Affymetrix® Mismatch (MM) Probes: Useful After All

Robert M. Flight
Department of Anatomical Sciences and Neurobiology
University of Louisville
Louisville, Kentucky USA 40292
robert.flight@louisville.edu

Abdallah M. Eteleeb and Eric C. Rouchka
Department of Computer Engineering and Computer Science
University of Louisville
Louisville, Kentucky USA 40292
ametel01@louisville.edu; eric.rouchka@louisville.edu

***Abstract*—Affymetrix® GeneChip® microarray design defines probe sets consisting of 11, 16, or 20 distinct 25 base pair (BP) probes for determining mRNA expression for a specific gene, which may be covered by one or more probe sets. Each probe has a corresponding perfect match (PM) and mismatch (MM) set. Traditional analytical techniques have either used the MM probes to determine the level of cross-hybridization or reliability of the PM probe, or have been completely ignored. Given the availability of reference genome sequences, we have reanalyzed the mapping of both PM and MM probes to reference genomes in transcript regions. Our results suggest that depending of the species of interest, 66%-93% of the PM probes can be used reliably in terms of single unique matches to the genome, while a small number of the MM probes (typically less than 1%) could be incorporated into the analysis. In addition, we have examined the mapping of PM and MM probes to five different human genome projects, resulting in approximately a 70% overlap of uniquely mapping PM probes, and a subset of 51 uniquely mapping MM probes commonly found in all five projects, 24 of which are found within annotated exonic regions. These results suggest that individual variation in transcriptome regions provides an additional complexity to microarray data analysis. Given these results, we conclude that the development of custom chip definition files (CDFs) should include MM probe sequences to provide the most effective means of transcriptome analysis of Affymetrix® GeneChip® arrays.**

Keywords— Bioinformatics, microarray, probe set, custom definition files

I. INTRODUCTION

Oligonucleotide-based microarray technologies provide a methodology whereby a researcher can indirectly measure the expression level of an mRNA molecule being actively transcribed under a set of conditions by labeling a cDNA fragment that hybridizes to a complementary probe sequence specific to a particular transcript. Since their first use on customized cDNA arrays [1] in the mid-1990s, they have been used as the de-facto standard for measuring global transcriptional changes under differing conditions. While RNA-Seq [2] may eventually supplant microarrays as the method of choice, a large number of microarray experiments exist that have been deposited into publicly available repositories such as NCBI's Gene Expression Omnibus (GEO) [3] and EBI's ArrayExpress [4]. As a case in point, GEO contains 32,471 series as of 9/6/2012. The majority of the entries in GEO were performed on arrays designed by companies such as Affymetrix®, Inc. (Santa Clara, CA), Agilent Technologies, Inc. (Santa Clara, CA), Illumina®, Inc. (San Diego, CA), and GE Healthcare Lifesciences (Piscataway, NJ), with nearly half of the series (16,181) being performed on various Affymetrix® arrays.

The design of Affymetrix® GeneChip® arrays in particular provides for probe sets consisting of 11, 16, or 20 distinct 25 base pair (BP) probes, with each probe having a corresponding perfect match (PM) and mismatch (MM) probe. The PM and MM differ by the exchange of the complementary base at the 13th position in the probe. While MM probes were originally designed to account for signal in the PM resulting from non-specific cross-hybridization, they are often underutilized or completely ignored. Mismatch probes have been explored for use in long oligonucleotide arrays as well [5], but their utilization is limited to the Affymetrix® platform.

Affymetrix® provides a default GeneChip® analysis package known as the Micro Array Suite 5.0 (MAS 5.0) [6] that measures the signal intensity for a particular probe pair as:

$$signal = TukeyBiweight\{\log(PM_j - MM_j^*)\} \quad (1)$$

Where MM^* is a modified version of MM that is never bigger than the intensity value of the PM. The motivation behind the modified mismatch intensity MM^* is to report all probe-level intensities as positive values, and to remove the influence of the minority of probes where the MM intensity value is significantly higher than the corresponding PM intensity. In addition to the intensity signal, MAS 5.0 also produces a detection p-value which flags a transcript as "P" (present), "M" (marginal), or "A" (absent) based on the reliability of the probe set based on differences between PM and MM intensities.

Known issues in the use of PM and MM probe intensities to generate a single probe set intensity values led to the development of other approaches, including RMA [7] and GCRMA [8] which completely ignore the MM probes.

978-0-7695-4938-5/12 $26.00 © 2012 IEEE
DOI 10.1109/BioMedCom.2012.8

With the availability of individual probe and reference genome sequences, it is possible to re-map probes based on new sources of genome annotations. This allows custom Chip Description Files (CDFs) wherein probes are grouped into novel probe sets based on exon, transcript, and gene level annotation [9-22]. Most notable is the effort of the BrainArray group [10] which updates custom CDFs for a large number of Affymetrix® GeneChips® by creating probe sets based on annotated features such as Entrez Gene [23], Ensembl transcript, Ensembl gene, and RefSeq Gene [24]. Using custom CDFs has been shown to impact the reliability of expression analysis [10, 20-22, 25]. However, to the authors' knowledge, only the PM probe sequences are used when generating custom CDFs.

Based on the observation that a small, yet significant number of PM-MM probe pairs exist where the MM intensity is significantly increased over the PM intensity, our initial inclination was that these differences in intensities were not due to cross-hybridization or rogue probes alone. Therefore, keeping in mind that Affymetrix® probes have been designed according to continually evolving genome assemblies, we proceeded to analyze PM and MM probes across eight commonly studied species (Table I) by looking at PM and MM probes that uniquely map to the respective genome.

In addition to changing functional annotations, one potential problem area for microarray probe design is the presence of single nucleotide polymorphisms (SNPs) within a population. As the probes are designed using a reference genome or transcriptome, a "one-size fits all" approach has been taken for the probes on a particular array. However, SNPs are known to occur relatively frequently throughout the genome, with build 137 of dbSNP [26] containing over 53.5 million reference SNPs for the human genome. We have previously studied the effects of SNPs on Affymetrix® GeneChips® [27] showing that a large number of SNPs lie in the areas where microarray probes have been designed. This has been taken into account in the BrainArray's custom CDF files which incorporate SNP information. To study the effects that individual variation can play in microarray analysis, we looked at the mappings of PM and MM probes within five distinct publicly available assemblies of human genomes.

II. METHODS

A. Mapping of PM and MM Probes

Chromosomal-based genome assemblies were downloaded from the UCSC Goldenpath Genomes ftp server using an anonymous login (ftp://hgdownload.cse.ucsc.edu/goldenPath/) [28] for eight commonly studied species, including *C. elegans* (roundworm), *D. melanogaster* (fruit fly), *S. cerevisiae* (baker's yeast), *X. tropicalis* (western clawed frog), *D. rerio* (zebrafish), *M. musculus* (house mouse), *R. norvegicus* (brown Norway rat), and *H. sapiens* (human) (Table I). Genome indices were created using `bowtie-build` version 0.12.8 [29] with the default parameters. Perfect match (PM) probe sequences for Affymetrix® GeneChips® were obtained from Bioconductor (v 2.10) probe packages, which are constructed from data available in NetAffx [30] (Table II) with each new Bioconductor release. Mismatch (MM) probe sequences were constructed by replacing the 13th base in the supplied PM probe sequence with the complementary base. PM and MM probes were aligned to the indexed genomes using `bowtie` version 0.12.8 [29] with the parameters `-v 0` and `-a` which used together will report all valid probes matching with 100% identity.

B. Generation of Exons and Overlap

Exon regions were obtained from the UCSC genome browser as BED files with an entry for each exon. `mergeBed` from the `bedTools` suite was used to merge overlapping exons from multiple transcripts into single contiguous exons. These merged exons were used when defining overlaps of probe alignments with an exon. Probe and exon overlaps were defined as any type of overlap with at least 23 bases overlapping on the same strand. Overlaps were determined using `Genomic Ranges` version 1.8.7 [31].

C. DNA Microarray Data

For each GeneChip®, CEL files were downloaded from GEO for 20 random samples (with the exception of *S. cerevisiae* (12) and *X. tropicalis* (4), a list of GSMs is available upon request). Probe intensities were background corrected using the MAS background correction method implemented in Bioconductor. Depending on the application, intensities were log (base 2), square root transformed, or used as is.

D. Negative PM-MM Set

A PM-MM set of probes was considered to be negative if nine (two for *S. cerevisiae* and six for *X. tropicalis*) or more samples had a negative value for the difference in the PM-MM intensities. For examination of intensity distribution, any PM-MM pair with a negative difference greater than 1000 in one or more samples was considered and examined.

E. Probe Correlations

For each MM probe that uniquely overlapped one merged exon (designated as a true-match MM, (TM_{mm})), the correlation with all other MM probes in the probe set (mm)

TABLE I. GENOME ASSEMBLIES USED

Organism	Reference Assembly	Build Date
Caenorhabditis elegans	ce6	May 2008
Drosophila melanogaster	dm3	Apr. 2006
Saccaromyces cerevisiae	sc3	Apr. 2011
Xenopus tropicalis	xt3	Nov. 2009
Danio rerio	dr6	Dec. 2008
Mus musculus	mm10	Dec. 2011
Rattus norvegicus	rn4	Nov. 2004
Homo sapiens	hg19	Feb. 2009

TABLE II. AFFYMETRIX® GENECHIPS® USED

Organism	GeneChip® Name
Caenorhabditis elegans	*C. elegans* Genome
Drosophila melanogaster	*Drosophila* Genome 2.0
Saccaromyces cerevisiae	Yeast Genome 2.0
Xenopus tropicalis	*Xenopus tropicalis* Genome
Danio rerio	Zebrafish Genome
Mus musculus	Mouse Genome 430 2.0
Rattus norvegicus	Rat Genome 230 2.0
Homo sapiens	Human Genome U133 Plus 2.0

and the correlation with all other TM probes that also mapped uniquely to the same exon (if there were three or more other probes also mapped to the exon) was calculated (tm).

F. Human Variation

To gain an understanding of individual variation and the unique mapping of microarray probes, five whole genome assemblies were downloaded for the human genome [32-36] (Table III). Probes from the HGU133APlus2.0 Affymetrix® GeneChip® were aligned to each of these genomes using the methods previously described for mapping PM and MM probes.

III. RESULTS

A. Probes Matching Genomic Locations

Given the PM probe sequences and the inferred MM sequences, individual probes were mapped to the corresponding genome assembly as outlined in Methods. The percentage of perfect match probes mapping to the genome ranged from a low of 80% (*X. tropicalis*) to a high of 95% (*D. melanogaster*), with the exception of *S. cerevisiae* (Table IV). It must be noted that the lower percentage of *S. cerevisiae* matches (53%) is expected, as the Yeast Genome 2.0 GeneChip® contains probes for two yeast species, *S. cerevisiae* and *S. pombe*.

Affymetrix® probesets are given suffix definitions depending upon the uniqueness of the exemplar sequence used

TABLE III. WHOLE HUMAN GENOME SEQUENCING PROJECTS

Name	Abbr.	Assembly Identifier	Bioproject Number	Race
GRCh37	Hg19	420368	31257	Mixed
Hs_Celera_WGSA	Celera	281338	1431	Mixed[1]
HuRefPrime	JCVI	281188	19621	Caucasian
BGIAF	BGI	165398	42201	African
HsapALLPATHS1	HSAP1	238948	59877	Caucasian

[1]Celera assembly consists of one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians.

TABLE IV. GENECHIP® PROBES MAPPING TO REFERENCE GENOMES

		Number of Probes Mapped to Reference Genome		Number of Probes Mapping Uniquely to Genome	
Organism	Number of Probe Pairs	PM	MM	PM	MM
Ce	249,165	226,856	143	213,745	96
Dm	265,400	251,602	89	245,712	54
Sc	120,855	63,731	1	61,942	1
Xt	648,548	519,177	1,884	426,237	1,014
Dr	249,752	200,608	1,282	171,282	726
Mm	496,468	456,432	555	427,639	396
Rn	342,410	304,646	391	286,784	282
Hs	604,258	562,673	1,094	521,642	608

to design a probe set. A designation of `_at` indicates the probe set perfectly matches a single transcript; `_a_at` probe sets only perfectly match transcripts of the same gene; `_s_at` perfectly match multiple transcripts for the same gene family; and `_x_at` indicates the probe set is identical or highly similar to other genes. One of the difficulties with these designations is that it relies upon a set of annotations at a particular point in time.

Analysis of the probes that map to the genome (Table IV) indicates that 82% to 98% of the mapped probes map uniquely to a single genomic location. The fact that a number of probes map to multiple locations is not to be unexpected due to the restrictions placed on probe set design. However, it is expected that those probes mapping to multiple locations would not be from the `_at` class of probes.

To determine the reliability of these probes with the fluctuation of unknown transcripts, those probes that map with 100% identity to two or more locations in the genome were considered. While these probes typically represent less than 10% of the total number of probes for a given GeneChip®, their classification could be important in detecting cross hybridization. One might expect that the greatest percentage

TABLE V. PROBE SET CLASSIFICATION OF PROBES PERFECTLY MATCHING MULTIPLE GENOMIC LOCATIONS

Organism	_x_at	_s_at	_at	_a_at	control
Ce	4,040	6,416	2,465	0	190
Dm	361	2,742	2,520	224	43
Sc	81	1,126	271	0	311
Xt	14,092	12,086	31,754	34,877	131
Dr	1,376	374	25,875	1,203	494
Mm	4,075	2,920	16,703	4,893	163
Rn	440	394	16,127	805	96
Hs	9,402	10,634	19,961	803	231

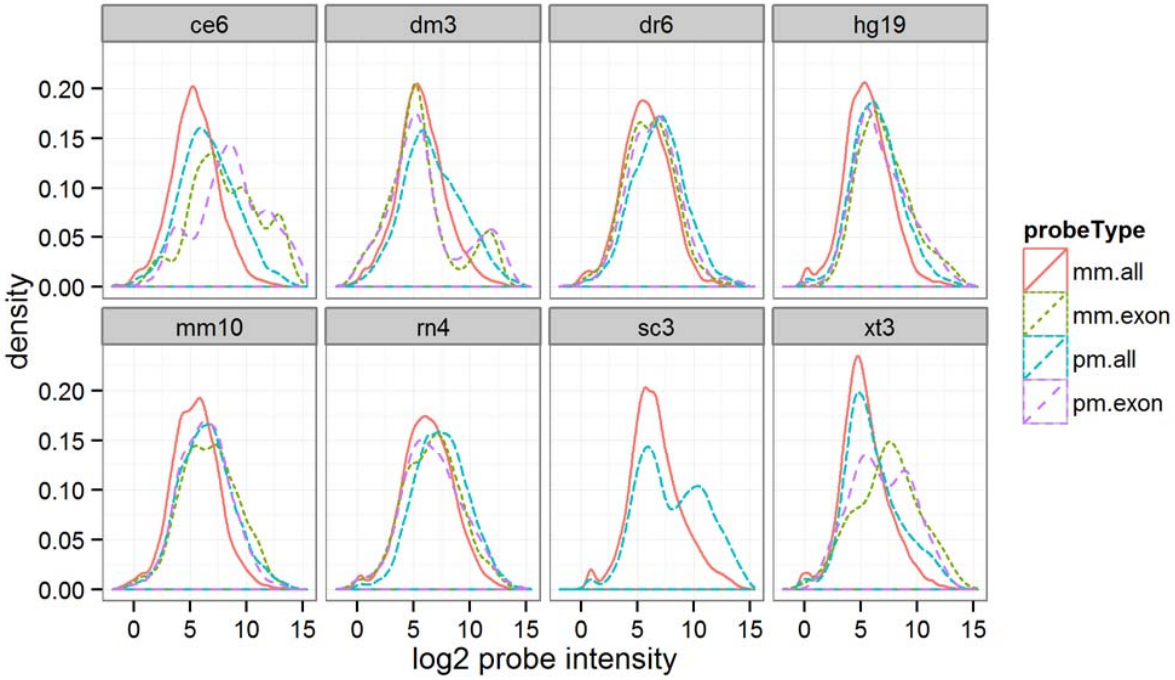

Figure 1. Density profile of probe intensities ($\log_2$). PM: perfect match, MM: mismatch. Mm.all: background MM intensities; pm.all: background PM intensities; mm.exon: intensities of MM probes in exonic regions; pm.exon: intensities of PM probes in exonic regions.

of these would be within the `_x_at` and `_s_at` classes. However, as Table V shows, the larger genomes actually contain the greatest percentage in the `_at` and `_a_at` classes, with anywhere from 18% (*S. cerevisiae*) to 91% *(R. norvegicus*) of the probes matching multiple locations belonging to the `_at` class. In addition, a small number of MM probes map to the genome as well. To better understand the effects this small set of MM probes might have on gene expression, we further reduced this to a smaller subset where the mapping was within exon regions. For these probes, we analyzed their signal intensities from random samples compared to the overall distribution of PM and MM intensities, and the distribution of PM and MM intensities within the corresponding exonic sequences (Figure 1). As these plots indicate, MM probes mapping within exonic regions closely follow the expression density of PM probes mapping within exonic regions, and are significantly shifted from the overall expression profiles of MM probes. These results suggest that while the number of these probes is small, they offer significant information that should not be ignored, and furthermore, can confound analyses where MM data is incorporated.

As some of these MM probes may bind to transcripts, we further considered those MM probes that uniquely mapped to exons (irrespective of whether the corresponding PM probe mapped zero, one or multiple times to exons or the full genome), examining the differences in signal intensity between the MM and its associated PM. In many cases there is a significant negative difference in the expression level of the PM-MM pair (an example is shown in Figure 2).

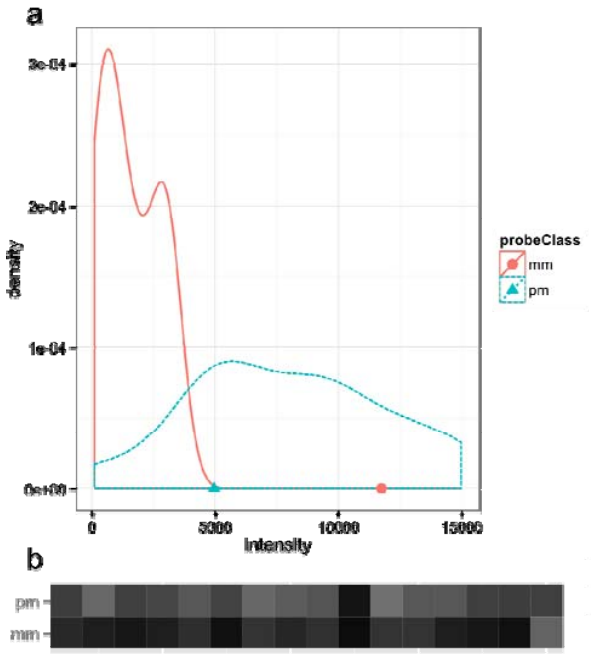

Figure 2. (a) Density of MM and PM probe set intensities not including the PM-MM pair that had a large negative difference. (b) Square root transformed intensities for each PM-MM pair. The negative difference pair is at the extreme right end of the figure. Intensities from the zebrafish probe set Dr.5545.1.S1_at, in GEO sample GSM604808.

TABLE VI. CORRELATIONS AND ANNOTATIONS OF TMMM PROBES BASED ON THE AFFYMETRIX® HGU133APLUS2.0 PROBE SET, AND TM PROBES TO COMMON EXONS

Correlation					Annotated	Exon
TM	MM	Probe ID	Annotated RefSeq	Exon RefSeq	Symbol	Symbol
0.87	0.53	209135_at.763.274	NM_001164750, NM_001164751, NM_001164752, NM_001164753, NM_001164754, NM_001164755, NM_001164756, NM_004318, NM_020164, NM_032466, NM_032467, NM_032468	NM_032466, NM_032468, NM_001164755, NM_001164754, NM_001164753, NM_001164752, NM_001164751	ASPH	ASPH
0.86	0.61	204041_at.895.14	NM_000898	NM_000898	MAOB	MAOB
0.84	0.60	206432_at.534.594	NM_005328	NM_005328	HAS2	HAS2
0.81	0.55	205004_at.1030.1044	NM_001173487, NM_001173488, NM_017544	NM_001173488, NM_001173487, NM_017544	NKRF	NKRF
0.81	0.78	201622_at.1129.72	NM_014390	NM_014390	SND1	SND1
0.79	0.82	235958_at.810.810	NM_213600, NR_033151	NM_213600, NR_033151	PLA2G4F	PLA2G4F
0.79	0.74	233052_at.882.6	NM_001206927, NM_001371	NM_001206927	DNAH8	DNAH8
0.77	0.36	206084_at.139.1104	NM_001207015, NM_001207016, NM_002849, NM_130846	NM_002849, NM_001207015, NM_130846, NM_001207016	PTPRR	PTPRR
0.75	0.40	217416_x_at.747.452		NM_152924, NM_007011		ABHD2
0.75	0.55	203646_at.468.274	NM_004109	NM_004109	FDX1	FDX1
0.74	0.61	210467_x_at.1162.450	NM_001166386, NM_001166387, NM_005367	NM_004988	MAGEA12	MAGEA1
0.72	0.64	207687_at.947.968	NM_005538	NM_005538	INHBC	INHBC
0.71	0.49	211741_x_at.916.1066	NM_021016	NM_002781, NM_001130014	PSG3	PSG5
0.67	0.41	211493_x_at.351.512	NM_001128175, NM_001198938, NM_001198939, NM_001198940, NM_001198941, NM_001198942, NM_001198943, NM_001198944, NM_001198945, NM_001390, NM_001391, NM_001392, NM_032975, NM_032978, NM_032979, NM_032980, NM_032981	NM_001198939, NM_001198940, NM_001390, NM_032975, NM_001198938, NM_001198944, NM_001198943, NM_001198942, NM_032980	DTNA	DTNA
0.66	0.55	219337_at.935.290	NM_017891	NM_017891	C1orf159	C1orf159
0.66	0.66	221351_at.390.354	NM_000524	NM_000524	HTR1A	HTR1A
0.65	0.55	238916_at.697.790	NR_028408	NR_028408	LOC400027	LOC400027
0.65	0.49	222221_x_at.912.300	NM_006795	NM_014600	EHD1	EHD3
0.64	0.42	203399_x_at.917.106	NM_021016	NM_002781, NM_001130014	PSG3	PSG5
0.64	0.78	201844_s_at.803.352	NM_012234	NM_012234	RYBP	RYBP
0.64	0.34	240239_at.132.1114	NM_001145343, NM_001145344, NM_001145345, NM_032838	NM_001145344, NM_001145343, NM_032838, NM_001145345	ZNF566	ZNF566
0.63	0.38	201220_x_at.905.256	NM_001083914, NM_001329, NM_022802	NR_003682	CTBP2	MGC70870
0.63	0.24	210835_s_at.906.256	NM_001083914, NM_001329, NM_022802	NR_003682	CTBP2	MGC70870
0.59	0.55	223485_at.745.316	NM_032304, NM_207112	NM_032304	HAGHL	HAGHL
0.59	0.45	206281_at.529.784	NM_001099733, NM_001117	NM_001117, NM_001099733	ADCYAP1	ADCYAP1
0.59	0.56	1562659_at.440.914	NR_033984	NR_033984	LOC400548	LOC400548
0.58	0.52	229852_at.330.998	NM_022787	NM_022787	NMNAT1	NMNAT1
0.57	0.25	1553901_x_at.258.150	NM_052852	NM_178558	ZNF486	ZNF680
0.56	0.53	220547_s_at.462.628	NM_019054	NM_019054	FAM35A	FAM35A
0.56	0.56	209811_at.257.468	NM_001224, NM_032982, NM_032983	NM_032982, NM_032983, NM_001224	CASP2	CASP2
0.54	0.60	219710_at.1132.106	NM_024577	NM_024577	SH3TC2	SH3TC2
0.50	0.38	223838_at.145.282	NM_025244, NM_182911	NM_182911, NM_025244	TSGA10	TSGA10
0.49	0.64	221691_x_at.746.320	NM_001037738, NM_002520, NM_199185	NR_036693, NM_001004419, NM_001197317, NM_001197318, NM_001197319, NM_013269	NPM1	CLEC2D
0.45	0.62	200724_at.575.1144	NM_001256577, NM_001256580, NM_006013, NR_026898	NM_001256580, NM_001256577, NM_006013	RPL10	RPL10
0.40	0.52	204431_at.1040.48	NM_001144761, NM_001144762, NM_003260	NM_003260, NM_001144761, NM_001144762	TLE2	TLE2
0.39	0.10	217547_x_at.860.982	NM_138330	NM_016220, NM_001013746	ZNF675	ZNF107
0.33	0.44	228128_x_at.1075.350	NM_002581	NM_002581	PAPPA	PAPPA

If these MM probes are grouped instead with the other probes within the transcriptional region for which they uniquely match (we have renamed these probes as "true match" (TM) probes since they truly match the region in the genome), there is a much better association between the probe intensities, as shown in Figure 3.

Figure 3. Plot of probe set intensities for zebrafish where the TM probes overlap with the same TM_{mm} probe in Figure 2.

Further analysis was performed to test the correlation of the TM probes with the expression levels of both the annotated probe group MM probes and with the TM-mapped transcript probes (Figure 4). The box plot in Figure 4 clearly indicates that the TM intensities more closely correlate with those from the group based on mapping to the same exon.

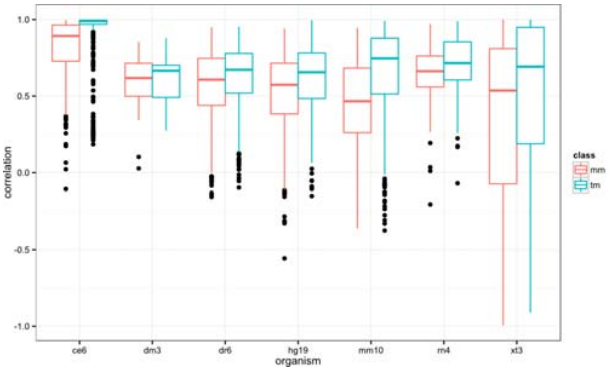

Figure 4. Box plot of correlations of the TM_{mm} with MM intensities of annotated probe set (red, MM) and TM intensities of custom probe sets based on shared mapping to exons (blue, TM).

To determine if the observed difference in the correlations from Figure 4 is due to measurement of different mRNA entities, we considered the MM probe both within its annotated location as well as within the new mapped location (TM_{mm}). The intensity of the MM probe was compared with the intensity of its corresponding neighbor probes (MM probes for the annotated location; PM probes for the new mapped location). An average correlation value between intensities was calculated on a per-exon probe basis. The resulting correlations are summarized in Table VI.

What is interesting is that with few exceptions, the transcripts and genes being measured by the probe sets are the same, implying that the TM_{mm} probe aligns to the same gene as its complimentary PM probe. Upon further examination it appears that many of the PM complements of the TM_{mm} do not align to the genome at all (data not shown). One possibility is that these probes are in regions that have seen changes in the reference sequence over the years, or that the original sequencing of the ESTs used to design the probe sets was of poor quality.

TABLE VII. NUMBER OF PROBES MAPPING UNIQUELY TO INDIVIDUAL HUMAN GENOMES

Assembly	Number of Uniquely Mapped Probes		
	Total	Perfect Match (PM)	Mismatch (MM)
Hg19	522,250	521,642	608
Celera	515,111	514,518	593
JCVI	530,213	529,569	644
BGI	469,973	469,714	259
HSAP1	522,480	521,922	558

Effects of Individual Variation

To gain an understanding of the effect of individual variation, the unique mapping of PM and MM probes to five distinct human genome assemblies was analyzed (Table VII). Four of the five projects have roughly the same number of uniquely mapped PM probes (within 2% variation). The fifth project (BGI) provides an exception to this trend. While there are a number of potential explanations for this (including sequence and assembly quality and coverage of the sequencing), one potential feature to be considered is the fact that this sequencing project involves the sequencing of an African individual, and it is the only project not to have a large component of the library consisting of Caucasian individuals.

While there is a large agreement for the number of probes, we also checked if the probes represented were consistent among all of the projects. A Venn diagram depicting the number of overlapping perfectly matching probes is given in Figure 5. As can be seen from this figure, a total of 422,279 probes uniquely map for all five assemblies. Of these, 422,119 are perfect match probes, indicating that 81% of the HGU133APlus2.0 perfect match probes are reliable in terms of their mapping to the genome for these assemblies. One of the interesting results is that there are 161 shared perfectly matching mismatch probes. Of these, 24 fall within RefSeq annotated exonic regions (results not shown), with 16 of the 24 showing higher correlation in the TM_{mm} assignments calculated in Table VI.

IV. CONCLUSION

MM probes are theoretically designed to capture background and non-specific binding. Alignment of the MM probes to the genome shows that in a very small percentage of cases, MM probes align uniquely to the genome in transcribed regions. Signal from these probes should be useful for quantifying true transcriptional events rather than for PM signal adjustment.

In addition, current custom CDF generation workflows ignore the MM probes during the probe alignment process. Given that some MM probes align to reference genomes, they should be considered for inclusion when creating custom CDFs. The utility of the probes may be limited due to variation among individuals.

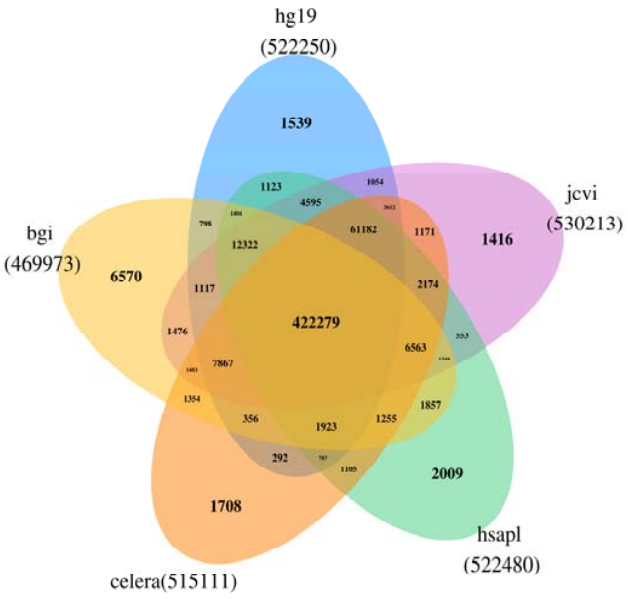

Figure 5. Venn diagram of overlapping perfectly matching Affymetrix® HGU133APlus2.0 probes to each of the five human genome assemblies.

ACKNOWLEDGEMENTS

This work was partially funded by National Institutes of Health (NIH) grant 8P20GM103436-12. Its contents are solely the responsibility of the authors and do not represent the official views of NIH or the National Institute of General Medical Sciences.

AVAILABILITY

All code used in this analysis is available at https://github.com/rmflight/affymm

REFERENCES

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science,* vol. 270, pp. 467-70, Oct 20 1995.

[2] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science,* vol. 320, pp. 1344-9, Jun 6 2008.

[3] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets--10 years on," *Nucleic Acids Res,* vol. 39, pp. D1005-10, Jan 2011.

[4] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar, and A. Brazma, "ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments," *Nucleic Acids Res,* vol. 39, pp. D1002-4, Jan 2011.

[5] Y. Deng, Z. He, J. D. Van Nostrand, and J. Zhou, "Design and analysis of mismatch probes for long oligonucleotide microarrays," *BMC Genomics,* vol. 9, p. 491, 2008.

[6] E. Hubbell, W. M. Liu, and R. Mei, "Robust estimators for expression analysis," *Bioinformatics,* vol. 18, pp. 1585-92, Dec 2002.

[7] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Res,* vol. 31, p. e15, Feb 15 2003.

[8] Z. Wu and R. A. Irizarry, "Stochastic models inspired by hybridization theory for short oligonucleotide arrays," *J Comput Biol,* vol. 12, pp. 882-93, Jul-Aug 2005.

[9] S. L. Carter, A. C. Eklund, B. H. Mecham, I. S. Kohane, and Z. Szallasi, "Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements," *BMC Bioinformatics,* vol. 6, p. 107, 2005.

[10] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data," *Nucleic Acids Res,* vol. 33, p. e175, 2005.

[11] W. C. de Leeuw, H. Rauwerda, M. J. Jonker, and T. M. Breit, "Salvaging Affymetrix probes after probe-level re-annotation," *BMC Res Notes,* vol. 1, p. 66, 2008.

[12] F. Ferrari, S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G. A. Danieli, and S. Bicciato, "Novel definition files for human GeneChips based on GeneAnnot," *BMC Bioinformatics,* vol. 8, p. 446, 2007.

[13] W. Langer, F. Sohler, G. Leder, G. Beckmann, H. Seidel, J. Grone, M. Hummel, and A. Sommer, "Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer," *BMC Genomics,* vol. 11, p. 676, 2010.

[14] H. Liu, B. R. Zeeberg, G. Qu, A. G. Koru, A. Ferrucci, A. Kahn, M. C. Ryan, A. Nuhanovic, P. J. Munson, W. C. Reinhold, D. W. Kane, and J. N. Weinstein, "AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets," *Bioinformatics,* vol. 23, pp. 2385-90, Sep 15 2007.

[15] J. Lu, J. C. Lee, M. L. Salit, and M. C. Cam, "Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays," *BMC Bioinformatics,* vol. 8, p. 108, 2007.

[16] A. G. Moll, M. T. Lindenmeyer, M. Kretzler, P. J. Nelson, R. Zimmer, and C. D. Cohen, "Transcript-specific expression profiles derived from sequence-based analysis of standard microarrays," *PLoS One,* vol. 4, p. e4702, 2009.

[17] F. Moreews, G. Rauffet, P. Dehais, and C. Klopp, "SigReannot-mart: a query environment for expression microarray probe re-annotations," *Database (Oxford),* vol. 2011, p. bar025, 2011.

[18] P. B. Neerincx, H. Rauwerda, H. Nie, M. A. Groenen, T. M. Breit, and J. A. Leunissen, "OligoRAP - an Oligo Re-Annotation Pipeline to improve annotation and estimate target specificity," *BMC Proc,* vol. 3 Suppl 4, p. S4, 2009.

[19] A. Risueno, C. Fontanillo, M. E. Dinger, and J. De Las Rivas, "GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs," *BMC Bioinformatics,* vol. 11, p. 221, 2010.

[20] R. Sandberg and O. Larsson, "Improved precision and accuracy for microarrays using updated probe set definitions," *BMC Bioinformatics,* vol. 8, p. 48, 2007.

[21] J. Yin, S. McLoughlin, I. B. Jeffery, A. Glaviano, B. Kennedy, and D. G. Higgins, "Integrating multiple genome annotation databases improves the interpretation of microarray gene expression data," *BMC Genomics,* vol. 11, p. 50, 2010.

[22] H. Yu, F. Wang, K. Tu, L. Xie, Y. Y. Li, and Y. X. Li, "Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data," *BMC Bioinformatics,* vol. 8, p. 194, 2007.

[23] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res,* vol. 39, pp. D52-7, Jan 2011.

[24] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy," *Nucleic Acids Res,* vol. 40, pp. D130-5, Jan 2012.

[25] X. Lu and X. Zhang, "The effect of GeneChip gene definitions on the microarray study of cancers," *Bioessays,* vol. 28, pp. 739-46, Jul 2006.

[26] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res,* vol. 40, pp. D13-25, Jan 2012.

[27] E. C. Rouchka, A. W. Phatak, and A. V. Singh, "Effect of single nucleotide polymorphisms on Affymetrix match-mismatch probe pairs," *Bioinformation,* vol. 2, pp. 405-11, 2008.

[28] T. R. Dreszer, D. Karolchik, A. S. Zweig, A. S. Hinrichs, B. J. Raney, R. M. Kuhn, L. R. Meyer, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, A. Pohl, V. S. Malladi, C. H. Li, K. Learned, V. Kirkup, F. Hsu, R. A. Harte, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. James Kent, "The UCSC Genome Browser database: extensions and updates 2011," *Nucleic Acids Res,* vol. 40, pp. D918-23, Jan 2012.

[29] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol,* vol. 10, p. R25, 2009.

[30] G. Liu, A. E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, and M. A. Siani-Rose, "NetAffx: Affymetrix probesets and annotations," *Nucleic Acids Res,* vol. 31, pp. 82-6, Jan 1 2003.

[31] H. Aboyoun, H. Pages, and M. Lawrence, "GenomicRanges: Representation and manipulation of genomic intervals. ," 1.8.7 ed.

[32] S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proc Natl Acad Sci U S A,* vol. 108, pp. 1513-8, Jan 25 2011.

[33] S. Istrail, G. G. Sutton, L. Florea, A. L. Halpern, C. M. Mobarry, R. Lippert, B. Walenz, H. Shatkay, I. Dew, J. R. Miller, M. J. Flanigan, N. J. Edwards, R. Bolanos, D. Fasulo, B. V. Halldorsson, S. Hannenhalli, R. Turner, S. Yooseph, F. Lu, D. R. Nusskern, B. C. Shue, X. H. Zheng, F. Zhong, A. L. Delcher, D. H. Huson, S. A. Kravitz, L. Mouchard, K. Reinert, K. A. Remington, A. G. Clark, M. S. Waterman, E. E. Eichler, M. D. Adams, M. W. Hunkapiller, E. W. Myers, and J. C. Venter, "Whole-genome shotgun assembly and comparison of human genome assemblies," *Proc Natl Acad Sci U S A,* vol. 101, pp. 1916-21, Feb 17 2004.

[34] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y. J. Chen, "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, pp. 860-921, Feb 15 2001.

[35] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y. H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and J. C. Venter, "The diploid genome sequence of an individual human," *PLoS Biol,* vol. 5, p. e254, Sep 4 2007.

[36] R. Li, Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li, G. Zhou, X. Zhu, H. Wu, J. Qin, X. Jin, D. Li, H. Cao, X. Hu, H. Blanche, H. Cann, X. Zhang, S. Li, L. Bolund, K. Kristiansen, H. Yang, and J. Wang, "Building the sequence map of the human pan-genome," *Nat Biotechnol,* vol. 28, pp. 57-63, Jan 2010.