DATE Analysis: A General Theory of Biological Change Applied to Microarray Data

David Rasnick
Chromosome Diagnostics, LLC, 883 Wood Street, Oakland, CA 94607

DOI 10.1002/btpr.239
Published online August 14, 2009 in Wiley InterScience (www.interscience.wiley.com).

In contrast to conventional data mining, which searches for specific subsets of genes (extensive variables) to correlate with specific phenotypes, DATE analysis correlates intensive state variables calculated from the same datasets. At the heart of DATE analysis are two biological equations of state not dependent on genetic pathways. This result distinguishes DATE analysis from other bioinformatics approaches. The dimensionless state variable F *quantifies the relative overall cellular activity of test cells compared to well-chosen reference cells. The variable* π_i *is the fold-change in the expression of the* ith *gene of test cells relative to reference. It is the fraction* ϕ *of the genome undergoing differential expression—not the magnitude* π*—that controls biological change. The state variable* ϕ *is equivalent to the control strength of metabolic control analysis. For tractability, DATE analysis assumes a linear system of enzyme-connected networks and exploits the small average contribution of each cellular component. This approach was validated by reproducible values of the state variables* F*, RNA index, and* ϕ *calculated from random subsets of transcript microarray data. Using published microarray data,* F*, RNA index, and* ϕ *were correlated with: (1) the blood-feeding cycle of the malaria parasite, (2) embryonic development of the fruit fly, (3) temperature adaptation of Killifish, (4) exponential growth of cultured* S. pneumoniae*, and (5) human cancers. DATE analysis was applied to aCGH data from the great apes. A good example of the power of DATE analysis is its application to genomically unstable cancers, which have been refractory to data mining strategies.* © 2009 American Institute of Chemical Engineers *Biotechnol. Prog.,* 25: 1275–1288, 2009
Keywords: metabolic, bioinformatics, state equations, control-analysis, data mining

Introduction

Analyzing large gene expression datasets is a relatively new area of data analysis with its own unique challenges.[1,2] Various statistical methods are used to sift through tens of thousands of data points searching for stable subsets of genes that are correlated with specific normal and abnormal phenotypes. The supervised and unsupervised statistical algorithms produce annotated lists of genes according to differences in expression.[3] The lists of genes are then assembled into genetic roadmaps that are thought to govern the specific phenotypes being investigated. This strategy comes from the general belief that a relatively small number of specific genes control certain normal and disease phenotypes. In spite of the promising initial results, this approach has not lived up to expectations, particularly with respect to cancer.[1,4–13]

The results to date indicate that the genetic roadmaps are not providing the rules governing the dynamic interplay between genotype and phenotype.[14] Knockout experiments, for example, have repeatedly shown that the whole animal, down to the cellular phenotype, is usually unaffected by the loss of one or a few genes[15,16] and when there are phenotypic consequences they are unpredictable.[17] Adding to the problem of associating subsets of genotype with phenotype, the "genetic signatures" generated from microarray experiments are highly unstable,[13,18] particularly for genomically unstable cancers.[6,10–12,19] A partial explanation for the instability was recently offered by Shi et al: "reproducibility has seldom been, but in the future should be, used as a crucial criterion to judge the validity of data analysis procedures."[13] According to Elser and Hamilton, "It seems that the only hope for creatively interrogating new data is to develop new, integrated theoretical frameworks to inform strategies for that interrogation."[20]

Metabolic control analysis (MCA) is a well-established integrated theoretical foundation upon which to construct a general theory of biological change.[21] Importantly, MCA explains why it is unlikely a unique, small subset of genes controls a specific macroscopic phenotype.[22] As Henrik Kacser observed, "But one thing is certain: to understand the whole you must look at the whole."[23] MCA provides the theoretical framework for studying the phenotype as a whole.

The past 30 years of applying MCA to glycolysis, the tricarboxylic acid cycle, photosynthesis, and the syntheses of fatty acids, urea, nucleotides, and amino acids has conclusively shown that even these relatively uncomplicated systems (where all components can be memorized) are rarely controlled by slow or rate-determining steps.[24–27] One of the

Correspondence concerning this article should be addressed to D. Rasnick at drasnick@mac.com.

fundamental discoveries of MCA is that even with a complete knowledge of the detailed properties of specific genes and gene products it is not possible to either predict or describe phenotypes at the cellular level and above in terms of a few individual genetic components. Yet, the field of bioinformatics is influenced by the belief that specific "genetic pathways" and "genetic programs" control or determine complex phenotypes.[28] An important example is the pursuit of oncogenes and tumor suppressor genes in cancer research,[29–32] which can be summarized as the search for the presumed rate-determining molecular steps in carcinogenesis. However, instead of rate limiting steps, experiments have repeatedly demonstrated control is distributed among the host of metabolic, catalytic, and regulatory components.[22,33–35] These cellular components are Rosen's particles of function.[36] While the fundamental principles of MCA are valid for all levels of cellular activity and phenotypic change, its practical application is limited to relatively small experimental systems of usually not more than 50 components. The methods and techniques of MCA are simply overwhelmed by the thousands of variables contributing to macroscopic phenotypes.

It became apparent in 2001 that the modification of metabolic control analysis used to investigate aneuploidy in cancer cells also applied to diploid phenotypes. This realization led to a general method of studying phenotypic change called DATE analysis, which comes from differentiation, adaptation, transformation, evolution. DATE analysis differs from MCA in that its essential task lies in the comparison of related phenotypes rather than in the precise definition or description of each. In place of tracking the kinetic particulars of thousands of individual cellular components, DATE analysis uses, instead, two biological equations of state to calculate their aggregate effects.[34,37,38] This approach makes it possible to analyze the phenotypic changes of whole cells, organs, and organisms.

The numerous cellular phenotypes of diploid and aneuploid species (cancer)[30] are due to the differential expression of fractions ϕ of the stable and unstable genomes, respectively. DATE analysis demonstrates the principle that it is the fraction ϕ of the genome undergoing differential expression, not the magnitude π of the differential expression, which controls phenotypic change.[38] Systems at least as complex as a cell are determined by tens of thousands of genes, gene products, and metabolites, each making a small contribution on the order of 10^{-5} to the macroscopic phenotype.[24,26,27,34,39,40] The small contribution of individual genes suggests their activities can be quantitatively treated as being equally important.[34,38,41] For tractability, DATE analysis assumes a linear system of enzyme-connected networks and exploits the small average contribution of each cellular component.[34,38] Nevertheless, the results apply equally to systems of interlocking pathways, cycles, feedback loops,[24,34] regulatory cascades,[42] and control of gene expression,[43] except that the formulations become more tedious.[26]

The first use of DATE analysis provided powerful theoretical support[38,44] for Theodor Boveri's hundred-year-old hypothesis[45] that the progression of aneuploidy is carcinogenesis. As originally formulated, Eqs. 1a and 1b describe the relationship between the state variables F, RNA index, and ϕ. Equation 1a gives the relative overall cellular (metabolic) activity F of a test sample compared to appropriate reference cells. For aneuploid phenotypes, Eq. 1a gives the relative value of F for aneuploid cells compared to their diploid counterparts. The state variable ϕ of aneuploid cells is the fraction of the genome that is aneuploid relative to normal diploid cells. The variable π is the average fold-change of the differential expression of a population of aneuploid cells. Equation 1b gives the state variable RNA index for the aneuploid phenotype compared to diploid precursors (RNA index is the total number of transcripts of aneuploid cells divided by the total number of transcripts of diploid cells from the same tissue type). The changes in RNA index are largely due to changes in the DNA content of aneuploid cells.[38,46–49] Since the state variables are relative values they are dimensionless.

$$\frac{1}{F_{\text{relative}}} = 1 - \phi + \frac{\phi}{\pi} \tag{1a}$$

$$\text{RNA}_{\text{index}} = 1 - \phi + \phi\pi \tag{1b}$$

DATE analysis has also been used to explain the Hayflick limit of cultured cells, the time-course of carcinogen-induced tumors in mice, the age distribution of human cancer, multidrug resistance, the lack of immune surveillance protecting against cancer, and the failure of cancer chemotherapy.[44,50]

Since the approach described here is so heavily dependent on the assumptions, formalisms, and general principles of metabolic control analysis, it is strongly recommended the reader become familiar with the 1981 paper by Kacser and Burns on "The Molecular Basis of Dominance"[34] in order to better understand the basis of DATE analysis.

Here is reported the application of DATE analysis to published transcript microarray data from: (1) the blood-feeding cycle of the malaria parasite, (2) embryonic development of the fruit fly, (3) temperature adaptation of Killifish, (4) exponential growth of cultured *Streptococcus pneumoniae*, and (5) human cancers. DATE analysis was applied to aCGH data from the great apes. The state variables F, RNA index, and ϕ were determined from microarray data from lymphoma and cancers of the breast, colon, kidney, ovary, pancreas, and stomach. The distribution entropy D, which is analogous to Shannon Entropy, is introduced as a measure of the entropy of microarray results presented in the form of a histogram. Acute environmental change and stress cause cells to increase the expression of some genes and simultaneously decrease expression of others in order to keep total levels of RNA and protein constant.[51–54] γ is introduced as a measure of this compensating differential expression of transcripts. γ also quantifies the genomic difference between test and reference aCGH data. When applied to cancer, D and γ quantify the genomic imbalance leading to the genetic instability of aneuploid cells.

Methods

Equations 2a and 2b are the general forms of the original state Eqs. 1a and 1b, first used to study aneuploid phenotypes. Importantly, Eqs. 2a and 2b apply to the phenotypic changes of aneuploid and diploid cells alike. Thus, aneuploidy is just a special case of DATE analysis.

$$\frac{1}{F_{\text{relative}}} = 1 - \sum \phi_i + \sum \frac{\phi_i}{\pi_i} \tag{2a}$$

$$\text{RNA}_{\text{index}} = 1 - \sum \phi_i + \sum \phi_i \pi_i \tag{2b}$$

The variable π_i is the fold-change of the ith transcript relative to an appropriate reference sample. The variable ϕ_i is

not restricted, as first described in 1999,[38] to gene expression of aneuploid cells, where $\pi_i \neq 1$, but is the fractional contribution each gene makes for any value of π_i, irrespective of ploidy. In agreement with its original definition,[38] ϕ without subscript is the state variable representing the fraction of genes with π values outside the normal range. For this report the normal range of π_i was empirically determined from comparisons of normal tissues of the same type and is defined as $0.5 \leq \pi_{\text{normal}} \leq 1.5$ because over 95% of transcripts of normal tissues were in this interval.

Since microarray experiments keep track of all genes individually, and in keeping with the principle that individual genes contributing to macroscopic phenotypes can be treated in the aggregate as being quantitatively equivalent,[34,38,41] Eqs. 2a and 2b can be expressed in the convenient form of Eqs. 3a and 3b, where n = total number of expressing genes and $\phi_i = 1/n$, giving $\Sigma\phi_i = 1$.

$$\frac{1}{F_{\text{relative}}} = \frac{1}{n}\sum\frac{1}{\pi_i} = \text{ave}\left(\frac{1}{\pi_i}\right) \tag{3a}$$

$$\text{RNA}_{\text{index}} = \frac{1}{n}\sum\pi_i = \text{ave}(\pi_i) \tag{3b}$$

Equation 3a shows that $1/F_{\text{relative}}$ is just the average of the $1/\pi_i$ values, giving F as the harmonic mean of the π_i values. Likewise, the average of π_i values gives the relative RNA index (Eq. 3b). In general, RNA index is the total number of transcripts of test cells divided by the total number of transcripts of reference cells from the same tissue type. When applied to aCGH data, π_i is the copy-number fold-change of the ith gene relative to an appropriate reference genome. Thus, π_i and $1/\pi_i$ are the central computational elements of DATE analysis for calculating the state variables F, RNA index, and ϕ from microarray data.

The general method of calculating the π_i values was as follows. Published expression and aCGH data used in this report were either downloaded from public sources, e.g., Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo and Stanford Microarray Database http://genome-www5.stanford.edu, or other specialty databases, e.g., for malaria http://plasmodb.org/plasmo. The rule was to analyze only those transcripts or genes for which un-flagged intensity values existed for both the test and reference samples. The best data were normalized fluorescence intensity, which is proportional to concentration. If the downloaded data were log transformed, then the antilog transformation was performed before analysis. The π_i values were then easily calculated by dividing the microarray value for each transcript (i) or gene (i) of the test sample by that for the reference sample. Then the inverse values were calculated. These values of π_i and $1/\pi_i$ were used to calculate F_{relative} and RNA index according to Eq. 3. The value of ϕ without subscript was calculated by dividing the number of transcripts outside the normal range of π_i values ($0.5 \leq \pi_{\text{normal}} \leq 1.5$) by the total number of transcripts.

Compensating differential expression

A defining property of diploid cells (and other cells with balanced genomes) is the relative amounts of DNA, total RNA, and total cellular protein remain constant.[51–54] For stable euploid phenotypes with RNA index = 1, Eq. 3b reduces to Eq. 4.

$$\sum\pi_i = n \tag{4}$$

Therefore, for $\pi_i \neq 1$, Eq. 4 demands values of π_i above and below one to maintain a constant RNA index = 1, which characterizes stable phenotypes. In other words, increases in the expression of different genes must be compensated by decreases in others in order to keep total levels of RNA and protein constant. It is this compensatory differential expression that regulates homeostasis when adapting to stress and acute changes in environment.

The well-known intercellular heterogeneity of a population of aneuploid cells is produced by random changes in the chromosome (DNA) content of individual aneuploid cells with each mitotic division.[38,44,55,56] Nevertheless, following mitosis the tight correlations—between the relative amounts of DNA and RNA, RNA, and total cellular protein—still hold for each individual aneuploid cell in a heterogeneous population of aneuploid cells.[46,57–61]

Lion and Gabriel have shown that F (the overall cellular activity of a cell) has a maximum value when total cellular enzyme concentration is constant,[62] i.e., constant RNA index. Dividing Eq. 3b, where RNA index = 1, by Eq. 3a gives the bounded Eq. 5, explained below.

$$0 < F = \frac{\sum\pi_i}{\sum\frac{1}{\pi_i}} \leq 1 \tag{5}$$

The ratio of the two sums in Eq. 5 is always less than or equal to one.[63] Thus, the relative overall cellular activity F of phenotypically stable cells with balanced genomes varies between zero and one during periods of compensating differential expression.

The state variable F may be greater than one for aneuploid cells.[38] Nevertheless, the ratio of the two sums in Eq. 5 is still less than one[63] for aneuploid cells. Thus, the two state variables F and RNA index are always out of balance for aneuploid cells, indicating reduced metabolic efficiency and viability relative to the diploid (balanced) state.[46,64]

Distribution entropy **D**

Shannon entropy[65] has been used to quantify the incredibly variable and complex karyotypes of cancer cells.[66] The same approach can be applied to the differential expression of microarray data. The distribution entropy D is given by Eq. 6.

$$D = -\sum_{i=1}^{n} p_i \ln p_i \tag{6}$$

D measures the distribution (entropy) of histogram data. The variable p_i ($0 \leq p \leq 1$) is the fraction of data placed in bin (i) of a histogram with n evenly spaced intervals of π. Figure 1 shows D is a function of bin size. Thus, the absolute value of D changes with bin size, nevertheless, the rank order of values of D for microarray experiments will not change if the bin size is held constant.

Even when the total RNA content of a cell or organism is constant and changes in F are barely perceptible, there can still be large changes in D. In such cases, large values and changes in D may indicate considerable stress leading to compensating differential expression needed to maintain homeostasis. Another interpretation is that large D indicates the

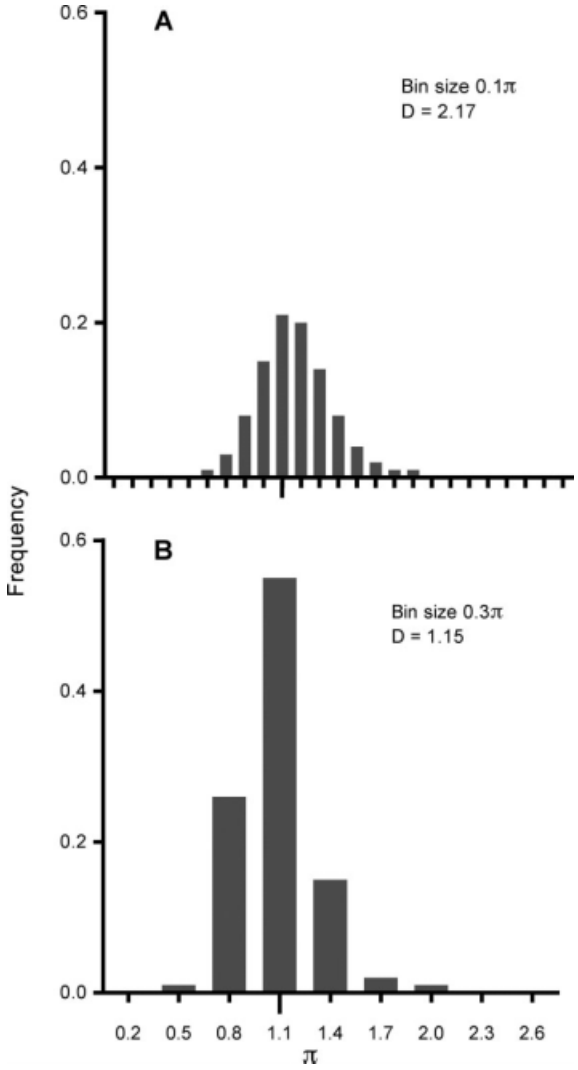

Figure 1. Distribution entropy *D* of histogram data is a function of bin size.

The results from 19,785 transcripts of normal skin[67] were processed according to Eq. 6 to give: (A) $D = 2.17$ at bin width 0.1π (π_i is the fold-change in the expression of the ith gene of test cells relative to reference cells). (B) $D = 1.15$ at bin width 0.3π. Thus, the absolute value of D changes with bin size, nevertheless, the rank order of values of D for microarray experiments will not change if the bin size is held constant. D is a measure of the entropy or spread of histogram data and a measure of genetic instability when applied to cancer cells.

generation of a metabolically heterogeneous population of cells, e.g., single-cell organisms and cancer. When applied to cancer, D is an objective measure of the genomic imbalance causing genetic instability and increased tumor malignancy.[38,68–71]

Results

Using the Kacser and Burns assumption that individual genes contributing to macroscopic phenotypes can be treated in the aggregate as being quantitatively equivalent,[34,38,41] it follows that for large datasets the state variables would be intensive variables because they would not depend on specific sets of genes. Thus, random subsets of transcript microarray data should reflect the whole and provide good estimates of the state variables F, RNA index, and ϕ.

Figure 2 shows that random subsets of transcript microarray data did indeed provide values that were very close to those determined using all the transcripts. Figure 2A shows that random subsets as small as 10% of 13,984 total transcripts measured for a ductal breast carcinoma[72] yielded reproducible (i.e., stable) values of RNA index, overall cellular activity F, and the aneuploid fraction ϕ, relative to normal breast. For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$ (as explained above). Error bars represent one standard deviation for three independent randomized series of the same transcript data. Likewise, Figure 2C shows random subsets down to 15–20% of 6 143 total transcripts produced acceptable estimates of F, RNA index, and ϕ for the 8 h point of fruit fly development relative to 0 h.[73] Figures 2B,D shows that plots of the cumulative sums of π_i and $1/\pi_i$ for random subsets of transcripts were linear. The solid and broken lines in Figures 2B,D are the least squares results for the cumulative sums of π_i and $1/\pi_i$, respectively. The linear results hold for random subsets from all microarray data examined in this report.

The results of Figure 2 demonstrate for large datasets that F, RNA index, and ϕ are indeed intensive state variables because they are not dependent on a perspicacious choice of specific genes. This result sets DATE analysis apart from conventional data mining approaches which seek to correlate specific sets of genes (i.e., extensive variables) with certain phenotypes.

Information added by using ordered values of π_i

As discussed in the introduction, the tens of thousands of interconnected cellular components made it necessary to assume for tractability each makes an equivalent small contribution to the macroscopic phenotype. Nevertheless, the expression of individual genes can vary considerably. The information lost by using random subsets of microarray data when calculating the intensive state variables F, RNA index, and ϕ (Figure 2) can be recovered by analyzing subsets of the same data ordered by increasing values of π_i (which implies decreasing values of $1/\pi_i$). α and β are introduced as measures of the information gained when comparing ordered vs. random values of $1/\pi_i$ and π_i, respectively.

While cumulative sums of random subsets were linear (Figures 2B,D), cumulative sums of subsets ordered by increasing values of π_i were not (Figure 3). The curved lines in Figure 3A are for the ordered data from the breast cancer patient of Figure 2 relative to normal breast. Similarly, Figure 3C is for the ordered data from fruit fly embryo at 8 h compared to 0 h. The curved lines for fruit fly embryo were due to compensating differential expression. The curved lines for the breast cancer were caused by a heterogeneous population of aneuploid cells. The area enclosed by the cumulative sums of ordered (curved lines) and random subsets (straight lines) of the same data are measures of the magnitude of breast cancer aneuploidy (Figure 3A) and the overall compensatory differential expression of fruit fly embryo (Figure 3C), respectively. The area enclosed by the broken lines is designated α (enclosed area of cumulative sums of $1/\pi_i$, ordered minus random). The area enclosed by the solid lines is designated β (enclosed area of cumulative sums of π_i, random minus ordered). Total area γ is the sum of α and β. For comparison, the enclosed areas for normal breast (Figure 3B) were much smaller as one would expect for a stable phenotype.

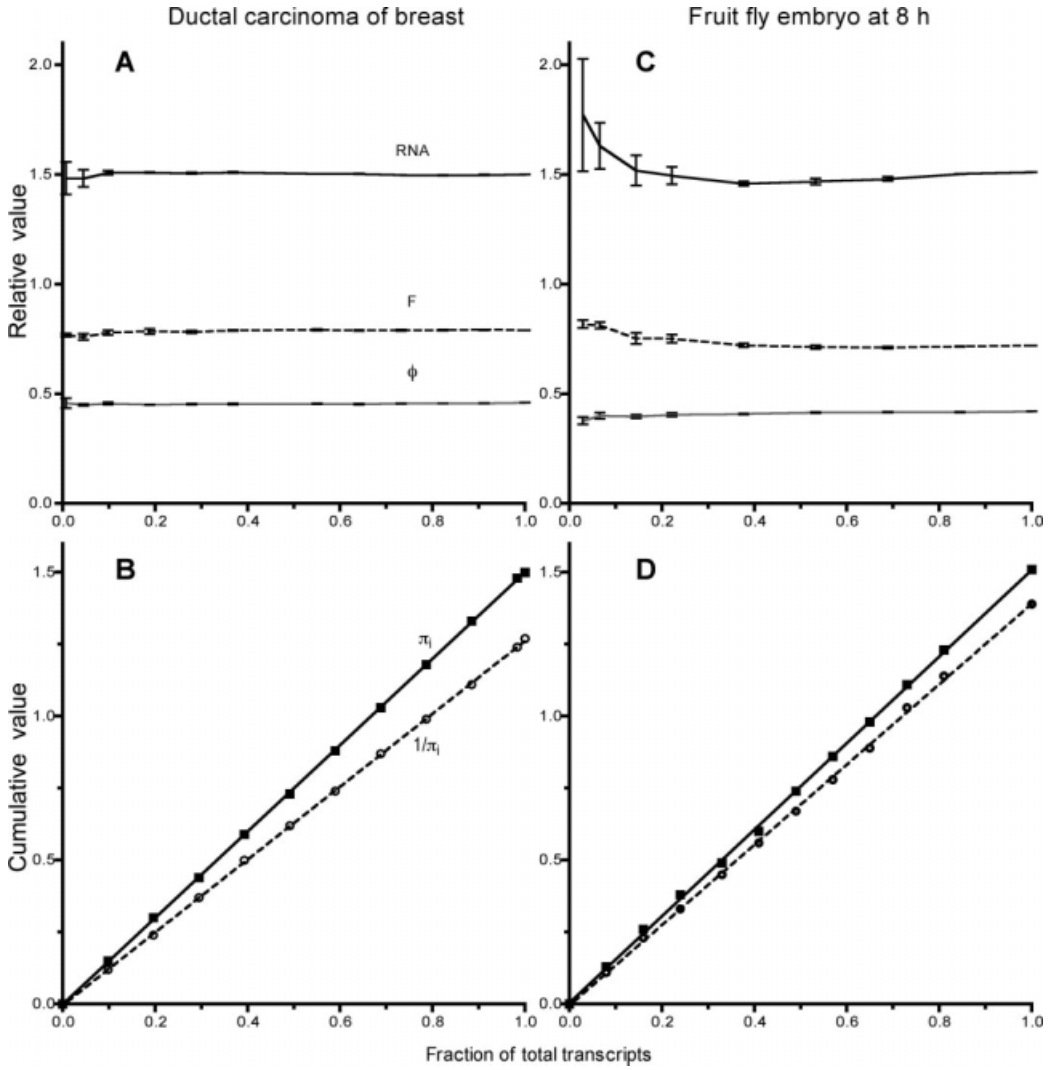

Figure 2. Random subsets of transcript microarray data reflect the whole.

The state variables are intensive variables because they are not dependent on specific sets of genes. This result is of fundamental importance and distinguishes DATE analysis from all other methods of bioinformatics. A: Random subsets down to 10% of 13,984 total transcripts from aneuploid ductal breast carcinoma[72] yielded reproducible values of the state variables F, RNA index, and ϕ (F is the dimensionless state variable quantifying the relative overall cellular activity of test cells compared to reference cells, RNA index is the total number of transcripts of test cells divided by the total number of transcripts of reference cells, ϕ is the fraction of the genome undergoing differential expression). For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$ based on comparisons of normal tissues of the same type (π_i is the fold-change in the expression of the ith gene of test cells relative to reference cells). Determination of D was based on a bin size of 0.1π (D is the distribution entropy of histogram data). B: Plots of the cumulative sums of π_i and $1/\pi_i$ for random subsets from the same transcript data were linear. C: Random subsets down to 15–20% of 6,143 total transcripts measured for the 8 h point relative to 0 h of diploid fruit fly development[73] produced acceptable estimates of F, RNA index, and ϕ. D: Plots of the cumulative sums of π_i and $1/\pi_i$ for random subsets of transcripts from the same transcript data were linear. The solid and broken lines (B, D) are the least squares results for the cumulative sums of π_i and $1/\pi_i$, respectively. Error bars represent one standard deviation for three independent random subsets.

Empirically it was found α and β often provide more information than D alone. For example, Figure 3D shows that a low level of compensating differential expression β produced a large perturbation α in the overall metabolic activity of diploid Killifish adapting to the 36 h point of the temperature cycle[74] compared to 0 h.

Differentiation and development

Example 1: State Variables Correlated with the Cyclical Differentiation of Plasmodium falciparum and Noncyclical Development of Drosophila melanogaster. Time-course studies are particularly amenable to DATE analysis because the results at any time point can represent a suitable reference for other time points. The blood-feeding cycle of the malaria parasite *P. falciparum* is 48 h.[75] The differentiation cycle starts when the merozoite stage of the parasite invades red blood cells to form the ring stage. The differentiation cycle was mirrored by the DATE results shown in Figure 4A, using merozoite as reference. Since the differentiation process produces very different phenotypic stages, there were, not surprisingly, dramatic changes in all of the state variables that reverted to the merozoite values at the end of the cycle.

Between 17 and 29 h the parasite was in the maturation phase, called the trophozoite stage, and experienced the largest fraction of differential expression ($\phi > 0.7$) relative to the merozoite stage (Figure 4A). For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$. It is during the trophozoite stage the parasite digests most of the hemoglobin. In the schizont stage, at the end of feeding, the parasite prepares for reinvasion of new red blood cells by replicating

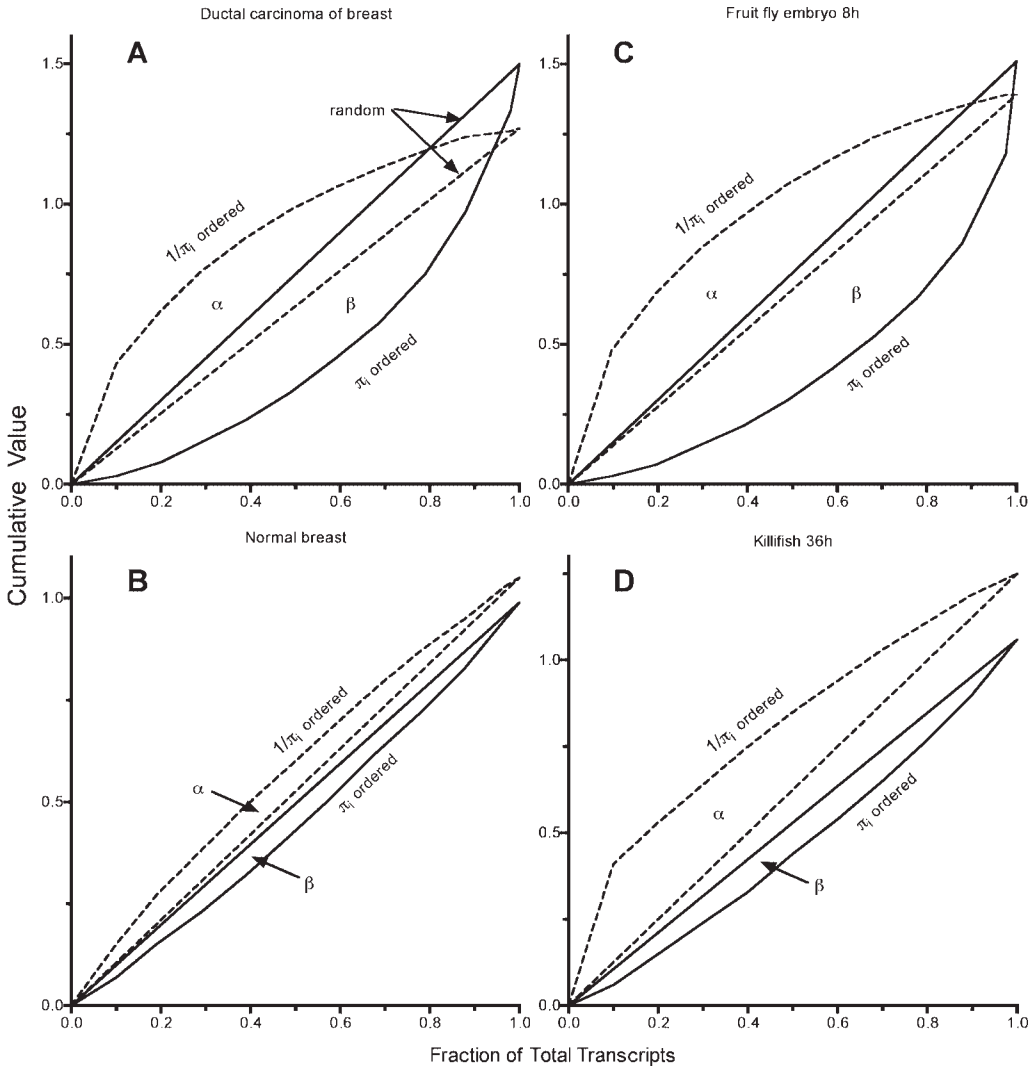

Figure 3. Information added by using ordered values of π_i and $1/\pi_i$.

Averaging random values of π_i and $1/\pi_i$ made it possible to calculate the state variables F, RNA index, and ϕ but at the expense of losing information present in the differential expression of individual genes (π_i is the fold-change in the expression of the ith gene of test cells relative to reference, F is the dimensionless state variable quantifying the relative overall cellular activity of test cells compared to reference cells, RNA index is the total number of transcripts of test cells divided by total number of transcripts of reference cells, ϕ is the fraction of the genome undergoing differential expression). However, this very useful information can be recovered by analyzing ordered values of $1/\pi_i$ (decreasing) and π_i (increasing), respectively. The curved broken lines were cumulative sums from ordered subsets of $1/\pi_i$. The curved solid lines were cumulative sums from ordered subsets of π_i. The straight broken lines were cumulative sums from random subsets of $1/\pi_i$. The straight solid lines were cumulative sums from random subsets of π_i. Enclosed areas α and β (see text) are measures of the information gained when comparing ordered vs. random values of $1/\pi_i$ and π_i. A: The same breast cancer patient as in Figure 2A. B: Normal breast. C: Fruit fly 8 h embryo. D: The areas α and β need not be symmetrical. The relatively small spread in transcripts represented by the area β produced a much larger effect α for the diploid Killifish adapting to the 36 h point of a 24 h temperature cycle.[74]

and dividing to form up to 32 new merozoites. Then the process repeats itself.

For sufficiently small incremental time steps, it is likely that the $n+1$ state is very similar metabolically and phenotypically to state n (i.e., $F_{n+1} \approx F_n$). Changes in RNA index are likewise expected to be relatively small, approximating the conditions of Eq. 5. Therefore, if the reference state is allowed to vary during differentiation and development, such that the reference state is the preceding one, i.e., state $n+1$ vs. state n, then F and RNA index are expected to be near one. Under this scenario significant transient changes were more readily detected. For example, when adjacent time points of the blood-feeding cycle of the malaria parasite were compared, RNA index and F oscillated about an average value close to one (Figure 4B; F not shown for clarity) and the average value of ϕ was 0.08 ± 0.05. Fourier analysis identified a cycle with a period of 2.3 h (sampling rate was 1 per h) for RNA index and F during the malaria parasite ring stage (Figure 4B; F not shown for clarity).

In contrast with malaria parasite differentiation, development of the fruit fly embryo was, of course, not cyclical (Figure 4C). A comparison of adjacent time points of fruit fly embryonic development demonstrated well-regulated, noncycling values of F and RNA index (Figure 4D), in agreement with Eq. 5, indicating a smooth developmental course.

Adaptation

Example 2: Killifish Adapting to Temperature Changes. Killifish live in small, isolated ponds in coastal desert and savanna regions of northern South America. They routinely experience wide daily fluctuations in temperature, oxygen

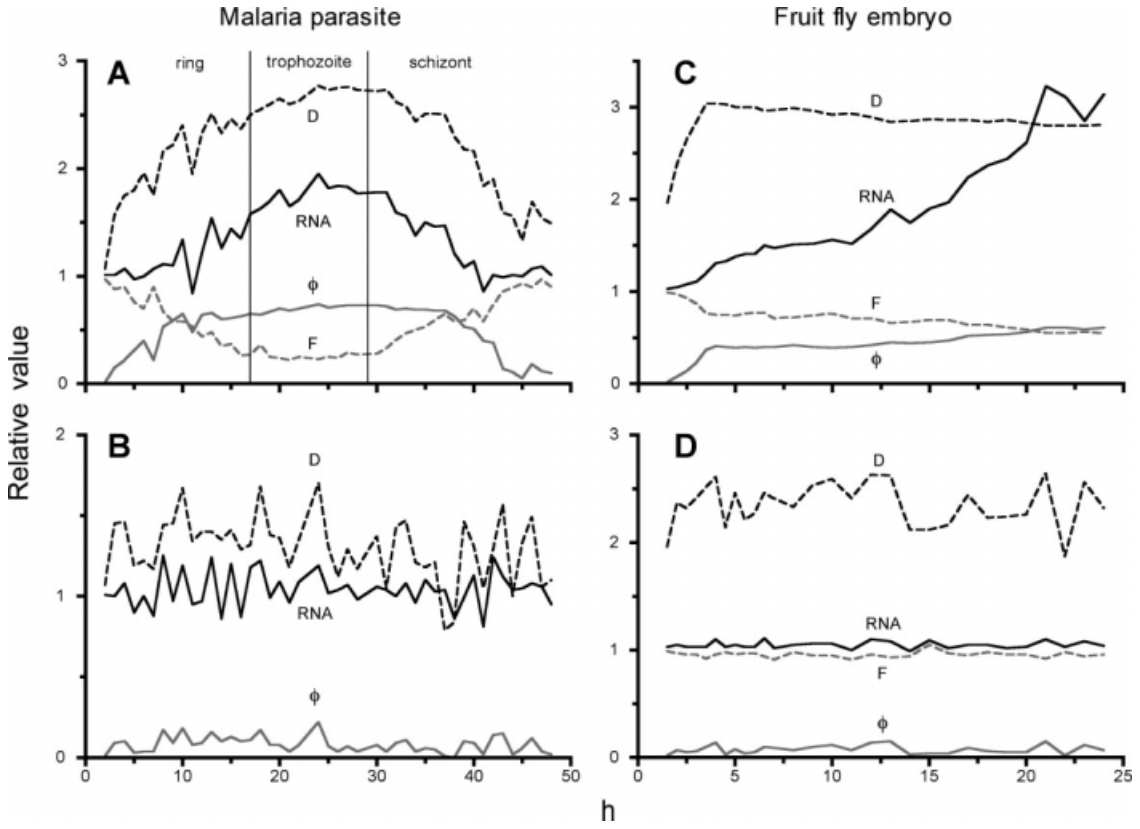

Figure 4. Differentiation of *P. falciparum* and embryonic development of *D. melanogaster*.

The state variable ϕ measures the fraction of the genome that is differentially expressing transcripts and F is the overall cellular activity relative to the beginning merozoite stage. For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$ based on comparisons of normal tissues of the same type (π_i is the fold-change in the expression of the ith gene of test cells relative to reference). A: The differentiation cycle of the malaria parasite starts when the merozoite stage invades red blood cells to form the ring stage. Over 70% of the genome ($\phi > 0.7$) was differentially expressed during the trophozoite stage compared to the merozoite stage. Determination of D was based on a bin size of 0.3π (D is the distribution entropy of histogram data). B: When adjacent time points were used for comparison, F and the RNA index oscillated about an average value near one (RNA index is the total number of transcripts of test cells divided by total number of transcripts of reference cells). Fourier transform analysis identified a cycle with a period of 2.3 h (sampling rate was 1 per h) for RNA index and F (not shown) during the malaria parasite ring stage. C: In contrast with the malaria parasite differentiation, the developmental course of the fruit fly embryo as expected was not cyclical. Determination of D was based on a bin size of 0.1π. D: A comparison of adjacent time points of fruit fly embryonic development showed well-regulated, noncycling values of F and RNA index near one. The time-course trajectories of the state variables may signify a biological principle analogous to the least action principle in physics.

availability, and pH. Temperatures may change over 20°C on a daily basis and may reach a high above 40°C.[74]

Figure 5A shows at constant temperature, F and RNA index of Killifish were stable at normal levels near one (4 h was reference). Interestingly, Figure 5A shows a dampened oscillation of the distribution entropy D, indicating the initial stress experienced by the Killifish at the start of the control experiment subsided over time. This example shows that the compensating differential expression of a stable phenotype could lead to unstable genetic signatures derived from conventional data mining strategies due to the dramatic changes in the expression of individual genes.

The Killifish temperature cycle experiment (Figure 5B) demonstrates the principle discussed earlier that even when global expression levels are constant over time, where average π_i is very close to one (i.e., RNA index = 1), an organism can experience a significant reduction in relative overall cellular activity (i.e., $F < 1$) when stressed. During the first 48 h of the experiment, Figure 5B clearly shows a 20 h cycle with F varying between 1 and 0.8 (broken gray line) as the temperature was cycled from 20 to 37°C every 24 h (0 h was reference). In contrast to the cycling of F, overall transcript levels (solid black line) did not change in a global manner, which indicates very tight regulation of steady state levels of mRNA during substantial changes in temperature.

The 20 h cycle of F observed during the first 48 h may have eventually synchronized to the 24 h temperature cycle. However, since the continuous 24 h cycle ended at 68 h, it was not possible to test this hypothesis using published data.

Example 3: Streptococcus pneumoniae Adapting During Exponential Growth. Figure 6 shows the time-course of *Streptococcus pneumoniae* adapting to experimentally induced exponential growth and to its abrupt end at 6 h (0 h was reference).[76] Figure 6A shows from 0 to 6 h there was little change in RNA index (solid black line) and a continuous, gradual decline in cellular activity F (broken gray line). However, there was a dramatic rise in D (Figure 6A) from the very beginning that did not dissipate as seen with the Killifish control (Figure 5A). The large increase in D during exponential growth indicates a growing metabolically heterogeneous population of bacteria during cell culture. This was reflected in the growing spread in the approximately symmetrical distribution of transcripts up to the end of log phase growth at 6 h (Figure 6B). After 6 h, the spread continued to increase but had become skewed with the center of mass shifted to the left, signifying a major reduction in the cellular activity of cultured *S. pneumoniae*, which was also reflected

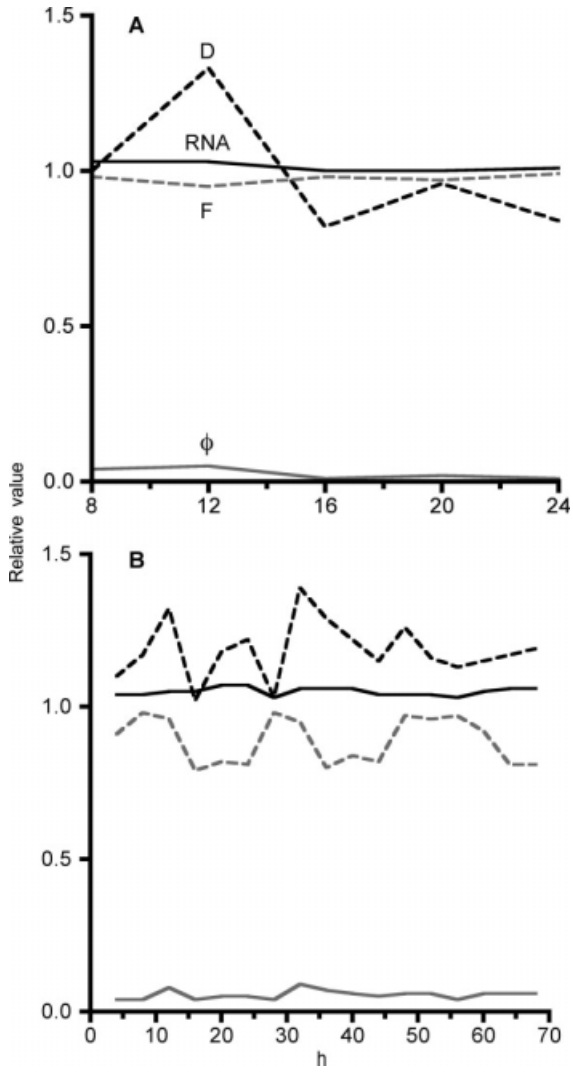

Figure 5. Killifish temperature cycling experiment.

The Killifish temperature cycle experiment demonstrated the principle that even when global expression levels are constant over time, where average π_i is very close to one (i.e., RNA index = 1: RNA index is the total number of transcripts of test cells divided by total number of transcripts of reference cells, π_i is the fold-change in the expression of the ith gene of test cells relative to reference), an organism can experience a significant reduction in relative overall cellular activity (i.e., $F < 1$) when stressed. A: Constant temperature control, 4 h was the reference. B: Cycling of the state variable F (broken gray line) as the temperature was cycled from 20 to 37°C every 24 h, 0 h was reference. Notice that ϕ (solid gray line) was generally well-below 10% (ϕ is the fraction of the genome undergoing differential expression). For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$ based on comparisons of normal tissues of the same type. Determination of D was based on a bin size of 0.1π (D is the distribution entropy of histogram data). The 20 h cycle of F observed during the first 48 h may have eventually synchronized to the 24 h temperature cycle. Since the data collected at 4 h intervals ended at 68 h, there was no way to test this hypothesis using published data.

in the sharp drop in F (Figure 6A). The large decline in F coupled with the pronounced rise in relative RNA index and ϕ indicated a massive phenotypic change in the bacteria commencing with the end of log phase growth. For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$.

Transformation (cancer)

Since Hansemann first observed chromosomal abnormalities over a hundred years ago in all of the epithelial cancers he investigated,[77] an overwhelming body of evidence has established an inseparable connection between cancer and aneuploidy.[78–81] By 1969, Albert Levan was confident enough to say, "there is safe evidence that carcinogenesis, as well as all stages of malignancy, is accompanied by chromosomal irregularities…."[82] But he went on to add that, "nothing is known, however, as to the significance of these chromosome irregularities in relation to the carcinogenic transformation." In other words, he raised the perennial question: is chromosomal imbalance (aneuploidy) a cause or consequence of cancer?

While leaving the question open, Levan acknowledged that aneuploidy satisfies at least one requirement of a cause: "Chromosome variation is an integrated part of tumor development from the earliest beginning of carcinogenesis to the latest progressive stages. Even before any malignancy has started chromosome variation in a normal tissue is generally associated with an increased tendency to cancer."[82]

Recently, we revived Theodor Boveri's somatic mutation theory[45] and have directly addressed the question of whether aneuploidy is the cause or a consequence of cancer. We[30,38,44,83] and others[71,84–86] have provided evidence that an imbalance in the number and composition of chromosomes (aneuploidy) is the underlying cause of cancer and is sufficient to explain all of the characteristic phenotypes and properties of cancer: anaplasia, autonomous growth, metastasis, abnormal cell morphology, DNA indices ranging from 0.5 to over 2, genetic instability, the high levels of membrane-bound and secreted proteins responsible for invasiveness and loss of contact inhibition, multi-drug-resistance, and the exceedingly long times of up to decades from carcinogen exposure to the appearance of cancer.

Example 4: F, RNA Index, and ϕ Determined for Six Human Cancers. Occasionally one can find useful microarray data for human cancer. The best data contain replicate runs and primary diploid reference samples for the same tissue type as the cancer. Analogous to flow cytometry, histograms of cancer transcript microarray data visualize aneuploidy. The six cancers shown in Figure 7—pancreas,[87] colon,[88] lymphoma,[89] breast,[90] stomach,[91] kidney[92]—were compared to normal tissue of the same type from which the cancer originated. By way of comparison, it was possible to compare the background spread in transcripts from normal tonsil and skin (graphs at top of Figure 7) because data from more than one sample were available.[67,89] However, these kinds of data are rarely part of micorarray results. Figure 7 shows that the normal tissues were characterized by a tight distribution of transcripts centered at RNA index = 1. In contrast to normal tonsil and normal skin, the distributions of cancer transcripts were all decidedly different and irregular compared to the respective normal tissues. The aneuploid fractions ϕ and RNA indices were characteristically large for all the cancers, indicating advanced malignancies.[38,44,50] For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$.

Example 5: D and γ Correlated with Invasive Ductal Carcinomas Stratified by Grade. Current laboratory diagnoses of cancer are based on interpretations that are unavoidably subjective. As Crum et al. state, much of practice in cytology and histology involves evaluating abnormal smears and biopsies under suboptimal circumstances or rendering diagnoses that are frequently based more on instinct than objective criteria.[93] Consequently, false positive and false

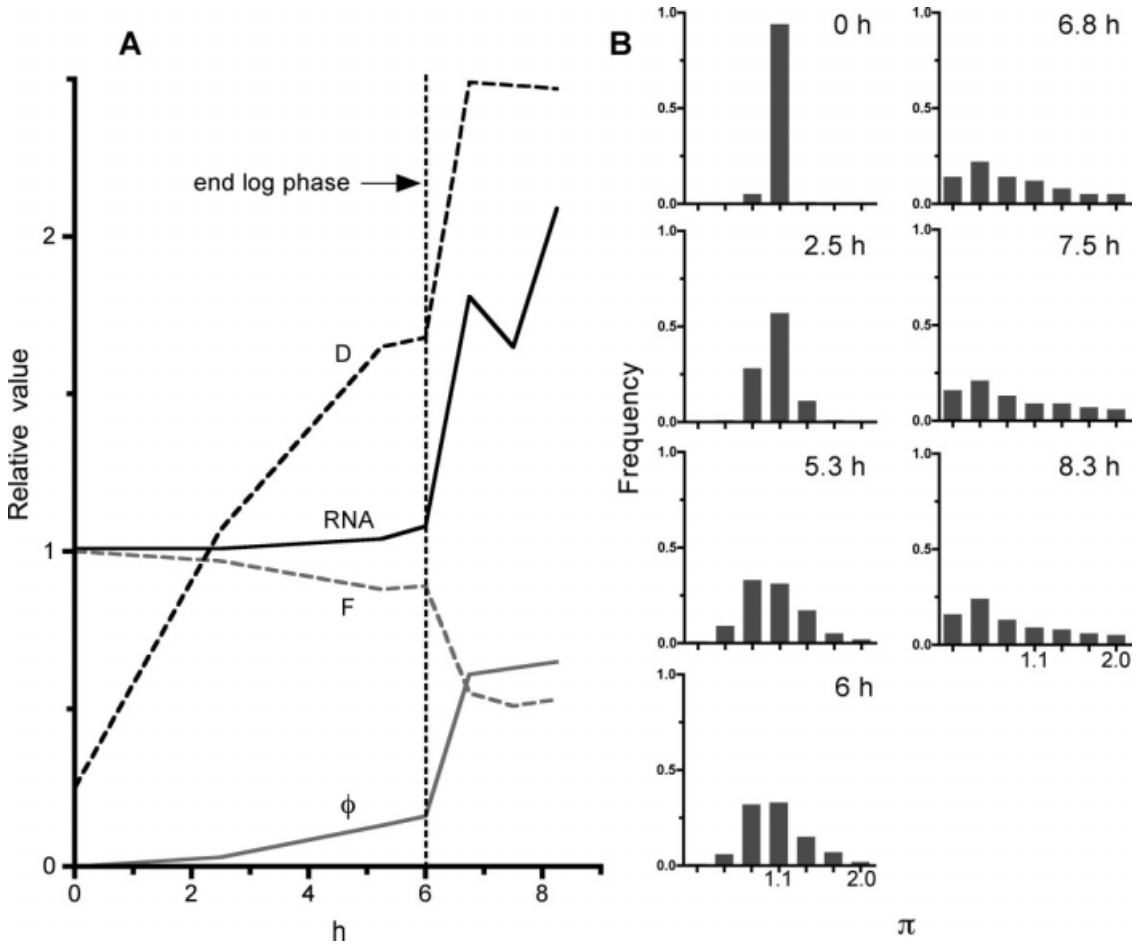

Figure 6. ***S. pneumoniae*** **adapting to exponential growth and its abrupt end at 6 h.**

A: The state variable ϕ was relatively small before the end of log phase (6 h), 0 h was reference (ϕ is the fraction of the genome undergoing differential expression). For determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$ based on comparisons of normal tissues of the same type (π_i is the fold-change in the expression of the ith gene of test cells relative to reference). At the end of log phase, there was a dramatic jump in ϕ at 6 h. There was also a jump in RNA index, implying a large increase in the synthesis of protein (RNA index is the total number of transcripts of test cells divided by total number of transcripts of reference cells). However, there was a pronounced reduction in F, signifying a massive imbalance in differential transcription (F is the dimensionless state variable quantifying the relative overall cellular activity of test cells compared to reference cells). The large increase in D during exponential growth indicated a growing population of metabolically heterogeneous bacteria (D is the distribution entropy of histogram data). B: The approximately symmetrical distribution of transcripts broadened until the end of log phase of growth at 6 h. After 6 h the spread continued to increase but had become skewed with the center of mass shifted to the left, signifying a major reduction in the overall cellular activity of the bacteria, which was also reflected by the sharp decline in F.

negative diagnoses are common.[94] DATE analysis offers a quantitative and objective means of characterizing both histological and cytological specimens.

DATE analysis was performed on the microarray data from 36 invasive ductal carcinomas of the breast for which there were clinical data.[72] The stated purpose of the Zhao et al. study was to determine if there were distinct genetic signatures distinguishing invasive ductal carcinoma from invasive lobular carcinoma. The authors did not correlate their results with clinical grade of the tumors. Because there was only one Grade 1 and no Grade 3 invasive lobular carcinoma patients, only the ductal carcinoma data were analyzed here.

Since genomic imbalance causes the genetic instability characteristic of invasive cancer,[38,55,70] the ductal carcinoma patients (represented by the black squares in Figure 8) were sorted along the horizontal axis by increasing values of D and γ, both measures of genomic imbalance of aneuploid cells. Grade 3 tumors were concentrated at high values of D and γ. With a notable exception the few examples of Grade 1 favored low values of D and γ. The Grade 1 tumor circled at the lower right of Figures 8A,B had the highest distribution entropy D of all the cancers. In my view this patient's tumor was likely misclassified and probably highly malignant. The Grade 2 tumors were disperse but tended to the left side of the graphs, with low and intermediate values of D and γ. It is likely intermediate Grade 2 is so uninformative as to be of little value.[95,96] This was recognized some years ago for cervical cancer when the intermediate category CIN-2 was eliminated.[97,98] Now there are only low and high grade cervical lesions. This simplified classification scheme has also been recommended for neoplastic lesions of esophagus, stomach, colon, and rectum.[99]

Evolution

Example 6: D and γ Recapitulate the Genetic Distance Separating the Great Apes. The state variables DNA index and F are not directly obtainable from array comparative genomic hybridization experiments (aCGH) because the data are typically normalized so that the modal ratio for the test genome as a whole is set to some standard value, typically 1.0 on a linear scale or 0.0 on a logarithmic scale, regardless of whether it is euploid, polyploid, or aneuploid.[100] Additional measurements, such as FISH (fluorescence in situ

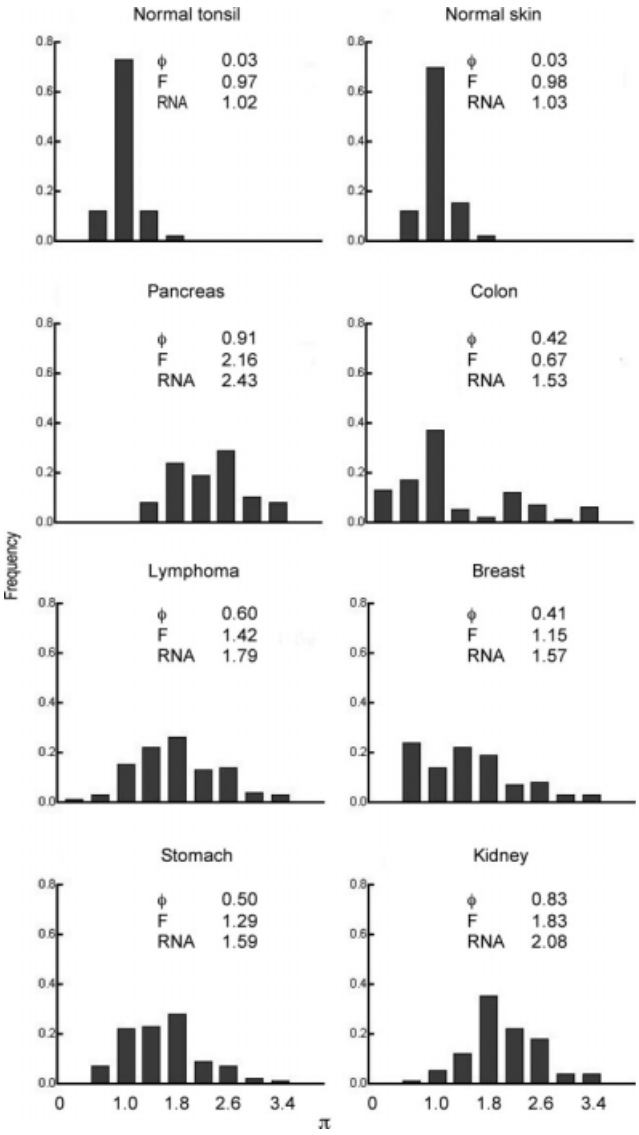

Figure 7. Six advanced human cancers.

Cancer cells are aneuploid and this genomic imbalance determines their properties. Normal tonsil and skin were characterized by a tight distribution of transcripts centered at RNA index = 1 (RNA index is the total number of transcripts of test cells divided by the total number of transcripts of reference cells). The cancers were compared to normal tissue of the same type from which each originated. In contrast to normal tonsil and skin (top graphs), the distribution of cancer transcripts were all decidedly different and irregular compared to normal tissues. The aneuploid fractions ϕ and RNA indices were characteristically large for all the cancers, indicating advanced malignancies.[38,44,50] For the determination of ϕ, the normal range of π was set to $0.5 \leq \pi \leq 1.5$ based on comparisons of normal tissues of the same type (ϕ is the fraction of the genome undergoing differential expression, π_i is the fold-change in the expression of the ith gene of test cells relative to reference).

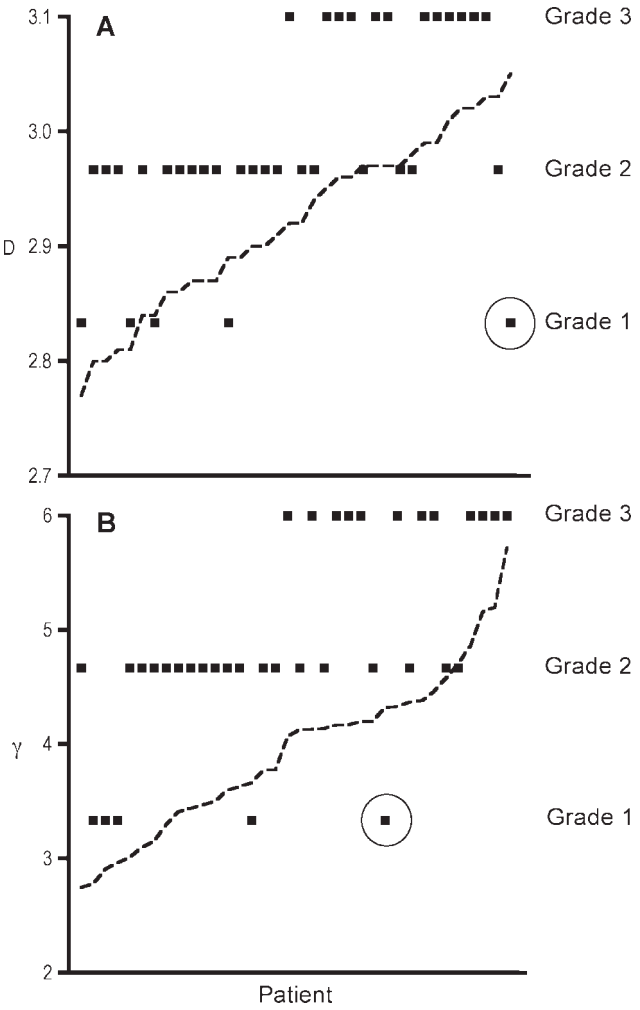

Figure 8. Invasive ductal carcinomas of the breast correlated with D and γ.

The solid squares represent a patient's tumor that had been graded 1, 2, or 3, for increasing severity. It is generally accepted that the most malignant cancers are the most genetically unstable. Both D and γ are measures of the genomic instability of cancer cells (D is the distribution entropy of histogram data, see text for definition of γ). Determination of D was based on a bin size of 0.1π (π_i is the fold-change in the expression of the ith gene of test cells relative to reference). A: Patients along the horizontal axis were sorted by increasing D (broken line). B: Patients were sorted by increasing γ (broken line). Grade 3 tumors were concentrated at high values of D and γ. The few examples of Grade 1 favored low values of D and γ. The Grade 1 tumor circled at the lower right of both graphs was likely misclassified and probably highly malignant. Grade 2 tumors were disperse but tended to the left side of the graphs, with low and intermediate values of D and γ. It is likely intermediate Grade 2 is so uninformative as to be of little value (see text).

hybridization) or flow cytometry, are needed to determine the DNA index of the test sample associated with the given ratio level. Even though DNA index and F cannot be calculated from aCGH data, D and γ are readily obtained since they are measures of the distribution or spread of comparative data.

Since all species draw from the same "dictionary" of genes,[34,101,102] the differences among species are due, in large part, to differences in copy numbers leading to differential orchestration and expression of the various genes. DATE analysis was applied to aCGH data from the great apes.[103] Figure 9 shows the great apes ordered along the horizontal axis according to increasing values of D, using human female as reference. Data from two women were compared to each other to give $D = 1.65$. Sorting the great apes by increasing γ (not shown) produced the same rank ordering of genetic distance relative to human. The order thus produced by DATE analysis (human, chimp, gorilla, orangutan) is identical with that accepted by many phylogeneticists.[103]

Discussion

Metabolic reactions have been preserved in bacteria, fungi, plants, and animals, essentially intact, through billions of

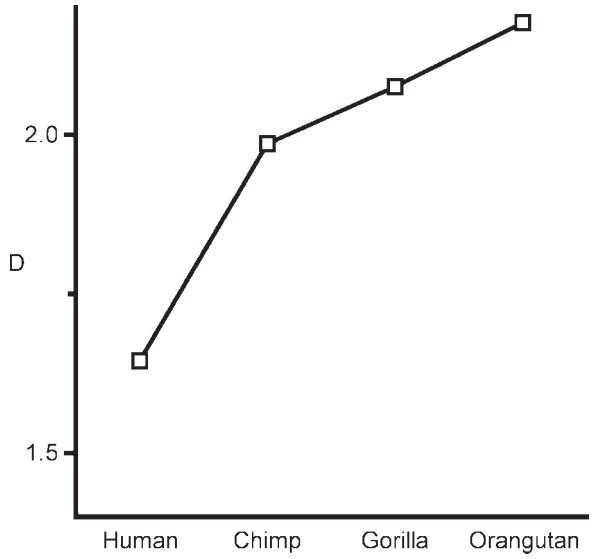

Figure 9. Phylogenetic order of the great apes correlated with *D* and γ.

The state variables DNA index and F are not directly obtainable from array comparative genomic hybridization experiments (aCGH). However, D and γ are readily obtained because they are measures of the distribution or spread of comparative data (D is the distribution entropy of histogram data, see text for definition of γ). The great apes[103] were ordered along the x-axis according to increasing values of D calculated from aCGH data. In this case D is a measure of the relative genetic distance separating the species, using human females as reference. Determination of D was based on a bin size of 0.1π (here π_i is the fold-change in copy number of the ith gene of test cells relative to reference). The two women were compared to each other giving $D = 1.65$ for the background spread of data. The phylogenetic order produced by DATE analysis was identical with that generally accepted by phylogeneticists. Sorting the great apes according to increasing γ (not shown) produced the same phylogenetic order.

years of evolution.[101] Yet, these immutable reactions produced the extraordinary variety of organisms that have populated the earth since life began. It is the particular orchestration and connectedness of these immutable reactions that make up life in all its diversity and are the subject of MCA and DATE analysis.

While the principles and insights of MCA are indeed profound, the application of its methods to the tens-of-thousands of interconnected components of whole cells is not feasible. Nevertheless, the historic work of Kacser and Burns on the molecular basis of dominance[34] contained the seeds of transforming MCA into a quantitative method of phenotypic analysis that can be applied to whole cells, organs, even organisms.

The theoretical basis of DATE analysis (with its foundation in MCA) sets it apart from all other bioinformatics approaches, which are fundamentally statistical in nature. DATE analysis differs from MCA in that its essential task lies in the comparison of phenotypes rather than in the precise definition or description of each. In the words of Thompson, "This process of comparison, of recognizing in one form a definite permutation or deformation of another, apart altogether from a precise and adequate understanding of the original 'type' or standard of comparison, lies within the immediate province of mathematics."[104]

Published microarray data were used to test the validity of the assumption, implicit in MCA, that individual genes contributing to macroscopic phenotypes can be treated as being quantitatively equivalent.[34,38,41] Figure 2 shows that random subsets of transcript microarray data, from as little as 10% of the whole, led to reproducible values of the intensive state variables F, RNA index, and ϕ. It appears that around 1,000 random transcripts are sufficient to generate accurate and stable values of F, RNA index, and ϕ.

Using random values of π_i and $1/\pi_i$ in order to calculate the state variables resulted in losing information contained in the expression of individual genes. However, the lost information can be recovered by analyzing ordered values of π_i (increasing) and $1/\pi_i$ (decreasing). α and β were introduced as measures of the information gained when comparing ordered vs. random values of $1/\pi_i$ and π_i, respectively. Since ordered values of π_i and $1/\pi_i$ are not linear, the whole dataset must be used to calculate α and β (Figure 3).

DATE analysis eliminates selection bias by using the whole microarray data to calculate state variables. The fact that the state variables F, RNA index, and ϕ are not dependent on a defined or unique set of specific genes contrasts sharply with the unstable "genetic signatures" generated by conventional data mining.[6,10,11,19] Aneuploidy and compensating differential expression are the major sources of the instability plaguing the genetic signatures derived from data mining. In addition, the massively interconnected metabolic networks make it highly unlikely macroscopic phenotypes are restricted to unique or characteristic expression profiles of specific genes.

The proper choice of reference cells or tissue is integral to DATE analysis. However, it was difficult to find published microarray data with the appropriate references. The choice of reference will depend on the specific experimental question being asked. For example, normal cervical cells should be the reference if one is studying cervical cancer. It is important that microarray experiments compare different samples of the same reference to each other to determine the background spread of data. It is also essential to include replicates of all the experiments. Regrettably, replicates have been rare in published microarray experiments, though this appears to be changing.

Time-course micorarray data are particularly amenable to DATE analysis because the results at any time point can represent a suitable reference for other time points. Indeed, half of the examples presented here were from time-course data: *Plasmodium falciparum*, Killifish, and *Streptococcus pneumoniae*.

Cycling of the state variables F, RNA index, and ϕ clearly mirrored the blood-feeding cycle of the malaria parasite. A moving reference (i.e., comparing adjacent time points) allowed the detection of a 2.3 h metabolic cycle during the ring stage of the parasite. Using a moving reference for the fruit fly embryo resulted in noncycling values of F and RNA index close to one, revealing a smooth, noncycling developmental course. Figures 4A,C invites the speculation that the time-course trajectories of the state variables may signify a biological principle analogous to the principle of least action in physics.[105]

The Killifish 24 h temperature cycle experiment demonstrated the general principle that stable phenotypes maintain tight regulation of RNA index = 1 when adapting to stress. While changes in RNA index were barely detectable, there was a pronounced cycling of the overall cellular activity F with a period of 20 h during the first 48 h of the experiment. The cycle of F may have eventually synchronized to the 24 h temperature cycle. Extending the temperature cycling

experiment to four or more complete 24 h cycles would be needed to test this hypothesis.

Streptococcus pneumoniae demonstrated the power of D (histogram entropy) to detect, during exponential growth, the production of a metabolically heterogeneous population of the single-cell organism. This process is reminiscent of, but different from, the rapid onset of intercellular heterogeneity produced by clonal expansion of aneuploid eukaryotic cells.[55,70,106] D and γ are objective and quantitative measures of the genomic imbalance of cancer cells resulting in widespread genetic instability.

The aCGH data for the great apes[103] provided the only evolution example available for DATE analysis. The DATE analysis results produced the generally accepted phylogenetic order of the great apes. DATE analysis promises to be an ideal tool for studying evolution because it provides an objective, quantitative means of comparing the genetic distances separating species.

Conclusions

These preliminary results are encouraging, especially because the data collection and this analysis were completely independent. DATE analysis is analogous to the state equations of thermodynamics in physics, while the kinetics details of metabolic control analysis (upon which DATE analysis is based) can be likened to statistical mechanics. Just as with thermodynamics and statistical mechanics, DATE analysis and metabolic control analysis are mutually consistent but serve different conceptual and experimental functions. DATE analysis eliminates selection bias by analyzing microarray data as a whole. For large datasets, the state variables F, RNA index, and ϕ are intensive variables, thus not dependent on a defined subset of specific genes. This result is of fundamental importance and distinguishes DATE analysis from conventional data mining, which seeks a stable and unique set of genes as either diagnostic tools or targets of drug therapy. Aneuploidy and compensating differential expression are the major sources of the instability plaguing the genetic signatures derived from conventional data mining. DATE analysis provides a robust strategy of correlating specific phenotypes with the state variables F, RNA index, and ϕ along with the measures of data dispersion D and γ. As a general theory of biological change, DATE analysis can be used to design, analyze, and interpret future microarray experiments.

Literature Cited

1. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*. 2003;95:14–18.
2. Macgregor PF, Squire JA. Application of microarrays to the analysis of gene expression in cancer. *Clin Chem*. 2002;48: 1170–1177.
3. Ochs MF, Godwin AK. Microarrays in cancer: research and applications. *BioTechniques*. 2003;Suppl:4–15.
4. Dunkler D, Michiels S, Schemper M. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur J Cancer*. 2007;43:745–751.
5. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*. 2005;365:488–492.
6. Michiels S, Koscielny S, Hill C. Interpretation of microarray data in cancer. *Br J Cancer*. 2007;96:1155–1158.
7. Reid JF, Lusa L, De Cecco L, Coradini D, Veneroni S, Daidone MG, Gariboldi M, Pierotti MA. Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *J Natl Cancer Inst*. 2005;97:927–930.
8. Eden P, Ritz C, Rose C, Ferno M, Peterson C. "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer*. 2004;40:1837–18341.
9. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*. 2003;362:1439–1444.
10. Koscielny S. Critical review of microarray-based prognostic tests and trials in breast cancer. *Curr Opin Obstet Gynecol*. 2008;20:47–50.
11. Pan K-H, Lih C-J, Cohen SN. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc Natl Acad Sci*. 2005;102: 8961–8965.
12. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*. 2007;99:147–157.
13. Shi L, Perkins RG, Fang H, Tong W. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr Opin Biotechnol*. 2008;19:10–18.
14. Bains W. The parts list of life. *Nat Biotechnol*. 2001;19:401–402.
15. Newman SA. Developmental mechanisms: putting genes in their place. *J Biosci*. 2002;27:97–104.
16. Shastry BS. Genetic knockouts in mice: an update. *Experientia*. 1995;51:1028–1039.
17. Sigmund CD. Viewpoint: are studies in genetically altered mice out of control? *Arterioscler Thromb Vasc Biol*. 2000;20: 1425–1429.
18. Miklos GL, Maleszka R. Microarray reality checks in the context of a complex disease. *Nat Biotechnol*. 2004;22:615–621.
19. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*. 2005;21:171–178.
20. Elser JJ, Hamilton A. Stoichiometry and the new biology: the future is now. *PLoS Biol*. 2007;5:e181.
21. Fell D. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J*. 1992;286:313–330.
22. Daran-Lapujade P, Rossell S, van Gulik WM, Luttik MAH, de Groot MJL, Slijper M, Heck AJR, Daran J-M, de Winde JH, Westerhoff HV, Pronk JT, Bakker BM. The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proc Natl Acad Sci*. 2007;104:15753–15758.
23. Kacser H. On parts and wholes in metabolism. In: Welch GR, Clegg JS, editors. *The Organizaton of Cell Metabolism*. New York: Plenum Press; 1986:327–337.
24. Kacser H, Burns JA. The control of flux. *Symp Soc Exp Biol*. 1973;27:65–104.
25. Fell D. *Understanding the Control of Metabolism*. London: Portland Press; 1997:301.
26. Kacser H, Burns JA. Molecular democracy: who shares the controls? *Biochem Soc Trans*. 1979;7:1149–1160.
27. Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem*. 1974;42:89–95.
28. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007;447:799–816.
29. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, Defazio A, Greaves MF, Green AR, Campbell

P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;446:153–158.
30. Duesberg P, Rasnick D. Aneuploidy, the somatic mutation that makes cancer a species of its own. *Cell Motil Cytoskeleton*. 2000;47:81–107.
31. Duesberg PH, Schwartz JR. Latent viruses and mutated oncogenes: no evidence for pathogenicity. *Prog Nucleic Acid Res Mol Biol*. 1992;43:135–204.
32. O'Neill GM, Catchpoole DR, Golemis EA. From correlation to causality: microarrays, cancer, and cancer treatment. *BioTechniques*. 2003;Suppl:64–71.
33. Fell D, Thomas S. Physiological control of metabolic flux: the requirement for multisite modulation. *Biochem J*. 1995;311: 35–39.
34. Kacser H, Burns JA. The molecular basis of dominance. *Genetics*. 1981;97:639–666.
35. Niederberger P, Prasad R, Miozzari G, Kacser H. A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast. *Biochem J*. 1992;287 (Part 2):473–479.
36. Rosen R. *Life Itself*, 1st ed. New York: Columbia University Press; 1991, Vol. 1:285.
37. Cornish-Bowden A. *Fundamentals of Enzyme Kinetics*. London: Portland Press; 2004.
38. Rasnick D, Duesberg PH. How aneuploidy affects metabolic control and causes cancer. *Biochem J*. 1999;340:621–630.
39. Kacser H. Recent developments beyond metabolic control analysis. *Biochem Soc Trans*. 1995;23:387–391.
40. Kacser H. The control of flux: 21 years on. *Biochem Soc Trans*. 1995;23:341–366.
41. Brown G. Total cell protein concentration as an evolutionary constraint on the metabolic control distribution in cells. *J Theor Biol*. 1991;153:195–203.
42. Kahn D, Westerhoff HV. Control theory of regulatory cascades. *J Theor Biol*. 1991;153:255–285.
43. Westerhoff HV, Koster JG, van Workum M, Rudd KE. On the control of gene expression. In: Cardenas AC-BML, editor. *Control of Metabolic Processes*. New York: Plenum Press; 1990.
44. Rasnick D. Auto-catalyzed progression of aneuploidy explains the Hayflick limit of cultured cells, carcinogen-induced tumours in mice, and the age distribution of human cancer. *Biochem J*. 2000;348:497–506.
45. Boveri T. *Zur Frage der Entstehung Maligner Tumoren*. Fischer: Jena; 1914.
46. Torres EM, Sokolsky T, Tucker CM, Chan LY, Boselli M, Dunham MJ, Amon A. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science*. 2007;317: 916–924.
47. Galitski T, Saldanha AJ, Styles CA, Lander ES, Fink GR. Ploidy regulation of gene expression. *Science*. 1999;285:251–254.
48. Masayesva BG, Ha P, Garrett-Mayer E, Pilkington T, Mao R, Pevsner J, Speed T, Benoit N, Moon CS, Sidransky D, Westra WH, Califano J. Gene expression alterations over large chromosomal regions in cancers include multiple genes unrelated to malignant progression. *Proc Natl Acad Sci USA*. 2004;101: 8715–8720.
49. Huettel B, Kreil DP, Matzke M, Matzke AJ. Effects of aneuploidy on genome structure, expression, and interphase organization in *Arabidopsis thaliana*. *PLoS Genet*. 2008;4: e1000226.
50. Rasnick D. Aneuploidy theory explains tumor formation, the absence of immune surveillance and the failure of chemotherapy. *Cancer Genet Cytogenet*. 2002;136:66–72.
51. Killander D, Zetterberg A. Quantitative cytochemical studies on interphase growth. *Exp Cell Res*. 1965;38:272–284.
52. Zetterberg A, Killander D. Quantitative cytochemical studies on interphase growth. II. Derivation of synthesis curves from the distribution of DNA, RNA and mass values of individual mouse fibroblasts in vitro. *Exp Cell Res*. 1965;39: 22–32.
53. Zetterberg A, Killander D. Quantitative cytophotometric and autoradiographic studies on the rate of protein synthesis during interphase in mouse fibroblasts in vitro. *Exp Cell Res*. 1965;40:1–11.
54. Epstein CJ. *The Consequences of Chromosome Imbalance: Principles, Mechanisms, and Models*. New York: Cambridge University Press; 1986:475.
55. Camps J, Ponsa I, Ribas M, Prat E, Egozcue J, Peinado MA, Miro R. Comprehensive measurement of chromosomal instability in cancer cells: combination of fluorescence in situ hybridization and cytokinesis-block micronucleus assay. *FASEB J*. 2005;19:828–830.
56. Losi L, Baisse B, Bouzourene H, Benhattar J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. *Carcinogenesis*. 2005;26:916–922.
57. Killander D. Intercellular variations in generation time and amounts of DNA, RNA and mass in a mouse leukemia population in vitro. *Exp Cell Res*. 1965;40:21–31.
58. Grade M, Hormann P, Becker S, Hummon AB, Wangsa D, Varma S, Simon R, Liersch T, Becker H, Difilippantonio MJ, Ghadimi BM, Ried T. Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. *Cancer Res*. 2007;67:41–56.
59. Grade M, Ghadimi BM, Varma S, Simon R, Wangsa D, Barenboim-Stapleton L, Liersch T, Becker H, Ried T, Difilippantonio MJ. Aneuploidy-dependent massive deregulation of the cellular transcriptome and apparent divergence of the Wnt/beta-catenin signaling pathway in human rectal carcinomas. *Cancer Res*. 2006;66:267–282.
60. Gao C, Furge K, Koeman J, Dykema K, Su Y, Cutler ML, Werts A, Haak P, Vande Woude GF. Chromosome instability, chromosome transcriptome, and clonal evolution of tumor cell populations. *Proc Natl Acad Sci USA*. 2007;104:8995–9000.
61. Aggarwal A, Leong SH, Lee C, Kon OL, Tan P. Wavelet transformations of tumor expression profiles reveals a pervasive genome-wide imprinting of aneuploidy on the cancer transcriptome. *Cancer Res*. 2005;65:186–194.
62. Lion S, Gabriel F, Bost B, Fievet J, Dillmann C, de Vienne D. An extension to the metabolic control theory taking into account correlations between enzyme concentrations. *Eur J Biochem*. 2004;271:4375–4391.
63. Cantrell DW. Pythagorean means. In: *From MathWorld—A Wolfram Web Resource (created by Weisstein EW)*. 2003. Available at: http://mathworld.wolfram.com/PythagoreanMeans.html.
64. Lindsley DL, Sandler L, Baker BS, Carpenter AT, Denell RE, Hall JC, Jacobs PA, Miklos GL, Davis BK, Gethmann RC, Hardy RW, Steven AH, Miller M, Nozawa H, Parry DM, Gould-Somero M. Segmental aneuploidy and the genetic gross structure of the Drosophila genome. *Genetics*. 1972;71:157–184.
65. Shannon CE, Weaver W. *The Mathematical Theory of Communication*. Urbana, Champaign: University of Illinois Press; 1949.
66. Castro MA, Onsten TT, de Almeida RM, Moreira JC. Profiling cytogenetic diversity with entropy-based karyotypic analysis. *J Theor Biol*. 2005;234:487–495.
67. Storz MN, van de Rijn M, Kim YH, Mraz-Gernhard S, Hoppe RT, Kohler S. Gene expression profiles of cutaneous B cell lymphoma. *J Invest Dermatol*. 2003;120:865–870.
68. Heppner G, Miller FR. The cellular basis of tumor progression. *Int Rev Cytol*. 1998;177:1–56.
69. Nowell PC. The clonal evolution of tumor cell populations. *Science*. 1976;194:23–28.
70. Duesberg P, Rausch C, Rasnick D, Hehlmann R. Genetic instability of cancer cells is proportional to their degree of aneuploidy. *Proc Natl Acad Sci USA*. 1998;95:13692–13697.
71. Ye CJ, Stevens JB, Liu G, Bremer SW, Jaiswal AS, Ye KJ, Lin MF, Lawrenson L, Lancaster WD, Kurkinen M, Liao JD, Gairola CG, Shekhar MP, Narayan S, Miller FR, Heng HH. Genome based cell population heterogeneity promotes tumorigenicity: the evolutionary mechanism of cancer. *J Cell Physiol*. 2009;219:288–300.

72. Zhao H, Langerod A, Ji Y, Nowels KW, Nesland JM, Tibshirani R, Bukholm IK, Karesen R, Botstein D, Borresen-Dale AL, Jeffrey SS. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol Biol Cell*. 2004;15:2523–2536.
73. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP. Gene expression during the life cycle of *Drosophila melanogaster*. *Science*. 2002;297:2270–2275.
74. Podrabsky JE, Somero GN. Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish *Austrofundulus limnaeus*. *J Exp Biol*. 2004;207:2237–2254.
75. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol*. 2003;1:e5.
76. Joyce EA, Kawale A, Censini S, Kim CC, Covacci A, Falkow S. LuxS is required for persistent pneumococcal carriage and expression of virulence and biosynthesis genes. *Infect Immun*. 2004;72:2964–2975.
77. Hansemann D. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virschows Arch Pathol Anat*. 1890;119:299–326.
78. Mitelman F. *Catalogue of Chromosome Aberrations in Cancer*, 4th ed. New York: Wiley-Liss; 1994.
79. Sandberg AA. The Chromosomes in Human Cancer and Leukemia, 2nd ed. New York: Elsevier Science Publishing; 1990.
80. Gebhart E, Liehr T. Patterns of genomic imbalances in human solid tumors (Review). *Int J Oncol*. 2000;16:383–399.
81. Mertens F, Johansson B, Hoglund M, Mitelman F. Chromosomal imbalance maps of malignant solid tumors: a cytogenetic survey of 3185 neoplasms. *Cancer Res*. 1997;57:2765–2780.
82. Levan A. Chromosome abnormalities and carcinogenesis. In: Lima-de-Faria A, editor. *Handbook of Molecular Cytology*. New York: American Elsevier Publishing; 1969:716–731.
83. Li R, Sonik A, Stindl R, Rasnick D, Duesberg P. Aneuploidy versus gene mutation hypothesis: recent study claims mutation, but is found to support aneuploidy. *Proc Natl Acad Sci USA*. 2000;97:3236–3241.
84. Liu P, Zhang H, McLellan A, Vogel H, Bradley A. Embryonic lethality and tumorigenesis caused by segmental aneuploidy on mouse chromosome 11. *Genetics*. 1998;150:1155–1168.
85. Heng HH, Bremer SW, Stevens J, Ye KJ, Miller F, Liu G, Ye CJ. Cancer progression by non-clonal chromosome aberrations. *J Cell Biochem*. 2006;98:1424–1435.
86. Heng HH, Stevens JB, Liu G, Bremer SW, Ye KJ, Reddy PV, Wu GS, Wang YA, Tainsky MA, Ye CJ. Stochastic cancer progression driven by non-clonal chromosome aberrations. *J Cell Physiol*. 2006;208:461–472.
87. Iacobuzio-Donahue CA, Maitra A, Olsen M, Lowe AW, van Heek NT, Rosty C, Walter K, Sato N, Parker A, Ashfaq R, Jaffee E, Ryu B, Jones J, Eshleman JR, Yeo CJ, Cameron JL, Kern SE, Hruban RH, Brown PO, Goggins M. Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am J Pathol*. 2003;162:1151–1162.
88. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res*. 2001;61:3124–3130.
89. Bohen SP, Troyanskaya OG, Alter O, Warnke R, Botstein D, Brown PO, Levy R. Variation in gene expression patterns in follicular lymphoma and the response to rituximab. *Proc Natl Acad Sci USA*. 2003;100:1926–1930.
90. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. *Nature*. 2000;406:747–752.
91. Chen X, Leung SY, Yuen ST, Chu KM, Ji J, Li R, Chan AS, Law S, Troyanskaya OG, Wong J, So S, Botstein D, Brown PO. Variation in gene expression patterns in human gastric cancers. *Mol Biol Cell*. 2003;14:3208–3215.
92. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau W-C, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. NCBI GEO: mining millions of expression profiles—database and tools. *Nucl Acids Res*. 2005;33 (Suppl 1):d562–d566.
93. Crum CP, Cibas ES, Lee KR. Pathology of Early Cervical Neoplasia, Vol. 22. New York: Churchill Livingstone; 1997:288.
94. Anthony PP. *Diagnostic Pitfalls in Histology and Cytopathology Practice*. London: Greenwich Medical Media; 1998: 125.
95. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, Lindahl T, Pawitan Y, Hall P, Nordgren H, Wong JEL, Liu ET, Bergh J, Kuznetsov VA, Miller LD. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*. 2006;66:10292–10301.
96. Ellsworth RE, Ellsworth DL, Love B, Patney HL, Hoffman LR, Kane J, Hooke JA, Shriver CD. Correlation of levels and patterns of genomic instability with histological grading of DCIS. *Ann Surg Oncol*. 2007;14:3070–3077.
97. National Cancer Institute. The 1988 Bethesda system for reporting cervical/vaginal cytological diagnoses. National Cancer Institute Workshop. *JAMA*. 1989;262:931–934.
98. Bollmann R, Bollmann M, Henson DE, Bodo M. DNA cytometry confirms the utility of the Bethesda system for the classification of Papanicolaou smears. *Cancer Cytopathol*. 2001;93: 222–228.
99. Schlemper RJ, Riddell RH, Kato Y, Borchard F, Cooper HS, Dawsey SM, Dixon MF, Fenoglio-Preiser CM, Flejou JF, Geboes K, Hattori T, Hirota T, Itabashi M, Iwafuchi M, Iwashita A, Kim YI, Kirchner T, Klimpfinger M, Koike M, Lauwers GY, Lewin KJ, Oberhuber G, Offner F, Price AB, Rubio CA, Shimizu M, Shimoda T, Sipponen P, Solcia E, Stolte M, Watanabe H, Yamabe H. The Vienna classification of gastrointestinal epithelial neoplasia. *Gut*. 2000;47:251–255.
100. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*. 2005;37 (Suppl):s11–s17.
101. Kornberg A. Ten commandments: lessons from the enzymology of DNA replication. *J Bacteriol*. 2000;182:3613–3618.
102. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000;25:25–29.
103. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, Karimpour-Fard A, Glueck D, McGavran L, Berry R, Pollack J, Sikela JM. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol*. 2004;2:e207.
104. Thompson D. *On Growth and Form: the Complete Revised Edition*. New York: Dover; 1992:1116.
105. Moore TA. Least-action principle. In: Rigden J, editor. *Macmillan Encyclopedia of Physics*, Vol. 2. New York: Simon & Schuster Macmillan; 1996:840.
106. Duesberg P, Li R. Multistep carcinogenesis: a chain reaction of aneuploidizations. *Cell Cycle*. 2003;2:202–210.

Manuscript received Sep. 10, 2008, and revision received Feb. 15, 2009.