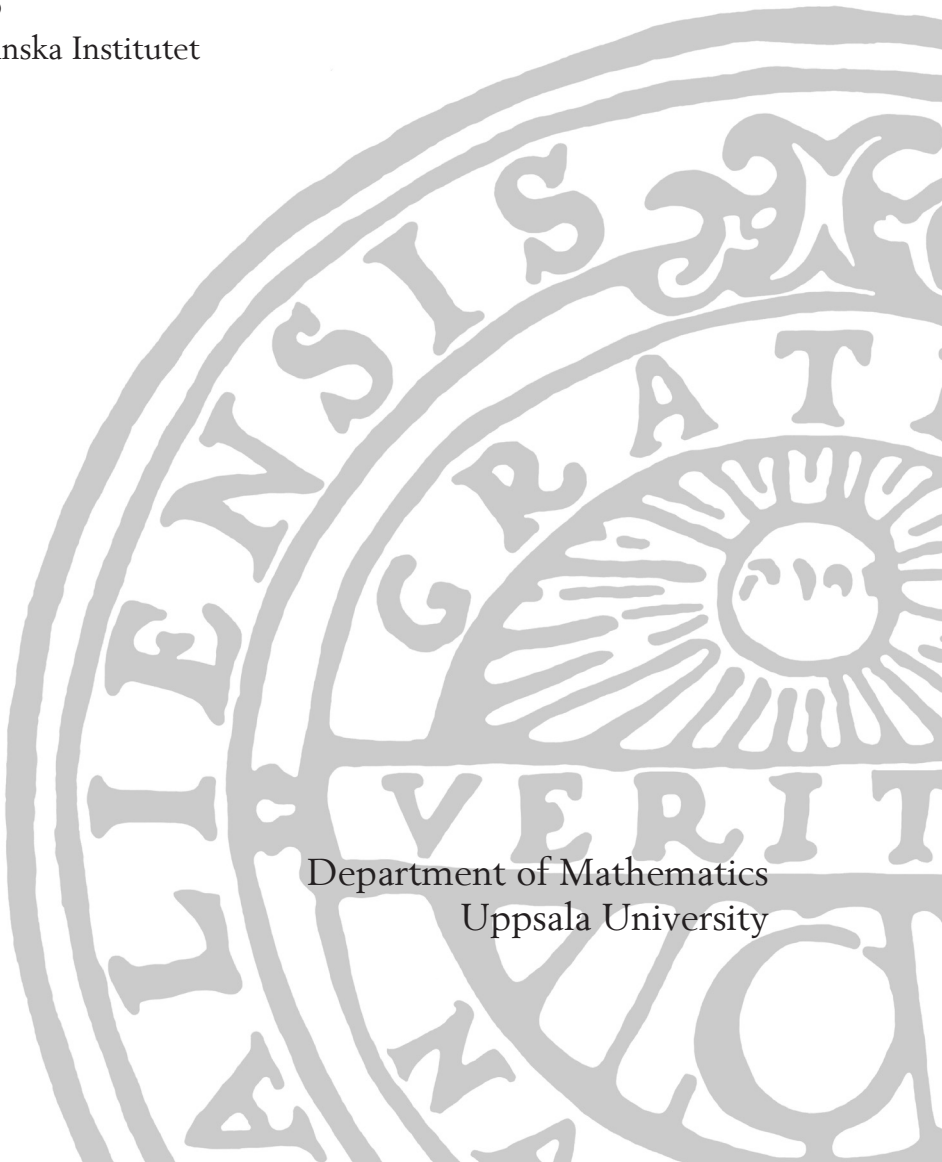
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2013:27

Likelihood Modelling of DNA Sequencing Data
Calibration and accuracy assessment

Olof Krantz

Examensarbete i matematik, 15 hp
Handledare Fredrik Jonsson, Karolinska Institutet
Examinator: Sven Erick Alm
Oktober 2013



Department of Mathematics
Uppsala University

Likelihood Modelling of DNA Sequencing Data

Calibration and accuracy assessment

Olof Krantz

October 18, 2013

Bioinformatics is the interdisciplinary field that focuses on handling large amounts of biological data. The amount of data collected by modern methods is so vast that new methods and tools for analysing, storing and organising are required, tools and methods to create knowledge from data.

Genotype likelihood modelling is a framework developed to accurately quantify the uncertainty in nucleotide sequencing and provide convenient statistical tools to extract knowledge from sequencing data. The *Uppsala Longitudinal Study of Adult Men* project is a longitudinal study. In ULSAM, which studies all available men born in Uppsala County between 1920 and 1924, the participants are investigated at the ages 50, 60, 70, 77, 82 and 88 years. The data acquired by the project has been used in many publications.

One of the aims of this report is to give an overview of the field of genome sequencing, which during the last decades has evolved rapidly. Another aim of the report was to give an overview of parts of the data from the ULSAM project and to use this data to evaluate the accuracy of genotype likelihood modelling.

1 Introduction

The structure of the DNA double helix was described in the early 1950s (Franklin et al. 2003; Watson and Crick 1953), yet the information contained within was still unknown.

In 1972 and 1976 the first complete gene and the first complete genome was published by Walter Fiers and his coworkers (Min Jou et al. 1972; Fiers et al. 1976). Right after that, in 1977, Frederick Sanger published a method for rapid DNA sequencing with chain-terminating inhibitors and also the first complete DNA genome, the genome of bacteriophage φX174 (Sanger et al. 1977a; Sanger et al. 1977b).

In 1987 the first fully automated DNA sequencer was ABI 370 which was produced by Applied Biosystems.

Several new methods were published during the 1990s, also known as "next generation" sequencing technologies. Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Tech-

nology in Stockholm published their method of pyrosequencing (Ronaghi et al. 1996). Massively parallel signature sequencing, MPSS, from Lynx Therapeutics (Zhou et al. 2006) and later a parallelized version of pyrosequencing from 454 Life Sciences (Margulies et al. 2005) are two early technologies considered to be "next generation" sequencing technologies. The technologies relies heavily on computer processing after reading fragmented and amplified genetic material to reassemble the fragments, known as aligning the reads.

After the completion of multiple reference genomes the task of cataloguing intra-species variation and associating genotype with phenotype was made simpler or even possible. What is known as *resequencing* is when already sequenced genetic material is sequenced again to catalogue variations and mutations. The sequencing of whole genomes is still an expensive process, one has thus aimed at lowering the cost of resequencing by limiting the scope of the process. Parts of special interest are selected using different technologies (Stratton 2008) and these fractions of the genome are then resequenced, giving high quality genomic maps of key parts of genomes.

Mutation take place whenever cells replicate. There are different types of mutations - additions and deletions for example. One of them is called single-nucleotide polymorphism, SNP. SNPs are widespread non-lethal mutations of a single nucleotide. The mutations can render proteins malfunctional or non-existant, but the mutations does not necessarily do anything directly. Sometimes they are associated with other genetic variations that in turn have a direct effect on the cells.

The detection of SNPs is enough for many applications. It is used in forensic analysis, measuring predisposition to disease or somatic mutations in cancers. With the aim of detecting if a variant is present or not the DNA microarray can be used.

The DNA microarray (Schena et al. 1995), or bio chip, is a plate made of class, plastic or silicon covered in complementary single-stranded DNA, complementary to the different sequences of interest. The amplified and labeled DNA sample is poured over the plate and the sample hybridize with the strands attached to the plate. When the plate is read the wells where DNA hybridized will emit light and those wells that are empty will remain dark. The position of the emitted light will determine which sequences that have hybridized and thus are contained in the sample.

Though having different approaches and methods the many multi-step processes of DNA-sequencing often have many steps in common. The samples need to be purified and amplified before the analysis can take place. The data reads created by the hardware has to be translated into genetic code, called base calling. After base calling the fragments of genetic code are aligned and analysed. Along all these steps are obstacles and problems that need to be overcome.

One of the major problems when assessing the genetic sequence of a sample is determining the quality of the results. In the base calling phase there are numerous ways to estimate the quality of each called base. A computer program which has had great impact on how we today measure quality is *phred*. phred is a computer program written in the late 1990's to accurately determine the most likely bases in a DNA-sequence given a chromatogram and assign to each one an accuracy value, q, accurately enough to eliminate the need for human supervision (Ewing et al. 1998).

The quality score, q, is calculated from the error probability, p, with

$$q = -10 \cdot \log_{10}(p),$$

thus, the larger q the more likely the base is called correctly (Ewing et al. 1998).

Though the algorithms for assigning quality scores have changed and differ between the methods of DNA sequencing, the phred scale is widespread and is also used for other parts of the process such as the genotype and variant calling, which are difficult due to ambiguity of data and the accuracy needed to do it correctly.

Tools and pipelines developed for processing of DNA sequencing data often distinguish between the aims of "variant calling" and "genotype calling". The former is referring to identification of variable sites and regions in the genome and the latter is referring to identifying individual samples' characteristics at the varying sites (Nielsen et al. 2011).

1.1 The structure of sequencing data

After base calling and aligning of the fragmented sequence reads the data can be displayed as seen in Figure 1. We see an example DNA-string where 4 reads have been aligned. The first line is a line consisting of the consensus sequence.

The mapped read depth, MRD, of a position is a measure of how many DNA sequence reads that after the aligning phase cover the current position. In Figure 1 the reads are aligned to a reference sequence, and the depth of a called genotype or variant is the number of reads that cover that particular position. The position 11 in the reference genome of Figure 1 has a read depth of 4, meanwhile position 1 has a read depth of 1.

```
GGGGGGTCAACTGGTGGGGGCC
GGGGGGTCAACTG
  GGGGTCAACTGGTGGGGGCC
    GGTCAAC
          CTGGTGG
```

Figure 1: Example of read depth of 4 over the base C. The underlined line is the previously established consensus sequence to which the read fragments are aligned.

The depth affects the quality of the read. High coverage gives more accurate genotype-/variant calling (Bentley et al. 2008). Thus it is important to make sure that read depth is deep enough and that it also is uniform across the genome.

Since sampling of sequence fragments is close to random the coverage is approximately Poisson distributed (Bentley et al. 2008). To see whether this holds one can compare the mean read depth, m, to the variance of the read depths, v. The Poisson distribution has a variance equal to it's mean and if $v > m$ the sample is over dispersed. In the case of read depths, due to certain dependencies on GC content this is only approximate and the variance is usually larger than the mean. The model might be too simple and not as fitting as believed and another more complex model might suit better.

To measure the coverage uniformity and quality of a read the inter-quartile range, IQR, can be utilized. The IQR is the difference between the lower and upper quartile of the read depths and gives an indication of the variance of the coverage.

As mentioned previously, the process of calling variants depends on many factors, among them prior information given by, if available, the reference genome about the given position and what the new aligned reads indicate.

In a diploid cell, the genome is in the form of two homologous chromosomes, meaning that there are two copies of the DNA in the cell. When the genes on the two copies are identical the

genes are called homozygotic, if they differ they are called heterozygotic.

When trying to genotype samples the result of a homozygote will contain only one genetic sequence while a heterozygote will have approximately half of the reads with one variant and the other half with a second variant, see Figure 2. It is impossible to know if a sample comes from a homozygotic or heterozygotic sequence beforehand and this has to be established after aligning using statistical methods.

(a) Homozygote

(b) Heterozygote

Figure 2: A fabricated heterozygote and a fabricated homozygote sample. The underlined line is the reference genome.

The risk of errors during the amplification, read and aligning phases imposes a great deal of uncertainty into the process. A heterozygote sample may very well give rise to results like the one seen in Figure 2a, and a homozygote might result in Figure 2b. To decrease the level of uncertainty you can increase the read depth.

Since the genomic material originates from a string of nucleic acids it is also important that the machinery does not introduce a systemic bias based on the position of the base along this string. By plotting the read depths against position one can visually discern any systemic biases based on position.

1.2 Genotype likelihood modelling

To answer the question whether reads that differ are from a heterozygotic sample or if they are just erroneous reads one approach used is likelihood modelling (Li 2011).

The idea is to determine how likely the observed data is given a genotype and a genomic position. Consider, for simplicity, a binary situation where the data can either equal the reference genome or to one alternative allele at a specific variant site (i.e. a biallelic SNP). Given sample data, let k denote the number of reads spanning the variant site. The observed data, $x = (x_1, \ldots, x_k)$, will then denote an event in the event space $\{0, 1\}^k$. Assuming sequencing and aligning errors are non-existing we either have the case of a heterozygotic genotype where all 2^k events are equally likely or we have the case of a homozygotic genotype were only the events $\{1\}^k$ or $\{0\}^k$ are possible.

We then examine the *likelihood* of the given data under the assumption of either of cases homozygotic with the alternative and reference allele and heterozygotic with both alleles. Let $g = 0, 1, 2$ denote the number of alternative alleles. The likelihood of data given the number of alternative alleles can be expressed as:

$$\mathcal{L}(g) = P(x \mid G = g) = \begin{cases} I\{x = (0, \ldots, 0)\}, & g = 0, \\ 2^{-k}, & g = 1, \\ I\{x = (1, \ldots, 1)\}, & g = 2. \end{cases} \tag{1}$$

This likelihood model can be extended to take into account errors made during the base calling and aligning phases. The error probability of *switch errors*, ε_j for $j = 1, \ldots, k$, the probability of a read produced being interpreted erroneously (reference interpreted as alternative and vice versa) are estimated by the algorithms utilised by programs such as phred. These programs base their estimates on the characteristics of the trace received from the sequencing machinery. phred was created to create quality scores based on a chromatogram, it used parameters such as peak spacing and peak resolution.

Assuming independence between reads, we may then extend model (1) into model (2):

$$\mathcal{L}(g) = P(Data \mid G = g) = \begin{cases} \prod_{x_j=1} \varepsilon_j \prod_{x_j=0} (1 - \varepsilon_j), & g = 0, \\ 2^{-k}, & g = 1, \\ \prod_{x_j=0} \varepsilon_j \prod_{x_j=1} (1 - \varepsilon_j), & g = 2. \end{cases} \tag{2}$$

Whenever *prior information* is available on the prevalence of the three genotypes in the study population, *genotype likelihoods* may be converted into so-called *genotype posterior probabilities* according to Bayes' rule:

$$P(G = g \mid Data) = P(Data \mid G = g) \times \frac{P(G = g)}{P(Data)} = \mathcal{L}(g) \times \frac{P(G = g)}{P(Data)}. \tag{3}$$

Since the posterior probabilities must sum to one, there is no need for prior information of the given data in order to convert given likelihoods into posterior probabilities.

The model makes the assumption that only two nucleotides are present at a given site. This is sub optimal but often sufficient (Li 2011) since triallelic variants are very uncommon in practice.

1.3 Calibration and verification of likelihood modelling

As stated previously, the algorithms used to assign quality values for our data are not the same as used in the original paper describing *phred*, but the importance of well calibrated models is still there.

This report will try to discern whether the posterior genotype probabilities, derived from likelihood estimates and prior information given by the tool chain, are well calibrated or not. Data from "next-generation" sequencing will be compared to data from SNP microarray sequencing, the observed error frequencies will be translated into error probabilities and compared to the given estimate of the posterior error probabilities.

The nature of the logarithmic quality measure makes observing and verifying the error rate of reads of high quality difficult, mostly due to the vast amount of reads needed. Therefore, in this report only reads of lower quality will be analysed.

2 Materials and Methods

2.1 Hardware and software

The "next-gen" sequencer used is the Illumina HiSeq 2000, it is a genome sequencer that uses sequencing by synthesis (SBS) technology. It uses solid phase amplification of the fragmented DNA sample. The cloned clusters are then sequenced using a proprietary four colour cyclic reversible termination method (Metzker 2010).

Before sequencing the samples the TruSeq Exome Enrichment Kit was used to amplify samples. It is a set of reagents and probes to amplify specific regions in a sample of DNA. The probe set spans more than 200000 exons. The targeted regions, genes or exons are isolated and amplified.

The resulting data from the sequencing machinery was processed using the Genome Analysis Toolkit, GATK, which is a software suite that processes data from *next generation* sequencers. The pipeline does read mapping, realignment around indels, genotyping and SNP discovery (McKenna et al. 2010).

The variant caller used is the Unified Genotyper from the GATK group. It uses variational bayesian models to call SNPs and indels from one or multiple samples (DePristo et al. 2011).

The Burrows-Wheeler Aligner is a piece of software written to efficiently align reads to a longer reference sequence, i.e. the human genome. It was used to align the genetic fragments. It uses two different algorithms, one for reads shorter than 200bp called bwa-short (Li and Durbin 2009) and one for longer reads up to 100kbp called bwa-SW (Li and Durbin 2010).

VCFtools is a program package designed for working with VCF files, such as those generated by the 1000 Genomes Project. The aim of VCFtools is to provide methods for working with VCF files: validating, merging, comparing and calculate some basic population genetic statistics.

SAM tools is a toolkit for manipulating SAM (Sequence alignment/Map) formatted data. SAM is a flexible and generic storage format for sequence alignment data.

R together with the libraries tikzDevice, ggplot2, reshape, plyr, binom, scales were used for the creation of graphics and the final statistical analysis.

2.2 Data

The *Uppsala Longitudinal Study of Adult Men* (Gustafsson 2013; ULSAM 2013), ULSAM, is a cohort study that was initiated 1970-73 when all 50 year old men in Uppsala County were invited to participate in the study. Two thousand (2322) persons agreed to participate and were examined. They were later reinvited for another examination at the age of 70. At this stage 1221 participated, the decrease in participation is mostly due to death and the fact that some participants moved out of Uppsala County. There has also been followups at the age of 50, 60, 70, 77, 82 and 88. Though the number of participants has decreased with the passing of time, a large portion of the subjects died and another large portion moved out of Uppsala county.

The participants answered a questionnaire regarding general as well as medical health. The examination was done in fasting state and blood samples were collected and blood pressure measured using a calibrated sphygmomanometer after 10 minutes of rest.

The ULSAM project has been approved by the Ethics committee of Uppsala University. The participants were informed and provided written consent.

This report studies 12 samples from the ULSAM project that were selected and sequenced using both microarray and "next-gen" techniques. SNPs detected during sequencing with the "next-gen" technique was compared to SNPs found using the more reliable, but less flexible, microarray chip. The chip contains information about roughly $1.6 \cdot 10^6$ variants and the data sequenced using the "next-gen" technology contains data of the samples' entire exomes. In this analysis only $2.4 \cdot 10^5$ of these were scrutinised. The reasons for this are *a)* that the statistical strength is not enough to identify variations unique to a sample in this small material and *b)* that only the samples' exome has been purified and sequenced, the exome makes up about 1% of the entire genome, and the chip does contain intron specific sequences as well.

Microarray sequencing

Illumina Omni2.5M and Illumina Metabochip was used for microarray genotyping. Quality control was performed on the samples and samples were excluded from the study based on genotype call rate < 95%, heterozygosity < 3 standard deviations, gender discordance, duplicated samples, identity-by-descent match and ethnic outliers. The variant calls were furthermore excluded if they were monomorphic SNPs, had a Hardy-Weinberg equilibrium P-value $< 1 \times 10^{-6}$, had a genotype call rate < 99.9% and MAF < 1%. Furthermore, SNPs were excluded if they had large position disagreements, did not map to genome, mapped more than once or had bad probe assays (Gustafsson 2013).

After quality control the genotyped data in part one of the ULSAM consisted of 1,179 samples and 1,621,833 SNPs.

Twenty four of these 1,179 samples were randomly selected for sequencing and the 12 best ones, based on number of reads, were selected. These 12 were also genotyped using "next-gen" sequencing. A summary of the data available for analysis can be seen in Table 1.

This data was used to create the prior information needed in the likelihood model described previously. The prior information consists of normalised allele frequencies from the all the 1,179 samples sequenced with the microarray technology.

"Next-gen" sequencing

Exome sequencing was done at the SNP & SEQ Technology Platform, Department of Medical Sciences, Uppsala University(*MOLMED*). Sequencing libraries where prepared with TruSeq DNA library preparation kit (Illumina Inc.) and the TruSeq Exome Enrichment Kit (Illumina Inc.) was used for exome capture on the libraries. The enriched libraries where sequenced on the Illumina HiSeq2000 using 100bp paired-end sequencing. BWA was used to align reads and The Genome Analysis Toolkit (GATK) with the UnifiedGenotyper was used for variant calling and to refine the data.

The following variables were available for each SNP.

- **Position** Reference position on the given chromosome.
- **Reference** Variant according to reference genome used when calling.
- **Alternative** Alternative allele.
- **Chip call** Number of alternative alleles for a given individual at the given position, according to microarray genotyping.

Property	Value
Mean mapped read depth, MRD	28.4
Number of samples	12.0
Entries removed	2905.0
Analysed SNPs	240384.0
SNPs per sample	20032.0

Table 1: Summary of data

- **"next-gen" call** Number of alternative alleles for a given individual at the given position, according to sequence based genotyping.
- **Quality value** Conditional genotype quality, encoded as a phred quality, for HiSeq2000 genotyping, the given individual and the given position.
- **Read depth** HiSeq2000 read depth for the given individual and the given position.
- **Priors** Prior information about each SNP used as weights in the likelihood model given as a 3-valued vector.
- **Likelihoods** Raw likelihoods used to call variants given as a 3-valued vector with the first value corresponding to the likelihood that 0 alleles differ from the reference genome, second value that 1 allele differs and the last value the likelihood that 2 alleles differ from the reference genome.

The data contains entries with multiple alternative alleles, these are ambiguous. These 2905 entries were therefore removed prior to analysis.

3 Results

The data analysed consists of samples from 12 individuals that were analysed as one to achieve enough data of the quality of interest.

3.1 Read depth

Investigating how read depth, D, is distributed over the length of the chromosomes was done by plotting D against position, see Figure 3c. The genome was sliced into 100 different slices, each covering an equal range of positions, and the mean read depth in each slice was plotted against the mean position of the corresponding slice. This was done because the plot otherwise would be unreadable. Figure 3c shows no sign of systemic bias.

To get a picture of how evenly the variants are covered by aligned reads we used a histogram showing the depths, seen in Figure 3b from one sample, looked at the IQR and also if it was overdispersed or not.

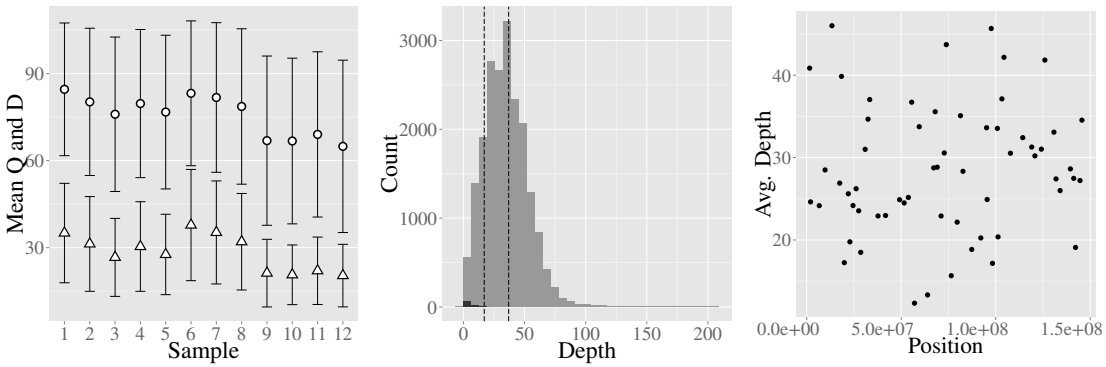
The difference between the 1st(17) and 3rd(37) quartile, the IQR, was 20 for the whole data set. The mean mapped read depth, m, was 28.4 and the variance of the depths, v, was 254 for the complete data set. This gives an overdispersion of read depths since the read depths were approximate Poisson distributed and $v > m$. The sample with the smallest variance had a sample

variance of 106 and the most varying sample had the variance 368. Mean variance per sample was 219.

The values mentioned above applied to the data set as a whole. If we broke it down to chromosome basis or sample basis the values differed slightly as seen in Table 2.

Property	Complete data set	Chromosome 8	Sample 7
IQR	20.0	19.0	19.0
MRD	28.4	28.5	35.2
Variance	254.0	214.0	316.0
Mean var. per sample	219.0	178.0	
Min. variance	106.0	76.4	
Max. variance	368.0	305.0	

Table 2: Comparison of properties of data from the entire data set, from data from a single chromosome and from a single sample.



(a) (b) (c)

Figure 3: *3a)* Bar plot showing individual samples' mean phred scaled quality± SD (circles) and mean read depth± SD (triangles). *3b)* Histogram showing distribution of read depths in sample 7. The dashed lines indicate the first and third quartile. IQR: 23 MRD: 35.24 VAR: 315.6. *3c)* Graph displaying the average depth at evenly spaced intervals in data from chromosome 8.

3.2 Genotype likelihoods

Due to the nature of things there will be errors in sequencing. To compensate for the errors we try to predict the error frequency for genotype calls. As a measure of quality for the call so they can be discarded if the quality is too low. The correctness of the prediction is important if it is to be used for excluding collected data. To verify that the predictions and observations match we compare them.

The model (1) gives 3 likelihoods for each genotype call, one assuming both alleles are equal to reference, one assuming one differing allele and lastly one that assumes that both alleles differ from the reference genome.

$$\mathcal{L}(g) = P(x \mid G = g) = \begin{cases} I\{x = (0, \ldots, 0)\}, & g = 0, \\ 2^{-k}, & g = 1, \\ I\{x = (1, \ldots, 1)\}, & g = 2. \end{cases}$$

The genotype with the highest likelihood will be considered the guess, this is written as $0, 1$ or 2 depending on the value g. Using data from the data microarray sequencing efforts as a golden standard we can determine which of the genotype calls that were made correctly, according to the microarray sequencing data.

The likelihoods computed are used to create posterior genotype probabilities using prior information, as seen in Equation 3, for each of the genotype calls. These posterior genotype probabilities can be compared to the actual observed allele frequencies in the data set. The observed frequencies are created, as mentioned previously, from comparing the microarray data with the "next-gen" genotype guesses.

g	Total	NND	Percentage
0	173587	2466	1.4
1	117159	16396	14.0
2	72355	2103	2.9
3	363101	20965	5.8

Table 3: Number of observations for each value of $g = 0, 1, 2$ or a combination of them (3). Total contains the values for the total number of genotype calls made with a phred scaled posterior genotype probability less than 30. The column NND, Number not displayed, is the number of values where the phred scaled posterior genotype probability was rounded to 0.

In table 3 and Figure 4 we see the results of comparing posterior genotype probabilities with observed allele frequencies to validate the accuracy of predictions. The observed frequency for a given probability is the fraction of correct genotype calls with an posterior genotype probability equal to the one of interest. In Figure 4 we have phred scaled posterior genotype probabilities on the horizontal axis and phred scaled observed allele frequency on the vertical axis. The red line denotes a perfect match between prediction and observation. Due to small sample size many genotype have a unique or very rare posterior genotype probability which results in very blunt observed allele frequencies, i.e. either 0 or 1 for those genotype calls that are data set-wide unique. To remedy this the data set was stratified, and the comparison made using the mean posterior genotype probability and mean observed allele frequency for each of the stratae. The stratification was made on the phred scaled posterior genotype probabilities and the length of the stratae was chosen as a trade off between resolution and certainty. The length 3 was chosen.

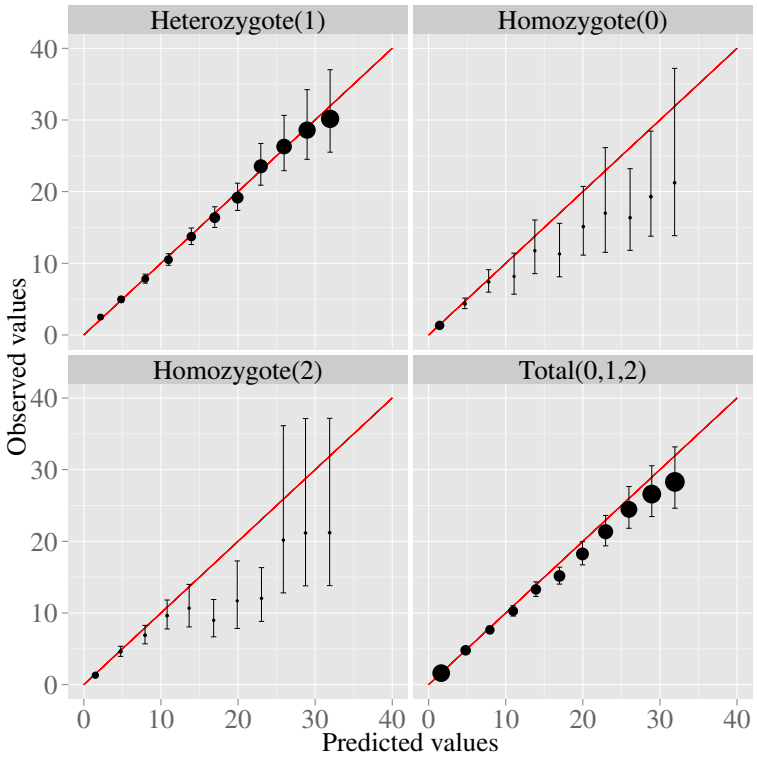


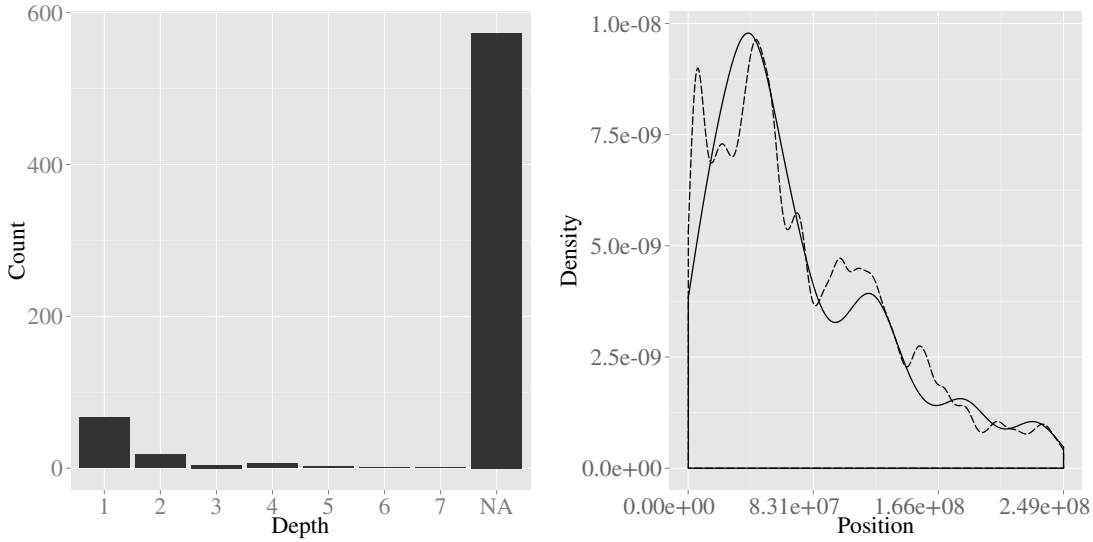
Figure 4: phred scaled observed genotype frequencies plotted against phred scaled probabilities greater than 0 predicted by the this likelihood model. Error bars denote 95% binomial confidence intervals, size denotes number of genotype calls used to average that point. *Upper left:* One allele differ from reference *Upper right:* Both alleles are equal to reference *Bottom left:* Both alleles are not equal to reference. *Bottom right:* All observed genotype frequencies vs. predicted genotype probabilities.

The data points where the phred scaled score was rounded to 0 has been omitted in Figure 4 because of the large discrepancy in number of observations, seen in table 3. The column *Total* consists of the genotype calls with a phred scaled posterior genotype probability of 30 or less and the column *NND* contains the number of genotype calls out of those in Total that had the phred scaled posterior genotype probability rounded to 0. The column *g* uses the same grouping as Figure 4 and is based on the genotype call made by the next-gen machinery. Only genotype calls with a phred-scaled quality value less or equal to 30 were plotted, for larger values the amount of data available is not enough. Error bars denote binomial 95% confidence intervals created using the Clopper-Pearson method. The size of points is corresponding to the number of observations in that particular strata, where the largest is composed of 3377 and the smallest only has 59 observations.

It is noted that the observed allele frequency and the posterior genotype probabilities correlate more closely when calculating the probabilities for heterozygotic genotypes than for homozygotic genotypes. The homozygotic posterior genotype probabilities show a tendency to be conserva-

tive, i.e. more correct calls are made than predicted. The overall correlation between posterior genotype probability and observed allele frequency was shown to not deviate from the line in Figure 4.

At certain base positions in the samples the HiSeq2000 has failed the genotype call and the result returned is *Not Available*, NA. Figure 5a shows read depth at those positions and Figure 5b shows how they were distributed throughout the genome. Figure 5a shows that the less coverage a position has, the harder it is to make a call, NA is interpreted as not being covered at all which. In Figure 5b we see that the density of failed variant calls followed that of successful variant calls. The density is smoother due to not being adapted to as many points as the successful one, only a few hundred compared to tens of thousands.



(a) Histogram of read depths where base calls have failed

(b) Density of failed base calls against base position. *Dashed:* Not failed *Solid:* Failed

Figure 5: Figures showing distribution of data points where variant calling for some reason failed, resulting in a variant call not being available, these points are plotted against position and depth. The histogram displays the number of failed calls at their respective depth, the density plots shows the density of failed and successful calls on top of each other.

4 Discussion

The microarray is an imperfect but widely used tool. It is used for diagnosing known genetic diseases and to categorize cancers (Hoopes 2008). One of its limitations is that you can not effectively do explorative genotyping, since the DNA sequence has to be known beforehand. This means that many hours of research has to be invested in determining which sequences to look for the first time. After the completion of large consensus genomes this need has decreased dramatically.

One of the goals of sequencing is to make it cheap enough to personalise treatments, to sequence patients' genomes in the same casual way they are sent to the radiology/imaging unit.

Since "next-gen" sequencing gives a broader and less focused view of a sample's genetic material than microarray sequencing it is of interest to see if the results from "next-gen" sequencing are as accurate as the clinically used microarray technology. Because if they are, "next-gen" sequencing can replace microarray sequencing as a diagnostic tool, which would simplify research and extend our knowledge of genetic disorders.

The 12 different samples are analysed in one single batch. This is motivated by them not being significantly different when it comes to reading depth and quality assessment, seen in Figure 3a.

4.1 Likelihood modelling - Calibration and accuracy

Whether you are creating consensus sequences, discovering new variants or diagnosing and determining correct treatments you rely on the correctness of data material you are basing your work on. One measure, that is widely used, of how correct genomic information is is the quality value assigned by the sequencer. The sequencer uses many variables when determining the quality of a base/variant call, such as agreement with existing consensus sequences, read depth and variance in between reads aligned over the current base. But regardless of what the sequencer bases the prediction on, regardless of how it does it, the correlation between predicted error rate and actual error rate is important, how well calibrated the sequencer is is important.

As seen in Figure 4, the heterozygotic base calls are well calibrated while the homozygotic have a tendency to drop below the expected line. In the model (2) we note that the estimated errors, ε_j for $j = 1, ..., k$, affect the homozygotic cases while leaving the heterozygotic cases entirely unaffected. The heterozygotic case treats all errors as if the read is from the other allele. This part of the model makes it more sensitive to errors and estimates from the read- and aligning phases in the homozygotic case than in the heterozygotic and results in more conservative estimates in homozygotic cases than in heterozygotic cases.

A few of the points' confidence intervals lie entirely outside of the expected range. This is most likely due to the sample size being to small to either take them all below the expected line or make all the confidence intervals cover the line. The method used to create the intervals is the Clopper-Pearson method which is based on the binomial distribution and not an approximation of it. The method gives wide intervals and it is argued that they cover more than the nominal 95%, on the other hand they are guaranteed to not cover less than 95%. As it is here, the overall impression is that the confidence intervals cover the expected line, and we err on the safe side.

The analysis of Figure 4 suggests that the overall accuracy of this likelihood model is adequate for error estimation and quality control. Figure 4 also indicates that the likelihood model is a sound model to estimate how likely different variants are, and that it performs particularly well when it tries to estimate how likely, or unlikely, the heterozygotic case is.

4.2 Depth, NA and systemic error

In this material the depth has a total inter quartile range of 20 and a mapped mean depth of 28.4. This could be considered good since the MRD of the failed ones and those with a quality value of 30 or less is 8.206 and 7.47, giving an indication that overall, the variant calls are well

covered. On the other hand, comparing these values with larger studies such as those published by the 1000 Genomes project (Abecasis et al. 2012; Abecasis et al. 2010) leads us to conclude that we are on below of what is usually called high coverage sequencing. In their reports they aim at 50-100 $\times$ exome coverage to be able to find rare single nucleotide polymorphisms. This is a work trying to map population wide variation and must therefore analyse large amounts of genomes and with a very high accuracy. After initial mappings have been done I suspect a lower read depth is required to determine the genotype of later samples.

Also, we note that the mean mapped read depth is much less than the sample variance. For an approximately Poisson distributed variable that gives that it is overdispersed. In this case perhaps a negative binomial distribution might suit better to model the behaviour of the read depth.

In Table 2 we see that there is little variation between the entire data set, a single chromosome and a single sample.

But of course, some bases are not covered. Those always result in a failed variant call. In Figure 5a we see that the majority of all incorrect calls are made when the position was not covered by any read; secondly, we see a clear negative correlation between read depth and rate of error. In Figure 5b we see how the incorrect variant calls are distributed against position, the distribution of correct and incorrect calls match each other well, this together with Figure 3c indicates that there is no existing systemic position-bias.

4.3 Data amount, quality and availability

The amount of data available for analysis imposes a restriction on the analysis, quality values scale logarithmically and to assess the calibration of quality values of 80 we would need 10^{79} observations, and since we only possess 240384 in total, regardless of quality value, we have opted to restrict ourselves to variant calls made with a quality value of 30 or less.

The data obtained using microarray technology is a analytic method as any other in, it is not perfect. The possibility that a number of variant calls have been falsely made and that it affects this study is very high due to the large amount of calls made. One might go as far as to ponder the possibility that if the sample used was very difficult to genotype at certain positions, that would increase the likelihood of results from both of the two different methods being incorrect at these positions. This information is not within reach for this study.

5 Acknowledgements

Thanks to Fredrik Jonsson for his advice and support, also thanks to Erik Ingelsson for supplying the data sets analysed data and comments, Yudi Pawitan and Sven Erick Alm for comments, and Olof Karlberg for supplementary information on “next-gen” sequencing.

References

Abecasis, G. R. et al. (2010). “A map of human genome variation from population-scale sequencing”. In: *Nature* 467.7319, pp. 1061–1073.

Abecasis, G. R. et al. (2012). “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422, pp. 56–65.

Bentley, D. R. et al. (2008). “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *Nature* 456.7218, pp. 53–59.

DePristo, M. A. et al. (2011). “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. In: *Nat. Genet.* 43.5, pp. 491–498.

Ewing, B. et al. (1998). “Base-calling of automated sequencer traces using phred. I. Accuracy assessment”. In: *Genome Res.* 8.3, pp. 175–185.

Fiers, W. et al. (1976). “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene”. In: *Nature* 260.5551, pp. 500–507.

Franklin, R. E. et al. (2003). “Molecular configuration in sodium thymonucleate. 1953”. In: *Nature* 421.6921, pp. 400–401.

Gustafsson, S. (2013). “Adiponectin: Genetic determinants and relations with subclinical cardiovascular disease”. Phd. Department of Medical Epidemiology and Biostatistics Karolinska Institutet.

Hoopes, L. (2008). *Genetic diagnosis: DNA microarrays and cancer.*

Li, H. (2011). “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data”. In: *Bioinformatics* 27.21, pp. 2987–2993.

Li, H. and R. Durbin (2009). “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14, pp. 1754–1760.

– (2010). “Fast and accurate long-read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 26.5, pp. 589–595.

Margulies, M. et al. (2005). “Genome sequencing in microfabricated high-density picolitre reactors”. In: *Nature* 437.7057, pp. 376–380.

McKenna, A. et al. (2010). “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”. In: *Genome Res.* 20.9, pp. 1297–1303.

Metzker, M. L. (2010). “Sequencing technologies - the next generation”. In: *Nat. Rev. Genet.* 11.1, pp. 31–46.

Min Jou, W. et al. (1972). “Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein”. In: *Nature* 237.5350, pp. 82–88.

MOLMED. *MOLMED*. URL: `http://molmed.medsci.uu.se/`.

Nielsen, R. et al. (2011). “Genotype and SNP calling from next-generation sequencing data”. In: *Nat. Rev. Genet.* 12.6, pp. 443–451.

Ronaghi, M. et al. (1996). “Real-time DNA sequencing using detection of pyrophosphate release”. In: *Anal. Biochem.* 242.1, pp. 84–89.

Sanger, F. et al. (1977a). “DNA sequencing with chain-terminating inhibitors”. In: *Proc. Natl. Acad. Sci. U.S.A.* 74.12, pp. 5463–5467.

Sanger, F. et al. (1977b). “Nucleotide sequence of bacteriophage phi X174 DNA”. In: *Nature* 265.5596, pp. 687–695.

Schena, M. et al. (1995). “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”. In: *Science* 270.5235, pp. 467–470.

Stratton, M. (2008). “Genome resequencing and genetic variation”. In: *Nat. Biotechnol.* 26.1, pp. 65–66.

ULSAM. *Uppsala Longitudinal Study of Adult Men*. URL: http://www2.pubcare.uu.se/ULSAM/index.htm.

Watson, J. D. and F. H. Crick (1953). "The structure of DNA". In: *Cold Spring Harb. Symp. Quant. Biol.* 18, pp. 123–131.

Zhou, D. et al. (2006). "Massively parallel signature sequencing". In: *Methods Mol. Biol.* 331, pp. 285–311.