Detecting genetic variation in microarray expression data

Jennifer A. Greenhall,^{1,2,8} Matthew A. Zapala,^{3,4,8} Mario Cáceres,^{1,5} Ondrej Libiger,⁴ Carolee Barlow,^{1,6} Nicholas J. Schork,^{3,4} and David J. Lockhart^{1,7,9}

¹The Salk Institute for Biological Studies, La Jolla, California 92037, USA; ²Neurosciences Graduate Program, School of Medicine, University of California, San Diego, California 92093, USA; ³Biomedical Sciences Graduate Program, School of Medicine, University of California, San Diego, California 92093, USA; ⁴Polymorphism Research Laboratory, Department of Psychiatry, University of California, San Diego, California 92093, USA; ⁵Genes and Disease Program, Center for Genomic Regulation (CRG-UPF), Barcelona 08003, Spain; ⁶Brain Cells, Inc., San Diego, California 92121, USA; ⁷Amicus Therapeutics, Cranbury, New Jersey 08512, USA

The use of high-density oligonucleotide arrays to measure the expression levels of thousands of genes in parallel has become commonplace. To take further advantage of the growing body of data, we developed a method, termed “GeSNP,” to mine the detailed hybridization patterns in oligonucleotide array expression data for evidence of genetic variation. To demonstrate the performance of the algorithm, the hybridization patterns in data obtained previously from SAMP8/Ta, SAMPIO/Ta, and SAMRI/Ta inbred mice and from humans and chimpanzees were analyzed. Genes with consistent strain-specific and species-specific hybridization pattern differences were identified, and ~90% of the candidate genes were independently confirmed to harbor sequence differences. Importantly, the quality of gene expression data was also improved by masking the probes of regions with putative sequence differences between species and strains. To illustrate the application to human disease groups, data from an inflammatory bowel disease study were analyzed. GeSNP identified sequence differences in candidate genes previously discovered in independent association and linkage studies and uncovered many promising new candidates. This approach enables the opportunistic extraction of genetic variation information from new or pre-existing gene expression data obtained with high-density oligonucleotide arrays.

[Supplemental material is available online at www.genome.org. GeSNP can be accessed at <http://porifera.ucsd.edu/~cabney/cgi-bin/geSNP.cgi>. The Affymetrix CEL files for the mouse studies and the human/chimpanzee array data have been submitted to GEO under accession nos. GSE6238 and GSE7540, respectively.]

High-density oligonucleotide arrays are used routinely to measure quantitative levels of gene expression and to screen thousands of genes for expression differences (Lipshutz et al. 1999; Lockhart and Winzler 2000; Lamb et al. 2006; Shi et al. 2006). Unlike cDNA microarrays that typically use a single spotted PCR product, Affymetrix oligonucleotide arrays are designed with multiple, different, sequence-specific DNA probes for each gene (Lockhart et al. 1996). The quantitative, multiple-probe hybridization patterns for each gene are reproducible, and the specific patterns depend on the sequence of the DNA or RNA molecules that bind (Fodor et al. 1991, 1993; Pease et al. 1994; Chee et al. 1996; Hacia et al. 1996; Wodicka et al. 1997). Several papers have highlighted the ability to mine gene expression data from high-density oligonucleotide arrays to find probes that behave unusually within a probe set (Li and Wong 2001) or to find genes with variant splice forms (Hu et al. 2001).

Although gene expression arrays were not designed to detect sequence differences, we reasoned that the underlying multiple-probe hybridization patterns could be retrospectively analyzed, probe by probe, to identify possible sequence differences between strains, species, or individuals. In addition, the identifica-

tion and removal or “masking” of probe pairs that target regions with sequence differences should produce more accurate expression results in comparisons between distinct genetic populations (Cáceres et al. 2003; Karaman et al. 2003; Khaitovich et al. 2004; Nagpal et al. 2004). Moreover, global genetic variation screening of this type, along with gene expression profiling, provides a valuable complement to other methods, such as direct candidate sequencing, SNP genotyping, and QTL analysis, for the identification of genes responsible for important phenotypes (Geschwind 2000; Grupe et al. 2001).

Here, we demonstrate the performance of a user-friendly Web-based program, “GeSNP” (available at <http://porifera.ucsd.edu/~cabney/cgi-bin/geSNP.cgi>) that can be used to identify potential sequence variation from gene expression data sets. This algorithm was used previously to identify sequence differences between three rare strains of inbred mice (Carter et al. 2005) and to improve the reliability of gene expression data by masking probe pairs that cover regions with sequence differences between humans and chimpanzees (Cáceres et al. 2003). Recently, this algorithm was also applied in an expression QTL (eQTL) study to exclude spurious *cis*-acting eQTLs due to probe-specific hybridization differences (Hovatta et al. 2007). Sequence variation that affects hybridization to the array probes and leads to false associations has been identified as a major problem in eQTL investigations (Peirce et al. 2006; Radcliffe et al. 2006). Sequencing of specific candidate genes has been used to minimize these false associations, but this approach is not practical on a genome scale

⁸These authors contributed equally to this work.

⁹Corresponding author.

E-mail dlockhart@amicustherapeutics.com; fax (609) 662-2001.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.6307307>.

(Hubner et al. 2005). In addition to our studies, similar techniques have also been used to identify sequence differences in prokaryotes and lower eukaryotes (Winzler et al. 1998; Borevitz et al. 2003; Albert et al. 2005; Hazen et al. 2005; Ronald et al. 2005; Gresham et al. 2006) and between primate species (Khaltovich et al. 2004). However, a fully functioning program for use with data from many different species and different Affymetrix arrays has not been made available.

Results

Development of the algorithm

On an Affymetrix oligonucleotide array, each gene is represented by a probe set, which is comprised of ~11–20 different oligonucleotide probe pairs that are designed to hybridize to specific regions of a gene. Each probe pair consists of a matched set of two 25-base oligonucleotide probes, a perfect match (PM) for the gene of interest and a mismatch (MM) containing a single nucleotide substitution in the middle of the probe (position 13). The MM serves as a measure of non-specific background binding and noise. The GeSNP algorithm compares the detailed hybridization patterns across the oligonucleotide probe pairs for a gene, after normalizing for expression level differences, in order to find probe pairs that show consistent, statistically significant differences between two sets of samples. The algorithm works as follows (see Fig. 1): First, the individual hybridization intensity values are extracted from the cell-by-cell intensity (CEL) file. The difference between the perfect match and the mismatch (PM – MM) intensities is calculated for each probe pair for each CEL file. In order to minimize false predictions of sequence differences due to inadequate hybridization signal, a probe set from a particular CEL file is excluded if <65% of the PM – MM values for the probe set are positive, indicating that the gene was not likely expressed at a detectable level. After eliminating data from samples that do not fulfill this criterion, the PM – MM values for all of the probe sets for each sample are globally scaled to compensate for gene expression differences. The scaling factor is calculated by dividing an arbitrary target value of 200 by the standard deviation of the PM – MM values for a probe set while ignoring the largest and smallest PM – MM values. Next, the scaled values for each sample group are averaged,

Figure 1. Detection of sequence differences using oligonucleotide array expression data. (A) Key steps in the GeSNP algorithm are described (left panels in boxes), and corresponding graphical illustrations of the SAMP10 data for MG-U74Av2 array probe set 98333_at, representing the gene ribosomal protein S18 (*Rps18*), are shown (right). Step 1 of the method is to extract data for a specific probe set from the CEL file. In step 2, the hybridization intensity difference between the perfect match and mismatch probe (PM – MM) for each probe pair (PP) is calculated. These values are then evaluated for inclusion in subsequent analyses as determined by passing pattern quality measures for detectable expression. The unscaled hybridization intensity values for *Rps18* are shown for all nine samples of the SAMP10 strain, where the PP number is indicated on the X-axis ranging from 1 to 16, and the PM – MM value is shown on the Y-axis. Next (step 3), the intensity patterns for each sample are individually scaled to a common value. The scaled PP differences are then averaged (step 4) to generate a single value and standard deviation for each PP. (B) For the *Rps18* probe set, the same analysis was performed for the nine SAMR1 samples, all of which passed the pattern quality measures for detectable expression. The average hybridization patterns with standard deviations obtained for SAMP10 (red line and squares) and SAMR1 (blue line and triangles) mice are shown. Using a *t*-value threshold of 6, the algorithm identified two PPs harboring putative sequence differences (black asterisks). Consistent with the hybridization pattern differences, DNA sequencing showed that each of these PPs indeed covered a region with a single base pair difference between the two strains. (C) The average hybridization signals with standard deviations are shown for the 96498_at probe set for the gene *Dmc1*, using the six files that passed the pattern quality measures for SAMP10 (red line and squares) and five files for SAMR1 (blue line and triangles). DNA sequencing identified no sequence differences between strains, consistent with the nearly identical, overlapping hybridization patterns (largest *t*-value of 2.4).

and an average and a variance are calculated for each probe pair in a probe set. To further reduce false positives, only probe sets for which at least four files in both sample groups have exceeded the pattern quality threshold are analyzed.

Figure 2. Sequence differences between SAM strains. GeSNP comparisons using the three SAM strains are shown. The Y-axis represents the number of SFPs occurring at a given t -value, and the X-axis shows several t -value thresholds. (Black squares) SAMP8 to SAMR1 comparison, (black triangles) SAMP10 to SAMR1 comparison, (gray circles) comparison of SAMP8 and SAMP10 to SAMR1, uncovering SAMP8 and SAMP10 common, shared differences. The divergence in the number of SFPs is consistent with the phylogenetic distance between strains that has been found by microsatellite studies (Xia et al. 1999) and SNP genotyping (Cervino et al. 2005). All the non-monomorphic SNPs (1907) of Cervino et al. (2005) were used to generate the phylogenetic tree shown inside the graph using the neighbor-joining option of the MEGA3 software (Kumar et al. 2004).

To identify statistically significant pattern differences, the Student's t -test using the separate variance formula was employed. The t -value for each probe pair (PP) was calculated as follows:

$$\frac{|\text{Average Scaled Difference (PP}_{av}, \text{group}_1) - \text{Average Scaled Difference (PP}_{av}, \text{group}_2)|}{\sqrt{\left(\frac{\text{Variance (PP}_{av}, \text{group}_1)}{n_1} + \frac{\text{Variance (PP}_{av}, \text{group}_2)}{n_2}\right)}}$$

where n_1 is the number of files in group 1 and n_2 is the number of files in group 2. Empirically, the P -value of the Student's t -test and P -values generated by permutation testing did not perform as well as the t -value in identifying confirmed sequence differences. In order to choose an appropriate t -value threshold for each comparison, a false positive comparison was created in which the two compared groups were equally distributed between two false positive groups, accounting for any subgroup bias. Probe pairs that are statistically significant in this randomized comparison represent the potential number of false positives at a particular threshold for the number of files compared (see Methods). Previous studies have referred to the probe pairs containing putative sequence differences as single feature polymorphisms (SFPs) (Winzeler et al. 1998), and for consistency we adopt this notation.

Identification of sequence differences among three rare mouse strains

We have been studying the aging of the mouse brain using three strains of senescence accelerated mice (SAM) developed in Japan as models of accelerated aging: SAMP8/Ta (SAMP8), SAMP10/Ta (SAMP10), and the control, aging-resistant strain, SAMR1/Ta

(SAMR1) (Takeda 1999). Initially, large-scale sequence or SNP information was not available for these strains, as is still the case for numerous other laboratory strains or crosses. To identify sequence differences among these three strains, MG-U74Av2 Affymetrix arrays hybridized to five hippocampal and four retinal samples of each mouse strain (Carter et al. 2005) were analyzed using the GeSNP algorithm. Figure 2 shows the number of SFPs identified at increasing t -value thresholds, and these numbers reflect the known genetic divergence among the three strains from microsatellite markers (Xia et al. 1999) and SNP genotyping (Cervino et al. 2005). The performance of the GeSNP algorithm was determined by sequencing 24 genes from two cortex samples of each strain and calculating the number of true positives, false positives, true negatives, and false negatives at various t -value thresholds. Table 1 shows the algorithm performance for the SAMP8 versus SAMR1 comparison. The true positive rate, also known as the positive predictive value, is the percentage of SFPs that are indeed true positives, and the detection rate, also known as the sensitivity, describes

the percentage of probe pairs covering sequence differences that are identified as SFPs by the algorithm. When results are analyzed by probe pair, a t -value of 6 yields an 89% true positive rate and a 75% detection rate. However, most false negatives and false positives at the probe pair level are contained within a probe set harboring true positives, and hence the performance at the probe set level yields a 100% true positive rate and a 100% detection rate for almost all t -values shown. A small number of false positives is acceptable in this type of screen as the purpose is to identify a manageable number of candidates for follow-up studies with conventional DNA sequencing. For the SAMP8 versus

Table 1. GeSNP performance, SAMP8 versus SAMR1 mouse strains

	$t \geq 5$	$t \geq 6$	$t \geq 7$
Probe pair analysis			
TP	24	24	21
FN	8	8	11
TN	365	367	369
FP	5	3	1
TPR	0.83	0.89	0.95
DR	0.75	0.75	0.66
Probe set analysis			
TP	7	7	6
FN	0	0	1
TN	19	19	19
FP	0	0	0
TPR	1.00	1.00	1.00
DR	1.00	1.00	0.86

(TP) True positive; (FN) false negative; (TN) true negative; (FP) false positive; (TPR) true positive rate (the positive predictive value), $[TP/(TP+FP)]$; (DR) detection rate (the sensitivity), $[TP/(TP+FN)]$.

Table 2. GeSNP performance, human versus chimpanzee

	$t \geq 5$	$t \geq 6$	$t \geq 7$
Probe pair analysis			
TP	54	42	36
FN	44	56	62
TN	395	414	422
FP	36	17	9
TPR	0.60	0.71	0.80
DR	0.55	0.43	0.37
Probe set analysis			
TP	19	16	16
FN	4	8	9
TN	8	10	10
FP	6	3	2
TPR	0.76	0.84	0.89
DR	0.83	0.67	0.64

(TP) True positive; (FN) false negative; (TN) true negative; (FP) false positive; (TPR) true positive rate (the positive predictive value), $[TP/(TP+FP)]$; (DR) detection rate (the sensitivity), $[TP/(TP+FN)]$.

SAMR1 and SAMP10 versus SAMR1 probe sets containing one or more SFPs with a t -value ≥ 5 , see Supplemental Tables 1 and 2. Although not discussed here in detail, in addition to single nucleotide differences, the GeSNP algorithm is also able to detect insertions/deletions and splice variants.

Identification of sequence differences between species

In order to further validate the algorithm and extend its applicability to interspecies comparisons, gene expression data from 10 humans (*Homo sapiens*) and seven chimpanzees (*Pan troglodytes*) obtained with human HG-U95Av2 Affymetrix arrays were analyzed (Enard et al. 2002; Cáceres et al. 2003). A total of 28 arrays hybridized to human samples and 21 to chimpanzee samples were compared, and the sequences of 24 human and chimpanzee genes (represented by 37 different probe sets) were examined. All but one of the genes were selected for sequencing because they appeared to be differentially expressed between primate brains in the initial analysis of the array data (Cáceres et al. 2003). The performance of the algorithm in the human–chimpanzee comparison is shown in Table 2 (see Supplemental Table 3 for probe sets containing SFPs with a t -value ≥ 5). At a t -value threshold of 6, 16 of 19 probe sets (84% true positive rate) or 42 of 59 probe pairs (71% true positive rate) identified as SFPs were independently confirmed to contain sequence differences (Table 2). Similarly, 414 of 431 probe pairs (96% specificity) covering identical sequence between species did not contain SFPs. In addition, taking into account the quality of the current genome sequences, comparable results were obtained based on a genome-wide analysis of the available human and chimpanzee genome sequence (Supplemental Table 7).

Compared with the analyses of closely related mouse strains, the numbers of false negatives are higher and the detection rates are lower in this interspecies comparison. The greater sequence variation in non-isogenic groups and the resulting larger statistical variance can yield a smaller t -value not identified as significant, especially for more subtle hybridization differences. Consistent with this observation, ~50% of the false negatives between humans and chimpanzees are due to the sequence difference lying within the first five or last five nucleotides of the probe sequence. Increasing the number of individuals with array data in both the human and chimpanzee groups is likely to in-

crease the detection of these smaller hybridization changes. Furthermore, data from a larger number of individuals will improve the resolution of the results, decreasing the number of false positives due to intraspecies SNPs and sharpening the peaks for the true positives.

Comparison of the GeSNP algorithm with other methods

The performance of the GeSNP algorithm was compared with the algorithm of Ronald et al. (2005), which uses a different approach for background subtraction and normalization and was previously used to identify sequence differences in gene expression data between yeast strains. James Ronald and Leonid Kruglyak kindly provided us with their C++ program implementing both the normalization procedure of Irizarry et al. (2003) and the PerfectMatch of Zhang et al. (2003), which is a positional-dependent-nearest-neighbor model that uses probe target nucleotide sequence and position to determine background binding. Comparing t -values from each algorithm, the GeSNP algorithm performed better for a test data set comparing SAMP10 and SAMR1 at all thresholds (Table 3). At t -value thresholds of 6 and 7, the improvement in performance was significant using a χ^2 test ($P = 0.0027$ and $P = 0.024$, respectively). These results show that in identifying confirmed sequence differences, the GeSNP algorithm is more robust and accurate than similar computational methods.

Improving the quality of array-based gene expression data

Sequence differences in gene regions covered by the oligonucleotide probes can affect the quantitative measurement of expression levels. This effect is especially important when mRNA from one species is interrogated with arrays designed for a different species. The ability to identify probes that cover regions with sequence differences and eliminate them from the analysis is essential for accurate gene expression quantification (Cáceres et al. 2003; Karaman et al. 2003; Khaitovich et al. 2004). For example, in the comparison of the human and chimpanzee gene expression data, we identified 246 probe sets that initially appeared to be expressed at different levels between human and

Table 3. SAMP10 versus SAMR1 performance, GeSNP versus Ronald et al. (2005)

	GeSNP			Ronald et al. (2005)		
	$t \geq 5$	$t \geq 6$	$t \geq 7$	$t \geq 5$	$t \geq 6^a$	$t \geq 7^a$
Probe pair analysis						
TP	21	21	20	10	8	8
FN	8	8	9	11	13	13
TN	377	387	388	314	315	320
FP	12	2	1	12	11	6
TPR	0.75	0.91	0.95	0.45	0.42	0.57
DR	0.72	0.72	0.69	0.48	0.38	0.38
Probe set analysis						
TP	7	7	7	6	5	5
FN	1	1	1	2	2	2
TN	18	19	19	18	18	18
FP	1	0	0	1	2	2
TPR	0.88	1.00	1.00	0.86	0.71	0.71
DR	0.88	0.88	0.88	0.75	0.71	0.71

(TP) True positive; (FN) false negative; (TN) true negative; (FP) false positive; (TPR) true positive rate (the positive predictive value), $[TP/(TP+FP)]$; (DR) detection rate (the sensitivity), $[TP/(TP+FN)]$.

^aSignificantly different using a χ^2 test.

chimpanzee brains (Cáceres et al. 2003). However, once the signals for the probe pairs predicted to have sequence differences were ignored or “masked,” 53 of these probe sets were no longer scored as being differentially expressed. One of the genes, *CTNNA1*, was independently determined by quantitative RT-PCR to be expressed at the same level in humans and chimpanzees (Cáceres et al. 2003). Sequence variation also affected probe hybridization and led to false assignments of differential expression in the SAM comparisons. For example, *Fgf1* gene expression levels appeared lower in SAMP10 mice than in SAMR1 mice. However, GeSNP identified sequence differences in four *Fgf1* probe pairs. Sequencing confirmed that nucleotide differences were covered by the *Fgf1* probes and further led to the identification of a functionally important mutation outside of the probe set region (Carter et al. 2005). Quantitative RT-PCR showed no gene expression difference between SAMP10 and SAMR1 mice, and after masking the affected probe pairs, *Fgf1* was no longer considered differentially expressed. Finally, sequence variation that affects probe hybridization was identified in another mouse study comparing C57BL/6J with 129S6/SvEvTac mice for the gene *Kcnab2* (Sandberg et al. 2000). Therefore, although the occurrence of this phenomenon is more frequent in interspecies studies, the potential effects on gene expression measurements due to sequence differences should be routinely investigated.

Identification of disease-causing mutations in human disease groups

In order to test the ability of the GeSNP algorithm to identify sequence differences between human disease populations, human gene expression data from a study on inflammatory bowel disease were analyzed (Burczynski et al. 2006). Data for peripheral blood samples on Affymetrix HG-U133A arrays were obtained from GEO accession number GSE3365 (<http://www.ncbi.nlm.nih.gov/projects/geo/>) and directly from the investigators. The aim of the study was to identify gene expression signatures from peripheral blood mononuclear cells that could discriminate between two common inflammatory bowel diseases, Crohn's disease and ulcerative colitis. Both disorders are thought to result from genetic and environmental factors that lead to an abnormal immune response in the gastrointestinal tract, with Crohn's disease having a larger genetic component than ulcerative colitis (Sartor 2006).

The data of 59 Crohn's disease patients, 26 ulcerative colitis patients, and 42 healthy controls were analyzed with the GeSNP algorithm in order to identify potential sequence differences between these groups. Several previously identified Crohn's disease and ulcerative colitis candidate susceptibility genes showed SFPs based on the GeSNP analysis, including *SLC22A4*, identified in linkage and association studies (Ma et al. 1999; Rioux et al. 2000, 2001; Giallourakis et al. 2003; Waller et al. 2006; among others) and showing functional genetic variants (Peltekova et al. 2004), and *TLR4* (Franchimont et al. 2004; Gazouli et al. 2005; Noble et al. 2006) and *IL1RN* (Tountas et al. 1999; Carter et al. 2001), both of which have been associated with Crohn's disease and ulcerative colitis in certain populations. In addition, many promising new candidates were also identified (see Supplemental Tables 4–6), including two interesting but as yet unimplicated genes that could be involved in inflammatory bowel disease pathogenesis, *VIL2* and *HMGB1*, and also *F2RL1*, which could be implicated in the differences between Crohn's disease and ulcerative colitis.

Discussion

The results demonstrate that the GeSNP algorithm can identify sequence differences using array-based gene expression data. The approach is general to several Affymetrix gene expression array types and applicable to the analysis of data obtained in different populations of genetically distinct individuals, including humans. With most array designs, the sequence coverage for each gene is incomplete. Usually 100–400 bases of sequence are interrogated for each gene since there are typically 11–20 probe pairs per gene, the probes are 25 bases in length, some of the probes are overlapping, and sequence differences that result in mismatches near the probe ends (e.g., the five bases at either end) are not expected to lead to consistently measurable hybridization differences (Pease et al. 1994; Chee et al. 1996). Nonetheless, this approach allowed us to take advantage of previously existing data, obtained initially for other purposes, to search in a broad and unbiased way for genetic differences without the need for any additional experiments. The GeSNP program can be used not only to identify small sequence differences, such as single-base substitutions, but also larger deletions or insertions and genes with different splice forms (Winzeler et al. 1998; Hu et al. 2001; Li and Wong 2001). We further illustrated the additional information that can be generated with publicly available data files that contain detailed clinical or phenotypic information. Using GeSNP, we identified several well-known inflammatory bowel disease candidate genes and many new, promising candidates that are consistent with the disease pathophysiology. Thus, this analysis method can be used to complement gene expression and other more traditional studies to accelerate the identification of genes that may mediate important diseases and phenotypes.

In addition to the identification of genetic variants, the analysis methods described here may find their most immediate application in improving array performance and enabling arrays designed for one strain or species to be used more broadly. We have used this technique successfully in the past to improve the quality of gene expression data by masking probes that cover regions with potential sequence differences in both mouse (Carter et al. 2005) and human studies (Cáceres et al. 2003). Identifying sequence variation that may influence hybridization patterns and lead to incorrect results is even more critical in eQTL analysis. There is growing interest in using eQTL studies to discover loci genetically associated with gene expression differences and to determine transcriptional regulatory networks (Schadt et al. 2003; Bystrykh et al. 2005; Chesler et al. 2005). However, SNPs within a probe region that affect expression results might be in linkage disequilibrium with a marker SNP and lead to a false eQTL association with the marker (Peirce et al. 2006). As eQTL studies become even more prominent, methods that minimize false positive associations will be increasingly important.

Moreover, as new array designs become more widely used, the GeSNP algorithm could have a much larger impact. For example, Affymetrix recently released the exon arrays to interrogate all putative exons in a genome. The human array contains 1.4 million probe sets with four PM probes per set (5.6 million probes, 140 million nucleotides). Assuming that half of the probe sets pass the pattern quality measure of detectable expression and that only 50% of the covered nucleotides provide information due to probe sequence overlap and lower sensitivity to differences at the probe ends, analysis with the GeSNP algorithm could yield information on ~35 million bases of sequence. However, because specific MMs for each probe are not part of the new exon

array design, the pattern quality control and background subtraction techniques would need to be modified in order to apply GeSNP to this new array type.

In summary, the GeSNP algorithm allows for the unbiased, opportunistic extraction of sequence variation information from array-based gene expression data. This information can be used to improve the quality of gene expression and eQTL analyses and to identify potential disease-causing genes in human disease populations. The GeSNP source code and a Web-based program are available for public implementation.

Methods

Computer software

The algorithm was written in standard ANSI C++ and compiled to run on UNIX. The extensively commented source code is available for download from Supplemental materials at the *Genome Research* Web site and the GeSNP Web site, <http://porifera.ucsd.edu/~cabney/cgi-bin/geSNP.cgi>. In addition, the GeSNP Web site hosts a user-friendly Web-based tool that allows users to upload their expression data in two predefined groups and obtain results online. A user manual and example data are also available at the Web site. The GeSNP program outputs a text file for each comparison with the following columns: Probe set, Probe pair, pssp (probe set with probe pair number appended at the end), N1 (number of files included in group one), Mean1, Var1 (variance of group 1), N2 (number of files included in group two), Mean2, Var2 (variance of group 2), and *t*-value. In the *t*-value column, the value "NaN" indicates that there were less than two files included in one or both groups and a *t*-value could not be calculated. Probe sets where both groups had zero files passing the pattern quality measures are not included in the output.

False positive estimation and choosing a threshold

Because the *P*-value of the Student's *t*-test and a permuted *P*-value with 100,000 permutations did not perform as well as the *t*-value alone in correctly identifying sequence differences, we developed a method to obtain an approximate, "predicted" true positive rate in order to determine an appropriate *t*-value threshold. First, a false positive comparison is generated, where the two groups of interest are equally distributed into two false positive groups, accounting for any subgroup bias in tissue type, race, gender, or prominent diseases. Ideally, no differences should be identified between these two randomized groups. The number of PPs exceeding a *t*-value threshold for the false positive comparison yields an estimated number of false positives for the specific files and number of samples being compared. Subtracting the number of estimated false positives at a given *t*-value from the number of putative SFPs, then dividing by the number of SFPs, yields a predicted true positive rate. The larger the number of independent samples for a comparison, the more accurate the results, assuming no subgroup bias is introduced. For studies within a species with homozygous loci, at least four independent samples should be used in each group. For studies within outbred populations, use of at least 10 independent samples per group is advisable.

Analysis of SFPs

Result files were filtered in Microsoft Access according to the minimum number of files ($N1$ and $N2 \leq 4$) and the *t*-value threshold. In the supplemental tables, the data are organized by probe set to illustrate important summary information. The number of SFPs in a probe set and the largest *t*-value of these SFPs

(with at least one positive mean) are shown. A larger number of SFPs within a probe set, a greater *t*-value, and/or multiple probe sets representing a single gene provide increased confidence that a true sequence difference exists for that gene. Annotation files were downloaded from Affymetrix (<http://www.affymetrix.com/analysis/index.affx>). Additional information on candidate genes was obtained from NCBI's Entrez (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>), OMIM (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>), and PubMed (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>).

Since increasing the number of files can improve analytical power, we also compared combined groups. For example, we combined ulcerative colitis and Crohn's disease samples and compared this group with control samples. While all the SFPs in common to both the ulcerative colitis versus normal and Crohn's disease versus normal lists appear in the combined comparison, additional probe sets are identified as SFPs. These probe sets may have more subtle hybridization differences (such as sequence differences near the ends of the probes) that are enhanced with the larger number of files.

Implementation of the Ronald et al. (2005) algorithm

James Ronald and Leonid Kruglyak kindly provided their C++ program (pdse.cpp) implementing some of the normalization procedure of Irizarry et al. (2003) and the PerfectMatch of Zhang et al. (2003). We then wrote a program in MATLAB to follow the remaining methods outlined in Ronald et al. (2005). Using the output of pdse.cpp for the SAMP10 versus SAMR1 comparison, the MATLAB program divided the observed intensity by the expected intensity (while expected intensity >100) and then calculated group means, group variances, and the *t*-values between groups.

RNA preparation and cDNA synthesis for sequence confirmation

Total RNA was prepared using TRIzol Reagent (Gibco/BRL) following the manufacturer's recommended protocol. For SAM strains, RNA was extracted from the cortex of at least two separate mice for each strain. Standard protocols were used for the generation of cDNA from RNA. Primers were designed to amplify the regions defined by the Affymetrix probe set target sequences of the selected genes, which can be downloaded from the Affymetrix Analysis Center Web site. Standard PCR reactions were performed on an Applied Biosystems GeneAmp PCR System 9700, and PCR products were purified using the recommended procedures for the QIAquick PCR purification kit protocol or the QIAquick gel extraction kit protocol (Qiagen). All sequencing was performed by the Salk Institute Sequencing Core. The sequences of human genes were obtained from GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/index.html>). Chimpanzee sequences were described in Cáceres et al. (2003) and were also obtained from NCBI's GenBank. For the global comparison of the HG-U95Av2 Affymetrix array probes to the human and chimpanzee genomes, we used the sequence assembly versions HG18 (NCBI Build 36.1) and PanTro2 (Build 2, Version 1). Probe sequences were aligned to the genome sequences using MegaBlast, and only probes with 100% identity over the 25 nucleotides were selected.

Acknowledgments

We thank Stephen Heinemann of the Salk Institute for support, Charles Abney for Web programming assistance, Eva Mitter of IMC for technical assistance, Sebastian Fuchs for sequencing assistance, Jo Del Rio for preliminary analysis, James Thomas for

help with the human and chimpanzee genome analysis, and Svante Pääbo, Todd Carter, and Michael Burczynski for providing data files. J.A.G. was supported by a generous gift from the Lewin family and the Sprint Corporation, the NIH Neuroplasticity of Aging Training Grant (5 T32 AG00216), and the National Defense Science and Engineering Graduate Fellowship. M.C. was supported by an EMBO Long-Term Fellowship, a Salk Institute Innovation Grant, and the Ramón y Cajal Program (Ministerio de Educación y Ciencia, Spain). Additional funding was provided by the DOD grant DAMD17-99-1-9561 and the Frederick B. Rentschler Developmental Chair to C.B.

References

- Albert, T.J., Dailidiene, D., Dailide, G., Norton, J.E., Kalia, A., Richmond, T.A., Molla, M., Singh, J., Green, R.D., and Berg, D.E. 2005. Mutation discovery in bacterial genomes: Metronidazole resistance in *Helicobacter pylori*. *Nat. Methods* **2**: 951–953.
- Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E., and Chory, J. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13**: 513–523.
- Burczynski, M.E., Peterson, R.L., Twine, N.C., Zuberek, K.A., Brodeur, B.J., Casciotti, L., Maganti, V., Reddy, P.S., Strahs, A., Immermann, F., et al. 2006. Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J. Mol. Diagn.* **8**: 51–61.
- Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., Wiltshire, T., Su, A.I., Vellenga, E., Wang, J., Manly, K.F., et al. 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics.' *Nat. Genet.* **37**: 225–232.
- Cáceres, M., Laucher, J., Zapala, M.A., Redmond, J.C., Kudo, L., Geschwind, D.H., Lockhart, D.J., Preuss, T.M., and Barlow, C. 2003. Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl. Acad. Sci.* **100**: 13030–13035.
- Carter, M.J., di Giovine, F.S., Jones, S., Mee, J., Camp, N.J., Lobo, A.J., and Duff, G.W. 2001. Association of the interleukin 1 receptor antagonist gene with ulcerative colitis in Northern European Caucasians. *Gut* **48**: 461–467.
- Carter, T.A., Greenhall, J.A., Yoshida, S., Fuchs, S., Helton, R., Swaroop, A., Lockhart, D.J., and Barlow, C. 2005. Mechanisms of aging in senescence-accelerated mice. *Genome Biol.* **6**: R48. doi: 10.1186/gb-2005-6-6-r48.
- Cervino, A.C., Li, G., Edwards, S., Zhu, J., Laurie, C., Tokiwa, G., Lum, P.Y., Wang, S., Castellini, L.W., Lusis, A.J., et al. 2005. Integrating QTL and high-density SNP analyses in mice to identify *Insig2* as a susceptibility gene for plasma cholesterol levels. *Genomics* **86**: 505–517.
- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., and Fodor, S.P. 1996. Accessing genetic information with high-density DNA arrays. *Science* **274**: 610–614.
- Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., et al. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* **37**: 233–242.
- Enard, W., Khaitovich, P., Klose, J., Zollner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., et al. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**: 340–343.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., and Solas, D. 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**: 767–773.
- Fodor, S.P., Rava, R.P., Huang, X.C., Pease, A.C., Holmes, C.P., and Adams, C.L. 1993. Multiplexed biochemical assays with biological chips. *Nature* **364**: 555–556.
- Franchimont, D., Vermeire, S., El Housni, H., Pierik, M., Van Steen, K., Gustot, T., Quertinmont, E., Abramowicz, M., Van Gossum, A., Deviere, J., et al. 2004. Deficient host-bacteria interactions in inflammatory bowel disease? The toll-like receptor (TLR)-4 Asp299gly polymorphism is associated with Crohn's disease and ulcerative colitis. *Gut* **53**: 987–992.
- Gazouli, M., Mantzaris, G., Kotsinas, A., Zacharatos, P., Papalambros, E., Archimandritis, A., Ikononopoulos, J., and Gorgoulis, V.G. 2005. Association between polymorphisms in the Toll-like receptor 4, CD14, and CARD15/NOD2 and inflammatory bowel disease in the Greek population. *World J. Gastroenterol.* **11**: 681–685.
- Geschwind, D.H. 2000. Mice, microarrays, and the genetic diversity of the brain. *Proc. Natl. Acad. Sci.* **97**: 10676–10678.
- Giallourakis, C., Stoll, M., Miller, K., Hampe, J., Lander, E.S., Daly, M.J., Schreiber, S., and Rioux, J.D. 2003. IBD5 is a general risk factor for inflammatory bowel disease: Replication of association with Crohn disease and identification of a novel association with ulcerative colitis. *Am. J. Hum. Genet.* **73**: 205–211.
- Gresham, D., Ruderfer, D.M., Pratt, S.C., Schacherer, J., Dunham, M.J., Botstein, D., and Kruglyak, L. 2006. Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* **311**: 1932–1936.
- Grupe, A., Germer, S., Usuka, J., Aud, D., Belknap, J.K., Klein, R.F., Ahluwalia, M.K., Higuchi, R., and Peltz, G. 2001. In silico mapping of complex disease-related traits in mice. *Science* **292**: 1915–1918.
- Hacia, J.G., Brody, L.C., Chee, M.S., Fodor, S.P., and Collins, F.S. 1996. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat. Genet.* **14**: 441–447.
- Hazen, S.P., Borevitz, J.O., Harmon, F.G., Prunedo-Paz, J.L., Schultz, T.F., Yanovsky, M.J., Liljegen, S.J., Ecker, J.R., and Kay, S.A. 2005. Rapid array mapping of circadian clock and developmental mutations in *Arabidopsis*. *Plant Physiol.* **138**: 990–997.
- Hovatta, I., Zapala, M.A., Broide, R.S., Schadt, E.E., Libiger, O., Schork, N.J., Lockhart, D.J., and Barlow, C. 2007. DNA variation and brain region-specific expression profiles exhibit different relationships between inbred mouse strains: Implications for eQTL mapping studies. *Genome Biol.* **8**: R25. doi: 10.1186/gb-2007-8-2-r25.
- Hu, G.K., Madore, S.J., Moldover, B., Jatke, T., Balaban, D., Thomas, J., and Wang, Y. 2001. Predicting splice variant from DNA chip expression data. *Genome Res.* **11**: 1237–1245.
- Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., et al. 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**: 243–253.
- Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**: e15. doi: 10.1093/nar/gng015.
- Karaman, M.W., Houck, M.L., Chemnick, L.G., Nagpal, S., Chawannakul, D., Sudano, D., Pike, B.L., Ho, V.V., Ryder, O.A., and Hacia, J.G. 2003. Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. *Genome Res.* **13**: 1619–1630.
- Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigle, S., Do, H.H., Weiss, G., Enard, W., et al. 2004. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* **14**: 1462–1473.
- Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief. Bioinform.* **5**: 150–163.
- Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al. 2006. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**: 1929–1935.
- Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* **98**: 31–36.
- Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. 1999. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**: 20–24.
- Lockhart, D.J. and Winzler, E.A. 2000. Genomics, gene expression and DNA arrays. *Nature* **405**: 827–836.
- Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, K.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**: 1675–1680.
- Ma, Y., Ohmen, J.D., Li, Z., Bentley, L.G., McElree, C., Pressman, S., Targan, S.R., Fischel-Ghodsian, N., Rotter, J.I., and Yang, H. 1999. A genome-wide search identifies potential new susceptibility loci for Crohn's disease. *Inflamm. Bowel Dis.* **5**: 271–278.
- Nagpal, S., Karaman, M.W., Timmerman, M.M., Ho, V.V., Pike, B.L., and Hacia, J.G. 2004. Improving the sensitivity and specificity of gene expression analysis in highly related organisms through the use of electronic masks. *Nucleic Acids Res.* doi: 10.1093/nar/gnh048.
- Noble, C., Nimmo, E., Gaya, D., Russell, R.K., and Satsangi, J. 2006. Novel susceptibility genes in inflammatory bowel disease. *World J. Gastroenterol.* **12**: 1991–1999.
- Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., and Fodor, S.P. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci.* **91**: 5022–5026.
- Peirce, J.L., Li, H., Wang, J., Manly, K.F., Hitzemann, R.J., Belknap, J.K., Rosen, G.D., Goodwin, S., Sutter, T.R., Williams, R.W., et al. 2006.

- How replicable are mRNA expression QTL? *Mamm. Genome* **17**: 643–656.
- Peltekova, V.D., Wintle, R.F., Rubin, L.A., Amos, C.I., Huang, Q., Gu, X., Newman, B., Van Oene, M., Cescon, D., Greenberg, G., et al. 2004. Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat. Genet.* **36**: 471–475.
- Radcliffe, R.A., Lee, M.J., and Williams, R.W. 2006. Prediction of cis-QTLs in a pair of inbred mouse strains with the use of expression and haplotype data from public databases. *Mamm. Genome* **17**: 629–642.
- Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., McLeod, R.S., Griffiths, A.M., Green, T., Brettin, T.S., Stone, V., Bull, S.B., et al. 2000. Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *Am. J. Hum. Genet.* **66**: 1863–1870.
- Rioux, J.D., Daly, M.J., Silverberg, M.S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., et al. 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* **29**: 223–228.
- Ronald, J., Akey, J.M., Whittle, J., Smith, E.N., Yvert, G., and Kruglyak, L. 2005. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.* **15**: 284–291.
- Sandberg, R., Yasuda, R., Pankratz, D.G., Carter, T.A., Del Rio, J.A., Wodicka, L., Mayford, M., Lockhart, D.J., and Barlow, C. 2000. Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl. Acad. Sci.* **97**: 11038–11043.
- Sartor, R.B. 2006. Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nat. Clin. Pract. Gastroenterol. Hepatol.* **3**: 390–407.
- Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
- Shi, L., Reid, L.H., Jones, W.D., Shippey, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y., et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**: 1151–1161.
- Takeda, T. 1999. Senescence-accelerated mouse (SAM): A biogerontological resource in aging research. *Neurobiol. Aging* **20**: 105–110.
- Tountas, N.A., Casini-Raggi, V., Yang, H., Di Giovine, F.S., Vecchi, M., Kam, L., Melani, L., Pizarro, T.T., Rotter, J.I., and Cominelli, F. 1999. Functional and ethnic association of allele 2 of the interleukin-1 receptor antagonist gene in ulcerative colitis. *Gastroenterology* **117**: 806–813.
- Waller, S., Tremelling, M., Bredin, F., Godfrey, L., Howson, J., and Parkes, M. 2006. Evidence for association of OCTN genes and IBD5 with ulcerative colitis. *Gut* **55**: 809–814.
- Winzler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J., et al. 1998. Direct allelic variation scanning of the yeast genome. *Science* **281**: 1194–1197.
- Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H., and Lockhart, D.J. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15**: 1359–1367.
- Xia, C., Higuchi, K., Shimizu, M., Matsushita, T., Kogishi, K., Wang, J., Chiba, T., Festing, M.F., and Hosokawa, M. 1999. Genetic typing of the senescence-accelerated mouse (SAM) strains with microsatellite markers. *Mamm. Genome* **10**: 235–238.
- Zhang, L., Miles, M.F., and Aldape, K.D. 2003. A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* **21**: 818–821.

Received January 23, 2007; accepted in revised form April 23, 2007.

Detecting genetic variation in microarray expression data

Jennifer A. Greenhall, Matthew A. Zapala, Mario Cáceres, et al.

Genome Res. 2007 17: 1228-1235 originally published online July 3, 2007

Access the most recent version at doi:[10.1101/gr.6307307](https://doi.org/10.1101/gr.6307307)

Supplemental Material

<http://genome.cshlp.org/content/suppl/2007/07/03/gr.6307307.DC1>

References

This article cites 54 articles, 21 of which can be accessed free at:

<http://genome.cshlp.org/content/17/8/1228.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Focused on your science.

USC
SCIENTIFIC
SIDE BY SCIENCE

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>