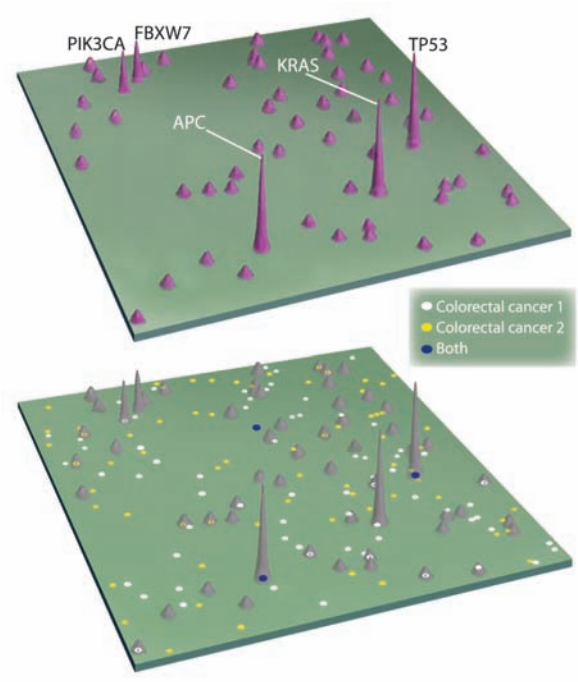
The Genomic Landscapes of Human Breast and Colorectal Cancers

Laura D. Wood,[1*] D. Williams Parsons,[1*] Siân Jones,[1*] Jimmy Lin,[1*] Tobias Sjöblom,[1*†] Rebecca J. Leary,[1] Dong Shen,[1] Simina M. Boca,[1,2] Thomas Barber,[1‡] Janine Ptak,[1] Natalie Silliman,[1] Steve Szabo,[1] Zoltan Dezso,[3] Vadim Ustyanksky,[3] Tatiana Nikolskaya,[3,4] Yuri Nikolsky,[3] Rachel Karchin,[5] Paul A. Wilson,[5] Joshua S. Kaminker,[6] Zemin Zhang,[6] Randal Croshaw,[7] Joseph Willis,[8] Dawn Dawson,[8] Michail Shipitsin,[9] James K. V. Willson,[10] Saraswati Sukumar,[11] Kornelia Polyak,[9] Ben Ho Park,[11] Charit L. Pethiyagoda,[12] P. V. Krishna Pant,[12] Dennis G. Ballinger,[12] Andrew B. Sparks,[12§] James Hartigan,[13] Douglas R. Smith,[13] Erick Suh,[13] Nickolas Papadopoulos,[1] Phillip Buckhaults,[7] Sanford D. Markowitz,[14] Giovanni Parmigiani,[1‖] Kenneth W. Kinzler,[1‖] Victor E. Velculescu,[1‖] Bert Vogelstein[1‖]

AUTHORS' SUMMARY

How many genes are mutated in a human tumor? Answering this question would have seemed like science fiction just a decade ago. However, as a result of advances in technology, we have been able to answer this question in breast and colorectal cancers: There are ~80 DNA mutations that alter amino acids in a typical cancer. Examining the overall distribution of these mutations in different cancers of the same type leads to a new view of cancer genome landscapes: They are composed of a handful of commonly mutated gene "mountains" but are dominated by a much larger number of infrequently mutated gene "hills."

The current study expands upon previous work (*1*) and includes analysis of the sequences of 20,857 transcripts from 18,191 human genes, including the great majority of those that encode proteins. The genes were sequenced in 11 breast and 11 colorectal cancers. Any gene that was mutated in the tumor but not in normal tissue from the same patient was analyzed in 24 additional tumors. Selected genes were further analyzed in another 96 colorectal cancers to better define their mutation frequency and aid subsequent bioinformatic analyses.



A two-dimensional map of genes mutated in colorectal cancers, in which a few gene "mountains" are mutated in a large proportion of tumors while most "hills" are mutated infrequently. The mutations in two individual tumors are indicated on the lower map. Note that most mutations are outside hills or mountains and may be harmless.

Statistical analyses suggested that most of the ~80 mutations in an individual tumor were harmless and that <15 were likely to be responsible for driving the initiation, progression, or maintenance of the tumor. Though the numbers of mutant genes in breast and colorectal cancers were similar, the particular genes that were mutated were quite different, as were the type of mutations found. For example, mutations converting 5′-CpG to 5′-TpG were much more frequent in colorectal than in breast cancers, indicating differences in mutagen exposure or DNA-repair processes.

The mutational landscapes of cancers can be shown on a map on which each gene is represented at a single point (see figure for the landscape for colorectal cancers). The heights of the peaks reflect the mutation frequency of each gene. A few gene "mountains" are mutated in a large proportion of tumors; most genes are mutated in <5% of tumors and are represented as "hills" in the figure. In the lower panel, the mutated genes in two colorectal tumors are indicated by differently colored dots. The mutated genes in the two tumors overlap to only a small extent. These differences are likely to be the basis for the wide variations in tumor behavior and responsiveness to therapy.

Historically, the focus of cancer research has been on the gene mountains, in part because they were the only alterations that could be identified with available technologies. However, our data show that the vast majority of mutations in cancers do not occur in such mountains. This new view of cancer is consistent with the idea that a large number of mutations, each associated with a small fitness advantage, drive tumor progression (*2*). It is the "hills" and not the "mountains" that dominate the cancer genome landscape.

Are these landscapes hopelessly complex? The large number of "hills" actually reflects alterations in a much smaller number of cell signaling pathways. Indeed, pathways rather than individual genes appear to govern the course of tumorigenesis (*3*). Accordingly, we devised methods to classify mutant genes into commonly altered pathways. Disruption of a pathway by mutation in any one of its genetic components would presumably lead to similar changes in growth. The <15 driver mutations in an individual tumor likely reflect alterations in a similar number of pathways.

Sequencing alone cannot definitively determine whether a specific gene "hill" actually contributes to tumor formation. We therefore used various bioinformatic and structural analyses to help determine which were pathogenic. Integration with functional studies will also be essential; indeed, several of the candidate cancer genes identified in our study have been independently implicated in tumorigenesis through functional studies reported by others.

In sum, our results make it clear that it is now "easy" to identify the genetic alterations in cancers on a genome-wide scale. It is much more difficult to elucidate the precise role of these alterations in tumorigenesis. The compendium of genetic changes in individual tumors provides new opportunities for individualized diagnosis and treatment of cancer. Taking advantage of these opportunities is the major challenge for the future.

Summary References

1. T. Sjöblom *et al.*, *Science* **314**, 268 (2006).
2. N. Beerenwinkel *et al.*, *PLoS Comput. Biol.* **3**, e225 (2007).
3. B. Vogelstein, K. W. Kinzler, *Nat. Med.* **10**, 789 (2004).

FULL-LENGTH ARTICLE

Human cancer is caused by the accumulation of mutations in oncogenes and tumor suppressor genes. To catalog the genetic changes that occur during tumorigenesis, we isolated DNA from 11 breast and 11 colorectal tumors and determined the sequences of the genes in the Reference Sequence database in these samples. Based on analysis of exons representing 20,857 transcripts from 18,191 genes, we conclude that the genomic landscapes of breast and colorectal cancers are composed of a handful of commonly mutated gene "mountains" and a much larger number of gene "hills" that are mutated at low frequency. We describe statistical and bioinformatic tools that may help identify mutations with a role in tumorigenesis. These results have implications for understanding the nature and heterogeneity of human cancers and for using personal genomics for tumor diagnosis and therapy.

Discovery of the genes mutated in human cancer has provided key insights into the mechanisms underlying tumorigenesis and has proven useful for the design of a new generation of targeted approaches for clinical intervention (*1*). With the determination of the human genome sequence and improvements in sequencing and bioinformatic technologies, systematic analyses of genetic alterations in human cancers have become possible (*2*–*4*).

Using such large-scale approaches, we recently studied the genomes of breast and colorectal cancers by determining the sequence of the Consensus Coding Sequence (CCDS) genes, a collection of the best-annotated protein-coding genes (*5*). In this study, we have extended these analyses to include examination of all of the Reference Sequence (RefSeq) genes. The RefSeq database is a comprehensive, nonredundant collection of annotated gene sequences that represents a consolidation of gene information from all major gene databases (*6*). The RefSeq database is believed to include the great majority of human gene sequences and represents the gold standard in the field.

Sequencing strategy. The first step in our approach was the design of primers that would permit polymerase chain reaction (PCR)-based amplification and analysis of coding exons in the RefSeq database. Of the 20,857 transcripts in the RefSeq database (representing 18,191 distinct genes), 14,661 transcripts were included in the CCDS set. These CCDS genes were in general not evaluated again; the only exceptions were a small subset in which particular regions of interest had been difficult to amplify and for these, new PCR primers were designed. For the remaining 6196 Refseq transcripts, 125,624 primers were designed and used to amplify the coding exons. The entire list of primers used to amplify the exons of the RefSeq genes (including the CCDS genes) is provided in table S1.

The primers were used to PCR-amplify and sequence the DNA from 11 breast and 11 colorectal cancers, as well as DNA from matched normal tissues of two patients. The samples used for this analysis were the same as those used in the previous study of CCDS genes (*5*). The sequence data from this Discovery Screen were assembled and evaluated using stringent quality criteria (*7*), resulting in successful analysis of 93% of targeted amplicons. We used bioinformatic and experimental strategies to distinguish germline variants and artifacts of PCR or sequencing from true somatic mutations (fig. S1). Genetic alterations found in the two normal samples and those present in SNP databases were removed and sequence traces of the remaining potential alterations were visually inspected to remove false-positive calls in the automated analysis. After these steps, the amplicons of the remaining alterations were re-amplified from the tumor DNA (to ensure reproducibility) and from DNA of matched normal tissue (to remove unannotated germline variants). Finally, the putative somatic mutations were examined "in silico" (by computer analysis) to ensure that the alterations did not occur as a result of mistargeted amplification of related regions of the genome (*7*).

To further evaluate the genes with somatic mutations in the Discovery Screen, we determined their sequence in a Validation Screen of 24 additional samples of the same tumor type in which the mutation was originally identified. Methods similar to those noted above were used to exclude germline variants, PCR and sequencing artifacts, and alterations due to mistargeted amplification of related genomic regions. Amplicons with putative somatic mutations were re-amplified in DNA from the tumor and from matched normal tissues to determine whether the alterations were truly somatic.

Somatic mutations. Combining the data from the current analysis with those previously obtained in CCDS genes, we found that 1718 genes (9.4% of the 18,191 genes analyzed) had at least one nonsilent mutation in either a breast or colorectal cancer (Table 1 and table S3). The great majority of alterations were single-base substitutions (92.7%), with 81.9% resulting in missense changes, 6.5% resulting in stop codons, and 4.3% resulting in alterations of splice sites or untranslated regions immediately adjacent to the start and stop codons (Table 1). The remaining somatic mutations were insertions, deletions, or duplications (7.3%). The mutation spectrum of colorectal cancers differed from that of breast cancers, and these spectra were similar to those observed in the previous CCDS study and in other analyses (*4*, *5*). In this study, we analyzed the nature of the nonsynonymous mutations in more detail and found a very large excess of C to T transitions at 5′-CpG-3′ in colorectal cancers, representing 19 times as many as expected from the representation of 5′-CpG-3′ sites in the coding regions of the genome. Similarly, there was a marked excess of G to C transversions at 5′-GpA-3′ sites in breast cancers, representing 4.5 times as many as expected (*7*).

Passenger mutation rates. The somatic mutations found in cancers are either "drivers" or "passengers" (*4*). Driver mutations are causally involved in the neoplastic process and are positively selected for during tumorigenesis. Passenger mutations provide no positive or negative selective advantage to the tumor but are retained by chance during repeated rounds of cell division and clonal expansion.

We used two independent methods to estimate the passenger mutation rates in the analyzed cancers. First, we evaluated 23.8 Mb of chromosome 8 in 11 colorectal cancer samples similar to those used in the Discovery Screen. This was performed with high-density oligonucleotide microarrays containing every possible single-base pair substitution. The tumors used for this analysis each had only one allele of chromosome 8 [i.e., they showed loss of heterozygosity (LOH)], rendering the detection of sequence alterations sensitive and reliable. A total of 151 somatic mutations were identified in 262 Mb of tumor DNA, and all but one of these were located in noncoding regions. Thus, there were a total of 0.6 noncoding mutations per Mb analyzed (95%

[1]The Ludwig Center for Cancer Genetics and Therapeutics and The Howard Hughes Medical Institute at The Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA. [2]Department of Biostatistics, Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA. [3]GeneGo, St. Joseph, MI 49085, USA. [4]Vavilov Institute of General Genetics, Moscow, Russia. [5]Department of Biomedical Engineering, Institute of Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA. [6]Department of Bioinformatics, Genentech, San Francisco, CA 94080, USA. [7]Department of Pathology and Microbiology, The Center for Colon Cancer Research, and The South Carolina Cancer Center, Division of Basic Research, The University of South Carolina, School of Medicine, Columbia, SC 29229, USA. [8]Department of Pathology and Ireland Cancer Center, Case Western Reserve University and University Hospitals of Cleveland, Cleveland, OH 44106, USA. [9]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA. [10]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center at Dallas, Dallas, TX 75390, USA. [11]The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD 21231, USA. [12]Perlegen Sciences, Mountain View, CA 94043, USA. [13]Agencourt Bioscience Corporation, Beverly, MA 01915, USA. [14]Department of Medicine and Ireland Cancer Center, Case Western Reserve University and University Hospitals of Cleveland, and Howard Hughes Medical Institute, Cleveland, OH 44106, USA.

*These authors contributed equally to this work.
†Present address: Department of Genetics and Pathology, Uppsala University, SE-751 85 Uppsala, Sweden.
‡Present address: Lilly Research Labs, Eli Lilly and Company, Indianapolis, IN 42685, USA.
§Present address: Complete Genomics, Sunnyvale, CA 94085, USA.
‖To whom correspondence should be addressed. E-mail: gp@jhu.edu (G.P.); kinzlke@jhmi.edu (K.W.K.); velculescu@jhmi.edu (V.E.V.); vogelbe@welch.jhu.edu (B.V.)

Table 1. Summary of somatic mutations. UTR, untranslated region. ND, not determined because synonymous mutations were not evaluated in the RefSeq genes analyzed in (*5*).

Tumor type	Screen	Gene set	Mutated genes	Coding changes						Noncoding changes	Total mutations
				Missense	Nonsense	Insertion	Deletion	Duplication	Synonymous	Splice site or UTR	
Colorectal cancers	Discovery	This study	325	237	14	0	8	0	93	12	364
		All RefSeq	848	722	48	4	27	18	ND	30	942
	Validation	This study	88	81	9	1	2	2	30	6	131
		All RefSeq	183	197	34	4	14	5	ND	15	299
Breast cancers	Discovery	This study	460	304	26	2	28	1	131	14	506
		All RefSeq	1137	909	64	5	78	3	ND	53	1243
	Validation	This study	62	52	3	0	3	0	19	2	79
		All RefSeq	167	153	11	2	15	2	ND	7	209

confidence interval: 0.52 to 0.64 mutations/Mb). Because only one copy of chromosome 8 was analyzed in these studies, the noncoding mutation rate per diploid genome was inferred to be 1.2 mutations/Mb. We then performed detailed LOH analyses of the 11 tumors used in the Discovery Screen using 317,503 polymorphisms. An average of 16% of polymorphic alleles showed LOH. It is known from studies of human genetic variation that the frequency of nonsynonymous (amino acid–changing) mutations is approximately half that of mutations in noncoding regions (*8*, *9*). After correcting for LOH and the difference in mutation rates between noncoding and nonsynonymous mutations, these analyses result in an estimated passenger mutation rate of 0.55 nonsynonymous mutations per Mb of tumor DNA in colorectal cancers (*7*). We consider this a minimum estimate as the ratio of mutations in noncoding regions to nonsynonymous mutations in coding regions is likely to be higher in the germ line than in tumors because of greater negative selection for mutations in coding regions in the germ line. Although we have not directly measured mutation rates in noncoding sequences in breast cancers, Stephens *et al*. have estimated that the rate of nonsynonymous mutations in breast cancers is 0.33 per Mb, and we used this as our minimum estimate for this tumor type (*10*).

Estimates of the passenger mutation rates were also obtained through the quantification of synonymous (silent) missense mutations in this study. Because most synonymous changes are expected to be biologically inert and thereby not selected for or against during tumorigenesis, such changes can be used as a tool to estimate passenger mutation rates (*11*). The analysis of synonymous mutations provided two estimates of the nonsynonymous mutation rate (*7*). One estimate was based on the ratio of nonsynonymous to synonymous mutations observed in the human germ line (*8*, *9*). The second estimate was derived by calculating the expected ratio of nonsynonymous to synonymous changes after accounting for codon usage of RefSeq genes and the different mutation spectra observed in colorectal and breast cancers. We considered this estimate to be a maximum because it did not take into account that nonsynonymous mutations that retard cell growth will be selected against during tumorigenesis.

Evaluating mutated genes. The mutational data obtained can be used to identify candidate cancer genes (*CAN*-genes) that are most likely to be drivers and are therefore most worthy of further investigation. In this study, we considered a gene to be a *CAN*-gene if it harbored at least one nonsynonymous mutation in both the Discovery and Validation Screens and if the total number of mutations per nucleotide sequenced exceeded a minimum threshold (*7*). Using these criteria, we identified a total of 280 *CAN*-genes, equally distributed between colorectal and breast cancers (table S4, A and B, respectively). The 280 *CAN*-genes listed in table S4, A and B, included most of the 191 *CAN*-genes identified in Sjöblom *et al*. (*5*) but differed by virtue of the inclusion of 114 new *CAN*-genes identified in the additional 6196 transcripts sequenced, the removal of data from a breast tumor with an abnormally high passenger mutation rate, the use of an experimental rather than statistical definition of *CAN*-genes, and additional evaluation of mutations in samples that had undergone whole-genome amplification (*7*).

It is reasonable to assume that genes that are mutated more frequently than predicted by chance are more likely to be drivers. In this study, we used a more sophisticated version of a metric, called the cancer mutation prevalence (CaMP) score, to rank genes by the number and nature of the mutations observed (table S4, A and B). To assess the likelihood that each of these genes is mutated at a frequency higher than the passenger mutation rate, we devised a method based on Empirical Bayes simulations (*7*). Though the likelihoods depend on the passenger rates (table S4, A and B), the rankings of the genes by CaMP scores are similar regardless of the assumed passenger mutation rates (rank correlations > 0.9). CaMP scores thereby provide priorities for future studies that are independent of many of the assumptions required to calculate passenger probabilities.

To determine the mutation prevalence of a subset of *CAN*-genes with more precision, we analyzed 40 *CAN*-genes in a separate cohort of 96 patients with colorectal cancers (*7*). The genes chosen were in biological pathways of interest to our groups and included those ranked 1st to 119th by CaMP scores. Colorectal cancers, rather than breast tumors, were chosen because more purified tumor tissues of this type were available. Twenty-five of the 40 genes (62%) were found to be mutated in at least one of the 96 cancers and, as predicted from our data and simulations, most were mutated in 5% or less of the cancers (table S5). The remaining 15 *CAN*-genes were not mutated in any of the additional 96 cancers studied, but this finding is still compatible with these genes being mutated in a low but significant fraction of tumors; the evaluation of more colorectal tumors than the 131 included in our study would be necessary to exclude this possibility.

Additional analyses of mutated genes. Mutation frequency is not the only type of information that can help determine whether a mutated gene is worthy of further evaluation. The analyses of the predicted effects on protein function can add independent evidence helpful for prioritization of specific genes and mutations for future research. For example, mutations producing stop codons, out-of-frame insertions or deletions, or splice site abnormalities are very likely to interfere with the normal function of the gene product (tables S3 and S4). To evaluate missense changes, we used two sequence-based methods for evaluating the probability that a specific alteration would have a deleterious effect on protein function: Sorting Intolerant from Tolerant (SIFT) and LogR.E-values based on Pfam domains (*7*). These probabilities are listed for each evaluable mutation identified in our study in table S3. For each *CAN*-gene, the number of missense mutations that were predicted to disrupt function in a statistically significant manner is included in table S4.

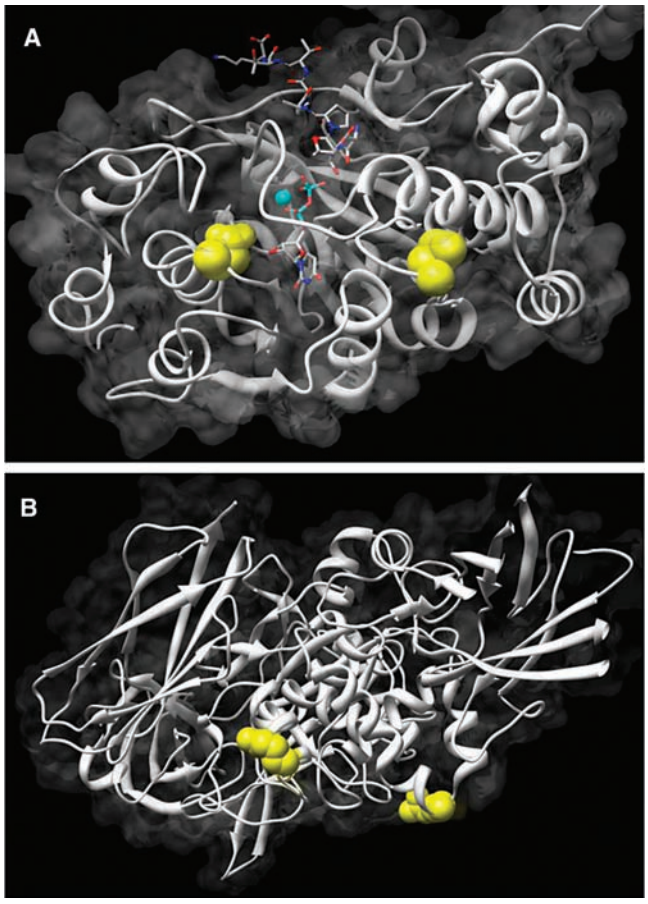


Fig. 1. Clustering of somatic mutations in protein structures. Individual somatic mutations were mapped onto structural homology models on the basis of known crystal structure information. Homology models were built with MODPIPE (*33*) and graphics were created with UCSF Chimera software (*34*). Yellow spheres indicate mutated residues. (**A**) Two somatic mutations in the glycosylation enzyme *GALNT5* occur in residues on different sides of the enzyme active site. Stick models indicate enzyme substrates. (**B**) Three somatic mutations in the transglutaminase *TGM3* located at nearby surface regions of the protein (two mutations are present at the same residue on the right-hand side).

Predictions about the functional effects of mutations can also be made at the structural level. We generated structural models for 622 of the RefSeq gene mutations from x-ray crystallography or nuclear magnetic resonance spectroscopy of their encoded or related proteins (*12*, *13*). Some of the models were intriguing in that they showed clustering of mutations around active sites of proteins or near an interface residue (examples in Fig. 1). We also used LS-SNP software (*14*) to predict the likelihood that each mutation would destabilize the protein, interfere with the formation of a domain-domain interface, or have an effect on protein-ligand binding (table S3, summarized for *CAN*-genes in table S4).

Finally, we identified a number of mutations that occurred at locations identical to those of genes involved in hereditary human diseases or that clustered at adjacent locations in the cancers analyzed. Such alterations are likely to have functional effects on these proteins. These included the R360W mutation (substitution of arginine 360 with tryptophan) in the *RET* tyrosine kinase, corresponding to an identical loss-of-function germline change in Hirschsprung disease (*15*). Likewise, the R1624W mutation in the *PKHD1* gene in colorectal cancer is identical to that observed in polycystic kidney disease, a syndrome that has neoplastic features (*16*). The T745M mutation (substitution of threonine 745 with methionine) in the cell adhesion gene *CRB1* gene is identical to one that has been shown to be a cause of retinitis pigmentosa (*17*). In addition to these examples, we identified 126 mutations in 39 proteins that occurred within a distance of 10 amino acids from one another. In particular, mutations in at least two independent tumors occurred in the *DTNB*, *EDD1*, *GNAS*, and *TGM3* genes at exactly the same residue, implicating that region as vital to the protein's potential tumorigenic function.

Analysis of mutated pathways. It is becoming increasingly clear that pathways rather than individual genes govern the course of tumorigenesis (*1*). Mutations in any of several genes of a single pathway can thereby cause equivalent increases in net cell proliferation. Accordingly, we devised a method to determine whether the genes within specific pathways were mutated more often than predicted by chance. The resultant "pathway CaMP" score incorporated the total number of mutations from all genes within each group, the number of different genes mutated, the combined sizes of the genes in each group, and the total number of tumors examined (table S6) (*7*).

Using this metric, we analyzed a highly curated database (Metacore, GeneGo, Inc.) that includes human protein-protein interactions, signal transduction and metabolic pathways, and a variety of cellular functions and processes. By including the number of mutated genes in addition to the total number of mutations as parameters, we excluded pathways that simply contained one gene that was mutated at high frequency (e.g., pathways containing only *TP53* mutations). There were 108 pathways that were found to be preferentially mutated in breast tumors. Many of the pathways involved phosphatidylinositol 3-kinase (PI3K) signaling (Fig. 2 and table S6B). Mutations in *PIK3CA* are frequent in multiple tumor types, including breast cancers (*18*–*21*). In this study, we identified mutations not only in *PIK3CA*, but also previously unreported mutations in *GAB1*, *IKBKB*, *IRS4*, *NFKB1*, *NFKBIA*, *NFKBIE*, *PIK3R1*, *PIK3R4*, and *RPS6KA3*, implicating both the PI3K pathway in general and nuclear factor κB (NF-κB) signaling in particular in breast tumorigenesis. Within the 38 colorectal cancer pathways that appeared to be mutated in a statistically significant manner, there were also many that centered on PI3K (table S6A). The pathway components mutated in colorectal cancers differed from those in breast, with mutations found in *IRS2*, *IRS4*, *PIK3R5*, *PRKCZ*, *PTEN*, *RHEB*, and *RPS6KB1* in addition to *PIK3CA*. Additional pathways altered in colorectal cancer were related to cell adhesion, the cytoskeleton, and the extracellular matrix (table S6A), supporting the idea that interactions between the cancer cell and the extracellular environment are important steps in the neoplastic process.

Finally, there were nine examples of mutated genes whose protein products were predicted to interact with other mutated genes more often than predicted by chance. The average number of mutant gene products with which these nine mutant genes interacted was 25 (table S6). These results illustrate the potential utility of pathway-based analyses and highlight a variety of different gene groups and pathways that can help focus further investigations on these tumor types.

The genomic landscapes of colorectal and breast cancers. The colorectal and breast cancers analyzed in the Discovery Screen contained a median of 76 and 84 nonsilent mutations in RefSeq genes, respectively (table S2). The number of mutations per tumor was similar among colorectal tumors (ranging from 49 to 111) but was more variable in breast cancers (varying from 38 to 193). The number of mutated *CAN*-genes per tumor averaged 15 and 14 in colorectal and breast cancers, respectively.

The "landscapes" of typical colorectal and breast cancer genomes are depicted in Fig. 3. In these landscapes, every RefSeq gene is represented by a point on a two-dimensional map corresponding to its chromosomal position, and all mutated genes in that tumor are indicated by a dot. The relief feature of the map is provided by the *CAN*-genes with the 60 highest CaMP scores (table S4). Just as topographical maps contain geological features of varying elevations, the cancer genome landscape consists of relief

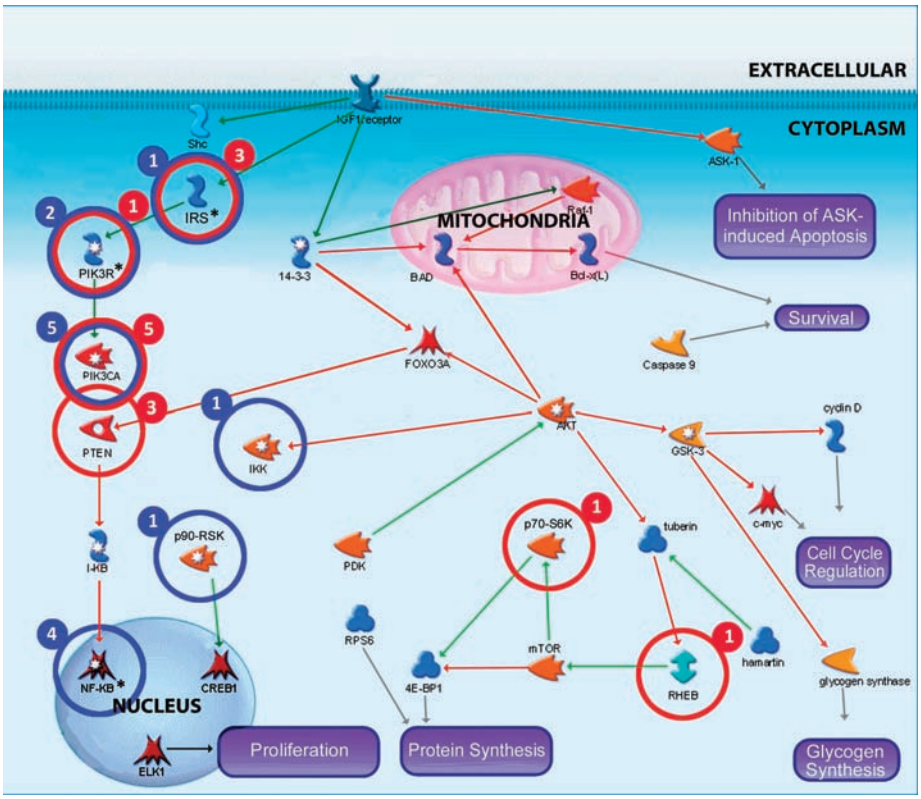


Fig. 2. PI3K pathway mutations in breast and colorectal cancers. The identities and relationships of genes that function in PI3K signaling are indicated. Circled genes have somatic mutations in colorectal (red) and breast (blue) cancers. The number of tumors with somatic mutations in each mutated protein is indicated by the number adjacent to the circle. Asterisks indicate proteins with mutated isoforms that may play similar roles in the cell. These include insulin receptor substrates IRS2 and IRS4; phosphatidylinositol 3-kinase regulatory subunits PIK3R1, PIK3R4, and PIK3R5; and NF-κB regulators NFKB1, NFKBIA, and NFKBIE.

features (mutated genes) with heterogeneous heights (determined by CaMP scores). There are a few "mountains" representing individual *CAN*-genes mutated at high frequency. However, the landscapes contain a much larger number of "hills" representing the *CAN*-genes that are mutated at relatively low frequency. It is notable that this general genomic landscape (few gene mountains and many gene hills) is a common feature of both breast and colorectal tumors.

Discussion. The results reported here add to those published previously (*5*) in several important ways. First, we report the sequences of an additional 5168 genes in 22 tumors. These new data provide a much more complete picture of the cancer genome, allowing us to formulate landscapes of breast and colorectal tumors (Fig. 3). We predict that the key features of this landscape—a few gene mountains interspersed with many gene hills—will prove to be a general feature of most solid tumors. Second, we present data on noncoding and synonymous mutations in addition to nonsynonymous mutations. As well as providing information useful for estimating the passenger rate, the data in table S2 show that passenger rates vary considerably from tumor to tumor, undoubtedly determined by their intrinsic mutability and the number of generations and bottlenecks through which they have evolved. Third, we present more sophisticated methods for identifying and classifying genes with more mutations than predicted by the passenger rate (table S4). Fourth, we present a variety of tools based on gene products' sequence and structure, as well as their inclusion in certain pathways, that can help identify mutated genes that are most deserving of further attention (Figs. 1 and 2 and tables S3, S4, and S6). These tools can be used to prioritize the research that follows cancer genome-sequencing efforts.

In terms of such research, it is important to note that sequence data can inform other, independent approaches to the study of cancer genes. For example, chromodomain helicase DNA binding domain 5 (*CHD5*) was recently proposed to be a tumor suppressor on the basis of its functional properties and copy-number alterations (*22*). We identified somatic mutations in this gene in breast tumors; the combined data strongly support a role for this gene in tumorigenesis. Similarly, the NF-κB pathway member *IKBKE* was recently suggested to be a breast cancer oncogene on the basis of functional and expression studies (*23*). We found somatic mutations in several additional components of this signaling pathway (Fig. 2), reinforcing its importance in breast cancers. The transglutaminase (TGM) enzymes have recently been implicated in invasion and metastasis (*24*), and we identified multiple somatic mutations in *TGM3* in colorectal cancers (Fig. 1). Additionally, a high-throughput retroviral insertional mutagenesis screen in mouse mammary tumor virus (MMTV)-induced mammary tumors in mice identified 33 common insertion sites as potential oncogenes (*25*); we found 7 of these 33 genes to be mutated in breast cancers. Given the entirely independent nature of these screens (insertional mutagenesis in mouse versus mutational analysis of human genes), the overlap of these results is remarkable.

Historically, the focus of cancer research has been on gene mountains, in part because they were the only alterations identifiable with available technologies. The ability to analyze the sequence of virtually all protein-encoding genes in cancers has shown that the vast majority of mutations in cancers, including those that are most likely to be drivers, do not occur in such mountains and emphasize the heterogeneity and complexity of human neoplasia. This new view of cancer is consistent with the idea that a large number of mutations, each associated with a small fitness advantage, drive tumor progression (*26*). But is it possible to make sense out of this complexity? When all the mutations that occur in different tumors are summed, the number of potential driver genes is large. But this is likely to actually reflect changes in a much more limited number of pathways, numbering no more than 20 (*1*). This interpretation is consistent with virtually all screens in model organisms, which have generally shown that the same phenotype can arise from alterations in any of several genes. Other recent studies lend support to this interpretation. For example, sequencing studies of the kinome in large numbers of tumors have shown that specific kinases are sometimes mutated in a small fraction of tumors of a given type (*4*, *10*, *27*–*29*). We cannot be certain that the bulk of the low-frequency mutations observed in our study are not passengers. However, in the kinome studies, the position of mutations within the activation loop and the demonstrated effects of the target residues on kinase function unambiguously implicate many of these rare mutations as drivers. Similarly, recent analyses of myelomas suggest that there are multiple genes, each mutated in a small proportion of tumors, that can alter the same signal transduction pathway (*30*, *31*). Furthermore, some of the low-frequency mutations observed in our study, such as activating mutations in the guanine nucleotide binding protein GNAS and a homozygous nonsense mutation in BRCA1-associated protein (BAP1), are likely to be functional (table S3). These examples, in addition to those in table S6, bolster the argument that infrequent mutations can be drivers and that they function through pathways that are already known.

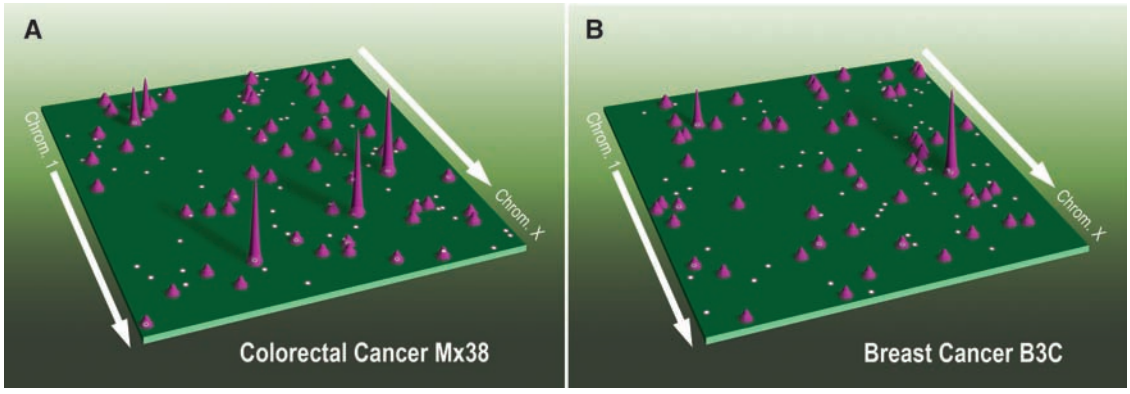


Fig. 3. Cancer genome landscapes. Nonsilent somatic mutations are plotted in two-dimensional space representing chromosomal positions of RefSeq genes. The telomere of the short arm of chromosome 1 is represented in the rear left corner of the green plane and ascending chromosomal positions continue in the direction of the arrow. Chromosomal positions that follow the front edge of the plane are continued at the back edge of the plane of the adjacent row, and chromosomes are appended end to end. Peaks indicate the 60 highest-ranking *CAN*-genes for each tumor type, with peak heights reflecting CaMP scores (*7*). The dots represent genes that were somatically mutated in the individual colorectal (Mx38) (**A**) or breast tumor (B3C) (**B**) displayed. The dots corresponding to mutated genes that coincided with hills or mountains are black with white rims; the remaining dots are white with red rims. The mountain on the right of both landscapes represents *TP53* (chromosome 17), and the other mountain shared by both breast and colorectal cancers is *PIK3CA* (upper left, chromosome 3).

Regardless of whether this pathway-centric interpretation is correct, it is clear that the "easy" part of future cancer genome research will be the identification of genetic alterations. The vast majority of subtle mutations in individual patient's tumors can now be identified with existing technology (Fig. 3), making personal cancer genomics a reality. Though understanding the precise role of these genetic alterations in tumorigenesis will be more challenging, opportunities for exploiting such personal genomic data on cancers are already apparent. For example, many of the genes altered in breast cancers appear to affect the NF-κB pathway (table S6), suggesting that drugs targeting this pathway could be efficacious in breast cancers with such mutations (*30*, *31*). Furthermore, our data indicate that individual breast and colorectal cancers each contain ~80 amino acid–altering mutations that are absent in all normal cells, providing a wealth of opportunities for personalized immunotherapy. Finally, any mutation identified in an individual cancer, whether driver or passenger, can be used as an exquisitely specific biomarker to guide patient management (*32*).

References and Notes

1. B. Vogelstein, K. W. Kinzler, *Nat. Med.* **10**, 789 (2004).
2. P. A. Futreal *et al.*, *Nat. Rev. Cancer* **4**, 177 (2004).
3. A. Bardelli, V. E. Velculescu, *Curr. Opin. Genet. Dev.* **15**, 5 (2005).
4. C. Greenman *et al.*, *Nature* **446**, 153 (2007).
5. T. Sjöblom *et al.*, *Science* **314**, 268 (2006).
6. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **35**, D61 (2007).
7. See supporting material on *Science* Online.
8. M. Cargill *et al.*, *Nat. Genet.* **22**, 231 (1999).
9. M. K. Halushka *et al.*, *Nat. Genet.* **22**, 239 (1999).
10. P. Stephens *et al.*, *Nat. Genet.* **37**, 590 (2005).
11. J. V. Chamary, J. L. Parmley, L. D. Hurst, *Nat. Rev. Genet.* **7**, 98 (2006).
12. R. Karchin, Structural models of mutants identified in breast cancers; http://karchinlab.org/RefSeqMutants/breast.html.
13. R. Karchin, Structural models of mutants identified in colorectal cancers. http://karchinlab.org/RefSeqMutants/colorectal.html.
14. R. Karchin *et al.*, *Bioinformatics* **21**, 2814 (2005).
15. S. Bolk *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 268 (2000).
16. L. F. Onuchic *et al.*, *Am. J. Hum. Genet.* **70**, 1305 (2002).
17. A. I. den Hollander *et al.*, *Nat. Genet.* **23**, 217 (1999).
18. Y. Samuels *et al.*, *Science* **304**, 554 (2004).
19. K. E. Bachman *et al.*, *Cancer Biol. Ther.* **3**, 772 (2004).
20. D. K. Broderick *et al.*, *Cancer Res.* **64**, 5048 (2004).
21. J. W. Lee *et al.*, *Oncogene* **24**, 1477 (2005).
22. A. Bagchi *et al.*, *Cell* **128**, 459 (2007).
23. J. S. Boehm *et al.*, *Cell* **129**, 1065 (2007).
24. M. Satpathy *et al.*, *Cancer Res.* **67**, 7194 (2007).
25. V. Theodorou *et al.*, *Nat. Genet.* **39**, 759 (2007).
26. N. Beerenwinkel *et al.*, *PLoS Comput. Biol.* **3**, e225 (2007).
27. A. Bardelli *et al.*, *Science* **300**, 949 (2003).
28. D. W. Parsons *et al.*, *Nature* **436**, 792 (2005).
29. R. K. Thomas *et al.*, *Nat. Genet.* **39**, 347 (2007).
30. C. M. Annunziata *et al.*, *Cancer Cell* **12**, 115 (2007).
31. J. J. Keats *et al.*, *Cancer Cell* **12**, 131 (2007).
32. F. Diehl, L. A. Diaz Jr., *Curr. Opin. Oncol.* **19**, 36 (2007).
33. R. Sanchez, A. Sali, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 13597 (1998).
34. E. F. Pettersen *et al.*, *J. Comput. Chem.* **25**, 1605 (2004).

35. We thank J. Lutterbaugh, E. Lawrence, and L. Beard for assistance with cell culture and DNA preparation; K. Makowski and the Agencourt sequencing team for assistance with automated sequencing; C.-S. Liu and the SoftGenetics team for their assistance with mutation detection analyses; and D. H. Nguyen of the Johns Hopkins Department of Art as Applied to Medicine for the artwork in Fig. 3. Supported by The Virginia and D. K. Ludwig Fund for Cancer Research; NIH grants CA121113, CA 43460, CA 57345, CA62924, GM07309, RR017698, P30-CA43703, CA109274, GM070219, and CA112828; NCI Division of Cancer Prevention contract HHSN261200433002C; Department of Defense grant DAMD17-03-1-0241; NSF grant DMS034211; The Pew Charitable Trusts; The Palmetto Health Foundation; The Maryland Cigarette Restitution Fund; The State of Ohio Biomedical Research and Technology Transfer Commission; The Clayton Fund; The Blaustein Foundation; The National Colorectal Cancer Research Alliance; the Strang Cancer Prevention Center; the Division of Cancer Prevention of the National Cancer Institute; the Avon Foundation; The Flight Attendant's Medical Research Institute; the V Foundation for Cancer Research; the Summer Running Fund; and The Palmetto Health Foundation. Under separate licensing agreements between the Johns Hopkins University and Genzyme Corporation and Exact Sciences Corporation, K.W.K., B.V., and V.E.V. are entitled to a share of royalties received by the University on sales of products related to research described in this paper. These authors and the University own Genzyme stock, which is subject to certain restrictions under University policy. The terms of these arrangements are managed by the Johns Hopkins University in accordance with its conflict-of-interest policies.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1145720/DC1
Materials and Methods
Statistical Analysis Package
Fig. S1
Tables S1 to S6
References

29 May 2007; accepted 1 October 2007
Published online 11 October 2007;
10.1126/science.1145720
Include this information when citing this paper.

The Genomic Landscapes of Human Breast and Colorectal Cancers

Laura D. Wood, D. Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjöblom, Rebecca J. Leary, Dong Shen, Simina M. Boca, Thomas Barber, Janine Ptak, Natalie Silliman, Steve Szabo, Zoltan Dezso, Vadim Ustyanksky, Tatiana Nikolskaya, Yuri Nikolsky, Rachel Karchin, Paul A. Wilson, Joshua S. Kaminker, Zemin Zhang, Randal Croshaw, Joseph Willis, Dawn Dawson, Michail Shipitsin, James K. V. Willson, Saraswati Sukumar, Kornelia Polyak, Ben Ho Park, Charit L. Pethiyagoda, P. V. Krishna Pant, Dennis G. Ballinger, Andrew B. Sparks, James Hartigan, Douglas R. Smith, Erick Suh, Nickolas Papadopoulos, Phillip Buckhaults, Sanford D. Markowitz, Giovanni Parmigiani, Kenneth W. Kinzler, Victor E. Velculescu and Bert Vogelstein (October 11, 2007)
Science **318** (5853), 1108-1113. [doi: 10.1126/science.1145720] originally published online October 11, 2007

Editor's Summary

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.