Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

# An overview of autosomal STRs and identity SNPs in a Norwegian population using massively parallel sequencing

Maria Martin Agudo [a,b], Håvard Aanes [a], Michel Albert [a], Kirstin Janssen [c], Peter Gill [a,b], Øyvind Bleka [a,*]

[a] Department of Forensic Sciences, Oslo University Hospital, Oslo, Norway
[b] Department of Forensic Medicine, Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[c] Centre for Forensic Genetics, UiT The Arctic University of Norway, Norway

ARTICLE INFO

*Keywords:*
Norwegian population
Massively parallel sequencing
Allele frequencies
Population genetics
Autosomal STRs
Identity SNPs

ABSTRACT

In recent years, probabilistic genotyping software has been adapted for the analysis of massively parallel sequencing (MPS) forensic data. Likelihood ratios (LR) are based on allele frequencies selected from populations of interest. This study provides an outline of sequence-based (SB) allele frequencies for autosomal short tandem repeats (aSTRs) and identity single nucleotide polymorphisms (iSNPs) in 371 individuals from Southern Norway. 27 aSTRs and 94 iSNPs were previously analysed with the ForenSeq™ DNA Signature Prep Kit (Verogen). The number of alleles with frequencies less than 0.05 for sequenced-based alleles was 4.6 times higher than for length-based alleles. Consistent with previous studies, it was observed that sequence-based data (both with and without flanks) exhibited higher allele diversity compared to length-based (LB) data; random match probabilities were lower for SB alleles confirming their advantage to discriminate between individuals. Two alleles in markers D22S1045 and Penta D were observed with SNPs in the 3′flanking region, which have not been reported before. Also, a novel SNP with a minor allele frequency (MAF) of 0.001, was found in marker TH01. The impact of the sample size on minor allele frequency (MAF) values was studied in 88 iSNPs from Southern Norway ($n = 371$). The findings were then compared to a larger Norwegian population dataset ($n = 15,769$). The results showed that the smaller Southern Norway dataset provided similar results, and it was a representative sample. Population structure was analyzed for regions within Southern Norway; $F_{ST}$ estimates for aSTR and iSNPs did not indicate any genetic structure. Finally, we investigated the genetic differences between Southern Norway and two other populations: Northern Norway and Denmark. Allele frequencies between these populations were compared, and we found no significant frequency differences (p-values > 0.0001). We also calculated the pairwise $F_{ST}$ values per marker and comparisons between Southern and Northern Norway showed small differences. In contrast, the comparisons between Southern Norway and Denmark showed higher $F_{ST}$ values for some markers, possibly driven by distinct alleles that were present in only one of the populations. In summary, we propose that allele frequencies from each population considered in this study could be used interchangeably to calculate genotype probabilities.

## 1. Introduction

In the field of forensic genetics, the advancement of massively parallel sequencing (MPS) has made it possible to analyze larger marker panels, that includes autosomal short tandem repeats (aSTRs), Y-STRs and X-STRs, and single nucleotide polymorphisms (SNPs) [1,2], along with access to complete sequence information. This has resulted in gain of information [1,3–6].

The future of forensic analysis might be shaped by the routine use of MPS technology in casework [7]. However, to achieve this, it is crucial to create population-specific allele frequency databases, which are necessary to compute likelihood ratio (LR) values [8]. Probabilistic genotyping software is used to evaluate the weight of evidence if a person of interest is a contributor to an evidence trace profile [9,10]. MPS technology has expanded data formats for STR genotypic information beyond the traditional capillary electrophoresis (CE) designation

[*] Corresponding author.
*E-mail address:* oeyble@ous-hf.no (Ø. Bleka).

https://doi.org/10.1016/j.fsigen.2024.103057
Received 15 September 2023; Received in revised form 27 February 2024; Accepted 28 April 2024
Available online 3 May 2024

of the alleles (length-based), allowing full sequences of nucleotides to be designated from the repeat region with or without flanking regions. Statistical models built for CE data have proven to be useful for MPS data, offering increased resolution [11,12]. Recently, inspired by other probabilistic genotyping models [13,14], EuroForMix [15] was extended to incorporate the full sequence information from MPS and to better account for stutter artefacts [16,17]. To calculate LR values based on these models, it is essential to incorporate a frequency database based on MPS data, with consistent genotypic data formatting [18]. Allele frequencies will vary according to use of different data formats, which could impact the LR values calculated. The purpose of the study is to characterize allele frequencies for 27 aSTR loci from the Verogen ForenSeq™ DNA Signature Prep Kit (Verogen, San Diego, USA) for the Southern Norwegian population, and to evaluate the discriminatory advantage of alternative sequence-based (SB) formats over the traditional length-based (LB) nomenclature [4,6,18–20]. Additionally, we investigate genetic differences between Southern Norway, Northern Norway and Denmark, addressing the possibility to utilize each other's databases. Denmark is a neighbouring country to Norway making it interesting to investigate potential genetic differences. Finally, the effect of the sample size on minor allele frequencies (MAF) was evaluated for 88 out of the 94 iSNPs included in the ForenSeq™ DNA Signature Prep Kit.

## 2. Materials and methods

### 2.1. Ethical declaration

This study was approved by the Data Protection Officer (DPO) at Oslo University Hospital with case numbers 20/16593 and 20/16592.

### 2.2. Population datasets

#### 2.2.1. Southern Norway

The Southern Norway population dataset was generated using the ForenSeq™ DNA Signature Prep Kit (Verogen, San Diego, USA) and initially consisted of sequences from 385 individuals [17,21], originating from different regions; East, West, Middle, and South Norway, as well as the cities of Oslo and Bergen [21] (see Supplementary Fig. S1).

Only 371 samples contained sufficient DNA extract to perform CE analysis and subsequently the genotype concordance analysis between CE and MPS was based on those. Suitable samples were quantified with PowerQuant®System (Promega) and amplified, following the manufacturer's protocol, with PowerPlex®Fusion 6 C (Promega). Fragments were detected with the Applied Biosystems™ 3500 Series Genetic Analyzer (Thermo Fisher Scientific). Genotypic data was retrieved using GeneMapper™ ID-X Version 1.6 (Thermo Fisher Scientific). The MPS dataset included aSTRs and identity SNPs (iSNP) genotype information and it was obtained from the reports created by the ForenSeq™ Universal analysis software (UAS) version 2.0 [22]. Final genotype tables were created with R (version 4.2.2).

To obtain the genotype information, sequences from the UAS reports were classified into alleles based on minimum read counts and heterozygote balance (Hb) value thresholds. For aSTRs, minimum read counts were set to 30; if two alleles were present, a minimum heterozygote balance (Hb) threshold of 0.3 was applied. For iSNPs, individual genotype data were extracted using UAS default analytical and interpretation thresholds of, respectively, 1.5% and 4.5% of the total read counts in the locus, were used to report the genotypes [21]. This corresponds to 11 and 30 reads when low total read counts in the locus (maximum of 650 counts) are observed. The repeat region and the 5′and 3′flanking regions were defined following the description from Gettings et al. [18] and STRseq [23]. The final genotype tables included full sequences of aSTR and additional processing was required to obtain a suitable sequence format of the repeat region and to separate it from the 5′and 3′flanking regions. LUSstrR (https://github.com/oyvble/LUSstrR), an R implementation of the tool lusSTR (https://github.com/bioforensics/lusSTR) [24], was used to obtain three different data formats (levels) employed in this study. These levels were used to compare sequence vs length-based alleles, each containing different degrees of information ranging from more to less informative:

1) *Sequence-based (SB) with flanks*: allele sequences with 5flank and 3flank (variation determined by the repeat region and e.g., SNPs in flanks).

2) *Sequence-based without flanks*: allele sequences without 5flank and 3flank (variation determined only by the repeat region).

3) *Length-based (LB)*: alleles designations based on the number of repeat units obtained from the UAS reports.

#### 2.2.2. Additional populations

To explore potential genetic differences of Southern Norway with other populations, two additional datasets were incorporated in our study. Both datasets were selected because they are geographically close to Southern Norway and were generated using the Verogen ForenSeq™ DNA Signature Prep Kit. These are as follows:

1. The dataset from **Northern Norway** contains genotype information from 600 random individuals living in Bodø or Tromsø at the time of sampling. All individuals have Norwegian ancestry, e.g., have four Norwegian grandparents (self-reported). The analysis of these samples was approved by the Norwegian Agency for Shared Services in Education and Research (SIKT) with reference number 194297, and details about sample collection, preparation and sequencing analyses are further described in [25]. Analytical threshold, interpretation threshold and stutter filter for autosomal STR-markers were applied according to the recommendations from the manufacturer. Any sample with dropout or ambiguous results in any of the markers that are included in the ForenSeq™ DNA Signature Prep Kit was re-sequenced and the concordance between length-based genotypes from UAS and genotypes obtained with the AmpFLSTR™ NGM SElect™ PCR Amplification Kit (Thermo Fisher Scientific) and capillary electrophoresis was checked: 540 samples analyzed with both MPS and CE were used, and three discordant observations were observed (99.99% concordance). As for the Southern Norway dataset, the Northern Norway dataset also included sequenced-based allelic information, both with and without flanks. In this study, genotypic and allelic information could only be made available per marker because data exchange between research groups is limited by the general data protection regulation.

2. The **Danish** dataset (**Denmark**) was published earlier and contains allele frequencies from 363 individuals of Danish ancestry [20], but genotype information per marker was not available. Only the repeat region of the sequences was used to perform the comparison, since not all the SNPs in the flanking regions of Southern Norway were observed in the published table of frequencies, e.g., SNP rs7789995 in marker D7S820.

#### 2.2.3. Quality control

*2.2.3.1. Concordance analysis CE vs MPS.* We compared length-based genotypes from MPS and CE data for the 371 samples in the Southern Norway dataset. Only overlapping aSTRs between the two kits were evaluated. Any discordances were further analyzed using STRait Razor version 3.0 [26], and FASTQ files from UAS were aligned using BWA [27] version 0.7.17-r1198-dirty. Bam files were visualized using Integrative Genomics Viewer (IGV) version 2.16.0 [28] to explore the relevant regions.

*2.2.3.2. Relatedness and duplicates.* Identical by state (IBS) analysis was used to detect possible duplicated samples and relatedness in the Southern Norway dataset. Genotypes were pairwise compared between all samples. Further screening for related individuals using the data from the CE analysis was performed with Familias [29,30] version 3.2.9, and the module Blind Search as detailed in [31]. Related candidate pairs were further investigated using MPS data.

### 2.3. Analysis of the Southern Norway dataset

#### 2.3.1. Frequencies and distinct alleles in autosomal STRs

Alleles and genotypes for the MPS sequences with flanks were retrieved per-marker from the UAS excel report “Project Autosomal STR Flanking Region Report”. An in-house R script was used to define the genotype pairs, and these were confirmed by comparing the length-based conversion of the alleles provided in the UAS report with the genotypes obtained from the CE analysis. The resulting frequency table was curated and submitted to STRidER, the STRs for identity ENFSI Reference database [32], for online publication (accession number: STR000387). Furthermore, the number of distinct alleles and allele frequencies for the three different levels of information previously explained in Section 2.2.1 were compared.

#### 2.3.2. Population genetics of autosomal STRs

Observed heterozygosity ($H_{obs}$) and gene diversity (GD), also known as expected heterozygosity ($H_{exp}$), measures the genetic variation in the population [33,34]. R package *adegenet* [35] was used to analyze $H_{obs}$ and GD, in the three information levels from Section 2.2.1. Hardy-Weinberg Equilibrium (HWE) was assessed with the function **HWPerm.Mult** from the R package *HardyWeinberg* [36]. **HWPerm.Mult** implements permutations tests for HWE in multiallelic markers [37]. P-values were adjusted using Benjamini-Hochberg (BH) correction. A global Weir and Cockerham’s $F_{ST}$ estimate was calculated per-marker with the function **wc** from the package *hierfstat* [38].

We calculated the average random match probability (RMP) per-marker using STRAF [39] in sequence-based data with flanks. Next, the difference in information gain (IG) was calculated between the three allele formats, i.e. the discriminatory change from using length-based information to using the sequence-based information (with or without flanks, Section 2.2.1). The overall information gain across all markers was assessed by calculating the expected random match probability (ERMP) on a $\log_{10}$ scale [40].

To infer possible population structure of Southern Norway, genotypes were divided into subpopulations based on the regions of origin of the individuals [21]; South, West, East, Central and the cities of Oslo and Bergen. Calculations were performed with modified R scripts from STRAF [39] and *adegenet* [35]. Weir and Cockerham’s pairwise per-marker $F_{ST}$ [41] values were compared for the different regions and visualized with multidimensional scaling (MDS). Genetic divergences were also compared with Nei’s Distance and visualized with MDS.

#### 2.3.3. Identity SNPs

We conducted similar analyses to those performed on aSTR for the 94 identity SNPs included in ForenSeq™ DNA Signature Prep Kit. As described in Section 2.2.1, UAS default thresholds were applied to obtain the genotypes. Suboptimal performance with these settings have been described before for particular iSNPs [6,20,42] and an improvement of reported genotypes by increasing the analytical threshold to 100 reads [20]. As the primary aim of the iSNPs study was not the characterization of frequencies, settings were not changed. Consequently, low read counts and high heterozygosity imbalance led to loss of several single alleles resulting in partial genotypes for some markers. STRAF was used to calculate per-marker observed heterozygosity ($H_{obs}$); within population gene diversity (GD); RMP and HWE p-values. Population substructure was assessed with iSNPs in combination with aSTRs, by considering the six regions: Oslo (capital), Bergen (city), South, West, East, and Middle and pairwise $F_{ST}$ values were calculated with STRAF.

To investigate the effect of the sample size on minor allele frequency (MAF) values, we compared SNPs from the 371 samples from Southern Norway, to values obtained from a larger dataset of 15,769 samples [43]. Two-sided 95% Clopper-Pearson confidence intervals for both datasets were calculated to illustrate sample size effects. In addition, a two-proportion Z-test was applied for testing for MAF differences between the two datasets. Non-overlapping confidence intervals should obtain small p-values from this test.

### 2.4. Comparisons with additional populations

With this study, genetic differences were compared between Southern Norway, Northern Norway, and Denmark. The full genotypes were not available for the last two populations, and accordingly, $F_{ST}$ estimates were calculated using in-house implementation where only it is necessary to input locus-specific information. This implementation consisted of biased [44] and unadjusted unbiased F-statistics [41], together with the adjusted formula from the function **thetaWC.pair** in R package *FinePop* [45].

Allele frequency similarity between populations was assessed with Fisher’s exact test. Frequencies of all allele sequences with flanks were compared per-marker between the two populations. Bonferroni adjustment was applied to correct for the number of multiple comparisons (i.e. number of distinct alleles tested): 505 for Southern Norway vs Northern Norway and 493 for Southern Norway vs Denmark. The initial p-value (significance level $\alpha_0 = 0.05$) was adjusted with the number of comparisons ($C$): $\alpha_1 = \alpha_0 / C$. Significant differences were considered if the p-values were below the updated significance level of $\alpha_1 = 0.0001$.

## 3. Results and discussion

### 3.1. Southern Norway dataset

#### 3.1.1. Quality control

The data obtained from MPS and CE analyses was subjected to a quality evaluation in order to ensure an optimal allele frequency database.

*3.1.1.1. Concordance analysis between CE and MPS data.* Of the total 8162 genotypes compared between CE and MPS, there were 15 discordances observed in 14 different samples (4% of the total samples) which leaves a concordance of 99.8%. Nine of the genotype differences corresponded to locus drop-outs in markers D22S1045 and Penta E, which has been also observed in previous studies [3,4,46]. For two different observations in markers D22S1045 and D16S539, no reads mapped alleles 19 and 13 in the MPS data, whilst these alleles were present in CE. None of the analyses with STRait Razor or visualization of the sequencing reads with IGV indicated the presence of the discrepant alleles, hence supplementary analysis would be needed for these two cases. One discordance was further evaluated for the marker Penta D as it was observed on three occasions (frequency of 0.008). The presence of the deletion rs536566765 (A>-) in the 3flanking region is interpreted as allele 13.4 with CE, but the repeat region in the full MPS sequence contains 14 repetitions, which would be interpreted as an allele 14. As previously discussed [3,47–49], such discrepancies may lead to inconsistencies in reported genotypes when different technologies are applied to the same sample, therefore the technology used must be taken into consideration when comparing casework samples that are analyzed by both CE and MPS.

*3.1.1.2. Relatedness and duplicates.* The results of the IBS analysis indicated that were no duplicates in the MPS dataset (i.e. samples sharing 54 sequence-based alleles). Out of approximately 68000 pairwise comparisons, one pair showed 28 shared alleles (this was the maximum observed). Supplementary Fig. S2 depicts shared allele counts. In parallel, the CE dataset was analysed with Familias Blind search module (propositions are $H_1$: Individuals are full-siblings vs $H_2$: Individuals are unrelated). One pairwise comparison obtained a moderately high LR $= 4 \times 10^5$; whereas the analysis with Pedigrees (propositions are $H_1$: Individuals are half-siblings vs $H_2$: Individuals are unrelated) gave a much lower LR = 3 $\times 10^3$. We further explored additional MPS information, and the two potentially related samples were compared with sequence-based information from 24 Y-STRs and 7 X-STRs profiles from ForenSeq™ DNA Signature Prep Kit (data not shown). It was found that alleles were only shared for two Y-STRs markers. In addition, analysis of sequence-based data revealed that five common length-based autosomal alleles were no longer shared. According to these results, the proposition that they could be siblings or half-siblings was discounted since only two Y-STRs and no X-STRs alleles were shared. This finding highlights the importance of using full sequences rather than the traditional number of repeats from CE, because of the higher level of information contained in the former.

### 3.1.2. *Alleles in autosomal STR markers*

*3.1.2.1. Frequencies.* Supplementary Table S1 and Supplementary Table S2 provide details of allele frequencies for, respectively:

a) Sequences including flanking regions as described in [18].

b) Sequences containing only repeat regions without flanking regions.

*3.1.2.2. SNPs in 5′ and 3′ flanking regions.* SNPs present in the 5′ and 3′ flanking regions were examined. These are listed in Supplementary Table S1. We observed 17 SNPs in total and compared them to those described in previous studies [4,18–20,49]. Four alleles with SNPs in flanking regions were neither described before nor reported in the STRseq catalog [23], namely:

1. SNP rs558394048 T>C (MAF = 0.001) in 3′ flanking region of allele [ATT]12 ACT [ATT]2 in marker D22S1045. MAF for European population was queried in 1000 Genomes through Ensembl [50] and it agreed with the results obtained in our study (MAF = 0.001).

2. SNPs rs1045120447 A>C (MAF = 0.001) and rs186259515 A>G (MAF = 0.007) in 3′ flanking region of allele [AAAGA]12 in Penta D marker. The MAF value for the European population in rs1045120447 was not available in 1000 Genomes, whilst for rs186259515 MAF = 0.009.

3. SNP rs1986487517 G>C (MAF = 0.001) in 3′ flanking region of allele [AGAT]14 in marker D20S482. It was found in the dbSNP database [51].

4. We observed one sequence with a G>T change in the 5' flanking region (MAF = 0.001) of the allele [AATG]6 in the TH01 marker (see Supplementary Fig. S3). This SNP was searched for in the dbSNP database, but no known SNP was found. Ensemble Variant Effect Predictor (VEP) [52] was also accessed using the Human Genome Variation Society (HGSV) [53] nomenclature (NC_000011.10:g.2171084 G>T (GRCh38) and NC_000011.9:g. 2192314 G>T (GRCh37)), and no effect on proteins was reported by experimental evidence or prediction.

*3.1.2.3. Comparison between sequenced-based and length-based alleles.* The number of distinct alleles per maker was compared between the three information levels described in Section 2.2.1 (Fig. 1). D12S391, D21S11, D2S1338 and D1S1656 loci exhibited the largest total number of distinct alleles. In contrast, TPOX, D4S2408 and D17S1301 had the lowest number of distinct alleles. In summary, 23 of the 27 aSTR (85.2% of the aSTRS) showed greater allele variability in the sequence-based format, compared to the length-based, and as illustrated in Fig. 1, no additional alleles were observed in TPOX, D17S1301, CSFPO and Penta E. The pattern of the increased number of distinct alleles for the Southern Norway population in this study is comparable to previous studies [12–14,16,30,31,34].

Allele frequencies were calculated for the three information levels outlined in Section 2.2.1 and the distributions were represented in density curves in Fig. 2. At lower frequency values, below 0.05, we observed a higher density of sequence-based alleles with 4.6 times more observations compared to length-based alleles. This area of high density corresponds in a great extension (62.9% of the alleles) to isoalleles, where one length-based allele with a higher frequency gives rise to two or more sequence-based alleles, each with lower frequencies. For frequencies exceeding 0.05, the distribution of distinct allele formats exhibited similar patterns. Overall, the distributions of the two sequence-based formats were comparable.

### 3.1.3. *Population genetics using autosomal STRs*

*3.1.3.1. HWE, heterozygozity and information gain.* A summary with the most characteristic population parameters such as the number of distinct alleles ($N_{all}$), $H_{obs}$ and GD values, for sequence-based alleles with and without flanks, and length-based alleles, is included in Supplementary Table S3. Markers D12S391, D1S1656, D21S11 and D2S1338 reported the highest $H_{obs}$ in sequence-based data with flanks, with values > 0.9 (highlighted in green in Supplementary Table S3), and as shown in the previous section, had the highest number of distinct alleles (Fig. 1). Fig. S4 illustrates relative changes in $H_{obs}$ between sequence-based genotypes and length-based for all loci. The greatest relative difference in $H_{obs}$ was observed in marker D9S1122 which was 19.4% greater than length-based data, similar to the findings of Delest et al. [16]; followed by markers D5S818 (16.1%) and D3S1358 (10%). The differences in $H_{obs}$ for markers D9S1122, D5S818 and D3S1358 may be explained by an increase of allele diversity paired with similar frequencies of SB alleles compared to LB alleles.

HWE analysis with exact tests (Supplementary Table S4) was performed separately for sequence-based data with flanks (Benjamini-Hochberg corrected p-values were in interval [0.695, 0.988]) and without flanks (Benjamini-Hochberg corrected p-values were in interval [0.818, 0.909]). Results were in Hardy-Weinberg Equilibrium, as adjusted p-values did not indicate a significant deviation of observed frequencies from expected frequencies for any of the markers.

The discriminatory capacity of sequence-based data with flanks was assessed by calculating the random match probability. Supplementary Fig. S5 illustrates the relationship between the $N_{all}$ (orange circles), $H_{obs}$ (light green), GD (blue) and average per locus PM (or RMP) (yellow). Similar to results reported in [18], the plot shows a trend where $H_{obs}$ and GD increase with the number of distinct alleles, whilst the RMP decreases (see all values in Supplementary Table S5 and additional figure). The expected RMP across all loci on log10 scale (ERMP) for sequences with and without flanks, and for length-based format was calculated as -40.2, -38.51 and -34.52 respectively. The information gain values showed that there is a small benefit (IG=1.044) of two orders of magnitude by including flanking regions in the analysis compared to that without flanks, whereas comparison of length-based with sequence-based (with and without flanks) showed benefit of six and four orders of magnitude (IG = 1.165 and IG = 1.116), respectively. These results are in agreement with previous studies [3,4,6,46], and indicate that sequence-based allele format improves the discrimination between individuals.

*3.1.3.2. Population structure.* Genetic similarities within Southern Norway were calculated based on Nei's Distance and Weir and Cockerham's pairwise $F_{ST}$ estimates, and these were represented with MDS in Supplementary figures Fig. S6 and Fig. S7. Additionally, $F_{ST}$ values were calculated per-marker in search for a relationship of allele pairs between the different areas of Southern Norway. Supplementary Table S6a shows pairwise per-marker $F_{ST}$ estimates ranging from -0.0031 to 0.006, with an average value of 0.002. The overall per-marker $F_{ST}$ values for Southern Norway (Supplementary Table S6b) varied from -0.003 to 0.012. As recommended in [54,55] to overcome variability per-locus, the average across all loci was calculated as 0.001. These values are lower compared to locus-specific and overall $F_{ST}$ estimates reported in Buckleton et al. [54] for Caucasian populations.

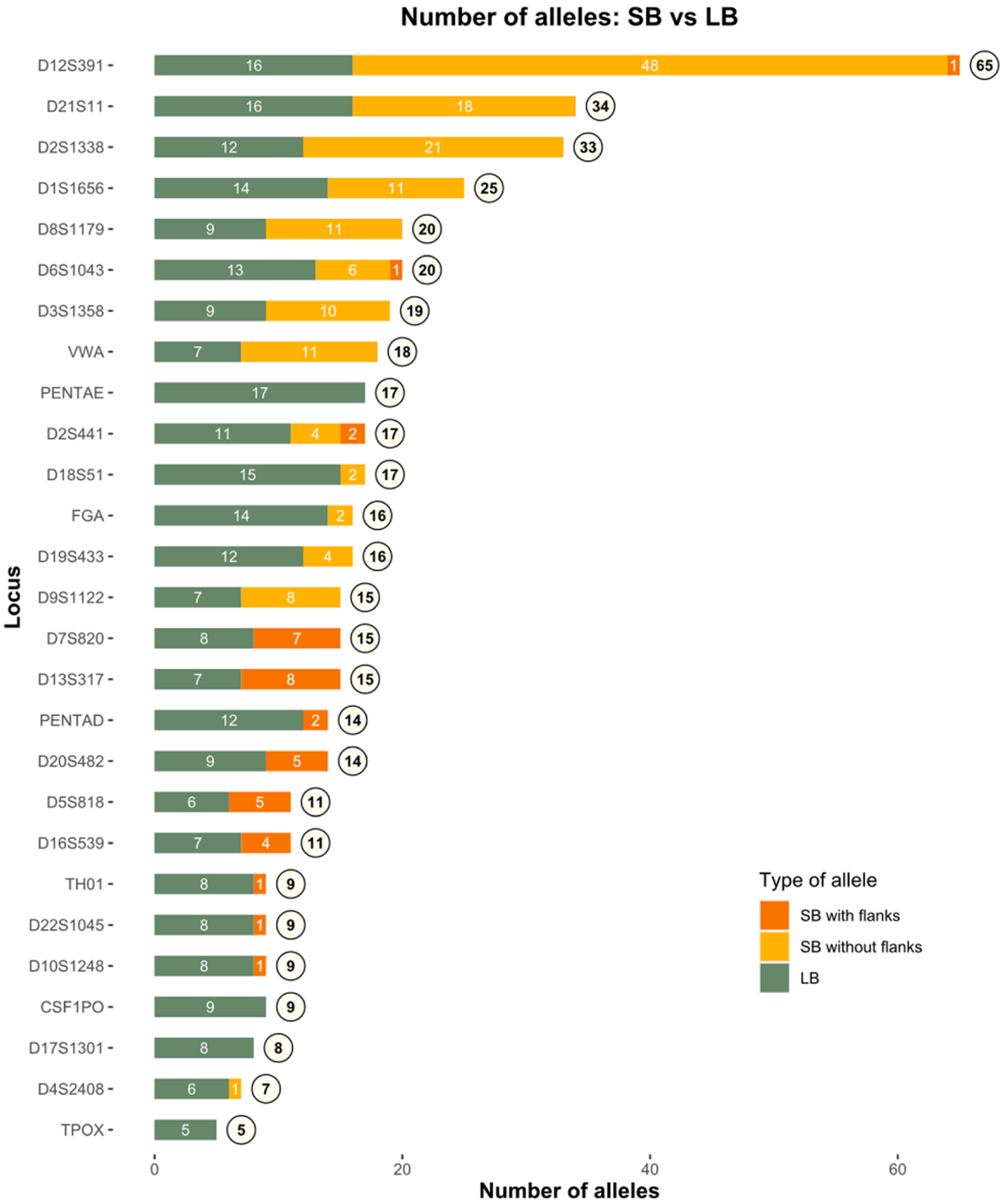

**Fig. 1. Number of alleles.** Comparison of the number of distinct length-based (LB) alleles (green segments), and sequence-based (SB) alleles. SB with flanking regions corresponds to the orange segments and without flanking regions to the dark yellow segments. The number in the circles at the end of the stacked bars indicates the total number of alleles.

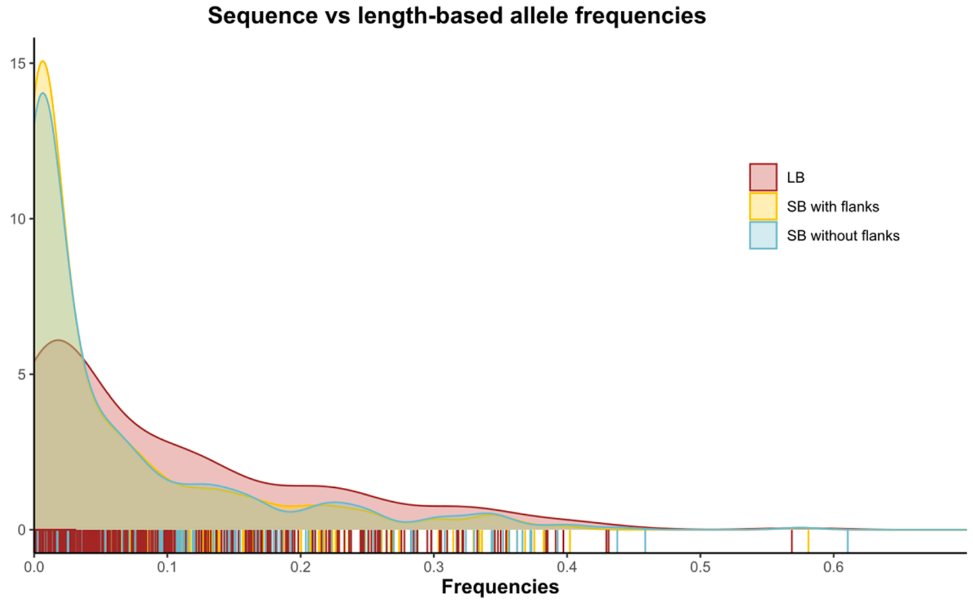

**Fig. 2. Distribution of allele frequencies.** Comparison of frequencies calculated from sequence-based (SB) alleles with flanking regions (yellow), without flanking regions (blue) and length-based (LB) frequencies (red).

In the MDS plots (Supplementary figures Fig. S6 and Fig. S7), areas of Bergen, Oslo and South were slightly divergent from the West, East and Middle areas of Norway. Results here should be interpreted with caution, as literature suggests genetic closeness of Oslo with counties close to the Oslofjord [43]. It has also been proposed that contemporary migration from neighboring agricultural regions to the city [56] has played a role in influencing the genetic landscape of this area. Nevertheless, in this study, the use of 27 aSTRs did not indicate any substructure of Southern Norway. Similar results were described in [57], where the analysis of different regions in UK with forensic aSTRs did not show a clear population substructure.

### 3.1.4. *Identity SNPs*

As mentioned in Section 2.2.1, UAS default settings were applied for the analysis of iSNPs and limitations were accounted for in downstream analysis. In all samples, at least 87.2% of the 94 iSNPs were typed. There was a failure to type iSNP rs1736442 (see Supplementary Fig. S8) in 55.3% of the samples. It was followed by iSNPs rs1031825, rs2920816, rs7041158, rs338882, rs1357617, rs719366 and rs1493232. Performance issues have been previously identified for some of these iSNPs [20,42].

*3.1.4.1. Population genetics analyses.* Results per-marker for the main population genetics parameters in the Southern Norway dataset are summarized in Supplementary Table S7. Lowest Benjamini-Hochberg adjusted p-values (0.012) for testing the Hardy-Weinberg equilibrium were observed for rs1736442, rs1031825, rs2920816, rs7041158, rs338882, rs1357617, rs719366 and rs1493232. Deviations for these iSNPs may be explained by the high loss of alleles/genotypes in the dataset, due to poor technical performance. The second lowest p-values were observed for iSNPs rs2056277 (0.047) and rs2342747 (0.042); the same assumption could explain a low p-value in rs2342747 given the 1.3% of allele/genotype loss in the dataset. For the remaining 84 iSNPs there is no evidence of deviation from the HWE. Compared with previous studies from Delest et al. [6] and Hussing et al. [20], seven iSNPs were commonly found that were not in Hardy-Weinberg equilibrium: rs1736442, rs1031825, rs2920816, rs7041158, rs338882, rs719366 and rs2342747. Hussing et al. [20] also suggested genotyping accuracy as the cause of the deviation. The mean $H_{obs}$ for the 94 iSNPs was 0.43. For 18 iSNPs $H_{obs}$ values were above 0.5, the SNP biallelic maximum heterozygosity value, and within the interval [0.501, 0.550]. Meanwhile, the mean of the GD was 0.45 and for six iSNPs, values approximated 0.5. In Supplementary Table S7 the range of $H_{obs}$ is represented for all 94 iSNPs together with the gene diversity (GD) and RMP values (combined RMP = 1.49E-37).

Supplementary Table S8a summarises the pairwise $F_{ST}$ estimates per iSNP in the six regions of Southern Norway; observed values were in the interval [-0.028, 0.022]. Supplementary Table S8b contains the matrix of pairwise $F_{ST}$ values between the different subpopulations. Comparable to the findings for aSTRs, there was no indication of genetic differences between the subpopulations. Additionally, iSNPs and aSTRs were combined and $F_{ST}$ values were plotted with MDS (Supplementary Table S8c and Fig. S9), and no clear evidence of structure was observed.

*3.1.4.2. Frequency comparison with a larger dataset.* The results of the MAF (see Supplementary Table S9 for values in Southern Norway) comparisons of 88 iSNPs from ForenSeq™ DNA Signature Prep Kit between the datasets “Southern Norway iSNPs” (n=371) and “All Norway iSNPs” (n=15,769) are summarized in Supplementary Table S10. We observed that for seven iSNPs (8%), p-values were approximately 0 (<< 0.01): rs1031825, rs1493232, rs1736442, rs8037429, rs873196, rs876724 and rs891700, indicating that there were significant differences between MAF values. In Fig. 3, the same iSNPs were observed to have non-overlapping confidence intervals. However, iSNPs rs1031825 and rs1736442 were highlighted in the previous section as they were not typed in a great percentage of samples (see Supplementary Fig. S8) leading to deviations from the HWE. This could explain some of the significant MAF differences between the two datasets. The relatively small number of differences indicates that the MAF estimates from the Southern Norway dataset (n=371) would be representative of the whole Norwegian population.

## 3.2. *Comparison with additional populations*

Genetic diversity was investigated among different populations by assessing the relationship of frequencies within a population genetics context. Accordingly, several implementations of the $F_{ST}$ calculation (see Section 2.4) were applied to compare sequence-based alleles with flanks between the Southern Norway dataset with Northern Norway (see Fig. 4 and Supplementary Table S11a). The biased $F_{ST}$ estimates were

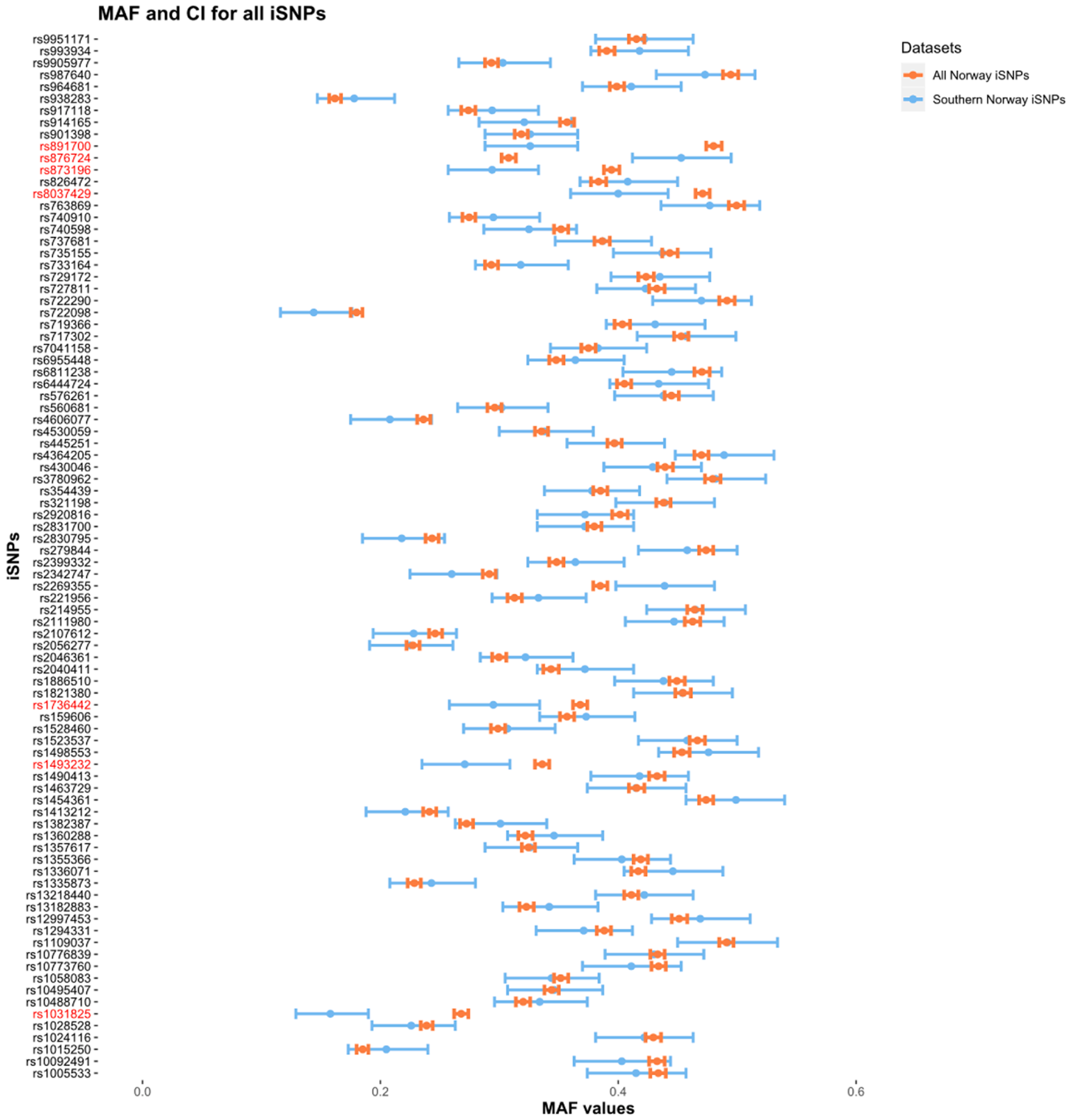

**Fig. 3. Comparison of minor allele frequency (MAF) values between Southern Norway iSNPs and All Norway iSNPs.** In the plot were represented MAF values (dots) from 88 iSNPs and the correspondent two-sided 95% Clopper-Pearson confidence intervals for the "Southern Norway iSNPs" in light blue colour and the "All Norway iSNPs" in orange colour.

similar for all loci and the average was 0.001. However, the unbiased implementation showed more variation, and the highest value was found in marker D12S391 (0.038). Using the unbiased adjusted method (sample correction) instead, the values reduced to an average of 0.006, and the $F_{ST}$ value of D12S391 decreased to 0.019. Only sequence-based allele (repeat region) frequencies, without genotype frequencies were available for the Danish dataset. Hence biased $F_{ST}$ calculations were conducted and compared against Southern Norway (Supplementary Table S11b). The average $F_{ST}$ was calculated as 0.088. $F_{ST}$ values above 0.15 were found in nine markers, ranging from 0.151 in marker D1S1656 to 0.319 in D21S11. These results are reflected in Fig. 4, as we can see differences in the distribution of the estimates for comparisons between North and Southern Norway, and Southern Norway and Denmark. For Northern and Southern Norway, the overall $F_{ST}$ values are smaller and only a minor shoulder is observed to the right of the density curves for the unbiased calculations, which may correspond to the marker D12S391. Furthermore, $F_{ST}$ estimates for Southern Norway and Denmark are quite disperse, as the wide distribution in Fig. 4 shows.

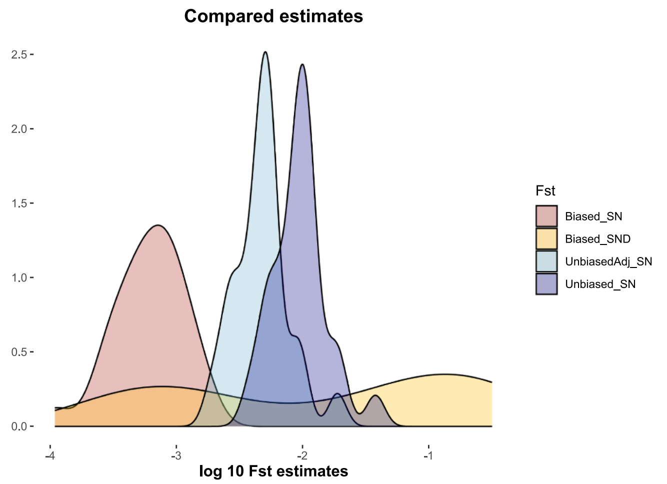

**Fig. 4. Density curves of per-marker $F_{ST}$ estimates for pairwise compared populations.** Biased_SND (yellow) uses a biased estimator and compares the Southern Norway with the Danish population (sequences without flanks). Biased_SN (light red), UnbiasedAdj_SN (light blue) and Unbiased_SN (purple) compares Southern Norway with Northern Norway (sequences with flanks) using a biased, adjusted unbiased and unbiased estimators respectively.

These differences can be due to the type of estimator implemented for each comparison. Interestingly, the markers exhibiting the highest $F_{ST}$ values were those in which a significant number of alleles were exclusively present in one of the populations. For example, in the comparison of Southern Norway vs Denmark, the marker D21S11 displayed 21 distinct alleles that were exclusively present in either the Southern Norway or the Denmark dataset. Similarly, in Southern Norway vs Northern Norway comparison, marker D12S391, exhibited 20 alleles that were specific to one of the populations.

Similarity between allele frequency values was assessed between the different populations. Supplementary Fig. S10 represents all p-values from Fisher's exact tests between Southern Norway and Danish population allele frequency pairs and the lowest p-value ($< 0.001$) is observed in marker D21S11. Southern Norway and Northern Norway frequencies exact test p-values are shown in Supplementary Fig. S11. Only for marker D4S2408 it was observed a p-value $< 0.001$. In both pairwise comparisons, none of the p-values were under the Bonferroni adjusted critical p-value 0.0001, meaning that no evidence for differences in frequencies between populations was found with this study. This result implies that it would be possible for Denmark and Norway (Southern and Northern datasets) to use the same frequency database, since differences between them appear to be negligible.

## 4. Conclusion

Population data is needed in order to have a complete implementation of MPS analyses into routine laboratory casework. By utilizing the Norwegian allele frequencies calculated in this study, it is possible to compute genotype probabilities that are representative of the entire population. These probabilities can then be employed in probabilistic genotyping software to facilitate weight of evidence calculations.

Information gain was evaluated for the different allele formats; results confirmed the advantages offered by using sequence-based alleles, compared to traditional length-based nomenclature. There was an increase of the number of distinct alleles in sequence-based data, and the expected RMP decreased six orders of magnitude. In this study, despite observing a modest reduction of RMP when adding flanking regions in the analysis, we still consider that the information contained in the flanks could be of relevance for the resolution of complex cases.

It is not always possible to count on large datasets for studying populations. Obtaining samples sometimes entails difficulties and MPS sequencing is still expensive. With the comparison of the MAF values for selected 88 iSNPs, we suggest that a sample of 371 individuals is sufficiently representative of the allele diversity in the population.

Understanding the genetic structure of populations is of relevance, not only in evolutionary terms, but also in forensics. However, more markers than those autosomal STRs and iSNPs used in this study would be necessary to obtain a better understanding of the genetic substructure within Southern Norway. Two different statistical approaches were applied to compare Southern Norway with neighbouring populations. With the first approach, allele frequencies values were compared one by one. For the second approach, the pairwise $F_{ST}$ values were calculated to investigate differences in genetic diversity or coancestry levels between the populations. In summary, no indication of differences in allele frequencies or genetic diversity between these populations was found. Consequently, it is probable that each of the frequency databases considered in this study could be used interchangeably, even if there is an apparent high $F_{ST}$ at some loci, driven by alleles present only in one of the populations.

## CRediT authorship contribution statement

**Maria Martin Agudo:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Håvard Aanes:** Conceptualization, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Michel Albert:** Investigation. **Kirstin Janssen:** Data curation, Formal analysis, Writing – review & editing. **Peter Gill:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Øyvind Bleka:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work is part of the PhD of Maria Martin Agudo and the project is funded by the Department of Forensic Sciences, Oslo University Hospital, Norway. We would like to thank Dr. Berit Myhre Dupuy for donating the samples that were used in this study and Dr. Morten Mattingsdal for sharing with us MAF data of SNPs in the Norwegian population. Also, we would like to thank Erik Francisco Bergseth and Professor Thore Egeland for their help with the sample analysis with Familias and their constructive discussions. Finally, we would like to acknowledge the technical support of Arne Roseth.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2024.103057.

## References

[1] F. Casals, R. Anglada, N. Bonet, R. Rasal, K.J. van der Gaag, J. Hoogenboom, et al., Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations, Forensic Sci. Int. Genet. 30 (2017 Sep) 66–70.

[2] P.A. Barrio, Ó. García, C. Phillips, L. Prieto, L. Gusmão, C. Fernández, et al., The first GHEP-ISFG collaborative exercise on forensic applications of massively parallel sequencing, Forensic Sci. Int. Genet. 49 (2020 Nov) 102391.

[3] N.M.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016 Nov) 214–226.

[4] L. Devesse, L. Davenport, L. Borsuk, K. Gettings, G. Mason-Buck, P.M. Vallone, et al., Classification of STR allelic variation using massively parallel sequencing and

assessment of flanking region power, Forensic Sci. Int. Genet. 48 (2020 Sep) 102356.

[5] A. Alonso, P.A. Barrio, P. Müller, S. Köcher, B. Berger, P. Martin, et al., Current state-of-art of STR sequencing in forensic genetics, Electrophoresis 39 (21) (2018 Nov) 2655–2668.

[6] A. Delest, D. Godfrin, Y. Chantrel, A. Ulus, J. Vannier, M. Faivre, et al., Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit, Forensic Sci. Int. Genet. 47 (2020 Jul) 102304.

[7] M.M. Foley, F. Oldoni, A global snapshot of current opinions of next-generation sequencing technologies usage in forensics, Forensic Sci. Int. Genet. 63 (2023 Mar) 102819.

[8] C.D. Steele, D.J. Balding, Choice of population database for forensic DNA profile analysis, Sci. Justice 54 (6) (2014 Dec) 487–493.

[9] M.D. Coble, J.A. Bright, Probabilistic genotyping software: An overview, Forensic Sci. Int. Genet. 38 (2019 Jan) 219–224.

[10] P. Gill, C. Benschop, J. Buckleton, Ø. Bleka, D. Taylor, A Review of Probabilistic Genotyping Systems: EuroForMix, DNAStatistX and STRmix™, Genes 12 (10) (2021 Sep 30) 1559.

[11] C.C.G. Benschop, K.J. van der Gaag, J. de Vreede, A.J. Backx, R.H. de Leeuw, S. Zuñiga, et al., Application of a probabilistic genotyping software to MPS mixture STR data is supported by similar trends in LRs compared with CE data, Forensic Sci. Int. Genet. 52 (2021 May) 102489.

[12] R.S. Just, J.A. Irwin, Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results, Forensic Sci. Int. Genet. 34 (2018 May) 197–205.

[13] S.B. Vilsen, T. Tvedebrink, P.S. Eriksen, C. Hussing, C. Børsting, N. Morling, Modelling allelic drop-outs in STR sequencing data generated by MPS, Forensic Sci. Int. Genet. 37 (2018 Nov) 6–12.

[14] K. Cheng, M. Lin, L. Moreno, J. Skillman, S. Hickey, D. Cuenca, et al., Modeling allelic analyte signals for aSTRs in NGS DNA profiles, J. Forensic Sci. 66 (4) (2021 Jul) 1234–1245.

[15] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int. Genet. 21 (2016 Mar) 35–44.

[16] Ø. Bleka, R. Just, M.M. Agudo, P. Gill, MPSproto: An extension of EuroForMix to evaluate MPS-STR mixtures, Forensic Sci. Int. Genet. 61 (2022 Nov) 102781.

[17] M.M. Agudo, H. Aanes, A. Roseth, M. Albert, P. Gill, Ø. Bleka, A comprehensive characterization of MPS-STR stutter artefacts, Forensic Sci. Int. Genet. (2022 May) 102728.

[18] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based U.S. population data for 27 autosomal STR loci, Forensic Sci. Int. Genet. 37 (2018 Nov) 106–115.

[19] P.A. Barrio, P. Martín, A. Alonso, P. Müller, M. Bodner, B. Berger, et al., Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power, Forensic Sci. Int. Genet. 42 (2019 Sep) 49–55.

[20] C. Hussing, R. Bytyci, C. Huber, N. Morling, C. Børsting, The Danish STR sequence database: duplicate typing of 363 Danes with the ForenSeq™ DNA Signature Prep Kit, Int. J. Leg. Med. 133 (2) (2019 Mar) 325–334.

[21] B.M. Dupuy, M. Stenersen, T.T. Lu, B. Olaisen, Geographical heterogeneity of Y-chromosomal lineages in Norway, Forensic Sci. Int. 164 (1) (2006 Dec) 10–19.

[22] Verogen. ForenSeq Universal Analysis Software Guide. 2018;182.

[23] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, et al., STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017 Nov) 111–117.

[24] Mitchell R., Sandage D. lusSTR [Internet]. Bioforensics; 2021. Available from: https://github.com/bioforensics/lusSTR.

[25] N.M. Salvo, K. Janssen, M.K. Kirsebom, O.S. Meyer, T. Berg, G.H. Olsen, Predicting eye and hair colour in a Norwegian population using Verogen's ForenSeq™ DNA signature prep kit. Forensic Sci. Int. Genet. 56 (2022 Jan) 102620.

[26] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Sci. Int. Genet. 30 (2017 Sep) 18–23.

[27] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009 Jul 15) 1754–1760.

[28] J.T. Robinson, J.P. Mesirov, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, et al., Integrative genomics viewer, Nat. Biotechnol. 29 (1) (2011) 24–26.

[29] D. Kling, A.O. Tillmar, T. Egeland, Familias 3 – Extensions and new functionality, Forensic Sci. Int. Genet. 13 (2014 Nov) 121–127.

[30] Egeland T., Mostad P.F., Mev B. Beyond traditional paternity and identification cases Selecting the most probable pedigree. Forensic Sci Int. 2000;

[31] E.F. Bergseth, A. Tillmar, P.J.T. Haddeland, D. Kling, Extended population genetic analysis of 12 X-STRs – Exemplified using a Norwegian population sample, Forensic Sci. Int. Genet. 60 (2022 Sep) 102745.

[32] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, et al., Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016 Sep) 97–102.

[33] B.S. Weir, Genetic Data Analysis II, Sinauer Associates, Sunderland, Massachusetts, 1996.

[34] M. Nei, Analysis of Gene Diversity in Subdivided Populations, Proc. Natl. Acad. Sci. 70 (12) (1973 Dec) 3321–3323.

[35] T. Jombart, *adegenet*: a R package for the multivariate analysis of genetic markers, Bioinformatics 24 (11) (2008 Jun 1) 1403–1405.

[36] Graffelman J. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. J Stat Softw [Internet]. 2015 [cited 2023 Feb 18];64(3). Available from: http://www.jstatsoft.org/v64/i03/.

[37] J. Graffelman, B.S. Weir, Multi-allelic exact tests for Hardy-Weinberg equilibrium that account for gender, Mol. Ecol. Resour. 18 (3) (2018 May) 461–473.

[38] J. Goudet, hierfstat, a package for r to compute and test hierarchical F-statistics, Mol. Ecol. Notes 5 (1) (2005 Mar) 184–186.

[39] A. Gouy, M. Zieger, STRAF—A convenient online tool for STR data evaluation in forensic genetics, Forensic Sci. Int. Genet. 30 (2017 Sep) 148–151.

[40] Ø. Bleka, R. Just, J. Le, P. Gill, An examination of STR nomenclatures, filters and models for MPS mixture interpretation, Forensic Sci. Int. Genet. 48 (2020 Sep) 102319.

[41] B.S. Weir, C.C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure, Evolution 38 (6) (1984 Nov) 1358.

[42] L. Davenport, L. Devesse, D. Syndercombe Court, D. Ballard, Forensic identity SNPs: Characterisation of flanking region variation using massively parallel sequencing, Forensic Sci. Int. Genet. 64 (2023 May) 102847.

[43] M. Mattingsdal, S.S. Ebenesersdóttir, K.H.S. Moore, O.A. Andreassen, T.F. Hansen, T. Werge, et al., The genetic structure of Norway, Eur. J. Hum. Genet. 29 (11) (2021 Nov) 1710–1718.

[44] S. Kitada, R. Nakamichi, H. Kishino, Understanding population structure in an evolutionary context: population-specific *F* ST and pairwise *F* ST, in: J. Ross-Ibarra (Ed.), G3 GenesGenomesGenetics, 11, 2021 Oct 19.

[45] Nakamichi R., Kishino H., Kitada S. FinePop: Fine-Scale Population Analysis [Internet]. 2018 [cited 2023 May 24]. Available from: https://cran.r-project.org/web/packages/FinePop/index.html.

[46] L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, D. Syndercombe Court, Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups. Forensic Sci. Int. Genet. 34 (2018 May) 57–61.

[47] J.D. Churchill, N.M.M. Novroski, J.L. King, L.H. Seah, B. Budowle, Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System, Forensic Sci. Int. Genet. 30 (2017 Sep) 81–92.

[48] C. Phillips, L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, et al., Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit, Electrophoresis 39 (21) (2018 Nov) 2708–2724.

[49] P. Hölzl-Müller, M. Bodner, B. Berger, W. Parson, Exploring STR sequencing for forensic DNA intelligence databasing using the Austrian National DNA Database as an example, Int. J. Leg. Med. 135 (6) (2021 Nov) 2235–2246.

[50] F. Cunningham, J.E. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, et al., Ensembl 2022, Nucleic Acids Res. 50 (D1) (2022 Jan 7) D988–D995.

[51] S.T. Sherry, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (1) (2001 Jan 1) 308–311.

[52] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, et al., The Ensembl Variant Effect Predictor, Genome Biol. 17 (1) (2016 Dec) 122.

[53] J.T. den Dunnen, R. Dalgleish, D.R. Maglott, R.K. Hart, M.S. Greenblatt, J. McGowan-Jordan, et al., HGVS Recommendations for the Description of Sequence Variants: 2016 Update, Hum. Mutat. 37 (6) (2016 Jun) 564–569.

[54] J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, B.S. Weir, Population-specific F values for forensic STR markers: A worldwide survey, Forensic Sci. Int. Genet. 23 (2016 Jul) 91–100.

[55] S.E. Aalbers, M.J. Hipp, S.R. Kennedy, B.S. Weir, Analyzing population structure for forensic STR markers in next generation sequencing data, Forensic Sci. Int. Genet. 49 (2020 Nov) 102364.

[56] D. Kristjansson, J. Bohlin, A. Jugessur, T.G. Schurr, Matrilineal diversity and population history of Norwegians, Am. J. Phys. Anthr. 176 (1) (2021 Sep) 120–133.

[57] T.I. Huszar, W.F. Bodmer, K. Hutnik, J.H. Wetton, M.A. Jobling, Sequencing of autosomal, mitochondrial and Y-chromosomal forensic markers in the People of the British Isles cohort detects population structure dominated by patrilineages, Forensic Sci. Int. Genet. 59 (2022 Jul) 102725.