Sequencing of high-complexity DNA pools for identification of nucleotide and structural variants in regions associated with complex traits

European Journal of Human Genetics (2012) 20, 77–83

www.nature.com/ejhg

ARTICLE

Sequencing of high-complexity DNA pools for identification of nucleotide and structural variants in regions associated with complex traits

Ghazal Zaboli[1,13], Adam Ameur[1,13], Wilmar Igl[1], Åsa Johansson[1], Caroline Hayward[2], Veronique Vitart[2], Susan Campbell[2], Lina Zgaga[3,11], Ozren Polasek[3], Gerd Schmitz[4], Cornelia van Duijn[5], Ben Oostra[6], Peter Pramstaller[7,8,9], Andrew Hicks[7], Tomas Meitinger[10], Igor Rudan[11,12], Alan Wright[2], James F Wilson[12], Harry Campbell[12] and Ulf Gyllensten*[,1], for the EUROSPAN Consortium

We have used targeted genomic sequencing of high-complexity DNA pools based on long-range PCR and deep DNA sequencing by the SOLiD technology. The method was used for sequencing of 286 kb from four chromosomal regions with quantitative trait loci (QTL) influencing blood plasma lipid and uric acid levels in DNA pools of 500 individuals from each of five European populations. The method shows very good precision in estimating allele frequencies as compared with individual genotyping of SNPs (r^2=0.95, $P<10^{-16}$). Validation shows that the method is able to identify novel SNPs and estimate their frequency in high-complexity DNA pools. In our five populations, 17% of all SNPs and 61% of structural variants are not available in the public databases. A large fraction of the novel variants show a limited geographic distribution, with 62% of the novel SNPs and 59% of novel structural variants being detected in only one of the populations. The large number of population-specific novel SNPs underscores the need for comprehensive sequencing of local populations in order to identify the causal variants of human traits.

European Journal of Human Genetics (2012) **20,** 77–83; doi:10.1038/ejhg.2011.138; published online 3 August 2011

Keywords: pooling; next-generation DNA sequencing; SOLiD; SNP; indels

INTRODUCTION

Genome-wide association studies (GWAS) have identified a large number of SNPs associated with human diseases and quantitative traits (QTs).[1] At present information for about 19 million SNPs is available in dbSNP (as of version 130).[2] The 300 000 to 1 000 000 SNPs that have been typed in association studies explains only a small fraction of the phenotypic variation of the traits.[3] One possible reason that most of the phenotypic variation remains unexplained is that the SNPs on the arrays are biased towards common variants[4] and analysis does not capture the effect of rare variants. The targeted sequencing by the HapMap 3 Consortium,[5] large-scale exome sequencing projects, such as that of 200 exomes of individuals of Danish origin studied using low-coverage DNA sequencing,[6] and the data generated through the pilot phase of the 1000 Genomes Project[7] have all emphasized the large number of previously undetected rare variants present in human populations. Most studies have included a limited number of individuals per population, and have therefore identified mainly sequence variants with a medium to low frequency (5–10%), while identification of SNPs with a frequency below that requires even larger sample size per population.

Next-generation DNA sequencing (NGS) technologies provide an opportunity to sequence complete genomes, exomes or large genomic regions, but the cost is still prohibitive for sequencing of thousands of individual genomes from a single population.[8] An alternative is to perform targeted sequencing of the DNA of pools of individuals. Ingman *et al*[9] used long-range PCR (LR-PCR) and Roche Genome Sequencher (454; Roche Applied Sciences, Werk Penzberg, Germany) for SNP detection and allele frequency estimation in DNA pools of 96 individuals and Out *et al*[10] used long-range PCR and a population pool of 287 individuals using Solexa Genome Analyzer (Illumina Inc., San Diego, CA, USA). Druley *et al*[11] sequenced 13 237 bases in a pool of 1111 individuals and showed a good precision in the frequency estimation for 14 validated SNPs. A limitation in pools sequencing is that individual genotypes cannot be deduced. Bansal *et al*[12] used amplicons generated from single individuals and pooled amplicons from 48 individuals for sequencing on the Illumina GA. However, handling of amplicons from single individuals becomes impractical for sequencing of large number of samples from a population. Prabhu and Peer[13] proposed that individual genotypes in a pool could be identified using a series of overlapping pools to enable

[1]Department of Immunology, Genetics and Pathology, Rudbeck Laboratory, SciLifeLab Uppsala, Uppsala University, Uppsala, Sweden; [2]MRC, Human Genetics Unit, IGMM, Western General Hospital, Edinburgh, UK; [3]Andrija Stampar School of Public Health, Faculty of Medicine, University of Zagreb, Zagreb, Croatia; [4]Institute for Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Franz-Josef-Strauss-Allee, Regensburg, Germany; [5]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands; [6]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands; [7]Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), Bolzano, Italy; [8]Department of Neurology, General Central Hospital, Bolzano, Italy; [9]Department of Neurology, University of Lubeck, Lubeck, Germany; [10]Helmholtz Zentrum Munchen, Neuherberg, Munich, Germany; [11]Croatian Centre for Global Health, Faculty of Medicine, University of Split, Soltanska and Institute for Clinical Medical Research, University Hospital 'Sestre Milosrdnice', Vinogradska, Split, Croatia ; [12]Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh, UK

*Correspondence: Professor U Gyllensten, Department of Immunology, Genetics and Pathology, Rudbeck Laboratory, SciLifeLab Uppsala, Uppsala University, Uppsala SE-751 85, Sweden. Tel: +46 708 99 34 13; Fax: +46 18 471 49 31; E-mail: ulf.gyllensten@igp.uu.se

[13]These authors contributed equally to this work.

Received 6 January 2011; revised 7 June 2011; accepted 30 June 2011; published online 3 August 2011

identification of single alleles. Erlich *et al*[14] used barcoded pools for identification of the pool from which an allele was obtained and Shental *et al*[15] developed a method denoted compressed sensing (CS) that is based on using overlapping pools and suited for identification of rare alleles both in heterozygous or in homozygous state.

All present next-generation DNA sequencing technologies can be used for analysis of high-complexity DNA pools. The two-base encoding system SOLiD technology reduces the frequency of technical errors, potentially providing an advantage for sequencing of pools.[8] Here, we describe a method for detection of genetic variants in high-complexity DNA pools using the SOLiD technology. We used this method for identification of SNPs and structural variants (indels) in DNA pools of about 500 individuals from each of five populations. The regions sequenced were selected on the basis of associations previously found between SNPs in these regions and lipid (*LASS4*, *LIPC* and *ATP10D*)[16] or uric acid (*SLC2A9*) levels.[17]

METHODS

Populations and genomic regions

The cohorts studied stem from populations in Sweden, Italy, Scotland, Croatia and The Netherlands and are part of the European Special Population Research Network (EUROSPAN, http://www.eurospan.org). All participants gave their written informed consent.[18] The Northern Swedish Population Health Study (NSPHS) is a cross-sectional study conducted in the community of Karesuando north of the Arctic Circle in the Norrbotten County, Sweden.[19] The Orkney Complex Disease Study (ORCADES) is a longitudinal study in the Scottish archipelago of Orkney.[20] The VIS study is a cross-sectional study in the villages of Vis and Komiza on the Dalmatian island of Vis, Croatia.[21,22] The Micro-isolates in South Tyrol Study (MICROS) is a cross-sectional study carried out in the villages of Stelvio, Vallelunga and Martello, Venosta valley, South Tyrol, Italy.[23] The Erasmus Rucphen Family Study (ERF) is a longitudinal study of a population living in the Rucphen region, The Netherlands, since the 19th century.[24] We selected about 500 individuals with the lowest kinship coefficient from each population (450 from ERF; 480 from NPSHS and MICROS; 500 from ORCADES and VIS) and an equal amount of DNA from each individual was used to prepare five population pools with a final DNA concentration of 10 ng/μl. Naturally, the structure of the populations implies that the individuals included are not completely unrelated.

Four chromosomal regions (*LASS4*, *LIPC*, *ATP10D* and *SLC2A9*) were selected for population sequencing based on the associations previously found between SNPs in these regions and lipid levels (*LASS4*, *LIPC* and *ATP10D*)[16] or uric acid levels (*SLC2A9*).[17] The different regions are defined as follows, with coordinates in the hg18 genome assembly: (a) *LASS4* (LAG1 homolog ceramide synthase (4), chromosome 19, positions 8 179 257 to 8 234 302, 55 kb, (b) *LIPC* (hepatic lipase), chromosome 15, positions 56 450 000 to 56 550 000, 100 kb, (c) *ATP10D* (ATPase class V, type 10D), chromosome 4, positions 47 195 000 to 47 299 000, 104 kb and (d) *SLC2A9* (solute carrier family 2, member 9), chromosome 4, positions 9535 000 to 9655 000, 120 kb. The selected regions span the parts of the genome with SNPs showing the most significant associations in the previous GWAS.[16,17]

Enrichment and sequencing

The FastPCR software by Primer Digital Ltd (Helsinki, Finland) (2006–2009) was used for design of primers for long-range PCR (LR-PCR). Between 30 and 50 primer pairs were designed for each chromosomal region, with an average amplicon size of 2 kb and 50 to 100 nucleotides overlap. LR-PCR was carried out in the Veriti thermal cycler (ABI, ABI-Life Technologies, Carlsbad, CA, USA) using 50 ng DNA from the pool in a reaction volume of 100 μl, containing 5x HF buffer, 200 mM dNTPs, 12 μM of each primer (http://www.sigma.com/oligos) and 1 unit of Phusion high-fidelity polymerase (Finnzymes, Vantaa, Finland). A two step PCR was performed with an initial denaturation for 30 s at 98°C followed by 30 cycles of denaturation for 10 s at 98°C and extension for 90 s at 72°C and a final extension at 72°C for 10 min. The PCR products were then subjected to electrophoresis on 1.5% agarose gels (Roche Diagnostic GmbH, Mannheim, Germany) and visualized by ethidium bromide staining.

Amplicons were purified using the Qiagen Clean-up kit (QIAquick, QIAGEN Nordic, Sweden) for PCR products and the concentration was determined using NanoDrop (NanoDrop Technologies, Thermo Fisher Scientific, Wilmington, DE, USA). An equal copy number of each amplicon across all four chromosomal regions was used to prepare an amplicon pool for each population. The amplicon pool from a population was used to generate a fragment library and this was used to generate 50 bp sequence reads on the SOLiD 3 instrument by Applied Biosystems (http://www.appliedbiosystems.com). Sequence reads are deposited in the EBI Sequence Read Archive at (SRA) under the study accession number ERP000249.

PCR errors

Each amplicon was generated by PCR from 100 ng genomic DNA (equivalent to 30 000 copies of a single-copy gene or on average 30 copies of each homolog in the population sample) from the population pool. With a sequence coverage of 10^4 per base and reported per-base error rate of the Phusion DNA polymerase of 4.4×10^{-7} (see[25]), the expected error rate per nucleotide is 4.4×10^{-3}. In the sequencing library, preparation amplicons were amplified only for a few cycles (n=3). This could contribute to a bias with respect to the allelic products that were sequenced and we therefore developed an analysis strategy that takes into account the number of reads with different unique starting points in the calling of variants.

Computational strategy to identify SNPs and indels

The different steps of the analysis are described in the sections below.

1. Alignment of sequence reads
2. SNP analysis
   (a) Pre-processing and calculating UVAM scores
   (b) Calling significant SNPs
   (c) Estimating allele frequencies
3. Indel analysis
   (a) Split-read alignment and SplitSeek analysis
   (b) Estimating allele frequencies

Alignment of sequence reads. The SOLiD reads were aligned to a reference consisting of the DNA sequence in the *ATP10D*, *LASS4*, *LIPC* and *SLC2A9* regions with an additional 10 kb upstream and downstream of each region, using the SOLiD system analysis pipeline tool (corona lite) and allowing for up to four mismatches for each 50 bp read. Valid adjacent mismatches were counted as one mismatch instead of two. The reads that mapped to the target regions were processed for SNP detection, while the unmapped reads were analyzed for structural variants (indels).

SNP analysis. The SNP calling strategy can be divided into the two main steps outlined below. To reduce the effect of variation in coverage between different regions and populations, an independent analysis was performed for each region in each of the cohorts. The program code for the SNP analysis is available upon request.

Pre-processing and calculating UVAM scores The General Feature Format (GFF) Conversion Tool available from the SOLiD software development community (http://solidsoftwaretools.com) was used to extract the nucleotide sequence and mismatches for the aligned reads. To avoid pileups of primer sequence reads at the ends of our amplicons, we filtered out all reads mapping to the exact same position and with the same orientation as one of the primers used in the LR-PCR. Custom programs were implemented to extract the coverage and valid adjacent mismatches for each position in the sequenced regions. We also counted the number of uniquely placed reads containing a valid adjacent mismatch for every position, a value that we labeled 'unique valid adjacent mismatches' (UVAM). UVAM is a robust measurement for SNP detection, as it is unaffected by pileups of identical reads and insensitive to variations in coverage. For reads with a length 50 bp the UVAM values range between 0 and 100, where zero means that no valid adjacent mismatches were detected at the position. A maximum value of 100 means that there were 50 reads with unique starting points on each of the two strands, which contain a

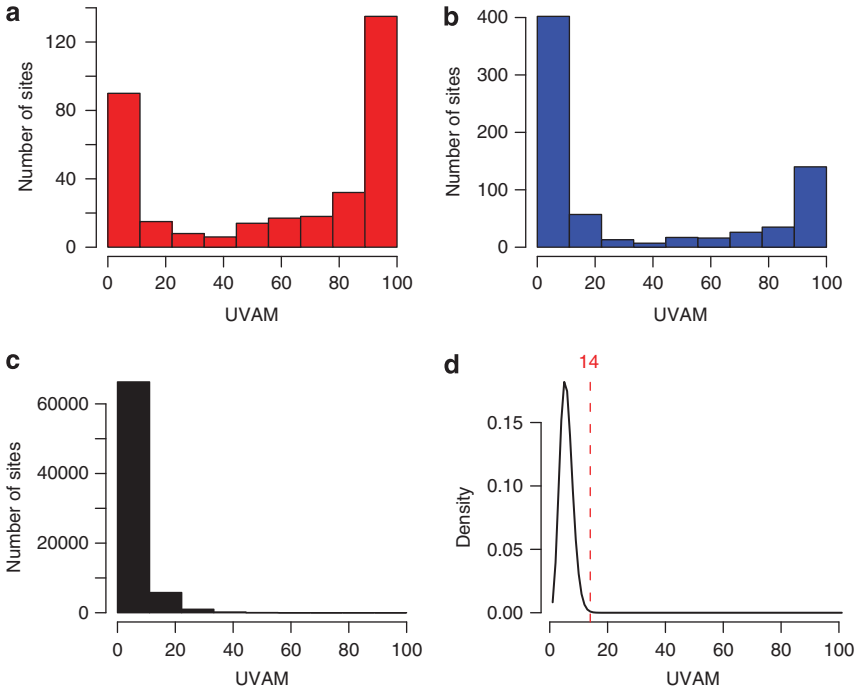

Figure 1 UVAM distributions and cut-off for SNP detection for the ATP10D region in the ERF cohort. The UVAM score is a value representing the number of uniquely placed reads that contain a SNP. The three first panels show the UVAM scores for (**a**) all positions where a SNP has been reported in dbSNP; (**b**) all positions where a SNP has been reported in the 1000 Genomes Project data; (**c**) all other positions, that is, where no SNPs are reported in dbSNP or in the 1000 Genomes Project; (**d**) estimated background binomial distribution for UVAM scores. The binomial distribution is used as a model for the UVAM scores at positions where there are no SNPs. There is a high similarity between the theoretical distribution in (**d**) and the empirical results in (**c**). The dotted red line shows a cut-off level, calculated from the binomial distribution, corresponding to a FDR of 0.01.

valid adjacent mismatch at the given position. As SNPs are reported as valid adjacent mismatches in the SOLiD system, true SNPs sequenced with high coverage will be associated with high UVAM values.

Calling significant SNPs We utilized information about the localization of known SNPs to determine a UVAM cut-off value in the sequenced regions. First, we extracted the previously computed UVAM scores at all positions reported in dbSNP and in the 1000 Genomes Project. The UVAM distribution at known SNP positions in dbSNP and 1000 Genomes Project is shown for the *ATP10D* region in the ERF cohort in Figures 1a and b, respectively. The remaining bases in the sequenced region have a distribution of UVAM scores that looks fundamentally different (Figure 1c), as the vast majority of such sites are not polymorphic. This implies that the values at those positions largely reflect background noise, something we can take advantage of in order to construct a model for the UVAM distribution at non-SNP sites. Assuming that s is the total sum of UVAM scores over n non-SNP sites, we modeled the probability of having a UVAM score of at least k by a binomial approximation of the hypergeometric distribution.

$$\text{prob}(\text{UVAMscore} \geq k) = \sum_{i=k}^{s} \binom{s}{i} (1/n)^i (1-1/n)^{s-i}$$

The formula above thus models the background distribution of UVAM scores. This theoretical distribution of UVAM scores for the *ATPD10* region is seen in Figure 1d and is in very good agreement with the empirical values. On the basis of the *P*-values we selected a cut-off corresponding to a false discovery rate (FDR) of 0.01. In the example in Figure 1d, this implied that SNPs at positions with a UVAM score of at least 14 were called as significant. To increase the confidence in the SNPs that were detected as novel at the 0.01 FDR level, three additional criteria were enforced for detection of novel SNPs; (i) at least 95% of the reads must contain the same alternative nucleotide, (ii) a higher number of valid adjacent mismatches than all other types of mismatches combined, (iii) at least 20% of the reads from each of the two strands, and (iv) a total coverage of at least 1000 reads.

Estimating allele frequencies From the initial mapping, we extracted the average coverage of each predicted indel, r_c, a value that corresponds to the number of full-length (50 bp) reads that were aligned at the position. If the number of reads that instead support the predicted insertion or deletion are denoted by the variable r_i, we then estimated the allele frequency for each indel by the ratio $r_i/(r_i+r_c)$.

Indel analysis. Split-read alignment and SplitSeek analysis All reads that could not be aligned to the sequenced regions in the alignment step above (step 1) were used to search for small insertions and deletions by applying the *SplitSeek* method, which has previously been used to find splice junctions and indels in RNA-seq data.[26] We first aligned the previously unmapped reads using the SOLiD split-read alignment program (http://solidsoftwaretools.com), with the same settings as in the RNA-seq study, and the alignment results were used as input to *SplitSeek*.[26] We required each indel to be supported by at least 10 reads from both strands and the maximum length for deletions was set to 500 bp.

Estimating allele frequencies From the initial mapping, we could extract the average coverage of each predicted indel, r_c, which corresponded to the number of reads that map to the reference sequence. If we denoted the number of reads that instead support the predicted insertion or deletion by r_i, we could then estimate the allele frequency for each indel by the ratio $r_i/(r_i + r_c)$.

Annotation of SNPs and indels

Version 130 of dbSNP was used for SNP analysis.[2] SNP data from the 1000 Genomes Project[27] were downloaded from their site (http://www.1000genomes.org) using SNP data files with release date 2009_04.

Experimental validation of novel SNPs

To independently verify the presence of novel, low frequency, SNP, we selected a set of novel SNP and genotyped these in the 480 individual DNA samples from the NSPHS cohort that were used to generate the DNA pools, using TaqMan

genotyping assays (Applied Biosystems) and the 7900 HT Fast Real Time PCR system (Applied Biosystems), according to the manufacturer's instructions.

RESULTS

Analysis overview

We sequenced a total of 286 470 bases (*ATP10D* 74 207 bp, *LASS4* 34 559 bp, *LIPC* 92 226 bp, *SLC2A9* 85 478 bp) out of the about 397 kb of the four regions. The parts not covered represent repeated regions that proved difficult to amplify with long-range PCR. Approximately 300 million reads for the five populations mapped uniquely to the four chromosomal regions (Supplementary Table S1). To avoid biases due to oversequencing of reads containing primer sequences, we removed all reads matching to primer sequences, leaving 263 millions reads for further SNP analysis. To identify SNPs in the DNA pools, we developed a computational procedure for SNP and indel identification (see Methods for details). For each position, we computed the number of uniquely placed reads that contain a SNP, denoted unique valid adjacent mismatch (UVAM) scores, and these were used for SNP calling procedure (Figure 1). Indels were detected through split read alignment and analysis using the SplitSeek method.[26] As an example the *ATPD10* region in the NSPHS cohort when viewed through the UCSC genome browser[28] is shown in Supplementary Figure S1. In total, 99% of the region under the amplicons has a sequence coverage of 100 fold, 95% a sequence coverage of 1000 fold and 44% a sequence coverage of 10 000 fold (Figure 2). This corresponds to a one-fold coverage per chromosome in the pool for 95% of the sequenced positions, and a 10-fold coverage per chromosome for 44% of the sites.

SNP detection and accuracy in allele frequency estimation

To determine our ability to detect SNPs and accurately estimate the allele frequency, we used genotypes from the Illumina Infinium HumanHap300v2 array.[16] Out of 49, 48 SNPs located in the sequenced regions were found in at least one of the five cohorts, corresponding to an analytical sensitivity of 98%. The single SNP (rs11666866) that could not be detected is located near the end of a PCR primer sequence, where it was not possible to obtain a sufficient number of uniquely mapping reads. Of 505 imputed SNPs in the sequenced regions, only 13 SNPs (3%) were not found in any of the five populations. These SNPs are either located at the end of the primer sequences or in the regions with extremely low coverage.

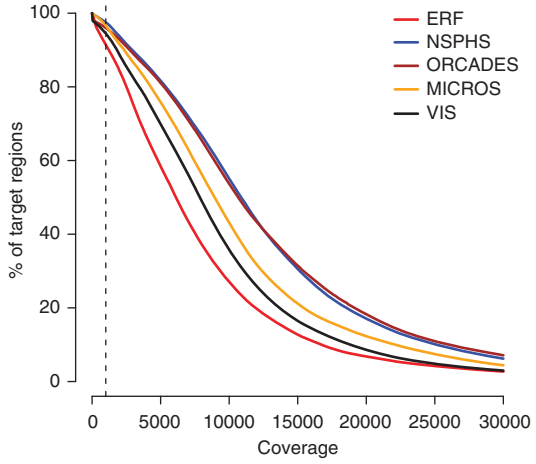

Figure 2 Sequence coverage per base for the four sequenced chromosomal regions. The colored lines show the percentage of bases that exceed the coverage level on the *x* axis. One separate line is drawn for each of the population cohorts.

The correlation between the allele frequencies estimated from the DNA pools and from individual genotypes for the 48 SNPs on the Illumina array is very high (r^2=0.95) ($P<10^{-16}$) (Figure 3a). There is also a good correlation between the frequency of imputed SNPs and those sequenced (r^2=0.91) (Figure 3b). The exception are two SNPs (rs11736479 and rs7696092), which appear as outliers in all populations, and both have very low frequencies in HapMap CEU (MAF=0 and MAF=0.01) and the rsq value from MACH is <0.30 for all populations. Rsq is a quality measure by MACH, illustrating the squared correlation between imputed and true genotypes. Typically, a cut-off of 0.30 will flag most of the poorly imputed SNPs. To examine the effect of sequence coverage on the precision of the allele frequency estimate, we studied SNPs with a coverage of >10 000 fold (n=1059) (Figure 3c) and <2000 fold (n=186) (Figure 3d). The correlation even at lower sequence coverage is still very high (r^2=0.89). The results show that method can be used to estimate the frequency of SNPs in a pool of DNA samples and is useful at different sequence coverage levels.

Experimental validation of novel SNPs

A set of novel SNPs detected in the NSPHS cohort were genotyped using TaqMan genotyping assays on the individual DNA samples. We selected the SNPs uniformly across the allele frequency spectrum, with predicted MAF from about 8% to below 1%. All 12 SNPs genotyped were identified (see Supplementary Table S2), and the correlation between the allele frequency estimated from the DNA pools and from individual genotyping is very high (r^2=0.95) (Figure 3e). These results shows that our method can identify novel SNPs with high sensitivity for novel/rare SNPs and predict accurate allele frequencies for rare SNPs with MAF at 5% or below.

SNP and indel discovery

We identified 1884 SNPs in the five population cohorts and four chromosomal regions (Table 1). Of the SNPs, 17% (318/1884) are not present in dbSNP version 130 or in the 1000 Genomes Project data (release date 2009_04). Among structural variants 61% (63/103) are not present in dbSNP (Table 1). In all, 81% (257/318) of novel SNPs occur at a frequency of 5% or less (Supplementary Figure S2a). Of the SNPs with a frequency of up to 5% and identified in all five populations, 37% (257/697) are not available in any database (Table 2). As many as 81% of the indels with a frequency up to 5% are not available in the public databases (Supplementary Figure S2b). Some deletions have lengths up to 100 bp or more but the majority of them are small (1–5 bp) (Supplementary Figure S3). A similar analysis for insertions is presently not possible due to technical reasons. The complete lists of all SNPs and indels detected in each of the five populations are available as supplementary data files.

Comparisons between populations

The majority of SNPs are found in more than one of the cohorts. Of all SNPs in the five cohorts, 63% are common to all cohorts and 19% are found in only one cohort (Figure 4a). Of the novel SNPs, only 6% is present in all cohorts and about 62% is only found in one (Figure 4b). In total, 93–136 novel SNPs are found per cohort (Table 1). In all, 46% (323/697) of SNPs with frequency of 5% or less and 67% (177/263) with a frequency of 1% or less are found in only one of the five cohorts (Table 2). Of all indels, 38% are common to all cohorts (Figure 4c), while among the novel indels 21% are common to all cohorts (Figure 4d). Of all indels, 40% are found in

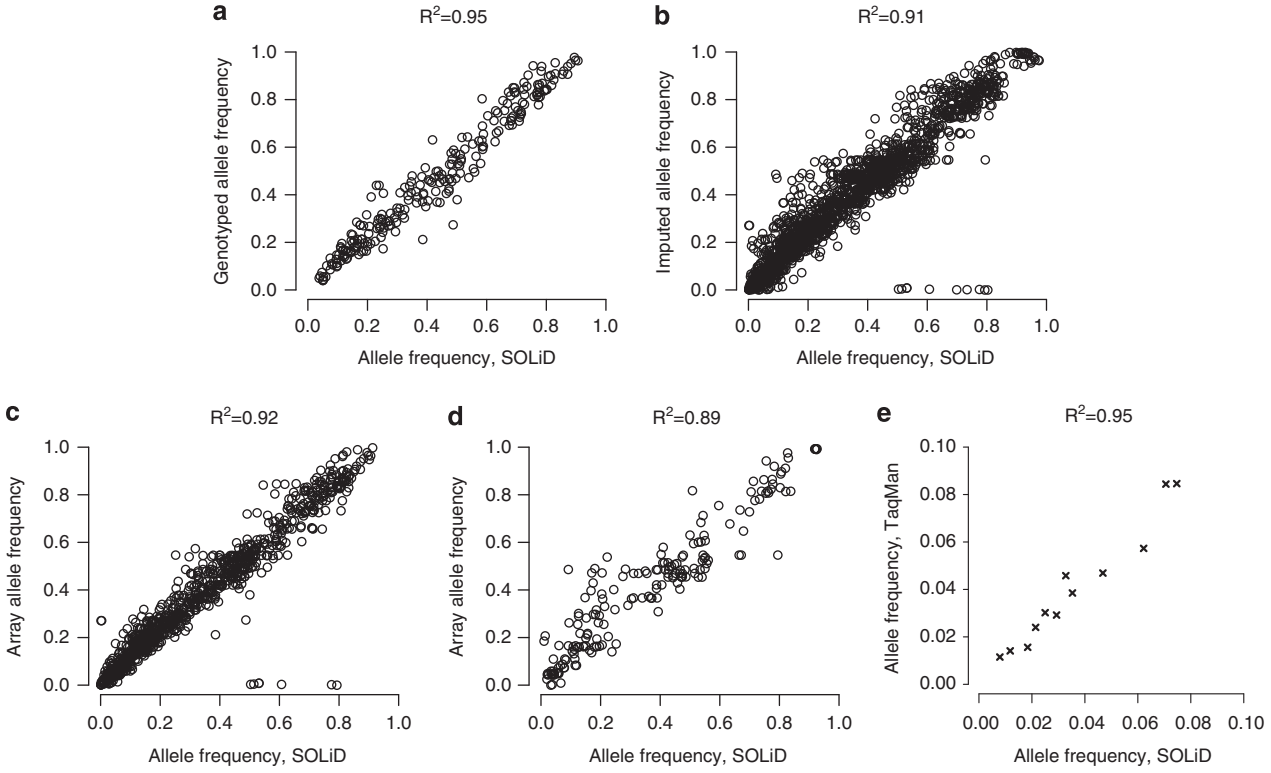

Figure 3 Correlation between the allele frequencies estimated by sequencing of pools (*x* axes) and from individual genotyping (*y* axes). Correlation for (**a**) SNPs genotyped on the array (n=240); (**b**) SNPs where the array frequencies were inferred by imputation (n=2217); (**c**) SNPs with at least 10 000 fold coverage (n=1059) and where the array frequencies were inferred by imputation; and (**d**) SNPs with at most 2000-fold coverage (n=186) and where the array frequencies were inferred by imputation; (**e**) comparison of the allele frequencies for of a set of novel low frequency SNPs detected by the method and by individual TaqMan genotyping in 480 individuals from the NSPHS cohort.

Table 1 Summary of the number of SNPs and indels found in the sequencing of DNA pools from the five EUROSPAN populations

	Number of detected variants						*Number of variants in one population only*					
	ERF	*MIC*	*NSP*	*ORC*	*VIS*	*Total*	*ERF*	*MIC*	*NSP*	*ORC*	*VIS*	*Total*
SNPs												
Total	1431	1488	1504	1439	1426	1884	73	62	118	51	57	361
In dbSNP 130	1121	1153	1159	1130	1132	1217	4	6	18	5	11	44
In 1000 genomes	1219	1269	1310	1256	1252	1459	17	12	51	22	19	121
Novel	128	136	114	97	93	318	52	44	49	24	27	196
Indels												
Total	53	58	69	63	57	103	4	10	12	8	7	41
In dbSNP 130	32	32	38	31	32	40	0	0	3	0	1	4
Novel	21	26	31	32	25	63	4	10	9	8	6	37

Abbreviations for population cohorts: MIC, MICROS; NSP, NSPHS; ORC, ORCADES.

only one of the cohorts. When only considering novel indels, the corresponding percentage is 59%.

DISCUSSION

We have shown that long-range PCR and deep DNA sequencing can be used to identify SNPs and structural variation in high-complexity DNA pools. Pooling of DNA samples substantially increases the efficiency of sequencing studies[29] and has previously been used to sequence selected regions in pools of DNA samples or pools of amplicons.[9,11,12,30]

Using an analysis strategy that require a SNP or indel to be found in several reads with different alignment positions, in combination with using the SOLiD two-base encoding, we show that low frequency variants can be identified with high confidence. The allele frequencies estimated from the DNA pools show very good correlation with frequencies based on individual genotypes, as well as imputed SNPs. Validation of candidate novel SNPs also supported our procedure for SNP identification. For detection of indels, we used a method initially developed to study splice variation in RNA-seq data.[26] As only one algorithm was used for indel detection and the predicted indels were

not experimentally validated, we cannot exclude that some of them may represent false positives. However, the algorithm used is only able to detect indels of a certain size, thus inherently underestimating the total number. Also, we used stringent criteria in the variant calling (ie, number of reads supporting an indel), which will reduce the number of false positive calls. Another factor that affects the number of novel SNPs and indels detected is sequence coverage. The average sequence coverage ranges from 6000 fold to 13 000 fold with 95% of the bases having a coverage of 1000 fold in a DNA pool of 1000 chromosomes. This means that we are likely not to have identified all genetic variants present, and deeper sequencing would have resulted in additional variants. This is supported by our results, where we find that the average coverage for rare variants (MAF $<1\%$) is about 16 000×, while it is lower (around 10 000×) among the other SNPs.

We studied populations from different parts of Europe, encompassing much of the genetic diversity present on the European continent. In the sequencing of 100 kb in 692 samples from 10 global populations, the HapMap 3 Consortium found that 77% of SNPs are not in dbSNP (build 129) and of these 99% had a MAF $<5\%$.[5] By comparison, 26% of all SNPs in our cohorts are not in dbSNP and of novel SNPs, 81% had a MAF $<5\%$ (Tables 1 and 2). The low-coverage sequencing in the 1000 Genomes project similarly reported an overall frequency of novel SNPs of 54% and novel indels of 57%.[7] The novel low frequency variants tend to be found only in a limited set of populations and 62% of novel SNPs and 59% of structural variants are found in only one of our five cohorts. Among SNPs with a MAF $<0.5\%$, 88% were found in only one of our cohorts, as compared to 37% in the HapMap 3 study.[5] In the 1000 Genomes Project pilot study, 25% of SNPs already reported in dbSNP (vers129) were detected in only one of the population panels, while as many as 84% of novel SNPs was found in a single panel.[7] The difference seen between studies in the frequency of variants that is found only in a single population/panel may reflect both the sample size and extent of geographic region covered. The 1000 Genomes Project and HapMap 3 Consortium panels include sampling from the major human populations groups (ie, African, European and Asian populations), but with a more limited sample size per population (about 50–100 individuals), while our five populations have been sampled at a much higher depth.

Table 2 Detection of SNPs with different frequencies in the total material

Allele frequency (%)	*Total SNPs*	*In dbSNP*	*In 1000 genomes, not in dbSNP*	*Novel (% of total)*	*Present only in one population (% of total)*
≤0.1	10	1	9	0 (0)	8 (80)
≤0.5	140	32	92	16 (11)	103 (74)
≤1	263	56	144	63 (24)	177 (67)
≤2	434	91	195	148 (34)	242 (56)
≤3	539	112	227	200 (37)	281 (52)
≤5	697	166	274	257 (37)	323 (46)
≤10	929	325	310	294 (32)	350 (38)
≤20	1136	493	336	307 (27)	357 (31)
>20	748	724	13	11 (1)	4 (1)
Total	1884	1217	349	318 (17)	361 (19)

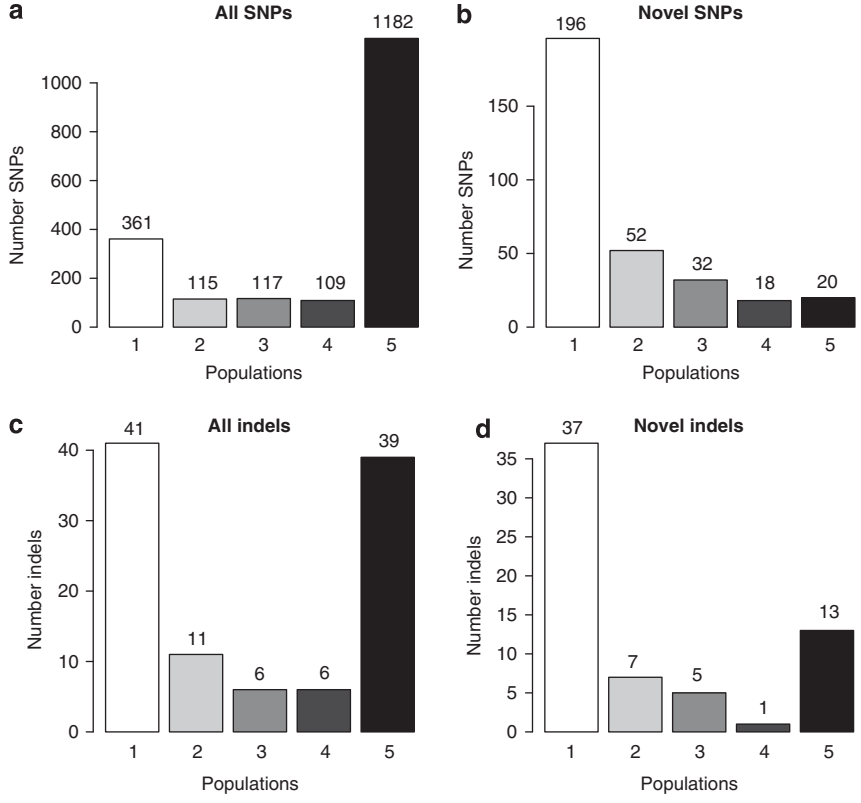

Figure 4 Number of overlapping SNPs and indels between the five populations. The five bars in each of the panels indicate the number of populations, in which the variants are found. The panels show the results for (**a**) all detected SNPs; (**b**) novel SNPs, that is, those not in dbSNP or the 1000 Genomes Project; (**c**) all detected in-dels; (**d**) novel indels, that is, those not in dbSNP.

The observation that over 50% of the novel SNPs in our study show a very restricted geographic distribution attests to the large amount of genetic variation present in local populations within the European continent.

Extrapolating from our results to a genome-wide scale this corresponds to about 3.7 million novel SNPs (between 1.1–1.6 million per population) and 0.7 million indels (between 0.2–0.4 million indels per population). The 1000 Genomes Project low coverage sequencing identified about eight novel million SNPs in all three population panels, with 1.7–5.1 novel million SNPs per panel and 0.75 million novel indels, with 0.3–0.5 million per panel. The 1000 Genomes Project cover about 86% of the genome, including both genic and intergenic regions and span major human population groups. It is therefore surprising that deep sequencing of populations from a much more limited geographic area can yield similar high numbers of genetic variants. Given the large number of novel low frequency variants identified, many of which show a limited distribution, future association studies based on cohort sequencing will have to consider careful population matching of cases and controls. Until the sequencing technology becomes affordable to perform sequencing of thousands of individuals per population, DNA pools remains a practical alternative.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The SOLiD DNA sequencing and Taqman genotyping was performed by the Uppsala Genome Center, funded by the Knut and Alice Wallenberg Foundation (CMS), The Swedish Natural Sciences Research Council (SNISS) and Science for Life Laboratory, Uppsala. This work was supported by the following grants and agencies: Swedish Medical Sciences Research Council, the Foundation for Strategic Research (SSF), the Linneaus Centre for Bioinformatics (LCB), European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947), The Netherlands Organisation for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043), European Commission FP7 grant LipidomicNet (2007-202272), NWO, ErasmusMC and the Centre for Medical Systems Biology (CMSB), the Ministry of Health and Department of Educational Assistance, University and Research of the Autonomous Province of Bolzano, and the South Tyrolean Sparkasse Foundation, the Scottish Executive Health Department and the Royal Society, the Medical Research Council UK, Ministry of Science, Education, and Sport of the Republic of Croatia (number 108-1080315-0302), Deutsche Forschungsgemeinschaft, the German Federal Ministry of Education and Research in the context of the German National Genome Research Network and Cardiogenics (EU-funded integrated project LSHM-CT- 2006-037593), the GSF-National Research Centre for Environment and Health funded by the German Federal Ministry of Education and Research and of the State of Bavaria.

1 Plomin R, Haworth CM, Davis OS: Common disorders are quantitative traits. *Nat Rev* 2009; **10**: 872–878.
2 Sherry ST, Ward MH, Kholodov M *et al*: dbSNP: the NCBI database of genetic variation. *Nucleic acids Res* 2001; **29**: 308–311.
3 Lango Allen H, Estrada K, Lettre G *et al*: Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; **467**: 832–838.
4 Frazer KA, Murray SS, Schork NJ, Topol EJ: Human genetic variation and its contribution to complex traits. *Nature Rev* 2009; **10**: 241–251.
5 Altshuler DM, Gibbs RA, Peltonen L *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
6 Li Y, Vinckenbosch N, Tian G *et al*: Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 2010; **42**: 969–972.
7 Durbin RM, Abecasis GR, Altshuler DL *et al*: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
8 Voelkerding KV, Dames SA, Durtschi JD: Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 2009; **55**: 641–658.
9 Ingman M, Gyllensten U: SNP frequency estimation using massively parallel sequencing of pooled DNA. *Eur J Hum Genet* 2009; **17**: 383–386.
10 Out AA, van Minderhout IJ, Goeman JJ *et al*: Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mut* 2009; **30**: 1703–1712.
11 Druley TE, Vallania FL, Wegner DJ *et al*: Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 2009; **6**: 263–265.
12 Bansal V, Harismendy O, Tewhey R *et al*: Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res* 2010; **20**: 537–545.
13 Prabhu S, Pe'er I: Overlapping pools for high-throughput targeted resequencing. *Genome Res* 2009; **19**: 1254–1261.
14 Erlich Y, Chang K, Gordon A *et al*: DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res* 2009; **19**: 1243–1253.
15 Shental N, Amir A, Zuk O: Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Res* 2010.
16 Hicks AA, Pramstaller PP, Johansson A *et al*: Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genetics* 2009; **5**: e1000672.
17 Vitart V, Rudan I, Hayward C *et al*: SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* 2008; **40**: 437–442.
18 Mascalzoni D, Janssens AC, Stewart A *et al*: Comparison of participant information and informed consent forms of five European studies in genetic isolated populations. *Eur J Hum Genet* 2010; **18**: 296–302.
19 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.
20 McQuillan R, Leutenegger AL, Abdel-Rahman R *et al*: Runs of homozygosity in European populations. *Am J Hum Genet* 2008; **83**: 359–372.
21 Rudan I, Campbell H, Rudan P: Genetic epidemiological studies of eastern Adriatic Island isolates, Croatia: objective and strategies. *Collegium Antropologicum* 1999; **23**: 531–546.
22 Vitart V, Biloglav Z, Hayward C *et al*: 3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia. *Eur J Hum Genet* 2006; **14**: 478–487.
23 Pattaro C, Marroni F, Riegler A *et al*: The genetic study of three population microisolates in South Tyrol (MICROS): study design and epidemiological perspectives. *BMC Med Genet* 2007; **8**: 29.
24 Aulchenko YS, Heutink P, Mackay I *et al*: Linkage disequilibrium in young genetically isolated Dutch population. *Eur J Hum Genet* 2004; **12**: 527–534.
25 Frey B: SB. *Biochemica* 1995; **2**: 34–35.
26 Ameur A, Wetterbom A, Feuk L, Gyllensten U: Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol* 2010; **11**: R34.
27 Kaiser J: DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science (New York, NY)* 2008; **319**: 395.
28 Kent WJ, Sugnet CW, Furey TS *et al*: The human genome browser at UCSC. *Genome Res* 2002; **12**: 996–1006.
29 Sham P, Bader JS, Craig I, O'Donovan M, Owen M: DNA Pooling: a tool for large-scale association studies. *Nat Rev* 2002; **3**: 862–871.
30 Craig DW, Pearson JV, Szelinger S *et al*: Identification of genetic variants using barcoded multiplexed sequencing. *Nat Methods* 2008; **5**: 887–893.

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