BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 3. 259–277
Advance Access publication April 8, 2009
doi:10.1093/bib/bbp018

Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips

William B. Langdon, Graham J. G. Upton and Andrew P. Harrison

Submitted: 19th December 2008; Received (in revised form): 3rd March 2009

Abstract

The reliable interpretation of Affymetrix GeneChip data is a multi-faceted problem. The interplay between biophysics, bioinformatics and mining of GeneChip surveys is leading to new insights into how best to analyse the data. Many of the molecular processes occurring on the surfaces of GeneChips result from the high surface density of probes. Interactions between neighbouring adjacent probes affect their rate and strength of hybridization to targets. Competing targets may hybridize to the same probe, and targets may partially bind to more than one probe. The formation of these partial hybrids results in a number of probes not reaching thermodynamic equilibrium during hybridization. Moreover, some targets fold up, or cross-hybridize to other targets. Furthermore, probes may fold and can undergo chemical saturation. There are also sequence-dependent differences in the rates of target desorption during the washing stage. Improvements in the mappings between probe sequence and biological databases are leading to more accurate gene expression profiles. Moreover, algorithms that combine the intensities of multiple probes into single measures of expression are increasingly dependent upon models of the hybridization processes occurring on GeneChips. The large repositories of GeneChip data can be searched for systematic effects across many experiments. This data mining has led to the discovery of a family of thousands of probes, which show correlated expression across thousands of GeneChip experiments. These probes contain runs of guanines, suggesting that G-quadruplexes are able to form on GeneChips. We discuss the impact of these structures on the interpretation of data from GeneChip experiments.

Keywords: *Affymetrix Genechips; expression measures; probe correlations; G-quadruplexes*

INTRODUCTION

One of the most popular forms of microarray is the Affymetrix GeneChip. However, analysing the millions of data points simultaneously provided by each GeneChip is not straightforward. We provide an overview of recent insights into the biophysics and bioinformatics of GeneChip technology. We begin by briefly reviewing the Affymetrix GeneChip and move on to discussing the physics of the technology, describing the various types of molecular interactions which potentially occur on GeneChips. We focus on the evidence for probe–probe interactions and the effects of competitive hybridization between targets binding to several probes and probes binding to several targets. We then discuss the bioinformatics associated with GeneChips, particularly the annotation of probes and the algorithms that have been developed to collate the results from

Corresponding author. Andrew P. Harrison, Department of Mathematical Sciences and Department of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK. Tel: +44 1206 872964; Fax: +44 1206 873043; E-mail: harry@essex.ac.uk

William Langdon is a Software Engineer with over 100 publications. He has almost 30 years experience of developing insightful analysis tools in Industry and Academia.

Graham Upton is a Professor of Applied Statistics at the University of Essex. With more than 100 publications, he is the co-author of *The Oxford Dictionary of Statistics*.

Andrew P. Harrison is a lecturer in Mathematical Bioinformatics at the University of Essex. He has worked on the analysis of Affymetrix GeneChips since 2003. His original research training was in Astrophysics.

multiple probes in a probeset into a single measure of gene activity. These algorithms usually assume that a probe affinity is constant from experiment to experiment. We continue by describing our discovery of a family of thousands of probes that show correlated expression across GeneChip datasets. We conclude by discussing how the existence of this family raises questions about several of the basic assumptions behind GeneChip analysis, such as the size of overlap in sequence expected to result in cross-hybridization, and whether the affinity of these probe sequences are always the same.

AFFYMETRIX TECHNOLOGY

An Affymetrix GeneChip consists of a high-density array of *in situ* synthesized oligonucleotides [1]. Transcripts for each gene are detected through probes of 25 bases. Each gene is measured by 11 or more perfect match (PM) sequences of 25 bases. Each PM also has an associated mis-match (MM) probe that is identical to the PM except for the central base (position 13) being set to the complementary base of the PM.

Each probe is covalently attached to a siloxane layer through a linker, with Affymetrix originally choosing a hexaethylene glycol linker [2]. The probe is grown via the process of photolithography in which light-directed oligonucleotide synthesis selectively removes terminal protecting groups in predefined locations by exposure to light through masks. The photochemistry used by Affymetrix results in the 3′ end of the probe being tethered to the surface and the 5′ end free [3]. The growth of each probe occurs in a series of steps with measurements by Affymetrix [2] showing the stepwise synthesis efficiency increasing over the first approximately six steps, quickly progressing to ∼92–95%. The synthesis of probes is curtailed at 25 bases and so ∼20% of all the probes ($0.92^{25} = 0.13$; $0.95^{25} = 0.28$) will reach this length. The density of initiation sites is high, $\sim 5 \times 10^{17}\,\mathrm{m}^{-2}$ [4], resulting in full-length probes being only ∼3 nm apart. The full extension of a phosphate backbone is 0.7 nm/base repeat [5] so a full-length 25-mer probe, including a flexible linker attached to the surface, may be ∼20 nm in length. It follows that probes can readily come into contact with each other since they are much longer than their separation distances.

In the standard Affymetrix protocol, following synthesis of cRNA, the product is fragmented into short lengths. Affymetrix [1] suggest the length of fragments lie between 30 and 400 bases, whereas [6] suggest between 50 and 200. Thus, the target RNA is considerably longer than the probes to which it sticks. Figure 1 shows the idealized hybridization between the probe 209649_at_pm5 and an example fragment from its target transcript.

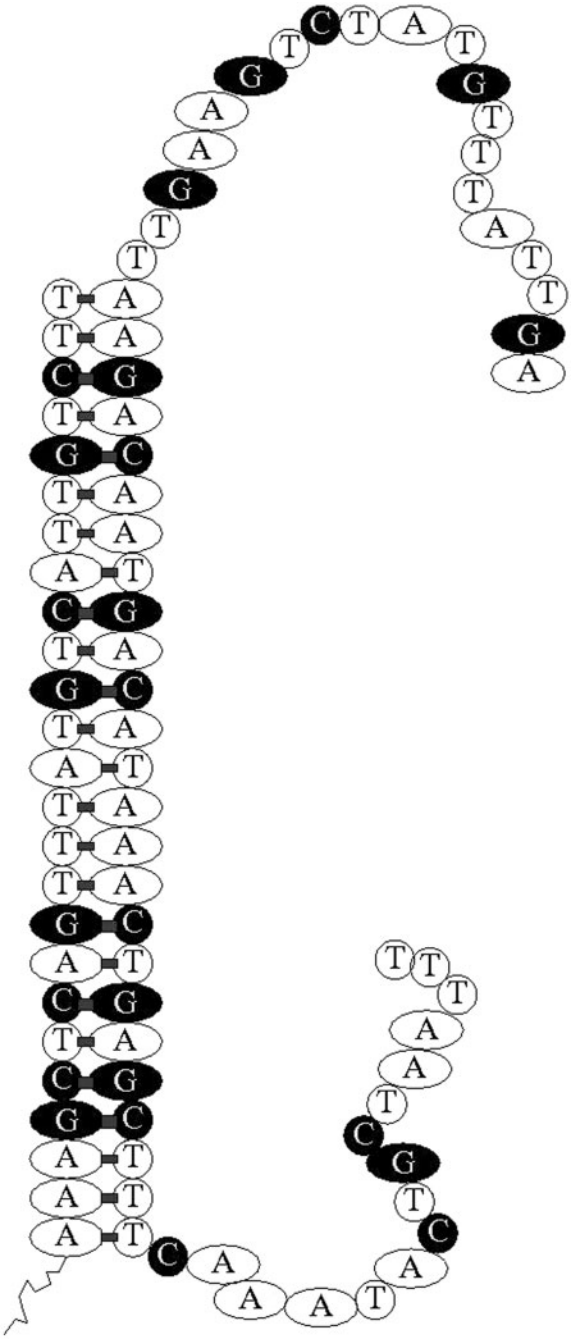

Figure 1: The ideal hybridization between the probe 209649_at_pm5 and its target.

The fragmented RNA is hybridized to the GeneChip for 16 h, after which the array undergoes a series of washing steps. In the standard protocol,

there are two post-hybridization wash steps [7]: a low-stringency wash, followed by a high-stringency wash in which the salt concentration of the buffer is decreased and the temperature is increased. There then follows staining with streptavidin phycoerythrin, washing in a non-stringent buffer, staining with anti-SAPE antibody and washing again in non-stringent buffer before scanning. Several wash cycles are used in order to increase the chances of washing away non-specific duplexes.

THE BIOPHYSICS OF GENECHIPS

Affymetrix have released several 'spike-in' experiments in order to help analysts develop a better understanding of GeneChip technology. These calibration experiments consist of GeneChips being treated with mRNA at carefully controlled concentrations, e.g. a Latin-Square experiment, which uses 42 HGU-133A GeneChips with triplicate measurements of 14 different concentrations for 42 different transcripts. The analysis of such experiments show that the number of targets attached to oligonucleotide microarrays results from the interplay between a number of processes. These include the strength of hybridization, which in turn is affected by the density of probes, the competitive hybridization between multiple targets, the folding of probes and targets, saturation resulting from a lack of probes or target available to hybridize, and which duplexes avoid desorption. The analysis of hybridization to GeneChips, and how to interpret the observed signals in terms of biophysical processes and target concentrations, is a multi-faceted problem [8].

The interplay between the surface density of probes and probe–target and probe-probe interactions

The kinetics of duplex formation in solution is well understood, with the rate-limiting step being the formation of a nucleus containing a few base pairs [9]. Once formed, the nucleus grows by adding base pairs faster than they are dissociated, and the strands 'zip-up'. The size of the nucleus is estimated to be ~5 bp for oligomers containing only AT/U pairs and ~2 bp for oligomers containing at least two GC pairs [9]. Nearest-neighbour models provide a close approximation for the sequence dependence of duplex stability in solution [10]. In particular, the Independent Nearest-Neighbour Hydrogen Bonding model makes the assumption that the stability of a base pair depends upon its adjacent base pairs. The nearest-neighbours 5′AT3′/3′TA5′, 5′TA3′/3′AT5′ and 5′AA3′/3′TT5′ are considered different because the combinations have different stacking energies, resulting from van der Waals' interactions and hydrogen bonds between the base-pair stacks. The stability of a helix then depends upon the base composition of the helix forming the nearest-neighbour interactions as well as the terminal base pairs. Also, cations in solution, such as sodium and magnesium ions, act to reduce the repulsive Coulombic interactions between the negatively charged backbone phosphates and help to determine the stability and folding kinetics of nucleic acids in solution [11].

There are considerable differences between the rates and efficiencies of hybridization in solution compared to that on a microarray [12], with surface hybridization rates 20–40 times slower than solution-phase rates for identical sequences and conditions [13]. The surface probe density of microarrays plays a central role in regulating the hybridization of targets to probes [14]. In low-probe density regimes essentially 100% of probes can be hybridized with relatively quick kinetics whereas at higher densities efficiencies drop and the kinetics of hybridization are also slower [14]. It is evident that not all probes are able to hybridize to targets on high-density oligonucleotide arrays, as Affymetrix [15] report saturating adsorption densities of target RNA to be less than 10% of the probe surface density.

Microarrays carry a high-charge density, due to the phosphate backbones of the nucleic acids, and this acts to decrease the stability of duplexes [16]. This interpretation is supported by the significant differences observed in the thermodynamics of duplex hybridization for solution and surfaces, with a suppression and broadening of the duplex melting curve observed for surface hybridization [15]. There is also the presence of steric crowding at high-probe density [17]. Furthermore, when the surface probe density becomes high enough, a polyelectrolyte brush results from the mutual crowding of the nucleic-acid tails of transcripts hybridized to probes [18]. This brush acts to lower the hybridization efficiency for the largest probe densities. Another explanation for why the observed energy of duplex formation is observed to be smaller on microarrays than in solution is that only a limited fraction of duplexes become fully zipped during hybridization on GeneChips [7].

Crowded conditions on the array surface enable probes to come into contact [6]. A model of the hybridization dynamics of surface-attached DNA oligomers [19] shows that, as probe molecules interact more strongly, fewer nucleation sites become accessible, and binding rates are diminished relative to those in solution. The effective association rate for hybridization is measured to decrease with probe density, σ, and the models indicate a scaling of $\sigma^{-1.8}$ [19]. This means that duplexes will be more stable as the probe density drops and will become less stable as the probe density increases. The number of bound hybrids is of direct relevance to the interpretation of microarray data because it is this value that is responsible for the light detected for each cell. A model of the number of hybrids expected for different probe densities [16] shows a sharp rise in the number occurring within a narrow range of probe densities. According to the model, the narrow range occurs because at low probe densities there are a limited number of probes available, and at high-probe densities the majority of probes cannot hybridize because of the electrostatic repulsion [16]. These models of hybridization dynamics [16, 19] indicate that regions on GeneChips which have reduced densities of probes will result in more hybrids, and these hybrids will be more strongly bound. In contrast, regions of excess probe-density will result in lower numbers of hybrids and they are more likely to be weakly bound. Subtle changes in local probe density will be likely to result in changes in the light reported from the cells containing the probes.

Preheating a microarray surface has a positive effect on hybridization efficiency [14, 17]. We expect that this is due to the preheating dissociating many probe–probe interactions, freeing up probes to bind to target. Probe–probe associations may involve only a few residues but these may still be able to compete with the formation of probe–target duplexes [15]. Hydrogen bonding and stacking of bases from two fixed strands is likely to incur a small loss of conformational entropy, there will be a smaller entropy loss from counterion release, and probe–probe interactions do not incur the RT ln(target) terms that apply to probe–target equilibrium [15]. The formation of probe–probe duplexes has been modelled in 1D, but the two-dimensional monomer–dimer model is not tractable, and no solutions exist to establish whether groups of probes will undergo phase transitions in their population of duplexes and higher-order interactions [6]. All-atom molecular-dynamics simulations of DNA tethered to a surface [20] show a stable tilt in the direction of DNA, suggestive of DNA entering a colloid state [21]. A possible transition to the colloid state in the simulations is induced by surface-induced solvent activity changes, particularly salt-induced DNA–DNA attractions [20]. We are not aware of simulations for single-stranded DNA, but we suggest that they will show related effects to those seen by [20], in particular, the formation of probe–probe interactions, possibly resulting in phase transitions.

Hybridization is usually considered in terms of Watson–Crick base pairing. However, biomolecular crystal structures contain examples of different types of base interactions [22]. Purines and pyrimidines provide three edges for hydrogen-bonding interactions: the Watson–Crick edge; the Hoogsteen edge for purines and the C–H edge for pyrimidines; the Sugar edge. A given edge for one base can interact, in principle, with any of the three edges of another base. Moreover, in RNA structures, there is evidence for base–base interactions involving the *cis* or *trans* orientation of the glycosyl bonds, i.e. the ribose sugars are on the same sides, or opposites sides, of a line joining the interacting edges. This gives a total of 12 alternative base-pairing geometries.

The existence of these non-Watson–Crick pairings opens up the possibility that higher-order interactions between probes may also be prevalent on GeneChips. Of particular interest is the Hoogsteen hydrogen-bonded guanine (G)-tetrad, a planar motif, which has been observed in telomeres [23]. G-quadruplexes result from the hydrophobic stacking of several tetrads and are thermally stable (T_m typically >90°C) [24], with RNA quadruplexes being more stable than their DNA counterparts [25]. Each tetrad is held together by eight hydrogen bonds while a central cation forms cation–dipole interactions with eight guanines, thereby reducing the repulsion of the central oxygen atoms [25]. GeneChip probes containing multiple guanines in a row have abnormal binding behaviour compared with other probes and do not covary with other probes that interrogate the same gene [26]. It has been argued that because probes are immobilized on GeneChips, it is not possible for them to form quadruplexes amongst themselves [26]. By contrast, we would suggest that since probes are immobilized, in close-proximity, and running in parallel, this

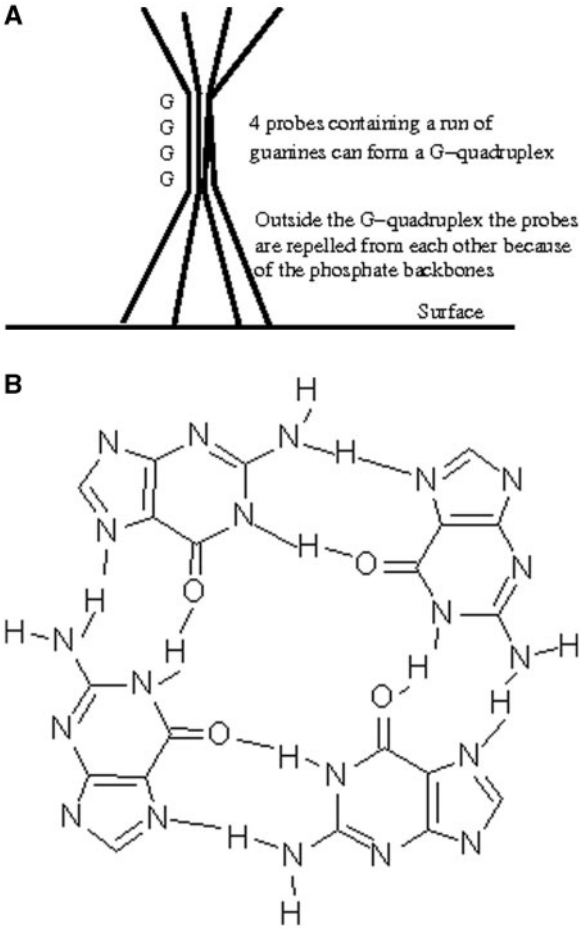

Figure 2: Four adjacent probes, each containing a run of contiguous guanines, may associate and result in a G-quadruplex. (a) The four probes in profile. (b) Looking down on the bonds between the four probes. Note that the guanines within the quadruplex all face in and so cannot bind to the target.

provides an ideal opportunity for four probes containing a run of contiguous guanines to associate and form a stable quadruplex (Figure 2). We note that it must be the probes with long runs of Gs that are abnormal, not the target RNA (which will contain a complementary run of Cs). The possibility of a G-quadruplex forming on the surface of GeneChips has also been suggested by [27].

Competitive hybridization

There is evidence that the multiplicity of non-specific targets binding partially to probes interferes with the formation of specific-target duplexes. A study by Affymetrix [15] of the effects of probe lengths on hybridization on high-density oligonucleotide arrays reports that the signal for a short probe (10-, 12-, 14-, 16- and 18-mer) exceeds the equilibrium signal for a 20-mer probe at 25°C. Similarly, [4] also report that a different 20-mer probe showed smaller absorbed target density than did an 18-mer and 16-mer at 22°C. This is contrary to the expectation that longer probes should produce more stable hybrids than shorter probes and suggests that the probes may be folding or interacting with each other at ~22–25°C [Shingo Suzuki, private communication]. Indeed, another study of hybridization onto oligonucleotide probes with variable lengths [28], at higher temperatures (45°C), shows intensities increasing with longer probe lengths. Furthermore, Glazer and colleagues' own measurements [4] show that the discrepancy between the 18-mers and 20-mers is reduced considerably at 45°C.

The absorbed target density decreases with temperature and changes over time, in the experiments of [15]. Furthermore, the melting curves for target concentrations above 100 nM have dips and step-structures [15]. These experiments suggest that structural reorganization is likely occurring at the array surface and that many probes may be involved in probe–probe interactions [15]. Glazer *et al*. [4] also find an 'overshoot' in absorbed target at high concentrations, with a large amount of target binding rapidly before desorbing to a final plateau. This effect was associated with the high surface density of probes, and the fact that probe–target hybrids can nucleate in many places, but can only fully hybridize when there is a run of complementarity outside the nucleation site [4]. The results of the experiments of [4] and [15] suggest that some targets form bridging interactions linking two probes together (Figure 3a). As hybridization approaches equilibrium most of the partially bound targets become displaced, because one of the complementary targets is able to hybridize to most of the probe [4].

A model of partially bound targets has also been suggested by [7], who see a pronounced wash effect on the intensities. Probes that have a large intensity, immediately prior to washing, see little fall off during the wash cycle, whereas probes with low intensity have a much larger post-wash reduction [7]. These results suggest that, in the earliest stages of hybridization, many probes hybridize partially to non-specific targets and these partial hybrids are stable enough to obstruct other hybridization sites [7]. Over time, non-specific targets are replaced

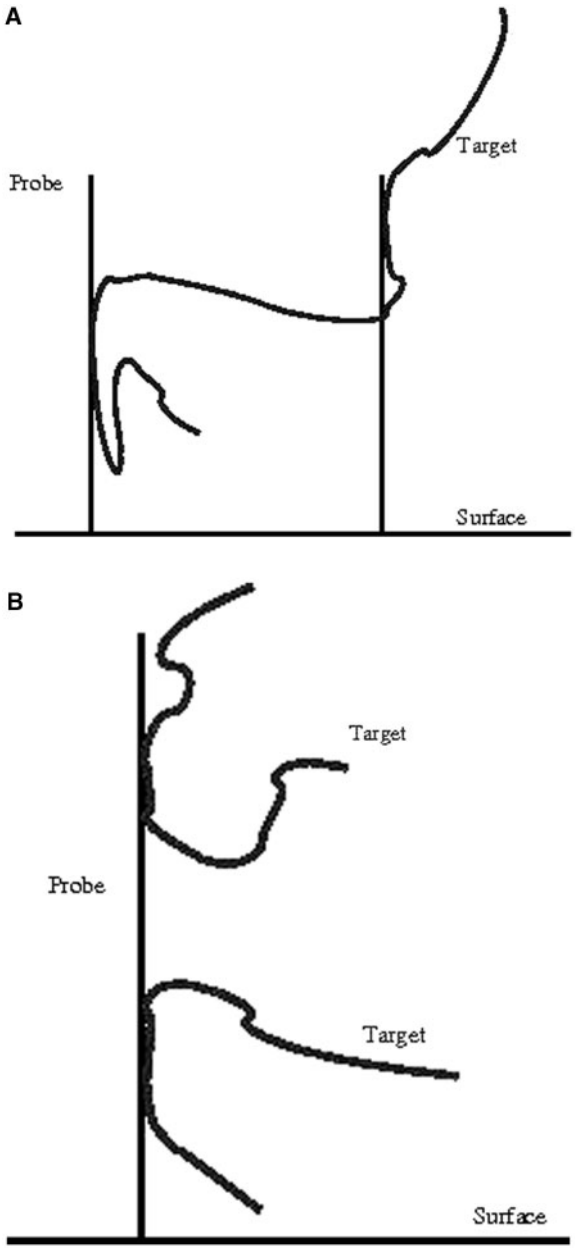

Figure 3: The ratio of probe:target may deviate from 1:1. (a) A target piece of RNA may hybridize to more than one target. (b) More than one piece of RNA may hybridize to the same target.

by specific targets, with the rate of replacement dependent upon the probe sequence. Support for this interpretation comes from experiments studying hybridization of multiple targets to the same probe [29]. When several targets are able to bind to the same probe, a high concentration, low-affinity, species dominates the earliest stages of hybridization. A low concentration high-affinity species then acts to displace the initial hybrid during a second competitive phase [29].

The time-scales of replacement may be much longer than the duration of the hybridization stage of GeneChips and so equilibrium may not be reached for some probes [7]. This is to be expected as models of multiplex hybridization [30, 31] indicate that the presence of multiple species extends the time to reach equilibrium. Moreover, increasing the hybridization stage from 16 to 40 h results in the lowest-intensity probes experiencing less of a reduction in their intensities following the washing stages [7]. These results are in good agreement with the non-equilibrium interpretation of hybridization, because theoretical models show that the time to equilibrium increases as the relative concentrations of the specific target drops [31, 32].

It has been suggested that only the probe–targets, which are fully bound are able to survive a stringent wash [7]. However, an alignment overlap between probes and targets of 10–16 nt is sufficient to result in correlations between spiked-in transcripts and cross-hybridizing probes [33], demonstrating that partial hybrids exist, and are able to remain bound through the washing stages. Furthermore, as we discuss later, we have recently discovered large correlations between probes whose sequences only have a relatively small runs of guanines in common [34]. We suggest that all of these probes are able to nucleate cross-hybridizing transcripts efficiently. The small sequence of overlap in these transcripts must enable the hybridization of these transcripts to survive the washing stage. Moreover, models of the effects of partial zippering between probes and targets suggest that partial binding is prevalent following the wash-cycle [35].

The biophysics of competitive hybridization between several targets that are able to nucleate and partially hybridize to one or more probes underpins the interpretations of [4] and [7]. A model of oligonucleotide replacement [36] has found the activation energy for the displacement pathway to be about one-third that for the dissociative pathway for oligonucleotides in free solution. It has been assumed that the relative energies for these two pathways will be the same on microarrays as in solution [4]. However, the simulations of [4] show that changes to either the dissociation coefficient of the secondary target or to the replacement coefficient, have the same effect on the time course of the adsorbed target density. These results indicate that the ratio of the relative energies of displacement and dissociation on microarrays

is not tightly constrained, suggesting that the rate of replacement may be very sensitive to slight differences in the array synthesis and/or hybridization conditions [4]. Moreover, under physiological conditions, the kinetics of the exchange process for oligonucleotides shorter than 12 nt is dominated by dissociation of the duplex. The replacement of one duplex by another duplex drives the kinetics for sequences longer than 12 nt [36]. Because GeneChip probes are only 25 bases in length at most, it is impossible that any probe will be bound by two different targets which are both hybridized by more than 12 nt. At first glance, it is therefore probable that the dissociative pathway regulates the interactions between targets and probes on GeneChips. We cannot be sure, however, as the dominant hybridization biophysics may be different between physiological conditions and that expected on a chip. We note that, if the kinetics of replacement on GeneChips dominates the removal of duplexes for sequences shorter than 12 nt, the rate of replacement is likely to be dependent on the sequence composition. We expect that this will result in the existence of hot-spot motif sequences that form partial hybrids that then cannot be displaced by other partial hybrids. The existence of such hot-spot motifs will then regulate the population of partially bound targets at the end of the hybridization stage.

A further issue affecting the rates of nucleation along each probe results from the asymmetry of a microarray, with one end of each probe attached to the surface and the other end free. Experiments indicate that the position along the probe, as well as changes in probe density, act to modify the efficiency of nucleation and hybridization [17]. A study of the hybridization of an 18-mer target to a 25-mer probe compared hybridization to the 18 bases at the tethered end with hybridization to the 18 bases at the free end [17]. At sparse probe coverage densities of $1.5 \times 10^{16}\,\mathrm{m}^{-2}$ the targets hybridize at the same rate, whereas at a higher probe coverage density of $3 \times 10^{16}\,\mathrm{m}^{-2}$, the 18-mer binding at the free end of the probe hybridizes several times faster than the target binding to the surface end of the probe [17]. Theoretical models of hybridization dynamics [19] are in reasonable agreement with the findings of [17] and indicate that nucleation sites near the grafted ends of the probes are least accessible. Moreover, there is a sharp dielectric discontinuity at a GeneChip surface, with the dielectric constant in the siloxane layer being ~10, much lower than in the aqueous environment from where the target absorbs [15]. The energetics of duplex formation depend strongly on the dielectric constant and ionic strength of the immediate environment, and so there is expected to be a difference in hybridization between different ends of the probes [15].

Target–target interactions, target folding and probe folding

Suzuki *et al.* [28] hybridized target oligonucleotides in both the presence and absence of a complex background of cDNA produced from *E. coli* total RNA. As the spiked-in target concentration was raised, the signals with the background were smaller than without the background. This suggests that the background increases the amount of target–target interactions in solution and this acts to reduce the density of targets able to hybridize to probes [28]. This results in some probes showing saturated behaviour, not because there are no probes available to hybridize but because there are no targets available to hybridize. Several groups have modelled the competitive bulk-hybridization between targets in solution [8, 37–39], and show that this leads to improvements to the fits to spiked-in data.

Probe and target secondary structure affect hybridization kinetics [13]. The Vienna package [40] and mfold [41] can be used to calculate the self-folding energies of nucleic acids, which will depend upon the sequence composition of the fragments. Secondary structure in the targets expressed from *Brucella suis 1330* are likely pervasive [42], with a significant fraction of target found in double-stranded conformations at high temperature. Moreover, an analysis of GeneChips predicted that target folding may affect the signal from ~10% of GeneChip probes [43]. Furthermore, in several cases it has been verified that probes, which are expected to hybridize to targets that are expected to form stable folds do indeed have a low-intensity signal [43]. However, because the target fragments come in a range of sizes, there is likely a population of RNA target folds interacting with each probe [42]. Fifty nucleotides have been chosen as a representative size of the population of transcripts hybridizing to the chip [44], whereas [43] postulated that the most efficient hybridization will occur for the smallest fragments. Hybridization of duplexes occurs through a nucleation event followed by zippering and so [44]

looked for runs of four or five contiguous unpaired bases, which could result in the formation of a nucleus. It was concluded that folding may not be a critical factor in explaining hybridization intensities, but [44] acknowledged that their treatment was simplistic with respect to a full analysis, which requires a study of how partially folded targets hybridize. Moreover, the analysis of [44] ignores the problem of getting the RNA hairpin to the surface. Furthermore, an RNA–RNA interaction is stronger than a RNA–DNA interaction [43], and so it may become energetically unfavourable for the unzippering of RNA to occur so as to enable the zippering of the DNA-RNA duplex, contrary to the model of [44].

Probes are dynamic and may undergo folding if they contain regions that are self-complementary. It is difficult to model this accurately on GeneChips, because probes are tethered, and also in close proximity to other probes. This results in steric hindrances and external forces that are usually missing from algorithms to calculate minimum free-energies. In spite of these complications, inclusion of energy terms to simulate probe folding leads to better fits to spike-in experiments [39]. However, an attempt to include RNA folding [45] into the background calculation of a model of hybridization [46] resulted in only a slight (∼3%) change to the predicted intensities.

Probes chemically saturate and undergo differential desorption in the washing stage

GeneChips can undergo chemical saturation [47]. Early models [48] of GeneChip saturation used the classic thermal equilibrium Langmuir adsorbtion isotherm [49], in which there is assumed to be a single-binding energy regulating hybridization. However, [4] reports that the hybridization isotherms can be fit better using a Sips model [50], a generalization of the Langmuir model that allows multiple-binding energies during adsorption. The Sips model makes more physical sense as GeneChips contain a range of interactions involving targets and probes. Peterson *et al.* [17] also find that binding isotherms for mismatched targets can only be fit by assuming a Sips model, whereas [6], in contrast, argue that their fits to experimental data provide little evidence in favour of the complex Sips model over the simple Langmuir model. Moreover, as mentioned previously, the effects of bulk hybridization [8, 37–39] also provide isotherms in good agreement with the spike-in data.

According to the theory of Langmuir adsorption isotherms, probes within a GeneChip should saturate at the same intensity level. However, several groups have noted that saturation at the same concentration does not happen in practice, e.g. [7], and the asymptotic signature at high concentrations is sometimes lower for a mismatch feature than for the PM partner [6]. The difference in saturation intensity is not surprising if sequences have differing numbers of fluorescent markers, but the discrepancies go beyond this. Models [6, 44] suggest that the desorption of targets during the washing stage explains the differences in saturating intensities. Experimental support for this interpretation has come from [7], who explored the effects of the stringent wash, target concentration and hybridization time on the final microarray signal. A further complication to interpreting the results after washing has been identified by [51], who report several examples in which specific target duplexes dissociate faster than some non-specific duplexes, with temperature and duplex sequence affecting the relative dissociation rates of PM and MM duplexes.

AN OVERVIEW OF BIOINFORMATICS TECHNIQUES APPLIED TO GENECHIPS

The annotation of probes

The name of each GeneChip is associated with the relevant build of UniGene [52] used for probe selection, e.g. the Human GeneChip HG-U133A is based on build 133 of UniGene. The Affymetrix probe design uses sequences from databases such as UniGene, which are then clustered into groups, which represent distinctive transcripts. Either a consensus sequence, assembled from all the member sequences from a cluster, or an exemplar sequence of one of the members is then used as the source from which probes are chosen. The 'target' sequence is contained within the exemplar/consensus sequence and runs from the first base of the most 5′ probe and ends with the last base of the most 3′ probe.

The Affymetrix protocol [53] uses an oligo(dT) primer, complementary to a poly-Adenylation tail, for reverse transcription. At each step during reverse transcription, the process may stop and produce no more transcript, which means that the 3′ end

occurs most often in the resulting population of target sequences. Probes are therefore chosen to be towards the 3′ end of the sequence. They are usually selected to be no more than 600 bases upstream of a poly-Adenylation site.

The mapping from probes into biology is crucial. A Chip Definition File (CDF) contains information about which probes should be associated together into biologically meaningful probe sets. A significant factor in microarray analysis is the procedure by which the multiple probe values within a probeset are condensed together into an expression measure [54]. Affymetrix annotate their GeneChips by aligning the probes to databases of genomic and transcript sequences [55]. Affymetrix claim that over 70% of the probesets on the latest Human and Mouse GeneChips have at least 9 of the 11 probes in a probe set matching perfectly to a transcript. This leaves many probes that do not align to the transcripts for which they were designed. The need to identify and remove such spurious probes has led to considerable effort by the community to generate alternative CDFs, e.g. [56]. Moreover, many of the databases contain information which changes over time and probeset definitions should change so as to follow the most reliable knowledge [55, 56]. The updated probeset definitions provide better precision and accuracy compared to the original Affymetrix definitions [57].

Probes that map to different exons may show differential regulation, due to alternative splicing [58], and therefore should not be treated as measuring a single transcript. Mapping probes to transcripts, instead of genes, minimizes the effects of having multiple transcripts per gene [59]. Furthermore, mappings of probes to exon/intron structure provide a clearer view of what each probeset is measuring [60].

A further complication to the interpretation of GeneChip data is the potential of hybridization between a probe and transcripts from several genes. Indeed, there are many probes that map to multiple transcripts containing the full 25 contiguous bases [61]. Both [61] and [62] have argued that this multiple targeting results in correlations between the expression values for each of the separate probesets, even though there may be little biological significance in the correlation. However, [63] disagreed, suggesting instead that many of the large-scale correlations seen in microarray data are due to biological causes.

A number of probes have intensities that are correlated with the concentrations of the spiked-in transcripts in the Affymetrix Latin-Square experiment [33]. The correlation is likely due to those probes cross-hybridizing with one of the spiked-in transcripts, with small overlaps in sequences corresponding to runs of 10–16 nt responsible for the cross-hybridization [33]. Unfortunately, because there are only a small number of spiked-in transcripts, it is only possible to use the Latin-Square experiment to explore a relative limited amount of this type of cross-hybridization.

Models and algorithms used to derive expression measures

The development of algorithms to calibrate GeneChip data is an active field. A full calibration includes normalization, correcting for a fluorescent background, correcting for a background from non-specific binding and condensing the multiple probe intensities into an expression measure, a single measure of gene activity. The biggest differences in the sets of genes reported to be differentially expressed results from how the expression measure is calculated [54]. A number of algorithms that calculate expression measures have been benchmarked through the use of a web tool [64, 65], based around analysing the Affymetrix Spike-In Latin-Square data. The benchmarking shows that removing the signal from non-specific binding has the largest impact on performance [65].

All Affymetrix chips of a given design are created almost equal, and so the intensities across several experiments can be modelled by assuming that they result from both how sticky a probe is, and how much transcript was dropped onto the chip in each experiment. A collection of chips can be used together to estimate the probe affinities, and popular algorithms using this approach include dChip [66] and RMA [67]. The RMA algorithm assumes that all the PM signal is related to expression of the genes and thus ignores the effects of non-specific binding. However, each probe detects not only transcripts for which it is designed, but also RNA with similar sequences to these transcripts. Additionally, the general ‘stickiness’ of a probe results in it hybridizing non-specifically to the genomic background. Correspondingly, there is a non-linear response between RNA concentration and the fluorescent signature [47], with a background apparent at the lowest concentrations and saturation occurring

at the highest concentrations. In both the low-concentration and the high-concentration regimes, fold-changes in the raw light intensity (measured) do not transform into fold-changes in RNA concentration (of interest).

The original goal of the mismatch probe was to measure background. Some of the earliest releases of the Affymetrix Microarray Analysis Suite software (versions 4 and below) calculated an expression measure using an 'Average Difference', taking the mean of the PM–MM values for a given probeset. However, this was shown to be a poor measure of expression level [66] and so the average difference approach was modified in MAS5 [68]. The replacement, called Signal, calculates the Tukey bi-weighted mean of the logs of PM-MM. However, around 30% of probe pairs have MM greater than PM [69], which means that logs cannot be taken. For these cases, Signal defines an adjusted value for the MM so as to ensure that the log is calculated for a positive value.

Given the prevalence of MM values that exceed their PM counterparts, many authors [67, 70–72] regard the MM value as being an unreliable measure of the background. One reason is that the MM probe is likely to pick up a shadow of the gene-specific signal being detected by PM. The size of the central base is also a significant factor in determining the relative strengths of the PM and MM signal [46]. If the central base in the PM is a purine (G or A), then the corresponding position in the mismatch is a pyrimidine (C or T), and the position in the target RNA will also be a pyrimidine. Pyrimidines are relatively small and so there is little steric hindrance stopping hybridization between the mismatch probe and the target RNA. By contrast, if the central base in the PM is a pyrimidine, then the corresponding position in the mismatch and target will be a purine. As purines are relatively large, this means that there is considerable steric hindrance stopping hybridization between the mismatch probe and the target RNA. If the central base of the PM is a purine, then the ratio of PM/MM is smaller than if the central base of the PM is a pyrimidine.

The need to correct the background contribution to a probe's signal has led to alternative methods, which aim to model the effects of specific and non-specific hybridization based on probe sequences in conjunction with models of hybridization biophysics. The positional-dependent-nearest-neighbour model of hybridization [69] is based on the nearest-neighbour philosophy, but due to the effects of surface hybridization a different weight is assigned to each nucleotide position: a position close to the surface may have different properties to a position away from the surface. The signal for each probe was modelled as arising from a combination of gene-specific binding, non-specific binding of RNA fragments and a uniform fluorescent background signal. The model required fitting parameters for stacking energies for 16 sequence pairs (since, for example, the GC and CG sequences have different stacking energies), 24 positional weight parameters, parameters for the number of RNA molecules for each gene, the number of molecules of RNA contributing to the non-specific binding and the background signal. However, since there are $\sim 10^5$ probes on the array, there is little danger of overfitting [69]. A model of hybridization biophysics used for probe design by Affymetrix [73] does not use the nearest-neighbour approach, but instead fits a model in which the base type and position are central to binding stability. Their model also covers the unfavourable interactions of consecutive hairpins, non-consecutive hairpins and G-quartets associated with G-quadruplexes. The signal for each probe is modelled assuming that it results from a mixture of target-probe duplexes and a non-specific background, similar in form to that of [69]. Naef and Magnasco [46] developed a similar model to [69], neglecting nearest-neighbour effects and assuming that the affinity depends upon the base type and position along the probe. Yeast control RNA has been hybridized to a Human array by [74], enabling them to measure the non-specific background directly and to develop their algorithm GCRMA around a statistical model of this background measurement. GCRMA uses sequence information to describe the non-specific binding variation, favouring a model similar to that suggested by [46], which ignores nearest-neighbour information, over the nearest-neighbour model advocated by [69].

The model fits of [46, 69, 73, 74] all show that the affinity of a cytosine at a designated position in a probe is greater than having a guanine at this position, and a thymine is more sticky than an adenine. It was initially suggested that the asymmetry is related to the labelling of biotin to pyrimidines during the preparation of GeneChips, with the biotin label then interfering with the hybridization process [46]. However, it has been argued [75] that the asymmetry results from RNA being hybridized

to DNA on GeneChips, and thus it is not surprising that an A–U binding is different to a T–A binding. Similarly, there is a difference between C (DNA) binding to G (RNA) compared with G (DNA) binding to C (RNA). Naef *et al.* [76] have since agreed with [75] that this is the most likely cause of the asymmetry. However, [76] also provide evidence that modifying the protocol from labelling both pyrimidines, C and U, to just labelling C, results in substantial differences in the PM–MM distributions, consistent with their earlier suggestions [46].

The affinity of a probe can be estimated from its sequence composition [46, 69, 73, 74]. The resulting affinity is a composite of how strongly the target hybridizes to the probe as well as how much of the hybridized target avoids being dissociated during the washing phase [44], and both of these terms are described in terms of free energy [6]. The sum of the light produced from targets bound to any probe on a GeneChip will consist of large numbers of fragments that do not fully hybridize to the probe, yet avoid being removed in the washing stages. Models [8, 35] indicate that many hybrids are partially zipped up and that targets are most likely to be unbound towards the ends of the probes. The existence of such a fall-off in hybridization fraction may provide the mechanism for why models of hybridization such as [69] have their fitted weights of relative affinity falling off towards the end of the probe.

USING LARGE SURVEYS OF GENECHIPS TO HELP INTEGRATE BIOPHYSICS AND BIOINFORMATICS OF THE TECHNOLOGY

The popularity of GeneChips has led to thousands of publications and large, and rapidly growing, repositories of microarray data, such as the Gene Expression Omnibus (GEO [77]). This provides a glut of data available to analyse. The surveys of GeneChip data enable the identification of probes that are behaving in an unexpected manner. Biophysical processes are responsible for some of this behaviour and so identifying a common factor behind why groups of probes are behaving unexpectedly will help to shed light on these processes. Moreover, identifying and removing those probes that do not reliably measure the same thing as other probes measuring a gene will result in a more precise estimate of gene expression. Exons are believed to behave as 'atoms' of transcription, being either included or not, within a transcript. Thus looking at groups of probes that map to the same exon, and only this one exon, provides good biological controls. Each of these probes should show high correlation with the other members of the group as they are all measuring the same entity. Probe-pairs showing poor correlation are therefore indicative of unexpected behaviour in at least one of the probes.

We have begun to study the correlations between probes that map to the same exon [34]. We downloaded almost 40 000 Affymetrix GeneChip CEL files from GEO, normalized the data, transformed all intensities onto a log scale and then correlated these signals across all examples of a chip with a given design (e.g. HGU-133 Plus 2). All the correlations for a given exon are collated into a matrix, which is colour-coded according to the correlation value. All of the correlation matrices can be obtained from (http://bioinformatics.essex.ac.uk/users/wlangdon). Figure 4 shows an example of one such correlation matrix, for Ensembl exon ENSE00001121710. This exon is part of the SFRP1 gene (secreted frizzled-related protein 1), and the probes are derived from the 202036_s_at and 202037_s_at probe-sets. All of the Perfect-Match probes are closely correlated, except for Probe 7.

There is a large family, containing thousands of probes taken from thousands of probesets, which show large correlations across many experiments in GEO [34]. Probe 7 in ENSE00001121710 is a member of this family and so it is measuring something in common with many other probes (Figure 5). It is NOT measuring the same thing as the other probes for ENSE00001121710. Probe 7 has sequence AGGGGAGAGGCATTGCCTTCT CTGC, which contains a run of four contiguous guanines. The large family of correlated outlier probes all have a similar run of contiguous guanines [34]. This run was named the G-spot. For the HG-U133 Plus 2 design, there are more than 30 000 such probes (Table 1).

We find that the strength of the typical correlation between probes increases with the length of their G-spot. It is also dependent upon the location of the G-spots within the probes, with the largest average correlations resulting from probe pairs with G-spots at their 5′ ends, the end that is free. As an example, for the HG-U133A data, [34] show that about 80% of the more than 50 000 pairs of probes

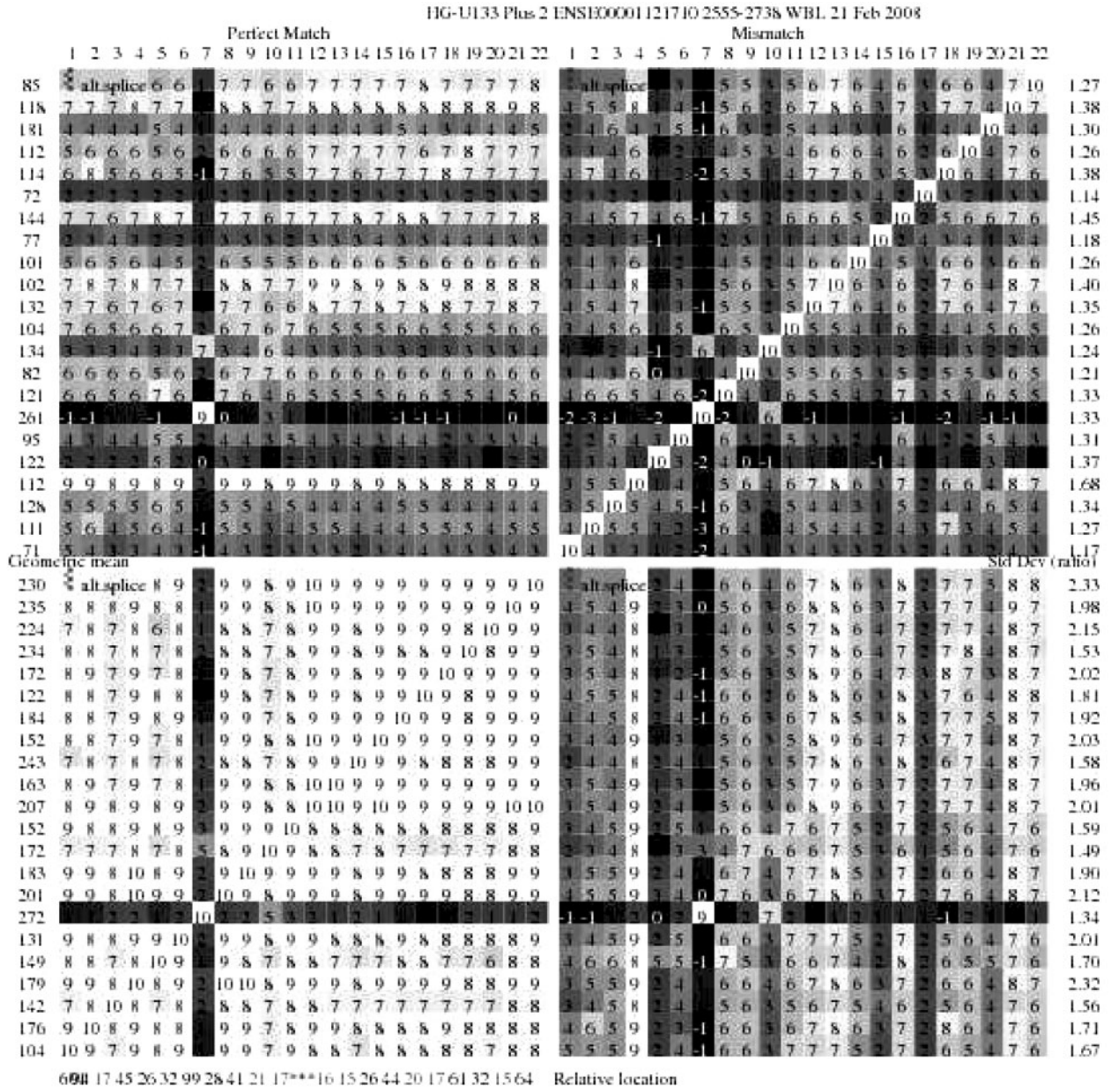

Figure 4: Correlation matrix for all the probes that map uniquely to the exon ENSE00001121710. The lower left quadrant represents the correlations between pairs of perfect match probes. The upper right quadrant represents the correlations between the pairs of mismatch probes. The top left quadrant represents the correlations between the perfect match and mismatch probes. The matrix is diagonally symmetric. The numbers in each of the matrix elements is 10× the correlation for that pair—hence along the diagonal the values are 10, which represent the perfect correlation between a probe and itself. The numbers to the left of each of the rows represent the geometric mean signal for each probe within GEO. The numbers to the right measures the standard deviation of the log intensities.

that have the G-spot at the free end have correlations exceeding 0.5. Pairs of probes with G-spots at the same distance from the 5′ end also typically show enhanced correlations over those pairs of probes, which differ in the location of their respective G-spots.

We suggest that the correlations in intensity for probes containing runs of guanine result from the formation of G-quadruplexes [34]. Each cell that contains one group of these probes will have large numbers of single-stranded oligonucleotides in close proximity, all running in parallel and all containing a run of guanines. This provides an ideal situation for four probes to associate and form a G-quadruplex. The formation of G-quadruplexes will result in the probes involved all having their bases pointing

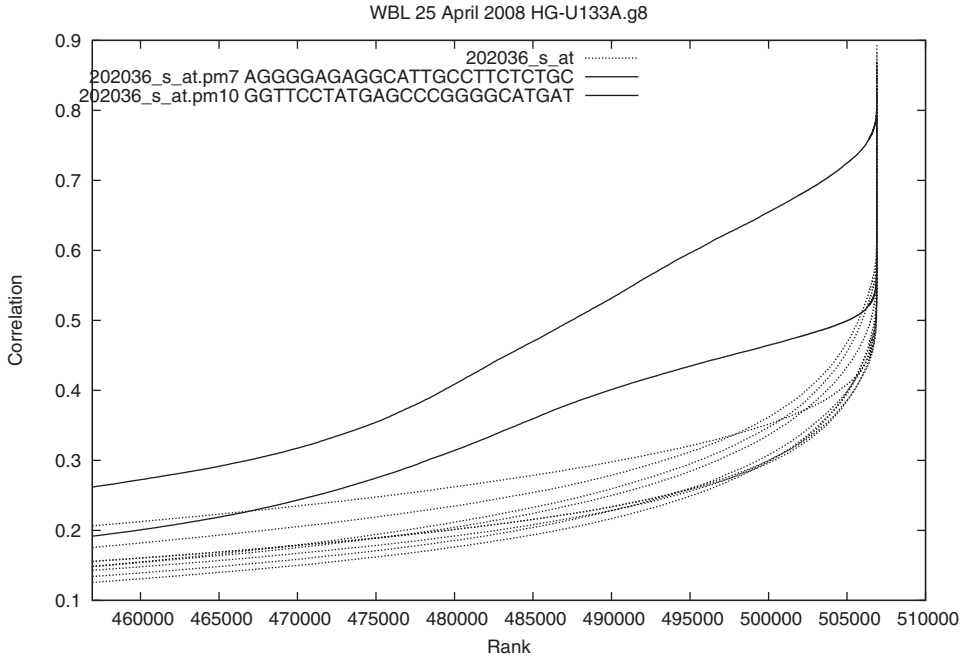

Figure 5: The correlations in the logs of intensities between the probes in probe-set 202036_s_at and all the other probes in thousands of GeneChips. Probes containing runs of four or more contiguous guanines are found to correlate highly with thousands of other probes.

Table 1: The number of probes containing a contiguous run of four or more bases

Chip ID	Species	Target	GGGG	CCCC	AAAA	TTTT
HG-U133 Plus2	Homo sapiens	RNA	32547	32256	32474	72250
ATH-121501	Arabidopsis	RNA	6680	5188	9939	21118
Drosophila-2	Drosophila	RNA	64	7974	15057	25237
Exon 1.0 ST	Homo sapiens	RNA	199985	310884	379623	388654
SNP 6.0	Homo sapiens	DNA	3588	291821	874472	986073

inwards within the quadruplex, and thus these cannot hybridize to the target. However, probe density plays a significant role in regulating the rate of duplex formation, and as probes interact more strongly the formation of nucleation sites available is modified [19]. The formation of the quadruplex will free up space in its immediate environment thereby enabling an increase in the kinetics and strength of nucleation and hybridization to other probes spatially adjacent to the quadruplex [34] (Figure 6).

THE IMPLICATION OF THE FAMILY OF PROBES WITH RUNS OF CONTIGUOUS GUANINES SHOWING CORRELATED BEHAVIOUR

We see thousands of probes, taken from many biologically unrelated genes, undergoing correlated changes in intensity across thousands of experiments in GEO. These correlated probes all have a run of guanines in common, and we associate the source of the correlation with the formation of G-quadruplexes resulting from probe–probe interactions on the surface of GeneChips. The discovery of this family has several implications for analysing the data from GeneChips.

Although the run of contiguous guanines makes up a small fraction of the 25 bases within these probes, fragments of RNA will hybridize particularly efficiently to this region in probes, which are not bound up in G-quadruplexes—the reduced probe density in the immediate lateral environment of the quadruplex acts to both increase the rate of hybridization and also increase the strength of the hybrid. The stability of nucleation and partial zippering around the small number of bases in the G-spot is greater than the binding to other

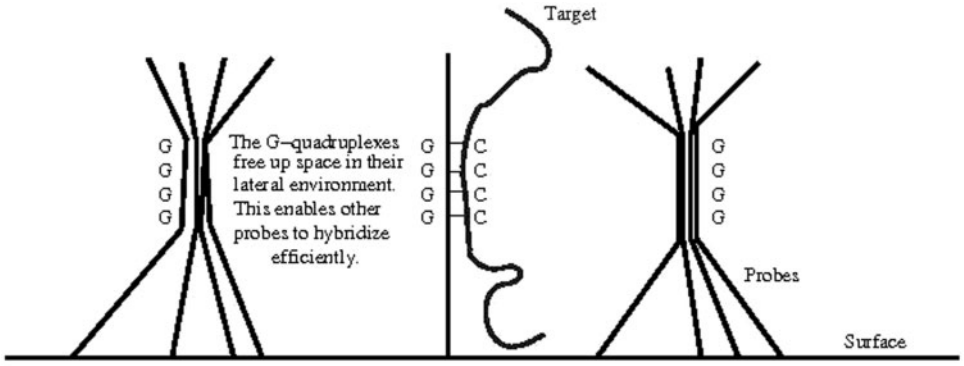

Figure 6: If groups of probes form G-quadruplexes, the remaining probes are able to hybridize efficiently, because their runs of guanines are in a region of low probe density.

nucleation sites for competing targets on the probes. This means that it is difficult for other oligonucleotides to displace the bound duplex in the G-spot region, even though these other oligonucleotides can result in a larger sequence of overlap when more fully hybridized. This will lead to longer times to reach equilibrium for these probes.

As many target transcripts can nucleate in the G-spot region, probes containing a G-spot are particularly prone to cross-hybridization. This has implications for algorithms attempting to establish which probes undergo cross-hybridization. Previously, bioinformatics studies of multiple-targeting have concentrated on large overlaps between sequences [55, 56]. However, there is evidence for 10–16 nucleotide sequence overlaps that were sufficient to result in cross-hybridization between spike-in transcripts and probes [33]. The evidence from the G-spot probes pushes the size limit down even further, because we find correlations between G-spot probes occurring when there is only an overlap of four guanines in common. These findings have implications for modelling the stringency of the washing of GeneChips, as the strength of the duplex in the G-spot region is clearly sufficient to last through the washing stages.

The existence of the correlations in expression, across many experiments in GEO, suggests that there is something in these experiments which causes thousands of G-spot probes to change together. We associate these changes in signal with the formation of G-quadruplexes in G-spot probes, and so there must be something that is causing the stability of G-quadruplexes to change from experiment to experiment. Within a G-quadruplex the ionic radius of the central cation determines how effectively the tetrads are stabilized [78] and for alkali metal cations the affinity order is $K^+ \gg Na^+ > Rb^+ > Cs^+ \gg Li^+$ and for the earth alkali the order is $Sr^{2+} \gg Ba^{2+} > Ca^{2+} > Mg^{2+}$. On the other hand, bivalent transition metal cations such as Mn^{2+}, Co^{2+} and Ni^{2+} destabilize quadruplexes containing potassium [24]; this is likely to be due to nucleophilic atoms within the guanines coordinating to the cation, preventing formation of the Hoogsteen hydrogen bonds that stabilize the quadruplex [78]. Molecular crowding also helps to induce quadruplex formation [79]. Furthermore, ethanol has recently been shown to be a better inducer of quadruplexes than even potassium cations [80], possibly due to three effects: molecules of ethanol bind at sites in the tetraplex to stablize it; the change in dielectric constant caused by ethanol diminishes the repulsion between phosphate chains; ethanol enhances the affinity for cations. We note that changes in concentration of different cations, pH and even ethanol concentration will show subtle variations from experiment to experiment due to random and, quite likely, small differences in the use of the protocol when running the GeneChips. Another feature that might affect the chip-to-chip variation in the extent of quadruplex formation is the life-history of the chip prior to being run in the experiment. A chip in a cold and dark environment for a long time may well form lots of G-quadruplexes on its surface, whereas a chip that is heated strongly immediately prior to being used is expected to have fewer G-quadruplexes.

A further implication of our findings is there may be other families of probes, which undergo probe-probe interactions in a coherent manner. Runs of Gs result in a quadruplex with the greatest

stability [81]. However, C-tetrads [82], T-tetrads [83] and A-tetrads [84] have been seen within G-quadruplexes. A number of motifs are therefore able to result in stacks of tetrads, including the CGG repeats involved in fragile-X syndrome [85]. Another type of quadruplex is the I-motif structure [86], which results from two parallel duplexes containing runs of cytosine and protonated cytosine, inserted head to tail. This motif may be possible on a GeneChip but will require pairs of probe duplexes to bend back towards each other. Furthermore, groups of three to seven strands are now known to result in complexes [87], including triplet-forming complexes containing two bases the same and one complementary base. Such triplets may be relevant for describing the hybridization between a probe, a strand of target RNA and an adjacent probe. If quadruplexes and other probe–probe structures exist, then these may be observable through techniques such as scanning probe microscopy (SPM) [88]. Solved structures of quadruplexes [82] indicate that the quadruplex helices take up one chirality, either left- or right-handed, and so a given probe quadruplex will take up a helicity dependent upon the bases causing the helix. The chirality should be evident from SPM images, as will the number of tethered probes forming the helix.

Wu *et al.* [26] have already identified that probes which contain runs of contiguous guanines show abnormal affinities. It is possible to remedy the affinity models to include terms for the presence of quadruplexes, as suggested by [73]. However, [73] used a fixed free-energy change to describe the impact of a quadruplex that will not vary from experiment to experiment, and so the affinities will also not change from experiment to experiment. This is in agreement with the proposed biophysics behind many expression measures, which assume the affinity is constant for all experiments, and that changes in intensity for a probe are due to changes in RNA concentrations. However, we find that most of the signal from probes containing runs of guanines results from cross-hybridization and shows large changes across the many experiments in GEO. When there are lots of G-quadruplexes, the remaining adjacent probes will be particularly effective at hybridizing, because of the reduced surface density of probes in the immediate environment of the adjacent probes. When there are few or no G-quadruplexes then the probes will be less effective at hybridizing because of the increased surface density of probes. This demonstrates that the affinity for non-specific hybridization cannot be treated as a constant for these probes. We have shown there are several viable mechanisms within the Affymetrix protocol that could induce correlated G-quadruplex formation. Furthermore, this correlation would be across the chip rather than within individual probe-sets.

G-spot probes are usually highly correlated and so when one of the probes has a high intensity, it is likely that other G-spot probes will similarly have high intensities. If a G-spot probe has a high value, and other probes in the probe set have high values (because the gene is well expressed), then the G-spot probe will not be excluded in calculations of overall gene expression. Moreover, and crucially, it will act to alter the detection of outliers within the expression measure calculations. The misleading G-spot values will be those that appear to tentatively support the values seen by others in their probe set, even though this support is coincidental [34]. We therefore advise that G-spot probes should not be included in the calculations of expression. This can be done efficiently through modifying the CDFs that result from increasingly sophisticated bioinformatics pipelines, e.g. [56].

As an illustration of the potential benefit of using a revised CDF file with G-spot probes omitted, we present Figure 7. We have used the affycomp facility [65] to benchmark the RMA procedure omitting background correction (since this appears to be the best of the RMA family of corrections) against the known changes expected from the Latin-Square spike-in experiment provided by Affymetrix. We generated customized CDFs after removing probes containing either 4Gs, 5Gs, the same number of probes as the 4G set but chosen randomly ignoring their sequences, and a similar set of random probes but with same population size as the 5G probes. The experiment used the HG-U133A array for which there are 16 744 probes containing the GGGG sequence. Of these, 3538 contain the longer GGGGG sequence. Masking probes implies losing information, which would diminish performance. To properly assess the advantage of masking the G-spot probes, we therefore superimpose curves that plot the cumulating numbers of false and true positives. The uppermost curve shows that the best result results from removing the GGGGG probes. This result is superior (for these data), because it involves the removal of fewer probes—it should

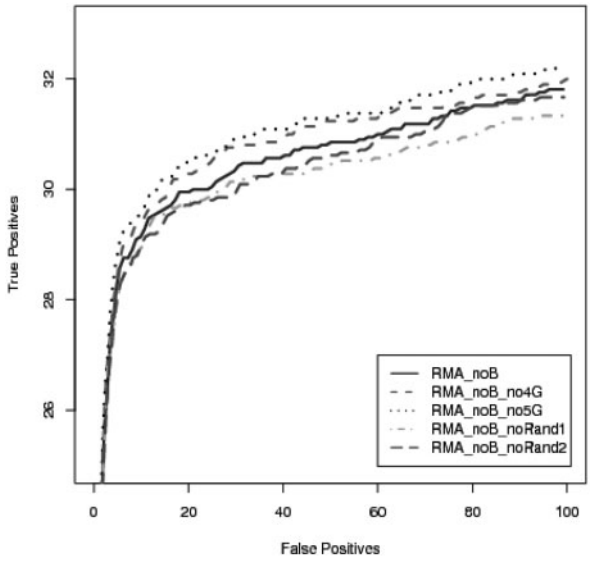

Figure 7: The impact of G-spots on the ability of RMA to detect True-Positives and to discriminate False-Positives, from changes in the Latin-Square spike-in experiment of Affymetrix, using the analysis tools within Affycomp [65]. We have chosen RMA without background subtraction, as this appears to be more accurate than RMA with background subtraction [65]. The several lines result from: using RMA with all probes available; RMA_no4G (excluding 16744 probes—~6.8% of total number of probes); RMA_no5G (excluding 3538 probes—~1.5% of total number of probes); RMA_no Random1 (excluding a random sample of 16744 probes); RMA_noRandom2 (excluding a random sample of 3538 probes).

be compared with the curve marked 'Rand2' in the diagram, since this shows the result of removing an equivalent number of probes at random. Comparison of the curves relating to the removal of the GGGG probes and an equivalent number (the 'Rand1' curve) of other probes shows the greater improvement that results from removal of the GGGG probes is offset to an extent by the general loss of information. Nevertheless, removal is an improvement on no action and the graph gives an idea of the improvement.

Table 1 shows the number of probes containing contiguous runs of four or more guanines (and cytosine, thymine and adenine for comparison) in several GeneChip designs. The design of Human and Arabidopsis 3′ GeneChips resulted in orders of magnitude more probes that have the potential to form G-quadruplexes than did the design for Drosophila. Moreover, the Human exon array contains a significant number of probes containing runs of guanine. Their existence is likely to confound the analysis of this type of data, as there are only four probes per exon. Whereas the genotyping array has only a very small of probes containing guanine runs.

We are not aware of studies that have identified the role of G-quadruplexes in modifying the data from post-genomic technologies, other than in microarrays, which we have described here. However, we expect that the formation of G-quadruplexes, and likely other structures, will occur in any post-genomic experiment that utilizes single-stranded nucleic acids of the same sequence being held in close proximity.

SUMMARY

The study of Affymetrix GeneChips is an active field, bringing together many disciplines ranging from physics to genomics. However, the breadth of the science makes it difficult to keep abreast of the developments in each of these fields. We hope this review goes some way towards bringing disparate information to light. We also hope that this paper will help users of GeneChips and also the scientists in the machine-learning and systems-biology communities, who wish to better understand how calibration issues limit what can be inferred from GeneChip data.

GeneChips are very popular, and this popularity has led to many observations in the public domain. Mining of this data indicates that there are a range of systematic biases in the raw data, which can be traced to the biophysics of the technology. Probes on GeneChips are able to come into physical contact, which modifies the environment of neighbouring probes. Furthermore, the hybridization between probes and a heterogeneous population of transcripts in an RNA sample results in many different types of interactions on GeneChips. Informatics developments are also leading to better tools in which biological knowledge can be used to aid the analysis of the data. However, care still needs to be taken when averaging signals from multiple probes. A large family of probes, each containing a run of contiguous guanines, shows correlated expression across thousands of GeneChip experiments. We suggest the existence of this family is associated with the formation of G-quadruplexes on the surface of GeneChips. The correlations suggest that a small-sequence overlap, containing

the run of guanines, is sufficient for hybridization between a probe and a target, with the resulting hybrid being sufficiently stable that it avoids dissociation during the washing cycle.

Acknowledgements

We are grateful for discussions with Manolo Arteaga-Salas, Renata Camargo, Julian Dow, Neil Graham, Mike Hubank, Sean May, Joanna Rowsell, Olivia Sanchez-Graillet, Hugh Shanahan, Maria Stalteri and Shingo Suzuki. We are also grateful for all the users of GeneChips who have deposited their data in repositories such as GEO. We apologize for not being able to fully reference the large and rapidly expanding literature on the many aspects of GeneChip analysis in this review.

Key Points

- There has been considerable recent progress in understanding the molecular interactions occurring at the surfaces of GeneChips.
- Models of hybridization on GeneChips and data-mining of GeneChip surveys are beginning to converge on a consistent view.
- Probe–probe interactions may be prevalent on GeneChips, resulting in local changes to probe density, which feed into changes in hybridization efficiency.
- There exists a family of thousands of probes in many biologically unrelated genes, yet which show large correlations across GeneChip surveys. Probes in this family all contain a run of guanines, and they have abnormal affinities. It is possible that G-quadruplexes are forming on the surface of GeneChips.
- Partial hybrids of only four bases on some probes may be stable enough to survive the wash cycles.

FUNDING

BBSRC (BB/E001742/1) grant (to W.B.L.).

References

1. Affymetrix. The structure, function and applications of GeneChip microarrays. Technical report, 2005.
2. McGall G, Barone A, Diggelmann M, *et al*. The efficiency of light-directed synthesis of DNA arrays on glass substrates. *J Am Chem Soc* 1997;**119**:5081–90.
3. Pawloski A, McGall G, Kuimelis R, *et al*. Photolithographic synthesis of high-density DNA probe arrays: challenges and opportunities. *J Vacuum Sci Technol* 2007;**25**:2537–46.
4. Glazer M, Fidanza J, McGall G, *et al*. Kinetics of oligonucleotide hybridization to photolithographically patterned DNA arrays. *Anal Biochem* 2006;**358**:225–38.
5. Lemeshko S, Powdrill T, Belosludtsev Y, *et al*. Oligonucleotides form a duplex with non-helical properties on a positively charged surface. *Nucleic Acids Res* 2001;**29**:3051–8.
6. Burden C, Pittelkow Y, Wilson S. Adsorption models of hybridisation and post-hybridisation behaviour on oligonucleotide microarrays. *J Phys: Condens Matter* 2006;**18**:5545–65.
7. Skvortsov D, Abdueva D, Curtis C, *et al*. Explaining differences in saturation levels for Affymetrix GeneChips. *Nucleic Acids Res* 2007;**35**:4154.
8. Binder H. Thermodynamics of competitive surface adsorption on DNA microarrays. *J Phys: Condens Matter* 2006;**18**:S491–523.
9. Turner D. Conformational changes. In: Bloomfield V, Crothers D, Tinoco I (eds). *Nucleic Acids; Structures, Properties and Functions*. Sausalito, CA: University Science Books, 2000.
10. SantaLucia J. A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbour thermodynamics. *Proc Natl Acad Sci USA* 1998;**95**:1460–5.
11. Tan Z-J, Chen S-J. Nucleic acid helix stability: effects of salt concentration, cation valence and size, and chain length. *Biophys J* 2006;**90**:1175–90.
12. Levicky R, Horgan A. Physicochemical perspectives on DNA microarray and biosensor technologies. *Trends Biotechnol* 2005;**23**:143–9.
13. Gao Y, Wolf L, Georgiadis R. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Res* 2006;**34**:3370–7.
14. Peterson A, Heaton R, Georgiadis R. The effect of surface probe density on DNA hybridization. *Nucleic Acids Res* 2001;**29**:5163–8.
15. Forman J, Walton I, Stern D, *et al*. Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesised oligonucleotide arrays. In: Leontis NB, SantaLucia J (eds). *Molecular Modeling of Nucleic Acids* (ACS Symposium Series, Vol. 682). Washington, DC: American Chemical Society, 1998, 206–28.
16. Vainrub A, Pettitt B. Coulomb blockage of hybridization in two-dimensional DNA arrays. *Phys Rev E* 2002;**66**:041905.
17. Peterson A, Wolf L, Georgiadis R. Hybridization of mismatched or partially matched DNA at surfaces. *J Am Chem Soc* 2002;**124**:14601–7.
18. Halperin A, Buhot A, Zhulina E. Brush effects on DNA chips: thermodynamics, kinetics and design guidelines. *Biophys J* 2005;**89**:796–811.
19. Hagan M, Chakraborty A. Hybridization dynamics of surface immobilized DNA. *J Chem Phys* 2004;**120**:4958–68.
20. Wong KY, Pettitt BM. A study of DNA tethered to a surface by an all-atom molecular dynamics simulation. *Theor Chem Acc* 2001;**106**:233–5.
21. Podgornik R, Strey H, Gawrisch K, *et al*. Bond oriental order, molecular motion, and free energy of high-density DNA mesophases. *Proc Natl Acad Sci USA* 1996;**93**:4261–6.
22. Lescoute A, Westhof E. The interaction networks of structured RNAs. *Nucleic Acids Res* 2006;**34**:6587.
23. Burge S, Parkinson G, Hazel P, *et al*. Quadruplex DNA: sequence topology and structure. *Nucleic Acids Res* 2006;**34**:5402–15.
24. Blume S, Guarcello V, Zacharias W, *et al*. Divalent transition metal cations counteract potassium-induced quadruplex assembly of oligo(dG) sequences. *Nucleic Acids Res* 1997;**25**:617–25.

25. Mergny J-L, De Cian A, Ghelab A, *et al*. Kinetics of tetramolecular quadruplexes. *Nucleic Acids Res* 2005;**33**: 81–94.
26. Wu C, Zhao H, Baggerly K, *et al*. Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bioinformatics* 2007;**23**: 2566–72.
27. Walton S, Mindrinos M, Davis R. Analysis of hybridization of the molecular barcode GeneChip microarray. *Biochem Biophys Res Commun* 2006;**348**:689–96.
28. Suzuki S, Ono N, Furusawa C, *et al*. Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. *BMC Genom* 2007;**8**:373.
29. Bishop J, Wilson C, Chagovetz A, *et al*. Competitive displacement of DNA during surface hybridization. *Biophys J* 2007;**92**:L10–12.
30. Bishop J, Chagovetz A, Blair S. Kinetics of multiplex hybridization: mechanisms and implications. *Biophys J* 2008; **94**:1726–34.
31. Bhanot G, Louzoun Y, Zhu J, *et al*. The importance of thermodynamic equilibrium for high-throughput gene expression arrays. *Biophys J* 2003;**84**:124–35.
32. Bishop J, Blair S, Chagovetz A. A competitive kinetic model of nucleic acid surface hybridization in the presence of point mutants. *Biophys J* 2006;**90**:831–40.
33. Wu C, Carta R, Zhang L. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res* 2005;**33**:e84.
34. Upton G, Langdon W, Harrison A. G-spots cause incorrect expression measurement in Affymetrix microarrays. *BMC Genom* 2008;**9**:613.
35. Deutsch J, Liang S, Narayan O. Modeling of microarray data with zippering. Preprint, q-bio.BM/ 0406039.
36. Reynaldo L, Vologodskii A, Neri B, *et al*. The kinetics of oligonucleotide replacements. *J Mol Biol* 2000;**297**: 511–20.
37. Carlon E, Heim T. Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays. *Phys A* 2006;**362**:433–49.
38. Halperin A, Buhot A, Zhulina E. Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophys J* 2004;**86**:718–30.
39. Ono N, Suzuki S, Furusawa C, *et al*. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. *Bioinformatics* 2008;**24**:1278–85.
40. Hofacker I. Vienna RNA secondary structure server. *Nucleic Acids Res* 2003;**31**:3429–31.
41. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;**31**: 3406–15.
42. Ratushna V, Weller J, Gibas C. Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genom* 2005;**6**:31.
43. Heim T, Tranchevent L, Carlon E, *et al*. Physics-based analysis of Affymetrix microarray data. *J Phys Chem B* 2006; **110**:22786–95.
44. Held G, Grinstein G, Tu Y. Relationship between gene expression and observed intensities in DNA microarrays—a modelling study. *Nucleic Acids Res* 2006;**34**:e70.
45. Gharaibeh R, Fodor A, Gibas C. Software note: using probe second structure information to enhance Affymetrix GeneChip background estimates. *Comput Biol Chem* 2007; **31**:92–8.
46. Naef F, Magnasco M. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E* 2003;**68**:011906.
47. Chudin E, Walker R, Kosaka A, *et al*. Assessment of the relationship between signal intensities and transcript concentrations for Affymetrix GeneChips. *Genome Biol* 2001;**3**:research0005.1–0005.10.
48. Hekstra D, Taussig A, Magnasco M, *et al*. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res* 2003;**31**:1962–8.
49. Langmuir I. The constitution and fundamental properties of solids and liquids. *J Am Chem Soc* 1916;**38**:2221–95.
50. Sips R. On the structure of a catalyst surface. *J Chem Phys* 1948;**16**:490–5.
51. Pozhitkov A, Stedtfeld R, Hashsham S, *et al*. Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. *Nucleic Acids Res* 2007;**35**:e70.
52. Pontius J, Wagner L, Schuler D. UniGene: a unified view of the transcriptome. In: *The NCBI Handbook*. Bethesda, MD: National Centre for Biotechnology Information, 2003.
53. Affymetrix, The New GeneChip IVT Labeling Kit: optimized protocol for improved results, 2004, Part No. 701466 *Rev*. 21.
54. Harrison A, Johnston C, Orengo C. Establishing a major cause of discrepancy in the calibration of Affymetrix GeneChips. *BMC Bioinform* 2007;**8**:195.
55. Stalteri M, Harrison A. Comparisons of annotation predictions for Affymetrix GeneChips. *Appl Bioinform* 2006;**5**:237.
56. Dai M, Wang P, Boyd A, *et al*. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005;**33**:e175.
57. Sandberg R, Larsson O. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinform* 2007;**8**:48.
58. Stalteri M, Harrison A. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinform* 2007;**8**:13.
59. Lu J, Lee J, Salit M, *et al*. Transcript-based redefinition of grouped oligonucleotide probesets using AceView: high-resolution annotation for microarrays. *BMC Bioinform* 2007;**8**:109.
60. Rambaldi D, Felice B, Praz V, *et al*. Splicy: a web-based tool for the prediction of possible alternative splicing events from Affymetrix probeset data. *BMC Bioinform* 2007; **8**(Suppl. 1):S17.
61. Okoniewski M, Miller C. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinform* 2006;**7**:276.
62. Casneuf T, Van de Peer Y, Huber W. In situ análisis of cross-hybridisation on microarrays and the inference of expresión correlation. *BMC Bioinform* 2007;**8**:461.
63. Klebanov L, Chen L, Yakovlev A. Revisiting adverse effects of cross-hybridization in Affymetrix gene expression data: do they matter for correlation analysis? *Biol Direct* 2007;**2**:28.

64. Cope L, Irizarry R, Jaffee H, *et al*. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 2004;**20**:323–31.

65. Irizarry R, Wu Z, Jaffee H. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 2006;**22**: 789–94.

66. Li C, Wong W. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* 2001;**98**:31–6.

67. Irizarry RA, Hobbs B, Collin F, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;**4**:249–64.

68. Affymetrix. Microarray Suit Users Guide, 5th edn. Santa Clara, CA: Affymetrix, 2001.

69. Zhang L, Miles M, Aldape K. A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol* 2003;**21**:818–21.

70. Naef F, Lim DA, Patil N, *et al*. From features to expression: high-density oligonucleotide array analysis revisited. *Genome Biol* 2002;**3**:RESEARCH0018.

71. Sásik R, Calvo E, Corbeil J. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics* 2002;**18**:1633–40.

72. Irizarry R, Bolstad B, Collin F, *et al*. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;**31**:e15.

73. Mei R, Hubbell E, Bekiranov S, *et al*. Probe selection for high-density oligonucleotide arrays. *Proc Natl Acad Sci USA* 2003;**100**:11237.

74. Wu Z, Irizarry R, Gentleman R, *et al*. A model-based background adjustment for oligonucleotide expresión arrays. *J Am Stat Assoc* 2004;**99**:909–17.

75. Carlon E, Heim T, Wolterink J, *et al*. Comment on Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E* 2006;**73**: 063901.

76. Naef F, Wijnen H, Magnasco M. Reply to comment on solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E* 2006;**73**:063902.

77. Barrett T, Suzek T, Troup D, *et al*. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res* 2005;**33**(Database issue):D562–6.

78. Galezowska E, Gluszynska A, Juskowiak B. Luminescence study of G-quadruplex formation in the presence of Tb^{3+} ion. *J Inorg Biochem* 2007;**101**:678–85.

79. Kan Z-Y, Yao Y, Wang P, *et al*. Molecular crowding induces telomere G-quadruplex formation under salt-deficient conditions and enhances its competition with duplex formation. *Angew Chem Int Ed* 2006;**45**:1629.

80. Vorlíčková M, Bednářová K, Kypr J. Ethanol is a better inducer of DNA guanine tetraplexes than potassium cations. *Biopolymers* 2006;**82**:253–60.

81. Gros J, Rosu F, Amrane S, *et al*. Guanines are a quartet's best friend: impact of base substitutions on the kinetics and stability of tetramolecular quadruplexes. *Nucleic Acids Res* 2007;**35**:3064–75.

82. Patel P, Bhavesh N, Hosur R. NMR observation of a novel C-tetrad in the structure of SV40 repeat sequence GGGCGG. *Biochem Biophys Res Commun* 2000;**270**:967–71.

83. Patel P, Hosur R. NMR observations of T-tetrads in a parallel stranded DNA quadruplex formed by *Saccharomyces cerevisiae* telomere repeats. *Nucleic Acids Res* 1999;**27**: 2457–64.

84. Patel P, Koti A, Hosur R. NMR studies on truncated sequences of human telomeric DNA: observations of a novel A-tetrad. *Nucleic Acids Res* 1999;**27**:3836–43.

85. Fry M, Loeb L. The fragile X syndrome $d(CGG)_n$ nucleotide repeats form a stable tetrahelical structure. *Proc Natl Acad Sci USA* 1994;**91**:4950.

86. Malliavin T, Gau J, Snoussi K, *et al*. Stability of the I-motif structure is related to the interactions between phosphodiester backbones. *Biophys J* 2003;**84**:3838–47.

87. Sühnel J. Beyond nucleic acid base pairs: from triads to heptads. *Biopolymers (Nucleic Acid Sequences)* 2002;**61**:32–51.

88. Meunier V, Kalinin S, Lambin Ph. Theory of scanning probe microscopy of carbon nanostructures. In: Kalinin S, Goldberg B, Eng L, Huey D (eds). *Scanning-probe and Other Novel Microscopies of Local Phenomena in Nanostructured Materials*. Warrendale, PA: Mater. Res. Soc. Symp. Proc. 838E, 2005, 012.1.1.