nature biotechnology

Microarray reality checks in the context of a complex disease

George L Gabor Miklos & Ryszard Maleszka

A problem in analyzing microarray-based gene expression data is the separation of genes causally involved in a disease from innocent bystander genes, whose expression levels have been secondarily altered by primary changes elsewhere. To investigate this issue systematically in the context of a class of complex human diseases, we have compared microarray-based gene expression data with non-microarray-based clinical and biological data about the schizophrenias to ask whether these two approaches prioritize the same genes. We find that genes whose expression changes are deemed to be of importance from microarrays are rarely those classified as of importance from clinical, *in situ*, molecular, single-nucleotide polymorphism (SNP) association, knockout and drug perturbation data. This disparity is not limited to the schizophrenias but characterizes other human disease data sets. It also extends to biological validation of microarray data in model organisms, in which genome-wide phenotypic data have been systematically compared with microarray data. In addition, different bioinformatic protocols applied to the same microarray data yield quite different gene sets and thus make clinical decisions less straightforward. We discuss how progress may be improved in the clinical area by the assignment of high-quality phenotypic values to each member of a microarray-assigned gene set.

Microarrays of diverse types[1–3] are seen as the new divining rods of molecular medicine[4]. The recent availability of advanced commercial RNA expression platforms and the milestone of nearly 5,000 published microarray papers[5] demonstrate the rapid inroads made by this technology into basic and applied research. Hundreds of microarray papers attest to their proven track record at both the genomic and RNA expression levels[1]. Their use in comparative genomic hybridizations and in SNP detection is undisputed. They are also a powerful taxonomic tool because of their ability to discriminate between very different cell types, such as cancers, as well as between the effects of various drugs on cells and between the effects of different pathogens on cells. However, their value as classifiers of data is quite different from their ability to determine which genes and networks are causative in a particular disease or phenomenon, an area that requires biological validation and extensive bioinformatics. Concomitant with the production of these massive data sets, there has been a proliferation of sophisticated bioinformatic protocols for examining gene expression data[6]. These algorithms filter the raw data in different ways and generate gene sets that then form the basis for years of further diagnostic, clinical and pharmaceutical investment. At the moment, it is not at all clear whether bioinformatic protocols are robust enough to prioritize key genes in disease networks.

To evaluate the power of microarrays and their attendant bioinformatic protocols to generate and filter raw data, we have examined microarray-based gene expression data in the context of one of the most debilitating human conditions, the schizophrenias. The schizophrenias are devastating psychiatric conditions[7–17]. They afflict at least 1% of the US population, resulting in annual costs exceeding US $65 billion; worse yet, all currently available antipsychotic drugs only alleviate the symptoms by decreasing dopaminergic transmission. The etiology and pathogenesis of the schizophrenias are unknown.

We chose this complex disease because there is a large amount of molecular, cellular, neuroanatomical, clinical, functional magnetic resonance imaging (fMRI), neuropsychiatric and drug data that can be examined in the context of microarray data. Our aim was to answer the simple question: do microarray and non-microarray data analyses highlight similar, or different, gene sets?

How the data sets compare

A total of 89 genes whose RNA expression levels were deemed to be significantly altered in the prefrontal cortices of schizophrenic patients, relative to matched controls, as examined on Affymetrix (Santa Clara, CA, USA) microarrays are detailed in **Table 1** (ref. 18). As a result of their particular bioinformatic protocols, Hakak *et al.*[18] conclude that genes involved in myelination not only provide a new area for future studies but also demonstrate the use of microarray-based gene expression profiling data for providing insights into the etiology of the schizophrenias. This 89-member gene set thus represents a typical microarray-based output on which future clinical and pharmaceutical commitments might be made.

Table 2 illustrates 49 genes whose RNA expression levels were deemed to be significantly altered in the prefrontal cortices of a different group of schizophrenic patients as examined on diverse spotted microarray platforms using cDNA probes[19–23]. As a result of their particular bioinformatic protocols, the groups of Mirnics and Levitt[19,22,23], Vawter[20] and Bahn[21] conclude that genes involved in presynaptic functions, specific metabolic alterations and lipoprotein

George L. Gabor Miklos is at Secure Genetics, 81 Bynya Road, Palm Beach, Sydney, NSW, Australia 2108 and Human Genetic Signatures. Ryszard Maleszka is at the Research School of Biological Sciences, Visual Sciences, Australian National University, Canberra, ACT 0200, Australia.
e-mail: gmiklos@securegenetics.com

Published online 30 April 2004; doi:10.1038/nbt965

Table 1 Differential gene expression in chronic schizophrenic versus normal individuals (measured with Affymetrix HuGeneFl photolithographic platforms)

Upregulated		Downregulated
↑*AMPH*	↑*NPY*	↓*ACTA2*
↑*ATP1A3*	↑*NEFL*	↓*CNP*
↑*AF1Q*	↑*NELL1*	↓*CSRP1*
↑*APP*	↑*NELL2*	↓*ERBB3*
↑*ATP6V1B2*	↑*NDN*	↓*FOLH1*
↑*ATPB2*	↑*NEF3*	↓*GSN*
↑*BACH*	↑*NPTX1*	↓*GPR37*
↑*CCK*	↑*NDUFA5*	↓*HSPA2*
↑*CALM3*	↑*OLFM1*	↓*MAL*
↑*CNR1*	↑*OXCT*	↓*MAG*
↑*CLTB*	↑*PCSK1*	↓*RNASE1*
↑*CCND2*	↑*PRKCB1*	↓*SLC14A1*
↑*CRYM*	↑*PRKAR2B*	↓*SLC31A2*
↑*C5orf13*	↑*PPP3CB*	↓*SEPP1*
↑*DSCR1L1*	↑*PFN2*	↓*TXNIP*
↑*ELAVL4*	↑*PCP4*	↓*TIP1*
↑*FAM3C*	↑*RIT2*	↓*TF*
↑*GAP43*	↑*RPS4Y*	
↑*GABRA1*	↑*RGS7*	
↑*GAD1*	↑*RTN1*	
↑*GAD2*	↑*RIMS3*	
↑*GLRB*	↑*SLC12A5*	
↑*GOT1*	↑*STMN2*	
↑*HSPH1*	↑*SNRK*	
↑*HPCAL1*	↑*SGNE1*	
↑*HPRT1*	↑*SERPINI1*	
↑*HSPA1B*	↑*SST*	
↑*HT2626*	↑*SCG2*	
↑*ITPR1*	↑*SNCB*	
↑*KNCK1*	↑*SMARCA2*	
↑*KIAA0252*	↑*TA-LRRP*	
↑*KNS2*	↑*TAC1*	
↑*LMO4*	↑*UBE2V2*	
↑*MARCKS*	↑*VDAC1*	
↑*MMD*	↑*23682*	
↑*MDH1*	↑Germ mRNA	

Genes shown are considered to be significantly altered in their expression levels (1.4-fold or greater changes), according to the particular bioinformatic protocols used by Hakak *et al.*[18]. Gene symbols denote National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA) LocusLink nomenclature derived from accession numbers provided by the authors. ↑, increased, and ↓, decreased mRNA expression in schizophrenic patients relative to appropriate controls.

Table 2 Differential gene expression in schizophrenic versus normal individuals (measured on spotted microarrays)

Upregulated	Downregulated	
↑*APOL1*	↓*ACLY*	↓*TIMM17A*
↑*CAP2*	↓*ASNS*	↓*UCHL1*
↑*CREBL2*	↓*ATP5A1*	↓*USP9X*
↑*ENO2*	↓*ATP6V1C1*	↓*USP14*
↑*PIK3CD*	↓*CALM3*	
↑*PIK4CA*	↓*CRYM*	
↑*PRKACB*	↓*DLD*	
↑*PRKCB1*	↓*DYRK1A*	
↑*RFXANK*	↓*GAD1*	
	↓*GNL1*	
	↓*GOT1*	
	↓*GOT2*	
	↓*GRIA1*	
	↓*GRIA2*	
	↓*IDH3*	
	↓*MAP3K1*	
	↓*MDH1*	
	↓*MINK*	
	↓*NARS*	
	↓*NF2*	
	↓*NSF*	
	↓*OAT*	
	↓*OAZIN*	
	↓*OXCT*	
	↓*PLP1*	
	↓*POR*	
	↓*POU6F1*	
	↓*PRKX*	
	↓*PSMA1*	
	↓*RGS4*	
	↓*SERPINI1*	
	↓*SLC10A1*	
	↓*SYN2*	
	↓*SYNGR1*	
	↓*SYNJ1*	
	↓*SYT5*	

Genes shown are considered to be significantly altered in their expression levels (>1.9-fold changes or high *Z* scores), according to the particular bioinformatic protocols used by the authors[19–23]. The data derive from spotted microarray platforms from both commercial and custom-made sources (UniGEM-V, UniGEM-V2 platforms from Incyte (Palo Alto, CA, USA) and custom-made microarrays). Study designs were different between different investigators as were the bioinformatic protocols used to analyze the data. Gene symbols denote NCBI LocusLink nomenclature derived from the accession numbers provided by the authors[19–23]. ↑, increased, and ↓, decreased mRNA expression in schizophrenic patients relative to appropriate controls.

pathways are important contributors to the etiology and pathogenesis of the schizophrenias.

Table 3 shows data from almost 100 studies using non-microarray approaches (see **Supplementary Table 1** online). These approaches have implicated 97 genes in the schizophrenias on the basis of genetic analysis, RNA assays and protein measurements. At the genetic level, correlations were found on the basis of the following approaches: differences in allele frequencies at a particular locus between schizophrenic and normal individuals, linkage between certain chromosomal regions and phenotypic manifestations of the schizophrenias, and translocation breakpoints within or near to a gene and associated phenotypic effects. At the RNA level, measurements involved differences in RNA expression between brain regions of schizophrenic and control individuals on the basis of *in situ* hybridization of labeled probes to tissue sections of particular brain regions, reverse transcription polymerase chain reaction (RT-PCR) on tissue samples from the brains of schizophrenic versus normal individuals and cDNA library subtractive hybridization experiments between libraries made from tissues of schizophrenic and normal individuals. At the protein level, the differences in expression between schizophrenic and control individuals were measured using antibodies incubated to brain tissue sections; western blotting immunohistochemistry; enzyme-linked immunosorbent assay (ELISA) to tissue homogenates from schizophrenic and normal brains; binding of labeled ligands or drugs to

tissue sections of brains and then *in situ* visualization; measurements of protein levels by electrophoresis, by radioimmunoassay or using monoclonal antibodies in various human tissues such as brain, cerebrospinal fluid and blood; or classical differences in enzyme activity from schizophrenic and normal individuals. The hundreds of authors associated with these studies variously conclude that genes involved in the glutamatergic, serotonergic and dopaminergic pathways, in calcium signaling, in ion-channel processes, in G-protein signaling, in glycoprotein and lipoprotein metabolism, and in transcription factors are the important entities in the schizophrenias. No unifying hypothesis has yet emerged.

A comparison of the gene sets given precedence from samples analyzed on Affymetrix and spotted platforms is given in **Tables 1** and **2**. Of the 89 genes of **Table 1** and the 49 genes of **Table 2**, only 8 genes are common to both gene sets (**Fig. 1a**). Surprisingly, expression of 7 of the 8 common genes is found to be increased in the samples analyzed on the Affymetrix platform, whereas expression of all 7 is decreased in the samples tested on the spotted platforms (*CALM3*, *CRYM*, *GAD1*, *GOT1*, *MDH1*, *OXCT* and *SERPINI1*). Only one gene's product, that of *PRKCB1*, overlaps with a consistent directionality.

We next compared the combined 138 microarray entries of **Tables 1** and **2** with the 97 entries of **Table 3**, which derive from diverse non-microarray sources. We found only 8 genes in common that show consistent expression changes between the microarray and 'non-microarray' gene sets, these being *ERBB3*, *GAP43*, *GRIA1*, *GRIA2*, *MAG*, *PLP1*, *SYN2* (**Fig. 1b**), and only 2 genes that are common to all three gene sets: glutamic acid decarboxylase (*GAD1*) and malate dehydrogenase (*MDH1*). These two particular genes have been extensively studied by different researchers[22,24] in the schizophrenias, and their levels of expression as measured by a variety of the techniques described earlier are consistently decreased in all previous single-gene publications and on spotted arrays. Surprisingly, in the tests on Affymetrix platforms, they show increased expression.

We further asked whether microarray-based gene sets determined from other brain regions of schizophrenic individuals, such as the cerebellum, middle temporal gyrus and entorhinal cortex[20], overlapped with the non-microarray gene set of **Table 3**. The overlap was again small. Therefore, irrespective of the brain region studied, there is a notable difference between gene sets deemed important from different types of expression microarrays and those deemed important by non-microarray approaches.

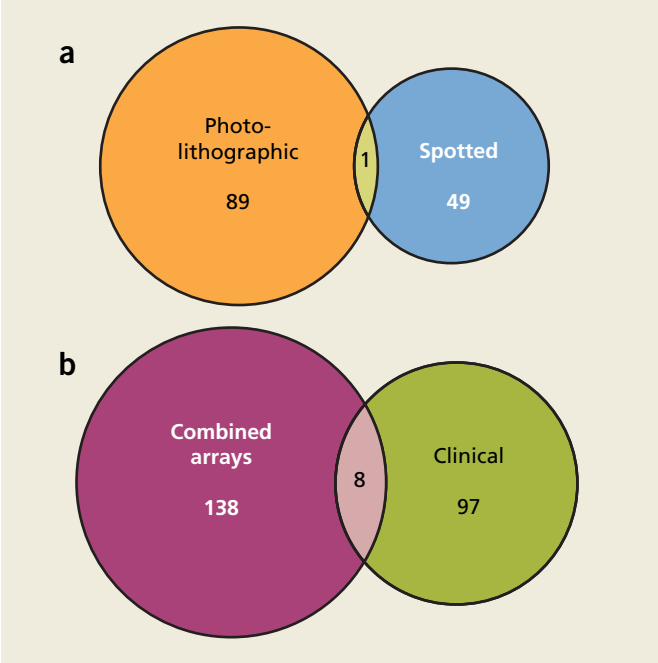

Figure 1 Venn diagram illustrating the overlap between genes prioritized by different methodologies. (**a**) Data from photolithographic arrays[18] shown in orange versus data from spotted microarrays[19–23] shown in blue. The relevant probes on the spotted microarrays had representative members for most of the same transcripts on the photolithographic microarray. (**b**) Combined microarray data from photolithographic and spotted arrays (magenta) versus data derived from curation of the literature (green).

Sources of heterogeneity

Given the disparity in the data from **Tables 1–3**, what variables could account for the differences in these gene sets and what improvements can be implemented to allow more reproducible application of microarrays in the clinical setting?

Tissue sample complexity. Tissue samples from a normal human cortex and deeper brain structures, each with its different proportion of cell types drawn from cortical-cerebellar-thalamic circuitry, are heterogeneous. Even more so are post mortem samples from the brains of schizophrenic individuals, in which the size of the thalamus can be reduced and in which there can be neuronal loss in the subnuclei of the medial dorsal nucleus, in the cingulate cortex, hippocampal formation, entorhinal cortex and parahippocampal gyrus[25,26]. These findings may be due to developmental events before or after birth or to progression of pathology after the onset of symptoms. Furthermore, the subsyndromes of the schizophrenias, analyzed by functional MRI, show anatomically distinct psychophysiological associations between subsyndromal symptom scores and neuroanatomical areas[11,17]. These altered cytoarchitectures are the clinical starting material for microarray analyses.

It is known that considerable neuroanatomical differences occur in the prefrontal cortices of normal individuals and even between monozygotic twins[27]. Large between-individual variations have been found by MRI in other brain regions of normal subjects and in diseased and medicated patients. For example, positron emission tomography (PET) studies using radioligands show the differences in serotonin receptor densities that can occur between schizophrenic patients taking different medications[9]. Molecular data show low concentrations of synapsin-1 protein in the hippocampus of some schizophrenic patients, whereas others have normal concentrations[28]. This variability is illustrated in the non-microarray data of **Table 3**, where, for example, synaptophysin (*SYP*) gene expression can be increased, decreased or remain unchanged. For proteolipid protein-1, some schizophrenic patients show decreases, others increase and still others remain largely unchanged[23]. Similarly, for synaptojanin-1, synaptotagmin-5 and synapsin-2, expression changes are found in only some subjects[19]. Similar findings apply to SNP data, in which many of the associations identified with the schizophrenias, such as that of *NOTCH4*, are not confirmed between independent studies[29]. It should be emphasized that none of the 89 genes of **Table 1** are differentially expressed in all schizophrenic patients[18].

It is important to note that post mortem brain tissue samples will always consist of a pooled population of heterogeneously expressing neurons and support cells, a point that has been repeatedly emphasized in the literature[19,22,23]. For example, *in situ* data show that GAD1 is significantly decreased in cortical layers 3–5 of the prefrontal cortex of schizophrenic patients, a decrease that is restricted to only a particular subset of γ-aminobutyric acid (GABA) neurons. GAD1 is relatively unaltered in most GABA neurons but is reduced to below the

Table 3 Candidate genes deemed significant in schizophrenia in published reports

Upregulated	Downregulated	Unchanged[a]	SNP association[b]
↑*ADRB1*	↓*CALB1*	***CAMK2****	+*APOE*
↑*APOD*	↓*CCK*	***CHGA****	+*CCKAR*
↑*APOL2*	↓*CCKAR*	***CHGB****	+*CNR1*
↑*APOL4*	↓*CCKBR*	***DRD2****	+/–*COMT*
↑*ATF2*	↓***CHGA****	***GABRA1****	+*CTLA4*
↑*CALCYON*	↓***CHGB****	***GABRA5****	+*DAO*
↑***CAMK2****	↓*CLDN11*	***GABRG2****	+*DAT1*
↑*CHGA*	↓*CPLX1*	*MBP*	+*DPYSL2*
↑*CREB*	↓*CPLX2*	***NOS****	+/–*DTNBP1*
↑*DNMT1*	↓*DISC1*	***NRG1****	***+/–DRD2****
↑***DRD2****	↓***DRD2****	***SNAP25****	***+/–DRD3****
↑*ELK1*	↓***DRD3****	***SYP****	+/–*DRD4*
↑*FREQ*	↓*DRD4*		+*G72*
↑***GABRA1****	↓*ERBB3*		***+GRIK3****
↑***GABRA5****	↓***GABRG2****		***+/–HTR2A****
↑*GAP43*	↓↓↓↓*GAD1*		+*HTR5A*
↑*GRM5*	↓↓↓*GRIA1*		+/–*KCNN3*
↑*JUN*	↓↓↓*GRIA2*		+*L1CAM*
↑*MPZL1*	↓*GRIK2*		+*LTA*
↑*NCAM1*	↓***GRIK3****		+/–*MAOA*
↑*NRGN*	↓*GRIK4*		***+NOS****
↑*RTN4*	↓*GRIK5*		+/–*NOTCH4*
↑*S100B*	↓*GRIN1*		***+NPY****
↑*SEMA3A*	↓*GSK3A*		***+/–NRG1****
↑***SNAP25****	↓***HTR2A****		+*PHOX2B*
↑***SYP****	↓*L1CAM*		+*PCQAP*
↑***SYN1****	↓*MAG*		+/–*PLA2G4A*
↑***VSNL1****	↓*MBP*		+*PPP3CC*
	↓*MDH1*		+/–*PRODH*
	↓*MOG*		+*RAI1*
	↓***NOS****		+/–*RGS4*
	↓*NPAS3*		+/–*TNF*
	↓***NPY****		+*UFD1L*
	↓*NTRK2*		+*ZNF74*
	↓*OLIG1*		
	↓*OLIG2*		
	↓*PLP1*		
	↓*PVALB*		
	↓*RAB3A*		
	↓↓↓*RELN*		
	↓*SHMT1/2*		
	↓*SLC1A2*		
	↓***SNAP25****		
	↓*SOX10*		
	↓*SYP*		
	↓***SYN1****		
	↓*SYN2*		
	↓*TF*		
	↓***VSNL1****		
	↓*WKL1*		

*, entry appears in more than one column. Candidate genes were described by multiple approaches at the genetic, RNA and protein levels (see text for further details). ↑, increased, and ↓, decreased mRNA or protein expression in schizophrenic patients relative to appropriate controls. Multiple icons (*e.g.*, ↓↓↓) indicate that data are from different groups or different studies. [a]Unchanged based on mRNA or protein levels measured using molecular, *in situ* or enzymological assays. [b]Data on polymorphisms (SNPs, deletions or additions) are denoted by (+) following the gene symbol if the polymorphism was inferred to be involved in a predisposition to schizophrenia (directly or indirectly) or as (–) if an independent study failed to replicate the findings in a different group of individuals. In different cases, there was insufficient information available to distinguish between the contributions of paralogs or different splice forms to increases or decreases in expression levels measured using antibody staining; there was insufficient information on the loss or gain of particular cell populations in the regions being studied by *in situ* analyses; or there was no information on increases or decreases owing to stability or degradation of the mRNA or protein products. (For further details, see **Supplementary Table 1**.)

threshold of detection in others[30]. Furthermore, *in situ* data also show that the expression of the calcium-binding protein VSNL1 is decreased in pyramidal neurons but increased in interneurons of the hippocampus, and this shift is only observed in the left hippocampus[31].

The brain regions of each schizophrenic patient, even before medication, undoubtedly have RNA expression profiles that are a function of a unique neuroanatomical cytoarchitecture and unique genetic background. Thus between-individual variability will impinge heavily on any molecular analyses, microarray or otherwise, that involve tissue samples containing different proportions of different cell types.

Platform comparisons. Different types of gene expression platforms (microarrays, real-time PCR, northern blots, differential display, serial analysis of gene expression and total gene expression analysis), each with their own strengths, throughput capacities and levels of reproducibility, could be used to validate each others' outputs. This has not yet been done on a genome-wide basis, mainly because several more urgent technical issues await resolution within each particular platform[1,32–38]: First, RNA outputs from the same gene can be quite different when assayed by different groups[39]; second, RNA expression levels estimated from different probes interrogating the same transcript are highly variable[1,40], even within the same microarray type; third, the task of distinguishing between different isoforms of the same gene is not trivial, in that the short form of the GABRG2 receptor is markedly reduced in the prefrontal cortex of schizophrenic patients, whereas the long form is not significantly changed[41]; fourth, deconvoluting the cross-hybridizations of paralogous gene products (as exemplified by the complexities of granzyme B and granzyme H expression) is not straightforward[34].

Molecular confirmation of microarray-based expression levels is currently approached by testing a small number of genes by real-time PCR, as has been carefully done in the schizophrenias and in bipolar disorder[42]. Unfortunately, comparing a small number, or even a few hundred genes between microarray and PCR platforms is not equivalent to statistical validation. Validation between technological platforms is most meaningful when both the outputs are the result of genome-wide prioritizations. For the aforementioned example, molecular validation requires comparing the top 100 genes given precedence from a 30,000-gene microarray, say, with the top 100 genes given precedence from 30,000 real-time PCRs on the same sample. Molecular validation can be considered robust when different platforms, using the same sample and treated by the same bioinformatic protocol, prioritize very similar gene sets.

Bioinformatic comparisons. Biologists and clinicians are not always fully cognizant of the intricate details and extensive manipulations that have been performed on their raw microarray data. Most researchers proceed with their most highly ranked genes based on readily available data processing packages or with the sophisticated palette of algorithms implemented by their local bioinformatics group[40,43–56]. For example, it was concluded from extensive analyses that more than 700 human genes are expressed in a cell cycle–specific manner and that the analyses provided clues to biological function for hundreds of previously uncharacterized human genes. Yet different bioinformatic protocols applied to the same data show that most of the cyclicities could arise from coincidental arrangements of measurement error and biological variation[57]. We discuss these problems in the Perspectives section later on.

Validating microarray data

Microarray-based gene validation requires independent biological and clinical information. Ideally, genome-wide phenotypic data are required to accomplish this, but these data are only partially available

from model organisms[58–63]. The most advanced validations to date have been in *Saccharomyces cerevisiae*, and the overlap between microarray-selected and biologically selected data is low[59]; thus this disparity is not a peculiarity of complex multicellular organisms. These particular yeast data cannot be emphasized strongly enough.

In contrast, there are currently insufficient genome-wide phenotypic data from humans to allow meaningful validation of microarray-based gene sets. Thus, until pragmatic inroads can be made into network analyses, biological value assignments must be made on a gene-by-gene basis with guidance from less complex systems.

There is a widely held assumption from RNA expression profiling that genes showing the greatest difference in expression between two conditions are likely to be the most important biologically[64]. However, extensive genetic, molecular and phenotypic data show quite the opposite: an alteration of twofold or more in the expression level of a gene, for example, may be no more an indicator of phenotypic importance than a change of only 0.5-fold. For example, the abundance of xanthine dehydrogenase, acid phosphatase and various esterases and dehydrogenases in *Drosophila melanogaster* can be decreased to below 1% of normal while the phenotype remains relatively unperturbed under laboratory conditions, whereas dopa decarboxylase can be increased by an order of magnitude while the phenotype remains wild type. In contrast, a reduction of <50% in myosin heavy chain output has a marked effect on muscle phenotype[65]. In mice, deletion of vimentin, a major component of the intermediate filament network, has no known effects on phenotype or reproductive ability under laboratory conditions[66], nor does knockout of many members of the extracellular matrix protein family, the tenascins[67]. In humans, approximately 20% of normal individuals in the general population completely lack α-actinin-3 (ref. 68). Furthermore, individuals with $1/100^{th}$ to $1/1{,}000^{th}$ of the normal serum concentration of albumin are largely phenotypically normal and may live a full life span[69], even though they are missing the most abundant plasma protein that transports hormones, metals, therapeutic drugs, ligands and bilirubin and maintains blood volume and colloid osmotic pressure. The foregoing sample of data predicts that some genes with small effects on phenotype[62] (innocent bystanders) will be found in the important lists of microarray data, whereas other genes with large effects on phenotype will remain hidden in the bulk of microarray data. The extent to which important genes remain hidden and innocent bystanders are given importance is currently unknown. However, it is this type of phenotypic information that is crucial to an improved evaluation of microarray-based genes.

Although gene-by-gene validations are currently the best pragmatic approach in humans, the reality is that gene products are mere parameters in networks[70,71], and that it is network fluxes and network equilibria that need to be understood. Perturbation of any network (be it intracellular, cell-cell, organ-organ or networks of any nervous system) will invariably result in a series of changes that percolate through the system until some form of stability is reached at what have been termed attractor basins by mathematicians[72]. At its most devastating extremes, initial perturbations lead to the types of complex neuropsychiatric disorders evaluated in this article. To deconvolute such illnesses, the highest quality biological and clinical values must be assigned to microarray-based genes. This is obligatory to predict either the manner in which the same networks in different human beings behave after a perturbation, or the manner in which different networks lead to similar phenotypes.

To understand complex human diseases, expression microarrays must be harnessed for time-series data generation, as they have been for model systems[73–76]. The basic rules need to be gleaned from an evaluation of network fluxes[77] network robustness[78], network topology[79], network noise[80] and network oscillations[76]. Such data are badly needed to successfully dissect the interactions between each of the different layers within an organism[81–85] all the way from the molecular level to the cognitive level[86] and to allow network reconstruction[87,88], to discern network regulatory properties[89,90] or to reprogram corrupted networks. Without clinical time series data, progress in therapeutics and drug development is likely to be severely compromised. The initial challenge, therefore, is to find matching molecular, cellular and clinical criteria that will allow filtering of the innocent bystander gene products (those whose effects on network fluxes are phenotypically minimal in a given genetic background) away from those that are at hypersensitive network nodes (perturbation of which has far-reaching consequences for phenotype). This task can be accelerated if knowledge can be gained of the types of human gene products and their expression levels, as exemplified by α-actinin-3 and serum albumin. Such data allow biologically realistic values to be attached to individual members of microarray gene sets.

Perspectives

Our analysis shows that for complex diseases, such as the schizophrenias, a considerable discrepancy exists between the differentially expressed genes identified on photolithographically synthesized oligonucleotide arrays and those identified on cDNA microarrays. Moreover, these microarray-based gene sets barely overlap with the non-microarray-based gene sets, and there is little indication of which source provides the more robust information about the etiology and pathogenesis of disease. We have also compared spotted microarray-based gene sets with non-microarray-based gene sets from Alzheimer patient samples (G.L.G.M., data not shown), and once again there was only a small overlap between the two sets of genes. Thus, we contend that the absence of overlap that we have identified in the schizophrenias is unlikely to be a peculiarity of that class of complex diseases.

The discrepancies between gene sets obtained using different technologies probably reflect the different medication histories and ages of the schizophrenic patients[18], the different terminal medical conditions[91], the differences in probes sets between platforms and the different bioinformatic protocols. It is also possible that none of the genes described in **Tables 1–3** is causative and that their altered activities faithfully represent and report on the metabolic status of a brain at the time of death of a particular schizophrenic individual. Certain viral and bacterial infections, acting in particular genetic backgrounds, may also initiate diverse processes that lead to common disease endpoints[13–15]. However, whatever initial perturbations and subsequent changes alter the cytoarchitectures and chemistries of schizophrenic individuals, the etiology and pathogenesis remain unknown.

This general problem of data evaluation using different bioinformatic protocols has recently been systematically addressed by the Critical Assessment of Microarray Data Analysis (CAMDA) group[6], a forum similar to the Genome Annotation Assessment Group that stringently evaluated standards for genomic annotation in the context of genetic, molecular and biological data sets[92]. At the 2003 CAMDA meeting, more than 150 statisticians, computer scientists and biologists, including representatives from both academia and GlaxoSmithKline), analyzed the same four microarray data sets, which included two different Affymetrix platforms as well as spotted arrays[6,53–56]. In one such analysis, a particular set of 26 genes was identified whose expression levels offered prognostic information on the survival of lung cancer patients. This group of genes was distilled from the microarray data using a particular suite of algorithmic protocols[50]. However, none of these 26 genes appeared in the top 100 gene set

published by the original authors. These microarray data were also independently analyzed by two different bioinformatic groups from GlaxoSmithKline (Stevenage, UK, and Collegeville, PA, USA)[51], again employing different bioinformatic protocols. They found 33 genes that assigned lung cancer patients into groups based on survival, but again, most of these 33 genes were not deemed to have importance in the original publications. Finally, analysis of spotted arrays using yet different bioinformatic protocols also identified genes correlated with clinical outcome[93], but again, different gene sets overlapped variably with the group identified in the original spotted microarray papers. Thus, different analyses of the same microarray data lead to different gene prioritizations.

We conclude that in this current state of differing bioinformatic prioritizations (each of which undoubtedly has some superior attributes, but none of which can *a priori* be ranked as more powerful than any other without recourse to independent biological and clinical data), it may be prudent to evaluate microarray data by a diverse suite of bioinformatic protocols and to assign potential clinical relevance only to those genes that are reproducibly selected by all protocols.

Different types of microarrays provide user-friendly genome-wide and transcriptome-wide initiation into biological systems. Some, such as SNP-oriented genotyping microarrays, are relatively straightforward[94]. Others, such as those measuring gene expression levels in alternatively spliced transcriptomes, are still problematic. The uncertainty associated with sets of genes that have been highly ranked from microarray data stems from the use of cutting-edge technology to report on conditions of remarkable biological complexity involving only dual comparisons at single time points, namely normal versus diseased contexts. To understand the etiology and the ongoing changes in disease requires time series analyses of samples from the individual patient to compare the outputs of any sophisticated technology with treatments of therapeutic relevance.

In summary, to validate clinical and therapeutic utility, data selected by bioinformatics must pass the following two-part test: First, in each different genetic background studied, innocent bystander genes and those genes causally involved in the phenomenon under investigation must be separable; second, those genes or gene products that might be manipulated to return a perturbed system to a normal state must be identified.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank Alberto Ferrus for very helpful discussions and Rob Dunne, David Mitchell and Glenn Stone for clarification of bioinformatic protocols related to expression microarray data deconvolution.

COMPETING INTEREST STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. The Chipping Forecast II. *Nature Genet.* **32** (Suppl.), 465–552 (2002).
2. Taussig, M.J. & Landegren, U. Progress in antibody arrays. *Targets* **2**, 169–176 (2003).
3. Ziauddin, J. & Sabatini, D.M. Microarrays of cells expressing defined cDNAs. *Nature* **411**, 107–110 (2001).
4. Garber, K. Gene expression tests foretell breast cancer's future. *Science* **303**, 1754–1755 (2004).
5. Holzman, T. & Kolker, E. Statistical analysis of global gene expression data: some practical considerations. *Curr. Opin. Biotechnol.* **15**, 52–57 (2004).
6. Wigle, D., Tsao, M. & Jurisica, I. Making sense of lung-cancer gene-expression profiles. *Genome Biol.* **5**, 309 (2004).
7. Andreasen, N.C., Arndt, S., Alliger, R., Miller, D. & Flaum, M. Symptoms of schizophrenia. Methods, meanings and mechanisms. *Arch. Gen. Psychiatry* **52**, 341–351 (1995).
8. Andreasen, N.C. A unitary model of schizophrenia. Bleuler's "fragmented Phrene" as schizencephaly. *Arch. Gen. Psychiatry* **56**, 781–787 (1999).
9. Sedvall, G. & Farde, L. Chemical brain anatomy in schizophrenia. *Lancet* **346**, 743–749 (1995).
10. Harrison, P.J. The neuropathology of schizophrenia. A critical review of the data and their interpretation. *Brain* **122**, 593–624 (1999).
11. McCarley, R.W. *et al.* MRI anatomy of schizophrenia. *Biol. Psychiatry* **45**, 1099–1119 (1999).
12. Andreasen, N.C. Schizophrenia: the fundamental questions. *Brain Res. Rev* **31**, 106–112 (2000).
13. Yolken, R.H., Karlsson, H., Yee, F., Johnston-Wilson, N.L. & Torrey, E.F. Endogenous retroviruses and schizophrenia. *Brain Res. Rev.* **31**, 193–199 (2000).
14. Rothermundt, M., Arolt, V. & Bayer, T.A. Review of immunological and immunopathological findings in schizophrenia. *Brain Behav. Immun.* **15**, 319–339 (2001).
15. Karlsson, H. *et al.* Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc. Natl. Acad. Sci. USA* **98**, 4634–4639 (2001).
16. Freedman, R. Schizophrenia. *N. Engl. J. Med.* **349**, 1738–1749 (2003).
17. Honey, G.D. *et al.* The functional neuroanatomy of schizophrenic subsyndromes. *Psychol. Med.* **33**, 1007–1018 (2003).
18. Hakak, Y. *et al.* Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proc. Natl. Acad. Sci. USA* **98**, 4746–4751 (2001).
19. Mirnics, K., Middleton, F.A., Marquez, A., Lewis, D.A. & Levitt, P. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* **28**, 53–67 (2000).
20. Vawter, M.P. *et al.* Application of cDNA microarrays to examine gene expression differences in schizophrenia. *Brain Res. Bull.* **55**, 641–650 (2001).
21. Mimmack, M.L. *et al.* Gene expression analysis in schizophrenia: reproducible upregulation of several members of the apolipoprotein L family located in a high-susceptibility locus for schizophrenia on chromosome 22. *Proc. Natl. Acad. Sci. USA* **99**, 4680–4685 (2002).
22. Middleton, F.A., Mirnics, K., Pierri, J.N., Lewis, D.A. & Levitt, P. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J. Neurosci.* **22**, 2718–2729 (2002).
23. Pongrac, J., Middleton, F.A., Lewis, D.A., Levitt, P. & Mirnics, K. Gene expression profiling with DNA microarrays: advancing our understanding of psychiatric disorders. *Neurochem. Res.* **27**, 1049–1063 (2002).
24. Guidotti, A. *et al.* Decrease in reelin and glutamic acid decarboxylase 67 (GAD67) expression in schizophrenia and bipolar disorder. *Arch. Gen. Psychiatry* **57**, 1061–1069 (2000).
25. Popken, G.J., Bunney, W.E., Potkin, S.G. & Jones, E.G. Subnucleus-specific loss of neurons in medial thalamus of schizophrenics. *Proc. Natl. Acad. Sci. USA* **97**, 9276–9280 (2000).
26. Thompson, P.M. *et al.* Mapping adolescent brain change reveals dynamic wave of accelerated grey matter loss in very early-onset schizophrenia. *Proc. Natl. Acad. Sci. USA* **98**, 11650–11655 (2001).
27. Rajkowska, G. & Goldman-Rakic, P.S. Cytoarchitectonic definition of prefrontal areas in the normal human cortex: II. Variability in locations of areas 9 and 46 and relationship to the Talairach coordinate system. *Cereb. Cortex* **5**, 323–337 (1995).
28. Browning, M.D., Dudek, E.M., Rapier, J.L., Leonard, S. & Freedman, R. Significant reductions in synapsin but not synaptophysin specific activity in the brains of some schizophrenics. *Biol. Psychiatry* **34**, 529–535 (1993).
29. McGinnis, R.E. *et al.* Failure to confirm *NOTCH4* association with schizophrenia in a large population-based sample from Scotland. *Nature Genet.* **28**, 128–129 (2001).
30. Volk, D.W., Austin, M.C., Pierri, J.N., Sampson, A.R. & Lewis, D.A. Decreased glutamic acid decarboxylase67 messenger RNA expression in a subset of prefrontal cortical γ-aminobutyric acid neurons in subjects with schizophrenia. *Arch. Gen. Psychiatry* **57**, 237–245 (2000).
31. Bernstein, H.-G.C.A. *et al.* Hippocampal expression of the calcium sensor protein visinin-like protein-1 in schizophrenia. *Neuroreport* **13**, 393–396 (2002).
32. Chudin, E. *et al.* Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.* **3**, research0005.1z–0005.10 (2001).
33. Churchill, G.A. Fundamentals of experimental design for cDNA microarrays. *Nature Genet.* **32** suppl. Suppl., 490–495 (2002).
34. Kothapalli, R., Yoder, S.J., Mane, S. & Loughran, T.P. Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22 (2002).
35. Kuo, W.P. *et al.* Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
36. Li, J., Pankratz, M. & Johnson, J.A. Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol. Sci.* **69**, 383–390 (2002).
37. Barczak, A. *et al.* Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.* **13**, 1775–1785 (2003).
38. Fan, J., Tam, P., Vande Woude, G. & Ren, Y. Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl. Acad. Sci. USA* **101**, 1135–1140 (2004).
39. Mills, J.C. & Gordon, J.I. A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Res.* **29**, E72–2 (2001).
40. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).
41. Huntsman, M.M., Tran, B.-V., Potkin, S.G., Bunney, W.E. & Jones, E.G. Altered ratios of alternatively spliced long and short γ2 subunit mRNAs of the γ-amino butyrate type A receptor in prefrontal cortex of schizophrenics. *Proc. Natl. Acad. Sci. USA* **95**, 15066–15071 (1998).
42. Tkachev, D. *et al.* Oligodendrocyte dysfunction in schizophrenia and bipolar disorder.

Lancet **362**, 798–805 (2003).
43. Kiiveri, H.T. A Bayesian approach to variable selection when the number of variables is very large. Institute of Mathematical Statistics, Lecture Notes, Monograph Series. **40**, 127–143 (2003).
44. Moler, E.J. *et al.* Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genomics* **4**, 109–126 (2000).
45. Staudt, L.M. & Brown, P.O. Genomic views of the immune system. *Annu. Rev. Immunol.* **18**, 829–859 (2000).
46. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
47. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
48. Lemon, W.J., Liyanarachchi, S. & You, M. A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biol.* **4**, R67 (2003).
49. Somorjai, R.L., Dolenko, B. & Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* **19**, 1484–1491 (2003).
50. Morris, J.S., Yin, G., Baggerly, K.A., Wu, C. & Zhang, L. Pooling information across different studies and oligonucleotide microarray chip types to identify prognostic genes for lung cancer. *Methods Microarray Anal.* in the press. **III**, 1–16 (2004).
51. Robb, L., Stephens, R. & Coleman, J. Application of survival and multivariate methods to gene expression data combined from two sources. *Methods Microarray Anal. III* (2004), in the press.
52. Jung, S.-H., Owzar, K. & George, S. Associating microarray data with a survival endpoint. *Methods Microarray Anal. III* (2004), in the press.
53. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790–13795 (2001).
54. Garber, M.E. *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA* **98**, 13784–13789 (2001).
55. Wigle, D.A. *et al.* Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.* **62**, 3005–3008 (2002).
56. Beer, D.G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.* **8**, 816–824 (2002).
57. Shedden, K. & Cooper, S. Analysis of cell cycle–specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl. Acad. Sci. USA* **99**, 4379–4384 (2002).
58. Gerdes, S.Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684 (2003).
59. Birrell, G.W. *et al.* Transcriptional response of *Saccharomyces cerevisiae* to DNA-damaging agents does not identify the genes that protect against these agents. *Proc. Natl. Acad. Sci. USA* **99**, 8778–8783 (2002).
60. Ooi, S.L., Shoemaker, D.D. & Boeke, J.D. DNA helicase gene interaction network define using synthetic lethality analyzed by microarray. *Nature Genet.* **35**, 277–286 (2003).
61. Jeong, H., Mason, S.P., Barabasi, A.-L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
62. Thatcher, J.W., Shaw, J.M. & Dickinson, W.J. Marginal contributions of nonessential genes in yeast. *Proc. Natl. Acad. Sci. USA* **95**, 253–257 (1998).
63. Miklos, G.L.G. & Rubin, G.M. The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**, 521–529 (1996).
64. Zhang, L. *et al.* Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272 (1997).
65. John, B. & Miklos, G.L.G. *The Eukaryote Genome in Development and Evolution* (Allen & Unwin, London, 1988).
66. Colucci-Guyon, E. *et al.* Mice lacking vimentin develop and reproduce without an obvious phenotype. *Cell* **79**, 679–694 (1994).
67. Erikson, H.P. A tenascin knockout with a phenotype. *Nature Genet.* **17**, 5–8 (1997).
68. Mills, M.A. *et al.* Differential expression of the actin-binding proteins, α-actinin-2 and -3 in different species: implications for the evolution of functional redundancy. *Hum. Mol. Genet.* **10**, 1335–1346 (2001).
69. Watkins, S. *et al.* Analbumenia: three cases resulting from different point mutations in the albumin gene. *Proc. Natl. Acad. Sci. USA* **91**, 9417–9421 (1994).
70. Stock, R.P. & Bialy, H. The sigmoidal curve of cancer. *Nature Biotechnol.* **21**, 13–14 (2003).
71. Kacser, H. & Burns, J.A. The molecular basis of dominance. *Genetics* **97**, 639–666 (1981).
72. Wuensche, A. Basins of attraction in network dynamics. in *Modularity in Development and Evolution* (eds. Schlosser, G. & Wagner, G.P.) 1–17 (Chicago University Press, Chicago, Illinois, 2004).
73. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
74. Arkin, A., Shen, P. & Ross, J. A test case of correlation metric construction of a reaction pathway from measurements. *Science* **277**, 1275–1279 (1997).
75. Dewey, T.G. From microarrays to networks: mining expression time series. *Drug Discov. Today* **7**, S170–S175 (2002).
76. Klevecz, R.R., Bolen, J., Forrest, G. & Murray, D.B. A genome-wide oscillation in transcription gates DNA replication and cell cycle. *Proc. Natl. Acad. Sci. USA* **101**, 1200–1205 (2004).
77. Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N. & Barabasi, A.-L. Global organization of metabolic fluses in the bacterium *Escherichia coli*. *Nature* **427**, 839–843 (2004).
78. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
79. Barabasi, A.-L. *Linked: The New Science of Networks* (Persus Publishing, Cambridge, Massachusetts, USA, 2002).
80. Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–418 (2004).
81. Miklos, G.L.G. & Maleszka, R. Integrating molecular medicine with functional proteomics: realities and expectations. *Proteomics* **1**, 30–41 (2001).
82. Miklos, G.L.G. & Maleszka, R. Protein functions and biological contexts. *Proteomics* **1**, 169–178 (2001).
83. Strohman, R. Maneuvering in the complex path from genotype to phenotype. *Science* **296**, 701–703 (2002).
84. Carney, S.L. Leroy Hood expounds the principles, practice and future of systems biology. *Drug Discov. Today* **8**, 436–438 (2003).
85. Palsson, B. O. *In silico* biotechnology. Era of reconstruction and interrogation. *Curr. Opin. Biotechnol.* **15**, 50–51 (2004).
86. Miklos, G.L.G. Molecules to cognition: the latter-day lessons of levels, language and *lac*. Evolutionary overview of brain structure and function in some vertebrates and invertebrates. *J. Neurobiol.* **24**, 842–890 (1993).
87. Famili, I., Forster, J., Nielsen, J. & Palsson, B.O. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. USA* **100**, 13134–13139 (2003).
88. Liao, J.C. *et al.* Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA* **100**, 15522–15527 (2003).
89. Davidson, E.H., McClay, D.R. & Hood, L. Regulatory gene networks and the properties of the developmental process. *Proc. Natl. Acad. Sci. USA* **100**, 1475–1480 (2003).
90. Herrgard, M.J., Covert, M.W. & Palsson, B.O. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.* **13**, 2423–2434 (2003).
91. Li, J.Z. *et al.* Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Hum. Mol. Genet.* **13**, 609–616 (2004).
92. Reese, M.G. *et al.* Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10**, 483–501 (2000).
93. Jones, L., Ng, S.-K., Ambroise, C. & McLachlan, G. Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. *Methods Microarray Anal. III* (2004), in the press.
94. Matsuzaki, H. *et al.* Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14**, 414–425 (2004).