Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

____ _______________

Yasminka Aleksandra Jakubek Marinkovic Date

A Biochemical Model of Hybridization on DNA Microarrays and its Application to Single Nucleotide Polymorphism and Copy Number Variation Genotyping in Trisomy 21 Individuals

By

Yasminka Aleksandra Jakubek Marinkovic
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

____
David J. Cutler
Advisor

____
Madhuri R. Hegde
Committee Member

____
Carlos S. Moreno
Committee Member

____
Stephen T. Warren
Committee Member

____
Michael E. Zwick
Committee Member

Accepted:

____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

____
Date

A Biochemical Model of Hybridization on DNA Microarrays and its Application to Single Nucleotide Polymorphism and Copy Number Variation Genotyping in Trisomy 21 Individuals

By

Yasminka Aleksandra Jakubek Marinkovic
B.A., Cornell University, 2007

Advisor: David J. Cutler, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2014

Abstract

A Biochemical Model of Hybridization on DNA Microarrays and its Application to Single Nucleotide Polymorphism and Copy Number Variation Genotyping in Trisomy 21 Individuals

By Yasminka Aleksandra Jakubek Marinkovic

DNA microarrays have several uses in biological research. In the field of human genetics, they are used to characterize genome-wide patterns of variation. Affymetrix Genome-wide Human SNP array 6.0 microarrays genotype ~900,000 single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) across the genome. Analysis methods for genotyping arrays rely on statistical approaches to generate accurate data. Two recurrent problems with these methods are evident. The first is the existence of batch effects. The second lies in the fact that these approaches often discard a large fraction of the raw data from probes that systematically fail across experiments. In order to address and understand these problems, we developed a novel analysis method that is based on a low-level model of hybridization on microarrays. We model binding between all probe-DNA duplexes that form on the array. In addition we model errors in probe synthesis, hybridization conditions (temperature, salt concentration), and details of the experimental protocol (target concentration, target fragmentation, wash stringency, and scanner settings). We used this model to predict probe intensities. The average correlation between expected and observed intensities was 0.701 with a range of 0.88 to 0.55. In this model batch effects are caused by differences in probe synthesis efficiency, target concentration, target fragmentation, wash stringency, and scanner settings. We used this model to develop a SNP and CNV genotyping algorithm that explicitly models batch effects and cross-hybridization. Our approach allows for the individual analysis of chips and can call SNPs and CNVs on chromosomes of any ploidy. We used this approach to analyze Down syndrome and normal samples. A significant percentage (13%) of SNPs that are targeted by Affymetrix 6.0 have high levels of cross-hybridization. Each SNP call has a quality score (QS). SNPs on trisomic chromosomes had lower QS scores (57% with QS> 0.99) than SNPs on diploid chromosomes (84% with QS > 0.99). Our approach uses direct estimates of DNA concentration to call CNVs. We called an average of 50 CNVs per samples of which 68% are in known CNV regions. Using only first-principles our method detects genetic variants with a comparable accuracy to current approaches.

A Biochemical Model of Hybridization on DNA Microarrays and its Application to Single Nucleotide Polymorphism and Copy Number Variation Genotyping in Trisomy 21 Individuals

By

Yasminka Aleksandra Jakubek Marinkovic
B.A., Cornell University, 2007

Advisor: David J. Cutler, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2014

Table of Contents

Figure and Tables

Chapter 1:

General Introduction

Introduction

Correctly identifying genetic variation is an integral part of human genetics and is important for several fields of study including hereditary disorders, cancer, and population genetics [1-7]. There are many technologies available for the analysis of genome wide genetic variation, one of them being genotyping microarrays. These arrays have two general uses, the detection of single nucleotide polymorphisms (SNPs) and structural genomic variants [8]. Advancements in microarray technology have allowed the development of genotyping microarrays that can query hundreds of thousands of SNPs and copy number variants (CNVs) simultaneously. These arrays provide an efficient tool for the analysis of genetic variation in large-scale human genetic studies [2, 4-8]. One of the most commonly used types of genotyping arrays are Affymetrix Genome-Wide Human SNP 6.0 (Affy 6.0) microarrays, which can genotype close to two million markers across the entire genome. These arrays are analyzed using statistical methods that rely heavily upon empirical data [9, 10]. In this dissertation we present a low-level model of hybridization on microarrays. In this model binding on an array is the same as in solution, with differences arising due to experimental details of the array production and processing. We then applied this model to Affy 6.0 arrays and show that using only first principles we can call SNPs and CNVs with relatively high confidence. This unique approach allows us to call variants for chromosomes of any ploidy, which we did for chromosome 21 of Down syndrome individuals.

DNA Microarrays

DNA microarrays are used to detect nucleic acids via hybridization of complementary sequences of oligonucleotides [11, 12]. They are made up of probes, single stranded DNA molecules that are attached to the array surface. The array surface is a grid, where each square on the grid is a probe spot also known as a feature. Each spot is made up of probes that are manufactured to include the same sequence. Affymetrix manufactures chips using photolithography, in which probes are manufactured 3’ to 5’ by adding one base at a time [13, 14]. The 3’ end is attached to a linker that anchors the probe to the array surface. Target refers to the DNA/RNA in the hybridization solution that is placed onto the array. Target DNA is derived from the sample of interest. The basic steps in DNA array processing are target preparation, hybridization, washing, and scanning. The target is fragmented using sonication or restriction enzymes and then labeled with a fluorophore. The solution with the target is then placed onto the array surface and allowed to hybridize for several hours. After the hybridization step, the chip is washed using a solution with a low salt concentration. This solution causes weakly bound targets to disassociate from the probes. In the final step the chip is scanned and the intensity for each probe spot is recorded. The intensity reading at each probe spot provides information regarding the concentration of labeled target sequences that are complimentary to the probe. They also provide information regarding the type of target sequences in the solution. Microarrays are used to query genomic DNA as well as RNA [11]. RNA microarrays provide information regarding the steady-state transcript levels of different mRNA isoforms and of non-coding RNAs. DNA microarrays are used for re-sequencing, SNP typing, and CNV detection. There are other uses of DNA microarrays including the detection of species-specific sequences in a mixed sample and the identification of protein binding sites in the genome [11].

Genotyping Arrays

The first generation of genotyping arrays targeted approximately 1500 SNPs, while current technologies are capable of genotyping close to two million markers per sample [8]. One of the most widely used genotyping arrays are Affy 6.0 arrays, which query 900,000 SNPs. Additionally, they have 115,000 monomorphic probes that target known CNVs and 831,000 monomorphic probes that are distributed across the genome. The intensity information from both the CNV and SNP probes are used to call CNVs. Illumina genotyping arrays, like Affy 6.0, are also widely used for genome wide SNP and CNV detection; however, the underlying technology used by Illumina is quite different from the technology used by Affymetrix [8]. In Illumina genotyping arrays, probes are not attached to the array surface; instead they are attached to beads. Also target DNA is not fluorescently labeled, instead A, C, G, and T nucleotides, which are differentially labeled, are used for single base pair extension.

To understand how these two technologies work, let us look at the example of an arbitrary SNP with two alleles, A and C. Illumina arrays query the SNP by using probes that are 50 bases long. The 50-mer is a perfect match to the nucleotides on the 5' end of the SNP. During hybridization the target DNA binds the probe up to but not including the SNP base. This hybridization process is followed by single base pair extension when a G or T base is incorporated into the sequence. Since the nucleotides are differentially labeled the signal can be used to infer the genotype for the SNP. Affy 6.0 arrays work differently; probes are directly attached to the array surface and are organized into probe spots. Affymetrix probes are 25 bases long. For each SNP that Affymetrix queries, there are two unique probes. For a SNP with alleles A and C, one probe is a perfect match to the A allele and the other a perfect match to the C allele. The mismatch is placed towards the center of the probe. For an autosomal SNP, there are three possible genotypes (AA, AC, CC). The general idea is that target DNA with the A allele binds much more strongly to the A probe than to the C probe. Also, the C allele binds more strongly to the C probe. Therefore, the signal intensities at each probe spot are a function of the number of A and C alleles. Overall the experimental details of Affymetrix and Illumina arrays are quite different. Since our goal was to explicitly model hybridization and the experimental details we focused on only one type of technology, Affy 6.0 arrays.

Methods for SNP Genotyping in Affymetrix Arrays

The basic principle behind all array SNP genotyping methods is to use probe intensities to call genotypes [8]. For simplicity, we will refer to the alleles of a diploid SNP as A and B and the three possible genotypes as AA, AB, BB. Also, probe A refers to the probe that perfectly matches allele A and probe B perfectly matches the

B allele. The first methods for automated SNP detection were designed for the analysis of data from Variation Detection Arrays (VDAs), a predecessor to genotyping arrays. One of the first methods for VDA analysis was Adaptive Background genotype Calling Scheme (ABACUS), a statistical approach, which uses the intensity at individual SNPs from a single chip, to calculate likelihoods for each genotype [15]. In this model probe intensities follow a normal distribution. When calls do not reach a certain quality threshold they are dropped. There are two important observations regarding the calls that tend to be dropped by ABACUS. First, heterozygote calls are more likely to get dropped than homozygote calls. Second, there are a subset of SNPs that tends to be dropped across samples. The observation that heterozygotes are harder to call than homozygotes is intuitive given that it is easier to call a 2 to 0 allele ratio than it is to call a 1 to 1 ratio. Explanations for the second observation are not quite as intuitive and are probably due to more than one aspect of the microarray experiment. For the first two generations of genotyping arrays, Affymetrix developed the Modified Partitioning Around Medoids (MPAM) and the dynamic model-base algorithm (DM) methods for SNP genotyping [16, 17]. DM was based on ABACUS; it analyses SNPs individually and tends to drop heterozygotes. On the other hand MPAM, creates genotype clusters, AA, AB, BB based on the observed probe intensities for a SNP across all samples. Problems with this latter approach arise when the sample size is small and for SNPs with a low minor allele frequency. Improved versions of this “cluster” approach to SNP genotyping include the Robust Linear Model with Mahalanobis Distance Classifier (RLMM), the Corrected Robust Linear Model with Maximum

Likelihood Classification (CRLMM), the Bayesian Robust Linear Model with Mahalanobis Distance Classifier (BRLMM) and Birdseed, the default analysis algorithm for Affy 6.0 arrays [9, 10, 18]. Improvements include normalization of data to account for batch effects (CRLMM) as well as Bayesian models that use DM to assign a prior to clusters, which improves calling for SNPs with low minor allele frequency [8, 10]. These "cluster" approach and improvements by Affymetrix in probe selection have made it possible to call SNPs with high accuracy; however, across experiments a significant fraction of SNPs (between 20% to 33%) are dropped, because they do not pass quality control [4-7]. Overall the "cluster" approach to SNP genotyping is very practical; however, it has not helped us understand why some SNPs are harder to call. The answer to this question is one of the major focuses of the work presented in this dissertation.

CNV Detection using Affymetrix Genotyping Arrays

The standard approach to CNV calling in Affy 6.0 arrays is made up of two distinct algorithms, Canary (copy number analysis routine) and Birdseye [10]. These algorithms are part of Birdsuite, the standard analytical framework for Affy 6.0 data that also includes the previously described Birdseed algorithm for SNP detection.

Canary was developed for the detection of common CNVs, defined as those with greater than 1% frequency, the authors refer to these as copy number polymorphism (CNP) [10]. The rest of the CNVs are referred to as rare/ de novo and these are detected using Birdseye. Canary uses predefined sets of probes to call

CNPs. Both algorithms use intensity data across chips to call CNVs. This makes data analysis susceptible to batch effects.

In Solution DNA Binding

When two DNA molecules come together, there is a Gibbs free energy (ΔG) value associated with the binding reaction [19]. ΔG can be experimentally determined; however, it is impossible to calculate ΔG experimentally for the infinite number of theoretical DNA duplexes. The most practical approach is to estimate ΔG from the DNA sequences. The most widely used method to approximate ΔG for two DNA molecules binding in solution is the nearest-neighbor (NN) model [19]. In this model the stability of a base pair is dependent on the two neighboring nucleotides. There are 10 unique NN pairs that can form in a perfectly matched DNA duplex. They are (AA/TT), (AT/TA), (TA/AT), (CA/GT), (GT/CA), (CT/GA), (GA/CT), (CG/GC), (GC/CG), (GG/CC). Where the bases are listed (5' to 3' / 3' to 5'). Each pair has a ΔG value that can be calculated using the enthalpy (ΔH) and entropy (ΔS) for each NN pair, as well as the hybridization temperature, and salt concentration. The NN model also accounts for differences between terminal base pairs. There are two ΔG initiation values, one for terminal A•T and the other for terminal G•C. Using the NN model, the ΔG value for biding between 5' -ATACG -3' to 3' -TATGC -5' is the sum of ΔG values for (AT/TA), (TA/AT), (GT/CA), (CG/GC), A•T initiation, and G•C initiation. The predicted ΔG for the hybridization reaction can be used to calculate melting temperature and other measures of the affinity between the two molecules.

ΔG values have also been estimated for mismatches [20-24]. For example, when 5' -ACG -3' binds to 3' -TAC -5', there are two mismatch NN pairs, (AC/TA) and (CG/AC). Overall the NN model for DNA hybridization in solution provides an efficient way to predict binding affinity for sequences with and without mismatched bases.

DNA Microarray Observations that are not Explained by In Solution DNA Binding

There are two striking observations regarding microarray probe behaviors that are not explained by in-solution binding kinetics. The first one is the difference in intensity between a probe that binds the forward target and one that binds the reverse target. The forward and reverse targets are perfectly complementary DNA molecules and should not be confused with the +/- strand labeling often used to refer to "sense" and "antisense" RNA. In the ABACUS study, all probes had a forward and reverse probe spot. A striking observation from this study was the consistent difference between the two and the fact that the G rich probe consistently had lower intensity than the complimentary C rich probe [15]. The second observation that is not explained by in solution binding kinetics is the position dependent destabilizing effect of mismatches. Mismatches towards the center of the probe have a much larger destabilizing effect than mismatches towards the edges [25]. This is the reason why Affymetrix places the query base for a SNP close to the middle of the probe.

Batch Effects

Batch effects are differences in array behavior that are independent of the genotype of the sample [26]. Batch effects can arise when the same sample is processed in different facilities using chips that are manufactured on different dates, or handled by different scientists. By increasing overall variability, batch effects can decrease power to detect a biologically relevant signal [26]. They can also have the opposite effect when they lead to what appears to be a biologically relevant observation that is confounded with a batch effect. For example, this can happen when cases and controls are genotyped at different locations [26]. There are several approaches to try to remove batch effects. Statistical methods include principal components analysis, which can be used to account for differences in the processing date and time [26, 27]. A different, not mutually exclusive approach is to design experiments in such a way that samples that are processed in the same group (done by the same technician, at the same time and place) consist of both cases and controls. However, even when time and place are accounted for, there are other sources of batch effects that are not well understood [26].

Summary

We developed a low-level hybridization model of binding on Affymetrix arrays in order to gain a better understanding of the factors that underlie the differences between in solution binding and binding on arrays, batch effects, and the reason why a significant fraction (20-33%) of probes on Affy arrays fail. In this model binding happens the same as in solution. In addition we modeled the experimental details. These include probe-manufacturing errors, salt concentration and hybridization temperature. Furthermore, we model cross-hybridization by directly calculating the biding affinities between all probes and targets. In addition we model target concentration and fragmentation as well as scanner setting. After developing the model we sought to answer two major questions: 1) Can we use this approach to model binding on arrays? 2) Can we apply this model to type SNPs and CNVs in diploid and triploid chromosomes?

References

1. **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.
2. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E *et al*: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls**. *Nature* 2010, **464**(7289):713-720.
3. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* 2013, **45**(10):1113-1120.
4. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Magi R *et al*: **Association analyses of**

249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010, **42**(11):937-948.

5. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, Korenaga M, Hino K, Hige S *et al*: **Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C**. *Nat Genet* 2009, **41**(10):1105-1109.

6. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL *et al*: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1**. *Nat Genet* 2009, **41**(3):324-328.

7. Myocardial Infarction Genetics C, Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ *et al*: **Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants**. *Nat Genet* 2009, **41**(3):334-341.

8. LaFramboise T: **Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances**. *Nucleic Acids Res* 2009, **37**(13):4181-4193.

9. Lin S, Carvalho B, Cutler DJ, Arking DE, Chakravarti A, Irizarry RA: **Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays**. *Genome Biol* 2008, **9**(4):R63.

10. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K *et al*: **Integrated genotype calling and**

association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008, **40**(10):1253-1260.

11. Stoughton RB: **Applications of DNA microarrays in biology**. *Annu Rev Biochem* 2005, **74**:53-82.

12. Southern EM, Maskos U, Elder JK: **Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models**. *Genomics* 1992, **13**(4):1008-1017.

13. McGall GH, Barone AD, Diggelmann M, Fodor SPA, Gentalen E, Ngo N: **The Efficiency of Light-Directed Synthesis of DNA Arrays on Glass Substrates**. *Journal of the American Chemical Society* 1997, **119**(22):5081-5090.

14. Pirrung MC, Fallon L: **Proofing of photolithographic DNA synthesis methods. Fabrication of DNA microchips**. *Abstr Pap Am Chem S* 1997, **213**:362-ORGN.

15. Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA *et al*: **High-throughput variation detection and genotyping using microarrays**. *Genome Res* 2001, **11**(11):1913-1925.

16. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G *et al*: **Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays**. *Bioinformatics* 2005, **21**(9):1958-1963.

17. Liu WM, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder TB, Webster TA, Dong S, Liu G *et al*: **Algorithms for large-scale genotyping microarrays**. *Bioinformatics* 2003, **19**(18):2397-2403.

18. Rabbee N, Speed TP: **A genotype calling algorithm for affymetrix SNP arrays**. *Bioinformatics* 2006, **22**(1):7-12.

19. SantaLucia J, Jr.: **A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics**. *Proc Natl Acad Sci U S A* 1998, **95**(4):1460-1465.

20. Allawi HT, SantaLucia J, Jr.: **Thermodynamics and NMR of internal G.T mismatches in DNA**. *Biochemistry* 1997, **36**(34):10581-10594.

21. Allawi HT, SantaLucia J, Jr.: **Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects**. *Biochemistry* 1998, **37**(26):9435-9444.

22. Allawi HT, SantaLucia J, Jr.: **Thermodynamics of internal C.T mismatches in DNA**. *Nucleic Acids Res* 1998, **26**(11):2694-2701.

23. Allawi HT, SantaLucia J, Jr.: **Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA**. *Biochemistry* 1998, **37**(8):2170-2179.

24. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J, Jr.: **Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches**. *Biochemistry* 1999, **38**(12):3468-3477.

25. Duan F, Pauley MA, Spindel ER, Zhang L, Norgren RB, Jr.: **Large scale analysis of positional effects of single-base mismatches on microarray gene expression data**. *BioData Min* 2010, **3**(1):2.

26. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data**. *Nat Rev Genet* 2010, **11**(10):733-739.

27. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA: **A multilevel model to address batch effects in copy number estimation using SNP arrays**. *Biostatistics* 2010.

Chapter 2:

A Model of Binding on DNA Microarrays: Understanding the Combined Effect of Probe Synthesis Failure, Cross-Hybridization, DNA Fragmentation and other Experimental Details of Affymetrix Arrays

Yasminka A. Jakubek[1,2] and David J. Cutler[1,*]

[1] Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA

[2] Graduate Program in Genetics and Molecular Biology, Emory University, Atlanta, GA, 30322, USA

* To whom correspondence should be addressed. Tel: 404-727-5388; Fax: 404-727-3949; Email: djcutle@emory.edu

Published in BMC Genomics (2012; 13: 737. Published online Dec 27, 2012)

Abstract

Background

DNA microarrays are used both for research and for diagnostics. In research, Affymetrix arrays are commonly used for genome wide association studies, resequencing, and for gene expression analysis. These arrays provide large amounts of data. This data is analyzed using statistical methods that quite often discard a large portion of the information. Most of the information that is lost comes from probes that systematically fail across chips and from batch effects. The aim of this study was to develop a comprehensive model for hybridization that predicts probe intensities for Affymetrix arrays and that could provide a basis for improved microarray analysis and probe development. The first part of the model calculates probe binding affinities to all the possible targets in the hybridization solution using the Langmuir isotherm. In the second part of the model we integrate details that are specific to each experiment and contribute to the differences between hybridization in solution and on the microarray. These details include fragmentation, wash stringency, temperature, salt concentration, and scanner settings. Furthermore, the model fits probe synthesis efficiency and target concentration parameters directly to the data. All the parameters used in the model have a well-established physical origin.

Results

For the 302 chips that were analyzed the mean correlation between expected and observed probe intensities was 0.701 with a range of 0.88 to 0.55. All available chips

were included in the analysis regardless of the data quality. Our results show that batch effects arise from differences in probe synthesis, scanner settings, wash strength, and target fragmentation. We also show that probe synthesis efficiencies for different nucleotides are not uniform.

Conclusion

To date this is the most complete model for binding on microarrays. This is the first model that includes both probe synthesis efficiency and hybridization kinetics/cross-hybridization. These two factors are sequence dependent and have a large impact on probe intensity. The results presented here provide novel insight into the effect of probe synthesis errors on Affymetrix microarrays; furthermore, the algorithms developed in this work provide useful tools for the analysis of cross-hybridization, probe synthesis efficiency, fragmentation, wash stringency, temperature, and salt concentration on microarray intensities.

Background

DNA microarray chips consist of large numbers of probes, single stranded DNA molecules attached to a solid surface, that hybridize to nucleic acids [1]. Microarrays have several uses in DNA analysis including CNV detection [2-5], re-sequencing [6], SNP typing [7, 8] , detection of species specific DNA in complex samples [1], and identification of protein-DNA binding sites [1]. They are also used to assess transcript levels in samples of coding and non-coding RNA [1, 9, 10]. In the field of human genetics, DNA microarrays are used to investigate disease [4, 11, 12], to study variation [5, 8], and to detect variants in clinical samples [9].

Collections of identical probes are called probe spots or features. Each probe spot consists of many copies of identical single stranded DNA molecules. Many DNA array designs have multiple features querying the same target DNA. Often one set of features queries targets on the forward strand of the DNA while the other set queries targets on the reverse strand. The first step in a DNA microarray experiments is to isolate and amplify the target DNA or RNA. Next, the amplified target is fragmented and fluorescently labeled. The labeled target solution is then hybridized to the chip where target binds to probe DNA. Following hybridization the chip is washed in order to eliminate non-specific binding. Finally, the chip is scanned and the fluorescent intensity measured for each feature.

The fundamental assumption behind a DNA microarray experiment is that the intensity measure for a probe spot correlates to the concentration of target bound to that spot, which in turn correlates to the amount of target in the original solution [6, 13]. However, the relationship between observed intensity and target DNA composition is not straightforward. Known variables with microarrays include high variance in intensity between probe spots, high variance in a single probe spot's intensity between chips, as well as background (non-specific binding) intensity differences between chips. Several studies have focused on the binding kinetics of DNA molecules attached to a solid surface and on cross-hybridization [14-18]. These studies have helped illuminate some aspects of the microarray experiment; however, several fundamental observations of microarray behavior remain poorly understood. First, when an array contains probe spots for both the forward and reverse target, simple liquid phase kinetics predict that both probe spots ought to have the same binding affinities and therefore should have equal amounts of bound target DNA [19]. However, in practice forward and reverse strand probe spots usually have significantly different intensity measurements [20]. These differences in intensity are observed in both GC and AT rich probes as well as probes with and without nucleotide runs (20). Second, liquid phase kinetics predicts that mismatches anywhere in the oligo (other than in the last 3 bases) ought to have equal effects on binding [19, 21-25]. However, mismatches near the center of the probe have a stronger effect on probe intensity compared to mismatches towards the edges [26]. Third, chips manufactured on different days often have subtly different binding properties (so-called chip-effects); as do chips processed by different facilities or on different days (batch effects) [27, 28].

The goal of this study is to attempt to understand all of these aspects. To do so we develop a detailed model of the DNA microarray experiment and then use this model to predict probe intensities for seven different microarray designs.

The basic assumption behind this modeling approach is that the hybridization kinetics of DNA binding on a chip are fundamentally the same as liquid phase kinetics. Apparent differences between liquid phase predictions and microarray observations arise from the combined effect of different aspects of the microarray experiment. In the model we include the effect of probe sequence, cross-hybridization, nucleotide position, and hybridization conditions. We model the combined effect of these factors in one step rather than normalizing the data for each factor in a stepwise manner [35, 36]. Unlike previous studies we do not adjust binding strength for nucleotides based on their position on the probe [14]. Instead the "positional" effect of nucleotides arises naturally in our model as a side effect of target DNA fragmentation and microarray synthesis. In particular we assume that microarray synthesis is not perfect [29-31]; more specifically, we assume that during probe synthesis individual A, C, G, and T nucleotides fail to incorporate at different rates. We also model abasic sites on the probe sequence where the probabilities that A, C, G, and T nuclotides become abasic are not necessarily equal to each other. Consequently differences in synthesis efficiency/abasic sites between nucleotides explain why forward and reverse DNA probe spots often have significantly different intensities and explain chip-effects. Additionally we explicitly model target DNA concentration, hybridization temperature, mean fragmentation size, wash stringency [15, 32, 33], and microarray scanner settings [34] which together with errors along the probe sequence give rise to batch effects. Previous studies have reported that different probe sequences have different saturation intensities [32], under our model this is expected given that each probe spot consists of a "forest" of probes and that the number of probes capable of binding target DNA is sequence dependent. Furthermore, the strength of the wash impacts the final number of probe/target duplexes [32].

Our model consists of two parts. First, we calculate binding affinities for the probes and target DNA in the hybridization solution. To do so we use the Langmuir isotherm. This part of the model is independent from the chip intensity data and simply yields equilibrium constants for all possible target-probe complexes. In the second part of the model the binding data is used to predict probe intensity for Affymetrix arrays. In this step, we fit several parameters (probe synthesis efficiency, wash stringency, fragmentation, scanner's dynamic range, and target DNA concentration) to each individual chip and predict the probe intensities for that chip. For the analyzed data the average correlations ranged from 0.88 and 0.55. Our results show that the different bases (A, C, G, and T) do not incorporate into the probe with the same efficiency.

Methods

Our model begins with the assumption that DNA hybridization on a microarray is the same as DNA hybridization in solution, but that many of the experimental details previously ignored as well as other details of the microarray experiments must be explicitly modeled to account for the observed differences between solution and microarray.   In particular, we model target fragmentation, cross-hybridization, microarray synthesis imperfections, the effect of the wash, and the scanner's dynamic range. In the final model we must fit at least ten parameters specific to each microarray (eight parameters related to synthesis efficiencies, one parameter for the mean fragmentation size of the target, and one concentration parameter per target molecule). The model also includes four parameters that vary between batches of microarrays processed at the same time (one parameter for the wash common between chips, and three parameters related to the shape of the scanner's dynamic range).

K_{eq} Calculations

Our approach to target DNA/probe DNA hybridization is exhaustive. We begin by assuming that hybridization temperature, hybridization solution salt concentration, probe sequences, and target DNA sequences are known. In our model the target DNA consists of one or more DNA “segments” which have unknown concentrations. These DNA “segments” are user defined and can be PCR fragments, reduced representations of the genome, chromosomes, transcripts, whole genomes, or any other set of DNA sequences that accurately represents the segments produced by the experimental protocol. First we fragment the target at every possible position; thus, modeling all potential cut sites on the target DNA. We then allow each of the resulting target fragments to bind to every position on every probe. Thus in our model, a probe spot consists of a “forest” of bound target/probe complexes. The target molecules that are hybridized to the probes in a given probe spot are of differing lengths and are bound at differing start and end-positions of the probe sequence. Even though each spot consists of a complex assortment of probe/target complexes the underlying thermodynamics of each individual binding reaction is fundamentally the same as in solution and follows a Langmuir isotherm with nearest-neighbor kinetics [19, 37, 38].

In order to model target binding along all positions of a probe, we split the target into all possible sequences that are ^{3}2 base pairs (bp) (Figure 1). For each of these sequences we calculate the K_{eq} (equilibrium constant) values for the forward and reverse target sequences aligned to every position (Figure 2) of the probe. To calculate each K_{eq} value we first calculate the change in free energy, DG, for each of these probe-DNA duplexes. Where DG is

$$\Delta G = \Delta H - T\Delta S \tag{1}$$

DH is change in enthalpy, DS is change in entropy, and T is hybridization temperature. We calculate DG using the nearest neighbor model [37, 38]. The values for DH, and DS for perfect match and mismatch base pairs, initiation/termination GC and AT values, and [Na+] corrections for DG calculations come from [19, 21-25]. All the nearest neighbor values used to calculate changes in free energy are listed in the supplementary materials (Supplementary Table 1A). We then calculate using

$$K_{eq} = e^{\frac{-\Delta G}{RT}} \tag{2}$$

where R is the gas constant.

Fraction Bound Probes Calculations

After we calculate target DNA/probe DNA binding thermodynamics, for all possible target/probe combinations we use those K_{eq} values to calculate the fraction of bound probes, a, for each probe spot. To do so we assume the Langmuir isotherm

$$CK_{eq} = \frac{\alpha}{1-\alpha} \tag{3}$$

where C is equal to the target DNA concentration. Rearranging, we solve for a and get

$$\alpha = \frac{CK_{eq}}{CK_{eq} + 1} \tag{4}$$

In order to use this formula for a, two conditions have to be met. First, equilibrium for target/probe formation must be reached, and second, [C] >> [probe], such that D[C] due to target/probe binding is negligible.

Equation 4 holds for a single target sequence and a single probe spot. However, the experiment consists of long segments of DNA fragmented at random into a collection of targets of differing lengths with differing start and stop positions. Our first assumption is that the fragmentation process creates a uniform pattern of fragmentation, such that the probability that the target is "cut" between any two bases is equal. Thus, start and stop positions of every target DNA fragment are uniformly distributed, and the lengths of the fragments in between the cut sites are geometrically distributed. We use the mean fragment size, m, estimated from the data, to calculate p, the probability that the target is fragmented between any given pair of bases, and q = 1-p. Let $C_{i,j}$ be the concentration of target with a cut after the ith base, and another cut j bases later. The concentration of such fragment ($C_{i,j}$) is

$$C_{i,j} = C * p^2 q^{j-1} \tag{5}$$

where C is the overall concentration of the target molecule. For target molecules with a cut after base i, and extending at least j bases without a cut, but which continues an unknown number of bases past the end of the probe, we estimate the concentration as

$$C_i = \sum_{k=j}^{\infty} Cp^2 q^k = Cpq^j \tag{6}$$

Each of the $C_{i,j}$ fragments can bind at any position in a probe. Let $K_{eq\ (i,j,k)}$ equal to the K_{eq} when fragment $C_{i,j}$ binds starting at position k of the probe. For any one fragment (i,j) bound at a position k, the fraction of bound probes $a_{(i,j,k)}$ is

$$\alpha_{i,j,k} = \frac{C_{i,j,k} K_{eq(i,j,k)}}{C_{i,j,k} K_{eq(i,j,k)} + 1} \tag{7}$$

Equation 7 fails to model competitive hybridization. When we incorporate competition between targets we must sum all possible fragment and target duplexes [39] to get

$$\alpha = \frac{\sum_k \sum_i \sum_j C_{i,j,k} K_{eq(i,j,k)}}{\left(\sum_k \sum_i \sum_j C_{i,j,k} K_{eq(i,j,k)} \right) + 1} \tag{8}$$

When the target solution consists of n distinct DNA segments with different concentrations the fraction of bound probes for a given probe spot becomes

$$\alpha = \frac{\sum_n \sum_k \sum_i \sum_j C_{i,j,k,n} K_{eq(i,j,k,n)}}{\left(\sum_n \sum_k \sum_i \sum_j C_{i,j,k,n} K_{eq(i,j,k,n)} \right) + 1} \tag{9}$$

where C_n is equal to the concentration of the n^{th} DNA segment.

In our model competitive hybridization only takes place during the hybridization period. Following this hybridization period, the chip is washed with a low salt solution. We model the wash as a K_{eq} threshold. We assume that any probe-target complexes with K_{eq} values below this threshold come apart and no new target/probe complexes form during the wash period. We modify Formula 9 above to include the effect of the wash by excluding K_{eq} values < K_{eqW} threshold values in the numerator term. To do so we modify a

$$\alpha = \frac{\sum_n \sum_k \sum_i \sum_j C_{i,j,k,n} K_{eqW(i,j,k,n)}}{\left(\sum_n \sum_k \sum_i \sum_j C_{i,j,k,n} K_{eq(i,j,k,n)} \right) + 1} \tag{10}$$

where $K_{eqW(i,j,k)}$ are all K_{eq} values greater than K_{eqW}.

Equation 10 above describes probe/target binding when there are no errors during probe synthesis and consequently all oligos in a probe spot are identical in sequence and in length. However, errors in probe manufacturing are common [29-31] and will alter probe binding efficiencies. In our model we employ two distinct mechanisms of error: "truncation errors" and "abasic" sites (Table 1). We define truncation errors as the failure of a nucleotide to incorporate during the protection/deprotection stage of probe synthesis; consequently these errors cause the probe to be truncated at a given spot, with no further incorporation of bases. A site is said to be abasic, if the DNA backbone is present, but the nucleotide is not. Sites are assumed to become abasic some time after probe synthesis, but prior to hybridization. Due to both of these types of errors, probe spots consist of a heterogeneous mixture of probes. These spots have full-length probes (no errors) as well as probes with one or more errors (Table 1). Since probe synthesis starts at the 3' end, a probe with a truncation error will extend from the 3' end up to the last base that was successfully incorporated. Probes with abasic sites, on the other hand, vary from the full length probe only at the site where the nucleotide was lost. We further assume that the probability of an error is independent and identically distributed and fixed across a microarray. Let A_S, C_s, G_s, and T_s be the probabilities that A, C, G and T nucleotides are synthesized correctly (contain no truncation error) and A_B, C_B, G_B, and T_B be the probabilities that A, C, G, and T nucleotides do not subsequently lose their base. The probability that a probe is full length p_F (no errors) is

$$p_F = A_S^{nA} C_S^{nC} G_S^{nG} T_S^{nT} A_B^{nA} C_B^{nC} G_B^{nG} T_B^{nT} \tag{11}$$

where nA, nC, nG, and nT are the number of A, C, G, and T bases in the full length probe. Similarly the probability that a probe has exactly one abasic site at an A nucleotide is equal to

$$\frac{p_F(1 - A_B)}{A_B} \tag{12}$$

We calculate all other one error probabilities in a similar manner.

When we account for synthesis failure, we imagine the spot as composed of a forest of full length probes together with all possible manufacturing errors for that probe. We calculate a_x for each possible error, x, individually, and create an overall a by weighting each by the probability, p_x, that this particular error will occur. Thus,

$$\alpha = p_F\alpha_F + \sum_{x=1}^{l} p_x\alpha_x \tag{13}$$

where the sums are taken over all possible errors x. To simplify calculations, we assume that probes with two or more errors have little to no binding and an a_x value of zero.

Estimating Probe Spot Intensity

Equation 13 gives us the proportion of probes bound. By assumption, probes with more target bound will have greater florescent intensity when the probe is scanned. In general, the aim is to have a nearly linear relationship between the amount of probe bound and the observed florescent intensity. However, it is absolutely certain that at the limits of the scanner's dynamic range, a linear response is physically impossible [34]. In order to model this and other factors such as quenching and the dynamic range of the a/d converter, we assume that the relationship between observed florescent intensity and the proportion of probes bound follows a gompertz curve and the expected intensity is equal to

$$MAX * e^{(\log(\frac{MIN}{MAX})*(e^{-\alpha*GOMP}))} \tag{14}$$

where MIN is equal to the background intensity, MAX is equal to the linear cutoff for intensity, and GOMP is equal to the shape parameter of the gompertz curve (Figure 3). The user supplies the MIN, MAX, and GOMP values.

We fit four parameters for the truncation rates, four parameters for the rate of abasic sites, and a variable number of parameters for the target DNA segment concentrations. To do so we minimize the square difference between the observed and expected probe intensities across all probe spots on a chip, using Powell's method for numerical minimization [40]. In order to ensure that the algorithm did not find a sub-optimal solution, we ran several searches on the same chip. Each time the algorithm would start at a different part of parameter space and always find the same solution.

One Dimensional Nearest Neighbor and Initiation Termination Parameter Search

We performed a one-dimensional search for the NN and the initiation/termination parameters used in the K_{eq} calculations. In this search we maximized the observed mean correlation for a set of five chips each ran under four different combinations of wash/fragmentation parameter values.

The Effect of Mismatches on Probe Intensities

To understand the effect of mismatches under our model we compared the intensity of "perfect match" probes to all possible "mismatched probes." More specifically we calculated the expected intensity of 20,102 probes that align perfectly to the human reference sequence. Then for each of these 25 bp long probes we calculated the expected intensity for all sequences that have all possible one base pair differences. Therefore, for each "perfect match" probe there are 75 "mismatch" probes. The "perfect match" probe sequences come from the FMR1 chip design and the parameters used to calculate the expected intensities are the same as the parameters that were fit for chip number 19 FMR1 design (Supplementary Table 2).

Comparison of Forward and Reverse Strand Probe Intensities

We compared the observed and expected intensities for probes that perfectly matched the reference sequence. For every forward and reverse probe pair we calculated the log ratio of the forward/reverse observed intensities. We then calculated the average of this ratio across all 62 chips from the FMR1 design. We did the same for the expected intensity. Then for each probe pair we estimated the difference in nucleotide composition between forward and reverse as the sum of (A-T) and (G-C). Where A, C, G, and T are the number of A, C, G, and T bases in the forward strand probe. We used this value to create bins for the means of the ratios.

We then plotted the mean for the observed and expected ratios for each bin (Supplementary Figure 1).

Results

We used our model to predict probe intensities for seven Affymetrix re-sequencing array designs with 25 base pair long DNA probes that bind end labeled target DNA. The cwrs labeled designs come from [20]. These chips were manufactured and processed in 1999 and 2000. The seventh design, FMR1, chips come from [12]. They were manufactured and processed in 2009. The array designs have highly variable GC content, number of PCR products, and features. Mean correlations between expected and observed intensities, as well as the incorporation rates (1 - truncation rate), and the base retention rates (1 – abasic rate) for individual chips are listed in the supplementary material (Supplementary Table 2). For the 302 chips analyzed the mean correlations (Pearson) range from 0.881 to 0.550. For the FMR1 chip design the average correlation across all chips was 0.76. For these same chips the correlation for the log values was 0.73. Figure 4 shows the plot of the log (observed – observed mean) and log (expected – expected mean) values for chips 32 and 34 from the FMR1 design. The calculation times on a 2.4Ghz single core CPU for K_{eq} values for each design were in the range of 30 minutes to 2 hours per PCR product. These calculations happen only once per probes set design, and do not depend in any way on the observed intensities. The running time for each individual chip was in the range of 20 minutes to 3 hours. For all chips the hybridization temperature was set to 42°C, the minimum intensity for the scanner was set to 100, and the maximum intensity was set to 65536. For all but one chip, the GOMP shape value was fitted to 7. The other chip appeared to have a GOMP value of 6.5. We ran a couple of chips using different temperature and salt concentration values and found that the temperature given by the experimental protocol plus/minus 2 degrees and a salt concentration of one molar gave the best fits. We ran each chip using 30-48 different combinations of wash and fragmentation values, and then selected the values that gave the highest correlation. There is little chance for over fitting, because the number of parameters fit (16-23) for each chip is literally 3-4 orders of magnitude lower than the number of observations for each chip (80,428 - 231,776) (Table 2). Furthermore, the maximum intensity, and minimum intensity were never fit to the data, but were rather inferred from the experimental protocol. By design, the FMR1 chip had 512 probe spots replicated in two or more places on the chip. This allowed us to estimate the correlation in observed intensity between two spots with the same probe sequence, but different positions on a single chip. The average correlation in intensity between these replicated spots is 0.906, where the average is taken over 62 different chips.

For each chip design, the average correlation, the mean incorporation rates, and the mean base retention rates are listed in Table 1, Table 2, and Table 3 respectively. We used a Kruskal-Wallis rank sum test to determine whether the incorporations rates for each nucleotide are different between chip designs. The Kruskal-Wallis chi-squared values are 175, 72, 223, and 156 for A, C, G, and T incorporations rates respectively. Each test has six degrees of freedom and a p value $< 10^{-12}$. For A, C, G, and T base retention rates, the Kruskal-Wallis chi-squared values are 235, 154, 259, and 226 respectively. Each test has six degrees of freedom and a p value $< 10^{-12}$. We then compared the parameters for the cwrs design chips only. For A, C, G, and T incorporation rates as well as A, C, G base retention rates each test had five degrees of freedom and a p value $< 10^{-12}$. For T base retention rates the Kruskal-Wallis chi-square value was 16 and the p value was 0.006.

Intuitively, this is the primary explanation of the observed difference between forward and reverse probe intensities. If, for example, the C incorporation rates are much higher than G incorporation rates and they both have similar rates of base retention, and one probe (say the forward probe) is C rich, while the other G rich, then we will observe that the forward C rich probe will be much brighter than the complimentary reverse G rich probe. Across chip designs adenosine seems to be synthesized more efficiently than other bases. More specifically, for the FMR1 chip design adenosine synthesis is more efficient than thymidine synthesis and cytidine synthesis is more efficient than guanosine synthesis. For this chip design we looked at the ratio of forward and reverse strand intensities as a function of probe base pair composition (Supplementary Figure 1). For bins with more than 50 observations the observed and the expected intensity ratios followed the same trend. The deviations between the observed and the expected at the edges of the curve could be the results of small sample size and nucleotide runs that have binding affinities that deviated from those calculated using the nearest neighbor model.

To test the hypothesis that nearest-neighbor binding NN values have the same values as in solution binding [19, 21-25], we ask whether or not the overall correlation between predicted and observed intensity can be improved by varying any of the binding kinetic parameters. To do so, we conducted a one-dimensional search for DH for all ten possible perfect match NN values, and all 51 mismatch values that have been previously estimated [21-25]. For the perfect match NN values the DH values that maximized the intensity correlation were essentially equal to the values estimated by Santa-Lucia *et al* [19], and in no case was the best estimate more than 5% different from the reference values (Figure 5). On the other hand some of the DH values for the mismatches were modestly different from the reference values. The range of error for the mismatch values is much larger than that for the perfect matches; therefore, our estimates, which are based on thousands of observations, might represent a more accurate estimate of mismatch NN values for microarray and/or in solution binding. However, it is quite possible that these values are fundamentally confounded with manufacturing error in our model, and given the modest nature of the difference, we are not convinced that mismatch binding kinetics are substantially different from in solution binding. The linear search for optimal DH initiation/termination values also yielded results that were fundamentally the same as in solution kinetics. The best fit values for the mismatch

NN numbers can be found in the Supplementary Materials (Supplementary Table 1B).

One of the main goals of our study was to understand the effect of mismatches on probe intensity. In our model the binding affinity of a particular mismatch is independent of its position in the probe sequence. Therefore, we were interested to see if the other details in our model could explain the observation that mismatches towards the center of the probe have a larger effect on intensity compared to mismatches that are closer to the edges. Figure 6 shows the average effect of mismatches at each position on the probe. The distance of a mismatch from the edge of the probe correlates to the difference between the perfect match and mismatch intensities, with mismatches in the center having the largest effect. Mismatches towards the 3' end have a slightly larger effect than mismatches towards the 5' end.

Discussion

Our model for DNA hybridization on microarrays is comprehensive. It includes parameters that are specific to the chip design and to the processing protocol. These parameters are probe length, temperature, salt concentration, wash stringency, target DNA size, and the parameters that are related to the scanners dynamic range. Furthermore, the NN values can be adjusted to accommodate DNA-RNA and RNA-RNA binding. The algorithm that calculates K_{eq} values for probe target complexes is applicable to many array designs as well as to other binding reactions in thermodynamic equilibrium. Thus, this algorithm can provide valuable information on binding affinities and cross-hybridization for a wide array of applications including probe design.

Our approach to probe intensity calculations takes into account many of the "problems" inherent to microarrays such as batch effects and probe synthesis failure. Preparation of target DNA for microarray experiments often poses many technical challenges; consequently, the concentration of DNA products often varies between experiments as does the wash strength and average fragmentation size. Furthermore, probe synthesis efficiency varies between individual chips, and different batches of arrays. Our model directly tackles this problem by fitting the concentration as well as the nucleotide synthesis parameters for each chip. It is important to note that all the chips analyzed are Affymetrix chips with 25 bp long probes; therefore, the chip synthesis aspects of our model may be somewhat specific to the company's manufacturing process. Our model fits one parameter for the concentration of each target molecule, which limits its use for arrays that quantify transcript levels. However, the K_{eq} information from our model is applicable to any type of array.

For all the chips that were analyzed, the mean correlation between expected and observed probe intensities is 0.701, with average correlations for each chip ranging from 0.881 to 0.550. Furthermore, extreme intensity values do not dominate the correlation terms (Figure 4). For the FMR1 design chips the correlation for replicated probe spots (probe spots with the exact same probe sequence) is 0.906. Similar measurements for correlation between probe spots have been previously reported [41, 42]. Thus, if we view ~90% as the maximum possible correlation between expected and observed intensity, because observed intensity varies this much between replicated spots within a single chip, our model can be seen to be doing a very good, but not quite a perfect job of predicting probe intensities.

Our model has helped us understand some puzzling observations regarding microarrays: the difference in intensity between complimentary forward and reverse probes; the larger decrease in intensity for mismatches towards the center of a probe; and batch effects.

First, assuming simple liquid phase kinetics, probe spots that target the forward DNA strand should have equal amount of bound target DNA as the complementary spot that targets the reverse DNA strand. However, the intensity of the forward and reverse strands usually have systematically different intensities, with probes targeting the forward strand of a given genomic region being brighter/darker than the set of probes that target the reverse strand of the same genomic region [20]. Under our model this observation is simply the result of probe synthesis failure or sites becoming abasic after synthesis. This claim is supported by the fact that the A,C,G, and T synthesis parameters are statistically different from each other for each of the different chip designs (Supplementary Table 2). The estimates for A,C,G, and T incorporation rates (Table 3) are similar to previously published estimates of synthesis failure for Affymetrix arrays [29]. Furthermore, this explanation for the difference between forward and reverse strand probes is far more parsimonious than relating this difference to G-stacks [43], given that this difference is observed in A/T rich probes that have no stretches of the same amino acid. There are two factors that we did not model, but could partially contribute to the observed difference between forward and reverse probes. For one there are base analogues that are often used to manufacture probes, this difference can easily be incorporated into our model by revising the NN numbers used for the Keq calculations. Second there could be an entropy penalty to base pairs that are closer to the array surface.

In our model, probe synthesis failure along with wash and fragmentation differences explain batch effects for microarrays. Batch effects can be subdivided into two types, those that happen during manufacturing of the chip and those that happen during the processing of the chip. In our model, the former is explained by differences in the efficiency of probe synthesis, while the latter is explained by differences in the fragmentation, washing steps, and as the rate of abasic site formation.

Another previously puzzling observation is the correlation between the distance of a mismatch from the edge of the probe and its effect on probe intensity. Mismatches towards the center have a larger effect than mismatches towards the edges [26]. Under our model this observation is expected (Figure 6), and is the result of fragmentation of the target molecule. Intuitively, there are more fragments that can bind the center of a probe, than fragments that can only bind a single edge (Figure 1). Hence, if a target molecule contains a mismatch, its effect will be proportional to its distance from the middle of the probe. In our model, simple hybridization kinetics can explain these puzzling observations without the need to assign different weights along the probe sequence nor a penalty to probes with a mismatch.

Using our model we get an average correlation of ~70% between observed and expected probe intensity. This includes data from all probe spots regardless of their quality. Even so, our model comes close (0.881) on some chips, but never achieves our theoretical maximum correlation (0.906). From the manufacturing process up until the reading of probe spots intensities, the microarray experiment has several complex steps. Our model makes several “simplistic” assumptions that allowed us to develop efficient algorithms. In doing so we made several compromises. One of these assumptions is that probes with two or more errors do not bind target DNA. This assumption should have a relatively small effect on the correlation for 25 bp long probes; however, it is expected to have a larger effect on the correlation for longer probes. In our model we use the Langmuir isotherm to calculate the fraction of bound probes and do not take into account probe surface density [16, 17], non-equilibrium, and low target/probe ratio [39]. Theoretically, commercial arrays of 25 bp long probes should have reached equilibrium at the end of the hybridization step; however, equilibrium might not be achieved by the end of the washing period; therefore, our approach to modeling the wash step of the protocol is rather simplistic. Furthermore, in these arrays the target/probe ratio should be very large. When arrays deviate from this ideal scenario our model loses predictive power. It is important to note that even though the probe surface density is not directly modeled by our approach, the parameters we use to describe the scanners dynamic range can indirectly be used to adjust for microarrays with varying probe surface densities. This “adjustment” however has some limitations. The most obvious one being if increasing probe density affects mismatched sequences and matched sequences disproportionately [17].

Overall our data suggests that the Langmuir isotherm appropriately and efficiently models binding between probe and target DNA on a microarray; however, other more computationally intensive measurements for binding on arrays have been proposed [39, 44]. Furthermore, when we calculate target-probe binding we do not account for the known in solution effects of dangling ends [45] and the stabilizing effect of mismatches in the last three base pairs of a sequence [21-25]. We also do not model secondary structures that can form on arrays with long probes or arrays that hybridize to targets with extensive secondary structures, for example rRNA arrays [46].

Other details of the microarray experiment that are left out of our model are bleed-through between features and regional artifacts such as air bubbles, scratches, and miscellaneous particles. There are two sources of bleed-through between features: one, the probes at the edge of a feature may have a hybrid sequence due to incorporation of nucleotides during the synthesis of the neighboring probe; two, the scanner may be detecting light from neighboring features and falsely determining its origin. If this were going on, its effect would be most noticeable in probes that would otherwise be very dark, and appears to be present (Figure 4) in our data, where a substantial fraction of probes that are predicted to have very low intensities appear to have much higher than expected intensity.

Our approach to modeling probe synthesis failure also has some limitations. First, there is the possibility that the concentration parameters are confounded with the synthesis efficiency parameters. This can be a problem when dealing with G/C or A/T rich PCR products. Second, the synthesis efficiency of a nucleotide can potentially be dependent on its position on the probe [29].

For our calculations we assume that the target DNA has the reference sequence. This assumption is never completely valid because each individual almost surely has a unique sequence that may differ slightly or even significantly from the genomic reference sequence. The impact of this assumption on the correlations for the analyzed data depends on the type of genetic variation of the samples. When the target DNA only has SNPs and/or other one base pair changes, then the genetic variation is unlikely to have a large impact on the average correlation over the entire chip; however, if the target DNA has large CNVS and/or several CNVs then these genetic variants would be expected to have a significant effect on the average correlation for the chip.

Our model is most applicable to 25 bp long arrays that are designed to detect genetic variation. With this in mind, the obvious next step is to apply our model to SNP arrays with the goal being to better determined which genetic variants are present, by incorporating our model into a variant calling algorithm.

Author's Contributions

Both authors contributed equally to all aspects of this work and the preparation of this manuscript.

Acknowledgement

We thank the Warren and the Zwick labs for providing the microarray data. We would also like to thank the ELLIPSE Emory High Performance Computer cluster, which we used for this project.

Funding

This work was supported by the National Institutes of Health [R01 HG003461].

Supplementary Data

Supplementary Data are available Online: Supplementary Tables 1-2. Figure 1.

Computer Programs

The computer programs described in this paper are available at genome.emory.edu/faculty/cutler/software.html.

References

1. Stoughton RB: **Applications of DNA microarrays in biology**. *Annu Rev Biochem* 2005, **74**:53-82.

2. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays**. *Nat Genet* 1999, **23**(1):41-46.

3. Emanuel BS, Saitta SC: **From microscopes to microarrays: dissecting recurrent chromosomal rearrangements**. *Nat Rev Genet* 2007, **8**(11):869-883.

4. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E *et al*: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls**. *Nature* 2010, **464**(7289):713-720.

5. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE: **De novo rates and selection of large copy number variation**. *Genome Res* 2010.

6. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP: **Accessing genetic information with high-density DNA arrays**. *Science* 1996, **274**(5287):610-614.

7. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J *et al*: **Large-scale identification, mapping, and**

genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998, **280**(5366):1077-1082.

8. **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.

9. Ohira M, Oba S, Nakamura Y, Isogai E, Kaneko S, Nakagawa A, Hirata T, Kubo H, Goto T, Yamada S *et al*: **Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas**. *Cancer Cell* 2005, **7**(4):337-350.

10. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays**. *Nat Genet* 1999, **21**(1 Suppl):10-14.

11. Davila S, Wright VJ, Khor CC, Sim KS, Binder A, Breunis WB, Inwald D, Nadel S, Betts H, Carrol ED *et al*: **Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease**. *Nat Genet* 2010, **42**(9):772-776.

12. Collins SC, Coffee B, Benke PJ, Berry-Kravis E, Gilbert F, Oostra B, Halley D, Zwick ME, Cutler DJ, Warren ST: **Array-based FMR1 sequencing and deletion analysis in patients with a fragile X syndrome-like phenotype**. *PLoS One* 2010, **5**(3):e9476.

13. Southern EM, Maskos U, Elder JK: **Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models**. *Genomics* 1992, **13**(4):1008-1017.

14. Zhang L, Miles MF, Aldape KD: **A model of molecular interactions on short oligonucleotide microarrays**. *Nat Biotechnol* 2003, **21**(7):818-821.

15. Zhang Y, Hammer DA, Graves DJ: **Competitive hybridization kinetics reveals unexpected behavior patterns**. *Biophys J* 2005, **89**(5):2950-2959.

16. Peterson AW, Heaton RJ, Georgiadis RM: **The effect of surface probe density on DNA hybridization**. *Nucleic Acids Res* 2001, **29**(24):5163-5168.

17. Watterson JH, Piunno PAE, Wust CC, Krull UJ: **Effects of Oligonucleotide Immobilization Density on Selectivity of Quantitative Transduction of Hybridization of Immobilized DNA**. *Langmuir* 2000, **16**(11):4984-4992.

18. Wu C, Carta R, Zhang L: **Sequence dependence of cross-hybridization on short oligo microarrays**. *Nucleic Acids Res* 2005, **33**(9):e84.

19. SantaLucia J, Jr.: **A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics**. *Proc Natl Acad Sci U S A* 1998, **95**(4):1460-1465.

20. Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA *et al*: **High-throughput variation detection and genotyping using microarrays**. *Genome Res* 2001, **11**(11):1913-1925.

21. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J, Jr.: **Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches**. *Biochemistry* 1999, **38**(12):3468-3477.

22. Allawi HT, SantaLucia J, Jr.: **Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects**. *Biochemistry* 1998, **37**(26):9435-9444.

23. Allawi HT, SantaLucia J, Jr.: **Thermodynamics of internal C.T mismatches in DNA**. *Nucleic Acids Res* 1998, **26**(11):2694-2701.

24. Allawi HT, SantaLucia J, Jr.: **Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA**. *Biochemistry* 1998, **37**(8):2170-2179.

25. Allawi HT, SantaLucia J, Jr.: **Thermodynamics and NMR of internal G.T mismatches in DNA**. *Biochemistry* 1997, **36**(34):10581-10594.

26. Duan F, Pauley MA, Spindel ER, Zhang L, Norgren RB, Jr.: **Large scale analysis of positional effects of single-base mismatches on microarray gene expression data**. *BioData Min* 2010, **3**(1):2.

27. Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, Xu J, Chen JJ, Han T, Kaput J *et al*: **Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples**. *BMC Bioinformatics* 2008, **9 Suppl 9**:S17.

28. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA: **A multilevel model to address batch effects in copy number estimation using SNP arrays**. *Biostatistics* 2010.

29. McGall GH, Barone AD, Diggelmann M, Fodor SPA, Gentalen E, Ngo N: **The Efficiency of Light-Directed Synthesis of DNA Arrays on Glass**

Substrates. *Journal of the American Chemical Society* 1997, **119**(22):5081-5090.

30. Pirrung MC, Fallon L: **Proofing of photolithographic DNA synthesis methods. Fabrication of DNA microchips**. *Abstracts of Papers of the American Chemical Society* 1997, **213**:362-ORGN.

31. Forman Jonathan E, Walton Ian D, Stern D, Rava Richard P, Trulson Mark O: **Thermodynamics of Duplex Formation and Mismatch Discrimination on Photolithographically Synthesized Oligonucleotide Arrays**. In: *Molecular Modeling of Nucleic Acids.* vol. 682: American Chemical Society; 1997: 206-228.

32. Skvortsov D, Abdueva D, Curtis C, Schaub B, Tavare S: **Explaining differences in saturation levels for Affymetrix GeneChip arrays**. *Nucleic Acids Res* 2007, **35**(12):4154-4163.

33. Held GA, Grinstein G, Tu Y: **Relationship between gene expression and observed intensities in DNA microarrays--a modeling study**. *Nucleic Acids Res* 2006, **34**(9):e70.

34. Shi L, Tong W, Su Z, Han T, Han J, Puri RK, Fang H, Frueh FW, Goodsaid FM, Guo L *et al*: **Microarray scanner calibration curves: characteristics and implications**. *BMC Bioinformatics* 2005, **6 Suppl 2**:S11.

35. Binder H, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: theory and algorithm**. *Algorithms Mol Biol* 2008, **3**:12.

36. Binder H, Krohn K, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures**. *Algorithms Mol Biol* 2008, **3**:11.

37. Devoe H, Tinoco I, Jr.: **The stability of helical polynucleotides: base contributions**. *J Mol Biol* 1962, **4**:500-517.

38. Crothers DM, Zimm BH: **Theory of the melting transition of synthetic polynucleotides: Evaluation of the stacking free energy**. *Journal of Molecular Biology* 1964, **9**(1):1-9.

39. Halperin A, Buhot A, Zhulina EB: **On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions**. *Journal of Physics-Condensed Matter* 2006, **18**(18):S463-S490.

40. Press W, Teukolsky S, Vetterling W, Flannery B: **Numerical Recipes 3rd Edition: The Art of Scientific Computing**: Cambridge University Press; 2007.

41. Wang X, Ghosh S, Guo SW: **Quantitative quality control in microarray image processing and data acquisition**. *Nucleic Acids Res* 2001, **29**(15):E75-75.

42. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J *et al*: **Within the fold: assessing differential expression measures and reproducibility in microarray assays**. *Genome Biol* 2002, **3**(11):research0062.

43. Wu C, Zhao H, Baggerly K, Carta R, Zhang L: **Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays**. *Bioinformatics* 2007, **23**(19):2566-2572.

44. Vainrub A, Pettitt BM: **Theoretical aspects of genomic variation screening using DNA microarrays**. *Biopolymers* 2004, **73**(5):614-620.

45. Bommarito S, Peyret N, SantaLucia J, Jr.: **Thermodynamic parameters for DNA sequences with dangling ends.** *Nucleic Acids Res* 2000, **28**(9):1929-1934.

46. Mueckstein U, Leparc GG, Posekany A, Hofacker I, Kreil DP: **Hybridization thermodynamics of NimbleGen microarrays**. *BMC Bioinformatics* 2010, **11**:35.

TABLES

Table 1: Synthesis Errors

Probe Sequence	Type of Synthesis Error(s)
5’ $C^{12}T^{11}A^{10}C^{9}C^{8}G^{7}T^{6}A^{5}C^{4}C^{3}G^{2}T^{1}$ 3’	Full length probe (no error)
5’ $C^{8}G^{7}T^{6}A^{5}C^{4}C^{3}G^{2}T^{1}$ 3’	Incorporation error base 9
5’ $C^{8}G^{7}T^{6}\_^{5}C^{4}C^{3}G^{2}T^{1}$ 3’	Incorporation error base 9 and abasic site
5’ $C^{12}T^{11}A^{10}\_^{9}C^{8}G^{7}T^{6}\_^{5}C^{4}C^{3}G^{2}T^{1}$ 3’	Full length probe with two abasic sites
5’ $T^{6}\_^{5}C^{4}\_^{3}G^{2}T^{1}$ 3’	Incorporation error base 7 and two abasic sites
Table 1: (_) denotes an abasic site.	

Table 2: Correlations

Design	Number of Chips	Number of Probe Spots per Chip	Number of Parameters that were Fit to each Chip	Number of PCR products	Mean Correlation between Expected and Observed Probe Intensities
cwrs-07	40	230640	21	9	0.765 ± 0.038
cwrs-39	40	230240	20	8	0.605 ± 0.020
cwrs-51	40	226592	23	11	0.730 ± 0.021
cwrs-53	40	230432	21	9	0.714 ± 0.016
cwrs-63	40	229280	21	9	0.666 ± 0.028
cwrs-67	40	231776	18	6	0.629 ± 0.029
fmr1	62	80408	16	4	0.762 ± 0.081

Table 2 : Summary for each chip design.

Table 3: Incorporation Rates

Design	A incorporation	C incorporation	G incorporation	T incorporation
cwrs-07	0.985 ± 0.012	0.952 ± 0.017	0.916 ± 0.035	0.945 ± 0.018
cwrs-39	1.000 ± 0.000	0.950 ± 0.009	0.990 ± 0.017	0.944 ± 0.005
cwrs-51	0.999 ± 0.004	0.945 ± 0.015	0.995 ± 0.013	0.967 ± 0.008
cwrs-53	0.996 ± 0.007	0.943 ± 0.013	0.990 ± 0.015	0.973 ± 0.006
cwrs-63	0.983 ± 0.018	0.944 ± 0.010	0.943 ± 0.022	0.952 ± 0.009
cwrs-67	0.987 ± 0.012	0.931 ± 0.009	0.923 ± 0.016	0.951 ± 0.007
fmr1	0.967 ± 0.010	0.944 ± 0.015	0.908 ± 0.020	0.950 ± 0.006

Table 3: Mean incorporation rate for each nucleotide.

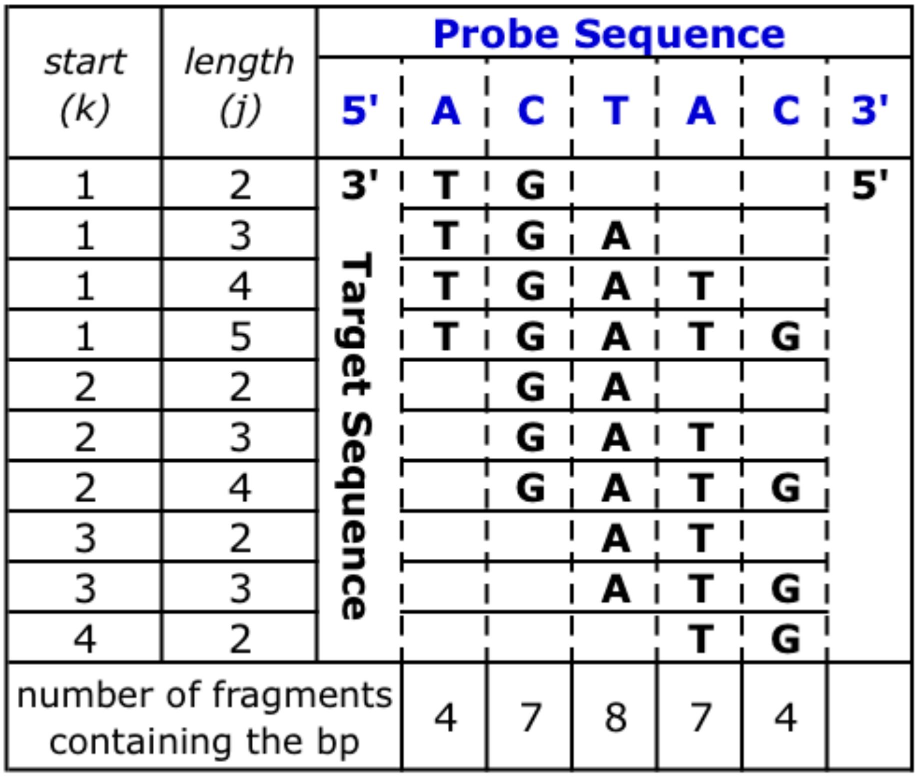
Table 4: Retention Rates

Design	A retention	C retention	G retention	T retention
cwrs-07	0.963 ± 0.028	0.981 ± 0.013	0.865 ± 0.036	0.998 ± 0.007
cwrs-39	0.882 ± 0.005	0.930 ± 0.011	0.791 ± 0.017	1.000 ± 0.001
cwrs-51	0.904 ± 0.010	0.961 ± 0.015	0.791 ± 0.014	1.000 ± 0.000
cwrs-53	0.901 ± 0.008	0.972 ± 0.015	0.809 ± 0.015	1.000 ± 0.000
cwrs-63	0.944 ± 0.025	0.959 ± 0.013	0.866 ± 0.029	0.998 ± 0.003
cwrs-67	0.938 ± 0.013	0.956 ± 0.011	0.884 ± 0.021	0.998 ± 0.006
fmr1	0.959 ± 0.014	0.941 ± 0.027	0.953 ± 0.029	0.962 ± 0.010

Table 4: Mean retention rate for each nucleotide. (Rate abasic site = 1 – retention

Figures

Figure 1



start (k)	length (j)	Probe Sequence						
		5'	A	C	T	A	C	3'
1	2	3'	T	G				5'
1	3	Target Sequence	T	G	A			
1	4		T	G	A	T		
1	5		T	G	A	T	G	
2	2			G	A			
2	3			G	A	T		
2	4			G	A	T	G	
3	2				A	T		
3	3				A	T	G	
4	2					T	G	
number of fragments containing the bp			4	7	8	7	4	

Figure 1 Legend: Probe and Target Sequences. Sample probe sequence in blue. All unique target sequences that are 2 base pairs or longer and that are perfect reverse complements of the probe in black. Columns 1 and 2 have the corresponding k and j values for each target sequence. The bottom row counts the number of times each probe position binds to a different target fragment.

Figure 2

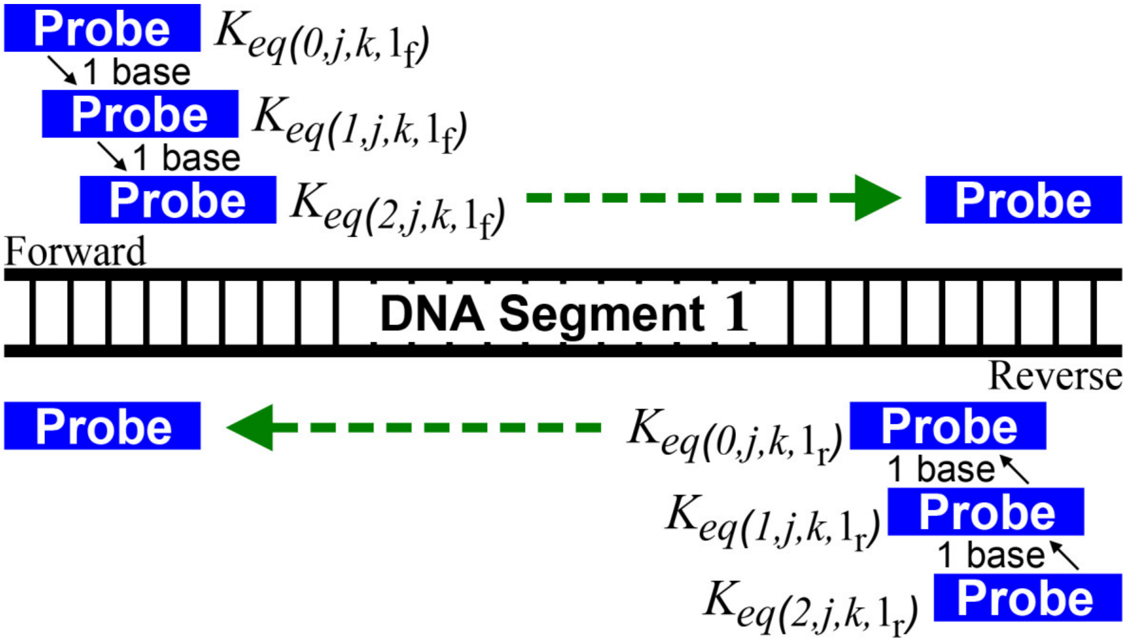


Figure 2 Legend: Probe-Target Binding. Ways in which a probe sequence is aligned to the target DNA. Only one target DNA segment is shown. The probe is aligned to both the forward and reverse target sequences along all positions (i) of the target DNA and for all appropriate j and k values.

Figure 3

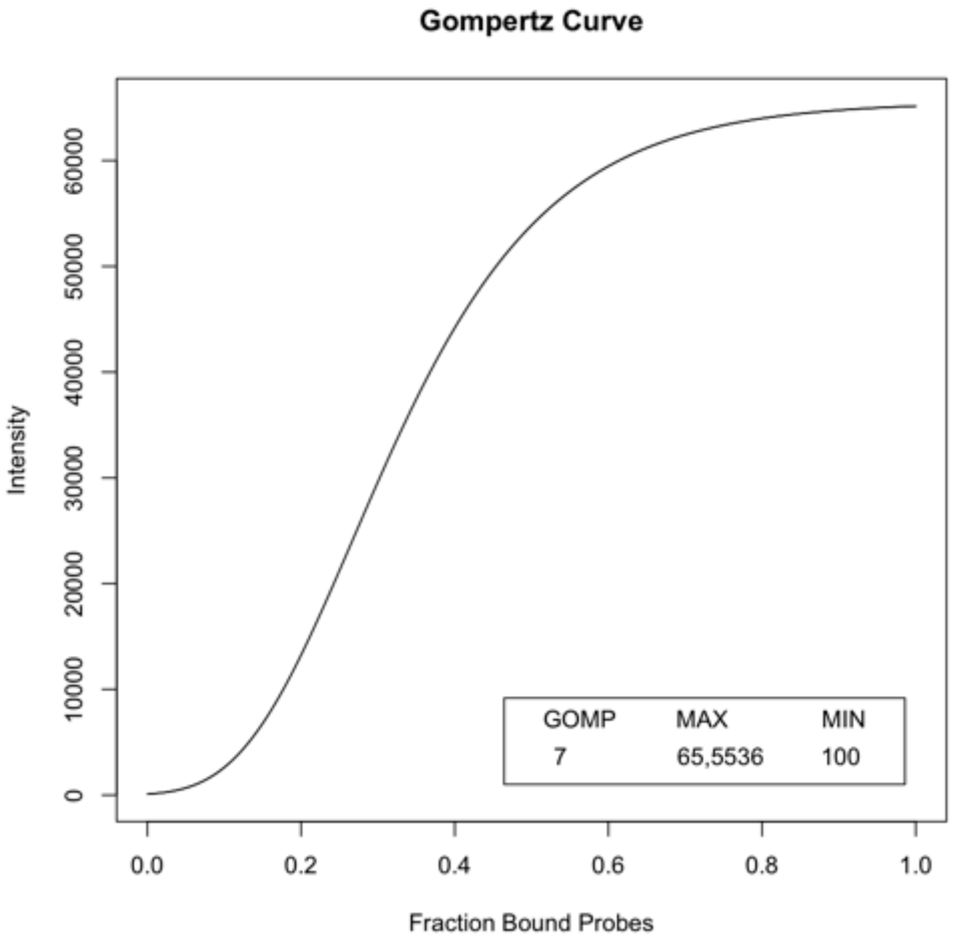
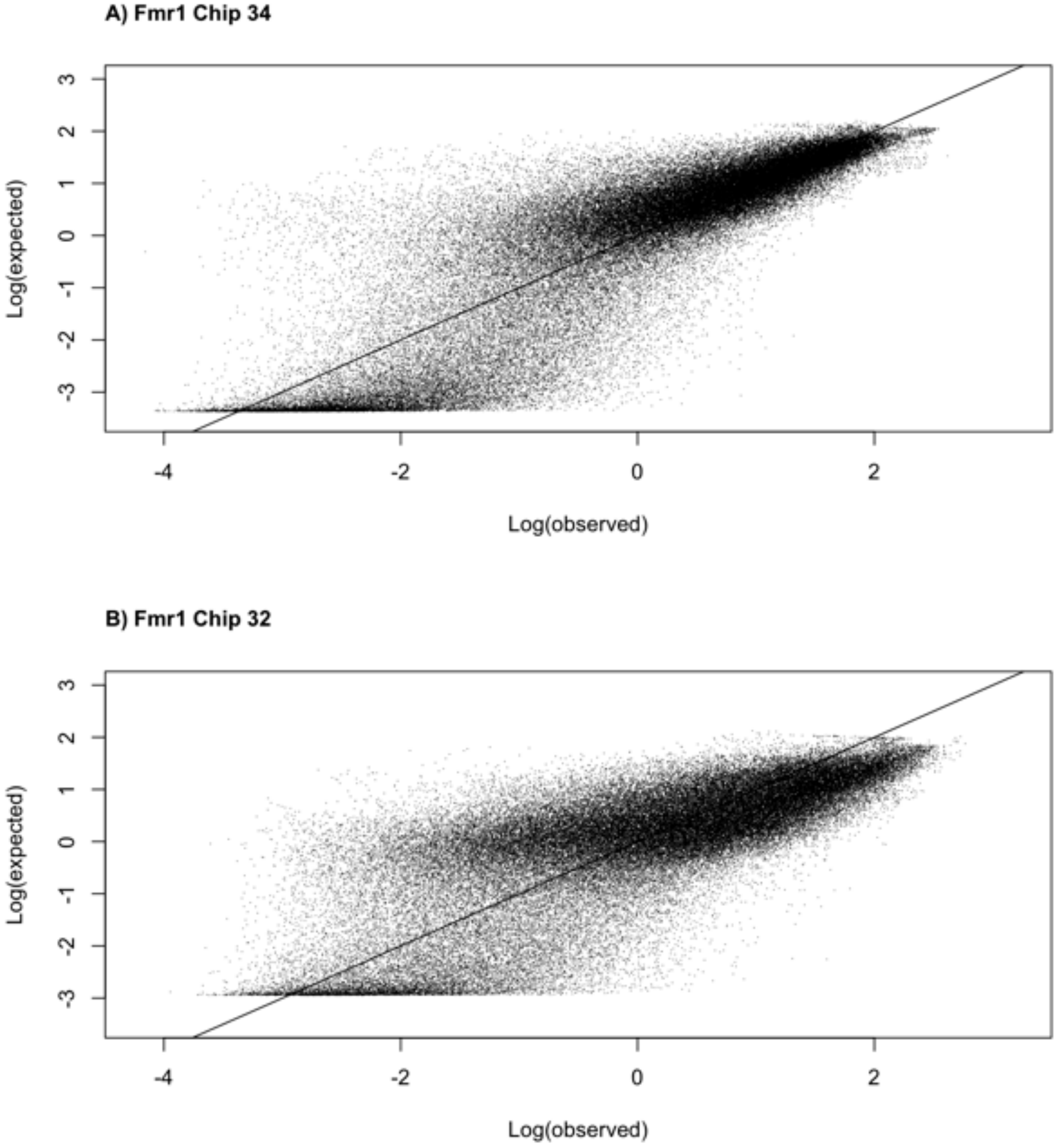


Figure 3 Legend: Gompertz Curve. Sample Gompertz Curve.

Figure 4



A) Fmr1 Chip 34

B) Fmr1 Chip 32

Figure 4 Legend: Observed and Expected Intensity Plots. Plots for the log observed and expected intensity values for two FMR1 chips. The intensity values are centered around their mean. A) Chip number 34 with an observed mean of 5,993

and an expected mean of 5,914. B) Chip number 32 with an observed mean of 3,054 and an expected mean of 3,058.

Figure 5

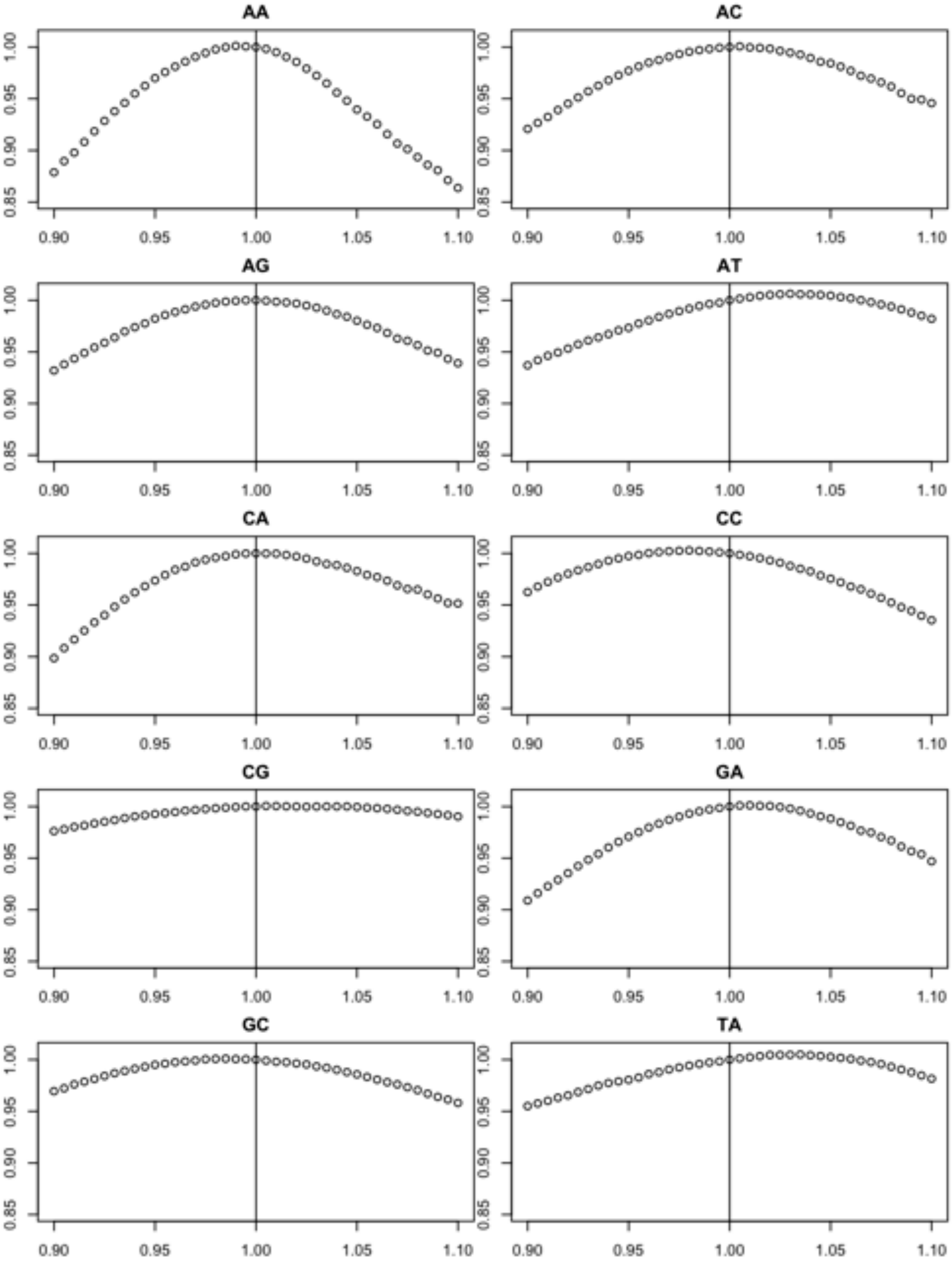


Figure 5 Legend: Nearest-Neighbor Parameter Search. Results for one-dimensional search for perfect match NN values. The x-axis is the ratio of assayed DH value divided by the DH in solution value [19]. Thus, x=1 are the in solution values. The y-axis has the corresponding mean correlation divided by the mean correlation when the DH values are set to the in solution value [19].

Figure 6

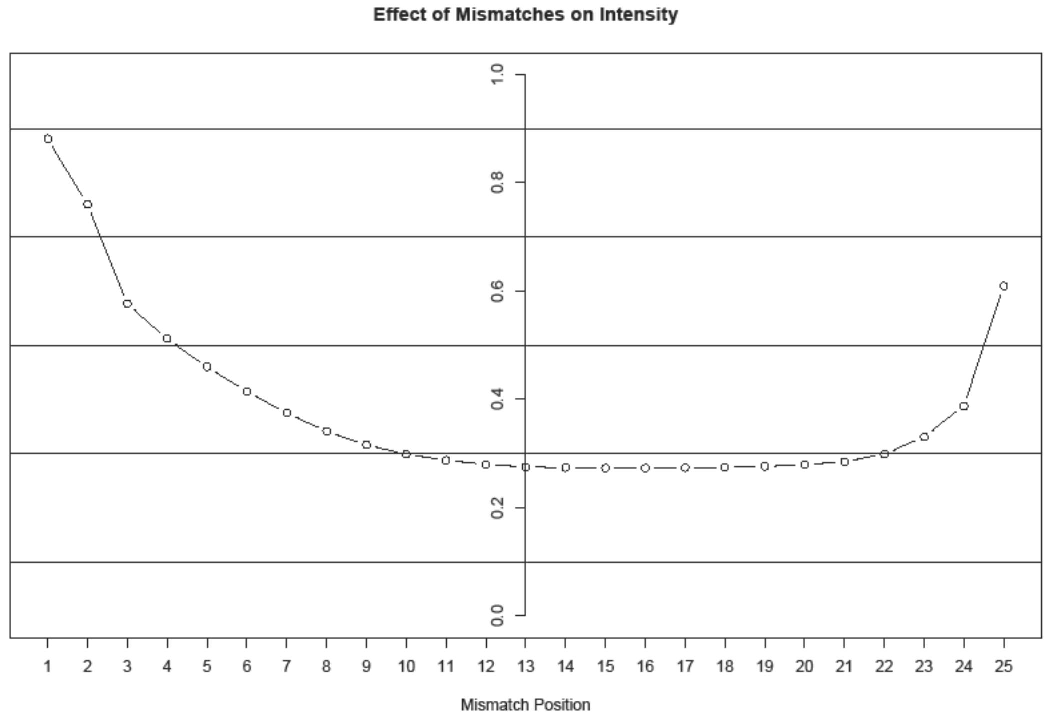
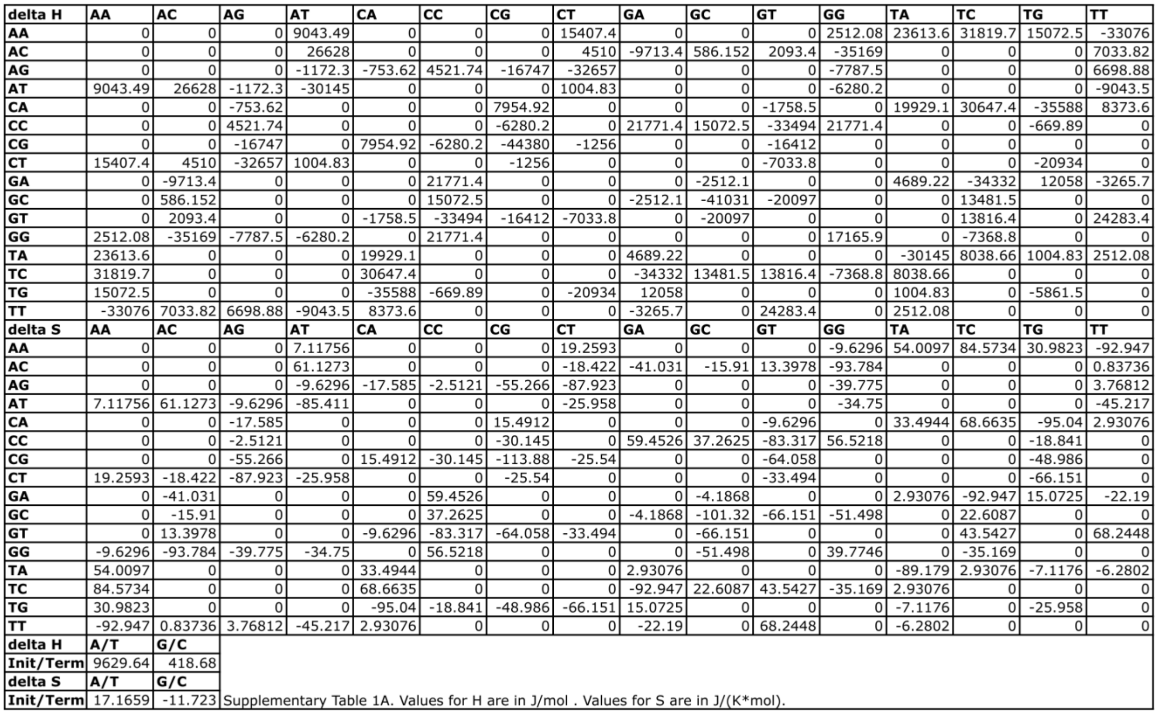


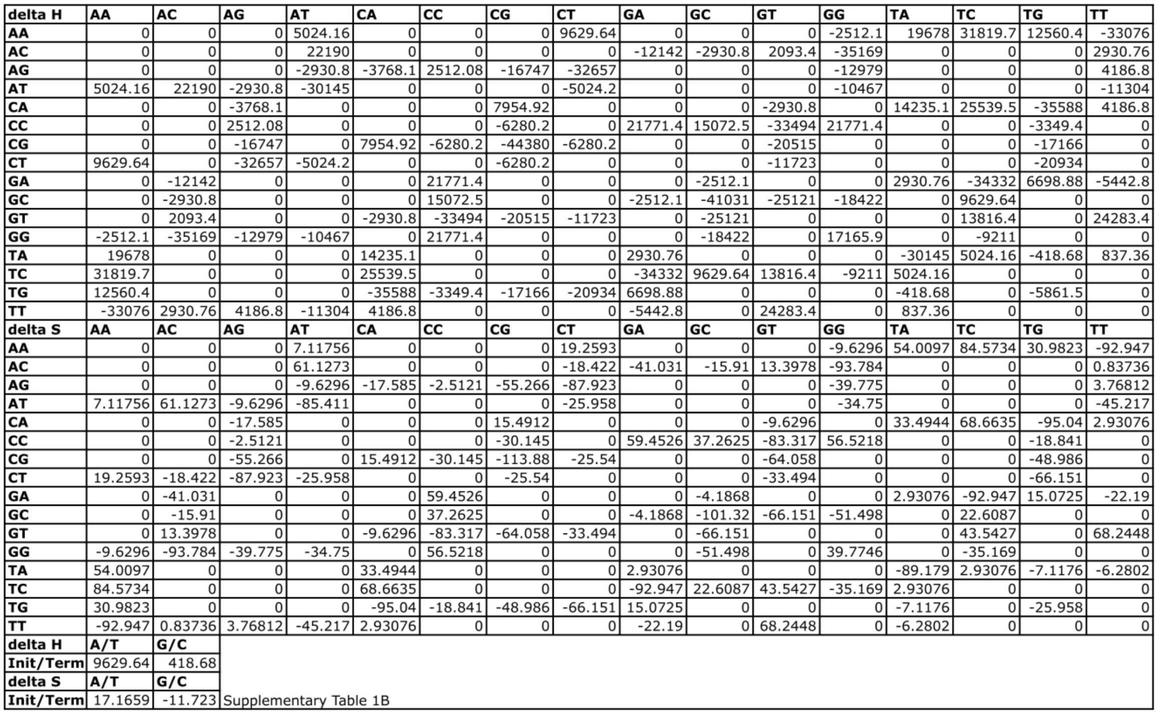
Figure 6 Legend: Predicted Effect of Mismatches on Intensity. A graph of the expected effect of each mismatch on intensity. On the x-axis is the position of the mismatch. On the y-axis is the mismatch probe intensity divided by the intensity of the corresponding probe with no mismatches, averaged across all probes.

Supplementary Information

Supplementary Table 1



delta H	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GT	GG	TA	TC	TG	TT
AA	0	0	0	9043.49	0	0	0	15407.4	0	0	0	2512.08	23613.6	31819.7	15072.5	-33076
AC	0	0	0	26628	0	0	0	4510	-9713.4	586.152	2093.4	-35169	0	0	0	7033.82
AG	0	0	0	-1172.3	-753.62	4521.74	-16747	-32657	0	0	0	-7787.5	0	0	0	6698.88
AT	9043.49	26628	-1172.3	-30145	0	0	0	1004.83	0	0	0	-6280.2	0	0	0	-9043.5
CA	0	0	-753.62	0	0	0	7954.92	0	0	0	-1758.5	0	19929.1	30647.4	-35588	8373.6
CC	0	0	4521.74	0	0	0	-6280.2	0	21771.4	15072.5	-33494	21771.4	0	0	-669.89	0
CG	0	0	-16747	0	7954.92	-6280.2	-44380	-1256	0	0	-16412	0	0	0	0	0
CT	15407.4	4510	-32657	1004.83	0	0	-1256	0	0	0	-7033.8	0	0	0	-20934	0
GA	0	-9713.4	0	0	0	21771.4	0	0	0	-2512.1	0	0	4689.22	-34332	12058	-3265.7
GC	0	586.152	0	0	0	15072.5	0	0	-2512.1	-41031	-20097	0	0	13481.5	0	0
GT	0	2093.4	0	0	-1758.5	-33494	-16412	-7033.8	0	-20097	0	0	0	13816.4	0	24283.4
GG	2512.08	-35169	-7787.5	-6280.2	0	21771.4	0	0	0	0	0	17165.9	0	-7368.8	0	0
TA	23613.6	0	0	0	19929.1	0	0	0	4689.22	0	0	0	-30145	8038.66	1004.83	2512.08
TC	31819.7	0	0	0	30647.4	0	0	0	-34332	13481.5	13816.4	-7368.8	8038.66	0	0	0
TG	15072.5	0	0	0	-35588	-669.89	0	-20934	12058	0	0	0	1004.83	0	-5861.5	0
TT	-33076	7033.82	6698.88	-9043.5	8373.6	0	0	0	-3265.7	0	24283.4	0	2512.08	0	0	0
delta S	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GT	GG	TA	TC	TG	TT
AA	0	0	0	7.11756	0	0	0	19.2593	0	0	0	-9.6296	54.0097	84.5734	30.9823	-92.947
AC	0	0	0	61.1273	0	0	0	-18.422	-41.031	-15.91	13.3978	-93.784	0	0	0	0.83736
AG	0	0	0	-9.6296	-17.585	-2.5121	-55.266	-87.923	0	0	0	-39.775	0	0	0	3.76812
AT	7.11756	61.1273	-9.6296	-85.411	0	0	0	-25.958	0	0	0	-34.75	0	0	0	-45.217
CA	0	0	-17.585	0	0	0	15.4912	0	0	0	-9.6296	0	33.4944	68.6635	-95.04	2.93076
CC	0	0	-2.5121	0	0	0	-30.145	0	59.4526	37.2625	-83.317	56.5218	0	0	-18.841	0
CG	0	0	-55.266	0	15.4912	-30.145	-113.88	-25.54	0	0	-64.058	0	0	0	-48.986	0
CT	19.2593	-18.422	-87.923	-25.958	0	0	-25.54	0	0	0	-33.494	0	0	0	-66.151	0
GA	0	-41.031	0	0	0	59.4526	0	0	0	-4.1868	0	0	2.93076	-92.947	15.0725	-22.19
GC	0	-15.91	0	0	0	37.2625	0	0	-4.1868	-101.32	-66.151	-51.498	0	22.6087	0	0
GT	0	13.3978	0	0	-9.6296	-83.317	-64.058	-33.494	0	-66.151	0	0	0	43.5427	0	68.2448
GG	-9.6296	-93.784	-39.775	-34.75	0	56.5218	0	0	0	-51.498	0	39.7746	0	-35.169	0	0
TA	54.0097	0	0	0	33.4944	0	0	0	2.93076	0	0	0	-89.179	2.93076	-7.1176	-6.2802
TC	84.5734	0	0	0	68.6635	0	0	0	-92.947	22.6087	43.5427	-35.169	2.93076	0	0	0
TG	30.9823	0	0	0	-95.04	-18.841	-48.986	-66.151	15.0725	0	0	0	-7.1176	0	-25.958	0
TT	-92.947	0.83736	3.76812	-45.217	2.93076	0	0	0	-22.19	0	68.2448	0	-6.2802	0	0	0
delta H	A/T	G/C														
Init/Term	9629.64	418.68														
delta S	A/T	G/C														
Init/Term	17.1659	-11.723	Supplementary Table 1A. Values for H are in J/mol . Values for S are in J/(K*mol).													



delta H	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GT	GG	TA	TC	TG	TT
AA	0	0	0	5024.16	0	0	0	9629.64	0	0	0	-2512.1	19678	31819.7	12560.4	-33076
AC	0	0	0	22190	0	0	0	0	-12142	-2930.8	2093.4	-35169	0	0	0	2930.76
AG	0	0	0	-2930.8	-3768.1	2512.08	-16747	-32657	0	0	0	-12979	0	0	0	4186.8
AT	5024.16	22190	-2930.8	-30145	0	0	0	-5024.2	0	0	0	-10467	0	0	0	-11304
CA	0	0	-3768.1	0	0	0	7954.92	0	0	0	-2930.8	0	14235.1	25539.5	-35588	4186.8
CC	0	0	2512.08	0	0	0	-6280.2	0	21771.4	15072.5	-33494	21771.4	0	0	-3349.4	0
CG	0	0	-16747	0	7954.92	-6280.2	-44380	-6280.2	0	0	-20515	0	0	0	-17166	0
CT	9629.64	0	-32657	-5024.2	0	0	-6280.2	0	0	0	-11723	0	0	0	-20934	0
GA	0	-12142	0	0	0	21771.4	0	0	0	-2512.1	0	0	2930.76	-34332	6698.88	-5442.8
GC	0	-2930.8	0	0	0	15072.5	0	0	-2512.1	-41031	-25121	-18422	0	9629.64	0	0
GT	0	2093.4	0	0	-2930.8	-33494	-20515	-11723	0	-25121	0	0	0	13816.4	0	24283.4
GG	-2512.1	-35169	-12979	-10467	0	21771.4	0	0	0	-18422	0	17165.9	0	-9211	0	0
TA	19678	0	0	0	14235.1	0	0	0	2930.76	0	0	0	-30145	5024.16	-418.68	837.36
TC	31819.7	0	0	0	25539.5	0	0	0	-34332	9629.64	13816.4	-9211	5024.16	0	0	0
TG	12560.4	0	0	0	-35588	-3349.4	-17166	-20934	6698.88	0	0	0	-418.68	0	-5861.5	0
TT	-33076	2930.76	4186.8	-11304	4186.8	0	0	0	-5442.8	0	24283.4	0	837.36	0	0	0
delta S	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GT	GG	TA	TC	TG	TT
AA	0	0	0	7.11756	0	0	0	19.2593	0	0	0	-9.6296	54.0097	84.5734	30.9823	-92.947
AC	0	0	0	61.1273	0	0	0	-18.422	-41.031	-15.91	13.3978	-93.784	0	0	0	0.83736
AG	0	0	0	-9.6296	-17.585	-2.5121	-55.266	-87.923	0	0	0	-39.775	0	0	0	3.76812
AT	7.11756	61.1273	-9.6296	-85.411	0	0	0	-25.958	0	0	0	-34.75	0	0	0	-45.217
CA	0	0	-17.585	0	0	0	15.4912	0	0	0	-9.6296	0	33.4944	68.6635	-95.04	2.93076
CC	0	0	-2.5121	0	0	0	-30.145	0	59.4526	37.2625	-83.317	56.5218	0	0	-18.841	0
CG	0	0	-55.266	0	15.4912	-30.145	-113.88	-25.54	0	0	-64.058	0	0	0	-48.986	0
CT	19.2593	-18.422	-87.923	-25.958	0	0	-25.54	0	0	0	-33.494	0	0	0	-66.151	0
GA	0	-41.031	0	0	0	59.4526	0	0	0	-4.1868	0	0	2.93076	-92.947	15.0725	-22.19
GC	0	-15.91	0	0	0	37.2625	0	0	-4.1868	-101.32	-66.151	-51.498	0	22.6087	0	0
GT	0	13.3978	0	0	-9.6296	-83.317	-64.058	-33.494	0	-66.151	0	0	0	43.5427	0	68.2448
GG	-9.6296	-93.784	-39.775	-34.75	0	56.5218	0	0	0	-51.498	0	39.7746	0	-35.169	0	0
TA	54.0097	0	0	0	33.4944	0	0	0	2.93076	0	0	0	-89.179	2.93076	-7.1176	-6.2802
TC	84.5734	0	0	0	68.6635	0	0	0	-92.947	22.6087	43.5427	-35.169	2.93076	0	0	0
TG	30.9823	0	0	0	-95.04	-18.841	-48.986	-66.151	15.0725	0	0	0	-7.1176	0	-25.958	0
TT	-92.947	0.83736	3.76812	-45.217	2.93076	0	0	0	-22.19	0	68.2448	0	-6.2802	0	0	0
delta H	A/T	G/C														
Init/Term	9629.64	418.68														
delta S	A/T	G/C														
Init/Term	17.1659	-11.723	Supplementary Table 1B													

Supplementary Table 2

Table 2

cwrs07	***Incorporation Rate***				**Base Rate**				
Chip	***A***	***C***	***G***	***T***	**A**	**C**	**G**	**T**	**Correlation**
1	*0.980*	*0.963*	*0.891*	*0.934*	0.973	0.972	0.881	1.000	0.819
2	*0.968*	*0.962*	*0.879*	*0.925*	0.998	0.972	0.900	1.000	0.812
3	*0.982*	*0.959*	*0.877*	*0.919*	0.999	0.976	0.895	1.000	0.806
4	*0.973*	*0.948*	*0.878*	*0.925*	0.992	0.991	0.879	1.000	0.778
5	*0.983*	*0.953*	*0.914*	*0.966*	0.982	0.989	0.899	1.000	0.766
6	*0.981*	*0.957*	*0.909*	*0.960*	0.979	0.986	0.886	1.000	0.778
7	*0.985*	*0.959*	*0.886*	*0.942*	0.992	0.976	0.907	1.000	0.802
8	*0.988*	*0.960*	*0.914*	*0.963*	0.943	0.984	0.842	1.000	0.764
9	*0.979*	*0.967*	*0.889*	*0.940*	0.987	0.972	0.876	1.000	0.797
10	*0.970*	*0.965*	*0.879*	*0.935*	1.000	0.970	0.884	1.000	0.804
11	*1.000*	*0.969*	*0.955*	*0.957*	0.932	0.983	0.863	0.987	0.764
12	*0.976*	*0.959*	*0.869*	*0.918*	1.000	0.962	0.915	1.000	0.817
13	*0.955*	*0.974*	*0.965*	*0.947*	0.984	0.984	0.847	0.987	0.769
14	*1.000*	*0.963*	*0.963*	*0.967*	0.925	0.979	0.826	1.000	0.758
15	*0.991*	*0.947*	*0.953*	*0.965*	0.930	1.000	0.829	1.000	0.732
16	*0.999*	*0.947*	*0.936*	*0.959*	0.926	1.000	0.833	1.000	0.773
17	*0.973*	*0.962*	*0.899*	*0.935*	0.982	0.973	0.876	1.000	0.773
18	*0.979*	*0.962*	*0.912*	*0.949*	0.954	0.978	0.848	1.000	0.764
19	*0.995*	*0.971*	*0.976*	*0.966*	0.935	0.953	0.820	1.000	0.722
20	*0.990*	*0.962*	*0.964*	*0.964*	0.937	0.972	0.828	1.000	0.730
21	*0.993*	*0.943*	*0.952*	*0.966*	0.939	0.991	0.832	1.000	0.735
22	*1.000*	*0.949*	*0.931*	*0.950*	0.928	0.977	0.851	1.000	0.727
23	*1.000*	*0.943*	*0.851*	*0.925*	0.997	0.976	0.962	0.964	0.828
24	*0.973*	*0.965*	*0.906*	*0.932*	0.999	0.967	0.900	1.000	0.786
25	*0.992*	*0.933*	*0.906*	*0.949*	0.943	1.000	0.852	1.000	0.730
26	*0.982*	*0.958*	*0.888*	*0.922*	0.977	0.963	0.877	1.000	0.785
27	*0.981*	*0.940*	*0.943*	*0.971*	0.959	0.997	0.848	1.000	0.742
28	*0.989*	*0.939*	*0.940*	*0.966*	0.945	1.000	0.844	1.000	0.720
29	*1.000*	*0.949*	*0.971*	*0.953*	0.918	0.980	0.809	1.000	0.743
30	*0.986*	*0.959*	*0.968*	*0.964*	0.946	0.968	0.828	1.000	0.721
31	*1.000*	*0.933*	*0.958*	*0.962*	0.926	0.999	0.825	1.000	0.725
32	*0.969*	*0.961*	*0.890*	*0.934*	1.000	0.970	0.886	1.000	0.785
33	*0.996*	*0.925*	*0.930*	*0.951*	0.932	0.996	0.829	1.000	0.678
34	*0.991*	*0.955*	*0.852*	*0.905*	1.000	0.960	0.951	0.988	0.824
35	*0.973*	*0.910*	*0.928*	*0.933*	0.956	0.992	0.824	1.000	0.705
36	*0.991*	*0.971*	*0.906*	*0.941*	0.943	0.968	0.851	1.000	0.788
37	*0.982*	*0.962*	*0.906*	*0.934*	0.983	0.983	0.905	1.000	0.775
38	*0.980*	*0.965*	*0.915*	*0.948*	0.967	0.982	0.883	1.000	0.825
39	*1.000*	*0.943*	*0.923*	*0.951*	0.928	0.993	0.840	1.000	0.764
40	*0.957*	*0.888*	*0.873*	*0.912*	0.963	0.999	0.858	0.982	0.697
Mean	***0.985***	***0.952***	***0.916***	***0.945***	**0.963**	**0.981**	**0.865**	**0.998**	**0.765**
Std. Dev.	***0.012***	***0.017***	***0.035***	***0.018***	**0.028**	**0.013**	**0.036**	**0.007**	**0.038**
Max.	***1.000***	***0.974***	***0.976***	***0.971***	**1.000**	**1.000**	**0.962**	**1.000**	**0.828**
Min.	***0.955***	***0.888***	***0.851***	***0.905***	**0.918**	**0.953**	**0.809**	**0.964**	**0.678**

cwrs39	*Incorporation Rate*				Base Rate				
Chip	*A*	*C*	*G*	*T*	A	C	G	T	Correlation
1	*1.000*	*0.954*	*1.000*	*0.949*	0.883	0.936	0.794	1.000	0.620
2	*1.000*	*0.944*	*0.999*	*0.943*	0.882	0.937	0.780	1.000	0.608
3	*1.000*	*0.960*	*0.975*	*0.946*	0.889	0.917	0.810	0.999	0.550
4	*1.000*	*0.953*	*0.975*	*0.948*	0.886	0.928	0.804	1.000	0.603
5	*1.000*	*0.949*	*0.997*	*0.949*	0.881	0.937	0.784	1.000	0.612
6	*1.000*	*0.956*	*0.998*	*0.941*	0.880	0.921	0.778	1.000	0.604
7	*1.000*	*0.950*	*1.000*	*0.939*	0.874	0.925	0.774	1.000	0.615
8	*1.000*	*0.946*	*0.986*	*0.946*	0.886	0.935	0.793	1.000	0.603
9	*1.000*	*0.945*	*1.000*	*0.947*	0.884	0.936	0.781	1.000	0.617
10	*1.000*	*0.955*	*1.000*	*0.939*	0.882	0.921	0.775	1.000	0.595
11	*1.000*	*0.953*	*0.997*	*0.948*	0.883	0.930	0.784	1.000	0.616
12	*1.000*	*0.968*	*0.981*	*0.956*	0.891	0.917	0.820	1.000	0.622
13	*1.000*	*0.948*	*0.997*	*0.944*	0.887	0.929	0.786	1.000	0.585
14	*1.000*	*0.943*	*1.000*	*0.949*	0.880	0.937	0.779	1.000	0.618
15	*1.000*	*0.948*	*0.994*	*0.946*	0.886	0.931	0.787	1.000	0.595
16	*1.000*	*0.951*	*0.999*	*0.948*	0.884	0.928	0.784	1.000	0.595
17	*1.000*	*0.949*	*1.000*	*0.950*	0.883	0.932	0.785	0.999	0.606
18	*1.000*	*0.955*	*0.998*	*0.951*	0.881	0.928	0.787	0.999	0.610
19	*1.000*	*0.940*	*0.990*	*0.942*	0.883	0.935	0.786	1.000	0.611
20	*1.000*	*0.943*	*0.996*	*0.949*	0.883	0.940	0.779	1.000	0.635
21	*1.000*	*0.952*	*0.994*	*0.947*	0.884	0.928	0.790	1.000	0.589
22	*0.998*	*0.954*	*0.972*	*0.939*	0.899	0.927	0.810	1.000	0.672
23	*1.000*	*0.959*	*0.996*	*0.946*	0.887	0.922	0.789	0.997	0.599
24	*1.000*	*0.926*	*0.928*	*0.943*	0.882	0.955	0.849	1.000	0.594
25	*1.000*	*0.953*	*1.000*	*0.946*	0.881	0.929	0.782	1.000	0.610
26	*1.000*	*0.951*	*1.000*	*0.946*	0.873	0.929	0.782	1.000	0.625
27	*1.000*	*0.951*	*0.990*	*0.942*	0.878	0.926	0.784	1.000	0.611
28	*1.000*	*0.958*	*1.000*	*0.934*	0.871	0.914	0.776	1.000	0.592
29	*1.000*	*0.939*	*0.933*	*0.937*	0.880	0.939	0.839	1.000	0.593
30	*1.000*	*0.937*	*0.953*	*0.936*	0.881	0.935	0.819	1.000	0.569
31	*0.999*	*0.951*	*0.997*	*0.934*	0.876	0.917	0.780	1.000	0.586
32	*1.000*	*0.954*	*1.000*	*0.937*	0.874	0.919	0.775	1.000	0.603
33	*1.000*	*0.943*	*0.982*	*0.937*	0.882	0.932	0.794	1.000	0.582
34	*1.000*	*0.919*	*0.983*	*0.948*	0.887	0.970	0.801	1.000	0.605
35	*1.000*	*0.962*	*0.997*	*0.931*	0.873	0.899	0.783	0.995	0.578
36	*1.000*	*0.955*	*0.994*	*0.947*	0.892	0.930	0.791	1.000	0.618
37	*1.000*	*0.948*	*0.998*	*0.948*	0.885	0.935	0.782	1.000	0.618
38	*1.000*	*0.958*	*0.994*	*0.949*	0.882	0.923	0.793	0.999	0.612
39	*1.000*	*0.949*	*0.996*	*0.946*	0.881	0.932	0.781	1.000	0.617
40	*1.000*	*0.950*	*0.997*	*0.945*	0.880	0.934	0.780	1.000	0.613
Mean	***1.000***	***0.950***	***0.990***	***0.944***	**0.882**	**0.930**	**0.791**	**1.000**	**0.605**
Std. Dev.	***0.000***	***0.009***	***0.017***	***0.005***	**0.005**	**0.011**	**0.017**	**0.001**	**0.020**
Max.	***1.000***	***0.968***	***1.000***	***0.956***	**0.899**	**0.970**	**0.849**	**1.000**	**0.672**
Min.	***0.998***	***0.919***	***0.928***	***0.931***	**0.871**	**0.899**	**0.774**	**0.995**	**0.550**

cwrs51	***Incorporation Rate***				**Base Rate**				
Chip	***A***	***C***	***G***	***T***	**A**	**C**	**G**	**T**	**Correlation**
1	*0.991*	*0.968*	*0.988*	*0.973*	0.919	0.938	0.795	1.000	0.743
2	*1.000*	*0.932*	*1.000*	*0.952*	0.889	0.959	0.776	1.000	0.718
3	*0.984*	*0.967*	*0.978*	*0.974*	0.930	0.939	0.804	1.000	0.732
4	*1.000*	*0.943*	*0.988*	*0.973*	0.913	0.966	0.793	1.000	0.736
5	*1.000*	*0.934*	*0.990*	*0.968*	0.899	0.978	0.789	1.000	0.739
6	*1.000*	*0.933*	*1.000*	*0.963*	0.900	0.971	0.783	1.000	0.718
7	*1.000*	*0.948*	*1.000*	*0.971*	0.904	0.957	0.786	1.000	0.740
8	*1.000*	*0.953*	*0.998*	*0.963*	0.891	0.956	0.779	1.000	0.749
9	*1.000*	*0.931*	*1.000*	*0.964*	0.907	0.966	0.784	1.000	0.691
10	*1.000*	*0.936*	*0.996*	*0.966*	0.899	0.976	0.785	1.000	0.732
11	*1.000*	*0.936*	*0.996*	*0.965*	0.896	0.974	0.785	1.000	0.734
12	*1.000*	*0.960*	*1.000*	*0.974*	0.913	0.949	0.791	1.000	0.733
13	*1.000*	*0.947*	*0.924*	*0.978*	0.910	0.965	0.867	1.000	0.724
14	*1.000*	*0.931*	*1.000*	*0.966*	0.904	0.976	0.786	1.000	0.720
15	*1.000*	*0.940*	*0.982*	*0.971*	0.905	0.971	0.809	1.000	0.738
16	*1.000*	*0.969*	*0.999*	*0.979*	0.909	0.939	0.791	1.000	0.766
17	*1.000*	*0.934*	*1.000*	*0.946*	0.881	0.961	0.780	1.000	0.718
18	*0.985*	*0.959*	*1.000*	*0.970*	0.919	0.948	0.788	1.000	0.734
19	*1.000*	*0.939*	*0.993*	*0.967*	0.903	0.969	0.794	1.000	0.731
20	*1.000*	*0.941*	*0.994*	*0.953*	0.903	0.951	0.792	1.000	0.720
21	*1.000*	*0.928*	*1.000*	*0.965*	0.898	0.973	0.786	1.000	0.712
22	*1.000*	*0.944*	*1.000*	*0.966*	0.899	0.959	0.784	1.000	0.724
23	*1.000*	*0.949*	*1.000*	*0.978*	0.905	0.968	0.788	1.000	0.749
24	*0.991*	*0.965*	*0.979*	*0.976*	0.926	0.935	0.810	1.000	0.697
25	*1.000*	*0.920*	*1.000*	*0.960*	0.897	0.979	0.786	1.000	0.694
26	*1.000*	*0.956*	*1.000*	*0.970*	0.904	0.955	0.793	1.000	0.741
27	*0.995*	*0.936*	*1.000*	*0.967*	0.905	0.975	0.792	1.000	0.722
28	*0.999*	*0.957*	*1.000*	*0.968*	0.904	0.952	0.792	1.000	0.724
29	*1.000*	*0.965*	*1.000*	*0.957*	0.899	0.941	0.796	1.000	0.712
30	*1.000*	*0.929*	*1.000*	*0.966*	0.907	0.979	0.791	1.000	0.721
31	*1.000*	*0.932*	*1.000*	*0.971*	0.906	0.976	0.790	1.000	0.727
32	*1.000*	*0.919*	*1.000*	*0.957*	0.892	0.970	0.777	1.000	0.700
33	*1.000*	*0.929*	*1.000*	*0.955*	0.903	0.966	0.783	1.000	0.705
34	*1.000*	*0.956*	*0.997*	*0.979*	0.910	0.964	0.793	1.000	0.755
35	*1.000*	*0.984*	*1.000*	*0.977*	0.919	0.913	0.795	1.000	0.725
36	*1.000*	*0.953*	*1.000*	*0.974*	0.903	0.958	0.784	1.000	0.749
37	*1.000*	*0.938*	*1.000*	*0.949*	0.884	0.958	0.776	1.000	0.727
38	*1.000*	*0.953*	*0.995*	*0.970*	0.895	0.965	0.787	1.000	0.767
39	*1.000*	*0.949*	*1.000*	*0.978*	0.908	0.968	0.784	1.000	0.796
40	*1.000*	*0.933*	*1.000*	*0.970*	0.907	0.974	0.784	1.000	0.730
Mean	***0.999***	***0.945***	***0.995***	***0.967***	**0.904**	**0.961**	**0.791**	**1.000**	**0.730**
Std. Dev.	***0.004***	***0.015***	***0.013***	***0.008***	**0.010**	**0.015**	**0.014**	**0.000**	**0.021**
Max.	***1.000***	***0.984***	***1.000***	***0.979***	**0.930**	**0.979**	**0.867**	**1.000**	**0.796**
Min.	***0.984***	***0.919***	***0.924***	***0.946***	**0.881**	**0.913**	**0.776**	**1.000**	**0.691**

cwrs53	***Incorporation Rate***				**Base Rate**				
Chip	***A***	***C***	***G***	***T***	**A**	**C**	**G**	**T**	**Correlation**
1	1.000	0.954	1.000	0.971	0.897	0.959	0.800	1.000	0.704
2	0.984	0.948	1.000	0.975	0.914	0.972	0.808	1.000	0.689
3	1.000	0.954	0.990	0.964	0.902	0.960	0.811	1.000	0.729
4	1.000	0.968	1.000	0.980	0.905	0.937	0.797	1.000	0.721
5	1.000	0.930	1.000	0.974	0.895	0.988	0.793	1.000	0.705
6	1.000	0.936	0.989	0.977	0.897	0.978	0.806	1.000	0.727
7	1.000	0.928	0.990	0.970	0.902	0.991	0.806	1.000	0.704
8	1.000	0.964	1.000	0.967	0.886	0.940	0.793	1.000	0.705
9	0.983	0.962	1.000	0.975	0.909	0.952	0.798	1.000	0.701
10	1.000	0.940	0.983	0.974	0.892	0.977	0.811	1.000	0.713
11	1.000	0.941	0.974	0.974	0.893	0.976	0.821	1.000	0.681
12	1.000	0.957	1.000	0.977	0.905	0.958	0.804	1.000	0.699
13	1.000	0.935	0.973	0.968	0.900	0.976	0.825	1.000	0.742
14	0.968	0.947	1.000	0.977	0.927	0.974	0.803	1.000	0.733
15	0.998	0.933	1.000	0.976	0.893	0.979	0.803	1.000	0.744
16	1.000	0.947	1.000	0.980	0.901	0.973	0.804	1.000	0.718
17	1.000	0.925	0.980	0.964	0.905	0.997	0.823	1.000	0.738
18	1.000	0.934	0.995	0.958	0.896	0.977	0.804	1.000	0.719
19	0.989	0.945	1.000	0.971	0.905	0.973	0.797	1.000	0.725
20	1.000	0.933	0.974	0.974	0.889	0.991	0.808	0.999	0.732
21	0.997	0.933	0.947	0.967	0.911	0.977	0.851	1.000	0.733
22	1.000	0.967	0.997	0.969	0.909	0.944	0.809	1.000	0.727
23	1.000	0.935	0.993	0.977	0.906	0.988	0.811	1.000	0.697
24	0.979	0.943	1.000	0.975	0.913	0.974	0.797	1.000	0.696
25	1.000	0.935	0.955	0.973	0.901	0.974	0.838	1.000	0.718
26	0.989	0.955	1.000	0.979	0.898	0.964	0.798	1.000	0.720
27	1.000	0.933	0.998	0.971	0.899	0.977	0.798	1.000	0.686
28	0.992	0.948	1.000	0.975	0.892	0.968	0.796	1.000	0.723
29	1.000	0.941	1.000	0.971	0.898	0.970	0.797	1.000	0.728
30	1.000	0.941	1.000	0.975	0.899	0.964	0.797	1.000	0.728
31	1.000	0.935	1.000	0.970	0.894	0.981	0.799	1.000	0.691
32	1.000	0.939	0.995	0.971	0.900	0.971	0.805	1.000	0.695
33	1.000	0.916	0.958	0.954	0.912	0.992	0.841	1.000	0.714
34	0.991	0.945	0.964	0.981	0.904	0.979	0.838	1.000	0.695
35	1.000	0.928	0.989	0.975	0.896	0.985	0.813	1.000	0.726
36	0.992	0.936	0.968	0.979	0.905	0.990	0.835	1.000	0.699
37	0.994	0.975	1.000	0.976	0.911	0.933	0.797	1.000	0.716
38	0.997	0.941	0.983	0.976	0.897	0.983	0.816	1.000	0.704
39	1.000	0.957	1.000	0.977	0.896	0.961	0.798	1.000	0.718
40	0.988	0.941	1.000	0.975	0.907	0.981	0.802	1.000	0.711
Mean	***0.996***	***0.943***	***0.990***	***0.973***	**0.901**	***0.972***	***0.809***	***1.000***	***0.714***
Std. Dev.	***0.007***	***0.013***	***0.015***	***0.006***	**0.008**	***0.015***	***0.015***	***0.000***	***0.016***
Max.	***1.000***	***0.975***	***1.000***	***0.981***	**0.927**	***0.997***	***0.851***	***1.000***	***0.744***
Min.	***0.968***	***0.916***	***0.947***	***0.954***	**0.886**	***0.933***	***0.793***	***0.999***	***0.681***

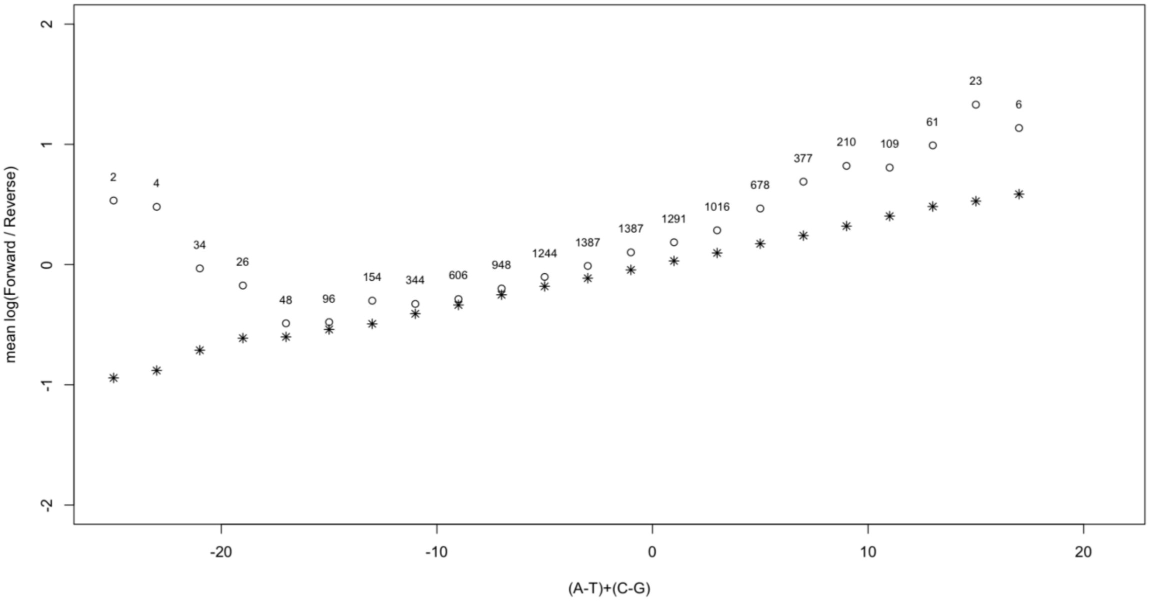
cwrs63	***Incorporation Rate***				**Base Rate**				
Chip	***A***	***C***	***G***	***T***	**A**	**C**	**G**	**T**	**Correlation**
1	*0.994*	*0.965*	*0.948*	*0.942*	0.948	0.933	0.870	1.000	0.646
2	*1.000*	*0.932*	*0.927*	*0.963*	0.910	0.972	0.884	1.000	0.614
3	*1.000*	*0.934*	*0.934*	*0.960*	0.920	0.973	0.867	1.000	0.644
4	*0.988*	*0.949*	*0.959*	*0.950*	0.936	0.949	0.843	1.000	0.655
5	*1.000*	*0.947*	*0.953*	*0.956*	0.924	0.949	0.852	0.998	0.667
6	*1.000*	*0.938*	*0.924*	*0.961*	0.906	0.962	0.875	1.000	0.608
7	*0.945*	*0.938*	*0.917*	*0.968*	0.987	0.978	0.907	1.000	0.618
8	*0.961*	*0.939*	*0.949*	*0.957*	0.959	0.968	0.856	1.000	0.667
9	*1.000*	*0.948*	*0.953*	*0.949*	0.927	0.950	0.851	1.000	0.679
10	*0.977*	*0.949*	*0.951*	*0.952*	0.943	0.962	0.856	1.000	0.680
11	*0.982*	*0.943*	*0.972*	*0.955*	0.925	0.962	0.821	1.000	0.630
12	*0.999*	*0.969*	*0.933*	*0.949*	0.939	0.930	0.885	1.000	0.692
13	*0.953*	*0.932*	*0.898*	*0.952*	0.998	0.983	0.927	1.000	0.621
14	*1.000*	*0.960*	*0.991*	*0.950*	0.928	0.941	0.818	0.997	0.675
15	*0.978*	*0.939*	*0.993*	*0.946*	0.936	0.967	0.806	1.000	0.678
16	*0.976*	*0.939*	*0.962*	*0.942*	0.955	0.963	0.847	1.000	0.709
17	*1.000*	*0.944*	*0.958*	*0.945*	0.931	0.955	0.844	0.995	0.683
18	*0.979*	*0.939*	*0.943*	*0.939*	0.956	0.955	0.865	1.000	0.689
19	*1.000*	*0.943*	*0.946*	*0.951*	0.925	0.956	0.858	0.999	0.674
20	*0.970*	*0.941*	*0.958*	*0.953*	0.950	0.962	0.844	1.000	0.687
21	*0.965*	*0.924*	*0.908*	*0.938*	0.984	0.964	0.916	0.984	0.692
22	*1.000*	*0.976*	*0.966*	*0.960*	0.930	0.922	0.845	1.000	0.651
23	*0.990*	*0.945*	*0.921*	*0.944*	0.959	0.952	0.906	0.992	0.718
24	*0.956*	*0.943*	*0.923*	*0.960*	0.980	0.973	0.896	1.000	0.652
25	*1.000*	*0.949*	*0.974*	*0.968*	0.908	0.951	0.828	0.993	0.660
26	*1.000*	*0.937*	*0.954*	*0.958*	0.908	0.960	0.847	1.000	0.626
27	*0.993*	*0.943*	*0.931*	*0.940*	0.943	0.945	0.874	0.996	0.697
28	*1.000*	*0.943*	*0.929*	*0.967*	0.901	0.961	0.870	1.000	0.660
29	*0.942*	*0.949*	*0.918*	*0.971*	0.993	0.967	0.905	1.000	0.637
30	*0.995*	*0.942*	*0.932*	*0.945*	0.942	0.958	0.879	0.997	0.638
31	*0.975*	*0.941*	*0.977*	*0.951*	0.936	0.963	0.818	1.000	0.657
32	*1.000*	*0.950*	*0.943*	*0.946*	0.924	0.949	0.857	1.000	0.677
33	*0.964*	*0.940*	*0.963*	*0.955*	0.957	0.965	0.841	1.000	0.673
34	*0.970*	*0.941*	*0.921*	*0.945*	0.968	0.965	0.896	1.000	0.676
35	*0.965*	*0.938*	*0.918*	*0.953*	0.952	0.963	0.884	1.000	0.687
36	*0.965*	*0.938*	*0.933*	*0.948*	0.970	0.966	0.876	1.000	0.661
37	*0.987*	*0.948*	*0.931*	*0.948*	0.960	0.950	0.888	0.989	0.710
38	*1.000*	*0.941*	*0.952*	*0.943*	0.927	0.959	0.847	1.000	0.667
39	*0.981*	*0.935*	*0.921*	*0.939*	0.966	0.973	0.907	1.000	0.709
40	*0.964*	*0.940*	*0.917*	*0.949*	0.969	0.969	0.898	1.000	0.656
Mean	***0.983***	***0.944***	***0.943***	***0.952***	**0.944**	**0.959**	**0.866**	**0.998**	**0.666**
Std. Dev.	***0.018***	***0.010***	***0.022***	***0.009***	**0.025**	**0.013**	**0.029**	**0.003**	**0.028**
Max.	***1.000***	***0.976***	***0.993***	***0.971***	**0.998**	**0.983**	**0.927**	**1.000**	**0.718**
Min.	***0.942***	***0.924***	***0.898***	***0.938***	**0.901**	**0.922**	**0.806**	**0.984**	**0.608**

cwrs67	*Incorporation Rate*				Base Rate				
Chip	*A*	*C*	*G*	*T*	A	C	G	T	Correlation
1	*0.987*	*0.940*	*0.924*	*0.945*	0.947	0.939	0.892	1.000	0.625
2	*0.991*	*0.960*	*0.961*	*0.957*	0.924	0.929	0.839	1.000	0.596
3	*1.000*	*0.932*	*0.938*	*0.956*	0.925	0.962	0.872	0.994	0.641
4	*1.000*	*0.953*	*0.957*	*0.957*	0.923	0.935	0.846	1.000	0.607
5	*0.988*	*0.923*	*0.919*	*0.943*	0.943	0.964	0.892	1.000	0.596
6	*1.000*	*0.934*	*0.908*	*0.955*	0.923	0.960	0.902	1.000	0.585
7	*1.000*	*0.939*	*0.927*	*0.956*	0.923	0.952	0.881	1.000	0.621
8	*0.995*	*0.929*	*0.919*	*0.946*	0.934	0.962	0.895	1.000	0.693
9	*1.000*	*0.928*	*0.918*	*0.946*	0.925	0.960	0.885	1.000	0.659
10	*1.000*	*0.941*	*0.922*	*0.966*	0.912	0.954	0.884	1.000	0.647
11	*0.998*	*0.917*	*0.906*	*0.945*	0.928	0.966	0.903	0.997	0.639
12	*0.989*	*0.942*	*0.921*	*0.947*	0.954	0.936	0.905	0.990	0.627
13	*0.980*	*0.935*	*0.922*	*0.951*	0.939	0.954	0.875	1.000	0.607
14	*0.974*	*0.916*	*0.887*	*0.941*	0.967	0.965	0.933	0.991	0.611
15	*0.993*	*0.921*	*0.901*	*0.941*	0.940	0.964	0.910	1.000	0.673
16	*0.980*	*0.935*	*0.945*	*0.956*	0.935	0.962	0.856	1.000	0.607
17	*1.000*	*0.933*	*0.926*	*0.952*	0.923	0.951	0.880	1.000	0.641
18	*0.967*	*0.930*	*0.923*	*0.959*	0.949	0.963	0.877	1.000	0.598
19	*1.000*	*0.928*	*0.935*	*0.953*	0.923	0.960	0.872	1.000	0.641
20	*1.000*	*0.931*	*0.920*	*0.950*	0.927	0.953	0.890	0.998	0.625
21	*0.990*	*0.927*	*0.907*	*0.945*	0.937	0.957	0.899	1.000	0.629
22	*0.984*	*0.923*	*0.936*	*0.950*	0.934	0.963	0.864	1.000	0.600
23	*0.985*	*0.918*	*0.909*	*0.943*	0.957	0.968	0.914	1.000	0.669
24	*0.989*	*0.916*	*0.899*	*0.940*	0.940	0.968	0.910	1.000	0.630
25	*0.997*	*0.927*	*0.932*	*0.949*	0.925	0.960	0.874	1.000	0.647
26	*1.000*	*0.931*	*0.919*	*0.948*	0.926	0.958	0.890	1.000	0.610
27	*0.965*	*0.932*	*0.955*	*0.963*	0.944	0.960	0.846	1.000	0.569
28	*0.970*	*0.930*	*0.928*	*0.955*	0.948	0.960	0.875	1.000	0.622
29	*0.983*	*0.925*	*0.914*	*0.946*	0.944	0.963	0.887	1.000	0.600
30	*0.975*	*0.929*	*0.916*	*0.952*	0.950	0.962	0.887	1.000	0.578
31	*0.981*	*0.933*	*0.915*	*0.941*	0.957	0.946	0.902	1.000	0.624
32	*0.991*	*0.919*	*0.903*	*0.939*	0.943	0.962	0.910	1.000	0.622
33	*0.967*	*0.933*	*0.920*	*0.962*	0.943	0.957	0.881	1.000	0.642
34	*0.984*	*0.935*	*0.914*	*0.943*	0.959	0.942	0.902	1.000	0.684
35	*0.987*	*0.930*	*0.927*	*0.951*	0.927	0.958	0.865	1.000	0.663
36	*0.982*	*0.937*	*0.915*	*0.951*	0.960	0.926	0.903	0.963	0.672
37	*0.963*	*0.937*	*0.943*	*0.961*	0.946	0.961	0.854	1.000	0.618
38	*0.963*	*0.936*	*0.942*	*0.962*	0.950	0.961	0.854	1.000	0.647
39	*1.000*	*0.934*	*0.937*	*0.954*	0.922	0.959	0.868	1.000	0.668
40	*0.973*	*0.928*	*0.920*	*0.954*	0.947	0.965	0.880	1.000	0.637
Mean	***0.987***	***0.931***	***0.923***	***0.951***	**0.938**	**0.956**	**0.884**	**0.998**	**0.629**
Std. Dev.	***0.012***	***0.009***	***0.016***	***0.007***	**0.013**	**0.011**	**0.021**	**0.006**	**0.029**
Max.	***1.000***	***0.960***	***0.961***	***0.966***	**0.967**	**0.968**	**0.933**	**1.000**	**0.693**
Min.	***0.963***	***0.916***	***0.887***	***0.939***	**0.912**	**0.926**	**0.839**	**0.963**	**0.569**

fmr1	***Incorporation Rate***				**Base Rate**				
Chip	***A***	***C***	***G***	***T***	**A**	**C**	**G**	**T**	**Correlation**
1	*0.955*	*0.933*	*0.887*	*0.938*	0.972	0.975	0.983	0.971	0.810
2	*0.963*	*0.944*	*0.885*	*0.943*	0.966	0.961	0.995	0.979	0.881
3	*0.967*	*0.955*	*0.920*	*0.955*	0.971	0.945	0.956	0.958	0.848
4	*0.970*	*0.939*	*0.908*	*0.949*	0.962	0.965	0.969	0.965	0.850
5	*0.972*	*0.936*	*0.875*	*0.953*	0.960	0.966	1.000	0.946	0.850
6	*0.955*	*0.930*	*0.884*	*0.944*	0.970	0.971	0.980	0.960	0.836
7	*0.956*	*0.933*	*0.895*	*0.943*	0.978	0.968	0.969	0.967	0.826
8	*0.960*	*0.943*	*0.898*	*0.949*	0.973	0.943	0.955	0.954	0.873
9	*0.962*	*0.939*	*0.891*	*0.948*	0.971	0.965	0.986	0.956	0.839
10	*0.964*	*0.943*	*0.896*	*0.950*	0.971	0.959	0.985	0.967	0.868
11	*0.965*	*0.948*	*0.907*	*0.951*	0.968	0.963	0.970	0.962	0.852
12	*0.989*	*0.972*	*0.934*	*0.959*	0.948	0.909	0.925	0.948	0.834
13	*0.979*	*0.968*	*0.912*	*0.961*	0.960	0.908	0.941	0.936	0.836
14	*0.949*	*0.928*	*0.878*	*0.941*	0.970	0.980	0.990	0.957	0.824
15	*0.985*	*0.928*	*0.926*	*0.956*	0.920	0.951	0.913	0.971	0.656
16	*0.957*	*0.921*	*0.882*	*0.940*	0.971	0.972	0.989	0.967	0.734
17	*0.969*	*0.949*	*0.916*	*0.948*	0.957	0.959	0.952	0.961	0.827
18	*0.971*	*0.941*	*0.909*	*0.953*	0.942	0.945	0.950	0.971	0.677
19	*0.962*	*0.942*	*0.891*	*0.948*	0.971	0.968	0.987	0.962	0.845
20	*0.974*	*0.938*	*0.906*	*0.948*	0.943	0.941	0.930	0.972	0.746
21	*0.972*	*0.952*	*0.916*	*0.954*	0.952	0.923	0.943	0.966	0.679
22	*0.994*	*0.999*	*0.981*	*0.961*	0.942	0.873	0.881	0.955	0.701
23	*0.961*	*0.941*	*0.899*	*0.943*	0.956	0.916	0.946	0.958	0.642
24	*0.972*	*0.945*	*0.923*	*0.950*	0.958	0.929	0.929	0.968	0.716
25	*0.974*	*0.937*	*0.921*	*0.949*	0.947	0.942	0.926	0.972	0.714
26	*0.957*	*0.932*	*0.879*	*0.943*	0.966	0.931	0.980	0.964	0.654
27	*0.975*	*0.955*	*0.920*	*0.960*	0.962	0.903	0.932	0.951	0.769
28	*0.964*	*0.937*	*0.911*	*0.946*	0.955	0.931	0.939	0.965	0.662
29	*0.974*	*0.960*	*0.930*	*0.955*	0.952	0.905	0.936	0.957	0.639
30	*0.972*	*0.969*	*0.933*	*0.953*	0.951	0.906	0.924	0.969	0.664
31	*0.960*	*0.941*	*0.905*	*0.945*	0.962	0.919	0.948	0.955	0.701
32	*0.960*	*0.937*	*0.877*	*0.942*	0.958	0.942	0.981	0.949	0.756
33	*0.968*	*0.943*	*0.907*	*0.951*	0.971	0.961	0.975	0.958	0.818
34	*0.968*	*0.941*	*0.908*	*0.955*	0.971	0.951	0.976	0.946	0.855
35	*0.972*	*0.952*	*0.927*	*0.950*	0.959	0.917	0.929	0.968	0.726
36	*0.974*	*0.950*	*0.923*	*0.954*	0.952	0.923	0.931	0.962	0.711
37	*0.960*	*0.925*	*0.911*	*0.939*	0.957	0.967	0.939	0.989	0.697
38	*0.971*	*0.961*	*0.930*	*0.958*	0.956	0.901	0.933	0.960	0.724
39	*0.990*	*0.951*	*0.924*	*0.956*	0.915	0.951	0.921	0.985	0.735
40	*0.963*	*0.944*	*0.896*	*0.947*	0.962	0.973	0.985	0.956	0.833
41	*0.966*	*0.949*	*0.896*	*0.951*	0.968	0.946	0.980	0.955	0.858
42	*0.959*	*0.949*	*0.872*	*0.944*	0.965	0.962	0.997	0.955	0.852
43	*0.966*	*0.946*	*0.912*	*0.952*	0.965	0.958	0.967	0.955	0.809
44	*0.973*	*0.947*	*0.919*	*0.954*	0.950	0.930	0.927	0.962	0.762
45	*0.965*	*0.957*	*0.907*	*0.952*	0.969	0.951	0.966	0.959	0.857
46	*0.970*	*0.975*	*0.963*	*0.962*	0.957	0.875	0.884	0.961	0.616
47	*0.972*	*0.961*	*0.927*	*0.961*	0.963	0.906	0.917	0.953	0.822

48	*0.959*	*0.945*	*0.883*	*0.952*	0.970	0.963	0.990	0.951	0.857
49	*0.991*	*0.922*	*0.918*	*0.939*	0.913	0.965	0.920	0.985	0.661
50	*0.956*	*0.910*	*0.886*	*0.954*	0.968	0.960	0.997	0.966	0.743
51	*0.965*	*0.947*	*0.915*	*0.953*	0.961	0.924	0.946	0.956	0.731
52	*0.970*	*0.942*	*0.907*	*0.953*	0.964	0.956	0.976	0.952	0.830
53	*0.954*	*0.919*	*0.892*	*0.953*	0.980	0.974	0.981	0.964	0.828
54	*0.978*	*0.944*	*0.904*	*0.954*	0.959	0.927	0.971	0.960	0.745
55	*0.961*	*0.941*	*0.914*	*0.948*	0.961	0.924	0.942	0.962	0.638
56	*0.965*	*0.930*	*0.908*	*0.945*	0.952	0.955	0.935	0.976	0.716
57	*0.955*	*0.932*	*0.891*	*0.936*	0.959	0.930	0.957	0.968	0.646
58	*0.952*	*0.964*	*0.917*	*0.939*	0.969	0.873	0.913	0.950	0.627
59	*0.976*	*0.933*	*0.908*	*0.949*	0.931	0.971	0.945	0.986	0.691
60	*0.975*	*0.948*	*0.917*	*0.952*	0.943	0.923	0.914	0.964	0.719
61	*0.963*	*0.945*	*0.899*	*0.945*	0.970	0.964	0.968	0.968	0.864
62	*0.970*	*0.945*	*0.909*	*0.953*	0.947	0.928	0.935	0.967	0.639
Mean	***0.967***	***0.944***	***0.908***	***0.950***	**0.959**	**0.941**	**0.953**	**0.962**	**0.762**
Std. Dev.	***0.010***	***0.015***	***0.020***	***0.006***	**0.014**	**0.027**	**0.029**	**0.010**	**0.081**
Max.	***0.994***	***0.999***	***0.981***	***0.962***	**0.980**	**0.980**	**1.000**	**0.989**	**0.881**
Min.	***0.949***	***0.910***	***0.872***	***0.936***	**0.913**	**0.873**	**0.881**	**0.936**	**0.616**

Supplementary Figure 1



Supplementary Figure 1 Legend: For each Forward/Reverse probe pair the mean of the log (Forward/Reverse) was calculated across all 62 chips from the FMR1 design. Probes are binned by base composition with each bin corresponding to the excess of A over T nucleotides plus the excess of C over G nucleotides on the Forward strand. The numbers on the figure correspond to the number of distinct features that are observed in the bin. Observed (O) and expected means (∗) are plotted for each bin.

Chapter 3:

SNP and CNV Detection in Normal and Trisomy 21 Individuals Using a First-Principles Approach

Yasminka A. Jakubek and David J. Cutler

Abstract

Affymetrix 6.0 (Affy 6.0) arrays are used for the analysis of common variants in large human genetic studies. They are used to simultaneously genotype single nucleotide polymorphism (SNPs) and copy number variants (CNVs). Current methods for SNP genotyping rely on data across samples. They are highly accurate; however, they tend to drop 20% to 33% of the targeted SNPs and have difficulty calling SNPs with a low minor allele frequency. Batch effects are a significant complication when analyzing these arrays. We have developed a SNP and CNV detection algorithm for Affy 6.0 arrays that is based on a low-level model of hybridization, which fully models cross-hybridization. In this approach chips are independently analyzed and batch effects are explicitly modeled. Our algorithm can genotype SNPs and CNVs on chromosomes of any ploidy, and each SNP call has a quality score (QS). We analyzed data from Down syndrome and normal samples. 13% of targeted SNPs show significant cross-hybridization. 84% of SNPs on diploid chromosomes and 57% of SNPs on trisomic chromosome 21 had QS > 0.99. We called an average of 50 CNVs per samples and 68% of the CNVs called were in the database of genomic variants (DGV). This data was previously analyzed; validation was attempted for 64 CNVs of which 59 were validated. Our method called only the 59 validated CNVs.

Introduction

DNA microarrays have many applications in the field of human genetics, including SNP typing, CNV detection, RNA profiling, and identification of protein binding sites[1]{Stoughton, 2005 #76}{Stoughton, 2005 #76}. {Stoughton, 2005 #73}{Stoughton, 2005 #59}One of the most useful applications of this technology is genotyping microarrays, because they provide an efficient tool for the analysis of common variation. Genotyping microarrays are used to simultaneously detect single nucleotide polymorphisms (SNPs) and copy number variants (CNVs)[1-6].

Affymetrix Genome-Wide Human SNP Array 6.0 (Affy 6.0) arrays are a type of genotyping array that is commonly used in human genetic studies [5, 7-10]. They are used to simultaneously type 906,600 SNPs together with 946,000 monomorphic probes useful for CNVs detection. The probes that are used to genotype SNPs are organized into probe sets, one per SNP. Each set is made up of two unique probe sequences, one for each SNP allele. The intensities for these probes are used to call the genotypes[11]. Several SNP genotyping algorithms have been developed for the analysis of SNP microarray data [4, 12-17]. The most accurate methods call genotypes by using prior empirical knowledge of where genotype clusters are likely to exist, together with experiment generated normalization algorithms [16]. Data from prior experiments together with all current chips is used to create clusters for each genotype. If we label the alleles A and B, there are three possible genotypes/cluster AA, AB, and BB. The observed intensities for probes A and B are assigned to one of the three clusters, and the genotype is called. Improvements upon this method include recalibrating the clusters and pre-processing the array data to correct for batch effects [12, 16-18]. Since these methods are based on empirical data from diploid autosomal chromosomes, they cannot be easily adapted for SNPs on chromosomes with a ploidy larger than two. These methods also make it difficult to call SNPs with a low minor allele frequency. In genotyping arrays, CNVs are called by using the observed probe intensities. These probe intensities are normalized and then used to call CNVs. Currently, the standard analysis method for Affy 6.0 arrays consists of two algorithms, Canary (copy number analysis routine) and Birdseye [12]. Intensity data from both monomorphic and SNP probes is used for CNV typing. The intensity for a SNP set is the sum of the intensities of the A and B allele. Canary is used to detect CNVs with a frequency greater than 1%. The algorithm compares log 2 probe intensities across all samples for predefined sets of probes that tag known CNVs. Birdseye is used to detect rare or de novo CNVs. Probe intensities across samples are used to identify contiguous probes that deviate from the standard diploid model, which is defined using data across samples. This and other CNV detection algorithms are susceptible to batch effects [12, 19].

There are two major problems that arise during the analysis of Affy 6.0 data. First, a significant fraction of SNPs (between 20% to 33%) are not called; this is true across different studies [7-10]. This problem is partially due to the fact that most samples are only processed once using only one array. Therefore, there are no independent replicates. The other problem is batch effects, in which probe intensities are affected by conditions that are unrelated to the genotype of the data [18-20]. In order to understand the causes for these observations, we previously developed a low-level model of hybridization on Affymetrix microarrays [20]. In this approach the predicted intensity for a probe spot is the function of the binding affinity between the probe and all targets and the specific hybridization conditions (target concentration, average target size, salt concentration, probe synthesis errors, wash stringency, and scanner settings). The correlation between observed probe intensities and the expected intensities under this model is approximately 70% [20]. This model also explains two striking differences between in-solution hybridization and hybridization on the array. The first on is that when an array has a probe spot for the forward target and a probe spot for the reverse target, these two probes seldom have the same intensity [4, 20]. The second is that mismatches towards the center of the probe have a much larger destabilizing effect than do mismatches near the edges [20, 21]. This is the reason why the query base is placed towards the center of the probe.

Our model explains these observations in terms of the experimental details. Binding between probe and target happen the same as in solution. Differences arise due to chip specific conditions. The difference between forward and reverse probes is the result of errors along the probe sequence. When manufacturing errors are not uniform for the four different nucleotides (A, C, G, T), then the forward and reverse probe have different binding affinities. The increase in the destabilizing effect of mismatches is due to target fragmentation. The target sequences that can bind a probe have different sizes, and bind the probe at different start and stop positions. Bases towards the middle of the probe are bound by target more often than bases towards the edges. The differences between in solution binding and binding on the array may also be due to physical link between the probe and the array surface; however, our previous work suggests that probe errors and fragmentation account for a large fraction of the observed intensities [20]. This model also helped us understand the underlying causes of batch effects in terms of differences between chips in probe error rates, DNA concentration, wash stringency, average target size, and scanner settings [20].

In the work presented here we applied our previously develop model to Affy 6.0 arrays. The purpose of this work is to understand problems with Affy 6.0 analysis and genotype SNPs and identify CNVs in Affy 6.0 array experiments without the use of empirical data. Our algorithm does not use reference data across different chips, but rather relies on the basic biochemistry of the array and the unique hybridization conditions for each individual experiment. Since the binding biochemistry of each probe spot is directly and independently modeled, the algorithm is capable of calling SNPs on chromosomes of any ploidy. Our approach assigns an estimate of accuracy for each SNP, the quality score (QS). The algorithm also produces direct estimates for the DNA concentration at each of the targeted genomic regions. These measures of local DNA concentration are relative to the ploidy of each individual chromosome. We use the log2 values of the local concentration (R) to call CNVs.

Our approach to genotyping on Affy 6.0 arrays is very different from current methods. It allows for individual analysis of each chip, and it explicitly models batch effects and cross-hybridization. For each chip we call 789,344 SNPs (87% of SNPs targeted by Affy 6.0) and do not call the remaining SNPs due to significant cross-hybridization to 8 or more genomic regions. We used our method to analyze data from 516 Down syndrome and 308 normal samples. Our results suggest that our approach has comparable accuracy to empirical approaches.

Methods

Low-Level Model of Hybridization on Microarrays

Microarrays consist of probes, single stranded DNA molecules attached to a microarray surface [1]. Probes are organized into probe spots. Each probe spot has thousands to millions of probes based on the same reference DNA sequence. The DNA that is hybridized onto the array is referred to as the target and it is fluorescently labeled. When a probe sequence is complimentary to a target then probe-target complexes form; consequently, the probe spot's intensity is a function of the number of these complexes [22, 23]. Ideally, the intensity for a probe spot would have a linear relationship to the copy number of the allele that perfectly complements the probe sequence. However, as is evident by chip effects as well as batch effects, this assumption does not consistently hold true [18, 20]. We have previously developed a low-level model of hybridization on the array that explains these inconsistencies in terms of the basic biochemistry of the array and the details of the experimental protocol [20]. It is summarized in the following paragraph.

The simplest model for the hybridization reaction at each probe spot is described by the Langmuir isotherm [20].

$$\propto = \frac{CK_{eq}}{CK_{eq} + 1} \tag{1}$$

The fraction of bound probes, a, is a function of C, the concentration of target DNA, and K_{eq}, the equilibrium constant for the reaction, which is a measure of the binding affinity between the target and probe sequences. We calculate all possible K_{eq} values for all probes and all target sequences. In our model we account for errors in probe synthesis by modeling two types of errors, base incorporation errors and abasic sites along the probe sequence, which together create probe spots with a heterogeneous mixture of sequences [24, 25]. Chips with different error rates for A, C, G, and T bases have different distributions of a values at each probe spot and consequently have different binding behavior. Other details of the experiment that we model are target DNA fragmentation and wash stringency, which also affect probe-target binding [20, 26, 27]. DNA is fragmented prior to hybridization; therefore, the concentration of available target DNA that can hybridize to a probe is a function of the global DNA concentration and the average length of the targets. The number of probe-target complexes that can form is also affected by the wash conditions. In the final step of the array protocol, the chip is washed with a low salt solution. The salt concentration and the duration of this wash step can vary between experiments and cause differences in the number of targets that disassociate from the probes. Overall this model helps explain how differences between chips in fragmentation, salt concentration, washing, and probe synthesis errors can cause batch effects, differences between chips that are unrelated to the genotype of the target DNA.

Application of our Model to Genotype Calling on Affymetrix Arrays

Our approach to genotype calling can be broken up into three general steps. We start by calculating K_{eq} values for every possible probe and target DNA fragment that can form on the array, including modeling bindings at every possible position along both probe and target sequences. This step allows us to form sets of probes and that bind the same genomic fragment. In the second step we fit chip-specific parameters that include target DNA fragmentation, four probe synthesis efficiency parameters (A, C, G, T), four parameters for the rate of abasic site formation (A, C, G, T), wash strength, global DNA concentration, minimum intensity, and maximum intensity. In the third step, we calculate expected probe intensity values. For SNP probes, expected values are calculated for every possible genotype. The genotype with the highest likelihood is called and a corresponding QS value is calculated for that call. Simultaneously, a log2 local DNA concentration value (R) is calculated for the genomic spot. This value is a modifier to the global DNA concentration for the chip and a measure of the copy number for that genomic region. These values are then input into a CNV calling algorithm, and used to call CNVs.

Binding Affinity Calculations for Genomic Fragments

In order to model all possible binding reactions on the array, we first start by creating a list of all genomic sequences on the array. To create this list we model the steps for genomic DNA preparation in the Affymetrix 6.0 protocol. These step include restriction enzyme digestion with NspI and StyI, PCR amplification, purification, and DNaseI fragmentation of the genomic DNA. We start by *in silico* restriction digesting the human genome (GRCh37) using NspI and StyI. We keep all genomic fragments that are between 180 and 1,210 base pairs (bp) long. We remove fragments with an N content in the reference sequence of more than 5%. We then model all SNPs present in dbSNP build 137 with a minor allele frequency of 1% or larger. For example if a genomic fragment has one SNP we model both the A and B allele. If there are two SNPs in a genomic fragment (SNP 1 has alleles A,B and SNP 2 has alleles C,D) we model all four genotypes (AC, AD, BC, and BD). Fragments in which more than 2% of the bases have SNPs are dropped, both to simplify the computational complexity of the problem, and because these regions often represent misassembly of the reference genome. We then “hybridize” all the probe sequences on the array to our list of genomic fragments. This step is carried out as described in [20], with a hybridization temperature of 49°C and one molar salt concentration. Briefly, we “cut” the genomic fragments at all possible positions in order to model the DNaseI fragmentation step in the array protocol. We then model bindings of each target DNA fragment to every probe starting at every position along the probe sequence. In this model a probe spot consists of target fragments of differing lengths that bind probes at different start and end positions. Therefore, for each target there are 300 K_{eq} values that describe binding at each 25 base probe spot. Even though each probe spot is made up of a heterogeneous mixture of target-probe complexes, each individual binding reaction is modeled as an in solution DNA-DNA complex using the Langmuir isotherm with nearest-neighbor kinetics.

After we calculate binding affinities between all probes and all genomic fragments, we organize the data into groups of probes that bind the same genomic fragment(s). Many of the probes do not cross hybridize and form neat probe–genomic fragment groups. For example, many of the monomorphic CNV probes only have significant binding to one fragment; therefore, information for that genome spot comes from one probe. For SNP probes that only bind one genome spot, the probe-genomic fragment group consists of the A and B allele probes and their binding affinities to two target sequences; these sequences have a one base difference at the SNP. Other probes have significant cross-hybridization to more than one genome spot. We sort these probe-genomic fragments by the genomic coordinate of the fragment with the strongest binding to the probes in the group. We define “strongest” as the fragment with the highest K_{eq} values (binding to probes). For a probe-genomic fragment group, the probe intensities are affected most by the genotype/copy number of the genome spot that contributes most to binding. It is important to note that even though we use the probe set to infer the genotype at one genomic location, cross-hybridization to the other genomic fragments is also used to calculate probe intensities. In the next step we remove probe groups that have more than eight genomic fragments that contribute to binding. We do so because we cannot confidently model such high levels of cross-hybridization. We use each of the resulting probe sets to call genotypes.

Parameter Fitting for Chips

In the second step we fit a set of parameters that describes the binding conditions for each experiment: global DNA concentration, probe errors, target fragmentation, wash stringency, and the dynamic range of the scanner (minimum and maximum intensity).

In order to fit parameters for a chip, we need to use a subset of observed probe intensities. For this purpose we selected SNP probes that do not have cross-hybridization. These probes account for 7,657 of the queried SNPs. Each of these SNPs can be categorized in terms of the A and B alleles for the SNP. There are six possible combinations of A/B alleles. They are A/C, A/G, A/T, C/G, C/T, G/T). In order to avoid any bias, the 7,657 SNPs that were used to fit the parameters included roughly the same number of each type of SNP. These SNPs were selected, because the genotype for the SNP can easily be inferred from the observed probe intensities. For each probe the ratio of the A probe intensity to the B probe intensity consistently falls in one of three categories: less than 0.5 (BB, homozygote); between 5/6 and 7/6 (AB, heterozygote); or greater than 2(AA, homozygote). Therefore, the genotype for each SNP is determined before the parameters are fit to the data.

We fit the chip specific parameters using a modified version of our previously developed algorithm [20]. In our previous work the wash, fragmentation, and salt parameters were input parameters. We improved our estimates for these parameters by directly fitting them to the data. The new parameter fitting procedure consists of four Powell iterations instead of one. The new fitting process works in the following way. In the first iteration we fit the wash, fragmentation, and a new salt parameter (X), which is used to model the stabilizing effect of salt on A and T bases. Each K_{eq} that describes binding between a probe and target is modified using the following equation.

$$K_{eq}X^{AT} \tag{2}$$

Where AT is the number of A and T bases in the probe that are complementary to the target. We fit X, wash, and fragmentation by minimizing the squared difference between expected and observed probe intensities. In the second iteration, we use Powell's method to fit four parameters: two that describe the global DNA concentration (NspI fragment concentration and StyI fragment concentration); and two parameters that describe the scanner's dynamic range (minimum and maximum intensity). This is done as described in our previous work, with two modifications: 1) We fit the parameters by maximizing the correlation between expected and observed probe intensities; 2) The expected intensity is a linear function of the fraction of bound probes as described in the following equation.

$$E\{intensity\} = (max - \mathrm{min})\,\alpha + min \tag{3}$$

Min is the minimum intensity, and max is the maximum intensity. In the third iteration we fit parameters that describe the errors along the probes sequence by minimizing the squared difference between the expected and observed intensities. Four parameters for incorporations errors for A, C, G, and T which cause probe truncation; and four that describe the rate of abasic site formation along the probes sequence. We explicitly model binding for the full-length probe and all possible probe sequences with one error. We then use the average difference between the K_{eq} values for the full-length probe and the K_{eq} values for probes with one error to estimate the effect of two and more errors on the final alpha value. We use this estimate because it would be computationally inefficient to model all possible errors and because probes with two and more errors contribute little to overall binding. In the fourth iteration, we re-fit the fragmentation, wash, and salt parameter.

Genotyping

For each chip, we use the parameters from the previous step and parameters for chromosome ploidy to call genotypes. The ploidy at each chromosome is an integer value provided by the user. We call genotypes for each genomic fragment using the observed probe intensities for the group. For each group we start by calculating the expected probe intensities for all possible SNP genotypes for that fragment. For example, a SNP with alleles A and B on an autosome has three possible genotypes (AA, AB, BB). Each probe has two sets of K_{eq} values, one for A and one for B. To get the expected probe intensity we multiply these K_{eq} values by the number of copies of each allele. For a SNP on a diploid chromosome A has 2,1,0 copies and B has 0,1,2 copies for each genotype AA, AB, and BB respectively. We then use the expected probe intensities to calculate the likelihood of each genotype. To calculate the likelihood for AA we use:

$$\log\big(li(AA)\big) \propto \sum_{p=1}^{n} \frac{(E\{pAA\}-Obs\{p\})^2}{2\sigma^2} \tag{4}$$

Where n is the number of probes, Obs{p} is the observed intensity for probe p, and E{pAA} is the expected intensity for probe p and genotype AA. In this step, we fit one parameter for the local DNA concentration (lc). The concentration for NspI and StyI is multiplied by lc. This parameter is fit to each genotype independently by doing a linear search for the value of lc that maximizes the likelihood function. We call the genotype with the highest likelihood and output the log2 of lc, R. This value, R, can be used in the same way as the log2 probe intensity values for CNV genotyping. In addition we report the posterior probability for the genotype, which we refer to as the quality score (QS). If the AA genotype is called then QS for that SNP is:

$$QS = \frac{li(AA)}{li(AA)+li(AB)+li(BB)} \tag{5}$$

In this model genotypes have a uniform prior. For a SNP on a chromosome with three copies we calculate four likelihood values, one for each possible genotype (AAA, AAB, ABB, BBB). The number of copies of A and B for are used to calculate the expected probe intensities. If the AAA genotype has the highest likelihood then QS for the call is:

$$QS = \frac{li(AAA)}{li(AAA)+li(AAB)+li(ABB)+li(BBB)} \tag{6}$$

This approach allows us to call genotypes on chromosomes of any ploidy and calculate R for CNV genotyping.

CNV Calling

We used our method to analyze data for male and female samples with and without Down syndrome. CNVs were called by analyzing R values with GADA, a CNV calling algorithm, with input parameters *a* = 0.2, T = 5, and M =6 (20). M is the minimum number of probes used to call a CNV. Parameter *a* is set to the default value and parameter T is set to 5 which is more stringent than the default T value of 4. Before we use GADA to call CNVs we normalize R values. First within chips, by dividing R values by the mean. We then normalize the values across chips. We ran the C version of GADA, which takes a list of R values for each chromosome and outputs a list of breakpoints with corresponding amplification values for each segment. A duplication is called if the amplification value is greater than:

$$\frac{1}{2}log_2(\frac{expected\ ploidy+1}{expected\ ploidy}) \tag{7}$$

A deletion is called if the amplification value is less than:

$$\frac{1}{2}log_2(\frac{expected\ ploidy-1}{expected\ ploidy}) \tag{8}$$

In cases where the expected ploidy is one, we set the deletion cut-off to -3.

Quality Control

To measure agreement between replicated probe spots we calculated the average coefficient of variation (variance/mean) for the 15,314 replicated probe spot intensity values used during the parameter-fitting step. We removed samples with an average coefficient of variation greater than 0.328 (mean for all the chips + 3 standard deviations) and those with more than 139 CNVs (mean + 3 standard errors). Our method does not drop calls, instead each call has a QS value; therefore, we used the number of SNP calls with QS > 0.9 for autosomal SNPs (chr 21 SNPs excluded) as a quality control measure for each chip. We removed samples when fewer than 87% (mean – 3 standard deviations) of SNPs have QS > 0.9.

Results

We calculated binding affinities for all probe sequences and all genomic targets on the Affy 6.0 array. There are 3.25 million target sequences from the genome, 5.8 million SNP probes, and 946,000 CNV probes. We removed probes that cross-hybridize to more than 8 genomic fragments. The probes that were removed account for 13% of targeted SNPs and 25% of CNV probes.

We used our method to call SNPs for samples with and without Down syndrome (Table 1). SNPs on autosomal chromosomes out of Hardy-Weinberg, those with a chi-square statistic greater than 20 (11.6%), were removed (Figure 1). We calculated the chi-square statistic using only data from individuals with two copies of chromosome 21. On average, SNP calls on diploid chromosomes had higher QS values than calls on trisomic chromosome 21; 84% of SNP calls with QS > 0.99 and 57% with QS > 0.99 respectively. Overall QS values were a good predictor of agreement between duplicates (Table 1 and Figure 2). When there are three copies of a SNP, QS values are more conservative (Table 1 and Figure 2). The data set included 110 trios, each had parents with two copies of chromosome 21 and a child with Down syndrome. We used these data to calculate the fraction of Mendelian consistent calls, which was 97.5% for SNP calls on autosomal chromosomes. QS values were a good predictor of Mendelian consistency (Table 2).

Next we focused on heterozygous calls for SNPs with three copies in order to understand how well our method does at detecting allele ratios of 2:1 and 1:2 for the different heterozygotes (AAB, BBA). Overall heterozygotes have lower QS values, with 53% of calls having QS > 0.90 (Table 3). When a heterozygote is called in one duplicate the same call is made in the opposite duplicate 79% of the time. If the call has a QS value greater than 0.9 there is agreement 88% of the time (Table 3). From these data we can conclude that there is some correlation between QS values and agreement between duplicates for the heterozygous calls; however, from these data we cannot directly estimate accuracy. In order to do so we looked at Mendelian inheritance of heterozygous SNPs calls for trisomic chromosome 21. For each trio we know which parent passed on two copies of chromosome 21. If the parent's genotype is known, there are eight genotype combinations where we have power to detect a genotype error in the child (Figure 3). We estimated accuracy for these heterozygous calls by using the data from all SNP calls with one of those genotype combinations and with QS > 0.9 for both parents (Table 4). The calls for these SNPs were accurate 91.2% of the time (QS > 0 for the child). For calls where the child had QS > 0.8, calls were 95% accurate.

After removing probes with significant cross-hybridization and those out of Hardy-Weinberg there are 1.35 million R values per chip. We called an average of 29 deletions and 21 duplications per sample. Of the CNVs called, 68% of them are in the database of genomic variants (DGV) (Table 5). For the CNVs called on chromosome 21, there appeared to be a different distribution of CNVs in DGV; however, approximately 200 of the calls mapped to one region in the p-arm that is not present in the latest genome build GRCh38. If those calls are removed, the remaining CNVs on chromosome 21 follow the same pattern as those on the rest of the autosomes. These data has been previously analyzed and validation was attempted for 64 CNVs, of length 40 kb or larger. We called all 59 CNVs that were validated with the Illumina HumanOmni2.5-8 bead chip, and did not call any of the CNV that failed to validate. Three of the validated CNVs were in chromosome 21.

We used the trio data to look at CNV inheritance. The fraction of CNVs in the parents provides an estimate of the product of the false positive and false negative rates. For every CNV called in the child we asked if it was present in one of the parents. CNVs in DGV are twice as likely to be in one of the parents (Table 6). Larger CNVs appear more frequently in the parents; however, this trend is not as defined for CNVs that are larger than 20kb.

Discussion

Our method allows for individual analysis of Affy 6.0 arrays. Since genotyping is independent across samples, this method can be used to analyze a single sample and

SNPs on chromosomes of ploidy larger than two This approach is completely independent of allele frequency and it explicitly models batch effects and cross-hybridization.

It is difficult to type SNPs in genomic regions with low complexity or those that are repetitive [28]. From our hybridization data, we estimate that 13% of Affy 6.0 SNP probes have high levels of cross-hybridization. This number is consistent with the observation that 20%-33% of SNPs fail to be reliably called across different studies [7-10]. Overall, CNV probes have more cross-hybridization than SNP probes (13% vs. 25%). Even after accounting for cross-hybridization there was a significant number of SNPs (12%) that were not in Hardy-Weinberg equilibrium (Figure 1). These SNPs might have cross-hybridization to genomic regions that are not in the genome build that we used for our analysis. Another possibility is that these SNPs are in a CNV. One possible way to correct this problem is to re-run our algorithm for SNPs within CNVs and model the ploidy of the CNV. A third possibility is that our hybridization model does a good job, but not perfect job of modeling hybridization for these probes. For example mononucleotide runs are not well modeled by NN thermodynamics [19].

Our method assigns a QS for each SNP call; this value is the posterior probability that the call is correct (Figure 2 and Tables 1-4). QS values for SNPs with three alleles tend slightly underestimate accuracy We call 84% of SNPs on diploid chromosomes and 57% of SNPs on trisomic chromosome 21 with high confidence, QS > 0.99. This is a total of approximately 600,000 SNP calls per chip with a QS > 0.99. Heterozygotes are more difficult to call than homozygotes. This is especially true for AAB and ABB calls (Table 3 and 4). However, a significant fraction of these heterozygotes are called with high confidence and they can be included in data analysis (Table 4). Some work has been done to type SNPs within CNVs [12]; however, accuracy for these calls can be significantly improved. We envision our analysis being particularly useful for the analysis of SNPs with a low minor allele frequency. Our method can also be used to call genotypes for other types of aneuploidy samples. These include other types of trisomy as well as cancer samples, where aneuploidy is common [29].

We call CNVs by using direct estimates of local DNA concentration for each probe set. These R values can be used in exactly the same way that conventional R values (log 2 probe intensity) are used for CNV typing. We call an average of 50 CNVs per sample. Close to 70% of deletions and duplications are in DGV (Table 5). Inheritance patterns for the CNVs indicate that our method for CNV detection has a substantial false positive and false negative rate. These rates are highest for small CNVs and those not in DGV (Table 6). We can improve upon our method by using the number of probes in a CNV and the GADA amplification value to assign a quality score to each CNV call.

The results of this work help validate our previously published hybridization model for Affymetrix arrays [20]. In it we use basic biochemical principals to exhaustively model hybridization on arrays. In addition we meticulously model the details of the array protocol. In this work we have applied this model and shown that we can use first principles to call genotypes from Affymetrix 6.0 data.

References

1. Stoughton RB: **Applications of DNA microarrays in biology**. *Annu Rev Biochem* 2005, **74**:53-82.
2. **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.
3. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J *et al*: **Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome**. *Science* 1998, **280**(5366):1077-1082.
4. Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA *et al*: **High-throughput variation detection and genotyping using microarrays**. *Genome Res* 2001, **11**(11):1913-1925.
5. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E *et al*: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls**. *Nature* 2010, **464**(7289):713-720.

6. Emanuel BS, Saitta SC: **From microscopes to microarrays: dissecting recurrent chromosomal rearrangements**. *Nat Rev Genet* 2007, **8**(11):869-883.

7. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Magi R *et al*: **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index**. *Nat Genet* 2010, **42**(11):937-948.

8. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, Korenaga M, Hino K, Hige S *et al*: **Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C**. *Nat Genet* 2009, **41**(10):1105-1109.

9. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL *et al*: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1**. *Nat Genet* 2009, **41**(3):324-328.

10. Myocardial Infarction Genetics C, Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ *et al*: **Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants**. *Nat Genet* 2009, **41**(3):334-341.

11. LaFramboise T: **Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances**. *Nucleic Acids Res* 2009, **37**(13):4181-4193.

12. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K *et al*: **Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs**. *Nat Genet* 2008, **40**(10):1253-1260.

13. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G *et al*: **Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays**. *Bioinformatics* 2005, **21**(9):1958-1963.

14. Liu WM, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder TB, Webster TA, Dong S, Liu G *et al*: **Algorithms for large-scale genotyping microarrays**. *Bioinformatics* 2003, **19**(18):2397-2403.

15. Rabbee N, Speed TP: **A genotype calling algorithm for affymetrix SNP arrays**. *Bioinformatics* 2006, **22**(1):7-12.

16. Lin S, Carvalho B, Cutler DJ, Arking DE, Chakravarti A, Irizarry RA: **Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays**. *Genome Biol* 2008, **9**(4):R63.

17. Carvalho B, Bengtsson H, Speed TP, Irizarry RA: **Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data**. *Biostatistics* 2007, **8**(2):485-499.

18. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA: **A multilevel model to address batch effects in copy number estimation using SNP arrays**. *Biostatistics* 2010.

19. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data**. *Nat Rev Genet* 2010, **11**(10):733-739.

20. Jakubek YA, Cutler DJ: **A model of binding on DNA microarrays: understanding the combined effect of probe synthesis failure, cross-hybridization, DNA fragmentation and other experimental details of affymetrix arrays**. *BMC Genomics* 2012, **13**:737.

21. Duan F, Pauley MA, Spindel ER, Zhang L, Norgren RB, Jr.: **Large scale analysis of positional effects of single-base mismatches on microarray gene expression data**. *BioData Min* 2010, **3**(1):2.

22. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP: **Accessing genetic information with high-density DNA arrays**. *Science* 1996, **274**(5287):610-614.

23. Southern EM, Maskos U, Elder JK: **Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models**. *Genomics* 1992, **13**(4):1008-1017.

24. McGall GH, Barone AD, Diggelmann M, Fodor SPA, Gentalen E, Ngo N: **The Efficiency of Light-Directed Synthesis of DNA Arrays on Glass Substrates.** *Journal of the American Chemical Society* 1997, **119**(22):5081-5090.

25. Pirrung MC, Fallon L: **Proofing of photolithographic DNA synthesis methods. Fabrication of DNA microchips**. *Abstr Pap Am Chem S* 1997, **213**:362-ORGN.

26. Held GA, Grinstein G, Tu Y: **Relationship between gene expression and observed intensities in DNA microarrays--a modeling study**. *Nucleic Acids Res* 2006, **34**(9):e70.

27. Skvortsov D, Abdueva D, Curtis C, Schaub B, Tavare S: **Explaining differences in saturation levels for Affymetrix GeneChip arrays**. *Nucleic Acids Res* 2007, **35**(12):4154-4163.

28. Flannick J, Korn JM, Fontanillas P, Grant GB, Banks E, Depristo MA, Altshuler D: **Efficiency and power as a function of sequence coverage, SNP array density, and imputation**. *PLoS computational biology* 2012, **8**(7):e1002604.

29. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* 2013, **45**(10):1113-1120.

Tables

Table 1

Data Summary		
	Counts	**Fraction of Data**
Samples	824	1.00
Trisomy 21 Samples	516	0.63
Diploid 21 Samples	308	0.37
Duplicates	44	0.05
Trisomy 21 Duplicates	25	0.03
Diploid 21 Duplicates	19	0.02
Trios	110	1.00
Trios with paternal origin for trisomy 21	5	0.05
Trios with maternal origin for trisomy 21	105	0.95
Duplicates:	**Fraction Agree**	**Fraction SNPs**
Autosomal* SNPs	0.935	1.00
Autosomal* SNPs QS product > 0.99	0.990	0.75
Trisomy Chr 21 SNPs	0.889	1.00
Chr 21 Trisomy SNPs QS product > 0.99	0.995	0.44
*Excluding chromosome 21		

Table 2

Fraction of Mendelian Consistent SNPs for each Bin and Fraction of Total SNPs in each Bin						
QS	0.75	0.90	0.95	0.96	0.97	0.98
Fraction Mendelian	0.829	0.936	0.966	0.976	0.980	0.985
Fraction of SNPs	0.100	0.061	0.044	0.014	0.019	0.028
QS	0.9900	0.9925	0.9950	0.9975	1.00	
Fraction Mendelian	0.990	0.993	0.994	0.996	0.999	
Fraction of SNPs	0.051	0.023	0.033	0.059	0.568	

Table 3

Duplicate Agreement Heterozygous Calls Trisomy 21					
QS*	**> 0.0**	**> 0.6**	**> 0.7**	**> 0.8**	**> 0.9**
Same Heterozygote	0.789	0.832	0.85	0.867	0.884
Different Heterozygote	0.077	0.059	0.05	0.041	0.033
Heterozygote Homozygote	0.133	0.109	0.1	0.092	0.083
Total Heterozygous Calls	151893	133588	120793	104798	81009
Data is for all heterozygotes. All heterozygous calls in duplicate I are compared to the call made in duplicate II (this call can be the same heterozygote, a different heterozygote, or a homozygote). QS* is for the heterozygote in duplicate I. Then heterozygote calls in duplicate II are compared to the call made in duplicate I. For these data points QS* is for the heterozygote in duplicate II.					

Table 4

<table>
<tr><th colspan="3">Accuracy for Heterozygous Trisomy 21 Calls</th></tr>
<tr><th>QS &gt;</th><th>Accuracy</th><th>Total Heterozygotes</th></tr>
<tr><td>0.0</td><td>0.91237</td><td>63702</td></tr>
<tr><td>0.5</td><td>0.91263</td><td>63476</td></tr>
<tr><td>0.6</td><td>0.92526</td><td>58402</td></tr>
<tr><td>0.7</td><td>0.93711</td><td>52997</td></tr>
<tr><td>0.8</td><td>0.94805</td><td>46603</td></tr>
<tr><td>0.9</td><td>0.95875</td><td>37141</td></tr>
</table>

Table 5

Fraction of Autosomal Deletions and Duplications in the Database of Genomic Variants

	Total Del.	Total Dup.	Frac. Del. in DGV	Frac. Dup. in DGV
Autosomes*	23348	17133	0.687	0.666
Chr 21 diploid	88	145	0.614	0.276
Chr 21 trisomy	210	180	0.348	0.294

*Excluding chromosome 21. Input parameters for GADA -a 0.2 -T 5 -M 6

Table 6

CNV Trio Data				
Deletions	**Not in DGV**		**In DGV**	
Size (kb)	Total	Fraction in Parents	Total	Fraction in Parents
< 20	599	0.033	706	0.263
20 - 50	240	0.075	573	0.410
50 - 100	78	0.167	276	0.406
100 - 200	28	0.107	260	0.446
> 200	45	0.200	224	0.438
Duplications	**Not in DGV**		**In DGV**	
Size (kb)	Total	Fraction in Parents	Total	Fraction in Parents
< 20	289	0.042	301	0.216
20 - 50	145	0.062	324	0.478
50 - 100	35	0.000	143	0.483
100 - 200	16	0.188	175	0.589
> 200	172	0.267	290	0.648
Data is for CNVs on autosomal chromosomes of children in trios (excluding chromosome 21). Database of Genomic Variants (DGV).				

Figures

Figure 1

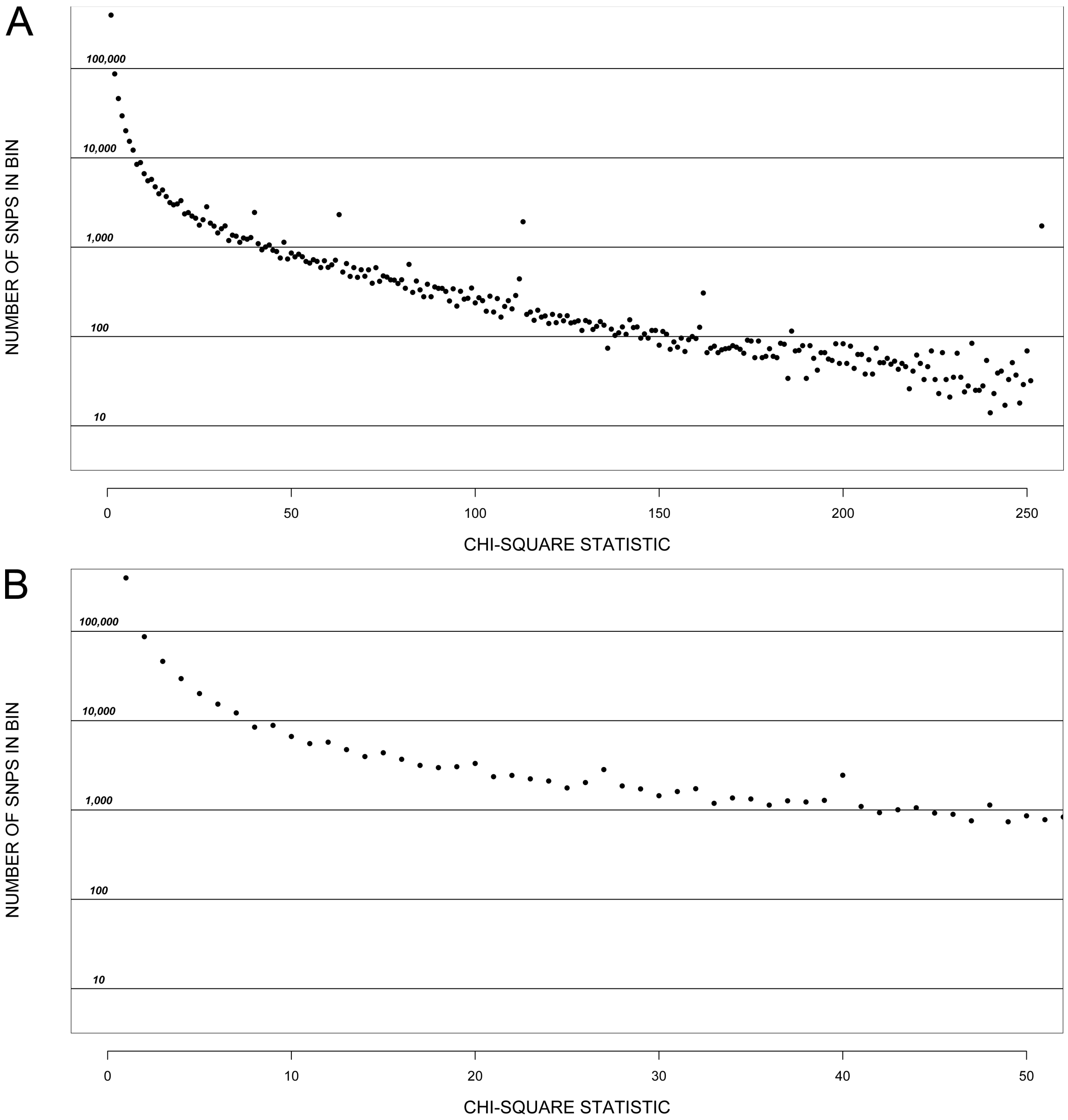


Figure 1: Distribution of Chi-Square Values for Hardy-Weinberg Equilibrium. For autosomal SNPs we calculated allele frequencies using data from the parents (308 individuals) and then used them to test for Hardy-Weinberg equilibrium. A)

Data shown for all autosomal SNPs. B) Data for autosomal SNPs with a chi-square value less than 53.

Figure 2:

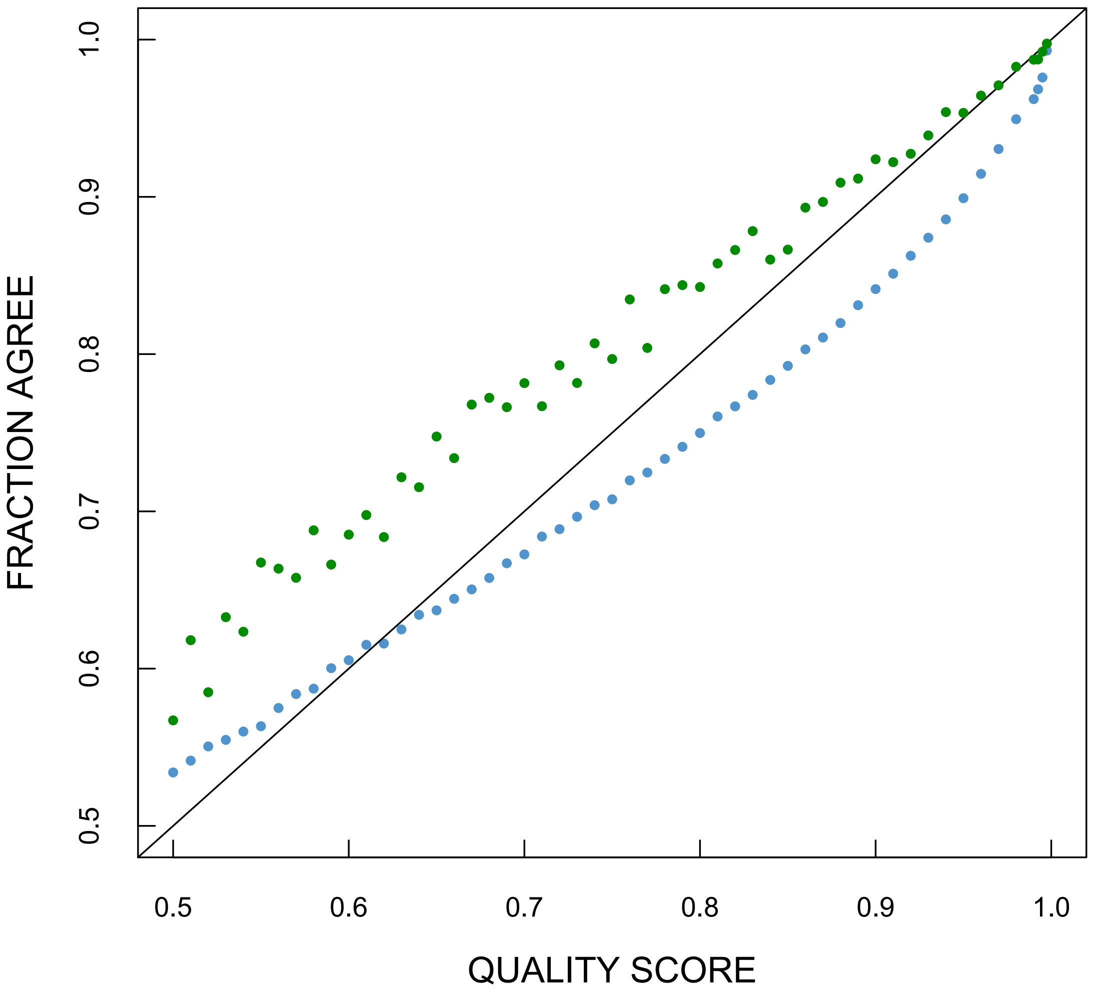


Figure 2: Duplicate Agreement. For SNPs in duplicated samples we plot the product of the QS values of the SNP calls on the x-axis and the fraction of times the calls agree on the y-axis. Green dots are for SNPs on chromosome 21 of duplicated Down syndrome samples. Blue dots are for SNPs on autosomal chromosomes (excluding chromosome 21) for all duplicates.

Figure 3:

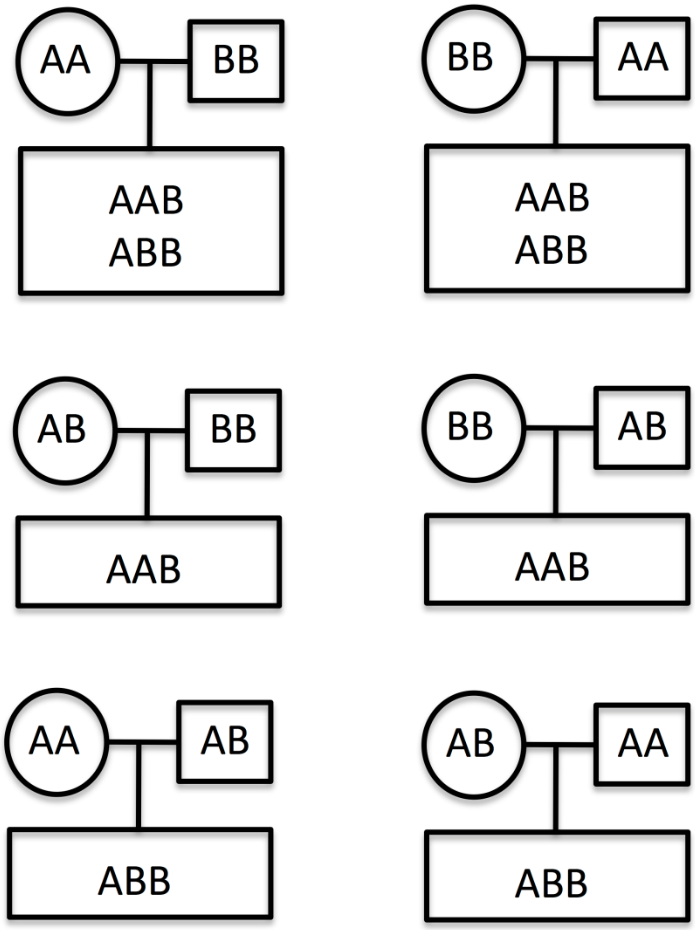


Figure 3: Genotype Combinations with Power. There are eight genotype combinations for which there is power to detect genotype errors for trisomy 21 heterozygous calls.

Chapter 4:

Conclusions

Our model for hybridization on arrays and its application to SNP and CNV detection has helped us gain a better understanding of microarray behavior. Our data shows that nearest-neighbor thermodynamics and the Langmuir isotherm provide a good approximation for hybridization on the array (Chapter 2, Table 2 and Figure 5). Other data that supports our model includes the difference in hybridization temperature between the data from chapter 2 and chapter 3. In our model we do not directly fit a parameter for the hybridization temperature. We assume that heating blocks for chips maintain the temperature that is specified in the microarray processing protocol. For the data in chapter one the temperature parameter was set to 42°C. For the chapter 3 data it was set to 49°C. For both sets of data the correlations between expected and observed probes intensities were high, which suggests that our model accurately models hybridization temperature.

Overall our approach has helped us understand the differences between binding in solution and on the array. Probe synthesis errors can have a significant effect on hybridization (Chapter 2, Table 3, Table 4 and Supplementary Figure 1). Differences in the error rates of A, C, G, and T bases along the probe sequence provide the most parsimonious explanation for differences between forward and reverse probes. Binding is also affected by target fragmentation, which is the reason why mismatches have a larger destabilizing effect when they are placed towards the center of the probe (Chapter 2, Figure 6). Other factors that also affect binding are salt concentration, target concentration, wash stringency, and scanner settings (Chapter2, Table 2).

The results presented here help us understand some aspects of Affymetrix probe design. The first generation of arrays had complementary forward and reverse probes [1]. However, more recent arrays tend have only the forward or the reverse probe. This is not surprising given the data that we present in Chapter 2, Supplementary Figure 1. Another aspect of probe design supported by our model is the placement of the query base towards the center of the probe. Data from Chapter 2, Figure 6 suggests that mismatches on positions 11 to 22 on a 25 nucleotide long probe will have the largest destabilizing effect on binding.

Perhaps the most striking validation for our model is the fact that we can use it to call SNPs and CNVs accurately (Chapter 3). Overall, our approach is not as practical as current statistical methods for SNP analysis. We call approximately 600,000 SNPs with high QS values; this is on the lower end compared to data from other studies [2-5]. However, our approach has three major advantages, it can call SNPs on chromosomes of any ploidy, it explicitly models batch effects, and SNP calls are not affected by allele frequency. One possible application of our method is to re-analyze genome-side association studies. A significant problem with these studies is a lack of reproducibility. It is possible that our method might remove spurious associations caused by batch effects or possibly cross-hybridization. The accuracy of our approach for CNV detection is comparable to the accuracy of current methods. It could be improved upon by assigning quality scores to the calls and by using more than one break-point detection algorithm for the analysis of the R values generated by our method. Another way to remove false positives is to call CNVs using both our approach and using the standard R values, log 2 probe intensity, and then searching for calls made by both methods. This approach appears to work well for large CNVs, as was shown in Chapter 3.

Accurate and cost-effective genetic variation detection is important across different fields of biological research [6]. There is no method that is both highly accurate and cost-effective for the analysis of all possible variants in the human genome. The most commonly used technologies for genome-wide variant detection are genotyping arrays and next-generation sequencing (NGS) [7]. Microarrays are best suited for the analysis of common SNPs and CNVs. On the other hand NGS is more expensive; however, it can detect both common and rare variation. A recent study analyzed agreement for SNP calls between Affy 6.0 and NGS [7]. Array errors were more common for SNPs near repetitive DNA while NGS errors were more common where there is a strand bias during sequencing [7]. Overall, CNV detection is less accurate than SNP detection. This is true for both genotyping arrays and NGS data. Both technologies do a poor job identifying small CNVs. One way to improve accuracy in variant detection is to use genotyping array data in combination with NGS [8]. This joint approach has two benefits. It can improve overall call accuracy and it can be used to improve quality control filters. In addition our approach has potential clinical applications for variant detection in polyploidy cells such as Down syndrome as well as polyploidy chromosomes from somatic cells such as those commonly found in cancer.

References

1. Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA *et al*: **High-throughput variation detection and genotyping using microarrays**. *Genome Res* 2001, **11**(11):1913-1925.
2. Myocardial Infarction Genetics C, Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ *et al*: **Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants**. *Nat Genet* 2009, **41**(3):334-341.
3. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Magi R *et al*: **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index**. *Nat Genet* 2010, **42**(11):937-948.
4. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, Korenaga M, Hino K, Hige S *et al*: **Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C**. *Nat Genet* 2009, **41**(10):1105-1109.

5. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL *et al*: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1**. *Nat Genet* 2009, **41**(3):324-328.

6. Ragoussis J: **Genotyping technologies for genetic research**. *Annual review of genomics and human genetics* 2009, **10**:117-133.

7. Flannick J, Korn JM, Fontanillas P, Grant GB, Banks E, Depristo MA, Altshuler D: **Efficiency and power as a function of sequence coverage, SNP array density, and imputation**. *PLoS computational biology* 2012, **8**(7):e1002604.

8. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* 2013, **45**(10):1113-1120.