# PathScan: A Tool for Discerning Mutational Significance in Groups of Putative Cancer Genes

Michael C. Wendl [1,2,3,]*, John W. Wallis [1,2], Ling Lin [1], Cyriac Kandoth [1], Elaine R. Mardis [1,2,4], Richard K. Wilson [1,2,4], and Li Ding [1,2]

[1]The Genome Institute, [2]Department of Genetics, [3]Department of Mathematics, and [4]Siteman Cancer Center, Washington University, St. Louis MO 63108, USA

Associate Editor: Prof. Alfonso Valencia

## ABSTRACT

**Motivation:**

The expansion of cancer genome sequencing continues to stimulate development of analytical tools for inferring relationships between somatic changes and tumor development. Pathway associations are especially consequential, but existing algorithms are demonstrably inadequate.

**Methods:**

Here, we propose the PathScan significance test for the scenario where pathway mutations collectively contribute to tumor development. Its design addresses two aspects that established methods neglect. First, we account for variations in gene length and the consequent differences in their mutation probabilities under the standard null hypothesis of random mutation. The associated spike in computational effort is mitigated by accurate convolution–based approximation. Second, we combine individual probabilities into a multiple–sample value using Fisher–Lancaster theory, thereby improving differentiation between a few highly–mutated genes and many genes having only a few mutations apiece. We investigate accuracy, computational effort, and power, reporting acceptable performance for each.

**Results:**

As an example calculation, we re–analyze KEGG–based lung adenocarcinoma pathway mutations from the Tumor Sequencing Project. Our test recapitulates the most significant pathways and finds others for which the original test battery was inconclusive are not actually significant. It also identifies the focal adhesion pathway as being significantly mutated, a finding consistent with earlier studies. We also expand this analysis to other databases: Reactome, BioCarta, Pfam, PID, and SMART, finding additional hits in ErbB and EPHA signaling pathways and regulation of telomerase. All have implications and plausible mechanistic roles in cancer. Finally, we discuss aspects of extending the method to integrate gene–specific background rates and other types of genetic anomalies.

**Availability:**

PathScan is implemented in Perl and is available from the Genome Institute at: `http://genome.wustl.edu/software/pathscan`.

*to whom correspondence should be addressed

## 1 INTRODUCTION

The Human Genome Project (HGP) recently culminated in the first composite human reference sequence (International Human Genome Sequencing Consortium, 2004) and researchers have since been vigorously building upon this result. Much of the work targets medical applications, such as in cancer genomics, and a significant fraction of the sequencing enterprise is now shifting in that direction (Sjöblom et al., 2006; Ley et al., 2008; Mardis et al., 2009; Shah et al., 2009; Ding et al., 2010; Berger et al., 2011). Indeed, instruments and automation have advanced to the point where the ability to sequence both the tumor and normal genomes from large numbers of patients is now emerging.

Such whole genome data should allow somatic mutations to be reliably separated from germline variations for further study. The subsequent, more difficult challenge then becomes one of differentiating functionally related somatic "driver" mutations from incidental "passenger" variants (Wood et al., 2007; Greenman et al., 2007). In the early stages of a project, this task usually manifests itself as a hypothesis testing problem on the mutational significance of genes or pathways. The intent is to filter an initially large collection of candidates down to a better targeted set that will be examined more comprehensively (Sjöblom et al., 2006; Wood et al., 2007). Concerns at this stage revolve mainly around false–positive and false–negative errors, i.e. instances where an irrelevant feature is accepted and where a true feature is overlooked, respectively.

Methods for statistical testing of cancer DNA sequence data are now actively being developed (Beroukhim et al., 2007), several of which are listed in Table 1. While there are certain subject–specific nuances in applying the statistical method to cancer sequence data, these examples all share the commonality of being founded upon well–established concepts from mathematical statistics. If history is any guide, we anticipate development of additional types of tests and a subsequent "toolbox" approach for statistical inference in cancer genomics studies (Ding et al., 2008).

While these remarks paint a pleasant picture of orderly development and application, the design of new statistical tools for cancer sequence has not been without its difficulties. Aside from the requirement that a test be constructed on a sound basis, i.e. the underlying hypothesis is scientifically relevant and the resultant p–value is a reliable indicator on this hypothesis, it must also satisfy a more utilitarian "computability" condition. That is, a test is unlikely to find broad application if it is inordinately difficult

**Table 1.** Representative hypothesis tests in cancer sequencing

| test | mathematical basis | reference |
|------|--------------------|-----------|
| CaMP[†] | binomial | Sjöblom et al. (2006) |
| log–likelihood | binomial | Getz et al. (2007) |
| Group–CaMP | binomial | Lin et al. (2007) |
| Greenman's test | Poisson | Greenman et al. (2006) |
| Ratio Test | Monte–Carlo | Stephens et al. (2005) |
| TRAB | Poisson/Gamma | Parmigiani et al. (2008) |

[†]Cancer Mutation Prevalence Score

to evaluate (Brown et al., 2001). This aspect is sometimes not sufficiently appreciated, as the odyssey of the Cancer Mutation Prevalence (CaMP) score (Sjöblom et al., 2006) so dramatically illustrates. CaMP was roundly criticized as *ipso facto* incorrect because of its use of probability mass values rather than tailed p–values (Forrest and Cavet, 2007; Getz et al., 2007; Rubin and Green, 2007). Its designers countered that the CaMP concept is sound, but conceded that its form makes it very difficult in practice to compute (Parmigiani et al., 2007). Their claim, that mass values are a legitimate substitute, clearly contravenes standard theory (Sokal and Rohlf, 1981).

Advances in cancer genomics will depend, among other things, on statistical tests that are sufficient in both rigor and economy. Given that a combinatorially large number and diversity of somatic events at the gene level tend to collapse at the pathway level, there is growing consensus that the search for drivers is best focused on the latter (Vogelstein and Kinzler, 2004; Wood et al., 2007; Lin et al., 2007; Beroukhim et al., 2007; Glaab et al., 2010; Vandin et al., 2010; Cerami et al., 2010; Efroni et al., 2011). It is for this scenario that we wish to propose a significance test that we call PathScan. Methodologically, it considers certain information that other tests neglect, specifically the distributions of both gene lengths within a pathway and of mutations among samples. We demonstrate that ignoring these factors greatly compromises resulting p–values. The test is similar to those in Table 1 in the sense that it relies on certain, well–established probability concepts. Consequently, we do not claim any particular mathematical significance, but feel rather that our contribution lies in a well–balanced combination of biological relevance, conceptual probity, and efficient algorithmic implementation. These characteristics should render it useful for actual application.

## 2 METHODS

Let a "test set" be any biologically–relevant collection of $m$ genes, $\gamma = \{g_1, g_2, \ldots, g_m\}$, e.g. the members of a pathway, and assume genomic samples have been sequenced such that somatic mutations can be distinguished from germline variation in these genes. We take the null hypothesis, $H_0$, as the scenario in which there is no association between the disease phenotype and $\gamma$ in a particular sample by virtue of the latter's somatic mutations having occurred according to some random process, itself characterized by an underlying background mutation rate, $\rho$. The test we propose is based on the observed number of genes $k$ that are mutated in $\gamma$.

PathScan resolves two fundamental issues not generally recognized by other methods. First, it accounts for variations in gene length and the consequent differences in their mutation probabilities under $H_0$. Second, it properly frames the multi–sample "overall p–value" for $\gamma$ in terms of its individual tests of $H_0$ for each sample. The details and implications of these two aspects are described below.

### 2.1 Probability Masses

Background rates in cancer genomes are typically estimated to be on the order of $10^{-6}$ per nucleotide (Wang et al., 2002; Stephens et al., 2005; Sjöblom et al., 2006; Greenman et al., 2007; Ding et al., 2008), meaning that somatic mutations will remain relatively rare under $H_0$. Most individual genes will have either 1 or 0 mutations with a probability approaching unity (see proof of Thm. 1 below). Consequently, we can reasonably dispense with the distribution of mutations within an instance of a gene, simply treating it as either mutated (one or more mutations) or not mutated (zero mutations). Under $H_0$, longer genes are more likely to be mutated than shorter ones (Getz et al., 2007), which leads to gene–specific, non–identical mutation probabilities.

THEOREM 1 (Bernoulli Mutation Probability for a Gene). If the length of gene $i$ is $L_i$, the Bernoulli probability that it is not mutated in a given sample is $b_i = \exp(-\rho L_i)$, where $\exp$ is the exponential function. Its mutation probability, $1 - b_i$, follows as an immediate corollary. □

PROOF. The probability that any particular independent position is not mutated is $1 - \rho$, so the probability of no mutations over all $L_i$ bases of the gene is $(1 - \rho)^{L_i}$. Given necessarily small and large values relative to unity of $\rho$ and $L_i$, respectively, the theorem follows directly from asymptotic approximation (Wendl and Barbazuk, 2005). The process is Poissonian for the number of mutations (Feller, 1968). For example, zero and one mutations have probabilities of $\exp(-\rho L_i)$ and $\rho L_i \exp(-\rho L_i)$, respectively, while the probability of two or more mutations is $1 - (1 + \rho L_i) \exp(-\rho L_i)$, a very small number for relevant values of $\rho$ and $L_i$. □

### 2.2 Single–Sample Test

The fact that the individual gene probabilities, $b_1, b_2, \ldots, b_m$, are not generally equivalent points to a significant computational issue: We cannot extract probability masses from a straightforward and very economical application of the Binomial Theorem. In the simplest, yet most naïve approach, we could instead consider all combinations explicitly via expansion. For example, if the random variable $K$ represents the event where exactly $k$ genes are mutated, then for $k = 0$ we have $P_{K=0} = b_1 b_2 b_3 \cdots b_m$, for $k = 1$

$$P_{K=1} = \left[ (1 - b_1) b_2 b_3 \cdots b_m \right] + \left[ b_1 (1 - b_2) b_3 \cdots b_m \right] + \cdots$$

and so forth. To evaluate the $k$–th probability mass in this fashion requires summing $\binom{m}{k}$ products, each having $m$ terms, where $\binom{m}{k}$ is the number of different combinations of $m$ different objects selected $k$ at a time (Feller, 1968). The number of multiplications and additions here will often be infeasible, for example $60 \cdot \binom{60}{10} > 10^{12}$. A more efficient procedure is given by the following expression.

THEOREM 2 (Exact Probability Mass for Single Sample). The probability mass characterizing the number of mutated genes mutated in a sample, $K$, can be expressed in factored form as

$$P_{K=k} = \exp\left(-\rho G\right) \sum_{i_1=1}^{m-k+1} \sum_{i_2=i_1+1}^{m-k+2} \cdots \sum_{i_k=i_{k-1}+1}^{m} \prod_{n=1}^{k} R_{i_n},$$

where $G = L_1 + L_2 + \cdots + L_m$ is the effective overall length of the genes in the test set and $R_i$ is the ratio $(1 - b_i)/b_i$. A special case is $P_{K=0} = \exp\left(-\rho G\right)$. □

PROOF. A straightforward combinatorial argument shows that the $k$–th mass expands as

$$
\begin{aligned}
P_{K=k} = & \left[\left(1 - b_1\right)\left(1 - b_2\right) \cdots \left(1 - b_k\right) b_{k+1}\, b_{k+2}\, \cdots\, b_m\right] \\
& + \cdots + \\
& \left[b_1\, b_2\, \cdots\, b_{m-k}\left(1 - b_{m-k+1}\right) \cdots \left(1 - b_m\right)\right],
\end{aligned}
$$

where there are $\binom{m}{k}$ such product terms. The theorem follows directly by factoring $b_1 b_2 \cdots b_m$ and rearranging the result. □

Although the factored form is cheaper to evaluate than simple expansion (see below), it still has limited ability to scale as $m$ and $k$ become large. One remedy to this problem is an approximation that exploits the mathematical concept of *convolution* (Feller, 1968). Assume the $m$ Bernoulli probabilities can be arranged into subsets, where all the values within each subset are similar (quantified below) to one another. If there are $j$ such subsets, or "bins", we can write the average bin Bernoulli probabilities for no mutations as $\{\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_j\}$, where the "hat" symbol denotes an average. Given our assumption, each of these values should be a reasonable characterization of each gene in its associated bin.

THEOREM 3 (Approximate Probability Mass for Single Sample). In a $j$–bin model, $K$ is the sum of the individual random variables associated with each bin: $K = K_1 + K_2 + \cdots + K_j$. A computationally–efficient form of the probability mass is obtained by convolution for $j \geq 2$ as

$$
\begin{aligned}
P_{K=k} = & \exp\left(-\rho G\right) \cdot \hat{R}_j^k \sum_{i_{j-1}=0}^{k} \binom{\mu_j}{k - i_{j-1}} \left(\frac{\hat{R}_{j-1}}{\hat{R}_j}\right)^{i_{j-1}} \times \\
& \sum_{i_{j-2}=0}^{i_{j-1}} \binom{\mu_{j-1}}{i_{j-1} - i_{j-2}} \left(\frac{\hat{R}_{j-2}}{\hat{R}_{j-1}}\right)^{i_{j-2}} \times \cdots \times \\
& \sum_{i_1=0}^{i_2} \binom{\mu_1}{i_1}\binom{\mu_2}{i_2 - i_1} \left(\frac{\hat{R}_1}{\hat{R}_2}\right)^{i_1},
\end{aligned}
$$

where the numbers of genes in each bin are $\mu_1, \mu_2, \ldots, \mu_j$, respectively, and satisfy the compatibility condition $\mu_1 + \mu_2 + \cdots + \mu_j = m$. It reduces to a simple binomial form (Feller, 1968) for the special case of $j = 1$, i.e. $P_{K=k} = \exp(-\rho G) \cdot \hat{R}_1^k \binom{m}{k}$, where $m = \mu_1$. □

PROOF. Divide the test set into $j$ bins having $\{\mu_1, \mu_2, \ldots, \mu_j\}$ genes, respectively, where $m = \mu_1 + \mu_2 + \cdots + \mu_j$. Assuming the variabilities of the gene sizes in each bin are not too large, the respective average gene lengths, $\{\hat{L}_1, \hat{L}_2, \ldots, \hat{L}_j\}$, and their

corresponding average bin probabilities for mutation $\{1 - \hat{b}_1, 1 - \hat{b}_2, \ldots, 1 - \hat{b}_j\}$ characterize the bins reasonably well. Under these circumstances, the numbers of mutations in each bin, represented by the random variables $\{K_1, K_2, \ldots, K_j\}$, follow a set of $j$ corresponding binomial distributions.

The random mutation variable for the overall test set is $K = K_1 + K_2 + \cdots + K_j$ and this is characterized by the convolution of the individual distributions (Feller, 1968). For $\kappa_1 + \kappa_2 + \cdots + \kappa_j = k$ observations, the convolution can be written

$$
\begin{aligned}
P_{K=k} = & \sum_{i_{j-1}=0}^{k} \sum_{i_{j-2}=0}^{i_{j-1}} \cdots \sum_{i_1=0}^{i_2} \binom{\mu_1}{i_1}\binom{\mu_2}{i_2 - i_1} \cdots \binom{\mu_j}{k - i_{j-1}} \times \\
& \left(1 - \hat{b}_1\right)^{i_1} \left(1 - \hat{b}_2\right)^{i_2 - i_1} \cdots \left(1 - \hat{b}_j\right)^{k - i_{j-1}} \times \\
& \hat{b}_1^{\mu_1 - i_1}\, \hat{b}_2^{\mu_2 - (i_2 - i_1)} \cdots \hat{b}_j^{\mu_j - (k - i_{j-1})}.
\end{aligned}
$$

Factor the product $\prod_{n=1}^{j} \hat{b}_n^{\mu_n}$ and recognize that this expression is simply $\exp(-\rho G)$. The theorem then follows directly by gathering terms of like powers and performing a final factoring. □

Theorems 2 and 3 should suffice for most cases of practical interest and their forms readily lend themselves to efficient recursive implementation (Cormen et al., 1990). However, if $m$ is especially large, then Poisson approximation (Feller, 1968) might also be applied.

COROLLARY 4 (Idealized Poisson Probability Mass). In the limiting case of a very large test set, where $\max(1 - b_1, 1 - b_2, \ldots, 1 - b_m)$ is very small, $P_{K=k}$ is Poisson–distributed with a mean $(1 - b_1) + (1 - b_2) + \cdots + (1 - b_m)$. This simply restates the limiting case of so–called *Poisson trials* (Feller, 1968). □

The above results are readily cast as tests of significance on a single sample. Specifically, the tailed p–value furnishes the probability of obtaining at least $k$ mutations in a given sample genome under the null hypothesis as

$$P_{K \geq k} = \sum_{i=k}^{m} P_{K=i} = 1 - \sum_{i=0}^{k-1} P_{K=i}, \tag{5}$$

where $H_0$ is rejected if $P_{K \geq k}$ is less than a user–chosen significance threshold, $\alpha$. The first expression is obviously more efficient if $k > m/2$, otherwise the second is cheaper.

### 2.3 Integration of Multiple Samples: The "Overall P–Value"

A single genomic sample actually represents just one test of $H_0$ for $\gamma$. Yet, the ability to sequence many genomes in the course of a project is now emerging, effectively enabling *multiple* tests on $H_0$. These multiple bits of information must be reduced in a rigorous way to an "overall p–value" for the pathway. The problem of integrating $n \geq 2$ such p–values is not new (Pearson, 1933; Fisher, 1938; Wallis, 1942; Lancaster, 1949). However, it is also not one for which mathematics yet furnishes a solution that is both exact and numerically efficient when the underlying distributions are discrete, as they are here. We will therefore resort to layering two classical results upon one another: Lancaster's continuity correction (Lancaster, 1949) applied to Fisher's transform (Fisher, 1938).

This combination furnishes reasonable approximations over a broad range.

## 2.4 Algorithm Description

The execution procedure is straightforward. A gene list representing $\gamma$ is constructed directly from any suitable database, e.g. KEGG (Kanehisa et al., 2010). In conjunction with an estimated background mutation rate, this list begets corresponding gene–specific Bernoulli values according to Thm. 1, which are then used to compute probability masses using Thms. 2 and/or 3, which in turn are collected as a significance test via Eq. 5. Each sample represents a single test of $H_0$ for that gene list through its count of observed mutations. P–values for several samples are subsequently combined into a single project–wide probability for that list using Fisher–Lancaster theory (Fisher, 1938; Lancaster, 1949). Multiple testing correction for many gene lists is subsequently applied via standard methods, such as the False Discovery Rate (FDR) calculation (Benjamini and Hochberg, 1995).

## 3 DISCUSSION

The basic idea of examining somatic events in the context of sets of genes using annotated databases is now a cornerstone of cancer genomics (Efroni et al., 2011; Berger et al., 2011). Mutational significance testing will play increasingly important roles as growing sequencing capacities allow for broader and deeper studies. Here, we formally characterize computational cost, approximation accuracy, and power; these are aspects that have all generally been left unexplored for new tests. We also compare our method to some of the other available tests and illustrate its application via two calculations for lung adenocarcinoma.

### 3.1 Computational Effort

Computational requirements for the various $P_{K=k}$ implementations can be systematically assessed to find how each scales with problem size (Supplementary Information). This analysis indicates that, regardless of the size of the test set and choice of implementation, computational cost will be minimal when the number of observed mutations is small. Incidentally, these cases are not typically of much biological interest because they will tend to fall outside standard ranges of statistical significance. However, costs grow at various rates with $k$ (Fig. 1). Although the factored result (Theorem 2) is more economical than brute–force expansion, the CPU requirements of both appear to rise too fast to be practical in many cases. Conversely, effort for the approximate $j$–bin solution (Theorem 3) grows much more slowly, as illustrated for $j = 1, 2, 3$. Note that this method does not depend upon $m$, since each bin behaves binomially, so it will tend to be tractable even for larger test sets.

### 3.2 Approximation Accuracy

The approximation method in Thm. 3 gathers genes into bins and uses the average bin length as a proxy for the individual lengths. The degree of error in this process depends upon the loss of resolution of the individual gene Bernoulli probabilities. For instance, the hypothetical multiset of gene lengths $\{3000, 8500, 8500, 3000\}$ can be partitioned into $\{3000, 3000\}$ and $\{8500, 8500\}$ without loss
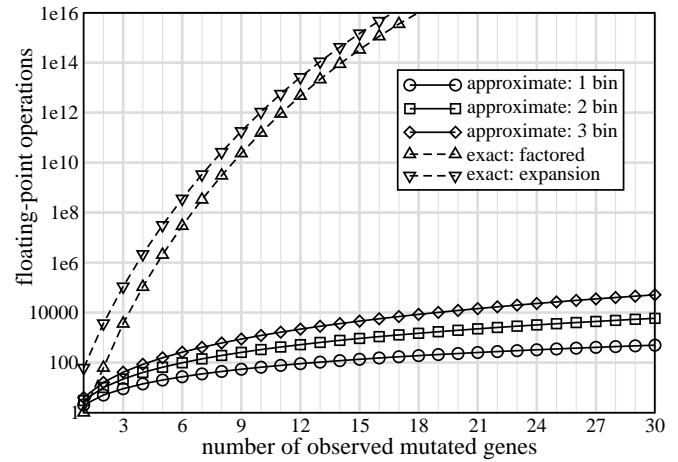


**Fig. 1.** Floating–point operations as a function of the number of observed mutations for a test set of $m = 60$ genes for exact (dashed curves) and approximate (solid curves) solutions. Given Eq. 5, all curves are symmetric about $m/2$, so the plot only shows data up to $k = 30$.

of resolution, i.e. it is exact. However, test sets will generally not contain such a fortuitous list of gene lengths, prompting the question of how to best partition a list of lengths. Optimal clustering in any given instance will produce $j$ subsets, not necessarily with equivalent numbers of elements, but with each subset having minimal size variation among its elements. The general problem for $m$ elements is not trivial (Xu and Wunsch, 2005).

Let us first sort the original lengths $L_1, L_2, \ldots, L_m$ into an ordered list $L_{(1)} \leq L_{(2)} \leq \cdots \leq L_{(m)}$. Optimization then requires determining how many bins should be created and where the boundaries between bins should be placed. While coding lengths of human genes vary from hundreds of nucleotides up to order $10^4$ nucleotides, the background mutation rate is generally not larger than order $10^{-6}$ per nucleotide. These observations suggest the accuracy of using approximation (Thm. 3) would not be a strong function of partitioning because variations in the Bernoulli probabilities would not vary wildly. In other words, sub–optimal partitions should not cause unacceptably large errors in calculated p–values.

We tested this hypothesis in a "naïve partitioning" experiment, where the number of bins is picked *a priori* and then the ordered lengths are divided as equally as possible among these bins. For example, for $j = 2$ one bin would contain all lengths up to $L_{(m/2)}$, with the remaining lengths going to the other bin. Fig. 2 shows results for representative small and large gene sets using 1 bin and 3 bin approximations. Plots are made for plausible background rate bounds of 1 and 3 mutations per Mb. P–values are over–predicted, with errors being sensitive to both the number of bins and the mutation rate. From a hypothesis testing perspective, error is most critical in the neighborhood of $\alpha$. Yet we generally will not have the luxury of knowing its magnitude here *a priori*, or by extension, whether a gene set has been misclassified according to our choice of $\alpha$. Evidently, error is readily controlled by small increases in $j$ without incurring significantly increased computational cost. This behavior will be especially important in two regards: for controlling
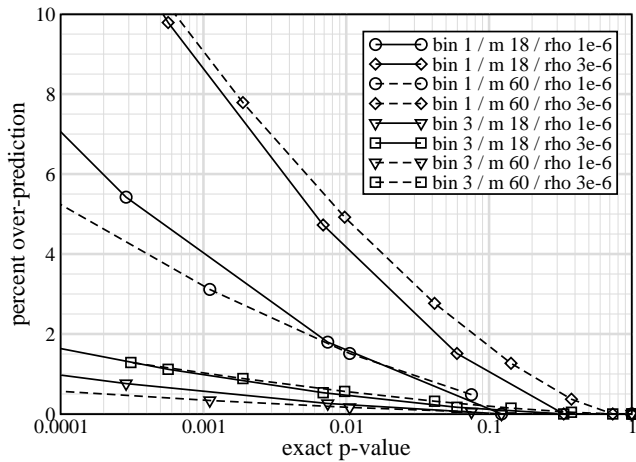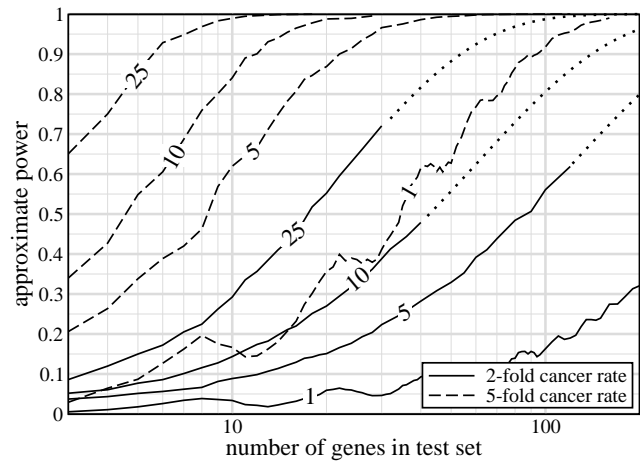
**Fig. 2.** Percent over–prediction of p–values from representative small ($m = 18$, solid curves) and large ($m = 60$, dashed curves) gene sets. Four scenarios are considered: $j = 1$ bin with background mutation rate of $\rho = 1$ per Mb (circles), $j = 1$ and $\rho = 3$ (diamonds), $j = 3$ and $\rho = 1$ (triangles), and $j = 3$ and $\rho = 3$ (squares). Test sets were generated with randomly selected lengths between 200 and 15,000 nucleotides.

**Fig. 3.** Estimated statistical power as a function of test set size for $\alpha = 1\%$. Solid and dashed curves represent assumed cancer mutation rates of two–fold and five–fold higher than the background rate ($\rho = 3$ per Mb), respectively. Dotted curves denote extrapolation beyond computational limits of the 2–fold results based on least–squares fitting (Supplemental Information). All calculations were made using the 3–bin approximate solution on randomly–generated test sets having gene lengths between 200 bp and 15 kb. Each datum indicates the average of 100 such trials.

the error contribution of any "outlier" genes having unusually long or short lengths, and for the "matrix problem" of testing many hypotheses using many genomes, where substantially lower adjusted values of $\alpha$ will be required (Benjamini and Hochberg, 1995). Note that Fig. 2 results are *simulated* in the sense that the gene lengths were chosen randomly. Errors realized in practice could be less if size variance is correspondingly lower. A good general strategy may be to always use at least 3–bin approximation in conjunction with naïve partitioning.

There is necessarily a second level of approximation in combining the sample–specific p–values from many genome samples into a single, project–wide value. These errors are not readily controlled at present because the fundamental mathematical theory underlying combined discrete probabilities remains incomplete. Moreover, obtaining any reliable assessment against true population–based probability values, i.e. via exact p–values and their subsequent exact "brute–force" combination, is computationally infeasible for realistic scenarios. It is important to observe that all tests leveraging data from multiple genomes will be faced with some form of this problem, though none evidently resolve, acknowledge, or perhaps even recognize it. The implications are substantial, as we demonstrate below when comparing to so–called pooled statistical methods.

### 3.3 Statistical Power Analysis

Power analysis is useful for characterizing the minimum conditions for which an effect would be reasonably detectable (Sokal and Rohlf, 1981). Recall, $H_0$ is based on the premise of a random underlying mutation rate $\rho$. Mutations should be more frequent in true cancer–related sets, implying we could estimate power based on the simple alternative hypothesis, $H_1$, of a process characterized by some elevated, cancer–specific mutation rate, $\overline{\rho}$ (Parmigiani et al., 2008). There is a necessary degree of speculation here, as we have

no reliable information regarding such a rate. It would presumably vary by tumor stage and grade, by cancer type, by individual, etc., so a given value would not be representative of cancer in any general sense. This echoes earlier points regarding the difficulties of accurately assessing power (Tarca et al., 2009). Let us simply examine this issue on the basis of a plausible lower bound, $\overline{\rho}/\rho = 2$, and a perhaps conservative upper bound, $\overline{\rho}/\rho = 5$, respectively. Calculation methodology is detailed in Supplemental Information.

Fig. 3 shows power curves at $\alpha = 1\%$ for several sample sizes, where $\rho$ is taken as 3 mutations per Mb. If differences in cancer and background rates are on the order of only two–fold, small pathways will remain undetectable unless sample size is extremely large. For example, a 10–gene pathway nets only about 30% power for 25 sequenced genomes. Conversely, the value jumps to almost 100% for a five–fold difference in mutation rates. Discovery of small pathways is clearly very sensitive to the true cancer mutation rate, which requires better characterization to make suitably accurate predictions. For larger gene sets, e.g. $m \geq 100$, extrapolation (Supplemental Information) suggests that power will be acceptable regardless of the cancer mutation rate. For example, sequencing 25 genomes is now basically feasible using next–generation instrumentation and this sample size puts power near 100%, even for fairly small differences between the cancer and background mutation rates.

### 3.4 Comparison to Other Tests

PathScan is admittedly not the first pathway test to be developed or applied for cancer analysis. So–called "pooled techniques" have already been used for quite some time. These procedures simply combine total mutations and total mutable positions in a gene set into single respective tallies, calculating significance directly from, for example, Fisher's test (mutation rate within pathway versus rate
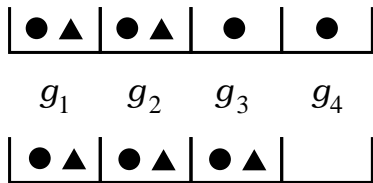
**Fig. 4.** Two mutation scenarios for $(n, m, b, k) = (2, 4, 0.5, 6)$. The top panel represents a $4 + 2$ configuration, i.e. all 4 genes mutated in one sample (circles) and only 2 genes mutated in the other sample (triangles), while the bottom panel is a $3 + 3$ configuration. Pooled statistics are unable to distinguish between these two scenarios, even though their significance values are appreciably different, i.e. $\approx 0.047$ for the top panel and $\approx 0.063$ for the bottom (Wallis, 1942).

outside of pathway) or binomial or Poisson distributions (observed mutation count in light of an estimated background rate). The Group–CaMP test (Table 1) is perhaps the most well–known of these tally methods (Lin et al., 2007). This elementary class of tests harbors a critical liability in the form of significant information loss that necessarily follows from discarding both the distribution of gene lengths in $\gamma$ and the distribution of mutations among samples. While the implications of the former are readily understood in terms of differing gene mutation probabilities (Thm. 1), the latter aspect is less apparent. Consider the following.

The essential problem is that simple tallies cannot distinguish between a few genes having multiple mutations versus many genes having only a few mutations apiece in a group of samples (Fig. 4). Let us borrow a common, but elementary example from the statistics literature (Wallis, 1942; Lancaster, 1949) to illustrate this point, i.e. $(n, m, b) = (2, 4, 0.5)$. Here, each gene has an equal mutation probability. Binomial pooling reduces this problem to a simple tallying scenario having a maximum $n \cdot m = 8$ potential mutations, where probability masses are $P_{K=k} = \binom{8}{k}/256$. For example, for $k = 6$, calculations return $P_{K \geq 6} \approx 0.145$. However, pooling is not actually able to distinguish differences in how mutations could be distributed among the samples. There are two possibilities here for $k = 6$: four mutations in one sample and two in the other, or three in each sample (Fig. 4), with the latter being about a third more probable.

This example has been solved exactly via enumeration (Wallis, 1942), from which we find the true p–value $P_{K \geq 6} \approx 0.184$. The explanation for the perhaps surprising difference is that there are actually several configurations having *fewer* than six mutations, which are nevertheless more significant than the $3 + 3$ configuration. These cases, $0 + 4$ and $1 + 4$, are necessarily omitted from the pooling calculation because of its loss of resolution. Combinatorial considerations indicate that such "out–of–rank" mutation probabilities multiply enormously as the numbers of genes and samples increase, implying increasingly large errors in the resulting p–values. Our opinion in light of this observation is that simple statistical pooling methods are no longer tenable.

### 3.5 Example Calculation

We applied PathScan to the Tumor Sequencing Project (TSP) lung adenocarcinoma gene set data in order to extend the calculations originally performed by Ding et al. (2008) to find significantly mutated KEGG pathways (Kanehisa et al., 2010). The TSP study used two tally–based tests: a binomial model based on a background rate of 3 mutations per Mb and a Fisher exact test of gene mutations within the subject pathway versus those outside the pathway. Our results are compared to these computations in Supplementary Table 1. The following cases have been redacted from this table: (1) known cancer pathways whose mutation lists are invariably dominated by a few known cancer genes, especially *TP53*, *KRAS*, and *EGFR*, (2) other pathways whose mutations likewise reside in just a single gene, (3) pathways having only a single mutation, and (4) pathways not appearing in the original TSP study. In all, 129 pathways were examined and we based our multiple–testing correction (Benjamini and Hochberg, 1995) on this figure.

Not surprisingly, our calculations recapitulate the same pathways found to be highly significant by the two TSP tests. Their p–values differ from one another by orders of magnitude and both differ similarly from our own results, which are generally much less extreme. Yet, all of these differences are vastly outweighed by the extent to which each p–value surpasses a standard 1% threshold. Significant members include the signaling pathways MAPK and mTOR.

The more relevant cases for our assessment purposes are ones for which the original TSP tests were *prima facie* inconclusive, i.e. where the two calculations disagreed. Importantly, the extent of these disagreements is always several orders of magnitude in the p–value. In other words, the ambiguity in such cases is not simply a result of how FDR is chosen, but instead reflects the inherent problems of tally–based tests we described above. For example, the binomial TSP test counted taste transduction (hsa04742) and Alzheimers disease (hsa05010) as significant, while the Fisher TSP test did not. The converse list includes Toll–like receptor signaling (hsa04620), Jak–STAT signaling (hsa04630), and leukocyte transendothelial migration (hsa04670). PathScan concludes none of these pathways is significant.

On the other hand, PathScan finds several of the previously inconclusive pathways to actually be significant, the biologically most interesting example being focal adhesion (hsa04510). Focal adhesions are large protein complexes linking the cell cytoskeleton with the extracellular matrix. They transmit regulatory signals affecting many cellular processes including motility, proliferation, differentiation, regulation of gene expression and cell survival. These functions immediately imply various possible physical relevancies to cancer. Moreover, this pathway has been found to be significantly affected in gene expression studies on prostate cancer (Huang and Chow, 2007), ovarian cancer (Crijns et al., 2009) and proliferative breast lesions (Emery et al., 2009).

We went a step further, expanding calculations for the mutated TSP gene list to 6 databases: BioCarta (Nishimura, 2001), KEGG (Kanehisa et al., 2010), PID (Schaefer et al., 2009), Pfam (Bateman et al., 2000), Reactome (Joshi-Tope et al., 2005), and SMART (Letunic et al., 2009). These resources collectively furnish 988 tests of individual gene groupings and we use this figure for our FDR correction. Table 2 shows 25 candidates that remain after a redaction process similar to that described above. In other words, these groupings all have both a significant FDR and a mutation list containing a good variety of gene hits. In addition to the KEGG focal adhesion pathway, there are several other notable hits that we highlight here.

**Table 2.** Significant Lung Adenocarcinoma Groupings from 6 Databases

| # | database | pathway description | FDR |
|---|----------|---------------------|-----|
| 1 | KEGG | hsa04010: MAPK signaling | 3.0e-42 |
| 2 | Pfam | PF07714: Pkinase Tyr | 5.9e-26 |
| 3 | SMART | SM00219: TyrKc | 2.0e-25 |
| 4 | Reactome | REACT 18266: Axon guidance | 1.8e-18 |
| 5 | KEGG | hsa04012: ErbB signaling | 6.5e-18 |
| 6 | KEGG | hsa04020: calcium signaling | 1.0e-12 |
| 7 | Pfam | PF07679: I-set | 3.8e-12 |
| 8 | Reactome | REACT 11061: signalling by *NGF* | 1.1e-11 |
| 9 | KEGG | hsa04144: endocytosis | 3.2e-10 |
| 10 | SMART | SM00408: IGc2 | 3.0e-09 |
| 11 | PID | regulation of telomerase | 3.5e-09 |
| 12 | KEGG | hsa04060: cytokine interaction | 5.4e-09 |
| 13 | KEGG | hsa04510: focal adhesion | 1.8e-08 |
| 14 | SMART | SM00060: FN3 | 7.8e-07 |
| 15 | Pfam | PF00041: fn3 | 7.8e-07 |
| 16 | BioCarta | h_her2Pathway | 8.7e-07 |
| 17 | PID | signaling events mediated by *PTP1B* | 1.7e-06 |
| 18 | PID | Thromboxane A2 receptor signaling | 3.3e-06 |
| 19 | KEGG | hsa04520: adherens junction | 2.0e-05 |
| 20 | PID | endothelins | 2.9e-05 |
| 21 | SMART | SM00409: IG | 1.9e-04 |
| 22 | KEGG | hsa04150: mTOR signaling | 3.6e-04 |
| 23 | SMART | SM00220: S_TKc | 2.8e-03 |
| 24 | PID | EPHA forward signaling | 0.008 |
| 25 | BioCarta | h_no1Pathway | 0.0094 |

The ErbB signaling pathway (Table 2, hit 5) is activated by extracellular growth factor binding to one of four structurally related receptor tyrosine kinases, EGFR, ERBB2, ERBB3 or ERBB4. Excessive ErbB signaling has been implicated in the development of a wide variety of solid tumors (Hynes and MacDonald, 2009). The large number of mutations in *ERBB4*, *EGFR* as well as in downstream genes such as *KRAS*, *PAK3*, and *PIK3CG* contribute to the highly significant p–value calculated by PathScan for this pathway. Replicative senescence, growth arrest caused by progressive shortening of telomeres during cell division, is thought to be bypassed in most tumors (Shay and Wright, 2002). The telomerase regulation pathway (hit 11) is significantly affected in the samples surveyed, showing mutations in pathway genes such as *EGFR*, *ATM*, *TERT*, and others. Notably, *TERT* has been found to be significantly amplified in lung adenocarcinoma (Ding et al., 2008). The EPHA pathway (hit 24) plays various roles in vertebrate and invertebrate development by regulating cell position and morphology. Disregulated EPHA signaling has been associated with cancers from breast, colon, prostate, and esophagus and the Eph receptors are a promising drug target (Pasquale, 2010). In addition to mutations in the receptors *EPHA1* through *EPHA7*, mutations in pathway genes such as *FYN* and *HCK* were identified in lung adenocarcinoma.

## 4 CONCLUSION

The statistical testing spectrum for cancer DNA sequence is growing rapidly. We have proposed a new method that considers the variable Bernoulli probabilities for differing gene sizes under the null hypothesis and systematically treats the combination of sample–specific p–values in order to obtain a population–based value for a large set of samples. In short, this procedure retains several important pieces of information that existing models discard. Moreover, the method accounts for these factors in a way that does not add significant computational liabilities by using the mathematical concepts of convolution and Fisher–Lancaster Theory.

The model is easily extended to more general scenarios. For example, growing bodies of data will allow increasingly accurate assignments of gene–specific background mutation rates, $\rho_i$. Yet, because we already assume gene–specific Bernoulli values, it is trivial to generalize Thm. 1 as $b_i = \exp\left(-\rho_i L_i\right)$ without incurring any net increase in CPU cost. This property will also permit integration of other phenomena, including copy number changes, structural variation, and methylation and expression changes, should it become possible in the future to assign meaningful Bernoulli values to such events. This last aspect is especially relevant. The cancers are a complex family of diseases and it will be important to broaden investigations to integrate all the types of aberrations that could be linked to a specific phenotype.

The integration problem is actually just one part of a broader research program of pathway analysis. For example, there remains no conclusive method to differentiate the action of a gene from that of a pathway in those cases where the mutation list is dominated by one gene. Although we have shown that PathScan suffers vastly less from this phenomenon than other tests, it does not fully solve this problem. Development of supplemental "exclusionary" tests that specifically examine distributions of mutations among member genes may be necessary. Moreover, models do not yet systematically account for relationships or conditioning that may exist between specific mutations in a network sense, i.e. considering the position and role of a mutated gene within its pathway, multiple gene functions, etc.

PathScan is applicable to any set of genes, $\gamma$, no matter how constructed, meaning it is useful both with pathway databases, as we have shown here, and in *de novo* network–building methods that use interaction databases. The latter must ultimately evaluate network significance in the context of the associated somatic events and often still resort to elementary tests (Glaab et al., 2010). Irrespective of method, any calculation is necessarily limited by whatever databases it uses (Vandin et al., 2010; Cerami et al., 2010). However, because the collective wealth of stored information continues to increase at a remarkable rate (Kanehisa et al., 2010), such concerns should diminish over time.

These observations all suggest that future methods will necessarily become more sophisticated and increasingly focused on the deeper aspects of cancer genomic analysis. We feel that PathScan represents an initial, though deliberate step in that direction.

## REFERENCES

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. (2000) The Pfam protein families database. *Nucleic Acids Research*, **28**, 263–266.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Berger, M. F., Lawrence, M. S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A. Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C. et al. (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214–220.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S. et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*, **104**, 20007–20012.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101–133.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S. and Sander, C. (2010) Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, **5**, paper no. e8918.

Cormen, T. H., Leiserson, C. E. and Rivest, R. L. (1990) *Introduction to Algorithms*. MIT Press, Cambridge MA.

Crijns, A. P. G., Fehrmann, R. S. N., de Jong, S., Gerbens, F., Meersma, G. J., Klip, H. G., Hollema, H., Hofstra, R. M. W., Meerman, G. J. T., de Vries, E. G. E. et al. (2009) Survival–related profile, pathways, and transcription factors in ovarian cancer. *PLoS Medicine*, **6**, 181–193.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L. et al. (2010) Genome remodelling in a basal–like breast cancer metastasis and xenograft. *Nature*, **464**, 999–1005.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B. et al. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.

Efroni, S., Ben-Hamo, R., Edmonson, M., Greenblum, S., Schaefer, C. F. and Buetow, K. H. (2011) Detecting cancer gene networks characterized by recurrent genomic alterations in a population. *PLoS ONE*, **6**, paper no. e14437.

Emery, L. A., Tripathi, A., King, C., Kavanah, M., Mendez, J., Stone, M. D., de las Morenas, A., Sebastiani, P. and Rosenberg, C. L. (2009) Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. *American Journal of Pathology*, **175**, 1292–1302.

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, New York NY, 3rd edition.

Fisher, R. A. (1938) *Statistical Methods for Research Workers*. Oliver and Boyd, London, 7th edition.

Forrest, W. F. and Cavet, G. (2007) Comment on "the consensus coding sequences of human breast and colorectal cancers". *Science*, **317**, 1500.

Getz, G., Höfling, H., Mesirov, J. P., Golub, T. R., Meyerson, M., Tibshirani, R. and Lander, E. S. (2007) Comment on "the consensus coding sequences of human breast and colorectal cancers". *Science*, **317**, 1500.

Glaab, E., Baudot, A., Krasnogor, N. and Valencia, A. (2010) Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics*, **11**, article no. 597.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Edkins, S. et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.

Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. and Easton, D. F. (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, **173**, 2187–2198.

Huang, D. and Chow, T. W. S. (2007) Indentifying the biologically relevant gene categories based on gene expression and biological data: An example on prostate cancer. *Bioinformatics*, **23**, 1503–1510.

Hynes, N. E. and MacDonald, G. (2009) ErbB receptors and signaling pathways in cancer. *Current Opinion in Cell Biology*, **21**, 177–184.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, E., Jasal, B., Gopinath, G. R., Wu, G. R., Matthews, L. et al. (2005) Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, **33**, D428–D432.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, **38**, D355–D360.

Lancaster, H. O. (1949) The combination of probabilities arising from data in discrete distributions. *Biometrika*, **36**, 370–382.

Letunic, I., Doerks, T. and Bork, P. (2009) SMART 6: Recent updates and new developments. *Nucleic Acids Research*, **37**, D229–D232.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M. et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66–72.

Lin, J., Gan, C. M., Zhang, X., Jones, S., Sjöblom, T., Wood, L. D., Parsons, D. W., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. et al. (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Research*, **17**, 1304–1318.

Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D. et al. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *New England Journal of Medicine*, **361**, 1058–1066.

Nishimura, D. (2001) BioCarta. *Biotech Software and Internet Report*, **2**, 117–120.

Parmigiani, G., Chen, S. and Velculescu, V. E. (2008) TRAB: Testing whether mutation frequencies are above an unknown background. *Statistical Applications in Genetics and Molecular Biology*, **7**, paper no. 11.

Parmigiani, G., Lin, J., Boca, S. M., Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Barber, T., Buckhaults, P., Markowitz, S. D. et al. (2007) Response to comments on "the consensus coding sequences of human breast and colorectal cancers". *Science*, **317**, 1500.

Pasquale, E. B. (2010) Eph receptors and ephrins in cancer: Bidirectional signalling and beyond. *Nature Reviews Cancer*, **10**, 165–180.

Pearson, K. (1933) On a method of determining whether a sample of size *n* supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, **25**, 379–410.

Rubin, A. F. and Green, P. (2007) Comment on "the consensus coding sequences of human breast and colorectal cancers". *Science*, **317**, 1500.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K. H. (2009) PID: The Pathway Interaction Database. *Nucleic Acids Research*, **37**, D674–D679.

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J. et al. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.

Shay, J. W. and Wright, W. E. (2002) Telomerase: A target for cancer therapeutics. *Cancer Cell*, **2**, 257–265.

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N. et al. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.

Sokal, R. R. and Rohlf, F. J. (1981) *Biometry*. W. H. Freeman and Company, New York NY, 2nd edition.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C. et al. (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature Genetics*, **37**, 590–592.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., Kim, C. J., Kusanovic, J. P. and Romero, R. (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.

Vandin, F., Upfal, E. and Raphael, B. J. (2010) Algorithms for detecting significantly mutated pathways in cancer. *Lecture Notes in Computer Science*, **6044**, 506–521.

Vogelstein, B. and Kinzler, K. W. (2004) Cancer genes and the pathways they control. *Nature Medicine*, **10**, 789–799.

Wallis, W. A. (1942) Compounding probabilities from independent significance tests. *Econometrica*, **10**, 229–248.

Wang, T. L., Rago, C., Silliman, N., Ptak, J., Markowitz, S., Willson, J. K. V., Parmigiani, G., Kinzler, K. W., Vogelstein, B. and Velculescu, V. E. (2002) Prevalence of somatic alterations in the colorectal cancer cell genome. *Proceedings of the National Academy of Sciences*, **99**, 3076–3080.

Wendl, M. C. and Barbazuk, W. B. (2005) Extension of Lander–Waterman theory for sequencing filtered DNA libraries. *BMC Bioinformatics*, **6**, article no. 245.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J. et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.

Xu, R. and Wunsch, D. (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, **16**, 645–678.