Master Thesis

DNA QUALITY CONTROL METRICS FOR SAMPLES USED FOR GENOTYPING ARRAYS

An analysis of 34,626 whole blood genomic DNA samples genotyped at the Estonian Genome Center

Submitted by

Steven Smit

For the academic degree of

Master of Science in Biobanking (MSc)

at the

Medical University of Graz

Executed at

Biobank Graz

Under the supervision of

Professor Berthold Huppertz

Submitted 2019

Tartu, 11.07.2019

Statutory Declaration

I declare on my honor that I have written this dissertation independently and without assistance, that no sources other than those cited were used and that the sources used verbatim or in substance have been marked as such.

Tartu, 11.07.2019 *eh Steven Smit*

Acknowledgements

First and foremost, I would like to thank professor Berthold Huppertz, my supervisor, for endless patience during all the delays with this thesis and for all the encouragement, assistance and advice during the writing!

I would also like to thank professor Andres Metspalu for offering me the opportunity to take part in this Master's course and for continuously urging me to finalize this thesis.

Further thanks belong to Kristi Läll, PhD for advice and consultations on statistical methods, Reedik Mägi, PhD for consultations and information on genotyping data quality control methods and Gabriele Hartl, Mag.(FH) for technical advice and assistance during the Master’s course.

Finally, I would like to thank my family and co-workers for the understanding and patience when I was busy with the Master's course.

Zusammenfassung

Hintergrund

Die Qualitätskontrolle von Biobankproben spielt eine wichtige Rolle bei der Sicherstellung von hochqualitativen Daten bei nachfolgenden Analysen. Gleichzeitig müssen die Kriterien der Qualitätskontrolle für die entsprechende Situation passend sein. Eine kosteneffektive Lösung, um Kriterien für die Qualitätskontrolle zu entwickeln und zu verifizieren, ist die retrospektive Analyse bereits vorhandener Daten in einer Biobank. Diese Studie hatte zum Ziel, diesen Ansatz umzusetzen und die Kriterien für die Qualitätskontrolle für DNA Proben zu verbessern, die für Microarray-basiertes Genotyping eingesetzt wurden.

Methoden

Die Ergebnisse der Qualitätskontrolle nach dem Genotyping von 34.626 Proben wurden kombiniert mit bereits vorliegenden spektrophotometrischen Daten und Daten des Probenhandlings der Biobank. Hinzu kamen Daten von retrospektiv durchgeführten Agarosegelelektrophoresen, um die relevantesten Kriterien für die Qualitätskontrolle zu identifizieren. Diese Kriterien sollten dazu genutzt werden, die Ergebnisse der Daten der Qualitätskontrolle aus dem Genotyping vorherzusagen.

Ergebnisse

Diese Analyse hat ergeben, dass es eine signifikante Verbindung zwischen dem Scheitern einer Probe bei der Qualitätskontrolle des Genotypings und der Degradation der Proben-DNA im Agarosegel gibt. Spektrophotometrisch gemessene A260/A230 Verhältnisse waren ebenso vorhersagend für potentielle DNA Degradationsprobleme. Ein signifikanter Anstieg der DNA Konzentration und ein erhöhtes A260/A280 Verhältnis war für Proben potentiell vorhersagend für eine DNA Degradation in spezifischen Proben. Ein weiterer Effekt, der mit dem Scheitern der Qualitätskontrolle in Verbindung gebracht werden konnte, bezog sich auf die Position einer Probe in den Lagerplatten und den Genotyping Arrayplatten.

Diskussion

Die Verwendung Proben-bezogener Daten in einer Biobank in Verbindung mit einer schnellen und preiswerten Testung einer Subgruppe von Proben erlaubt es, Qualitätsprobleme von Proben während des Handlings und Prozessierens zu identifizieren, bevor eine Analyse gemacht wird. Der Effekt der Probenposition in den Platten könnte mit einer Evaporation von Proben an bestimmten Positionen in Verbindung gebracht werden. Eine bessere Versiegelung der Platten während des Prozessierens könnte diesen Effekt reduzieren.

Fazit

Die retrospektive Qualitätskontrolle in Verbindung mit der Analyse bereits vorliegender Probendaten zeigte sich als eine wertvolle Methode, um die Kriterien für die Qualitätskontrolle von DNA Proben zu verbessern.

Abstract

Background

Quality control of samples in biobanks has a critical role in ensuring high quality data that is generated in downstream analyses. At the same time, quality control criteria need to be suitable for the specific situation. A cost-effective solution to creating and verifying quality control criteria would be to retrospectively analyse pre-existing data in biobanks. This study aimed at applying this approach to improve quality control criteria for genomic DNA samples used for microarray-based genotyping.

Methods

The results of post-genotyping data quality control of 34,626 samples were combined with pre-existing spectrophotometric and sample-processing related quality control data from the biobank, as well as retrospectively generated agarose gel electrophoresis results to identify most relevant quality control criteria that could have predicted the genotyping data quality control results of different samples.

Results

This analysis revealed a significant link between genotyping data quality control failure and the degradation of DNA samples observed in agarose gel electrophoresis. Spectrophotometrically measured A260/A230 ratios were also shown to be predictive of potential DNA degradation issues. Significant increases in DNA concentrations and A260/A280 ratios for samples were shown to be potentially predictive of DNA degradation in specific samples. An additional effect of sample positions on sample release plates and genotyping arrays on genotyping failure rates was identified.

Discussion

Using sample related data in a biobank plus some quick and inexpensive testing of a subgroup of samples allowed to identify quality issues during handling and processing of samples prior to analysis. The effect of sample positions could potentially be linked to sample evaporation from sample release plates and better sealing of sample release plates during processing could possibly be used to reduce this effect.

Conclusion

In conclusion, the retrospective quality control and analysis of pre-existing data was shown to be a viable method in improving quality control criteria for DNA samples.

Table of Contents

Glossary and Abbreviations

A230 – absorbance of sample at the 230 nm wavelength

A260 – absorbance of sample at the 260 nm wavelength

A280 – absorbance of sample at the 280 nm wavelength

Call-rate – a quality control criterion used in analysing genotyped data, defined as the ratio of successfully determined SNPs (single-nucleotide polymorphisms) to all SNPs present on the genotyping array. Typically, a threshold of 95% is used for sample exclusion.

GWAS – genome-wide association study

kb – kilobase, 1000 base pairs, a measure of nucleic acid molecule length

K_2EDTA - dipotassium ethylenediaminetetraacetic acid, an anticoagulant used in blood-drawing tubes

QC – quality control

SNP – single nucleotide polymorphism

SOP – standard operating procedure

TBE buffer – a buffer solution containing Tris (tris(hydroxymethyl)aminomethane), boric acid and EDTA (ethylenediaminetetraacetic acid) used to keep DNA water soluble and to inhibit nuclease activity

TE buffer – a buffer solution containing Tris (tris(hydroxymethyl)aminomethane) and EDTA (ethylenediaminetetraacetic acid) used to solubilize DNA and inhibit nuclease activity

List of figures

List of tables

1 Introduction

1.1 Background

Precision medicine (in some contexts, the term "personalized medicine" is also used interchangeably) is a growing field that has high expectations from both the healthcare system and patients regarding better disease treatment and prevention as well as lowered healthcare costs. Precision medicine is predicted to have a disruptive effect on the healthcare system and shift the healthcare strategies from intervention and disease management towards a more proactive management of disease risks and prevention (1).

While the development and application of precision medicine encompasses a very wide range of technologies and information, genomics and the use of genetic information obtained through sequencing or genotyping is among the most important parts of that range, either taken alone as in GWAS (genome-wide associaton studies) or as part of wider omics-based approaches. In addition to potential applications as part of precision medicine, the use of (multi)omics data (including genomic data) offers an enormous potential for a better understanding of the molecular mechanisms, processes and pathways discriminating health and disease (2).

While next-generation sequencing methods have significant advantages over microarray-based genotyping, the genome-wide microarray approach to generate genomic data is currently still around 1-2 orders of magnitude more economical when compared to whole exome sequencing or whole genome sequencing, making it the preferable method particularly in the case of large cohorts (2,3). This allows for the expectation that microarray genotyping will be performed on an increasing number of samples in the future, also increasing the importance of having well-defined quality control criteria that could be applied to the DNA samples used for such analyses, both in a biobanking and research context as well as in the healthcare system. In biobanking, this importance of well-defined quality control criteria also falls into the wider issue of a need for more

harmonization and standardization in order to provide high-quality samples and data to researchers (4).

Since the specific quality criteria for samples are often highly dependent on the downstream analyses, the requirements of the analyses that are expected to be performed on the samples would need to be taken into account. However, specific independent guidelines are hard to come by (3). As an example, the company's user guide for Infinium High-Throughput Screening genotyping arrays from Illumina only suggests that for the best performance, a minimum DNA concentration of 50 ng/µl is recommended (5). This lack of information likely leads service providers to rely on trial and error or anecdotal evidence, as using large numbers of samples to validate the suitable quality ranges beforehand would be expensive. In some cases, this could also lead to overly strict demands on samples, limiting the use of samples that could possibly provide useful results to researchers.

The quality control criteria most often applied for genomic DNA samples include (a) the assessement of DNA concentration, either specrophotometrically or fluorometrically; (b) the assessment of DNA purity by using spectrophotometrical absorbance ratios; (c) the electrophoretic assessment of DNA integrity, either qualitatively or quantitatively and (d) the performance of the sample in PCR amplification reactions (6–10). With the possible exception of the PCR amplification reactions where a binary classification (success/failure) could potentially be applied, the quality control criteria require interpretation using pre-existing guidelines regarding the suitable ranges of quality control results that the samples need to adhere to, in order to classify as suitable for use.

This leads to the implication that quality control criteria and the guidelines for their interpretation need to be verified separately for different types of samples and analyses in order to achieve the best performance. However, this is an inherently expensive process, especially in the case of biobanks where a wide range of potential downstream applications for the samples need to be considered and would thus require a large number proactive testing, which is likely not economically possible for biobanks.

In order to fulfill this need for verification of quality control criteria, a potential cost-effective solution could be to retrospectively use pre-existing data available in biobanks regarding already analysed samples, the quality or success rate of those analyses and the quality control results for the samples available at the biobank. This could enable the biobanks to improve the quality control criteria and their interpretation over time, leading to better guidelines regarding sample quality and suitability for different analyses. Publication of any such results could enable other biobanks, researchers and service providers to employ better quality control criteria in similar circumstances, as well as potentially improve the criteria further using data available to them.

This Master's thesis attemps such an approach to improve quality control criteria for genomic DNA samples used for microarray-based genotyping beyond the manufacturer specified minimum DNA concentration requirement.

1.2 Goal

The goal of this Master's thesis is to identify the most relevant quality control criteria for human genomic DNA samples used in whole-genome genotyping with Illumina BeadChip genotyping arrays. For this purpose, samples and quality control data of 34,626 participants of the biobank of the Estonian Genome Center (Institute of Genomics, University of Tartu, Estonia) were used without personally identifiable information, descriptions of DNA or states of health. The Estonian Genome Center has whole-genome genotyped these samples using Illumina Global Screening Array beadchips in 2017. Due to strict project timelines, only limited quality control was performed before the genotyping.

All genotyping data has been analysed in the post-genotyping quality control pipeline and the samples as well as biobank logs are available for further analysis. This provides a unique opportunity to use this large dataset to retrospectively identify relevant quality control criteria that could have been used to improve the genotyping success rate in this project and could potentially be applied to any future genotyping experiments.

1.3 Hypothesis

The hypothesis of this Master's thesis is that it is possible to retrospectively identify quality control measures that could have been applied to samples used in this thesis before genotyping in order to reduce genotyping failures, and that these quality control measures could be applied in future genotyping projects to achieve higher success rates.

2 Materials and Methods

2.1 Samples and processing

2.1.1 Description of samples

The population-based biobank of the Estonian Genome Center contains samples from about 52,000 participants recruited between 2002 and 2017 in accordance with the Estonian Human Genes Research Act and based on a broad informed consent form. Biological material stored from all participants includes extracted DNA, blood plasma and buffy coat cells (11).

DNA stored in the biobank has been extracted, using a manual DNA extraction protocol based on alcohol precipitation, from K_2EDTA (dipotassium ethylenediaminetetraacetic acid) stabilized whole blood. An external quality assessment study involving 197 laboratories in Europe found that 90% of respondents reported typically using K_2EDTA blood collection tubes (12). The blood samples have been stored and transported at +4°C and processed within 72 hours of samples collection. A recent study has found that although DNA yields slowly decline with longer blood sample storage times, there is no significant change in DNA purity and integrity when storing K_2EDTA stabilized blood samples up to 168 hours at +4°C (13). The detailed protocol used for DNA extraction can be found in Annex 1. Following isolation, DNA has been resuspended in TE buffer (Tris-EDTA buffer), measured spectrophotometrically and aliquotted in storage straws, with 10-14 straws of about 450 µl each per participant, and stored in the liquid phase of nitrogen in RCB600 liquid nitrogen storage vessels (Air Liquide). TE-buffer is used to limit pH variations that could lead to DNA degradation, stabilize the DNA double helix and reduce the activity of nucleases by chelating divalent cations (14,15). The average measured DNA concentration before aliquotting is 159 ng/µl and the median DNA concentration is 139 ng/µl (range of concentrations is 1 – 2932 ng/µl, standard deviation is 90 ng/µl). Prior to use in research projects, one of the DNA aliquots from each participant has been thawed, transferred to a storage tube, and placed in a -20°C automated storage system for

increased speed and ease of access to the samples. During the transfer to the -20°C automated store, all samples that had concentrations measured above 100 ng/µl before aliquotting and liquid nitrogen storage were diluted to an expected concentration range of 50 to 100 ng/µl based on the pre-aliquotting concentration measurements. However, large sample volumes and imperfect homogenization prior to aliquotting and storage (especially in higher concentration samples) led to imprecise inital concentration measurements in some samples. The importance of adequate mixing of DNA samples before spectrophotometric measurements has been stressed elsewhere as well (16). Also, some samples had measured concentrations below 50 ng/µl. Hence, a number of samples were still expected to be outside the 50 to 100 ng/µl concentration range when retrieved from the automated store and aliquotted for sample release.

Pseudonymized samples and data of the participants are accessible to researchers, following approval of the project application by a scientific committee and permission from the Ethics Review Committee on Human Research of the University of Tartu (17). In addition to outside researchers, basic research in human genetics and genomics of common diseases and traits of medical importance is also carried out by the Estonian Genome Center itself (18–20), with generated data later being available to other researchers.

Because no descriptions of DNA or states of health are used in this Master's thesis and only information classified as quality control data from the biobank is used, an ethics approval is not required for this thesis.

In 2017, the Estonian Genome Center carried out a project to genotype all biobank samples that did not have existing whole-genome array data from previous projects. This project included a total of 34,626 samples which were genotyped using the Illumina Infinium Global Screening Array whole-genome genotyping arrays with approximately 700,000 markers at the Genotyping and Sequencing Core Facility of the Estonian Genome Center.

2.1.2 Initial sample handling and genotyping

The work described in this section was performed prior to this Master's thesis at the Biobank Lab and the Genotyping and Sequencing Core Facility of the Estonian Genome Center in 2017.

Samples selected for genotyping were picked and retrieved from the automated -20°C storage system and aliquotted into 96-well microplates (sample release plates) in a volume of about 15 µl per sample. The aliquotting was performed on a Hamilton Microlab StarPlus liquid handling system integrated with the automated storage system in daily batches of about 960 to 1920 samples. No particular sample sorting criteria were used besides attempted grouping of samples with expected very low or very high concentrations into daily batches – due to the unreliability of some concentration measurements described in 2.1.1, this had limited success, and the majority of samples were sorted without any criteria by the automated storage system.

Since at the time, the Genotyping and Sequencing Core Facility required the sample concentrations to be in the range of 50-100 ng/µl, this was followed by concentration measurements of all samples on a NanoDrop ND-8000 spectrophotometer (Thermo Scientific). Samples that were above the suitable concentration range were manually diluted using TE buffer in their positions in the sample release plates. Samples that were below the suitable concentration range were moved from the microplate to a microtube labelled with the plate and position information of the sample and concentrated using a RVC 2-25 CDplus rotational vacuum concentrator (Martin Christ) at 1 mbar and +40°C. If initial DNA concentrations were too low to reach the required concentration range in a suitable volume, additional DNA solution from the same -20°C storage aliquot was added to the microtube. If the vacuum concentrating led to higher than expected final concentrations, samples were diluted again using Milli-Q water. In some cases where the 96-well microplate contained a large number of low concentration samples, the entire microplate was subjected to vacuum concentrating followed by diluting individual samples as necessary.

Due to time constraints, no additional quality control procedures were performed on the samples, and following the concentration measurements and adjustments, the DNA samples were transferred to the Genotyping and Sequencing Core Facility for genotyping. The genotyping was performed following the Illumina Infinium HTS assay workflow (5) and the BeadChips were scanned using an Illumina HiScan microarray scanner.

2.1.3 Genotyping data quality control

The work described in this section was performed prior to this Master's thesis at the Estonian Genome Center Science Center in 2017 and 2018.

Genotyping data from all samples was passed through the standard Genotyping Array QC (quality control) Pipeline of the Estonian Genome Center. The PLINK software package (21) was used for the analysis.

The *ped* and *map* files created for all samples at the Genotyping and Sequencing Core Facility were used as an input for the QC pipeline. The number of samples and markers in the analysis batches were counted and the format of sample ID values was verified.

As a first QC step, the sample call-rates were analysed and samples with call-rates below 95% were excluded. This exclusion step provided the main part of "genotyping data quality control failed" samples that were analysed in this Master's thesis.

Following this, the sample heterozygosity for both common and rare variants for all samples was analysed and documented, but no sample exclusions were made at this point. The data was then checked for duplicate samples and non-biobank samples that might have been present in the sample batches. No duplicate or non-biobank samples were used in this Master's thesis, so any exclusions in this QC step have already been removed from the initial dataset of 34,626 samples used in this work.

A gender check was then performed, comparing the reported gender for each sample with the gender determined from the genotype. Any samples with mismatches in the gender check were excluded. Further description of samples that failed the gender check is outlined in 3.2.

Following this, principal component plots of nationalities were then created for the analysed samples. Samples that clustered differently than expected were flagged for further analysis. Further description of samples that were excluded in this step is outlined in chapter 3.1.

The rest of the QC pipeline involved cleaning and reformatting data to prepare ready to use genotype data, which is not relevant in the context of this Master's thesis.

2.2 Data related to samples

The data described in section 2.2.1 was generated prior to this Master's thesis during initial sample processing at the Biobank Lab from 2002 to 2017 (described in 2.1.1), during the sample release procedures and genotyping at the Biobank Lab and the Genotyping and Sequencing Core Facility in 2017 (described in 2.1.2) and during the genotyping data quality control at the Estonian Genome Center Science Center in 2017 and 2018 (described in 2.1.3).

The data aggregation from different databases and logfiles, data cleaning and preparation for analysis was done as part of this Master's thesis.

2.2.1 Pre-existing data available for the samples

The total number of samples that were submitted to genotyping was 34,626. For all these samples, information regarding the spectrophotometric measurements taken prior to genotyping is available. This includes the estimated DNA concentrations as well as A260/A280 and A260/A230 ratios. For samples that were manually diluted (n=4229) or vacuum concentrated (n=4529), this fact was also logged.

2.2.1.1 Spectrophotometric data

Spectrophotometric data is commonly used for both DNA quantification as well as assessing DNA purity. Although fluorometric DNA quantification is often preferred by microarray genotyping and next-generation sequencing service providers (3), an external quality assessment study involving 197 laboratories in Europe found that 94.4% of respondents reported using spectrophotometric measurements to evaluate DNA concentration in samples (12). In spectrophotometric measurements, the absorbance of samples is measured over a range of wavelengths (for nucleic acids, the range is typically from 220 to 350 nm) in comparison to a blanking measurement using the buffer used for sample dissolution.

For quantification, the absorbance of the sample at 260 nm (A260) is used to estimate the concentration of DNA in the sample – the estimated concentration is calculated automatically by the NanoDrop Operating Software (Thermo Scientific) and provided with an accuracy of 0.01 ng/µl. Due to small variations in results even when measuring the same sample drop several times in succession, the concentration data has been logged in the biobank records with an accuracy of 1 ng/µl.

For DNA purity assessment, the ratios of absorbances of the sample between 260 nm and 280 nm (A260/A280) and 260 nm and 230 nm (A260/A230) are routinely used. DNA has an absorbance peak at 260 nm, so the A260/A280 and A260/A230 ratios can be used to detect contaminating molecules that have high absorbances at 280 nm and 230 nm, respectively. Contaminations increasing the absorbances at those wavelengths will decrease the respective ratios.

The generally accepted A260/A280 ratios for pure DNA samples are approximately 1.8 to 2.0, with ratios lower than that indicating potential contamination with proteins, phenol or other organic compounds, especially ones containing aromatic groups. High A260/A280 ratios are not generally considered as indicative of DNA quality issues, however they could indicate the presence of

RNA contamination (22,23). The A260/A280 ratio can also be affected by the pH of the solution (24).

The expected A260/A230 ratios for pure DNA samples are approximately 2.0 to 2.2, with ratios lower than that indicating potential contamination with carbohydrates, phenol, glycogen, guanidine or other salts (22). Others have also reported using a range A260/A230 ratios of 1.8 to 2.2 (25) or 1.4 to 2.0 (16) as representing sufficiently pure DNA, depending on the planned downstream applications.

2.2.1.2 Other data

Since the positions of samples in the 96-well microplates used for sample release in the Biobank Lab corresponded directly to final positions of samples on genotyping arrays in the Genotyping and Sequencing Core facility, the grouping of samples on different genotyping arrays and their positions on the arrays is available for analysis.

For all samples the calculated call-rate information is available, with higher call-rates indicating relatively better data quality, both for samples that passed genotyping data quality control and those that failed (call-rate less than 95%).

The biobank logs include some more information regarding the samples that can be used to more thoroughly analyse the sample quality, including the reported gender of the participant and the time of initial sample collection. The time of initial sample collection could be used to assess possible changes in sample quality over time.

2.2.2 Retrospective DNA integrity evaluation

The work described in this section was done as part of this Master's thesis.

2.2.2.1 Grouping samples

Following all sample exclusions (described in 3.1), the remaining samples that were included in quality control analysis were divided into two groups – samples that had passed all genotyping data quality control measures and samples that had failed. The samples that had failed genotyping data quality control were each assigned a positionally matched control to create a smaller subset of successfully genotyped samples to be used in retrospective DNA integrity evaluation alongside the failed samples.

The control samples were matched to the failed samples based on positions on sample release plates (and subsequently, positions on genotyping arrays) in order to avoid possible batch effects affecting the following analysis. Each 96-well sample release plate was analysed on four 24-position genotyping arrays in the Genotyping and Sequencing Core Facility, with samples from rows A+B, C+D, E+F and G+H each making up a single genotyping array (see also: Figure 13 in 3.7.3).

The control matching was done using the following procedure:

- if a failed sample was in a plate position G3, the matched control was selected from either position G2 or G4;
- if several consecutive positions had failed samples, like B5, B6, B7, the matched controls were selected from the consecutive positions before or after the failed samples: B2, B3, B4 or B8, B9, B10;
- if matching from consecutive positions was impossible due to other failed samples, the matched controls were selected from an immediately preceding or following plate row, whichever included samples that were genotyped on the same array. In this case failed samples in positions B3 and B4 could get matched controls also from positions A3 and A4, but not from C3 and C4.

2.2.2.2 Agarose gel electrophoresis

Following the creation of the groups of failed samples and their matched controls, the samples in both groups were analysed using agarose gel electrophoresis. The same sample aliquots from the -20°C automated storage system that were used for genotyping were used for all samples in agarose gel electrophoresis. This quality control method was selected because this is used as a standard quality control step to check DNA integrity under other circumstances at the Biobank Lab. Agarose gel electrophoresis is a routine and low-cost procedure for estimating DNA integrity and detecting DNA degradation that could possibly influence the performance of molecular tests, especially tests utilizing PCR amplification of genomic DNA (26).

2 µl of each sample was loaded on a 1% agarose gel with ethidium bromide in TBE buffer (Tris/Borate/EDTA buffer) using Orange DNA Loading Dye (6X) (Thermo Scientific). The gels were ran at 150V for 35 minutes, with "Lambda DNA/ EcoRI+HindIII Marker, 3, ready-to-use" (Thermo Scientific) DNA ladder as a size reference on each lane using a Compact XL gel electrophoresis equipment (Analytik Jena) and a Consort EV243 power supply (Consort BVBA). The gels were visualized and imaged on a GeneGenius Bio Imaging System (Syngene).

The gel electrophoresis images were visually analysed, and the samples were labelled as either (a) displaying a clear band of intact genomic DNA (estimated to have >20 kb (kilobase) fragment lengths using the DNA ladder as reference); (b) having degraded DNA represented as a highly smeared image on the gel, or (c) displaying no visible double-stranded DNA on the gel. Samples were labelled as having intact genomic DNA even if there was a smeared component on the image as well, as long as there was a recognisable band of high molecular weight DNA.

2.3 Statistical methods

All statistical analysis was done as part of this Master’s thesis.

Kruskal-Wallis H test was used in the case of more than 2 sets of ordinal data to test for significant differences in the distribution of values between at least 1 pair of sample sets. IBM SPSS Statistics software (version 1.0.0-3209) was used for these tests.

Mann-Whitney U test was used in the case of 2 sets of ordinal data to test for significant differences in the distribution of values between the 2 sample sets. IBM SPSS Statistics software (version 1.0.0-3209) was used for these tests.

Pearson's chi-squared test was used in the case of categorical data to test the significance of the observed differences in the distribution of samples from the expected distributions. Microsoft Excel software (version 1808) was used for these tests.

3 Results

3.1 Sample exclusions

Out of the total of 34,626 samples released for genotyping from the Biobank Lab, there were 53 samples, which had no information on the genotyping data quality control results returned from the Genotyping and Sequencing Core Facility.

While it cannot be ruled out that some of the issues with these samples were caused by low quality DNA samples, inspection of the positional data of the samples in question revealed that 48 out of the 53 samples in question were clustered on only 4 separate genotyping arrays, indicating that a possible issue with those arrays or their processing in the Genotyping and Sequencing Core Facility had led to the missing data issue. To avoid mixing up likely technical issues with actual DNA quality issues in further analysis, all the samples with no genotyping data quality control information (53 samples) as well as other samples processed on the 4 genotyping arrays where several samples had no data (a further 48 samples) were excluded from further analysis. Out of the 48 extra samples excluded from further analysis, 16 samples (33.3%) had genotyping data quality control issues logged as well, further confirming the genotyping array processing issues.

The second step of exclusion included 192 samples (two complete 96-well sample release plates), where the samples had otherwise good data (191 samples out of 192 passed all other quality control), but for an unidentified reason all the samples on these plates clustered separately from all other samples in the principal component plots in genotyping data quality control. No variable common to these two sample release plates was identified during sample processing, as the plates were not subsequent and each plate was also batched together with other sample release plates throughout the sample processing chain. While no specific issue was identified, the data from these samples was excluded from all downstream uses, and as such will be excluded for the purposes of the current DNA quality analysis as well.

3.2 Identifying genotyping data quality control issues

Out of the remaining 34,333 samples that were included in the DNA quality analysis, 33,122 samples (96.5%) passed all genotyping data quality control steps and were cleared for use in future research projects, while 1211 samples (3.5%) failed to pass the quality control.

Out of the 1211 samples that failed genotyping data quality control, 1008 samples had low call-rates (a call-rate lower than 95%), 67 samples had a mismatch between reported gender for the sample and the gender determined from the genotype data, and 136 samples had both low call-rates and a gender mismatch.

Since the genotype-based gender determination method is highly dependent on the quality of the genotyping data, the samples that had both low call-rates and a gender mismatch (136 samples) were treated as low call-rate samples for the purposes of subsequent analysis. The samples that otherwise passed genotyping data quality control but had a gender mismatch (67 samples) were treated as having successfully passed quality control for the purposes of subsequent DNA quality analysis.

Including the gender mismatch samples with good call-rates as having passed genotyping data quality control and the gender mismatch samples with low call-rates as having failed genotyping data quality control, for the purposes of this Master's thesis, the total number of failed samples out of 34,333 analysed samples was 1144 (3.33%).

3.3 Agarose gel electrophoresis

Agarose gel electrophoresis was performed on all 1144 failed samples and a further 1144 positionally matched control samples, as described in 2.2.2.2.

Typical gel electrophoresis images for failed samples and matched controls are shown in Figure 1 and Figure 2, respectively.

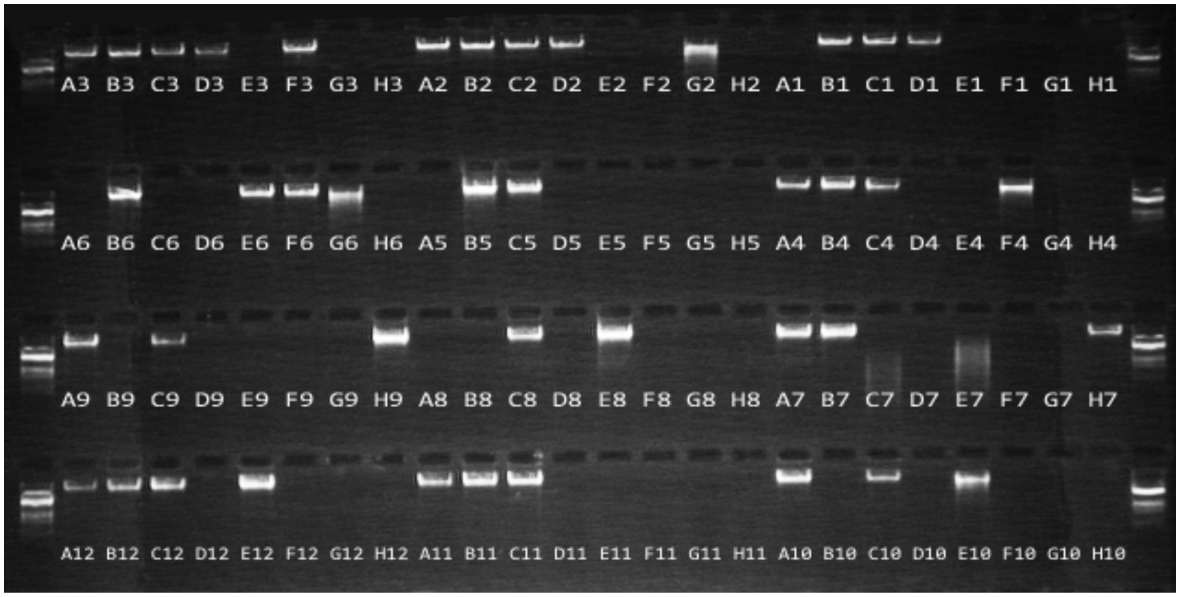

Figure 1. A typical gel electrophoresis image for failed samples. In this image, 41 samples were evaluated as being intact, 2 samples as degraded (labelled C7 and E7 in the image) and 53 samples as missing completely.

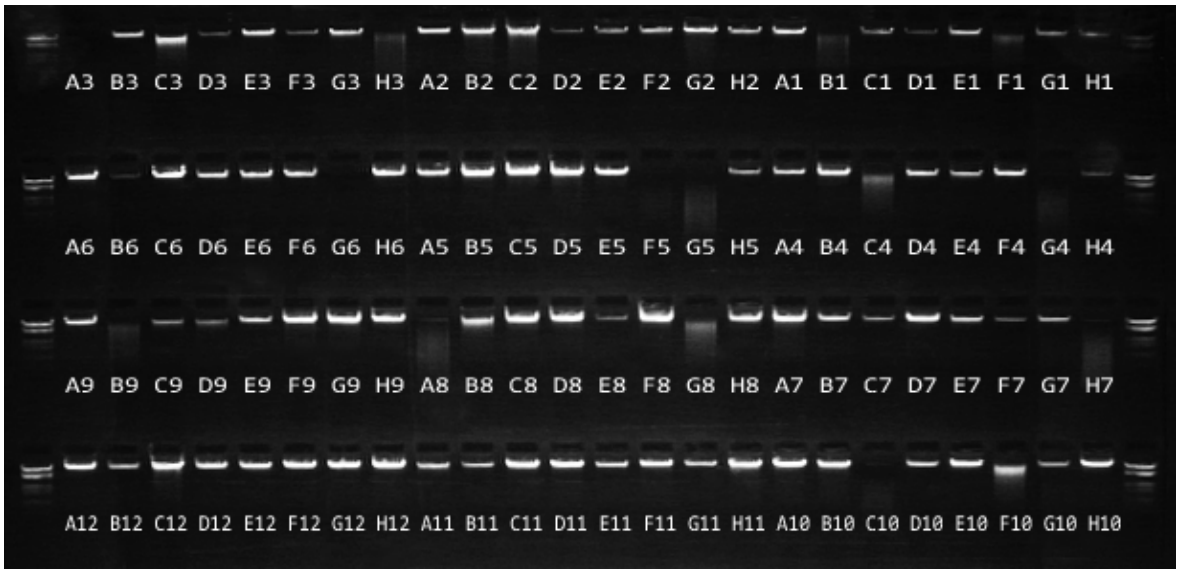

Figure 2. A typical gel electrophoresis image for matched control samples. In this image, 90 samples were evaluated as being intact, 4 samples as degraded (labelled G4, G5, F5 and C10 in the image) and 2 samples as missing completely (labelled A3 and G6 in the image). In some cases, the band of high molecular weight DNA is very poorly visible, and brightness and contrast adjustments were used in assessing the images.

The results from analysing all gel electrophoresis images are shown in Table 1.

	Genotyping failed samples		Matched control samples	
Agarose gel electrophoresis result	**Number of samples**	**Percentage**	**Number of samples**	**Percentage**
Intact	430	37.6%	1048	91.6%
Degraded	127	11.1%	73	6.4%
Missing	587	51.3%	23	2.0%
TOTAL	1144		1144	

Table 1. Agarose gel electrophoresis results.

A chi-squared test was used to test whether the differences in the distribution of samples between genotyping failed samples and matched controls were significant, and a significant difference was confirmed for all 3 agarose gel electrophoresis results groups ($p<0.001$ in all cases). The genotyping failed samples have significantly less intact samples and significantly more degraded and missing samples than the matched controls.

3.4 Comparing genotyping call-rate data with gel electrophoresis data

To further assess the impact of agarose gel electrophoresis results on the genotyping data quality, even in matched control samples, available genotyping call-rate data was used. Because call-rate data was already available for all samples, all other successfully genotyped samples were also used as an additional group of control samples (named as “other controls”) in addition to the positionally matched control samples used for agarose gel electrophoresis. For analysis, the samples were divided into the same 6 groups as described in Table 1 and the sample group of other controls was added. The descriptive statistics of the call-rate data for six groups of samples based on agarose gel electrophoresis results plus other controls and all samples are shown in Table 2 and average call-rates with 95% confidence intervals are shown in Figure 3.

	All samples	Intact failed samples	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Number of samples	33,723	427	127	587	1043	73	23	31,443
Minimum value	0.0000	0.1091	0.3499	0.0000	0.9500	0.9500	0.9506	0.9500
Maximum value	0.9965	0.9500	0.9486	0.9467	0.9964	0.9947	0.9961	0.9965
Average	0.9856	0.8596	0.8683	0.6239	0.9924	0.9901	0.9737	0.9943
Standard deviation	0.0583	0.1247	0.1118	0.1846	0.0074	0.0081	0.0136	0.0036
Median	0.9947	0.8972	0.9105	0.6018	0.9946	0.9931	0.9723	0.9948

Table 2. Descriptive statistics of call-rate data.

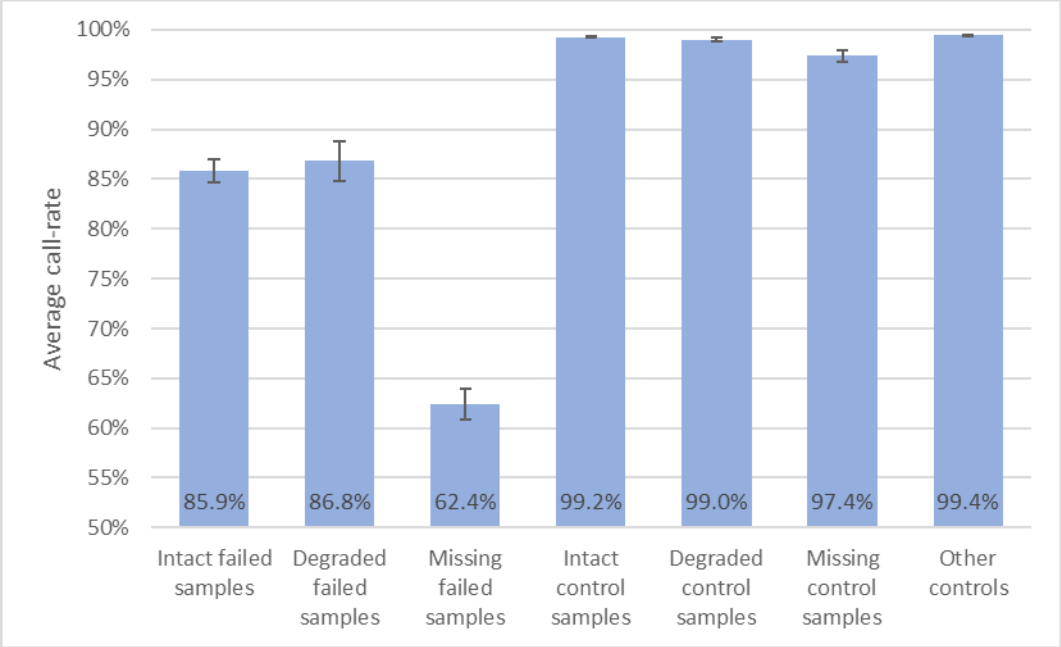

Figure 3. Average call-rates with 95% confidence intervals for different sample groups of agarose gel electrophoresis results and other controls.

Since the call-rate value of 0.95 as a cut-off point was the main distinguishing factor between genotyping failed samples and successfully genotyped samples, the differences in average call-rates between genotyping failed samples and matched control samples are evident. To determine whether there are further significant differences in call-rates between different groups of genotyping failed samples and different groups of matched control samples, a Kruskal-Wallis H test was used. The test confirmed that there is a significant difference in call-rates

between at least 2 groups of genotyping failed samples ($p<0.001$) as well as between at least 2 groups of matched control samples ($p<0.001$). Pairwise Mann-Whitney U tests were used to find the specific pairs of sample groups with significant differences in call-rate distributions in both genotyping failed samples and matched control samples. The results are shown in Table 3 and Table 4.

	Degraded failed samples	**Missing failed samples**
Intact failed samples	0.015	< 0.001
Degraded failed samples		< 0.001

Table 3. p-values from pairwise Mann-Whitney U tests between call-rates of different groups of genotyping failed samples.

	Degraded control samples	**Missing control samples**
Intact control samples	< 0.001	< 0.001
Degraded control samples		< 0.001

Table 4. p-values from pairwise Mann-Whitney U tests between call-rates of different groups of matched control samples.

The testing revealed significant differences in call-rates between all groups of genotyping failed samples as well as between all groups of control samples, although as seen from Table 2 and Figure 3, in the cases of some pairs, the actual differences between average call-rate values are very small.

3.5 Spectrophotometric data analysis

Because spectrophotometric data was also available for all samples, the group of other control samples introduced in 3.4 was used here as well, along with the continued use of sample grouping based on agarose gel electrophoresis results for both genotyping data quality control failed samples and positionally matched control samples in order to determine whether spectrophotometric data could be used as a proxy for agarose gel electrophoresis results.

3.5.1 Sample exclusions

Initial analysis of spectrophotometric data revealed that in 254 samples out of the total of 34,333 analysed samples, the A260/A230 ratios were reported as negative values due to the fact that the absorbance values at the 230 nm wavelength (A230) have been measured as negative compared to the blanking measurement (TE buffer was used as a blanking solution). This is unexpected, because absorbance values in correctly measured samples should always be higher than in the blanking measurement, and a usual quality control procedure is to check for higher than expected A230 values (translated into lower than expected A260/A230 ratios), which is associated with contaminants absorbing at 230 nm. In this context, the samples with negative A260/A230 ratios are more closely resembling samples with very high A260/A230 ratios (due to very low but positive A230 values), rather than samples with very low but positive A260/A230 ratios (due to high A230 values) – see Figure 4 below.

A further analysis of the positions of the 254 samples with negative A230 values revealed that 219 of those samples were positioned in a total of six 96-well plates measured in a single batch, indicating either a wrong blanking solution being used for that batch or some other issue with the blanking - for example dirty measurement pedestals causing a higher absorbance measurement for the blank, which was subsequently wiped away from the measurement pedestals with the blanking solution.

Because the indicated issue with blanking measurement in this sample batch would make all the spectrophotometric measurement values incorrect, the spectrophotometric data of all samples from those 6 plates was excluded from further analysis (a total of 575 samples, as 1 sample had already been previously excluded due to missing genotyping data).

The rest of the samples with negative A230 values (a total of 35 samples) were spread over a total of 25 different sample release plates, mostly measured in different batches, indicating issues with specific samples rather than measurement

batches. As described in 2.2.1.1, a likely explanation for samples having lower than expected A230 values could be that water has been used instead of TE buffer to dissolve or dilute the samples at some point in the sample processing chain either during the current experiment or initial sample processing at the Biobank Lab.

The spectrophotometric data from these 35 samples was also excluded from further analysis due to unreliability. After the exclusion of the 610 samples, the total number of samples included in the spectrophotometric data analysis was 33,723.

The 610 excluded samples included 3 samples that had failed genotyping data quality control (a 0.49% failure rate), the 33,723 samples in spectrophotometric data analysis included 1141 samples that had failed genotyping data quality control (a 3.38% failure rate). The excluded samples also included 5 positionally matched control samples that were intact on agarose gel electrophoresis images.

Due to a number of additional samples with very low, although positive, A230 values present in the spectrophotometric data even after the exclusions – 40 samples in the 0 to 0.05 range, 63 samples in the 0.05 to 0.10 range, 109 samples in the 0.10 to 0.15 range and so on - it is likely that a small part of the spectrophotometric data remains unreliable, although no further measurement batch effects that could be used for exclusions were identified.

Inspection of the full absorbance spectra of the samples with negative and very low positive A230 values revealed that while the trough in the spectrum at about 230 nm was significantly lower and slightly shifted towards lower wavelengths for these samples, the rest of the spectrum from about 250 nm and above appeared similar to samples with normal A230 values (Figure 4). This makes it likely that the unreliability in the samples with the outlier A230 values is mostly limited to the A260/A230 ratios, and the A260/A280 ratios and the estimated concentration values are less affected.

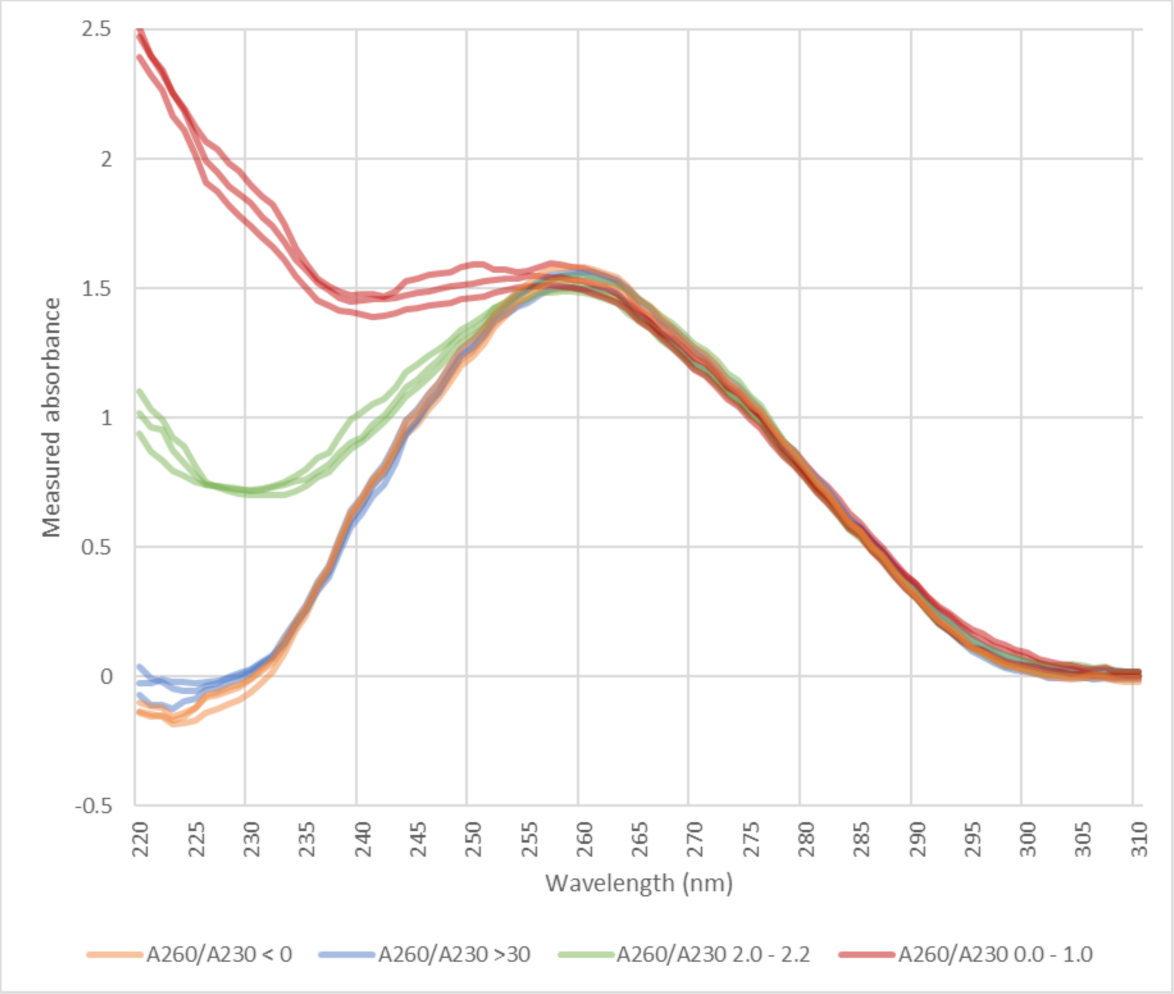

Figure 4. Absorbance spectra of selected samples with different A260/A230 ratio values. Three samples with similar measured concentration values are shown from each group. The similarity of samples with A260/A230 ratios that are negative (orange) or higher than 30 (blue) is visible. Samples with A260/A230 ratios in the 0.0 – 1.0 range (red), which can be associated with high contamination, show completely different absorbance spectra from the negative A260/A230 ratio samples. Samples with the A260/A230 values in the expected range of 2.0 – 2.2 (green) are shown for comparison. In the entire dataset, the transitions between the shown groups are continuous without distinguishable cut-off points.

Due to the reasons stated above and, because of the continuous nature of the range of A230 values, any cut-off point for further sample exclusion would be arbitrary, it was decided to continue the analysis of spectrophotometric data as is, keeping in mind that the data for samples with lower A230 values might at least be partly unreliable.

3.5.2 DNA concentration analysis

The descriptive statistics of the measured DNA concentration data for the six groups of samples based on agarose gel electrophoresis results plus other controls and all samples are shown in Table 5 and the average DNA concentrations with 95% confidence intervals are shown in Figure 5.

	All samples	Intact failed samples	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Number of samples	33,723	427	127	587	1043	73	23	31,443
Minimum value (ng/µl)	42	48	53	48	48	59	59	42
Maximum value (ng/µl)	103	100	98	101	101	98	96	103
Average (ng/µl)	75.8	76.2	80.4	80.9	75.8	78.9	78.4	75.6
Standard deviation (ng/µl)	11.5	11.3	7.4	7.9	11.0	8.3	7.5	11.6
Median (ng/µl)	77	77	80	80	77	79	79	77

Table 5. Descriptive statistics of measured DNA concentration data.

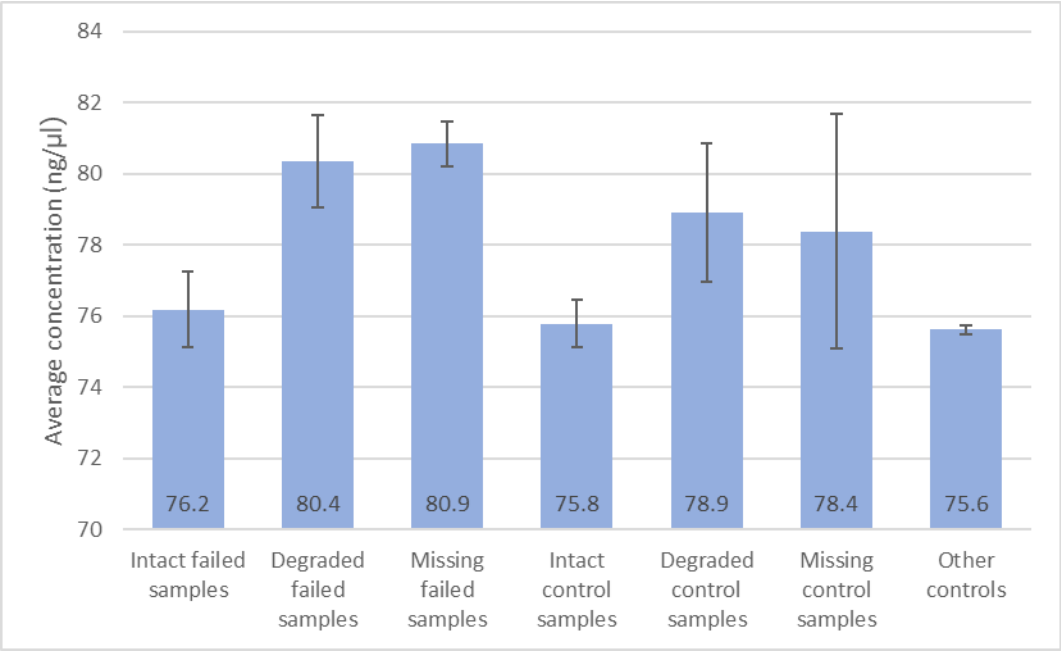

Figure 5. Average DNA concentrations with 95% confidence intervals for different sample groups of agarose gel electrophoresis results and other controls.

To determine whether there are significant differences in concentrations between different groups of samples, a Kruskal-Wallis H test was used. The test confirmed that there is a significant difference in concentrations between at least 2 groups of samples ($p<0.001$). Pairwise Mann-Whitney U tests were used to find the specific pairs of sample groups with significant differences in DNA concentrations. The results are shown in Table 6.

	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Intact failed samples	< 0.001	< 0.001	0.502	0.046	0.372	0.303
Degraded failed samples		0.615	< 0.001	0.333	0.275	< 0.001
Missing failed samples			< 0.001	0.129	0.166	< 0.001
Intact control samples				0.014	0.260	0.653
Degraded control samples					0.699	0.010
Missing control samples						0.237

Table 6. p-values from pairwise Mann-Whitney U tests between DNA concentrations of different groups of samples. p-values < 0.05 labelled with yellow, p-values < 0.001 labelled with green.

The analysis revealed significant differences between the sample groups of both intact samples and other controls, compared to sample groups of degraded or missing samples, except in the case of missing control samples, where the small number of samples in the group likely prevented reaching higher significance. No significant differences in DNA concentrations of other control samples and intact failed or intact control samples were found.

A repeated analysis that pooled together all samples that were missing or degraded on agarose gel electrophoresis (n=810) versus all other samples (n=32,913) in a Mann-Whitney U test also showed a significant difference between the concentrations in the two groups ($p<0.001$) with degraded and missing samples having an average concentration of 80.5 ng/µl and all other samples having an average concentration of 75.6 ng/µl.

3.5.3 A260/A280 ratio analysis

The descriptive statistics of the measured A260/A280 ratio data for the six groups of samples based on agarose gel electrophoresis results plus other controls and all samples are shown in Table 7 and the average A260/A280 ratios with 95% confidence intervals are shown in Figure 6.

	All samples	Intact failed samples	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Number of samples	33,723	427	127	587	1043	73	23	31,443
Minimum value	1.34	1.65	1.76	1.75	1.63	1.77	1.80	1.34
Maximum value	2.58	2.27	2.15	2.39	2.20	2.18	2.20	2.58
Average	1.86	1.87	1.91	1.92	1.87	1.90	1.90	1.86
Standard deviation	0.06	0.06	0.07	0.07	0.06	0.06	0.08	0.06
Median	1.86	1.86	1.91	1.92	1.87	1.90	1.90	1.86

Table 7. Descriptive statistics of measured A260/A280 ratio data.

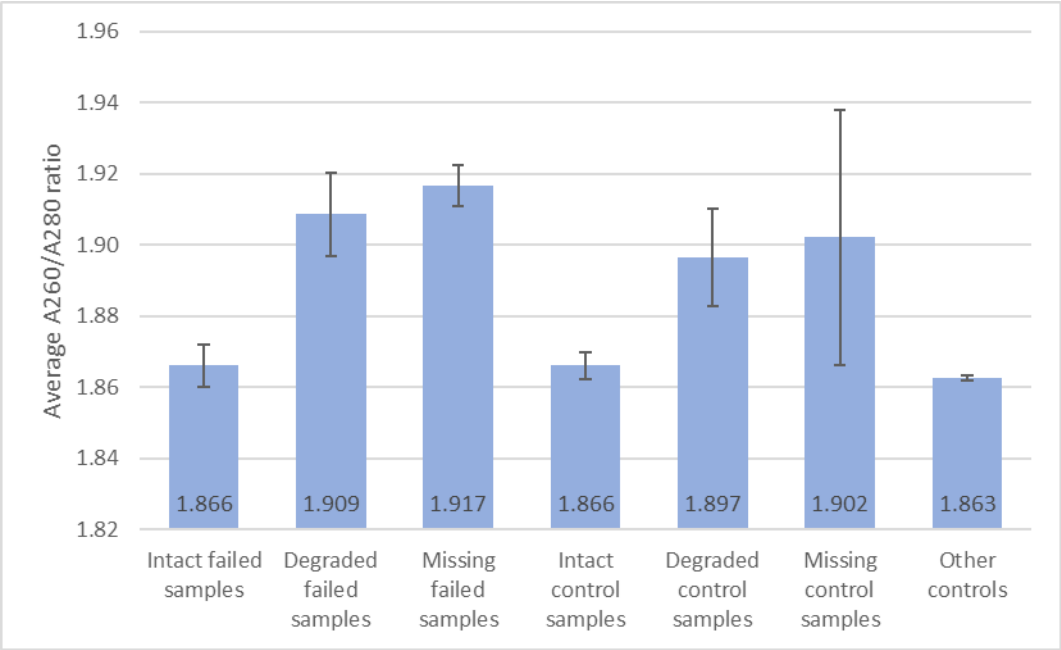

Figure 6. Average A260/A280 ratios with 95% confidence intervals for different sample groups of agarose gel electrophoresis results and other controls.

To determine whether there are significant differences in A260/A280 ratios between different groups of samples, a Kruskal-Wallis H test was used. The test confirmed that there is a significant difference in A260/280 ratios between at least 2 groups of samples ($p<0.001$). Pairwise Mann-Whitney U tests were used to find the specific pairs of sample groups with significant differences in A260/280 ratios. The results are shown in Table 8.

	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Intact failed samples	< 0.001	< 0.001	0.861	< 0.001	0.033	0.348
Degraded failed samples		0.145	< 0.001	0.185	0.357	< 0.001
Missing failed samples			< 0.001	0.004	0.092	< 0.001
Intact control samples				< 0.001	0.031	0.080
Degraded control samples					0.901	< 0.001
Missing control samples						0.016

Table 8. p-values from pairwise Mann-Whitney U tests between A260/A280 ratios of different groups of samples. p-values < 0.05 labelled with yellow, p-values < 0.001 labelled with green.

The analysis revealed significant differences between the sample groups of intact samples and other controls, compared to sample groups of degraded or missing samples – unlike with DNA concentration analysis, here even the missing control samples show some significant differences with intact samples and other controls. No significant differences in A260/A280 ratios of other control samples and intact failed or intact control samples were found.

A repeated analysis that pooled together all samples that were missing or degraded on agarose gel electrophoresis (n=810) versus all other samples (n=32,913) in a Mann-Whitney U test showed a significant difference between the average A260/A280 ratios in the two groups ($p<0.001$) with degraded and missing samples having an average A260/A280 ratio of 1.91 and all other samples having an average A260/A280 ratio of 1.86.

3.5.4 A260/A230 ratio analysis

The A260/A230 ratios of samples were found to have high variability and a large number of outliers even after the sample exclusions described in 3.5.1. While in the case of A260/A280 ratios, 86.1% of the samples were in the generally expected range of 1.8 to 2.0 (described in 2.2.1.1) and 98.5% of the samples were in the 1.7 to 2.0 range, using the generally accepted range of 2.0 to 2.2 for the A260/A230 ratios, only 13.0% of all samples were found to be in that range. 28.9% of the samples had A260/A230 ratios higher than 2.2 and 58.1% of the samples had A260/A230 ratios lower than 2.0.

The descriptive statistics of the measured A260/A230 ratio data for the six groups of samples based on agarose gel electrophoresis results plus other controls and all samples are shown in Table 9 and the average A260/A230 ratios with 95% confidence intervals are shown in Figure 7.

	All samples	Intact failed samples	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Number of samples	33,723	427	127	587	1043	73	23	31,443
Minimum value	0.10	0.29	0.66	0.31	0.24	1.12	1.36	0.10
Maximum value	279.21	14.71	2.7	68.81	12.49	2.49	8.82	279.21
Average	2.25	1.99	1.66	1.94	1.92	1.73	2.08	2.27
Standard deviation	3.29	1.08	0.35	2.81	0.96	0.24	1.47	3.38
Median	1.90	1.87	1.67	1.75	1.77	1.70	1.69	1.92

Table 9. Descriptive statistics of measured A260/A230 ratio data.

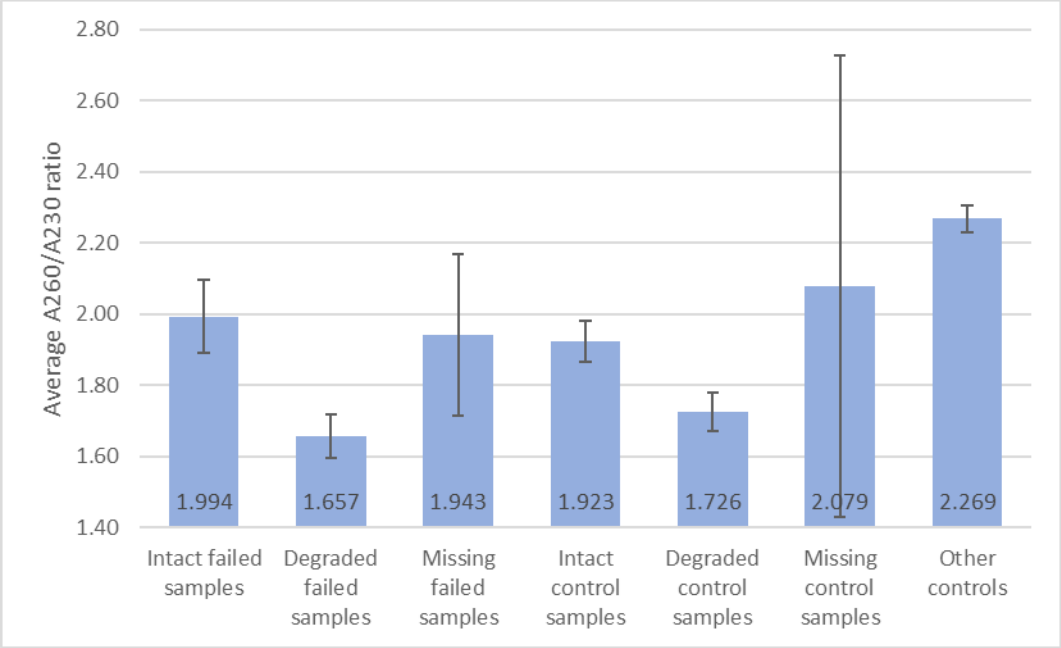

Figure 7. Average A260/A230 ratios with 95% confidence for different sample groups of agarose gel electrophoresis results and other controls.

To determine whether there are significant differences in A260/A230 ratios between different groups of samples, a Kruskal-Wallis H test was used. The test confirmed that there is a significant difference in A260/230 ratios between at least 2 groups of samples ($p<0.001$). Pairwise Mann-Whitney U tests were used to find the specific pairs of sample groups with significant differences in A260/230 ratios. The results are shown in Table 10.

	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Intact failed samples	< 0.001	0.001	0.012	0.001	0.219	0.003
Degraded failed samples		< 0.001	< 0.001	0.379	0.253	< 0.001
Missing failed samples			0.665	0.011	0.409	< 0.001
Intact control samples				0.036	0.633	< 0.001
Degraded control samples					0.631	< 0.001
Missing control samples						0.055

Table 10. p-values from pairwise Mann-Whitney U tests between A260/A230 ratios of different groups of samples. p-values < 0.05 labelled with yellow, p-values < 0.001 labelled with green.

Surprisingly, a significant difference in the distribution of A260/A230 ratios between other control samples (genotyping data quality control passed samples that were not included in agarose gel electrophoresis) and every other sample group (except missing control samples) is seen. The fact that a significant difference exists also with intact control samples indicates that matching agarose gel electrophoresis control samples positionally (as described in 2.2.2.1) has introduced an unplanned bias to the control samples.

A repeated analysis that pooled together all samples that were missing on agarose gel electrophoresis (n=610), all samples that were degraded on agarose gel electrophoresis (n=200) and all samples that were intact on agarose gel electrophoresis (n=1470) versus other control samples (n=31,443) for pairwise Mann-Whitney U tests was performed. The results are shown in Table 11.

	All degraded samples	All missing samples	Other controls
All intact samples	< 0.001	0.087	< 0.001
All degraded samples		< 0.001	< 0.001
All missing samples			< 0.001

Table 11. p-values from pairwise Mann-Whitney U tests between A260/A230 ratios of different groups of samples. p-values < 0.001 labelled with green.

3.5.4.1 Analysis of A260/A230 ratio subgroups

Although significant differences were identified between several sample groups, the high variability observed in the A260/A230 ratio data (standard deviations higher than averages in some sample groups), means that while the identified differences may be useful in helping to determine possible underlying causes of the degradation of samples and samples failing genotyping data quality control, this information regarding differences of average values is less useful in separating samples with possible issues based on A260/A230 ratio values. Due to this, a further analysis into the distribution of A260/A230 ratios between sample groups was carried out.

The A260/A230 ratios of all samples were divided into 5 subgroups of about 6,745 samples each, in order of measured A260/A230 ratios. The subgroups were not exactly equal in size, because the closest value changes in A260/A230 ratios were selected as group borders. The distribution of different sample groups in these A260/A230 ratio subgroups was then compared in order to find any possible deviations. The distribution of samples between subgroups is shown in Table 12 and the relative distributions are visualized in Figure 8.

	All samples	Intact failed samples	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples	Other controls
Total number of samples	33,723	427	127	587	1043	73	23	31,443
A260/A230 ratio <1.55	6747 (20.0%)	100 (23.4%)	31 (24.4%)	69 (11.8%)	238 (22.8%)	12 (16.4%)	3 (13.0%)	6294 (20.0%)
A260/A230 ratio 1.55 .. 1.79	6631 (19.7%)	83 (19.4%)	63 (49.6%)	274 (46.7%)	315 (30.2%)	43 (58.9%)	13 (56.5%)	5840 (18.6%)
A260/A230 ratio 1.80 .. 2.01	6735 (20.0%)	98 (23.0%)	18 (14.2%)	133 (22.7%)	204 (19.6%)	11 (15.1%)	3 (13.0%)	6268 (19.9%)
A260/A230 ratio 2.02 .. 2.49	6785 (20.1%)	72 (16.9%)	14 (11.0%)	91 (15.5%)	188 (18.0%)	7 (9.6%)	1 (4.3%)	6412 (20.4%)
A260/A230 ratio >2.49	6825 (20.2%)	74 (17.3%)	1 (0.8%%)	20 (3.4%)	98 (9.4%)	0 (0.0%)	3 (13.0%)	6629 (21.1%)

Table 12. Distribution of samples in different groups between A260/A230 ratio subgroups. The proportion of sample group samples in each A260A230 ratio subgroup is shown in brackets.

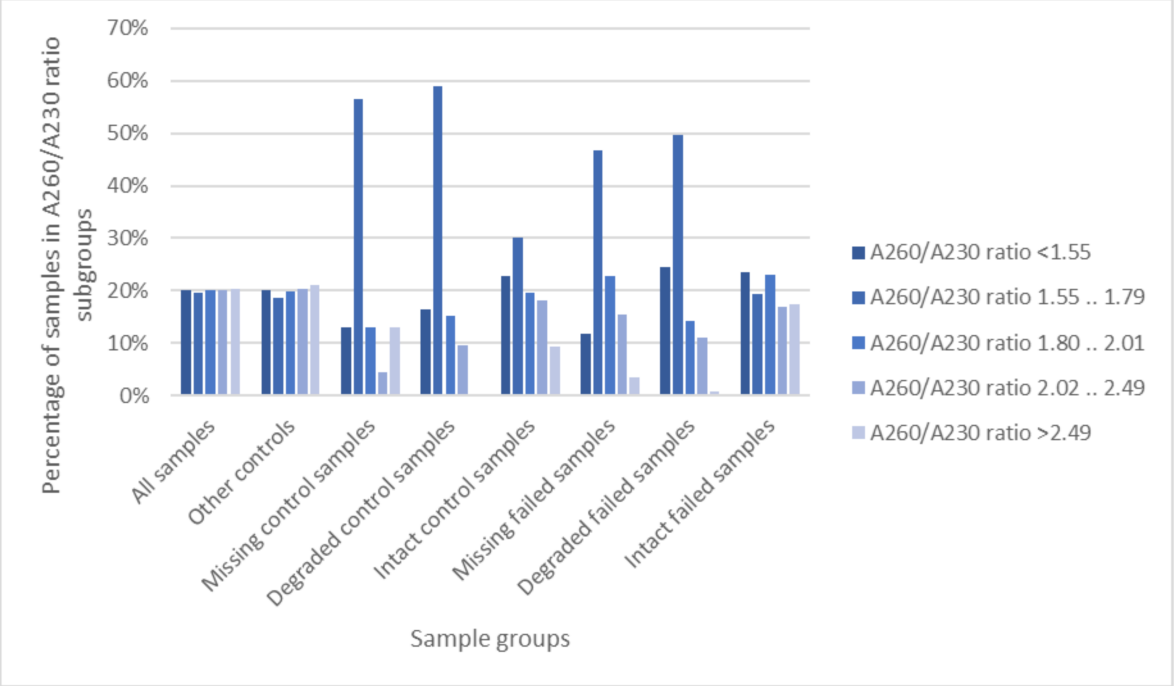

Figure 8. Relative distributions of samples among A260/A230 ratio subgroups.

A chi-squared test was used to test each sample group distribution between A260/A230 ratio subgroups to determine whether they differ significantly from the expected distribution based on the distribution of all samples. Significant differences ($p<0.001$) from the expected distributions (based on the group of all samples) were found in all sample groups, except intact failed samples ($p=0.077$).

Since the data indicates that all samples that were missing or degraded on agarose gel electrophoresis are significantly overrepresented in the A260/A230 ratio range between 1.55 and 1.79, a more granular inspection of the data is warranted. Surprisingly, the positionally matched control samples that were intact in agarose gel electrophoresis also show a significant tendency towards the two lowest A260/A230 ratio subgroups. No significant trend was found in genotyping data quality control failed samples that were intact in agarose gel electrophoresis.

Since the initial analysis shows a similar trend for all sample groups that were missing or degraded in agarose gel electrophoresis, these sample groups are pooled together (n=810) in the following analysis in order to establish whether A260/A230 ratios could be used as proxies for agarose gel electrophoresis results. Intact control samples (n=1043) and intact failed samples (n=427) are maintained as separate samples groups, as well as other control samples (n=31,443).

For a more granular analysis, smaller subgroups of A260/A230 ratios were created by sorting all samples in order of A260/A230 ratios along with the sample group designation of each sample and creating subgroups with a minimum size of 500 samples (the minimum subgroup size is intended to reduce random variation in data). Samples with each specific measured A260/A230 ratio in growing order were combined in a single subgroup, until the number of samples in the subgroup exceeded 500, after which a new subgroup was started with the samples with the next highest measured A260/A230 ratio. A total of sixty A260/A230 ratio subgroups were created, with an average size of 562 samples per subgroup.

Following this, the number of samples from each sample group present in each A260/A230 ratio subgroup was counted and an expected number of such samples was calculated by multiplying the size of the A260/A230 ratio subgroup by the

relative proportion of the sample group compared to all samples. By dividing the actual number of samples by the expected number of samples for each data point, the “relative abundance” of each sample subgroup in each A260/A230 ratio subgroup was determined.

For example, in the A260/A230 ratio subgroup of 1.66 to 1.67 with a size of 546 samples, the expected number of all missing and degraded samples would be 546 * (810 / 33,723) = 13.11. The actual number of all missing and degraded samples in that subgroup is 46, so the relative abundance is calculated as 46 / 13.11 = 3.51, meaning there are 3.51 times more missing and degraded samples in that A260/A230 ratio subgroup than would be expected with a uniform distribution.

The results of this analysis for the four groups of samples used in this analysis are shown in Figure 9.

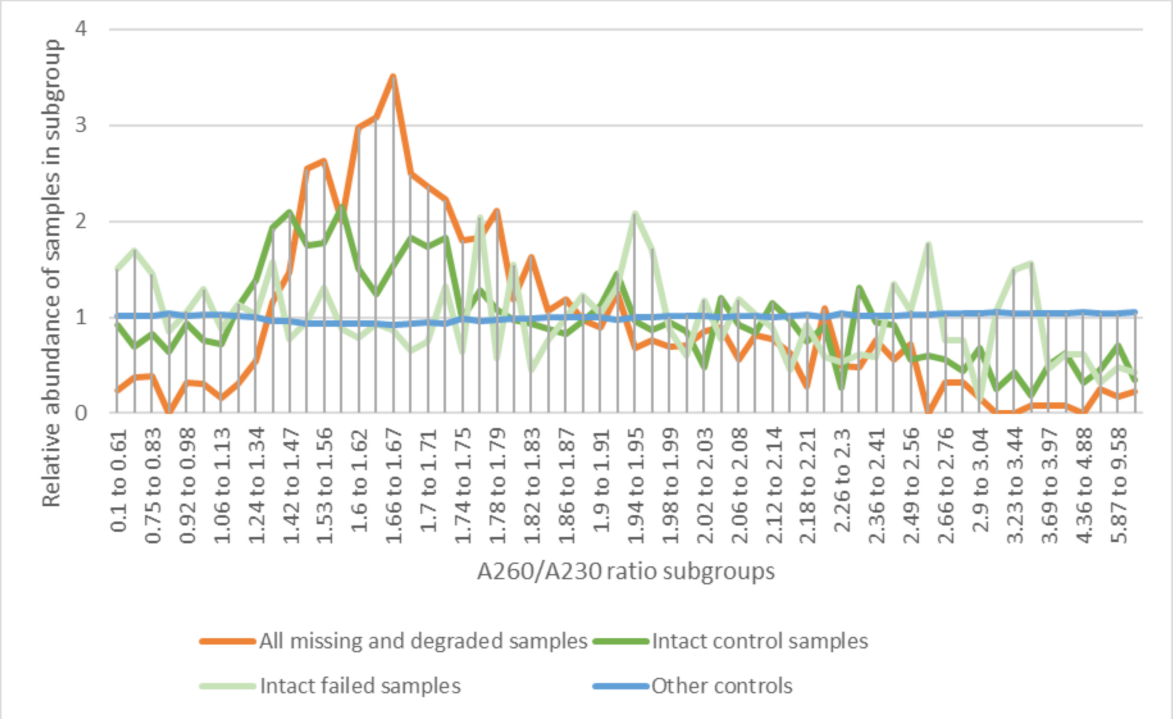

Figure 9. The relative abundance of samples in A260/A230 ratio subgroups for different sample groups. Not all subgroups are labelled on the axis to improve readability.

The graph for the group of all missing and degraded samples has a peak at the A260/A230 ratio subgroup of 1.66 to 1.67 with a relative abundance of 3.51. The relative abundance for all missing and degraded samples is higher than 1 over 18 continuous A260/A230 ratio subgroups covering the A260/A230 ratio range of 1.35 to 1.87, with a total of 11,097 samples (32.9% of all analysed samples) and 554 missing and degraded samples (68.4% of all missing and degraded samples).

The graph for the group of intact control samples has a less distinguished peak at the A260/A230 ratio subgroup of 1.57 to 1.59 with a relative abundance of 2.16. The relative abundance for intact control samples is higher than 1 over 13 continuous A260/A230 ratio subgroups covering the A260/A230 ratio range of 1.14 to 1.73, with a total of 7611 samples (22.6% of all analysed samples) and 393 intact control samples (37.7% of all intact control samples).

The graph for the group of intact failed samples has several lower and narrower peaks, but no clear trend is established – this was already predicted by the failure to find a significant difference from a uniform distribution using larger A260/A230 ratio subgroups above.

The graph for the group of other control samples, which is less variable due to considerably higher sample numbers, has a shallow trough where the relative abundance is lower than 1 throughout 17 A260/A230 ratio subgroups covering the A260/A230 ratio range of 1.24 to 1.83, with a total of 10,381 samples (30.8% of all samples) and 9240 other control samples (29.4% of all other control samples). This through coincides with the peaks in all missing and degraded sample group and the intact control sample group.

3.6 Sample collection time analysis

For the analysis of initial sample collection times and genotyping data quality control results, the sample exclusions used in spectrophotometric data analysis as described in 3.5.1 were not used. Rather, the total number of samples analysed is 34,333 with 1144 samples out of those having failed genotyping data quality control as described in 3.1 and 3.2.

3.6.1 Analysis of failed samples by collection year

For the analysis of initial sample collection times, the numbers of all samples and failed samples were counted for each sample collection year and the failure rates were calculated for each year. A chi-squared test was used to test the distribution of genotyping quality control failed and passed samples of each sample collection year for significant differences from the expected distribution based on the average failure rate. The results of the analysis are shown in Table 13 and the failure rates with 95% confidence intervals are shown in Figure 10.

Sample collection year	Total number of samples	Number of failed samples	Failure rate	p-value
2002	213	2	0.94%	0.055
2003	5470	51	0.93%	< 0.001
2004	838	10	1.19%	0.001
2005	3	0	0%	0.752
2007	1310	40	3.05%	0.573
2008	10,803	315	2.92%	0.004
2009	7052	466	6.61%	< 0.001
2010	8097	225	2.78%	0.002
2011	243	10	4.12%	0.502
2012	101	18	17.82%	< 0.001
2013	95	2	2.11%	0.512
2014	33	1	3.03%	0.924
2015	22	1	4.44%	0.755
2016	10	0	0%	0.564
2017	43	3	6.98%	0.190
TOTAL	**34,333**	**1144**	**3.33%**	

Table 13. The number of all analysed samples, genotyping data quality control failed samples and the failure rate by initial sample collection year. The chi-squared test p-values are shown in the last column. p-values < 0.05 labelled with yellow, p-values < 0.001 labelled with green.

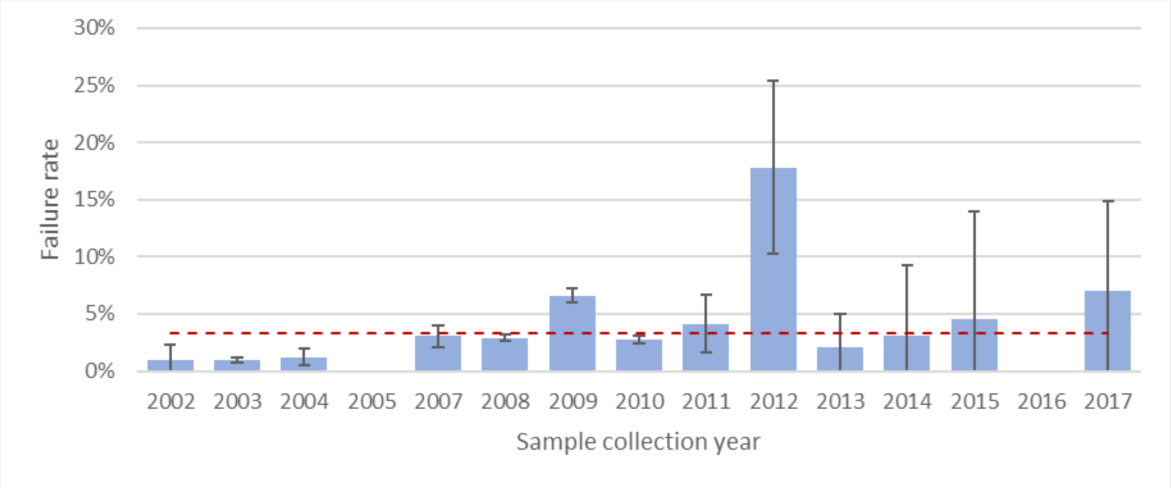

Figure 10. Failure rates with 95% confidence intervals for different sample collection years. The confidence interval calculations are based on the number of analysed samples and failure rates from each sample collection year and represent the expected failure rate for potential other samples from the same year. The dashed red line indicates the average failure rate over all samples of 3.33%.

The initial analysis suggests a difference in failure rates of samples collected in different years. Overall, samples collected from 2002 to 2005 have lower failure rates than samples collected from 2007 to 2017 (0.97% versus 3.89%, chi-squared test $p<0.001$).

These intervals coincide with the initial sample collection period in the biobank between 2002 and early 2005, a pause in sample collection until 2007 and a second sample collection period from 2007 to 2011, with relatively fewer samples collected from 2012 onwards (18).

Based on this information, the samples are pooled into two sample collection time groups for the following analysis steps – samples collected from 2002 to 2005 (total number of samples 6524 and number of failed samples 63) and samples collected from 2007 to 2017 (total number of samples 27,809 and number of failed samples 1081).

3.6.2 Sample collection time groups and agarose gel electrophoresis

The agarose gel electrophoresis results for samples in the 2 sample collection time groups are shown in Table 14.

Sample collection time group	Total number of samples	Samples in agarose gel electrophoresis	Intact failed samples	Degraded failed samples	Missing failed samples	Intact control samples	Degraded control samples	Missing control samples
2002-2005	6524	150	62	0	1	87	0	0
2007-2017	27,809	2138	368	127	586	961	73	23

Table 14. Agarose gel electrophoresis results by sample collection time group.

The agarose gel electrophoresis results show that in addition to a difference in genotyping quality control failure rate (0.97% vs 3.89%), there is an even larger difference in agarose gel electrophoresis issues between the two sample collection time groups.

The ratio of failed samples that are missing or degraded in agarose gel electrophoresis is 1.59% in the 2002-2005 sample collection time group and 65.96% in the 2007-2017 sample collection time group. In addition, there are no control samples that were missing or degraded in agarose gel electrophoresis in the 2002-2005 sample collection time group, while in the 2007-2017 sample collection time group 9.08% of the control samples were either degraded or missing.

Also visible in the data is a preference for the 2007-2017 sample collection time group for the positionally matched control samples, which was not expected, as the samples were chosen based on the positions in the sample release plates (as described in 2.2.2.1) without consideration for sample collection times. Nevertheless, while the 2007-2017 sample collection time group consists of 81.0% of all samples, 92.4% of positionally matched controls are from that time group. A chi-squared test shows that the distribution of control samples between the sample collection time groups is significantly different from the expected uniform distribution ($p<0.001$).

3.6.3 Sample collection time groups and spectrophotometric data

Since the analysis in 3.6.2 revealed that 99.9% (809 out of 810) of all samples that were degraded or missing in agarose gel electrophoresis are from the 2007-2017 sample collection time group, the sample grouping by agarose gel electrophoresis results that was used in 3.5 is not applied here, and the spectrophotometric data is analysed only in respect to the sample collection time groups. The sample exclusions based on spectrophotometric data issues described in 3.5.1 are still applied.

The spectrophotometric data for samples in the two sample collection time groups are shown in Table 15, Table 16 and Table 17.

	All samples	2002 to 2005 sample collection time group	2007 to 2017 sample collection time group
Number of samples	33,723	5930	27,793
Minimum value (ng/µl)	42	45	42
Maximum value (ng/µl)	103	103	103
Average (ng/µl)	75.8	76.6	75.6
Standard deviation (ng/µl)	11.5	12.8	11.2
Median (ng/µl)	77	77	77

Table 15. Descriptive statistics of measured DNA concentration data for the two sample collection time groups.

While the average concentrations of the two sample collection time groups are very similar, a Mann-Whitney U test still showed a significant difference between the concentrations ($p<0.001$).

	All samples	2002 to 2005 sample collection time group	2007 to 2017 sample collection time group
Number of samples	33,723	5930	27,793
Minimum value	1.34	1.44	1.34
Maximum value	2.58	2.19	2.58
Average	1.864	1.862	1.865
Standard deviation	0.062	0.059	0.063
Median	1.86	1.86	1.86

Table 16. Descriptive statistics of measured A260/A280 ratio data for the two sample collection time groups.

Again, while the differences between the average values for the 2 groups are very small, a Mann-Whitney U test shows a significant difference between the A260/A280 ratio values ($p=0.001$).

	All samples	2002 to 2005 sample collection time group	2007 to 2017 sample collection time group
Number of samples	33,723	5930	27,793
Minimum value	0.10	0.14	0.10
Maximum value	279.21	279.21	188.98
Average	2.245	4.092	1.851
Standard deviation	3.294	6.932	1.426
Median	1.9	3.39	1.83

Table 17. Descriptive statistics of measured A260/A230 ratio data for the 2 sample collection time groups.

A Mann-Whitney U test was used to confirm the significance of the difference between the A260/A230 ratio values of the two groups ($p<0.001$).

3.6.3.1 A260/A230 ratio histograms

Due to the very high difference of A260/A230 ratio values between the two sample collection time groups, the histograms of both groups were constructed to compare the differences in distribution between the groups. The histograms are shown in Figure 11.

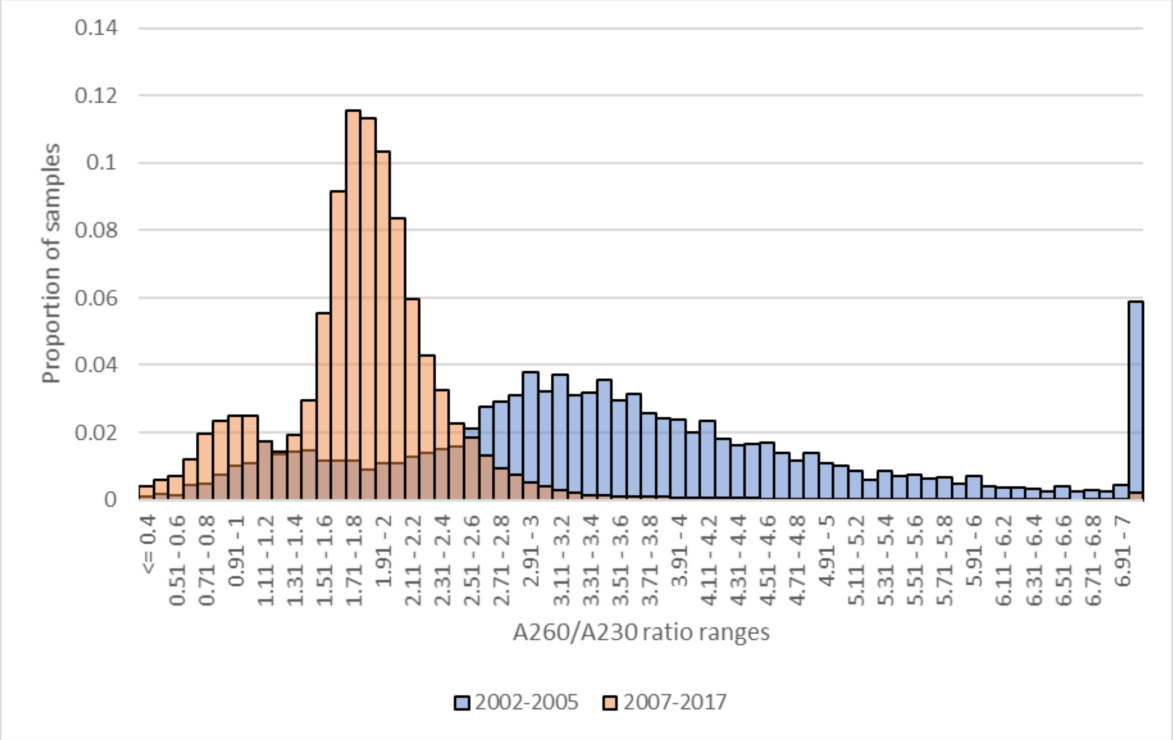

Figure 11. The histograms of A260/A230 ratio values for sample collection time groups. Bin sizes of 0.1 are used in the range of 0.4 – 7.0 and outliers are pooled together in the first and last bin. Separate histograms are used for each sample collection time group, so the sum of the heights of columns is 1 for each group, regardless of the different numbers of samples in groups.

The 2002-2005 sample collection time group has a peak at the 2.91 – 3.00 bin, with a long tail extending past the A260/A230 ratio value of 7.0. The 2007-2017 sample collection time group has a peak at the 1.71 – 1.80 bin, with a much shorter tail towards larger values.

Both sample collection time groups show an additional, smaller peak at lower A260/A230 ratio values, which could be associated with vacuum concentrated samples with higher buffer concentrations than the blanking solution used in spectrophotometric measurements. To verify this, all the A260/A230 ratios of all 4529 samples that were flagged as having been vacuum concentrated during sample processing were removed from the histograms. These histograms are shown in Figure 12.

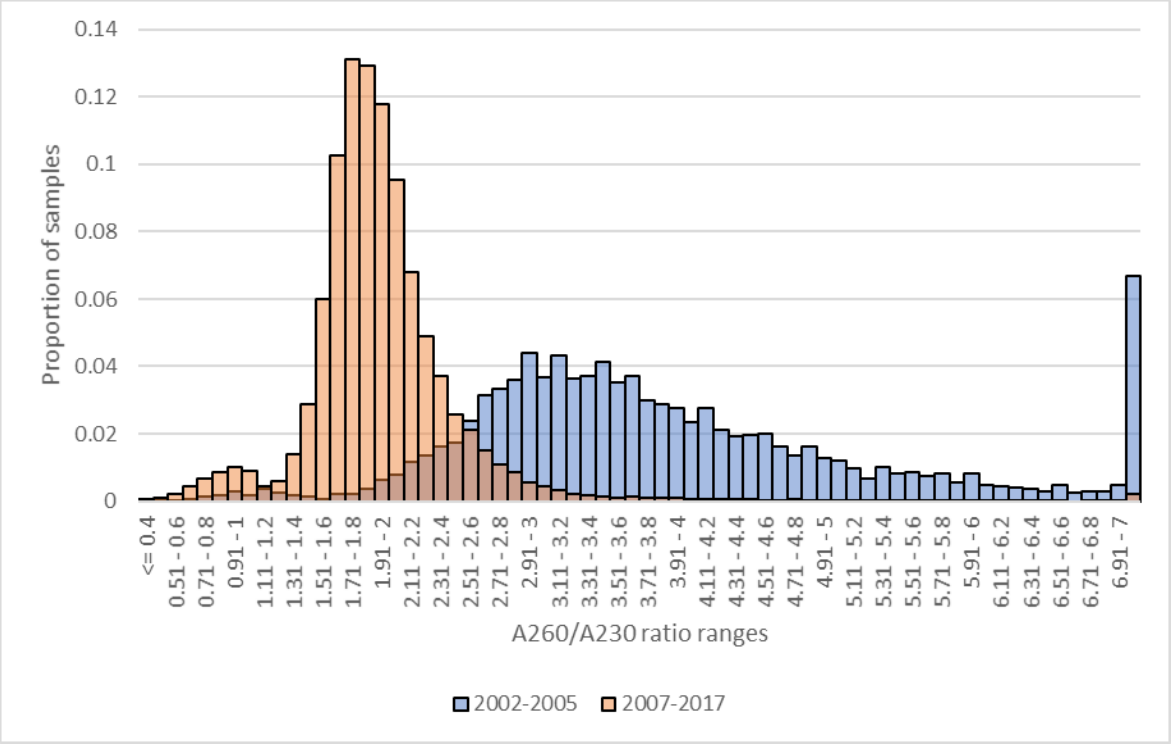

Figure 12. The histograms of A260/A230 ratio values for sample collection time groups without samples flagged as having been vacuum concentrated during sample processing. Bin sizes of 0.1 are used in the range of 0.4 – 7.0 and outliers are pooled together in the first and last bin. Separate histograms are used for each sample collection time group, so the sum of the heights of columns is 1 for each group, regardless of the different numbers of samples in groups.

Without vacuum concentrated samples, the peaks at lower A260/A230 ratio values are noticeably smaller, confirming the connection to vacuum concentrated samples, however slight peaks remain which need to be explained differently.

3.7 Analysis of sample positions on release plates and genotyping arrays

3.7.1 Sample exclusions

For the analysis of the possible effect of the positions of samples on the sample release plates and genotyping arrays on the genotyping data quality, the genotyping data quality control failed samples that were found to be missing or degraded on agarose gel electrophoresis were excluded, as for those samples the DNA degradation already presents a likely cause for failing genotyping data quality control, which could hide the possible effects analysed in this chapter.

Excluding the 587 missing and 127 degraded genotyping data quality control failed samples, the total number of samples in this analysis is 33,619, with 430 samples (1.28%) out of these failing genotyping data quality control.

3.7.2 Sample release plate position analysis

For this analysis, the numbers of all samples and failed samples were counted for each position of the 96-well sample release plates and the failure rates were calculated for each position. A chi-squared test was used to test the distribution of genotyping quality control failed and passed samples of each sample position, row and column for significant differences from the expected distribution based on the average failure rate. The failure rate data for all positions is shown in Table 18 and data by sample rows and sample columns in Table 19 and Table 20.

	1	2	3	4	5	6	7	8	9	10	11	12
A	**2.54%***	2.25%	2.01%	2.01%	1.13%	1.69%	**2.55%***	0.85%	2.30%	1.70%	1.70%	**4.29%***
B	**3.14%***	1.98%	1.42%	0.85%	0.29%	**0.00%***	0.57%	0.87%	0.85%	1.41%	0.85%	1.99%
C	0.85%	1.45%	0.57%	0.57%	0.57%	2.27%	**3.13%***	0.86%	0.29%	1.14%	1.16%	2.02%
D	1.71%	0.85%	0.57%	0.57%	0.28%	0.57%	0.28%	0.29%	0.58%	1.14%	**0.00%***	0.57%
E	**3.40%***	0.57%	1.45%	1.42%	0.57%	0.56%	2.00%	0.57%	0.29%	1.73%	1.14%	1.14%
F	**3.45%***	1.13%	1.43%	0.86%	0.29%	0.29%	0.57%	0.85%	**0.00%***	1.41%	0.57%	1.98%
G	1.16%	0.29%	0.58%	1.99%	0.57%	0.86%	1.42%	0.58%	0.86%	0.57%	1.43%	1.72%
H	**5.41%***	**2.60%***	1.71%	1.16%	1.14%	0.86%	1.15%	0.29%	1.15%	1.71%	1.44%	**2.87%***

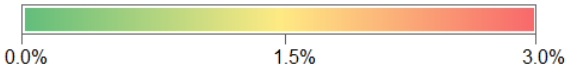

Table 18. The average failure rate for samples in all sample release plate positions. The failure rates are colour coded for readability. Values labelled with asterisk and underlined in bold (a total of 13 positions) differ significantly from the average failure rate of 1.28% based on a chi-squared test ($p<0.05$).

Plate Row	Total number of samples	Number of failed samples	Failure rate
A	4223	88	**2.08%***
B	4217	50	1.19%
C	4196	52	1.24%
D	4207	26	**0.62%***
E	4205	52	1.24%
F	4210	45	1.07%
G	4176	42	1.01%
H	4185	75	**1.79%***
TOTAL	**33,619**	**430**	**1.28%**

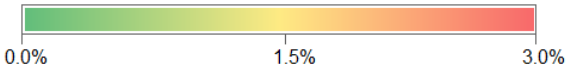

Table 19. The total number of samples, number of failed samples and the failure rate for samples by sample release plate rows. The failure rates are colour coded for readability. Values labelled with asterisk and underlined in bold differ significantly from the average based on a chi-squared test ($p<0.05$).

Plate column	Total number of samples	Number of failed samples	Failure rate
1	2803	76	**2.71%***
2	2809	39	1.39%
3	2790	34	1.22%
4	2800	33	1.18%
5	2806	17	**0.61%***
6	2808	25	0.89%
7	2804	41	1.46%
8	2790	18	**0.65%***
9	2794	22	**0.79%***
10	2811	38	1.35%
11	2804	29	1.03%
12	2800	58	**2.07%***
TOTAL	**33,619**	**430**	**1.28%**

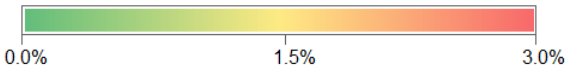

Table 20. The total number of samples, number of failed samples and the failure rate for samples by sample release plate columns. The failure rates are colour coded for readability. Values labelled with asterisk and underlined in bold differ significantly from the average based on a chi-squared test ($p<0.05$).

The data indicates that the failure rates for samples on the corners and edges of sample release plates could be higher than for samples in the middle positions of sample release plates. This was tested by grouping the data from the respective positions and using a chi-squared test to check the distribution of genotyping quality control failed and passed samples for each group of sample positions for significant differences from the expected distribution based on the average failure rate. The sample position group of corners includes sample positions A1, A12, H1 and H12 (a total of 4 positions). The sample position group of edges includes sample rows A and H and sample columns 1 and 12, except the sample positions A1, A12, H1 and H12 (a total of 32 positions). The sample group of middle positions includes all other sample positions (a total of 60 positions). The data is shown in Table 21.

Position group	Total number of samples	Number of failed samples	Failure rate
Corners	1404	53	**3.77%***
Edges (excl. corners)	11,203	191	**1.70%***
Middle positions	21,012	186	**0.89%***
TOTAL	**33,619**	**430**	**1.28%**

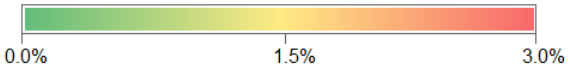

Table 21. The total number of samples, number of failed samples and the failure rate for samples in different sample position groups. The failure rates are colour coded for readability. Values labelled with asterisk and underlined in bold differ significantly from the average based on a chi-squared test ($p<0.05$).

In addition to significant differences from the overall average failure rate, further pairwise chi-squared tests confirmed significant differences in the distribution of genotyping data quality control failed and passed samples between all separate pairs of sample position groups compared to the expected distribution based on the average combined failure rate of each pair ($p<0.001$ in all cases), so the failure rate of the corner positions is significantly higher from both edge positions and middle positions and the failure rate of the edge positions is significantly higher from the middle positions.

A comparison of the failure rates of the positions on different edges of the sample release plates was carried out as well. Corner positions were excluded in all groups. A chi-squared test was used to test the distribution of genotyping quality control failed and passed samples of each sample position group for significant differences from the expected distribution based on the average failure rate over all groups. The data is shown in Table 22.

Position group	Total number of samples	Number of failed samples	Failure rate
A2 - A11	3519	64	1.82%
H2 - H11	3485	46	1.32%
B1 - G1	2098	48	**2.29%***
B12 - G12	2101	33	1.57%
TOTAL	**11,203**	**191**	**1,70%**

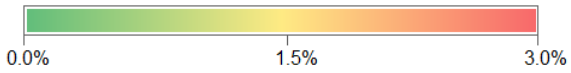

Table 22. The total number of samples, number of failed samples and the failure rate for samples for different sample release plate edges. The failure rates are colour coded for readability. Values labelled with asterisk and underlined in bold differ significantly from the average over all groups based on a chi-squared test ($p<0.05$).

3.7.3 Genotyping array position analysis

For this analysis, the positions of samples on 96-well sample release plates are converted to positions on 24-sample genotyping arrays, with samples in the positions A1-B12, C1-D12, E1-F12 and G1-H12 on sample release plates each going to positions A1-B12 on different genotyping arrays, see Figure 13.

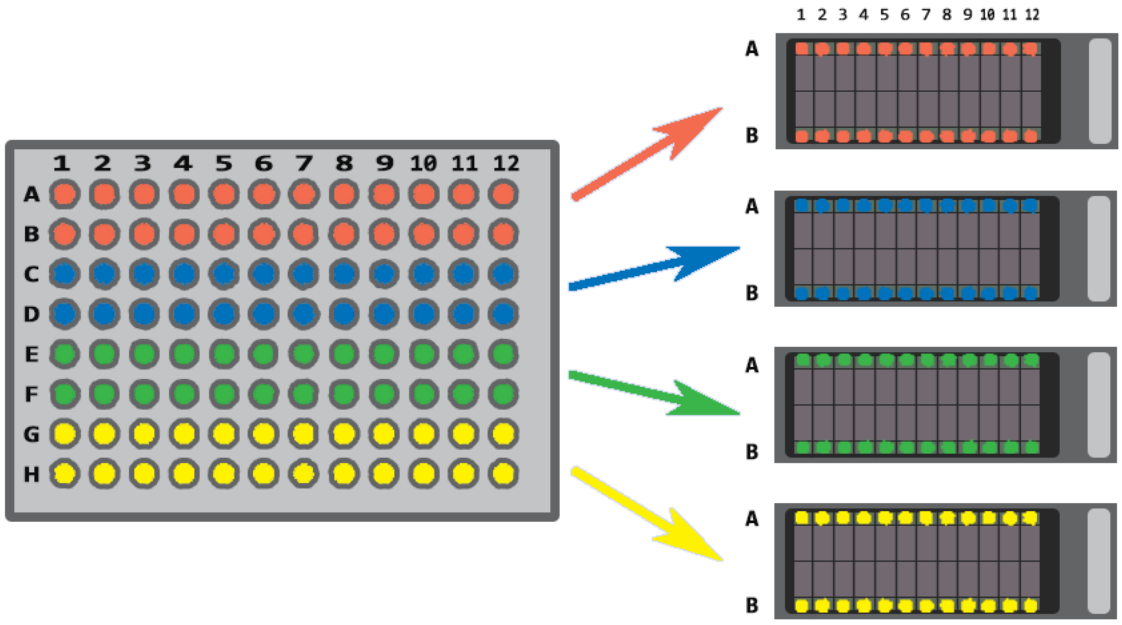

Figure 13. Illustration of the transfer of DNA samples from 96-well sample release plates onto 24-sample genotyping arrays during the genotyping workflow. Image of genotyping array adapted from illumina.com.

For each position on the genotyping arrays, the numbers of all samples and failed samples were counted and the failure rates were calculated. A chi-squared test was used to test the distribution of genotyping quality control failed and passed samples of each array position for significant differences from the expected distribution based on the average failure rate. This data is shown in Table 23.

Array position	Total number of samples	Number of failed samples	Failure rate	Failure rate	Number of failed samples	Total number of samples	Array position
A1	1403	28	**2.00%***	**3.43%***	48	1400	B1
A2	1401	16	1.14%	1.63%	23	1408	B2
A3	1391	16	1.15%	1.29%	18	1399	B3
A4	1400	21	1.50%	0.86%	12	1400	B4
A5	1405	10	0.71%	**0.50%***	7	1401	B5
A6	1409	19	1.35%	**0.43%***	6	1399	B6
A7	1406	32	**2.28%***	**0.64%***	9	1398	B7
A8	1397	10	0.72%	**0.57%***	8	1393	B8
A9	1395	13	0.93%	**0.64%***	9	1399	B9
A10	1399	18	1.29%	1.42%	20	1412	B10
A11	1397	19	1.36%	0.71%	10	1407	B11
A12	1397	32	**2.29%***	1.85%	26	1403	B12
TOTAL	**16,800**	**234**	**1.39%**	**1.17%**	**196**	**16,819**	**TOTAL**

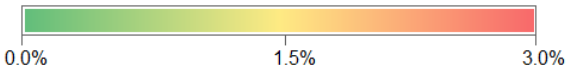

Table 23. The total number of samples, number of failed samples and the failure rate for samples in different genotyping array positions. A two-sided table is used to illustrate the adjacency of the A and B rows of the genotyping arrays. The failure rates are colour coded for readability. Values labelled with asterisk and underlined in bold (a total of 9 positions) differ significantly the average failure rate of 1.28% based on a chi-squared test ($p<0.05$).

As expected from the data in 3.7.2, the analysis reveals that the failure rate is significantly higher at both physical ends of the genotyping arrays (positions A1, B1, A12 and B12 - coinciding with columns 1 and 12 of the sample release plates, which showed significantly higher than average failure rates as well).

The A-row on the genotyping arrays had an average failure rate of 1.39% and the B-row had an average failure rate of 1.17%, but the difference is not significant (chi-squared test, $p=0.328$). However, when excluding the positions A1, B1, A12 and B12, with the significantly higher failure rates that can be linked to sample release plate columns 1 and 12, and only comparing the failure rates of genotyping array positions A2-A11 (1.24%) and B2-B11 (0.87%), the difference is significant (chi-squared test, $p=0.026$).

4 Discussion

4.1 Summary of results

The most significant result regarding DNA sample quality and its effects on genotyping data quality control results that was obtained in this Master's thesis is that there is a significant relation between the agarose gel electrophoresis results and the genotyping data quality. While no published results regarding DNA degradation and genotyping array results were identified, this is consistent with published results regarding DNA integrity and next-generation sequencing results with both naturally and artificially degraded DNA samples (7,27). This is also consistent with published results showing that DNA samples with poor agarose gel electrophoresis results give poorer results in TaqMan assay genotyping and PCR amplification (6).

More than 62.4% of samples that failed genotyping data quality control were either missing or degraded on agarose gel electrophoresis, while only 8.4% of control samples that passed genotyping data quality control had similar issues. The difference is even clearer when only completely missing samples are considered, with 51.3% of failed samples missing on agarose gel and only 2.0% of control samples showing up as missing on agarose gel.

The significant relation between agarose gel electrophoresis results and genotyping data quality is also reflected in the average call-rates of different sample groups, with samples that were missing on agarose gel having significantly lower call-rates in both failed samples as well as control samples. This indicates that even though some samples that were missing on agarose gel electrophoresis passed genotyping data quality control, the effect on sample quality is still noticeable.

The spectrophotometric data analysis revealed more significant differences between the missing and degraded samples compared to all other samples – the DNA concentrations for degraded and failed samples are significantly higher, as

well as A260/A280 ratios. The A260/A230 ratios were shown to be highly variable both overall and in different sample groups, but a detailed analysis of the distribution of all missing and degraded samples revealed a significant increase of failed samples in the A260/A230 ratio range of about 1.35 – 1.87, with a peak at the values of 1.66-1.67. Surprisingly, a weaker but similar trend was also identified for the group of intact control samples.

The analysis of initial sample collection time again showed significant differences between different sample groups, which can be generalized as samples collected from 2002-2005 having significantly lower failure rates (0.97%) than samples collected from 2007-2017 (3.89%), which also coincides with different sample collection periods in the biobank.

The analysis of agarose gel electrophoresis results of the two sample collection time periods confirmed that the difference in sample quality can also be seen in agarose gel electrophoresis results, with only one missing sample detected for the 2002-2005 sample collection period, whereas for the 2007-2017 sample collection period, 65.96% of genotyping failed samples and 9.08% of control samples were either missing or degraded. The same analysis also showed a significant preference for the 2007-2017 sample collection time group in the positionally matched control samples, indicating a bias in the control sample selection process.

Spectrophotometric data analysis for the two sample collection time groups showed significant but small differences in DNA concentration and A260/A280 ratios between the groups, and a very large and significant difference in the A260/A230 ratios – the 2002-2005 sample collection time group had an average A260/A230 ratio of 4.09 and the 2007-2017 sample collection time group had an average of 1.85. The differences were also visualized using histograms of the A260/A230 values of both groups, implying that a clear difference must exist between the two sets of samples.

4.2 Agarose gel electrophoresis

The results obtained clearly indicate that a large proportion of the genotyping data quality control failed samples are linked to partially or completely degraded DNA samples. While degraded and missing samples were also found in the positionally matched controls that were successfully genotyped, the difference with failed samples is very high and could account for a majority of the samples that failed genotyping data quality control.

The interpretation of degraded samples is quite straightforward, with reduction in DNA integrity being visible and is likely caused due to either biological contaminants in the sample or by outside physical factors such as freeze-thaw cycles or storage at elevated temperatures during sample processing. While some studies have shown that DNA samples in TE buffer are stable after a month at +4°C (14) or even several months at room temperature or more than 10 freeze-thaw cycles (28), and found no decrease in RT-PCR performance after being stored at -20°C for 1 year (29), it is likely that potential contaminations could cause more rapid DNA degradation as well. It has been reported that dilute and small volume DNA samples may have variable stability during storage (23).

The reasons for the apparent complete absence of double-stranded DNA in some samples could be as follows:

1. Similar to visibly degraded DNA samples, the reason could be DNA degradation to the point that the remaining DNA fragments are short enough to migrate through the used agarose gel rapidly and diffusely, reducing the fluorescence signal of ethidium bromide below the visibility threshold in the gel images. Due to the established link to genotyping data quality control failures, this interpretation is the most likely for most samples. However, since some samples that were completely missing on agarose gel electrophoresis images were found to have successfully passed genotyping data quality control, alternatives should be considered.

2. Another potential cause for samples being missing on agarose gel electrophoresis images could be a simple pipetting error in loading the DNA sample onto the gel. This could include both improper aspiration errors, where the DNA sample is not successfully aspirated into the pipetting tip, or improper dispensing errors, where the DNA sample is not correctly loaded into the gel well and is lost in the electrophoresis buffer. This cause would be unrelated to the actual DNA integrity of the sample.

3. A third potential cause for samples being missing on agarose gel electrophoresis images could be low concentrations of DNA in samples, which lead to lower fluorescence intensities. The visibility of DNA on gel electrophoresis images could then be further reduced if the sample is degraded, which leads to a more diffuse image. However, since all the DNA samples in this case had DNA concentration measurements and adjustments prior to agarose gel electrophoresis, this explanation would only apply if some samples were incompletely homogenized and a lower concentration part of the aliquot was loaded onto the agarose gel, making it unlikely to be a significant part of the observed results.

4.2.1 Potential for application

By analysing all samples on agarose gel electrophoresis prior to genotyping and then excluding all samples that were missing on gel images could have reduced the number of failed samples more than 50% - from 1144 to 587. Adding the degraded samples to the exclusion list could have achieved a further reduction in failed samples to 430 (a total reduction of 62.4%).

However, the rate of false positives for sample exclusion must be considered as well. Extrapolating the agarose gel electrophoresis results of the 1144 control samples to all other samples that passed genotyping data quality control, the expected results for the entire set of 34,333 samples can be estimated as 30,834 samples intact, 2245 samples degraded and 1254 samples missing on agarose gel electrophoresis. The estimated failure rates would then be 1.39% for intact samples, 5.66% for degraded samples and 46.81% for missing samples.

The preferred course of action here would depend on the context of the samples being analysed – while excluding samples with high expected failure rates would reduce unnecessary costs of unsuccessful genotyping experiments, it would also introduce a further reduction of the sample set size by falsely excluded samples, which may be an even more important consideration in some cases.

While excluding the 1254 samples with an expected failure rate of 46.81% might be the preferred choice in most cases, excluding a further 2245 samples with an expected failure rate of 5.66% might be a more difficult decision. However, if the excluded samples could be replaced with better quality aliquots of the same samples, this trade-off could be avoided. In the case of limited resources for genotyping and an availability of suitable samples, replacing samples with a potentially higher failure rate with different samples altogether could be an option as well.

In a biobanking context, this also highlights the importance of storing samples in several aliquots, so a working aliquot that is faced with potentially adverse conditions such as numerous freeze-thaw cycles or periods of elevated temperatures could be replaced with a higher quality aliquot when necessary.

4.2.2 Other considerations

While agarose gel electrophoresis is a simple and low-cost method for estimating DNA integrity, it can only provide a qualitative assessment for DNA integrity. There are commercially available automated electrophoresis instruments that provide a quantitative assessment as well, such as Agilent Tapestation, PerkinElmer LabChip GX Touch HT or Bio-Rad Experion. These automated instruments measure peak DNA fragment lengths and may also output an algorithmically quality score for the DNA integrity, along with an estimated DNA concentration value (30,31).

In the experience of the author and the Biobank Lab at the Estonian Genome Center, these methods do not offer a significant advantage in identifying degraded samples compared to experienced lab technicians assessing agarose gel

electrophoresis images. While conventional agarose gel electrophoresis does have limitations in resolving DNA molecules larger than 40-50 kb (32), the resolution appears to be sufficient to identify degraded DNA samples. The clear link between visually estimated genomic DNA quality from agarose gels, the quantified assessment of genomic DNA quality and the effect of DNA degradation on the quality of downstream analysis results has been shown elsewhere as well (8). However, others have found that using only visual estimations of electrophoresis results may have insufficient reproducibility when comparing different observers, and a combined approach of using both qualitative and quantitative quality estimations may prove better results (33). In analysing large numbers of samples, the automated and digital assignment of quality values to each sample could also significantly reduce the workload compared to manually documenting the agarose gels, processing and adjusting the images, labelling the samples, documenting the quality assessment of each sample and finally archiving all the images and data for future reference. Due to the mostly manual work required for agarose gel electrophoresis imaging and processing, the potential for human error must be considered at all points as well – particularly in cases where a significant number of samples need to be labelled as degraded or missing in images.

The added value of a high detection range and simultaneous DNA concentration estimation, when compared to agarose gel electrophoresis, could be considered as well – as discussed above, a potential failure in successfully loading a DNA sample on the agarose gel would be indistinguishable from a truly degraded sample on the gel images, while the concentration estimation as well as a detailed electropherogram provided by the automated electrophoresis instruments would enable to detect such issues and possibly prevent false exclusions of some samples. The automated electrophoresis instruments have been shown to enable potential detection of denatured ssDNA as well (34). The potential advantage would of course depend on the actual rate of issues in loading the agarose gel successfully, which could be estimated by re-running any samples found to be missing on agarose gel electrophoresis. Even a small potential reduction in the number of falsely excluded samples could increase the diagnostic power of checking DNA integrity even further.

4.3 Spectrophotometric data

4.3.1 Issues and sample exclusions

The analysis of spectrophotometric data was started with the exclusion of the data of a number of samples, due to the unexpected detection of negative A260/A230 ratio values in a total of 254 samples, which were caused by measured negative absorbance values at 230 nm for those samples. Since 219 of those samples were positioned in a single measurement batch of a total of 576 measured samples, an issue with the blanking measurement for that batch can be suggested as the likely cause of the unexpected A260/A230 ratio values for those samples.

Possible issues with blanking could be the following:

1. Instead of using a TE-buffer that had been used for dissolving the DNA samples, a wrong solution was used for the blanking measurement. The wrong blanking solution would have needed to have a higher than expected absorbance at the 230 nm wavelength, causing the DNA samples in TE buffer to have comparatively lower absorbances, which for some samples reached into negative values, causing the negative A260/A230 ratios.

2. The sample measurement pedestals were insufficiently cleaned before making the blanking measurement and the contamination present on the pedestals caused higher than expected absorbance at the 230 nm wavelength during the blanking measurement. When the blanking solution was then subsequently wiped away from the measurement pedestals before making the first sample measurements, the contamination could have been wiped away as well, returning the 230 nm absorbances to their normal values for any further measurements made in that batch. This would have also caused the measured DNA samples to have comparatively lower absorbances, leading to negative A230 values and A260/A230 ratios for some samples.

Should the second option be correct, it cannot be ruled out that the same issue could be present in some other sample measurement batches as well, although with less pronounced effects – lowering the measured A230 values for samples, but not as much as to cause negative values in any samples. Since an 8-channel NanoDrop instrument - which uses a separate blanking measurement for each channel - was used for the spectrophotometric measurements, it also cannot be ruled out that in some batches, only some individual sample rows could be affected by the issue.

Due to normal variations in A260/A230 ratios and the reduced effect of this issue on the A260/A280 ratios and estimated concentration values (as described in 3.5.1, Figure 4), this issue could be very difficult to notice during routine sample measurement, without specific efforts to detect it. Therefore, a careful cleaning of sample measurement pedestals of the spectrophotometer prior to blanking measurements as well as between consecutive sample measurements is essential in ensuring reliable spectrophotometric data.

It could be speculated that due to the difficulty to detect the nature of this issue, the same problem could be behind highly variable as well as negative A260/A230 ratios described elsewhere as well (9).

In the case of the further 35 samples that also displayed negative A260/A230 ratios, but where a similar batch effect was not found, it is possible that the issues are with the individual samples – a possible cause for negative absorbances at 230 nm could be water having been used for sample dilutions, instead of TE buffer, which could also cause lower than expected absorbance values when compared to a TE buffer blank. Out of these 35 samples, 8 are logged as having been vacuum concentrated during initial sample processing. If some or all of these 8 samples were vacuum concentrated due to a concentration measurement error – for example poorly homogenized samples giving a lower than actual concentration measurement – they would have then been diluted to volumes higher than initial using Milli-Q water, reducing the TE buffer concentration in the sample and potentially causing measured negative A230 values. Finally, potential measurement errors cannot be completely ruled out as well – while a large air

bubble inside the sample column during sample measurement usually causes measurement errors or easily noticeable changes in the absorbance spectra, it is possible that smaller air bubbles or other issues with the sample column could cause more subtle issues, including negative A260/A230 values.

4.3.2 DNA concentrations and A260/A280 ratios

The analysis of spectrophotometric data revealed significant differences in the estimated concentrations between different sample groups, with the all missing and degraded samples having about 6.5% higher average concentrations than all other samples - 80.5 ng/µl and 75.6 ng/µl, respectively.

The higher concentration values of missing and degraded samples could be linked to the fact that both single-stranded DNA and oligonucleotides have higher absorbances at the 260 nm wavelength than double-stranded DNA (35), and thus a DNA sample with a given quantity of DNA present would give higher absorbance readings and higher calculated DNA concentrations as the sample degrades into oligonucleotides or denatures. The increase in absorbance for single-stranded DNA has shown to be caused by the DNA hyperchromic effect (36), and a similar effect could be behind the increase in absorbance in oligonucleotides as well. Although some analysed samples were also vacuum concentrated or diluted during the initial sample handling phase described in 2.1.2, it appears that the effect of sample degradation on the concentration values was not masked by this.

While the differences in average concentrations are high and would thus be easily detected even while accounting for possible measurement inaccuracies inherent in the used NanoDrop ND-8000 instrument, the application of this difference in concentrations for identifying possibly degraded samples in the sample set used in this Master's thesis would have been prevented by the even higher inter-sample variability of DNA concentrations. The standard deviation of average concentrations over all samples was 11.5 ng/µl.

On the other hand, this detected effect of DNA degradation on spectrophotometrically measured DNA concentrations could possibly be applied in

other cases where the DNA concentrations of stored samples are known, to detect possible sample degradation over time by remeasuring the sample concentrations and looking for unexpectedly increased DNA concentrations. There are, however, a number of considerations that need to be taken into account, including (a) possible unrelated increase in sample concentrations through evaporation of water during sample storage; (b) changes in sample concentration due to precipitation of DNA during freezing or thawing the samples; (c) the need for well-homogenized samples where repeated measurements of the same sample could be assumed to be close to identical and (d) the need for well-calibrated instruments at both timepoints to get reliable measurements over potentially long storage times of samples.

The fact that no other significant effects regarding measured DNA concentrations and genotyping data quality control results were identified (intact failed samples had no significant difference from intact control samples or other control samples) suggests that the used 50-100 ng/µl DNA concentration range was within the optimal performance window of the genotyping arrays, and further studies could be useful to identify whether samples with a wider range of DNA concentrations could be suitable for genotyping arrays.

There was also a significant difference in A260/A280 ratios between different sample groups, with the all missing and degraded samples group having about 2.7% higher average A260/A280 ratios than all other samples - 1.91 and 1.86, respectively.

This difference should be evaluated in parallel with the differences in average concentrations since higher concentrations mean higher absorbances at the 260 nm wavelength. Since an increase is also seen in the average A260/A280 ratios, it follows that the effect of DNA degradation or denaturation on the absorbances of samples is lower at the 280 nm wavelength than at the 260 nm wavelength – the measured A280 values of samples have not increased as much as A260 values, leading to an increase of the average A260/A280 ratios.

Since the A260/A280 ratios of DNA samples should be independent of DNA concentration (pure DNA samples are expected to have A260/A280 ratios of 1.80 to 2.00 regardless of DNA concentration), it could be expected that in pure DNA samples the absorbances at both 260 nm and 280 nm are determined by the DNA content and thus a change in DNA concentration should not be expected to change the A260/A280 ratios – even more so if the difference in concentrations is relatively small. The literature on the effects of changing DNA concentrations on A260/A280 ratios is sparse, but in one study with 8 vacuum concentrated DNA samples, no significant effect of vacuum concentrating on the A260/A280 ratios was found (37). This was also confirmed by examining the concentrations and A260/A280 ratios of the samples being logged as having been vacuum concentrated in this study – the vacuum concentrated samples showed a similar increase in concentrations as missing and degraded samples (average concentration 80.1 ng/µl, a 6.7% increase compared to all other samples), but the change in A260/A280 ratios was significantly lower than in the missing and degraded samples (average A260/A280 ratio 1.87, a 0.5% increase compared to all other samples.

While this observed increase in average A260/A280 ratios for missing and degraded samples is significant and quite high, it is still lower than the standard deviation of average A260/A280 ratios over all samples (0.06), so taken alone and as a single measurement the effectiveness for identifying potentially degraded DNA samples in the sample set is limited, as was the case with DNA concentrations. Also, average A260/A280 ratio values for missing and degraded samples as well as all other samples are both well within the expected range of 1.80 to 2.00 for pure DNA samples and thus should not be indicative of any problems with contamination when viewed as individual measurements.

However, if previous spectrophotometric measurements for samples are available, this effect could potentially be combined with the method of detecting unexpected increases in DNA concentrations described above, to increase the potential detection power. By looking for samples that have not only increased concentrations but also increased A260/A280 ratios, DNA samples that have degraded during storage and handling could potentially be differentiated from

samples that have increased concentrations through water evaporation or other issues.

4.3.3 A260/A230 ratios

The initial analysis of A260/A230 ratios first revealed the high variability of these ratios between different samples – over all samples, the average A260/A230 ratio was 2.25, but the standard deviation was 3.29 and the overall range was 0.10 to 279.21 – which complicated the analysis. While significant differences were found between several sample groups, no clear conclusions could be made based on only average values, as samples that were missing on agarose gel electrophoresis had average A260/A230 ratios more similar to intact samples than to degraded samples. Also, the sample group of other control samples had significantly different A260/A230 ratios from matched controls that were intact on agarose gel electrophoresis, which was unexpected.

Further analysis by constructing A260/A230 ratio subgroups and comparing the distributions of different sample groups between those subgroups offered more insight into why the comparison of average values proved to be impractical, as it appeared that the samples that were missing or degraded on agarose gel electrophoresis were not associated with either the highest or lowest A260/A230 ratio values, but rather with a specific range of A260/A230 ratios (between about 1.35 and 1.87). Surprisingly, a similar association was also found for matched control samples that were intact on agarose gel electrophoresis.

Further insights into the A260/A230 ratios of samples and their connection to DNA quality was provided by the analysis of A260/A230 ratios, failure rates and agarose gel electrophoresis results of the two initial sample collection time groups. It was determined that samples collected between 2002 and 2005 had significantly lower genotyping data quality control failure rates, significantly lower sample degradation seen on agarose gel electrophoresis and significantly higher A260/A230 ratios, compared to samples collected between 2007 and 2017.

This leads to the suggestion that the genotyping data quality control failed samples that were missing or degraded on agarose gel electrophoresis (62.4% of all failed samples) could be directly linked to the 2007-2017 sample collection time group and the observed lower A260/A230 ratios for those samples. The lower A260/A230 ratios in turn suggest a possible issue with higher contamination carryover in the DNA extraction process for those samples, leading to higher absorbance in the 230 nm wavelength seen in spectrophotometric measurements. While lower than optimum A260/A230 ratios have been shown to have no effect on PCR amplification and SNP (single nucleotide polymorphism) genotyping immediately after DNA extraction in some cases (25), it is possible that over time contamination could lead to DNA degradation and detectable effects on downstream analyses. A study utilizing similar steps of using ammonium acetate and alcohol to precipitate DNA during DNA extraction as the samples used in this Master's thesis suggests that the low A260/A230 ratios could primarily be caused by ammonium acetate carryover, as ammonium acetate has high absorbance at 230 nm (38).

The histograms of A260/A230 values of the two sample collection time groups revealed that while the peaks of both histograms are indeed clearly different, both groups also have a small peak at lower A260/A230 values as well (Figure 11). These samples with lower A260/A230 values in both groups could be linked to vacuum concentrating, which causes the TE buffer concentrations to increase along with the DNA concentrations. Since absorbance at the 230 nm wavelength is associated with salts, among other substances, the increase in buffer concentration will lead to higher absorbance at 230 nm when compared to the regular TE buffer used for spectrophotometer blanking, which in turn leads to lower A260/A230 ratios. This was confirmed by excluding samples that were flagged as being vacuum concentrated during initial sample handling from the histograms, leading to a noticeable lowering of the secondary peaks for both sample collection time groups (Figure 12).

The secondary peaks on A260/A230 ratio histograms did not completely disappear though, requiring a different explanation than vacuum concentrating during initial sample processing. Some potential explanations for this could be proposed:

1. Some samples could have been inadvertently concentrated through water evaporation during sample processing, without the use of a vacuum concentrator. This could have been caused by leaving some sample release plates unsealed for a longer than usual time during sample processing.

2. Another potential cause for inadvertent water evaporation could be imperfect aluminium foil sealing of some sample release plates, leading to water evaporation, with the corners and edges of sample release plates being the most prone to evaporation.

3. There could be cases of vacuum concentrating samples in the biobank prior to sample release, with that information not having been made available or been considered during the initial sample processing phase.

With as many confounding factors detected and removed as possible, it could now be said that the samples that were missing or degraded on agarose gel electrophoresis (with a peak in the abundance of those samples at the A260/A230 ratio range of 1.66 to 1.67) are linked to samples collected between 2007 and 2017 (with a peak in A260/A230 ratio histogram at the 1.71 to 1.80 bin). The difference in the peak values – as well as the connection between lower A260/A230 values and potential contamination - could suggest that a further link between lower A260/A230 ratio values and sample degradation could be present within the 2007 to 2017 sample collection time group, but further studies would be needed to establish that.

4.4 Sample collection time groups

There was a significant bias between the two analysed sample collection time groups seen in (a) sample failure rates; (b) agarose gel electrophoresis results; (c) A260/A230 ratios as well as (d) distribution of positionally matched controls.

As discussed above, the combination of the first three results suggests that the 2007 to 2017 sample collection time group has higher contamination carryover in

the DNA extraction process, which leads to the degradation of some samples over time, which in turn leads to a higher failure rate for those samples in genotyping.

This difference in sample quality is surprising, because the SOPs (standard operating procedures) for sample collection, processing and storage used in both 2002 to 2005 and 2007 to 2017 were identical, which means that less obvious changes must be accountable for the significant change in DNA sample quality. A potential explanation could be provided by the fact that while the SOPs were unchanged, the period between 2005 and 2007 did include a stoppage in sample collection as well as restructuring and a change in governance for the biobank (18). This suggests that between the first and second sample collection period, the changes in governance or laboratory staff caused changes in laboratory practices or the interpretations of the SOPs, which in turn led to the differences in sample quality observed today.

This issue highlights the extreme importance of continuous and thorough monitoring of laboratory procedures as well as sample quality in a biobanking context, even if no immediate issues are apparent, because the samples potentially need to be stored for decades and still have reliable quality. While routine quality control and management procedures are in place at the Estonian Genome Center Biobank Lab as well, the complexities in the analysis of A260/A230 ratios in particular show that simple and straightforward methods, such as assessing the average values of quality control metrics, can be completely inadequate and insufficient in finding potential issues in large and diverse sets of samples.

This also stresses the importance of using the various data made available by projects involving a large number of biobank samples for concurrent quality analysis in the biobank. It is not obvious that a smaller set of samples and their genotyping data quality control results, or several smaller sets of samples in different projects would have enabled to find the significant differences in sample quality identified in this Master's thesis. It is likely that differences in other variables when using several smaller sets of analysed samples would have at least made the analysis more complicated and masked some potential issues.

The observed unplanned bias in positionally matched controls should be addressed as well. As shown by the data in 3.6.2, the samples from the 2002 to 2005 sample collection time group make up 19.0% of all samples, but only 7.6% of positionally matched controls. As the selection of control samples only relied on the positions of samples on sample release plates and genotyping arrays (as described in 2.2.2.1), this indicates that there was also a bias in the grouping of samples on sample release plates. This would also explain the uneven distribution of the A260/A230 values of intact control samples seen in 3.5.4.1, as the control samples are not evenly distributed but have a bias for the 2007 to 2017 sample collection time group and the associated distribution of A260/A230 vales.

While the randomization of samples by sample collection time was not a goal during sample release, it was nevertheless a default assumption that when no sample sorting criteria were applied, the result would be a random distribution of samples. The cause for this bias could be based on the fact that while samples that are frequently accessed from the automated storage system would have their positions randomized by numerous sample sortings, the samples used in this project were specifically the ones with no pre-existing genotyping data, possibly making them less likely to have been requested for other previous research projects as well. This would cause the samples to retain the positions they were initially loaded into the automated store in, along with the proximity to samples with similar sample collection times, as the initial loading of samples into the automated store was based on sample collection times. While not directly relevant in the context of sample and genotyping data quality, this is nevertheless a finding that should be kept in mind for future sample release projects where randomization could be more important.

4.5 Sample positions

As described in 3.7.1, the genotyping data quality control failed samples that were found to be missing or degraded on agarose gel electrophoresis were excluded for the purposes of analysing the effect of sample positions on genotyping data quality control results, as for those samples a likely cause for failure was already determined.

A separate analysis of the failure rates of samples in different positions on the sample release plates and samples in different positions on genotyping arrays was carried out, but since the two types of positions have a direct correspondence between one-another, any possible interpretations need to consider this correspondence as well.

4.5.1 Positions on sample release plates

The analysis of failure rates of samples in different positions on sample release plates revealed that all four corners of the sample release plates have significantly higher failure rates from the average, along with five more positions on the edges of the release plates and one position in the middle of the release plate. At the same time, three individual positions in the middle of the release plates had significantly lower failure rates from the average. This called for a further analysis grouping together samples in (a) the corners; (b) the edges and (c) the middle of sample release plates, which identified the failure rates as 3.77% for the corners, 1.70% for the edges and 0.89% for the middle positions of sample release plates.

The higher failure rates for the corners and edges of the sample release plates could be related to water evaporation caused by imperfect aluminium foil sealing of some sample release plates, where the corner and edge positions would be most prone to evaporation. This could have taken place during the time between making the sample release aliquots and performing the genotyping protocol, when the samples were stored at +4°C. Several issues could arise from the sample evaporation:

1. Evaporation of water from the DNA samples would lead to higher viscosities of the samples, making errors in pipetting the samples during the genotyping workflow more likely.

2. In the case of a very high proportion of the water evaporating, the resulting increase in DNA concentrations could bring the DNA concentrations outside the suitable range for the used genotyping arrays.

3. In the case of complete evaporation of the water in the sample position, the resulting desiccated DNA would be impossible to be successfully pipetted. The high throughput of samples in the Genotyping and Sequencing Core Facility could have increased the probability of this going unnoticed during the genotyping workflow.

Using a simplified assumption that the middle positions in the sample release plates with the 0.89% average failure rate represent a “background” failure rate of the overall genotyping workflow and possible unidentified DNA quality issues, the identified effect of corner and edge positions on the genotyping failure rate could account for a total of about 132 failed samples (11.5% of all failed samples and 30.7% of failed samples without detected DNA degradation).

If the identified effect is indeed caused by imperfect manual aluminium foil sealing of sample release plates, a potential solution to reduce if not eliminate this effect could be to use automated heat sealers to close the sample release plates at all times, to achieve a potentially more reliable sealing of all sample positions and reduce the possibility of human errors in manually sealing the sample release plates. This would, however, be associated with extra costs in the biobank, especially in the case where different steps of sample processing and quality control are carried out at different times and locations, requiring repeated sealing of the sample release plates and more heat-sealing equipment to be available in different locations.

4.5.2 Positions on genotyping arrays

After converting the sample positions on sample release plates to corresponding positions on genotyping arrays, the failure rates of different genotyping array positions were analysed. As predicted by the higher failure rates of sample release plate columns 1 and 12, both physical ends of the genotyping arrays (positions A1, B1, A12 and B12) that correspond to those sample release plate columns showed significantly higher failure rates from the average.

In order to determine whether the higher failure rates in sample columns 1 and 12 seen in the data should be attributed to sample release plates or the genotyping arrays, the failure rates of different sample position groups were analysed. If the higher failure rates were linked only to positions on the sample release plates, no significant differences between the failure rates of different edges of the sample release plates should be expected. However, the analysis showed that the sample positions B1-G1 (the left edge of the sample release plates) – corresponding to the column 1 on the genotyping arrays - had significantly higher failure rates than the other edges of the sample release plates. This would indicate that in addition to the effect of the sample release plate edges on the failure rates, an additional effect is added by the genotyping array positions A1 and B1 (one physical end of the genotyping arrays). No similar effect was seen for the positions B12-G12, indicating that the other physical end of the genotyping arrays does not have the same effect on genotyping failure rates.

Since the effect could potentially be linked to either one physical end of the genotyping arrays or one physical side of the sample release plates, any step in the entire sample processing and genotyping workflow could potentially be causing the detected increase in genotyping data quality control failure rates for column 1. If there is a significant left to right bias in manually sealing the sample release plates with aluminium foil seals, the cause could be the same as for the overall increase in failure rates on the edges and corners of the sample release plates described above, only slightly amplified for the left edge of the sample release plates. However, other potential causes throughout the sample processing and genotyping workflow (5) that might increase the failure rates on one end of the plates or genotyping arrays cannot be ruled out either – these could include:

1. Unexpected evaporation in any other sample processing step after the sample release phase

2. Variations in sample incubation conditions in the genotyping workflow

3. Variations in applying reagents to samples in the genotyping workflow

4. A bias in quality during the imaging of the genotyping arrays with the Illumina HiScan instrument

This effect, while significant, would nevertheless account for a relatively small number of failed samples. Using a simplified assumption that the increase in failure rate of 0.72 percentage points seen between the failure rate of positions B1-G1 and the average failure rate of all other edge positions could be applied to all samples in the column 1 of sample release plates, this difference in failure rates could account for a total of about 20 failed samples (1.7% of all failed samples and 4.7% of failed samples without detected DNA degradation). These samples would already be included in the previous estimation of about 132 failed samples linked to the edges and corners of sample release plates calculated above.

On the genotyping array row B, five middle positions B5 to B9 showed significantly lower than average failure rates, while on the genotyping array row A, a similar effect was not present but instead position A7 had a significantly higher than average failure rate.

While the average failure rates of row A (1.39%) and row B (1.17%) were not shown to differ significantly, a separate analysis excluding the positions A1, B1, A12 and B12 with the higher failure rates that can be linked to sample release plate corner and edge positions which could potentially mask any further differences between genotyping array rows, showed that positions A2-A11 had a significantly higher failure rate (1.24%) than positions B2-B11 (0.87%).

The reasons for the difference in failure rates between rows A and B on the genotyping arrays as well as the significantly higher failure rate for position A7 are difficult to estimate, but because the significant difference is between the two physical rows of the genotyping arrays, it is likely that the effect is introduced after the processed DNA samples are loaded onto the genotyping arrays. This includes the genotyping workflow steps of (a) DNA fragment hybridization; (b) washing of unhybridized DNA; (c) single-base extension and staining and (d) imaging of the genotyping arrays on an Illumina HiScan instrument (5). Each of these steps could

potentially introduce a difference between the two rows on genotyping arrays, leading to the observed difference in failure rates.

For the position A7 specifically, a potentially good explanation for the observed increase in failure rates could have been a faulty channel of a multi-channel pipette used during the genotyping workflow, but since the same 12-channel pipette is used for both genotyping array rows and a similar effect is not seen on the B-row, this cause can be ruled out.

Using a simplified assumption that the difference in failure rate of 0.37 percentage points seen between the positions A2-A11 and B2-B11 was masked by the higher failure rates of positions A1, B1, A12 and B12, and assuming that the difference could be applied to the entire A and B rows of the genotyping arrays, this difference in failure rates could account for a total of about 62 failed samples (5.4% of all failed samples and 14.4% of failed samples without detected DNA degradation). This is, however, based on a simplified assumption and, compared to the assumptions made above, it would likely be more difficult to identify and correct the cause of the increased failure rates observed.

4.5.3 Other considerations

While the analysis above looked into any possible systemic effects related to sample positions on release plates and genotyping arrays, another possible method could be to look into the individual genotyping arrays and the positions of any samples that failed genotyping data quality control. This type of analysis could potentially identify specific cases of sample handling errors which have led to the failure of several adjacent samples. However, with the systemic effects already identified and outlined above, it would likely be difficult to differentiate between systemic issues and specific pipetting or sample handling errors, so this analysis was not carried out in this Master's thesis.

4.6 Gender mismatch issues

Although further analysis of the samples with mismatches between reported genders for the sample and the gender determined from the genotype data is outside the scope of this work, all the samples were later re-genotyped using a new DNA aliquot from long-term storage.

From the 136 samples that had both a gender mismatch and a low call-rate, about 6% still had a similar result after re-genotyping (low call-rate and a gender mismatch), indicating an underlying issue with all the aliquots of that specific person. The rest of the samples passed genotyping data quality control successfully following the re-genotyping and none of the samples then had a confirmed gender mismatch issue, confirming that the gender mismatch issues were caused by low genotyping data quality, rather than an actual sample mixup.

From the 67 samples that had a gender mismatch but good genotyping data quality, about 25% had correct genders and about 75% still had a gender mismatch following the re-genotyping. The cases where a gender mismatch was resolved using a new sample aliquot most likely indicate a sample mixup issue in the biobank while transferring the samples from long-term storage to automated intermediate storage. The cases where the gender mismatch was not resolved with a new sample aliquot indicate either a mixup issue during initial sample collection and processing, or an underlying genetic issue with the biobank participant. This rate of unresolved gender issues (about 0.15% of all analysed samples) is comparable to the prevalence of sex-chromosome abnormalities in the general population, estimated by Samango-Sprouse et al. at 1 in 1439 or about 0.07% (39). However, the possibility of a sample mixup should also be considered. In medical testing, specifically pretransfusion testing of blood samples, a sample mixup rate of 0.5 to 0.8 sample tubes per 1000 (0.05 to 0.08%) has been reported (40). In genetic testing, one study identified that significant problems have been reported with 0.33% of tests performed, with the majority (60%) of the problems occuring in the pretest phase (41).

4.7 Conclusions

In conclusion, the results of this Master's thesis could be regarded as positive, with the identification of DNA degradation as a quality control factor heavily influencing the quality of the genotyping array results. This can be used in any future genotyping experiments to identify samples with significantly higher risks of failing genotyping data quality control. In addition to this, a specific sample collection period associated with samples that are significantly more likely to degrade over time was identified, based on the A260/A230 ratios of samples, offering the opportunity to divide samples into different quality control groups in the future. A potential proxy to agarose gel electrophoresis results was identified as well, with the increase in DNA concentrations and A260/A280 ratios correlating with observed DNA degradation.

In addition to this, an effect of sample positions on sample release plates and genotyping arrays on genotyping failure rates was seen as well, which could enable the improvement of sample processing protocols to avoid or reduce this effect in future sample releases via better sealing of sample release plates.

Based on this, the goal of the Master's thesis was achieved and the hypothesis confirmed.

Besides the specifically achieved results, this Master's thesis could also serve as an insight into the challenges involved with analysing the quality control data of large numbers of samples in a biobanking context, where the exclusion of samples always needs to be weighed against the potential loss of data from those samples. The numerous complexities encountered by the author during the preparation and handling of data and the writing of this Master's thesis also reconfirms the crucial necessity of employing a dedicated quality management specialist in a biobanking setting, as a superficial analysis of quality control data could easily lead to missing critical details and relations between quality control metrics and analysis results, as well as reduce the opportunities to use pre-existing data to improve quality related procedures.

5 Reference list

1. Ginsburg GS, Phillips KA. Precision medicine: From science to value. *Health Aff*. 2018;37(5):694–701.
2. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, Ferrari R. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Brief Bioinform*. 2018;19(2):286–302.
3. Psifidi A, Dovas CI, Bramis G, Lazou T, Russel CL, Arsenos G, Banos G. Comparison of eleven methods for genomic DNA extraction suitable for large-scale whole-genome genotyping and long-term DNA banking using blood samples. *PLoS One*. 2015;10(1):e0115960.
4. Kinkorová J. Biobanks in the era of personalized medicine: Objectives, challenges, and innovation. *EPMA J*. 2016;7(1):4.
5. Illumina Infinium HTS Assay Protocol Guide [Internet]. Illumina, Inc.; 2013 [cited 2019 Jul 9]. Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/humanomniexpress-24/infinium_hts_assay_protocol_user_guide_15045738_a.pdf
6. Hansen TVO, Simonsen MK, Nielsen FC, Hundrup YA. Collection of blood, saliva, and buccal cell samples in a pilot study on the Danish nurse cohort: Comparison of the response rate and quality of genomic DNA. *Cancer Epidemiol Biomarkers Prev*. 2007;16(10):2072–6.
7. Montgomery MC, Petraroia R, Weimer ET. Buccal swab genomic DNA fragmentation predicts likelihood of successful HLA genotyping by next-generation sequencing. *Hum Immunol*. 2017;78(10):634–41.
8. Permenter J, Ishwar A, Rounsavall A, Smith M, Faske J, Sailey CJ, Alfaro MP. Quantitative analysis of genomic DNA degradation in whole blood under various storage conditions for molecular diagnostic testing. *Mol Cell Probes*. 2015;29(6):449–53.
9. Mousquer GT, Maciel LP, Pompeu Saraiva AC, Dalla Costa ER, Rosa Rossetti ML. Validation of a quick and low-cost DNA extraction protocol applicable to long-stored blood samples. *Anal Biochem*. 2018;561–562:47–51.

10. Khare P, Raj V, Chandra S, Agarwal S. Quantitative and qualitative assessment of DNA extracted from saliva for its use in forensic identification. *J Forensic Dent Sci*. 2014;6(2):81.
11. Leitsalu L, Haller T, Esko T, Tammesoo M-L, Alavere H, Snieder H, Perola M, Ng PC, Mägi R, Milani L, Fischer K, Metspalu A. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol*. 2015;44(4):1137–47.
12. Malentacchi F, Pazzagli M, Simi L, Orlando C, Wyrich R, Hartmann CC, Verderio P, Pizzamiglio S, Ciniselli CM, Tichopad A, Kubista M, Gelmini S. SPIDIA-DNA: An External Quality Assessment for the pre-analytical phase of blood samples used for DNA-based analyses. *Clin Chim Acta*. 2013;424:274–86.
13. Palmirotta R, Ludovici G, De Marchis ML, Savonarola A, Leone B, Spila A, De Angelis F, Morte D Della, Ferroni P, Guadagni F. Preanalytical Procedures for DNA Studies: The Experience of the Interinstitutional Multidisciplinary BioBank (BioBIM). *Biopreserv Biobank*. 2011;9(1):35–45.
14. Visvikis S, Schlenck A, Maurice M. DNA extraction and stability for epidemiological studies. *Clin Chem Lab Med*. 1998;36(8):551–5.
15. El-Ashram S, Al Nasr I, Suo X. Nucleic acid protocols: Extraction and optimization. *Biotechnol Reports*. 2016;12:33–9.
16. Boesenberg-Smith KA, Pessarakli MM, Wolk DM. Assessment of DNA yield and purity: An overlooked detail of PCR troubleshooting. *Clin Microbiol Newsl*. 2012;34(1):1–6.
17. Geenivaramu.ee Data access [Internet]. Institute of Genomics, University of Tartu; [cited 2019 Jul 9]. Available from: https://www.geenivaramu.ee/en/biobank.ee/data-access
18. Kaasik A-T, Keis A, Metspalu A, Allik A, Mölder E, Leego E, Alavere H, Lilienthal K, Fischer K, Leitsalu-Moynihan L, Milani L, Tammesoo M-L, Väli-Täht M, Leego M, Hass M, Tamm R, Nikopensius T, Esko T, Haller T. Estonian Genome Center 2001-2011 [Internet]. Estonian Genome Center, University of Tartu; 2011 [cited 2019 Jul 9]. Available from: https://www.geenivaramu.ee/sites/default/files/geenivaramu/estoniangenomecenter.pdf

19. Tasa T, Krebs K, Kals M, Mägi R, Lauschke VM, Haller T, Puurand T, Remm M, Esko T, Metspalu A, Vilo J, Milani L. Genetic variation in the Estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *Eur J Hum Genet*. 2019;27(3):442–54.
20. Fischer K, Kettunen J, Würtz P, Haller T, Havulinna AS, Kangas AJ, Soininen P, Esko T, Tammesoo ML, Mägi R, Smit S, Palotie A, Ripatti S, Salomaa V, Ala-Korpela M, Perola M, Metspalu A. Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons. *PLoS Med*. 2014;11(2):e1001606.
21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75.
22. Matlock B. Assessment of Nucleic Acid Purity [Internet]. Thermo Fisher Scientific Inc. 2011 [cited 2019 Jul 9]. Available from: http://www.nanodrop.com/Library/T042-NanoDrop-Spectrophotometers-Nucleic-Acid-Purity-Ratios.pdf
23. Santella RM. Approaches to DNA/RNA extraction and whole genome amplification. *Cancer Epidemiol Biomarkers Prev*. 2006;15(9):1585–7.
24. Wilfinger WW, Mackey K, Chomczynski P. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *Biotechniques*. 1997;22(3):474–81.
25. Riemann K, Adamzik M, Frauenrath S, Egensperger R, Schmid KW, Brockmeyer NH, Siffert W. Comparison of manual and automated nucleic acid extraction from whole-blood samples. *J Clin Lab Anal*. 2007;21(4):244–8.
26. Malentacchi F, Ciniselli CM, Pazzagli M, Verderio P, Barraud L, Hartmann CC, Pizzamiglio S, Weisbuch S, Wyrich R, Gelmini S. Influence of pre-analytical procedures on genomic DNA integrity in blood samples: The SPIDIA experience. *Clin Chim Acta*. 2015;440:205–10.

27. Carboni I, Fattorini P, Previderè C, Ciglieri SS, Iozzi S, Nutini AL, Contini E, Pescucci C, Torricelli F, Ricci U. Evaluation of the reliability of the data generated by Next Generation Sequencing from artificially degraded DNA samples. *Forensic Sci Int Genet Suppl Ser*. 2015;5:e83–5.
28. Wu J, Cunanan J, Kim L, Kulatunga T, Huang C, Anekella B. Stability of Genomic DNA at Various Storage Conditions. *Int Soc Biol Environ Repos 2009 Annu Meet* [Internet]. 2009;12 [cited 2019 Jul 9]. Available from: https://www.colorado.edu/ecenter/sites/default/files/attached-files/seracare_stability_of_genomic_dna_at_various_storage_conditions_isber2009.pdf
29. Schuurman T, de Boer R, Patty R, Kooistra-Smid M, van Zwet A. Comparative evaluation of in-house manual, and commercial semi-automated and automated DNA extraction platforms in the sample preparation of human stool specimens for a Salmonella enterica 5'-nuclease assay. *J Microbiol Methods*. 2007;71(3):238–45.
30. Gassmann M, McHoull B. DNA Integrity Number (DIN) with the Agilent 2200 TapeStation System and the Agilent Genomic DNA ScreenTape Assay [Internet]. 2015 [cited 2019 Jul 9]. Available from: https://www.agilent.com/cs/library/applications/5991-5258EN.pdf
31. Wheeler T, Swenerton R, Zhu K, Singh R, Fathollahi B, Ho M. Genomic DNA Quality Assessment by an Automated Microchip Electrophoresis Platform. *J Biomol Tech*. 2013;24(Suppl):S56–S56.
32. Goering R V. Pulsed field gel electrophoresis: A review of application and interpretation in the molecular epidemiology of infectious disease. Infect Genet Evol. 2010;10:866–75.
33. Ciniselli CM, Pizzamiglio S, Malentacchi F, Gelmini S, Pazzagli M, Hartmann CC, Ibrahim-Gawel H, Verderio P. Combining qualitative and quantitative imaging evaluation for the assessment of genomic DNA integrity: The SPIDIA experience. *Anal Biochem*. 2015;479:60–2.
34. Fellermann H, Lopiccolo A, Kozyra J, Krasnogor N. In Vitro Implementation of a Stack Data Structure Based on DNA Strand Displacement BT - Unconventional Computation and Natural Computation. In: Amos M, CONDON A, editors. Cham: Springer International Publishing; 2016. p. 87–98.

35. Li X, Wu Y, Zhang L, Cao Y, Li Y, Li J, Zhu L, Wu G. Comparison of three common DNA concentration measurement methods. *Anal Biochem*. 2014;451(1):18–24.
36. D'Abramo M, Castellazzi CL, Orozco M, Amadei A. On the nature of DNA hyperchromic effect. *J Phys Chem B*. 2013;117(29):8697–704.
37. Sánchez I, Betsou F, Mathieson W. Does vacuum centrifugal concentration reduce yield or quality of nucleic acids extracted from FFPE biospecimens? *Anal Biochem*. 2019;566:16–9.
38. Qamar W, Khan MR, Arafah A. Optimization of conditions to extract high quality DNA for PCR analysis from whole blood using SDS-proteinase K method. *Saudi J Biol Sci*. 2017;24(7):1465–9.
39. Samango-Sprouse C, Kirkizlar E, Hall MP, Lawson P, Demko Z, Zneimer SM, Curnow KJ, Gross S, Gropman A. Incidence of X and Y chromosomal aneuploidy in a large child bearing population. *PLoS One*. 2016;11(8):e0161045.
40. Ravine D, Suthers G. Quality standards and samples in genetic testing. J Clin Pathol. 2012;65:389–93.
41. Hofgärtner WT, Tait JF. Frequency of problems during clinical molecular-genetic testing. *Am J Clin Pathol*. 1999;112(1):14–21.

Annex 1 – DNA extraction protocol

Preparation of white blood cells from 2 x 10 mL EDTA blood

1. Centrifugate the EDTA blood tubes for 6 minutes at 200 rcf to prepare the removal of excess blood plasma.
2. Discard the top fraction of blood plasma from each blood tube, taking care not to remove any of the middle buffy coat fraction.
3. Transfer the blood of each patient into a single 50 mL centrifuge tube. Rinse the EDTA blood tubes with the RBC solution and transfer that into the 50 mL centrifuge tube as well. Fill the 50 mL tube with the RBC solution to 40 mL.
4. Close the 50 mL tube, shake it vigorously by hand, and place it on ice for 10 minutes.
5. After 10 minutes, centrifugate the 50 mL tubes for 10 minutes at 600 rcf.
6. After centrifugation, carefully discard the supernatant without disturbing or losing the precipitate at the bottom of the tube, leaving 5 – 10 mm of solution at the bottom of the tube.
7. Add 25 mL of the RBC solution to the 50 mL tube, close the tube and shake it vigorously by hand.
8. Centrifugate the 50 mL tubes for 10 minutes at 500 rcf.
9. After centrifugation, carefully discard the supernatant again, without disturbing or losing the precipitate, leaving about 5 mm of solution at the bottom of the tube.
10. Add 2 mL of the RBC solution to the 50 mL tube and use a sterile pasteur pipette to suspend the white blood cells into the solution, creating a homogenous mix.
11. Add 20 mL of the WBC solution to the 50 mL tube and thoroughly vortex the solution, making sure a homogenous mass forms. If vortexing is not enough, use a 5 mL pipette to suspend the solution, in order to create a homogenous mass.

At this point, samples can be stored at room temperature for at least 1 week if necessary, to form larger batches.

DNA extraction

1. Incubate the tubes with the homogenized solution at +38ºC for 30 minutes.
2. After 30 minutes, visually check whether the cells have been lysed in the solution. If the cells have been lysed, place those tubes on ice to cool them to +20ºC in about 15 minutes.
   If some of the cells haven't been lysed, keep the tube in the incubator and check it again every 30 minutes. If necessary, some tubes can be left in the incubator overnight.
3. Add 8,5 mL of 10M ammonium acetate to the tubes with lysed cells and vortex the tubes to precipitate the protein-membrane complexes.
4. Visually check the formation of the precipitate, and if necessary, put the tubes on ice for 10 minutes to speed up and increase the precipitation.
5. Centrifugate the 50 mL tubes for 10 minutes at 2400 rcf.
6. During the centrifugation, prepare and label new 50 mL tubes and add 20 mL of isopropanol to each tube.
7. After centrifugation, pour the supernatant from each tube to a new tube containing the isopropanol. The precipitate containing protein will be left in the old tube and discarded.
8. Carefully shake the tubes with precipitating DNA for about 5 minutes, either one by one or on a rack. In the same time, visually observe the precipitation and formation of DNA strands.
9. Prepare and label new and empty 50 mL tubes and also 50 mL glass beakers with about 20 mL of 70% ethanol.
10. If the DNA strand has formed in the 50 mL tubes, use a 1 mL pipette to catch the DNA strand from the solution without completely aspirating it, but only holding from the end. Wash the DNA strand in 70% ethanol for about 5 minutes, then catch it again with a 1 mL pipette and place it into the new and empty 50 mL tube. Leave the tube open for about 5-15 minutes to have the residual ethanol evaporate.
11. After the 5-15 minutes, check if the ethanol has evaporated, and if not, use a pipette to carefully remove any ethanol left.

12. Add 6 mL (less, if higher concentrations are desired) of 1X TE buffer to the 50 mL tube containing the DNA strand.
13. Incubate the tubes at +56ºC for 30 minutes.
14. After incubation, place the tubes on a slow shaker for up to 7 days to allow the DNA to redissolve in the buffer. Remove the tubes after no DNA precipitate is visible in the tube. With higher concentrations, redissolving DNA will take a long time, so the visual inspection is necessary.
15. Once the DNA is in the solution, quality analysis can be performed – for example concentration and A260/A280, A260/A230 ratios on a Nanodrop spectrophotometer or gel electrophoresis to confirm intactness of the DNA.
16. DNA can be stored at +4ºC for shorter periods and at -20ºC for long term storage.

Solutions used in the protocol

RBC solution – for lysis of red blood cells

Contents:

- 155 mM NH_4Cl
- 10 mM $KHCO_3$
- 1 mM EDTA-NA_2 pH 8.0

For 1 liter:

- 8,3 g NH_4Cl
- 1,0 g $KHCO_3$
- 2 ml 0.5M EDTA-NA_2 pH 8.0
- milliQ water to 1 liter

WBC solution – for lysis of white blood cells

Contents:

- 10 mM Tris-HCl pH 8.0
- 25 mM EDTA-NA_2 pH 8.0
- 2% SDS

For 1 liter:

- 10 mL 1M Tris-HCl pH 8.0
- 50 mL 0.5M EDTA-NA_2 pH 8.0
- 20 mL 10% SDS
- milliQ water to 1 liter

10M ammonium acetate

For 1 liter:

- 770,8 g $NH_4C_2H_3O_2$
- milliQ water to 1 liter

TE buffer

Contents:

- 10 mM Tris-HCl ph 8.0
- 1 mM EDTA-NA_2 pH 8.0

For 1 liter:

- 10 mL 1M Tris-HCl pH 8.0
- 2 mL 0.5M EDTA-NA2 pH 8.0
- milliQ water to 1 liter

Stock solutions

0.5M EDTA-Na_2 ph 8.0

For 1 liter:

- 186,1 g EDTA-NA2
- ca 20 g NaOH for correcting pH
- milliQ water to 1 liter
- correct pH to 8.0

1M Tris-HCl ph 8.0

For 1 liter:

- 121,1 g Tris
- ca 42 mL concentrated HCl for correcting pH
- milliQ water to 1 liter
- correct pH to 8.0

10% SDS

For 1 liter:

- 100 g SDS
- milliQ water to 1 liter