Identification of Mislabeled Samples and Sample Mix-ups in Genotype Data Using Barcode Genotypes

Christian Theil Have, Emil Vincent Appel, Niels Grarup, Torben Hansen, and Jette Bork-Jensen

Abstract—**Undetected mislabeled samples may affect the results of genotype studies, particular when rare genetic variants are investigated. Mislabeled samples are not often detected during quality control and if they are detected, they are normally discarded due to a lack of a reliable method to recover the correct labels.**

Here we describe a statistical method which given a few extra independent genotypes (barcode genotypes) detects mislabeled samples and recovers the correct labels for sample mix-ups. We have implemented the method in a program (named Wunderbar) and we evaluate the reliability of the method on simulated data. We find that even with only a small number of barcode genotypes, Wunderbar is capable of identifying mislabeled samples and sample mix-ups with high sensitivity and specificity, even with a high genotyping error rate and even in the presence of dependency between the individual barcode genotypes.

To detect mislabeled samples, we calculate the probability that the discordance between genotypes in the data and in the independent genotypes can be attributed to random (non-mislabeling) genotyping errors. To identify mix-ups we calculate the probability of identifying the set of identical genotypes between sample x and sample y by chance. Based on this we calculate a mix-up confidence score with penalization for introducing mismatches in the proposed new label and adjustment for independency among the genotypes. This confidence score is used to identify probable mix-ups.

Index Terms—**Barcoding, genetics, quality control.**

I. Introduction

While absolutely essential, it may be challenging to ensure that genotypes are correctly coupled with the sample IDs when obtaining genotype data for a large number of samples. Often these data are in the form of single nucleotide polymorphism (SNP) arrays, but the concern about sample label quality control transcends the type of genotyping technology.

In the process of genotyping samples with SNP arrays there are a number of ways in which mislabeling may occur; Mislabeling or swapping of blood sample vials, pipetting errors, masterplate orientation, and database entry blunders, etc.

In June 2010, the personal genomics company 23andMe sent out the wrong genotypes to 96 people [1]. This caused major frustration among customers, some of which ended up with doubts about biological parenthood. The culprit was a flipped 96 well masterplate containing their DNA samples. Because of the severe consequences, a mix-up like this is not likely to go unnoticed in personal genomics. In academic studies, however, such errors could easily happen without being discovered, and if left unnoticed can lead to publication of false results. To avoid mislabeling, the genotype data must be carefully examined and compared against available phenotypic information. For instance, routine quality control often includes gender verification by comparing the genetic gender with the recorded phenotypic gender of the sample and a verification of the known family relations between samples [2]. Most datasets contain only unrelated samples in which case the similarity between the genotypes of family members cannot be used to distinguish samples. Furthermore, the gender check does not identify sample mismatches between samples with identical gender, and it cannot be used to detect which samples are swapped if there is more than one mismatch.

A much more reliable technique is to genotype a small subset of the genetic variants included in the array in the same samples using an independent and more economic platform. These genotypes serve as a molecular barcode that can be used to identify the sample. If the sample is correctly labeled, then the barcode genotypes will agree to the sample genotypes thereby minimizing possible genotype errors. The term *genetic barcoding* and the idea underlying it are well known within forensic genetics [3] and in species identification [4].

With a sufficient number of genotypic variants serving as genetic barcode, it is possible to detect the correct labels for mislabeled or swapped DNA samples. The number, quality and independence of the genotypes serving as barcodes are important. Sometimes, a few such genotypes just happen to be available from earlier studies of the same samples, but may be obtained using a different and possibly error-prone, and outdated technology.

We here present a statistical method which is capable of identifying mislabeled samples and mix-ups with high reliability even when a limited amount of suboptimal quality barcoding genotypes are available, and even if they are in partial linkage disequibrilium (LD). We have implemented this method in the program *Wunderbar*, which is freely available (GPL license). The program is compatible with the widely used genotype analysis tool Plink [5] and uses the Plink format for both array genotypes and barcoding genotypes. It is written in the Ruby programming language which runs on a variety of operating systems.

Several tools [6], [7] exist to identify mix-ups when GWAS data is accompanied by expression data. However, we are not aware of any other tools which are designed to facilitate the process of examining the concordance between barcode and array genotype data.

Manuscript received March 9, 2014; revised May 13, 2014. These authors contributed equally to this manuscript.

The authors are with the Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Copenhagen University, Denmark (e-mail: c.have@sund.ku.dk, vincent@sund.ku.dk, ngrp@sund.ku.dk, torben.hansen@sund.ku.dk, jbj@sund.ku.dk).

II. Approach

Our approach separately deals with the problem of detecting *single* mislabeled samples and the more complex problem of detecting mix-ups, e.g., samples that have been swapped. The former problem is a logical precedent of the latter problem, *e.g*., if two samples are swapped then they must be also mislabeled separately.

A. Testing for Mislabeled Samples

To test for mismatches, we compare the barcoding genotypes with array genotypes for each labeled individual in a dataset containing *n* samples.

When we identify discordance between barcode and array genotypes in an individual, we would like to know the probability that this mismatch is likely to happen due to random genotyping errors, as opposed to being a mislabeling. To reflect this, we introduce a *p*-value, which is a probability that the discordance can be attributed to random (non-mislabeling) genotyping errors. When the *p*-value is *less than* α, we consider the discordance to be caused by mislabeling. The higher the discordance between the barcode and array genotypes is in an individual, the less likely it is that the discordances can be attributed to chance. This is reflected in the *p*-value, and thus for a significant low value of α only individuals with a high degree of discordance between the barcode and genotype arrays is identified as a mislabeling.

To calculate the *p*-value, we first make the assumption that the frequency of array versus barcode mismatches for each SNP is estimates of the probabilities that a mismatch on that particular SNP occurs by chance. This assumption relies on the number of mismatches to be small relative to the total number of samples. We treat these estimates as independent probabilities and calculate the probability of a particular discordance in an individual as the product of the mismatch probabilities for each mismatched SNP (see equation 1).

We expect the array genotypes to include more SNPs than the barcode genotypes, but for barcoding purposes we only need to consider the set of SNPs that are in common between the array genotypes and the barcoding genotypes.

We use the notation, G^{array}, to describe the array genotypes. G^{array} is an $n \times m$ matrix where n is the number of individuals and m is the number of SNPs which present both on the array and in the barcode. The genotype of a particular individual i on a particular SNP j is given by

$$G_{ij}^{array},$$

where $0 < i < n$ and $0 < j < m$.

For each SNP the genotype is an integer in the set [0, 1, 2, 3] where 0 means that the genotype is unknown, 1 means homozygous wildtype, 2 means heterozygous and 3 means homozygous for the variant. We define the barcoding genotype matrix $G^{barcode}$ as the same n individuals and the same m SNPs, such that G^{array} and $G^{barcode}$ include the same set and order of SNPs and individuals.

L is a missingness matrix with the same dimension and order as G^{array} and $G^{barcode}$. The element in position i, j – denoted L_{ij} – is 1 whenever $G_{ij}^{sample} \neq 0$ $G_{ij}^{barcode} \neq 0$, i.e., when SNP j in individual i is present in both G^{array} and $G^{barcode}$.

The match matrix M is used to indicate whether samples have identical genotypes in both G^{array} and $G^{barcode}$. The element in position i, j (M_{ij}) of the match matrix M is 1 whenever $G_{ij}^{array} = G_{ij}^{barcode}$ and 0 otherwise. Oppositely, in the mismatch matrix M', M'_{ij} is 1 whenever $G_{ij}^{array} \neq G_{ij}^{barcode}$ and 0 otherwise.

The probability that a particular set of mismatches occur in a sample is calculated as the product of the relative frequencies of mismatching SNPs as detailed in equation 1 below:

$$P(mismatch_i) = \prod_{j=1}^{m} \left(M'_{ij} \frac{\sum_{k=1}^{n} M'_{kj}}{\sum_{k=1}^{n} L_{kj}} + M_{ij} \right) \qquad (1)$$

Note that this equation is conservative in the sense that if there are mislabeled samples, their mismatched SNPs will contribute towards increasing this probability and it is therefore expected to be a slight overestimate of the *real* chance probability.

Besides this *point probability* we are interested in the probability of getting as bad a mismatch as the one observed for a particular individual by chance, i.e, assuming that no mislabeling has occurred. This probability, which we denote $p_i^{mislabel}$, is a *p*-value which serves as a statistical test where the null hypothesis is to observe a particular set of mismatched genotypes *as extreme* as the ones observed in the data *by chance* in a correctly labelled sample. The point probability described in equation 1 is a measure of the chance mismatch extremity and $p_i^{mislabel}$ is hence the probability of seeing a probability as low as $P(mismatch_i)$.

To calculate $p_i^{mislabel}$, we use a random sampling procedure. In brief, the procedure samples a large number of random *P(mismatch)* probabilities according to the SNP mismatch frequencies observed in G^{array}. $p_i^{mislabel}$ is then equal to the lowest percentile of sampled probabilities to which $P(mismatch_i)$ belongs.

We reject the null hypothesis (chance mismatch) when $p_i^{mislabel} < \alpha$, for a sufficiently small α, and as a consequence, individual i is categorized as mislabeled. To select a reasonable value for α, one should consider the level of sensitivity that is required to detect mislabeled samples. An α value of 0.05 is expected to detect 19/20 mislabeled samples.

B. Testing for Mix-ups

A mix-up is when a mislabeled sample which is labeled as x is actually another sample y. Mix-ups can for example be in the form of swapped samples, where x is labeled as y and vice versa. We use the notation $x \rightarrow y$ to represent a mix-up, i.e., the case where the genotype of sample x corresponds to the genotypes of the sample labeled y in the barcoding data. In the following we present a method to identify mix-ups deemed statistically unlikely to occur by chance-similarities between the genotypes of a pair of samples. Compared to our method

for identifying mislabels, we here use a score instead of a p-value. The score incorporates the p-value for finding a mix-up by chance and a penalization for introducing mismatches in the proposed new label. Furthermore, we also incorporate a LD adjustment when calculating the p-values for mix-ups.

For each SNP *j* we calculate the relative frequencies of each genotype *g*:

$$R_j(G,g) = \frac{\sum_{i=1}^{n} H_{ij}(g)}{\sum_{g' \in \langle 1,2,3 \rangle} \sum_{i=1}^{n} H_{ij}(G,g')} \quad (2)$$

H is defined as a function of a genotype *g* and a position index *i, j*:

$$H_{ij}(G,g) = \begin{cases} 1 & if \quad G_{ij} = g \\ 0 & otherwise \end{cases} \quad (3)$$

For each pair of samples, *x* and *y*, we calculate the probability that the set of identical overlapping genotypes between sample *x* and sample *y* would occur by chance (the overlap probability). This is calculated as the product of genotype probabilities for all *matching* genotypes:

$$P_{overlap}(x \to y) = \prod_{j=1}^{m} F_j^{x \to y} \quad (4)$$

where,

$$F_j^{x \to y} = \begin{cases} R_j(G^{barocde}, G_{xj}^{array}) & if \quad G_{xj}^{array} = G_{yj}^{barcode} \neq 0 \\ 1 & otherwise \end{cases} \quad (5)$$

The equation used to calculate the probability assumes independence between the genotype frequencies of the SNPs. This may lead to an inflation when the independence assumption is incorrect, such as if the SNPs are in LD, or if the samples are related.

To correct for inflation, we calculate an inflation factor by means of a sampling procedure. Random subsets of the barcoding SNPs are repeatedly sampled along with genotypes from the array genotypes and it is measured how often the random subset genotypes match the sample in question *(x)*. The $P_{overlap}(x \to y)$ probability is then multiplied with the inflation factor to give a more realistic measure. The sampling algorithm is shown in pseudo-code below:

PROC LD-RATIO:
  LET *sumOfRatios* := 0
  Ω TIMES DO
    LET *numSnps* := random integer in interval [1-*m*]
    LET *snps* := *numSnps* random snps in {1,…,*m*}
    LET n_{count} := 0
    FOR *i* in 1 to *n*
      IF $\forall_{j \in snps} G_{xj}^{barcode} = G_{ij}^{array}$ THEN
        n_{count} := n_{count} + 1
      END IF
    END FOR
    LET $n_{expected} := \prod_{j \in snps} R(G^{array}, geno_j) \times n$,
    LET *sumOfRatios* := *sum_of_ratios* + $n_{count} / n_{expected}$
  RETURN $sumOfRatios / \Omega$

The procedure returns the mean of a series of ratios between observed random genotype subsets and expected counts of those genotype subsets in G^{array}. The expected number of occurrences $n_{expected}$ of a set of genotypes is the product of the genotype frequencies in $G^{array} \times n$.

The loop of procedure runs Ω times, which is a user adjustable parameter to the program. Increasing the number of loop iterations increases the precision of the ratio estimate, but it also significantly adds to the execution time, since this procedure has run for every potential mix-up, i.e., on the order of $n \times n$ times.

The overlap probability measure does not account for the fact that when we are allowed for arbitrary mismatches, then there is an exponential number of possible ways that one individual's barcode genotypes may overlap (partially match) with another individuals array genotypes. $P_{overlap}(x \to y)$ is the concrete probability for one of these exponentially many possible overlaps. To adjust this we apply a mean-field approximation, which consists of multiplying the LD-adjusted overlap probability by the number of possible ways that we can match two samples with at least as many overlapping/identical genotypes, e.g., allowing *k* mismatches, $1 \le k \le m$, then the number of possible ways to partially match two samples is expressed as the binomial series, $\sum_{1}^{k} \binom{m}{k}$. Intuitively, this may also be seen as a penalty on the score which reflects the negative evidence of having a number of mismatching genotypes.

Because of the nature of the mean-field approximation, the result is no longer a probability measure. Instead, we refer to it as the *mismatch-adjusted overlap measure*. The mean-field is a good approximation when *m* is small (e.g., less than 64), but it becomes increasingly inaccurate when *m* is large. Hence, using too many barcoding SNPs will negatively affect accuracy.

The mismatch-adjusted overlap measure is not in itself useful to indicate whether the overlap is due to chance or caused by a mix-up. To investigate this, we calculate yet another *p*-value, which indicates the probability of seeing a mismatch-adjusted overlap measure as extreme as or more extreme than a given one by chance. To calculate this *p*-value we employ a sampling procedure. Random "barcode" genotypes which mimic the genotype frequencies in the sample genotypes are created and matched to all array genotypes. For each match we record the *mismatch-adjusted overlap measure* of the match. With a large amount of such probabilities we are equipped to ask the question of which percentile a particular *mismatch-adjusted overlap* measure belongs too. The *p*-value – which is denoted $p_{x \to y}$ – is then the lowest percentile to which a *mismatch-adjusted overlap*

measure for the match $x \rightarrow y$ belongs.

Finally, we score and rank matches according to a combined probability measure which for each pair of samples takes into account the probability of each of the of samples being mislabeled as well as the probability that they are swapped:

$$score_{x \rightarrow y} = (1 - p_{x \rightarrow y}) \times (1 - p_x^{mislabel}) \times (1 - p_y^{mislabel}) \quad (6)$$

This is an approximate confidence score and by applying a cut-off on the score, we can classify mix-ups. The user can select an appropriate cut-off, e.g., 0.95%, which corresponds to a particular specificity/sensitivity trade-off. However, the score is an independent measure of the probability of a mix-up and classifications based on this measure may warrant further investigations. For instance, with a given cut-off value there may be several probable mix-ups with the same label.

III. Evaluation

We evaluate the Wunderbar program with regard to its ability to detect mislabeled samples as well as its ability to detect swaps. We perform the evaluation for a varying number of mislabeled/swapped samples and for a varying number of barcoding SNPs. Furthermore, we evaluate the effect of the LD-adjustment procedure.

In our experiments, we simulate a chip array of SNPs for 1000 individuals in the first two experiments and 100 individuals in the last experiment. For each individual, a SNP genotype is picked randomly as either homozygote wildtype, heterozygote, homozygote derived allele or as missing data. The chance of a missing genotype is 12.5% whereas each of the three possible alleles are equiprobable (~ 29.17%).

The barcoding array is created as an identical copy of the chip array. In this barcoding array we then simulate random mismatches with regard to the original chip array genotypes. This has been done in order to mimic realistic data which may have random genotyping errors. The chance of a genotype being changed to a mismatching genotype in the barcoding array is fixed at 1% in all experiments. For a particular mismatching genotype, there is a 16.67% chance that it will be changed to a missing genotype call and 41.67% chance that it will be substituted with either of the two alternative genotypes.

To simulate mix-ups we swap the genotypes of two samples on chip array, but not in the barcoding array.

The number of sampling iterations used to calculate the p-value was set to 10000 in all experiments. LD adjustment is not used in the first two experiments (Section III.A and III.B) and genotypes for different SNPs are generated independently of each other. In the final experiment in which we test the LD adjustment procedure, we simulate a chip array of SNPs in high LD (R^2 = 90%) and report the difference in accuracy with and without LD adjustment (Section III.C).

We use receiver operating characteristic (ROC) curves to report the results. We generate ROC curves by measuring the true and false positive rates for the top-*n* candidates produced by Wunderbar for all possible values of *n*.

A. Varying the Number of Barcoding SNPs

In this experiment we deliberately mislabel eight samples. Then we use from one and up to six of the SNPs in the barcoding array to attempt to identify the mislabeled samples.

The ROC curve in Fig. 1 displays the accuracy for identifying mislabeled samples and the ROC curve in Fig. 2 displays the accuracy for identifying mix-ups.

From the ROC curves in both Fig. 1 and Fig. 2, it is apparent that for more than three SNPs, Wunderbar has prediction power to identify the mislabeled samples as well as the mix-ups with almost perfect accuracy for the chosen number of samples (1000).

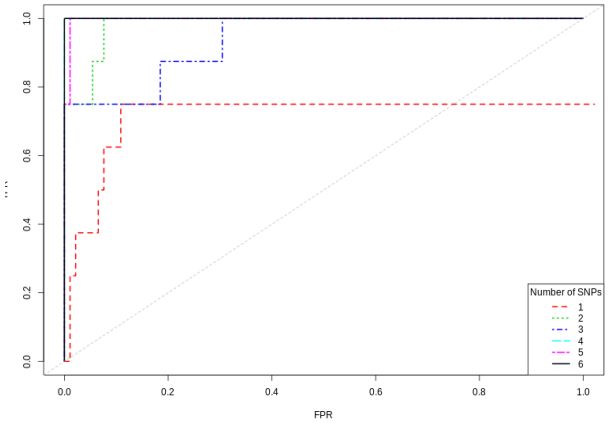

Fig. 1. ROC curve that displays the accuracy for identifying mislabeled samples with varying number of SNPs.

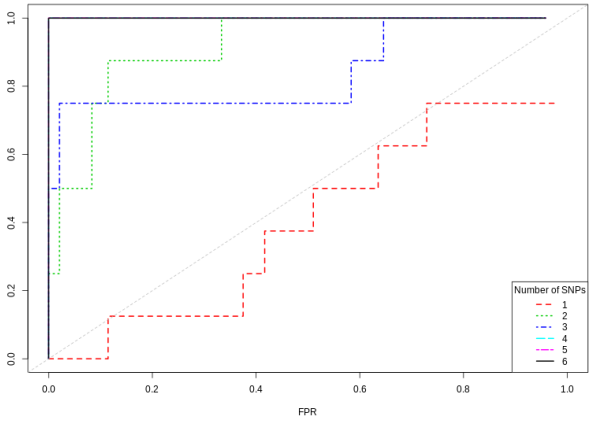

Fig. 2. ROC curve that displays the accuracy for identifying mix-ups with varying number of SNPs.

B. Varying the Number of Swaps

As a second experiment we keep the number of SNPs constant. We use five SNPs, which in the previous experiment is demonstrated to have sufficient predictive power. Then we vary the number of mislabeled samples in the dataset. All mislabeled samples in the data are in the form of pairs of swapped samples.

The ROC curve in Fig. 3 shows the accuracy for identification of mislabeled samples and the ROC curve in Fig. 4 shows the accuracy for identification of mix-ups. Even with a high percentage of mislabeled samples, Wunderbar reliably detects the samples that have been mislabeled. It also shows from Fig. 4, that identifying swaps is a more difficult problem and with 64 swaps (more that 5% of all samples) detection is not sufficiently accurate to be blindly trusted when only using

five SNPs. However, even with 64 swaps and only five SNPs the achieved level of accuracy shows that the confidence scores provided by the program are still very informative. Coupled with further knowledge (more SNPs or other sources of information), the confidence scores could still help to resolve mix-ups.

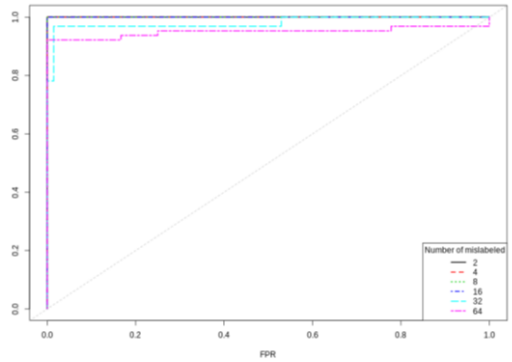

Fig. 3. ROC curve that displays the accuracy for identifying mislabeled samples with varying number of swaps.

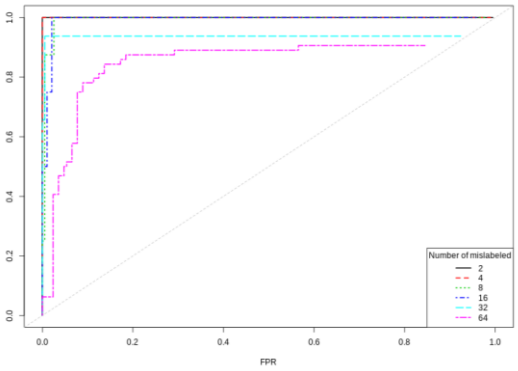

Fig. 4. ROC curve that displays the accuracy for identifying mix-ups with varying number of swaps.

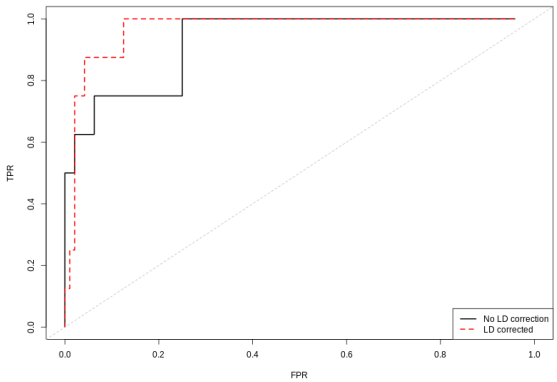

Fig. 5. ROC curve that displays the accuracy for identifying mix-ups with and without LD-correction. The ROC curve was produced for a dataset with four swaps, five barcoding SNPs, 100 samples. The LD-adjustment was run for 50 iterations for each pair of samples.

C. Experiment 3: Effect of LD-Adjustment

In the two previous experiments, we have been using SNPs that were independent. This might not be the case with actual barcoding SNPs, which may be left-overs from a previous study, perhaps of a narrow region in the genome. In such a case, the SNPs may be in LD. Here, we simulate a chip array of SNPs in high LD (R^2 = 90%) and we test how this affects the accuracy of Wunderbar with and without the LD-adjustment procedure.

Fig. 5 shows a ROC curve of the accuracy for detecting mix-ups with and without the LD-adjustment procedure. It can be observed that LD-adjustment procedure improves accuracy significantly.

IV. Conclusion

While Wunderbar is useful to detect potential sample mix-ups, it is difficult to determine where the mix-up has occurred. Barcoding genotypes only help to identify mix-ups that occur at the DNA sample level, but do not help to identify mix-ups that occur in phenotype registration. Furthermore, even at the DNA sample level, the approach does not reveal whether the mix-up occurred in the G^{array} or $G^{barcode}$. To reveal this, additional information is required such as phenotype information and its relation to the genotypes, the genotypic gender versus the reported gender and the verification of reported familiar relations between samples compared to the levels of shared genotypes. In addition, knowledge about the genotyping technology and the protocol for obtaining genotypes can provide useful pointers to resolve this issue. For instance, genotypes could be swapped due to flipped masterplates, as was the case for the 23andMe episode. Future versions of Wunderbar may try to incorporate and take advantage of such information. In particular knowledge of sex, and easily obtained phenotypes like eye and hair color [8] as well as blood type [9] and similar phenotypic information that have direct parallels on the genotypic level would be useful to incorporate in the model.

Wunderbar is available for download from: http://cth.github.io/wunderbar/

References

[1] D. MacArthur. (2014). Sample swaps at 23andMe: a cautionary tale. [Online]. Available: http://scienceblogs.com/geneticfuture/2010/06/07/sample-swaps-at-23andme-a-caut/.

[2] S. Turner, L. L. Armstrong, Y. Bradford, C. S. Carlson *et al.*, "Quality control procedures for genome-wide association studies," *Curr. Protoc. Hum. Genet.*, vol. 1, Jan. 2011.

[3] M. A. Jobling and P. Gill, "Encoded evidence: DNA in forensic analysis," *Nat. Rev. Genet.*, vol. 5, no. 10, pp. 739-751, Oct. 2004.

[4] P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. de Waard, "Biological identifications through DNA barcodes," *Proc. Biol. Sci.*, vol. 270, no. 1512, pp. 313-321, Feb. 2003.

[5] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559-575, Sept. 2007.

[6] E. E. Schadt, S. Woo, and K. Hao, "Bayesian method to predict individual SNP genotypes from gene expression data," *Nat. Genet.*, vol. 44, no. 5, pp. 603-608, May 2012.

[7] H. J. Westra, R. C. Jansen, R. S. Fehrmann, G. J. T. Meerman *et al.*, "MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects," *Bioinformatics.*, vol. 27, no. 15, pp. 2104-2111, Aug. 2011.

[8] P. Sulem, D. F. Gudbjartsson, S. N. Stacey, A. Helgason *et al.*, "Genetic determinants of hair, eye and skin pigmentation in Europeans," *Nat. Genet.*, vol. 39, no. 12, pp. 1443-1452, Dec. 2007.

[9] S. P. Yip, "Sequence variation at the human ABO locus," *Ann. Hum. Genet.*, vol. 66, pp. 1-27, Jan. 2002.

Christian Theil Have is a post doc at the Section for Metabolic Genetics in the Novo Nordisk Foundation Center for Basic Metabolic Research at Copenhagen University, Denmark. He received his MSc in computer science from IT University of Copenhagen, Denmark and PhD in computer science and bioinformatics from Roskilde University, Denmark.

Emil Vincent R. Appel is a scientific research assistant at the Section for Metabolic Genetics in the Novo Nordisk Foundation Center for basic metabolic research at Copenhagen University, Denmark. He got his M.Sc in bioinformatics from Copenhagen University, Denmark. He is an UCEAP student and Fulbright grantee at UC Berkeley 2012/13, USA.

Niels Grarup is an assistant professor at the Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Copenhagen University, Copenhagen, Denmark.

Torben Hansen is a professor at the Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Copenhagen University, Copenhagen, Denmark.

Jette Bork-Jensen is a post-doc at the Section for Metabolic Genetics in the Novo Nordisk Foundation Center for basic metabolic research at Copenhagen University, Denmark. She received her M.Sc in bioinformatics from Copenhagen University, Denmark and the Ph.D in health science from Copenhagen University, Denmark.