Surveying genome-wide levels of sequence divergence between species using comparative genomic hybridization: A proof-of-concept from *Drosophila*

Suzy C.P. Renn[1], Heather E. Machado[1], Kosha Soneji[2], Rob Kulathinal[3,4], Hans A. Hofmann[5]

[1]Department of Biology, Reed College, Portland, OR 97202
[2]Boston University School of Medicine, Boston MA 02118
[3]Department of Organismic & Evolutionary Biology, Harvard University, Cambridge, MA, 01730
[4]Department of Biology, Temple University, Philadelphia, PA, 19122
[5]Section of Integrative Biology, Institute for Molecular and Cellular Biology, Institute for Neuroscience, The University of Texas at Austin, Austin, TX, 78712

KEY WORDS: microarray, heterologous, adaptation, divergence, comparative genomics

ABSTRACT

Genome-wide analysis of sequence divergence among species offers profound insight regarding the evolutionary processes that shape species. When full-genome sequencing is not feasible for a broad comparative study, we propose the use of array-based comparative genomic hybridization (aCGH) in order to identify orthologous genes with high sequence divergence. Here we discuss experimental design, statistical power, success rate, sources of variation and potential confounding factors. We used a spotted PCR product microarray platform from *Drosophila melanogaster* to assess sequence divergence on a gene-by-gene basis in three fully sequenced heterologous species (*D. sechellia*, *D. simulans*, and *D. yakuba)*. Because full genome sequence is available for these species this study presents a powerful test for the use of aCGH as a tool to measure sequence divergence. We found a consistent and linear relationship between hybridization ratio and sequence divergence of the sample to the platform species. At higher levels of sequence divergence (< 92% sequence identity to *D. melanogaster*) ~84% of features had significantly less hybridization to the array in the heterologous species than the platform species, and thus could be identified as "diverged". At the lower levels of divergence (97% identity and greater), only 13% of genes were identified as diverged. While ~40% of the variation in hybridization ratio can be accounted for by variation in sequence identity of the heterologous sample to *D. melanogaster*, other individual characteristics of the DNA sequences also contribute to variation in hybridization ratio, as does technical variation. Therefore, evolutionary process and genomic architecture that shapes species diversity can be addressed through the use of aCGH.

INTRODUCTION

Comparison of genomic DNA sequence among closely related strains or species is a powerful approach with which to identify heterogeneity in evolutionary processes such as selection, mutation rates, and rates of introgression, as well as to unmask phylogenetic relationships. However, even with the recent advances in DNA sequencing technology and rapidly dropping costs, complete genome sequence data are not readily available for many closely related eukaryotes that serve as model systems for organismal evolution (but see Rokas and Abott 2009; Turner et al. 2009). As an alternative, comparative genomic hybridization (CGH) offers a means to estimate sequence divergence.

Although the use of genomic DNA (gDNA) hybridization for phylogenetic analyses and genome-wide estimation of sequence similarity dates to long before vast amounts of sequence

data became available (e.g. Sibley and Ahlquist, 1984; Templeton 1985), this approach has experienced a renaissance with the development of genomic tools, specifically microarrays. On a relatively coarse level, array-based CGH (aCGH) has been widely applied to identify chromosomal aberrations underlying cancer (for review, see Pinkel and Albertson, 2005). When gDNA isolated from a tumor is competitively hybridized against gDNA isolated from normal tissue, genomic regions that have been deleted in the genome of the tumor cells will fail to hybridize to the array features while genomic regions that have been duplicated (amplified) in the genome of the tumor cells will hybridize at a ratio of 2:1 (or greater). At a finer level of resolution, modifications of this technique allow microarray-based genotyping of single nucleotide polymorphisms within and between populations (e.g. *Arabidopsis*: West et al. 2006; stickleback fish: Miller et al. 2007). Array-based techniques can also be applied to genome-scale comparisons between closely related species (or strains) in order to conduct a (nearly) complete analysis of sequence divergence on a gene-by-gene basis.

Unlike microarrays designed for genotyping known polymorphisms (reviewed by Fan et al. 2006) or re-sequencing (e.g. *Arabidopsis*: Clark, R.M, et al. 2007; human: Hinds et al. 2005), microarrays designed for gene expression studies can also be used to compare the genomic content (in coding sequence) of more-or-less closely related species. In a typical experiment, gDNA from the platform species (from which the microarray was constructed) is compared on the array to gDNA from another (heterologous) species of interest. This technique has been used to reveal genomic regions likely involved in an organism's ability to inhabit a specific environment (*Chlamydia trachomatis* tissue specificity: Brunelle et al. 2004; *Sinorhizobium meliloti* root symbiont: Giuntini et al. 2005; *Clostridium difficile* host specificity: Janvilisri et al. 2009), pathogenicity (*Yersinia pesits*: Zhou et al. 2004; Hinchliffe et al. 2003 *Mycobacterium tuberculosis*: Kato-Maeda et al. 2001; *Vibrio cholerae*; Dziejman et al. 2002), genomic duplications and deletions associated with population divergence and speciation (*Anopheles gambiae*: Turner et al. 2005; Riehle et al. 2006), and genomic regions that differentiate humans from other primate species (Locke et al. 2003; Fortna et al. 2004). While most studies rely only on presence or absence metrics, a few studies have suggested the relationship between hybridization signal ratio using aCGH and nucleotide identity is roughly log-linear (Taboada et al. 2005; Brunelle et al. 2004). Using this relatively inexpensive approach, it is possible to identify rapidly evolving genes (*Paxillus involutus*: Le Quere et al. 2006) and in some cases lend insight to phylogenentic relationships (*Shewanella*: Murray et al. 2001; *Saccharomyces***:** Edwards-Ingram et al. 2004; *Salmonella*: Porwollik et al. 2002). While the majority of these examples derive from studies in microbes, the technique is amenable to any size genome. It must be noted, of course, that array-based comparisons do not reveal the actual genomic sequence for the novel gene of interest. Instead, an estimate of sequence identity is obtained at a price and effort far below that of whole genome sequencing.

In the present study, we provide a quantitative analysis of the relationship between hybridization ratio and sequence divergence using a cDNA microarray constructed for *D. melanogaster*. The availability of complete genome assemblies (Clark, A.G. et al. 2007) for *Drosophila melanogaster* as well as three other Drosophilid species, *D. simulans*, *D. sechellia* (2-3 MY diverged from *D. melanogaster*) and *D. yakuba* (10-15 MY diverged from *D. melanogaster* ; O'Grady and Markow, Fly 2009) provides us with a unique opportunity to understand the relationship

between DNA hybridization kinetics and sequence divergence. We show that sequence divergence between orthologous genes can be successfully detected for closely and not so closely related species. Approximately 40% of the variation in gDNA hybridization ratios can be explained by sequence divergence, as measured by nucleotide dissimilarity between sequences. Other sequence-specific characteristics also explain part of the variation in hybridization ratio, and become more prominent with increased sequence divergence. Similarly, technical variation increases with increasing sequence divergence; however this last source of variation can be overcome with increased replication.

MATERIALS AND METHODS

Array Production

We used a *Drosophila melanogaster* microarray with ~22,000 features containing PCR products (~500 base pairs long) generated from custom primers designed to predict open reading frames (Dopman and Hartl 2007; GEO platform number GPL6056). The microarray was printed on poly-L-lysine slides (Thermo Scientific) in a 48 pin format using an OmniGrid-100 arrayer (GeneMachines). Following hydration, snap drying and UV cross-linking, the slides were blocked with succinic anhydride and sodium borate in 1-Methyl-2-Pyrrolidinone, rinsed, dried according to standard procedure (Hedge et al. 2000) and stored dry until used.

Sample Preparation and aCGH

Isogenic *Drosophila melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba* strains (Dmel\y;cn;bw;sp, Dsim\w[501], Dsec\Robertson3C, Dyak Tai18E2) were obtained from the Tuscon *Drosophila* stock center (now known as the San Diego Drosophila Species Stock Center). Genomic DNA was isolated from ~ 100 *Drosophila* males of each stock according to a standard ProteinaseK/Phenol:Chloroform protocol. DNA quantity and purity was assayed (via Nanodrop 1000) prior to and after DNA size reduction using a Hydroshear (Genome Solutions/Digilab) with a standard orifice set to maximal possible shearing speed (13) for 20 cycles (maximal shearing speed varies with individual orifice). This treatment resulted in fragments of 500bp - 2Kbp as determined by gel electrophoresis, visualized with ethidium bromide. Two micrograms of sheared genomic DNA was fluorescently labeled through incorporation of Cy3 or Cy5 labeled dCTP (Amersham) in a Klenow fragment (Invitrogen; Bioprime) reaction of 35 microliters according to manufacturer's protocol. Labeled sample DNA was purified by size exclusion on YM-30 filters (Eppendorf) and appropriate samples were combined. Hybridizations proceeded for ~16 hours at 65 °C in a 3.4X SSC, 0.15% SDS, 1 mM DTT hybridization buffer. Male *D. melanogaster* samples were used in competitive hybridizations with two male *D. sechellia* samples, two male *D. simulans* samples and two male *D. yakuba* samples, incorporating dye swaps to account for dye bias. These aCGH hybridizations were analyzed for the ability to detect significantly diverged genes. An additional six *D. melanogaster* versus *D. yakuba* aCGH hybridizations were available in order to asses the effect of increased technical replication. For this power analysis, only genes located on the autosomes were used because a subset of the hybridizations involved *D. yakuba* female genomic DNA of the same strain.

Microarray Data Analysis

Hybridized arrays were scanned with an Axon 4000B scanner (Axon Instruments) using Genepix 5.0 software (Axon Instruments). All raw array data have been submitted to GEO database (dataseries number GSE18416 sample number GSM459056 -67 Features of poor quality (signal intensity < 2 standard deviations above background) and those of

potentially erroneous sequences (mismatch between initial PCR product sequence prediction and current *D. melanogaster* database; refseq_rna 12/2008) were excluded. Features were only considered in the analysis if they survived these technical filters on multiple arrays for a given species comparison. Raw data from Genepix was imported into R, and LIMMA (Linear Models for Microarray Data, Smyth 2005) was used to apply a background correction ("minimum") and within-array intensity normalization ("loess"). Because we expect the normalization of cross-species arrays to be affected by a substantial number of diverged genes in the non-platform species (van Hijum et al., 2008), we performed the within-array normalization using a set of ~1000 genes highly conserved (greater than ~95% sequence identity; determined with NCBI BLAST to Genbank) among *D. melanogaster D. simulans* and *D. yakuba*. A linear model was fitted to the data using "lmFit", and "eBayes" provided error shrinkage towards a pooled estimate of variation (Smyth 2004). Array features were tested for hybridization ratios that were significantly different from equal as assessed after a FDR multiple testing correction at $P < 0.1$ (Benjamini & Hochberg 1995).

Genomic Sequence Divergence

The sequences of the *D. melanogaster* probes were predicted by blasting primers from the *D. melanogaster* probe (GEO Profiles accession: GPL6056) against the *D. melanogaster* Release 5 assembly and searching for unique and proximal (within 600 base pairs of each other) targets. We queried the resulting sequences against the *D. simulans* and *D. yakuba* NCBI genomes (chromosome) using "megablast" (2009) and against the full chromosome sequence assemblies for *D. sechellia* downloaded from flybase.org (release 1.3). From each heterologous genome, the top BLAST hit to each array feature (threshold e-value -14) was used to obtain the percent similarity between the two sequences and the length of the alignment.

RESULTS AND DISCUSSION

Detection of reduced hybridization

In order to identify array features for which hybridization strength was reduced in each heterologous species, two direct comparisons to *D. melanogaster* were performed. After filtering for unusable array features (low quality or intensity), approximately 80% of the array features were available for analysis in each species, with slightly different numbers on each individual array. From these data we identified array features for which the genomic DNA hybridization signal for each species was reduced compared with *D. melanogaster*. As predicted by their phylogenetic relationship, the fraction of array features that showed a statistically significant reduction in genomic hybridization signal relative to *D. melanogaster* was similar for *D. sechellia* (45.4%) and *D. simulans* (55.8%), and was greater for the more distant *D. yakuba* (70.6%) ($P < 0.1$ FDR corrected) (Table 1). This result, a first for this degree of divergence among multicellular organisms with complex genomes, is consistent with that obtained by Edwards-Ingram et al. (2006); these authors showed that the "molecular taxonomy" of yeast (*Saccharomyces sensu stricto*) as determined by aCGH using a binary presence-

Table 1
Features identified as diverged from *D. melanogaster*

H. Species[a]	# Analyzed	P < 0.1 FDR	P < 0.05 FDR	P < 0.01 FDR
D. sechellia	18374	45%	38%	23%
D. simulans	16325	56%	34%	21%
D. yakuba	17724	71%	66%	58%

[a] The heterologous species used in the 2-array comparison with *D. melanogaster*

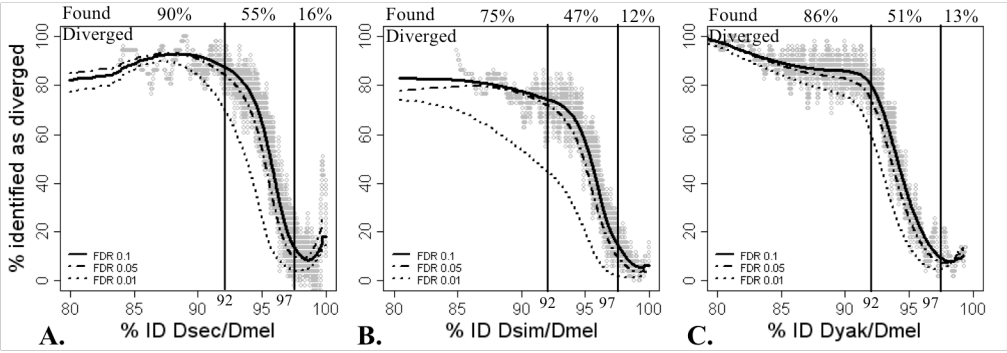

Fig. 1. Successful identification of sequence divergence in orthologs between *D. melanogaster* compared with *D. sechellia* (A), *D. simulans* (B) and *D. yakuba* (C).The percent of array features that were identified as "diverged" based on statistical analysis of hybridization ratio (y-axis) is reported as function of actual sequence divergence (x-axis). Grey points indicate the percentage for a moving window of 51 array features at $P < 0.1$, corrected for false discovery. Lowess-smoothed curves summarize these values for $P < 0.1$ FDR (solid), $P < 0.05$ FDR (dashed) and $P < 0.01$ FDR (dotted).

absence and parsimony-based method closely matched the phylogeny inferred from the complete genome sequences (Fitzpatrick et al. 2006). Similarly, the neighbor-joining and parsimony-based trees constructed with aCGH data from different *Salmonella* subtypes (Porowollik et al. 2002) correspond with the phylogenetic hypotheses inferred from genomic sequence (McQuiston et al. 2008).

Detection of sequence divergence

In order to test the ability to detect highly diverged sequences with aCGH, we BLASTed the predicted *D. melanogaster* probe sequences against the full genome assemblies of each of the heterologous species to provide a measure of sequence divergence for comparison to the array-based measures. The percent nucleotide similarity of the top BLAST hit for the heterologous species to the probe sequence is termed the "percent identity" (%ID). Therefore, a lower %ID represents greater sequence divergence between the heterologous species and *D. melanogaster* for that particular array feature. We asked to what extent statistical analysis of aCGH results recovered the actual sequence divergence between the species examined. Since our measure of %ID is dependent on current sequence data and BLAST results, for which a certain degree of error is possible, the following results concerning the ability to detect diverged genes should be interpreted as conservative estimates. The majority of the array features for which hybridization was significantly reduced in the heterologous species relative to *D. melanogaster* exhibited sequence divergence in the heterologous species being examined (Figure 1). On average, 84% of the orthologs that share less than 92 %ID to *D. melanogaster* showed significantly reduced hybridization. Orthologs between 92 - 97 %ID were detected less well, and at 97 %ID and greater an average of only 13% of features had significantly reduced hybridization. Fitting a logistic curve to these data, we estimate the "detectable sequence divergence level" as the %ID for which there is a 50% chance of a feature being called diverged by aCGH analysis (similar to power analysis techniques, eg. in Townsend 2004). The detectable sequence divergence level was similar for all three heterologous species (*D. sechellia*: 95.5 %ID; *D. simulans*: 94.7 %ID; *D. yakuba*: 94 %ID). The similarity in detectable sequence divergence level across these three species suggests that the hybridization ratio for a given array feature is more dependent on the

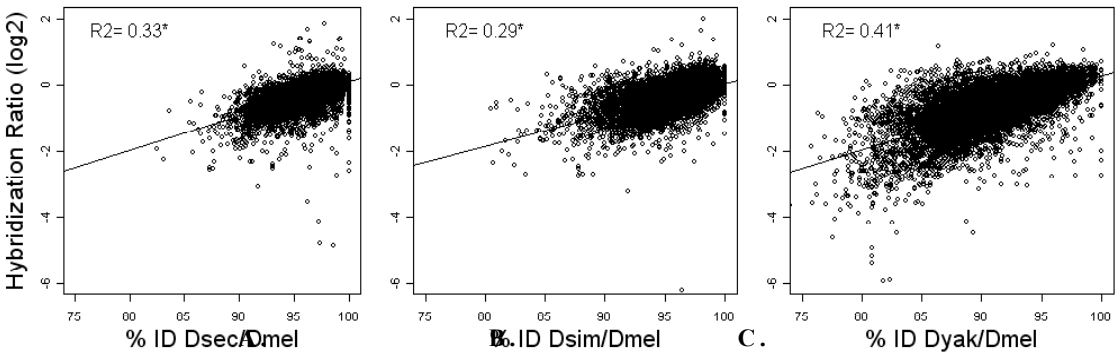

Fig. 2. Linear regression analysis of hybridization ratios vs. %ID of the heterologous species (A) *D. sechellia*, (B) *D. simulans*, and (C) *D. yakuba* relative to *D. melanogaster*.

individual sequence characteristics of that gene in that species than it is on the overall genome similarity between that species and the platform species. This hypothesis should be tested for genomes of substantially different size, or greater dissimilarity. These results demonstrate that aCGH is an effective method by which to detect sequence divergence.

Relationship between sequence divergence and hybridization ratio

The majority of aCGH studies, even in microbes, aim to identify only presence or absence of orthologs. Such studies generally employ one of two assignment strategies. The first strategy employs cut-off threshold compared to results obtained for a characterized strain or compared to other published results (Hatfield and Baldi 2002; Dopman & Hartl 2007). The second strategy analyzes each hybridization dataset according to its intrinsic experimental variability (e.g. GACK: Kim et al. 2002; GENCOM: Pin et al. 2006) in order to determine presence or absence. However, beyond a binary assignment, a more descriptive relationship between sequence divergence and hybridization ratio is possible. For example, Kim et al. (2002) defined a transition zone for those genes thought to be present but highly diverged. For two genes, studied in seven microbial species, a linear relationship was found between the log hybridization ratio and percent divergence (Murray et al. 2001).

In order to determine the extent to which hybridization ratio depicts the true underlying sequence divergence, we examined the correlation between these two measures. For all three species comparisons, a linear regression of %ID and hybridization ratio showed a strong and highly significant correlation (*D. simulans*: Multiple R^2= 0.2920, P < 2.2e-16; *D. sechellia*: Multiple R^2= 0.3257, P < 2.2e-16; *D. yakuba*: Multiple R^2= 0.4083, P < 2.2e-16) (Figure 2), with the data for *D. yakuba* showing the strongest correlation. The apparent decrease in

Table 2
Detection of divergence with increasing technical replication

# Arrays	# Analyzed	Diverged[a]	ID Prob. 50%[b]	R^2 (%ID/Hyb)[c]
2	15372	61%	92.8	0.4085
4	15851	75%	95.0	0.4382
6	16001	80%	95.7	0.4485
8	16060	83%	96.0	0.4530

Note- Increased technical replication leads to an increased power to detect array features that show a difference in the hybridization strength for *D. yakuba* vs. *D. melanogaster*.
This is a reduced dataset due to removal of features on the X chromosome.
[a] Percent of features identified as diverged at P <0.1 FDR
[b] %ID at which there is a 50% chance of a feature being identified as diverged

correlations (lower R^2) seen for *D. sechellia* and *D. simulans* compared to that for *D. yakuba* reflects the increased range of feature sequence divergence of *D. yakuba* orthologs, in addition to the contribution of technical variation rather than actual sequence variation for the less diverged orthologs that predominate in *D. sechellia* and *D. simulans*. Similar control studies conducted with species of *Bartonella* (Lindroos et al. 2005) or strains of *Chlamydia* (Brunelle et al. 2004) suggest that the linear relationship between log hybridization ratio and percent sequence divergence is appropriate only for orthologs with > 75 %ID. In the current study, the small number of array features for which %ID falls below this threshold precludes investigation in that range.

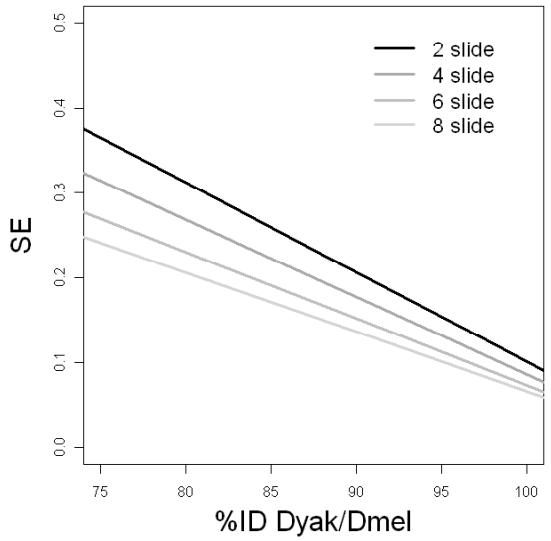

Fig. 3. The standard error of the fitted value for hybridization ratio of *D. yakuba* vs. *D. melanogaster* (linear regression). The standard error decreases with increased technical replication, particularly for orthologs of greater sequence divergence.

Sources of variation and statistical power

While technical variation in hybridization ratios exists even for within-species experiments (where sample and probe sequences are almost identical), previous work suggested that this variation increases as sequence identity decreases (e.g. Taboada et al. 2005; Gilad et al. 2005). Consequently, we expected greater variation in the hybridization ratios of the more diverged orthologs compared with the conserved ones. To test whether increased technical replication would allow us to minimize this source of technical variation, we focused on *D. yakuba*, because this species offers the greatest range of sequence divergence relative to *D. melanogaster*. We included six additional *D. yakuba* vs. *D. melanogaster* hybridization experiments (for a total of eight arrays) and repeated the analysis with all possible two, four, and six array combinations as well as using the full complement of eight heterologous aCGH experiments. For these analyses, genes located on the X chromosome were not included because some of the replicates included female gDNA. Therefore, these analyses rely on a reduced number of features relative to the *D. yakuba* vs. *D. melanogaster* analysis reported above. As expected, for a given statistical threshold, increased technical replication lead to an increase in the percentage of features that were detected as diverged and an increase in detectable sequence divergence level (Table 2). Interestingly, the R^2 value for the regression model of hybridization ratio on %ID was not substantially affected by the increased replication, yet the accuracy of sequence identity estimates for individual array features improved as demonstrated by the decreased standard error of the fitted value (Figure 3). This effect was stronger for array features of 70 – 80 %ID than for features of 90 – 95 %ID. This observation has important implications for experimental design. While orthologs of greater sequence divergence are more easily identified even with few technical replicates, an accurate estimate of sequence identity requires additional hybridizations. However, even with eight technical replicates only 45% of the variation in hybridization ratio

can be attributed to %ID. The remainder of the variation (i.e. deviation from the regression model) is likely due to the individual characteristics of the sequence differences between the heterologous species and the platform species.

Variation in DNA hybridization kinetics has been shown to increase with sequence divergence (Tahboda et al. 2005). Such variation can be caused by any of several physical characteristics of, or differences between, sample DNA and probe DNA, including presence/absence of introns, GC content, distribution of sequence variation, length of probe, and length of sequence alignment, in addition to %ID. While minor differences between technical replicates account for the proportion of the total variation that is due to technical error, the proportion of the total variation that is due to DNA hybridization kinetics manifests as a deviation from the regression model (Figure 4A). When DNA sequence characteristics produce increased hybridization strength, the hybridization ratio for the array feature in question is expected to be greater than the value predicted by the regression model for %ID. Conversely, when DNA sequence characteristics produce decreased hybridization strength, there is an expected decrease in hybridization ratio for that array feature. Even though it was not our goal here to devise an absolute metric that accounts for all possible sources of variation, we explored the relative contribution of technical variation and variation due to hybridization kinetics.

The magnitude of the variation due to DNA hybridization kinetics is measured by the standard deviation (SD) of the residuals from the regression model. This was estimated by dividing the median absolute residual by 0.6745, utilizing a property of normally distributed data to obtain an estimator of SD that is resistant to outliers. The size of technical error is measured by the SD of the fitted value for the hybridization ratio of each

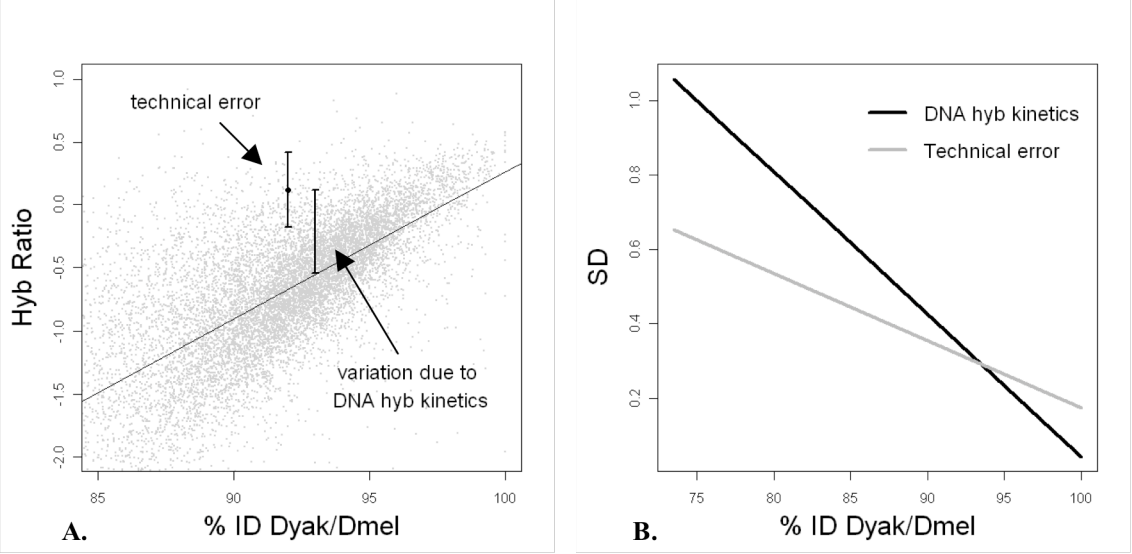

Fig. 4. Two sources of variation affect the quantitative prediction of sequence divergence from hybridization ratios. A) Schematic representation of the "technical error", the standard error of the fit of the hybridization ratios for that feature among technical replicates, and the variation due to physical characteristics of each probe and sample DNA (DNA hybridization kinetics), the deviation from the regression model of hybridization ratio vs. %ID. B) The calculated relative contributions of technical error (grey line) and variation due to DNA hybridization kinetics (black line) as a function of %ID of *D. yakuba* to *D. melanogaster.* At low levels of divergence, technical error predominates.

feature. Based on the 8-array *D. yakuba* vs. *D. melanogaster* dataset, we found that hybridization ratios for features representing conserved orthologs (greater than ~95 %ID) were more affected by technical error, whereas those of more diverged orthologs (less than ~95 %ID) were more strongly influenced by DNA hybridization kinetics (Fig. 4B). Both technical error and variation due to DNA hybridization kinetics increase with greater sequence divergence.

To demonstrate how specific factors other than %ID can contribute to DNA hybridization kinetics, we took into consideration GC content of the *D. melanogaster* probe (GC content), the length of the *D. melanogaster* probe (probe length), and the percent of the *D. melanogaster* probe length over which the heterologous sequence can be aligned (percent alignment length). We incorporated these variables into the linear regression model of hybridization ratios vs. %ID to *D. melanogaster* (hybridization ratio ~ %ID * (GC + length + % align)) of the 8-array *D. yakuba* and *D. melanogaster* dataset. Both GC content and percent alignment length were significant in the model, and there was a significant interaction effect of GC content and %ID (Table 3). To assess the relative effect size for each explanatory variable, the response and explanatory variables were standardized such that the mean of each variable was 0 with a standard deviation of 1. The effect size for GC content decreased with increasing %ID, yet it retained a positive effect on hybridization ratio (Fig. 5). There was also a positive effect of percent alignment length on hybridization ratio, although unaffected by %ID. At lower levels of divergence (greater than 85 %ID), percent alignment length had a greater relative effect size than GC content. Since high GC content is likely to produce a more stable bond between two DNA strands, it is not surprising that a higher GC content of the array probe would produce a stronger bond with a diverged sequence than would be seen at low GC content, with conserved GC regions contributing extra stability to the otherwise weak bond. It is interesting that this effect is even more pronounced at higher levels of divergence, supporting the hypothesis of stabilization of weak bonds.

It is also intuitive that percent alignment length would be positively correlated with the *D.*

Table 3
%ID and sequence characteristics vs. hybridization ratio regression

	Estimate	Std. Error	t value	Pr(>|t|)	Min[a]	Max[b]	1 SD[c]
(Intercept)	-21.7825	1.1968	-18.2001	5.43E-73			
%ID	0.2018	0.0130	15.4977	1.25E-53*	73.48	100	3.67
GC	0.1538	0.0166	9.2395	2.91E-20*	123	1839	150.47
Length	-0.0005	0.0006	-0.8412	0.4002	23.77	76.19	5.29
%align	0.0309	0.0084	3.6692	2.44E-04*	6.40	118.46	8.90
%ID:GC	-0.0015	0.0002	-8.0096	1.26E-15*			
%ID:length	1.88E-06	6.49E-06	0.2902	0.7716			
%ID:%align	-0.0002	9.16E-05	-1.7257	0.0844			

Note- Regression model: hybridization ratio ~ %ID * (%GC + length + % align)). The model includes the interaction of GC content, probe length, and percent alignment of *D. melanogaster* probe sequence to *D. yakuba* sequence.
[a] minimum value
[b] maximum value
[c] 1 standard deviation

yakuba vs. *D. melanogaster* hybridization ratio, as it is another measure of divergence that is not taken into consideration by %ID. While a complete alignment indicates low divergence, an incomplete alignment indicates substantial divergence. The absence of a strong correlation between relative alignment length and hybridization ratio may be due to insertions in the heterologous sequence that result in reduced hybridization strength and inflated relative alignment length (even above 100%). Since our goal here was not to devise a precise metric, we did not include such insertions and other possible explanatory variables (e.g. GC content of the heterologous sample) in our model. We see aCGH as an efficient and inexpensive method for identifying highly diverged genes among species for which little or no genomic sequence information is available. Direct sequence analysis (in multiple individuals) would be necessary to further investigate any genes of interest identified this way (e.g., synonymous vs. non-synonymous substitutions).

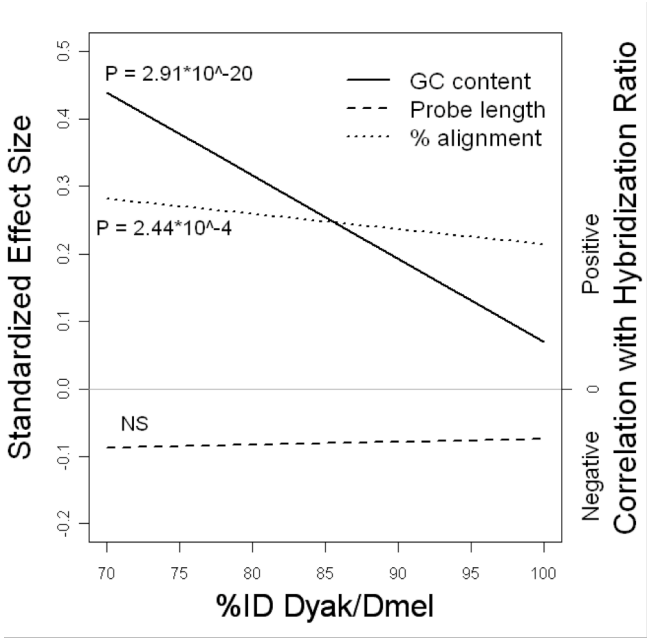

Fig. 5. Standardized effect sizes of probe characteristics with hybridization ratio and %ID of *D. yakuba* to *D melanogaster*. P-values for the partial correlations of GC content, probe length, and relative alignment length of probe sequence to *D. yakuba* sequence are f rom the (non-standardized) regression model: hybridization ratio ~ %ID * (GC + length + % alignment). There is a significant interaction between GC content and %ID ($P = 1.26*10^{-15}$) and a consistently positive effect (positively correlated with hybridization ratio).

CONCLUSIONS

The results presented here demonstrate that aCGH can reliably detect genes that are highly diverged in a species compared with one for which a microarray has been constructed. Our use of a *D. melanogaster* microarray to estimate sequence divergence on a gene-by-gene basis for three fully sequenced heterologous species allowed for a robust proof-of-principle for this approach. We found a consistent and linear relationship between array hybridization ratio and sequence divergence between the sample and the platform species. The level of sequence difference (in our study ~95 %ID) at which divergence can statistically be detected will depend on the power of the experiment. While the power can be increased with additional technical replicates, there will still be a subset of diverged genes that escape detection due to other specific hybridization kinetics of the sample DNA to the array feature. This technique is generally applicable, even though thresholds, correlation strengths, and appropriate divergence distances may differ for different arrays platforms (cDNAs, long or short oligonucleotides). As the number of microarray platforms available for non-traditional model species is continually growing (axolotl: Page et al. 2008; dolphin: Mancia et al. 2007; butterfly: Vera et al. 2008; Zhu et al. 2009; coral: Edge et al. 2008; many fish species: see Kassahn, 2008, for review; crustaceans: e.g. Dhar et al. 2003; Stillman et al 2006; Ki et al. 2009), we believe that researchers focusing on these model systems will continue to benefit from the relatively low cost of array hybridizations, rapid advances in

next-gen sequencing technology notwithstanding (Shendura et al. 2004; Turner et al. 2009), as *de novo* sequencing of a large number of complex eukaryotic genomes is still less cost effective.

ACKNOWLEDGEMENTS

We are grateful for the support of Dan Hartl (Harvard University) and members of his lab for assistance with microarray production and annotation. Albyn Jones (Reed College) consulted for statistical analysis. This work was supported by the Murdock Life Trust Foundation (S.C.P.R.) and by the Bauer Center for Genomics Research (H.A.H.).

LITERATURE CITED

Bay, L. K., K. E. Ulstrup, H. B. Nielsen, H. Jarmer, N. Goffard, B. L. Willis, D. J. Miller, and M. J. H. Van Oppen. 2009. Microarray analysis reveals transcriptional plasticity in the reef building coral *Acropora millepora*. Molecular Ecology **18**:3062-3075.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological **57**:289-300.

Brunelle, B. W., T. L. Nicholson, and R. S. Stephens. 2004. Microarray-based genomic surveying of gene polymorphisms in *Chlamydia trachomatis*. Genome Biology **5**:article42.

Clark, A.G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow, T.C. Kaufman, et al. Drosophila 12 Genomes Consortium. 2007. *Evolution of genes and genomes on the Drosophila* phylogeny. Nature. **450**:203-18.

Clark, R. M., G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T. T. Hu, G. Fu, D. A. Hinds, H. M. Chen, K. A. Frazer, D. H. Huson, B. Schoelkopf, M. Nordborg, G. Raetsch, J. R. Ecker, and D. Weigel. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science **317**:338-342.

Dhar, A. K., A. Dettori, M. M. Roux, K. R. Klimpel, and B. Read. 2003. Identification of differentially expressed genes in shrimp (*Penaeus stylirostris*) infected with White spot syndrome virus by cDNA microarrays. Archives of Virology **148**:2381-2396.

Dopman, E. B., and D. L. Hartl. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences of the United States of America **104**:19920-19925.

Dunn, B., R. P. Levine, and G. Sherlock. 2005. Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. BMC Genomics **6**:article524.

Dziejman, M., E. Balon, D. Boyd, C. M. Fraser, J. F. Heidelberg, and J. J. Mekalanos. 2002. Comparative genomic analysis of *Vibrio cholerae*: Genes that correlate with cholera endemic and pandemic disease. Proceedings of the National Academy of Sciences of the United States of America **99**:1556-1561.

Edge, S. E., M. B. Morgan, D. F. Gleason, and T. W. Snell. 2005. Development of a coral cDNA array to examine gene expression profiles in *Montastraea faveolata* exposed to environmental stress. Marine Pollution Bulletin **51**:507-523.

Edwards-Ingram, L. C., M. E. Gent, D. C. Hoyle, A. Hayes, L. I. Stateva, and S. G. Oliver. 2004. Comparative genomic hybridization provides new insights into the molecular taxonomy of the *Saccharomyces sensu stricto* complex. Genome Research **14**:1043-1051.

Fan, J. B., M. S. Chee, and K. L. Gunderson. 2006. Highly parallel genomic assays. Nature Reviews Genetics **7**:632-644.

Fitzpatrick, D. A., M. E. Logue, J. E. Stajich, and G. Butler. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology **6**:article99.

Fortna, A., Y. Kim, E. MacLaren, K. Marshall, G. Hahn, L. Meltesen, M. Brenton, R. Hink, S. Burgers, T. Hernandez-Boussard, A. Karimpour-Fard, D. Glueck, L. McGavran, R. Berry, J. Pollack, and J. M. Sikela. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biology **2**:937-954.

Gilad, Y., S. A. Rifkin, P. Bertone, M. Gerstein, and K. P. White. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Research **15**:674-680.

Giuntini, E., A. Mengoni, C. De Filippo, D. Cavalieri, N. Aubin-Horth, C. R. Landry, A. Becker, and M. Bazzicalupo. 2005. Large-scale genetic variation of the symbiosis-required megaplasmid pSymA revealed by comparative genomic analysis of *Sinorhizobium meliloti* natural strains. BMC Genomics **6**:article158.

Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snesrud, N. Lee and J. Quackenbus. 2000/ A concise guide to cDNA microarray analysis. Biotechniques **29**:548-50, 552-4.

Hinchliffe, S. J., K. E. Isherwood, R. A. Stabler, M. B. Prentice, A. Rakin, R. A. Nichols, P. C. F. Oyston, J. Hinds, R. W. Titball, and B. W. Wren. 2003. Application of DNA mcroarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. Genome Research **13**:2018-2029.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. Science **307**:1072-1079.

Janvilisri, T., J. Scaria, A. D. Thompson, A. Nicholson, B. M. Limbago, L. G. Arroyo, J. G. Songer, Y. T. Grohn, and Y. F. Chang. 2009. Microarray Identification of *Clostridium difficile* Core Components and Divergent Regions Associated with Host Origin. Journal of Bacteriology **191**:3881-3891.

Kassahn, K. S. 2008. Microarrays for comparative and ecological genomics: beyond single-species applications of array technologies. Journal of Fish Biology **72**:2407-2434.

Kato-Maeda, M., J. T. Rhee, T. R. Gingeras, H. Salamon, J. Drenkow, N. Smittipat, and P. M. Small. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Research **11**:547-554.

Ki, J. S., S. Raisuddin, K. W. Lee, D. S. Hwang, J. Han, J. S. Rhee, I. C. Kim, H. G. Park, J. C. Ryu, and J. S. Lee. 2009. Gene expression profiling of copper-induced responses in the intertidal copepod *Tigriopus japonicus* using a 6K oligochip microarray. Aquatic Toxicology **93**:177-187.

Kim, C. C., E. A. Joyce, K. Chan, and S. Falkow. 2002. Improved analytical methods for microarray-based genome-composition analysis. Genome Biol **3**:RESEARCH0065.

Le Quere, A., K. A. Eriksen, B. Rajashekar, A. Schutzenbubel, B. Canback, T. Johansson, and A. Tunlid. 2006. Screening for rapidly evolving genes in the ectomycorrhizal fungus *Paxillus involutus* using cDNA microarrays. Molecular Ecology **15**:535-550.

Lindroos, H. L., A. Mira, D. Repsilber, O. Vinnere, K. Naslund, M. Dehio, C. Dehio, and S. G. E. Andersson. 2005. Characterization of the genome composition of *Bartonella koehlerae* by microarray comparative genomic hybridization profiling. Journal of Bacteriology **187**:6155-6165.

Locke, D. P., R. Segraves, L. Carbone, N. Archidiacono, D. G. Albertson, D. Pinkel, and E. E. Eichler. 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Research **13**:347-357.

Mancia, A., M. L. Lundqvist, T. A. Romano, M. M. Peden-Adams, P. A. Fair, M. S. Kindy, B. C. Ellis, S. Gattoni-Celli, D. J. McKillen, H. F. Trent, Y. A. Chen, J. S. Almeida, P. S. Gross, R. W. Chapman, and G. W. Warr. 2007. A dolphin peripheral blood leukocyte cDNA microarray for studies of immune function and stress reactions. Developmental and Comparative Immunology **31**:520-529.

McQuiston, J. R., S. Herrera-Leon, B. C. Wertheim, J. Doyle, P. I. Fields, R. V. Tauxe, and J. M. Logsdon. 2008. Molecular Phylogeny of the *Salmonellae*: Relationships among Salmonella Species and Subspecies Determined from Four Housekeeping Genes and Evidence of Lateral Gene Transfer Events. Journal of Bacteriology **190**:7060-7067.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research **17**:240-248.

Murray, A. E., D. Lies, G. Li, K. Nealson, J. Zhou, and J. M. Tiedje. 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proceedings of the National Academy of Sciences of the United States of America **98**:9853-9858.

Page, R. B., S. R. Voss, A. K. Samuels, J. J. Smith, S. Putta, and C. K. Beachy. 2008. Effect of thyroid hormone concentration on the transcriptional response underlying induced metamorphosis in the Mexican axolotl (*Ambystoma*). BMC Genomics **9**:article78.

Pin, C., M. Reuter, B. Pearson, L. Friis, K. Overweg, J. Baranyi, and J. Wells. 2006. Comparison of different approaches for comparative genetic analysis using microarray hybridization. Applied Microbiology and Biotechnology **72**:852-859.

Pinkel, D., and D. G. Albertson. 2005. Comparative genomic hybridization. Annual Review of Genomics and Human Genetics **6**:331-354.

Porwollik, S., R. M. Y. Wong, and M. McClelland. 2002. Evolutionary genomics of *Salmonella*: Gene acquisitions revealed by microarray analysis. Proceedings of the National Academy of Sciences of the United States of America **99**:8956-8961.

Riehle, M. M., K. Markianos, O. Niare, J. N. Xu, J. Li, A. M. Toure, B. Podiougou, F. Oduol, S. Diawara, M. Diallo, B. Coulibaly, A. Ouatara, L. Kruglyak, S. F. Traore, and K. D. Vernick. 2006. Natural malaria infection in *Anopheles gambiae* is regulated by a single genomic control region. Science **312**:577-579.

Rokas, A., and P. Abbot. 2009. Harnessing genomics for evolutionary insights. Trends in Ecology & Evolution **24**:192-200.

Shendure, J. A., G. J. Porreca, and G. M. Church. 2008. Overview of DNA sequencing strategies. Curr Protoc Mol Biol **Chapter 7**:Unit 7.1.

Sibley, C. G., and J. E. Ahlquist. 1984. The Phylogeny of the Hominoid Primates, as Indicated by DNA-DNA Hybridization. Journal of Molecular Evolution **20**:2-15.

Smyth, G. K. 2005. Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397-420.

Smyth, G. K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat App Genet Mol Biol **3**:1–26.

Stillman, J. H., K. S. Teranishi, A. Tagmount, E. A. Lindquist, and P. B. Brokstein. 2006. Construction and characterization of EST libraries from the porcelain crab, *Petrolisthes cinctipes*. Integrative and Comparative Biology **46**:919-930.

Taboada, E. N., R. R. Acedillo, C. C. Luebbert, W. A. Findlay, and J. H. E. Nash. 2005. A new approach for the analysis of bacterial microarray-based Comparative Genomic Hybridization: insights from an empirical study. BMC Genomics **6**:1-10.

Templeton, A. R. 1985. The Phylogeny of the Hominoid Primates - a Statistical-Analysis of the DNA-DNA Hybridization Data. Molecular Biology and Evolution **2**:420-433.

Townsend, J. P. 2004. Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays. BMC Bioinformatics **5**:article13.

Turner, D. J., T. M. Keane, I. Sudbery, and D. J. Adams. 2009. Next-generation sequencing of vertebrate experimental organisms. Mammalian Genome **20**:327-338.

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin. 2005. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biology **3**:article285

van Hijum, S., R. J. S. Baerends, A. L. Zomer, H. A. Karsens, V. Martin-Requena, O. Trelles, J. Kok, and O. P. Kuipers. 2008. Supervised Lowess normalization of comparative genome hybridization data - application to lactococcal strain comparisons. BMC Bioinformatics **9**:article93.

Vera, J. C., C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski, and J. H. Marden. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology **17**:1636-1647.

West, M. A. L., H. van Leeuwen, A. Kozik, D. J. Kliebenstein, R. W. Doerge, D. A. St Clair, and R. W. Michelmore. 2006. High-density haplotyping with microarray-based expression and single feature polymorphism markers in *Arabidopsis*. Genome Research **16**:787-795.

Zhou, D. S., Y. P. Han, Y. Song, Z. Z. Tong, J. Wang, Z. B. Guo, D. C. Pei, X. Pang, J. H. Zhai, M. Li, B. Z. Cui, Z. Z. Qi, L. X. Jin, R. X. Dai, Z. M. Du, J. Y. Bao, X. Q. Zhang, J. Yu, J. Wang, P. T. Huang, and R. F. Yang. 2004. DNA microarray analysis of genome dynamics in *Yersinia pestis*: Insights into bacterial genome microevolution and niche adaptation. Journal of Bacteriology **186**:5138-5146.

Zhu, H., R. J. Gegear, A. Casselman, S. Kanginakudru, and S. M. Reppert. 2009. Defining behavioral and molecular differences between summer and migratory monarch butterflies. BMC Biology **7**:article14.