SNP-based quantitative deconvolution of biological mixtures: application to the detection of cows with subclinical mastitis by whole-genome sequencing of tank milk

Wouter Coppieters,[1] Latifa Karim,[1] and Michel Georges[2]

[1]Genomics Platform, GIGA Institute, University of Liège; [2]Unit of Animal Genomics, GIGA Institute & Faculty of Veterinary Medicine, University of Liège, 4000 Liège, Belgium

Biological products of importance in food (e.g., milk) and medical (e.g., donor blood-derived products) sciences often correspond to mixtures of samples contributed by multiple individuals. Identifying which individuals contributed to the mixture and in what proportions may be of interest in several circumstances. We herein present a method that allows to do this by shallow whole-genome sequencing of the DNA in mixed samples from hundreds of donors. We show the efficacy of the approach for the detection of cows with subclinical mastitis by analysis of farms' tank mixtures containing milk from as many as 500 cows.

[Supplemental material is available for this article.]

Mastitis, namely, the inflammation of the udder, is the most important health issue in dairy cattle. It is estimated to cost European farmers more than 1 billion euros per year in treatment and milk loss (Hogeveen et al. 2011). Upon inflammation, immune cells migrate into the udder and milk. Although milk from healthy cows typically contains less than 100,000 cells per milliliter of milk, these numbers (referred to as somatic cell counts [SCCs]) typically increase into the millions in case of mastitis. Before the manifestation of overt clinical symptoms, SCCs progressively increase in the milk of cows developing mastitis: SCCs ≥ 200,000 are typically considered to be a sign of pre- or subclinical mastitis. Both yield and quality of the milk of cows with subclinical mastitis is reduced (Schukken et al. 2003). Mastitis is routinely managed by periodically counting the SCC in milk samples to preemptively identify cows developing subclinical udder inflammation. As profit margins decrease, farmers tend to forgo milk testing, thereby compromising health management. Cost-effective alternatives for rapid detection of cows with subclinical mastitis are hence needed (Viguier et al. 2009).

The milk obtained from individual cows is typically collected in one or more large "milk tanks" on the farm before being shipped to dairy factories. We previously proposed that SCCs in the milk of individual cows could be estimated by measuring the allelic frequencies in the tank milk for sufficient numbers of SNPs, provided that all cows contributing milk to the tank be genotyped once for the corresponding variants. Thus, the proposed method would allow the identification of a minority of cows with subclinical mastitis by regularly analyzing a single sample containing a mixture of milk from all the cows on the farm, hence considerably reducing costs (Blard et al. 2012). Before around 2010, the estimation of breeding values to select the best dairy sires and dams used pedigree-based estimates of kinship. Since then, selection methods increasingly use genome-wide SNP information in a process referred to as "genomic selection" (GS) (Georges et al. 2019). As GS is becoming routine in dairy cattle (including for dams), herds that are fully genotyped with genome-wide SNP arrays are becoming standard, and the proposed method now feasible. However, as GS typically relies on the use of low-density SNP arrays, the basic method proposed by Blard et al. (2012) is only effective for small farms (100 or fewer cows). We herein show that by combining low-density SNP genotyping or shallow sequencing of the cows and tank milk's DNA with in silico genotype imputation, individual SCCs can be accurately determined and cows with subclinical mastitis effectively identified even in the largest farms (500 or more cows). The proposed method has the potential to improve the monitoring of udder health in dairy farms and to allow the tracing of the origin of bulk animal food products other than milk.

Results

Principle of the proposed method

Assume that cows and tank (i.e., the reservoir in which the milk of the cows is collected) milk are genotyped for a collection of SNPs. Assume that the interrogated SNPs are biallelic, each characterized by an *A* (say the allele of the reference genome) and a *B* allele (say the alternate allele). If all cows contribute identical amounts of DNA to the milk, the expected proportion of the *B* allele (commonly referred to as "*B*-allele frequency" when analyzing SNP array data particularly to search for copy number variants) in the tank milk corresponds to the frequency of the *B* allele in the farm's cow population. The actual DNA amount contributed by each cow depends on the volume of milk that she produced and its SCC. Unequal DNA contributions will cause slight departures from the expected *B*-allele frequencies in the tank milk.

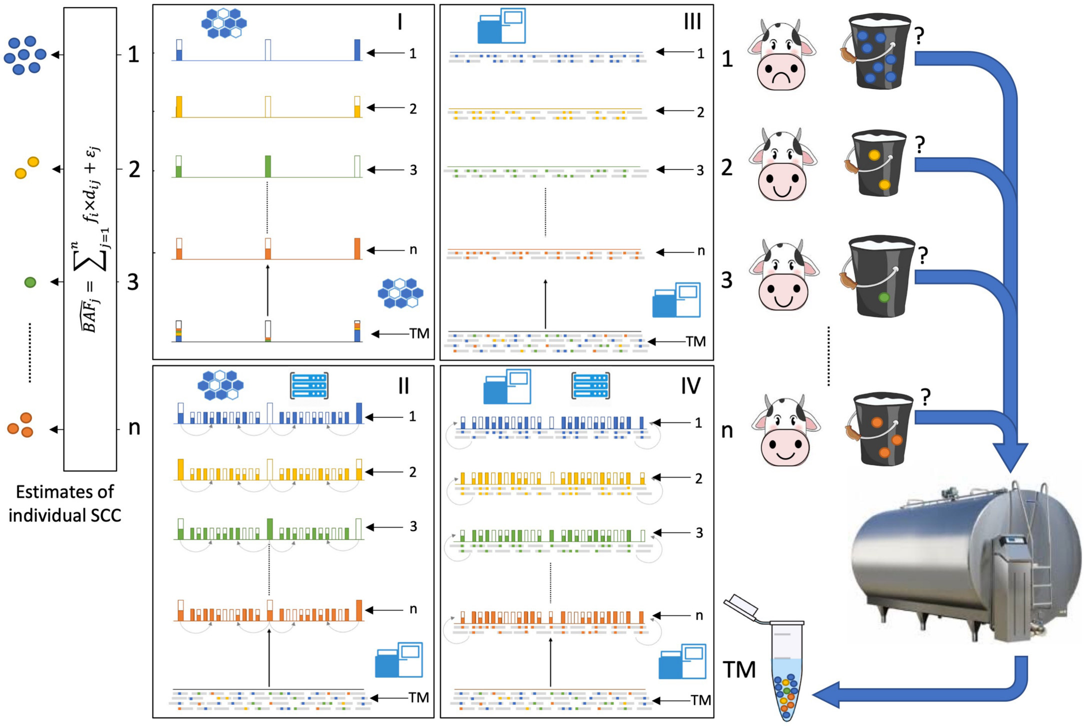
Corresponding author: michel.georges@uliege.be

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.256172.119.

Integrating these shifts over a large number of SNPs in conjunction with the known genotypes of individual cows allows for the estimation of the relative DNA contribution of each cow. This can, for instance, be achieved using a set of *m* linear equations in which the "*B*-allele frequency" of each SNP *j* (of *m*) is modeled as the sum (over *n* cows) of the products of the dosage of the *B* allele in the genotype of cow *j* (d_{ij}, known from her SNP genotype) multiplied by the proportion of DNA contributed by cow *i* (f_i) to the milk. The proportions of DNA contributed by each cow can then be estimated using, for instance, least square methods. Accounting for individual milk volumes and for the SCCs in the tank milk allows for the estimation of SCCs for individual cows (see Methods) (Fig. 1).

Evaluating the proposed method by simulation

We first evaluated the proposed method by simulation (see Methods). Genotyping the cows and the tank milk using 10,000 SNP arrays (i.e., low-density [LD] arrays as generally used in the context of GS) allowed for the accurate estimation of individual SCCs for farms with up to 100 cows ($r \geq 0.9$, where *r* is the correlation between real and estimated SCCs; scheme I). However, farms with more than 100 cows are increasingly common. Medium- (MD; e.g., 50,000) and high-density (HD, e.g., 700,000) SNP arrays would be needed for the approach to be effective in farms with 250 or more or 500 or more cows, respectively. Yet—being too expensive—this is presently not a viable proposition (Fig. 2A; Supplemental Table 1). We therefore envisaged a second scheme (II) in which the cows would still be genotyped with LD SNP arrays (as performed in practice) yet imputed (Marchini and Howie 2010) to the whole genome (8 million [M] SNPs in the simulations) using a sequenced reference population (e.g., Daetwyler et al. 2014), whereas the DNA of the tank milk would be genotyped by shallow whole-genome sequencing (SWGS). We found that in this scenario, sequencing the tank milk at a depth of 0.25× was sufficient for farms with 100 cows, 0.5× for farms with 250 cows, and 2× for farms with 500 cows (Fig. 2B). Accuracies were not significantly affected by the density of the SNP arrays, that is, the method performed as well with LD as with MD arrays (Supplemental Fig. 1). Anticipating further advances in sequencing technology, we also envisaged a scheme (III) in which both cows and tank milk would



$\widehat{BAF}_j = \sum_{j=1}^{n} f_i \times d_{ij} + \varepsilon_j$
1
2
3
n
Estimates of individual SCC
I
II
III
IV
1
2
3
n
TM
?
TM

Figure 1. Estimating somatic cell counts (SCCs) in the milk of individual cows by analyzing a sample of milk from the farm's tank. Cows 1 to *n* contribute different amounts of milk (buckets of various sizes in the figure) to the farm's tank. The milk contains somatic cells (shown as small spheres in the milk colored by cow) whose numbers reflect the health status of the cow's udder. Cow 1 has higher a SCC, an indicator of subclinical mastitis. SCCs are unknown upon milking (indicated by a question mark). Cows are individually genotyped once. In scheme I, this is performed using SNP arrays (illustrated by the mesh), yielding genotype information for the limited number of interrogated SNPs (high bars) that can be summarized by the *B*-allele frequency as shown (white: 0; half-colored: 0.5; fully colored: 1). SNP genotypes of individual cows are coded in the same colors as the SCCs. In scheme II, the genotypes of the interrogated SNPs are augmented by imputation (illustrated by the computer rack), yielding dosage information (*B*-allele frequency) for many more SNPs (small bars). In scheme III, cows are genotyped individually by shallow whole-genome sequencing (SWGS; illustrated by the sequencer). Sequence reads (gray lines) are aligned to the reference genome, and alternate alleles at SNP positions are highlighted as color-coded tics. The *B*-allele frequency at specific SNP positions is measured as the ratio of the number of reads with the *B* allele versus the total number of reads. In scheme IV, the genotype information from SWGS is augmented by imputation improving the accuracy of the *B*-allele frequency estimates for millions of SNPs (small bars). A small sample of milk (tank milk [TM]) is periodically (e.g., monthly or weekly) collected from the farm's tank. DNA is extracted from the TM and genotyped using SNP arrays (scheme I) or SWGS (schemes II, III, and IV). *B*-allele frequency for SNP *j* in the milk ($\widehat{BAF_j}$) is estimated from the ratio of fluorescence intensities when using SNP arrays or from the proportion of reads with *B* allele in SWGS. The SCCs of individual cows are estimated from a set of linear equations modeling $\widehat{BAF_j}$ as the sum of *B*-allele dosage (d_{ij}) multiplied by the proportion of the DNA in the tank contributed by cow *i* (f_i). The estimated proportions of DNA contributed by each cow correspond to the values of f_i's that minimize the sum of squared errors (ε_j) over all SNPs. The SCCs for individual cows, per se, can be estimated as $SCC_i = SCC_{tank} \times V_{tank} \times f_i / V_i$, where SCC_{tank} is the SCC measured in the farm's tank, and V_i / V_{tank} is the proportion of the milk volume contributed by cow *i*.

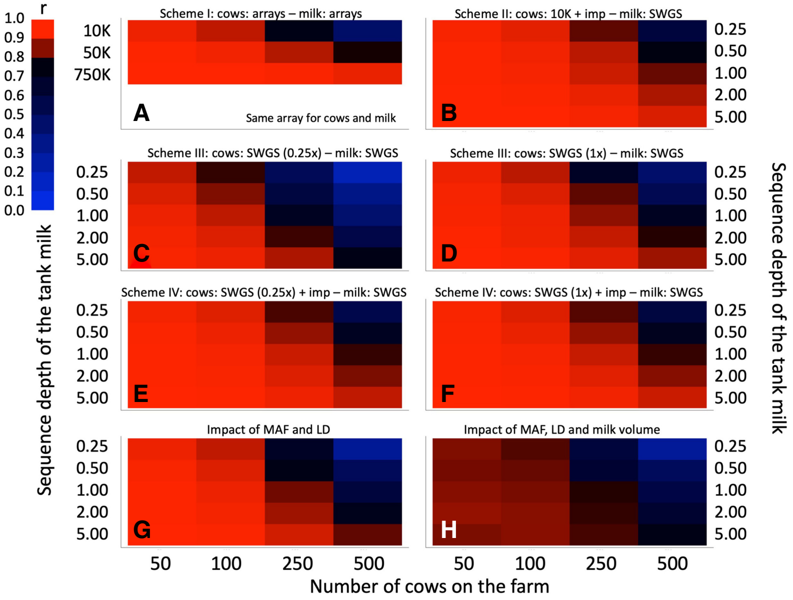


Figure 2. Evaluating the efficiency of the proposed approach by simulation. (*A*) Reference scheme I in which individual cows and tank milk are genotyped with the same array interrogating 10,000 (LD), 50,000 (MD), or 700,000 (HD) SNPs. (*B*) Scheme II in which individual cows are genotyped with a LD 10,000 SNP array and imputed to whole genome (8 M SNPs), whereas the tank milk is whole-genome sequenced at depth ranging from 0.25× to 5×. (*C*) Scheme III in which individual cows (0.25×) and tank milk (range: 0.25× to 5×) are genotyped by SWGS. (*D*) Same as *C* except that individual cows are sequenced at 1× depth. (*E*) Scheme IV in which individual cows are genotyped by SWGS (0.25×) followed by imputation to whole genome (8 M SNPs), and tank milk is genotyped by SWGS (range: 0.25× to 5×). (*F*) Same as *E* except that individual cows are sequenced at 1× depth. (*G*) Scheme in which the cow genotypes are sampled from a real data set and hence conform to reality with regard to distribution of MAF, LD, and relatedness. Genotypes of the cows are assumed to be known (very similar to II and IV) and tank milk genotyped by SWGS (range: 0.25× to 5×). (*H*) Same as *G* except that the milk volume is estimated with error. The color code used to quantify the correlations between predicted and real SCCs is shown. Corresponding numerical values are provided in Supplemental Table 1.

be genotyped by SWGS. We found that a 1× sequencing depth of the tank milk would be sufficient when combined with a 0.25× depth for 100 cows, whereas a 5× sequencing depth of the tank milk would be needed in combination with 0.25× depth for 250 cows and 1× depth for 500 cows (Fig. 2C,D). In scheme III, allelic dosage in the cows is directly measured from the number of alternative and reference alleles in the sequence reads. We further explored the effectiveness of augmenting the cow genotype information from SWGS by imputation (scheme IV). This proved to be effective, reducing the required sequence depth to 0.25× for tank milk and 0.25× for 100 cows, to 1× for tank milk and 0.25× for 250 cows, and to 5× for tank milk and 0.25× for 500 cows (Fig. 2E,F). The previous simulations make a number of assumptions that may not apply in the real world: (1) SNPs were sampled from a uniform distribution (i.e., rare and common SNPs equally represented); (2) SNPs were assumed to be in linkage equilibrium; (3) cows on the farm were assumed to be unrelated; and (4) milk volumes were assumed to be known without error. To more accurately mimic real conditions, we repeated the simulations by (1) sampling genotypes from a phased data set of 750 Holstein-Friesian whole-genome sequences (hence properly accounting for true MAF distribution, true linkage disequilibrium [LD], and relatedness; many of the sequenced animals were related as parent offspring or full- or half-sibs) and by (2) adding a normally distributed error with mean 0 and standard deviation of 5 L to the simulated milk volumes (assumed to be normally distributed with mean of 30 liter and standard deviation of 10 L). This error rate corresponds approximately to that expected when having to estimate the daily milk volume from the total lactation yield using a standard lactation curve (Atashi et al. 2019). We assumed in these simulations that the genotypes of the cows were known without error and that the milk was sequenced at a depth ranging from 0.25× to 5× as before. MAF, LD, and relatedness jointly had a relatively modest impact on the accuracy of the method, which could be compensated for by increasing the sequencing depth of the milk to fivefold and still allowing for accurate estimates even in farms with 500 cows. Estimating the milk volume with error had a more pronounced impact on the accuracy, making it difficult to reach a correlation reaching 0.9 in farms with 500 cows (Fig. 2G,H).

Real-world application of the proposed method

To test the feasibility of our method in the real world, we first collected cow (blood) and tank (milk) samples from a farm milking 133 Holstein-Friesian cows. When only using genotypes from the Illumina LD arrays (17,000 SNPs) for both cows and tank milk (scheme I), correlations between predicted and measured SCCs were 0.91 (or 0.79 when ignoring one cow with SCC > 3 million). We then imputed the cows to whole genome (13 M SNPs) using a reference population of approximately 750 whole-genome sequenced Holstein-Friesian animals, and sequenced the tank milk at ∼3.5× depth. The corresponding correlations (scheme II) were 0.97 (0.95) when using all sequence information or 0.96 (0.92) when down-sampling sequence information as low as 0.1× depth (Fig. 3A). We next performed a similar experiment on a farm milking 520 Holstein-Friesian cows. The correlation between predicted and measured SCCs was 0.78 (or 0.42 when ignoring 23 cows with SCC > 3 million) when only using information from the LD array for both cows and tank milk (scheme I). When imputing the cows to whole genome (13 M SNPs) and sequencing the milk at ∼3.5× depth (scheme II), the correlation increased to 0.89 (0.83). Down-sampling the sequence information to 0.1× depth reduced the correlation to 0.79 (0.57) (Fig. 3B).

As shown in both farms, correlation estimates are affected by SCC spread: Small numbers of cows with very high SCCs tend to inflate *r*. We therefore computed accuracies, computed as the proportion of correctly classified cows for different SCC thresholds, which is how farmers would likely use the information. It can be seen that for a threshold value of, for example, 500,000 SCCs, accuracies >0.85 were obtained when sequencing (scheme II) the tank milk at, respectively, 0.1× (133 cows) and 3.5× depth (520 cows). Thus—as predicted by the simulations—scheme I provided adequate precision for the farm with 133 cows but not for the farm with 520 cows. However, in this large farm, combining SWGS of

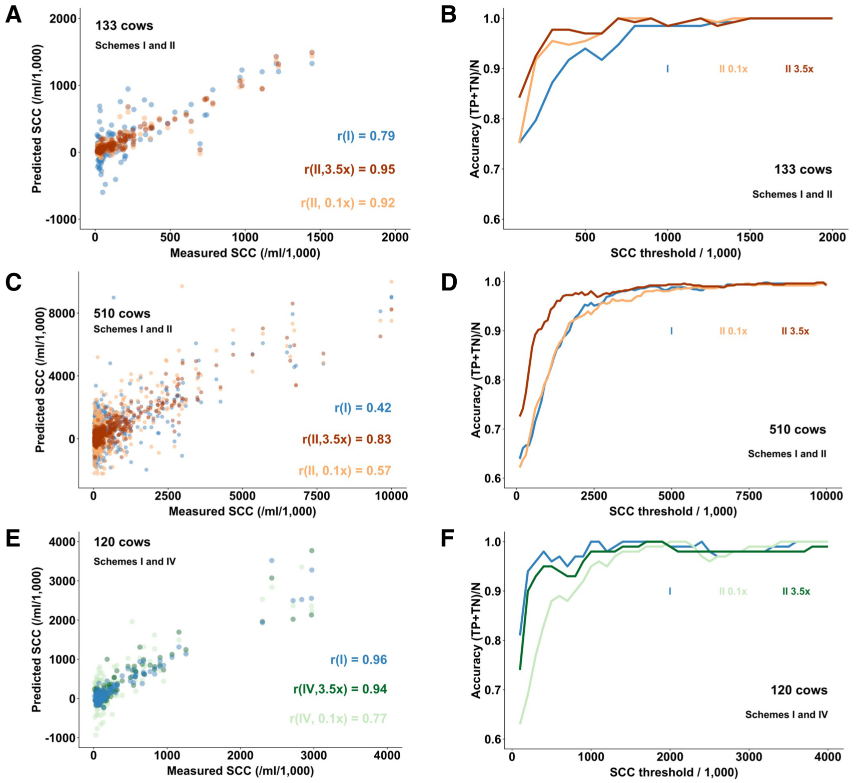


Figure 3. Correlation between predicted and measured SCCs in the milk of individual cows (*A*,*C*,*E*), as well as accuracies in classifying cows with SCCs above and below a chosen threshold value (*B*,*D*,*F*), in farms with 133 (*A*,*B*), 520 (*C*,*D*), and 120 (*E*,*F*) cows, using scheme I (blue), scheme II (red), or scheme IV (green). (Scheme I) Cows and tank milk genotyped with LD SNP arrays (17 K), no imputation. (Scheme II) Cows genotyped with LD array and imputed to 13 M SNPs; tank milk sequenced 3.5× (red) or 0.1× (orange). (Scheme IV) Cows genotyped by WGS (1×) and imputation to HD; tank milk sequenced at 3.5× (dark green) or 0.1× (light green).

the tank milk with whole-genome imputation of the cows (i.e., scheme II) was indeed effective (Fig. 3).

As costs per base pair continue to decline, sequencing is likely to replace array-based genotyping in the future. To test the feasibility of schemes III and IV (i.e., genotype the cows by SWGS rather than with SNP arrays, without (III) and with (IV) imputation), we collected samples from a farm with 120 Holstein-Friesian cows. All cows were genotyped with the Illumina LD array (17,000) as well as sequenced at average 1.08× depth (range: 0.26–1.73). The milk was sequenced at ~3.5× depth. The correlation between predicted and measured SCCs was 0.97 (or 0.96 when ignoring one cow with SCC > 3 million) under scheme I. Under scheme III, correlations were 0.82 (0.83) when sequencing the milk at 3.5× and 0.75 (0.76) when down-sampling the milk to 0.1×. We then imputed the sequenced cows to HD (770,000 SNPs) using a population of 800 reference animals genotyped with the HD array (scheme IV). The correlation increased to 0.93 (0.94) when sequencing the milk at 3.5× and to 0.83 (0.77) when down-sampling the milk to 0.1× (Fig. 3C). Accuracies at SCC threshold of 500,000 were 0.96 (scheme I), 0.95 (3.5×) and 0.80 (0.1×) (scheme II), 0.82 (3.5×) and 0.81 (0.1×) (scheme III), and 0.95 (3.5×) and 0.88 (0.1×) (scheme IV) (Fig. 3C). In summary, (1) combining cow genotyping using SNP arrays with genome-wide imputation with SWGS of tank milk allows for cost-effective identification of cows with subclinical mastitis even in farms with as many as 500 cows per milk tank, and (2) as sequencing costs continue to decline, arrays-based targeted SNP genotyping of the cows could be replaced by genotyping by SWGS and yield comparable results.

Monitoring SCC dynamics with the proposed method

Farmers typically measure individual SCCs once a month or less. Yet, SCCs may rapidly change. The SCC measured on the milk testing date may not be a reliable indicator of the cow's udder health during the intervening period. To examine the SCC dynamics over time, we collected 20 tank milk samples over a 100-d period (day −84 to +17 from day of milk testing) for the farm with 120 cows. Milk samples were genotyped using the Illumina LD array and individual SCCs estimated using scheme I. Figure 4A shows the SCCs predicted every 5 d on average for the 120 cows, sorted by SCCs measured on day 0 (milk testing day). Of note, the correlation between the SCCs measured on day 0 and the average of the SCC estimates for the 21 collection dates was low (r = 0.52) (Fig. 4B) and decreased rapidly with the number of days from milk testing day (Fig. 4C).

Discussion

We herein show that by combining array-based SNP genotyping and whole-genome imputation for the cows with SWGS of the tank milk, it is possible to accurately estimate SCCs for individual cows and hence effectively identify animals with subclinical mastitis even for tanks collecting milk for more than 500 cows, and this by performing a single analysis for the entire herd. Reagent costs to sequence a mammalian genome at onefold depth are now less than 20 euros, thus making this a cost-effective proposition. As a matter of fact, the method is being deployed in the field in several countries.

Implementing the method requires all cows on the farm to be genotyped. This will increasingly correspond to reality as genotyping costs continue to decrease and GS is more and more used for the selection of cows. In 2016, more than 1.2 million dairy cows had been reportedly genotyped in the US alone (Wiggans et al. 2017), and present worldwide numbers are likely 3 million or more. In addition, a reference population of a few hundred animals of the breed of interest that are either HD genotyped (700,000) or better whole-genome sequences are required for accurate imputation. Such reference populations are already available for the most important dairy cattle breeds (Daetwyler et al. 2014; Charlier et al. 2016) and could be easily generated for the remaining ones.

We show that SCCs are dynamic and rapidly change over time. SCCs measured on day 0 are poor indicators of SCC in previous and future weeks: Cows with high SCCs on the day of milk testing may have low SCCs a few days later (or earlier) and vice versa.

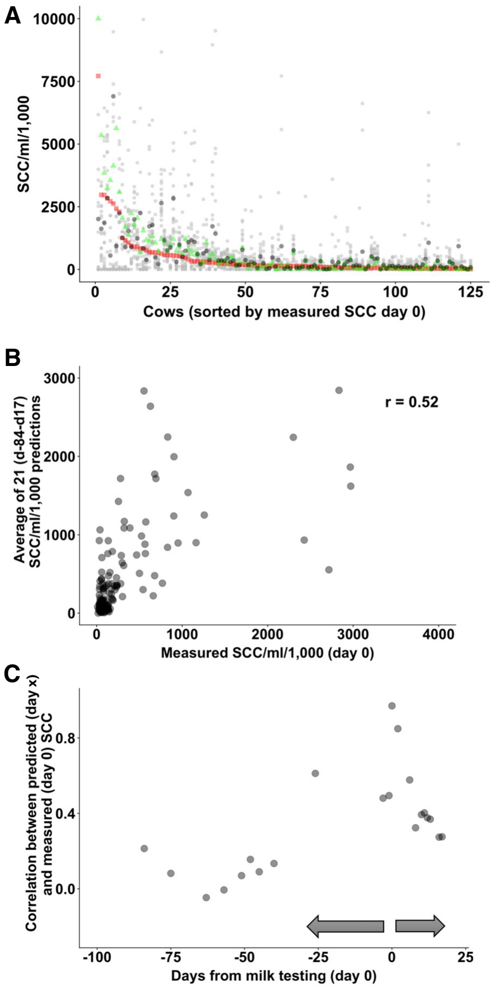


Figure 4. Evaluating SSC dynamics. (*A*) SCCs predicted using scheme I for 21 tank milk samples collected over a 100-d period from 125 cows total. Small gray circles indicate 20 predictions per cow; large gray circles, average of 21 measurements per cow; red square, SCCs measured on day 0; green triangle, SCC predictions on day 0. (*B*) Relationship between SCC values measured on day 0 and average of 21 predictions sampled over a 100-d period (days −84 to +17). (*C*) Correlations between measured (day 0) and predicted (day x) SCCs as a function of the number of days from day 0.

The proposed method would allow tighter monitoring of SCCs, hence improving udder health management. More frequent monitoring of SCCs for large number of cows may reveal interindividual differences with regard to SCC dynamics that may be correlated with mastitis resistance, heritable and hence amenable to selection including by GS.

Sequencing of the DNA in the tank milk allows simultaneous characterization of the tank's microbiome. As a matter of fact, ∼1% of reads in this study mapped to bacterial genomes. This information may be very useful both from a farm health management point of view as well as from a downstream dairy processing point of view. Whole-genome sequence data of bulk milk also informs about the herd frequency of functional variants such as casein variants affecting consumer health or processing properties (Brooke-Taylor et al. 2017) or variants causing inherited defects or embryonic lethality in cows (Georges et al. 2019). In many countries, it is not allowed to add milk from cows being treated with antibiotics to the tank. As suggested before, the proposed approach can be adapted to verify whether a specific cow did contribute milk to the tank or not (e.g., by testing the significance of the corresponding cow effect in the linear model) (Blard et al. 2012). The described method may have applications in tracing the origins of bulk animal food products other than milk, as well as in monitoring the composition of mixed-donor blood-derived transfusion products.

Methods

Simulated data

Reference scheme (I)

We simulated farms with n (25, 50, 100, 250, and 500) cows contributing milk to the tank. Cows were genotyped with SNP arrays for m (10,000, 50,000, or 750,000) markers without error. Minor allele frequencies (MAFs) were sampled from a uniform [0, 0.5] distribution, and genotypes from the corresponding Hardy–Weinberg distributions. SCCs of individual cows (SCC_i) were simulated by sampling values from a Weibull distribution with scale parameter $\alpha = 1$ and shape parameter $\beta = 2$ and by multiplying the ensuing value by 200,000. Exact B-allele frequencies of individual SNPs (BAF_j) in the milk were determined for each SNP j based on the combination of cellular contribution of the n cows to the milk and of their genotype. It was assumed that B-allele frequencies were estimated with a normally distributed error $N(0, 0.0025)$ (i.e., SE = 0.05), yielding $m\ \widehat{BAF_j}$.

Scheme II

Same setting as in the reference scheme with the following additions. For cows genotyped for 10,000 or 50,000 SNPs, we simulated imputation by augmenting the data to 8 M genotypes using an error model mimicking real, MAF-dependent imputation accuracy. The error model was constructed using a real data set for 800 unrelated Holstein-Friesian individuals that were genotyped for the Illumina 777K array. This data set was split into a set of 200 and a set of 600 individuals. The set of 200 was reduced first to the genotypes interrogated by the Illumina 10K (LD) array and then to the genotypes interrogated by the Illumina 50K SNP arrays. The reduced SNP sets were imputed back to the content of the Illumina 777K (HD) SNP array using the 600 individuals as reference population. The frequencies of imputing a given genotype depending on the real genotype were scored for MAF bins of 0.01 separately for the LD and 50,000 array data. We simulated genotyping-by-sequencing of tank milk as follows. For each of the 8 M SNP positions, we sampled local read depth ($r \in$ integers) from a Poisson distribution with mean C, where C is the average genome-wide coverage (0.25, 0.5, 1, 2, or 5). We then sampled r reads, each with a probability = BAF_j (computed as above) of being the B allele.

Scheme III

Individual SNP genotypes and tank B-allele frequencies (BAF_j) were generated as in scheme I (genotypes at 8 M SNP positions). It was assumed that milk tank was genotyped by SWGS at an average coverage of C (0.25, 0.5, 1, 2, or 5), and cows were genotyped by SWGS at average coverage of C (0.25, 0.5, or 1). Genotyping-by-sequencing of individual cows was simulated by (1) sampling, for each of 8 M SNP positions, local read depth ($r \in$ integers) from a Poisson distribution with mean C and (2) sampling r reads with probability 0, 0.5, or 1 to be the alternate allele (B) depending on the genotype of the cow (AA, AB, or BB). Genotyping-by-sequencing of the tank milk was performed as in scheme I.

Scheme IV

Scheme IV was identical to scheme III except that cow genotypes were generated at 8 M SNP position using a MAF-dependent and sequence depth–dependent imputation error model. The error model was constructed using available SWGS data down-sampled to 1× (176 cows) or 0.25× coverage (192 cows). The cows were imputed to HD (777,000 SNPs) using a reference population of 800 unrelated Holstein-Friesian individuals that were genotyped with the Illumina 777K array. At each of the 777,000 SNP positions, the likelihood of the sequence data under the three possible genotypes (AA, AB, and BB), was computed following the method of Chan et al. (2016), as

$$L(nr_A, nr_B|''\text{AA}'', \varepsilon) = \binom{nr_A + nr_B}{nr_B} \times (1-\varepsilon)^{nr_A} \times \varepsilon^{nr_B},$$

$$L(nr_A, nr_B|''\text{AB}'', \varepsilon) = \binom{nr_A + nr_B}{nr_B} \times 0.5^{(nr_A+nr_B)},$$

$$L(nr_A, nr_B|''\text{BB}'', \varepsilon) = \binom{nr_A + nr_B}{nr_B} \times (1-\varepsilon)^{nr_B} \times \varepsilon^{nr_A},$$

where nr_A (respectively, nr_B) is the number of A (respectively, B reads), and ε is the sequencing error rate set at 0.01. The corresponding $\log_{10} L$ was used as input for BEAGLE 4.0 (Browning and Browning 2009). Variant positions without sequence coverage in any of the 176 (192) cows (hence not imputed by BEAGLE 4.0) were dealt with in a second round of imputation using BEAGLE 5.0 (Browning et al. 2018). The imputation accuracy was evaluated in 0.01 MAF-bins by comparing imputed and real genotypes at the approximately 17,000 variant positions interrogated by the Illumina LD array.

Real data

Data set 1

We obtained a sample of tank milk from a farm in France milking 133 Holstein-Friesian cows. All had been genotyped with an Illumina LD array interrogating 17,000 SNPs using standard procedures. For all cows, genotypes were imputed to the whole genome using a reference population of 743 Holstein-Friesian animals sequenced at average depth of 15× (range: 4–48) and the BEAGLE software (v5.0) (Browning and Browning 2009), yielding allelic dosages for a total of 13 M SNPs. Individual milk records, including volume and SCCs (cells/mL) measured on the day of the sample collection, were obtained for all cows that had contributed milk to the tank. DNA was isolated from 1.5 mL tank milk using the NucleoMag kit (Macherey-Nagel). The tank milk DNA was first genotyped using the Illumina LD array interrogating 17,000 SNPs. An Illumina compatible NGS library was then prepared with 50 ng of genomic DNA using the KAPA HyperPlus kit (Roche). Sequencing was performed on a NextSeq 500 instrument (Illumina), yielding 63 million paired end reads of 2 × 75 bp, corresponding to a genome coverage of 3.5×. Reads were mapped to the bosTau8 genome build using BWA-MEM (Li 2013). Reference (R) and alternate (A) alleles were counted at 13 M SNP positions of the HD array using the Bam-ReadCount tool (https://github.com/genome/bam-readcount.git) for reads with a minimum mapping quality of 30.

Data set 2

We obtained samples of tank milk from a Belgian farm, including milk from 520 Holstein-Friesian cows. Milk volume and SCCs (cells/mL) measured on the same day were obtained for all cows that had contributed milk to the tank. All cows were genotyped with the Illumina LD array interrogating 17,000 SNPs using standard procedures and imputed to the whole genome using whole-genome sequence data (average depth: 15×; range: 4×–48×) from 743 Holstein-Friesian animals as a reference and the BEAGLE software (v5.0) (Browning and Browning 2009), yielding allelic dosages for a total of 13 M SNPs. DNA extraction from the tank milk samples and genotyping with the Illumina LD (17,000) array were conducted as for data set 1. For sequencing of the tank milk, an Illumina-compatible sequencing library was prepared using 12 ng of DNA and the Riptide high-throughput rapid library prep kit (iGenomx). The library was sequenced on an Illumina NextSeq 500 2 × 150 paired-end flow cell at 4× coverage.

Data set 3

We obtained samples of tank milk from a Belgian farm, including milk from 120 Holstein-Friesian cows. Milk volume and SCCs (cells/mL) measured on the same day were obtained for all cows that had contributed milk to the tank. All cows were genotyped with the Illumina LD array interrogating 17,000 SNPs using standard procedures, and imputed to the whole genome using whole-genome sequence data (average depth: 15×; range: 4–48) from 743 Holstein-Friesian animals as a reference and the BEAGLE software (v5.0) (Browning and Browning 2009), yielding allelic dosages for a total of 13 M SNPs. We additionally prepared an Illumina-compatible NGS library for each cow, using 12 ng of genomic DNA and the Riptide high-throughput rapid library prep kit (iGenomx). Libraries were sequenced on an Illumina NovaSeq S4 2150 paired-end flow cell at average 1.08× depth (range: 0.26–1.73). Cow genotype-by-sequencing data were imputed to HD (777,000) density using a reference population of 800 Holstein-Friesian animals genotyped with the bovine HD Illumina array (777,000 SNPs) and the BEAGLE software (v5.0) (Browning and Browning 2009), yielding allelic dosages for a total of 777,000 SNPs. DNA extraction from the tank milk samples, genotyping with the Illumina LD (17,000) array, and sequencing (coverage 4×) were conducted as for data sets 1 and 2.

Data set 4

In addition to obtaining a sample of tank milk on the day of the milk recording (i.e., yielding the SCC measured using with a cell counter) for the Belgian farm with 120 cows, we weekly collected an additional 11 tank milk samples before and 9 samples after, spanning a total period of ∼3 mo. The corresponding DNA samples were genotyped using the Illumina LD (17,000) array.

Statistical models

We defined a set of m linear equations of the form

$$\widehat{BAF_j} = \sum_{j=1}^{n} f_i \times d_{ij} + \varepsilon_j,$$

in which f_i is the proportion of the DNA in the tank milk contributed by cow i, d_{ij} is the "dosage" of the alternate allele A for cow i and marker j, and ε_j is the error term for marker j. When genotyping the tank milk with arrays, $\widehat{BAF_j}$ corresponds to the B-allele frequency estimated by Genome Studio (Illumina). When genotyping the tank milk by SWGS, $\widehat{BAF_j}$ corresponds to the proportion of A reads at the corresponding genome position. For cow genotypes obtained with arrays, d_{ij} corresponds to 0, 0.5, or 1 for genotypes AA, AB, and BB, respectively. For cow genotypes obtained by imputation, d_{ij} is the dosage of the B allele estimated by BEAGLE. For cow genotypes obtained by SWGS, $d_{ij} = 0.5 \times P(''AB''|nr_A, nr_B, q_j) + P(''BB''|nr_A, nr_B, q_j)$, where nr_A (respectively nr_B) is the number of A (respectively, B reads) for marker *j* and cow *i*, and q_j is the population frequency of the *B* allele of marker *j*.

$$P(''AB''|nr_A, nr_B, q_j) = \frac{2q_j(1-q_j) \times 0.5^{nr_A} \times 0.5^{nr_B} \times \dfrac{(nr_A+nr_B)!}{nr_A! \times nr_B!}}{(1-q_j)^2 \times 1^{nr_A} \times 0^{nr_B} + 2q_j(1-q_j) \times 0.5^{nr_A} \times 0.5^{nr_B} \times \dfrac{(nr_A+nr_B)!}{nr_A! \times nr_B!} + q_j^2 \times 0^{nr_A} \times 1^{nr_B}}$$

$$P(''BB''|nr_A, nr_B, q_j) = \frac{q_j^2 \times 0^{nr_A} \times 1^{nr_B}}{(1-q_j)^2 \times 1^{nr_A} \times 0^{nr_B} + 2q_j(1-q_j) \times 0.5^{nr_A} \times 0.5^{nr_B} \times \dfrac{(nr_A+nr_B)!}{nr_A! \times nr_B!} + q_j^2 \times 0^{nr_A} \times 1^{nr_B}}$$

For SNPs *j* without usable information for cow *i* (e.g., genotyping failure or no covering reads) d_{ij} was set at $\widehat{BAF_j}$.

The f_is were estimated by least square analysis, namely, by minimizing $\sum_{j=1}^{m} \varepsilon_j^2$. When the tank milk was genotyped by SWGS, we also performed a weighted least square analysis; namely, we estimated f_is by minimizing $\sum_{j=1}^{m} w_j \varepsilon_j^2$, where w_j is the coverage ($nr_A + nr_B$).

The SCC_is were calculated from the f_is

$$SCC_i = SCC_{tank} \times V_{tank} \times f_i / V_i,$$

where V_{tank} and V_i are the volumes of milk in the tank and contributed by cow *i*, respectively.

The accuracies of the predictions were measured by (1) the correlation (*r*) between real and estimated SCC_i, and/or (2) the ability to discriminate animals with SCCs above versus below a certain threshold value measured as $(T_P + T_N)/n$, where T_P stands for the number of true positives, T_N for the number of true negatives, and *n* for the total number of cows.

To test the effect of sequence depth on accuracy, we sampled reads overlapping SNP positions with probability *x*, such that $E(C \times x) = D$, where *D* is the desired sequence depth.

Data access

All sequence (FASTQ files) and genotype (VCF files) data used in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession number PRJEB38123/ERP121506 and to the European Variation Archive (EVA; https://www.ebi.ac.uk/eva/) under accession number PRJEB38336. Additional information to rerun the analyses are provided as Supplemental Table 2.

Competing interest statement

The proposed method is the subject of awarded (WO/2013/079289) and filed (PCT/EP2019/057628) patents.

Acknowledgments

We thank Jean-Bernard Davière, Pierre Lenormand, Bonny Van Ranst, Kristien Neyens, and Miel Hostens for providing the samples and information needed to conduct the experiments. This work was funded by the Unit of Animal Genomics and by the ERC DAMONA grant to M.G.

Author contributions: W.C. and M.G. designed experiments, analyzed data, and wrote the manuscript. L.K. performed experiments.

References

Atashi H, Salavati M, De Koster J, Ehrlich J, Crowe M, Opsomer G, GplusE consortium, Hostens M. 2019. Genome-wide association for milk production and lactation curve parameters in Holstein dairy cows. *J Anim Breed Genet* **137:** 292–304. doi:10.111/jbg.12442

Blard G, Zhang Z, Coppieters W, Georges M. 2012. Identifying cows with subclinical mastitis by bulk single nucleotide polymorphism genotyping of tank milk. *J Dairy Sci* **95:** 4109–4113. doi:10.3168/jds.2011-5178

Brooke-Taylor S, Dwyer K, Woodford K, Kost N. 2017. Systematic review of the gastrointestinal effects of A1 compared with A2 β-casein. *Adv Nutr* **8:** 739–748. doi:10.3945/an.116.013953

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84:** 210–223. doi:10.1016/j.ajhg.2009.01.005

Browning BL, Zhou Y, Browning SR. 2018. A one-penny imputed genome from next generation reference panels. *Am J Hum Genet* **103:** 338–348. doi:10.1016/j.ajhg.2018.07.015

Chan AW, Hamblin MT, Jannink J-L. 2016. Evaluating imputation algorithms for low depth genotyping-by-sequencing (GBS) data. *PLoS One* **11:** e0160733. doi:10.1371/journal.pone.0160733

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F, et al. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations comprising fertility in livestock. *Genome Res* **26:** 1333–1341. doi:10.1101/gr.207076.116

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46:** 858–865. doi:10.1038/ng.3034

Georges M, Charlier C, Hayes B. 2019. Harnessing genomic information for livestock improvement. *Nat Rev Genet* **20:** 135–156. doi:10.1038/s41576-018-0082-2

Hogeveen H, Huijps K, Lam TJGM. 2011. Economic aspects of mastitis: new developments. *N Z Vet J* **59:** 16–23. doi:10.1080/00480169.2011.547165

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997 [q-bio.GN].

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11:** 499–511. doi:10.1038/nrg2796

Schukken YH, Wilson DJ, Welcome F, Garrison-Tikofsky L, Gonzales RN. 2003. Monitoring udder health and milk quality using somatic cell counts. *Vet Res* **34:** 579–596. doi:10.1051/vetres:2003028

Viguier C, Arora S, Gilmartin N, Welbeck K, O'Kennedy R. 2009. Mastitis detection: current trends and future perspectives. *Trends Biotechnol* **27:** 486–493. doi:10.1016/j.tibtech.2009.05.004

Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. 2017. Genomic selection in dairy cattle: the USDA experience. *Annu Rev Anim Biosci* **5:** 309–327. doi:10.1146/annurev-animal-021815-111422

Received August 26, 2019; accepted in revised form June 19, 2020.