UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
INSTITUTE OF MOLECULAR AND CELL BIOLOGY
DEPARTMENT OF BIOINFORMATICS

Taavi Võsumaa

Development of a methodology for predicting the quality of resequencing probes

Masters thesis

Supervisor: Maido Remm, Ph.D.

TARTU
2008

Table of contents

Abbreviations and definitions

ACR – arcsine transformed call rate

APEX – Arrayed Primer Extension; a genotyping platform developed by Asper Biotech Ltd.

CFS – chance of false signal

CR – call rate

GLM – general linear models

LSI – logarithmic average standardised signal intensity

MSE – mean squared errors

P – the probability value or p-value

PERL – Practical Extraction and Reporting Language; a programming language

PHP – PHP Hypertext Preprocessor; a programming language

R^2 – coefficient of determination

RFU – Relative Fluorescent Units

SB-factor – sequence-based factor

SBS – sequencing-by-synthesis

SI – average standardised signal intensity

SNP – single nucleotide polymorphism

SS1 – type-1 sums of squares

SS2 – type-2 sums of squares

Tm – oligonucleotide melting temperature

INTRODUCTION

This study is divided into two parts: the literature overview and practical part. In the literature overview, a brief background is given about the importance of resequencing, available resequencing technologies and issues regarding the design of oligonucleotide probes – an essential component of many resequencing technologies. In the practical part, a statistical method for predicting the quality of resequencing probes is developed and the resulting algorithm is integrated into probe design software.

I LITERATURE REVIEW

1. Resequencing background

1.1. Importance of resequencing

DNA sequencing is important for gathering information about gene functions for biological and medical studies to explain molecular mechanisms and find causes and cures for diseases. However, as every living organism is genetically different and variance in phenotypes is vast, the gathering of such information doesn't end with the sequencing of just one individual of any species, but must be repeated again and again to discover relationships between genotypes and phenotypes – hence the term "resequencing".

For human genome, the first general draft has already been available since 2001. It is a consensus sequence of several human genomes, i.e. the Golden Path, which provides preliminary insight into the structure and function of the human genome and provides starting points for future studies and eases subsequent resequencing attempts, required for more detailed understanding of individual genomes.

However, the cost and throughput of current resequencing technologies are still not at a level where DNA resequencing could be made part of routine healthcare inspection – a broader goal that would require the cost of sequencing to drop to about $1000 per genome, which is comparable to the annual healthcare costs of an average US citizen (Service, 2006). Using today's standard capillary gel electrophoresis, the resequencing of a single individual

would take 30 instruments a full year and cost about $10 million (Bentley, 2006). Therefore, 10000-fold reduction in cost and increase in throughput would be necessary to make personalised medicine possible. Even 100 to 1000-fold improvements would make it feasible to study human genetic variation or sequence bacterial genomes at a much larger scale than it is possible at the moment.

To meet this demand many new resequencing technologies are being developed. In order to illustrate the diversity of resequencing approaches and summarise the advancements made in current technologies, a short description of different methods will be given.

1.2. Current resequencing technologies

Current resequencing methods use a variety of approaches, some of which are rather different from classic Sanger sequencing, and range from being in early piloting stages to full commercialisation. The following brief overview is not exhaustive, but will illustrate the diversity of approaches that can be used in a genome sequencing project.

1.2.1. Synthetic chain termination

The synthetic chain termination method involves the priming of target sequences with universal primers and then feeding the polymerase process with a mixture of regular- and labelled terminator-nucleotides. The terminator-nucleotides are added in such concentrations that all possible read lengths from one to more than a thousand are obtained, with a base-specific label in the 3' end. Finally, the sequence is deducted by separating the products by electrophoresis and identifying labels as they pass a detector while traversing through the gel (Figure 1). The classic example here is the Sanger sequencing. The advances in this field include the capillary sequencing instrument and microelectrophoresis (Paegel et al., 2003) – the former increases throughput by performing electrophoresis simultaneously in several hundred capillaries and the latter aims to reduce reagent volumes by the use of microfluidic devices.

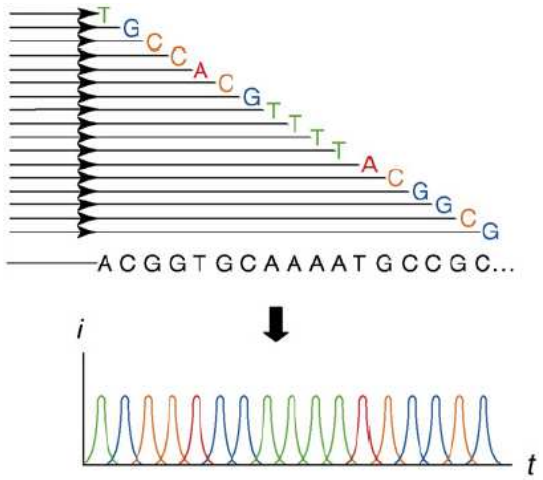


Figure 1. A scheme depicting the process of sequencing by synthetic chain termination (Bentley, 2006).

1.2.2. Sequencing-by-synthesis (SBS)

Sequencing-by-synthesis (SBS) methods, similarly to Sanger sequencing, rely on priming and replicating target DNA. However, here the individual bases are identified directly and a subsequent time-consuming electrophoresis step is not needed. SBS can be divided into distinct sub-methodologies: base-by-base SBS, pyrosequencing and single-molecule SBS in real time.

1.2.2.1. Base-by-base SBS

Base-by-base SBS is achieved by replicating array-bound template DNA with a mix of fluorescently labelled reversible terminator nucleotides that allow DNA to be sequenced one base at a time. Each incorporated nucleotide stops polymerase activity and the fluorescent signal can be read; then the blocking moiety, along with the label, is removed from the nucleotide and the next base is added. These steps are iterated until read limit or desired read length is reached. Finally, signal intensities are plotted with respect to cycle number and the DNA sequence can be deducted (Figure 2).

Such techniques are used by Illumina (*http://www.illumina.com*) in their Genome Analyzer and by Helicos BioSciences (*http://www.helicosbio.com*).

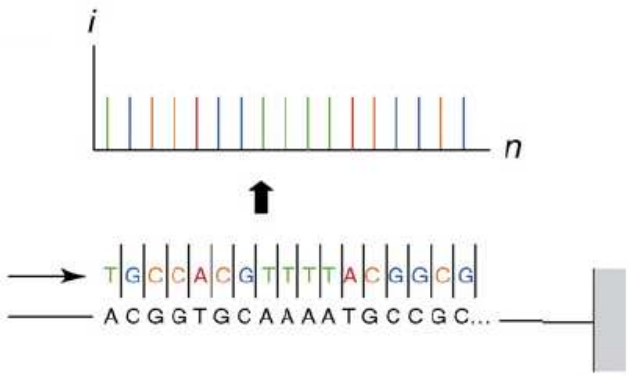


Figure 2. A scheme depicting the read signals from by-base-by base SBS (Bentley, 2006). Template (black) is bound to surface and distinctly labelled nucleotides are incorporated into growing strand one by one. Final sequence is determined by plotting the intensities of different signals (i) against the cycle number (n) (here only the prevalent signal is shown).

1.2.2.2. Pyrosequencing

Unique to this approach is the use of regular nucleotides, instead of labelled and/or terminator ones, and the use of chemoluminescence to detect the incorporation of bases. Firstly, template DNA is attached to solid beads, amplified and primed by universal primes for subsequent synthesis reactions. Each cycle consists of adding just one out of four didesoxynucleotides to the reaction mixture, to avoid mixing signals, and then measuring the light that is produced chemically from the released pyrophosphate. By measuring the relative intensity of light, compared to other signals, the amount of nucleotides, incorporated in each cycle, can be deduced (Figure 3).

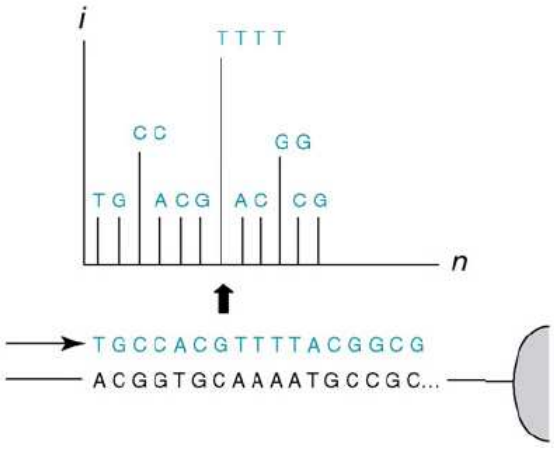


Figure 3. A scheme depicting the process of pyrosequencing (Bentley, 2006). The successful detection events of each cycle (n) and their respective intensities (i) are plotted together to form a flowgram. The intensity is proportional to the quantity of incorporated nucleotides.

1.2.2.3. Real-time SBS

Real-time SBS is an application, developed to sequence DNA in real time without interruptions, i.e. at the rate dictated by the DNA polymerase.

One such method is to observe the movement of polymerase along template DNA as it extends the nascent strand (Greenleaf and Block, 2006). The movement can be observed by using two polystyrene beads, one of which is attached to the polymerase and another to the 3' end of template DNA. The reaction is performed in four separate solutions where one of the four dNTPs is in lower concentrations. This makes the polymerase pause occasionally when it has to incorporate the nucleotide, which is in shortage. As a result, the stopping patterns of the polymerase are obtained in case of each limiting nucleotide and the original sequence can be deducted (Figure 4).

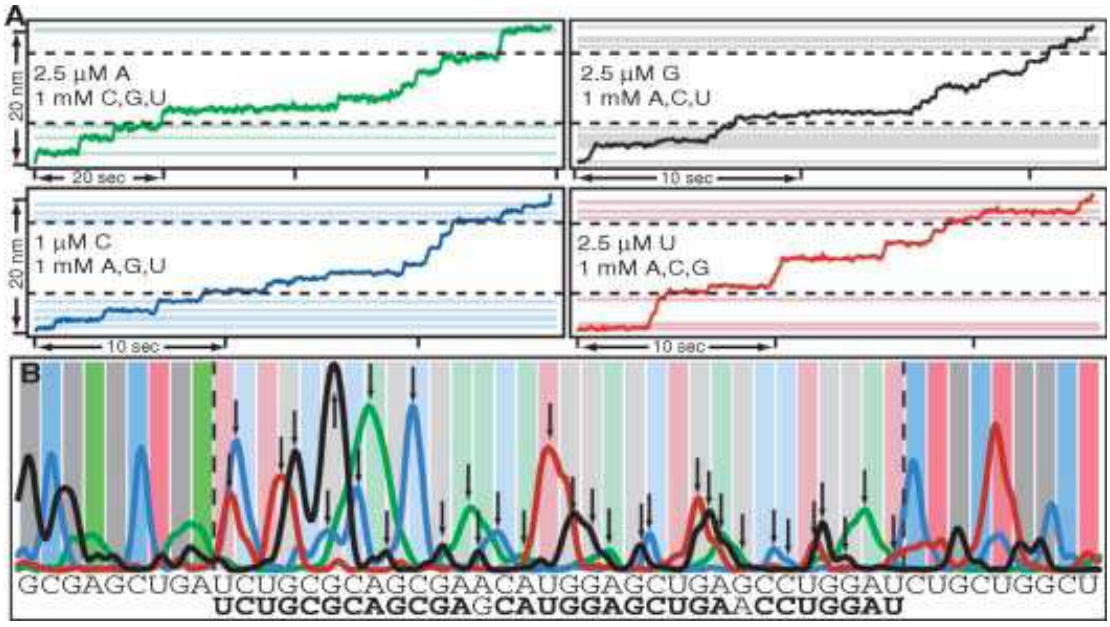


Figure 4. A scheme depicting the reads obtained from real-time SBS (Greenleaf and Block, 2006). Upper four diagrams show the motion of the polymerase when each of four different nucleotides is limited. The bottom diagram shows the final sequence.

There are other experimental real-time SBS methods that use surface-bound polymerase and fluorescently labelled nucleotides.

VisiGen (*http://visigenbio.com*) designed a fluorescent resonance energy transfer system that uses a donor attached to the polymerase and acceptors attached to the gamma-phosphates of each dNTP, so that incorporated nucleotides become excited and emit a base-specific signal.

Pacific Biosciences (*http://www.pacificbiosciences.com*) uses a miniature detection chamber in the reaction solution, which houses the polymerase and allows only minute amounts of reagents to flow through. All the gamma-phoshpate-labelled dNTPs in the chamber are constantly excited, but only those emission signals stand out that belong to the type of nucleotide that is currently being incorporated (i.e. A,C,G or T), because others flow through faster.

1.2.3. Sequencing-by-hybridisation

For sequencing by hybridisation, an oligonucleotide array is created to probe specific nucleotides in target DNA. A single base of target DNA is queried by four almost identical probes, located at different locations on the array, with a varying nucleotide only in the middle (Figure 5). Target DNA is labelled, hybridised onto the array, and the spots that subsequently yield high signal intensity, identify a specific base on the target sequence.

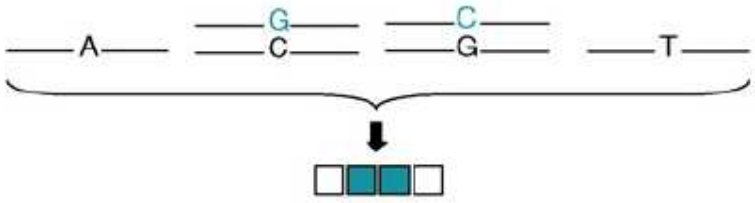


Figure 5. A scheme depicting sequencing by hybridisation (Bentley, 2006). Two out of a group four probes are hybridised to target DNA fragments and yield a signal, indicating a C/G heterzygote at a specific location on target DNA.

This method can also be used in conjunction with ligation, known as sequencing-by-ligation (see Section 1.2.4).

1.2.4. Sequencing-by-ligation

Sequencing-by-ligation combines hybridisation with ligation steps to query surface-bound target DNA. The process is initiated by binding universal anchor primers to the template DNA and then adding a mixture of oligonucleotides, composed of all possible 9-mer

combinations, with each having only their central nucleotide labelled. A matching oligonucleotide is hybridised to the template and then ligated to the anchor primer, identifying the central nucleotide. Then the label and part of the oligonucleotide is chemically removed and the steps are repeated, resulting in a sequence where each fifth nucleotide is known. To close the gaps, the process is repeated four more times, with anchor primers shifted by one position, until the whole template is sequenced (Figure 6).

A variation of this method, were two central nucleotides are labelled, is used in Applied Biosystem's SOliD sequencing platform (*https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=604416*, 25.05.2008).

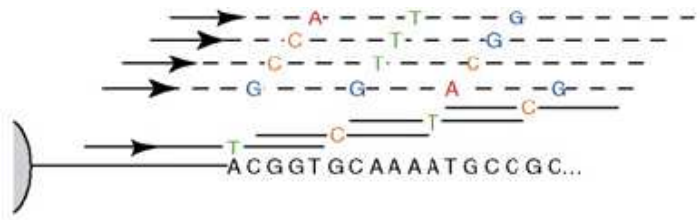


Figure 6. A scheme depicting the process of sequencing by ligation (Bentley, 2006). Arrows are the anchor primers, to which, step-by-step, labelled oligonucleotides are ligated.

1.2.5. Nanopore sequencing

Nanopore sequencing is based on measuring changes in electric impedance in the nanopore as negatively charged target DNA is moved through by the use of an electric field (Figure 7). Each base or a combination of successive bases that enter the ~1.5nm wide biological or synthetic nanopore obstruct the flow of ions in that pore and the resulting electric fingerprint can be used to determine DNA sequence.

Although nanopore sequencing is a theoretically promising method; potentially yielding long reads, requiring but a single template molecule and needing little reagents, its success has been limited. Main problems are the construction of stable pores (Rhee and Burns, 2007) and preventing the DNA from wobbling back and forth as it traverses the pore (Service, 2006).

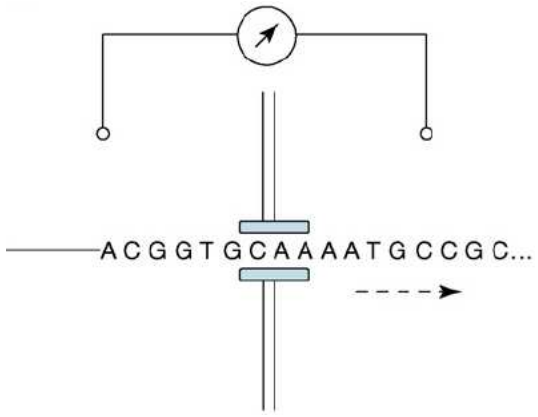


Figure 7. A scheme depicting the process of nanopore sequencing (Bentley, 2006). Nucleotides are identified by changes in the electric impedance of the nanopore as DNA is pulled through the pore by an electric field.

1.3. Resequencing probe design issues

Oligonucleotide probes serve as key components in several resequencing technologies by priming polymerase reactions or detecting bases by mismatch discrimination. Consequently, the performance of these technologies is dependent on the proper design of probes. It is especially relevant to methods that use large sets of distinct probes, in contrast to methods using few universal probes. The former methods, e.g. GenoVoxx (*http://www.genovoxx.de*) and ArraySBS (*http://www.ist-world.org/ProjectDetails.aspx?ProjectId=c369da46689e4058a1994cbc7a0bc6b2*, 25.05.2008), are generally designed to resequence specific genes or other areas of interest while the latter are geared towards whole-genome resequencing (e.g. Illumina Genome Analyser and Helicos Genetic Analysis System).

The nucleotide composition of probes affects both the hybridization and enzymatic efficiencies, which in turn affect experimental signal-to-noise ratios and, ultimately, the success or failure of base calls (Yuryev et al., 2004). Therefore, a careful design of probes is needed to ensure the success of resequencing experiments involving such oligonucleotides.

An effective probe interacts only with its intended target and yields a signal, which can be distinguished from background noise. To achieve that, a probe must not form dimers with itself or with other probes and it must not form secondary structures or hybridise to unintended regions of template DNA (Kaderali et al., 2003). Also, the probe-target duplex must be easily accessible to DNA polymerase, which requires that the 3' end of the probe forms a stable duplex over at least eight bases (Rychlik, 1995) and is composed of specific nucleotides (Onodera and Melcher, 2004). Failing to meet these criteria can lead to either the

false priming of probes (i.e. yielding a signal unrelated to the investigated nucleotide) or cause the attenuation of true signals.

In order to design high-quality probes, the characteristics responsible for probe efficiency must be identified, and used to upgrade existing probe selection principles. The challenges lie in discovering those characteristics and accurately assessing their significance.

Some researches have focused on specific factors to improve their probe or primer selection criteria, such as probe 3' end binding energies (Miura et al., 2005), probe 3' end nucleotide combinations (Onodera and Melcher, 2004) and possible number of non-specific binding sites and secondary structures, calculated in different ways (Rubin and Levy, 1996; Kaderali and Schliep, 2002); others have examined the effects of several factors together by using fuzzy logics (Haas et al., 1998) or statistical models (Yuryev et al., 2002; Benita et al., 2003; Andreson et al., 2008).

This illustrates the fact that there are many choices to be made in improving primer design, starting from picking the most promising factors up to using proper tools to assess their importance in final decision-making.

According to Yuryev, the use of statistical models is the best approach to successful design of probes for any given application, as it allows easy tuning of existing probe selection principles (i.e. re-evaluating the significance of various probe properties) for a specific platform or developing new criteria as long as the fundamental working principles of these applications are similar (Yuryev et al., 2002).

Straightforwardly speaking, statistical models, based on information about the system behaviour in certain conditions, help to predict how it performs in other conditions. They can be used to analyse which probe properties, and to what extent, are correlated with their success or failure in previous experiments. As the result, a statistical model, a formula, is created. The model can be used to predict the "goodness" of future probes by knowing the numeric values of all relevant probe properties.

Although the statistical models present a powerful toolkit for researchers, great care must be taken in choosing appropriate statistical procedures and properly preparing the data and interpreting the results. These procedures usually make specific assumptions about the data and if these are not met, e.g. because of lack of background information about data, the results may become biased and lead to wrong conclusions.

Additionally, despite their versatility and the relative ease by which they can accommodate changes in data and factor sets, when the platform, for which the solution is tailored, is replaced with a significantly different one, the needed data preparation, model testing and factor selection steps might have to be modified to an extent where it becomes

preferable to develop an alternative approach altogether. Also, there are no single universal model making rules to follow and it is the job of the researches to customise their analysis and make educated guesses to create models that best describe a particular system.

Consequently, in our study we tailored our own approach for designing better probes for a four-colour SBS method, using statistical models, along with custom data preparation, factor selection and model testing procedures.

II PRACTICAL PART

2. Aims of the study

The aim of the practical part of this study was to develop a methodology for predicting the quality of resequencing probes for use in four-colour sequencing-by-synthesis platform that employs large sets of probes to directly sequence specific target sequences; and integrate it into a probe design software. This task was divided into the following sub-goals: 1) find ways to normalise microarray signals for use in data analysis; 2) develop a statistical methodology for predicting the quality of resequencing probes; 3) find the best set of factors to predict the quality of APEX (Kurg et al., 2000) genotyping probes; 4) implement the resulting statistical models into probe design software.

The ability to predict probes' quality allows us to select probes that will more likely work well. That would reduce the need for trial-and-error experiments with several candidate probes – time-consuming and costly procedures. In addition, if we have to use probes that have undesirable qualities, e.g. the tendency of forming dimers, we can take the predicted quality of probes into account when interpreting the results.

We measured the quality of probes by their call rate and fluorescent signal intensity, acquired from microarray experiments. Call rate is the percentage of those experiments where the probe has been known to work correctly.

While call rate reflects probe quality the best, its use in statistical models is difficult because it has most of values near 100%. As a supporting predictor of probe quality, we chose fluorescent signal intensity. The predicted signal intensity indicates whether the probe's signal will likely be high enough to be distinguishable from the background noise.

As shown in Figure 8, there is no clear correlation between signal intensity and call rate, meaning that they do not duplicate each other's information and both have the potential to be independently used for probe quality prediction.

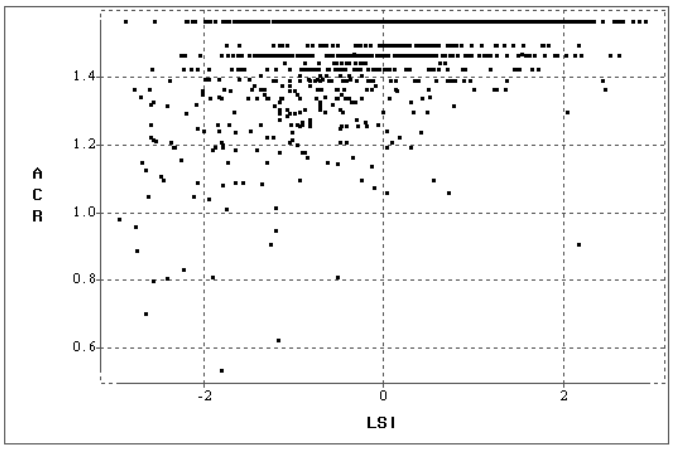


Figure 8. Scatter plot showing the correlation between transformed signal intensity (LSI) on the x-axis and transformed call rate (ACR) on the y-axis; the transformations are explained in section 3.2. Corresponding R^2 is 0.19.

For prediction of the probe quality, we used statistical methods implemented in *SAS* software package release 9.1.3 (SAS Institute Inc. 2004, Cary, NC, USA), mainly the GLM (General Linear Models Analysis) and MIXED (Mixed Linear Analysis of Variance) procedures which help to find relationships between characteristics of probe sequence properties and the call rate and signal intensity.

3. Data used in analyses

Data for the analysis was obtained from previous experiments with Asper Biotech' (Asper Biotech Ltd., Tartu, Estonia) APEX platform – a four-colour fluorescent genotyping system. Analysed data was collected from three separate sets, designated correspondingly *ABCR170*, *ABCR100* and *ARCAGE.* These genotyping experiments were a suitable source of data for the development of SBS probe quality prediction methods because molecular mechanics are similar in these systems. Even though the actual relationships between probe sequence and quality may differ, the statistical approach for discovering these relationships will likely remain the same.

Each dataset consisted of 96-243 microarrays, carrying several hundred distinct probes that were spotted in duplicates and hybridised with target DNAs (Table 1). Every dataset also included one or more control microarrays where no target-DNA was present, for identifying probes that gave false-positive signals by hybridising onto and extending itself.

For every probe on every microarray we had the following information: 1) probe sequence (including notes if a probe contained modified nucleotides); 2) fluorescent signal intensities produced by the probe in four separate channels (A, C, G, T; measured in RFUs – relative fluorescent units); 3) genotype detected by the probe (all genotypes were manually

determined by experts); 4) expert opinion on the probe's effectiveness (some probe that worked badly were marked as such).

Table 1. The total amount of data available to us. Every probe on each array detected a genotype. There were eight signals per genotype – all probes were in duplicates and each probe gave signals in four different channels. '*' relates to ABCR170 and ABCR100 datasets that used mostly the same probes, thus the total number of different probes is less.

	ABCR170	ABCR100	ARCAGE	TOTAL
MICROARRAYS	243	101	96	441
PROBES	890	901	463	2254 (1364*)
GENOTYPES	216270	91001	44448	351719
SIGNALS	1730160	728008	355584	2813752

3.1. Selecting reliable data

To make further analysis easier to interpret and to find clearer statistical relationships, we removed unreliable data from our original datasets, including microarrays, probes and single spots. See Table 2 for details about the amount of excluded data. Data was filtered out according to the following criteria:

1. Missing expert-determined genotypes – arrays, probes and single spots for which expert-confirmed genotype was unavailable, were excluded.
2. Heterozygous genotypes – in heterozygous states signals in independent channels have lower intensities than in homozygous state, since less target DNA is available for either channel. As there are not many heterozygous signals and they might make signal intensity prediction more uncertain, they were removed from analysis.
3. Probes for which its duplicated spots had highly different signal intensities were excluded. Signals were considered highly different when (a) their intensities differed by more than two times and both had more than ten RFUs (relative fluorescent units) or (b) their intensities differed by more than ten RFUs and at least one of the signals had the intensity of ten or less RFUs.
4. Probes containing chemically modified nucleotides were excluded.
5. Probes yielding high signals in control-arrays – if a probe gave a signal greater than 20 RFUs in any of the four channels of any control-array, it was considered prone to self-priming and was excluded.
6. Probes marked unreliable by experts were dropped.
7. Database conflicts – in some cases array and probe data that was originally kept in different databases could not be confidently matched, because their ID-s had been tampered with.

Table 2. The total amount of data excluded from analysis or how much is left. "*" means, that exact information is not available – single removed genotypes and signals are not counted.

	DATA REMOVED FROM ANALYSIS				
	ABCR170	**ABCR100**	**ARCAGE**	**TOTAL**	**TOTAL LEFT**
ARRAYS	73 (30%)	0	0	73 (10%)	368 (90%)
PROBES	472 (53%)	418 (46%)	33 (7%)	923 (41%)	1331 (59%)
GENOTYPES	145210 (63%)*	42218 (46%)*	3168 (7%)*	190596 (54%)*	161123 (46%)
SIGNALS	1151680 (63%)*	337744 (46%)*	25344 (7%)*	1514768 (54%)*	1298984 (46%)

Further, plotting the call rate and signal intensity values together revealed that in our dataset, there are several probes with very low average signal intensity but a very high call rate that formed a distinct group (Figure 9). These probes were considered unreliable as they clearly separate from others probes on the plot and having a high call rate with very low signal intensity is counter-intuitive. Probably some other information was used for making these calls, e.g. probe information from the other strand. Consequently, all probes with low average signal intensity (LSI < -3) were also removed from further analysis.

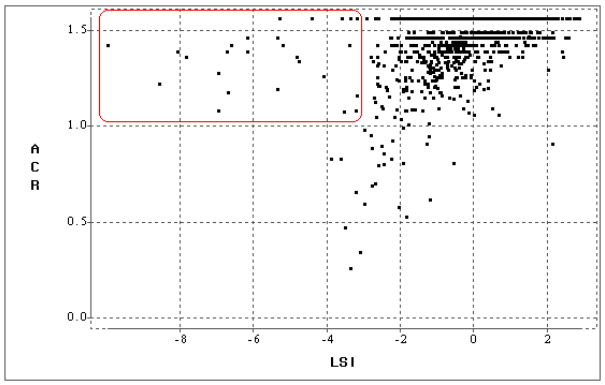


Figure 9. Scatter plot with log-transformed signal intensity (LSI) on the x-axis and arcsin-transformed call rate (ACR) on the y-axis (the transformations are described in section 3.2). Probes with low LSI and high ACR form a distinct group on the plot. These are considered noise and therefore all probes with LSI < -3 were removed from further analysis.

3.2. Calculation and transformation of call rate and signal intensity

Prior to statistical analysis, independent factors were defined and calculated for each probe. Probe's call rate (abbreviated as CR) was calculated as the percentage of arrays where the probe yielded a successful (according to lab specialists) genotype call, out of all arrays where the probe was present:

$$CR = \frac{\text{number of arrays where the probe worked}}{\text{number of arrays where the probe was present}} \times 100$$

CR is an estimate of the probability that genotype will be determined correctly. Theoretically, it is a random variable multiple to a variable having binomial distribution. In the following analyses, we use statistical methods where the normal distribution is preferred and for this, we apply the arcsine transformation to CR to make its distribution more normal:

$$ACR = \arcsin\left(\sqrt{\frac{CR}{100}}\right)$$

This transformation decreases dependence of CR variance on CR mean value, but the distribution of ACR converges to a normal distribution with a reasonable speed only when the probability of correct call is close to ½. Because we have call rates close to 100% (Figure 10), the normality cannot be achieved.

Although there is some effect, the arcsine transformation of CR did not improve the distribution greatly, as it is impossible to normalise distributions that have most of the values near one (Figure 11).

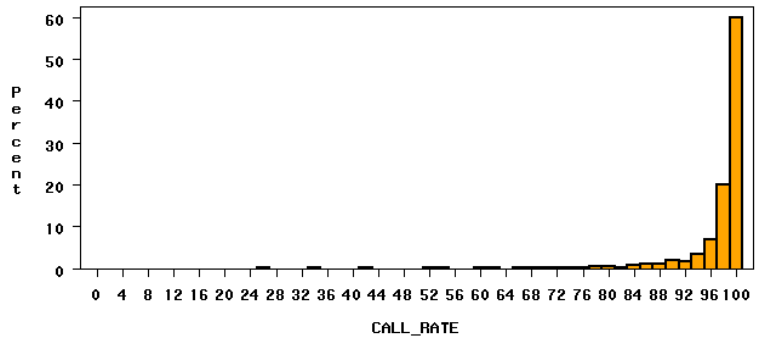


Figure 10. The distribution of call rates before the normalisation.

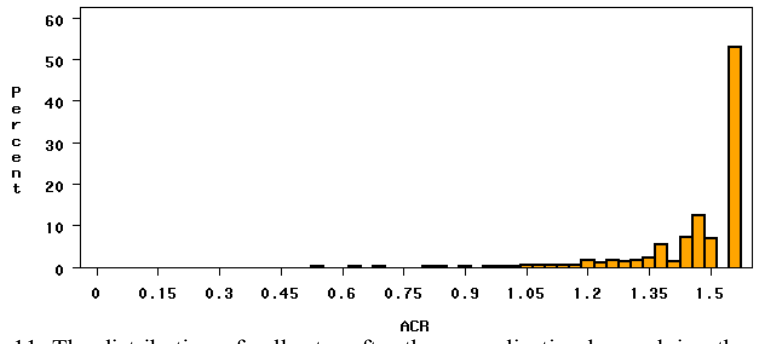


Figure 11. The distribution of call rates after the normalisation by applying the arcsine transformation.

Probe's signal intensity (SI), was calculated as the average signal intensity over all arrays where the probe had a signal in the channel that corresponded to the expert-called genotype (1). All signals had been previously normalised (see Section 4.1).

$$\mathrm{SI} = \frac{\text{sum of probe's normalised signal intensities from all arrays}}{\text{number of arrays where the probe gave a signal}} \times 100 \qquad (1)$$

Distribution of SI is clearly asymmetric (Figure 12) but after the logarithmic transformation (with base of 2) its distribution becomes almost normal (Figure 13).

$$\mathrm{LSI} = \log_2(\mathrm{SI}/100) \qquad (2)$$

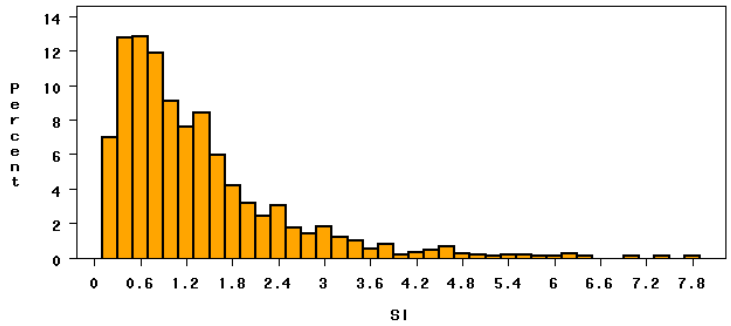


Figure 12. The distribution of non-logarithmic signal intensities (SI) of all probes in our data.

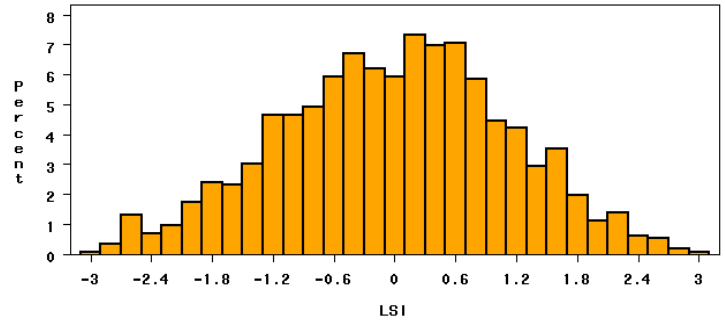


Figure 13. The distribution of logarithmic signal intensities (LSI) of all probes in our data.

4. Results

Using pre-filtered reliable data from APEX genotyping experiments we found a method to normalise microarray signals, tailored a statistical approach to predicting resequencing probe quality, found the best model to predict the quality of APEX genotyping probes, tested its universality and implemented the model into a probe designing software.

4.1. Normalisation of microarray signals

To make signal intensities read from different microarrays and different channels more comparable, it has to be ensured that different microarrays and the type of channel where the signal was produced have no disturbing effect on the signal intensity.

Signal intensity distributions, calculated for each channels separately, using the median signal intensities of every microarrays, revealed that the intensities vary greatly between microarrays as well as between different channels (Figure 14). For these calculations only the previously filtered reliable data, including no heterozygous signals, was used. Median was used instead of arithmetic mean to reduce the effect of particularly strong or weak signals (outliers). For example, signals generated by the incorporation of cytosine were frequently over two times lower than those of other nucleotides**;** and the median signals of different microarrays range from as low as six RFU-s to more than 100 RFU-s. Therefore, signals had to be normalised to make them usable in further analyses.

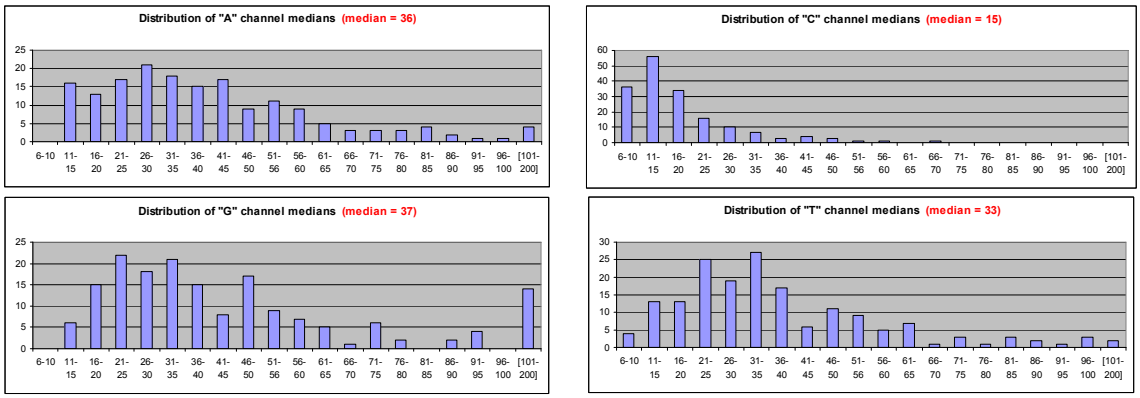


Figure 14. Four tables showing the distribution of median signals of all ABCR170 microarrays in each of four channels.

The normalisation of signals was achieved by dividing each signal in a specific channel on an array by a percentile of all signals of the same channel on that array. As can be seen from Table 3, if no normalisation was used, then 23% of the signal variance was solely the result of different microarrays having different signal levels. By using the 90^{th} percentile for normalisation, the effect of arrays to signals was reduced to 0.16% of total variance, eliminating it almost entirely. The 90^{th} percentile was chosen amongst the 70^{th}, 75^{th}, 80^{th}, 85^{th},

90^{th} and 95^{th} percentiles into the final normalisation formula, as it reduced microarray effect on probe signals the most:

$$\text{normalised signal} = \frac{\text{raw signal of array X, channel Y}}{\text{90th percentile signal of array X, channel Y}}$$

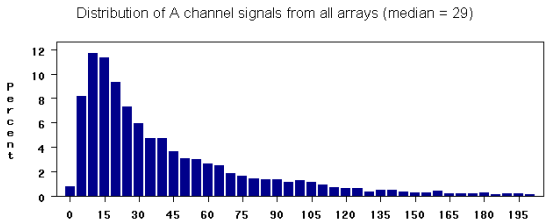
Percentiles less than 70 could not be used since most of the signals in every channel of a microarray are zeroes.

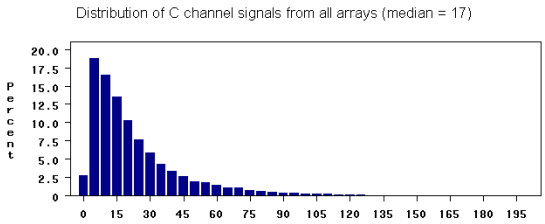
It should be noted that the 90^{th} percentile signal was calculated using each and every signal on that array, not only the ones that we preserved as reliable. This is because we want this normalisation method to be usable and work well in the case when we do not have any expert's notes, indicating which signals can be treated as true positives. In addition, the 90^{th} percentile, like the median, should be relatively free from the effects of particularly high or low signals.

Table 3. This table shows how much of the probe signal variance is caused by microarrays after performing various normalisations. The results were obtained by SAS's MIXED procedure, which performs mixed-type analysis of variance.

NORMALISATION METHOD	MICROARRAY COMPONENT OF SIGNAL VARIANCE
None	22.97 %
70^{th} percentile	5.77 %
75^{th} percentile	2.89 %
80^{th} percentile	1.65 %
85^{th} percentile	0.53 %
90^{th} percentile	0.16 %
95^{th} percentile	0.40 %

Figure 15 and Figure 16 show the distribution of all signals of all arrays in four different channels before and after normalisation step, respectively. Here, the distributions individual probe signals are provided, instead of array medians, as using medians can result in loss of information.





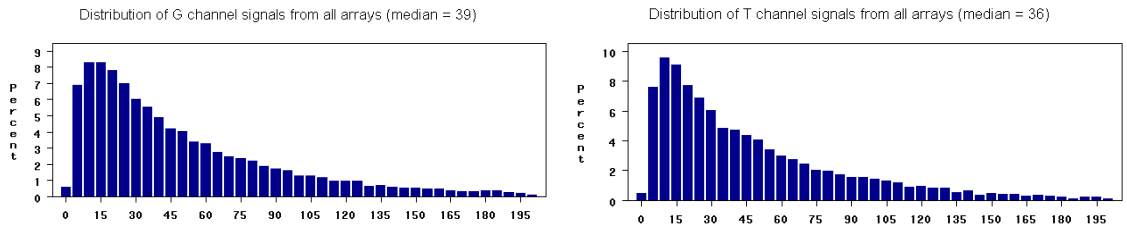


Figure 15. Distribution of signals of all ABCR170 microarrays in each of four channels before normalisation Very high signals (> 200 RFU-s) are not shown on these graphs to enhance appearance.

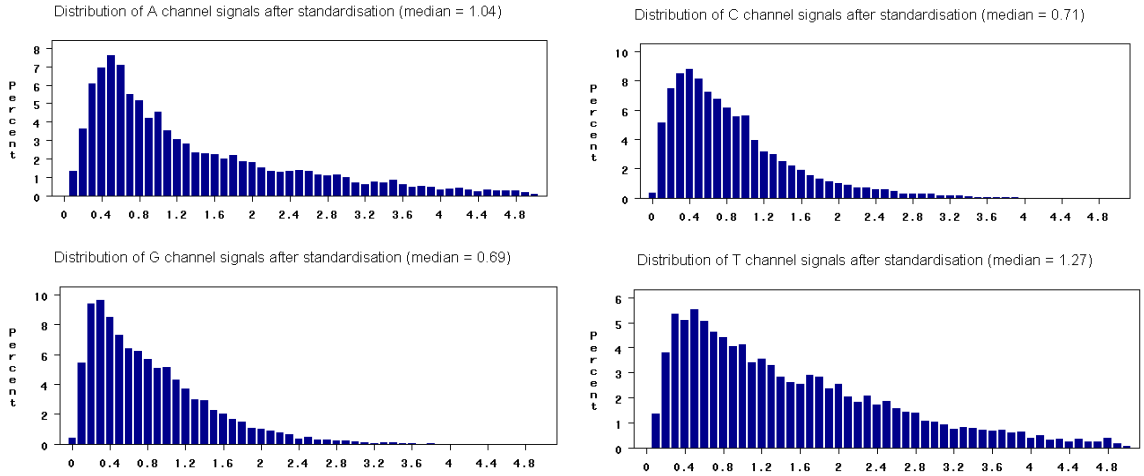


Figure 16**.** Distribution of median signals of different arrays in four different channels after normalisation by the 90th percentile. Very high signals (greater than 5.0) are not shown on these graphs to enhance appearance. Only data from ABCR170 dataset was used here.

4.1.1. Spatial signal intensity patterns

We also checked for spatial signal intensity patterns on the microarray to see whether some locations on the microarray glass had frequently strong or weak signals and if spatial signal normalisation was required. This could arise for example when target DNA is unevenly distributed on the surface of the array or if some spotting needles spot concentrations of probes differently. In this case, the raw signal intensities from specific areas of microarrays should be normalised to compensate for the above-mentioned technical defects. To analyse this we created several 'pseudoarray' images where stronger than median signal intensities were highlighted – green squares were two times higher and red square four times higher than the median of current chip and channel (Figure 17). Two different colours were used in case the potential signal patterns only emerged at certain intensity levels. However, visual examination of the pseudoarrays didn't reveal such patterns and no further analysis was performed.

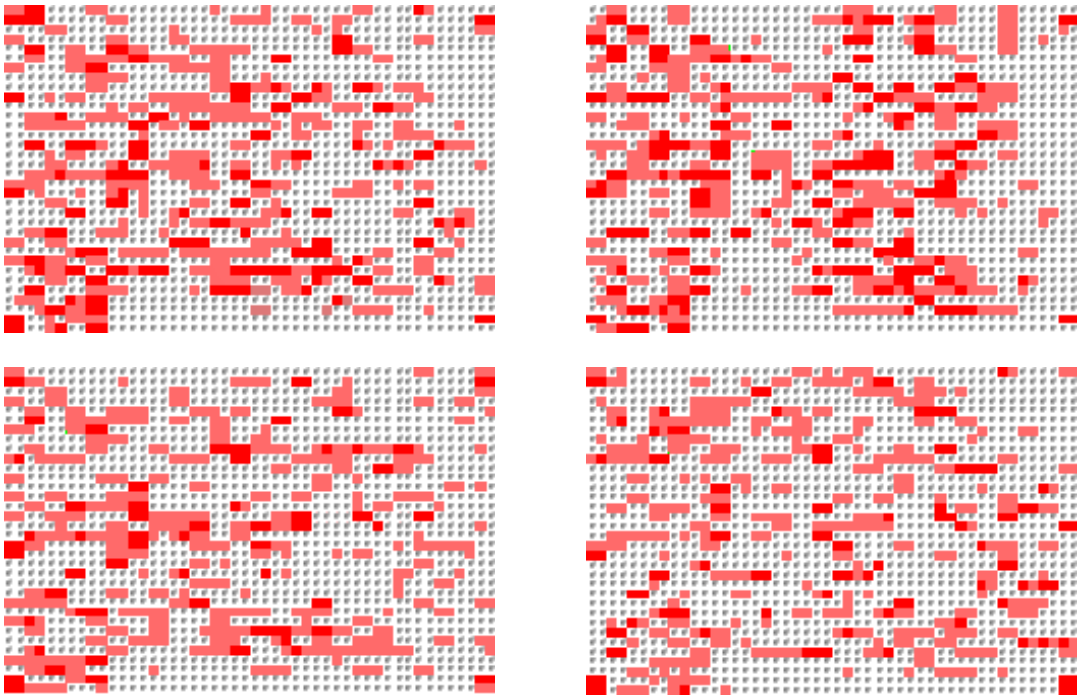


Figure 17. Four selected pseudoarray images that were created to search for spatial signal intensity patterns. Each image displays the signal layout of an array from the ABCR170 set. Each square denotes one spot on microarray. Light red squares have signal intensity higher than two times of the array median and dark red squares have intensity higher than four times of the array median. Summed intensities over all channels are shown.

4.2. Development of a statistical methodology for probe quality prediction

To predict oligonucleotide probe quality we used statistical methods that can be used to find relationships between probes' call rates, signal intensities and the nucleotide composition of their sequences, which make all the probes different from each other. For that task, general linear analysis, realised in SAS GLM procedure, was selected. This procedure can also perform the classical analysis of variance (ANOVA) where all the factors are of categorical type, and the pure classical regression analysis, where all the factors are numerical. Such analysis gives information about how well one or more independent variables that can be calculated from a probe's sequence (called hereafter the sequence-based factors or SB-factors) can predict the dependent variables – call rate and signal intensity of the probe.

GLM is suitable for our analysis as it can use both numerical and categorical variables for prediction and data does not have to be balanced – i.e. the number of observations for all levels of a categorical factor does not need to be equal.

The quality of resulting models will be characterised by its prediction power, measured in R^2 and p-values, which indicate the level of confidence that the model as a whole and each independent variable (factor) in it is actually significant for the prediction of probe

quality. We set the significance level α for all analyses equal to 0.001; relationships having greater p-value than α were considered unproven as they may have arisen by chance.

4.3. Search for the best set of factors for the prediction of call rate and signal intensity

Two groups of factors were used for the prediction of probe quality: sequence-based factors (SB-factors) and auxiliary factors. The SB-factors are calculated directly from the probe's sequence and can therefore be used to predict probe quality before real experiments are made. The auxiliary factors are calculated using data other than probe's sequence and they are known only after the experiments and therefore cannot be used for predicting the probe quality. They can be used for explaining the probe quality and, consequently, to locate other components of the sequencing system, which could be improved to ensure that probes work better – e.g. the efficiency of PCR reactions prior to sequencing.

Multiple SB-factor candidates were calculated and their effect on both, ACR and LSI was tested using general linear analysis. For numerical factors, e.g. dG (the Gibbs free energy of probe-template interaction), the effects of up to the 3^{rd} order polynomial were tested. The descriptions and effects of tested factors on both ACR and LSI are listed in Supplementary Table 1 (SB-factors) and Supplementary Table 2 (auxiliary factors). The list of most significant factors is provided below separately, as part of the final models.

Out of all SB-factors, the optimal set was chosen separately for call rate and signal intensity prediction. An optimal set is one that has minimal number of factors, while retaining maximum prediction power. Not all factors are required for maximum prediction power, because many of them are not independent having strong correlation with others.

A selection of factors is necessary for three reasons: firstly, a model with fewer factors can be calculated faster; secondly, simpler models are easier to understand and interpret; thirdly, having too many factors in a model compared to the amount data may result in over-fitting – i.e. reported R^2 will be greater than it actually is. For example, when all of our SB-factors were used for LSI prediction at once, the R^2 was reported as 100%, which clearly does not reflect true prediction power of the model.

For finding the optimal set of factors, a custom model-making algorithm had to be devised, as there are no standard methods for the data and factors that we used.

We have tailored a stepwise model-making algorithm that seeks out the factor with the greatest effect and then systematically adds new factors that provide the most additional information. Statistical analysis was performed by using general linear type-1 analyses (SS1). SS1 gives a p-value for every factor in the factor set, indicating whether this factor gives

additional information to the model, when all preceding factors have already been taken into account. This guarantees that every subsequent factor we include into the model adds to our prediction power. Initially, all factors were additionally tested with type-2 variance analysis (SS2), which evaluates the effect of every factor by removing it and analysing whether this decreases the model quality. However, for all models where SS1 p-values were below threshold, also SS2 p-values were below threshold, indicating that such double-checking was not necessary. Overview of the algorithm is described in Table 4. This procedure was automated by a PERL script, which took factors from a preset list, sent them to SAS software for analysis, and based on returned data, sent a new set of factors, until the overall model could not be improved anymore.

Alternatively, to the step-wise factor additions, we tried analysing all permutations of factors, where the positions of the factors in the model were also shuffled. Since by using SS1 the factor p-values are position-dependent, a scenario could arise where the highest R^2 scoring factor would render two other factors statistically insignificant, which together on their own could have a greater R^2 and still have significant p-values. The ACR models obtained by this method showed that this method could give models with greater R^2, but only slightly. Since this method was also very time-consuming (calculations for ACR took several days; LSI analyses run for more than one week and were cancelled) it was not investigated any further and the previously described method was adapted instead.

Table 4. Overview of the final factor selection algorithm.

OVERVIEW OF THE FINAL FACTOR SELECTION ALGORITHM

1. Cycle through all factors available for analysis.
2. Add each factor temporarily into the model
3. Measure their R^2 and p-values.
4. Select the factor with the best R^2, where p-value was at most 0.001.
5. Add that factor permanently into the model and remove it from further cycles.
6. Repeat from step 1, until there are no more factors with p-value ≤ 0.001.

This algorithm was used to find the best set of factors for both LSI and ACR prediction. The resulting LSI model had six factors and could predict 28% of LSI; ACR model had three factors and could predict 5.2% of ACR (Table 5). Figures 18 and 19 illustrate the correlation between the predicted values and experimental values for ACR and LSI models, respectively.

Table 5. Sets of SB-factors that explain the variability of LSI and ACR the most. The R^2 describes the model's total prediction power. Descriptions of the factors are provided in Supplementary Table 1.

MODEL BUILT USING SB-FACTORS ONLY					
DEPENDENT VARIABLE		**MODEL R^2**	**DEPENDENT VARIABLE**		**MODEL R^2**
LSI		28%	**ACR**		5.2%
POS	**FACTORS**	**P-VALUE**	**POS**	**FACTORS**	**P-VALUE**
1.	N1N2	<.0001	1.	dG13	<.0001
2.	N4N5	<.0001	2.	any×any×any	<.0001
3.	dG15	<.0001	3.	dG13×dG13	0.0005
4.	N3N6	0.0001			
5.	dG15×dG15	<.0001			
6.	any×any×any	<.0001			

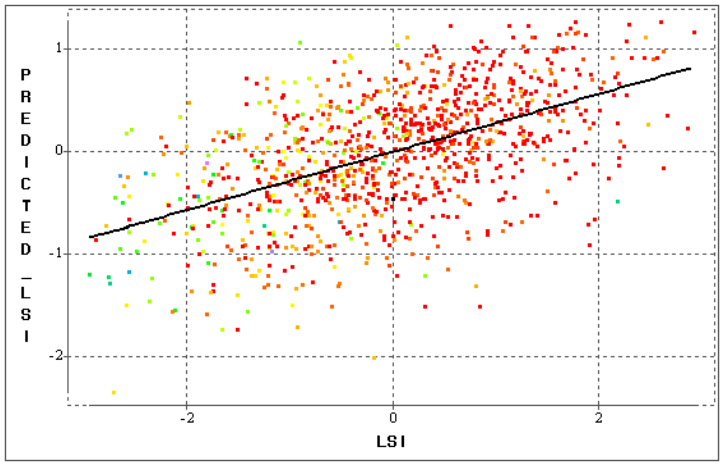


Figure 18. A scatterplot visualising our model's ability to predict probe signal from its sequence. On the x-axis there are experimentally measured LSI-s, and on the y-axis there are LSI-s predicted by our model. Black line is the fit curve (R^2 = 0.2840). The colours of the dots denote probes' ACR-s, where read mean the highest and blue the lowest ACR-s.

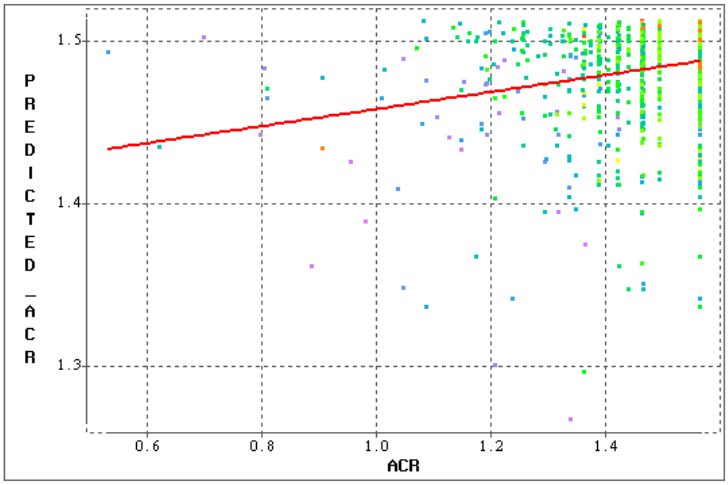


Figure 19. A scatterplot visualising our model's ability to predict call rate from its sequence. On the x-axis there are experimentally measured ACR-s, and on the y-axis there are ACR-s predicted by our model. Black line is the fit curve (R^2 = 0.0523). ACR of 1.57 is equal to call rate of 100%. The colours of the dots denote probes' LSI-s, where red means the highest and blue the lowest LSI-s.

Two more models were created to estimate our ability to explain the total variance of call rate and signal intensity. Here in addition to the SB-factors, the auxiliary factors were also included. These models have significantly better correlation with ACR and LSI, 31% and 45% respectively (Table 6). Comparison with Table 5 indicates that about 26% of ACR and 17% of LSI variability comes from PCR and sample preparation steps.

Table 6. This table shows the sets of factors that explain the variability of LSI and ACR the most. The related models are not usable for prediction as the levels of auxiliary factors like PCR and 'sample' cannot be calculated for future probes. Descriptions of the factors are provided in Supplementary Tables 1 and 2.

MODEL BUILT USING ALL FACTORS					
DEPENDENT VARIABLE		**MODEL R^2**	**DEPENDENT VARIABLE**		**MODEL R^2**
LSI		45%	**ACR**		31%
POS	**FACTORS**	**P-VALUE**	**POS**	**FACTORS**	**P-VALUE**
1.	N1N2	<.0001	1.	PCR	<.0001
2.	PCR	<.0001	2.	sample	<.0001
3.	N4N5	<.0001	3.	dG13	<.0001
4.	N3N6	<.0001	4.	dG13×dG13	<.0001
5.	dG16	<.0001	5.	Crank×Crank×Crank	0.0001
6.	dG16×dG16	<.0001			
7.	end1×end1×end1	<.0001			
8.	dG3×dG3×dG3	0.0002			

To further analyse the selected factors we created graphs that describe their effect on the dependent variables in detail (Suppl. Figure 1 and 2). The graphs show that dG values have optimums, near -22 kcal/mol and -19 kcal/mol for dg15 and dg13, respectively. The 'any' factor has positive correlation with both call rate and signal intensity, which is expected, as probes with very negative 'any' values should form dimers, which yield no signals. N1N2, N4N5 factor graphs (Suppl. Figure 1) show that probes containing C and G nucleotides generally tend to yield higher signals than those containing A and T nucleotides, with AA and TT combinations yielding the lowest signals

In addition, the distribution of every factor used in our models was compared to their distribution in a randomly generated dataset of 10000 randomly chosen genomic probes (Suppl. Figure 3). Every factor can only reliably predict the quality of future probes if their values are in the same intervals as the ones they were trained on – i.e. our dataset. As can be seen from Supplementary Figure 3, the distribution of factors in our dataset is similar to those of random genomic probes. The only notable difference is the higher percentage of C and G nucleotides in our probes, compared to the random genomic probes, but this likely because our probes are designed specifically on genes, which are known for their greater GC content.

4.4. Model testing

To verify that the factors selected by our statistical methodology are not strictly specific to one dataset but can be also used in case of other datasets, we split our original data into two subsets, used one for model making and then compared the model's prediction power by applying it on both subsets.

First, the data was split into two subsets: A (*ABCR170* and *ARCAGE* data) and B (*ABCR100* data). Then, using only data of subset B, we estimated a new set of factors and the corresponding model that predicts probe quality the best in B. After that, this model was applied to both subsets A and B and we measured the mean squared errors (MSE), which shows the mean deviance of experimental values from predicted values, and found their ratio for both ACR and LSI prediction ($Ratio_{ACR}$ and $Ratio_{LSI}$, correspondingly):

$$\mathrm{MSE} = \frac{1}{\text{degrees of freedom}} \sqrt{\sum (\text{experimental value} - \text{predicted value})^2}$$

$$\mathrm{Ratio}_{\mathrm{ACR}} = \mathrm{MSE}_{\mathrm{B,\,ACR}} / \mathrm{MSE}_{\mathrm{A,\,ACR}} = 0.89$$

$$\mathrm{Ratio}_{\mathrm{LSI}} = \mathrm{MSE}_{\mathrm{B,\,LSI}} / \mathrm{MSE}_{\mathrm{A,\,LSI}} = 0.98$$

For both ACR and LSI, the ratio was near one (0.89 and 0.98 respectively) meaning that the model found by our method worked almost equally good on both datasets and was not strictly specific to the data from which it was estimated (calculated) and that it can be used more universally.

4.5. Implementation of the probe design algorithm

As a part of this study, a resequencing probe design software, called SBS Designer, was realised in PERL and PHP programming languages for the practical implementation of our probe quality prediction models and improvement of previous software. So far only the call rate prediction model has been included, since the best way to use signal intensity for probe selection is under consideration.

Here we give a short description of the program's input, output and probe selection algorithm.

SBS Designer can either evaluate the quality of existing probes or design new probes for a DNA sequence (Figure 20). The sequence can be provided in three different ways:

entered manually into web text-fields; loaded from a text file (in FASTA format); or searched for by using chromosome coordinates or gene ID.

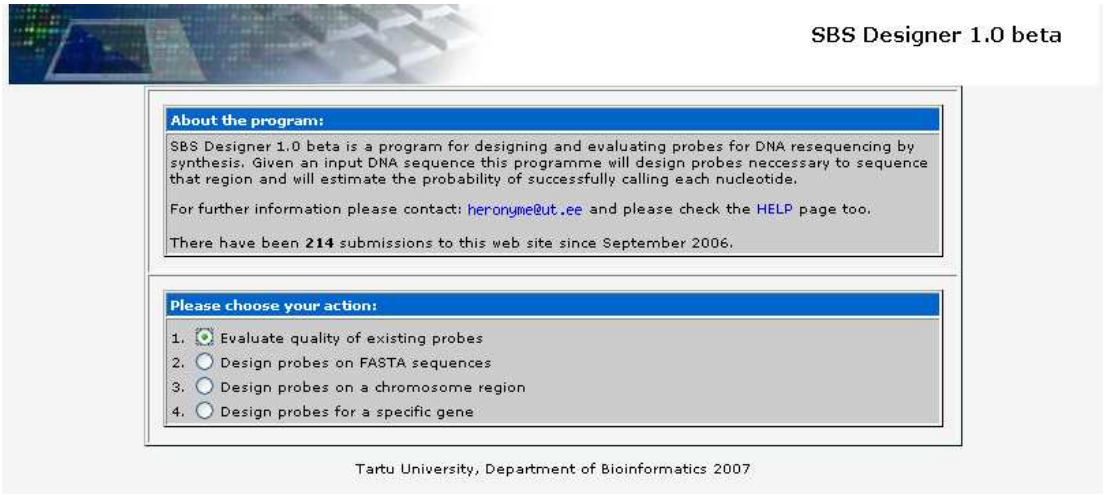


Figure 20. Screenshot of SBS Designer main page where one can choose to analyse existing probes or have the program design new ones, using three different types of input.

As the result, the programme will produce a list of probes that can be used to resequence the given DNA region, accompanied by statistics, such as predicted call rate (that is calculated using the ACR prediction model), melting temperature (Tm), length, genomic positions, etc (Figure 21). Also a visualisation of the DNA region will be provided that will show which nucleotides have the requested coverage (i.e. how many probes will sequence that nucleotide) and which don't, along with reasons why enough probes couldn't be designed for that region (e.g. the area contained repeats or probe's predicted call rate was too low).

PROBE ID	SEQ ID	STRAND	START POS	END POS	LENGTH	Tm ?	PREDICTED CALL-RATE ?	DUST SCORE ?	CFS SCORE ?	5'-PROBE SEQUENCE-3'
iO9zmzQRez6b9Yw2C7dx-1	Sequence #1	antisense	1	15	15	35	99.35	8	11	GACAAACAAACAACG
iO9zmzQRez6b9Yw2C7dx-2	Sequence #1	antisense	2	16	15	35	99.09	8	16	CGACAAACAAACAAC
iO9zmzQRez6b9Yw2C7dx-3	Sequence #1	antisense	3	17	15	39	99.06	7	7	CCGACAAACAAACAA
iO9zmzQRez6b9Yw2C7dx-4	Sequence #1	antisense	4	18	15	42	99.45	6	7	GCCGACAAACAAACA
iO9zmzQRez6b9Yw2C7dx-5	Sequence #1	antisense	5	19	15	45	99.49	4	16	CGCCGACAAACAAAC
iO9zmzQRez6b9Yw2C7dx-6	Sequence #1	antisense	6	20	15	49	99.49	3	7	GCGCCGACAAACAAA
iO9zmzQRez6b9Yw2C7dx-7	Sequence #1	antisense	7	21	15	53	99.58	2	7	CGCGCCGACAAACAA
iO9zmzQRez6b9Yw2C7dx-8	Sequence #1	antisense	8	22	15	52	99.53	2	7	ACGCGCCGACAAACA
iO9zmzQRez6b9Yw2C7dx-9	Sequence #1	antisense	9	23	15	51	99.41	2	16	AACGCGCCGACAAAC
iO9zmzQRez6b9Yw2C7dx-10	Sequence #1	antisense	10	24	15	53	99.41	2	7	CAACGCGCCGACAAA
iO9zmzQRez6b9Yw2C7dx-11	Sequence #1	antisense	11	25	15	56	99.41	2	7	GCAACGCGCCGACAA
iO9zmzQRez6b9Yw2C7dx-12	Sequence #1	antisense	12	26	15	60	99.28	3	7	CGCAACGCGCCGACA
iO9zmzQRez6b9Yw2C7dx-13	Sequence #1	antisense	13	27	15	61	98.18	4	16	GCGCAACGCGCCGAC
iO9zmzQRez6b9Yw2C7dx-14	Sequence #1	antisense	14	28	15	61	97.87	4	7	AGCGCAACGCGCCGA
iO9zmzQRez6b9Yw2C7dx-15	Sequence #1	antisense	15	29	15	62	97.06	4	18	CAGCGCAACGCGCCG

Figure 21. Screenshot depicting part of a results page, which shows the probes that our algorithm has designed for the resequencing of a DNA region (region not shown). Probes coloured in yellow have some characteristics below recommended thresholds but are still considered eligible for sequencing.

The probe design algorithm works by first designing probe for all nucleotide over the input DNA region. The probes can either be of fixed length (minimum is 15) or have uniform melting temperatures, in which case such probe length will be automatically chosen that is closest to the desired Tm value. Probes can be designed on both strands of DNA or only for either one. Then all unusable probes are filtered out according to following criteria: too low predicted call rate (< 95 by default), high 'dust' score (> 20; see Suppl. Table 1 for details on 'dust' score), high CFS value (>50; see Suppl. Table 1 for details on CFS score), probe is on common repeats (optional), probe contains SNPs (optional). After the filtration the sequencing ranges for every remaining probe are calculated – i.e. how many nucleotides they can sequence. The range can either be set to a fixed number by the user, or calculated by a formula from the probe's call rate. Finally, starting from the 5' end of input DNA region, the algorithm will then step-by-step remove all probes, while ensuring that the coverage of all nucleotides in its range will remain above threshold. The remaining probes comprise to final set of probes designed for sequencing the input DNA region. This algorithm is subject to future improvements as it currently doesn't guarantee the selection of the optimal set of probes.

SBS Designer is currently in beta stage and can be previewed at *http://bioinfo.ut.ee/sbsdesigner*.

5. Discussion

The ability of our ACR and LSI models to predict the quality of probes is rather modest, having correlations of 5.2% (for ACR) and 28% (for LSI) between the experimental and predicted values of our APEX data. This reflects the fact that probe sequence is only a minor factor affecting microarray results and other factors (success of PCR reaction and sample preparation efficiency) have higher effect on the results. However, applying our statistical models in addition to existing probe choice principles should still improve the microarray results – the gain may be slight, but overlooking it would be unjustified.

The lack of call rate prediction power can be explained by the fact that it has rather small variance in our data – over 50% of probes have a call rate of 100 and over 90% have call rate above 90 (Figure 10; Section 3.2). With so little variance in the data, it is hard to find good relationships between probe sequence and call rate. The lack of call rate variance is probably due to APEX genotyping probes having been optimised over time, whereby poor-quality probes were removed.

Signal intensity prediction can be used to homogenise the probe signal strengths on microarrays or employed as a score for quality estimation. For instance, if several alternatives are available, the probes with estimated signal strengths of less than -3 could be removed or replaced with other probes, since signals this weak usually do not result in a good call rate. However, this threshold must not be used too stringently, as many potentially good probes could be mistakenly discarded along with bad candidates, since the signal prediction power was not very high.

While the 3' end nucleotide combinations (factors N1N2, N4N5, N3N6) were the strongest predictors of signal intensity and their importance regarding polymerase reactions has been noted previously, we propose that their effects need further confirmation. Firstly, it is hard explain the biological relevance of these specific combinations and their states. Secondly, when we removed some of these factors, other nucleotide combinations or single nucleotides emerged as important – but to a lesser extent. Thirdly, Andreson and colleagues also investigated primer 3' bases in regard to PCR failure rate and found that while they did exhibit some correlation, their use quickly resulted in model over-fitting, and were consequently removed from final analysis (Andreson et al., 2008). We however, chose to keep them, as we had no strong statistical grounds for their removal. Perhaps larger datasets and more thorough model testing would reveal if they are artificial factors or not.

The 'dust' and CFS factors, although considered important, did not have statistical relationships with probe quality in our data. This is probably because all probes used in APEX genotyping had been previously filtered in regards of 'dust' and CFS values to an extent that they no longer had any effect on call rate or signal intensity. Such filtration was confirmed by comparing 'dust' and CFS distribution in our dataset with those of 10000 random genomic probes (Suppl. Figure 4). In addition, when implementing our prediction models the same filtration has to be redone on all probes, for our models to be valid.

In addition to SB-factors, the effects of the auxiliary factors, e.g. PCR and sample, were also measured. Knowing their effects was not necessary for this study as they cannot be used for prediction of probe quality, but can be used to explain additional sources of variation afterwards.

The effect of PCR turned out to be significant for both call rate and signal intensity (15% and 11% correlation respectively), whilst the sample factor affected only call rate (18%). It confirms the fact that the PCR step is one of the most crucial elements of genotyping / resequencing systems, also reported by others (Kaderali et al., 2003), and further efforts should be made to select good PCR primers to ensure the success of the whole system.

The great impact of the sample factor on call rate indicates that something was done systematically differently when the experiments of the three array sets were conducted; by identifying and controlling those conditions, call rate could be improved by another amount that is comparable the effect of PCR. That effect was verified by eliminating the effects of PCR and using only ABCR170 and ABCR100 datasets, which have mostly the same probes: the relationship between sample and ACR remained above 11%.

Finally, regardless of the prediction power of current models, it should be noted that these were tailored specifically for the APEX method, but our methodology allows these models to be easily recalibrated for other systems when provided with corresponding experimental data.

SUMMARY

The aim of this study was to develop a methodology for predicting probe quality from its sequence. The analyses were performed using data from APEX genotyping experiments that employ four-channel fluorescent detection.

As the result, specific procedures were developed for signal intensity normalisation, statistical model creation, factor selection, and model testing.

A 90^{th} percentile-based normalisation technique was devised that reduced array-specific signal intensity variation from 23% to 0.16% of total variation.

The prediction of probe quality reflecting variables, "call rate" and "signal intensity," was accomplished by tailoring an automated model making algorithm, which uses general linear analyses to find the best set of probe quality predicting factors out of more than 50 variables, which were calculated from probe sequences. The final models had 28% and 5.2% correlation with signal intensity and call rate, respectively. However, a total of 45% and 31% of signal intensity and call rate could be explained by including other known experiment-related factors, which are not related to probe sequence and thus cannot be used for probe selection.

Cross-validation of the model was performed using data subsets, which confirmed that the models are not strictly specific to our data.

Finally, the call rate prediction model was implemented into the probe selection algorithm of our probe design software.

Metoodika resekveneerimisproovide kvaliteedi ennustamiseks

Taavi Võsumaa

Kokkuvõte

Antud töö eesmärgiks oli välja töötada metoodika resekveneerimisproovide kvaliteedi ennustamiseks.

Resekveneerimine on oluline organismide geneetilise mitmekesisuse uurimiseks ja haigusgeenide identifitseerimiseks. Paljude resekveneerimistehnoloogiate lahutamatuks komponendiks on DNA proovid, mille õige disain on vajalik hea tööedukuse tagamiseks.

Kuigi proovide disaini on palju uuritud, on varasemate autorite tööde tulemused ning tänapäeva resekveneerimismeetodite tööpõhimõtted üksteisest piisavalt erinevad, et oleks võimalik piirduda üldistatud reeglitega. Seetõttu tuleb proovidisaini printsiipe kohandada antud platvormi tarbeks ning vajadusel lisada uusi kriteeriumeid. Üks paremaid vahendeid disainireeglite väljatöötamiseks on statistilised mudelid ja neid on kasutatud ka antud töös.

Töö tulemusena töötati välja protseduurid DNA proovide *call rate*'i ja signaali tugevuse ennustamiseks nelja kanaliga sünteesil põhineva resekveneerimise tarbeks: signaaliintensiivsuste normaliseerimine, statistilite meetodite kasutus, parimate kvaliteeti mõjutavate faktorite valik ja mudeli testimine. Lisaks koostati loodud mudelit rakendav tarkvara.

Signaaliintensiivsused normaliseeriti algandmete 90. protsentiili suhtes, mille tulemusel vähenes kiibist sõltuv signaali varieeruvus 23 protsendilt 0.16 protsendini kogu varieeruvusest.

Proovide *call rate*'i ja signaali tugevuse ennustamiseks loodi automaatne mudelikoostamise algoritm, mis valib üldist lineaarset analüüsi kasutades välja rohkem kui 50 faktorikandidaadi hulgast parima komplekti. Lõplike mudelite korrelatsioon signaali tugevuse ja *call rate*'i-ga oli vastavalt 28% ja 5.2%. Võttes kasutusele ka nn. kontrollimatud faktorid, mida proovide disainimisel kasutada ei saa, suudeti seletada vastavalt 45% ja 31% signaali tugevuse ja *call rate*'i varieeruvusest.

Mudeleid testiti andmete erinevaid alamhulki kasutades ning leiti, et ühe andmeosa peal treenitud mudel töötab hästi ka treenimiseks mittekasutatud andmete ennustamiseks, mis kinnitab mudeli universaalsust.

Proovide kvaliteeti hindavate mudelite rakendamiseks loodi ka spetsiaalne resekveneerimisproovide disainimise tarkvara, SBS Designer, mille beeta-versioon asub aadressil *http://bioinfo.ut.ee/sbsdesigner*.

ACKNOWLEDGEMENTS

For the completion of this work I wish to thank my supervisor Maido Remm – for decisive remarks, background information and keeping me generally pointed in the right direction. Also, thanks go to Priit Palta for friendly guidance and critical reading of the thesis. Very special thanks go to one of the best Estonian biostatisticians, Tõnu Möls, for warm and motivating support, suggestions and excellent advice in the field of statistics.

REFERENCES

Andreson, R., Möls, T., Remm, M. (2008). Predicting failure rate of PCR in large genomes. Nucleic Acids Res., accepted.

Benita, Y., Oosting, R. S., Lok, M. C., Wise, M. J., Humphery-Smith, I. (2003). Regionalized GC content of template DNA as a predictor of PCR success. Nucleic Acids Res. 31: e99.

Bentley, D. R. (2006). Whole-genome re-sequencing. Curr. Opin. Genet. Dev. 16: 545-552.

Greenleaf, W. J., Block, S. M. (2006). Single-molecule, motion-based DNA sequencing using RNA polymerase. Science 313: 801.

Haas, S., Vingron, M., Poustka, A., Wiemann, S. (1998). Primer design for large scale sequencing. Nucleic Acids Res. 26: 3006-3012.

Kaderali, L., Deshpande, A., Nolan, J. P., White, P. S. (2003). Primer-design for multiplexed genotyping. Nucleic Acids Res. 31: 1796-1802.

Kaderali, L., Schliep, A., (2002). Selecting signature oligonucleotides to identify organisms using DNA arrays. Bioinformatics 18: 1340-1349.

Kaplinski, L., Andreson, R., Puurand, T., Remm, M. (2005). MultiPLX: automatic grouping and evaluation of PCR primers. Bioinformatics 21: 1701-1702.

Kurg, A., Tonisson, N., Georgiou, I., Shumaker, J., Tollett, J., Metspalu, A. (2000). Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. Genet. Test. 4: 1-7.

Miura, F., Uematsu, C., Sakaki, Y., Ito, T. (2005). A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences. Bioinformatics 21: 4363-4370.

Onodera, K., Melcher, U., 2004. Selection for 3' end triplets for polymerase chain reaction primers. Molecular and cellular probes 18: 369-372.

Paegel, B. M., Blazej, R. G., Mathies, R. A. (2003). Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis. Curr. Opin. Biotechnol. 14: 42-50.

Rhee, M., Burns, M. A. (2007). Nanopore sequencing technology: nanopore preparations. Trends Biotechnol. 25: 174-181.

Rubin, E., Levy, A. A. (1996). A mathematical model and a computerized simulation of PCR using complex templates. Nucleic Acids Res. 24: 3538-3545.

Rychlik, W. (1995). Priming efficiency in PCR. BioTechniques 18: 84-86, 88-90.

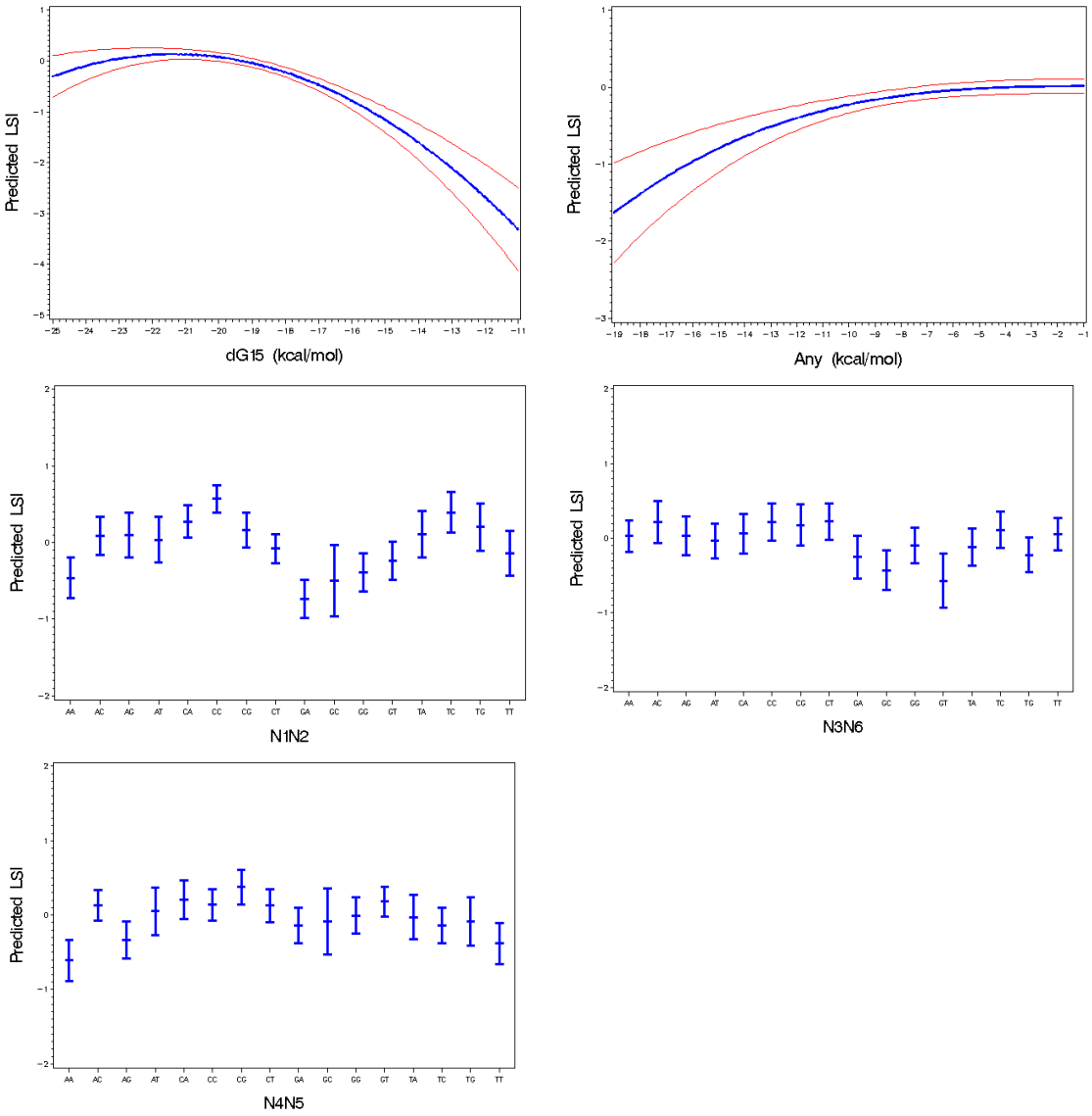
SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. USA 95: 1460-1465.

Service, R. F. (2006). Gene sequencing. The race for the $1000 genome. Science 311: 1544-1546.

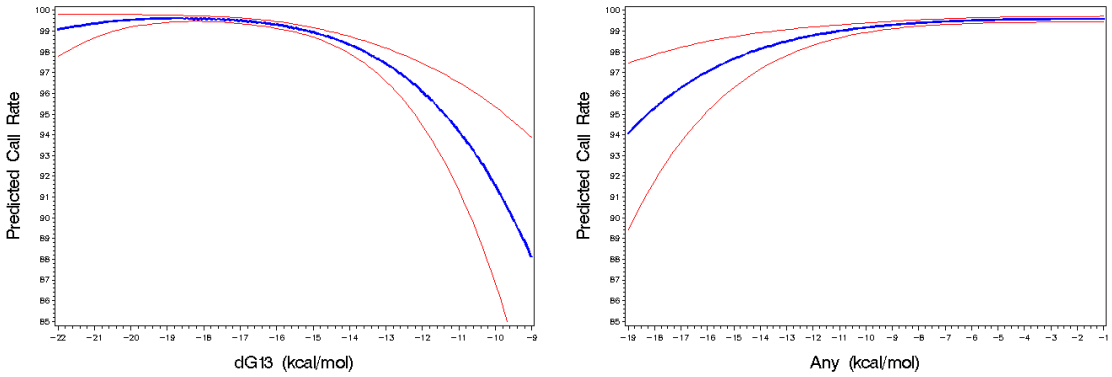
Yuryev, A., Huang, J., Pohl, M., Patch, R., Watson, F., Bell, P., Donaldson, M., Phillips, M. S., Boyce-Jacino, M. T. (2002). Predicting the success of primer extension genotyping assays using statistical modelling. Nucleic Acids Res. 23: e131.

Yuryev, A., Huang, J., Scott, K. E., Kuebler, J., Donaldson, M., Phillips, M. S., Pohl, M., Boyce-Jacino, M. T. (2004). Primer design and marker clustering for multiplex SNP-IT primer extension genotyping assay using statistical modelling. Bioinformatics 20: 3526-3532.

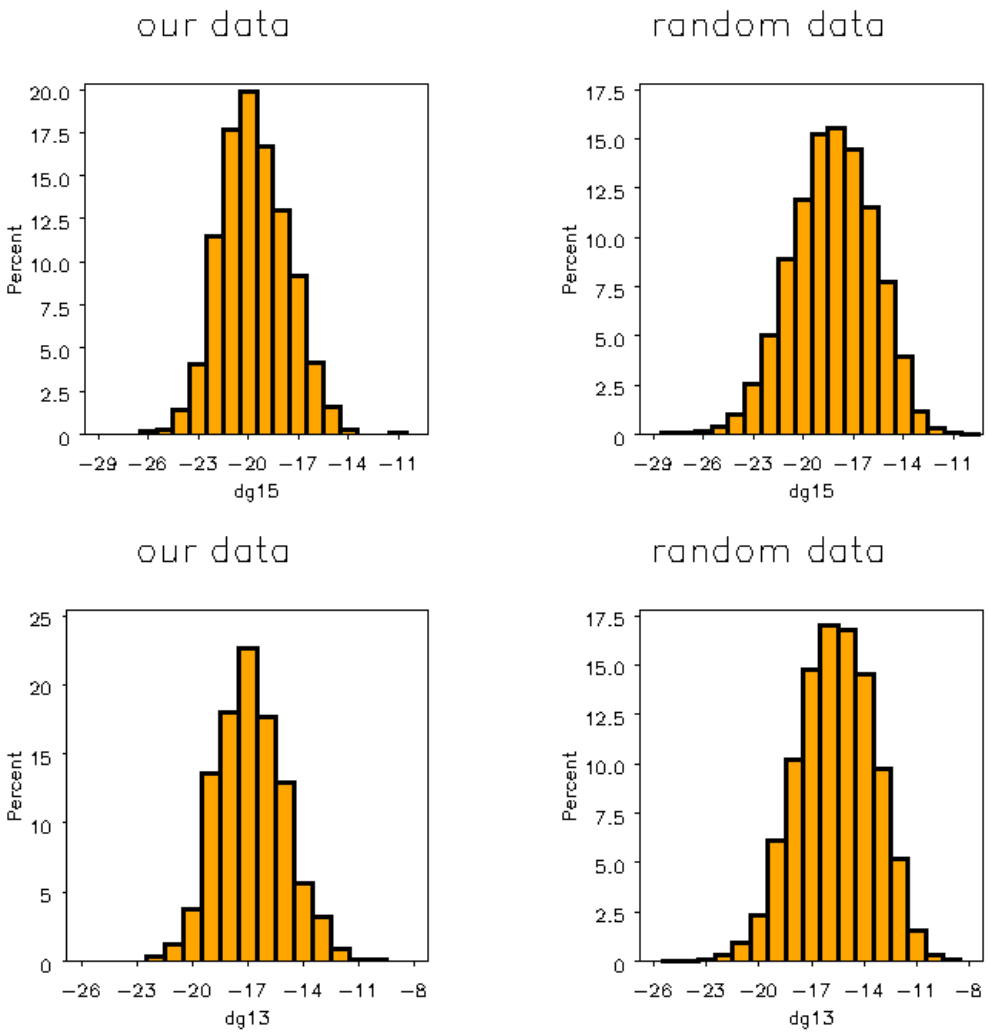
APPENDIX

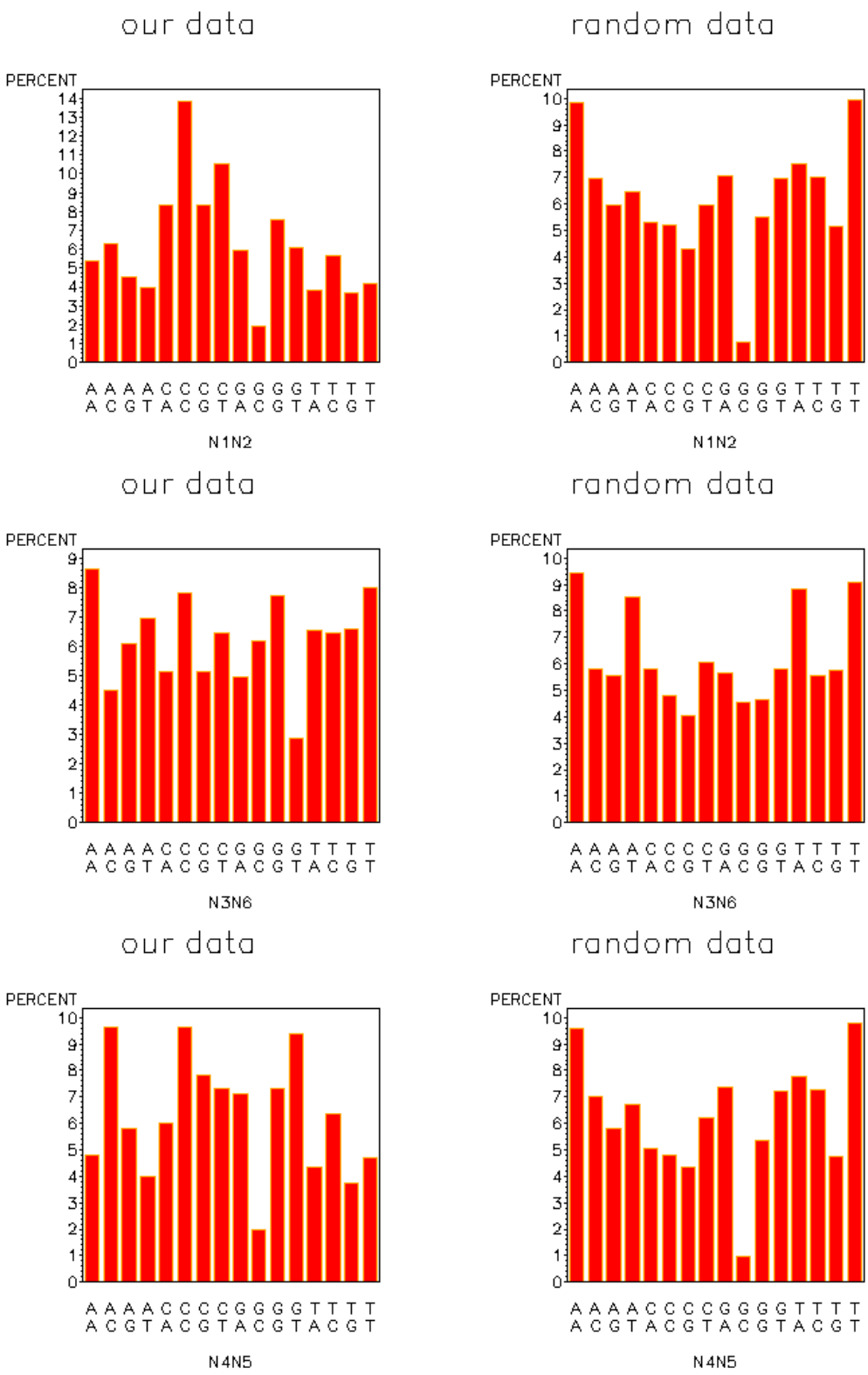


Supplementary Figure 1. Predicted effects of factors dG15, 'any', N1N2, N3N6 and N4N5 on LSI, using the LSI prediction model. When the influence of a factor was analysed, the levels of other factors were fixed to their mean levels. The blue line or the centre of the bar is the predicted value; red lines or the top and the bottom of the bar are the upper and lower 95% confidence limits for the expected mean LSI value, respectively.

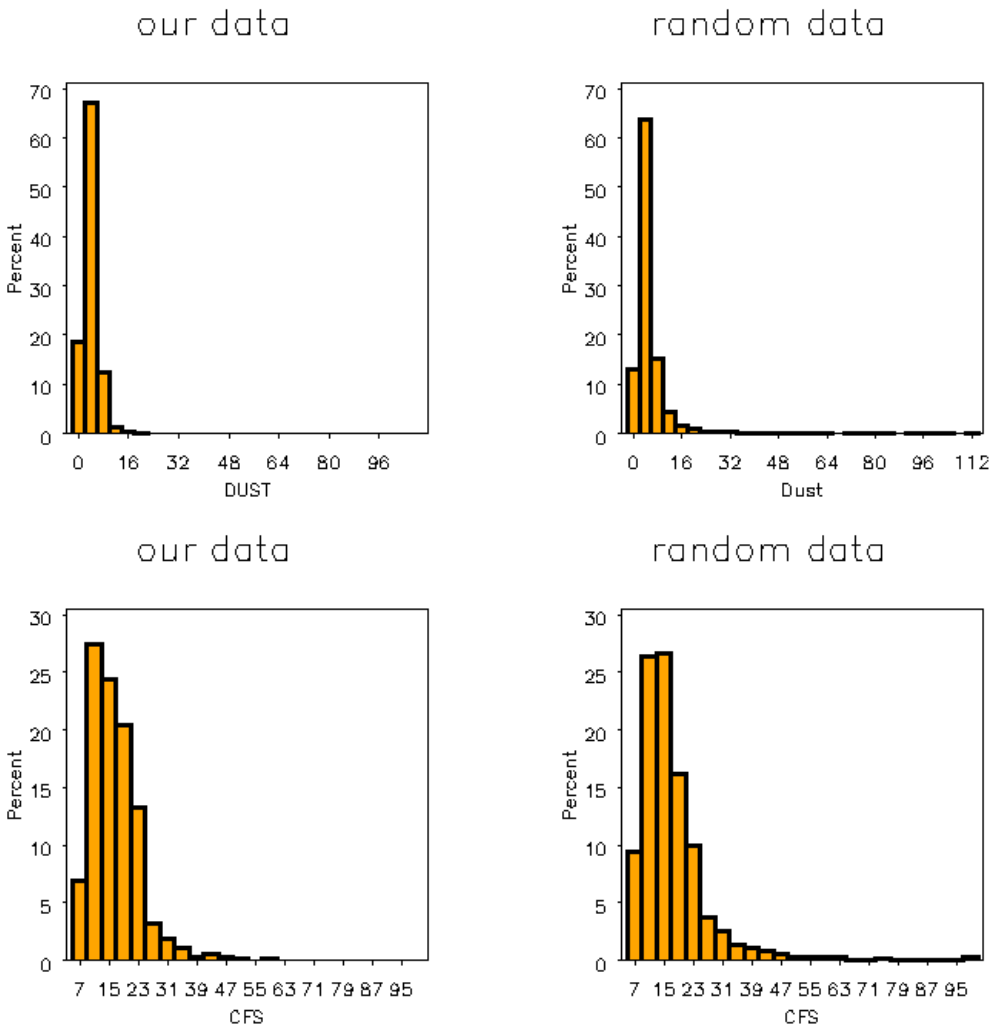


Supplementary Figure 2. Predicted effects of factors dG13 and 'any' on ACR, using the ACR prediction model. While each factors was analysed, the effects of every other factor were fixed to their mean levels. The blue line is the predicted value; the upper and bottom red lines are the upper and lower 95% confidence limits, respectively.





Supplementary Figure 3. Distribution of factor levels, used in LSI and ACR prediction models, in our dataset and in a random dataset, composed of 10000 randomly selected probes from the human genome.



Supplementary Figure 4. Distribution of CFS and 'dust' factor values in our dataset and in a random dataset composed of 10000 randomly selected probes from the human genome. As one can see, the probes in our dataset had been previously filtered by eliminating probes with high CFS and 'dust' values.

Supplementary Table 1. Descriptions and effects of sequence-based factors on ACR and LSI. Type – type of the factor: either categorical (C) or numeric (N); Values – range of possible values the factor can have ("continuous" designates that any value is theoretically possible); R^2 – percent of variability explained by factor (prediction power); P – the p-value for the significance of the factors.

SEQUENCE-BASED FACTORS							
FACTOR	**DESCRIPTION**	**TYPE**	**VALUES**	**LSI**		**ACR**	
				R^2	P	R^2	P
dG15	The Gibbs free energy of probe's last 15 bases binding to its target DNA. All dG-s are calculated using FASTAGREP (available from http://bioinfo.ebc.ee/download/), which utilises nearest-neighbour DNA thermodynamics (SantaLucia, 1998)	N	continuous	5.4%	<.0001	1.4%	<.0001
dG14	The Gibbs free energy for the last 14 bases	N	continuous	5.4%	<.0001	1.6%	<.0001
dG13	The Gibbs free energy for the last 13 bases	N	continuous	5.4%	<.0001	1.8%	<.0001
dG12	The Gibbs free energy for the last 12 bases	N	continuous	5.0%	<.0001	1.8%	<.0001
dG11	The Gibbs free energy for the last 11 bases	N	continuous	4.3%	<.0001	1.5%	<.0001
dG10	The Gibbs free energy for the last ten bases	N	continuous	3.6%	<.0001	1.1%	0.0006
dG9	The Gibbs free energy for the last nine bases	N	continuous	3.0%	<.0001	0.9%	0.0015
dG8	The Gibbs free energy for the last eight bases	N	continuous	3.3%	<.0001	1.1%	0.0004
dG7	The Gibbs free energy for the last seven bases	N	continuous	3.6%	<.0001	1.1%	0.0006
dG6	The Gibbs free energy for the last six bases	N	continuous	3.6%	<.0001	0.8%	0.0035
dG5	The Gibbs free energy for the last five bases	N	continuous	3.5%	<.0001	0.8%	0.0034
dG4	The Gibbs free energy for the last four bases	N	continuous	2.1%	<.0001	0.8%	0.0039
dG3	The Gibbs free energy for the last three bases.	N	continuous	1.3%	0.0001	0.5%	0.0355
dust	A score reflecting the complexity of a probe's sequence. A probe containing many simple repeats has a high 'dust' score, while a probe containing various different nucleotide combinations has a low 'dust' score (*ftp://ftp.ncbi.nlm.nih.gov/pub/tatusov/dust/*)	N	0, 1, 2, 3…n	0.3%	0.0684	0.0%	0.7051
total_Aprc	Percent of A nucleotides in the whole probe sequence	N	0-1	0.7%	0.1634	0.6%	0.1031
Factors related to the 3' end of probes							
N1	Type of the nucleotide at the 3' end of the probe. Can be either A, C, G or T	C	A, C, G, T	4.5%	<.0001	0.4%	0.2084
N2	2nd nucleotide from the 3' end of the probe	C	A, C, G, T	4.6%	<.0001	0.5%	0.1163
N3	3rd nucleotide from the 3' end of the probe	C	A, C, G, T	2.4%	<.0001	0.1%	0.8191
N4	4th nucleotide from the 3' end of the probe	C	A, C, G, T	2.8%	<.0001	0.7%	0.0394
N5	5th nucleotide from the 3' end of the probe	C	A, C, G, T	1.1%	0.0049	0.6%	0.0610
N6	6th nucleotide from the 3' end of the probe	C	A, C, G, T	0.3%	0.3590	0.1%	0.6392
N1N2	Combination of the 1st and 2nd nucleotide, counting from the 3' end of the probe	C	AA,AC,AG,AT,CA…TT	9.8%	<.0001	2.1%	0.0540
N1N3	Combination of the 1st and 3rd nucleotide	C	AA,AC,AG,AT,CA…TT	7.5%	<.0001	1.4%	0.3852
N1N4	Combination of the 1st and 4th nucleotide	C	AA,AC,AG,AT,CA…TT	9.7%	<.0001	2.5%	0.0143
N1N5	Combination of the 1st and 5th nucleotide	C	AA,AC,AG,AT,CA…TT	6.8%	<.0001	1.5%	0.3233

N1N6	Combination of the 1st and 6th nucleotide	C	AA,AC,AG,AT,CA…TT	5.8%	<.0001	1.5%	0.3206
N2N3	Combination of the 2nd and 3rd nucleotide	C	AA,AC,AG,AT,CA…TT	9.8%	<.0001	2.1%	0.0639
N2N4	Combination of the 2nd and 4th nucleotide	C	AA,AC,AG,AT,CA…TT	8.7%	<.0001	2.3%	0.0332
N2N5	Combination of the 2nd and 5th nucleotide	C	AA,AC,AG,AT,CA…TT	7.3%	<.0001	2.3%	0.0314
N2N6	Combination of the 2nd and 6th nucleotide	C	AA,AC,AG,AT,CA…TT	5.3%	<.0001	1.2%	0.5706
N3N4	Combination of the 3rd and 4th nucleotide	C	AA,AC,AG,AT,CA…TT	5.7%	<.0001	1.7%	0.1741
N3N5	Combination of the 3rd and 4th nucleotide	C	AA,AC,AG,AT,CA…TT	4.2%	<.0001	1.2%	0.5594
N3N6	Combination of the 3rd and 6th nucleotide	C	AA,AC,AG,AT,CA…TT	4.2%	<.0001	0.8%	0.8561
N4N5	Combination of the 4th and 5th nucleotide	C	AA,AC,AG,AT,CA…TT	5.9%	<.0001	2.5%	0.0197
N4N6	Combination of the 4th and 6th nucleotide	C	AA,AC,AG,AT,CA…TT	4.3%	<.0001	1.4%	0.3497
N5N6	Combination of the 5th and 6th nucleotide	C	AA,AC,AG,AT,CA…TT	2.5%	0.0183	1.6%	0.2194
Aprc	A nucleotide percentage among last six bases	N	0-1	3.4%	<.0001	0.5%	0.0179
Cprc	C nucleotide percentage among last six bases	N	0-1	7.7%	<.0001	0.6%	0.0096
Gprc	G nucleotide percentage among last six bases	N	0-1	1.2%	0.0002	0.1%	0.3976
Tprc	T nucleotide percentage among last six bases	N	0-1	0.6%	0.0093	0.1%	0.2233
Arank	Weighed score of A nucleotide presence among last six bases. 1st position from the 3' end gives six points, 2nd position gives five points, and so on	N	0, 1, 2 … 21	3.3%	<.0001	0.7%	0.0042
Crank	Weighed score of C nucleotide presence among last six bases	N	0, 1, 2 … 21	9.1%	<.0001	0.7%	0.0036
Grank	Weighed score of G nucleotide presence among last six bases	N	0, 1, 2 … 21	2.7%	<.0001	0.0%	0.6908
Trank	Weighed score of T nucleotide presence among last six bases	N	0, 1, 2 … 21	0.5%	0.0211	0.4%	0.0319
Factors related to predicted primer dimers and self-priming							
end1	These factors reflect three types of strongest primer-primer binding energies, calculated using MultiPLX 2.0 software (Kaplinski et al., 2005). First, the maximum free energy of a probe binding itself or identical probe in such a way that the 3' end of one of the probes is bound and other is unbound	N	continuous	0.9%	0.0012	0.8%	0.0024
end2	The maximum free energy of a probe binding itself or identical probe in such a way that the 3' ends of both probes are bound and can give a signal	N	continuous	1.0%	0.0007	0.7%	0.0046
any	The maximum free energy of a probe binding another identical nearby probe in such a way that the 3' ends of both probes remain unbound and therefore no signal can be given	N	continuous	1.1%	0.0004	0.9%	0.0010
CFS	Chance (probability, %) of getting false signal in genotyping due to self-priming. The formula for calculating that score is specifically designed for APEX oligos based on nearest-neighbour DNA thermodynamics (SantaLucia, 1998). The score can vary between 0-100 (0 being the best result)	N	0, 1, 2 … 100	0.6%	0.0084	0.1%	0.2849

Factors related to mismatches and SNPs							
mms	Number of known mismatches between the probe and its target	N	0, 1, 2 … 25	1.8%	<.0001	0.0%	0.9082
snps	Number of SNP in the probe sequence	N	0, 1, 2 … 25	0.0%	0.7839	0.0%	0.8796
last_mms	Position of the mismatch nearest to the 3' end of the probe: 1 when at the 3' end; 26 when there were no mismatches. In our data all probes were 25 bases long	N	1, 2, 3 … 26	2.8%	<.0001	0.1%	0.2150
last_snp	Position of the SNP nearest to the 3' end of the probe	N	1, 2, 3 … 26	0.0%	0.6256	0.0%	0.8046

Supplementary Table 2. Descriptions and effects of the auxiliary factors on ACR and LSI. Type – type of the factor: either categorical (C) or numeric (N); Values – range of possible values the factor can have ("continuous" designates that any value is theoretically possible); R^2 – percent of variability explained by factor (prediction power); P – the p-value for the significance of the factors.

AUXILIARY FACTORS							
FACTOR	**DESCRIPTION**	**TYPE**	**VALUES**	**LSI**		**ACR**	
				R^2	P	R^2	P
PCR	Category of the PCR reaction that amplified the probe's target DNA. This factor allows us to analyse the impact of PCR reactions to the efficiency of the probes that are associated with it	C	1, 2, 3 … 49	11.4%	<.0001	15.1%	<.0001
PCR_Tprc	Percent of T nucleotides in the PCR product that contains the probe's target DNA	N	0-1	0.4%	0.1634	0.3%	0.2838
PCR_len	Length of the PCR product that contains the probe's target DNA	N	1, 2, 3…n	0%	0.9319	0.1%	0.5737
sample	Category of the dataset where that probe belongs to: *ABCR170*, *ABCR100* or *ARCAGE*	C	1, 2, 3	0.3%	0.2140	17.6%	<.0001
dG25	The Gibbs free energy of probe's last 25-16 bases binding to its target DNA. These are not considered for use in quality prediction model since we wish to predict quality for probes minimally 15 nucleotides long and for such probes dG16-dG25 can't be calculated.	N	continuous	2.7%	<.0001	0.8%	0.0034
dG24	The Gibbs free energy for the last 24 bases	N	continuous	3.0%	<.0001	0.7%	0.0051
dG23	The Gibbs free energy for the last 23 bases	N	continuous	3.5%	<.0001	0.8%	0.0032
dG22	The Gibbs free energy for the last 22 bases	N	continuous	4.5%	<.0001	0.9%	0.0012
dG21	The Gibbs free energy for the last 21 bases	N	continuous	5.0%	<.0001	1.1%	0.0005
dG20	The Gibbs free energy for the last 20 bases	N	continuous	4.9%	<.0001	1.2%	0.0004
dG19	The Gibbs free energy for the last 19 bases	N	continuous	4.8%	<.0001	1.1%	0.0006
dG18	The Gibbs free energy for the last 18 bases	N	continuous	4.8%	<.0001	1.1%	0.0006
dG17	The Gibbs free energy for the last 17 bases	N	continuous	5.2%	<.0001	1.2%	0.0003
dG16	The Gibbs free energy for the last 16 bases	N	continuous	5.4%	<.0001	1.3%	0.0002