Molecular Evolutionary Analysis of Cancer Cell Lines

Yan Zhang, Michael J. Italia, Kurt R. Auger, Wendy S. Halsey, Stephanie F. Van Horn, Ganesh M. Sathe, Michal Magid-Slav, James R. Brown, and Joanna D. Holbrook

Abstract

With genome-wide cancer studies producing large DNA sequence data sets, novel computational approaches toward better understanding the role of mutations in tumor survival and proliferation are greatly needed. Tumors are widely viewed to be influenced by Darwinian processes, yet molecular evolutionary analysis, invaluable in other DNA sequence studies, has seen little application in cancer biology. Here, we describe the phylogenetic analysis of 353 cancer cell lines based on multiple sequence alignments of 3,252 nucleotides and 1,170 amino acids built from the concatenation of variant codons and residues across 494 and 523 genes, respectively. Reconstructed phylogenetic trees cluster cell lines by shared DNA variant patterns rather than cancer tissue type, suggesting that tumors originating from diverse histologies have similar oncogenic pathways. A well-supported clade of 91 cancer cell lines representing multiple tumor types also had significantly different gene expression profiles from the remaining cell lines according to statistical analyses of mRNA microarray data. This suggests that phylogenetic clustering of tumor cell lines based on DNA variants might reflect functional similarities in cellular pathways. Positive selection analysis revealed specific DNA variants that might be potential driver mutations. Our study shows the potential role of molecular evolutionary analyses in tumor classification and the development of novel anticancer strategies.

Introduction

The advent of lower-cost, higher-throughput DNA sequencing technologies is ushering in a new era of cancer genomics or oncogenomics (1–4). Recent large-scale mutational analyses of cancer cell lines and clinical samples have revealed complex and highly heterogeneous variation even among tumors derived from the same tissue type (5). Such findings have important implications for the clinical treatment of cancer patients. For example, a large-scale sequencing project of cDNAs from 90 tyrosine kinase genes across 254 cell lines found that colon cancer cell lines had similar mutations in the epidermal growth factor receptor kinase domain that rendered non–small cell lung carcinoma patients responsive to the inhibitor gefitinib, thus suggesting another potential indication for this anticancer drug (6). Although the cost of generation of DNA sequence data has greatly decreased, these data sets present formidable bioinformatic challenges, creating a critical need for the development of new analytic approaches to help understand the dependence of the cancer cell on mutations and the relationships among different tumor types.

One significant objective of molecular cancer genomic studies is to attempt to distinguish driver mutations responsible for tumor proliferation and survival from coincidental passenger mutations resulting from relaxed DNA repair and replication fidelity (5). Another frequent study goal is the classification of tumors based on various molecular data, including DNA variants, transcriptional profiling, epigenetic signatures, or combinations thereof. Interestingly, although tumorigenesis has been viewed as an evolutionary process influenced by natural selection processes (7), there has been little application of molecular evolutionary approaches in the analysis of cancer mutations and understanding the relationships of diverse tumors at the DNA sequence level.

Here, we describe the evolutionary relationships of several hundred cancer cell lines and characterize potential mutational patterns using both phylogenetic and positive selection analytic approaches. We owe much of our current understanding of cancer biology to the study of established cell lines, and it is expected that further functional validation of discoveries from the oncogene will rely on many of the same cell lines. For example, the NCI-60 panel of cell lines has been extensively characterized at both the phenotypic and the genotypic level (8, 9). Recently, Lorenzi et al. (10) reported on DNA fingerprinting of the NCI-60 cell line panel, which
suggested that a few cell lines might have common donor origins. However, phylogenetic classification based on DNA mutational patterns of any collection of cell lines has not been previously reported. The basis for any molecular evolutionary analysis is a consistent and biologically relevant multiple sequence alignment of nucleotide or protein sequences from the relevant taxa. Although larger-scale genome-level rearrangements are often associated with tumorigenesis, we focused on single nucleotide synonymous and non-synonymous substitutions or point mutations as the most tractable genetic variant for phylogenetic analyses. Point mutations are highly important in the modulation of many different tumorigenic pathways, by either activating or inhibiting enzymatic activity of specific proteins, thereby causing clinical resistance to new cancer drugs targeting specific proteins such as kinases (11, 12). Because for any particular tumor gene only one or a few DNA variants may occur, we constructed concatenated multiple sequence alignments composed of DNA variant codons from several hundred genes collected across hundreds of tumor cell lines. Phylogenetic analysis of concatenated data sets has been used previously to study other complex evolutionary questions, such as the origin of eukaryotes (13) and the universal tree of life (14). To our knowledge, this is the first application of rigorous phylogenetic analyses to cancer mutation data sets.

Figure 1. Flowchart of cancer cell line gene variant analysis. In-house sequencing project determined a total of 2,777 variant nucleotides (compared with GenBank RefSeq) in 55 genes across 353 human tumor cell lines. Additional mutations for those cell lines were obtained from two public data sources: COSMIC (nucleotides; ref. 15) and Tykiva (amino acids only; ref. 6). A single multiple sequence alignment was constructed from concatenated mutated codons that differed from RefSeq. An amino acid multiple sequence alignment was also constructed from translated codons and additional amino acid mutations identified in the Tykiva database. Both nucleotide and amino acid sequence alignments were used for subsequent phylogenetic and positive selection analysis (see Materials and Methods).
Materials and Methods

DNA Sequencing

Our core data set for phylogenetic analyses was derived from novel DNA sequencing of exons from 55 selected oncogenes, or genes otherwise implicated in tumor proliferation, from 353 cancer cell lines. Supplementary Table S1 lists the cell lines and their tissue of origin along with the genes sequenced in this study. All cell lines were obtained from either the American Type Culture Collection or other public repositories. Both cDNAs and genomic DNA were sequenced. Total RNA from cancer cell lines was isolated using a modified RNeasy kit (Qiagen, Inc.) and converted into cDNA using the First-Strand cDNA Synthesis kit using oligo(dT) primers (Roche Diagnostics). Genomic DNA was prepared using the Maxwell-16 DNA Purification kit (Promega). For genomic DNA, all exons were PCR amplified by designing oligonucleotide primers within flanking intronic regions. Approximately 2 kb of 5′-untranslated region and 1 kb of 3′-untranslated region areas were also covered by sequencing. PCR primers were tagged with M13 universal forward and reverse sequencing primer sequences. All primers have been tested on Human Genomic DNA (Promega) and BD qPCR Human Reference cDNA, random-primed (BD Biosciences), respectively, before using on cell line samples. PCRs were carried out using HotStarTaq DNA polymerase (Qiagen). DNA was amplified for 35 cycles at 95°C for 20 s, 55°C for 30 s, and 72°C for 45 s, followed by 7-min extension at 72°C.

Before DNA sequencing, all PCR products were purified using AmPure (Agencourt Bioscience Corp.). Direct sequencing of purified PCR products was done with BigDye Terminator Cycle Sequencing kit (v1.1; Applied Biosystems) followed by purification using CleanSeq (Agencourt Bioscience). Sequencing was done using an Applied Biosystems Genetic Analyzer 3730XL. All sequence data were assembled and analyzed using Aligner software (CodonCode Corp.), and sequence variants were confirmed by independent PCR amplifications and sequencing.

Construction of Multiple Sequence Alignments

Figure 1 shows a flow diagram of the DNA sequence analysis pipeline. For the cell lines we sequenced, comparisons to the National Center for Biotechnology Information (NCBI) Human RefSeq (August 2008) revealed 2,777 different nucleotide variations. No distinction was made between germline single nucleotide polymorphisms (SNP) or somatic tissue tumor-specific mutations (herein collectively called “variants”). To augment our sample with additional gene sequences, overlapping cell line collections were identified in two public repositories of cancer mutations. Of the 353 cell lines sequenced, 229 were also found to have nucleotide-level mutations recorded in the database COSMIC (15), a comprehensive source of cancer mutations maintained by the Sanger Centre, from which we extracted an additional 922 mutations across 452 genes. The Tykiva Database (6) of the Bioinformatics Institute of Singapore provides amino acid sequences of cancer cell line mutations (DNA sequences are unavailable). A total of 133 amino acid variants in 59 genes across 60 common cell lines were obtained from the Tykiva Database.

Because each of the databases used different wild-type human reference sequences to determine tumor cell line variants (called here dbRef), we adopted a protocol to standardize variants against a common wild-type human reference. As our standard wild-type or wtRef for new DNA sequences generated in this study as well as sequences imported from cancer mutation databases, we used human gene sequences from the NCBI Reference Sequence collection (August 2008). From the COSMIC database, we extracted information of the exact position of each reported point mutation, its corresponding dbRef type nucleotide, and its position in wtRef sequences. To ensure the correctness of the data extraction, we compared the extracted dbRef nucleotide with the corresponding wtRef nucleotide. Furthermore, we confirmed that the extracted three-nucleotide codons were identical in both wtRef and dbRef sequences. Any differences were investigated and rejected if unresolved. If the wtRef and dbRef codons were identical, then the variant codon or amino acid recorded in that database was retained. For the inclusion of a codon in our multiple sequence alignments, at least one variant had to be observed in one of the cell lines. For heterozygous genes, we used the variant allele to increase phylogenetic signal.

To build the concatenated sequence alignments, we added the variant codons in sequential order such that each aligned nucleotide position was homologous across all cell lines. For those genes with missing data, the wtRef codon was used to build a complete multiple sequence alignment without gaps. As the outgroup, a complete wtRef codon sequence was added to the cell line multiple sequence alignment. This produced a multiple sequence alignment of 3,252 nucleotides for 353 cell lines plus nucleotide wtRef. Before phylogenetic analysis, cell lines with 100% identity in nucleotide sequence were reduced to a single representative (see Results and Discussion for a list of identical cell lines). Thus, phylogenetic analyses were done on the distinct sequences from 321 cell lines, including the wtRef sequence.

A similar approach was used to construct the amino acid sequence alignment with some additional steps. First, all variant codon positions from our DNA sequence data as well as that of the COSMIC database were translated to the corresponding amino acid. An additional 133 amino acid variants found in 59 genes were added from the Tykiva resource. The wtRef nucleotide sequence was also translated into amino acids. After consolidating cell lines with identical amino acid sequences, the final protein multiple sequence alignment was 1,170 amino acids.
for 292 cell lines, including an amino acid sequence version of wtRef.

Phylogenetic Analysis

For the nucleotide sequence alignments, maximum likelihood (ML) tree topologies were constructed using the software GARLI v0.96 (16, 17). Estimation of rate heterogeneity was done with the $\gamma$ distribution model. Sixty initial runs of GARLI were carried out to ensure similar likelihood scores were reached. Subsequently, another 60 runs were made to evaluate the consistency of tree topologies with improved parameters to increase the intensity of the searches (attachments per taxon = 200, gendethreshfortopopterm = 80,000, number of precollapse reductions = 40, other parameters default). Nucleotide sequence phylogenies were also reconstructed using Bayesian posterior probabilities (BP) as implemented by the software MrBayes v3.0B4 (18, 19). Bayesian analysis also used the $\gamma$-distributed rate model with six discrete rate categories. Markov chains were run for $10^8$ generations, burn-in values were set for $10^7$ generations, and trees were sampled every 100 generations.

We constructed amino acid–based phylogenetic trees using BP and distance neighbor-joining methods. BP was done as described for nucleotide alignments but with the mixed model for the amino acid rate matrix. Neighbor-joining trees were based on pairwise distances between amino acid sequences using the programs NEIGHBOR and PROTDIST (Dayhoff option) of the PHYLIP 3.6 package. All reconstructed phylogenetic trees were visualized using the programs TreeView v1.6.6 (20) and Dendroscope v2.2.2 (21).

mRNA Expression and Pathway Analysis

To determine the relative differences in mRNA expression between clade A (see Results and Discussion) and other cell lines, we analyzed mRNA microarray profiles previously generated for these cell lines by GlaxoSmithKline, which are available from the public repository caBIG (22). Transcript profiling data as well as background information for each cell line microarray experiment sample are available online. Because large numbers of peripheral blood cell lines fall outside of clade A, inclusion of peripheral blood cell lines could bias results toward genes that simply differentiate solid tumor from peripheral blood cancer cell lines. Thus, we focused our gene expression analysis exclusively on solid tumor cell lines. For the statistical analysis of gene expression data, we used partial least squares discriminant analysis (PLS-DA; ref. 23) for class comparison (24) to identify genes with the most discriminating probe sets between clade A and non–clade A solid tumor cell lines. PLS-DA has been widely used for biomarker discovery and biological process elucidation, particularly for cancer transcriptomic studies (25–28). PLS-DA analysis was done using SIMCA-P+ 11.5 (Umetrics). Cross-validation was done according to default software settings, except for increasing the number of permutations from 7 to 100. A total of 398 probe sets (PLS-DA gene set) were selected based on their high variable importance for the projection (VIP) values. As a cross-check of PLS-DA results, the data were also analyzed using Student’s $t$ test. The $P$ value was corrected by Benjamini-Hochberg procedure with a false discovery rate of <0.05 as the significance threshold (29). Pathway enrichment analyses were conducted with MetaCore (GeneGo, Inc.; ref. 30).

Selection Pressure Analysis

The branch-site model (31, 32) implemented in the CODEML program from the PAML package (33, 34) was used to test for positive selection. We tested each of the branches on the cell line phylogeny, treating each in turn as the foreground branch, with all the other branches specified as background branches. Likelihood ratio tests were done with the Bonferroni correction for multiple testing (35). The alternative branch-site model has four codon site categories: the first two for sites evolving under purifying selection and neutral selection on all the lineages and the additional two categories for sites under positive selection on the foreground branch. The null model restricts sites on the foreground lineage to be undergoing neutral evolution. Each branch-site model was run thrice. At least two of three replicate runs of each model should converge at or within 0.001 of the same log-likelihood value. Runs that did not converge indicated problems with the data and were rerun until convergence was obtained or else reported as a convergence problem.

Results and Discussion

We constructed an initial data set from an in-house DNA sequencing effort of 55 known oncogenes and tumor suppressor genes from 353 cancer cell lines (Fig. 1). The selected cell lines represent a significant proportion of tumor cell lines commonly used in cancer research, including the National Cancer Institute collection. Multiple tissue sources were represented in this cancer cell line sample, with the top four tumor types being breast ($n = 32$), colon ($n = 26$), lung ($n = 38$), and lymphomas ($n = 88$; classified as peripheral blood in Supplementary Table S1).

Among the cell lines we sequenced, comparisons to Human RefSeq (retrieved from NCBI human genome build 36.3, August 2008) revealed 2,777 different nucleotide variations that were either germline SNPs or somatic tissue tumor-specific mutations, herein collectively called variants. To augment our sample with further gene sequences, overlapping cell line collections were identified in two public repositories of cancer mutations. The Sanger Centre COSMIC (15) resource, a comprehensive database of cancer mutations, had 229 of the 353 cell lines
Figure 2. Phylogenetic tree of tumor cell lines based on DNA sequences. Best ML (ML value = −ln19311.744) tree of 320 unique tumor cell lines outgroup rooted with the human RefSeq (wtRef). Based on a multiple sequence alignment of 3,252 nucleotides derived from the concatenation of variant codons, the tree was reconstructed using the GARLI package (see Materials and Methods; ref. 16). Cell lines are color coded by tumor tissue type as listed, along with legend abbreviations, in Supplementary Table S1. Identical cell lines are listed at terminal nodes separated by commas. Solid red bar indicates the terminal node for the clade A subtree, which was consistently obtained in 60 randomized, replicate ML analyses. Red arc line shows the range of cell lines included in clade A. Also indicated are other nodes supported in 50% to 69% (+) and 70% to 100% (*) replications. Those nodes supported by 0.8 to 1.0 probabilities in Bayesian tree reconstruction using the software MrBayes (19) are marked with “!” Solid blue bar indicates an example of a well-supported subcluster of cell lines from diverse tissues of origin (discussed in text). Supplementary Fig. S1 is a PDF version of the same phylogenetic tree in vertical phylogram format.
from which we extracted an additional 922 mutations. The Tykiva Database (6) provided amino acid sequences with 133 additional variants for those cancer cell lines (DNA sequences are unavailable).

For every gene, codon or amino acid residues that differed from the respective “wild-type” Human RefSeq (herein called wtRef) were included in the nucleotide or protein multiple sequence alignments, respectively. To
build a complete data matrix, wherever the variant codon sequence was unavailable for a particular cell line, the wtRef codon sequence was assumed. Codons were then concatenated into a single multiple sequence alignment composed of 3,252 nucleotides for 353 cell lines. For proteins, the codons were translated from the nucleotide alignment with additional amino acid variants from the Tykiva database to give a final alignment of 1,170 amino acids. Among all cancer cell lines, the proportion of different nucleotides and amino acids ranged from 0% to 7.9% and 0% to 6.6%, respectively.

Six cell lines had nucleotide sequences that were 100% identical to the wtRef sequence. There were another 16 identity groups, ranging from two to six cell lines, several of which included multiple tumors types (Supplementary Table S2). For example, tumor cell lines originating from patients with Hodgkin’s disease (L-428 and RPMI 6666), chronic myeloid leukemia (MEG-01), squamous cell cervical carcinoma (SiHa), breast ductal carcinoma (UACC-812), and retinoblastoma (Y79) all share identical nucleotide sequences. In our cell line sample, MDA-MB-435 and M14 as well as U251 and SNB-19 were also previously reported to have identical DNA fingerprints and thus could have common donor origins (10). For those occurrences of two or more identical cell lines, a single representative was selected for the multiple sequence alignments. The resulting phylogenetic data sets of 321 unique nucleotides and 292 unique protein sequences included the respective nucleotide or amino acid wtRef outgroup (Supplementary Tables S3 and S4, respectively).

Phylogenetic analysis of the cancer cell line data set is challenging because many gene variants were represented in low frequencies in the overall data set. The pattern of nucleotide variation and the large number of cell lines meant certain phylogenetic reconstruction methods were less suitable for determining tree topologies, in part due to their sensitivity to errors involving unequal rates of evolutionary changes among lineages. Therefore, we used the ML method as implemented by the software GARLI (16), which has a rapid algorithm allowing for multiple reiterations to optimize parameters and assess tree topology robustness.

The best phylogenetic tree (lowest ML value) for combined new DNA sequence and public nucleotide data is shown in Fig. 2 (shown in Supplementary Fig. S1 as a vertical phylogram with branch lengths). A similar tree topology was recovered in separate phylogenetic analyses restricted to the new DNA sequences generated in this study alone (data not shown). The most striking aspect of the tree is the overall lack of clustering with respect to tissue of origin or cancer type, which suggests that common nucleotide variants can occur in multiple tumor types. In several clusters, cell lines derived from liquid tumors, such as lymphomas and leukemias, co-occurred with solid tumors, suggesting that both of these broad tumor types might be dependent on common mutations for tumorigenesis and metastasis. For example, one statistically well-supported subcluster (indicated in Fig. 2 by a solid blue bar) is composed of 11 tumor cell lines, including 7 lymphomas (lymphocytes and lymphoblasts), 2 ovarian tumors (adenocarcinoma of endometrium ovarian tissue), and 2 myeloid leukemias (bone marrow).

Extensive DNA sequence heterogeneity in common tumor types sampled from different patients or cell lines has been previously observed in large-scale cancer genome surveys. Such studies have highlighted pathways apparently important for the transformed phenotype. For example, a large-scale gene survey of glioblastomas found higher than expected incidences of ERBB2 mutations, which had been previously thought to be more commonly associated with lung, gastric, and colon cancers (4). Other profiling methods, such as DNA fingerprints, have often been unable to distinguish cell line by tissue of origin (10). Our study shows that phylogenetic analysis can assist in quantifying and visualizing the complex variability of the cancer genome, thus facilitating further exploration of relationships between particular collections of tumor samples.

An exceptional phylogenetic structure in all ML runs was a cluster composed of up to 91 cell lines, which we will refer to as clade A (indicated in Fig. 2 and the best ML scored clade A subtree shown in Fig. 3). The clustering of 83 of these cell lines was consistently supported in >80% of 60 separate ML replicates (Supplementary Table S5). The existence of the clade A node, as well as several other internal nodes, was also significantly supported in BP phylogenetic reconstructions. All main tumor types as well as several rare types are represented in clade A. To determine if the nucleotide tree topology reflected nonsynonymous changes, we also reconstructed a protein-based phylogenetic tree (Supplementary Fig. S2) for all cell lines with amino acid sequences (320 cell lines and 1,170 amino acids). Similar to the nucleotide tree, very few clusters in the protein tree were composed of cell lines with common tissue types. Although some associations among cell lines in the nucleotide tree were not recapitulated in the protein tree, the majority of clade A cell lines clustered together, which suggests that most variants defining this group of cell lines are nonsynonymous changes and thereby could reflect potentially functional changes at the protein level.

As shown in the clade A tree, several cell lines, particularly those derived from lung, breast, and colon cancers, have especially long branches reflecting more numerous DNA variants. Higher levels of variation can result from either elevated mutational events in those tumor cell lines or sampling bias toward particular tumor types. We feel that the latter explanation is less likely because gene sequence data were available across a wide spectrum of tumor tissue types. Thus, the longer branches leading to various lung cancers, for example, are reflective of the higher mutation rate in that particular tumor type for the genes studied. This does not exclude the possibility that there are other genes yet to be sequenced, which are commonly mutated in other cancers. Previous studies suggest that early- and late-stage
tumors can accumulate different mutations, which could also account for differences in mutation rates (36, 37). However, in our study, we examined cell lines that, for the most part, likely best model late-stage cancers. Additionally, some methods of phylogenetic tree reconstruction are highly sensitive to taxa with unequal mutation rates and exceptional long branches will tend to cocluster as tree artifacts (38). By using ML methods, which are

<table>
<thead>
<tr>
<th>Gene</th>
<th>Amino acid position</th>
<th>Codon</th>
<th>Variant occurrence</th>
<th>Ratio clade A: total cell lines</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>No. clade A cell lines</td>
<td>Total no. cell lines</td>
</tr>
<tr>
<td>AKT1</td>
<td>242</td>
<td>GAG</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>AURKA</td>
<td>31</td>
<td>TTT</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>AURKA</td>
<td>57</td>
<td>ATT</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>AURKB</td>
<td>295</td>
<td>TCC</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>BAD</td>
<td>96</td>
<td>CGC</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>BIRC5</td>
<td>152</td>
<td>GAG</td>
<td>40</td>
<td>42</td>
</tr>
<tr>
<td>BUB1B</td>
<td>349</td>
<td>CAA</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>BUB1B</td>
<td>388</td>
<td>GCG</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>31</td>
<td>AGC</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>CENPE</td>
<td>81</td>
<td>ACT</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>CENPE</td>
<td>338</td>
<td>GTA</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>CENPE</td>
<td>675</td>
<td>CAG</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>CENPE</td>
<td>1,535</td>
<td>TTT</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>CENPE</td>
<td>1,911</td>
<td>AGC</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>CENPE</td>
<td>2,090</td>
<td>ACG</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>CHFR</td>
<td>254</td>
<td>CCA</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>CHFR</td>
<td>528</td>
<td>TTG</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>CHFR</td>
<td>539</td>
<td>GTG</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>FOXO3A</td>
<td>53</td>
<td>GCC</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>FRAP1</td>
<td>479</td>
<td>GAT</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>FRAP1</td>
<td>999</td>
<td>AAC</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>FRAP1</td>
<td>1,577</td>
<td>GCG</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>FRAP1</td>
<td>1,851</td>
<td>AGC</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>FRAP1</td>
<td>2,303</td>
<td>CTG</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>INCENP</td>
<td>36</td>
<td>GAG</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>INCENP</td>
<td>120</td>
<td>GTT</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>INCENP</td>
<td>506</td>
<td>ATG</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>MAP2K2</td>
<td>64</td>
<td>GTC</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>MAP2K2</td>
<td>151</td>
<td>GAC</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>MAP2K2</td>
<td>220</td>
<td>ATA</td>
<td>51</td>
<td>52</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>1,047</td>
<td>CAT</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>PIK3CD</td>
<td>936</td>
<td>TAC</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>PIK3CG</td>
<td>327</td>
<td>GAT</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>PIK3CG</td>
<td>442</td>
<td>TCC</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>PIK3CG</td>
<td>675</td>
<td>AGC</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>PIK3R1</td>
<td>73</td>
<td>TAC</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>PIK3R1</td>
<td>326</td>
<td>ATG</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>PIK3R2</td>
<td>313</td>
<td>CCC</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>PIK3R2</td>
<td>637</td>
<td>AGT</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>RPS6KB2</td>
<td>269</td>
<td>TTC</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>RPS6KB2</td>
<td>420</td>
<td>GCC</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>TSC1</td>
<td>322</td>
<td>ATG</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>TSC1</td>
<td>445</td>
<td>GAA</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>TSC2</td>
<td>860</td>
<td>TTT</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>TSC2</td>
<td>1,667</td>
<td>GAT</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>TSC2</td>
<td>1,732</td>
<td>TCG</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>
less sensitive to long-branch effects, we have minimized the introduction of this bias in our phylogeny.

Clade A cell lines had several unique variant combinations distinguishable from other cell lines (Table 1). As an example, the gene RPS6KB2 encodes ribosomal protein S6 kinase (70 kDa, polypeptide 2 isoform), which functions downstream of the rapamycin-sensitive mammalian target signaling pathway involved in cell growth and proliferation (39). RPS6KB2 is also one of the most variable genes in our analysis having both synonymous and nonsynonymous nucleotide variants. Clade A included all 58 cell lines with one particular RPS6KB2 variation, A420V. This variant, a known human polymorphism (RS13859), is also suggested to be under positive selection pressure from our analysis (described below).

Another example is the gene BIRC5, a member of the inhibitor of apoptosis gene family, which negatively regulates apoptotic cell death. There are multiple splice variants of BIRC5, including survivin-2B, which is highly expressed during fetal development and in many tumors but not expressed in normal adult tissue (40). The BIRC5 variant E152K, which occurred in 42 cell lines, including 40 clade A members, also tested significant for positive selection. Many nonsynonymous variants that occurred in 10 or more cell lines (across all 353 cell lines) were represented disproportionately higher (>50%) in clade A cell lines. These variants occurred in genes involved with chromosome segregation and cell proliferation (AURKA, BUB1B, INCENP, CENPE, CDKN1A, and CHFR), phosphoinositide 3-kinase/mammalian target of rapamycin pathway genes (TSC1, PIK3CA, PIK3CG, PIK3R1, and PIK3R2), the oncogene KRAS, and the tumor suppressor TP53.

We looked for significant differences in mRNA expression between clade A and other cell lines using data previously released to the caBIG database (22) by GlaxoSmithKline. PLS-DA was used to identify genes with the most discriminating probe sets between clade A and non–clade A solid tumor cell lines. A total of 398 probe sets (PLS-DA gene set) were selected based on their high VIP values. Pathway enrichment analyses on the PLS-DA gene set and clade A commonly variant genes returned many of the same pathways (Table 2). In addition, regulation of translation initiation was significantly enriched in PLS-DA gene expression set ($P = 2.706E−03$), which is a downstream pathway for several variant genes.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>$P$</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apoptosis and survival_BAD phosphorylation</td>
<td>1.32E−09</td>
<td>Variants</td>
</tr>
<tr>
<td>Cell cycle_The metaphase checkpoint</td>
<td>5.76E−10</td>
<td>Variants</td>
</tr>
<tr>
<td>Cell cycle_Chromosome condensation in prometaphase</td>
<td>0.002715</td>
<td>Variants</td>
</tr>
<tr>
<td>Cell cycle_Role of APC in cell cycle regulation</td>
<td>0.01052</td>
<td>Variants</td>
</tr>
<tr>
<td>Signal transduction_AKT signaling</td>
<td>0.03767</td>
<td>Expression</td>
</tr>
<tr>
<td>Development_PIP3 signaling in cardiac myocytes</td>
<td>1.47E−15</td>
<td>Variants</td>
</tr>
<tr>
<td>Development_HGF-RI signaling</td>
<td>7.83E−15</td>
<td>Variants</td>
</tr>
<tr>
<td>G protein signaling_EDG5 signaling</td>
<td>0.0368</td>
<td>Variants</td>
</tr>
<tr>
<td>G protein signaling_G Protein α-12 signaling pathway</td>
<td>0.01146</td>
<td>Variants</td>
</tr>
<tr>
<td>Development_EDG3 signaling pathway</td>
<td>0.00108</td>
<td>Expression</td>
</tr>
<tr>
<td>Development_Mu-type opioid receptor signaling</td>
<td>0.006602</td>
<td>Expression</td>
</tr>
<tr>
<td>Immune response_Role of DAP12 receptors in NK cells</td>
<td>0.03142</td>
<td>Variants</td>
</tr>
<tr>
<td>Neurophysiologic process_HTR1A receptor signaling in neuronal cells</td>
<td>0.007976</td>
<td>Variants</td>
</tr>
</tbody>
</table>

---

Table 2. Pathways enriched in MetaCore for both clade A genes (variants) and PLS-DA gene set (expression) based on hypergeometric $P$ value of <0.05

---

9 https://cabig.nci.nih.gov/caArray_GSKdata/
<table>
<thead>
<tr>
<th>Tissue</th>
<th>Positive selected sites</th>
<th>SNP ID</th>
<th>Gene</th>
<th>Occurrence in cell lines</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Clade A</td>
<td>All</td>
</tr>
<tr>
<td>Brain</td>
<td>I391M</td>
<td>rs3729680</td>
<td>PIK3CA</td>
<td>10 49</td>
</tr>
<tr>
<td></td>
<td>R273H*/Y/C</td>
<td>rs28934576*</td>
<td>TP53</td>
<td>8 24</td>
</tr>
<tr>
<td>Breast</td>
<td>E152K</td>
<td>rs18011376</td>
<td>BUB1B</td>
<td>17 17</td>
</tr>
<tr>
<td></td>
<td>H1047R/L</td>
<td>rs28934576</td>
<td>PIK3CA</td>
<td>9 15</td>
</tr>
<tr>
<td>Colon</td>
<td>E152K</td>
<td>rs11540652*</td>
<td>TP53</td>
<td>7 28</td>
</tr>
<tr>
<td></td>
<td>G12V/C/D/A</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td></td>
<td>G13D</td>
<td>rs1042522*</td>
<td>TP53</td>
<td>7 28</td>
</tr>
<tr>
<td>Bone</td>
<td>R72P*</td>
<td>rs11540652*</td>
<td>TP53</td>
<td>7 28</td>
</tr>
<tr>
<td>Lung</td>
<td>G12V/CS/R/F</td>
<td>rs1801376</td>
<td>BUB1B</td>
<td>17 17</td>
</tr>
<tr>
<td></td>
<td>G131M</td>
<td>rs3729680</td>
<td>PIK3CA</td>
<td>10 49</td>
</tr>
<tr>
<td></td>
<td>R72P</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td>Peripheral tissues</td>
<td>G13D/A/V/L</td>
<td>RS1042522</td>
<td>KRAS</td>
<td>4 9</td>
</tr>
<tr>
<td></td>
<td>G12D/V/N/C</td>
<td>rs11554290†</td>
<td>NRAS</td>
<td>1 5</td>
</tr>
<tr>
<td></td>
<td>Q61L/K/H†</td>
<td>rs11540652*</td>
<td>TP53</td>
<td>7 28</td>
</tr>
<tr>
<td></td>
<td>I391M</td>
<td>rs3729680</td>
<td>PIK3CA</td>
<td>10 49</td>
</tr>
<tr>
<td></td>
<td>R72P*</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td>Prostate</td>
<td>E152K</td>
<td>rs11540652*</td>
<td>TP53</td>
<td>7 28</td>
</tr>
<tr>
<td></td>
<td>Q349R</td>
<td>rs1801376</td>
<td>BUB1B</td>
<td>17 17</td>
</tr>
<tr>
<td></td>
<td>M506T</td>
<td>rs2277283</td>
<td>INCENP</td>
<td>19 19</td>
</tr>
<tr>
<td></td>
<td>I391M</td>
<td>rs3729680</td>
<td>PIK3CA</td>
<td>10 49</td>
</tr>
<tr>
<td></td>
<td>R72P</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td>Kidney</td>
<td>I391M</td>
<td>rs3729680</td>
<td>PIK3CA</td>
<td>10 49</td>
</tr>
<tr>
<td>Skin</td>
<td>V600E/D</td>
<td>rs1154290†</td>
<td>NRAS</td>
<td>4 19</td>
</tr>
<tr>
<td></td>
<td>Q61K/R*</td>
<td>rs17847825</td>
<td>PIK3CG</td>
<td>18 18</td>
</tr>
<tr>
<td></td>
<td>M326I</td>
<td>rs13859</td>
<td>RPS6KB2</td>
<td>58 58</td>
</tr>
<tr>
<td></td>
<td>A420V</td>
<td>rs1011320</td>
<td>PIK3R2</td>
<td>10 10</td>
</tr>
<tr>
<td></td>
<td>P313S</td>
<td>rs1011320</td>
<td>PIK3R2</td>
<td>10 10</td>
</tr>
<tr>
<td>Ovarian</td>
<td>R72P</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td>Pancreas</td>
<td>G12V/D/C</td>
<td>rs1801376</td>
<td>BUB1B</td>
<td>17 17</td>
</tr>
<tr>
<td></td>
<td>M326I</td>
<td>rs3730089</td>
<td>PIK3R1</td>
<td>24 24</td>
</tr>
<tr>
<td></td>
<td>G72P</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td></td>
<td>G72P*</td>
<td>rs11540652*</td>
<td>TP53</td>
<td>7 28</td>
</tr>
<tr>
<td></td>
<td>I255N/T</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td></td>
<td>M322T</td>
<td>rs3729680</td>
<td>PIK3CA</td>
<td>10 49</td>
</tr>
<tr>
<td></td>
<td>G72P</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td></td>
<td>G72P*</td>
<td>rs11540652*</td>
<td>TP53</td>
<td>7 28</td>
</tr>
<tr>
<td>Other</td>
<td>G12V/D/R/C</td>
<td>rs1042522</td>
<td>TP53</td>
<td>34 144</td>
</tr>
<tr>
<td></td>
<td>I391M</td>
<td>rs1801376</td>
<td>BUB1B</td>
<td>17 17</td>
</tr>
</tbody>
</table>

**NOTE:** Based on PAML likelihood scores (see Supplementary Methods File 1), all tumor tissue types, where genes had amino acid sites under significant positive selection (BP > 0.95), are shown. Bladder and liver tumor-derived cell lines were also tested but do not have any significant positive selection sites.

*Positive selection sites, which are also sites of known SNPs and their corresponding SNP ID.

†For NRAS, positive selection site Q61 is a site of a known SNP, but mutations at this site were not of the SNP variant (Q61R).
that they may also confer cancer susceptibility in poly-
cellular growth advantage. Therefore, we hypothesize
normal tissues, although our analysis suggests that they
are known germline SNPs as identified in the NCBI
dbSNP database (Table 3). These SNPs are present in
occurring in 33% of primary malignant melanomas.
activating mutations at V600 position of the
in 73% of primary pancreatic ductal carcinomas and
activating mutations at G12 position of the
that are highly represented in COSMIC (15), such as
iations conferring growth or survival advantage on cancer
The differential mRNA expression of genes in clade A
cell lines relative to other cell lines suggests that phyloge-
etic clustering of tumor cell lines based on DNA variants
might recapitulate functional similarities in cellular path-
ways. We can also exclude the possibility that the genes
commonly mutated in clade A are substantially per-
turbed by other mutational changes (e.g., large-scale
genomic aberrations such as deletions, amplifications,
or rearrangements) in the wild-type cell lines. If this
was the case, we would not expect differential pathway
gene expression patterns between the two groups.

Determining the critical mutations that contribute
to the malignant cellular phenotype is challenging.
Beerenwinkel et al. (41) estimated that colorectal cancers
may carry ~100 nonsynonymous mutations, of which
perhaps as few as 3 may be sufficient for developing can-
cer. These critical mutations are commonly termed driver
mutations. Conversely, those mutations that are ac-
cumulated by the tumor cell but do not seem to confer a
growth advantage are called passenger mutations (these
can occur in an accelerated manner as a consequence of
impaired DNA repair and apoptosis processes). Evolu-
tionary methods may have the capability to discriminate
between passenger and driver mutations (42). In an effort
to highlight putative driver mutations within the phylog-
eny, we classified cell lines by tissue of origin. Nucleotide
alignments for each tissue group were analyzed to detect
evidence for evolution influenced by positive selection
pressures using the PAML algorithms (33, 34). We found
significant evidence for the influence of positive selection
pressure acting on several sites in the concatenated
alignment for most tissue groups.

For those tissue groups that returned a significant test
for positive selection, a Bayesian analysis was used to
identify the sites responsible for the signal with posterior
probability of >0.95. These sites might be driver muta-
ations conferring growth or survival advantage on cancer
cells (Table 3). Included are known oncogenic mutations
that are highly represented in COSMIC (15), such as
activating mutations at G12 position of the KRAS gene
in 73% of primary pancreatic ductal carcinomas and
activating mutations at V600 position of the BRAF gene
occurring in 33% of primary malignant melanomas.

Interestingly, 12 of the 20 positively selected mutations
are known germline SNPs as identified in the NCBI
dbSNP database (Table 3). These SNPs are present in
normal tissues, although our analysis suggests that they
are selected for in cell culture and possibly confer some
cellular growth advantage. Therefore, we hypothesize
that they may also confer cancer susceptibility in poly-
morphic individuals—a hypothesis supported by two
of the SNP variants. The variant Q349R in the gene
BUB1B is associated with chronic lymphocytic leukemia
incidence (43). The P72R polymorphism in the gene TP53
has been shown to affect risk of cervical cancer (44), and a
meta-analysis of all types of cancer estimated a pooled
weak cancer risk in p53-P72 homozygotes (45, 46). The
frequencies of some of variants in the cancer cell lines
are higher than expected for human populations accord-
ing to HapMap. This could be due to the cancer cells hav-
ing somatically mutated to the other allele during cell
line culture or by originating from individuals with
germline SNPs more likely to develop tumors. Further
cautions should be used in the interpretation of positive
selection by the branch-site method used by PAML,
which has been recently challenged as overestimating
sites with potential functional changes (47). Thus, further
sequencing of the positively selected polymorphic
sites in tumors seems warranted to determine the clinical
relevance of these variants.

Our phylogenetic and positive selection analysis
approaches to the study of DNA variants in human
on breast cancer cell lines have several important caveats. First, it
is optimal to have a complete data matrix with sequence
variant data available across all tumor samples. Given
the low frequency of individual gene mutations, char-
acterization of variants across hundreds of genes is
required for robust phylogenetic trees. Fortunately, such
gene-wide DNA sequence data are becoming avail-
able through several large-scale cancer genome projects
now under way. Second, although point mutations are
very important in tumorigenesis, larger-scale DNA aber-
rations are also key mutational events in many cancers.
Potentially, reconstruction of comprehensive phylogenies
of tumor genomes might be possible using combined da-
ta models of both point mutations and genomic aber-
rations as has been done for mixed DNA sequence and
morphologic data sets (48). Third, sequenced genes in
our data set are biased toward signal transduction path-
ways known to be perturbed in cancer. Broader DNA se-
quencing surveys involving additional genes may well
alter the tree topology. Finally, the genetic variation
between cancer cell lines, as well as reflecting variation
in human tumors, could also be the result of germline
differences between the cell line donors and changes
acquired during cell culture.

Here, we show that evolutionary analyses can illumi-
nate commonalities in tumorigenic mechanisms among
diverse cancers—this finding has several important im-
portant implications for future cancer research. First, phyloge-
tic trees provide a direct and visual means for understand-
ing the relationships between tumor types based on
complex data sets as well as a statistical basis for their
classification. Previous results have been reported in tab-
ular formats (i.e., mutated protein kinases; ref. 6), which
are less tractable for identifying tumor type relationships
across large multiple gene variant data sets. Second, tu-
more cell lines are key tools in the preclinical screening of
potential compounds and biological agents for anticancer activity. By knowing the homologous relationships of tumor cell lines based on genome-wide variant data, more rational selection of cell lines can be made for drug screening campaigns. Furthermore, by screening cell lines from different tissues of origin that might cluster together because of similar DNA variation patterns, new indications for novel drugs can be discovered preclinically. Finally, sophisticated DNA-based diagnostic tests to identify specific target gene mutations are being increasingly used to develop clinical treatment regimens for cancer patients (49). As these data are collected across more genes and patients, evolutionary approaches can be used to classify clinical tumor samples based on molecular similarities and potentially guide therapeutic decisions.

References

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
We thank C. Traini, E. Thomas, K. Allen, and A. Hughes for DNA sequencing; D. Zwickl, N.J. Wickett, P.K. Wall, and Jeremy M. Martin for computational consulting; C.W. deAmphilius for support of Y. Zhang; and R. Wooester and three anonymous reviewers for their comments on this manuscript. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 6/11/09; revised 11/11/09; accepted 12/22/09; published OnlineFirst 2/2/10.

290 Mol Cancer Ther 9(2) February 2010 Molecular Cancer Therapeutics

Downloaded from mct.aacrjournals.org on June 24, 2017. © 2010 American Association for Cancer Research.
35. Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 2007;24:1219–28.