding NS to the hg38 reference.

An average of 42.5 SNVs per kb was detected in the NS and their frequency distribution is shown in Figure 4B. Some of these SNVs are likely to represent true novel genetic variation in the cohort, but a fraction may also originate from errors in the Swe1 and Swe2 assemblies. Surprisingly, the addition of NS to the reference had a large effect on variant calls of the autosomes and sex chromosomes, where 134,851 SNVs disappeared (outside of centromeric regions) when the extended reference was used (Figure 4C). These SNVs originate from reads that preferentially align to a NS and can be considered as false positives in hg38. Interestingly, we also found a substantial number of SNVs (n = 72,157) which were gained on the hg38 chromosomes when using the extended reference. These gained SNVs in hg38 have overall lower allele frequencies compared to the lost SNVs (see Figure 4D).

Finally, we investigated SNVs that were consistently lost or gained in hg38 for at least 5% of the 200 SweGen samples when using the extended reference (see Figure 5A). Only a small number of SNVs (n = 823) were gained on the hg38 chromosomes in at least 5% of the samples when using the extended reference. However, 26,724 SNVs were lost in at least 5% samples when appending NS to the hg38 reference. These consistently lost SNVs have an uneven distribution over the genome, with the highest peak on chrY and smaller peaks on several other chromosomes. Global annotation of the consistently lost SNVs showed that 7130 (27%) of these are present in version 147 of dbSNP. For the consistently gained SNVs, only 130 (16%) are present in dbSNP, suggesting that these SNVs are more difficult to detect using the hg38 reference alone. A total of 109 consistently lost SNVs were located in a coding sequence of a gene, but none of the consistently gained SNVs were in coding regions. Figure 5B shows an example region on chr17 where the NS improved the alignment of Illumina WGS data for two SweGen individuals, resulting in the removal of around 100 false positive SNVs, and importantly, the discovery of seven novel SNVs that were previously masked by the mis-aligned reads. A second example is shown in Figure 5C where a region on chrY with about 1000× coverage and many dubious SNVs are cleaned up when NS are appended to hg38. In a third example, as illustrated by the genome browser view of the *FRG2C* locus, the hg38+NS reference improves alignments in coding regions (see Figure 5D).

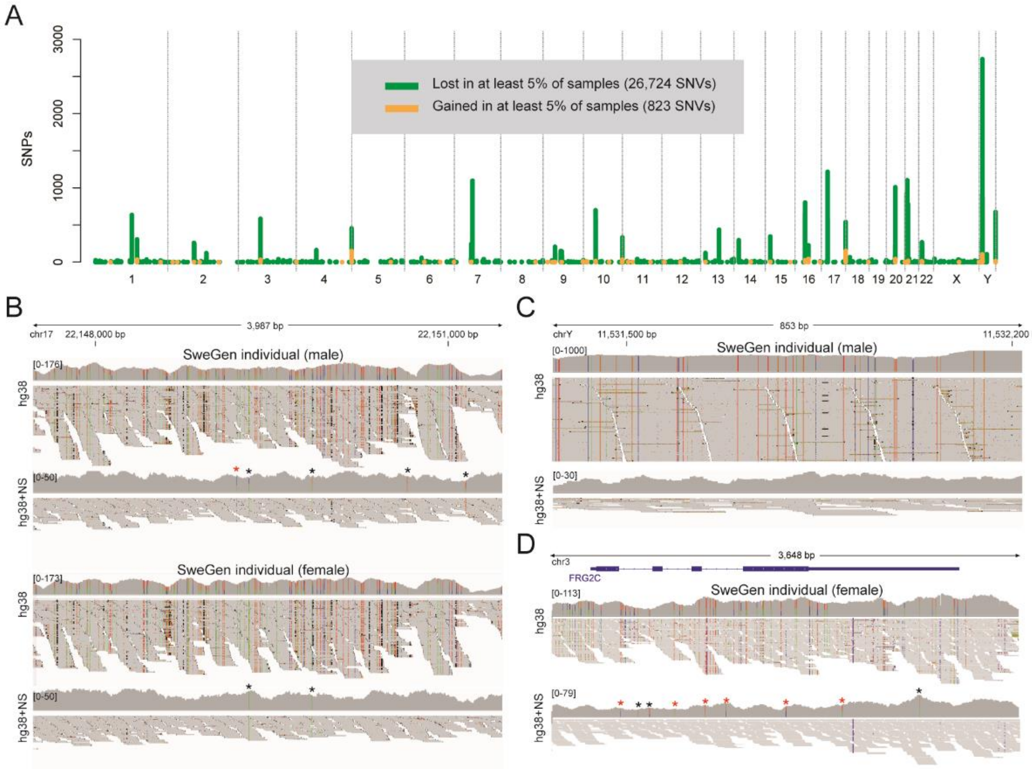

Figure 5. A novel reference gives improved alignment and SNV calling of SweGen WGS data. (**A**) Genomic distribution of SNVs that are lost (green) and gained (orange) when NS are appended to the hg38 reference. Only non-centromeric SNVs that are lost/gained in at least 5% of the 200 SweGen samples are shown in this figure. (**B**) An IGV [31] view of Illumina reads for two representative SweGen samples at a region on chr17, where some SNVs are lost and others are gained when using the hg38+NS reference. Illumina data is shown for a male and a female (not the same individuals as Swe1 and Swe2). Both for the male and female, the coverage decreases over the region when NS are appended to hg38, and about 100 (homozygous) false positive SNV calls are lost in each of the samples. Only five heterozygous SNVs where found for the male individual when the novel reference was used, and two homozyogous SNVs for the female (marked by asterisks '*'). A red asterisk indicates a gained SNV that is not detected in hg38. (**C**) An example region

on chrY where the coverage was reduced from almost 1000× to below 30× when using hg38+NS, and where a large number of SNVs were lost. Only data for the male individual is shown in this panel. (**D**) Improved alignment and SNV calling over the *FRG2C* locus on chromosome 3. A large number of SNVs were lost, and six SNVs were gained (red asterisks '*'), in the female SweGen sample. Some of the lost and gained SNVs are located in the coding sequences of *FRG2C*.

4. Discussion

Swe1 and Swe2 represent two of the most complete individual human de novo assemblies produced to date. On average, the primary contigs contain 2.87 Gb of unambiguous (non-N) sequence per individual, which is similar to the Chinese HX1 [14] and the Korean AK1 [15] individuals, and 133 million more bases than what could maximally be assembled in any of the 150 Danish genomes [10]. Even though our Swedish assemblies are comparable to HX1 and AK1, both in terms of quality and completeness, they are still not entirely complete. To a large extent, this incompleteness can be explained by the current limitations of the sequencing and optical mapping technologies used in this study. Not even the longest DNA molecules in our data can bridge repeats that are spanning over several mega bases, and this implies that our assemblies will break down at such loci. Eventually, the Swe1 and Swe2 assemblies could be further improved by BAC sequencing of specific regions [15], by chromosome interaction mapping [32], by linked-reads from 10× Genomics [15,17], or long reads from nanopore-based sequencing [33]. By combining data from these technologies, it would also be possible to phase haplotypes, and to generate diploid sequences over large parts of the Swe1 and Swe2 genomes.

With the new sequencing technologies, it is possible to assemble complete human genomes starting directly from tissues or blood, instead of using cell line samples that have been the traditional source of DNA for human reference genomes. This is of importance since it should resolve the potential issue with genomic aberrations introduced during cell line transformation and long-term culturing [34]. In this study, we have therefore chosen to compare the assemblies of Swe1 and Swe2 to the Chinese HX1 genome, which was also obtained from a blood sample, rather than the Korean AK1 genome, which was based on cell line DNA.

The Swedish de novo assemblies reveal a large amount of NS not present in the GRCh38 reference. Over 5 Mb of NS were overlapping between our two Swedish genomes and the Chinese HX1, and this likely represents a valid human DNA sequence. In addition, 1.36 Mb of male-specific sequence was found (i.e., overlapping between Swe1 and HX1) and a substantial fraction of these sequences could be anchored to the Y chromosome. We estimate the total amount of sequence missing from GRCh38 to be at least 6 Mb, which corresponds to about 0.2% of the size of the human genome. Another interesting class of NS are the ~1.5 Mb found in Swe1 and Swe2, but not in HX1. Several of these potentially population-specific NS are clustered at certain genomic regions, such as chr17 (see Figure 3C). Some of these regions could have been targets for selection during human evolution, although this needs to be investigated further.

At this point, there is no evidence for the presence of functional elements within the NS, although preliminary data suggest that some of them are actively transcribed (data not shown). A more thorough analysis would be required to shed light on the functional relevance of these NS in the different cell types in the human body. For example, this could involve searching for open reading frames, conserved sequences, expressed mRNAs, enhancers, and transcription factor binding motifs. It might even be possible to leverage mass spectrometry data to search for peptide sequences missing from the current version of the human genome. Although the functional analysis is highly relevant, it would be a major undertaking that falls outside the scope of this present study.

Our results show that the NS can be used to construct a new version of the human genome reference that improves the analysis of population-scale Illumina WGS data. On average, 10,898 SNVs per individual were lost, and 75,035 SNVs per individual were gained in 200 SweGen samples when appending the NS to GRCh38, with some of this variation affecting the coding sequences of known genes. Because of the stringent filtering options used in our analysis (see Figure 4A), these numbers should be seen as a conservative estimate of the novel variation that could be resolved using an improved reference. In addition, since many regions still show poor alignments for SweGen data also when using hg38+NS (data not shown), it is likely that our reference could be further improved and customized for the Swedish population. This could for example be done by flipping genetic variants so that the common alleles in the Swedish population are represented in the reference sequence, as this is an approach that has been suggested to improve the alignments of population specific NGS data [35]. However, the benefits of an improved reference are likely to be even stronger for other, non-European, population groups that were poorly represented in the original assembly of GRCh38 [9].

5. Conclusions

In conclusion, despite all efforts to refine the human genome since its original release in 2001 [36], our results indicate that substantial improvements could still be made, not least to represent specific population groups, by the de novo assembly of representative human genomes from different populations.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, The Swe1 and Swe2 raw sequence data and assembly files will be made available during 2018 from a local Swedish installation of the European Genome-phenome Archive (EGA) (https://www.ebi.ac.uk/ega) that is now being implemented at Uppsala University and SciLifeLab. The dataset has the following doi:10.17044/NBIS/G000006. In addition, the following data files are made available as supplementary material: Supplementary Data S1: Novel sequences in Swe1 and Swe2; Supplementary Data S2: BLAST hits in novel sequences. Figure S1. Length distribution of structural variants detected in Swe1 and Swe2. (A) Lengths of insertions (red) and deletions (blue) in Swe1 and Swe2 ranging from 50 bp to 1 kb. (A) Lengths of insertions (red) and deletions (blue) in Swe1 and Swe2 ranging from 1 kb to 10 kb. Table S1. SMRT-sequencing data overview. Table S2. Results of FALCON *de novo* assembly. Table S3. Overview of hybrid scaffolding of PacBio data using BioNano optical maps. Table S4. Alignment results of PacBio data to hg38. Table S5. Structural variation results for Swe1, Swe2 and HX1. Table S6. Statistics for NS in Swe1 and Swe2. Table S7. Repeat contents for NS is Swe1 and Swe2. Table S8. GC contents for NS is Swe1 and Swe2. Table S9. BLAST results for NS. Table S10. Overlap of NS between Swe1, Swe2 and HX1. Table S11. Amount of NSs that could be anchored to hg38.

Author Contributions: A.A. and U.G. conceived the study. H.C., M.M., P.O., I.B., J.D., S.H., F.V., J.N. and A.A. analyzed the data. I.H. and S.H. optimized and performed wet lab experiments. A.A., L.F. and U.G. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Funding: 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).

Acknowledgments: PacBio SMRT sequencing was performed by the National Genomics Infrastructure (NGI), which is hosted by SciLifeLab in Uppsala. Optical maps were generated at BioNano (CA, USA). Computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under projects b2015225 and sens-2016003. MM and PO were financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab.

Conflicts of Interest: The authors declare that they have no competing interests

References

1. Ameur, A.; Dahlberg, J.; Olason, P.; Vezzi, F.; Karlsson, R.; Martin, M.; Viklund, J.; Kahari, A.K.; Lundin, P.; Che, H.; et al. SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur. J. Hum. Genet.* **2017**, *25*, 1253–1260.
2. Boomsma, D.I.; Wijmenga, C.; Slagboom, E.P.; Swertz, M.A.; Karssen, L.C.; Abdellaoui, A.; Ye, K.; Guryev, V.; Vermaat, M.; van Dijk, F.; et al. The Genome of the Netherlands: Design, and project goals. *Eur. J. Hum. Genet.* **2014**, *22*, 221–227.
3. Fakhro, K.A.; Staudt, M.R.; Ramstetter, M.D.; Robay, A.; Malek, J.A.; Badii, R.; Al-Marri, A.A.; Abi Khalil, C.; Al-Shakaki, A.; Chidiac, O.; et al. The Qatar genome: A population-specific tool for precision medicine in the Middle East. *Hum. Genome Var.* **2016**, *3*, 16016.

4. Gudbjartsson, D.F.; Helgason, H.; Gudjonsson, S.A.; Zink, F.; Oddson, A.; Gylfason, A.; Besenbacher, S.; Magnusson, G.; Halldorsson, B.V.; Hjartarson, E.; et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **2015**, *47*, 435–444.
5. Nakatsuka, N.; Moorjani, P.; Rai, N.; Sarkar, B.; Tandon, A.; Patterson, N.; Bhavani, G.S.; Girisha, K.M.; Mustak, M.S.; Srinivasan, S.; et al. The promise of discovering population-specific disease-associated genes in South Asia. *Nat. Genet.* **2017**, *49*, 1403.
6. Wong, L.P.; Ong, R.T.; Poh, W.T.; Liu, X.; Chen, P.; Li, R.; Lam, K.K.; Pillai, N.E.; Sim, K.S.; Xu, H.; et al. Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.* **2013**, *92*, 52–66.
7. Consortium, U.K.; Walter, K.; Min, J.L.; Huang, J.; Crooks, L.; Memari, Y.; McCarthy, S.; Perry, J.R.; Xu, C.; Futema, M.; et al. The UK10K project identifies rare variants in health and disease. *Nature* **2015**, *526*, 82–90.
8. Telenti, A.; Pierce, L.C.; Biggs, W.H.; di Iulio, J.; Wong, E.H.; Fabani, M.M.; Kirkness, E.F.; Moustafa, A.; Shah, N.; Xie, C.; et al. Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11901–11906.
9. Schneider, V.A.; Graves-Lindsay, T.; Howe, K.; Bouk, N.; Chen, H.C.; Kitts, P.A.; Murphy, T.D.; Pruitt, K.D.; Thibaud-Nissen, F.; Albracht, D.; et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **2017**, *27*, 849–864.
10. Maretty, L.; Jensen, J.M.; Petersen, B.; Sibbesen, J.A.; Liu, S.; Villesen, P.; Skov, L.; Belling, K.; Theil Have, C.; Izarzugaza, J.M.; et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **2017**, *548*, 87–91.
11. Ross, M.G.; Russ, C.; Costello, M.; Hollinger, A.; Lennon, N.J.; Hegarty, R.; Nusbaum, C.; Jaffe, D.B. Characterizing and measuring bias in sequence data. *Genome Biol.* **2013**, *14*, R51.
12. Ameur, A.; Kloosterman, W.P.; Hestand, M.S. Single-molecule sequencing: Towards clinical applications. *Trends Biotechnol.* **2018**. doi:10.1016/j.tibtech.2018.07.013.
13. Chaisson, M.J.; Huddleston, J.; Dennis, M.Y.; Sudmant, P.H.; Malig, M.; Hormozdiari, F.; Antonacci, F.; Surti, U.; Sandstrom, R.; Boitano, M.; et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **2015**, *517*, 608–611.
14. Shi, L.; Guo, Y.; Dong, C.; Huddleston, J.; Yang, H.; Han, X.; Fu, A.; Li, Q.; Li, N.; Gong, S.; et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **2016**, *7*, 12065.
15. Seo, J.S.; Rhie, A.; Kim, J.; Lee, S.; Sohn, M.H.; Kim, C.U.; Hastie, A.; Cao, H.; Yun, J.Y.; Kim, J.; et al. De novo assembly and phasing of a Korean human genome. *Nature* **2016**, *538*, 243–247.
16. Pendleton, M.; Sebra, R.; Pang, A.W.; Ummat, A.; Franzen, O.; Rausch, T.; Stutz, A.M.; Stedman, W.; Anantharaman, T.; Hastie, A.; et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **2015**, *12*, 780–786.
17. Mostovoy, Y.; Levy-Sakin, M.; Lam, J.; Lam, E.T.; Hastie, A.R.; Marks, P.; Lee, J.; Chu, C.; Lin, C.; Dzakula, Z.; et al. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods* **2016**, *13*, 587–590.
18. Wong, K.H.; Levy-Sakin, M.; Kwok, P.Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat. Commun.* **2018**, *9*, 3040.
19. Chin, C.S.; Alexander, D.H.; Marks, P.; Klammer, A.A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E.E.; et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **2013**, *10*, 563–569.
20. Zheng-Bradley, X.; Streeter, I.; Fairley, S.; Richardson, D.; Clarke, L.; Flicek, P. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience* **2017**, *6*, 1–8.
21. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. *Genome Biol.* **2004**, *5*, R12.
22. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; von Haeseler, A.; Schatz, M.C. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **2018**, *15*, 461–468.
23. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421.
24. Wang, K.; Li, M.; Hakonarson, H. Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164.
25. Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311.

26. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2015**, *44*, D733–D745.
27. Genomes Project, C.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74.
28. Chin, C.S.; Peluso, P.; Sedlazeck, F.J.; Nattestad, M.; Concepcion, G.T.; Clum, A.; Dunn, C.; O'Malley, R.; Figueroa-Balderas, R.; Morales-Cruz, A.; et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **2016**, *13*, 1050–1054.
29. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
30. Bennett, H.M.; Mok, H.P.; Gkrania-Klotsas, E.; Tsai, I.J.; Stanley, E.J.; Antoun, N.M.; Coghlan, A.; Harsha, B.; Traini, A.; Ribeiro, D.M.; et al. The genome of the sparganosis tapeworm *Spirometra erinaceieuropaei* isolated from the biopsy of a migrating brain lesion. *Genome Biol.* **2014**, *15*, 510.
31. Thorvaldsdottir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192.
32. Bickhart, D.M.; Rosen, B.D.; Koren, S.; Sayre, B.L.; Hastie, A.R.; Chan, S.; Lee, J.; Lam, E.T.; Liachko, I.; Sullivan, S.T.; et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **2017**, *49*, 643–650.
33. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338.
34. Redon, R.; Ishikawa, S.; Fitch, K.R.; Feuk, L.; Perry, G.H.; Andrews, T.D.; Fiegler, H.; Shapero, M.H.; Carson, A.R.; Chen, W.; et al. Global variation in copy number in the human genome. *Nature* **2006**, *444*, 444–454.
35. Yuan, S.; Johnston, H.R.; Zhang, G.; Li, Y.; Hu, Y.J.; Qin, Z.S. One Size Doesn't Fit All—RefEditor: Building Personalized Diploid Reference Genome to Improve Read Mapping and Genotype Calling in Next Generation Sequencing Studies. *PLoS Comput. Biol.* **2015**, *11*, e1004448.
36. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.