UNIVERSITÄT LEIPZIG
Fakultät für Mathematik und Informatik
Institut für Informatik

Sequence specific probe signals on SNP microarrays

Diplomarbeit

Vorgelegt von:
Glomb, Torsten (geb. am 11.02.1981)
Studiengang Informatik
Studienrichtung Bioinformatik

Betreuer:
PD Dr. rer. nat. habil. Hans Binder
Interdisziplinäres Zentrum für Bioinformatik (IZBI)
Universität Leipzig

Leipzig, 09. Mai 2010

Abstract

Single nucleotide polymorphism (SNP) arrays are important tools widely used for genotyping and copy number estimation. This technology utilizes the specific affinity of fragmented DNA for binding to surface-attached oligonucleotide DNA probes. This thesis contemplates the variability of the probe signals of Affymetrix GeneChip SNP arrays as a function of the probe sequence to identify relevant sequence motifs which potentially cause systematic biases of genotyping and copy number estimates.

The probe design of GeneChip SNP arrays affords the identification of different sources of intensity modulations such as the number of mismatches per duplex, perfect match and mismatch base pairings including nearest neighbors and base triples and their position along the probe sequence. Probe sequence effects are estimated in terms of triple motifs with central matches and mismatches including all combinations of possible base pairings. The probe/target interactions on the chip can be decomposed into nearest neighbor contributions which correlate well with free energy terms of DNA/DNA-interactions in solution. The effect of mismatches is about twice as large as that of canonical pairings. Runs of guanines (G) and the particular type of mismatch pairings formed in cross-allelic probe/target duplexes constitute sources of systematic biases of the probe signals with consequences for genotyping and copy number estimates. The poly-G effect seems to be related to the crowded arrangement of probes which facilitates complex formation of neighboring probes with at least three adjacent G's in their sequence.

The applied method of "triple averaging" represents a model-free approach to estimate the mean intensity contributions of different sequence motifs which can be applied in calibration algorithms to correct signal values for sequence effects. Rules for appropriate corrections of the probe intensities are suggested.

Acknowledgments

Ich möchte mich bei allen bedanken, die mich bei der Durchführung meiner Diplomarbeit unterstützt haben und mir geholfen haben diese abzuschließen. Besonderer Dank geht an Herrn Dr. Binder vom IZBI für die Möglichkeit das Thema bearbeiten zu können, seine Geduld, seine ausgezeichnete Betreuung und seine hervorragende Unterstützung, die Diplomarbeit zu Ende zu führen. Außerdem danke ich Herrn Dr. Hasenclever vom IMISE und Herrn Dr. Ahnert vom BBZ, die mir zu Beginn der Diplomarbeit geholfen haben, das Thema auf den richtigen Weg zu bringen.

Von ganzem Herzen geht mein Dank an meine Familie, die mich die ganze Zeit hinweg enorm unterstützt hat und nicht aufgehört hat an mich zu glauben. Danke!

Contents

Chapter 1
Introduction

About 90% of all human genetic variation is made up by single nucleotide polymorphisms (SNPs) [1]. Usually, they are not causal for defective genes, therefore rarely the cause of diseases but as SNPs are often linked to genetic regions or genes (normal or defective) and are inherited along with them, they can be used as markers for these regions or genes. This has driven the parallel developments of dense SNP marker maps and technologies for high-throughput SNP genotyping which are used in genetic association studies to understand the background of different phenotypes, such as disease risk or variable drug response [2]. Several technologies are available for SNP genotyping such as primer extension (e.g. MALDI-TOF), amplification with allele specific primers (e.g. TaqMan™) or hybridization of PCR products to microarrays (e.g. Affymetrix SNP arrays). Sham *et al.* [3] gives an comparative overview.

The genotyping platform provided by Affymetrix interrogates hundreds of thousands of biallelic human SNPs on a single microarray. This technology utilizes the specific affinity of fragmented DNA to build bimolecular complexes with surface-attached oligonucleotide probes of complementary sequence and subsequent optical detection of the bound fragments using fluorescent markers. Two allele-specific sets of probes interrogate each of the two SNP alternatives to determine genotype and the copy number. Ideally, the probe intensities are directly related to the abundance of the respective allele. The ratio and sum of the two allele specific signals then simply provide naive measures of the genotype and the copy number, respectively. In reality, the measured probe signals however strongly depend on the sequence context of each SNP which is given by the particular probe sequence of 25 nucleotides and their interactions with the target fragments. Moreover, sequence similarity of both allele-specific probes gives rise to cross-allelic hybridization causing mutual correlations of the intensities in a SNP-specific fashion. The naive analysis is therefore highly inaccurate for most of the probes due to the sequence bias of the signals.

A number of calibration methods have been developed to transform biased probe-level signals into reliable genotyping and copy number information (→ e.g. [2,4-13]). These preprocessing algorithms are however not without limitations due to insufficient corrections that have implications for downstream analyses (→ e.g. [6] for a critical overview). Given the vast number of genotypes being produced, a systematic bias, even if very small, may lead to spurious association signals [14].

The successful correction of raw probe signals for parasitic effects requires identification and understanding of the main sources of signal variation. The main purpose of this thesis is to analyze the variability of the probe signals as a function of the probe sequence, to identify relevant sequence motifs which significantly modulate the probe signals and to quantify their effect in the context of genotyping and copy

number estimation. This issue has a high impact on signal correction because the identified sequence motifs constitute potential building blocks for improved calibration models.

The presented approach is important also in a more general context: DNA/DNA duplex formation is the basic mechanism that is used not only on SNP arrays but also on other array types such as resequencing [15] and expression arrays (gene or exon-related and whole genome tiling arrays) of newer generations. It has been demonstrated that thermodynamic models for hybridization, which take the sequence dependent probe affinities into account are capable to significantly reduce the signal fluctuation between probes interrogating the same target [16-19].

Knowledge of the underlying physical process is however still lacking in many details despite the recent progress in this field (→ e.g. [20-25]). Particularly, surface hybridization differs in many respects from oligonucleotide duplexing in solution (→ e.g. [26-28]). Systematic studies on oligonucleotide interactions on microrrays are therefore required to tackle selected problems such as signal anomalies of poly-guanine runs [29,30], the specific effect of mismatched base pairings [17,28,31] and/or the positional dependence of interaction strengths [22,32].

This approach takes advantage of the probe design used on GeneChip SNP arrays and of the target composition of fractionated genomic DNA hybridized on the arrays which enable us to deduce the base pairings in the probe/target complexes producing a particular probe intensity. Making use of the hundreds of thousand signal values per SNP array it allows to extract specific intensity contributions of selected short sequence motifs of 2-4 adjacent nucleotides by appropriate averaging. The obtained motif-specific intensity contributions characterize the stability of the involved base pairings which include all relevant combinations of canonical Watson-Crick and mismatch pairings. Finally, the systematic analysis of different motifs such as triples of adjacent bases XBY, X,B,Y∈{A,C,G,T}, where B can form canonical or mismatch pairings and X and Y refer to the neighboring WC pairs, allows to identify those which account for significant signal variations.

Previously an analogous chip study using intensity data of expression arrays to characterize base pair interactions in DNA/RNA hybrid duplexes was performed [33] and improved algorithms for signal calibration and quality control [18,19] were developed. Note that, compared with expression arrays, SNP arrays are better suited to study base pair interactions because probe/target duplexes are typically less contaminated with nonspecific target fragments of unknown sequence and because genomic copy numbers are less variable than mRNA transcript concentrations.

My thesis is laid out as follows: Chapter 2 presents probe and sequence characteristics and, particularly, explains the used classification criteria to assign the probe intensities to different interaction modes. In Chapter 3, I analyze different factors, in terms of triple averages, which affect the probe intensities such as the optical and nonspecific background, the number of mismatches and their positional dependence along the sequence, signal contributions due to different base triples and their symmetry relations, as well as single and tandem mismatches. In addition, I decompose the triple terms into nearest neighbor terms and compare the results with thermodynamic nearest neighbor parameters characterizing DNA/DNA interactions in solution. In Chapter 4 I discuss the stability of different mismatches, discover the

possible origin of the “poly-G” effect and illustrate and estimate the systematic genotyping and copy number errors in a SNP-specific fashion. Finally, I suggest rules for selecting appropriate sequence motif to adequately correct the probe signals for sequence effects which might serve as the basic modules of improved calibration methods.

Chapter 2
Specifics of SNP microarray data

This chapter explains how the SNPs are represented on the array and considers the hybridization process also in terms of thermodynamic relations. Probes and base pairings are categorized according to their characteristics and hybridization behavior to develop a probe selection scheme for a proper analysis of the SNP microarray data.

2.1 SNP data

Sample intensity data of the Affymetrix GeneChip Mapping 100K set and supplementary files were downloaded from the supplier's website[1]. This sample data set was specially designed for the development and evaluation of low-level analysis methods for genotyping and copy number estimation from probe intensity data (→ e.g. [7]). From the available data the intensity data of sample NA06985_Xba_B5_4000090 coming from the Mapping 100K HapMap Trio Dataset was used, as well as library and annotation information (probe sequences, fragment lengths and GC-content of the targets, GCOS genotype calls).

Note: The library information can not be used without limitations in the context of low-level analyses as they contain erroneous data (→ Appendix A).

2.2 Probe design for SNP analysis

SNP arrays are intended to determine genotype and copy number of hundreds of thousands biallelic SNP loci in one measurement. I specify each SNP by its two alternative nucleotides in the sense DNA strand of allele A and allele B using the convention B_A/B_B. $B_A/B_B \in \{A/C,A/G,A/T,C/G,C/T,G/T\}$ represents the six SNP types considered on GeneChip microarrays. These SNP types are either complementary (cSNP: A/T, C/G) in terms of Watson-Crick (WC) base pairings or non-complementary (ncSNP) otherwise.

On Affymetrix GeneChip Mapping 100K set arrays, each allele is interrogated by ten perfect match (PM) probes, the 25meric oligonucleotides that perfectly match the genomic target sequence on its sense or antisense DNA strand (→ Figure 1 for illustration). Three to seven PM probes refer to the sense strand, thus the remaining seven to three probes refer to the antisense strand. The SNP position of each probe is shifted by different offset values relative to the center interrogation position, $\delta \in \{-4,\ldots,0,\ldots,+4\}$. Ten probes with different offsets and target strandedness are realized to probe each allele.

[1] http://www.affymetrix.com/support/technical/byproduct.affx?product=100k

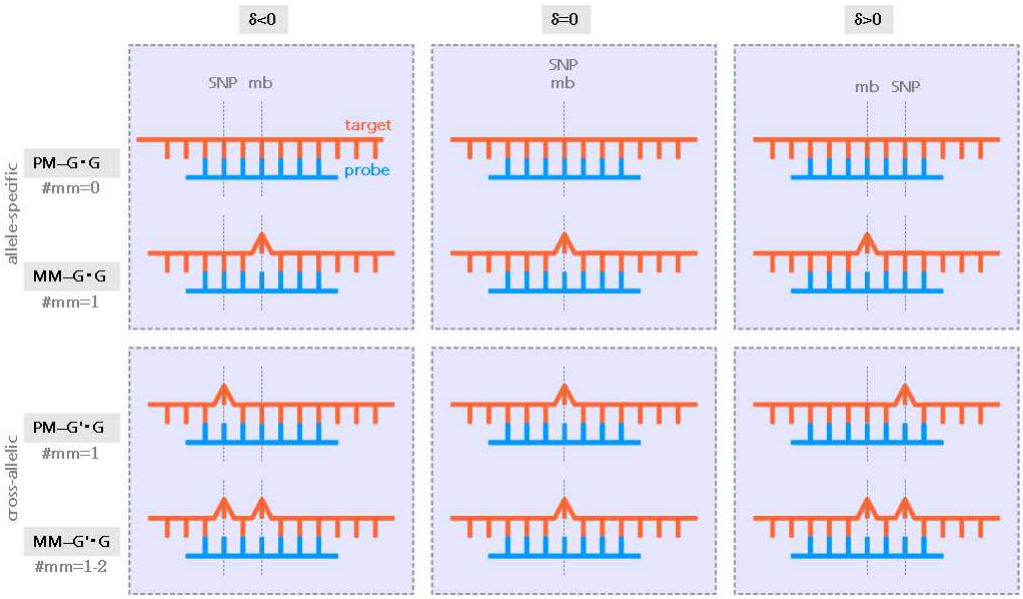


Figure 1: Probe design and hybridization modes for SNP detection. *Each column illustrates a probe quartet which consists of two PM/MM probe pairs interrogating either targets of allele G=A (or B) or targets of allele G'=B (or A). Only allele G is assumed to be present in the genomic target. It hybridizes to the probes of both allele sets forming either specific or cross-allelic duplexes, respectively (→ see also the reaction scheme 11). The three selected probe quartets differ in the offset value δ of the SNP position relatively to the middle base (mb) of the probe. Mismatch pairings are indicated by the bulges. Their number varies between #mm=0 and #mm=2 in dependance on the probe type, hybridization mode and offset position. A SNP is interrogated by probe sets using ten offset positions providing thus 10 probe quartets.*

Each PM probe has one corresponding mismatch (MM) probe of identical sequence except the modified middle base. This modification is intended to drastically reduce specific binding of the respective target to the MM probe to estimate the contribution of nonspecific background hybridization to the respective PM probe intensity. As standard, the middle base is substituted by its WC complement, i.e. A↔T, C↔G.

Unlike the standard substitution of the middle base in MM probes, I figured out that Affymetrix modifies the middle base in MM probes of cSNPs with offset δ=0 according to the non-complementary replacements A↔G and T↔C to avoid cross-allelic binding of target sequences of the alternative allele to the proper MM (→ see below).

Hence, each SNP is interrogated by a set of 20 PM/MM probe pairs, i.e. in total 40 probes. They split into two subsets of 10 probe pairs for each allele which I will term 'allele set'. The allele sets yield to redundant information by forming pseudo-replicates using probes for both strand directions as well as slight differences in the probe sequence to increase the accuracy of genotyping and copy number estimation.

Both allele sets of one SNP have the same offset values, thus, each particular offset δ is probed by one probe pair for each allele. These two probe pairs, i.e. four probes, build up the so-called probe quartet (→ Figure 1).

2.3 Hybridization modes

During the hybridization process DNA fragments attach to the probes of a given SNP array. Basically three hybridization modes (h) can be observed considering a SNP of a heterozygous genotype, i.e. both alleles A and B of a particular SNP are present in the hybridization solution. These three modes can be described by three coupled reactions for each probe,

$$\begin{array}{lcll} P-G+G & \Leftrightarrow & P-G\cdot G & \text{allele-specific} \\ P-G+G' & \Leftrightarrow & P-G\cdot G' & \text{cross-allelic} \\ P-G+N & \Leftrightarrow & P-G\cdot N & \text{nonspecific} \end{array} , \qquad (1)$$

where P−G denotes the probes P∈{PM,MM} which are designed to interrogate targets of allele G∈{A,B}. G'∈{B,A} designates targets of the alternative allele in respect to G.

In the allele-specific mode (S-mode) probes form duplexes with the intended targets of the type P−A•A and P−B•B, respectively. Targets of allele A and B differ in only one base at the SNP site, so that probes also bind targets of the alternative allele of the type P−A•B and P−B•A, respectively. This is called the cross-allelic hybridization mode (C-mode). The nonspecific hybridization mode (N-mode) includes all the remaining hybridization results of probes with genomic fragments not referring to a specific SNP. The results are of the type P−A•N and P−B•N, where N subsumes all nonspecific fragments with non-zero affinity to the particular probe.

In the S-mode a PM probe completely matches the target sequence whereas in the C-mode it mismatches the target at the SNP position. The respective MM probe mismatches the target in the S-mode at the middle position. In the C-mode it mismatches the target either only at the middle position (δ=0) or at both the middle and the SNP position (δ≠0) (→ Figure 1). The respective base pairings are specified below.

2.4 Thermodynamic considerations

The hybridization reaction of each probe of a heterozygous SNP is described by the three coupled equations shown in Eq. 1. The measured intensity I^{P-G} obtained from each probe interrogating a particular allele G∈{A,B} represents a superposition of contributions due to the allele-specific (S), cross-allelic (C) and the nonspecific (N) hybridization modes. In addition, an optical background intensity I^{O} caused by the dark signal of the scanner and by residual fluorescent marker not attached to target fragments contributes to the measured intensity,

$$I^{P-G} = I^{P-G\cdot G} + I^{P-G\cdot G'} + I^{P-G\cdot N} + I^{O} . \qquad (2)$$

The nonspecific and optical background contributions are, on the average, independent of the probe type, i.e. $I^{PM-G\cdot N} \approx I^{MM-G\cdot N} \approx I^{N}$. Both contributions are combined into one mean background intensity

$$I^{BG} = I^{N} + I^{O} . \qquad (3)$$

Its fraction and the fraction of nonspecific hybridization,

$$x^{P-G,BG}=\frac{I^{BG}}{I^{P-G}} \text{ and } x^{P-G\cdot N}=\frac{I^{N}}{I^{P-G}-I^{O}}, \tag{4}$$

define the percentage of background intensity in the total signal and the percentage of nonspecific hybridization signal in the total signal after correction for the optical background, respectively.

The S- and C-hybridization modes refer to probe/target duplexes of the type P–G•G=P–A•A, P–B•B and P–G•G'=P–A•B, P–B•A, respectively. The intensity contributions are directly related to the fraction of probe oligomers occupied by targets T of one allele, the so-called partial occupancy of the probe [26],

$$I^{P-G\cdot T}\approx M\cdot\Theta^{P-G\cdot T} \text{ with } \Theta^{P-G\cdot T}\equiv\frac{[P-G\cdot T]}{[P-G]} \text{ and } T\in\{A,B,N\}. \tag{5}$$

The squared brackets denote the concentrations of dimerized, [P–G•T], and of total, [P–G]=[P–G•A]+[P–G•B]+[P–G•N], probe oligomers. M is the proportionality constant which transforms the dimensionless occupancy into intensity units. It has the meaning of the maximum intensity observed if all probe oligomers are dimerized with the respective targets. The partial occupancy is given to a good approximation by the hyperbolic function of the so-called binding strength $[X^{P-G\cdot T}]$

$$\Theta^{P-G\cdot T}\equiv\frac{X^{P-G\cdot T}}{1+X^{P-G}}, \tag{6}$$

where

$$X^{P-G\cdot T}=K^{P-G\cdot T}\cdot[P-G\cdot T] \text{ and } X^{P-G}=X^{P-G\cdot A}+X^{P-G\cdot B}+X^{P-G\cdot N} \tag{7}$$

are the partial and the total binding strengths of the hybridization. $[K^{P-G\cdot T}]$ is the binding constant of the respective reaction in Eq. 1 characterizing the association of the targets T to the probes interrogating allele G.

Eq. 6 transforms far from saturation ($[X^{P-G}\ll 1]$) into the linear approximation

$$\Theta^{P-G\cdot T}\approx X^{P-G\cdot T} \tag{8}$$

The partial probe occupancy and thus also the respective intensity component (Eq. 5) are directly related to the total target concentration of genomic copies of the respective allele in the hybridization solution according to Eqs. 8 and 7. Furthermore, the proportionality constant M is given by the respective binding constant. Thus, the respective intensity component can be assumed as

$$I^{P-G\cdot T}\propto K^{P-G\cdot T}\cdot[P-G\cdot T]=X^{P-G\cdot T}. \tag{9}$$

In consequence, the intensity contributions are directly related to the respective number of probe/target duplexes in a first order approximation,

$$I^{P-G\cdot T}\propto[P-G\cdot T], \tag{10}$$

i.e. $I^{P-G\cdot A}\sim[A]\sim CN_A$ and $I^{P-G\cdot B}\sim[B]\sim CN_B$. $K^{P-G\cdot T}$ varies from probe to probe in a sequence specific fashion and from SNP type to SNP type depending on the particular substitution of the allele bases. The intensity measures are consequently biased by these allele specific factors.

2.5 Homozygous present and homozygous absent probes

In case of heterozygous genotypes three different target types compete for duplex formation (Eq. 1). In the special case of homozygous genotypes only targets of one allele are present in the hybridization solution. Therefore, the types of competing targets reduce to two ones, namely nonspecific and either allele-specific or cross-allelic targets. Targets of the present allele hybridize specifically to their respective probes (homozygous present probes) or in cross-allelic mode to the probes interrogating the alternative allele (homozygous absent probes), i.e.

$$\begin{array}{lcll} P-G+G & \Leftrightarrow & P-G \cdot G & \text{homozygous present (hp)} \\ P-G'+G & \Leftrightarrow & P-G' \cdot G & \text{homozygous absent (ha)} \end{array} . \qquad (11)$$

Hence, Eq. 2 simplifies regarding homozygous present probes with $I^{P-G \cdot G'}=0$ or regarding homozygous absent probes with $I^{P-G \cdot G}=0$.

2.6 Base pairings in probe/target duplexes

Considering the allele-specific and the cross-allelic hybridization modes only base pairings at two sequence positions need to be taken into account, namely the pairings at the middle base and the SNP position of the probe. The remaining base pairings are consistently WC base pairs. The SNP position is shifted by the offset value δ with respect to the middle base. For δ=0 the SNP and the middle base position are identical.

In the S-mode the PM probes form WC base pairs with the respective target throughout the whole probe sequence including the two positions of interest, i.e. perfect matching of probe and target (→ Figure 1 and Table 1). However, in the C-mode a mismatching base pair is introduced at the SNP position. MM probe/target duplexes always contain a mismatch base pair at the middle position. Thus, upon C-hybridization, a MM probe contains two mismatches, one at the middle position

Table 1: Hybridization modes, probe attributes and interaction groups.

Hybridization mode	Probe attributes			Interaction groups				
				Ab-group[2]				no. of
	probe type	SNP offset δ	base position[1]	At	Aa	Ag	Ac	mismatches #mm[3]
allele-specific P–G•G	PM	all	mb/SNP	✘				0
	MM	=0	mb/SNP		✘		✘	1
		≠0	mb		✘			
		≠0	SNP	✘				
cross-allelic P–G•G'	PM	=0	mb/SNP		✘	✘	✘	
		≠0	mb	✘				
		≠0	SNP		✘	✘	✘	
	MM	=0	mb/SNP			✘	✘	
		≠0	mb		✘			2
		≠0	SNP		✘	✘	✘	

[1] Base pairings formed at the center position of the 25meric probe sequence (mb) or at the SNP position (SNP) which is shifted by δ base positions relatively to the center position. The mb and SNP positions are consequently identical for δ=0.

[2] Base pairings are classified into four Ab-groups as follows: At-group (At, Ta, Gc, Cg); Aa-group (Aa, Tt, Gg, Cc); Ag-group (Ag, Tc, Ga, Ct); Ac-group (Ac, Tg, Gt, Ca). Lower case letters refer to the target.

[3] Number of mismatches per probe/target duplex

and the other at the SNP interrogating position. An exception are MM probes with δ=0 because middle and SNP position are the same. A special case exists for MM probes with δ=±1 referring to so-called tandem mismatches of two adjacent mismatched base pairs. For |δ|>1 the two mismatches are separated by at least one WC base pairing.

The duplex formation of the two probe types P∈{PM,MM} with targets coming from the six biallelic SNP types B_A/B_B upon specific or cross-allelic hybridization enables an all-embracing analysis of the full set of 16 possible DNA/DNA base pairings occurring on SNP microarrays (→ Table 1 and Table 2). The pairings are classified into canonical Watson-Crick pairs (referred to as At-group; upper and lower case letter refer to the probe and target sequences, respectively) and three groups of mismatch pairings (Aa, Ag and Ac-group). The notation of the interaction groups Ab is chosen in agreement with the respective pairing formed by an adenine of the probe sequence (→ Table 2). The mismatch groups contain self-complementary (At-group: Aa, Tt, Gg, Cc), self-paired (Ag-group: Ag, Tc, Ga, Ct) and cross-paired (Ac-group: Ac, Tg, Gt, Ca) pyrimidines and purines, respectively. These groups are invariant with respect to the target strandedness because complementary substitutions do not change the group membership.

The probe/target duplexes formed in the nonspecific hybridization mode are not specified by means of number and type of the mismatches. Nevertheless, the

Table 2: Base pairings in probe/target duplexes at the middle and SNP position of the probe sequence[1].

Position	SNP offset δ	SNP type B_A/B_B	PM base B	Base pairing Bb				Probes[2]	
				S-mode (P–G•G)		C-mode (P–G•G')		number	percent
				PM	MM	PM	MM		
k=13 (mb)	δ≠0		T	Ta	Aa	Ta	Aa	52.940	26,2
			A	At	Tt	At	Tt	52.800	26,2
			C	Cg	Gg	Cg	Gg	33.008	16,3
			G	Gc	Cc	Gc	Cc	33.627	16,7
							total	172.375	85,4
k=13+δ (SNP)	δ≠0	A/C	T/G	Ta/Gc		Tc/Ga		14.683	7,3
		G/T	C/A	Cg/At		Ct/Ag		11.224	5,6
		A/G	T/C	Ta/Cg		Tg/Ca		60.547	30,0
		C/T	G/A	Gc/At		Gt/Ac		60.585	30,0
		A/T	T/A	Ta/At		Tt/Aa		9.273	4,6
		C/G	G/C	Gc/Cg		Gg/Cc		16.063	8,0
							total	172.375	85,4
k=13+δ (mb/SNP)	δ=0	A/C	T/G	Ta/Gc	Aa/Cc	Tc/Ga	Ac/Ca	2.537	1,3
		G/T	C/A	Cg/At	Gg/Tt	Ct/Ag	Gt/Tg	1.956	1,0
		A/G	T/C	Ta/Cg	Aa/Gg	Tg/Ca	Ag/Ga	10.393	5,1
		C/T	G/A	Gc/At	Cc/Tt	Gt/Ac	Ct/Tc	10.265	5,1
		A/T	T/A	Ta/At	Ca/Gt	Tt/Aa	Ct/Ga	1.597	0,8
		C/G	G/C	Gc/Cg	Ac/Tg	Gg/Cc	Ag/Tc	2.777	1,4
							total	29.525	14,6

[1] Interaction groups are highlighted as follows: At, Aa, Ag, Ac. Base pairings are given for the sense strand only. Pairings of the antisense strand can be obtained by the WC complement (At-group) or the bond-reversal (mismatch groups) of the base B.

[2] Only probes referring to homozygous SNP loci are selected (41,629 out of a total of 58,960 loci, ~70.1%) and used in further analysis. Note that the probes with δ≠0 (85.4% of all used probes) are used twice, considering the sequence motifs about the middle base (k=13) and about the SNP base (k=13+δ). The remaining 14.6% of probes refer to δ=0. The probes with offset δ≠0 split into 27.6% (55,634) with |δ|=1; 14.2% (28,742) with |δ|=2; 14.0% (28,355) with |δ|=3 and 29.5% (59,644) with |δ|=4.

sequence effect can be described in terms of the properties of canonical WC pairings [33,34]. This seems to contradict the fact that nonspecific duplexes are by definition destabilized at minimum by one, but typically by more mismatch pairings. On the other hand, these mismatch effects are averaged out by calculating mean binding characteristics of WC interactions (At-group) which stabilize the nonspecific duplexes.

2.7 Interaction modes

The previous sections point out that probe/target duplexes are characterized by a series of probe attributes:

- probe type $P \in \{PM,MM\}$,
- probe sequence,
- middle base $B_{13} \in \{A,T,G,C\}$,
- strand direction $d \in \{s,as\}$,
- SNP type (B_A/B_B),
- SNP offset $\delta \in \{-4,\ldots,0,\ldots,+4\}$ and
- hybridization mode $h \in \{S,C,N\}$.

Each particular combination of the hybridization mode with a set of probe attributes unambiguously determines the interaction mode between probe and target. It is characterized by

(i) the base pairing at the SNP and the middle position, which includes all 16 pairwise combinations of nucleotides, where four of them form WC pairings whereas the remaining 12 form mismatches;

(ii) WC pairings at the remaining positions of the probe sequence;

(iii) the mutual shift δ between the middle base and the SNP interrogating position by up to four bases in both directions;

(iv) varying numbers of mismatches #mm per duplex going from #mm=0 (for P=PM and h=S) to #mm=2 (P=MM, h=C, $\delta \neq 0$);

(v) varying relative positions of paired mismatches (#mm=2) which are either separated by at least one WC base pair ($|\delta|>1$) or form tandem mismatches ($|\delta|=1$).

The interaction modes directly affect the probe intensities. Vice versa, the probe intensities are related to the amount of bound DNA target fragments which, in turn, depends on the stability of the duplexes and thus on the binding constant of the respective interaction mode. Knowing both interaction mode and binding constant then allows to compute the genotype call and copy number of a given SNP.

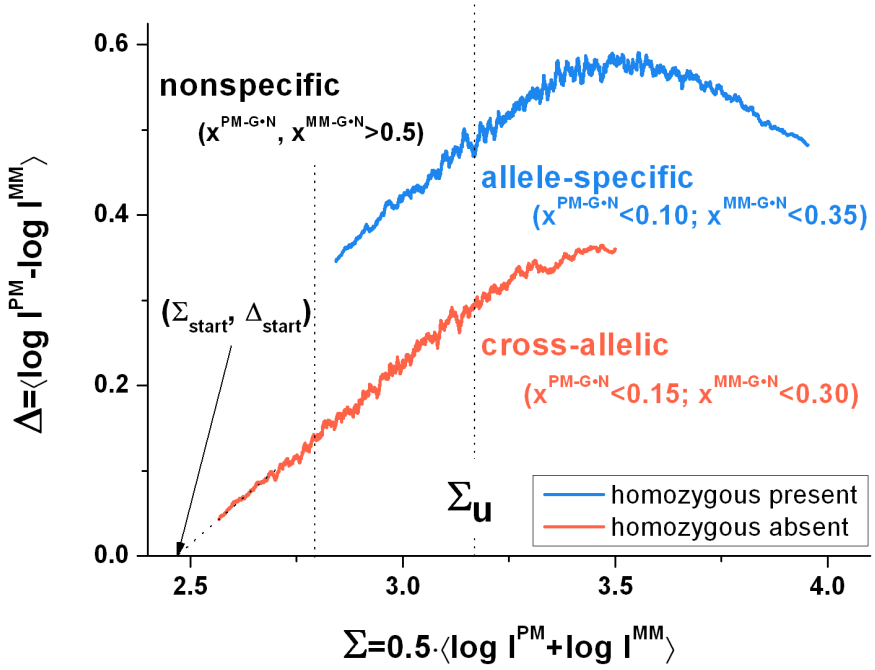


Figure 2: Classification of probe intensities according to their hybridization mode. *So-called hook curves are plotted for homozygous absent (ha) and homozygous present (hp) probes. The 'start' coordinates of the hook curve are given by the intersection of the extrapolated ha hook with the abscissa. The intensity fraction per probe due to nonspecific binding depends on the hook coordinates (→ Eq. 12). The right vertical line refers to* $(\Sigma\text{-}\Sigma_{start})=0.7$. *It is used as threshold for probe selection to characterize the interaction modes upon allele-specific (S) and cross-allelic (C) hybridization. Above this threshold, probe intensities are distorted, on the average, by a contribution of nonspecific hybridization of less than 20%. The fraction of nonspecific binding slightly differs between the PM and MM probes as indicated in the figure.*

2.8 Probe selection for intensity analysis

First of all, I use the genotype call information provided by Affymetrix for the analyzed sample array to select only probes of homozygous SNPs, i.e. 41,629 homozygous out of a total of 58,960 SNPs (~70.1%). The advantage of homozygous SNPs is that the hybridization mode is either allele-specific or cross-allelic for homozygous present and homozygous absent alleles, respectively (→ Eq. 11). Therefore, the signal of homozygous probes are not superposed by the signal of the alternative allele.

Further selection criteria considering nonspecific hybridization are applied using the hook plot (→ [18,19] and Figure 2). Probe sets with relatively large contribution of nonspecific hybridization, $x^{P-G\cdot N}>0.5$ (→ Eq. 4), are characterized by small coordinate values Σ and Δ. Both coordinates increase with decreasing $x^{P-G\cdot N}$ and level off at a peak with vanishing contributions of nonspecific hybridization, $x^{P-G\cdot N}\approx 0$ (→ Figure 2).

The logarithmic fraction of probe intensity due to nonspecific hybridization can be estimated using the coordinate differences with respect to the starting point of the hook curve [19],

$$\log x^{P-G\cdot N} \approx -\left((\Sigma-\Sigma_{start})\pm\frac{1}{2}(\Delta-\Delta_{start})\right), \tag{12}$$

where ± refer to P=PM (+) and P=MM (–), respectively. The fraction $x^{P-G\cdot N}$ consequently depends on the probe type with $x^{PM-G\cdot N}<x^{MM-G\cdot N}$ for Σ=constant. As selection criterion ("Hook" criterion) a threshold of $(\Sigma-\Sigma_{start})>0.7$ is applied to obtain allele sets with an average nonspecific intensity contribution of less than 20%, i.e. $x^{N}<0.2$ with $\log x^{N}_{\text{allele set}}=0.5\cdot\langle\log x^{PM-G\cdot N}+\log x^{MM-G\cdot N}\rangle_{\text{allele set}}$. This implies that the selected allele sets originate at least to 80% either from specific or cross-allelic hybridization.

The strand direction d does not affect the strength of the respective base pairings provided that sequence motifs from both the s- and the as-strands are considered in the same direction. Probe sequences are prepared for further analyses to be in 5'-3' direction, i.e. the probes' strandedness can be neglected.

Note that the hook plot obtained from the SNP array data lacks the horizontal starting range observed typically for expression arrays as a characteristic signature of "absent" probes without complementary targets. Hence, nonspecific hybridization to a smaller degree contributes to the signal intensities of SNP arrays compared with expression arrays in agreement with previous results [22].

2.9 Triple averaged intensities and sensitivities

The standard triple is defined as a string of three consecutive bases (XBY: X,B,Y∈{A,T,G,C}) in 5'-3' direction of the probe sequence. The middle base B is either the middle base or the SNP base of the probe sequence. Lateral bases form WC pairs with the respective target bases whereas B forms base pairings Bb, b∈{a,t,g,c}, according to the interaction group of the selected probe, Ab∈{At,Aa,Ag,Ac} (→ Table 1, Table 2).

Triple averages are calculated as log-mean of probes with a certain interaction group Ab of the central base B, a certain position k∈{2,...,24} of B in the probe sequence and the triple motif XBY of interest,

$$\log I^{P-T\cdot G}_{(Ab,k)}(XBY)=\left\langle \log I^{P-T\cdot G}_{(Ab,k,XBY)}\right\rangle, \tag{13}$$

with T∈{G,G'} for hp and ha probes, respectively. Triple averages regarding the offset value δ, with k=13+δ, are defined accordingly as

$$\log I^{P-T\cdot G}_{(Ab,\delta)}(XBY)=\left\langle \log I^{P-T\cdot G}_{(Ab,\delta,XBY)}\right\rangle, \tag{14}$$

Hence, probes can be chosen for triple average analyses by their attributes (→ Figure 3: part a gives an overview).

A series of nested means can be generated by averaging over one or more of the attributes Ab, k, δ or XBY, e.g. averaging over the offset positions can be performed with

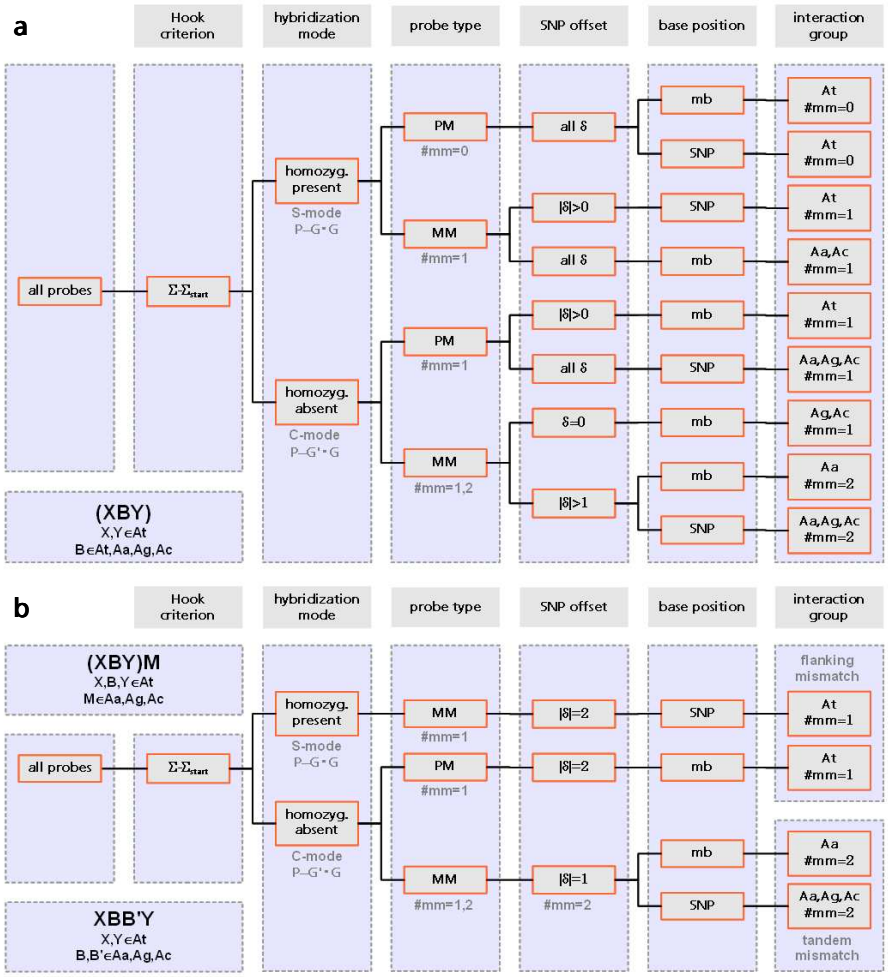


Figure 3: Probe selection for triple averaging. *Standard triples (XBY) are selected according to the scheme shown in part a. The interaction mode of the center base of the triple is defined by the chosen hybridization mode, the probe attributes (type, offset) and the position of B (SNP or middle base, mb) in the probe sequence. The interaction mode determines the base pairing formed by B with the target according to one of the four Ab-groups, At, Aa, Ag, Ac (→ Table 1 and Table 2), and the total number of mismatches per probe/target duplex, #mm. Part b shows special selections of triples with one flanking mismatch or of tandem mismatches.*

$$\log \mathrm{I}_{(\mathrm{Ab})}^{\mathrm{P-T \cdot G}}(\mathrm{XBY}) = \left\langle \log \mathrm{I}_{(\mathrm{Ab,\,XBY})}^{\mathrm{P-T \cdot G}} \right\rangle \tag{15}$$

and, in addition, averaging over the triple motifs to get the mean intensity per interaction group can be achieved with

$$\log \mathrm{I}_{(\mathrm{Ab})}^{\mathrm{P-T \cdot G}} = \left\langle \log \mathrm{I}_{(\mathrm{Ab})}^{\mathrm{P-T \cdot G}}(\mathrm{XBY}) \right\rangle . \tag{16}$$

The triple sensitivity specifies the deviation of a triple average from an appropriate mean value over all triples (→ [35] and see below) such as

$$\mathrm{Y}_{(\mathrm{Ab})}^{\mathrm{P-T \cdot G}}(\mathrm{XBY}) = \log \mathrm{I}_{(\mathrm{Ab})}^{\mathrm{P-T \cdot G}}(\mathrm{XBY}) - \left\langle \log \mathrm{I}_{(\mathrm{Ab})}^{\mathrm{P-T \cdot G}}(\mathrm{XBY}) \right\rangle . \tag{17}$$

Chapter 3
Triple average analysis

The stability of a particular base pair in an oligonucleotide duplex is significantly influenced by the two adjacent base pairings, i.e. by one on each side of the selected base pair (→ Figure 8: part a and b). The approach of 'Triple averaging' [33] accounts for the effect of the sequence on the probe intensities using triples of neighboring bases.

This chapter is aimed on classifying triples with common properties at first and to study them concerning background contributions, positional dependances, their sensitivities, mismatch stabilities, symmetry relations, adjacent WC pairings, tandem and flanking mismatches and, finally, in terms of nearest neighbor contributions.

3.1 Classification of triples

At first, I systematically analyze all triples of the sample array data regarding the relevant combinations of probe type P and hybridization mode (P−T•G), interaction group Ab, offset value δ and position π∈{mb,SNP} of B (data not shown).

$$\log \mathrm{I}^{\mathrm{P-T\cdot G}}_{(\mathrm{Ab},\delta,\pi)}(\mathrm{XBY}) = \left\langle \log \mathrm{I}^{\mathrm{P-T\cdot G}}_{(\mathrm{Ab},\delta,\pi,\mathrm{XBY})} \right\rangle , \tag{18}$$

After comparing and joining the results it turned out that there are 8 main groups of 64 standard triples regarding the mb and SNP position (→ Figure 4: part a, Table 2), namely an At-group with #mm=0 (At_0), an At-group (At_1), an Aa-group (Aa_1), an Ag-group (Ag_1) and an Ac-group (Ac_1) with #mm=1, and an Aa-group (Aa_2), an Ag-group (Ag_2) and an Ac-group (Ac_2) with #mm=2 (→ Appendix B). Triples in these groups are almost independent of their position in the probe sequence (→ see below). They cover about 85.1% of all the examined triples. The remaining 14.9% of the triples originate from MM−G, PM−G' and MM−G' probes with δ∈{−1,1}. These probes contain flanking or tandem mismatches and are subject to positional dependances (→ see below). The classification of the triples can also be applied for the probes.

The different triples of each group give rise to considerable variability of the intensity values (→ Figure 4: part a). The standard deviation of the whole set of the 64 triples of the At_0- and At_1-group is $\mathrm{sd}(\log \mathrm{I}(\mathrm{XBY}))=0.04$ and 0.06, respectively, but more than twice as large for the mismatch groups Aa_1 (sd=0.14), Ag_1 (sd=0.15) and Ac_1 (sd=0.10) (→ Table 3). Hence, mismatch pairings with adjacent WC pairs give rise to considerably larger variation of duplex stability than mere triples of WC pairs.

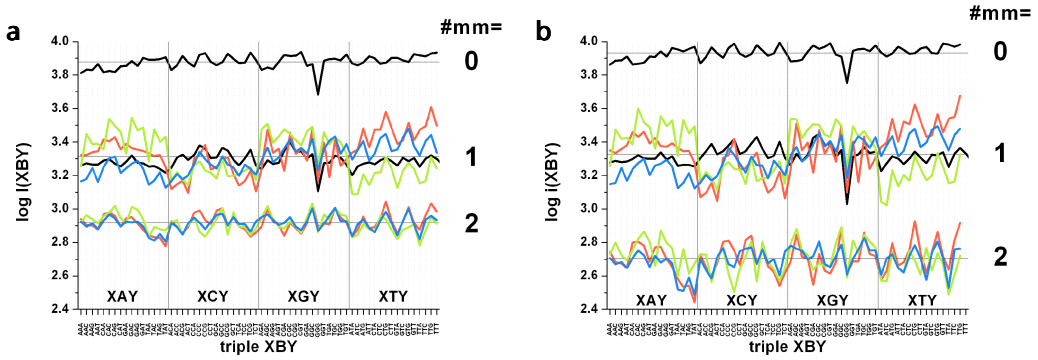


Figure 4: Triple-averaged probe intensities of different interaction groups and with different #mm. *Part a shows the 64 standard triples of the groups At_0, At_1, Aa_1, Ag_1, Ac_1, Aa_2, Ag_2, Ac_2. Triples are sorted accordingly to their central base pairing Bb. The groups arrange consequently to their number of mismatches per duplex #mm. See the text for the mean log-intensities regarding the number of mismatches. The mean difference between probes with #mm=0 and #mm=1 is ~0.58 and between probes with #mm=1 and #mm=2 ~0.39. Groups with #mm=1 have distinct triple log-intensities whereas groups with #mm=2 have virtually the same triple log-intensities. This is due to background contributions (→ see below and part b). Part b shows the groups of part a after background corrections. The mean log-intensities regarding the number of mismatches are for #mm=0 ~3.94, for #mm=1 ~3.32 and #mm=2 ~2.70. Accordingly, the differences between probes with #mm=0 and #mm=1 probes with #mm=1 and #mm=2 have changed after background correction to ~0.61 and 0.63, respectively.*

The triple intensities of each group vary about a certain mean log-intensity $\log I$ (→ compare with Eq. 16) according to $\#mm \in \{0,1,2\}$ with $Ab=At_0$, $Ab \in \{At_1,Aa_1,Ag_1,Ac_1\}$ and $Ab \in \{Aa_2,Ag_2,Ac_2\}$, respectively. The related $\log I$ are ~3.88 for #mm=0, ~3.30 for #mm=1 and ~2.91 for #mm=2 and the related variabilities are 0.04, 0.10 and 0.05. The low variability of groups with #mm=2 and the similar developing of their triple values in Figure 4 (part a) indicate contributions due to the optical and nonspecific background discussed in the next section.

3.2 Background corrections

The mean intensity level decreases for increasing #mm (→ Figure 4: part a). But independently of the number of mismatches per duplex it could be assumed that similar base-specific effects occur on the different mismatch levels. Figure 5 (solid symbols) shows for each interaction group the correlation between the triple averaged log-intensities of duplexes with #mm=k and duplexes with #mm=k+1, i.e. for duplexes differing by one mismatch. Triple data of the groups At_0 and At_1 arrange rather parallel to the diagonal line. However, the triple data of the Aa-, Ag- and Ac-groups do not. They are poorly correlated unlike the data of the At-groups.

The addressed triple average intensities contain contributions due to the optical and nonspecific background (→ Eqs. 2 and 3). Moreover, probe intensities saturate at large transcript concentrations and/or binding constants K_{duplex}. The mean probe intensities regarding #mm can be described by the hyperbolic function of the respective mean binding constant $K_{duplex}(\#mm)$ [35,36]

$$I(\#mm) \approx \left(\frac{I^{sat} \cdot c \cdot K_{duplex}(\#mm)}{1 + c \cdot K_{duplex}(\#mm)} + I^{BG} \right) \tag{19}$$

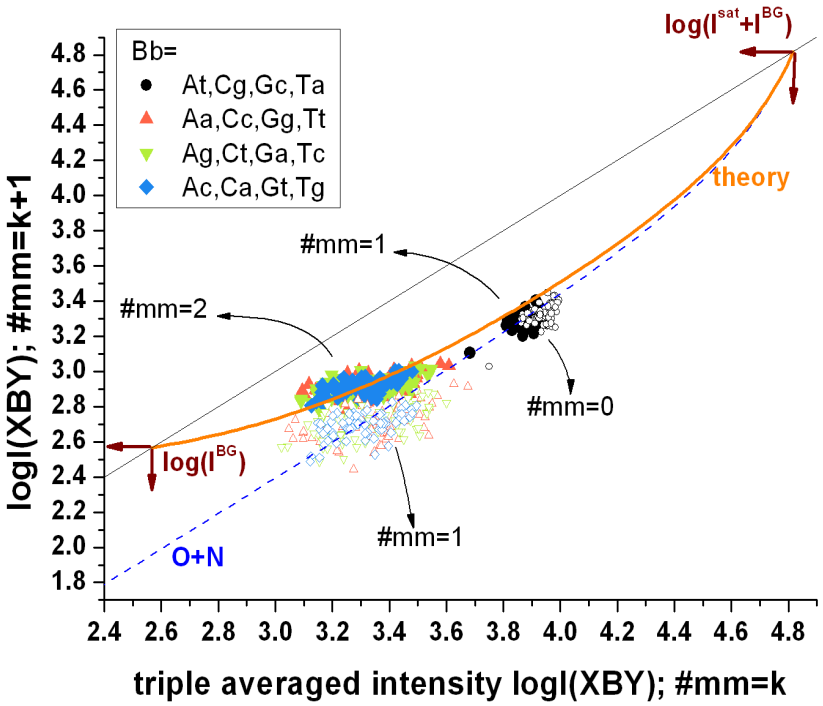


Figure 5: Background contributions. *The triple averages were correlated for #mm=0-versus-#mm=1 and #mm=1-versus-#mm=2, i.e. At_0 vs. At_1, Aa_1 vs. Aa_2, Ag_1 vs. Ag_2, Ac_1 vs. Ac_2. The data do not group in parallel with respect to the diagonal owing to the residual background intensity. Its consideration predicts the grouping of the data along the thick theoretical curve which was calculated using Eq. 19 and the equations in Appendix C. This curve intersects the diagonal line at the background and saturation intensities, $\log I^{BG}\approx 2.57$ and $\log I^{max}=\log(I^{sat}+I^{BG})\approx 4.82$, respectively. Correction of the intensities for the optical background and the nonspecific background improves the linear correlation of the data (→ curve "O+N").*

where I^{sat} denotes the saturation intensity at strong binding, $c\cdot K_{duplex}>>1$ and c is the transcript's concentration.

Assuming a factorial change of the binding constant per mismatch, $K_{duplex}(\#mm+1)=K_{duplex}(\#mm)/s$ (→ Figure 4: horizontal lines in part a and Figure 6: right axis) and a varying value of $c\cdot K_{duplex}(\#mm)$ in the limits $0<c\cdot K_{duplex}(\#mm)<\infty$ gives the theoretical relation between the mean intensities of duplexes which differ by one mismatched pairing (→ Figure 5). The theoretical curves intersect the diagonal line (x=y) at low and high intensities. The lower intersection point designates the mean background intensity I^{BG}, i.e. $\lim_{c\cdot K_{duplex}(\#mm)\to 0} I(\#mm)=I^{BG}$, whereas the upper one indicates the intensity $I^{max}=I^{sat}+I^{BG}$, i.e. $\lim_{c\cdot K_{duplex}(\#mm)\to \infty} I(\#mm)=I^{sat}+I^{BG}$, because Eq. 19 assumes independence of background and saturation levels from the number of mismatches. Eq. 19 predicts a significant deviation from the linear relation between the intensities of probes with #mm and #mm+1. The thick curve in Figure 5 was calculated using Eq. 19 and the equations in Appendix C. $\log I^{max}\approx 4.82$ can be obtained from the data by taking the logarithm of the maximum intensity of

the PM and MM probes and s≈3.79 as ratio I(#mm=0)/I(#mm=1) taking I(#mm) from the classification shown in Figure 4. This reveals a residual background intensity of $\log I^{BG} \approx 2.57$. It explains the lack of linear correlation between the experimental triple data of '#mm=0-versus-#mm=1' and especially of '#mm=1-versus-#mm=2'.

The used mean background intensity I^{BG} refers to the optical and nonspecific contributions according to Eq. 3. The intensities will be corrected for this contribution as (→ Figure 5: open symbols)

$$i = I - I^{BG} . \tag{20}$$

The respective theoretical curve 'O+N' runs parallel with the diagonal at decreasing intensities.

The correction progressively reduces the mean intensity level for #mm=2 (→ Figure 4: part b). The triple-specific effect is almost negligible for #mm≤1 but it affects the results for #mm=2. To avoid potential perturbations in the data of groups with #mm≤1, their intensities will be left unchanged whereas the corrected intensities of the groups with #mm=2 substitute their respective uncorrected intensities. Anyway, all intensities will be denoted as i (→ Eq. 20).

3.3 Mismatch effect

The allele-specific and cross-allelic modes include perfect matched and mismatched probe/target duplexes with zero to two mismatch pairings at the mb and/or SNP position (→ Table 1). The number of mismatches contained in a duplex results from the combination of hybridization mode, probe type, interaction group and the offset value (→ Table 1 and Table 2).

The presented results of the triple's classification (→ Figure 4) show that the number of mismatched base pairings per duplex (#mm) is the dominant factor which affects the mean intensity of the triples and, hence, of the probes (→ Figure 4: horizontal lines). The logarithmic intensity ratio can be approximated as function of #mm by

$$\frac{\log i(\#mm)}{\log i(0)} \propto \frac{\log K_{duplex}(\#mm)}{\log K_{duplex}(0)} \approx x \cdot \left(1 - \gamma \cdot \left(1 - x^2\right)\right) \quad \text{with} \quad x = 1 - \frac{\#mm}{25} , \tag{21}$$

where $i(\#mm) = \log I(\#mm) - I^{BG}$ is the background corrected mean intensity of probes with #mm mismatches; $K_{duplex}(\#mm)$ denotes the respective mean binding constant; x is the fraction of WC pairs in the duplex and γ is a fitting constant depending on the hybridization conditions [37].

The logarithmic intensity ratio can also be estimated from the assumption of additive contributions of each base pair using $\log K_{duplex}(\#mm) \approx \log K_{duplex}(0) - \#mm \cdot \delta\varepsilon$, where $\log K_{duplex}$ is the mean binding constant and δε is its mean incremental penalty (in units of $\log K_{duplex}$) if one WC pairing is substituted by a mismatch. This approach predicts an exponential decay of the intensity as a function of the number of mismatches, $i(\#mm) \approx i(0) \cdot 10^{-\#mm \cdot \delta\varepsilon}$, which transforms into

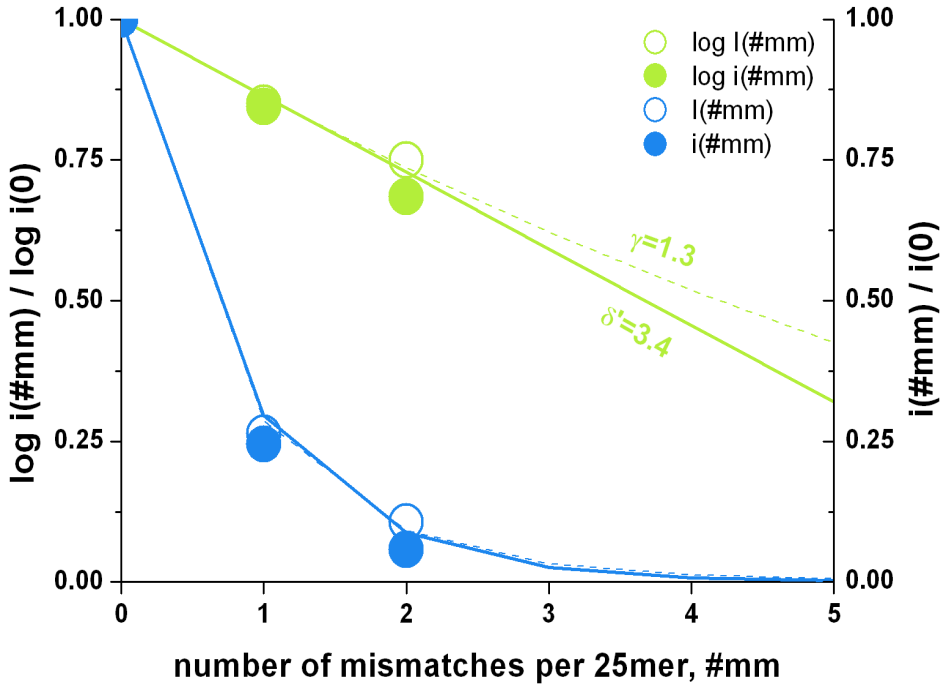


Figure 6: Mean effect of the number of mismatches (#mm). *Relative decrease of the mean probe intensity as a function of #mm (symbols). The curves are calculated using Eqs. 21 and 22. The data are shown in logarithmic (left axis, upper plots) and linear (right axis) scale without (open symbols) and with (solid symbols) background correction.*

$$\frac{\log i(\#mm)}{\log i(0)} \propto 1-\delta'\cdot(1-x) \quad \text{with} \quad \delta' = \frac{\delta\varepsilon}{\log K_{duplex}(0)/25}\,, \tag{22}$$

using the logarithmic form as in Eq. 21. The constant δ' is given by the ratio between the incremental penalty due to the mismatch and $\log K_{duplex}(0)/25$, which has the meaning of the mean additive contribution of one WC pairing to $\log K_{duplex}(0)$. Both alternative functions given by Eqs. 21 and 22 are virtually not distinguishable for #mm<3 (→ Figure 6). The functions predict that one mismatch reduces the probe intensity to about one fourth of its perfect match value. More than two mismatches consequently decay the intensity to tiny values of less than 5%. A decay rate δ'>3 indicates that the intensity penalty due to the first two mismatches markedly exceeds the average intensity contribution of a single WC pairing in the perfect matched probe/target duplexes. Simple balance considerations imply that δ' has to decrease with increasing number of mismatches as predicted by Eq. 21 (→ Figure 6: theoretical curves).

3.4 Positional dependance

During the classification of the data it became apparent that the triples are basically independent of their position in the probe sequence. Figure 7 shall clarify this by

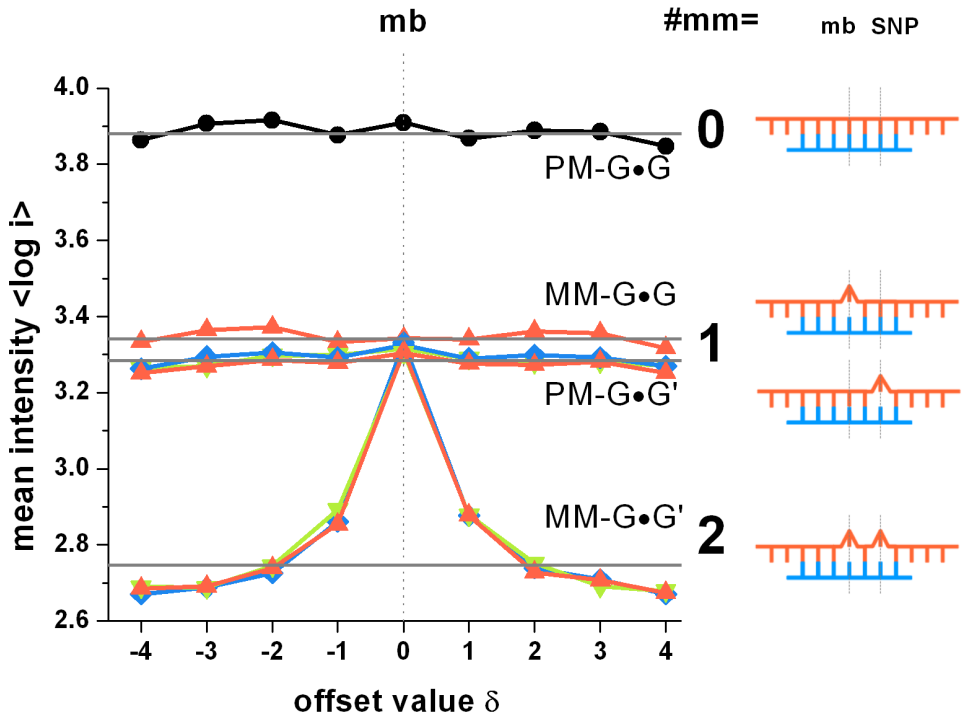


Figure 7: Averaged log-intensities for probes of different Ab-groups and offset positions. *Mean probe intensity, averaged over all probes with a given SNP offset (→ see also the sketches to the right of the figure), as a function of the offset value δ with respect to the middle base mb for probes with different number of mismatches per probe/target duplex (#mm=0-2). Virtually no significant effect of the offset position can be observed for single mismatches within the relevant range $|\delta|\leq 4$. Contrarily, the mean intensity decreases with increasing separation between double mismatches (#mm=2) where one is located at the center of the probe (middle base, mb) and the second one at the offset position δ. Note that both mismatches merge into one for δ=0. The homozygous absent data (P–G'•G) were separately calculated for the three groups of mismatches, Aa, Ag and Ac. The respective curves are almost identical.*

showing the log-intensity averages of the Ab-groups regarding the offset value. They are well separated by their number of mismatching base pairs.

The allele-specific and cross-allelic modes include perfect matched and mismatched probe/target duplexes with zero to two mismatches at the mb and/or the SNP position (→ Table 1). The number of mismatches results from the combination of hybridization mode, probe type, interaction group and the offset value (→ Table 1 and Table 2). To assess the influence of the offset value on the probe intensities I calculate the triple log-intensity averages $\log \mathrm{I}^{\mathrm{P-T\cdot G}}_{(\mathrm{Ab},\delta)} = \langle \log \mathrm{I}^{\mathrm{P-T\cdot G}}_{(\mathrm{Ab},\delta)}(\mathrm{XBY})\rangle_{\mathrm{XBY}}$ for homozygous present (T=G) and homozygous absent (T=G') probes (→ Figure 7 and Eq. 14).

PM–G•G and MM–G•G represent probe/target duplexes of the present allele with the intended probes (hp mode). PM–G•G duplexes contain exclusively WC base pairings (At-group) independently of the offset δ. MM–G•G duplexes contain one mismatching base pair of the Aa-group at the mb position except for duplexes of probes with δ=0 from complementary SNP types, i.e. $\mathrm{B_A/B_B}\in\{\mathrm{A/T,C/G}\}$. They form a mismatching base pair of the Ac-group at the mb position. Intensity averages of

PM−G•G and MM−G•G duplexes (δ≠0 and Ab≠Ac) represent pseudo-replicates, respectively. The scattering of the respective intensity averages about their means indicates the variability of the different probe ensembles formed for each offset value.

PM−G'•G and MM−G'•G represent probe/target duplexes of the present allele with probes intended for the cross allele (ha mode). PM−G'•G duplexes always contain one mismatching base pair at the SNP position either from the Ag-, Ac- or Aa-group. The averaged intensities refer to the shift of the mismatch relatively to the middle base. The position of the single mismatch weakly affects the mean intensity in the range of the SNP offset positions. MM−G'•G duplexes of probes with δ≠0 contain one mismatching base pair of the Aa-group at the mb position and one mismatch at the SNP position either from the Ag-, Ac- or Aa-group. Both mismatches are separated by δ−1 WC pairings in between. The intensity averages decrease with increasing distance of the mismatches (→ Figure 7). This trend indicates that the destabilizing effect of the mismatches is smaller for tandem mismatches (|δ|=1), it slightly decreases for a single intermediate WC pairing (|δ|=2) and essentially levels off for more WC pairings in between (|δ|>2). MM−G'•G duplexes of probes with δ=0 have only one mismatch pairing either of the Ac-group ($B_A/B_B \in \{A/C,G/T\}$) or the Ag-group ($B_A/B_B \in \{A/G,C/T,A/T,C/G\}$).

The results of MM−G•G and PM−G'•G agree with previous studies which show that the destabilizing effect of single mismatches is almost constant over a broad range in the middle part of short-length oligonucleotide duplexes and decreases only for the last 4-6 base positions near the ends of the probe sequence [38-40].

Positional dependance of single base and triple motifs

The PM probes form exclusively WC pairs in homozygous present PM−G•G duplexes. To study the positional effect of WC base pairings over the whole sequence length, I calculated mean log-intensities for all these duplexes containing a certain base B∈{A,T,G,C} at each position k∈{1,...,25} of the probe sequence (→ Figure 8: part a). The obtained positional-dependent log-intensity averages only weakly vary about their total mean. The base-specific differences essentially disappear towards the 3' end of the probes (k>23) which is attached to the chip surface (→ see also Figure 8: part c).

The homozygous present duplexes of the MM probes, MM−G•G, also form predominantly WC pairings except the middle base which forms mismatches of the Aa- or the Ac-group. The single base averaged intensities of these mismatches vary to a much larger degree about their mean compared to the WC pairings (→ Figure 8: arrow in part a). The strong mismatch effect extends also to the flanking bases at adjacent positions k=12 and 14 (→ see also Figure 8: part b). This justifies the 'Triple averaging' approach.

Part b of Figure 8 shows the single base positional dependance of homozygous absent PM probes (PM−G'•G) for different offset values δ of the SNP position which forms a mismatch pairing in the probe/target duplexes. As for the MM, the SNP position exhibits a larger spread of the single base values about their mean compared with the WC pairings at the remaining sequence positions. They represent averages over mismatches of the Aa-, Ag- and Ac-group in contrast to the mismatches of the middle base coming from the Aa-group shown in part a of Figure 8. The data clearly

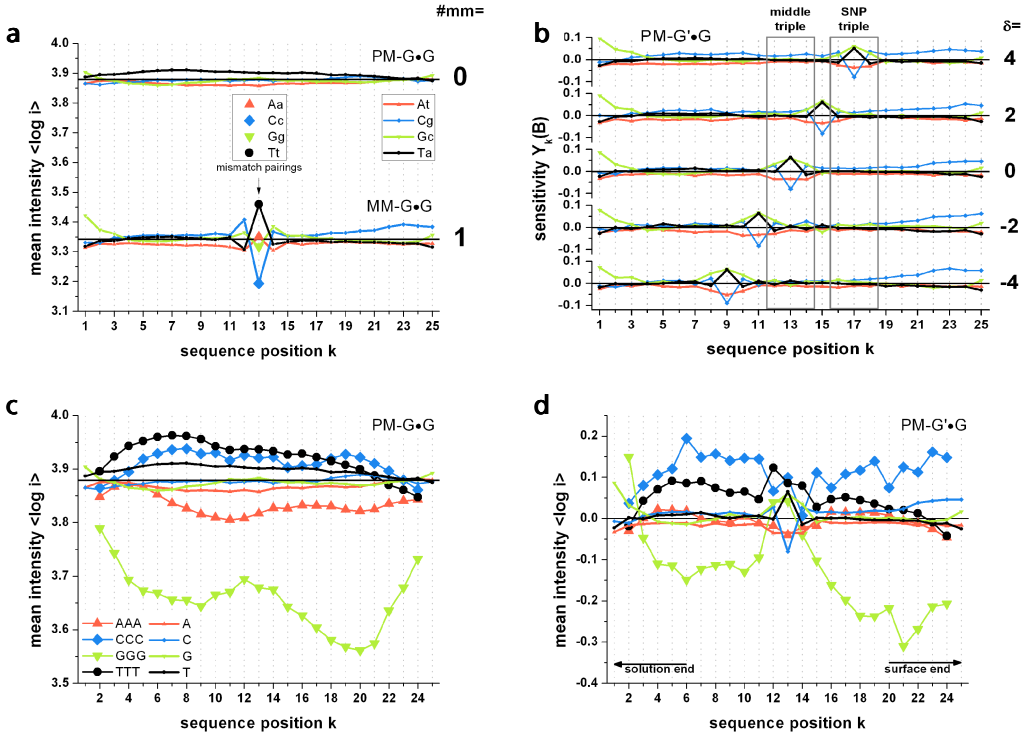


Figure 8: Positional dependence of the probe intensities. *Part a: Single base data of allele-specific (S-mode) PM and MM probes. Each data point was calculated as average log-intensity over all probes of the considered group with the indicated base at position k of the probe sequence. It is associated either with WC pairings or with mismatched pairings at the middle base (mb) position of the MM probes. These mismatches give rise to markedly larger variability of the intensities than the WC pairings do at the remaining positions. Part b shows the positional dependence of the sensitivity (deviation of the log-intensity from the mean over all probes of the respective group) of cross-allelic PM probes (C-mode) with different offsets of the SNP. The base at the SNP position forms a mismatched pairing which shifts along the sequence according to the offset. Note that the mismatch values are averages over all groups (Aa, Ag, Ac) whereas the mismatches in part a refer to the Aa-group only. Part c enlarges the single base curves for PM–G•G shown in part a. In addition, mean log-intensity values were calculated for homologue triples along the probe sequence (position k refers to the center base of the triples). The mean log-intensities slightly increase for AAA, CCC and TTT compared with the single base averages but markedly decrease for triple guanines. Part d shows the respective single base and triple values for the cross-allelic PM probes with offset δ=0 shown in part b. Comparison with part c indicates subtle differences of the curves at positions which refer to WC pairings in both situations. For example, triple guanine motifs give rise to relatively large intensities near the solution end of the probe and also the cytosines (C- and especially CCC-motifs) are associated with largest intensities for most of the WC pairings in part d whereas thymines give rise to largest intensities in part c.*

reflect the shift of the mismatch pairing with changing offset position of the SNP. The profiles remain nearly invariant at the remaining sequence positions.

To estimate the effect of longer sequence motifs I calculated intensity averages of probes possessing homologue triples, i.e. runs of three consecutive bases of the same type at a certain sequence position (→ Figure 8: part c and d). The specific effect of these motifs clearly exceeds that of the single bases, especially for runs of triple-G. These GGG-motifs systematically reduce the probe intensities by a factor of ~$10^{-0.1}$-$10^{-0.3}$≈0.8-0.5 compared with the mean intensity for most of the sequence positions. In

contrast, the mean effect of a single G is almost negligible. The GGG-effect essentially disappears at the mismatch position in the middle of the probe sequence (→ Figure 8: part d). The similar "buckled" shape of the GGG-profile in the middle of the probe sequence of PM−G•G duplexes (part c) probably indicates a certain small fraction of misassigned genotypes in the selected sub-ensemble of homozygous present probes.

Comparison of part c and d of Figure 8 reveals also more subtle differences between the profiles at positions which refer to WC pairings in both, the PM−G•G (part c) and PM−G'•G (part d) duplexes. Firstly, the triple TTT provides the largest intensities for the former duplexes whereas the triple CCC becomes largest in PM−G'•G duplexes. Moreover, the effect of cytosines progressively increases towards the surface end in PM−G'•G duplexes whereas it apparently disappears in the data obtained from PM−G•G duplexes. Secondly, the intensity effect due to guanines begins with positive values at the solution end of PM−G'•G duplexes (k=2) and then steeply decreases to negative values.

It is known that the sequence profiles are sensitive to factors such as the optical background correction and saturation [31,41]. Large and small intensities are prone to saturation and background effects, respectively, which differently affect the specific signal. Saturation, for example, limits large probe intensities and therefore reduces the relative effect of strong base pairings because probes containing such motifs are most affected by this effect. The relative small single and triple cytosine values in the profiles of PM−G•G duplexes can be attributed to selectively stronger saturation of probes containing these motifs. Contrarily, in the PM−G'•G duplexes saturation is much less relevant owing to the smaller average level of probe occupancy and intensity. The different response of triple guanines and cytosines near the solution and surface ends of the probe seems puzzling and will be addressed in the next chapter.

3.5 Triple sensitivities

Triple sensitivities provide a measure of the sequence specific influence of the pairing of the central base and their nearest neighbors on the probe intensities in terms of deviation from the mean intensity of the respective interaction group.

The four groups At_0, Aa_1, Ag_1 and Ac_1, are chosen for the analysis of the triple sensitivities and further analyses. At_0 represents perfect matching probe/target duplexes formed by WC base pairings. The three other groups represent probe/target duplex containing a single mismatch of the respective mismatch interaction group (→ Table 1). At_1 is neglected as it regards WC pairings in single mismatch probe/target duplexes that are superposed by effects originating from the mismatch base pairs. The groups with #mm=2 are also neglected as they are prone to be biases by background contributions that could remain after background correction to a higher degree than in groups with #mm=1. Furthermore, the effects due to the two mismatches interfere with each other.

The triple averaged and background corrected intensities are used to calculate the 64 triple sensitivity values for each of the four groups At_0, Aa_1, Ag_1 and Ac_1 (→ Eq. 17, Appendix B). The sensitivity values of At_0 is related to the mean of the triples of the At_0-group whereas the sensitivities of the groups Aa_1, Ag_1 and Ac_1 are related to the

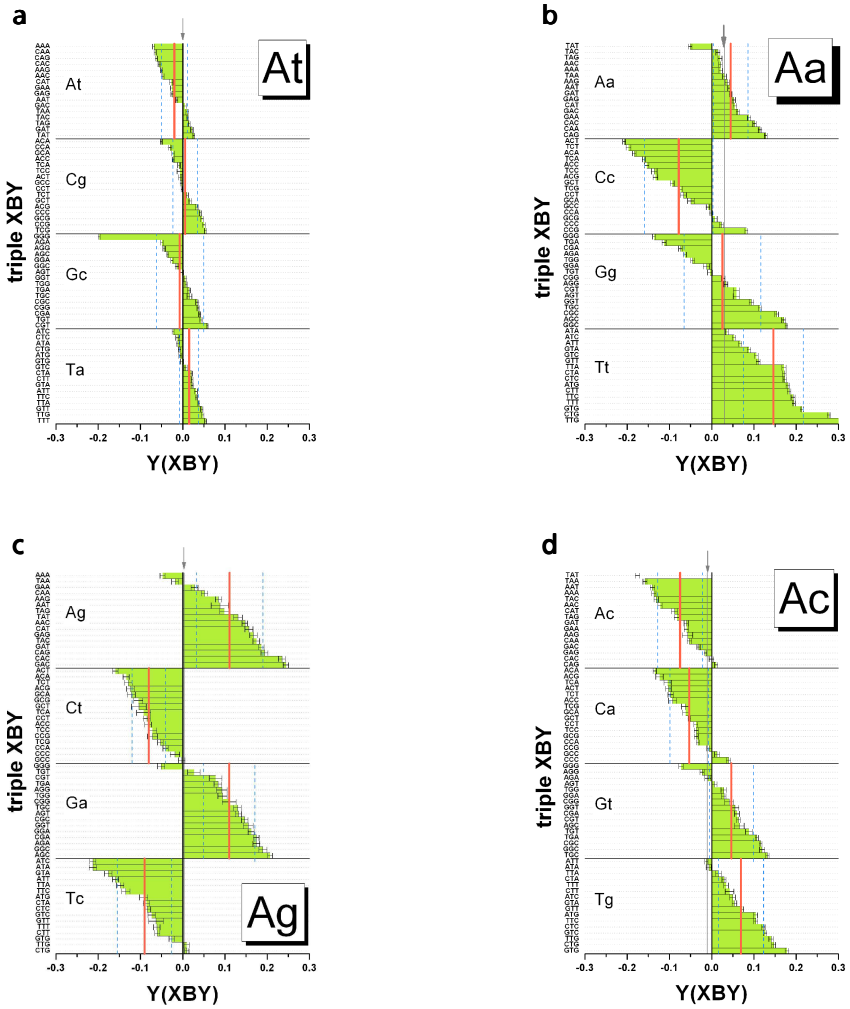


Figure 9: Triple averaged sensitivities. *The sensitivity values of At_0, Aa_1, Ag_1 and Ac_1 are calculated using Eq. 17 relative to the average log-intensities of either the perfect match probes (At_0) or all single mismatched probes (At_1,Aa_1,Ag_1,Ac_1, #mm=1). They are ranked with increasing sensitivity for each center base B forming base pairings with the target according to their interaction group as indicated in the figure by upper (probe) and lower (target) case letters. Averages of the particular groups (→ arrows) and of the central base pairs are shown by vertical solid lines. The vertical dashed lines indicate the standard deviation of the triple values about the mean related to the central base (→ see also Table 3). The mean and the standard deviation estimate the stability of the respective pairing Bb and the effect of flanking WC pairs, respectively. The error bars indicate the standard error of the triple sensitivities.*

total mean of the triples of the four groups with #mm=1 (At_1, Aa_1, Ag_1, Ac_1). Figure 9 summarizes the sensitivity data.

Most sensitivities of the At_0-group (WC pairings) scatter rather tightly around their mean indicating an only moderate sequence effect. Only the 'GGG' triple is quite conspicuous as it causes a relatively large intensity penalty. A 'GGG' motif reduces the intensity on the average by a factor of about $10^{-0.2}\approx 0.63$ compared with the mean intensity. The considered triples refer to offset positions $|\delta|\leq 4$ around the

middle base. The full positional dependance of 'GGG' indicates a yet stronger intensity drop for sequences containing a 'GGG' triple towards the ends (→ Figure 8, part d).

Importantly, the 'GGG' penalty is in contradiction to complementary rules because the complementary 'CCC' motif shows completely different sensitivity properties. Triple C's give rise to the opposite effect as they amplify the intensity by a factor of about $10^{+0.04}\approx 1.1$. This result is discussed below.

The sensitivities of the remaining three groups are subject to an increased variability of the triple data that considerably result from the substitution of the central WC base pairing by a mismatch. The mean variability of each group was estimated as standard deviation of all 64 triple sensitivities of each group (→ Table 3). It more than doubles for the mismatch groups Aa_1, Ag_1, Ac_1 (sd=0.09–0.13) compared to At_0 (sd=0.04). Single mismatches can modify the intensity by a factor between $\sim 10^{-0.25}\approx 0.55$ and $\sim 10^{+0.25}\approx 1.8$. This result generalizes the trend which is illustrated in Figure 8 (part d) for the special case of mismatches of the Aa-group in the middle of the probe sequence.

3.6 Mismatch stability

The mean sensitivity of all triples with a given middle base B provides a measure of the average stability of the respective mismatch pairing Bb (→ Figure 9: red lines). The Aa_1-, Ag_1- and Ac_1-groups show the relations Cc<Gg≈Aa<Tt, Tc≈Ct<Ag≈Ga and Ac≈Ca<Gt≈Tg, respectively. This confirms the expected symmetries for bond reversals Bb↔B^rb^r in symmetrical DNA/DNA interactions, i.e. $Y_{Ab}(Bb)\approx Y_{Ab}(B^rb^r)$, e.g.

Table 3: Sources of variability of triple motifs and of tandem mismatches.

Interaction group[1]	At_0	Aa_1	Ag_1	Ac_1
triples[2]	0.04±0.0005	0.12±0.0005	0.13±0.001	0.09±0.0005
3'/5' asymmetry[3]	0.03	0.11	0.07	0.05
complementary asymmetry (without GGG)[4]	0.07 (0.05)	0.10 (0.08)	0.08 (0.06)	0.06 (0.05)
NN residuals (without GGG)[5]	0.02 (0.01)	0.02 (0.02)	0.03 (0.02)	0.01 (0.01)
flanking mismatches[6]		0.03	0.03	0.02
tandem mismatches (XY)[7]		0.02 (0.04)	0.015 (0.05)	0.02 (0.05)
tandem mismatches (BB')[7]		0.07 (0.13)	0.08 (0.10)	0.06 (0.07)
tandem mismatches (YB'/B'Y)[7]		0.03 (0.02)	0.06 (0.08)	0.04 (0.05)

[1] Variability estimates are separately calculated as standard deviation for each Ab-interaction group: $SD=\sqrt{\langle \Delta^2 \rangle_{Ab}}$.

[2] Variability of triple averages with respect to the group-mean: $\Delta = Y_{Ab}(XBY) - \langle Y_{Ab}(XBY) \rangle_{Ab}$; it estimates the variability of interactions due to the choice of the triple; the standard error refers to the variability of the probe level data of each interaction group.

[3] Variability of triple averages after 3'/5' transformation: $\Delta = Y_{Ab}(XBY) - Y_{Ab}(YBX)$.

[4] Variability of triple averages after complementary transformation: $\Delta = Y_{Ab}(XBY) - Y_{Ab}(Y^cB^rX^c)$; the values in the brackets are obtained after omitting the GGG-motif.

[5] Variability of residual values after reduction of the model rank NNN→NN: $\Delta = \Delta^{res}_{Ab}$ (→ Eq. 28).

[6] Variability due to flanking mismatches: $\Delta = \Delta^{flank}_{Ab}$ (→ Eq. 26).

[7] Variability due to quadruplet motifs with tandem mismatches XBB'Y/YB'BX with B∈Aa and B'∈Aa,Ag,Ac. The SDs were calculated with respect to the average over the three groups ($\Delta(XY) = \langle Y_{Ab}(XBB'Y) \rangle_{BB'} - \langle \langle Y_{Ab}(XBB'Y) \rangle_{BB'} \rangle_{Ab}$ and $\Delta(BB') = \langle Y_{Ab}(XBB'Y) \rangle_{XY} - \langle \langle Y_{Ab}(XBB'Y) \rangle_{XY} \rangle_{Ab}$) and with respect to the total mean over all couples (values in the brackets; $\Delta(XY) = \langle Y_{Ab}(XBB'Y) \rangle_{BB'} - \langle \langle Y_{Ab}(XBB'Y) \rangle_{BB'} \rangle_{Ab,XY}$ and $\Delta(BB') = \langle Y_{Ab}(XBB'Y) \rangle_{XY} - \langle \langle Y_{Ab}(XBB'Y) \rangle_{XY} \rangle_{Ab,BB'}$).

Tc↔Ct and Ac↔Ca. Note that, in contrast, DNA/RNA interactions are asymmetrical in solution [42] and on microarrays [31,33].

A comparison of the mean sensitivities for each central mismatch pairing of the three mismatch groups gives the following ranking of the stability of mismatch pairings:

Tc(−0.10)≤**Ct**(−0.09)≤**Cc**(−0.08)≈**Ac**(−0.08)≤**Ca**(−0.06)<0
0<**Gg**(+0.03)≤**Aa**(+0.05)≈**Gt**(+0.05)≤**Tg**(+0.08)<**Ga**(+0.12)≈**Ag**(+0.12)<**Tt**(+0.16) (23)

The numbers in the brackets are the respective mean sensitivities of each mismatch pairing averaged over the 16 combinations of adjacent bases (standard error: ~±0.02).

Other authors report similar rankings of the stability of single mismatches in DNA/DNA oligomer duplexes which are obtained from hybridization studies on surfaces (microarrays or special solid supports) or in solution:

Gg≤**Ca<Ct≈Cc≈Gt≈Aa**<Ac≈Tc≤**Ga<Tg≤Tt<Ag** (microarray, [28])
Ct≈Cc≤Ca≤Ac≤Aa≈Tc≈Ga≤**Gt**<Gg<**Tt<Ag≈Tg** (microarray, [38])
Ac≈Tc≈Tt≈Aa<**Ag≈Tg** (solid support, [43], only selected pairings are studied)
CC≤AC≤TC≤AA≈TT<**GA**≈GT<GG (solution, [44]) (24)

In solution both dimerized oligonucleotides are equivalent as indicated by the two capital letters which assign the pairing.

Basic agreement of the reference studies with the ranking given by Eq. 23 is underlined and bold. Accordingly, the consensus ordering of the microarray studies comprises Ct, Ca, Cc as low stability mismatches, Ag, Tg, Tt as high stability mismatches and Gt and Aa at the intermediate position. A major difference between the previous rankings lies in the assignment of Gg which is the least stable in the study of Naiser *et al.* [28] and one of the most stable mismatches in the study of Wick *et al.* [38]. In the ranking given by Eq. 23 it is assigned as having intermediate stability. Figure 9 shows large variability of triples with a central Gg mismatch around zero. Imbalanced triple selection in studies using a limited number of oligonucleotides therefore are prone to lead to biased results where the apparent stability of Gg can vary between large and low values in dependance on the particular realization of probe/target duplexes containing a Gg mismatch. The total probe number of the studied SNP array (10^6) largely exceeds the probe number used in previous studies by about three orders of magnitude (10^3 [38] and $2\text{-}3 \cdot 10^3$ [28]). Comparison of the different rankings of mismatch stabilities obtained from microarray and solution data reveals disagreement especially for GG, GT and TT motifs. These differences possibly indicate additional or alternative explanations for the inconsistent chip rankings which will be discussed below.

It must be pointed out that the reported references [28,38] estimate mismatch stabilities by directly comparing the intensities of MM and PM probes. However, this refers to the stability difference between the mismatch pairing and the respective WC pairing. The ranking found (→ Eq. 23) uses the mean stability of all considered single base mismatches as reference level which is independent of the particular triple. The relatively small variability of the single base averages of the At-group (→ Figure 9: red lines of At-group) however show that the explicit use of the WC sensitivity as

reference essentially does not change the ranking of mismatch stabilities in the data set. Direct comparison with the reference data is therefore adequate.

3.7 Symmetry relations

Stacking interactions and probe/target imbalances can be analyzed using two symmetry relations of the triples, namely 3'/5' reversal and probe/target complementarity, with

$$XBY \Leftrightarrow YBX \text{ and } XBY \Leftrightarrow Y^{c}B^{r}X^{c}, \tag{25}$$

respectively. The sequence motifs are given in 5' to 3' order which corresponds to reading from the end of the sequence towards the glass. The superscripts 'c' and 'r' denote complementary nucleotide letters in the special case of WC pairings, e.g. A^{c}=T, and bond-reversals for the more general situation which includes also mismatch pairings, e.g. A^{r}=G or A^{r}=A for mismatches of the Ag- or Aa-group, accordingly. Triple sensitivities shown in Figure 9 are used to evaluate the symmetry relations.

Perfect 3'/5' symmetry of the triples is expected if the base pairings are independent of their nearest neighbors and is characterized by Y(XBY)=Y(YBX). Stacking interactions between adjacent nucleotides however make an essential contribution to the stability of DNA/DNA duplexes [45,46]. The change of stacking contributions after strand reversal is governed by the different stereochemistry of 3'/5' and 5'/3' strand directions in the duplexes. The deviation from the perfect 3'/5' symmetry relation thus estimates the effect of stacking interactions in the considered triples.

In contrast, the complementarity relation keeps the strand direction unchanged. Perfect complementarity of the triples is expected if both interacting strands are physically equivalent and if their reactivity is not selectively perturbed by parasitic reactions such as intramolecular folding and/or bulk dimerization [26]. It is characterized by Y(XBY)=Y($Y^{c}B^{r}X^{c}$). Duplexing experiments in solution typically use oligonucleotides of equal length and of low propensity for intramolecular folding and self-interactions. A quite different situation is encountered on microarrays as the reacting partners are highly asymmetric in length and conformational freedom. Firstly, the probes are attached to the chip surface whereas the targets are dissolved in the supernatant solution with consequences for their reactivity. Interactions, for example, depend on the position of the nucleotide letter in the probe sequence owing to their attachment to the chip surface which gives rise to positional dependent constraints of probe/target interactions [22,26]. Secondly, the length of the targets exceeds that of the probes typically by more than one order of magnitude which markedly enhances their propensity for intramolecular folding and intermolecular duplexing reactions in solution in a sequence-dependent fashion with consequences for their effective interactions with the probes. Hence, deviations from perfect complementarity are expected to detect imbalanced probe/target interactions due to the asymmetric nature of the hybridization reaction on microarrays.

The triple sensitivities shown in Figure 9 are ordered decreasingly for each group (→ Figure 10, thick lines) together with the reordered values according to the symmetry relations (→ Eq. 25 and Figure 10: symbols). The scatter width of the symbols around the ranked triples in terms of their standard deviation defines a kind of asymmetry funnels (→ Figure 10: dotted lines). The widths of the funnels

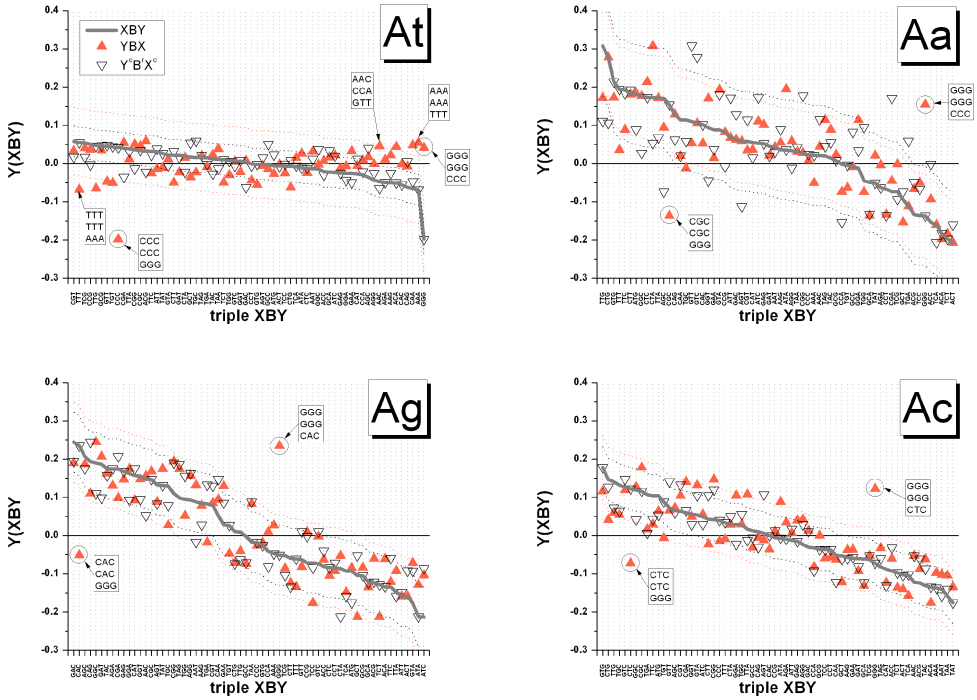


Figure 10: Symmetry relations of triple interactions. *The triple sensitivities, Y(XBY), of each interaction group are ranked in decreasing order and shown by thick lines. For each base triple three sensitivity values are shown according to Eq. 25 to reveal 3'/5' asymmetry, Y(YBX), and complementarity, Y(Y°B'X°), respectively (symbols are assigned in the figure). The abscissa labels indicate the XBY triple. The triples in the boxes indicate examples of triples whose sensitivity values reveal considerable asymmetry, for example XBY/YBX/Y°B'X°=AAC/CAA/GTT of the At-group. Note that GGG-motifs are highly non-complementary in all four interaction groups. Note also the markedly different widths of the scattering funnels of the different interaction groups given by their standard deviation (→ see dotted lines and also Table 3) indicating that the stacking terms and/or asymmetry of interactions are differently modulated by the central mismatch (→ see text). For symmetry reasons some of the asymmetries' differences vanish (e.g. 3'/5' asymmetry of GGG/GGG/CAC of the Ag-group).*

(→ Table 3, standard deviations) characterize the mean asymmetry of the triple interactions of the respective interaction group. Vanishing funnel widths are expected for perfect probe/target symmetries.

Both, 3'/5' and complementary asymmetries roughly behave identically. They are, by far, smallest for the At-group and largest for the Aa-group which agrees with the ranking of the variability of the triple sensitivities between the groups. Also the sd values roughly agree (→ Table 3) which indicates independence of triple sensitivities after symmetry transformation. Hence, the central mismatch coming from the Aa-group is obviously most effected by stacking interactions and complementary asymmetries among the considered groups. This causes the largest variability of the associated probe intensities. It is important to keep in mind that only mismatches of the Aa-group are used in the basic principles of the "PM-MM probe strategy" [47]. Though, the given results imply that this design principle seems to be suboptimal with regard to the rather high variability of mismatch stability. This effect introduces

additional noise to the MM intensities which are actually intended to correct the PM signals for background contributions.

Examples for symmetry relations are explicitly indicated in Figure 10. The respective triples XBY/YBX/$Y^cB^rX^c$ are given within the boxes whereas the abscissa labels just indicate the XBY triple. For example, the combination AAC/CAA/GTT taken from the At-group shows marked 3'/5'-asymmetry beyond the limits of the mean scattering funnel. The largest complementary asymmetry by far is associated with triple-G motifs in the probe sequence for all interaction groups (→ Figure 10: solid triangles surrounded by circles). They make a contribution of up to 30% to the mean variability of the respective interaction groups (→ Table 3). Note in this context that the GGG-motifs are characterized by the weakest interactions either among all 64 triples (→ Figure 9: At-group) or among the 16 triples with a central G (→ Figure 9: Aa-, Ag- and Ac-groups). This effect will be further discussed below.

3.8 Adjacent WC pairings

The triples XBY considered in this analysis, i.e. mb and SNP triples, consistently have two WC pairings adjacent to the central perfect match or mismatch B. The two WC pairings considerably modulate the strength of the central base pair interaction as assumed by the triple averaging approach. For example, the ratio of two triple sensitivities with a central Cc mismatch (Aa-group) flanked either by two C's or by two A's is about $Y(CCC)/Y(ACA)|_{Aa} \approx 10^{+0.2} \approx 1.60$ whereas the respective sensitivity ratio of the triples with a central Cg pair (At-group) is only $Y(CCC)/Y(ACA)|_{At} \approx 10^{+0.1} \sim 1.25$.

The influence of the adjacent WC pairings can be estimated by averaging the triple sensitivities of the At-group and the mismatch groups over the central base B, $Y^{ad}_{Ab}(XY) = \frac{1}{2}\langle Y_{Ab}(XBY) + Y_{Ab}(YBX)\rangle_B$. The values rank in good agreement with the expected mean stability of single nucleotide canonical DNA/DNA interactions, C≈G>A≈T [45] (→ Figure 11). A small systematic trend between the mismatch Ab-groups can be noticed, i.e. Aa>Ag≈Ac (→ Figure 11: symbols), as well as the decreasing variability of the data with decreasing mean.

3.9 Tandem and flanking mismatches

Tandem mismatches appear in duplexes of MM probes interrogating for the absent allele, i.e. MM−G'•G. They denote two adjacent mismatches in probes with $|\delta|=1$. Flanking mismatches, however, appear in MM−G•G and PM−G'•G duplexes and denote mismatches queued to standard triples containing no mismatch.

Both motifs are separately analyzed to estimate their specific effect on the probe intensities in comparison with the standard triples. Special selection criteria for triples with a flanking mismatch or probes with tandem mismatches are given in Figure 3, part b.

Tandem mismatches

Tandem mismatches are present in duplexes of MM probes of the absent allele (MM−G'•G) with offset $\delta=1$ or $\delta=-1$. One mismatch is located at the middle position of the probe sequence and is a member of the Aa-group. The other mismatch is

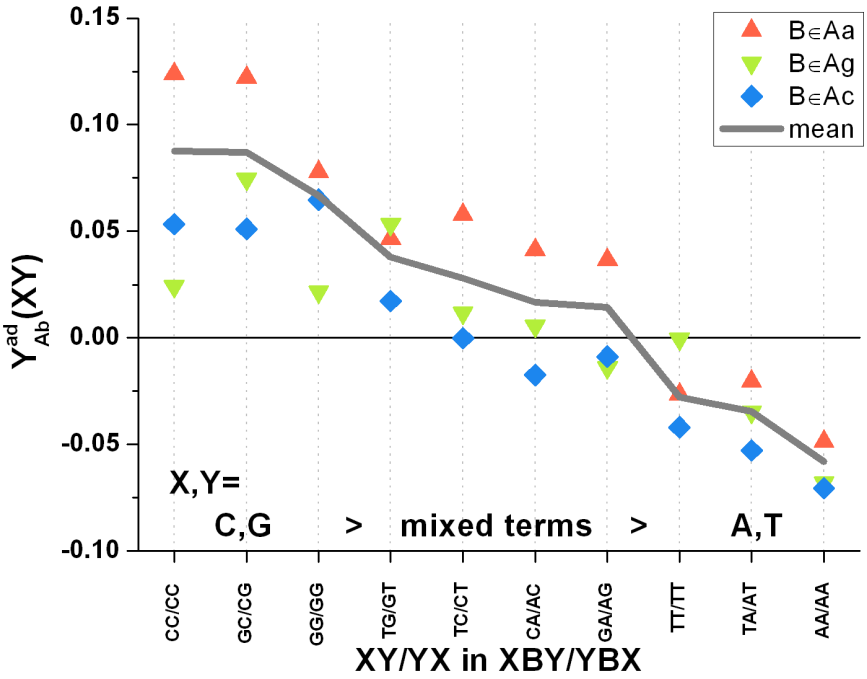


Figure 11: The effect of adjacent WC pairings in triples with a central mismatch. *Mean sensitivity values are calculated as averages of triple sensitivities over the central mismatch for each mismatch group shown in Figure 9. The obtained values characterize the mean effect of the couple XY in the triple XBY. They are ranked with decreasing mean of all three mismatch groups. It shows that X,Y=C and G give rise to largest sensitivities and standard deviation about the mean whereas adjacent X,Y=A and T cause smaller sensitivities and variability about the mean.*

located at the SNP position adjacent to the middle position and is member of either the Aa-, Ag- or Ac-group (→ Figure 12: part b, sketch). Similar to the definition of triples, the two neighboring mismatches together with their adjacent WC pairings form quadruplets YB'BX and XBB'Y for δ=−1 and δ=1, respectively. According to this convention the strand direction is ignored. B defines the mismatch of the Aa-group at the middle position whereas B' defines the mismatch of the Aa-, Ag- or Ac-group at the SNP position. The lateral WC pairings are denominated by X and Y. The necessity for quadruplet motifs to specify the stability of two adjacent mismatches was discussed previously in [48].

The sensitivities of the quadruplets are calculated using the background corrected intensities according to the three possible interaction groups of B', i.e. Ab∈{Aa,Ag,Ac}, regarding the mean log-intensity of probes having two mismatches (#mm=2) with at least one mismatch in between, $Y_{Ab}(XBB'Y)=\log\left(i^{MM-G'\cdot G}_{(Ab,|\delta|=1)}(XBB'Y)\right)-\left\langle \log i^{MM-G'\cdot G}_{(Ab,|\delta|>1)}\right\rangle$. The mean of the obtained sensitivities of the tandem mismatches are positive (→ Figure 12: part a and b, dashed lines). This indicates their larger stability in comparison to double mismatches separated by at least one WC pairing.

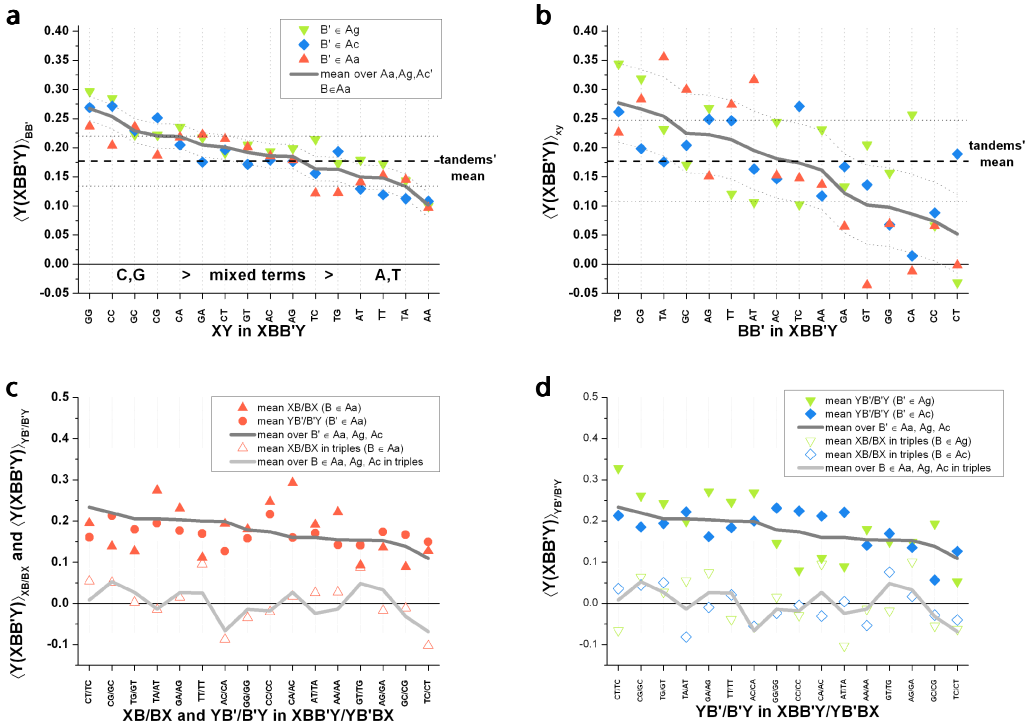


Figure 12: Sensitivities of quadruplets (XBB'Y) composed of central tandem mismatches BB' and edging WC pairings, X and Y. *The quadruplets are analyzed in terms of independent duplets of the WC couples XY (part a), of tandem mismatches BB' (part b) and of mixed NN couples XB/BX and YB'/B'Y (part c and d). Note that B refers to the Aa-group whereas B' to the Aa-, Ag- or Ac-group (→ see legends in the figure). Along the x-axis the respective pairings are ordered with decreasing mean sensitivity which is averaged over the three groups Aa, Ag and Ac of B' (→ thick decaying curve). Part a and b: The central tandem mismatches formed by B and B' cause considerably larger scattering than the adjacent WC pairings formed by X and Y. The thin dotted curves running parallel to the thick line illustrate the standard deviation of the dots about their mean (→ see also Table 3). In part c and d the respective NN terms derived from the triple motifs with single mismatches (→ Eq. 27 and Figure 15 below) are shown for comparison. The open symbols show the NN terms of the respective interaction groups and the thick gray line their mean value.*

For each alternative of B', the 256 possible quadruplet combinations can be reduced to 2•16 values by averaging either over the lateral WC bases XY or the tandem mismatches BB', $\langle \mathrm{Y_{Ab}(XBB'Y)} \rangle_{XY}$ or $\langle \mathrm{Y_{Ab}(XBB'Y)} \rangle_{BB'}$, respectively. The obtained values thus characterize the effect of the lateral base couples XY (→ Figure 12: part a) and of the tandem mismatch couples BB' (→ Figure 12: part b) on the corresponding probe sensitivities, respectively. The couples of lateral WC bases X and Y cause considerable smaller variability of the probe sensitivities than the couples of adjacent mismatches (→ Figure 12: part a). The standard deviations of the BB' couples exceeds that of the XY couples roughly by a factor of three (→ Table 3). The ratio decreases to about two regarding the scattering about the mean of the three Ab-groups, i.e. the scattering about the decaying line in Figure 12, part a and b. Hence, the particular couple of mismatches BB' mainly modulates the intensities of the probes whereas the lateral WC pairings X and Y give just rise to moderate intensity variations. This result agrees with the properties of triples with a

central mismatch discussed above, where the main source of probe intensity variation was also attributed to the central mismatch.

In part a of Figure 12 the lateral WC pairs rank according to $X,Y=C,G > X=C,G;Y=A,T > X=A,T;Y=C,G > X,Y=A,T$ and thus similar to the adjacent WC pairs of single mismatches (→ previous section and Figure 11). Both sets of mean sensitivities (→ Figure 11 and Figure 12: part a, thick lines) correlate with a regression coefficient of R=0.92.

Part b of Figure 12 indicates that a particular sensitivity value strongly depends on the combination of the two mismatches. For example, the combination BB'=CT of a relatively weak stability on the average varies between large and small sensitivities for C∈Ac and C∈Ag, respectively.

Alternatively, the quadruplets can be decomposed into two consecutive NN contributions according to XBB'Y→XB+B'Y and YB'BX→YB'+BX for δ=1 and δ=−1 by the averages $\frac{1}{2}\left(Y_{Aa}(XB)+Y_{Aa}(BX)\right)$ and $\frac{1}{2}\left(Y_{Ab}(B'Y)+Y_{Ab}(YB')\right)$, respectively (→ Figure 12: part c and d). These contributions characterize mixed contributions in accordance with the NN decomposition of the standard triples discussed below. The NN terms of both decompositions correlate with a regression coefficient of R=0.69. This result suggests that quadruplets with central tandem mismatches can be decomposed to a rough approximation into two NN terms that can be estimated also from triple data.

Flanking mismatches

Triples with flanking mismatches of the type (XBY)M (B∈At, M∈{Aa,Ag,Ac}) are selected according to the scheme shown in part b of Figure 3. These triples refer to SNP offset positions |δ|=2. The log-intensities of the respective probes are compared with the respective values of the standard triples XBY without a flanking mismatch, i.e. |δ|=3, to assess the effect of the flanking mismatch M,

$$\Delta^{\text{flank}}(XBY)=\langle \log i(XBY)\rangle_{|\delta|=3}-\langle \log i(XBY)\rangle_{|\delta|=2}\,. \qquad (26)$$

This difference estimates the mean intensity increment of the standard triple without flanking mismatches relative to that with flanking mismatches. The nomenclature assigns nucleotide Y to the position adjacent to the mismatch which flanks the triple, (XBY)M. This neighborhood relation can be realized for the triples (XBY)M and M(YBX), i.e. with the mismatch facing towards the 3' or the 5' end of the probe, respectively. In addition, in the probe and target sequence according to the complementary condition $M(YBX)\rightarrow(X^cB^cY^c)M^r$ where the superscript c denotes the WC complement and r the respective bond reversal. These, in total four, options such as (CGT)M, M(TGC), (GCA)M and M(ACG), are averaged to provide the mean effect of the flanking mismatch adjacent to Y and Y^c on the selected triple.

Figure 13 shows that the obtained mean excess values are consistently positive for Y=A,T and negative for Y=C,G. Hence, a mismatched pairing either stabilizes or destabilizes the adjacent triple in dependance on the neighboring base Y. The effect is, however, relatively weak and amounts to a few percentage of the probe intensities.

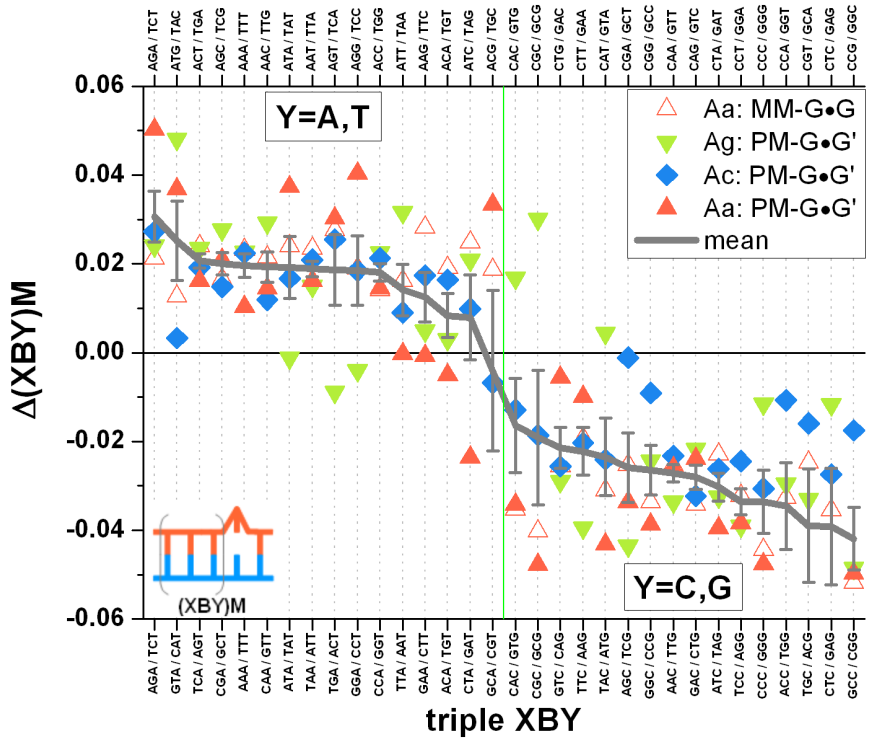


Figure 13: Excess sensitivities of triples with flanking mismatches (→ Eq. 26). *The respective probes with flanking triples are selected according to Figure 3, part b. Neglecting 3′/5′ and probe/target asymmetries, each value is calculated as mean value over the four triples indicated at the lower and upper x-axes for each mismatch group (→ symbols). The combination of triples shown at the lower axis denote the complements (XBY)M/(X°B°Y°)M and that at the upper axis M(YBX)/M(Y°B°X°). The thick line refers to the total mean over all three mismatch groups M∈Aa,Ag,Ac. The excess values are consistently positive and negative (except one) for adjacent Y=A,T and Y=C,G, respectively.*

3.10 Nearest neighbor approach

The triples are now decomposed into nearest neighbor (NN) terms in analogy to the NN free energy contributions in models describing the stability of DNA/DNA oligonucleotide duplexes in solution [45,46] and references cited therein) to see if the obtained NN terms are adequate to represent the triple terms and if they show similar characteristics as NN terms obtained in solution studies.

Nearest neighbor terms

The triple averaged sensitivities of each interaction group, $Y_{Ab}(XBY)$, can be decomposed into two nearest neighbor (NN) terms, $Y_{Ab}(XB)$ and $Y_{Ab}(BY)$, and two single base boundary contributions as

$$Y_{Ab}(XBY) = Y_{Ab}(X\underline{B}) + Y_{Ab}(\underline{B}Y) + \tfrac{1}{2}\left(Y_{Ab}(X) + Y_{Ab}(Y)\right) \qquad (27)$$

using Single Value Decomposition (SVD) [49]. The underlined B denotes the central base of the respective triple in the argument of the NN terms to avoid confusion in symmetry relations discussed below. The single base boundary terms consider the mean effect of the bases adjacent to the middle base. The triple data of each

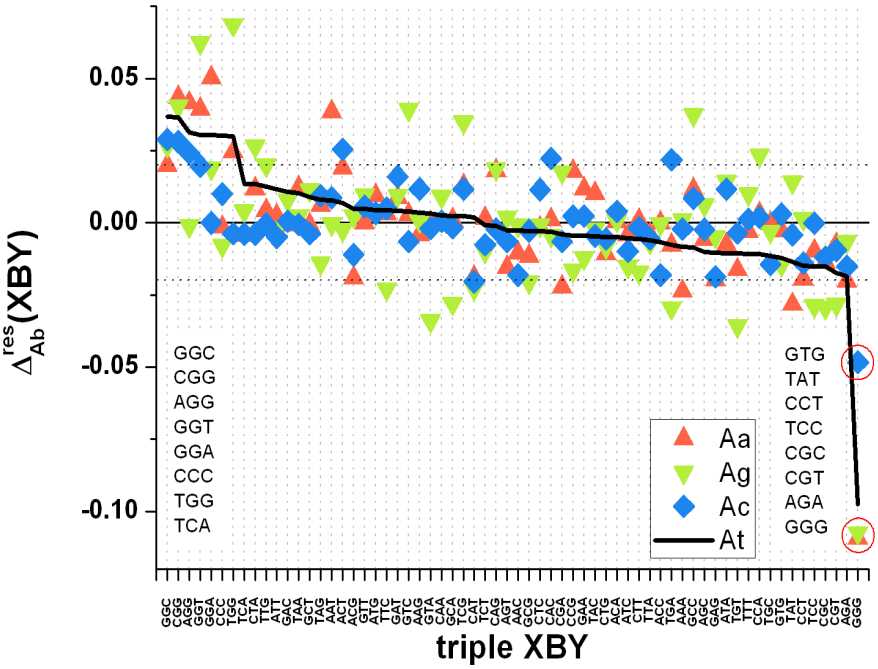


Figure 14: Residual sensitivity after decomposition of the triple sensitivities into NN terms (→ Eq. 28). *The symbols refer to the mismatch interaction groups. The triples are ranked with decreasing residual contributions of the At-group. The horizontal dashed lines mark the average standard deviation of the data about the abscissa. The two NNN lists indicate the largest positive (left list) and negative (right list) residual values of the At-group. Note that the triple GGG provides by far the largest (negative) residual contribution (→ red circles). Positive contributions are obtained for triples containing the couple GG which indicates that the respective NN terms underestimate their contribution to the triple sensitivities.*

interaction group thus define a system of 64 linear equations which is solved by multiple linear regression to determine a total of 32 NN terms and 8 boundary contributions (→ see also [33]).

The usability of the decomposition (→ Eq. 27) can be validated by means of the residual contribution

$$\Delta^{res}_{Ab}(XBY)=Y_{Ab}(XBY)-\left(Y_{Ab}(X\underline{B})+Y_{Ab}(\underline{B}Y)+\tfrac{1}{2}\left(Y_{Ab}(X)+Y_{Ab}(Y)\right)\right), \qquad (28)$$

which estimates the degree of additivity of the triple terms, i.e. the reliability of decomposition of the triples into NN terms. In the absence of interactions affecting the triple terms, vanishing residuals are expected, i.e. $\Delta^{res}_{Ab}(XBY)=0$. However, deviations from the additivity assumption (→ Eq. 27) seem to be more realistic due to the propensity of selected sequence motifs of probes and/or targets for intramolecular folding and for formation of special intermolecular complexes.

Figure 14 shows the residuals of all 64 triples per interaction group obtained after decomposition of the triple terms into nearest neighbor contributions. The standard deviation of each group is considerably smaller compared to that obtained from the

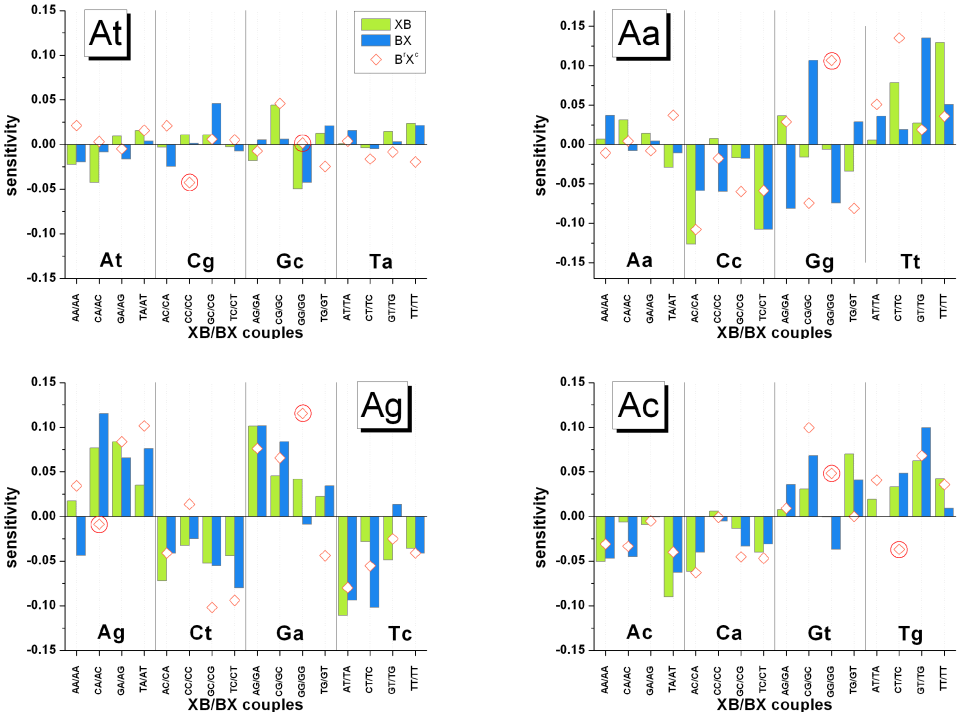


Figure 15: Nearest neighbor (NN) sensitivity terms of the four interaction groups. *The NN terms are calculated via decomposition of the triple terms using SVD (→ Eq. 27) where the base couples are ordered with respect to the center base B of the triples. The base couples are indicated as abscissa labels XB/BX (left/right bar, respectively). The symbols are the sensitivities after applying the complementary transformation to the NN terms, XB→B'X^c. NN terms related to GG motifs are indicated by red circles. They strongly deviate from the complementary condition.*

asymmetry relations (→ Table 3). This result indicates that most of the triples are additive with respect to NN terms to a good approximation.

However, motifs containing couples of adjacent GG are prone to positive deviations from additivity indicating that the respective GG term systematically underestimates the contribution of two adjacent guanines to the triple term. On the other hand, runs of three guanines, GGG, give rise to the strongest negative residual terms in all interaction groups. As the triple sensitivities $Y_{Ab}(GGG)$ are negative in all interaction groups (→ Figure 9), the observed residuals again indicate that the respective sum of two GG terms underestimates their contribution to the absolute value of the triple sensitivity, i.e. $|2\cdot Y(GG)+Y(G)|<|Y(GGG)|$ Hence, non-additivity of the considered triples is mainly introduced by the GG couples that underestimate their contribution to the respective triple terms.

Figure 15 separately shows the obtained NN terms for each interaction group and for each central base pairing of the respective triples. The NN terms are paired as XB/BX (left/right bar) to illustrate the 3'/5'-asymmetry with respect to the common base B forming the mismatch pair in the Aa-, Ag- and Ac-group. Comparison of the respective left and right bars essentially confirms the 3'/5'-asymmetry of the triple sensitivities discussed above. Likewise, the Aa and At-group show the largest and

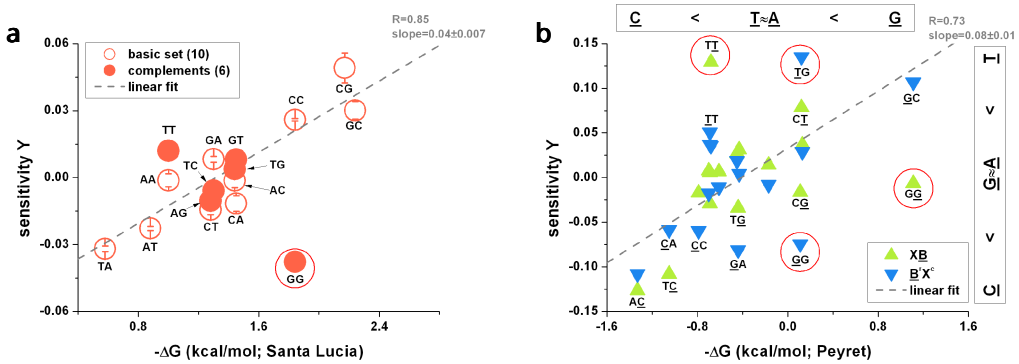


Figure 16: Comparison with solution data. *The figure shows the sensitivity NN terms of the At- (part a) and Aa-groups (part b) obtained in this thesis (→ Eq. 27) with NN-stacking free energy terms for DNA/DNA duplexes in solution taken from [46] and [44], respectively. The dashed diagonal lines are linear regressions using all NN data except GG (At-group) and in addition except TT and TG (Aa-group) which are marked by red circles (regression coefficients and slopes are given in the figure). Part a: Each NN sensitivity of couple XY was calculated as the mean value averaged over the two sensitivities with arguments XY and XY shown in Figure 15. The difference between these paired values is shown by the error bars which typically do not exceed the size of the symbol. The basic set of 10 independent terms is indicated by open circles. Part b: The complementary couples XB and B'X^c are shown by different triangles. Only selected NN motifs are assigned. The apparent mean stabilities of the mismatched pairings rank differently for chip (→ vertical bar) and solution (→ horizontal bar) data.*

smallest asymmetries, respectively. The NN data also reveal that most of the highly asymmetric base couples of the Aa-group, e.g. AC/CA, CC/CC, AG/GA, CG/GC and GG/GG, are associated with cytosines and guanines at the mismatch position.

Comparison with free energy terms describing duplexing in solution

The 32 NN couples of the At-group can be further reduced to 16 NN terms making use of the symmetry relation $Y_{At}(XY)\approx Y_{At}(\underline{X}\underline{Y})$ which however only applies to the At-group due to the equivalence of the two WC pairings associated with the nucleotide letters. Part a of Figure 16 correlates the obtained 16 averaged terms, $Y_{At}(XY)=0.5\cdot\left(Y_{At}(X\underline{Y})+Y_{At}(\underline{X}Y)\right)$, with the ten NN-free energy terms estimated in solution studies [46]. The data well correlate with a regression coefficient of R=0.85 if one ignores the GG-couple (→ Figure 16: part a, regression line). Their sensitivity values distinctly deviates in negative direction in agreement with the qualitative discussion of the residual contributions given above (→ Figure 14). The relatively large difference $Y_{At}(CC)-Y_{At}(GG)>0.06$ indicates that the complementarity between CC and GG is clearly disrupted. On the other hand, the sensitivity values of the remaining complementary couples (XY/Y^cX^c=AA/TT, CT/AG, TC/GA, AC/GT and CA/TG; see full and open symbols) are relatively close each to another (mean difference $|Y(XY)-Y(Y^cX^c)|\approx 0.01$) which justifies utilization of the complementarity condition to a good approximation. The linear regression coefficient slightly improves (R=0.92) after averaging over the complementary couples. Hence, except GG-motifs, the interactions of canonical WC pairings estimated from the probe intensities of SNP GeneChip microarrays correlate in acceptable agreement on a relative scale with free energies in solution.

Part b of Figure 16 shows an analogous correlation plot for the NN terms of the Aa-group where the solution free energies were taken from [44]. The 32 NN sensitivity terms split into 16 basic terms $Y_{Aa}(X\underline{B})$ and 16 complementary terms $Y_{Aa}(\underline{B}^{r}X^{c})$. As for the At-group, the double-guanine terms strongly deviate from the regression line and where excluded from the linear fit (R=0.65). Additional exclusion of double-thymines and $\underline{T}G$ further increases the regression coefficient (R=0.73) which indicates satisfactory correlation between solution free energy data and most of the NN sensitivities.

Note that the mean stability of self-complementary mismatches rank according to CC<TT≈AA<GG in solution but according to Cc<Gg≈Aa<Tt on the chip (→ Figure 12). Hence, Gg pairings apparently loose and Tt pairings gain stability on the chip. The stability ranking of the other mismatches except Gt essentially agrees for solution and chip data (→ see above).

Chapter 4
Tackling sequence effects

It is fundamental to understand the characteristics of SNP microarrays and the underlying processes to correct probe intensities for possible biases and, hence, to get reliable results. In my thesis I analyzed probe intensities of a 100K GeneChip SNP array regarding selected sequence motifs forming well-defined WC and mismatch base pairs in the probe/target duplexes.

This chapter gives a summary of the results, conclusions and a possible forecast on future applications of the thesis' results.

4.1 Sources of intensity modulation

The particular probe design of the GeneChip SNP arrays enables to disentangle different sources of intensity modulations such as the number of mismatches per duplex, the particular perfect match or mismatch base pairs, their neighbors, their position along the probe sequence and the relative position of a possible second mismatch. Triples of subsequent nucleotides centered about the middle base of the probe and/or about the SNP base have been chosen as the basic sequence motif. I calculated averages of the log-intensities of thousands of probes with identical triple motifs to average out the effect of the remaining sequence. These averages are measures of the stability of the selected triple in the probe sequence with the corresponding base triple in the target sequence. The former triple is defined by the probe sequence whereas the target triple can be deduced from the hybridization mode and the SNP type. I analyzed the averaged log-intensities, their difference to selected reference values, i.e. their sensitivity, and their variability in subsets of triple motifs. In addition to triple motifs, I also considered special sequence motifs such as flanking mismatches adjacent to the triples and tandem mismatches which were analyzed in terms of quadruplets including the edging WC pairs.

Various potential sources of intensity modulations have been analyzed for their impact on selected probe intensities. It turned out that

(i) the number of mismatches per probe/target duplex causes the largest effect of intensity modulation. Each mismatch changes the logarithmic intensity by $-\Delta\log I \approx 0.4-0.6$ in the case of uncorrected probe intensities. This means an intensity decrease by a factor of $0.25<F<0.4$ per mismatch. The background corrected logarithmic probe intensities change by $-\Delta\log i \approx 0.6$ with each additional mismatch, hence an intensity decrease by a factor of about $F \approx 0.25$

(ii) the effect of a mismatch is strongly modulated by the adjacent WC base pairings. They give rise to a mean logarithmic difference of $\Delta\log i \approx \pm 0.11$ or, equivalently, an average modulation factor of $0.8<F<1.3$ (→ Table 3). Selected

motifs cause even larger changes of about $\Delta\log i=\pm 0.3$ (→ Figure 9) which are almost comparable in magnitude with alterations in the number of mismatches (→ (i)).

(iii) even a WC base pair is modulated by its adjacent WC base pairings (→ Figure 9, Table 3). However, the mean variability due to these sequence effects is markedly smaller than the effect in triples with a central mismatch ($\Delta\log i\approx\pm 0.04$; $0.9<F<1.1$; → (ii)).

(iv) runs of three guanines in the probe sequence forming WC pairings represent a special motif which decreases the intensity to an exceptionally strong extent ($-\Delta\log i=0.2-0.4$; $0.4<F<0.6$; → Figure 8: part d). Also mismatched duplexes with runs of guanines possess relative small intensity values which are virtually incompatible with expected interaction symmetries in DNA/DNA duplexes (→ Figure 9).

(v) duplexes with tandem mismatches are more stable than those with two mismatches which are separated by at least one WC pair ($\Delta\log I\approx+0.18$ and $F\approx 1.5$).

(vi) flanking mismatches adjacent to the considered triples only weakly modulate their intensities ($\Delta\log I\approx\pm 0.025$; $0.94<F<1.06$).

(vii) the positional dependance of triple-averaged intensities along the probe sequence is relatively weak (→ Figure 8: part a, c and d). The sequence-specific effect progressively disappears towards the ends of the probe sequence at the final 3'-5' sequence positions for most of the motifs. Triple-G motifs partly deviate from this rule. Along the whole sequence they markedly reduce the intensity. In mismatched duplexes one observes the opposite effect at the probe end facing towards the supernatant solution.

(viii) small intensity values, e.g. from probes with two mismatches (MM−G'•G, $\delta\neq 0$), and large intensity values are especially prone to background and saturation effects, respectively (→ Figure 5). Appropriate background correction considers the optical background and in parts nonspecific hybridization as well. Saturation can be considered using the hyperbolic adsorption law (→ Eq. 19).

In the course of the analyses the request for a reduction of the considered sequence motifs arose to speed up the computation. I utilized symmetry relations and/or decomposition of the triples into nearest neighbor terms in analogy with interaction models for oligonucleotide duplexes in solution. This approach disclosed that

(ix) triples of WC pairs (At-group) can be reasonably well decomposed into NN terms which also meet the complementary condition to a good approximation and correlate well (R=0.85) with the independent NN free energy terms derived from duplex data in solution [45,46]. GGG-motifs strongly deviate from these properties and must be considered separately.

(x) the triples with a central mismatch (Aa-, Ag- and Ac-group) can also be decomposed to a good approximation into NN terms except special motifs containing at least doublets of guanines. The mismatch motifs partly obey the

symmetry relations, however, with larger residual variability compared with WC pairs. Comparison with NN terms of solution free energies [44] indicates satisfying correlation for most of the motifs (R=0.73). Runs of guanines and partly also thymine-containing motifs deviate from the expected behavior in negative and positive direction, respectively.

(xi) tandem mismatches can be decomposed into two NN terms referring to a combination of mismatch and WC base pairs. These values well correlate (R=0.69) with the NN terms obtained from the triple data suggesting to use a unified set of NN terms (→ (x)). However, the systematically larger stability of tandem mismatches compared with duplexes containing two mismatches with at least one WC pair in between has to be considered.

Relation to thermodynamics

The intensity of microarray probes is directly related to the effective bimolecular association constant of duplex formation, K^{duplex}, after correction for possible present parasitic effects such as the optical background, nonspecific hybridization and saturation (→ Eq. 2). The bimolecular effective association constant is a function of different association constants characterizing typical molecular interactions on microarrays such as dimerization of unfolded probes and targets (P•T, P•P, T•T) and unimolecular folding (P-fold, T-fold) [26] (→ see also [50]), i.e.

$$\begin{aligned} &K_{duplex} \approx K^{P \cdot T} \cdot F_{array} \text{ with} \\ &F_{array} = F_{surface} \cdot \left\{ \left(1 + K^{P-fold} + \sqrt{K^{P \cdot P}[P]}\right) \cdot \left(1 + K^{T-fold} + \sqrt{K^{T \cdot T}[T]}\right) \right\}^{-1}, \end{aligned} \tag{29}$$

where $F_{surface}<1$ is a factor taking into account surface effects such as electrostatic and entropic repulsions which effectively reduce target concentrations near the array surface. [P] and [T] denote probe and target concentrations, respectively.

According to Eq. 29, the effective constant of duplex formation is reduced by the factor $F_{array}<1$ compared with the association constant $K^{P \cdot T}$. Hence, folding and/or self-dimerization of probe and/or target get relevant at $1 < \left(K^{P-fold} + \sqrt{K^{P \cdot P}[P]}\right)$ for the probe and/or $1 < \left(K^{T-fold} + \sqrt{K^{T \cdot T}[T]}\right)$ for the target.

Stacking interactions are mainly governed by the pairings formed between the nucleotides and their nearest neighbors in the target and probe sequences. The decomposition of the corrected intensity into different interaction modes associated with single target types enables assignment of the probe sequence to canonical and mismatch base pairings with the target. The analyzed triple motifs represent a reasonable choice to study stacking interactions on an elementary level. Note that also the reduction factor F_{array} depends on the probe and target sequences, however in a more subtle fashion since, e.g., folding reactions comprise longer sequence motifs.

The duplex association constant can be multiplicatively decomposed into a triple related factor which modulates the total (average) contributions

$$\begin{aligned} &K_{duplex}(XBY) \approx k_{duplex}(XBY) \cdot K_{duplex}(\#mm) \text{ with} \\ &k_{duplex}(XBY) = k^{P \cdot T}(XBY) \cdot f_{array}(XBY) \text{ and} \\ &\log K_{duplex}(\#mm) = \left\langle \log K^{P \cdot T}(\#mm) + \log F_{array}(\#mm) \right\rangle_{Ab,\delta,XBY} \end{aligned}, \tag{30}$$

where the notations are used as introduced above. The triple related terms are denoted by lower case letters. The overall mean of the association constant mainly depends on the number of mismatches in the duplex, #mm. Modulation factor and mean value are decomposed into stacking and array terms using Eq. 29). Hence, the effective duplex association constant decomposes into a series of nested factors which consider triple motifs, stacking interactions and array specifics in different combinations.

Comparison with Eq. 17 and considering the direct relation between the corrected intensity and K_{duplex} provides the relation between the analyzed observables and the binding constants,

$$\begin{aligned} Y(XBY) &= \log k_{duplex}(XBY) = \log k^{P\cdot T}(XBY) + \log f_{array}(XBY) \\ \log i(\#mm) &\approx \log K_{duplex}(\#mm) + const. \end{aligned} \quad . \tag{31}$$

The logarithm of the association constant defines the stacking free energy of the duplex, $\Delta G^{P\cdot T} \sim -\log K^{P\cdot T}$, which applies also to the triple terms, i.e. $\Delta g^{P\cdot T}(XBY) = \Delta G^{P\cdot T}(XBY) - \langle \Delta G^{P\cdot T} \rangle \sim -\log k^{P\cdot T}(XBY)$. This definition and Eq. 31 result in

$$\begin{aligned} Y(XBY) &\propto -\Delta g^{P\cdot T}(XBY) + \log f_{array}(XBY) \\ \log i(\#mm) &\propto \langle -\Delta G^{P\cdot T}(\#mm) \rangle + \langle \log F_{array} \rangle + const. \end{aligned} \quad . \tag{32}$$

Hence, the triple averaged sensitivities are related to the deviation of the stacking free energy due to the considered triple from its mean value. This term is however distorted by an array term originated by folding, self-duplexing of target and probe and by specific surface effects. The former contributions are also functions of the sequence position of the chosen triple which is not explicitly expressed in Eq. 32 for sake of convenience. Note also that imperfect probe synthesis potentially reduces the real length of the oligomers in a motif-specific fashion with possible consequences for the observed triple sensitivities [26].

Sensitivity and free energy are opposing variables, i.e. larger stability of interactions is associated with higher sensitivity Y but less (more negative) free energy ΔG. After decomposition into NN terms, the correlation between the estimates from chip data and solution data taken from the literature have been found to be acceptable for most of the motifs (→ Figure 16). Thus, I conclude that chip effects are of small importance on the average, i.e. $\Delta g^{P\cdot T}(XBY) \gg \log f_{array}(XBY)$, and stacking free energies well represent the relation between the particular terms on a relative scale.

The proportionality constant in Eq. 32 is estimated by the slope of the regression lines in Figure 16. Their values are with $(0.4–0.8)\cdot 10^{-1}$ roughly one order of magnitude smaller than the proportionality constant predicted by the thermal energy $\sim(1/RT\cdot\ln 10) \approx 0.7$ ($T \approx 40\,°C$). In [26] it was argued that nonlinear (in logarithmic scale, as, e.g., predicted by Eq. 29) and sequence dependent contributions to $\log f_{array}(XBY)$ can cause proportionality constants less than unity. Sequence independent sources of intensity variability such as the fragment length of the genomic targets [22,51] (not considered here) are potential causes of the downscaling of the proportionality constant. Interestingly, the proportionality constant obtained for the mismatch pairings (Aa-group) exceeds that for the WC pairings (At-group) by

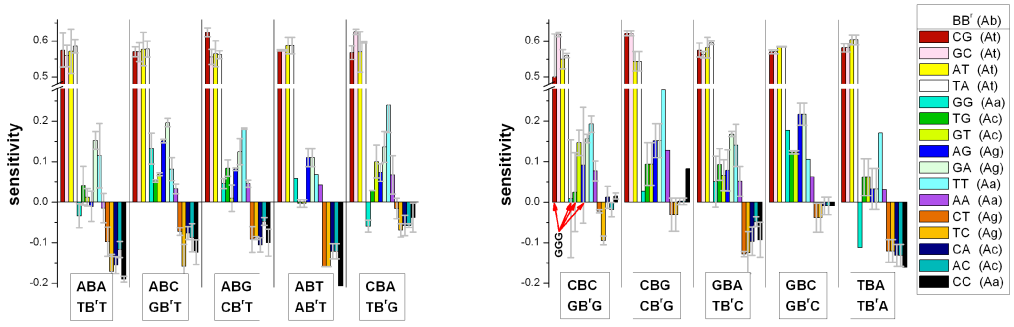


Figure 17: Stability of mismatch motifs. *Relative stabilities of the 10 possible contexts of complementary triples containing the 16 possible central base pairings (mismatches or Watson-Crick base pairs, → see legend in the figure). The sensitivities of the pairs of complementary triples XBY/Y'B'X' (B'=B') are averaged using the triple data shown in Figure 9. The error bars indicate the difference between the individual values and thus they quantify the deviation from complementary symmetry. The form of the bar diagram was chosen in correspondence with Figure 3 in [44] which ranks the stacking free energies of each triple in solution duplexes with decreasing stability (from left to right for each triple). The mean log-intensity increment of one mismatched pairing (see Figure 2) was added to the triple values of the At-group to compare the stabilities of WC and mismatched pairings in a unique scale. The sensitivities of the four triple combinations in the GGG context are exceptionally small (→ red arrows).*

a factor of two (→ Figure 16: part a and b). This difference suggests that the larger sensitivity response of the probes to mismatch pairings (compared with WC pairings) is not simply related to the variability of the respective stacking free energies but includes other effects related to the array technology.

Mismatch stability

The mismatch stability is influenced amongst others by the efficiency of the formation of hydrogen bonds between the paired nucleotides and their strengths, by steric factors such as the size of the aromatic moiety, i.e. one ring in pyrimidines (C, T) and two rings in purines (G, A), as well as stacking effects associated with nearest neighbors.

The stabilities of most of the mismatch pairs (→ Eq. 23) rank similarly as the results of previous chip and solution studies (→ Eq. 24). Figure 17 shows the detailed stability trend in all 10 possible contexts of triple pairings XBY/Y^{c}B'X^{c} with all 16 possible BB' pairs referring to Bb^{r} and B'b'^{r} with b^{r}=B' and b'^{r}=B, respectively. The figure was designed on Figure 3 in [44] which compares the stabilities of mismatches in solution (→ Eq. 24). The error bars indicate the difference between the individual values and thus quantify the deviation from probe/target complementarity (→ see also Figure 10). As in solution, the overall trend in pairing stability is context dependent. Essentially two groups of larger and weaker mismatch stabilities can be clearly distinguished for BB': (TT, GA, AG, GT, TG, AA, GG) > (AC, CA, CT, CC, TC) (→ detailed ranking in Eq. 23). In 5 of 10 contexts of complementary triples TT has the highest mismatch stability followed by GA which has the highest mismatch stability in the remaining cases. Mismatch pairs containing cytosine are predominantly of weaker stability (AC,CA,CT,CC,TC) with TC having the lowest in 5 contexts. The CTG/CTG triple pairing is the most stable of all the possible mismatch triple pairings whereas the ACT/ACT pairing, as in the solution study, is the weakest

one. Most of the triple intensities are modulated by the bases adjacent to their central base following the mean trend shown in Figure 11 (G,C>A,T), i.e. CBG, GBC, CBC and GBG strengthen the mismatch stability whereas ABT, TBA, ABA and TBT weaken it on average. Triple pairings with a GGG triple (→ Figure 17: red arrows) show extreme perturbation from probe/target complementarity as indicated by the large error bars, i.e. independently of the interaction group Ab, triple pairings with a GGG triple in the probe strand consistently decrease the mismatch stability. However, with a GGG triple in the target strand the mismatch stability is increased. Some other triple pairings also show a perturbation of probe/target complementarity but smaller by far (→ Figure 10).

The number of hydrogen bonds and their strengths definitely have an influence on the mismatch stabilities but the triple sensitivities (→ Figure 9) and the considerations of the triple pairings (→ Figure 17) allow no clear conclusions to be drawn about number and strength of hydrogen bonds between non-canonical bases. The stacking free energies in Figure 3 of [44] together with the comments of Peyret *et al.* [44] also yield no paradigm related to this. Ikuta *et al.* [43] raise the hypothesis that stable mismatch base pairs such as GT or GA form two hydrogen bonds and only slightly disrupt the structure of the oligonucleotide-DNA duplex. Further they suppose that unstable mismatched base pairs such as CT or CA significantly disrupt the duplex structure due to the small size of the pyrimidine/pyrimidine pairing or the disability to form at minimum two hydrogen bonds because of the lack of imino protons. Also the self-complementary single ringed CC mismatch has a low stacking propensity and forms only one hydrogen bond [44]. This rationalizes the low stability of the mismatches formed by cytosines in agreement with the chip data.

The second self complementary single ringed TT mismatch is, in contrast to CC, however stabilized by two hydrogen bonds [44] and shows high stacking potential. The two purine/purine self complementary mismatches GG and AA have intermediate stability compared to TT and CC. GG forms either one or two (weak) hydrogen bonds [44]. AA has only weak or no hydrogen bonding [52]. One expects therefore the stability series CC<AA≈GG<TT which is confirmed by the chip data and in [38]. Solution experiments [44] and that of others [28] slightly disagrees with the chip data (→ see also Eqs. 23 and 24). An analogous low stability of GG mismatches on microarrays compared with solution data was reported for DNA/RNA hybridizations [39]. It has been concluded that thermodynamic properties of oligonucleotide hybridization are by far not yet understood and not suited to assess probe quality.

Poly-guanin motifs

The triple average analysis shows that the low stability of Gg mismatches is accompanied with triple-G motifs in the probe sequence. The stability of central Gg pairings in the context of adjacent bases others than G, on the other hand, roughly agrees with the predictions from solution data (→ Figure 16). Runs of guanines are not only associated with low intensities in triples of the Aa-group, but also with mismatch pairings of the Ag- and Ac-group, and even with central WC pairings of the At-group.

The analyses reveal the following effects of triple-G motifs in the probe sequence on the observed probe intensities:

(i) The GGG-effect is non-complementary, i.e. the complementary triples (e.g. CCC for perfect matches) does not show exceptional small intensities as probes with GGG do.

(ii) Exceptional small intensities are observed for triple-G motifs with a central perfect match or central mismatch independent of the nominal pairing of the central base (→ Figure 17, arrows indicate the GGG-associated motifs in CBC/GB'G with BB'=CG, GG, TG, AG).

(iii) The effect is non-additive, i.e. the intensity drop due to GGG is inconsistent with the decomposition into GG contributions in the context of all triple motifs.

(iv) The effect depends on the sequence position that is typically smaller near the ends of the probe sequence (→ Figure 8).

(v) For probes with one mismatch pairing one observes, in contrast to (iv), that probes with terminal GGG at the solution end gain intensity, i.e. the sign of the effect reverses compared with the remaining sequence positions.

(vi) The intensity drop due to one triple-G corresponds roughly to 50% of the intensity loss due to one mismatched pairing (→ Figure 8)

The observations (i) and (ii) strongly indicate that the triple-G effect is not associated with the nominal base pairings deduced from the binding mode, otherwise equal intensity changes could be expected for complementary sequence motifs. Observation (iii) indicates that the effect exceeds the range of stacking interactions with the nearest neighbors. Observation (vi) shows that the magnitude of the effect is markedly large compared with the variability due to other base-specific effects but smaller than variability due to single mismatches.

To have a closer look on the properties of runs of homologous motifs I calculated the mean sensitivity for runs of one to five identical bases, e.g. G, GG, …, GGGGG, averaged over all sequence positions of homozygous present PM probes (P–G•G, → Figure 18 and Figure 8). The sensitivities of all considered runs fit along straight lines with similar absolute values of their slope for adenines, thymines and cytosines (→ Figure 18). The slope characterizes the mean sensitivity change which, in turn, estimates the stability change per WC pair in the motif compared with the mean stability of all canonical base pairs. The absolute value of the increment of single- and double-G motifs roughly agrees with that of the other bases (→ Figure 18). It however steeply increases by more than one order of magnitude for poly-G of length greater than two. Obviously, this change cannot be attributed to the incremental effect of additional WC pairings in agreement with observations (i) and (ii) but, instead, it supposably reflects the formation of another structural motif accompanied with an increased intensity penalty per additional guanine in the run.

Previous studies also reported abnormal intensity responses of probes containing multiple guanines in a row (called G-runs or G-stacks) compared with other probes in different chip assays including Affymetrix expression and SNP arrays [22,29,30,53,54]. It was found in agreement with the results of this thesis that the effect is asymmetric with respect to complementary C-stacks [54,29] and depends on the sequence position of the stack with a very strong amplitude at the solution-end position [29]. Note that on expression arrays poly-G containing probes show the

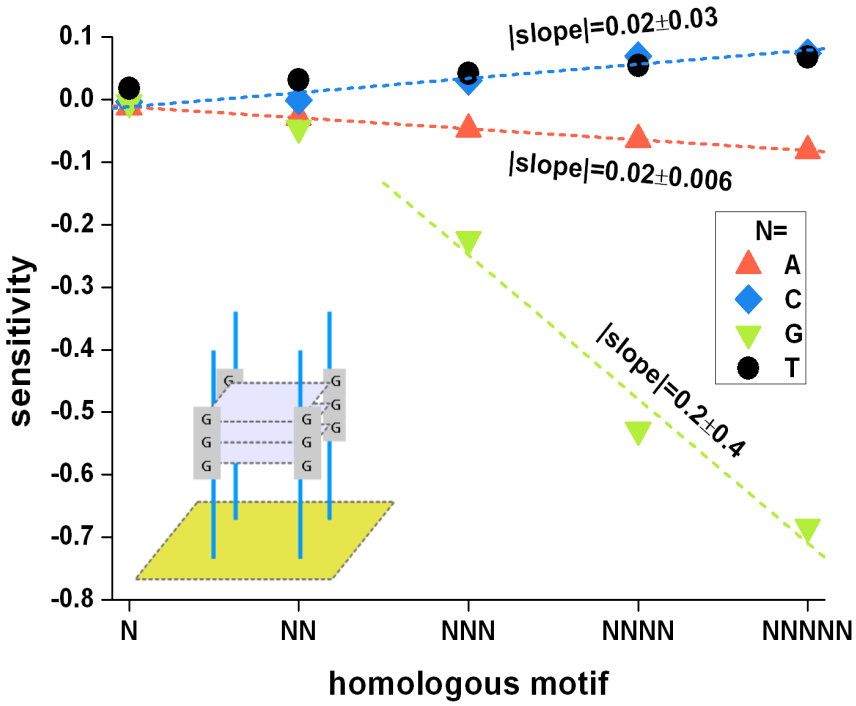


Figure 18: Sensitivities of runs of identical bases. *The sensitivity values are averaged over all sequence positions of homologous motifs of length 1 to 5 of homozygous present probes (PM-G•G, → see also Figure 3: part b). Adenines, cytosines and thymines follow straight lines. Their slope is related to the mean stability increment per additional WC pairing in the runs. For guanines the absolute value of the slope drastically increases by more than one order of magnitude for longer poly-G runs exceeding two adjacent G. This effect is attributed to the formation of stacks of at minimum three G-tetrads (G4). The sketch within the figure illustrates the structure of parallel quadruplexes formed by four neighbored probe oligomers with GGG-runs at the same sequence position. They are assumed to aggregate into three G4-layers.*

opposite tendency as on the studied SNP arrays: They shine relatively bright with intensities exceeding the expected signal level [29,55]. This opposite trend of abnormal strong intensities is associated with nonspecific hybridization [29].

The structural reason of the poly-G effect has been assigned concordantly to the propensity of poly-G motifs to arrange in stacks of stable molecular bundles of guanine tetrads. These structures potentially affect the efficiency of oligonucleotide synthesis and/or the hybridization of the target sequences to the appropriate probes and account for the abnormal performance of G-runs on the array [29,30,53,54]. Each G-tetrad is held together by eight Hoogsteen hydrogen bonds and is further stabilized by monovalent cations reducing the electrostatic repulsion between the nucleotides. At least three of such planar G-tetrads usually stack together forming very stable complexes by re-folding of one DNA strand with several poly-G motifs [56,57] or by aggregation of several DNA strands each containing one poly-G motif (parallel G-quadruplexes, → Figure 18: sketch). It has been conclusively argued that probe oligomers in close proximity containing poly-G motifs at the same sequence position are prone to aggregate into such parallel G-quadruplexes in the crowded conditions

on the surface of high density microarrays [30,58]. The length of 25-meric probes (~22 nm) largely exceeds the average separation between neighboring oligonucleotides on such arrays (~3 nm). This enables complexation of four adjacent probe strands as schematically illustrated in Figure 18. The onset of the stronger sensitivity decrement per additional guanine for triple-G motifs shown in Figure 18 supports the hypothesis that tree layers of G-tetrads represent the minimum for stable G-quadruplexes.

As mentioned above, there are two dimensions which potentially affect the performance of probes containing poly-G motifs: firstly, their ability to be correctly synthesized on an array, and secondly the ability of correctly synthesized probes to bind its target.

The first option: The GeneChip arrays are fabricated by *in situ* light-directed combinatorial synthesis on the surface of the array which is prone to produce 5'-truncated products but not internal deletions [59-61]. One can suggest that the synthesis yield per nucleotide is reduced in poly-G runs of length greater than two compared with the average synthesis yield possibly because the formation of G-quadruplexes between neighboring probes affects photo-deprotection of the partly synthesized oligonucleotides. As a result of incomplete synthesis the oligonucleotide features are contaminated with probe sequences which are truncated at the nominal position of the poly-G motif. The probability and thus also the number of such truncated probes is expected to increase with the length of the poly-G motif according to the synthesis yield per additional guanine. Truncated probes of length less than 22-20 nucleotides can be assumed to act as weak binders for the targets. Their binding affinity roughly refers to that of full-length probes with more than two mismatches (→ Figure 6 and also [26]). The truncated oligomers only weakly contribute to the intensity of the probe spots in mixtures with full length probes at low and intermediate target concentrations. As a result, the observed intensity drop of poly-G containing probe sequences is the result of the reduced number of full length probe oligomers in the respective probe spots. Their fraction can be approximately estimated by assuming proportionality between the intensity drop and the remaining number of full length probes of about $10^{Y(GGG)}$<0.4-0.5 for GGG motifs (with Y(GGG)=−0.2…−0.3; → Figure 6 and Figure 18). This fraction is equivalent with the effective synthesis yield per additional G of 40%-50% which roughly halves the number of remaining full length probes according to our data. The general effect of incomplete probe synthesis on the hybridization of microarrays has been discussed in [62] and [26].

Also the second option of modified target binding to correctly synthesized probes provides a tentative explanation of the GGG-effect [58]. It assumes that complex formation between the probe oligomers effectively blocks the involved probe strands and this way reduces the amount of free binding sites accessible for the targets. In consequence, their effective association constant is expected to decrease (→ Eq. 29). The probe-probe interaction term in Eq. 29 simply assumes bimolecular interactions between the probes. Substitution by an appropriate higher-order interaction term considering the stoichiometry of quadruplex formation, the proximity relations and fixation of the probes on the chip surface is expected to modify the respective contribution but leaves the expected trend unchanged.

Note that both discussed potential interpretations of the GGG-effect give rise to a common cause of the observed small intensity values, namely the reduced number of available binding sites for target binding either via truncation or via complexation of part of the probe oligomers. Both interpretations are compatible with observations (i) and (ii) because the reduced amount of full-length probes and probe-probe complexes are independent of the respective complementary target sequence upon allele-specific hybridization and independent of the respective mismatch target sequence upon cross-allelic hybridization. Also the onset of the increased sensitivity increment per additional guanine for triple-G motifs shown in Figure 18 supports both hypotheses because stable G-quadruplexes of the probes are assumed to affect synthesis and hybridization as well.

Tethering of the involved oligonucleotides to the surface and zippering effects towards both ends of the probes are expected to modify their propensity for G-tetrad formation in a positional dependent fashion in analogy to the positional dependance of base pairings in probe/target dimers [22,26,32,63-65] This trend provides a rational for effect (iv). However, probe-probe interactions modulate target binding by the array factor $F_{array}<1$ (→ Eq. 29). The GGG profile of homozygous absent probes (P−G'•G, → Figure 8: part d) shows the typical characteristics of the mismatch pairing in the middle of the sequence. This result indicates that a certain fraction of the oligomers of the respective probe spot is not fixed in G-tetrads but, instead, form specific dimers with the cross-allelic or allele-specific target as expected for the respective hybridization mode. This result is in agreement with both hypotheses discussed because incomplete synthesis and probe-probe complexes reduce but not prevent specific hybridization.

The suggested mechanisms explains the decreased intensity of probes containing runs of consecutive guanines. The effect (v) however seems puzzling because terminal poly-Gs increase the intensity of the respective probes, instead. On expression arrays even much stronger intensity gains for terminal poly-G containing probes are observed [29,55]. For expression arrays this opposite trend of abnormal strong intensities is clearly associated with nonspecific hybridization. I suppose that G-rich probes are able to form G-quadruplexes of different stoichiometry together with nonspecific targets containing longer runs of guanines in a positional dependent fashion with a strong bias towards the solution end of the probe. For SNP arrays the relative contribution of nonspecific hybridization is relatively weak compared with expression arrays, which explains the relative weak effect of bright poly-G motifs near the solution end of the probe sequences. Also the fact that effect (v) becomes evident only for relatively weak signals of probes forming at least one mismatched pairing is compatible with an additive contribution due to nonspecific binding (Eq. 2). At larger probe intensities, nonspecific binding becomes insignificant compared to specific binding. Upton *et al.* suggested yet another alternative mechanism which increases the intensity of poly-G containing probes by local opening of regions in the vicinity of quadruplexes [30].

In summary, the data support the hypothesis that runs of consecutive guanines facilitate the formation of stable G-quadruplexes between neighboring probes which in final consequence reduce the number of probe oligomers available for target binding via two alternative mechanisms, firstly, the reduced synthesis yield of full length probes and/or, secondly, the formation of complexes of neighboring full-length

probes. Both hypotheses are compatible with the observed intensity drop of probes containing runs of guanines on SNP arrays.

About 11% of all probes on the studied 100K GeneChip SNP array contain at least one GGG-run and nearly 30% of the allele sets contain at least one of these probes. Hence, GGG-runs are relatively common on SNP arrays and the discussed effect has to be considered in appropriate correction methods.

4.2 Consequences of sequence effects

Probe intensities of SNP arrays are intended to accurately and completely measure SNP genotype and copy number variants. Each SNP locus is interrogated by two sets of probes for each of the two possible alleles (B_A, B_B). Ideally, the probes for both alleles provide a binary signal $[f_A,f_B]$ of allele fractions with $[f_A,f_B]=[1,0]$ or $[0,1]$ in the simple case of homozygous genotypes and $[f_A,f_B]=[0.5,0.5]$ in case of heterozygous genotypes. The fractions are defined as $f_A=CN_A/CN$ and $f_B=CN_B/CN=1-f_A$ where CN_A and CN_B denote the copy numbers of the respective alleles and $CN=CN_A+CN_B$ the total copy number of both alleles. However, real probe signals are distorted by sequence specific allelic imbalances as both alternative alleles differ in only one nucleotide at the SNP interrogating position. Their close similarity gives rise to strong correlations between the two allele signals. This potentially causes biased genotype and copy number estimation and in final consequence lead to systematic errors in downstream analyses such as genome wide association studies [66,14]. A number of calibration methods have been developed to transform biased probe level signals into reliable genotyping and copy number information, e.g. [2], [5-7] and [9-11]. These preprocessing algorithms are however not without limitations due to insufficient corrections that have implications for downstream analyses (→ see e.g. [6] for a critical overview). Given the vast number of genotypes being produced, a systematic bias, even if very small, may lead to spurious association signals [14].

Incremental binding strength of the probes

As reference level of the probe intensities, I assume the mean intensity of all probe/target duplexes of homozygous genotypes on the chip having one mismatch pair (#mm=1), namely MM–G•G, PM–G•G' and MM–G•G' (δ=0). The mean binding strength is $X_{\#mm=1}=K_{\#mm=1}\cdot\langle[P-G\bullet T]_{\#mm=1}\rangle$ (→ Eq. 7) where $K_{\#mm=1}$ denotes the mean binding constant of the reference probe/target duplexes and $\langle[P-G\bullet T]_{\#mm=1}\rangle$ is the mean concentration of all genomic fragments of both alleles interrogated by the reference probes.

The partial binding strength of the binding reactions (→ Eq. 7),

$$X^{P-G\bullet T}=\delta X^{P-G\bullet T}\cdot X_{\#mm=1}\,, \tag{33}$$

scales with the mean binding strength of the reference level and the incremental contribution

$$\delta X^{P-G\bullet T}=\delta K^{P-G\bullet T}\cdot R^{P-G\bullet T} \tag{34}$$

which considers the sequence and SNP specifics of probe/target interactions in the respective interaction mode. $\delta K^{P-G\bullet T}$ represents the increment of the binding

constant relative to the reference value $K_{\#mm=1}$ due to, e.g. the particular sequence of the chosen probe. The stoichiometrical ratio is defined by

$$R^{P-G\cdot T}=R_{CN}\cdot f_T \text{ with } R_{CN}=\frac{[A]+[B]}{\langle [P-G\cdot T]_{\#mm=1}\rangle}\sim\frac{CN}{CN_{\#mm=1}} \tag{35}$$

where R_{CN} signifies the ratio of the total copy number of the SNP interrogated by the particular probe divided by the mean copy number of all reference probes, and f_T is the fraction of genomic copies of the alleles $T\in\{A,B\}$, i.e. $f_A=CN_A/CN$ and $f_B=CN_B/CN=1-f_A$, respectively. The copy numbers of the individual alleles, CN_T, meet the condition of material balance: $CN=CN_A+CN_B$. For a homozygous absent, heterozygous or homozygous present diploid allele A one would expect $f_A=0$, 0.5 or 1, respectively.

The total incremental binding strength of the S- and C- hybridization modes is $\delta X^{P-G}=\delta X^{P-G\cdot G}+\delta X^{P-G\cdot G'}$. Making use of Eqs. 34 and 35 and of symmetry conditions one obtains the incremental binding strengths for the probes interrogating allele A and B as a function of the relative copy number and of the fraction of the genomic copies,

$$\begin{aligned}
\delta X^{P-A}&=R_{CN}\cdot\left(f_A\cdot\delta K^{P-A\cdot A}+\left(1-f_A\right)\cdot\delta K^{P-A\cdot B}\right)\\
\delta X^{P-B}&=R_{CN}\cdot\left(f_B\cdot\delta K^{P-B\cdot B}+\left(1-f_B\right)\cdot\delta K^{P-B\cdot A}\right)\\
&=R_{CN}\cdot\left(f_A\cdot\delta K^{P-B\cdot A}+\left(1-f_A\right)\cdot\delta K^{P-B\cdot B}\right)\\
\delta X^{P-G}&=R_{CN}\cdot\left(f_G\cdot\delta K^{P-G\cdot G}+\left(1-f_G\right)\cdot\delta K^{P-G\cdot G'}\right)
\end{aligned} \tag{36}$$

The incremental binding strengths of the probes δX^{P-G} estimate the maximum measurable allele-specific effect and thus the sensitivity of the method whereas the relation between δX^{P-A} and δX^{P-B}, their so-called crosstalk, characterizes the specificity to discriminate between the genotypes. Both characteristics are governed by two incremental binding constants for each probe which modulate the signals and thus also their sensitivity and specificity in a probe specific fashion. This probe-specificity depends on the bases of the particular SNP type and on the neighboring nucleotides in the probe sequence context.

Sequence effect estimation

In the triple average analysis made in this thesis I considered the triple sequences from the probe point of view designated by XBY. The following consideration will be made from the genomic target point of view designated as xby. The target triple can be obtained from the probe/target duplex YBX/xby using $xby=X^cB^rY^c$ where b is the WC complement or a mismatch base at the SNP position in the target sequence. x and y denote the neighbor bases and are the WC complements of X and Y. The target point of view allows for a direct connection between the target allele letters (b_A, b_B) and the respective SNP type B_A/B_B.

The effect of the base pairing formed by a particular SNP allele and its nearest neighbors within the probe/target duplex can be also examined in terms of triple average sensitivities (→ compare with Eq. 17),

$$Y(Y^cB^rX^c)=Y(xby)=\log i(xby)-\log i_{\#mm=1}\,, \tag{37}$$

taking the mean log intensity, $\log i_{\#mm=1}$, averaged over all probes of the selected chip with one mismatch pair (#mm=1) in their probe/target duplexes as reference level. The intensities are corrected for the nonspecific and optical background indicated by the lower case letter (→ Eq. 20). According to Eq. 37, the triple sensitivities constitute the mean difference between the log intensities of the respective probes which form duplexes with the target triple motif xby (→ Eq. 15) and the mean of the intensities of the reference set.

Using the Eqs. 9 and 7, one can write $i_{\#mm=1} \approx X_{\#mm=1}$ and analogously for the triple related intensities $i(xby) \approx i_{\#mm=1} \cdot \delta X^{P-G}(xby)$. The latter equation transforms into $i(xby) \approx i_{\#mm=1} \cdot \delta X^{P-G \cdot T}(xby)$ for the special case of homozygous present (P–G•G) and homozygous-absent (P–G•G') probes. Insertion into Eq. 37 and assuming that the mean copy numbers in the sub-ensembles of probes with a given triple motif equals the copy number of all considered probes, one gets the triple specific log-increments of the binding constants in the allele specific and cross allelic duplexes of the PM and MM probes,

$$\begin{aligned} &\log \delta K^{PM-G \cdot G}(xby) \approx Y_{At}(xby) + \log(s_{01}) \quad \text{with } s_{01} \equiv i_{\#mm=0}/i_{\#mm=1} \\ &\log \delta K^{PM-G \cdot G'}(xby) \approx Y_{Ab}(xby) \end{aligned}, \tag{38}$$

and

$$\begin{aligned} &\log \delta K^{MM-G \cdot G}(xby) \approx Y_{Ab}(xby) \\ &\log \delta K^{MM-G \cdot G'}(xby) \approx Y_{Ab}(xby) - \log(s_{12}) \quad \text{with } s_{12} \equiv i_{\#mm=1}/i_{\#mm=2} \\ &\log \delta K^{MM-G \cdot G'}(xby) \approx Y_{Ab}(xby) \quad \text{for offset } \delta = 0 \end{aligned}, \tag{39}$$

respectively, with Ab∈{Aa,Ag,Ac}. The terms $\log(s_{01})$ and $\log(s_{12})$ consider the difference between the reference level and the mean intensities of probes with no mismatches in the S-mode (PM–G•G, #mm=0) or probes with two mismatches in the C-mode (MM–G•G', #mm=2), respectively (→ Table 1, Table 2).

SNP specific intra- and inter-allelic correlations

Both alternative alleles differ by only one nucleotide in the position which corresponds to the SNP locus in the genome. Their close similarity gives rise to strong correlations between the two allele signals. The resulting intrinsic crosstalk pattern is specific for each of the six possible SNPs due to the sequence context of the respective SNP. It can be obtained using the triple sensitivities (→ Eq. 37) and the incremental binding strengths given by Eq. 36. Accordingly, each background-corrected and normalized probe signal constitutes the superposition of contributions due to allele-specific (P–G•G) and cross-allelic (P–G•G') hybridization modes. For both alleles these modes give rise to four contributions of the incremental binding strength, $\delta X^{P-G \cdot T}$, of the P–A•A, P–A•B, P–B•B and P–B•A complexes for both PM (→ Eq. 38) and MM probes (→ Eq. 39) as illustrated in Figure 19 using the example of the triple context of the C/T-SNP. The four hybridization terms are functions of the triple sensitivities of the SNP loci.

Figure 20 shows the 12 intra- (→ part a) and 12 inter- (→ part b) allelic correlation patterns induced by the six SNPs probed on the GeneChip arrays, $B_A/B_B \in$ {A/C, G/T, A/G, C/T, A/T, C/G}. Note that the coordinate origin defines the

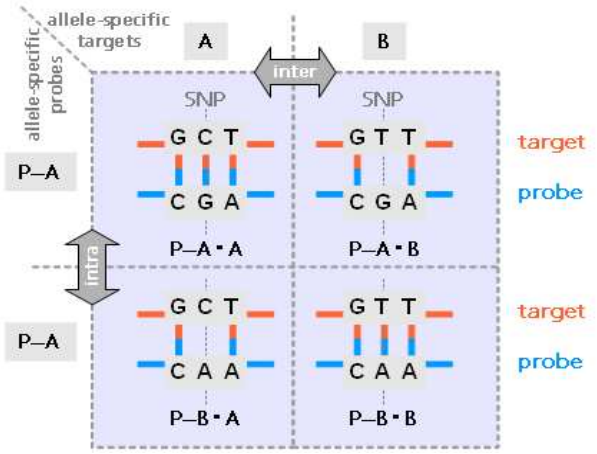


Figure 19: Triple-related interaction modes in probe/target duplexes. *The two probes interrogate either allele A or B of the C/T-SNP as example. The figure shows the base pairings at the SNP position and of the two adjacent bases upon allele-specific (P–A•A and P–B•B) and cross-allelic (P–A•B and P–B•A) hybridization. The intra-allelic correlation is governed by the same target triples whereas the inter-allelic correlation is characterized by the same probe triples.*

unbiased relation between the paired signals because the sensitivity values are centered about zero according to their definition (→ Eq. 37). The considerable scatter of the triple data shown in Figure 20 about the origin consequently indicates their systematic bias due to particular base pairing at the SNP position and their nearest neighbors.

The correlation between two probe/target duplexes of the same target allele constitutes the intra-allelic crosstalk, i.e. PM–G•G ↔ PM–G'•G, whereas the correlation between two probe/target duplexes with different target alleles constitutes the inter-allelic crosstalk, i.e. PM–G•G ↔ PM–G•G' (→ see Figure 19 for illustration). The intra- and inter-allelic crosstalks consequently characterize the intrinsic correlations between the two hybridization modes of duplexes having the same triple either in the target or the probe sequence, respectively.

Neglecting the GGG-motifs, the inter-allelic crosstalk reveals strong positive and negative correlation between allele-specific and cross-allelic signals with regression coefficients R≈±0.46 for six of the alleles, weak positive and negative correlation for four of the alleles with R≈±0.24 and no or negligible correleation for two of the alleles (→ Figure 20: part b). The predominant strong correlation can be probably attributed to the identical base triples in the probe sequence. The central WC pair in the allele-specific duplex changes to a mismatch in the cross-allelic duplex due to the altered allele base at the SNP position of the target sequence (→ Figure 19).

For interactions of complementary symmetry one expects analogous correlation pattern of intra- and inter-allelic crosstalks for swapped SNP bases. The data clouds in part a and b of Figure 20 indeed meet this prediction in a first order approximation. However, deviations from this symmetry are also evident. The intra-allelic crosstalk reveals considerably smaller regression coefficients with no or negligible correlation for five of the alleles. Two of the alleles show moderate positive correlation with R≈0.37 and one allele shows weak negative correlation with R≈–0.25. Only three of the alleles show strong positive and negative correlation (R≈0.58 and

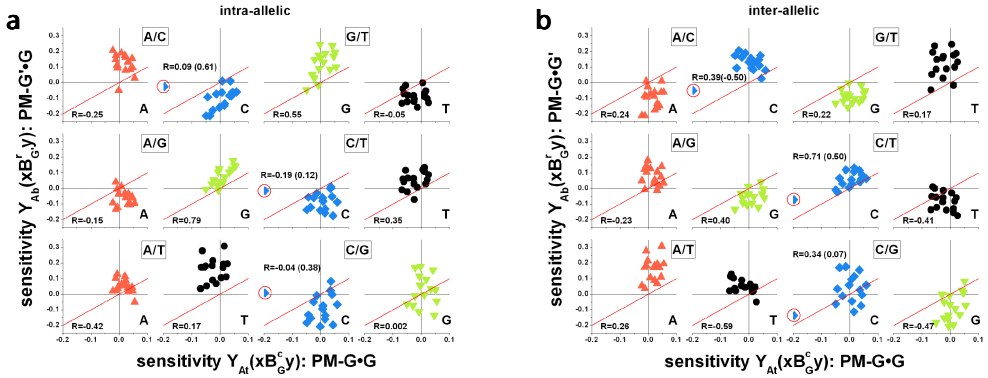


Figure 20: Correlation patterns of the triple sensitivities. *Referring to the six possible SNP-types B_A/B_B interrogated on GeneChip SNP arrays part a and part b show the intra- and inter-allelic correlation patterns. The diagonal dotted lines are shown as guide for the eye to visualize the correlation in each plot. The numbers are the respective regression coefficients (regression lines are not shown) whereas the values in brackets are obtained after omitting the GGG-motif. The intra-allelic correlations refer to the P–G•G versus P–G'•G hybridizations of targets of the same allele to probes interrogating different alleles and the inter-allelic correlations refer to P–G•G versus P–G•G' hybridizations of targets from different alleles to the same probe (→ see also Figure 19 and Eq. 37). Note the symmetry between part a and b upon swapping the SNP base by its WC complement. The sensitivities of GGG-motifs are indicated by red circles.*

R≈−0.42) and one allele shows very strong correlation (R≈0.79) (→ Figure 20: part a). The change of the center base in the probe triples obviously causes smaller correlations in a series of cases compared with the change of the center base in the target triples.

Especially GGG-motifs in the probe sequence clearly violate the expected symmetry for reasons discussed above (→ Figure 20: red circles). This GGG-bias has been attributed to self-complexing of the probes via quadruplex formation (→ see above and [58]) which gives rise to effectively asymmetric probe/target interactions. The biases in the inter-allelic correlation pattern is due to GGG-motifs in both the allele-specific and the cross-allelic probe sequences whereas in the intra-allelic correlation pattern GGG-motifs occur solely in allele-specific probe sequences. The underlying asymmetry of probe/target interactions thus leads to asymmetrical correlations between the allele-specific and cross-allelic signals for both considered situations. In other words, the different correlations observed for intra- and inter-allelic crosstalks are consequences of the lack of complementarity of part of the interaction modes.

Basic crosstalk between the alleles

The incremental allele-specific probe signals are weighted superpositions of allele-specific and cross-allelic contributions with weighting factors given by the fraction of targets of the respective allele (→ Eq. 36). They can be decomposed into a sequence independent basic component and into a SNP specific component. The basic component is given by Eq. 36 and Eq. 38 for PMs or Eq. 39 for MMs with vanishing triple sensitivities, i.e. $Y(xby)\to 0$,

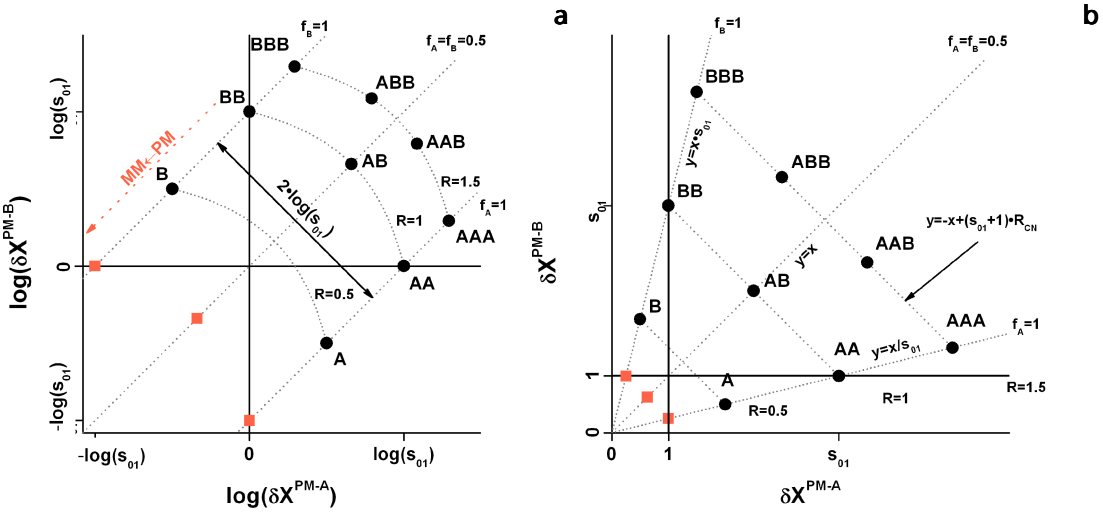


Figure 21: Basic crosstalk between the allele-specific PM probe signals. *Part a is logarithmic and part b in linear scale. Each genotype is defined by the parameter couple (R_{CN}, f_G) which unambiguously transforms into the relative intensity coordinates (δX^{P-A},δX^{P-B}) (→ Eq. 40). Special cases referring to mono- (haploid), bi- and triploid genotypes, (R_{CN}=0.5, $f_A \in \{1,0\}$), (1, {1,1/2,0}) and (1.5, {1,1/3,2/3,1}), respectively, are shown as solid dots (→ dotted lines, Eqs. 41 and 42). The width of the band included between the iso-f_G lines for f_A=1 and f_B=1 is ~2•log(s_{01}) where s_{01} denotes the mean intensity gain between probe/target duplexes without and with one mismatched pairing. Its value determines the available range of probe intensities for 0≤f_G≤1. The three red squares refer to the relative signals of the MM probes for R_{CN}=1. They are shifted by 1/s_{01} downwards along the iso-f_G lines (→ see also dotted arrow).*

$$\begin{aligned}\delta X_0^{PM-G} &= R_{CN}\cdot\left(f_G\cdot(s_{01}-1)+1\right)\\ \delta X_0^{MM-G} &= R_{CN}\cdot\left(f_G\cdot(s_{12}-1)+1\right)/s_{12}\\ \delta X_0^{MM-G} &= R_{CN} \quad \text{if offset } \delta=0\end{aligned}. \tag{40}$$

Each particular combination of the parameter values (R_{CN}, f_G) transforms into a couple of allele-specific incremental signals, (δX_0^{P-A},δX_0^{P-B}) which can be illustrated as a point in the "crosstalk" coordinate system shown in Figure 21.

The parametric Eq. 40 directly transforms into functions of the relative allele signal:

with f_G=const.,

$$\delta X^{P-B}\big|_{f_G=\text{const.}} = p\cdot\delta X^{P-A} \quad \text{with} \quad p=\frac{1+(1-f_A)\cdot(s-1)}{1+f_A\cdot(s-1)}=\begin{cases}s & \text{for } f_A=0\\ 1 & \text{for } f_A=f_B=0.5\\ 1/s & \text{for } f_A=1\end{cases}, \tag{41}$$

$$\text{and } s=\begin{cases}s_{01} & \text{for P=PM}\\ s_{12} & \text{for P=MM}\\ 1 & \text{for P=MM, offset } \delta=0\end{cases}$$

and with R_{CN}=const.,

$$\delta X^{PM-B}\big|_{R_{CN}=const.} = (1+s_{01})\cdot R_{CN} - \delta X^{PM-A}$$
$$\delta X^{MM-B}\big|_{R_{CN}=const.} = \frac{(1+s_{12})\cdot R_{CN}}{s_{12}} - \delta X^{MM-A} \cdot \quad (42)$$

Eqs. 41 and 42 define the framework of iso-f_G and iso-R_{CN} lines which are shown in Figure 21 in the case of PM probes. The iso-f_G lines run parallel to the diagonal of the coordinate axes in logarithmic scale (→ Figure 21: part a). The lines for f_A=1 and f_B=1 mark a band of width $2 \cdot \log(s_{01})$. It determines the potential range of genotypes ($0 \le f_G \le 1$). The parameter s_{01} consequently determines the basic crosstalk between the probe signals of both alleles (→ Eq. 38). The crosstalk decreases with increasing value of s_{01}, i.e. with increasing specificity of the allele-specific signals compared with the cross-allelic ones. In turn, it increases the bandwidth of the potential intensity range for different genotypes and thus also the discrimination power for intermediate f_G values. In linear scale the relevant range is given by a symmetric section about the diagonal (→ Figure 21: part b).

The basic crosstalk pattern of the PM probes transforms into that of the MM probes by rescaling the coordinate axes according to

$$\delta X^{MM-G} \rightarrow \delta X^{PM-G} \cdot \frac{f_G \cdot (s_{12}-1)+1}{s_{12} \cdot (f_G \cdot (s_{01}-1)+1)} \quad (43)$$

(→ Eq. 40). As I consider background corrected intensities the terms s_{01} and s_{12} are approximately equal, i.e. $s_{01} \approx s_{12}$. Thus, Eq. 43 simplifies to $\delta X^{MM-G} \rightarrow \delta X^{PM-G}/s_{12}$. In other words, the crosstalk between the MM probes is roughly shifted by $R_{CN} \rightarrow R_{CN}/s_{12}$ compared with that of the PM probes.

SNP-specific crosstalk between the alleles

The normalized specific probe signals can be obtained for each of the six considered SNPs assuming diploid (CN=2) homozygous ($f_A \in \{1,0\}$) or heterozygous (f_A=0.5) genotypes using the set of 256 triple sensitivity values (→ Figure 9) and

$$\delta X_S^{PM-G} = \delta X^{PM-G} - \delta X_0^{PM-G}$$
$$\delta X_S^{MM-G} = \delta X^{MM-G} - \delta X_0^{MM-G} \cdot \quad (44)$$

The resulting crosstalk patterns are shown in Figure 22 for each of the SNPs. These patterns characterize the correlation between the PM probe signals of both alleles caused by the change of the SNP base in the context of its nearest neighbors. The SNP assignments refer to the sense strand. The respective crosstalk of the antisense strand is simply given by substituting the SNP base by its WC complement.

The different shapes and sizes of the obtained data clouds show that the crosstalk between the alleles is considerably modulated by the particular sequence context of each SNP. This SNP-specific bias is given by the deviation of the respective SNP-specific signal from the expected basic crosstalk of the respective genotype (→ Figure 21). Especially GGG-motifs produce marked outliers in the crosstalk pattern (red circles). The scatter width of the data clouds is inversely related to the ability of the method to differentiate between the different genotypes. This resolution power clearly varies from SNP to SNP: For example, the data clouds of the three

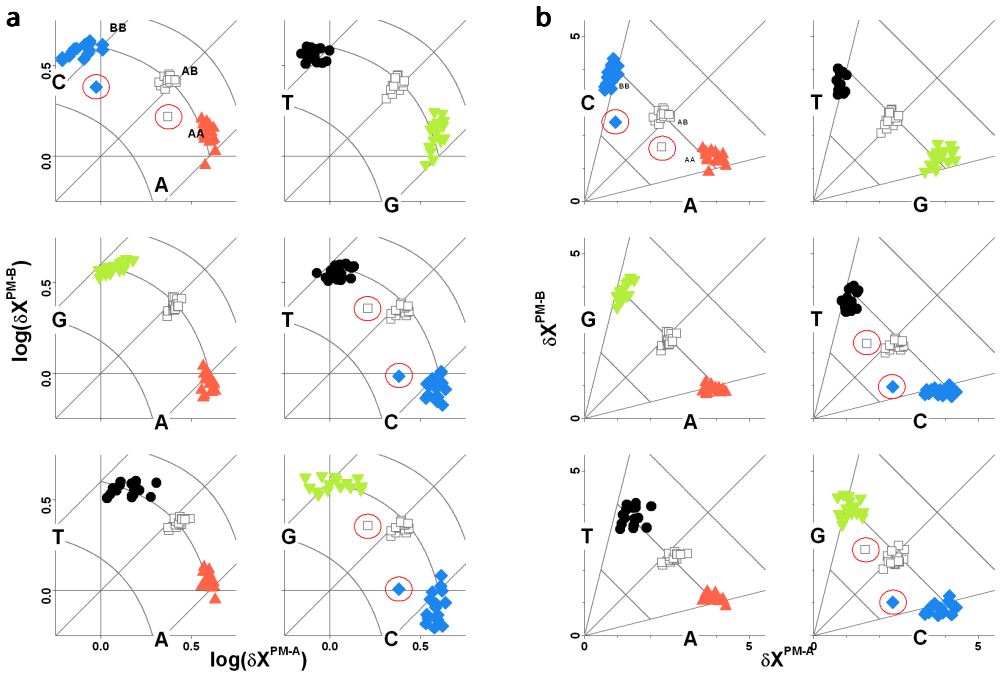


Figure 22: Specific crosstalk of the six SNPs considered on GeneChip arrays. *The normalized probe signals are calculated using Eq. 36 and the values of the respective triple sensitivities as superpositions of contributions due to the intra- and inter-allelic hybridization modes (→ Figure 20). The full symbols refer to the two homozygous genotypes and the open symbols to the heterozygous genotype. Signals associated with GGG-motifs are indicated by red circles. The network of dotted lines refers to the iso-f_A and iso-R_{CN} lines shown in Figure 21. Deviations between the SNP-specific signals and the basic crosstalk points given by the intersection of the respective iso-f_A and iso-R_{CN} lines cause biased genotyping estimates of f_G and R_{CN} (→ Figure 23 below).*

considered genotypes ($f_A \in \{1,0.5,0\}$) are more compact for the A/G-SNP than for the A/C one.

Figure 23 shows boxplots of the logarithmic signal difference of both alleles for all considered SNPs in the special cases of homozygous and heterozygous genotypes as a measure of the SNP-specific signal bias. It clearly correlates for $f_A=f_B=0.5$ and $f_A=f_B=1$. However, the signal ratios also show specific differences between the two selected allele fractions. For example, the median bias of the SNPs A/C and G/T change the sign between homozygous and heterozygous genotypes. This trend reflects the different contributions of the cross-allelic terms to the allele-specific signals for the inter-allelic crosstalk patterns (→ Eq. 36). Hence, the respective bias obviously depends on the particular SNP type and also on the allele fraction f_G.

SNP biased genotyping and copy number estimation

Essentially, each SNP is characterized by its individual crosstalk pattern shown in Figure 22, which, if left uncorrected, will produce biased genotyping and copy number estimates. Deviations between the iso-f_G lines are expected to cause predominantly biased relative allele abundances whereas deviations between the iso-R_{CN} curves are

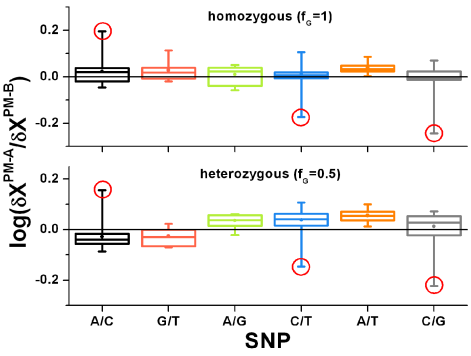


Figure 23: Boxplot of the SNP-specific signal bias. *The bias is calculated as the logged ratio of the SNP-specific signals of both alleles for homozygous (f_G=1) and heterozygous (f_G=0.5) genotypes. The red circles denote outliers which are associated with GGG-motifs. The horizontal lines indicate unbiased signals.*

expected to affect first of all the copy number estimates. The respective bias can be assessed by solving the two equations given by the Eqs. 41 and 42 for G∈{A,B} with respect to R_{CN} and f_G after replacing the SNP-dependent ('real') signals with the SNP-independent ('ideal') ones, i.e.

$$\begin{aligned} &CN^{PM}=\frac{2}{s_{01}+1}\cdot\left(\delta X^{PM-A}+\delta X^{PM-B}\right);\ CN^{MM}=\frac{2\cdot s_{12}}{s_{12}+1}\cdot\left(\delta X^{MM-A}+\delta X^{MM-B}\right) \\ &f_G^P=\frac{s}{s-1}\cdot\left(RAS^{P-G}-RAS^{P-G'}/s\right)\quad \text{with } RAS^{P-G}=\frac{\delta X^{P-G}}{\delta X^{P-A}+\delta X^{P-B}} \\ &\text{and } s=\begin{cases} s_{01} & \text{for P=PM} \\ s_{12} & \text{for P=MM} \\ 1 & \text{for P=MM, offet } \delta=0 \end{cases} \end{aligned} \quad . \qquad (45)$$

Accordingly, the copy number and the allele fraction are directly related to the sum of the allele specific signals and the relative allele signals (RAS), respectively. Note that the parameter s corrects the RAS values for the SNP crosstalk. In the limit of $s\to\infty$, i.e. for infinitely large specificity of the allele signals, one gets $f_G=RAS^{P-G}$, i.e. the allele-specific signal directly estimates the relative abundance of the respective allele.

The boxplots in Figure 24 (part a and b) show the median genotyping parameters f_A and CN of all 18 SNP-related triple values and their distribution in terms of their interquartile range. The systematic bias of the SNP-dependent values is given by their difference relative to the 'ideal' genotyping data of diploid homozygous and heterozygous genotypes, CN=2 and $f_A\in\{1,0.5,0\}$. For example, a thymine in the context of the A/T-SNP overestimates the copy number and underestimates the allele abundance whereas a thymine in the context of the G/T-SNP shows an opposite tendency.

The GGG-bias gives rise to outliers in the f_G estimates only (→ Figure 24: part a, red circles). This is due to the fact that the GGG-related signals deviate predominantly perpendicular to the iso-f_G lines of the basic signals (→ Figure 21) which, in turn, is a consequence of the exorbitantly weak sensitivity of the GGG-motifs in the allele-specific hybridization mode (→ Figure 20).

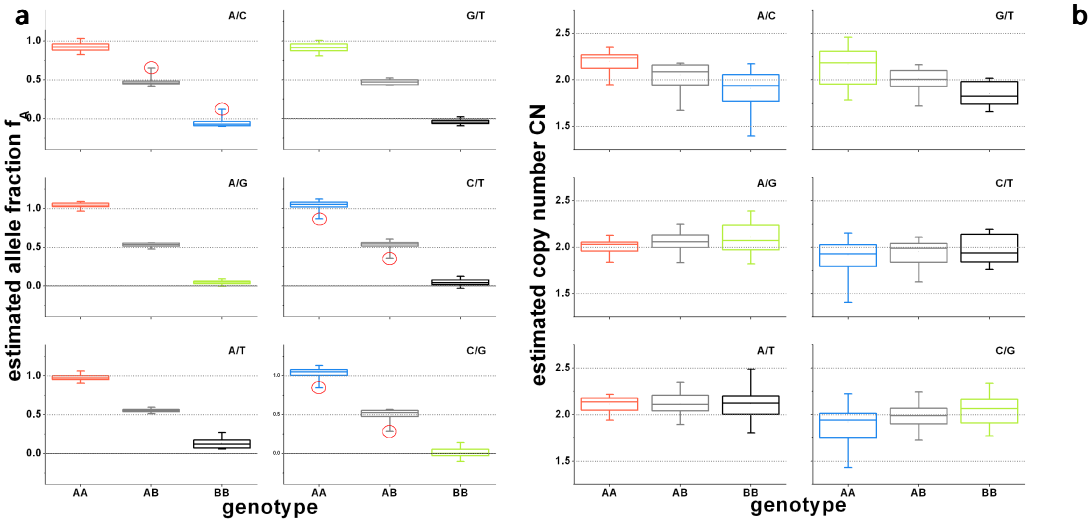


Figure 24: Boxplots of the estimated genotype parameters. *The parameters f_A (allele fraction, part a) and CN (copy number, part b) were estimated using Eq. 45 and the SNP-specific probe signals shown in Figure 22. Red circles indicate the outliers associated with GGG-motifs. The data refer to diploid homozygous and heterozygous genotypes, i.e. CN=2 and $f_A \in \{1,0.5,0\}$.*

To compare the bias between the different SNPs and its alleles on one hand, and between the f_G and CN estimates on the other hand I calculate the relative bias as the difference of the SNP-specific genotyping estimates and their respective ideal values (→ Figure 25). The resulting ranking of the SNP-dependent biases indicates that cytosine alleles are prone to overestimate f_G and to underestimate CN whereas guanine alleles show the opposite trend. Note also that the majority of the alleles that underestimate f_G overestimate CN and contrariwise. The G and T alleles of the G/T-SNP are associated with strong positive and negative biases of CN, respectively, as well as the A and C alleles of the A/C-SNP. A reversed behavior can be observed for the estimated allele fraction f_G. Small biases are noticed for the alleles G (C/G-SNP) and T (G/T-SNP) with respect to CN and for the alleles A (A/T-SNP) and G (C/G-SNP) with respect to f_G.

Combining PM and MM probe signals

Probe signals of PM and MM probes are usually combined to reduce cross allelic biases in genotyping algorithms [6,8,67] such as calculating the relative allele signals of the PM-MM difference, $RAS^G = \Delta^G / (\Delta^G + \Delta^{G'})$ with $\Delta^G = I^{PM-G} - 0.5 \cdot (I^{MM-G} + I^{MM-G'})$ [8]. The mismatch probes are intended to correct the PM probe intensities for the background intensity due to the optical background and nonspecific hybridization (→ Eq. 3) because these parasitic contributions are similar for PM and MM probes in a first order approximation (→ e.g. [67]). The middle base of a MM probe is given by the WC complement of the middle base of its respective PM probe. Only MM probes of complementary SNPs with offset δ=0 are an exception from the given rule. Both probe signals consequently produce the same correlation patterns as the intra-allelic correlations of the A/T- and C/G-SNPs shown in Figure 20 (part a). Taking the plot as "MM−G•G vs. PM−G•G" plot one has to consider the WC complement of the allele

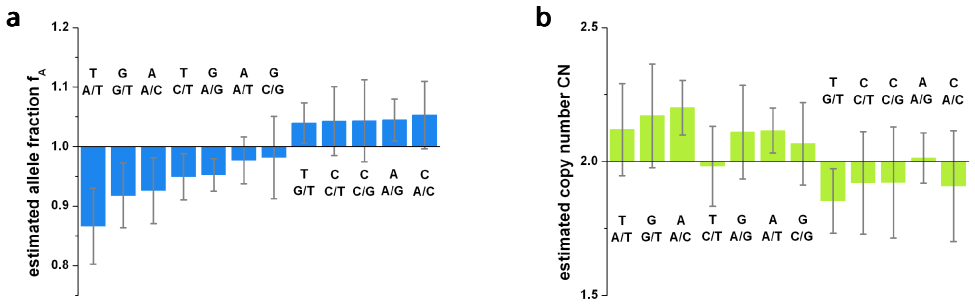


Figure 25: The mean bias for each allele. *It means the mean difference between SNP-specific estimates and the ideal values referring to diploid homozygous present genotypes (f_G=1, CN=2). The error bars indicate the standard deviation of the 16 triple values per allele. The f_G values in part a are ranked in increasing order. In part b the same ranking is used to indicate the negative correlation between f_G and CN. The allele and the respective SNP are given within the figure.*

letters as the middle bases of the PM probe, e.g. considering allele A gives "MM−A•A vs. PM−T•A". It is obvious to see that the combination of PM and MM in the common way introduces additional variability into the combined signal.

The sequence motif of the middle triple gives rise to PM/MM variability whereas the sequence motif of the SNP triple induces the P−G/P−G' variability (P∈{PM,MM}). Both motifs are virtually independent of each other for probes with $|\delta|>2$ (without overlap of the triples) and partly for probes with $|\delta|=1$ (overlap of the triples in one position)(→ Figure 7, Figure 4). To a good approximation one can therefore assume that the bias due to the middle triple independently combines with that of the SNP triple and, this way, causes the marked inflation of the total scatter width of combined signals of the PM and MM probes.

In consequence, the possible improvement of combinations of PM and MM signals due to background corrections is counterbalanced by the inflation of the bias due to sequence effects. Note also that the MM probe design in terms of self-complementary mismatches (Aa-group) appears insufficient as the sensitivities of the Aa-group indicate a relatively large variability compared with the alternative mismatch groups Ag and Ac (→ Figures 9 and 20: part a). PM-only approaches of genotyping and copy number estimation algorithms which ignore MM probe signals are therefore developed and recommended (→ e.g. [2,6,13]).

Correcting probe intensities for sequence effects

The SNP-specific sequence bias transforms into systematic errors of the genotyping characteristics derived from the signals of single probes. Note that the sequence-context of a partial SNP and consequently also the respective bias is essentially very similar for all probes of a selected probe set addressing the same SNP. As a consequence, the averaging of the probe signals into set-related allele values only weakly reduces the systematic signal error after the summarization step. SNP arrays differ in this aspect from expression arrays where the sequences of the set of probes interrogating the expression of the a gene or exon can be usually chosen independently.

A central task of the preprocessing of SNP probe signals is consequently their correction for sequence effects and in particular for SNP-specific biases. The detailed presentation and verification of an appropriate algorithm is beyond the scope of this work. The results of this systematic study however enable to identify relevant sequence motifs which significantly modulate the probe intensities. The intensity contributions of such motifs constitute the building blocks of an appropriate intensity model. In particular the results suggest the following rules for sequence correction of SNP probe intensities:

(i) Sequence effects due to WC pairings between probe and target are well approximated using nearest neighbor (NN) terms in analogy with accepted NN-free energy models for oligonucleotide duplexing in solution [46].

(ii) The anisotropy of probe/target interactions due to the fixation of the probes at the chip surface and end-opening (zippering effects) [26,62] requires the consideration of the positional dependance of the interactions in a motif-specific fashion, i.e. separately for each NN-combination of nucleotide letters. The assumption of a generic shape function which applies to all motifs seems suboptimal [22,11].

(iii) The modulation of probe intensities by mismatched pairings can be considered using triple motifs which consist of the central mismatch and the two adjacent WC pairings.

(iv) Nominal base pairings according to (i) and (iii) can be deduced from the hybridization mode of the respective probes which, in turn, provides selection criteria of the probes for parameter estimation. The mean intensity penalty owing to one and two mismatches can be estimated from the respective group of probes.

(v) Runs of triple guanines (GGG) represent a special motif which markedly modulates the intensities of the respective probes. The underlying effect does not originate from probe/target (pairwise) interactions but obviously results from the formation of collective complexes presumably of four neighboring probes. Therefore it affects essentially all probes with triple G-motifs independently of the hybridization mode.

(vi) Also tandem mismatches represent a special motif of MM probes with a modified intensity penalty compared with other MM probes possessing two mismatches with at least one WC pairing in between. This sequence effect can be taken into account in a first order approximation by decomposing the quadruple formed by the tandem mismatch and the two adjacent WC pairings into two NN terms referring to a WC and a mismatch pairing, respectively, or more roughly, by explicitly considering the two adjacent WC pairings.

(vii) The shift of mismatch motifs by a few sequence positions about the middle base of the probe and the effect of flanking mismatches adjacent to triples with a central mismatch can be neglected to a good approximation.

(viii) Background intensity contributions, i.e. optical background and “chemical” background due to nonspecific hybridization, should be considered especially for probes forming at least two mismatch pairings.

Established preprocessing algorithms for GeneChip SNP arrays explicitly consider the mean intensity penalty per mismatch [68,69] or, in addition, the single-base related positional effect [6]. The authors of the latter work conclude from their results that, after correction, "…the sequence effect is reduced but can be further improved". The presented results of this thesis clearly show that effects which are not taken into account in this model, namely the particular mismatch and its sequence context, the contribution of nearest neighbor stacking interactions and of triple-G runs, considerably modulate the probe intensities. It can be expected that their explicit consideration will further improve genotyping based on SNP microarrays.

My analysis has focused on sequence effects. Note for sake of completeness that an elaborated correction algorithm should also consider additional sources of intensity variation not taken into account here, such as the fragment length and the GC-content of the targets [6,51] and nonlinear effects due to saturation of the probes at large transcript concentrations [35,36,70], nonspecific hybridization [71] and/or bulk depletion of the targets [72,73].

4.3 Summary and conclusions

Mismatch pairings and runs of poly G-motifs in the probe sequence formed in cross-allelic probe target duplexes are the main sources of signal variability on SNP arrays giving rise to the loss of accuracy in genotyping estimates with consequences for downstream analyses. The sequence dependence of DNA/DNA-interactions must be considered in appropriate calibration methods of the probe intensities to obtain accurate genotyping and copy number estimates. The poly-G effect seems to be related to the crowded arrangement of probes on high density oligonucleotide arrays which facilitates complex formation of neighboring probes and this way reduces the amount of free probes available for target binding. The probe/target interactions on the chip can be decomposed into nearest neighbor contributions which well correlate with the respective free energy terms describing DNA/DNA-interactions in solution where the effect of mismatches is about twice as large as that of canonical pairings. Triple averaging represents a model-free approach to estimate the mean intensity contributions of different sequence motifs which can be applied in improved calibration algorithms to correct signal values for sequence effects.

Bibliography

1. Collins FS, Brooks LD, Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation**. *Genome Res* 1998, **8(12)**:1229-1231.
2. Rabbee N, Speed TP: **A genotype calling algorithm for affymetrix SNP arrays**. *Bioinformatics* 2006, **22(1)**:7-12.
3. Sham P, Bader JS, Craig I, O'Donovan M, Owen M: **DNA Pooling: a tool for large-scale association studies**. *Nat Rev Genet* 2002, **3(11)**:862-871.
4. Affymetrix: **BRLMM: an Improved Genotype Calling Method for the GeneChip® Human Mapping 500K Array Set**. *Affymetrix White Paper* 2006, www.affymetrix.com.
5. Affymetrix: **BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array**. *Affymetrix White Paper* 2007, www.affymetrix.com.
6. Carvalho B, Bengtsson H, Speed TP, Irizarry RA: **Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data**. *Biostat* 2007, **8(2)**:485-499.
7. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen M-m, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S: **Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays**. *Bioinformatics* 2005, **21(9)**:1958-1963.
8. Liu W-m, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder TB, Webster TA, Dong S, Liu G, Jones KW, Kennedy GC, Kulp D: **Algorithms for large-scale genotyping microarrays**. *Bioinformatics* 2003, **19(18)**:2397-2403.
9. Korn J, et al.: **Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs**. *Nature Genetics* 2008, **40**:1253-1260.
10. Lamy P, Andersen CL, Wikman FP, Wiuf C: **Genotyping and annotation of Affymetrix SNP arrays**. *Nucl Acids Res* 2006, **34(14)**:e100.
11. Shen F, Huang J, Fitch K, Truong V, Kirby A, Chen W, Zhang J, Liu G, McCarroll S, Jones K, Shapero M: **Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes**. *BMC Genetics* 2008, **9(1)**27.
12. Nicolae DL, Wu X, Miyake K, Cox NJ: **GEL: a novel genotype calling algorithm using empirical likelihood**. *Bioinformatics* 2006, **22(16)**:1942-1947.
13. Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA: **SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays**. *Bioinformatics* 2007, **23(1)**:57-63.
14. Richard J.L. Anney EK, Colm T. O'Dushlaine, Jessica Lasky-Su, Barbara Franke, Derek W. Morris, Benjamin M. Neale, Philip Asherson, Stephen V. Faraone, Michael Gill: **Non-random error in genotype calling procedures: Implications for family-based and case-control genome-wide association studies**. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2008,

147B(8):1379-1386.
15. Hacia JG: **Resequencing and mutational analysis using oligonucleotide microarrays**. *Genetics* 1999, **21**:42 - 47.
16. Bruun GM, Wernersson R, Juncker AS, Willenbrock H, Nielsen HB: **Improving comparability between microarray probe signals by thermodynamic intensity correction**. *Nucl Acids Res* 2007, **35(7)**.
17. Seringhaus M, Rozowsky J, Royce T, Nagalakshmi U, Jee J, Snyder M, Gerstein M: **Mismatch oligonucleotides in human and yeast: guidelines for probe design on tiling microarrays**. *BMC Genomics* 2008, **9**.
18. Binder H, Krohn K, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures**. *Algorithms for Molecular Biology* 2008, **3:11**.
19. Binder H, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: Theory and algorithm**. *Algorithms for Molecular Biology* 2008, **3:12**.
20. Burden CJ: **Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed**. *Physical Biology* 2008, **5(1)**016004.
21. Ferrantini A, Allemeersch J, Van Hummelen P, Carlon E: **Thermodynamic scaling behavior in genechips**. 2009, **10**.
22. Zhang L, Wu C, Carta R, Zhao H: **Free energy of DNA duplex formation on short oligonucleotide microarrays**. *Nucl Acids Res* 2006, gkl1064.
23. Binder H, Bruecker J, Burden CJ: **Non-specific hybridization scaling of microarray expression estimates - a physico-chemical approach for chip-to-chip normalization**. *J Phys Chem B* 2009, **113(9)**:2874-95.
24. Binder H, Preibisch S: **GeneChip microarrays - signal intensities, RNA concentrations and probe sequences**. *J Phys Cond Mat* 2006, **18**:537-566.
25. Binder H: **Probing gene expression - sequence specific hybridization on microarrays**. In: *Bioinformatics of Gene Regulation II*. Edited by Kolchanov N, Hofestaedt R: Springer Sciences and Business Media; 2006: 451-466.
26. Binder H: **Thermodynamics of competitive surface adsorption on DNA microarrays**. *Journal of Physics-Condensed Matter* 2006, **18(18)**:S491-S523.
27. Halperin A, Buhot A, Zhulina EB: **On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions**. *J Phys Cond Mat* 2006, **18**:463-490..
28. Naiser T, Ehler O, Kayser J, Mai T, Michel W, Ott A: **Impact of point-mutations on the hybridization affinity of surface-bound DNA/DNA and RNA/DNA oligonucleotide-duplexes: Comparison of single base mismatches and base bulges**. *BMC Biotechnology* 2008, **8(1)**:48.
29. 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.
30. Upton G, Langdon W, Harrison A: **G-spots cause incorrect expression measurement in Affymetrix microarrays**. *BMC Genomics* 2008, **9(1)**:613.
31. Binder H, Preibisch S, Kirsten T: **Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays**. *Langmuir* 2005, **21**:9287-9302.
32. Naiser T, Kayser J, Mai T, Michel W, Ott A: **Position dependent mismatch discrimination on DNA microarrays - experiments and model**. *BMC Bioinformatics* 2008, **9**:509.
33. Binder H, Kirsten T, Hofacker I, Stadler P, Loeffler M: **Interactions in**

oligonucleotide duplexes upon hybridisation of microarrays. *J Phys Chem B* 2004, **108(46)**:18015-18025.
34. Binder H, Preibisch S: **Specific and non-specific hybridization of oligonucleotide probes on microarrays**. *Biophys J* 2005, **89**:337-352.
35. Binder H, Kirsten T, Loeffler M, Stadler P: **The sensitivity of microarray oligonucleotide probes - variability and the effect of base composition**. *J Phys Chem B* 2004, **108(46)**:18003-18014.
36. Held GA, Grinstein G, Tu Y: **Modeling of DNA microarray data by using physical properties of hybridization**. *Proc Natl Acad Sci USA* 2003, **100(13)**7575-7580.
37. Marcelino LA, Backman V, Donaldson A, Steadman C, Thompson JR, Preheim SP, Lien C, Lim E, Veneziano D, Polz MF: **Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data**. *Proc Natl Acad Sci USA* 2006, **103(37)**:13629-13634.
38. Wick L, Rouillard J, Whittam T, Gulari E, Tiedje J, Hashsham S: **On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes**. *Nucl Acids Res* 2006, **34(3)**:e26.
39. Pozhitkov A, Noble P, Domazet-Loso T, Nolte A, Sonnenberg R, Staehler P, Beier M, Tautz D: **Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted**. *Nucl Acids Res* 2006, **34(9)**:e66.
40. Lee I, Dombkowski AA, Athey BD: **Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray**. *Nucl Acids Res* 2004, **32**:681-690.
41. Heim T, Wolterink JK, Carlon E, Barkema GT: **Effective affinities in microarray data**. *J Phys Cond Mat* 2006, **18**:S525-S536.
42. Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M, Sasaki M: **Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes**. *Biochem* 1995, **34(35)**:11211-11216.
43. Ikuta S., Takagi K., Bruce Wallace R., K. I: **Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs**. *Nucl Acids Res* 1987, **15(2)**:797-811.
44. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J: **Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A•A, C•C, G•G, and T•T Mismatches**. *Biochem* 1999, **38**:3468-3477.
45. The thermodynamics of DNA structural motifs: **Annual Review of Biophysics and Biomolecular Structure**. *Annual Review of Biophysics and Biomolecular Structure* 2004, **33**:415-440.
46. SantaLucia J: **A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbour thermodynamics**. *Proc Natl Acad Sci USA* 1998, **95**:1460-1505.
47. Affymetrix: **GeneChip Arrays Provide Optimal Sensitivity and Specificity for Microarray Expression Analysis**. *User Guide* 2001, .
48. Fish D, Horne M, Brewood G, Goodarzi J, Alemayehu S, Bhandiwad A, Searles R, Benight A: **DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison**. *Nucl Acids Res* 2007, **35(21)**:7197-7208.
49. Press WH, Flannery BP, Teukolsky SA, Vetterling WT: **Numerical Recipes**. New York: Cambrigde University Press; 1989
50. Matveeva OV, Shabalina SA, Nemtsov VA, Tsodikov AD, Gesteland RF, Atkins JF: **Thermodynamic calculations and statistical correlations for oligo-probes**

design. *Nucl Acids Res* 2003, **31**:4211-4217.
51. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S: **A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays**. *Cancer Res* 2005, **65(14)**:6071-6079.
52. Arnold FH, Wolk S, Cruz P, Tinoco I: **Structure, dynamics, and thermodynamics of mismatched DNA oligonucleotide duplexes d(CCCAGGG)2 and d(CCCTGGG)2**. *Biochemistry* 1987, **26(13)**:4068–4075.
53. Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians F, Shen M, Lu G, Fang J, Liu W, Ryder T, Kaplan P, Kulp D, Webster T: **Probe selection for high-density oligonucleotide arrays**. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 2003, **100(20)**:11237-11242.
54. Sharp AJ, Itsara A, Cheng Z, Alkan C, Schwartz S, Eichler EE: **Optimal design of oligonucleotide microarrays for measurement of DNA copy-number**. *Hum Mol Genet* 2007, **16(22)**:2770-2779.
55. Fasold M, Preibisch S, Binder H: **The GGG-bias of GeneChip expression data - analysis and correction**. *submitted* 2009, .
56. Rachwal PA, Brown T, Fox KR: **Effect of G-Tract Length on the Topology and Stability of Intramolecular DNA Quadruplexes**. *Biochem* 2007, **46(11)**:3036-3044.
57. Sühnel J: **Beyond nucleic acid base pairs: From triads to heptads**. *Biopolymers* 2002, **61(1)**:32-51.
58. Langdon WB, Upton GJG, Harrison AP: **Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips**. *Briefings in Bioinformatics* 2009, **10(3)**:259-277.
59. McGall GH, Barone AD, Diggelman M, Fodor SPA, Gentalen E, et al.: **The Efficiency of Light-Directed Synthesis of DNA Arrays on Glass Substrates**. *Journal of the American Chemical Society* 1997, **119**:5081-5090.
60. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, et al.: **Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists**. *Proc Natl Acad Sci USA* 1996, **93(24)**:13555–13560.
61. Pirrung MC, Fallon L: **Proofing of photolithographic DNA synthesis with 3',5'-dimethoxybenzoinyloxycarbonyl-protected deoxynucleoside phosphoramidites**. *Journal of Organic Chemistry* 1998, **63**:241-246.
62. Naiser T, Kayser J, Mai T, Michel W, Ott A: **Stability of a Surface-Bound Oligonucleotide Duplex Inferred from Molecular Dynamics: A Study of Single Nucleotide Defects Using DNA Microarrays**. *Physical Review Letters* 2009, **102**:218301–218304.
63. Naef F, Magnasco M: **Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays**. *Phys Rev E* 2003, **68**:011906.
64. Deutsch JM, Liang S, Narayan O: **Modeling of microarray data with zippering**. *arXiv:q-bioBM/0406039 v1* 2004, .
65. Zhang L, Miles M, Aldape K: **A model of molecular interactions on short oligonucleotide microarrays**. *Nat Biotechnol* 2003, **21(7)**:818 - 821.
66. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JMM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA: **Population structure, differential bias and genomic control in a large-scale, case-control association study**. *Nature Genetics* 2005, **37(11)**:1243-1246.
67. Xiao Y, Segal MR, Yang YH, Yeh R-F: **A multi-array multi-SNP genotyping**

algorithm for Affymetrix SNP microarrays. *Bioinformatics* 2007, **23(12)**:1459-1467.

68. LaFramboise T, Harrington D, Weir BA: **PLASQ: A Generalized Linear Model-Based Procedure to Determine Allelic Dosage in Cancer Cells from SNP Array Data**. *Harvard University Biostatistics Working Paper Series* 2006, **44**:1-28.
69. LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, Harrington D, Sellers WR, Meyerson M: **Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis**. *PLoS Computational Biology* 2005, **1**:e65.
70. Burden CJ, Pittelkow YE, Wilson SR: **Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays**. *Statistical Applications in Genetics and Molecular Biology* 2004, **3**:35.
71. Binder H, Bruecker J, Burden CJ: **Non-specific hybridization scaling of microarray expression estimates - a physico-chemical approach for chip-to-chip normalization**. *Journal of Physical Chemistry B* 2009, **113**:2874–2895.
72. Burden C, Binder H: **Physico-chemical modelling of target depletion during hybridisation on oligonulceotide microarrays**. *Phys. Biol.* 2010, **7(1)**.
73. Suzuki S, Ono N, Furusawa C, Kashiwagi A, Yomo T: **Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays** . *BMC Genomics* 2007, **8**:373.

List of Figures

List of Tables

Appendix A
Precaution with Affymetrix data

Affymetrix offers on their website several support materials[1] for the Human Mapping 100K Set such as library files (CDF files), sample data (CEL files), alignment, annotation and sequence files. Some materials are required by the Genotyping Console™ (GTC) software provided by Affymetrix as handy tool for the analysis of their microarrays. The additional materials can also be used for raw data analyses as made in this thesis. However, one pitfalls has to be taken into account.

Errors in CDF files

The data of the CDF files have to be taken with a pinch of salt if analyses or algorithms depend on the information of the strand direction such as getting the probe base at the SNP or middle position. As described in the thesis the probe intensities are sequence specific and any appropriate calibration method would be impaired by erroneous sequence information.

Table A.1: Data extracted by AffxFusion SDK from CDF file of Mapping 50K Xba 240 array.

SNP_ID	PM_X	PM_Y	PM_middle_base	Direction	Offset	Allele
SNP_A-1641765	737	1419	G	1	0	G

Table A.1 shows the data of a probe of an A/G-SNP named SNP_A-1641765 extracted by the AffxFusion SDK from the CDF file of the Mapping 50K Xba 240 array where PM_X and PM_Y designate the physical position on the chip and direction '1' means that the probe interrogates the sense strand. The offset value δ=0 and the SNP's allele is 'G'. As the probe interrogates the allele G of the SNP with δ=0 the base at the middle position of the PM probe should be the WC complement of the allele namely a C and not a G.

The data in the sequence file is shown in Table A.2. Here the same probe interrogates the antisense (=reverse; 'r') strand of the target. Thus the given middle base G can be taken as correct.

Table A.2: Data extracted from sequence file of Mapping 50K Xba 240 array.

SNP_ID	PM_X	PM_Y	Offset	Sequence	Direction	PM/MM	Allele
SNP_A-1641765	737	1419	0	TTTAACCCTCCA G TAGAACAAAGAG	1	PM	G

The question is: Is the data in both the CDF file and the sequence file incorrect or just in the CDF file. To answer this question I searched the annotation file for the

[1] Materials discussed here are as at April 18, 2010.

SNP to get the NCBI dbSNP reference ID which is rs10490928 and the flanking sequence of the SNP with 25 bases on each side of the SNP position which is

agttcagaagagatttaaccctcca[A/G]tagaacaaagagaagagacttgctg.

The Genome Version (NCBI Build 36.1, March 2006) and the dbSNP Version (NCBI dbSNP Build 126, May 2006) can also be found. Searching on NCBI for the reference ID gives the sequence from Perlgen used by Affymetrix (ss23580976 in section "Submitter records for this RefSNP Cluster") which is

cactaagttcagaagagatttaaccctcca[A/G]tagaacaaagagaagagacttgctggccag.

Comparing the three sequences, one can see that the probe sequence from the sequence file equals the others so that the CDF file information must be wrong.

TTTAACCCTCCA G TAGAACAAAGAG
agttcagaagagatttaaccctcca[A/G]tagaacaaagagaagagacttgctg
cactaagttcagaagagatttaaccctcca[A/G]tagaacaaagagaagagacttgctggccag

From all 1.179.200 probe pairs on the Mapping 50K Xba 240 array 544.000 probe pairs (~ 46%), i.e. 27.200 SNPs, are affected by this problem. The entries of the target strandedness of the corresponding probe pairs are switched for these SNPs. Also the Expos entries in the CDF file that could be used to calculate the offset positions are incorrect for these SNPs. The same problem occurs on the Mapping 50K Hind 240 array. Here are 571.400 out of 1.144.880 probe pairs (~ 50%) affected, i.e., 28.570 SNPs.

The erroneous CDF files does not affect the analysis made by the Affymetrix GTC software as the strandedness is ignored by the algorithms for the Human Mapping 100K sets and later chip generations. Only the MPAM algorithm used for the Human Mapping 10K arrays could be impaired if the library file of that generation is also faulty.

Raw data analyses that make use of the target strand direction contained in the CDF files could be prone to erroneous results. It is more safe to use the information of the sequence file.

Appendix B
Classification of triples

The intensities of the 64 triples of the Ab-groups found in 3.1 (Classification of triples) are given by the following equations:

At_0: $\log I_{At_0}(XBY) = \log I^{PM-G\cdot G}_{(Ab=At)}(XBY)$

At_1: $\log I_{At_1}(XBY) = \frac{1}{2}\cdot\left(\log I^{MM-G\cdot G}_{(Ab=At,\,\delta\neq(-1,0,1),\,\pi=SNP)}(XBY) + \log I^{PM-G'\cdot G}_{(Ab=At,\,\delta\neq(-1,0,1),\,\pi=mb)}(XBY)\right)$

Aa_1: $\log I_{Aa_1}(XBY) = \frac{1}{2}\cdot\left(\log I^{MM-G\cdot G}_{(Ab=Aa)}(XBY) + \log I^{PM-G'\cdot G}_{(Ab=Aa)}(XBY)\right)$

Ag_1: $\log I_{Ag_1}(XBY) = \frac{1}{2}\cdot\left(\log I^{PM-G'\cdot G}_{(Ab=Ag)}(XBY) + \log I^{MM-G'\cdot G}_{(Ab=Ag,\,\delta=0)}(XBY)\right)$

Ac_1: $\log I_{Ac_1}(XBY) = \frac{1}{3}\cdot\left(\log I^{MM-G\cdot G}_{(Ab=Ac)}(XBY) + \log I^{PM-G'\cdot G}_{(Ab=Ac)}(XBY) + \log I^{MM-G'\cdot G}_{(Ab=Ac,\,\delta=0)}(XBY)\right)$

Aa_2: $\log I_{Aa_2}(XBY) = \log I^{MM-G'\cdot G}_{(Ab=Aa,\,\delta\neq(-1,0,1))}(XBY)$

Ag_2: $\log I_{Ag_2}(XBY) = \log I^{MM-G'\cdot G}_{(Ab=Ag,\,\delta\neq(-1,0,1))}(XBY)$

Ac_2: $\log I_{Ac_2}(XBY) = \log I^{MM-G'\cdot G}_{(Ab=Ac,\,\delta\neq(-1,0,1))}(XBY)$

Appendix C
Background correction

The background contribution I^{BG} can be obtained solving Eq. 19.

$$I(\#mm) \approx \left(\frac{I^{sat} \cdot c \cdot K_{duplex}(\#mm)}{1 + c \cdot K_{duplex}(\#mm)} + I^{BG} \right) \qquad \left| x = c \cdot K_{duplex}(\#mm) \right.$$

$$\left| I^{sat} = I^{max} - I^{BG} \right.$$

$$\approx \left(\frac{\left(I^{max} - I^{BG}\right) \cdot x}{1 + x} + I^{BG} \right)$$

$$\rightarrow x \approx \frac{I(\#mm) - I^{BG}}{I^{max} - I(\#mm)}$$

$$I(\#mm+1) \approx \left(\frac{\left(I^{max} - I^{BG}\right) \cdot x/s}{1 + x/s} + I^{BG} \right) \qquad \left| s = \frac{I(\#mm)}{I(\#mm+1)} \right.$$

$$\approx \frac{I^{max} \cdot I^{BG} - I^{max} \cdot I^{BG} \cdot s - I^{max} \cdot I(\#mm) + I^{BG} \cdot s \cdot I(\#mm)}{I^{BG} - I^{max} \cdot s + I(\#mm) \cdot s - I(\#mm)}$$

$$\rightarrow I^{BG} \approx \frac{I^{max} \cdot \left(-I(\#mm) + s \cdot I(\#mm+1)\right) + \left(1 - s\right) \cdot I(\#mm) \cdot I(\#mm+1)}{I^{max} \cdot \left(-1 + s\right) - s \cdot I(\#mm) + I(\#mm+1)}$$

Appendix D
Publication

A part of the present thesis has been published in:

Binder H, Fasold M, Glomb T: **Mismatch and G-Stack Modulated Probe Signals on SNP Microarrays**. *PLoS ONE* 2009, 4(11):e7862. doi:10.1371/journal.pone.0007862

and is attached below along with the supplemental material.

OPEN ACCESS Freely available online

PLoS one

Mismatch and G-Stack Modulated Probe Signals on SNP Microarrays

Hans Binder*, Mario Fasold, Torsten Glomb

Interdisciplinary Centre for Bioinformatics, Universität Leipzig, Leipzig, Germany

Abstract

Background: Single nucleotide polymorphism (SNP) arrays are important tools widely used for genotyping and copy number estimation. This technology utilizes the specific affinity of fragmented DNA for binding to surface-attached oligonucleotide DNA probes. We analyze the variability of the probe signals of Affymetrix GeneChip SNP arrays as a function of the probe sequence to identify relevant sequence motifs which potentially cause systematic biases of genotyping and copy number estimates.

Methodology/Principal Findings: The probe design of GeneChip SNP arrays enables us to disentangle different sources of intensity modulations such as the number of mismatches per duplex, matched and mismatched base pairings including nearest and next-nearest neighbors and their position along the probe sequence. The effect of probe sequence was estimated in terms of triple-motifs with central matches and mismatches which include all 256 combinations of possible base pairings. The probe/target interactions on the chip can be decomposed into nearest neighbor contributions which correlate well with free energy terms of DNA/DNA-interactions in solution. The effect of mismatches is about twice as large as that of canonical pairings. Runs of guanines (G) and the particular type of mismatched pairings formed in cross-allelic probe/target duplexes constitute sources of systematic biases of the probe signals with consequences for genotyping and copy number estimates. The poly-G effect seems to be related to the crowded arrangement of probes which facilitates complex formation of neighboring probes with at minimum three adjacent G's in their sequence.

Conclusions: The applied method of "triple-averaging" represents a model-free approach to estimate the mean intensity contributions of different sequence motifs which can be applied in calibration algorithms to correct signal values for sequence effects. Rules for appropriate sequence corrections are suggested.

Citation: Binder H, Fasold M, Glomb T (2009) Mismatch and G-Stack Modulated Probe Signals on SNP Microarrays. PLoS ONE 4(11): e7862. doi:10.1371/journal.pone.0007862

Editor: Cameron Neylon, University of Southampton, United Kingdom

Received August 21, 2009; **Accepted** October 19, 2009; **Published** November 17, 2009

Funding: This work has been funded in whole or in part with funds from the State of Saxony. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: binder@izbi.uni-leipzig.de

Introduction

Genomic alterations are believed to be the major underlying cause of common diseases such as cancer [1]. These alterations include various types of mutations, translocations, and copy number variations. Single nucleotide polymorphisms (SNPs) are the most abundant type of polymorphism in the human genome. With the parallel developments of dense SNP marker maps and technologies for high-throughput SNP genotyping, SNPs have become the polymorphic genetic markers of choice for genetic association studies which aim at discovering the genetic background of different phenotypes. Microarray platforms are capable of parallel genotyping of hundreds of thousands of SNPs in one measurement. To date this high throughput technology is therefore routinely performed to get comprehensive genome wide information about the genetic variability of individuals in genome wide association studies.

The microarray technology utilizes the specific affinity of fragmented DNA to form duplexes with surface-attached oligonucleotide probes of complementary sequence and subsequent optical detection of bound fragments using fluorescent markers. The measured raw probe intensities are subject to large variability, and depend not only on the abundance of allelic target sequences, but also on other factors such as the sequence dependent probe binding affinity. The successful correction of raw probe signals for such parasitic effects is essential to obtain exact genotyping estimates. It requires identification and understanding of the main sources of signal variation on the arrays.

The main purpose of this paper is to analyze the variability of probe signals of Affymetrix GeneChip SNP arrays as a function of the probe sequence and to identify relevant sequence motifs which significantly modulate the probe signals. Such sequence motifs constitute potential building blocks for improved calibration methods which aim at correcting probe signals for sequence effects.

The discovery of characteristic sequence motifs using SNP arrays is also important in a more general context: DNA/DNA duplex formation is the basic molecular mechanism of functioning not only of SNP arrays but also of other array types such as re-sequencing [2] and different expression arrays (gene- or exon-related and whole genome tiling arrays) of newer generations. It has been demonstrated that thermodynamic models of hybridiza-

tion taking into account such sequence-dependent effects are capable to significantly reduce signal fluctuation between probes interrogating the same target [3–6]. Knowledge of the underlying physical process is however still lacking in many details despite the recent progress in this field (see, for example, [7–12]). Particularly, surface hybridization is different from oligonucleotide duplexing in solution (see e.g. [13–15]). Systematic studies on oligonucleotide interactions on microrrays are therefore required to tackle selected problems such as signal anomalies of poly-guanine runs [16,17], the specific effect of mismatched base pairings [4,15,18] and/or the positional dependence of interaction strengths [9,19].

The presented analysis takes special advantage of the probe design used on GeneChip SNP arrays. Particularly, this technology uses 25meric oligonucleotide probes corresponding to a perfect match for each of the two allele sequences. In addition, a mismatch probe is synthesized for each allele to detect non-specific binding. Combination of this information with the target composition of fractionated genomic DNA used for hybrization on the arrays enables us to deduce the base pairings in the probe/target complexes producing a particular probe intensity. Making use of the hundreds of thousands signal values per SNP array allows us to extract specific intensity contributions of selected short sequence motifs of two-to-four adjacent nucleotides via appropriate averaging. The obtained motif-specific intensity contributions characterize the stability of the involved base pairings which include all relevant combinations of canonical Watson-Crick and mismatched pairings. Finally, the systematic analysis of different sequence motifs such as triples of adjacent bases allows us to identify those which account for significant signal variations.

We previously performed an analogous chip study using intensity data of expression arrays to characterize base pair interactions in DNA/RNA hybrid duplexes [20] which in final consequence enabled us to develop an improved algorithm for signal calibration and quality control [5,6]. Note that, compared with expression arrays, SNP arrays are even better suited to study base pair interactions because probe/target-duplexes are typically less contaminated with non-specific target fragments of unknown sequence and because genomic copy numbers are less variable than mRNA-transcript concentrations.

The paper is laid out as follows: Section 2 sets out the method and, particularly, explains the classification criteria used to assign the probe intensities to different interaction modes. In Section 3, we analyze different factors which affect the probe intensities such as the number of mismatches, the optical and non-specific background, signal contributions due to different sequence motifs such as different base triples, single and tandem mismatches and their positional dependence along the sequence. In addition we assess symmetry relations of the motifs, their decomposition into nearest neighbor terms and compare the results with thermodynamic nearest neighbor parameters characterizing DNA/DNA interactions in solution. In Section 4 we discuss the stability of different mismatches and discover the possible origin of the "poly-G" effect. Finally, we suggest rules for selecting appropriate sequence motif to adequately correct the probe signals for sequence effects which might serve as the basic ingredient of improved calibration methods.

Methods

Probe design for SNP detection

SNP arrays intend to determine genotype and copy numbers of hundreds of thousands of bi-allelic single nucleotide polymorphism-(SNP) loci in one measurement. Let us specify each SNP by the alternative nucleotides in the sense DNA-strand of allele A and allele B using the convention B_A/B_B, where $B_A/B_B \in \{A/C, A/G, A/T, C/G, C/T, G/T\}$ stands for one of six SNP types considered on GeneChip SNP microarrays. These SNP types are either complementary (cSNP: A/T, C/G) for substitutions of complementary nucleotides or non-complementary (ncSNP) otherwise.

On Affymetrix 100k GeneChips, each allele is interrogated by ten perfect match (PM)-probes, the 25meric sequence of which perfectly matches the genomic target-sequence at the selected SNP position (see Figure 1 for illustration). The probes differ in their SNP position which is shifted by different offsets relative to the middle base, $\delta \in \{-4,\ldots,0,\ldots,+4\}$. Between three and seven of the PM probes refer to the sense strand and the remaining seven to three probes refer to the antisense strand.

Each PM-probe is paired with one mismatch (MM)-probe of identical sequence except the middle base which intends to estimate the contribution of non-specific background hybridization to the respective PM-probe intensity. Note that the mismatched pairing noticeably reduces specific binding of the respective target to the MM probes compared with the respective PM-probe. The middle base is substituted by its Watson-Crick complement as standard (for example A$\leftrightarrow$T) except for the probes interrogating cSNPs with offset $\delta = 0$, i.e. in the middle of the probe sequences. The non-complementary replacements A$\leftrightarrow$G and T$\leftrightarrow$C are realized in this special case to avoid inter-allelic specific binding to the MM (see below).

Taken together, each allele of each SNP is probed by a set of 20 PM/MM probe pairs. These, in total 40 probe split into two subsets of 10 probe pairs for each allele which we will term 'allele-set'. Each allele-set consists of probes with the SNP interrogation position placed at the sense and antisense strands and moving the 25meric probe sequence up and down the target sequence with respect to the SNP locus by different offsets to improve the accuracy of genotyping and copy number estimates.

Both allele sets use the same offset positions. Therefore each particular offset, δ, is probed by one probe pair for each allele. These four probes (i.e. two PM/MM-pairs) addressing each offset position make up the so-called probe-quartet referring to the same 25-meric segment of the target genome (see Figure 1).

Hybridization modes on SNP arrays

SNP microarrays are hybridized with fragmented genomic DNA representing the targets for the probes attached on the chip surface. Let us consider one SNP locus of a heterozygous genotype: The hybridization solution of genomic DNA consequently contains targets of both alleles A and B. The hybridization reactions can be described by three coupled equations for each probe,

$$\begin{aligned} &P-G+G \rightleftarrows P-G\bullet G && \text{allele-specific (S)} \\ &P-G+G' \rightleftarrows P-G\bullet G' && \text{cross-allelic (C)} \\ &P-G+N \rightleftarrows P-G\bullet N && \text{non-specific (N)} \end{aligned} \quad , \qquad (1)$$

where P-G (P = PM, MM) denotes the probes which are designed to interrogate targets of allele G = A, B. G' = B, A are the targets of the respective alternative allele.

In the allele-specific hybridization mode (called S-mode) the probes bind the target which they intend to detect via duplex formation of the type P-A•A and P-B•B, respectively. In the cross-allelic hybridization mode (C-mode) the probes bind targets of the alternative allele in duplexes of the type P-A•B and P-B•A, respectively. The considered probes also bind non-specific genomic fragments not referring to the selected SNP. Such non-

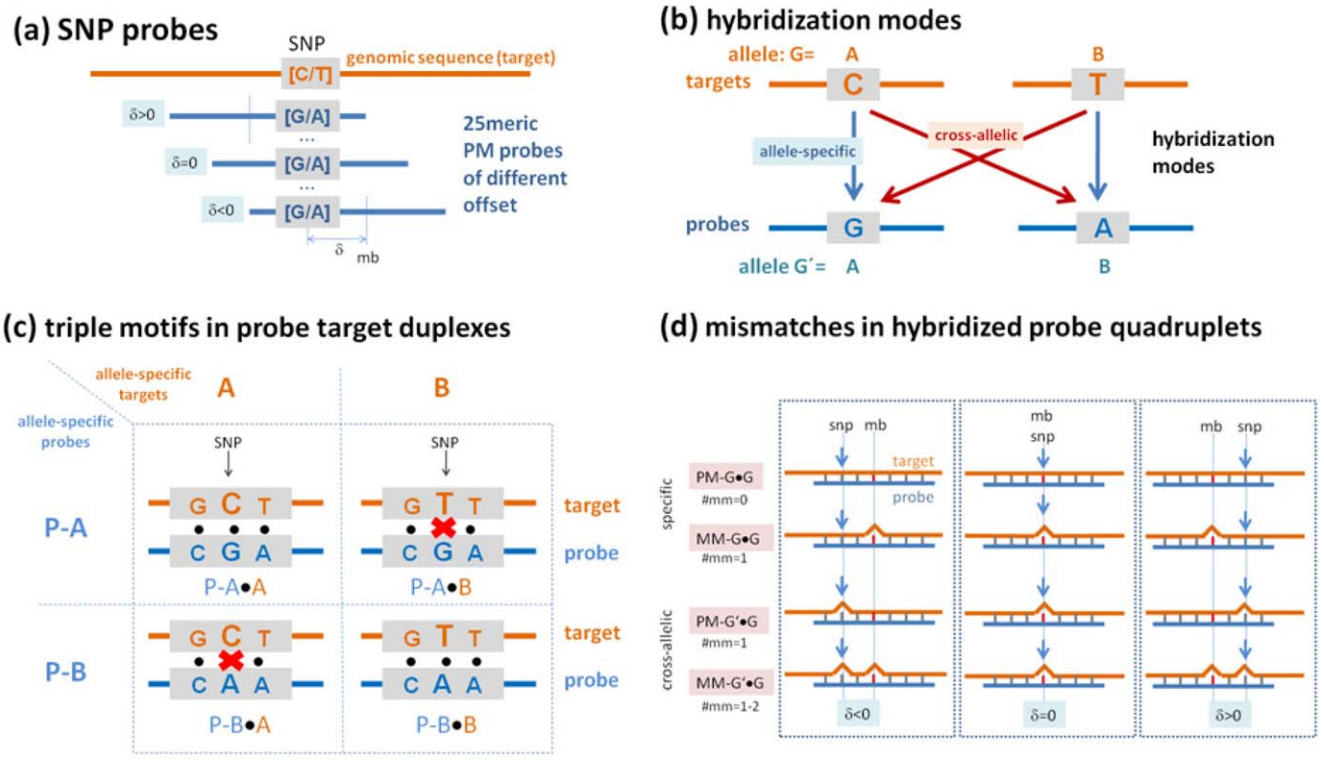


Figure 1. Probe design and hybridization modes for SNP detection. (a) Each SNP (for example [C/A]) is probed by 25meric probes of complementary sequence. Different offsets δ of the SNP position relative to the middle base (mb) of the probe sequence are used. In addition, each PM probe is paired with one MM probe the middle base of which mismatches the target sequence (not shown). (b) The allele-specific probes intend to detect the respective targets via allele-specific binding which however competes with cross-allelic hybridization of targets of the alternative allele (see also the reaction equation Eq. (6)). (c) Both hybridization modes give rise to four different types of probe/target duplexes formed by the two allele-specific probes. The figure shows the respective base pairings for a selected SNP-triple which consists of the SNP [C/T] and its nearest neighbors. Mismatched non-canonical pairings are indicated by crosses. (d) Each box includes one probe-quartet which consists of two PM/MM-probe pairs interrogating either targets of allele G = A or targets of allele G′ = B and vice versa (i.e. G = B and G′ = A). Only targets of one allele are assumed to be present as in the sample. They hybridize to the probes of both allele sets forming either specific or cross-allelic duplexes, respectively. The three selected probe quartets differ in the offset δ of the SNP position (see arrows and part a of the figure) relatively to the middle base of the probe. The different combinations give rise to different numbers and positions of mismatched pairings which are indicated by the bulges. Their number varies between #mm = 0 and #mm = 2 in dependence on the probe type, hybridization mode and offset position. Complete probe-sets use 10 probe quartets.
doi:10.1371/journal.pone.0007862.g001

specific duplexes are of the type P-A•N and P-B•N where N subsumes all non-specific target sequences with non-zero affinity to the selected probe.

In the S-mode the PM probes completely match the target sequence whereas in the C-mode the PM-sequence mismatches the target at the SNP position. The respective MM probes mismatch the target either only at the middle position (S-mode) or at both the middle and the SNP position (C-mode). The respective base pairings are specified below.

The measured intensity of each probe represents the superposition of contributions originating from the three hybridization modes, and from the optical background caused by the dark signal of the scanner and by residual fluorescent markers not attached to target-fragments,

$$I^{P} = I^{P,S} + I^{P,C} + I^{P,N} + I^{O}. \quad (2)$$

In a first order approximation, the intensity-contributions are directly related to the respective number of probe/target-duplexes (indicated by the square brackets),

$$I^{P,S} \propto [P - G\bullet G], \; I^{P,C} \propto [P - G\bullet G'] \text{ and } I^{P,N} \propto [P - G\bullet N]. \quad (3)$$

The non-specific and optical background contributions used in Eq. (2) are, on the average, independent of the probe type (e.g., $I^{PM,N} \approx I^{MM,N}$). We combine both contributions into one mean background intensity

$$I^{BG} = I^{P,N} + I^{O}. \quad (4)$$

Its fraction and the fraction of non-specific hybridization,

$$x^{P,BG} \equiv I^{BG}/I^{P} \quad \text{and} \quad x^{P,N} \equiv I^{N}/(I^{P} - I^{O}), \quad (5)$$

define the percentage of background intensity in the total signal and the percentage of non-specific hybridization signal in the total signal after correction for the optical background, respectively.

Homozygous-present and homozygous–absent probes

Three types of targets compete for duplex formation with each probe in the general case considered in Eq. (1). In the special case of homozygous genotypes only targets of one allele are present in the hybridization solution. As a consequence, the types of competing targets per probe reduce to two ones, namely non-specific and either allele-specific or cross-allelic targets. Particu-

larly, the probes targeting the present allele hybridize specifically (homozygous-present probes) whereas the probes interrogating the alternative allele hybridize in the cross-allelic mode (homozygous-absent probes), i.e.

$$\begin{array}{ll} P-G+G\rightleftharpoons P-G\bullet G & \text{homozygous-present (hp)} \\ P-G'+G\rightleftharpoons P-G'\bullet G & \text{homozygous-absent (ha)} \end{array}. \quad (6)$$

Eq. (2) applies to the special situations of homozygous-present and -absent hybridizations with $I^{P,C}=0$ and $I^{P,S}=0$, respectively (see Figure 1 for illustration).

Matched and mismatched base pairings in probe/target duplexes

In this section we specify the base pairings formed in the probe/target duplexes at two selected sequence positions, namely that of the SNP- and that of the middle-base of the probe sequence. The SNP position is shifted by the offset δ with respect to the middle base. SNP- and middle-base are consequently identical for $\delta=0$.

In the specific hybridization mode the PM probes perfectly match the respective target-allele forming Watson-Crick (WC) pairings along the whole probe sequence including the two selected positions (Figure 1 and Text S1). Contrarily, one mismatched pairing occurs at the SNP position of the PM probe upon cross-allelic hybridization. The MM probe always forms a mismatched pairing at the middle position and, upon C-hybridization, also at the SNP position. For $\delta\neq0$ the MM-duplexes contain consequently two mismatches with the special case $\delta=\pm1$ referring to so-called tandem-mismatches of two adjacent mismatched pairings. For $|\delta|>1$ the two mismatches are separated by at least one WC pairing. The MM form only one mismatch in the C-hybridization mode for $\delta=0$ because the mismatched SNP position equals the middle base.

The assignment of the specific and cross-allelic hybridization modes to the six probed bi-allelic SNP types B_A/B_B (see above) and the two probe types (P = PM, MM) provides the full set of 16 possible base pairings in the probe/target duplexes at their SNP- and/or middle-position (see Text S1). We classify the pairings into canonical Watson-Crick pairs (referred to as At-group; upper and lower case letter refer to the probe and target sequences, respectively), and three groups of mismatches (Aa-, Ag- and Ac-group). The notations of the groups are chosen in agreement with the respective pairing formed by an adenine in the probe sequence (see Text S1 for the details). The mismatched groups refer to self-complementary pairings (Aa-group: Aa, Tt, Gg, Cc), to self-paired (Ag-group: Ag, Tc, Ga, Ct) and cross-paired (Ac-group: Ac, Tg, Gt, Ca) pyrimidines and purines, respectively. Note that these groups are invariant with respect to the strand direction because complementary substitutions do not change the group membership.

The number and the type of the mismatches are not specified in probe/target duplexes formed in the non-specific hybridization mode. Nevertheless, the sequence effect can be described in terms of the properties of canonical WC pairings [20,21]. This result seems to contradict the fact that non-specific duplexes are per definition destabilized at minimum by one, but typically by more mismatched pairings. Note however, that these mismatch-effects are averaged out by calculating mean binding characteristics of WC-interactions (At-group) which stabilize the non-specific duplexes.

Interaction modes

As discussed in the previous subsections, the probe/target duplexes are characterized by the hybridization mode (h = S, C, N) and a series of probe attributes: probe-type (P = PM, MM), probe sequence and middle base (B_{13} = A,T,G,C), strand direction (d = s, as), SNP type (B_A/B_B), and SNP offset ($\delta=-4,\ldots,0,\ldots,+4$). Each particular combination of the hybridization mode with a set of probe attributes unambiguously determines the interaction mode between probe and target. It is characterized by

(i) the base pairing at the SNP position and at the middle position, which includes all 16 pairwise combinations of nucleotides, 4 of which form WC pairings and 12 of which are mismatches;

(ii) Watson-Crick pairings at the remaining positions of the probe sequence;

(iii) the mutual shift between the middle and the SNP base by up to four bases in both directions (δ);

(iv) different numbers of mismatches per duplex varying between #mm = 0 (for P = PM and h = S) and #mm = 2 (P = MM and h = C, only $\delta\neq0$);

(v) different relative positions of paired mismatches (#mm = 2) which are either separated by at least two WC pairings ($|\delta|>1$) or form tandem-mismatches ($|\delta|=1$).

The design of SNP GeneChips thus enables us to study how these interaction modes affect the probe intensities in a systematic way. Vice versa, the probe intensities are related to the amount of bound DNA-targets which, in turn, depends on the stability of the duplexes and thus on the binding constant of the respective interaction mode. Knowledge of the binding constant and of the interaction mode then allows us to compute the genotype call and copy number of a given SNP.

SNP array data

Intensity-data of the 100k GeneChip SNP array and supplementary files were downloaded from suppliers website (https://www.affymetrix.com/support/technical/sample_data/hapmap_trio_data.affx). This data set was specially designed for the development and evaluation of low-level analysis methods for genotyping and copy number estimation from probe intensity data (see, e.g., [22]). Particularly we analyzed array NA06985_Xba_B5_4000090 taken from the Mapping 100k HapMap Trio Dataset (100K_trios.xba.1.zip) including library- and annotation information (probe sequences, fragment lengths and GC-content of the targets, GCOS-genotype calls). We use the genotypes provided by Affymetrix for the array data and select only homozygous SNP loci for further analysis (41,629 homozygous out of 58,960 total loci, ~70.1%). In this special case the hybridization mode is either specific or cross-allelic for homozygous-present and homozygous-absent alleles, respectively (see Eq. (6)).

The data are further filtered to remove probe intensities which are dominated by nonspecific hybridization by more than $x^{P,N}>0.2$ (Eq. (5)). These selection criteria are chosen from the hook plot of the chip data which is briefly described in the supporting text (see Text S1 and also refs. [5,6]). This special type of analysis characterizes the hybridization quality of each chip. Interestingly, the data obtained reveal that nonspecific hybridization contributes to the signal intensities of SNP arrays to a smaller degree compared with expression arrays in agreement with previous results [9]. This difference can be rationalized by the smaller heterogeneity of genomic DNA copies (with respect to their sequences and fragment-lengths) and especially by the smaller range of copy number variations compared with the range of variation of mRNA-transcript concentrations. The latter values can cover several orders of magnitude whereas the former ones typically change by a factor of less than ten.

The intensity data are corrected for the optical background intensity and for residual non specific hybridization before further analysis as described in Text S1.

Triple averaged intensities and probe sensitivities

We previously used the so-called 'triple-averaging' approach to estimate the effective strength of base pairings in probe/target duplexes on GeneChip expression arrays [20]. This approach analyzes the effect of the sequence on the probe intensities using triples of neighboring bases. It accounts for the fact that the strength of a selected base pair interaction in oligonucleotide duplexes is significantly modulated by the two adjacent pairings on both sides of the selected base.

Let us define the standard triple as the string of three consecutive bases (xBy) in 5′→3′-direction of the probe sequence (x,B,y∈A,T,G,C) where the nearest neighbors (x, y) of the central base B form Watson-Crick pairs in the duplexes with the targets. The position of the triples along the probe sequence was chosen in such a way that its central base (B) agrees either with the middle base (mb) or with the SNP base (see Figure 1c for illustration). The triple is consequently centered about the middle base of the probe ($\delta = 0$) or shifted by δ sequence positions up or downwards ($\delta \neq 0$). The hybridization mode and the probe attributes unambiguously define the base pairing of the center base, Bb (b∈a,t,g,c), according to the selected interaction group, Ab = At, Aa, Ag or Ac (see Text S1). The Ab-group can be chosen by applying appropriate criteria of probe selection.

So-called triple averages of the intensity are calculated as log-mean over all probes within the classes defined by the interaction group of the central base (Ab = At, Aa, Ag or Ac), by the triple motif xBy at offset position ($\delta = -4,\ldots,0,\ldots,+4$) and by the number of mismatches per duplex (#mm = 0, 1 or 2)

$$\log \mathrm{I}^{\mathrm{P-T\bullet G}}_{(\mathrm{Ab},\delta,\#\mathrm{mm})}(\mathrm{xBy}) = \left\langle \log \mathrm{I}^{\mathrm{P-T\bullet G}} \right\rangle_{\mathrm{p\in class}}, \qquad (7)$$

with T = G,G' for hp- and ha-probes, respectively. A series of nested mean values can be generated by averaging over one or more of the attributes given by class = (Ab, δ, #mm). For example, $\log \mathrm{I}_{(\mathrm{Ab},\#\mathrm{mm})}(\mathrm{xBy}) = \left\langle \log \mathrm{I}^{\mathrm{P-T\bullet G}} \right\rangle_{\mathrm{p\in class},\delta}$ denotes averaging over the offset positions δ and $\log \mathrm{I}_{(\mathrm{Ab},\#\mathrm{mm})} = \left\langle \log \mathrm{I}(\mathrm{xBy}) \right\rangle_{\mathrm{xBy}}$ refers in addition to averaging over the triple motifs xBy to get the mean intensity per interaction group.

The triple sensitivities are defined as the deviation of the triple-averaged intensity from an appropriately chosen mean value over all triples (see below and [23]), e.g.,

$$\mathrm{Y}_{(\mathrm{Ab},\delta,\#\mathrm{mm})}(\mathrm{xBy}) = \log \mathrm{I}_{(\mathrm{Ab},\delta,\#\mathrm{mm})}(\mathrm{xBy}) - \left\langle \log \mathrm{I}_{(\mathrm{Ab},\delta,\#\mathrm{mm})} \right\rangle_{\mathrm{xBy}}. \qquad (8)$$

It is reasonable to assume that the strand direction does not affect the strength of the respective base pairings. In our analyses we therefore pool the probes which are assigned to the same interaction mode independently of their strand direction (d = s, as) assuming that the respective genotypes are properly assigned on both strands.

Tandem and flanking mismatches

Special selection criteria for triples with one flanking mismatch and of tandem mismatches are given in the scheme shown in Text S1. The former motif is characterized by the usual standard triple as defined in the previous section which is however flanked on one side by a mismatched pairing, i.e. w(xBy)m (w∈At; m∈Aa,Ag,Ac). Tandem mismatches are two adjacent mismatches present in homozygous-absent duplexes of the MM-probes with SNP offset positions $|\delta| = 1$. Both motifs were separately analyzed to estimate the specific effect of flanking and of tandem mismatches in comparison with the standard triples.

Results

SNP offset position and the number of mismatches

The specific and cross-allelic hybridization modes include perfect matched and mismatched probe/target duplexes with up to two mismatched pairings at the SNP- and/or mb-position (see Text S1). To study the effect of the number of mismatched pairings, #mm, and the effect of the SNP offset position, δ, on the intensities we calculate the log-intensity averages, $\log \mathrm{I}^{\mathrm{P-T\bullet G}}_{(\#\mathrm{mm},\delta)} = \left\langle \log \mathrm{I}^{\mathrm{P-T\bullet G}} \right\rangle_{\mathrm{Ab,xBy}}$, for each SNP offset of homozygous-present (T = G) and absent probes (T = G', see part a of Figure 2 and Eq. (7)).

The SNP base of each probe forms a WC pairing in P-G•G duplexes (hp-mode). The respective averaged intensities per SNP position are consequently pseudo-replicates of different sub-ensembles of probes referring to the same interaction mode, namely perfectly-matched (PM-G•G) or single-mismatched (MM-G•G) probe/target duplexes (see the schematic drawings in panel a of Figure 2). The scattering of the respective data about their mean thus reflects the variability of the obtained intensity averages in the different sub-ensembles of probes.

In P-G'•G duplexes (allele absent/ha-mode) the SNP base forms a mismatched pairing. The averaged intensities consequently refer to the shift of the mismatch relative to the middle base. For the PM probes (PM-G'•G) the position of the respective single mismatch only weakly affects the mean intensity in the relevant range of SNP offsets (panel a of Figure 2). This result is in agreement with previous studies which show that the destabilizing effect of single mismatches is almost constant over a broad range in the middle part of short-length oligonucleotide duplexes and decreases only for the last 4–6 base positions near the ends of the probe sequence [24–26].

In contrast, the MM-probes form two mismatches in the homozygous-absent mode (MM-G'•G) at the SNP- (for $|\delta| > 0$) and at the middle position. Both mismatches are separated by $(\delta - 1)$ WC pairings in-between. The observed mean intensity decreases with increasing distance between the mismatches (panel a of Figure 2). This trend indicates that the destabilizing effect of the mismatches is small for neighboring tandem mismatches ($|\delta| = 1$); it slightly increases for a single intermediate WC pairing ($|\delta| = 2$) and it essentially levels off for more WC pairings in between ($|\delta| > 2$).

The presented results show that the number of mismatched pairings per duplex (#mm) is the most relevant factor which affects the mean intensity of the probes (see the horizontal lines in Figure 2, panel a). The logarithmic-intensity ratio can be approximated as function of #mm by [27]

$$\frac{\log \mathrm{I}(\#\mathrm{mm})}{\log \mathrm{I}(0)} \propto \frac{\log \mathrm{K}_{\mathrm{duplex}}(\#\mathrm{mm})}{\log \mathrm{K}_{\mathrm{duplex}}(0)} \approx \mathrm{x}\left(1-\gamma(1-\mathrm{x}^2)\right) \quad \text{with} \quad \mathrm{x} = 1 - \frac{\#\mathrm{mm}}{25}, \qquad (9)$$

where $\mathrm{I}(\#\mathrm{mm}) = \left(\mathrm{I}_{\#\mathrm{mm}} - \mathrm{I}^{\mathrm{BG}}\right)$ is the background corrected intensity of probes with #mm mismatches; $\mathrm{K}_{\mathrm{duplex}}(\#\mathrm{mm})$ denotes the respective mean association constant of probe/target duplexes with #mm mismatches; x is the fraction of WC pairings in the duplex and γ is a fit-constant depending on the hybridization conditions.

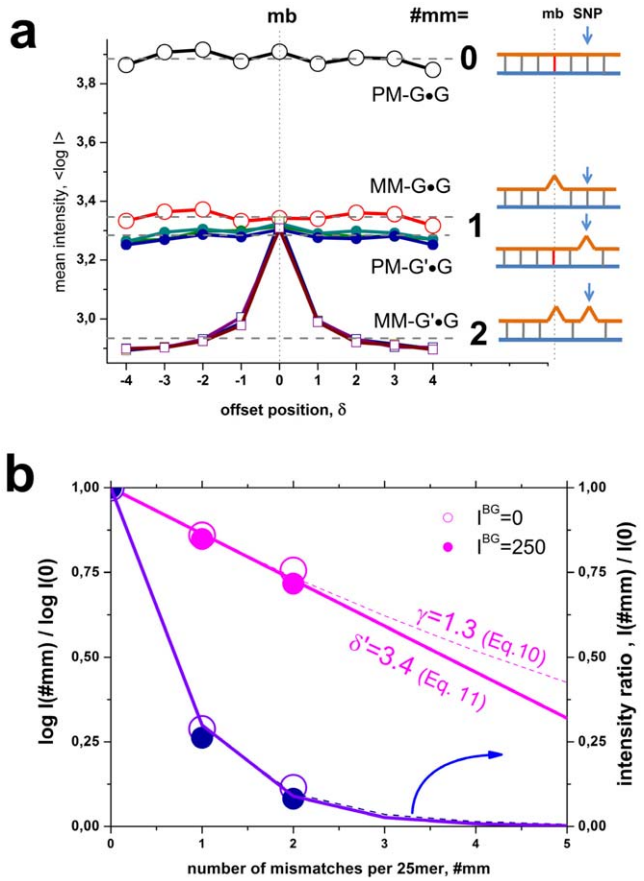


Figure 2. SNP offset and number of mismatches. Averaged log-intensities for probes of different mismatch-groups and offset-positions, (panel a) and mean effect of the number of mismatches (#mm) on the observed intensity (panel b). Panel a: Mean probe intensity (averaged over all probes with a given SNP offset, see the arrow in the schematic drawing in the right part for illustration) as a function of the offset-position of the mismatch with respect to the middle base (δ) for different number of mismatches per probe/target duplex (#mm = 0…2). Virtually no significant effect of the offset-position was observed for single mismatches within the relevant range $|\delta|<5$. Contrarily, the mean intensity decreases with increasing separation between double mismatches (#mm = 2) where one is located in the centre of the probe (middle base, mb) and the second one at offset position δ. Note that both mismatches merge into one for $\delta = 0$. The homozygous-absent data (P-G'•G) were separately calculated for the three groups of mismatches, Aa, Ac and Ag: The respective curves are almost identical. Panel b: Relative decrease of the mean probe intensity as a function of #mm (symbols). The curves are calculated using Eqs. (9) and (10). The data are shown in logarithmic (left axis, upper data) and linear (right axis) scale without (open symbols) and with (solid symbols) background correction.
doi:10.1371/journal.pone.0007862.g002

An alternative, simple "mismatch"-function results from the assumption of additive contributions of each base pairing, $\log \mathrm{K}_{\mathrm{duplex}}(\#\mathrm{mm}) \approx \log \mathrm{K}_{\mathrm{duplex}}(0) - \#\mathrm{mm}\cdot\delta\varepsilon$, where $\log \mathrm{K}_{\mathrm{duplex}}$ is related to the free energy of duplex stability and δε is its mean incremental penalty (in units of $\log \mathrm{K}_{\mathrm{duplex}}$) if one substitutes one WC pairing by a mismatch. This approach predicts an exponential decay of the intensity as a function of the number of mismatches, $\mathrm{I}(\#\mathrm{mm}) \approx \mathrm{I}(0)\cdot 10^{-\#\mathrm{mm}\cdot\delta\varepsilon}$, which transforms into

$$\frac{\log \mathrm{I}(\#\mathrm{mm})}{\log \mathrm{I}(0)} = 1 - \delta'\cdot(1-\mathrm{x}) \quad \text{with} \quad \delta' = \frac{\delta\varepsilon}{\log \mathrm{K}_{\mathrm{duplex}}(0)/25}, \tag{10}$$

using the logarithmic form as in Eq. (9). The constant δ′ is given by the ratio between the incremental penalty due to the mismatch and $\log \mathrm{K}_{\mathrm{duplex}}(0)/25$, which has the meaning of a mean additive contribution of one WC pairing to $\log \mathrm{K}_{\mathrm{duplex}}(0)$. Panel b of Figure 2 shows that both alternative functions given by Eqs. (9) and (10) are virtually not distinguishable for #mm<3. They can be used to extrapolate the intensity values to #mm>2 in a rough approximation. The data show that one and two mismatches reduce the intensity to about 25% and 10% of its initial value, respectively. Eqs. (9) and (10) predict that more than two mismatches decay the intensit to tiny values of less than 5% of its value for perfect matched duplexes. The estimated value of the decay rate δ′>3 in Eq. (10) indicates that the intensity penalty due to the first two mismatches markedly exceeds the average intensity contribution of a single WC pairing in the perfect matched probe/target duplexes. Simple balance considerations imply that δ′ has to decrease with increasing number of mismatches as predicted by Eq. (9) (see also the theoretical curves in part b of Figure 2).

Positional dependence of single base- and triple-motifs

The PM probes form exclusively WC pairings in homozygous-present PM-G•G duplexes. We calculated log-mean intensities for all these duplexes containing a certain base (B = A,T,G,C) at each position k = 1…25 of the probe sequence to study the positional effect of WC-base pairings over the whole sequence length (see lines in panel a of Figure 3). The obtained positional-dependent log-intensity averages only weakly vary about their total mean. The base-specific differences essentially disappear towards the right end of the probes (k>23) which is attached to the chip surface (see also panel c of Figure 3).

Also the homozygous-present duplexes of the MM-probes, MM-G•G, form predominantly WC pairings except the middle base which forms mismatches of the Aa-interaction group. The single base averaged intensities of these mismatches vary to a much larger degree about their mean compared to the WC pairings (see the arrow in panel a of Figure 3). The strong mismatch effect extends also to the flanking bases at adjacent positions k = 12 and 14.

Panel b of Figure 3 shows the single-base positional dependence of homozygous-absent PM probes (PM-G'•G) for different offsets δ of the SNP which forms a mismatched pairing in the probe/target duplexes. As for the MM, the SNP position exhibits a larger spread of the single-base values about their mean compared with the WC pairings at the remaining sequence positions. They represent averages over mismatches of the Aa-, Ag- and Ac-type in contrast to the Aa-type mismatches of the middle base shown in panel a of Figure 3. The data clearly reflect the shift of the mismatched pairing with changing offset position of the SNP. The profiles remain nearly invariant at the remaining sequence positions.

To estimate the effect of longer sequence motifs we calculated intensity-averages of probes possessing "homo"-triples, i.e. runs of three consecutive bases of the same type at a certain sequence position (see panel c and d of Figure 3). The specific effect of these motifs clearly exceeds that of the single bases, especially for runs of triple G: These GGG-motifs systematically reduce the probe intensities by a factor of $\sim 10^{-0.2}-10^{-0.4} \approx 0.6-0.4$ compared with the mean intensity for most of the sequence positions. In contrast, the mean effect of a single G is almost negligible. The GGG-effect essentially disappears at the mismatch position in the middle of the probe sequence (see panel d of Figure 3 which shows profiles of PM-G'•G probes with $\delta = 0$). The similar "buckled" shape of the GGG-profile in the middle of the probe sequence of PM-G•G duplexes (panel c) probably indicates a certain small fraction of incorrectly assigned genotypes in the selected sub-ensemble of homozygous present probes.

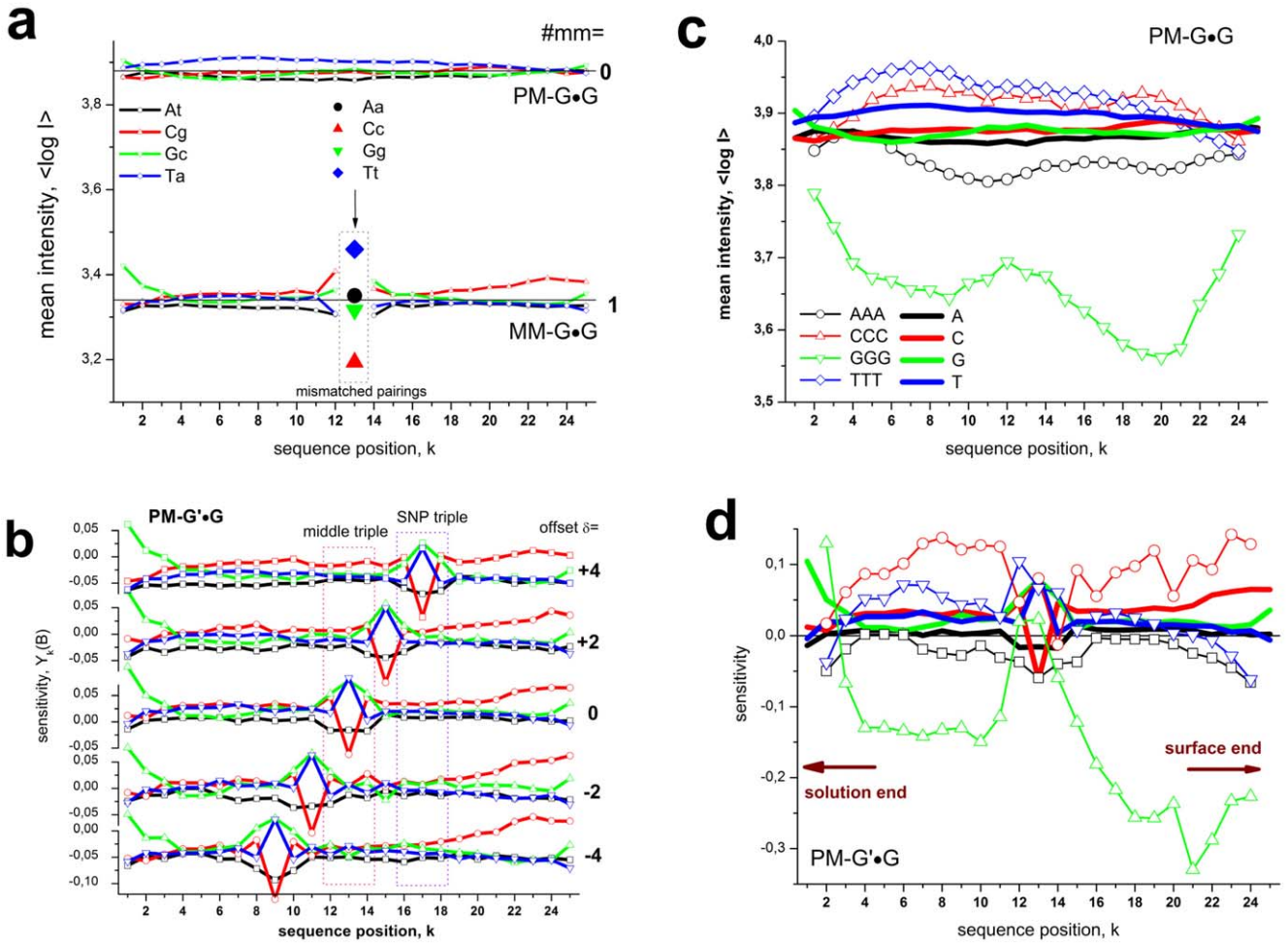


Figure 3. Positional dependence of the probe intensities. Panel a: Single base data of allele-specific (S-mode) PM and MM probes. Each data point was calculated as log-intensity average over all probes of the considered class with the indicated base at position k of the probe sequence. It is associated either with WC pairings or with mismatched pairings at the middle base (mb)-position of the MM. These mismatches give rise to markedly larger variability of the intensities than the WC pairings do at the remaining positions. Panel b shows the positional dependence of the sensitivity (deviation of the log-intensity from its mean over all probes of the class) of cross-allelic PM probes (C-mode) with different offsets of the SNP. The base at the SNP position forms a mismatched pairing which shifts along the sequence according to the offset. Note that the mismatch-values are averages over all groups (Aa, Ag, Ac; see Text S1) whereas the mismatches in part a of the figure refer to the Aa-group. Panel c enlarges the single-base curves for PM-G•G shown in panel a. In addition, mean log-intensity values were calculated for homo-triples along the probe sequence (the position k refers to the center base of the triples). The mean log-intensities slightly increase for AAA, CCC and TTT compared with the single-base averages but markedly decrease for triple guanines. Panel d shows the respective single-base and triple values for the cross-allelic PM data for offset $\delta = 0$ shown in panel b. Comparison with panel c indicates subtle differences of the curves at positions which refer to WC pairings in both situations: For example, triple-guanines motifs give rise to relatively large intensities near the surface end of the probe and also the cytosines (C- and especially CCC-motifs) are associated with largest intensities for most of the WC pairings in part d whereas thymines give rise to largest intensities in part c.
doi:10.1371/journal.pone.0007862.g003

Comparison of panel c and d of Figure 3 reveals also more subtle differences between the profiles at positions which refer to WC pairings in both, the PM-G•G (panel c) and PM-G'•G (panel d) duplexes: Firstly, triple TTT provide the largest intensities for the former duplexes whereas triple CCC become largest in PM-G'•G duplexes. Moreover, the effect of cytosines progressively increases towards the surface end in PM-G'•G duplexes whereas it apparently disappears in the data obtained from PM-G•G duplexes. Secondly, the intensity effect due to guanines begins with positive values at the solution end of PM-G'•G duplexes ($k = 1$) and then steeply decreases to negative values.

It is known that the sequence profiles are sensitive to factors such as the optical background correction and saturation [18,28] (see also below). Large and small intensities are prone to saturation and background effects, respectively, which differently affect the specific signal. Saturation, for example, limits large probe intensities and therefore reduces the relative effect of strong-base pairings because probes containing such motifs are most affected by saturation. The relative small single- and triple- cytosine values in the profiles of PM-G•G duplexes can be attributed to selectively stronger saturation of probes containing these motifs. Contrarily, in the PM-G'•G duplexes saturation is much less relevant owing to the smaller average level of probe occupancy and intensity. The different response of triple guanines and cytosines near the solution and surface ends of the probe seems puzzling and will be addressed in the discussion section.

Triple sensitivities

In the next step we neglect the positional dependence of probe intensities and address the sequence-specific effect of base pairings

in triple motifs centered about the middle and SNP base of the probes.

The triple averaged and background corrected intensities were used to calculate the 64 triple-sensitivity values for each of the four interaction groups, (Eq. (8)). Particularly, we selected the homozygous-absent PM probes (PM-G'•G) with one mismatched pairing at SNP position and used the base-triples centered about the middle base (At-group) and about the SNP base (Aa, Ag, Ac group, see Text S1). All intensities of probes with offset-positions $|\delta|>1$ were log-averaged. The sensitivity values were related to the total mean of all used PM-G'•G probes irrespective of the particular interaction group, i.e.,

$$Y_{Ab,\#mm=1}(xBy) = \langle \log I_{(Ab,|\delta|>1,\#mm=1)}(xBy) - \langle \log I_{(Ab,|\delta|>1,\#mm=1)}(xBy)\rangle_{Ab,\delta,xBy}\rangle_{\delta} \quad (11)$$

Figure 4 summarizes the obtained sensitivity data which provide a measure of the specific effects of the pairing of the central base and of their nearest neighbors in terms of the deviation from the mean over the respective group of probes.

Most of the sensitivities of the At-group (WC pairings) relatively tightly scatter about their mean indicating an only moderate sequence effect. The 'GGG'-triple however strongly deviates from this rule; it causes a relatively large intensity penalty: One 'GGG'-motif give rise to the reduction of the intensity on the average by a factor of about $10^{-0.2}\sim0.63$ compared with the mean intensity. The triples considered refer to offset positions $|\delta|\leq4$ about the middle base. The full positional dependence of 'GGG' (Figure 3, part d) actually indicates a stronger intensity drop for sequence positions halfway to the ends. Importantly, the 'GGG'-penalty is in contradiction to complementary rules because the complementary 'CCC'-motif reveals completely different sensitivity-properties: Triple C's gives rise to the opposite effect; i.e. they amplify the intensity by a factor of about $10^{+0.1}\sim1.25$. We will discuss this puzzling result below.

The substitution of the central WC pairing by mismatches considerably increases the variability of the triple data. The mean variability of each interaction group was estimated in terms of the standard deviation of all 64 combinations of each group (Table 1): Its value more than doubles for the mismatched groups (SD = 0.09–0.13) compared with the WC-group (SD = 0.04). Single mismatches can modify the intensity by a factor between $\sim10^{-0.25}=0.55$ and $\sim10^{+0.25}=1.8$. This result generalizes the trend which is illustrated in Figure 3a for the special case of mismatches of the Aa-group in the middle of the probe sequence.

Mean mismatch stability

The mean sensitivity over all triples with a given middle base B provides a measure of the average stability of the respective mismatched pairing Bb (see the red lines in Figure 4). For the Aa-, Ag- and Ac-groups one gets the relations Cc<Gg≈Aa<Tt, Tc≈Ct<Ag≈Ga and Ac≈Ca<Gt≈Tg, respectively. They confirm the expected symmetries for bond reversals Bb→B^rb^r in symmetrical DNA/DNA interactions, i.e. $Y_{Ab}(Bb)\approx Y_{Ab}(B^rb^r)$ (for example for Tc→Ct and Ac→Ca). Note that, in contrast, DNA/RNA interactions are asymmetrical in solution [29] and on microarrays [18,20].

Comparison of the mean sensitivity values for each central pairing of all three mismatch-groups provides the following ranking of the stability of mismatched pairings:

$$\mathbf{Tc}(-0.10)\leq\mathbf{Ct}(-0.09)\leq\mathbf{Cc}(-0.08)\approx\mathbf{Ac}(-0.08)\leq\mathbf{Ca}(-0.06)<0 \quad (12)$$

$$0<\mathbf{Gg}(+0.03)\leq\mathbf{Aa}(+0.05)\approx\mathbf{Gt}(+0.05)\leq\mathbf{Tg}(+0.08)<\mathbf{Ag}(+0.12)\approx\mathbf{Ga}(+0.12)<\mathbf{Tt}(+0.16)$$

The numbers in the brackets are the respective mean sensitivities for each mismatched pairing averaged over the 16 combinations of adjacent bases (standard error: $\sim\pm0.02$).

Other authors report similar rankings of the stability of single-mismatches in DNA/DNA-oligomer duplexes which are obtained from hybridization studies on surfaces (microarrays or special solid supports) or in solution:

$$\mathrm{Gg}\leq\mathbf{Ca}<\mathbf{Ct}\approx\mathbf{Cc}\approx\mathbf{Gt}\approx\mathbf{Aa}<\mathrm{Ac}\approx\mathrm{Tc}\leq\mathbf{Ga}<\mathbf{Tg}\leq\mathbf{Tt}<\mathbf{Ag}$$

(array, [15])

$$\mathbf{Ct}\approx\mathbf{Cc}\leq\mathbf{Ca}\leq\mathbf{Ac}\leq\mathbf{Aa}\approx\mathrm{Tc}\approx\mathrm{Ga}\leq\mathbf{Gt}<\mathrm{Gg}<\mathbf{Tt}<\mathbf{Ag}\approx\mathbf{Tg} \quad (13)$$

(array, [24]).

$$\mathbf{Ac}\approx\mathbf{Tc}\approx\mathrm{Tt}\approx\mathrm{Aa}<\mathbf{Ag}\leq\mathbf{Tg}$$

(support, [30], only selected pairingsare studied)

$$\mathbf{CC}\leq\mathbf{AC}\leq\mathbf{TC}\leq\mathbf{AA}\approx\mathrm{TT}<\mathbf{GA}\approx\mathrm{GT}<\mathrm{GG}$$ (solution, [31])

In solution, both dimerized oligonucleotides are equivalent as indicated by the two capital letters which assign the pairing.

Basic agreement of the reference studies with our ranking is highlighted using bold letters. Accordingly, the consensus-ordering of the array-studies comprises Ct, Ca, Cc as low stability mismatches; Ag, Tg, Tt as high stability mismatches and Gt and Aa at the intermediate position. A major difference between the previous rankings occurs for Gg which is the least stable in the study of Naiser et al. [15] and one of the most stable mismatches in the study of Wick et al. [24]. Our data plead for intermediate stability. Inspection of Figure 4 reveals the large variability of triples with a central Gg-mismatch about zero. Imbalanced triple selection in studies using a limited number of oligonucleotides therefore are prone to lead to biased results where the apparent Gg-stability can vary between large and low values in dependence on the particular realization of probe/target-duplexes containing a Gg-mismatch. The total probe number of the studied SNP array (10^6) exceeds the probe number used in previous studies by about three orders of magnitude (10^3 [24] and $2–3\times10^3$ [15]). Comparison of the different rankings of mismatch strength obtained from chip and solution data reveals disagreement especially for GG, GT and TT motifs. These differences possibly indicate additional or alternative explanations for the inconsistent chip rankings which will be discussed below.

Note also that the reported references [15,24] estimated mismatch-stabilities by directly comparing the intensities of MM and PM probes, which refers to the stability difference between the mismatched pairing and the respective WC pairing. Our ranking uses the mean stability of all considered single-base mismatches as reference level which is independent of the particular triple. The relatively small variability of the single-base averages of the At-group (see the red lines for the At-group in Figure 4) however show that the explicit use of the WC-sensitivity as reference essentially does not change the ranking of mismatch-stabilities in our data set. Direct comparison with the reference data is therefore adequate within the error limits.

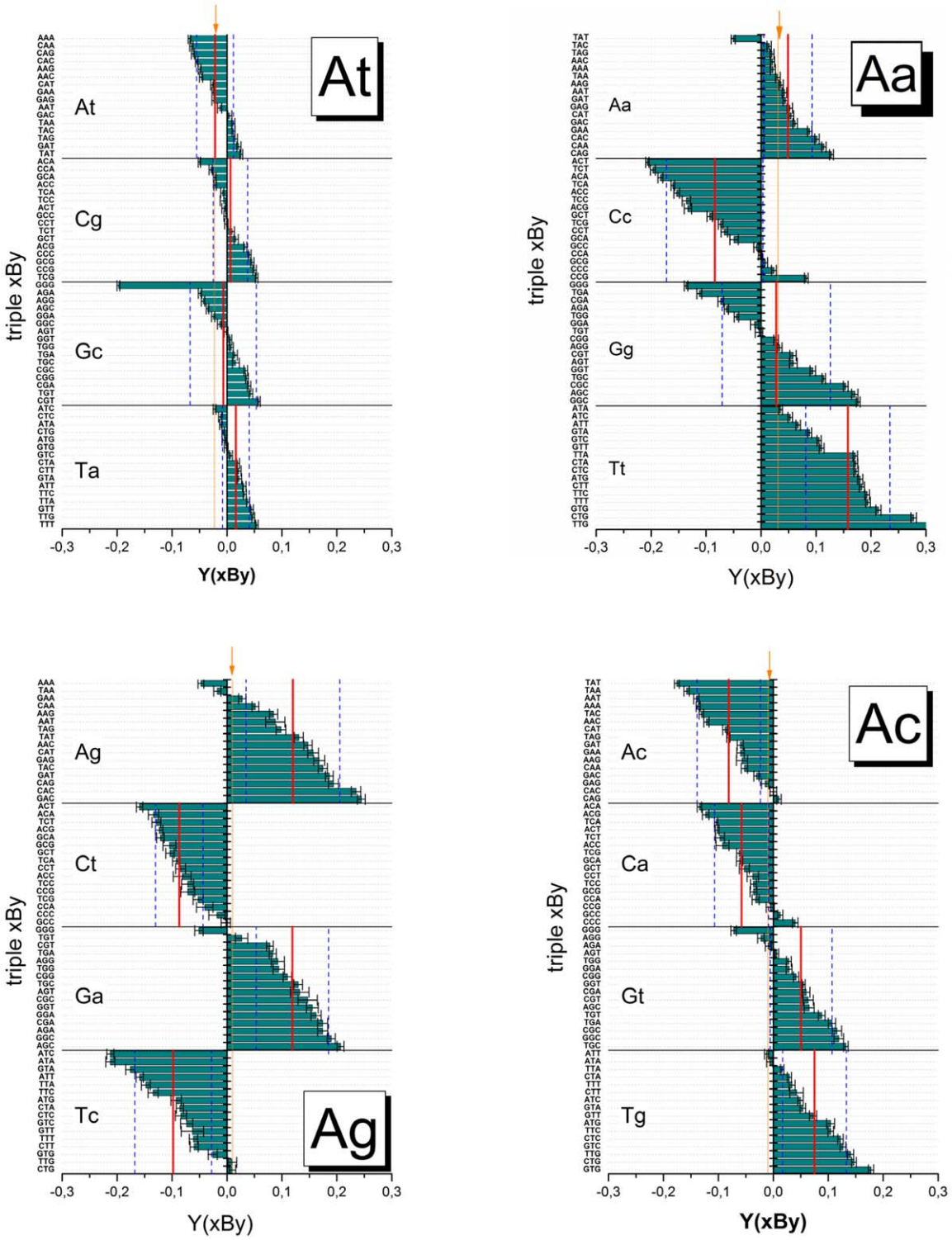


Figure 4. Triple averaged sensitivities. The triple values are calculated using (Eq. (8)) and ranked with increasing sensitivity for each center base B forming matched (group At) and different mismatched (groups Aa, Ag and Ac) pairings with the target as indicated in the figure by upper (probe) and lower (target) letters. The sensitivity-values are calculated relative to the total log-average of all single-mismatched probes of the chip. Sub-averages of the interaction groups (see arrows) and of the central base pairings are shown by vertical solid lines. The vertical dashed lines indicate the standard deviation of the triple values about the central-base related mean (see also Table 1). The mean and the standard deviation estimate the stability of the respective pairing Bb and the effect of flanking WC pairings, respectively. The error bars indicate the standard error of the triple sensitivities.

doi:10.1371/journal.pone.0007862.g004

Table 1. Sources of variability of triple motifs and of tandem mismatches.

Interaction group[a]	At	Aa	Ag	Ac
Base pairings (Watson Crick or mismatches)	**WC pairings: At, Cg, Gc, Ta**	**self complementary mismatches: Aa, Cc, Gg, Tt**	**self paired mismatches: Ag, Ct, Ga, Tc**	**cross paired mismatches: Ac, Ca, Gt, Tg**
triples[b]	0.04±0.001	0.12±0.0005	0.13±0.001	0.09±0.0005
3′/5′-asymmetry[c]	0.03	0.11	0.07	0.05
complementary asymmetry (without GGG)[d]	0.05 (0.02)	0.10 (0.08)	0.08 (0.06)	0.06 (0.05)
NN-residual[e]	0.02 (0.02)	0.02 (0.02)	0.03 (0.02)	0.01 (0.01)
flanking mismatches[f]		0.04	0.03	0.02
tandem mismatches (xy)[g]		0.02 (0.033)	0.015 (0.047)	0.013 (0.044)
tandem mismatches (BB′)[g]		0.06 (0.07)	0.06 (0.10)	0.05 (0.08)
tandem mismatches (yB′/B′y)[g]		0.05	0.08 (0.055)	0.05

[a]variability estimates are separately calculated as standard deviation for each Ab-interaction group: $SD = \sqrt{<\Delta^2>_{Ab}}$.
[b]variability of the triple averages with respect to the group-mean: $\Delta = Y_{Ab}(xBy) - <Y_{Ab}(xBy)>_{Ab}$; it estimates the variability of interactions due to the choice of the triple; the standard error refers to the variability of the probe level data of each interaction group.
[c]variability of the triple averages after 3′/5′-transformation: $\Delta = Y_{Ab}(xBy) - Y_{Ab}(yBx)$.
[d]variability of the triple averages after complementary-transformation: $\Delta = Y_{Ab}(xBy) - Y_{Ab}(x^cB^ry^c)$; the values in the brackets are obtained after omitting the GGG-motif.
[e]variability of the residual values after reduction of the model rank NNN→NN: $\Delta = \Delta^{res}{}_{Ab}$ (see Eq. (17)).
[f]variability due to flanking mismatches: $\Delta = \Delta^{flank}{}_{Ab}$ (see Eq. (15)).
[g]variability due to quadruplet motifs with tandem mismatches (xBB′y)/(yB′Bx) with B∈Aa and B′∈Aa,Ag,Ac. The SD were calculated with respect to the average over the three groups ($\Delta(xy) = <Y_{Ab}(xBB'y)>_{BB'} - <<Y_{Ab}(xBB'y)>_{BB'}>_{Ab}$ and $\Delta(BB') = <Y_{Ab}(xBB'y)>_{xy} - <<Y_{Ab}(xBB'y)>_{xy}>_{Ab}$) and with respect to the total mean over all couples (values in the brackets; $(\Delta(xy) = <Y_{Ab}(xBB'y)>_{BB'} - <<Y_{Ab}(xBB'y)>_{BB'}>_{Ab,xy}$ and $\Delta(BB') = <Y_{Ab}(xBB'y)>_{xy} - <<Y_{Ab}(xBB'y)>_{xy}>_{Ab,BB'})$.
doi:10.1371/journal.pone.0007862.t001

Symmetries

The triple sensitivities shown in Figure 4 can be examined with respect to two simple symmetry-relations, namely 3′/5′-reversal and probe/target-complementarity,

$$xBy \rightarrow yBx \quad \text{and} \quad xBy \rightarrow y^cB^rx^c, \qquad (14)$$

respectively (sequence motifs are ordered in 5′-3′ direction). The superscripts "c" and "r" denote complementary nucleotide letters in the special case of WC pairings (e.g., $A^c = T$) and bond-reversals for the more general situation which includes also mismatched pairings (e.g. $A^r = G$ and $A^r = A$ for mismatches of the Ag and Aa groups, respectively).

Perfect 3′/5′-symmetry of the triple sensitivities (i.e. $Y(xBy) = Y(yBx)$) is expected if the base pairings are independent of their nearest neighbors. Stacking interactions between adjacent nucleotides however make an essential contribution to the stability of DNA/DNA-duplexes [32,33]. The change of stacking contributions after strand-reversal is governed by the different stereochemistry of 3′/5′ and 5′/3′ strand directions in the duplexes. The deviation from the perfect 3′/5′-symmetry relation thus estimates the effect of stacking interactions in the considered triplets.

In contrast, the complementarity relation keeps the strand direction unchanged. Perfect complementarity of the triple sensitivities (i.e. $Y(xBy) = Y(y^cB^rx^c)$) is expected if both interacting strands are physically equivalent and if their reactivity is not selectively perturbed by parasitic reactions such as intramolecular folding and/or bulk dimerization [13]. For example, duplexing experiments in solution typically use oligonucleotides of equal length and of low propensity for intramolecular folding and self-interactions. A very different situation occurs on microarrays because the reacting partners are highly asymmetric in length and conformational freedom: Firstly, the probes are attached to the chip surface whereas the targets are dissolved in the supernatant solution with consequences for their reactivity. For example, the interactions depend on the position of the nucleotide letter in the probe sequence owing to their attachment to the chip surface which gives rise to positional dependent constraints of probe/target interactions [9,13]. Secondly, the length of the targets exceeds that of the probes typically by more than one order of magnitude which markedly enhances their propensity for intramolecular folding and intermolecular duplexing reactions in solution in a sequence-dependent fashion with consequences for their effective interactions with the probes. Hence, deviations from perfect complementarity are expected to detect imbalanced probe/target interactions due to the asymmetric nature of the hybridization reaction on microarrays.

Figure 5 re-plots the triple sensitivities shown in Figure 4 in decreasing order for each group (see thick line in each panel) together with the values which are re-ordered according to the symmetry-relations Eq. (14) (see symbols). We calculate the scatter width of the symbols about the ranked xBy-triples in terms of their standard deviation which defines a sort of "asymmetry" funnel shown by dashed curves in Figure 5. The widths of the funnels (the respective standard deviations are given in Table 1) characterize the mean asymmetry of the triple interactions of the respective interaction group. Note that for perfect symmetries one expects vanishing funnel widths.

Both, 3′/5′- and complementary asymmetries roughly behave in parallel. They are, by far, smallest for the At-group and largest for the Aa-group which agrees with the ranking of the variability of the triple sensitivities between the groups. Also the SD values roughly agree (see Table 1) which indicates independence of triple sensitivities after symmetry transformation.

Hence, the effect of the central mismatch of the Aa-group is obviously most modulated by stacking interactions and complementary asymmetries among the considered groups causing largest variability of the associated probe intensities. Note that just this

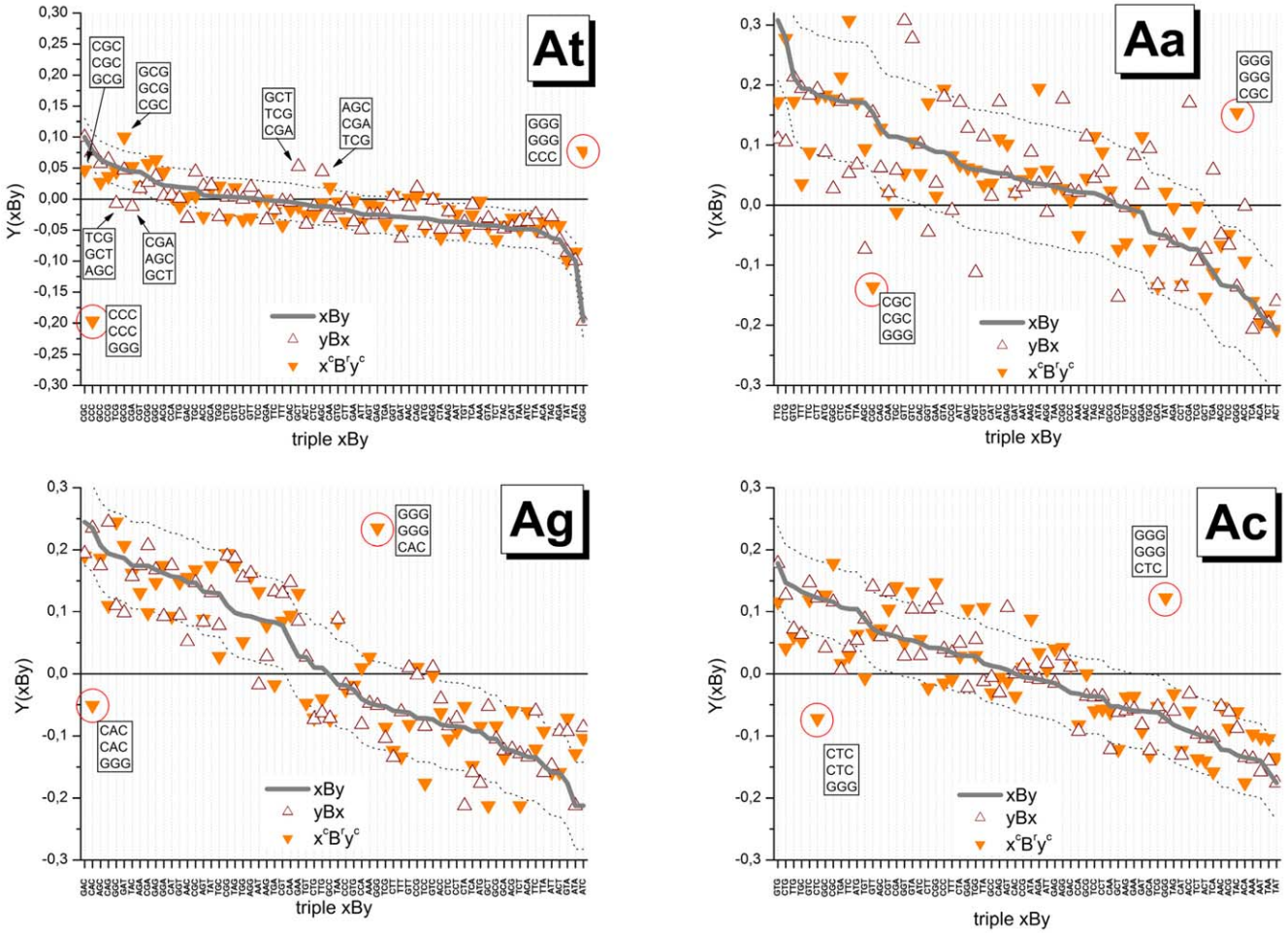


Figure 5. Symmetry relations of triple interactions. The triple sensitivities, Y(xBy), of each interaction groups are ranked in decreasing order and shown by thick lines. For each base-triple three sensitivity values are shown according to Eq. (14) to reveal 3′/5′-asymmetry, Y(yBx), and complementarity, $Y(y^cB^rx^c)$, respectively (symbols are assigned in the figure). The abscissa labels indicate the xBy-triple. The letter-triples in the boxes indicate special triples the sensitivity values of which reveal considerable asymmetry, for example $xBy/yBx/y^cB^rx^c$ = TCG/GCT/AGC of the At-group. Note that GGG-motifs are highly non-complementary in all four interaction groups. Note also the markedly different widths of the scattering funnels of the different interaction groups given by their standard deviation (see dotted lines and also Table 1) indicating that the stacking terms and/or asymmetry of interactions are differently modulated by the central mismatch (see text). For symmetry reasons part of the asymmetries differences vanish (e.g. 3′/5′-asymmetry of GGG/GGG/CAC).
doi:10.1371/journal.pone.0007862.g005

type of self-complementary mismatches was selected to design MM probes on microarrays of the GeneChip-type. Our results suggest that this design seems suboptimal because it is associated with a relatively high variability of mismatch stability. The effect introduces additional noise into the MM intensities which intend to correct the PM signals for background contributions.

Examples for symmetry relations are explicitly indicated in Figure 5 (the respective triples $xBy/yBx/y^cB^rx^c$ are given within the boxes, the abscissa labels indicate the xBy-triple only): For example, the combination AGC/CGA/TCG taken from the At-group shows marked 3′/5′-asymmety beyond the limits of the mean scattering funnel. The data clearly show that the by far largest complementary asymmetries are associated with triple-G motifs in the probe sequence for all interaction groups (see solid triangles surrounded by the circles). They make a contribution of up to 50% to the mean variability of the respective interaction groups (Table 1). Note in this context that the GGG-motifs are characterized by the weakest interactions either among all 64 triples (At-group) or among the 16 triples with a central G (Aa-, Ag- and Ac- groups, see Figure 4). This effect will be further discussed below.

Adjacent WC pairings

The context of adjacent WC pairs considerably modifies the effect of the central mismatch: For example, the ratio of two triple-sensitivities with a central Cc-mismatch (Aa-group) flanked either by two C's or by two A's is about $Y(CCC)/Y(ACA)|_{Aa} \approx 10^{+0.2} \sim 1.6$ whereas the respective intensity ratio for the triples with a central Cg-pair (At-group) is only $I(CCC)/I(ACA)|_{At} \approx 10^{+0.1} \sim 1.25$.

To generalize this result we average the triple sensitivities of each mismatch group over the central base, $Y^{ad}_{Ab}(xy) = \frac{1}{2}\langle Y_{Ab}(xBy) + Y_{Ab}(yBx)\rangle_{B=A,C,G,T}$. The obtained mean sensitivities characterize the effect of the WC pairings adjacent to the mismatched pairing. The values rank in good agreement with the expected mean stability of single-nucleotide canonical DNA/DNA interactions, C≈G>A≈T [32] (see

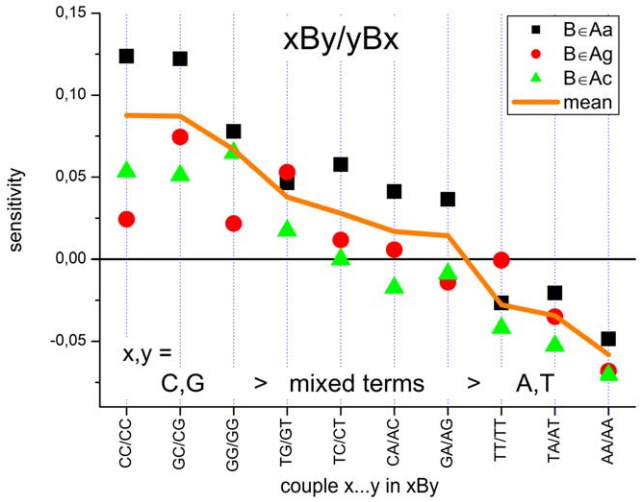


Figure 6. The effect of adjacent WC pairings in triples with a central mismatch. Mean sensitivity values were calculated as averages over triple sensitivities shown in Figure 4 for each Ab-group over the central mismatch. The obtained values characterize the mean effect of the couple xy in the triple xBy. They are ranked with decreasing mean of all three mismatch groups. It shows that x,y = C and G give rise to largest sensitivities and standard deviation about the mean whereas adjacent x,y = A and T cause smaller sensitivities and variability about the mean.
doi:10.1371/journal.pone.0007862.g006

Figure 6). Note also the small systematic trend between the Ab groups, Aa>Ag≈Ac, and the decreasing variability of the data with decreasing mean.

Tandem mismatches

Tandem mismatches occur in homozygous-absent duplexes of the MM-probes (MM-G'•G) with SNP offsets $\delta = +1$ and -1 (see Text S1). They consist of a mismatch of the Aa-group at the middle position of the probe sequence and a second mismatch of the Aa-, Ag- or Ac-group at the adjacent SNP position (see the sketch in Figure 7, panel a). The tandem mismatches are analyzed together with the adjacent WC pairs forming the quadruplets (yB'Bx) and (xBB'y) for $\delta = -1$ and +1, respectively (where x, y, B and B' denote the respective nucleotide bases in the probe sequence). According to this convention we ignore the strand direction: B defines the mismatch of the Aa-group and B' the mismatch of the Aa-, Ag or Ac-type and x and y form the edging WC pairings adjacent to B and B', respectively. The need for considering quadruplet-motifs (tandem mismatch and flanking WC pairs) to specify the stability of two adjacent mismatches was discussed previously [34].

We calculate the sensitivities of all possible combinations for each of the three possible options of B' (referring either to the Aa-, Ag- or Ac-group) using the background-corrected intensities relatively to the mean log-intensity of the probes with two mismatches (#mm = 2) with at least one WC pairing in-between, $Y_{Ab}(xBB'y) = \log(I_{Ab,\#mm=2,|\delta|=1}(xBB'y)) - \langle \log(I) \rangle_{\#mm=2,|\delta|>1}$ (see also part a of Figure 2).

The average values of the obtained sensitivities of the tandem mismatches are positive (see the horizontal dashed lines in part a and b of Figure 7) which reflects their larger stability compared with the double mismatches which are separated by at least one WC pair.

The 16^2 possible quadruplet combinations were reduced to 2×16 values for each of the three possible pairings of B' by calculating the average either over the edging WC pairings xy or over the mismatches BB', $\langle Y_{Ab}(xBB'y) \rangle_{xy}$ and $\langle Y_{Ab}(xBB'y) \rangle_{BB'}$, respectively. We consider all 16 combinations of xy and BB' in xBB'y because both members of each couple are not equivalent (B'∈Aa, Ag, Ac and B∈Aa). The obtained values thus characterize the effect of the edging base couples xy (part a of Figure 7) and of the mismatch couples BB' (part b) on the corresponding probe sensitivities, respectively. In addition we decompose the quadruplets in two consecutive NN-contributions according to xBB'y→xB+B'y/yB'Bx→yB'+Bx by calculating the averages $\frac{1}{2}\langle Y_{Aa}(xB) + Y_{Aa}(Bx) \rangle_{B'y}$ and $\frac{1}{2}\langle Y_{Ab}(B'y) + Y_{Ab}(yB') \rangle_{xB}$, respectively (see part c and d of Figure 7), which characterize mixed combinations of WC- and mismatched pairings in accordance with the NN-decomposition of the standard triples applied in the next section.

The couples of edging bases x and y cause considerable smaller variability of the probe sensitivities than the couples of adjacent mismatches (compare part a and b of Figure 7). The standard deviations of the latter group exceeds that of the former group roughly by the factor of two (see Table 1). This ratio actually increases to about three if one calculates the scattering about the mean of the three Ab-groups (i.e. the scattering about the decaying line in the figure). Hence, the particular couple of mismatches BB' mainly modulates the intensities of the probes whereas the edging WC pairings give rise to only moderate intensity variations. This result agrees with the properties of triples with a central mismatch discussed above. The main source of probe intensity variation was also attributed to the central mismatch in this case.

Part a of Figure 7 shows that the adjacent WC pairs rank according to x,y = C,G>x = G,C; y = A,T>x = A,T; y = G,C>x,y = A,T and thus in the similar order as the adjacent WC pairs of single mismatches (see previous section and Figure 6). Both sets of mean sensitivities (thick lines in Figure 6 and Figure 7, panel a) correlate with a regression coefficient of R = 0.57.

Part b of Figure 7 indicates that the particular sensitivity value strongly depends on the combination of mismatches. For example, the combination BB' = CT of, on the average, relatively weak stability varies between large and very small sensitivities for C∈Ac and C∈Ag, respectively.

Alternatively, we decomposed the quadruplets with the central tandem mismatch into two consecutive NN-terms as described above (Figure 7, panel c and d). These NN-terms can be compared with NN-terms which were obtained after decomposition of the triple sensitivities into two NN-terms as described in the next section (compare with thick blue lines and open symbols in Figure 7, panel c and d). Both data sets correlate with regression coefficient R = 0.69. This result suggests that quadruplets with central tandem mismatches can be decomposed to a rough approximation into two NN-terms which can be estimated also from triple data.

Flanking mismatches

Triples with flanking mismatches of the type w(xBy)m (B∈At; "w" and "m" denote a WC- and a mismatched pairing, respectively, i.e. w∈At and m∈Aa,Ag,Ac) were selected according to the scheme shown in Text S1. These triples refer to SNP offset positions $|\delta| = 2$. To assess the effect of the flanking mismatch "m" we compare the log-intensities of the respective probes with the respective values of the neighboring standard triples w(xBy)w without flanking mismatch (offset $|\delta| = 3$),

$$\Delta^{\text{flank}}(xBy) = \langle \log I(xBy) \rangle_{|\delta|=3} - \langle \log I(xBy) \rangle_{|\delta|=2}. \quad (15)$$

This difference estimates the mean intensity increment of the

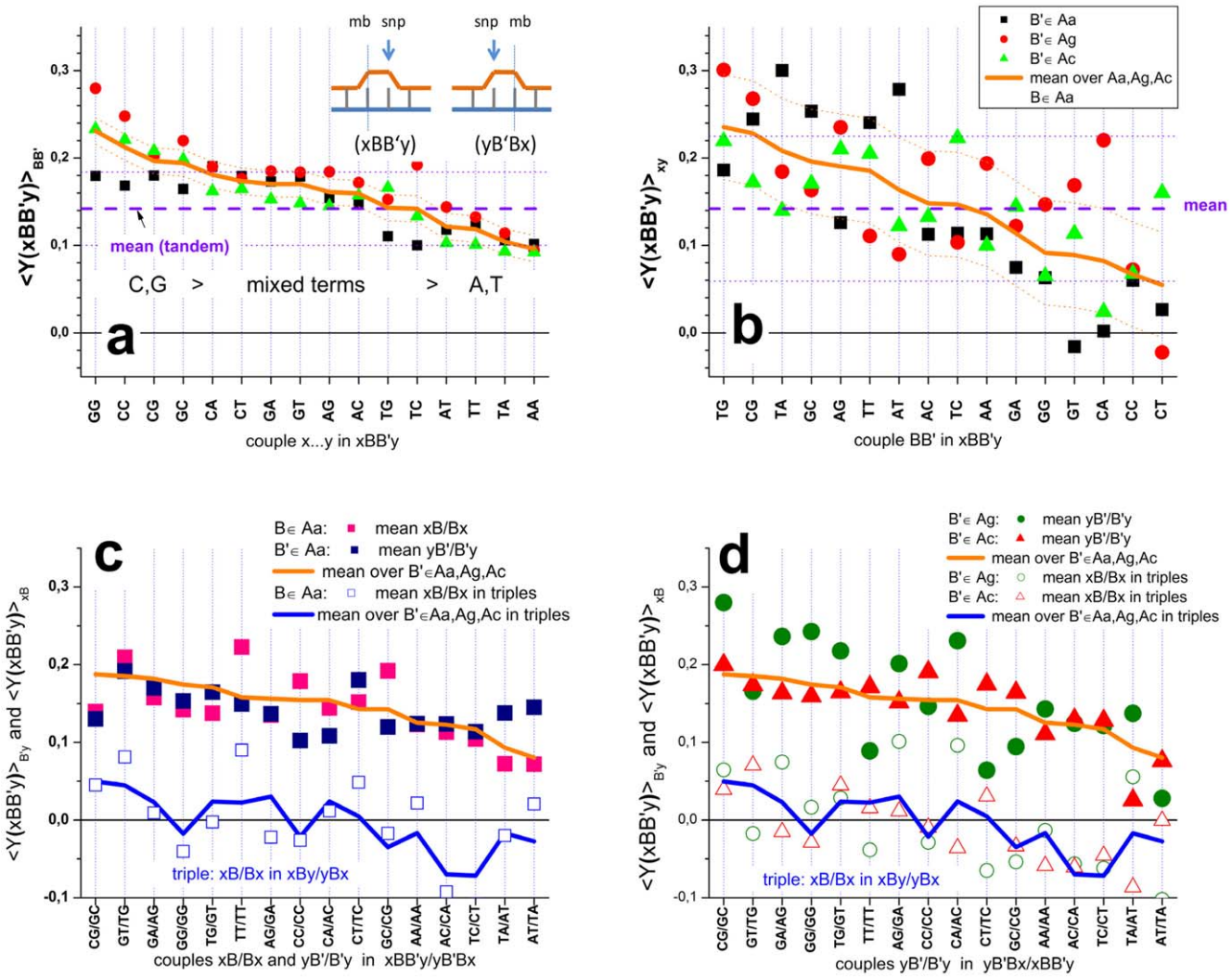


Figure 7. The sensitivities of quadruplets (xBB'y) composed of central tandem mismatches BB′ and edging WC pairings, x,y. The quadruplets were analyzed in terms of independent duplets of the WC-couples xy (part a), of tandem mismatches BB′ (part b) and of mixed NN-couples xB/Bx and yB′/B'y (part c and d). Note that B refers to the Aa-group whereas B′ to the Aa-, Ag- or Ac-group (see legends in the figure). Along the x-axis the respective pairings are ordered with decreasing mean sensitivity which is averaged over the three groups Aa, Ag and Ac of B′ (see the thick decaying curve). Part a and b: The central tandem mismatches formed by B and B′ cause considerably larger scattering than the adjacent WC pairings formed by x and y. The thin dotted curves running parallel to the thick line illustrate the standard deviation of the dots about their mean (see also Table 1). In part c and d the respective NN-terms derived from the triple motifs with single mismatches (see Eq. (16) and Figure 10 below) are shown for comparison (the open symbols show the NN-terms of the respective interaction groups and the thick blue line their mean value).
doi:10.1371/journal.pone.0007862.g007

standard triple without flanking mismatches relative to that with flanking mismatches. Our nomenclature assigns nucleotide 'y' to the position adjacent to the mismatch which flanks the triple, (xBy)m. This neighborhood-relation can be realized for the triples (xBy)m and m(yBx), i.e. with the mismatch facing towards the 3′ or the 5′ end of the probe, respectively; and, in addition, in the probe and target sequence according to the complementary condition $m(yBx) \rightarrow (x^cB^cy^c)m$ (the superscript "c" denotes the WC-complement). These, in total four options (for example (CGT)m, m(TGC), (GCA)m, m(ACG)) are averaged to provide the mean effect of the flanking mismatch adjacent to 'y' and 'y^c' on the selected triple.

Figure 8 shows that the obtained mean excess values are consistently negative for y = C,G and positive for y = A,T. Hence, a mismatched pairing either stabilizes or destabilizes the adjacent triple in dependence on the neighboring base y. The effect is however relatively weak and amounts to a few percent of the respective probe intensity.

Nearest neighbor terms

In analogy with the NN free energy contributions in models describing the stability of DNA/DNA-oligonucleotide duplexes in solution (see [32,33] and references cited therein) we decompose each triple-averaged sensitivity of each interaction group, $Y_{Ab}(xBy)$, into two nearest neighbor (NN) terms, $Y_{Ab}(x\underline{B})$ and $Y_{Ab}(\underline{B}y)$, and two single-base boundary contributions according to

$$Y_{Ab}(xBy) = Y_{Ab}(x\underline{B}) + Y_{Ab}(\underline{B}y) + \frac{1}{2}(Y_{Ab}(x) + Y_{Ab}(y)) \quad (16)$$

using Single Value Decomposition (SVD) [35]. The underlined letter denotes the central base of the respective triple in the

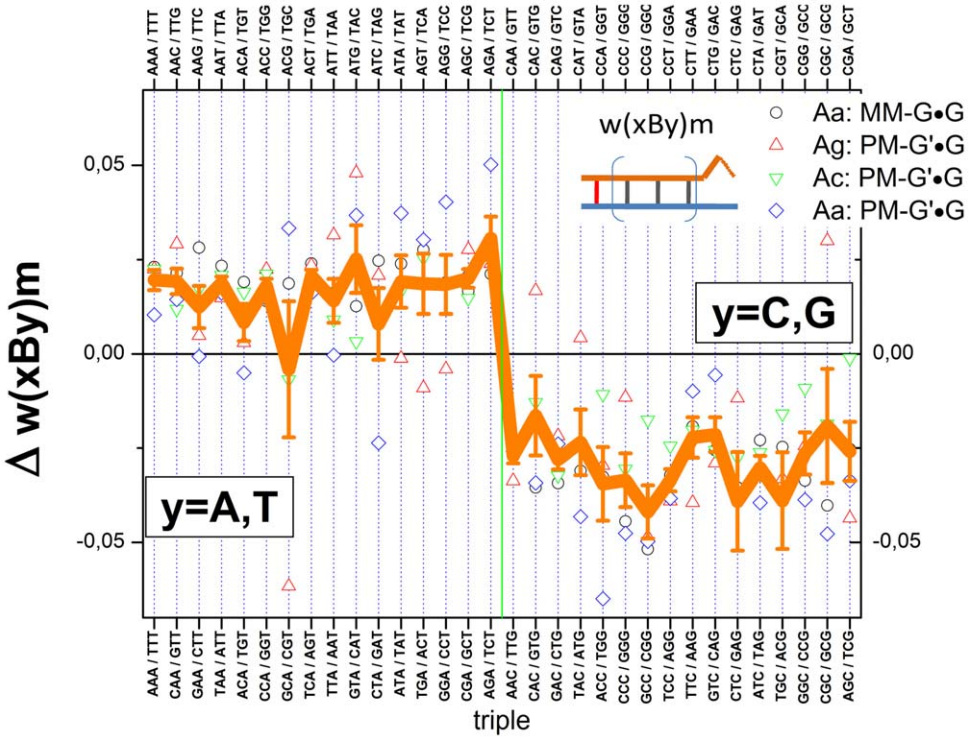


Figure 8. Excess sensitivities of triples with flanking mismatches (Eq. (15)). The respective probes with flanking triples are selected according to Text S1. Neglecting 3′/5′- and probe/target-asymmetries, each value is calculated as mean value over the four triples indicated at the lower and upper x-axes for each mismatch group (symbols; see legend for assignments). The combination of triples shown at the lower axis denote the complements w(xBy)m/w($x^cB^cy^c$)m and that at the upper axis m(yBx)w/m($y^cB^cx^c$)y. The thick line refers to the total mean over all three mismatch groups m∈Aa,Ag,Ac. The excess values are consistently positive and negative for adjacent y = A,T and y = C,G, respectively.
doi:10.1371/journal.pone.0007862.g008

argument of the NN-terms to avoid confusion in symmetry relations discussed below. The single-base boundary terms consider the mean effect of the bases adjacent to the triple. The triple data of each interaction group thus define a system of 64 linear equations which was solved by multiple linear regression to determine in total 8 boundary and 32 NN terms (see also [20]).

We first examined the adequacy of the decomposition (Eq. (16)) in terms of the residual contribution

$$\Delta_{\mathrm{Ab}}^{\mathrm{res}}(\mathrm{xBy}) = \mathrm{Y}_{\mathrm{Ab}}(\mathrm{xBy}) - \left(\mathrm{Y}_{\mathrm{Ab}}(\mathrm{x}\underline{\mathrm{B}}) + \mathrm{Y}_{\mathrm{Ab}}(\underline{\mathrm{B}}\mathrm{y}) + \frac{1}{2}(\mathrm{Y}_{\mathrm{Ab}}(\mathrm{x}) + \mathrm{Y}_{\mathrm{Ab}}(\mathrm{y}))\right), \quad (17)$$

which estimates the degree of additivity of the triple NNN-model, i.e., the reliability of decomposition of the triples into nearest neighbor NN-terms. In the absence of interactions affecting next nearest neighbors, one expects vanishing residuals, $\Delta_{\mathrm{Ab}}{}^{\mathrm{res}}(\mathrm{xBy}) = 0$. Especially the propensity of selected sequence motifs for intramolecular folding of the probes and/or the targets and also for the formation of special intermolecular complexes are expected to involve longer runs of subsequent nucleotides causing deviations from the additivity assumption (Eq. (16)).

Figure 9 shows the residuals of all 64 triples per interaction group obtained after decomposition of the NNN-terms into nearest neighbor contributions. The standard deviation of each group is considerably smaller compared to that obtained from the asymmetry relations (see Table 1). This result indicates that most of the triples are additive with respect to NN-terms to a good approximation.

However, motifs containing couples of adjacent GG are prone to positive deviations from additivity indicating that the respective

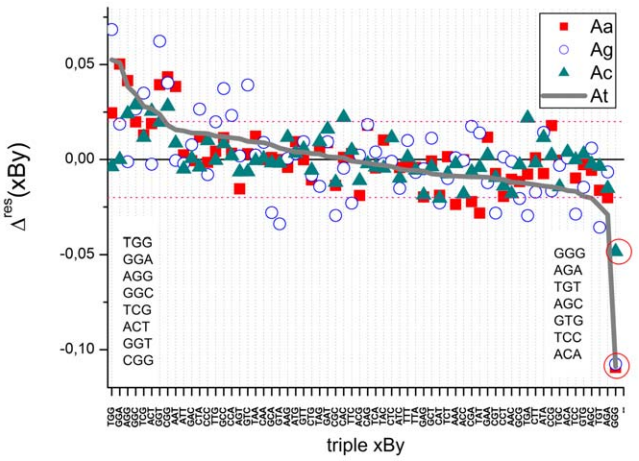


Figure 9. Residual sensitivity after decomposition of the triple sensitivities into NN-terms (Eq. (17)). The symbols refer to the mismatched interaction groups. The triples are ranked with decreasing residual contributions of the At-group. The horizontal dashed lines mark the average standard deviation of the data about the abscissa. The two NNN-lists indicate the largest positive (left list) and negative (right list) residual-values of the At-group. Note that triple GGG provides by far the largest (negative) residual contribution (see red circles). Positive contributions are obtained for triples containing the couple 'GG' which indicates that the respective NN-terms underestimate their contribution to the triple sensitivities.
doi:10.1371/journal.pone.0007862.g009

GG-term systematically underestimates the contribution of two adjacent guanines to the triple term. On the other hand, runs of three guanines, 'GGG', give rise to the strongest negative residual terms of all interaction groups. The triple sensitivities $Y_{Ab}(GGG)$ are negative for all interaction groups (see Figure 4). The observed residuals thus again indicate that the respective sum of two GG-terms underestimates their contribution to the absolute value of the triple sensitivity, i.e. $2\,|Y(GG)| < |Y(GGG)|$. Hence, non-additivity of the considered triples is mainly introduced by GG-couples, the NN-terms of which underestimate their contribution to triple terms containing adjacent GG.

Figure 10 separately shows the obtained NN-terms for each interaction group and for each central base pairing of the respective triples. The NN-terms are combined according to the convention $x\underline{B}/\underline{B}x$ (left/right bar) which estimates the 3′/5′ asymmetry with respect to the common base B forming the mismatched pairing in the Aa-, Ag- and Ac-groups. Comparison of the respective left and right bars essentially confirms the 3′/5′-asymmetry data of the triple sensitivities discussed above, namely that the Aa- and At-groups show the largest and smallest asymmetries, respectively. The NN-data in addition reveal that most of the highly asymmetric base couples of the Aa-group (e.g., $A\underline{C}/\underline{C}A$, $C\underline{C}/\underline{C}C$, $A\underline{G}/\underline{G}A$, $C\underline{G}/\underline{G}C$) are associated with guanines and cytosines at the mismatch position.

Comparison with free energy terms describing duplexing in solution

The 32 NN-couples of At-groups can be further reduced to 16 NN-terms making use of the symmetry-relation $Y_{At}(X\underline{Y}) \approx Y_{At}(\underline{X}Y)$ which however only applies to the At-group due to the equivalence of the two WC pairings associated with the nucleotide letters. Part a of Figure 11 correlates the obtained 16 averaged terms, $Y_{At}(XY) = 0.5 \cdot (Y_{At}(X\underline{Y}) + Y_{At}(\underline{X}Y))$, with the ten NN-free energy terms estimated in solution studies [33]. The data well correlate with a regression coefficient of $R = 0.85$ if one ignores the GG-couple (see regression line in Figure 11). Its sensitivity value distinctly deviates in negative direction in agreement with the qualitative discussion of the residual contributions given above (see Figure 9). The relatively large difference $Y_{At}(CC) - Y_{At}(GG) > 0.06$ indicates that the complementarity between CC and GG is clearly disrupted. On the other hand, the sensitivity values of the remaining complementary couples (XY/Y^cX^c = AA/TT, CT/AG, TC/GA,

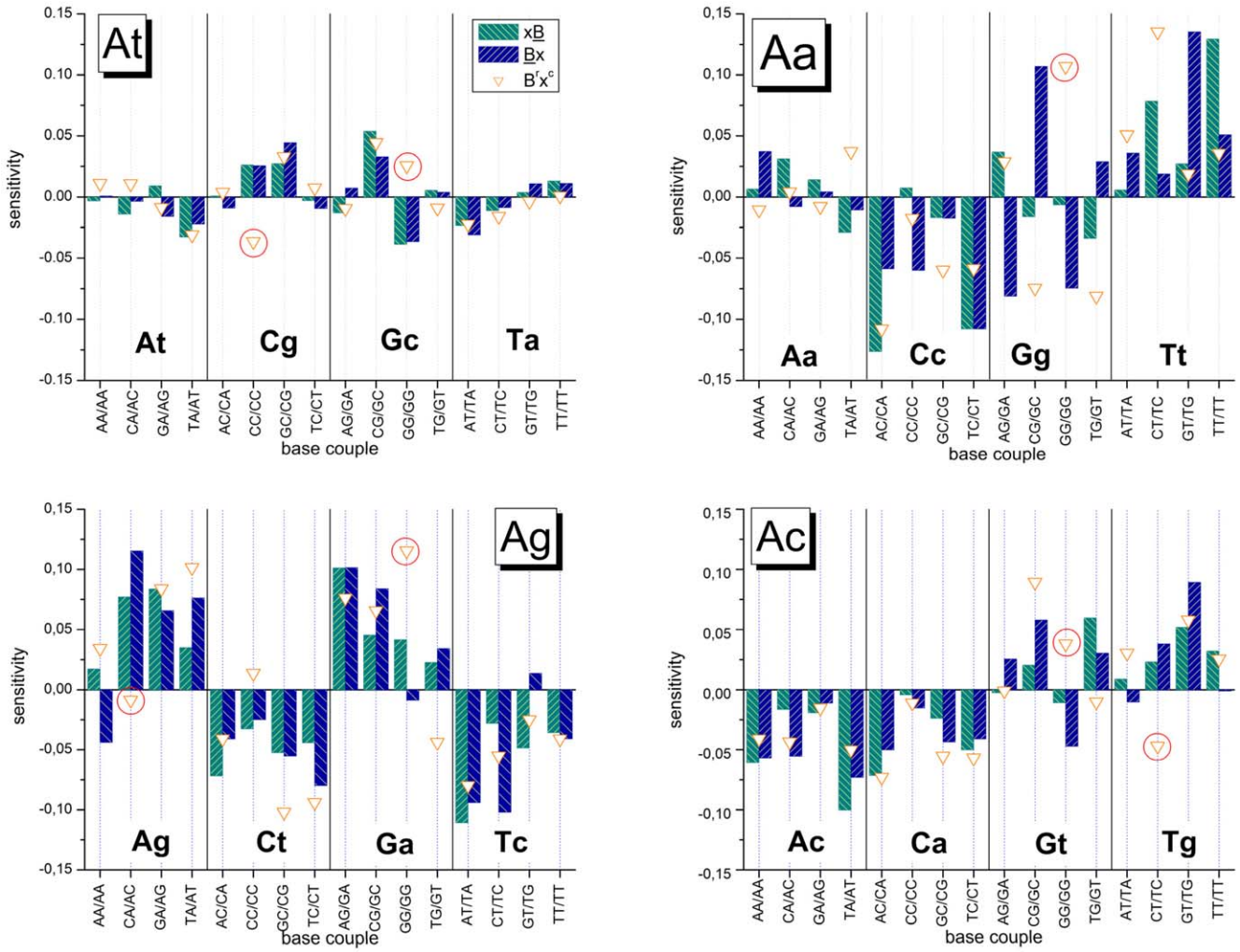


Figure 10. Nearest neighbor (NN) sensitivity terms of the four interaction groups. The NN-terms are calculated via decomposition of the triple terms using SVD (Eq. (16)) where the base couples are ordered with respect to the centre base B of the triples. The base couples are indicated as abscissa labels $x\underline{B}/\underline{B}x$ (left/right bar, respectively). The symbols are the sensitivities after applying the complementary transformation to the NN-terms, $xB \rightarrow B^rx^c$. NN-terms related to 'GG'-motifs are indicated by red circles. They strongly deviate from the complementary condition.
doi:10.1371/journal.pone.0007862.g010

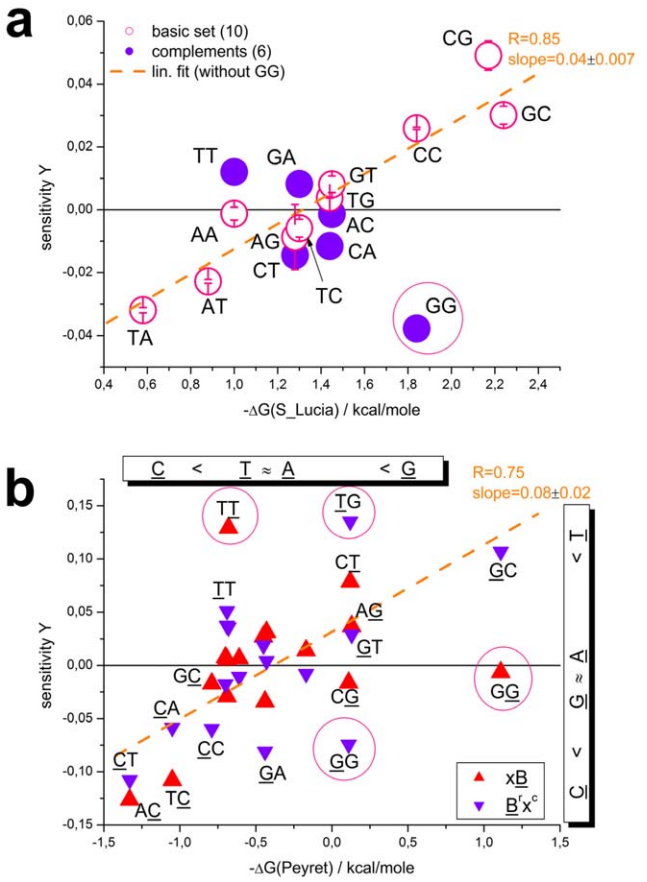


Figure 11. Comparison with solution data. The figure shows the sensitivity NN-terms of the At- (part a) and Aa- (part b) groups obtained in this study (Eq. (16)) with NN-stacking free energy terms for DNA/DNA-duplexes in solution taken from ref. [33] and [31], respectively. The dashed diagonal lines are linear regressions using all NN-data except the double-guanine terms (At-group) and in addition except TT and TG (Aa-group) which are included in red circles (regression coefficients and slopes are given in the figure). Panel a: Each NN-sensitivity of couple XY was calculated as the mean value averaged over the two sensitivities with arguments $\underline{XY}$ and $\underline{XY}$ shown in Figure 10. The difference between these paired values is shown by the error bars which typically do not exceed the size of the symbol. The basic set of 10 independent terms is indicated by open circles. Panel b: The complementary couples $x\underline{B}$ and $\underline{B^r}x^c$ are shown by different triangles. Only selected NN-motifs are assigned. The apparent mean stabilities of the mismatched pairings rank differently for chip (see vertical bar) and solution (horizontal bar) data.
doi:10.1371/journal.pone.0007862.g011

AC/GT and CA/TG; see full and open symbols) are relatively close each to another (mean difference $|Y(XY)-Y(Y^cX^c)|\approx 0.01$) which justifies utilization of the complementarity condition to a good approximation. The linear regression coefficient slightly improves ($R = 0.92$) after averaging over the complementary couples. Hence, except GG-motifs, the interactions of canonical WC pairings estimated from the probe intensities of SNP GeneChip microarrays in acceptable agreement correlate on a relative scale with free energies in solution.

Part b of Figure 11 shows an analogous correlation plot for the NN-terms of the Aa-group where the solution free energies were taken from ref. [31]. The 32 NN-sensitivity terms split into 16 basic terms $Y_{Aa}(x\underline{B})$ (open symbols) and 16 complementary terms $Y_{Aa}(\underline{B^r}x^c)$ (solid symbols). As for the At-group, the double-guanine terms strongly deviate from the regression line and were excluded from the linear fit ($R = 0.65$). Additional exclusion of double-thymines further increases the regression coefficient ($R = 0.75$) which indicates satisfactory correlation between solution free energy data and most of the NN-sensitivities. A recent study also reports clear correlation between solution and array estimates of hybridization free energies using a specially designed Agilent microarray containing sets of PM and MM probes with #mm = 1 and 2 mismatches upon duplexing [36].

Note that the mean stability of self-complementary mismatches rank according to CC<TT≈AA<GG in solution but according to Cc<Gg≈Aa<Tt on the chip (see Figure 7). Hence, Gg-pairings apparently loose and Tt-pairings gain stability on the chip. The stability-ranking of the other mismatches except Gt essentially agrees for solution and chip data (see above).

Discussion

In this study we analyzed the probe intensities taken from a 100k GeneChip SNP array in terms of selected sequence motifs forming well defined WC- and mismatched base pairing in the probe/target duplexes. The particular probe design of these GeneChip SNP arrays enables one to disentangle different sources of intensity modulations such as the number of mismatches per duplex, the particular matched or mismatched base pairings, their nearest and next-nearest neighbors, their position along the probe sequence and the relative position of a second mismatch. As the elementary sequence motif we chose triples of subsequent nucleotides centered about the middle base of the probe and/or about the SNP base and calculate log-averages of the intensities over thousands of probes with identical motifs to average out the effect of the remaining sequence. These averages are measures of the stability of the base pairings formed by the selected triple in the probe sequence with the corresponding base triple in the target sequence. The former triple is defined by the probe sequence whereas the target triple can be deduced from the genotype and the hybridization mode. We analyzed the log-averaged intensities, their difference to selected reference values, the so-called sensitivity, and their variability in subsets of triple-motifs. In addition to triple motifs, we also consider special motifs such as flanking mismatches adjacent to the triples and tandem mismatches which were analyzed in terms of quadruplets including the edging WC pairings.

The first question of our analyses addresses the impact of different interaction motifs on the observed probe intensities. It turns out that

a) the number of mismatches per probe/target-duplexes exerts the largest effect which modulates the intensity. One mismatch is associated with the logarithmic intensity change of $-\delta\log I = 0.5$–0.6 which is equivalent with the decrease of the intensity by a reduction factor of about $F = 0.3-0.25$ per mismatch.

b) the effect of mismatches is strongly modulated by the adjacent WC pairings which give rise to a mean logarithmic increment of $\sim\delta\log I = \pm 0.1$, or equivalently, with an average modulation factor of $0.8<F<1.25$ (see Table 1). Selected motifs cause larger log-increments of $\delta\log I = \pm 0.3$ (see Figure 4) which are almost comparable in magnitude with the mean mismatch effect (see a).

c) duplexes with tandem mismatches are more stable than double mismatches which are separated by at least one WC pairing ($\delta\log I \approx +0.1$ and $F \approx 1.25$).

d) flanking mismatches adjacent to the considered triples only weakly modulate their intensities ($|\delta\log I|<0.025$; $0.95<F<1.05$).

e) the mean variability due to sequence effects in triples of WC pairings is markedly smaller than the effect in triples with a central mismatch ($\delta\log I = \pm 0.05$; $0.9<F<1.1$; compare with b).

f) runs of three guanines in the probe sequence forming nominally WC pairings represent a special motif which decreases the intensity to an exceptionally strong extent ($\delta\log I = -0.2 - -0.35$; $F = 0.6 - 0.45$). Also mismatched duplexes with runs of guanines possess relative small intensity values which are virtually incompatible with expected interaction symmetries in DNA/DNA-duplexes.

g) the positional dependence of triple-averaged intensities along the probe sequence is relatively weak (see Figure 3 part a, c and d). The sequence-specific effect progressively disappears towards the ends of the probe sequence at the final 3–5 sequence-positions for most of the motifs. Triple 'GGG'-motifs partly deviate from this rule: Along the whole sequence they markedly reduce the intensity. In mismatched duplexes one observes the opposite effect at the probe end facing towards the supernatant solution.

h) especially small (e.g., for probes with two mismatches, $\#mm = 2$) and large intensity values are prone to background and saturation effects, respectively (see Text S1). Appropriate background corrections should consider the optical background and partly also non-specific hybridization. Saturation can be considered using the hyperbolic adsorption law (see supporting file Text S1).

Our analyses also address the question whether the number of considered sequence motifs can be reduced by utilizing symmetry relations and/or by decomposing the triple averages into nearest neighbor terms in analogy with interaction models for oligonucleotide duplexes in solution. It turned out that

i) triples of WC pairings (At-group) can be reasonably well decomposed into NN-terms which also meet the complementary condition to a good approximation and correlate well ($R = 0.85$) with the independent NN-free energy terms derived from duplex-data in solution [32,33]. GGG-motifs strongly deviate from these properties and must be considered separately.

j) also the triples with a central mismatch (Aa-, Ag- and Ac-group) to a good approximation decompose into NN-terms except special motifs containing at least doublets of guanines. The mismatch motifs partly obey the symmetry relations, however, with larger residual variability compared with WC pairings. Comparison with NN-terms of solution free energies [31] indicates satisfactory correlation for most of the motifs ($R = 0.75$). Runs of guanines and partly also thymine-containing motifs deviate from the expected behavior in negative and positive direction, respectively.

k) tandem mismatches can be decomposed into two NN-terms referring to a combination of mismatched and WC pairings. These values well correlate ($R = 0.59$) with the NN-terms obtained from the triple data suggesting to use a unified set of NN-terms (see j). For tandem mismatches one has however to consider their systematically larger stability compared with duplexes containing two mismatches which are separated by at least one WC pairing.

In the following subsections we discuss the physical origin of selected effects more in detail and derive rules for appropriate correction of parasitic intensity errors to obtain unbiased genotyping estimates.

Relation to thermodynamics

The intensity of microarray probes is directly related to the effective association constant for duplexing, $\sim K_{duplex}$ after correction for parasitic effects (or their neglect, if justified) such as the optical background, non-specific hybridization and saturation (see Eq. (2)). The effective association constant is a function of different reaction constants characterizing relevant molecular processes such as the bimolecular stacking of unfolded probes and targets (P•T, P•P, T•T), and their unimolecular folding propensities (P-fold, T-fold) [13] (see also [37]), i.e.

$$K_{duplex} \approx K^{P\bullet T} \cdot F_{array} \quad \text{with}$$
$$F_{array} = F_{surface} \cdot \left\{ (1 + K^{T-fold} + \sqrt{K^{T\bullet T}[T]}) \cdot (1 + K^{P-fold} + \sqrt{K^{P\bullet P}[P]}) \right\}^{-1}, \quad (18)$$

where $F_{surface} < 1$ is a factor taking into account surface effects, such as electrostatic and entropic repulsions which effectively reduce target concentrations near the array surface. According to Eq. (18), the effective constant of duplex formation is reduced by the factor $F_{array} < 1$ compared with the stacking interaction constant $K^{P\bullet T}$. Folding and/or self-dimerization of probe and/or target become relevant at $1 < \left(K^{P-fold} + \sqrt{K^{P\bullet P}[P]} \right)$ for the probe (substitute P→T for the target).

Stacking interactions are mainly governed by the pairings formed between the nucleotides in the target and probe and their nearest-neighbors along the sequence. The decomposition of the corrected intensity into different interaction modes associated with single target-types enables assignment of the probe sequence to canonical and mismatched base pairings with the target. We analyzed triple motifs which represent a reasonable choice to study stacking interactions on an elementary level. Note that also the reduction factor F_{array} depends on the probe and target sequences, however in a more subtle fashion because, for example, folding reactions comprise longer sequence motifs.

The duplex-association constants can be multiplicatively decomposed into a triple-related factor which modulates the total (average) contributions

$$K_{duplex} \approx k_{duplex}(xBy) \cdot K_{duplex}(\#mm) \quad \text{with}$$
$$k_{duplex}(xBy) = k^{P\bullet T}(xBy) \cdot f_{array}(xBy) \quad \text{and} \quad , \quad (19)$$
$$\log K_{duplex}(\#mm) = \langle \log K^{P\bullet T} + \log F_{array} \rangle_{Ab,\delta,xBy}$$

where we use the notations introduced above. The triple related terms are denoted by lower case letters. The overall mean of the association constant mainly depends on the number of mismatches in the duplex, #mm. The modulation factor and the mean value are decomposed into stacking and array terms using Eq. (18). Hence, the effective duplex association constant decomposes into a series of nested factors which consider triple motifs, stacking interactions and array specifics in different combinations.

Comparison with Eq. (8) and considering the direct relation between the corrected intensity and K_{duplex} provides the relation between the analyzed observables and the binding constants,

$$Y(xBy) = \log k_{duplex}(xBy) = \log k^{P\bullet T}(xBy) + \log f_{array}(xBy)$$
$$\log I(\#mm) \approx \log K_{duplex}(\#mm) + \text{const.} \quad (20)$$

The logarithm of the association constant defines the stacking free energy of the duplex, $\Delta G^{P\bullet T} \sim -\log K^{P\bullet T}$, which applies also to the triple terms, i.e., $\Delta\Delta G^{P\bullet T}(xBy) = \Delta G^{P\bullet T}(xBy) - \langle \Delta G^{P\bullet T} \rangle \sim -\log k^{P\bullet T}$. With this definition and Eq. (20) one finds

$$Y(xBy) \propto -\left(\Delta\Delta G^{P\bullet T}(xBy) + \log f_{array}(xBy) \right)$$
$$\log I(\#mm) \propto -\left(\left\langle \Delta G^{P\bullet T}(\#mm) \right\rangle + \left\langle \log F_{array} \right\rangle \right) + \text{const.} \quad (21)$$

Hence, the triple-averaged sensitivities are related to the deviation of the stacking free energy due to the considered triple from its mean

value. This increment is however distorted by an "array"-term caused by folding, self-duplexing of target and probe and by specific surface effects. The former contributions are also functions of the sequence position of the chosen triple which is not explicitly expressed in Eq. (21) for sake of convenience. Note also that imperfect probe synthesis potentially reduces the real length of the oligomers in a motif-specific fashion with possible consequences for the observed triple sensitivities [13].

The sensitivity and free energy change into opposite directions, i.e. larger stability of interactions is associated with larger Y but smaller (more negative) ΔG. After decomposition into NN-terms we found acceptable correlation between the estimates from chip data and solution data taken from the literature for most of the motifs (see Figure 11). We conclude that chip effects are of inferior importance on the average (i.e. $\Delta\Delta G^{P\bullet T}(xBy) >> \log f_{array}(xBy)$). Stacking free energies therefore well reproduce the relation between the particular terms on a relative scale.

The proportionality constant in Eq. (21) is estimated by the slope of the regression lines in Figure 11. Their values are with $(0.4–0.8)\cdot 10^{-1}$ roughly one order of magnitude smaller than the proportionality constant predicted by the thermal energy $\sim 1/(RT\cdot\ln 10)\approx 0.7$ ($T\approx 40^{\circ}C$). We previously argued that non-linear (in logarithmic scale, as, e.g., predicted by Eq. (18)) and sequence dependent contributions to $\log(f_{array}(xBy))$ can cause proportionality constants less than unity [13]. Sequence-independent sources of intensity variability such as the length-dependent yield of the genomic targets after PCR-amplification [9,38] not-considered here are potential causes of the downscaling of the proportionality constant. Interestingly, the proportionality constant obtained for the mismatched pairings (Aa-group) exceeds that for the WC pairings (At-group) by the factor of two (compare part a and b of Figure 11). This difference suggests that the larger sensitivity-response of the probes to mismatched pairings (compared with WC pairings) is not simply related to the variability of the respective stacking free energies but includes other effects related to the array technology.

Mismatches

The stabilities of most of the mismatched pairings (Eq. (12)) rank in similar order as the results of previous chip and solution studies (Eq. (13)). Figure 12 shows the detailed stability trend in all 10 possible contexts of complementary triples with all 16 possible pairings of BB' (accordingly, the couples BB' refer to the pairings $B\bullet b^{r}$ and $B'\bullet b'^{r}$ with $b^{r} = B'$ and $b'^{r} = B$, respectively). Our figure was designed similar to Figure 3 in ref. [31] which ranks the central bases according to its mismatch stability in solution (Eq. (13)). Essentially two groups of larger and weaker stabilities can be clearly distinguished for BB': (TT,GA,GA;GT,TG,AA,GG)> (CT,TC,CA,AC,CC), respectively (see also the detailed ranking in Eq. (12)). Hence, mismatched pairings formed by cytosines are consistently of weaker stability. Most of the triples are modulated by the nearest neighbors of the central base (x…y) which follows the mean trend shown in Figure 6 (i.e., (x…y) = G,C>A,T). As an exception, adjacent WC pairings however only weakly affect the triples with the central mismatches BB' = TT and GA.

The stability of mismatched pairings is governed by the propensity of the paired nucleotides to form hydrogen bonds (e.g., two bonds (T, A) versus three bonds (G, C) in canonical WC pairings), by steric factors such as the size of the aromatic moiety (one ring of the pyrimidines (C,T) versus two rings of the purines (G, A)) as well as stacking effects associated with nearest neighbors.

Stable mismatched base pairs such as GT or GA form two H-bonds and only slightly disrupt the structure of the oligonucleotide-DNA duplex. In particular, the former purine/pyrimidine mismatch GT is usually slightly more stable than the latter purine/purine mismatch GA because a two-ringed guanine better fits with a single-ringed thymine than with a double ringed adenine [30]. On the other hand, unstable mismatched base pairs such as CT or CA significantly disrupt the duplex structure due to the small size of the pyrimidine/pyrimidine pairing or the disability to form at minimum two H-bonds because of the lack of imino protons [30]. Also the self complementary single ringed CC mismatch has a low stacking propensity and forms only one H-bond. This rationalizes the low stability of the mismatches formed by cytosines in agreement with our chip data.

The second self complementary single ringed TT mismatch with low stacking propensity is, in contrast to CC, however stabilized by two H-bonds. The two purine/purine self complementary mismatches GG and AA have a relatively high stacking

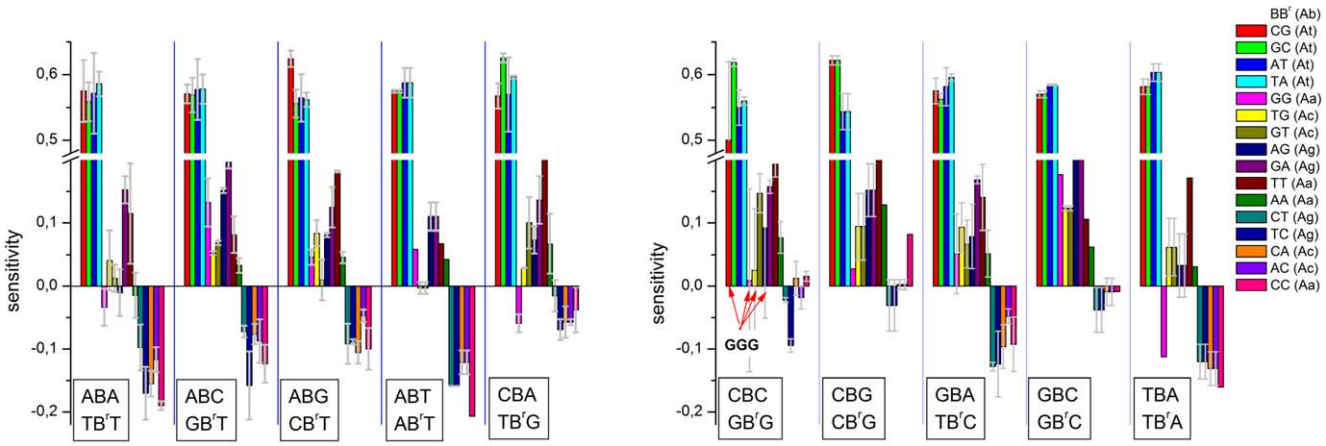


Figure 12. Stability of mismatch motifs. Relative stabilities of the 10 possible contexts of complementary triples containing the 16 possible central base pairings (mismatches or Watson-Crick base pairs, see legend in the figure). The sensitivities of the pairs of complementary triples xBy/$y^{c}B^{r}x^{c}$ ($B^{r} = B'$) are averaged using the triple data shown in Figure 4. The error bars indicate the difference between the individual values and thus they quantify the deviation from complementary symmetry. The form of the bar diagram was chosen in correspondence with Figure 3 in ref. [31] which ranks the stacking free energies of each triple in solution-duplexes with decreasing stability (from left to right for each triple). The mean log-intensity increment of one mismatched pairing (see Figure 2) was added to the triple-values of the At-group to compare the stabilities of WC- and mismatched pairings in a unique scale. The sensitivities of the four triple-combinations in the GGG-context are exceptionally small (see the red arrows).
doi:10.1371/journal.pone.0007862.g012

potential and form either two (GG) or only one (AA) H-bond. One expects therefore the stability-series AA≈TT<GG which is confirmed in solution experiments [31] but disagrees with our chip data and that of others [15] (see also Eqs. (12) and (13)). Especially GG mismatches are apparently much less stable than expected. An analogous low stability of GG mismatches on microarrays compared with solution data was reported for DNA/RNA hybridizations [25]. It has been concluded that thermodynamic properties of oligonucleotide hybridization are by far not yet understood and not suited to assess probe quality.

Poly-guanine motifs

Consideration of the neighboring bases shows that the apparent low stability of Gg-mismatches is accompanied with triple G-motifs in the probe sequence. These runs of guanines are associated with low intensities in triples with both, central WC-(At-group) and mismatched (Aa-, Ag- and Ac-group) pairings. The stability of central Gg-pairings in the context of adjacent 'non-G'-bases, on the other hand, roughly agrees with the predictions from solution data (see Figure 11).

Our analyses reveal the following effects of triple-G on the observed probe intensities:

(i) The GGG-effect is non-complementary, i.e. the complementary triples (e.g. CCC for perfect matches) don't show exceptionally small intensities as probes with GGG do.

(ii) Exceptional small intensities are also observed for triple-G with central mismatches independent of the nominal pairing of the central base (see the arrows in Figure 12 which indicate the GGG-associated motifs BB' = CG, GG, TG, AG in CBC/GB'G).

(iii) The effect is non-additive, i.e. the intensity drop due to GGG is inconsistent with the decomposition into GG-contributions in the context of all triple-motifs.

(iv) The effect depends on the sequence position being typically smaller near the ends of the probe sequence (see Figure 3).

(v) For probes with one mismatched pairing one observes, in contrast to (iv), that terminal GGG at the solution end of the probes gain intensity, i.e. the sign of the effect reverses compared with the remaining sequence positions.

(vi) The intensity drop due to one triple-G corresponds roughly to 50% of the intensity loss due to one mismatched pairing (see Figure 3).

The observations (i) and (ii) strongly indicate that the triple-G effect is not associated with the nominal base pairings deduced from the binding mode because otherwise one expects equal intensity changes for complementary sequence motifs. Observation (iii) indicates that the effect exceeds the range of stacking interactions with the nearest neighbors. Observation (vi) shows that the magnitude of the effect is relatively large compared with the variability due to other base-specific effects but smaller than the variability due to single mismatches.

To get further insight into the properties of poly-G motifs we calculated the mean sensitivity for runs of identical bases of length one to five, e.g. G, GG,…,GGGGG averaged over all sequence positions of homozygous-present PM-probes (PM-G•G, see Figure 13 and also Figure 3). The sensitivities of all considered runs fit along straight lines with similar absolute values of their slope for adenines, thymines and cytosines (see Figure 13). The slope characterizes the mean sensitivity increment per nucleotide in the run which, in turn, estimates the stability gain (or loss) upon formation of one additional WC pairing in the probe/target

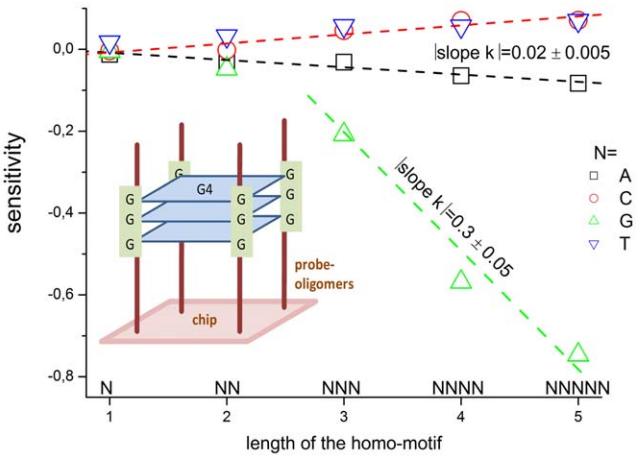


Figure 13. Sensitivities of runs of identical bases. The sensitivity values are averaged over all sequence positions of homo-motifs of length 1 to 5 of homozygous present probes (PM-G'•G, see also Figure 3). Adenines, cytosines and thymines follow straight lines the slope of which is related to the mean stability increment per additional WC pairing in the runs. For guanines the absolute value of the slope drastically increases by more than one order of magnitude for longer poly-G runs exceeding two adjacent G. This effect is attributed to the formation of stacks of at minimum three G-tetrads (G4, see the sketch within the figure which illustrates the structure of a parallel quadruplex formed by four neighbored probe oligomers with GGG-runs at the same sequence position; they are assumed to aggregate into three G4-layers).
doi:10.1371/journal.pone.0007862.g013

duplexes compared with the mean stability of all canonical base pairings. The absolute value of the increment agrees roughly with that of the other bases for single- and double-G (see Figure 13). It however steeply increases for poly-G of length greater than two by more than one order of magnitude. Obviously this change of the slope cannot be attributed to the incremental effect of additional WC pairings in agreement with observations (i) and (ii) but, instead, it presumably reflects the formation of another structural motif accompanied with an increased intensity penalty per additional guanine per run.

Previous studies also reported abnormal intensity responses of probes containing multiple guanines in a row (called G-runs or G-stacks) compared with other probes in different chip assays including Affymetrix expression and SNP arrays [9,17,39–41]. It was found in agreement with our results that the effect is asymmetric with respect to complementary C-stacks [40,41] and depends on the sequence position of the stack with a very strong amplitude at the solution-end position [41]. Note that on expression arrays poly-G containing probes show the opposite tendency as on the studied SNP arrays: They shine relatively bright with intensities exceeding the expected signal level [41,42]. This opposite trend of abnormal strong intensities is associated with non-specific hybridization [41].

The structural rationale behind the poly-G effect has been concordantly assigned to the propensity of poly-G motifs to arrange into stacks of stable molecular bundles of guanine tetrads. These structures potentially affect the efficiency of oligonucleotide synthesis and/or the hybridization of the probes to their target sequences accounting for the abnormal performance of G-runs on the array [17,39–41]. Each G-tetrad is held together by eight Hoogsteen-hydrogen bonds and further stabilized by monovalent cations reducing the electrostatic repulsion between the nucleotides. At minimum three of such planar G-tetrads usually stack

together forming very stable complexes via re-folding of one DNA-strand with several poly-G motifs [43,44] or via aggregation of several DNA-strands with one poly-G motif in each of them (parallel G-quadruplexes, see the sketch in Figure 13). It has been conclusively argued that probe oligomers in close proximity containing poly-G motifs at the same sequence position are prone to aggregate into such parallel G-quadruplexes in the crowded conditions on the surface of high density microarrays [17,45]. The length of 25-meric probes (~22 nm) largely exceeds the average separation between neighboring oligonucleotides on such arrays (~3 nm) which enables complexation of four adjacent probe strands as schematically illustrated in Figure 13. The onset of the stronger sensitivity decrement per additional guanine for triple-G motifs shown in Figure 13 supports the hypothesis that tree layers of G-tetrads represent the minimum motif for stable G-quadruplexes.

As mentioned above, there are two dimensions which potentially affect the performance of probes containing poly-G motifs: firstly, their ability to be correctly synthesized on an array, and secondly the ability of correctly synthesized probes to bind its target.

Let us discuss the first option. The GeneChip arrays are fabricated by *in situ* light-directed combinatorial synthesis on the surface of the array which is prone to produce 5′-truncated products but not internal deletions [46–48]. On can suggest that the synthesis yield per nucleotide is reduced in poly-G runs of length greater than two compared with the average synthesis yield possibly because the formation of G-quadruplexes between neighboring probes affects photo-deprotection of the partly synthesized oligonucleotides. As a result of incomplete synthesis the oligonucleotide features are contaminated with probe sequences which are truncated at the nominal position of the poly-G motif. The probability and thus also the number of such truncated probes is expected to increase with the length of the poly-G motif according to the synthesis yield per additional guanine. Truncated probes of length less than 22−20 nucleotides can be assumed to act as weak binders for the targets. Their binding affinity roughly refers to that of full-length probes with more than two mismatches (see Figure 2b and also ref. [13]). The truncated oligomers only weakly contribute to the intensity of the probe spots in mixtures with full length probes at low and intermediate target concentrations. As a result, the observed intensity drop of poly-G containing probe sequences is the result of the reduced number of full length probe oligomers in the respective probe spots. Their fraction can be approximately estimated by assuming proportionality between the intensity drop and the remaining number of full length probes $\sim 10^{Y(GGG)} \approx 0.4$–$0.5$ for GGG motifs (with $Y(GGG) = -0.2 \ldots -0.3$; see Figure 2b and Figure 13). This fraction is equivalent with the effective synthesis yield per additional G of 40%–50% which roughly halves the number of remaining full length probes according to our data. The general effect of incomplete probe synthesis on the hybridization of microarrays has been discussed in refs. [49] and [13].

Also the second option of modified target binding to correctly synthesized probes provides a tentative explanation of the GGG-effect [45]. It assumes that complex formation between the probe oligomers effectively blocks the involved probe strands and this way reduces the amount of free binding sites accessible for the targets with consequences for their effective association constant which is expected to decrease (see Eq. (18)). The probe-probe interaction term in Eq. (18) assumes simply bimolecular interactions between the probes. Substitution by an appropriate higher-order interaction term which considers the stoichiometry of quadruplex formation, the proximity relations and the fixation of the probes on the chip-surface is expected to modify the respective contribution but leaves the expected trend unchanged.

Note that both discussed potential interpretations of the GGG-effect give rise to a common cause of the observed small intensity values, namely the reduced number of available binding sites for target binding either via truncation or via complexation of part of the probe oligomers. Both interpretations are compatible with our observations (i) and (ii) because the reduced amount of full-length probes and also probe-probe complexes are independent of the respective complementary target sequence upon allele-specific hybridization and independent of the respective mismatched target motif upon cross-allelic hybridization. Also the onset of the increased sensitivity increment per additional guanine for triple-G motifs shown in Figure 13 supports both hypotheses because stable G-quadruplexes of the probes are assumed to affect synthesis and hybridization as well.

Tethering of the involved oligonucleotides to the surface and zippering effects towards both ends of the probes are expected to modify their propensity for G-tetrad formation in a positional dependent fashion in analogy with the positional dependence of base pairings in probe/target dimers [9,13,19,50–52]. This trend provides a rationale for effect (iv). Note however that probe-probe interactions modulate target binding via the array-factor $F_{array} < 1$ (Eq. (18)). The GGG-profile of homozygous-absent probes (PM-G'•G, see part d of Figure 3) shows the typical characteristics of the mismatched pairing in the middle of the sequence. This result indicates that a certain fraction of the oligomers of the respective probe spot form specific dimers with the cross-allelic or allele-specific target as expected for the respective hybridization mode. This result is in agreement with both hypotheses discussed because incomplete synthesis and probe-probe complexes reduce but not prevent specific hybridization.

The suggested mechanisms explain the decreased intensity of probes containing runs of consecutive guanines. The effect (v) however seems puzzling because terminal poly-G's increase the intensity of the respective probes, instead. On expression arrays one even observes much stronger intensity gains for poly-G containing probes [41,42]. This opposite trend of abnormal strong intensities is clearly associated with non-specific hybridization. We suggest that G-rich probes are able to form G-quadruplexes of different stoichiometry with non-specific targets containing longer runs of guanines in a positional dependent fashion with a strong bias towards the solution end of the probe. For SNP arrays the relative contribution of non-specific hybridization is relatively weak compared with expression arrays (see Text S1), which explains the relatively weak effect of bright poly-G motifs near the solution end of the probe sequences. Also the fact that effect (v) becomes evident only for relatively weak signals of probes forming at minimum one mismatched pairing is compatible with an additive contribution due to non-specific binding (Eq. (2)). At larger probe intensities, non-specific binding becomes less important compared to specific binding. For completeness we notice that Upton et al. suggested an alternative mechanism which increases the intensity of poly-G containing probes via local opening of regions in the vicinity of quadruplexes [17].

In summary, our data support the hypothesis that runs of consecutive guanines facilitate the formation of stable G-quadruplexes between neighboring probes which in final consequence reduce the number of probe oligomers available for target binding via two alternative mechanisms, firstly, the reduced synthesis yield of full length probes and/or, secondly, the formation of complexes of neighboring full-length probes. Both hypotheses are compatible with the observed intensity drop of probes containing runs of guanines on SNP arrays.

GGG-runs are relatively common on SNP arrays: About 11% of all probes on the studied 100k GeneChip SNP arrays contain at minimum one triple GGG motif and nearly 30% of the allele-sets contain at minimum one of these probes. We conclude that the discussed effect cannot be neglected in appropriate correction methods.

Correcting probe intensities for sequence effects

The SNP-specific sequence bias transforms into systematic errors of the genotyping characteristics derived from the signals of single probes. Note that the sequence-context of a partial SNP and consequently also the respective bias is essentially very similar for all probes of a selected probe set addressing the same SNP. As a consequence, the averaging of the probe signals into set-related allele values only weakly reduces the systematic signal error after the summarization step. SNP arrays differ in this respect from expression arrays where the sequences of the set of probes interrogating the expression of the same gene or exon can be chosen independently to a larger degree.

One central task of the preprocessing of signals of SNP probes is consequently their correction for sequence effects and in particular for SNP-specific biases. The detailed presentation and verification of an appropriate algorithm is beyond the scope of the present work and will be given elsewhere. The results of our systematic study however enable to identify relevant sequence motifs which significantly modulate the probe intensities. The intensity contributions of such motifs constitute the building blocks of an appropriate intensity model. In particular our results suggest the following rules for sequence correction of SNP probe intensities:

(i) Sequence effects due to WC pairings between probe and target are well approximated using nearest-neighbor (NN) motifs in analogy with accepted NN-free energy models for oligonucleotide-duplexing in solution [33].

(ii) The anisotropy of probe/target interactions due to the fixation of the probes at the chip surface and end-opening (zippering effects) [13,49] requires the consideration of the positional dependence of the interactions in a motif-specific fashion, i.e. separately for each NN-combination of nucleotide letters. The assumption of a generic shape function which applies to all motifs seems suboptimal [9,53].

(iii) The modulation of probe intensities by mismatched pairings can be considered using triple-motifs which consist of the central mismatch and the two adjacent WC pairings.

(iv) Nominal base pairings according to (i) and (iii) can be deduced from the hybridization mode of the respective probes which, in turn, provides selection criteria of the probes for parameter estimation. The mean intensity penalty owing to one and two mismatches can be estimated from the respective class of probes.

(v) Runs of triple guanines (GGG) represent a special motif which markedly modulates the intensities of the respective probes. The underlying effect does not originate from probe/target (pairwise) interactions but obviously results from the formation of collective complexes presumably of four neighboring probes. Therefore it affects essentially all probes with triple G-motifs independently of the hybridization mode.

(vi) Also tandem mismatches represent a special motif of MM-probes with a modified intensity penalty compared with other MM-probes possessing two mismatches with at least one WC pairing in-between. This sequence effect can be taken into account in a first order approximation by decomposing the quadruple formed by the tandem mismatch and the two adjacent WC pairings into two NN-terms referring to a WC- and a mismatched pairing each, or more roughly, by explicitly considering the two adjacent WC pairings.

(vii) The shift of mismatch motifs by a few sequence positions about the middle base of the probe and the effect of flanking mismatches adjacent to triples with a central mismatch can be neglected to a good approximation.

(viii) Background intensity contributions (optical background and "chemical" background due to non-specific hybridization) should be considered especially for probes forming at least one mismatched pairing.

Established preprocessing algorithms for GeneChip SNP arrays explicitly consider the mean intensity penalty per mismatch [54,55] or, in addition, the single-base-related positional effect [56]. The authors of the latter work conclude from their results that, after correction, '…the sequence effect is reduced but can be further improved'. Our results clearly show that effects which are not taken into account in this model, namely the particular mismatch and its sequence context, the contribution of nearest neighbor stacking interactions and of triple-G runs, considerably modulate the probe intensities. We expect that their explicit consideration will further improve genotyping based on SNP microarrays.

Our present analysis has focused on sequence effects. Note for sake of completeness that an elaborated correction algorithm should also consider additional sources of intensity variation not taken into account here, such as the fragment length and the GC-content of the targets [38,56] and non-linear effects due to saturation of the probes at large transcript concentrations [23,57,58], non-specific hybridization [10] and/or bulk depletion of the targets [59,60].

Summary and Conclusions

Single mismatched pairings formed in cross-allelic probe target duplexes and runs of poly G-motifs in the probe sequence are, with the exception of the number of mismatches per duplex, the main sources of signal variability on SNP arrays. These effects must be considered in appropriate calibration methods of the probe intensities to improve the accuracy of genotyping and copy number estimates. The poly-G effect seems to be related to the crowded arrangement of probes on high density oligonucleotide arrays which facilitates the formation of G-quadruplexes between neighboring probes and this way reduces the amount of free probes available for target binding either via incomplete synthesis of full length oligomers and/or via complexation of full length probes. The probe/target interactions on the chip can be decomposed into nearest neighbor contributions which in most cases well correlate with the respective free energy terms describing DNA/DNA-interactions in solution. The effect of mismatches is about twice as large as that of canonical pairings for unknown reasons. Triple-averaging represents a model-free approach to estimate the mean intensity contributions of different sequence motifs which can be applied in improved calibration algorithms to correct signal values for sequence effects.

Supporting Information

Supporting Text S1 Hybridization modes and base pairings for probe selection. The supporting text provides an overview about the hybridization modes, probe attributes and interaction groups; about base pairings in probe/target duplexes at the middle and

SNP position of the probe sequences; and how probes are selected for triple-averaging (including the 'hook' criteria and background correction).
Found at: doi:10.1371/journal.pone.0007862.s001 (0.44 MB PDF)

Author Contributions

Conceived and designed the experiments: HB. Analyzed the data: HB TG. Contributed reagents/materials/analysis tools: MF TG. Wrote the paper: HB TG.

References

1. Weinberg RA (1996) How cancer arises. Scientific American 275: 62–70.
2. Hacia JG (1999) Resequencing and mutational analysis using oligonucleotide microarrays. Nature Genetics 21: 42–47.
3. Bruun GM, Wernersson R, Juncker AS, Willenbrock H, Nielsen HB (2007) Improving comparability between microarray probe signals by thermodynamic intensity correction. Nucleic Acids Research 35.
4. Seringhaus M, Rozowsky J, Royce T, Nagalakshmi U, Jee J, et al. (2008) Mismatch oligonucleotides in human and yeast: guidelines for probe design on tiling microarrays. BMC Genomics 9.
5. Binder H, Krohn K, Preibisch S (2008) "Hook" calibration of GeneChip-microarrays: chip characteristics and expression measures. Algorithms for Molecular Biology 3: 11.
6. Binder H, Preibisch S (2008) "Hook" calibration of GeneChip-microarrays: Theory and algorithm. Algorithms for Molecular Biology 3: 12.
7. Burden CJ (2008) Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed. Physical Biology 5: 016004.
8. Ferrantini A, Allemeersch J, Van Hummelen P, Carlon E (2009) Thermodynamic scaling behavior in genechips. BMC Bioinformatics 10.
9. Zhang L, Wu C, Carta R, Zhao H (2006) Free energy of DNA duplex formation on short oligonucleotide microarrays. Nucleic Acids Research. gkl1064.
10. Binder H, Bruecker J, Burden CJ (2009) Non-specific hybridization scaling of microarray expression estimates - a physico-chemical approach for chip-to-chip normalization. Journal of Physical Chemistry B 113: 2874–2895.
11. Binder H, Preibisch S (2006) GeneChip microarrays - signal intensities, RNA concentrations and probe sequences. Journal of Physics Condensed Matter 18: S537–S566.
12. Binder H (2006) Probing gene expression - sequence specific hybridization on microarrays. In: Kolchanov N, Hofestaedt R, eds. Bioinformatics of Gene Regulation II: Springer Sciences and Business Media. pp 451–466.
13. Binder H (2006) Thermodynamics of competitive surface adsorption on DNA microarrays. Journal of Physics Condensed Matter 18: S491–S523.
14. Halperin A, Buhot A, Zhulina EB (2006) On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions. Journal of Physics Condensed Matter 18: S463–S490.
15. Naiser T, Ehler O, Kayser J, Mai T, Michel W, et al. (2008) Impact of point-mutations on the hybridization affinity of surface-bound DNA/DNA and RNA/DNA oligonucleotide-duplexes: Comparison of single base mismatches and base bulges. BMC Biotechnology 8: 48.
16. Wu C, Zhao H, Baggerly K, Carta R, Zhang L (2007) Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. Bioinformatics 23: 2566–2572.
17. Upton G, Langdon W, Harrison A (2008) G-spots cause incorrect expression measurement in Affymetrix microarrays. BMC Genomics 9: 613.
18. Binder H, Preibisch S, Kirsten T (2005) Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. Langmuir 21: 9287–9302.
19. Naiser T, Kayser J, Mai T, Michel W, Ott A (2008) Position dependent mismatch discrimination on DNA microarrays - experiments and model. BMC Bioinformatics 9: 509.
20. Binder H, Kirsten T, Hofacker I, Stadler P, Loeffler M (2004) Interactions in oligonucleotide duplexes upon hybridisation of microarrays. Journal of Physical Chemistry B 108: 18015–18025.
21. Binder H, Preibisch S (2005) Specific and non-specific hybridization of oligonucleotide probes on microarrays. Biophysical Journal 89: 337–352.
22. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, et al. (2005) Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. Bioinformatics. bti275.
23. Binder H, Kirsten T, Loeffler M, Stadler P (2004) The sensitivity of microarray oligonucleotide probes - variability and the effect of base composition. Journal of Physical Chemistry B 108: 18003–18014.
24. Wick L, Rouillard J, Whittam T, Gulari E, Tiedje J, et al. (2006) On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes. Nucleic Acids Research 34: e26.
25. Pozhitkov A, Noble P, Domazet-Loso T, Nolte A, Sonnenberg R, et al. (2006) Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. Nucleic Acids Research 34: e66.
26. Lee I, Dombkowski AA, Athey BD (2004) Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. Nucleic Acids Research 32: 681–690.
27. Marcelino LA, Backman V, Donaldson A, Steadman C, Thompson JR, et al. (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. Proc Natl Acad Sci U S A 103: 13629–13634.
28. Heim T, Wolterink JK, Carlon E, Barkema GT (2006) Effective affinities in microarray data. Journal of Physics Condensed Matter 18: S525–S536.
29. Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, et al. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry 34: 11211–11216.
30. Ikuta S, Takagi K, Bruce Wallace R, K. I (1987) Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs. Nucleic Acids Research 15: 797–811.
31. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J (1999) Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal AA, CC, GG, and TT Mismatches. Biochemistry 38: 3468–3477.
32. SantaLucia J, Hicks D (2004) The thermodynamics of DNA structural motifs. Annual Review of Biophysics and Biomolecular Structure 33: 415–440.
33. SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbour thermodynamics. Proc Natl Acad Sci U S A 95: 1460–1505.
34. Fish D, Horne M, Brewood G, Goodarzi J, Alemayehu S, et al. (2007) DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. Nucleic Acids Research 35: 7197–7208.
35. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989) Numerical Recipes. New York: Cambrigde University Press.
36. Hooyberghs J, Van Hummelen P, Carlon E (2009) The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. Nucleic Acids Research 37: e53.
37. Matveeva OV, Shabalina SA, Nemtsov VA, Tsodikov AD, Gesteland RF, et al. (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. Nucleic Acids Research 31: 4211–4217.
38. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, et al. (2005) A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays. Cancer Res 65: 6071–6079.
39. Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians F, et al. (2003) Probe selection for high-density oligonucleotide arrays. Proc Natl Acad Sci U S A 100: 11237–11242.
40. Sharp AJ, Itsara A, Cheng Z, Alkan C, Schwartz S, et al. (2007) Optimal design of oligonucleotide microarrays for measurement of DNA copy-number. Hum Mol Genet 16: 2770–2779.
41. Wu C, Zhao H, Baggerly K, Carta R, Zhang L (2007) Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. Bioinformatics 23: 2566–2572.
42. Fasold M, Preibisch S, P S, Binder H (2009) The GGG-bias of GeneChip expression data - analysis and correction. submitted.
43. Rachwal PA, Brown T, Fox KR (2007) Effect of G-Tract Length on the Topology and Stability of Intramolecular DNA Quadruplexes. Biochemistry 46: 3036–3044.
44. Sühnel J (2001) Beyond nucleic acid base pairs: From triads to heptads. Biopolymers 61: 32–51.
45. Langdon WB, Upton GJG, Harrison AP (2009) Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips. Brief Bioinform 10: 259–277.
46. McGall GH, Barone AD, Diggelman M, Fodor SPA, Gentalen E, et al. (1997) The Efficiency of Light-Directed Synthesis of DNA Arrays on Glass Substrates. Journal of the American Chemical Society 119: 5081–5090.
47. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, et al. (1996) Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci U S A 93: 13555–13560.
48. Pirrung MC, Fallon L (1998) Proofing of photolithographic DNA synthesis with 3′,5′-dimethoxybenzoinyloxycarbonyl-protected deoxynucleoside phosphoramidites. Journal of Organic Chemistry 63: 241–246.
49. Naiser T, Kayser J, Mai T, Michel W, Ott A (2009) Stability of a Surface-Bound Oligonucleotide Duplex Inferred from Molecular Dynamics: A Study of Single Nucleotide Defects Using DNA Microarrays. Physical Review Letters 102: 218301–218304.
50. Naef F, Magnasco M (2003) Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. Physical Review E 68: 011906.
51. Deutsch JM, Liang S, Narayan O (2004) Modeling of microarray data with zippering. arXiv: q-bioBM/0406039 v1.
52. Zhang L, Miles M, Aldape K (2003) A model of molecular interactions on short oligonucleotide microarrays. Nature Biotechnology 21: 818–821.
53. Shen F, Huang J, Fitch K, Truong V, Kirby A, et al. (2008) Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. BMC Genetics 9: 27.

54. LaFramboise T, Harrington D, Weir BA (2006) PLASQ: A Generalized Linear Model-Based Procedure to Determine Allelic Dosage in Cancer Cells from SNP Array Data. Harvard University Biostatistics Working Paper Series 44: 1–28.
55. LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, et al. (2005) Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis. PLoS Computational Biology 1: e65.
56. Carvalho B, Bengtsson H, Speed TP, Irizarry RA (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostat 8: 485–499.
57. Held GA, Grinstein G, Tu Y (2003) Modeling of DNA microarray data by using physical properties of hybridization. Proc Natl Acad Sci U S A 100: 7575–7580.
58. Burden CJ, Pittelkow YE, Wilson SR (2004) Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays. Statistical Applications in Genetics and Molecular Biology 3: 35.
59. Burden C, Binder H (2009) Physico-chemical modelling of target depletion during hybridisation on oligonulceotide microarrays. submitted.
60. Suzuki S, Ono N, Furusawa C, Kashiwagi A, Yomo T (2007) Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. BMC Genomics 8: 373.

Supporting Text S1

Mismatch- and G-stack modulated probe signals on SNP microarrays

Hans Binder[1*], Mario Fasold[1], Torsten Glomb[1]

[1] Interdisciplinary Centre for Bioinformatics, Universität Leipzig, Leipzig, Germany

Numbering of reference refers to the reference list of the main paper

Hybridization modes and base pairings for probe selection

1. Hybridization modes, probe attributes and interaction groups

Hybridization mode	Probe attributes			Interaction groups				
	type	SNP offset δ[1]	base position[1]	Ab-group [2] At	Aa	Ag	Ac	no. of mismatches #mm[3]
Specific (S) P-G•G	PM	all	mb	x				**0**
		all	SNP	x				
	MM	≠0	mb		x			**1**
		=0	mb/SNP			x	x	
		≠0	SNP	x				
cross-allelic (C) P-G'•G	PM	≠0	mb	x				
		all	SNP		x	x	x	
	MM	=0	mb/SNP			x	x	
	MM	≠0	mb		x			**2**
		≠0	SNP		x	x	x	

[1] Base pairings formed at the center position of the 25meric probe sequence (mb…middle base) or at the SNP position (SNP) which is offset by δ base positions relatively to the center position. The mb- and SNP positions are consequently identical for δ=0.

[2] Base pairings are classified into four Ab-groups (b = a,t,g,c) as follows: At-group (At, Ta, Gc, Cg); Aa-group (Aa, Tt, Gg, Cc); Ag-group (Ag, Tc, Ga, Ct); Ac-group (Ac, Tg, Gt, Ca). Lower case letters refer to the target.

[3] Number of mismatches per probe/target duplex

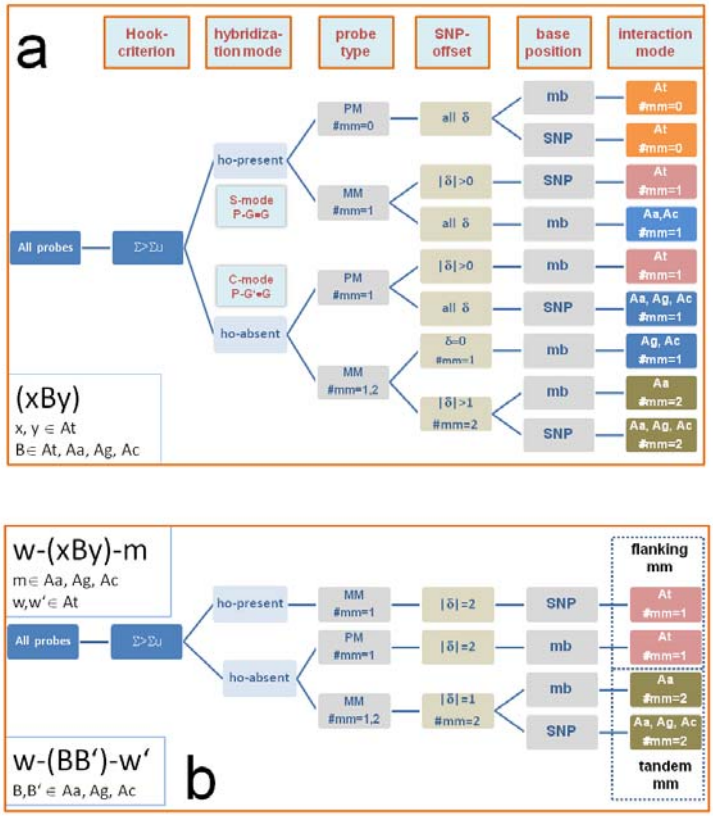
2. Base pairings in probe/target duplexes at the middle and SNP position of the probe sequences [a]

Position	SNP offset	SNP type	PM-base B	Base pairing Bb S-mode (P-G•G) PM	S-mode (P-G•G) MM	C-mode (P-G'•G) PM	C-mode (P-G'•G) MM	Probes [b] number	Probes [b] percent
k=13 (mb)	δ≠0		T	*At* Ta	*Aa:* Aa	*At:* Ta	*Aa:* Aa	52,940	26.2
			A	*At:* At	*Aa:* Tt	*At:* At	*Aa:* Tt	52,800	26.2
			C	*At:* Cg	*Aa:* Gg	*At:* Cg	*Aa:* Gg	33,008	16.3
			G	*At:* Gc	*Aa:* Cc	*At:* Gc	*Aa:* Cc	33,627	16.7
							total	172,375	85.4
k=13+δ (SNP)	δ≠0	[A/C]	T/G	*At:* Ta/Gc		*Ag:* Tc/Ga		14,683	7.3
		[G/T]	C/A	*At:* Cg/At		*Ag:* Ct/Ag		11,224	5.6
		[A/G]	T/C	*At:* Ta/Cg		*Ac:* Tg/Ca		60,547	30.0
		[C/T]	G/A	*At:* Gc/At		*Ac:* Gt/Ac		60,585	30.0
		[A/T]	T/A	*At:* Ta/At		*Aa:* Tt/Aa		9,273	4.6
		[C/G]	G/C	*At:* Gc/Cg		*Aa:* Gg/Cc		16,063	8.0
							total	172,375	85.4
k=13+δ (mb/SNP)	δ=0	[A/C]	T/G	*At:* Ta/Gc	*Aa:* Aa/Cc	*Ag:* Tc/Ga	*Ac:* Ac/Ca	2,537	1.3
		[G/T]	C/A	*At:* Cg/At	*Aa:* Gg/Tt	*Ag:* Ct/Ag	*Ac:* Gt/Tg	1,956	1.0
		[A/G]	T/C	*At:* Ta/Cg	*Aa:* Aa/Gg	*Ac:* Tg/Ca	*Ag:* Ag/Ga	10,393	5.1
		[C/T]	G/A	*At:* Gc/At	*Aa:* Cc/Tt	*Ac:* Gt/Ac	*Ag:* Ct/Tc	10,265	5.1
		[A/T]	T/A	*At:* Ta/At	*Ac:* Ca/Gt	*Aa:* Tt/Aa	*Ag:* Ct/Ga	1,597	0.8
		[C/G]	G/C	*At:* Gc/Cg	*Ac:* Ac/Tg	*Aa:* Gg/Cc	*Ag:* Ag/Tc	2,777	1.4
							total	29,525	14.6

[a] interaction groups (At, Aa, Ag, Ac) are indicated in leading cursive letters. Note that the probes interrogate each SNP on its sense and antisense strand with mutually complementary sequences. Consequently pairs of complementary letters B and B^c are realized in each probe set giving rise to different combinations of base pairings in the PM and MM probes.

[b] only probes referring to homozygous SNP loci are selected (41,629 out of 58,960 total loci, ~70.1%) and used in further analysis. Note that the probes with δ≠0 (85.4% of all used probes) are used twice, considering the sequence motifs about the middle base (k=13) and about the SNP base (k=13+δ). The remaining 14.6% of probes refer to δ=0. The probes with offset δ≠0 split into 27.6% (55,634) with |δ|=1; 14.2% (28,742) with |δ|=2; 14.0% (28,355) with |δ|=3 and 29.5% (59,644) with |δ|=4.

3. Probe selection for triple-averaging



Standard triples (xBy) are selected according to the scheme shown in part a: The interaction mode of the center base of the triple is defined by the chosen hybridization mode, the probe attributes (type, offset) and the position of 'B' (SNP- or the middle base, mb) in the probe sequence. The interaction mode determines the base pairing formed by 'B' with the target according to one of the four Ab-groups, At, Aa, Ag, Ac (see the Tables above), and the total number of mismatches per probe/target duplex, #mm. Part b shows special selections of triples with one flanking mismatch or of tandem mismatches.

4. *'Hook' criteria for probe selection*

Selection criteria considering non-specific hybridization are chosen from the hook-plot of the chip-data (see ref. [5,6] and also the figure). Briefly, the intensities of each probe pair are transformed according to $\Delta=\langle\log(I^{PM}/I^{MM})\rangle_{allele\text{-}set}$ and $\Sigma=0.5\ \langle\log(I^{PM}\cdot I^{MM})\rangle_{allele\text{-}set}$ (the angular brackets denote averaging over the respective allele-set), plotted into Δ-versus-Σ coordinates and smoothed using a sliding window of ~ 500 data points. Probe-sets with relatively large contribution of non-specific hybridization, $x^{P,N}>0.5$ (see Eq. (5)), are characterized by small coordinate-values Σ and Δ. Both coordinates increase with decreasing x^N and level-off at a peak for vanishing contributions of non-specific binding, $x^{P,N}\approx 0$ (see the figure below).

The logarithmic-fraction of the probe-intensity due to non-specific hybridization can be estimated using the coordinate differences with respect to the starting point of the hook curve [6]

$$\log x^{P,N} \approx -\left((\Sigma-\Sigma_{start})\pm\tfrac{1}{2}(\Delta-\Delta_{start})\right) \quad , \qquad \text{(E1)}$$

where the sum and the difference refer to P=PM(+) and MM(-), respectively. The fraction $x^{P,N}$ depends on the probe type with $x^{PM,N}<x^{MM,N}$ for Σ=const. Practically, a threshold of $(\Sigma-\Sigma_{start})>0.7$ is applied to obtain allele sets with an average nonspecific intensity contribution of less than 20%, i.e. $\langle x^N\rangle_{allele\ set}<0.2$ with $\langle\log(x^N)\rangle_{allele\text{-}set}= 0.5\langle\log(x^{PM,N})+\log(x^{MM,N})\rangle_{allele\text{-}set}$. This implies that the selected allele sets originate at least to 80% either from specific or cross-allelic hybridization.

Note that the hook-plots obtained from SNP arrays lack the horizontal starting range observed typically for expression arrays as a characteristic signature of "absent" probes without complementary targets. Non-specific hybridization to a smaller degree contributes to the signal intensities of SNP arrays compared with expression arrays in agreement with previous results [9]. This difference can be rationalized in terms of the smaller heterogeneity of genomic DNA-copies (in terms of sequence and fragment-length) and especially of the smaller range of copy number variations compared with the range of variation of mRNA-transcript concentrations. The latter can cover several orders of magnitude whereas the former typically change by less than the factor of ten.

Trivially, the strand direction does not affect the strength of the respective base pairings provided that sequence motifs from both, the s- and the as-strands, are considered in the same direction. In our analyses we therefore pool the probes which are assigned to the same interaction mode independently of their strand direction (d=s, as) assuming that the respective genotypes are properly assigned on both strands.

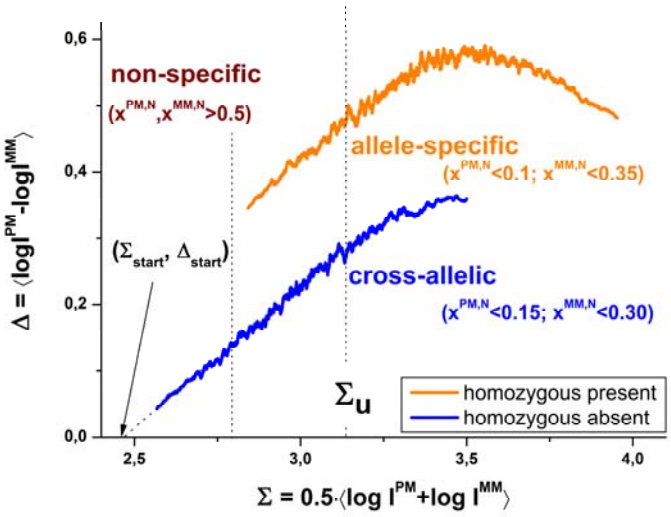


Figure: **Classification of probe-intensities according to their hybridization mode**. So-called hook curves are plotted for homozygous-absent (ha) and -present (hp) probes referring to cross-allelic and allele-specific hybridization modes, respectively. The 'start' coordinates of the hook curve are given by the intersection of the extrapolated ha-hook with the abscissa. The intensity fraction per probe due to non-specific binding depends on the hook coordinates (see Eq. (E1)). The right vertical line refers to $(\Sigma - \Sigma_{start})>0.7$. It was used as threshold for probe selection to characterize the interaction modes upon specific (S) and cross-allelic (C) hybridization. Above this threshold, probe intensities are distorted, on the average, by a contribution of non-specific hybridization of less than 20%. The fraction of non-specific binding slightly differs between the PM and MM probes as indicated in the figure.

5. Background correction and saturation effects

The figure (panels a and b) shows triple averaged mean intensities for all 64 standard triples with centre pairings taken from the At-group (WC pairings) and from the Aa-group (self complementary pairings, see also the next section). The data refer either to #mm=0 and 1 mismatches per duplex (At-group) or to #mm=1 and 2 (Aa-group). The mean intensity level decreases with increasing #mm as discussed in the previous section. The different triples of each class give rise to considerable variability of the intensity values. The standard deviation of the whole set of 64 triples of the At-group is SD(logI)=0.041 and 0.045 for #mm=0 and 1, respectively (part a of the figure), but more than twice as large for the mismatched Aa- (SD=0.12; part b of the figure), Ag- (SD=0.13) and Ac-groups (SD=0.09) for #mm=1 (see also Table 1). Hence, mismatched pairings with adjacent WC pairs give rise to considerably larger variation of duplex stability than triples of WC pairs.

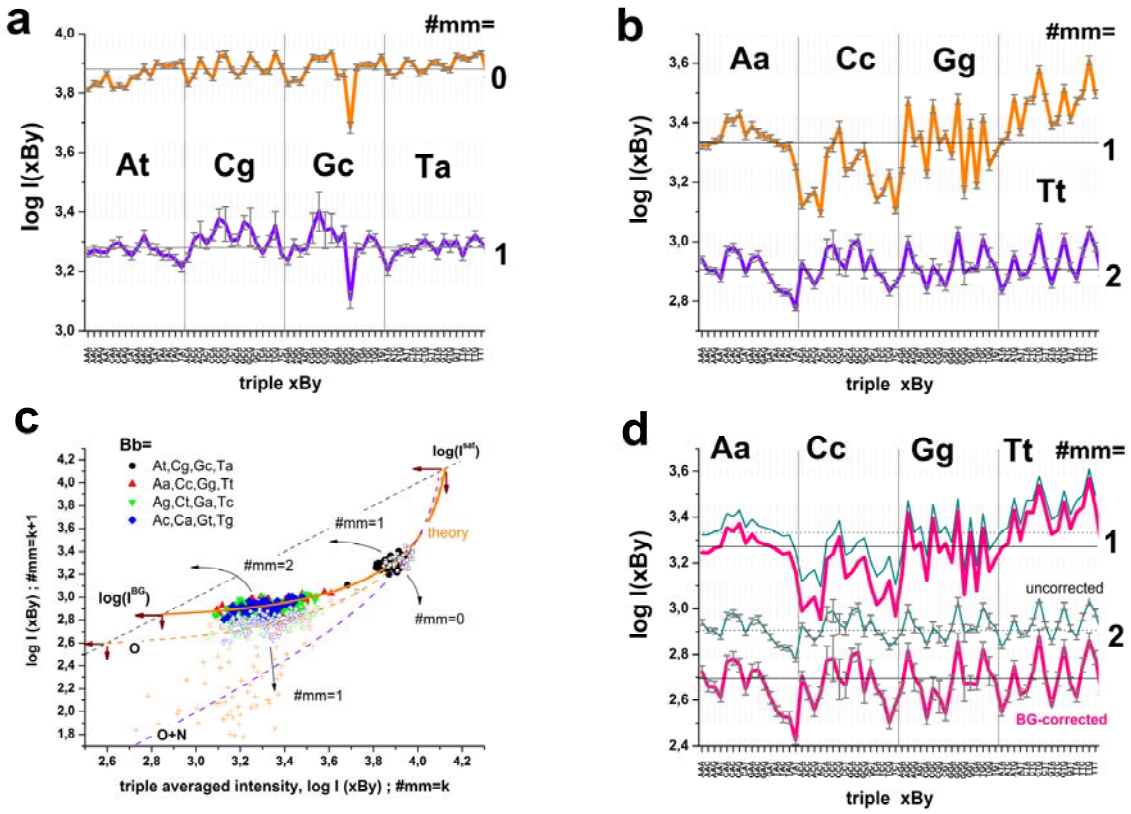


Figure: **Triple-averaged probe intensities and background contribution**. Panel a and b show the 64 triple averaged log-intensities of the perfect match- (At-group) and self complementary mismatch- (Aa-group) pairings. The data refer to different numbers of total mismatches per duplex (#mm, see the figure; the triples are sorted according to their central pairing Bb). These triple averages were correlated for #mm=0-versus-1 and #mm=1-versus-2 in panel c. Here also data for the mismatch-groups Ag and Ac are added. The data do not group in parallel with respect to the diagonal owing to the residual background intensity. Its consideration predicts the grouping of the data along the thick theoretical curve which was calculated using Eq. (E2) with g=11. This curve intersects the diagonal line at the background and saturation intensities, $\log I^{O}=2.85$ and $\log I^{sat}=4.1$, respectively. Correction of the intensities for the optical background (curve "O") slightly improves the linear correlation between the intensities, especially for #mm=1-versus-2 (open symbols). Consideration of the non-specific background ($\log I^{N}=2.6$) further improves linear correlation, however also inflates variation of the data (see also curve "O+N"). Panel d shows the triple-data of the Aa-group before (thin lines) and after (thick lines) background-correction using Eq. (E3).

In general, one expects the similar base-specific effect independently of the total number of mismatches per duplex. To assess this assumption we correlate the triple averaged log-intensities for #mm=k with that for #mm=k+1, i.e. for duplexes which differ by one mismatched pairing (see part c of the figure). Especially the triple-data of the mismatched groups (Aa, Ag, Ac) do not group in parallel with respect to the diagonal line. This behavior indicates poor correlation (solid symbols, see

also part b of the figure which shows the data for the Aa-group with #mm=1 and 2) in contrast to the data of the At-group data (#mm=0, 1; part a of the figure).
The discussed intensities contain contributions due to the optical and non-specific background (see Eqs. (2) and (4)). Moreover, the intensities saturate at large transcript concentrations and/or binding constants $K_{duplex}(\#mm)$. Let us describe the probe intensities by the hyperbolic function of $K_{duplex}(\#mm)$ [23,57]

$$I(\#mm) \approx \left(\frac{I^{sat} \cdot c \cdot K_{duplex}(\#mm)}{1 + c \cdot K_{duplex}(\#mm)} + I^{BG} \right) \quad . \qquad \text{(E2)}$$

I^{sat} denotes the saturation intensity at strong binding, $c \cdot K_{duplex} >> 1$, c is the transcript concentration.
Assuming a factorial increment of the binding constant per mismatch, $K_{duplex}(\#mm+1) = K_{duplex}(\#mm)/g$ (see right axis in Figure 2, panel b), and varying "c $K_{duplex}(0)$" in the limits $0 < c \cdot K_{duplex}(0) < \infty$ we get the theoretical relation between the mean intensities of duplexes which differ by one mismatched pairing (see the curves in panel c of the figure). The theoretical curves intersect the diagonal line (y=x) at low and high intensities at $I=I^{BG}$ and $I=I^{sat}$, respectively, because Eq. (E2) assumes that background and saturation levels are not affected by the number of mismatches. Eq. (E2) predicts significant deviation from the linear relation between the intensities for #mm and #mm+1. The thick curve in panel b of the figure was calculated assuming a residual background intensity of $\log I^{BG} \approx 2.85$. It explains the lack of linear correlation between the experimental triple data for #mm=0-versus-1 and especially of #mm=1-versus-2.
The used background refers to the optical and non-specific contributions according to Eq. (4). To estimate the optical background we simply select 1% smallest intensity probes of the array, calculate their log-intensity average ($\log I^{O}=2.39$), and correct the intensities for this contribution, $I^{corrO} = I - I^{O}$ (see open symbols in panel c of the figure). The dashed curve labeled with "O" refers to these data containing a contribution due to non-specific background intensity of about $\log I^{N} \approx 2.65$. Intensity data which are corrected for both contributions, $I^{corrO+N} = I - I^{BG}$, are shown by the small crosses. The respective theoretical curve labeled "O+N" runs parallel with the diagonal line at decreasing intensities.
The total background correction markedly inflates the variability of the data at small intensities. This effect is well known from microarray analyses as the consequence of diverging log-transformed data at vanishing argument. To avoid this trend it is common practice to confine the corrected data to a lower limit, for example by adding a small constant value to the corrected intensities. We also apply this modification using ($\log I^{N} - o$) with o= 0.6 instead of $\log I^{N}$.
So far we estimated the mean optical and non-specific background levels which apply to all probes of the chip. The background contribution due to non-specific hybridization is governed by the binding reaction of non-specific transcripts (see Eq. (1)). It consequently depends on the probe sequence and thus it is specific for each probe. We previously showed that non-specific hybridization is basically characterized by Watson-Crick pairing [18]. Final background correction of the triple averaged intensities was therefore applied in a sequence specific fashion using

$$I^{corr}(xBy) = I(xBy) - I^{O} - I^{N} \cdot 10^{-o+Y_{At}(xBy)} \qquad \text{(E3)}$$

where $Y_{At}(xBy)$ is the sensitivity of the respective triple of the At-group (see Eq. (8)).
This correction progressively reduces the mean intensity level for #mm=1 and #mm=2 (see Figure 2, part b and the figure above, part d). The triple-specific effect is almost negligible for #mm≤1 but it affects the results for #mm=2.

Erklärung

Hiermit versichere ich, Torsten Glomb, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Leipzig, den 09.05.2010