Assessing the Performance of Different High-density Tiling Microarray Strategies for Mapping Transcribed Regions of the Human Genome

Olof Emanuelsson[1], Ugrappa Nagalakshmi[2], Deyou Zheng[1], Joel S. Rozowsky[1], Jiang Du[3], Zheng Lian[4], Alexander E. Urban[2], Viktor Stolc[5], Sherman Weissman[4], Michael Snyder[1,2,*], Mark Gerstein[1,3,*]

[1] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven CT 06520-8114, USA

[2] Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven CT 06520-8103, USA

[3] Department of Computer Science, Yale University, New Haven, CT 06520-8285, USA

[4] Department of Genetics, Yale University School of Medicine, New Haven, CT 06520–8005, USA

[5] Center for Nanotechnology, NASA Ames Research Center, Moffett Field, CA 94035, USA

[*] To whom correspondence should be addressed: E-mail: michael.snyder@yale.edu (M.S.), mark.gerstein@yale.edu (M.G.)

Running title: High-density Tiling Microarray Strategies

ABSTRACT

Genomic tiling microarrays have recently become a popular approach for interrogating the transcriptional activity of large regions of the genome in an unbiased fashion. There are a number of key variables affecting the behavior of these arrays: probe length, mismatch probes, the experimental protocol, the number of replicates, and the overall genomic tiling density. Here we assess the role of these variables as they are manifest in a number of different platforms. First, we analyze the degree to which the transcription measured in several published tiling-array experiments agrees with established gene annotation on human chromosome 22. We observe that the transcription from very high-density tiling arrays correlates substantially better with annotation than that from other array types. Next, we perform an in-depth analysis of the transcription mapping performance in the ENCODE region of the human genome using the two main high-density array platforms. To best enable this comparison, we hybridize identical biological samples to the arrays. We then develop and evaluate a number of ways of scoring the arrays and segmenting the genome sequence into transcribed and non-transcribed regions, with the aim of making the platforms most comparable to each other. Finally, we develop a comparison approach based on looking for agreement with known annotation. Overall, we find that the most important variable in the experiment is simply the raw number of counts: higher density of the oligonucleotides on the array and the use of more array replicates, which enables greater statistical resolution in the scoring process, are crucial for the performance in terms of agreement with known annotation. In concordance with previous studies, our experiments reveal a significant amount of novel transcription (outside of known genes), often supported by both platforms. We validate experimentally a subset of the novel transcribed regions and find that a majority of the assayed regions from both technologies are, in fact, transcribed. [[SHOULD WE SAY AFFY IS BETTER?]]

NOTES:

Gene Expression Omnibus (GEO) accession numbers are provided for the array data. Supplementary material is provided.

INTRODUCTION

Mapping transcribed regions of the human genome in an unbiased fashion is a crucial step towards understanding at a molecular level the organization of hereditary information and the specific functions of each human cell or tissue type. To this end, a number of approaches using genomic tiling microarrays have been tested and published over the last few years (Kapranov et al., 2002; Rinn et al., 2003; Schadt et al., 2004; Bertone et al., 2004; Cheng et al., 2005). While the strategies differ substantially in most of their details, they all share a basic array design idea: to construct an array whose probes (the molecules attached to the microarray at the manufacturing) cover all of the non-repetitive sequence of the genome or genomic region under investigation.

There are five main papers describing tiling microarray-based transcript mapping of regions or the whole of the human genome, employing four different microarray platforms that all have proven valuable for mapping transcribed regions.

(i) Kapranov et al. (2002) used a high density oligonucleotide array design containing perfect match probes of length 25 bp and corresponding mismatch probes with a single mismatch to the human genome sequence. The arrays were synthesized *in situ* (directly on the supporting array material) using physical masks. In Kapranov et al., the arrays covered chromosomes 21 and 22 with probe starting positions spaced every 35 bp (genomic distance), and were hybridized with samples representing 11 cell lines. The data was later reanalyzed (Kampa et al., 2003) and a more sophisticated approach to genomic segmentation was introduced. We refer to this setup as the Affymetrix (or Affy, for short) tiling array platform.

(ii) Rinn et al. (2003) mapped transcribed regions of chromosome 22 with an array of PCR products (amplicons), tiled basically end-to-end with an amplicon (probe) size range of 300-1,400 bp. This array represents the PCR tiling array platform and was hybridized with placenta poly(A)+ RNA (Rinn et al., 2003) and later with RNA from two cell lines (White et al., 2004).

(iii) Schadt et al. (2004) used tiling arrays where the probes were phosphoramitide synthesized and attached to the array by the Agilent ink-jet technology (Shoemaker et al., 2001). They tiled the entire non-repetitive parts of chromosomes 20 and 22 with 60-mers spaced on average every 30 bp.

(iv) Bertone et al. (2004) used oligonucleotide microarrays with lower densities and longer oligonucleotides (36 bp, spaced every 46 bp) than Kapranov et al. to map transcribed regions of the whole non-repetitive part of the human genome. The arrays lack mismatch probes and are synthesized *in situ* using maskless technologies developed by NimbleGen Systems. This approach has also been used to map transcribed regions of rice (Li et al., 2005) , fruit fly (Stolc et al., 2004), and arabidopsis (Stolc et al., 2005). We refer to this as the MAS (maskless array synthesis) tiling array platform (Singh-Gasson et al, 1999).

(v) Cheng et al. (2005) used an updated version of the Affy platform with an even tighter spacing of the probes, every 5 bp, and covering 10 chromosomes (approximately 30%) of the human genome. Transcript maps were generated for polyadenylated cytosolic RNA from eight cell lines (and for one of these cell line, also non-polyadenylated RNA).

Several parameters affect the outcome of a particular tiling experiment. These five papers represent different choices in array design and manufacturing, RNA extraction and hybridization conditions, and data processing methods.

The array design parameters include the length and genomic spacing of the probes, as well as considerations concerning the use of mismatch probes or, depending on what experimental protocol is used, whether to cover one or both genomic strands. For the oligonucleotide tiling experiments referenced above, the probe length varies between 25-70 bases. Arrays with longer probes can be hybridized at a higher temperature and are thus supposedly less prone to pick up cross-hybridization, while providing a less detailed hybridization map (sensitivity/resolution trade-off). The genomic spacing of the probes is measured between probe initiation points, and could be anything from the smallest possible distance of one single base up to the length of the probe, or even a little more (there is no agreed upper spacing limit as to where an array design ceases to be a tiling array). At the design stage it is important to minimize potential cross-hybridization, self-pairing, and other probe sequence artifacts such as DNA secondary structure formation (SantaLucia and Hicks, 2004). For instance, genomic regions considered as repeats (by, e.g., RepeatMasker (A.F.A. Smit and P. Green, unpublished)) are usually omitted from the design due to potential cross-hybridization, and if some flexibility is allowed in the design process, probes may be chosen so as to achieve better probe thermodynamics. In arrays interrogating annotated genes only this is a feasible approach (Mathews et al, 1999; Hughes et al, 2001; Rouillard et al, 2003), but for tiling arrays with high genomic density the options for probe optimization are usually limited. Factors pertaining to the manufacturing include what technique is used to attach the probes to the supporting material, the spacing of the probes on the array slide or chip (the so-called feature density), and how easily new array designs can be implemented (the so-called design flexibility).

The experimental protocols for extraction, labeling, and hybridization of the RNA sample to the array vary considerably. Choosing what type of target RNA (tissue or cell line; poly(A)+ or total RNA; etc.) to extract and hybridize, and what reactions and conditions, such as temperature and salt concentration, to use in the hybridization will have an effect on the result. The number of technical and biological replicates is another crucial choice, where more replicates possibly enables greater certainty and detail in the interpretation of the results.

Once the tiling arrays have been designed, manufactured, and hybridized with labeled RNA and the corresponding hybridization intensities have been extracted, there are a number of ways to transform the raw intensities into a score for each probe. This is usually done employing statistical methods such as a sign test or the t-test, possibly within in a genomic window and including replicate arrays in the process. Exactly what methods are available also depends on the design features of the array, such as the presence of mismatch probes. The segmentation of the genome into transcribed and non-transcribed regions is then performed based on the scores.

Our goal is to assess different tiling microarrays used for transcription mapping, an area where no detailed comparison so far has been performed. Previous work on comparing the outcome of gene-based microarrays include papers by Tan et al (2003), Jarvinen et al (2004), Mah et al (2004), Park et al (2004), and Yauk et al (2004). Most of these indicate differences in the gene expression results from different microarray platforms, which have been attributed to differences in data processing or inadequate choice of comparison metrics (Larkin et al, 2005).

Here, we set out to identify the strengths and weaknesses of different tiling microarray platforms, with particular focus on technologies that are able to provide a detailed and comprehensive transcription map of large genomic regions.

We start by performing a pilot study on a set of already published chromosome 22 transcription data from several tiling microarray platforms. This study indicated that amplicon (PCR) arrays are less suitable to pick up existing annotation when compared to high-density oligonucleotide arrays.

We then describe how we enabled a direct comparison of the two oligonucleotide-based platforms with a demonstrated ability to cover entire genomes at a high genomic probe density with a reasonable number of arrays, MAS (Bertone et al, 2004) and Affy (Kapranov et al, 2002): We hybridized identical biological samples to the arrays and developed a unified data processing pipeline. Our microarray experiments are presented in detail, and we describe a number of statistically based scoring schemes and various genomic segmentation methods, some novel and some that have been used previously. We assess the different scoring and segmentation procedures, and the use of replicate arrays, and we also address the significance of the probe length, the genomic density of probes, and mismatch probes.

We compare the results from the two oligonucleotide tiling microarray strategies with each other and with the recently generated Gencode gene annotation (http://genome.imim.es/gencode/), and discuss possible implications for further transcription mapping studies.

RESULTS AND DISCUSSION

APPROACH

Preliminaries

The ultimate goal of our study is to provide guidance when choosing strategy for the multiple tissue whole-genome transcription mapping of the human genome that will take place in next phase of the ENCODE project (The ENCODE Project Consortium, 2004), which aims at elucidating the functional role of every nucleotide base in the human genome. We achieve this goal by bringing the concepts presented in the works published by Kapranov et al. (2002), Rinn et al. (2003), Bertone et al. (2004), and Cheng et al. (2005), into a unified framework to make the approaches as comparable as possible. Specifically, we compared two different high-density oligonucleotide array formats, MAS and Affy. They were hybridized with placenta poly(A)+ and NB4 total RNA samples resulting in five different sets of data (Table 1).

Except for the pilot study, the present study was performed on regions ENm001-ENm011 of the human genome defined by the ENCODE consortium. Currently, the ENCODE project is in its pilot phase, focusing on 44 regions that together encompass 30 Mb of the human genome (whereof ENm001-ENm011 covers 11.6Mb or 39%). Roughly 15 Mb correspond to well-studied gene loci such as the CFTR and beta-globin loci, and the other 15 Mb consist of regions randomly chosen by stratifying the genome according to the density of known genes and the degree of non-exonic conservation. In the ENm001-011 regions, the total number of known Gencode genes is 264, corresponding to 1342 different splice variants, for an average of 5.1 splice variants per gene. The number of unique exons (requiring at least one of the start and end positions to differ) is 4147, which translates to 6.7 exons per transcript (8938 non-unique exons).

We follow the nomenclature of Royce et al. (2006), i.e., *probes* refer to the molecules attached to the microarray at the time of manufacturing, and *target* or *sample* is the RNA extracted from a biological entity (tissue or cell line) and which is allowed to hybridize to the probes on the microarray.

Pilot study: comparison of public chromosome 22 tiling data

We carried out an initial comparison of previously published array-based transcription maps of the (gene-dense and well-annotated) chromosome 22 generated from PCR-based tiling arrays (Rinn et al., 2003; White et al., 2004) and two oligonucleotide tiling array platforms (Bertone et al., 2004; Kapranov et al., 2002) (Figure 1). These data sets were generated from 15 separate experiments (tissues or cell lines), representing three different microarray platforms (PCR, MAS, and Affymetrix). We used the RefSeq annotation (Pruitt et al., 2005) as a benchmark since Gencode annotation is not yet available for the entire chromosome 22. For each experiment, we measure the consistency between gene annotation and transcribed regions identified by individual studies. The transcription data from oligonucleotide arrays (MAS and Affymetrix) clearly agrees better with the RefSeq exon annotation than the data from PCR arrays, an observation that holds true across all cell lines and tissues (Figure 1). Thus, although the results have to be interpreted with some care since they were not obtained using the same biological samples or scoring schemes, we conclude that PCR-based arrays are clearly less useful for a detailed transcription mapping study, in part because of their lower genomic resolution. PCR-based arrays also have a significantly lower feature resolution on the array compared with arrays with *in-situ* synthesized probes. Therefore, we focus our subsequent experimental and analysis efforts on the oligonucleotide tiling microarrays.

Array design and experimental outline

An oligonucleotide array containing 36 bp oligonucleotides that tile both strands of the nonrepetitive sequence of the ENCODE regions end-to-end (allowing some positional shifts to reduce self-complementarity) was prepared using maskless photolithography, MAS (maskless array synthesis). The MAS arrays cover both strands of the ENCODE regions ENm001.ENm011 (11.6 Mbases). An Affymetrix ENCODE array, which covers one strand of the entire ENCODE region on one array, tiled with 25-mer oligonucleotides with an average distance between oligonucleotide starts of 21 bases was obtained from the manufacturer. This array has both perfect match (PM) and mismatch (MM) probes.

In total, five different hybridization experiments were carried out (Table 1): Two different RNA targets (placenta poly(A)+ RNA and NB4 total RNA) were hybridized to the two different array types, MAS and Affymetrix. The Affymetrix arrays were hybridized according to the manufacturer's recommendation (experiment id: Affy). The maskless arrays were hybridized using two different experimental protocols, the protocol described in Bertone et al., 2004 (experiment id: MAS-B) and a variant of the manufacturer's recommended protocol (experiment id: MAS-N). The protocols differ at many stages, e.g., the Bertone protocol generates cDNA while the other MAS protocol generates cRNA, which is also chemically fragmented prior to hybridization (see

Methods for a full description). The placental RNA was hybridized using both MAS protocols whereas the NB4 sample was only hybridized according to the MAS-N protocol (Table 1).

Generating comparable maps of transcriptionally active regions (TARs)

To bring the data generated using the two different technologies into a comparable form, we developed ways of applying similar post-processing procedures to the data. Specifically, we applied a framework for scoring hybridization intensities and for segmenting the tiles into transcribed and non-transcribed regions. This framework constitutes an extension of previously published methods for scoring of hybridization intensities and segmentation of genomic regions in an oligonucleotide tiling microarray transcription mapping experiment.

(i) Development of consistent scoring schemes

For each spot on the microarrays, a hybridization intensity was collected. It is usually advantageous, for oligonucleotide tiling arrays, to aggregate the intensities from probes that are adjacent to each other in genomic space (Kampa et al., 2003; Cheng et al., 2005; Royce et al., 2005). This is done by applying a sliding genomic window encompassing multiple probes and converting the intensities within the window into a score, which is assigned to the probe in the middle of the window. Using a windowed approach intuitively makes sense as we are ultimately interested in obtaining a set of regions whose intensities are significantly higher than the background, and we would expect those regions to be of the same length as exons (150-200 bp on average, depending on exon type) rather than of single probes (25-36 bp in this study).

We developed new ways of scoring the MAS array hybridization data, and describe these in a framework of three levels of scoring: (a) single probe intensities, (b) robust statistics within a sliding window, and (c) robust statistics using paired data within a sliding window (Cawley et al., 2004). In addition to the sliding window approach, the use of replicate arrays is beneficial in order to reduce the hybridization noise; the more replicates the greater the statistical power in the scoring. Hence, it is important to construct scoring models that easily can take into account replicate experiments.

(a) Single probe intensities
This is simply using the raw intensities from the arrays, possibly normalized. By wisely choosing methods and parameters to deal with the genomic segmentation (see below) it is possible to obtain reasonable results from this approach (Bertone et al., 2004). Both intra- and inter-array normalization of the microarray data may be important, depending on the actual structure of the data, to insure commensurability of the data used (Royce et al., 2005).

(b) Robust non-parametric statistics within a sliding window
We employed the sign test for scoring the MAS array data. It is an attractive test since it is statistically robust, which is important because the array data are inherently noisy, and it does not assume that the data are normally distributed. By comparing each intensity within a sliding genomic window of a specified size with the array median, it will yield a measure or a score of how significant the intensities are (see Methods for details). It is easy to include multiple replicates in this scheme: each probe is simply compared to the median intensity of its own array, and no inter-array normalization is necessary. On the other hand, the number of available score levels is restricted due to the discreteness introduced by the counting (it is a binomial distribution), and it may not be sufficient in situations where discerning the top scores (say, top 5%) from the near-top scores is important.

With an average genomic spacing of 36 bp between the starts of two adjacent probes, a window size of 160 bp encompasses five probes (allowing for some positional shifts). Using a smaller window (90 bp) results in insufficient statistical resolution (not enough score levels available). A larger window size (240bp) enables greater statistical resolution, but is not suitable to pick up transcribed regions the size of an average exon (150-200 bp). Clearly, the resolution of the array (genomic spacing of probes) is important in order to obtain both statistical power and reasonably distinct borders between transcribed and non-transcribed regions.

(c) Robust non-parametric statistics using paired data within a sliding window (Affymetrix)
When there is paired data available, such as the perfect match (PM) and mismatch (MM) probe intensities of Affymetrix tiling arrays, the paired Wilcoxon signed rank test is a more powerful option than the standard sign test. It was first used with tiling microarrays by Cawley et al. (2004) to score ChIP-chip data, and it is also immediately applicable to transcription data as is shown in Kampa et al. (2004) and Cheng et al. (2005). All pairwise PM-MM differences within the window are calculated and a p-value, which essentially measures how significantly the distribution of PM-MM differences is skewed to either side around zero, is calculated. While this is analogous to the standard sign test (above), it has considerably greater statistical power and it can also be argued that the mismatch setup is able to account for at least some of the cross-hybridization occurring on the array. For the transcription mapping reported in Kampa et al. (2004) and Cheng et al. (2005), the corresponding point estimate, called the pseudomedian, was used instead.

Given that the maskless arrays did not contain proper mismatch probes, we tried using the complementary strand oligo of the MAS arrays as a “mismatch” probe (to mimic the PM/MM setup). This use of this approach, which we call the Fwd-Rev scoring, is justified on the MAS-B (placenta) data, since the correlation between forward and reverse strand probes is close to the correlation between PM and MM probes for the Affy placenta data (Table 2, and below).

The input data type (e.g.: PM only, or PM and MM; the number of replicates), the algorithm (e.g.: the standard sign test), and, if applicable, the corresponding genomic window size (e.g.: 160 nt) together specify a scoring scheme.

(ii) Segmentation of genomic regions

Having obtained one score value per oligonuclotide probe, the next step is to construct a transcription map based on these scores, or, in other words, to segment the genomic regions into transcribed and non-transcribed regions. A couple of specific names have been used to denote transcribed regions derived from microarray tiling experiments: the term “transcriptionally active region” (TAR) was first used by Rinn et al. (2003) and also in Bertone et al. (2004), while “transfrag”, which stands for transcription unit or fragment, was introduced by Kampa et al. (2004) and later used in Cheng et al. (2005). Here, we will mostly use the term TAR, which pertains to any kind of tiling array-derived transcribed region regardless of overlap with genes, exons, or other genomic features.

Maxgap/minrun segmentation
In Bertone et al. (2004), transcribed regions were generated by demanding at least 5 adjacent probes with a raw intensity in the top 10% of all intensities of that slide; any such region was called a TAR. Thus, the threshold above which to consider a probe “positive” was the intensity value corresponding to the 90th percentile, and any probe that was below the threshold immediately terminated the transcribed region. In the Affymetrix series of publications (Cheng et al., 2005; Kapranov et al., 2002), the threshold for generating TARs was based on setting a maximum false

positive rate of the hybridization levels of negative bacterial controls, thus enabling an optimized percentile cutoff for each array set and biological sample. Furthermore, gaps were allowed, such that a maximum stretch of a certain number of nucleotides (called maximal gap, or maxgap for short) with a score below the threshold was allowed between probes whose scores were above the cutoff. Typically, the maxgap parameter allows one or two probes to be below the cutoff, while still being incorporated into the TAR. The total length of a TAR is then demanded to be of at least a certain length (a minimal run, or minrun), which depending on the tiling density and the actual value of the minrun parameter corresponds to at least two, and possibly more, probes for each TAR or transfrag. The threshold, the minrun value, and the maxgap value are the three parameters that need to be specified for the maxgap/minrun segmentation algorithm.

HMM segmentation

As an alternative to the maxgap/minrun segmentation, a hidden Markov model (HMM) (Rabiner, 1989; Li et al., 2005; Ji and Wong, 2005) was employed to predict and score TARs, given either the raw intensity data of each probe, or the derived scores (see above). Each probe can be in one of the four HMM states (TAR, non-TAR, and two other intermediate transition states), emitting the assigned intensity/score (i.e., the emission spectrum is continuous). The parameters of the HMM can be estimated by learning from the sequences of probes which fall into regions with known transcription characteristics (e.g., according to gene annotation). The HMM can then be applied to sequences of probes bearing the same scoring protocol to determine the most likely corresponding state sequence, in order to identify TARs. For the HMM segmentation approach, the decoding algorithm (e.g.: Viterbi decoding) needs to be specified.

Description of comparison pipeline

We have analyzed the five microarray tiling experiments (Table 1) at multiple stages throughout the data processing pipeline.

(i) The first comparison is on the level of the raw data. A necessary but not sufficient condition for a successful microarray experiment is satisfactory hybridization reproducibility between technical replicates. We calculate a simple correlation coefficient between the hybridization intensities of technical replicates within each platform, to assess the level of basic experimental reproducibility.

(ii) Before proceding to the next level, a decision has to be made as to what TAR sets to compare. As outlined in the previous section about TAR map generation, this includes decisions about what scoring algorithm, what segmentation method, and what corresponding parameter settings to use for each of the five experimental data sets.

(iii) The resulting sets of transcribed regions are compared with each other, both within and across microarray platforms and biological samples, and the degree of overlap between detected transcription and exon annotation is measured. There are two main measurements: how much of the known annotated exons are covered by a detected transcribed region ("sensitivity"), and the degree to which the detected transcription falls within known exons ("positive predictive value", PPV). The PPV is defined as the number of nucleotides in TARs that overlap with exonic regions, divided by the total number of nucleotides in the TAR set (in papers focused on gene finding this is sometimes referred to as "specificity"). The sensitivity is defined as the number of nucleotides in annotated exons that overlap with TARs, divided by the total number of nucleotides in annotated exons. For the annotation comparison, we have chosen to use the Gencode annotation. The Gencode project is a part of ENCODE and aims at finding all potential protein-coding gene candidates in the ENCODE regions, and to experimentally verify or reject these candidates using paired-exon PCR and RACE. Although not perfect, we believe that the Gencode set of genes constitutes the most comprehensive and accurate gene annotation available for the ENCODE regions. Furthermore, we assess the degree of evolutionary conservation of the TARs by matching them with a set of conserved regions. The overlap is calculated both for TARs that fall within known transcribed regions (i.e., within Gencode exons) and for TARs that fall outside of those regions (novel TARs).

(iv) The transcription status was assessed for all annotated splice variants of all known genes in the Gencode annotation of regions ENm001-ENm011 (accepting the exons with labels "VEGA_known", "VEGA_Novel_CDS", ")VEGA_Novel_transcript_gencode_conf", and "VEGA_Putative_gencode_conf"). For each splice variant, the scores of all exon overlapping probes were collected and the transcription status assessed using the sign test (by comparing each individual score with the median score) (Bertone et al, 2004). A gene was considered transcribed if at least one of its splice variants was deemed transcribed at the chosen significance level. We also assess the multi-exon coherence of each transcript, based on the median intensity score of each exon within the transcript.

(v) As a final step, we performed experimental validation of the results in placenta (MAS-B and Affy), using reverse transcriptase-PCR of a subset of the novelTARs, both TARs unique to either platform and TARs that were present in both. We also performed experimental validation of a number of genes with differing transcription status in the two platforms, including as a negative control a set of genes that were considered off in both platforms. PCR primers were chosen using Primer3 (Rosen and Skaletsky, 2000) so that any primer pair only had exactly one genomic match.

OUTCOMES

TAR generation pipeline

For the minrun/maxgap segmentation, the minrun parameter was set to 50 bp, which means that the minumun length of a TAR is 50 bp. Thus, at least four probes have to be included in an Affymetrix TAR and three for the MAS array TARs. Roughly 5% of all unique Gencode exons are shorter than 50 nucleotides and the loss in sensitivity due to this parameter setting will be comparable.

The maxgap parameter was set to 50 for the Affymetrix data and 80 for the MAS data, effectively allowing the inclusion in a TAR of a probe whose score is below the score threshold, if it is flanked immediately on both sides by probes with scores above the threshold. Other maxgap settings were tested but the results did not improve in terms of gene annotation agreement (data not shown).

As expected, the segmentation threshold directly influences the size of the resulting TAR sets (Figure 2a). The step-like pattern in this figure is because of the finite number of available scores.

Different scoring schemes give different results, and also differ from the results obtained using single probe intensity scores (Figure 2b and, in Supplementary material, Figures S2, S4, and S5). As is clear from Figure 2b and Figure S2a, the standard sign test provides the best performance for MAS-B (placenta) if a sensitivity at or above 25% is required, but its improvement in terms of PPV when increasing the segmentation threshold is modest. The Fwd-Rev scoring is more sensitive to the choice of threshold, and it performs better than the sign test scoring (Figure 2b) for segmentation thresholds above the 94th percentile. For the standard sign test, we tried applying different weights to the probes within a window, e.g., multiplying each score with a discretized Gaussian, but no improvement in performance was recorded (Figure S2 in Supplementary material).

The more elaborate scoring models using replicates outperform the single probe intensity scoring, as is seen for both Affymetrix and MAS data in Figure 2d and Figure S3, where the resulting TAR sets are compared to the Gencode annotation. For the Affymetrix data, it is also clear that the use of mismatch probes improves performance, in particular for the NB4 total RNA experiment (Figure S3). These two points (elaborate scoring with replicates and the advantage of using mismatches) are further illustrated in Figure 2e, where the positive predictive values (PPV) of the TAR sets when choosing a segmentation threshold that corresponds to a sensitivity of 30% have been plotted.

The analysis of different segmentation algorithms reveals that the TAR sets generated by the non-parametric HMM segmentation (Viterbi decoding) are biased towards a high sensitivity (Figure 2b and Figure S2) where it performs on par with the minrun/maxgap segmentation algorithm.

Results from comparison pipeline

(i) Replicate comparison of unprocessed hybridization intensities

For the MAS arrays, we obtained Pearson correlation coefficients of 0.83 and 0.96 for placenta using Bertone and our variant of the manufacturer's protocol, respectively, measured on pairwise comparison of the raw hybridization intensities of the arrays. The corresponding result for Affymetrix was 0.96, and NB4 results were similar (Table 2). We also note that the correlation of PM and MM probes for Affy placenta is close to the correlation of Fwd and Rev strand probes for MAS-B.

(ii) Choose TAR sets to include in comparison

For each of the five experiments, the best-performing scoring and segmentation algorithm was chosen, and the segmentation threshold was tuned to generate TAR sets of roughly equal size, as measured in number of bases (Table 3; circled data points in Figure 2a and Figure S1). To enable a comparison of the TAR sets, MAS array TARs on the two strands were merged into one set of non-stranded TARs for each biological sample (TARs from the Affymetrix arrays do not have strand information). For MAS, the standard sign test scoring was chosen, and for Affymetrix, the Wilcoxon signed rank test (pseudo-median). The segmentation thresholds range from the 87th to the 93rd percentiles for the various sets. The resulting sizes of the sets included in the subsequent comparison is in the range of 629kbp to 701kbp, yielding between 2545 and 4674 TARs. The total number of bases in exons in the analyzed regions (ENm001-ENm011) is 1001kbp, which means that the chosen threshold levels can yield at most a sensitivity of 0.63-0.70. This is a theoretical upper bound on the sensitivity reflecting the fact that not all genes are transcribed at all times in all tissues. Assuming further that approximately 25% of the detected transcription is novel and/or noncoding, and thus not present in the Gencode annotation, we expect to see a sensitivity not surpassing 0.47-0.53 while the positive predictive value has an upper bound at 0.75. The TAR length distributions are similar for the five sets in that they are all unimodal and decay roughly exponentially (Figures S6-S7).

(iii) Compare TAR sets with each other and to Gencode annotation and conserved regions

The overlap between the placenta and NB4 TAR sets from the different experiments is shown in Figure 3a. As a measure of the overlap between the sets we calculate a ratio $R = |\cap|/|\cup|$ for each pairwise comparison. For two sets that agree completely, R equals 1. We find that the MAS-B placenta TAR set agrees better with the Affy placenta TAR set (R=0.22, overlap size is 236k) than the MAS-N TAR set does with Affy (R is 0.17 for placenta and 0.16 for NB4). Between 62% and 72% of the nucleotides in the placenta TAR sets are exclusive to a particular experiment (pairwise comparison MAS-B vs. Affy and MAS-N vs. Affy). For the NB4 total RNA TAR sets, the overlap was 177k (Figure 3a), meaning that more than 70% of the nucleotides in either NB4 TAR set are exclusive to that set.

The overlaps across the different biological samples but within the experimental technologies are larger than the overlaps within the biological sets across experimental technologies, Figure 3b. An extreme example is the placenta MAS-N and NB4 MAS-N sets which agree much better (dashed-dotted brown line in Figure 3b; R=0.67) than NB4 MAS-N and NB4 Affy (solid black line in Figure 3b; R=0.16). Restricting the overlap calculations to the subset of TARs that overlap conserved or exonic regions, or both (i.e., moving to the right in Figure 3b; see also Figure S8), yields higher values of R for the within-biological sets comparisons (black lines) and for the within-Affy comparisons (solid brown line), but not for the within-MAS comparisons. Thus, there is no enrichment for conserved regions or known genes within the common parts of the MAS TAR sets.

The agreement with annotation (Gencode exons) is better for the Affy sets than for the MAS sets, both placenta and NB4. While similar sensitivity levels are achievable, the Affy TAR sets reach significantly higher PPVs (Figure 2d and Figure S3). Likewise, the agreement is larger for the placenta sets (MAS-B, MAS-N, and Affy) than for the NB4 sets (MAS-N and Affy) (Table 3 and Figure 2d and Figure S3). The NB4 sets agree less with annotation than the placenta sets from the same platform. The NB4 is a total RNA sample and thus also introns may be labeled and show up as transcribed, in part explaining this behavior.

Most Gencode unique exons are either fully covered by a TAR (>90% of exon nucleotides overlap with a TAR) or not covered at all (<10%), as is seen by the bimodal distribution in Figure 4. This is true for all TAR sets (Figure S9). Looking separately at the 5' and 3' exons, we notice a 3' bias for the Affymetrix poly(A)+ data (but not or the MAS data (Figure S9)), detecting 32% of the 3' exons and 25% of the 5' exons entirely.

The TAR sets were also compared with a set of conserved elements, generated from the union of conserved regions called by Threader Blockset Aligner (TBA) (Blanchette et al., 2004) and MLagan (Brudno et al., 2003) (Table 6 and Figure S10). The union set of conserved elements corresponds to a little over 10% of the human ENCODE regions. We investigated for all five TAR sets the degree to which the TARs overlapped with a conserved region and find that most of the novel (intergenic) TARs are not conserved according to this measure. Only 7-8% of Affymetrix novel (intergenic) TARs and 1-2% of MAS novel TARs overlap fully (>90%) with the conserved regions.

(iv) Transcription status of known genes and exons

The transcription status of all known splice variants (transcripts) of the 264 Gencode genes in the regions ENm001-ENm011 was assessed (Table 5) as outlined in the Approach section above. A gene is considered as transcribed if at least one of its transcripts is transcribed at significance level $p<0.001$. If a transcript has less than 10 probes it will be unable to reach a p-value below 0.001, and if this is true for all splice variants of a gene, that gene is in the "Too few probes" category (Table 5). In total 158 (69.3%) genes are considered transcribed according to both platforms, and 221 (83.7%) according to at least one platform (210 in MAS-B, 169 in Affy), and similar percentages are obtained on the transcript level. The coherence of transcription of exons is assessed and the results are in Table 6. In general, the multi-exon coherence (all exons on) is larger for the Affy sets. Furthermore, there is a significant enrichment of multi-exon coherence in transcripts that are considered as transcribed in both platforms (this is true for both placenta and NB4 data). One example is the Affy NB4 sets for which in total 15.5% of all transcripts have all their exons transcribed, while 26.9% of the transcripts that are on in both the Affy and the MAS-N NB4 sets have all their exons transcribed.

(v) Experimental validation of novel TARs and known genes

Experimental validation of the microarray transcription data is crucial to the interpretation of the results. First, we chose 96 novel placenta TARs to verify employing reverse transcriptase PCR; 39 TARs that were exclusively found on the MAS platform (placenta MAS-B), 37 that were exclusively found on the Affymetrix platform, and 20 that were common to both. All TARs were required to not overlap with any known Gencode gene, and Fifteen ???????? of the Affymetrix-only and ?????? of the MAS-only validation targets showed bands of the right size on the gel, while ??????? of the common validation targets showed bands of the right size. In total, [[NEED TO UPDATE THE NUMBERS...]] ????/???? (????%) of both the investigated Affymetrix novel TARs and of the investigated MAS novel TARs were verified with this RT-PCR approach, Table 6. [WAITING FOR RESULTS]

Second, a number of known genes were also validated, using the same experimental method and criteria for PCR primer picking as above. Genes that were completely off (i.e., none of its splice variants was considered transcribed according to the criteria outlined in (iv)) according to one of the platforms but not the other were assessed (in total 17 On-MAS-B/Off-Affy genes and 8 Off-MAS-B/On-Affy genes) as well as 18 genes that were completely off in both, Table 6. The results indicate ... ... [WAITING FOR RESULTS]

DISCUSSION

In this work we have attempted to assess the suitability of two oligonucleotide tiling microarray strategies for transcription mapping in human. We tried to overcome the inherent differences between the approaches through using the same biological samples and a unified scoring and TAR generation procedure, and we have produced and compared several sets of transcribed regions. We conclude that many factors are significant for the outcome of the experiments. Here, we discuss further some key parameters and findings.

Experimental considerations, array noise, and technical replicates consistency

In the comparison between the two microarray tiling platforms, the Affymetrix strategy yielded TARs that better agreed with the Gencode annotation (Figures 2d, 2e, 4, S3). An explanation for this would be a higher noise level for MAS arrays. The Pearson's correlation coefficients between raw intensities of technical replicates of MAS and Affymetrix arrays (Table 2) are strikingly similar, though, which points to the MAS array noise being systematic rather than random. This could for instance be due to probe sequence artifacts, sample contamination (after it was split into two aliquots for the MAS and Affymetrix experiments), sub-optimal hybridization parameters or protocol-dependent labeling artifacts (Nazarenko et al, 2002). A comparison of the NB4 total RNA and placenta poly(A)+ sets within the two platforms as presented in Tables 2-4 and 6, and in Figure 3, reveals that the coherence between the platforms, and between the results of each platform and annotation, is larger for the poly(A)+ sets. For total RNA, also introns may be labeled and it is thus not surprising that the agreement of TARs with gene annotation is lower for total RNA sets than for poly(A)+ sets given that the evaluation of TAR sets was based on TAR overlap with exons only, and given that the TAR sets for comparison were generated in order to be of similar size regardless of RNA type. While this explains the worse performance of the NB4 sets compared to placenta, it does not explain the differences in performance between MAS and Affy arrays. A thorough investigation of the behavior of different types of RNA is presented in Cheng et al (2005). A related issue is the cross-hybridization. For a transcription experiment, the amount of different RNA species present in the poly(A)+ or total RNA sample is large. Even though the probes, in our case, only represent part of the ENCODE regions, the target RNA is derived from the entire genome. Thus, cross-hybridization is potentially present at rather high levels. Even if an array appears to be flawless (even hybridization over the array surface, no scratches or dust particles), the derived results are sometimes significantly worse, in terms of, e.g., overlap with annotated genes (for transcription mapping arrays), than other arrays hybridized with the same sample RNA at the same time and under the same conditions. A suggestion of an intra-array measure was recently presented by Kim et al (2006).

Genomic spacing and length of probes

The spacing, or "genomic density", of the probes is directly related to the size of the objects the experiment is aimed at detecting via the size of the sliding genomic window that often is used in the scoring procedure. To gain a reasonable statistical power in the scoring and genomic segmentation approaches, the window must encompass several probes, and the greater the genomic distance between the probes the greater the required window size. In a transcription mapping experiment, a window that is significantly larger than the average size of an exon is not desired since it would likely contain both probes that represent actual transcription (exons) and probes that belong to truly non-transcribed regions (introns). In addition, the genomic probe spacing affects how well the endpoints of the discovered objects can be defined. The theoretical uncertainty of where a transcribed region starts and ends depends to some extent on the experimental (e.g., fragmentation) and hybridization (e.g., temperature) conditions, but its upper limit is the genomic distance between two adjacent probes. Initially, we hypothesized that the probe length would have a significant effect on the results, where longer probes would result in cleaner data since longer probes are more specific to a unique genomic location and thus less sensitive to cross-hybridization. We did not see any such effect, though this is to some extent dependent on the hybridization conditions, longer probes can be hybridized at a higher temperature with the same expected sensitivity. It is also true that at a particular temperature longer oligonucleotide probes will allow RNA with a greater number of mismatches to hybridize to the oligo, thus increasing the risk of cross-hybridization. An interesting idea is to construct an array with probes of variable length, such that the melting temperatures of all probe sequences on the array lie in a very narrow range (isothermal array).

The role of mismatch probes

Using mismatch probes generates significantly better results (Figure 2c, 2d and Figure S3), in particular when the target RNA is total RNA as opposed to poly(A)+ RNA. We speculate that the hybridization noise levels, for instance from labeled introns, are higher for total RNA experiments and that the MM probes then add crucial information when discerning true hybridization from background or cross-hybridization. The Affy arrays have mismatch probes, the MAS arrays don't. A simulation of mismatch probes is to use the reverse strand probes, that are present only on the MAS arrays, as mismatch probes, and this approach yields indeed improved performance for the MAS-B data as opposed to single intensities, but generally not compared with standard sign test scoring (below). It is our belief that the use of true mismatch probes is a straightforward way to significantly improve the signal-to-noise ratio of oligonucleotide tiling arrays.

Scoring and replicates

We tried several scoring approaches for our array data and found that there were significant differences between using single probe intensities and a probe score based on replicate arrays. Using the standard sign test scoring was ultimately deemed the best way of scoring the MAS data, while the best way of Affy scoring was to use the pseudomedian of the Wilcoxon ranked sign test. We tried, unsuccessfully, to improve the MAS scoring through putting more weight on the probes in the middle of the window. For MAS-B data, using a Fwd-Rev scoring algorithm (Wilcoxon, pseudomedian), as a surrogate for a true PM-MM scoring, improved agreement with annotation for high maxgap/minrun segmentation thresholds. Exploring the Affy data showed that the sign test performed quite badly on PM data only, while it approached the Wilcoxon test performance on PM-MM data (Figure 2c). In fact, the PM-only scoring of Affy with the standard sign test (ie., identical scoring as the MAS sign test) resulted in a behavior very similar to that of the MAS – a relatively low agreement with annotation, and reduced impact of increasing the segmentation threshold (*PPV* rather insensitive to threshold increases). This indicate that mismatches can be very useful. We see that reducing the genomic density of the Affy array to 50% (exclude every other probe on the arrays before scoring) is detrimental to the performance as well.

Segmentation into transcribed/non-transcribed regions

We also tried different segmentation methods. The maxgap/minrun algorithm that was first presented in Kampa et al. (2003) and that we have used here is a straightforward way of using neighboring information to partition the genomic region under study into transcribed and non-transcribed regions. Different parameter settings (threshold for calling a probe "positive", maxgap, minrun) were tried and evaluated. We also tried a more sophisticated approach by using hidden Markov models, HMMs, and observed similar results. The HMM uses a continuous emission spectrum and can be classified as a kind of peak-fitting algorithm. A possible improvement of the HMM segmentation may be realized by using a Generalized Hidden Markov Model (GHMM, a.k.a. Hidden semi-Markov Model), where the length distribution of transcribed regions is explicitly modeled, instead of an HMM in the algorithm. Further scoring can be done by computing the posterior probabilities $P(\pi_i = k \mid x)$ for the predicted states on probes, where π_i is the state of the *i*th probe, *k* is the predicted state, and *x* is the whole intensity/score sequence (the emitted values). These scoring data indicate the confidence in every single prediction and may be used to refine the TAR prediction results obtained by using an HMM method.

Assessing the transcription of known genes

The overlap of genes considered transcibed by both platforms is substantial, and the multi-exon coherence is clearly higher for the Affy sets than for the corresponding MAS sets. The figures pertaining to the transcription of individual splice variants presented in Table 6 (in parenthesis) are probably overestimates since the overlap between different variants of the same gene often is substantial, and wouldn't necessarily be detected by our sign test based transcription assessment of the transcripts. Unless each exon/intron junction is experimentally tested (using, e.g., RACE), it is impossible to know the true transcription status of each variant of a gene. The same reasoning is applicable to the column labeled "Some exons" in Table 6: several of the transcripts ending up in this column are most likely not transcribed at all. Here, a reliable quantitative measure of the transcription of each exon would be helpful. Basing the on/off calls for the exons on a sign-test generated p-value is a problem since most exons are not long enough and thus don't have enough probes to yield significant p-values, which is why we chose to use the median score for each exon.

Conclusions

In its current form, the Affymetrix tiling microarray platform is better than the MAS platform for detailed transcription mapping of the human genome in the sense that the agreement of the TARs with known annotation is larger (Figure 2), and also in terms of coherent transcription of all exons in multiple-exon transcripts (Table 6). From our study, we attribute this foremost to the use of mismatches and a higher genomic density of the probes, while we cannot entirely exclude the effects of the differing labeling and hybridization protocols. On the other hand, the two technologies are roughly equal in their ability to detect novel transcription as indicated by our experimental validation of novel TARs (checking for false positives) [[CHECK THIS WHEN THE RESULTS ARRIVE]] (Table 7). The overlap of genes or transcripts that are considered transcribed by the different platforms is substantial, and again the experimental validation of the differences (checking for false negatives) does not indicate that either platform is superior [[CHECK THIS WHEN THE RESULTS ARRIVE]]. Furthermore, the MAS technology allows for rapid manufacturing of customized designs and cost-effective production of small array series which may also be of importance in the choice of oligonucleotide microarray platform. We conclude that oligonucleotide tiling microarrays are suitable to detect novel transcribed regions, and that the use of replicates and statistically based scoring schemes significantly improves the performance for all investigated oligonucleotide tiling microarray-based transcription mapping experiments.

METHODS

Array designs

Affymetrix arrays

Arrays were designed and manufactured by Affymetrix, Inc., using a physical mask. Probes are 25 bp long with an average genomic spacing of 21 bp, and they cover one genomic strand with the exception of repeat regions. Each probe is present in a “perfect match” and a “mismatch” version. The mismatch probe contains a single substitution at the middle probe position (A->T, T->A, C-

>G, G->C). The probes were originally designed from Ncbi v31 of the human genome build, and mapped forward to Ncbi v34 (hg16) using the LiftOver tool at the UCSC Genome Browser. Each array contains in total approximately 1,400,000 features.

MAS arrays

Arrays were designed by us and manufactured by NASA using a NimbleGen maskless array synthesizer. Probes are 36 bp long with an average genomic spacing of 36 bp, in principle tiled end-to-end. Positional shifts were allowed to avoid self-complementarity at the probe ends (defined as at least 4 consecutive complementary nucleotides within the 6 5'/3' nucleotides). The probes cover both genomic strands with the exception of repeat regions, as defined by RepeatMasker (A.F.A. Smit and P. Green, unpublished). The design was done on the Ncbi v34 of the human genome build, and each array contains almost 390,000 features.

RNA extraction and array hybridization

In total three different placenta poly(A)+ biological batches and three different NB4 total RNA batches were used, with the number of technical array replicates according to Table 1.

Cell culture

The human NB4 cells were cultured in RPMI medium containing 20mM L-glutamine (Media Tech) and supplemented with 10% fetal bovine serum (Invitrogen), 100 IU/mL penicillin (Media Tech) and 100µg/mL Streptomycin (Media Tech). Cells were maintained at 37^0C under 5% CO_2/95% air in a humidified incubator.

RNA samples

Total RNA from the human NB4 cells was extracted using Qiagen RNA extraction kit according to the manufacturer’s instructions. Human placental poly (A)+m RNA was purchased from Ambion (Austin, TX).

Protocols

A detailed description of all three experimental protocols (MAS-B, MAS-N, Affy) is available in the Supplementary material. The MAS-N protocol yields in-vitro transcribed, biotin-labeled single-stranded cRNA, which is fragmented to an average size of 50-200 bp before hybridization. The MAS-B protocol yields Cy3-aminoallyl-labeled single-stranded cDNA (no fragmentation). The Affymetrix protocol yields end-labeled (bio-ddATP) double-stranded cDNA which is fragmented to an average size of 50-100 bp before hybridization.

Scoring schemes

To obtain the desired statistical resolution, MAS array scoring was done pooling the data from all three biological samples (for both placenta and NB4). For placenta data, this corresponds to seven measurements for each probe, and six for NB4. For Affymetrix, three technical replicates were used, corresponding to six measurements for each probe (three perfect match and three mismatch probes).

Sign test using array median intensity

The intensity of every probe within the window is compared to the median intensity of the slide and assigned a '1' if it is above and '0' otherwise. The number of ones within the window is counted and the probability *p* of finding at least this number of '1's under the null hypothesis that half of the probes should be above the median, is calculated. The score assigned to the probe in the middle of the window is then defined as score = -log(*p*). Window sizes of 90, 160, and 240 bp were tried, choosing 160 (corresponding to 5 probes) for the data to be presented in this study. No inter-array normalization is performed since each intensity is compared to the median intensity on its own array only.

Paired Wilcoxon signed rank sum test

Inter-array normalization is undertaken through dividing each intensity with the array median (median normalization). Within a window, all pairwise differences between the intensities of a perfect match probe and its corresponding mismatch probe are calculated and ranked. A sign is assigned to each rank number depending on whether the PM or the MM intensity was greater, and a p-value is calculated from the sum of this signed ranking (keeping track of the rank sum of all negative ranks and the rank sum of all positive ranks). The p-value, which is a measure of how significantly the distribution of PM-MM differences is skewed to either side around zero, can then be used to compute the final score for the probe in the middle of the window (Kampa et al, 2004; Royce et al., 2005). A related measure called the pseudomedian has been used (Cheng et al., 2005). The pseudomedian, or the Lehmann-Hodges estimator, which is a point estimator, is obtained by taking the median value of all the pairwise averages of PM-MM values within the sliding window. The Affymetrix scores used in this study were calculated by Affymetrix using a window size of 101 nucleotides, corresponding to on average 5 probes in the window.

Paired Wilcoxon for MAS arrays

The paired Wilcoxon signed rank test was applied as above, with the exception that the probe corresponding to the reverse strand of the exact same genomic locus was used as a mismatch probe in the calculations (instead of a designed mismatch probe).

Segmentation of genomic regions

Maxgap/minrun segmentation

The transcribed regions were generated from scored data. The maxgap parameter was set to 50 for Affymetrix data and 80 for MAS data. The minrun parameter was set to 50 for both approaches, thus at least four probes have to be included in an Affymetrix TAR and three for the maskless arrays. Other maxgap/minrun parameter settings were also tested (data not shown). We evaluated segmentation thresholds of 70-99th percentile. The parameters were chosen to in the end generate TAR sets of roughly similar size both in terms of the number of nucleotides within TARs.

HMM segmentation

The raw data were pre-scored by other methods (see above section) before being processed by HMM segmentation scheme. The emission distributions of the four-state HMM for each dataset were learned according to the scores of those probes which fall into known gene regions, where the score characteristics in the exon regions were used to estimate the parameters for the TAR state, and those in the intron regions for the non-TAR state. The parameters for the two intermediate transition states were obtained by investigating those probes containing both exon and intron regions. These emission distributions were then fitted with mixed-Gaussian distributions to generate a continuous model. The transition probabilities of the HMM were learned in a similar way. The Viterbi algorithm was utilized to identify TARs.

Choosing primer pairs for validation

Primer pairs were generated using Primer3 (Rozen and Skaletsky, 2000) (Supplemental material, Table S1). Primers assessing novel TARs were required to define a genomic region that did not overlap with any Gencode gene and the TARs required to have a minimum length of 120 bp. When assessing known genes, the exon with the highest p-value based transcription score was chosen. Primer3 settings were as default or more stringent, e.g., the GC content within 35-65%, and primer size was forced to be between 20 and 28 nucleotides, and the resulting PCR products were chosen to be between 100 and 200 bp. Validation candidates were checked using UCSC In Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr) against the Ncbi v35 human genome build (the latest at time of writing) to ensure that exactly one potential PCR product was possible; those that generated no or multiple hits were discarded. Three regions that did not contain any verified or predicted transcription were chosen to act as negative controls. The experimental protocol of the PCR validation is described in detail in Supplementary Material.

Accession numbers of array designs and hybridization experiments

The MAS ENCODE array platform has GEO accession number GPL2105; the corresponding data series has GEO accession number GSE2720 (placenta and untreated NB4). The Affymetrix anti-sense ENCODE array platform has GEO accession number GPL1789; the corresponding data series has accession number GSE2671 (placenta) and GSE2679 (untreated NB4).

ACKNOWLEDGEMENTS

O.E. is supported by a Knut and Alice Wallenberg Foundation postdoctoral fellowship. We acknowledge support from the NIH (1U01HG003156-01).

FIGURE LEGENDS

Figure 1:

Comparing human chromosome 22 transcription data sets with gene annotation. Transcription data sets were derived from previously published studies. They were generated from three different microarray platforms: PCR (red squares), MAS (blue diamond), and Affymetrix (green circles); in total 15 separate experiments (tissues or cell lines), each represented by a point in the figure . We used the RefSeq annotation as a benchmark to assess the quality of the data from each experiment. *X-axis*, the fraction of exonic probes that were identified to be transcribed in individual experiments (sensitivity). *Y-axis*, the fraction of transcribed probes overlapping with an exon (specificity). The PCR tiling array data were from placenta, fibroblast and B-cells (Rinn et al., 2003; White et al., 2004), the MAS data from liver (Bertone et al., 2004), and the Affymetrix sets were collected from Kapranov et al. (2002) representing 11 different cell lines. *Arrow*, the Affymetrix data from U87 cell line is not representative since a long section of chromosome 22 is identified as transcriptionally silent, suggesting this particular experiment probably did not work or something unusual about U87.

Figure 2:

(**A**) Number of nucleotides in placental TARs as a function of segmentation threshold (percentiles). TARs were generated with the maxgap/minrun algorithm based on the scored hybridization intensity data using a genomic window and technical replicates: MAS-B scored with standard sign test (*green*); MAS Fwd-Rev scoring using reverse strand as “mismatch” (*orange*), pseudomedian; Affymetrix scored using pseudomedian from PM-MM (*blue*). The data points corresponding to the data sets used in the Comparison section are circled: thresholds are 90th percentile for Affy and 91th percentile for MAS-B (sign test scoring). *X-axis*, the percentile score threshold for calling a probe “positive”. *Y-axis*, the number of nucleotides in TARs (in megabasepairs). The dashed line corresponds to the number of nucleotides in exons in the analyzed region (1,001,238 nts). (**B**) Positive predictive value (*PPV*) versus sensitivity for three different ways of scoring and segmenting the MAS-B data, varying the segmentation threshold from 70th percentile (to the right in the figure) to 99th percentile (to the left) for the MAS-B set scored with standard sign test (*green*); scored using reverse strand as “mismatch” (*orange*), pseudomedian; and the result from HMM segmentation (Viterbi decoding) of sign test-scored data (*grey triangle*). *Sensitivity* (*x*-axis), defined as the percentage of bases in Gencode exonic regions that are covered by a TAR. *PPV* (*y*-axis), defined as the percentage of bases in the TARs that overlap with a Gencode exonic region. (C) *PPV* versus sensitivity for two different ways of scoring the placenta Affy data: Wilcoxon signed rank test (*blue circles*), and standard sign test (*large cyan triangles* using PM-MM values; *yellow squares* using PM values only). The result from reducing the genomic density of the Affy array with 50% (ie., removing the data from every second probe) is also shown (*small cyan triangles* (PM-MM values uses)). (**D**) *PPV* versus sensitivity for MAS-B and Affy placenta data, varying the segmentation threshold from 70th percentile (to the right in the figure) to 99th percentile (to the left). The average results of TARs generated from raw intensities from single arrays for Affy (PM only (*blue squares*), and PM-MM (*blue triangles*)) and MAS-B (*green squares*) are plotted, as well as scored results for Affy (*blue circles*) and MAS-B (*green circles*). *Sensitivity* (*x*-axis), and *PPV* (*y*-axis), defined as above. The data points corresponding to the data sets used in the Comparison section are circled. The hatched area marks where a sensitivity of 30% is achieved for the various sets. (**E**) *PPV* for placental TAR sets when choosing a segmentation threshold that yields approximately 30% sensitivity (hatched area in (B)). Note that the actual sensitivity varies slightly between the sets.

Figure 3:

TAR set agreement. (**A**) Overlap of TAR sets, measured in number of overlapping nucleotides (kilobases). All three placenta TAR sets (MAS-B, MAS-N, Affy) and both NB4 TAR sets (MAS-N and Affy) . *R* is a measure of the size of the overlap. $R = |\cap| / |\cup|$ (calculated pairwise for the three placenta TAR sets). (**B)** Size of TAR set overlap, expressed in *R*, for comparisons within biological samples but across different array platforms (*black* lines), and comparisons within array platforms but across the biological samples (*brown* lines). Values in leftmost column of the graph are calculated with no further constraints. Second column, only TARs overlapping with conserved regions are included. Third column, only TARs overlapping with Gencode exons are included. Fourth column, only TARs overlapping with both conserved and exon regions included.

Figure 4:

Distribution of Gencode exon coverage by placenta TARs: all exons (MAS-B, g*reen squares,* and Affy, *blue squares*); 5' exons (Affy, *blue circles*); 3' exons (Affy, *blue triangles*). *X-axis*, the fraction to which an exon is covered by a TAR; 0.0-1.0 split up in 10 bins. *Y-axis*, the percentage of exons covered by a TAR to the fraction represented on the x-axis.

TABLES

Table 1

Outline of hybridization experiments.

--Experiment ID--		Sample	Number of technical replicates	Hybridization protocol
			batch1+batch2+batch3	
Placenta	MAS-B	Poly(A)+	3+2+2	Bertone et al. (2004)
Placenta	MAS-N	Poly(A)+	3+2+2	derived from manufacturer's
Placenta	Affy	Poly(A)+	3+2+2	manufacturer's
NB4	MAS-N	total RNA	2+2+2	derived from manufacturer's
NB4	Affy	total RNA	2+2+2	manufacturer's

Table 2

Correlation of hybridization intensities. Average of absolute values of Pearson correlation coefficients (R), calculated from unprocessed hybridization intensities (excluding internal standards and grid alignment probes). Between arrays: between technical and between biological replicates. Within arrays: for Affymetrix arrays, between corresponding perfect match (PM) and mismatch (MM) values; for MAS arrays, between probes representing forward/leading (Fwd) and reverse/lagging (Rev) strands of the same genomic location.

--Experiment ID--		---------------- Between arrays --------------		---------- Within arrays -------------	
		techn. repl.	biol. repl.	PM vs MM	Fwd vs Rev
Placenta	MAS-B	0.829	0.820	-	0.627
Placenta	MAS-N	0.955	0.953	-	0.046
Placenta	Affy	0.961	0.937	0.774	-
NB4	MAS-N	0.959	0.957	-	0.045
NB4	Affy	0.981	0.983	0.917	-

Table 3

Characteristics of TAR sets used in comparison (data for ENCODE regions ENm001-ENm011)

--Experiment ID--		Scoring method and		--Number of TARs and nucleotides--				Mean/	Gencode cmp.	
		Segmentation parameters		----Stranded------		---Unstranded---		Median	Sens.	PPV
		threshold/minrun/maxgap		#TARs	#bases	#TARs	#bases	length	(%)	(%)
Placenta	MAS-B	Sign test win.160	91/50/80	4079	955k	2545	684k	269/180	24.6	35.9
	MAS-N	Sign test win.160	92/50/80	3853	768k	3248	701k	216/144	22.3	31.7
	Affy	PM-MM P-median	90/50/50	-	-	3694	629k	170/105	37.0	58.6
NB4	MAS-N	Sign test win.160	93/50/80	3520	697k	2936	632k	216/144	19.1	30.2
	Affy	PM-MM P-median	87/50/50	-	-	4674	629k	135/91	26.5	41.8
Gencode exonic						2563	1018k	2482	1001k	403

Table 4

Percentage of genic and intergenic TARs that overlap with conserved regions (>90% of TAR length within conserved region) or that do not overlap with conserved regions (<10% of TAR length within conserved region).

			Intergenic TARs		**Genic TARs**	
--Experiment ID--		**Overlaps conserved region:**	**NO**	**YES**	**NO**	**YES**
Placenta	MAS-B		84%	2%	50%	8%
Placenta	Affy		77%	8%	41%	24%
NB4	MAS-N		85%	1%	54%	8%
NB4	Affy		79%	7%	49%	19%

Table 5

Transcribed placental genes (and in brackets: transcripts) in MAS-B and Affy experiments. A gene is considered as transcribed if at least one of its transcripts (splice variants) is transcribed at significance level p<0.001, using the sign test to score all transcripts. Probes that to at least 50% are within an exon of the transcript are considered. If a transcript/gene has less than 10 probes it will be unable to reach a p-value below 0.001, and is in the "Too few probes" category.

			Affy	
		Yes	**No**	**Too few probes**
	Yes	158 (871)	37 (155)	15 (58)
MAS-B	**No**	9 (13)	24 (61)	7 (13)
	Too few probes	2 (51)	7 (36)	5 (44)

[Total number of genes: 264; Total number of transcripts: 1342]

Table 6

Multi-exon coherence of transcripts with more than one exon. Percentage of transcripts where *all*, *some*, or *no* exons are considered transcribed according to median intensity of each exon. Exons were called on/off based on their median intensity compared with the experiment-specific score thresholds used for TAR generation (segmentation), specified in Table 3. The transcription status of entire transcripts was generated as in Table 5.

--Experiment ID--		**All exons**	**Some exons**	**No exons**
All transcripts (1298 transcripts):				
Placenta	MAS-B	6.3%	60.7%	33.0%
Placenta	MAS-N	3.4%	56.8%	39.8%
Placenta	Affy	36.8%	47.6%	15.6%
NB4	MAS-N	1.9%	51.7%	46.4%
NB4	Affy	15.5%	60.9%	23.7%
Intersection of placenta MAS-B and Affy transcribed transcripts (869):				

Placenta	MAS-B	7.8%	68.9%	23.2%
Placenta	Affy	50.1%	48.3%	1.6%
Intersection of NB4 MAS-N and Affy transcribed transcripts (543):				
NB4	MAS-N	2.9%	72.9%	24.1%
NB4	Affy	26.9%	61.9%	10.7%

Table 7

Results of experimental validation (reverse transcriptase PCR) in placenta of 142 regions: 96 novel TARs, 43 exons from known genes, 3 negative controls.

Set	**Transcription status**		**Overlap with**	**Number**	**Number**
	MAS-B	**Affy**	**known exon?**	**assessed**	**positive**
TAR	on	off	no	39	?
TAR	off	on	no	37	?
TAR	on	on	no	20	?
Genes	on	off	yes	17	?
Genes	off	on	yes	8	?
Genes	off	off	yes	18	?
Neg ctrl	off	off	no	3	?

REFERENCES

Albert, T.J., Norton, J., Ott, M., Richmond, T., Nuwaysir, K., Nuwaysir, E.F., Stengele, K.P., and Green, R.D. 2003. Light-directed 5'-->3' synthesis of complex oligonucleotide microarrays. *Nucl. Acids Res.* **31:** e35

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242-2246

Bertone, P., Trifonov, V., Rozowsky, J.S., Schubert, F., Emanuelsson, O., Karro, J., Kao, M.-Y., Snyder, M., and Gerstein, M. 2005. Design Optimization Methods for Genomic DNA Tiling Arrays. *Genome Res.* accepted for publication

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., and Miller, W. 2004. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.* **14:** 708-715

Brudno, M., Do, C., Cooper, G., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:**721-731

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499-509

Chen, J., Sun, M., Kent, W.J., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R.Z., and Rowley, J.D. 2004. Over 20% of human transcripts might form sense-antisense pairs. *Nucl. Acids Res.* **32:** 4812-4820

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149-1154

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636-640

Horak, C., and Snyder, M. 2002. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350:** 469-83

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., et al. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19:** 342-347

Jarvinen, A.-K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O.P., and Monni, O. 2004. Are data from different gene expression microarray platforms comparable? *Genomics* **83:** 1164-1168

Ji, H., and Wong, W.H. 2005. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21:** 3629-3636

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14:** 331-342

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916-919

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15:** 987-997

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap ,C.C., Suzuki, M., Kawai, J., et al. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564-1566

Kim, K., Page, G.P., Beasley, T.M., Barnes, S., Scheirer, K.E., Allison, D.B. 2006. A proposed meetric for assessing the measurement quality of individual microarrays. *BMC Bioinformatics* **7**: e35

Larkin, J.E., Frank, B.C., Gavras, H., Sultana, R., and Quackenbush, J. 2005. Independence and reproducibility across microarray platforms. *Nat. Methods* **2:** 337-343

Li, L., Wang, X., Xia, M., Stolc, V., Su, N., Peng, Z., Tongprasit, W., Li, S., Wang, J., Wang, X., Deng, X.W. 2005. Tiling microarray analysis of rice chromosome 10 to identify the transcriptome and relate its expression to chromosomal architecture. *Genome Biol.* **6:** R52

Li, W., Meyer, C.A., Liu, X.S. 2005. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21 (suppl. 1):** i274-i282

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. 1999. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21:** 20-24

Mah, N., Thelin, A., Lu, T., Nikolaus, S., Kuhbacher, T., Gurbuz, Y., Eickhoff, H., Kloppel, G., Lehrach, H., Mellgard, B., Costello, C.M., and Schreiber, S. 2004. A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol. Genomics* **16:** 361-370

Mathews, D.H., Burkard, M.E., Freier, S.M., Wyatt, J.R., and Turner, D.H. 1999. Predicting oligonucleotide affinity to nucleic acid targets. *RNA* **5:** 1458-1469

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E., and Ecker, J.R. 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85:** 1-15

Nazarenko, I., Pires, R., Lowe, B., Obaidy, M., and Rashtchian, A. {Effect of primary and secondary structure of oligodeoxyribonucleotides on the fluorescent properties of conjugated dyes. 2002. *Nucl. Acids Res.* **30:** 2089-2195

Nuwaysir, E.F., Huang, W., Albert, T.J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J.P., Ballin, J., et al. 2002. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* **12:** 1749-1755

Park, P.J., Cao, Y.A., Lee, S.Y., Kim, J.W., Chang, M.S., Hart, R., and Choi, S. 2004. Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J. Biotechnol.* **112:** 225-245

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.* **33:** D501-D504

Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev.* **17:** 529-540

Rouillard, J.M., Zuker, M., and Gulari, E. 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucl. Acids Res.* **31:** 3057-3062

Royce, T.E., Rozowsky, J.S., Bertone, P., Samanta, M., Stolc, V., Weissman, S., Snyder, M., and Gerstein, M. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet* . **21:** 466-475

Royce, T.E., Rozowsky, J.S., Luscombe, N.M., Emanuelsson, O., Yu, H., Zhu, X., Snyder, M., and Gerstein, M. 2006. Extrapolating traditional DNA microarray statistics to the tiling and protein microarrays technologies. *Methods Enzymol.* accepted for publication

Rozen, S., and Skaletsky, H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds. S. Krawetz and S. Misener), pp. 365-386. Humana Press, Totowa, NJ, USA

Schadt, E.E., Edwards, S.W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K.W., Russell, A., Li, G., et al. 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* **5:** R73

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922-927

Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., and Cerrina, F. 1999. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.* **17:** 974-978

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E., Bussemaker, H.J., and White, K.P. 2004. A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* **306:** 655-660

Tan, P.K., Downey, T.J., Spitznagel, E.L. Jr., Xu, P., Fu, D., Dimitrov, D.S., Lempicki, R.A., Raaka, B.M., and Cam, M.C. 2003. Evaluation of gene expression measurements from commercial microarray platforms. *Nucl. Acids Res.* **31:** 5676-5684

Van Gelder, R.N., von Zastrow, M.E., Yool, A., Dement, W.C., Barchas, J.D., and Eberwine, J.H. 1990. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. USA* **87:** 1663-1667

White, E.J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E.J., Weissman, S., Gerstein, M., Groudine, M., Snyder, M., and Schübeler, D. 2004. DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc. Natl. Acad. Sci. USA* **101**: 17771-17776

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., et al. 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302:** 842-846

Yauk, C.L., Berndt, M.L., Williams, A., Douglas, G.R. 2004. Comprehensive comparison of six microarray technologies. *Nucl. Adicds Res.* **32:** e124

Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21:** 379-386

FIGURES