

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
16 April 2009 (16.04.2009)

PCT

(10) International Publication Number
WO 2009/049166 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12Q 1/02* (2006.01)

Halkirk Way, Bel Air, Maryland 21015 (US). **VOGEL-STEIN, Bert** [US/US]; 3700 Breton Way, Baltimore, Maryland 21208 (US).

(21) International Application Number:
PCT/US2008/079526

(74) Agent: **KAGAN, Sarah A.**; Banner & Witcoff, Ltd., 1100 13th Street, N.W., Suite 1200, Washington, DC 20005-4051 (US).

(22) International Filing Date: 10 October 2008 (10.10.2008)

(25) Filing Language: English

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(26) Publication Language: English

(30) Priority Data:
60/960,733 11 October 2007 (11.10.2007) US
12/247,464 8 October 2008 (08.10.2008) US

(71) Applicant (*for all designated States except US*): **THE JOHNS HOPKINS UNIVERSITY** [US/US]; 5th Floor, 100 N. Charles Street, Baltimore, MD 21201 (US).

(72) Inventors; and

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(75) Inventors/Applicants (*for US only*): **WOOD, Laura D.** [US/US]; 701 Cathedral St., Apt. 51, Baltimore, Maryland 21201 (US). **PARSONS, D. Williams** [US/US]; 4553 Hemlock Cone Way, Ellicott City, Maryland 21042 (US). **JONES, Sian** [GB/US]; 111 W. Centre Street, Apt 305, Baltimore, Maryland 21201 (US). **LIN, Jimmy** [US/US]; 820 N. Broadway, Baltimore, Maryland 21205 (US). **SJÖBLOM, Tobias** [SE/SE]; Kungsgatan 71, SE-753 21 Uppsala (SE). **BARBER, Thomas** [US/US]; 298 Ashbourne Drive, Noblesville, Indiana 46060 (US). **PARMIGIANI, Giovanni** [US/US]; 337 Warren Avenue, Baltimore, Maryland 21230 (US). **VELCULESCU, Victor** [US/US]; 14064 Big Branch Drive, Dayton, Maryland 21036 (US). **KINZLER, Kenneth W.** [US/US]; 1403

Published:

- *without international search report and to be republished upon receipt of that report*
- *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

(54) Title: GENOMIC LANDSCAPES OF HUMAN BREAST AND COLORECTAL CANCERS

(57) Abstract: Human cancer is caused by the accumulation of mutations in oncogenes and tumor suppressor genes. To catalogue 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.

WO 2009/049166 A2

GENOMIC LANDSCAPES OF HUMAN BREAST AND COLORECTAL CANCERS

- [01] This invention was made using grant funds from the U.S. government. Under the term of the grants, the U.S. government retains certain rights in the invention. Grants used include NIH grants CA 43460, CA 57345, CA 12113, and CA 62924.
- [02] A sequence listing is provided on a single compact disc. The compact disc contains a file named templst.txt. The file is 22695 kb and was created October 03, 2008. The content of the compact disc is incorporated herein.

TECHNICAL FIELD OF THE INVENTION

- [03] This invention is related to the area of cancer characterization. In particular, it relates to breast and colorectal cancers.

BACKGROUND OF THE INVENTION

- [04] 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).
- [05] 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 the current study, we have extended these analyses to include examination of all of the Reference Sequence (RefSeq) genes. The RefSeq database is a comprehensive, non-redundant 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.

- [06] There is a continuing need in the art to identify genes and patterns of gene mutations useful for identifying and stratifying individual patients' cancers.

SUMMARY OF THE INVENTION

- [07] According to one embodiment of the invention a method is provided for diagnosing breast cancer in a human. A somatic mutation in a gene or its encoded cDNA or protein is determined in a test sample relative to a normal sample of the human. The gene is selected from the group consisting of those listed in **Fig. 10 (Table S4B)**. The sample is identified as breast cancer when the somatic mutation is determined.
- [08] A method is provided for diagnosing colorectal cancer in a human. A somatic mutation in a gene or its encoded cDNA or protein is determined in a test sample relative to a normal sample of the human. The gene is selected from the group consisting of those listed in **Fig. 9 (Table S4A)**. The sample is identified as colorectal cancer if the somatic mutation is determined.
- [09] A method is provided for stratifying breast cancers for testing candidate or known anti-cancer therapeutics. A CAN-gene mutational signature for a breast cancer is determined by determining at least one somatic mutation in a test sample relative to a normal sample of a human. The at least one somatic mutation is in one or more genes selected from the group consisting of **Fig. 10 (Table S4B)**. A first group of breast cancers that have the CAN-gene mutational signature is formed. Efficacy of a candidate or known anti-cancer therapeutic on the first group is compared to efficacy on a second group of breast cancers that has a different CAN-gene mutational signature. A CAN gene mutational signature which correlates with increased or decreased efficacy of the candidate or known anti-cancer therapeutic relative to other groups is identified.
- [10] A method is provided for stratifying colorectal cancers for testing candidate or known anti-cancer therapeutics. A CAN-gene mutational signature for a colorectal cancer is determined by determining at least one somatic mutation in a test sample relative to a normal sample of the human. The at least one somatic mutation is in one or more

genes selected from the group consisting of **Fig. 9 (Table S4A)**. A first group of colorectal cancers that have the CAN-gene mutational signature is formed. Efficacy of a candidate or known anti-cancer therapeutic on the first group is compared to efficacy on a second group of colorectal cancers that has a different CAN-gene mutational signature. A CAN gene mutational signature is identified which correlates with increased or decreased efficacy of the candidate or known anti-cancer therapeutic relative to other groups.

- [11] A method is provided for characterizing a breast cancer in a human. A somatic mutation in a gene or its encoded cDNA or protein is determined in a test sample relative to a normal sample of the human. The gene is selected from the group consisting of those listed in **Fig. 10 (Table S4B)**
- [12] Another method provided is for characterizing a colorectal cancer in a human. A somatic mutation in a gene or its encoded cDNA or protein is determined in a test sample relative to a normal sample of the human. The gene is selected from the group consisting of those listed in **Fig. 9 (Table S4A)**.
- [13] These and other embodiments which will be apparent to those of skill in the art upon reading the specification provide the art with additional methods and tools for better managing cancer treatment.

BRIEF DESCRIPTION OF THE DRAWINGS

- [14] Fig. 1 Clustering of somatic mutations in protein structures. Individual somatic mutations were mapped onto structural homology models based on 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).

- [15] 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 nuclear factor kappa-B regulators NFKB1, NFKBIA, and NFKBIE.
- [16] Fig. 3. Cancer genome landscapes. Non-silent 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) or breast tumor (B3C) 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).
- [17] **Fig. 4. (fig. S1.)** Schematic of the experimental and bioinformatic approaches used in the study
- [18] **Fig. 5. Table 1.** Summary of somatic mutations
- [19] **Fig. 6. Table S1.** Primers used for PCR amplification and sequencing
- [20] **Fig. 7. Table S2.** Distribution of somatic mutations in individual tumors
- [21] **Fig. 8. Table S3.** Somatic mutations discovered in RefSeq genes

- [22] **Fig. 9 (Table S4A)** Colorectal *CAN*-genes
- [23] **Fig. 10 (Table S4B)** Breast *CAN*-genes
- [24] **Fig. 11. Table S5.** Summary of mutation prevalence study
- [25] **Fig. 12. Table S6A.** Gene groups and pathways preferentially mutated in colorectal cancers
- [26] **Fig. 13. Table S6B.** Gene groups and pathways preferentially mutated in breast cancers

DETAILED DESCRIPTION OF THE INVENTION

- [27] The inventors have developed methods for characterizing breast and colorectal cancers on the basis of gene signatures. These signatures comprise one or more genes which are mutated in a particular cancer. The signatures can be used as a means of diagnosis, prognosis, identification of metastasis, stratification for drug studies, and for assigning an appropriate treatment.
- [28] According to the present invention a mutation, typically a somatic mutation, can be determined by testing either a gene, its mRNA (or derived cDNA), or its encoded protein. Any method known in the art for determining a somatic mutation can be used. The method may involve sequence determination of all or part of a gene, cDNA, or protein. The method may involve mutation-specific reagents such as probes, primers, or antibodies. The method may be based on amplification, hybridization, antibody-antigen reactions, primer extension, etc. Any technique or method known in the art for determining a sequence-based feature may be used.
- [29] Samples for testing may be tissue samples from breast or colorectal tissue or body fluids or products that contain sloughed off cells or genes or mRNA or proteins. Such fluids or products include breast milk, stool, breast discharge, intestinal fluid. Preferably the same type of tissue or fluid is used for the test sample and the normal sample. The test sample is, however, suspected of possible neoplastic abnormality, while the normal sample is not suspect.

- [30] Somatic mutations are determined by finding a difference between a test sample and a normal sample of a human. This criterion eliminates the possibility of germ-line differences confounding the analysis. For breast cancer, the gene (or cDNA or protein) to be tested is any of those shown in **Fig. 10. Table S4B**. Any somatic mutation may be informative. Particular mutations which may be used are shown in **Fig. 8 (Table S3)**. For colon cancer, the gene (or cDNA or protein) to be tested is any of those shown in **Fig. 9. Table S4A**. Any somatic mutation may be informative. Particular mutations which may be used are shown in **Fig. 8 (Table S3)**.
- [31] The number of genes or mutations that may be useful in forming a signature of a breast or colorectal cancer may vary from one to twenty-five. At least two, three, four, five, six, seven, ten, fifteen, twenty, or more genes may be used. The mutations are typically somatic mutations and non-synonymous mutations. Those mutations described here are within coding regions. Other non-coding region mutations may also be found and may be informative.
- [32] In order to test candidate or already-identified therapeutic agents to determine which patients and tumors will be sensitive to the agents, stratification on the basis of signatures can be used. One or more groups with a similar mutation signature will be formed and the effect of the therapeutic agent on the group will be compared to the effect of patients whose tumors do not share the signature of the group formed. The group of patients who do not share the signature may share a different signature or they may be a mixed population of tumor-bearing patients whose tumors bear a variety of signatures.
- [33] Efficacy can be determined by any of the standard means known in the art. Any index of efficacy can be used. The index may be life span, disease free remission period, tumor shrinkage, tumor growth arrest, improvement of quality of life, decreased side effects, decreased pain, etc. Any useful measure of patient health and well-being can be used. In addition, *in vitro* testing may be done on tumor cells that have particular signatures. Tumor cells with particular signatures can also be tested in animal models.

- [34] Once a signature has been correlated with sensitivity or resistance to a particular therapeutic regimen, that signature can be used for prescribing a treatment to a patient. Thus determining a signature is useful for making therapeutic decisions. The signature can also be combined with other physical or biochemical findings regarding the patient to arrive at a therapeutic decision. A signature need not be the sole basis for making a therapeutic decision.
- [35] An anti-cancer agent associated with a signature may be, for example, docetaxel, paclitaxel, topotecan, adriamycin, etoposide, fluorouracil (5-FU), or cyclophosphamide. The agent may be an alkylating agent (e.g., nitrogen mustards), antimetabolites (e.g., pyrimidine analogs), radioactive isotopes (e.g., phosphorous and iodine), miscellaneous agents (e.g., substituted ureas) and natural products (e.g., vinca alkyls and antibiotics). The therapeutic agent may be allopurinol sodium, dolasetron mesylate, pamidronate disodium, etidronate, fluconazole, epoetin alfa, levamisole HCL, amifostine, granisetron HCL, leucovorin calcium, sargramostim, dronabinol, mesna, filgrastim, pilocarpine HCL, octreotide acetate, dexrazoxane, ondansetron HCL, ondansetron, busulfan, carboplatin, cisplatin, thiotepa, melphalan HCL, melphalan, cyclophosphamide, ifosfamide, chlorambucil, mechlorethamine HCL, carmustine, lomustine, polifeprosan 20 with carmustine implant, streptozocin, doxorubicin HCL, bleomycin sulfate, daunirubicin HCL, dactinomycin, daunorubicin citrate, idarubicin HCL, plimycin, mitomycin, pentostatin, mitoxantrone, valrubicin, cytarabine, fludarabine phosphate, floxuridine, cladribine, methotrexate, mercaptopurine, thioguanine, capecitabine, methyltestosterone, nilutamide, testolactone, bicalutamide, flutamide, anastrozole, toremifene citrate, estramustine phosphate sodium, ethinyl estradiol, estradiol, esterified estrogens, conjugated estrogens, leuprolide acetate, goserelin acetate, medroxyprogesterone acetate, megestrol acetate, levamisole HCL, aldesleukin, irinotecan HCL, dacarbazine, asparaginase, etoposide phosphate, gemcitabine HCL, altretamine, topotecan HCL, hydroxyurea, interferon alpha-2b, mitotane, procarbazine HCL, vinorelbine tartrate, E. coli L-asparaginase, Erwinia L-asparaginase, vincristine sulfate, denileukin diftitox, aldesleukin, rituximab, interferon alpha-2a, paclitaxel, docetaxel, BCG live (intravesical), vinblastine sulfate, etoposide, tretinoin, teniposide, porfimer sodium, fluorouracil, betamethasone sodium phosphate and betamethasone acetate, letrozole,

etoposide citrororum factor, folinic acid, calcium leucouorin, 5-fluorouracil, adriamycin, cytoxan, or diamino-dichloro-platinum.

- [36] The signatures of CAN genes according to the present invention can be used to determine an appropriate therapy for an individual. For example, a sample of a tumor (e.g., a tissue obtained by a biopsy procedure, such as a needle biopsy) can be provided from the individual, such as before a primary therapy is administered. The gene expression profile of the tumor can be determined, such as by a nucleic acid array (or protein array) technology, and the expression profile can be compared to a database correlating signatures with treatment outcomes. Other information relating to the human (e.g., age, gender, family history, etc.) can factor into a treatment recommendation. A healthcare provider can make a decision to administer or prescribe a particular drug based on the comparison of the CAN gene signature of the tumor and information in the database. Exemplary healthcare providers include doctors, nurses, and nurse practitioners. Diagnostic laboratories can also provide a recommended therapy based on signatures and other information about the patient.
- [37] Following treatment with a primary cancer therapy, the patient can be monitored for an improvement or worsening of the cancer. A tumor tissue sample (such as a biopsy) can be taken at any stage of treatment. In particular, a tumor tissue sample can be taken upon tumor progression, which can be determined by tumor growth or metastasis. A CAN gene signature can be determined, and one or more secondary therapeutic agents can be administered to increase, or restore, the sensitivity of the tumor to the primary therapy.
- [38] Treatment predictions may be based on pre-treatment gene signatures. Secondary or subsequent therapeutics can be selected based on the subsequent assessments of the patient and the later signatures of the tumor. The patient will typically be monitored for the effect on tumor progression.
- [39] A medical intervention can be selected based on the identity of the CAN gene signature. For example, individuals can be sorted into subpopulations according to their genotype. Genotype-specific drug therapies can then be prescribed. Medical

interventions include interventions that are widely practiced, as well as less conventional interventions. Thus, medical interventions include, but are not limited to, surgical procedures, administration of particular drugs or dosages of particular drugs (e.g., small molecules, bioengineered proteins, and gene-based drugs such as antisense oligonucleotides, ribozymes, gene replacements, and DNA- or RNA-based vaccines), including FDA-approved drugs, FDA-approved drugs used for off-label purposes, and experimental agents. Other medical interventions include nutritional therapy, holistic regimens, acupuncture, meditation, electrical or magnetic stimulation, osteopathic remedies, chiropractic treatments, naturopathic treatments, and exercise.

[40] We report the sequences of an additional 5,168 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. We also present data on non-coding and synonymous mutations in addition to non-synonymous mutations. As well as providing information useful for estimating the passenger rate, the data in table S2 shows 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. We also present more sophisticated methods for identifying and classifying genes with more mutations than predicted by the passenger rate Figs. 9, 10, (table S4). Additionally, 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, 2, 8, 9, 10, 12, 13 (tables S3, S4, S6)). These tools can be used to prioritize the research that follows cancer genome sequencing efforts.

[41] 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 based on 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- κ B pathway member IKBKE was recently suggested to be a breast cancer oncogene based on 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 MMTV-induced mammary tumors in mice identified 33 common insertion sites as potential oncogenes (25); we found seven of these 33 genes to be mutated in breast cancers. Given the entirely independent nature of these screens (insertional mutagenesis in mouse vs. mutational analysis of human genes), these results are remarkable.

- [42] Historically, the focus of cancer research has been on the 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). And 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.

[43] 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 patients' 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- κ B pathway (Figs. 12, 13; 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 an average of ~90 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).

[44] The above disclosure generally describes the present invention. All references disclosed herein are expressly incorporated by reference. The disclosure of international application PCT/US07/017866 filed August 13, 2007, is expressly incorporated by reference. A more complete understanding can be obtained by reference to the following specific examples which are provided herein for purposes of illustration only, and are not intended to limit the scope of the invention.

EXAMPLES

EXAMPLE 1-- Sequencing strategy.

- [45] 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 6,196 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.
- [46] 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 to ensure that the alterations did not occur as a result of mistargeted amplification of related regions of the genome (7).
- [47] 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. Similar methods 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.

EXAMPLE 2-- Somatic mutations.

- [48] 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 non-silent 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 the current study we analyzed the nature of the non-synonymous mutations in more detail and found a very large excess of C to T transitions at 5'-CpG-3' in colorectal cancers, representing 19-fold more than 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 fold more than expected (7).

EXAMPLE 3-- Passenger mutation rates.

- [49] 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.

[50] 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 eleven 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), 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 non-coding regions. Thus, there were a total of 0.6 non-coding mutations per Mb analyzed (95% CI: 0.52 to 0.64 mutations/Mb). Because only one copy of chromosome 8 was analyzed in these studies, the non-coding 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 non-coding regions (8, 9). After correcting for loss of heterozygosity and the difference in mutation rates between non-coding and nonsynonymous mutations, these analyses result in an estimated passenger mutation rate of 0.55 nonsynonymous mutations per Mb tumor DNA in colorectal cancers (7). We consider this a minimum estimate because the ratio of mutations in non-coding regions to non-synonymous mutations in coding regions is likely to be higher in the germline than in tumors due to greater negative selection for mutations in coding regions in the germline. Although we have not directly measured mutation rates in non-coding sequences in breast cancers, Stephens et al. have estimated that the rate of non-synonymous mutations in breast cancers is 0.33 per Mb and we used this as our minimum estimate for this tumor type (10).

[51] Estimates of the passenger mutation rates were also obtained through the quantification of synonymous (silent) missense mutations in the current study. As the majority of 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 non-synonymous mutation rate (7). One estimate was

based on the ratio of non-synonymous to synonymous mutations observed in the human germline (8, 9). The second estimate was derived by calculating the expected ratio of non-synonymous 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 the fact that nonsynonymous mutations that retard cell growth will be selected against during tumorigenesis.

EXAMPLE 4-- Evaluating mutated genes.

- [52] 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 the current 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 (tables S4A and B, respectively). The 280 CAN-genes listed in tables S4A 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 6,196 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).
- [53] It is reasonable to assume that genes that are mutated more frequently than predicted by chance are more likely to be drivers. In the current 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 (tables S4A 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 new method based on Empirical Bayes' simulations (7). Though the likelihoods depend on the passenger rates (tables S4A 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.

- [54] 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 biologic pathways of interest to our groups and 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.

EXAMPLE 5-- Additional analyses of mutated genes.

- [55] 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, two sequence-based methods for evaluating the probability that a specific alteration would have a deleterious effect on protein function were employed, 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.

- [56] Predictions about the functional effects of mutations can also be made at the structural level. We were able to generate structural models for 622 of the RefSeq gene mutations from X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy of their encoded 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).
- [57] Finally, we were able to identify 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 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 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.

EXAMPLE 6-- Analysis of mutated pathways.

- [58] 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).

- [59] 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 PI3K signaling (Fig. 2 and table S6B). Mutations in PIK3CA are frequent in multiple tumor types, including breast cancers (18-21). In the current 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 NF- κ B signaling in particular in breast tumorigenesis. Within the 38 significant colorectal cancer pathways that appeared to be mutated in a statistically significant manner, there were also many that centered on PI3K (Fig. 12; table S6A). The pathway components mutated in colorectal cancers differed from those in breast, with mutations found in IRS2, IRS4, PIK3R5, PRK CZ, 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 (Fig. 12; table S6A), supporting the idea that interactions between the cancer cell and the extracellular environment are important steps in the neoplastic process.
- [60] 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 (Fig. 12, Fig. 13; table S6A and 6B). 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.

EXAMPLE 7-- The genomic landscapes of colorectal and breast cancers.

- [61] The colorectal and breast cancers analyzed in the Discovery Screen contained an average of 77 and 101 non-silent 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.
- [62] The "landscapes" of typical colorectal and breast cancer genomes are depicted in Fig. 3. In these landscapes, every RefSeq gene is given a location on a 2-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 (Figs. 9, and 10; table S4). Just as topographical maps contain geological features of varying elevations, the cancer genome landscape consists of relief 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.

References for the foregoing examples and disclosure.

The disclosure of each reference cited is expressly incorporated herein.

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. Sjoblom *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 Computational Biology*, in press (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).

EXAMPLE 8-- Supporting Online Material**Materials and Methods**

- [63] Gene selection. The Reference Sequence database (RefSeq) represents a curated sequence database of 20,857 transcripts from 18,191 unique genes (as of March 2006; <http://www.ncbi.nlm.nih.gov/RefSeq/>). The Consensus Coding Sequence (CCDS) database represents a subset of the genes included in the RefSeq database (<http://www.ncbi.nlm.nih.gov/CCDS/>). All transcripts and genes in the CCDS database are contained within the RefSeq database; however, the RefSeq database contains an additional 6,196 transcripts (from 5,168 unique genes) that are not included in CCDS. We previously sequenced the transcripts included in the CCDS database (S1). In the current study we determined the sequence of the coding regions (exons plus four bases of adjacent introns or untranslated regions) of the remaining 6,196 transcripts. We excluded transcripts that were located at multiple locations in the genome as a result of gene duplication as well as those located on the Y chromosome. The combined dataset of all 18,191 genes in RefSeq (including those genes in CCDS that were analyzed previously) was used for the analysis and conclusions described in the text.
- [64] Bioinformatic resources. RefSeq gene and transcript coordinates (release 16, March 2006), human genome sequences, and single nucleotide polymorphisms were obtained from the UCSC Santa Cruz Genome Bioinformatics Site (<http://genome.ucsc.edu>). Homology searches in the human and mouse genomes were performed using the BLAST-like alignment tool BLAT (S2) and In Silico PCR (<http://genome.ucsc.edu/cgi-bin/hgPcr>). All genomic positions correspond to UCSC Santa Cruz hg17 build 35.1 human genome sequence. The -3.4 million single nucleotide polymorphisms (SNPs) of db SNP (release 125) that were validated through the HapMap project (S3) were used for automated removal of known polymorphisms.
- [65] Primer design. Primers for PCR amplification and sequencing of each coding exon were designed as described previously (S1), with the exception that additional manual curation was performed to determine the correct reading frame of a subset of RefSeq

genes. Briefly, primer pairs were generated using Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). with forward and reverse PCR primers located no closer than 50 bp to target exon boundaries. Exons larger than 350 bp were divided into multiple overlapping amplicons. PCR products were designed to range in size from 300 to 600 bp and primer pairs were filtered using UCSC In Silico PCR to exclude pairs yielding more than a single product. A universal sequencing primer (M13 forward, 5'-GTAAAACGACGGCCAGT-3'; SEQ ID NO: 131,069) was appended to the 5' end of the primer in the pair with the smallest number of mono- and dinucleotide repeats between itself and the target exon. For convenience, all forward and reverse primer sequences used in the previous and current study are listed in table S1 (SEQ ID NO: 1-131,068, respectively).

- [66] DNA samples, PCR amplification, and sequencing. DNA samples from ductal breast carcinoma cell lines, primary breast tumors, colorectal cancer cell lines and xenografts, and matched normal tissue or peripheral blood were obtained as described previously (S1). In brief, the samples used in the Colorectal Cancer Discovery Screen were cell lines (three) or xenografts (eight), each developed from a liver metastasis of a different patient. The eleven samples used in the Breast Cancer Discovery screen were cell lines obtained from ATCC with the following ATCC 10 numbers: B1 C = Hs 578T; B2C = HCC1008; B3C = HCC1954; B4C = HCC38; B5C = HCC1143; B6C = HCC1187; B7C = HCC1395; B8C = HCC1599; B9C = HCC1937; B10C = HCC2157; B11 C = HCC2218 (see table S2). We chose the tumors used in the Discovery Screen on the following bases. First, the colorectal I cancer samples were all late-stage tumors derived from liver metastases because such tumors contain all the mutations found in early stage tumors, but the converse is not true. We wished to gain a picture of the genomic landscapes of fully progressed neoplasms rather than of intermediate stages. The genes identified through this analysis can in the future be analyzed in early stage tumors to determine their timing with respect to the neoplastic process. Another reason to study metastatic cancers is that these are the only ones that are lethal. Similarly, most of the breast cancers represented the most aggressive type (estrogen receptor negative, progesterone receptor negative, and ERBB2 negative)(S1). These tumors are the most difficult to manage clinically as they are often refractory to therapy. Another reason underlying the choice of the breast cancers

is that these are the only publicly available cell lines, to our knowledge, for which corresponding normal cells are also available (through ATCC). This availability provides positive controls for mutation analysis by other groups and will facilitate functional studies in the future. The samples used in the Colorectal Cancer Validation Screen were xenografts or cell lines derived from advanced cases (but not necessarily metastatic sites). The 96 samples used for further mutational analysis of 40 CAN-genes were xenografts derived from cancers of various stages. The samples used in the Breast Cancer Validation Screen were primary breast tumors microdissected using laser capture (S1). Whole genome amplification, performed as previously described (S1), was used to generate sufficient quantities of DNA for Validation Screen samples when required. PCR and sequencing reactions (including the monitoring of DNA sample identity) were performed as described previously (S1). All samples were obtained in accordance with the Health Insurance Portability and Accountability Act (HIPAA).

- [67] Mutation discovery screen. RefSeq exons were amplified and sequenced in 11 colorectal cancer samples, 11 breast cancer samples, and two matched normal DNA samples. Mutational analysis was performed as described previously (S1). In brief, mutational analysis was performed for all coding exonic sequences and the flanking four base pairs (bp) of intronic or UTR sequences using Mutation Surveyor (Softgenetics, State College, PA; <http://www.softgenetics.com>) coupled to a relational database (Microsoft SQL Server). Only amplicons meeting stringent quality criteria were analyzed: at least 75% of the tumor samples had to have Phred quality scores of :0:20 in :0:90% of the bases within the target region of each amplicon. In the amplicons that passed these quality criteria, three groups of mutations were removed: nonsynonymous changes in tumor samples identical to changes in the two normal DNA samples, known single-nucleotide polymorphisms (db SNP entries previously validated by the HapMap project), and false positive artifacts that could be eliminated by visual inspection of chromatograms. Somatic synonymous mutations were not removed from analysis in the current study, though they were removed in our previous study of CCDS genes. Following mutational analysis, each putative mutation was independently reamplified in both tumor DNA (to eliminate artifacts) and in DNA from normal tissue from the same patient (to eliminate germ line variants). To

exclude the possibility that putative somatic mutations were caused by amplification of homologous but non-identical sequences, BLAT (S2) was used to search the human genome for related exons. For samples from xenografts, BLAT was used to similarly search the mouse genome to exclude the possibility that a putative mutation actually represented a homologous mouse sequence.

- [68] Mutation validation screen. Every gene in which a nonsynonymous mutation was found in the Discovery Screen was further analyzed by amplification and sequencing of 24 additional tumor samples of the same tissue type. All RefSeq transcript variants were investigated for each gene of interest. Mutation detection, confirmation of alterations, and determination of somatic status was performed as described above, with the exception that all germline variants previously observed in the normal DNA samples of the Discovery Screen were excluded as possible somatic mutations. All somatic mutations observed in the Discovery and Validation Screens (including synonymous changes) are reported in table S3.
- [69] Mutations in non-coding sequences. To determine the rate of mutations in noncoding sequences in colorectal cancers, we used variant detection oligonucleotide microarrays. We selected tumors that had lost heterozygosity for all or nearly all of chromosome 8p. This loss of heterozygosity enhances the sensitivity of mutational analysis in microarrays because the great majority of mutations in these tumors will be homozygous (i.e., without the "noise" emanating from the wild type allele (S4)). The publicly available chromosome 8p sequence was masked for repeats using RepeatMasker (<http://www.repeatmasker.org/>), and oligonucleotide probes were designed to query each nucleotide position in the 23.79 Mb of non-repetitive 8p sequence, as previously described (S4, S5). Chromosome 8p was amplified as 3840 minimally overlapping -10 kb regions from each of eleven tumor samples using long range PCR as described (S4). Labeled PCR products were hybridized and the arrays scanned as previously described (S4). The mutations identified were then validated by individual genotyping on arrays and confirmed by dideoxy sequencing.
- [70] Analysis of loss of heterozygosity. Loss of heterozygosity (LOH) was evaluated in the Discovery screen colorectal cancers using Illumina's HumanHap300 Genotyping

BeadChip arrays. Genotype and intensity data were collected for over 317,000 polymorphic sites in each sample. The single nucleotide polymorphism (SNP) loci used in this assay were taken from the International HapMap Project and were selected for regions of the genome that are highly conserved or in close proximity to a gene. Using Illumina BeadStudio software, the normalized intensity values (log R ratio) and normalized genotype calls (B allele frequency) were plotted by genomic position across the entire genome. Regions that had undergone LOH were identified by an extended stretch of homozygous genotype calls (B allele frequencies of >0.9 or <0.1). For small regions of homozygous genotype calls (≤ 5 Mb) we also looked for a corresponding decrease in intensity (decreased log R ratio). Base positions of LOH boundaries were identified as the genomic location of the first heterozygous SNP on either side of the LOH region. On average, 16% of the tumors' genomes were found to harbor LOH.

- [71] Estimation of passenger mutation rates. The combination of somatic mutation detection with microarrays and LOH analyses described above was used to derive one estimate of passenger mutation frequencies in colorectal cancers, termed the "External" rate. This was determined to be 0.55 nonsynonymous mutations/Mb ($= 1.2$ mutations per Mb non-coding diploid DNA $\times 0.5$ nonsynonymous mutations per mutation in non-coding DNA $\times$ the fraction of diploid tumor DNA $[1 - 0.16] + 0.6$ mutations per Mb non-coding haploid DNA $\times 0.5$ nonsynonymous mutations per mutation in non-coding DNA $\times$ the fraction of haploid tumor DNA $[0.16]$, i.e., $0.55 = [1.2 \times 0.5 \times [1 - 0.16] + 0.6 \times 0.5 \times 0.16]$). As noted in the text, the External rate for breast cancers was assumed to be 0.33 nonsynonymous mutations/Mb.
- [72] To estimate the passenger mutation rates from the synonymous mutations discovered in the current study, we first determined the expected nonsynonymous to synonymous mutation ratios. These were estimated in two ways. First, we calculated this ratio based on coding SNPs identified in previous sequencing studies (S6) (S7). The ratio of nonsynonymous (NS) to synonymous (S) mutations in these studies was 1.02. This ratio may be an underestimate of the true passenger mutation rate because the selection against NS mutations may be more stringent in the germ line than during tumor development. We therefore also determined the NS:S ratio from the data

described in the current study in a manner similar to that previously described (88). In brief, context-specific mutation rates were used to determine the expected frequency of mutations that would create NS vs. S mutations. Each nucleotide of each codon was mutated in silico to determine whether a particular change would result in a NS or S change, thereby accounting for all possible changes to all bases of each codon. The fraction of changes resulting in NS and S alterations were adjusted to account for the type of base that was mutated, the base change that resulted from the mutation, the immediate 5' and 3' neighbors to the mutated base, and codon usage. Through analysis of all RefSeq genes, we determined that the expected NS:S ratios were 2.41 and 2.65 in colorectal and breast cancers, respectively. As noted in the text, these theoretical estimates provide an upper bound to the true mutation rate because they do not take into account the fact that nonsynonymous mutations that retard cell growth will be selected against during tumorigenesis.

- [73] The products of these ratios and the observed synonymous mutation rates in each screen yielded two different estimates of the passenger mutation rates, termed "SNP-based" and "NS/S-based," respectively. For example, the rate of synonymous mutations in the colorectal cancer Discovery Screen was 0.97 mutations/Mb. The SNP-based passenger rate was therefore estimated to be 0.99 NS mutations/Mb ($=0.97 \times 1.02$) while the NS/S-based passenger rate was 2.35 NS mutations/Mb ($=0.97 \times 2.41$). In the breast cancer Discovery screen, the rate of synonymous mutations was 1.37, leading to SNP- and NS/S-based passenger rates of 1.40 and 3.62 NS mutations/Mb, respectively. Different rates of synonymous mutations were observed in the various screens employed in our study, likely reflecting biologic differences in the samples analyzed. In the colorectal cancer Validation screen, the SNP- and NS/S-based passenger rates were estimated to be 1.44 and 3.41 NS mutations/Mb, respectively. In the breast cancer Validation screen, the SNP- and NS/S based passenger rates were estimated to be 0.74 and 1.91 NS mutations/Mb, respectively.
- [74] Computational analysis of mutations. Each missense mutation was analyzed by calculating a Sorting Intolerant From Tolerant (SIFT) probability (S9) and a 10gRE-value score (S10). SIFT was installed and run locally and only probabilities from variants with a median sequence information of <3.25 are listed in table S3.

Alignment files were generated using the October 2006 UniProt database. Mutations with a SIFT score ~ 0.05 are associated with a false positive rate of 20% (S9). Pfam-based LogRE-value scores were derived from expect values provided by the HMMER 2.3.2 software. The Is mode was used to search against the Pfam protein family database. LogRE-value scores were calculated as $\log_{10}(\text{Evariant}/\text{Ecanonical})$ only for canonical domains with expect values less than 1. In cases where multiple Pfam domains were found to overlap a single variant, the domain with the largest (i.e., least significant) LogRE-value score was used.

- [75] Structural modeling of mutations. For each somatic missense mutant identified in a breast or colorectal tumor sample, we applied a protocol developed for the LS-SNP large scale SNP annotation web service (S11). The UCSC Genome Browser API library was used to extract all human UniProt protein sequences that aligned with the genomic address of each mutant. Protein structure homology models for each sequence were then built with MODPIPE and MODELLER (S12-15). The MODPIPE pipeline identifies x-ray crystal structures ("templates") of proteins homologous to each protein sequence of interest by building a PSI-BLAST profile (using 10 iterations and E-value cutoff of 0.0001) and aligning the profile to a library of candidate template sequence profiles with IMPALA (S16, 17). Homology models are built with MODELLER for all sequence-template matches with statistically significant alignments (E-value < 0.0001). Amino acid residues that are near binding surfaces (at the interface of the protein and its ligand or at the interface between two protein domains) are often functionally important. Therefore, each template protein structure was checked for positions that are within a short distance of small molecule ligands ($< 5.0 \text{ \AA}$) or adjacent protein domains ($< 6 \text{ \AA}$) using the LIGBASE and PIBASE databases (S18, 19). All missense mutants that aligned to one of these "ligand-binding" or "domain interface" amino acid residues in the template structure were identified using the sequence-template alignments constructed by MODPIPE. If a missense mutation aligned to a binding or interface residue in a template protein structure, it was annotated as a binding or interface residue.
- [76] The LS-SNP score was calculated by a soft margin support vector machine trained on disease and neutral mutations annotated in UniProt (S15) with predictive features

described previously (S11). Negative LS-SNP scores predict a deleterious missense mutant while positive scores predict a neutral missense mutant. The absolute value of the score provides a confidence measure for the prediction. In a three-fold cross-validation test, the classifier yielded a false positive rate of 33%.

[77] Differences in CAN-genes between Sjoblom et al. and the current study. Sjoblom et al. reported a total of 191 CAN-genes while 280 CAN-genes are reported in this study. This difference is due to the following factors:

1. A major difference was that we discovered 114 new CAN-genes among the RefSeq genes analyzed in the current study. These genes were not included in the CCDS gene database and were not analyzed in the Sjoblom et al. study.
2. One of the breast cancers used in the Validation cohort of both Sjoblom et al. and the current study (BB23) was found to have more than six times the average number of synonymous mutations and more than ten times the average number of total mutations identified in the other breast cancers, presumably due to a higher passenger mutation rate. Because of the greater difficulty in interpreting the significance of mutations in tumors with abnormally high passenger mutation rates, we excluded all mutations identified in this tumor. This was a conservative measure, as a subset of these could have contributed to tumorigenesis.
3. CAN-genes were defined differently than in Sjoblom et al. In the current study, CAN-genes were simply defined as those in which at least one nonsynonymous mutation was discovered in both the Discovery and Validation Screens and whose length-dependent mutation rate exceeded a threshold (see section on Statistical Analyses of CAN-genes below). This definition emphasizes that CAN-genes are simply candidates that require further evaluation to implicate them as causal contributors to neoplasia. A new statistical method to determine the likelihood that each CAN-gene is mutated at greater frequency than expected by chance is presented in the current study (see below). However, the frequency of mutation among tumors

is not the only criterion that can be used to help assess the relevance of mutations in cancers. Other bioinformatic methods to help prioritize CAN-genes for future research are described in the text and in tables S4 and S5.

4. Whole genome amplification (WGA) with ϕ 29 polymerase was used in both Sjoblom et al. and in the current study to obtain sufficient DNA for samples in the Validation Screen. However, we recently found that WGA can produce a small fraction of artifactual mutations, even when as many as five WGA reactions are pooled together and used as templates for PCR (as was always employed in our studies). Analogous problems with WGA have recently been independently observed by others (S20, S21). We therefore confirmed mutations present in WGA samples by analyzing non-amplified samples from the same tumors whenever possible and excluded those that could not be confirmed from tables S3 and S4.

[78] Statistical Analyses of CAN-genes. The statistical analyses focused on quantifying the evidence that the mutations in a gene reflect an underlying mutation rate that is higher than the passenger rate (S22-25). The basis of this quantification was an Empirical Bayes analysis (S26) comparing the experimental results to a reference distribution representing a genome composed only of passenger genes. This was obtained by simulating mutations at the passenger rate in a way that precisely replicated the two-stage experimental design. Specifically, for the Discovery phase, we considered each gene in turn and simulated the number of mutations of each type from a binomial distribution with success probability equal to the context-specific passenger rate. The number of available nucleotides in each context was the number of successfully sequenced nucleotides for that particular context and gene in the samples studied in the Discovery Screen. When considering base pair substitution mutations, we considered only nucleotides-at-risk, i.e., those nucleotides that could result in a non-synonymous mutation when altered. For example, missense mutations at the third position of many codons would not result in a nonsynonymous mutation so were excluded from consideration. For all genes in which at least one mutation was generated in this simulation, the process was repeated, this time with the number of samples used in the Validation Screen. In the simulations employing the SNP- and

NS/S-based passenger rates, different passenger mutation rates were used in the Validation and Discovery stages of the simulations for the reasons described above ("Estimation of passenger mutation rates" section). We finally applied to the simulated data the same threshold that was applied to the experimental data, that is, we included only genes whose mutation rates were >15 and >6 mutations per Mb of successfully sequenced nucleotides for genes whose coding exons were greater or less than 10 kb, respectively.

- [79] Using these simulated datasets, we evaluated the passenger probabilities for each of the CAN genes. In Sjoblom et al., we calculated a false discovery rate (FOR) for groups of genes that had CaMP scores above a threshold. The FOR estimates the proportion of true passenger genes among a group of genes which may contain both passengers and nonpassengers. In contrast, the passenger probabilities calculated here (tables S4A and S4B) represent statements about specific genes rather than about groups of genes. The passenger probability is therefore more informative, when considering individual genes, than the false discovery rate. It is obtained via a logic related to that of likelihood ratios: the likelihood of observing a particular score in a gene if that gene is a passenger is compared to the likelihood of observing it in the real data. The gene-specific score used in our analysis was based on the Likelihood Ratio Test (LRT) for the null hypothesis that, for the gene under consideration, the mutation rates are all the same as the passenger mutation rates. To obtain this score, we simply transformed the LRT to $s = \log_{10}(\text{LRT})$. Higher scores indicate evidence of mutation rates above the passenger rates. The approach for evaluating passenger probabilities is the same as that described in Efron and Tibshirani (S21). Specifically, for any given score s , $F(s)$ represents the proportion of simulated genes with score higher than s in the experimental data, F_0 is the corresponding proportion in the simulated data, and p_0 is the estimated overall proportion of passenger genes (discussed below). The variation across simulations is small but nonetheless we generated and collated 1600 datasets to estimate F_0 . We then numerically estimated the density functions f and f_0 corresponding to F and F_0 and calculated, for each score s , the ratio $P_0 f_0(s) / f(s)$, also known as "local false discovery rate" (S26). Density estimation was performed using the function "density" in the R statistical

programming language (S27) with default settings. An open source R package for performing these calculations is available from the authors as well as from Science.

- [80] The passenger probability calculations depend on an estimate of p_0 , the proportion of true passengers. Our implementation seeks to give an upper bound to p_0 and thus provide conservatively high estimates of the passenger probabilities. We start by constructing histograms of the observed and simulated values of $10g(\text{LRT})$ for all genes in RefSeq, using bins of one unit. Consider the bin ranging from 0 to 1, which is composed mostly of genes with no mutations. Suppose that there are 1000 experimental genes and 1050 simulated genes in that bin. The 1000 genes include both passengers and non-passengers, while the 1050 genes should contain only passengers. Thus we can conclude that the number of passengers in the simulated set is too large and that p_0 is at most 1000/1050. Because this argument can be applied to all bins, we can estimate P_0 to be the reciprocal of the largest ratio between the simulated and observed bin counts. Estimates of p_0 were found to be stable over a wide range of bin sizes. This method is an adaptation of the approach proposed in Efron and Tibshirani (S26). In their approach, bin counts are modeled as a function of the scores using Poisson regression. In our case, a similar smoothing was achieved more simply by binning similar score values. We also constrained the passenger probabilities to change monotonically with the score by starting with the lowest values and recursively setting values that decrease to the next value to their right. A detailed mathematical account of the main analytic techniques used is provided in (S28).
- [81] The cancer mutation prevalence (CaMP) score was introduced in (S1) and described in additional detail in (S28). For each CAN-gene, we calculated the probability p_g of observing its exact mutation profile given the assumed passenger mutation rate. The mutation profile of a gene refers to the numbers of each of the 25 context-specific types of mutations in that gene (e.g., C to T transition mutations at 5'-CpG-3' sites are one type). The CaMP score is defined as the negative log of p_g divided by the relative rank of p_g among the CAN-genes. For visualization purposes in Fig. 3, all genes with CaMP scores <9 , as determined with the SNP-based passenger rate, were represented as hills of the same dimension. The CaMP scores calculated for each colorectal and breast CAN-gene are provided in tables S4A and B, respectively. To

compute CaMP scores in the SNP- and NS/S based passenger rate scenarios, we defined the pg as the product of two separate binomials for the two stages.

- [82] Analysis of mutation prevalence study. As described in the text, we experimentally tested 40 CAN-genes in a separate cohort of 96 cancers. Finding several additional mutations in these genes can provide strong evidence that they are mutated at rates higher than the passenger rate. Because the process of selection of these 40 genes for further study could not be easily represented in terms of mutation counts, it was difficult to generate reference distributions such as the ones used to compute passenger probabilities for the Discovery and Validation Screens. We therefore chose an analytic method that was insensitive to the selection process. In table S5, we report the a posteriori probability that the mutation rate for each gene studied was above the passenger rate. For this we used an Empirical Bayes estimate of the probability of the gene being a passenger to be the prior. This was constructed as for table S4A. For each of the 40 genes in the mutation prevalence study, we then computed a Bayes Factor, based on the results of the mutation prevalence study alone, for the hypothesis that the gene was mutated at the passenger mutation rate. Computation of the Bayes Factor requires specification of a prior distribution of mutation rates that corresponds to the alternative hypothesis. To construct this distribution, we assumed that, for each of non-passenger gene, the 25 non-passenger mutation rates followed Gamma distributions. These are further assumed to have the same shape parameter and scale parameters set so that the mean non-passenger rates are equal to the corresponding passenger mutation rates multiplied by a single scaling factor common to all contexts. The shape parameter and the scaling factor were estimated empirically from the set of CAN genes as follows. Drawing from the probabilities in table S4 we randomly assigned each gene to a true status of either passenger or non-passenger. We then fit, by maximum likelihood, a Poisson-gamma model in which mutations had a Poisson distribution and gene-specific mutation rates had a gamma distribution. Finally, Bayes' rule was used to combine the prior and Bayes Factor into the posterior probabilities reported in table S5. This method controlled for multiple testing via the prior distribution.

[83] Analysis of mutated gene pathways and groups. Four types of data were obtained from the MetaCore database (GeneGo, Inc., St. Joseph, MI): pathway maps, Gene Ontology (GO) processes, GeneGo process networks, and protein-protein interactions. The memberships of each of the 20,857 transcripts in these categories were retrieved from the databases using RefSeq identifiers. In GeneGo pathway maps, 21,252 relations were identified, involving 5,175 transcripts and 362 pathways. For Gene Ontology processes, a total of 33,797 pairwise relations were identified, involving 11,473 transcripts and 2,809 GO groups. For GeneGo process networks, a total of 27,312 pairwise relationships, involving 8,157 transcripts and 115 processes, were identified. The predicted protein products of each mutated gene were also evaluated with respect to their physical interactions with proteins encoded by other mutated genes as inferred from the MetaCore database. For each group in each of these four categories (pathways, GO Processes, GeneGo process networks, and protein-protein interactions), transcripts were combined into genes and several statistics were then calculated. First, we calculated the total number of nucleotides within each group that were successfully sequenced in our study. The total number of NS mutations observed in the study in each category was then tallied. The number of NS mutations observed, the number of nucleotides successfully sequenced, and the passenger mutation rates were then used to evaluate the probability of observing as many mutations as observed in the group, or more, using a binomial distribution (group P-value). The passenger mutation rate used for these calculations was the average of the estimates for the Discovery Screen (1.56 nonsynonymous mutations/Mb for both colon and breast; see above section on "Estimation of passenger mutation rates"). The group P-values for observing the number of mutations were calculated in the R statistical environment and subsequently corrected for multiplicity employing the Benjamini- Hochberg algorithm (S29) with an alpha of 0.05.

[84] We next determined whether any of the groups found to be significant in terms of the total number of mutations in the group were also significant with regards to the number of mutated genes. This second stage excluded groups in which one or a few genes in the group (such as TP53 or APC) accounted for most of the mutations in that group. For each group, we counted the number of genes sequenced and the number of genes mutated in the study. The significance of association between belonging to a

group and being a CAN-gene was assessed with a chi-square test using an alpha of 0.05. Because this second stage considered only those groups that were found to be statistically significant in terms of the total number of mutations (as described in the paragraph above), no further penalties for multiple comparisons were applied. Groups that were statistically significant in both analyses (i.e., by total number of mutations and by total number of genes with mutations) are listed in table S6.

Supplemental References for Example 8

- S1. T. Sjoblom et al., *Science* 314, 268 (2006).
- S2. W. J. Kent, *Genome Res* 12, 656 (2002).
- S3. I. H. Consortium, *Nature* 437, 1299 (2005).
- S4. N. Patil et al., *Science* 294, 1719 (2001).
- S5. M. Chee et al., *Science* 274, 610 (1996).
- S6. M. Cargill et al., *Nat Genet* 22, 231 (1999).
- S7. M. K. Halushka et al., *Nat Genet* 22, 239 (1999).
- S8. C. Greenman, R. Wooster, P. A. Futreal, M. R. Stratton, D. F. Easton, *Genetics* 173, 2187 (2006).
- S9. P. C. Ng, S. Henikoff, *Nucleic Acids Res* 31, 3812 (2003).
- S10. R. J. Clifford, M. N. Edmonson, C. Nguyen, K. H. Buetow, *Bioinformatics* 20, 1006 (2004).
- S11. R. Karchin et al., *Bioinformatics* 21, 2814 (2005).
- S12. R. Sanchez, A. Sali, *Proc Natl Acad Sci USA* 95, 13597 (1998).
- S13. A. Sali, T. L. Blundell, *Journal of Molecular Biology* 234, 779 (1993).
- S14. R. M. Kuhn et al., *Nucleic Acids Res* 35, 0668 (2007).
- S15. C. H. Wu et al., *Nucleic Acids Res* 34, 0187 (2006).
- S16. S. F. Altschul et al., *Nucleic Acids Research* 25, 3389 (1997).
- S17. A. A. Schaffer et al., *Bioinformatics* 15, 1000 (1999).
- S18. A. C. Stuart, V. A. Ilyin, A. Sali, *Bioinformatics* 18, 200 (2002).
- S19. F. P. Davis, A. Sali, *Bioinformatics* 21, 1901 (2005).
- S20. J. G. Paez et al., *Nucleic Acids Res* 32, e71 (2004).
- S21. J. J. Corneveaux et al., *Biotechniques* 42, 77 (2007).
- S22. A. F. Rubin, P. Green, *Science* 317, 1500 (2007);

- www.sciencemag.org/cgi/content/full/317/5844/1500c.
- S23. G. Getz et al., Science 317, 1500 (2007);
www.sciencemag.org/cgi/content/full/317/5844/1500b.
- S24. Forrest, G. Cavet, Science 317, 1500 (2007);
www.sciencemag.org/cgi/content/full/317/5844/1500a
- S25. G. Parmigiani et al., Science 317, 1500 (2007);
www.sciencemag.org/cgi/content/full/317/5844/1500d.
- S26. B. Efron, R. Tibshirani, Genet Epidemiol. 23, 70 (2002).
- S27. R. Ihaka, R. Gentleman, Journal of Computational and Graphical Statistics 5, 299 (1996).
- S28. G. Parmigiani et al., <http://www.bepress.com/jhubiostatlpaper126/> (2006).
- S29. Y. Benjamini, Y. Hochberg, Journal of the Royal Statistical Society. Series B (Methodological) 57 289-300 (1995).

We claim:

1. A method of diagnosing breast cancer in a human, comprising the steps of:
determining in a test sample relative to a normal sample of the human, a somatic mutation in a gene or its encoded cDNA or protein, said gene selected from the group consisting of those listed in **Fig. 10 (Table S4B)**.
identifying the sample as breast cancer when the somatic mutation is determined.
2. The method of claim 1 wherein the mutation is selected from those shown in **Fig. 8 (Table S3)**.
3. The method of claim 1 wherein the test sample is a breast tissue sample or a suspected breast cancer metastasis.
4. The method of claim 1 wherein the normal sample is a breast tissue sample.
5. A method of diagnosing colorectal cancer in a human, comprising the steps of:
determining in a test sample relative to a normal sample of the human, a somatic mutation in a gene or its encoded cDNA or protein, said gene selected from the group consisting of those listed in **Fig. 9 (Table S4A)**;
identifying the sample as colorectal cancer when the somatic mutation is determined.
6. The method of claim 5 wherein the mutation is selected from those shown in **Fig. 8 (Table S3)**.
7. The method of claim 5 wherein the test sample is a colorectal tissue sample or a suspected colorectal cancer metastasis.
8. The method of claim 5 wherein the normal sample is a colorectal tissue sample.
9. A method to stratify breast cancers for testing candidate or known anti-cancer therapeutics, comprising the steps of:

determining a CAN-gene mutational signature for a breast cancer by
determining at least one somatic mutation in a test sample relative to a normal sample of a

human, wherein the at least one somatic mutation is in one or more genes selected from the group consisting of **Fig. 10 (Table S4B)**;

forming a first group of breast cancers that have the CAN-gene mutational signature;

comparing efficacy of a candidate or known anti-cancer therapeutic on the first group to efficacy on a second group of breast cancers that has a different CAN-gene mutational signature;

identifying a CAN gene mutational signature which correlates with increased or decreased efficacy of the candidate or known anti-cancer therapeutic relative to other groups.

10. The method of claim 9 wherein the at least one mutation is selected from those shown in **Fig. 8 (Table S3)**.
11. The method of claim 9 wherein the test sample is a breast tissue sample.
12. The method of claim 9 wherein the normal sample is a breast tissue sample.
13. The method of claim 9 wherein the CAN-gene mutational signature comprises at least 2 genes selected from **Fig. 10. Table S4B**.
14. The method of claim 9 wherein the CAN-gene mutational signature comprises at least 3 genes selected from **Fig. 10. Table S4B**.
15. The method of claim 9 wherein the CAN-gene mutational signature comprises at least 4 genes selected from **Fig. 10. Table S4B**.
16. The method of claim 9 wherein the CAN-gene mutational signature comprises at least 5 genes selected from **Fig. 10. Table S4B**.
17. The method of claim 9 wherein the CAN-gene mutational signature comprises at least 6 genes selected from **Fig. 10. Table S4B**.
18. The method of claim 9 wherein the CAN-gene mutational signature comprises at least 7 genes selected from **Fig. 10. Table S4B**.
19. A method to stratify colorectal cancers for testing candidate or known anti-cancer therapeutics, comprising the steps of:

determining a CAN-gene mutational signature for a colorectal cancer by determining at least one somatic mutation in a test sample relative to a normal sample of the human, wherein the at least one somatic mutation is in one or more genes selected from the group consisting of **Fig. 9 (Table S4A)**;

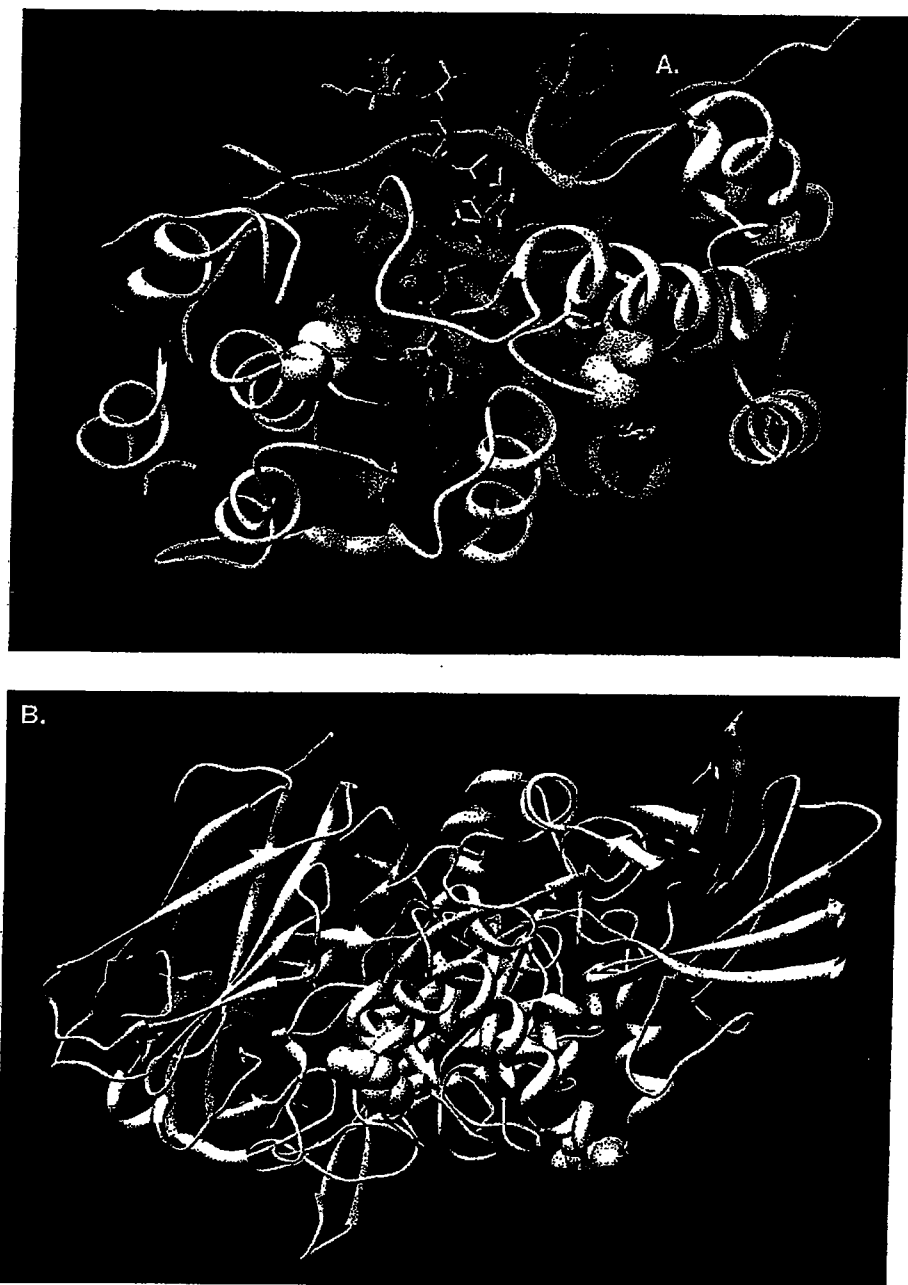
forming a first group of colorectal cancers that have the CAN-gene mutational signature;

comparing efficacy of a candidate or known anti-cancer therapeutic on the first group to efficacy on a second group of colorectal cancers that has a different CAN-gene mutational signature;

identifying a CAN gene mutational signature which correlates with increased or decreased efficacy of the candidate or known anti-cancer therapeutic relative to other groups.

20. The method of claim 19 wherein the at least one mutation is selected from those shown in **Fig. 8 (Table S3)**.
21. The method of claim 19 wherein the test sample is a colorectal tissue sample.
22. The method of claim 19 wherein the normal sample is a colorectal tissue sample.
23. The method of claim 19 wherein the CAN-gene mutational signature comprises at least 2 genes selected from **Fig. 9 (Table S4A)**.
24. The method of claim 19 wherein the CAN-gene mutational signature comprises at least 3 genes selected from **Fig. 9 (Table S4A)**.
25. The method of claim 19 wherein the CAN-gene mutational signature comprises at least 4 genes selected from **Fig. 9 (Table S4A)**.
26. The method of claim 19 wherein the CAN-gene mutational signature comprises at least 5 genes selected from **Fig. 9 (Table S4A)**.
27. The method of claim 19 wherein the CAN-gene mutational signature comprises at least 6 genes selected from **Fig. 9 (Table S4A)**.
28. The method of claim 19 wherein the CAN-gene mutational signature comprises at least 7 genes selected from **Fig. 9 (Table S4A)**.
29. A method of characterizing a breast cancer in a human, comprising the steps of:
determining in a test sample relative to a normal sample of the human, a somatic mutation in a gene or its encoded cDNA or protein, said gene selected from the group consisting of those listed in **Fig. 10 (Table S4B)**.
30. The method of claim 29 wherein the mutation is selected from those shown in **Fig. 8 (Table S3)**.
31. The method of claim 29 wherein the test sample is a breast tissue sample or a suspected breast cancer metastasis.

32. The method of claim 29 wherein the normal sample is a breast tissue sample.
33. A method of characterizing a colorectal cancer in a human, comprising the steps of:
determining in a test sample relative to a normal sample of the human, a somatic mutation in a gene or its encoded cDNA or protein, said gene selected from the group consisting of those listed in **Fig. 9 (Table S4A)**.
34. The method of claim 33 wherein the mutation is selected from those shown in **Fig. 8 (Table S3)**.
35. The method of claim 33 wherein the test sample is a colorectal tissue sample or a suspected colorectal cancer metastasis.
36. The method of claim 33 wherein the normal sample is a colorectal tissue sample.



Figs. 1A and 1B

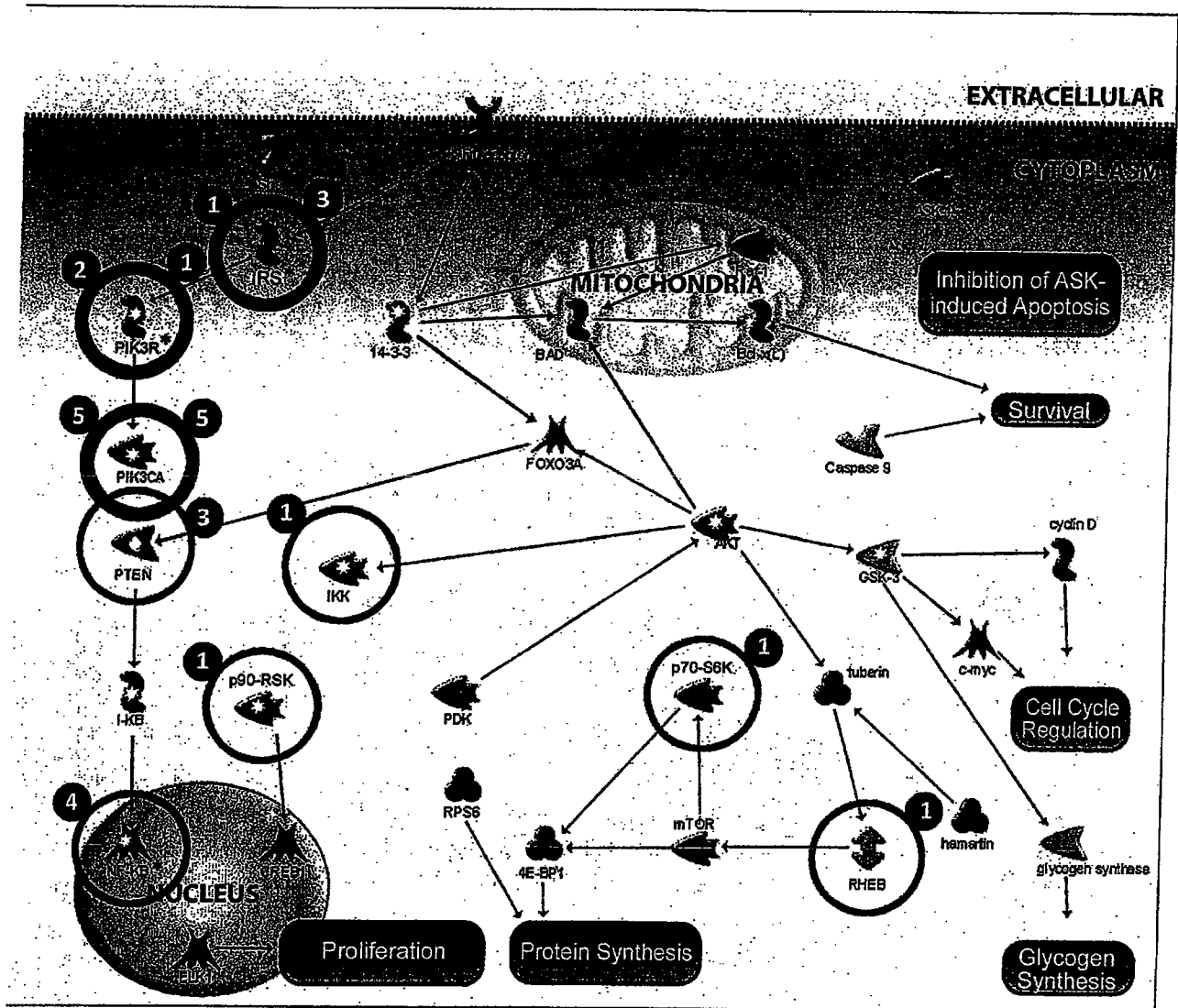
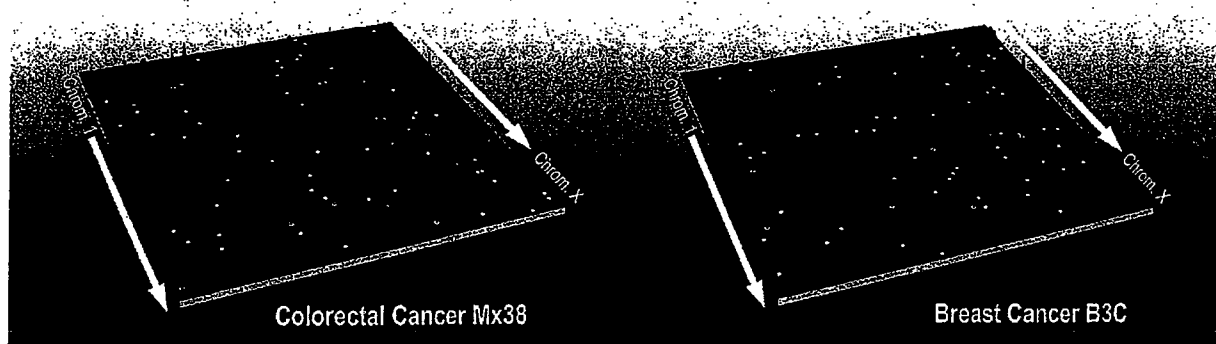


Fig. 2.

**Fig. 3.**

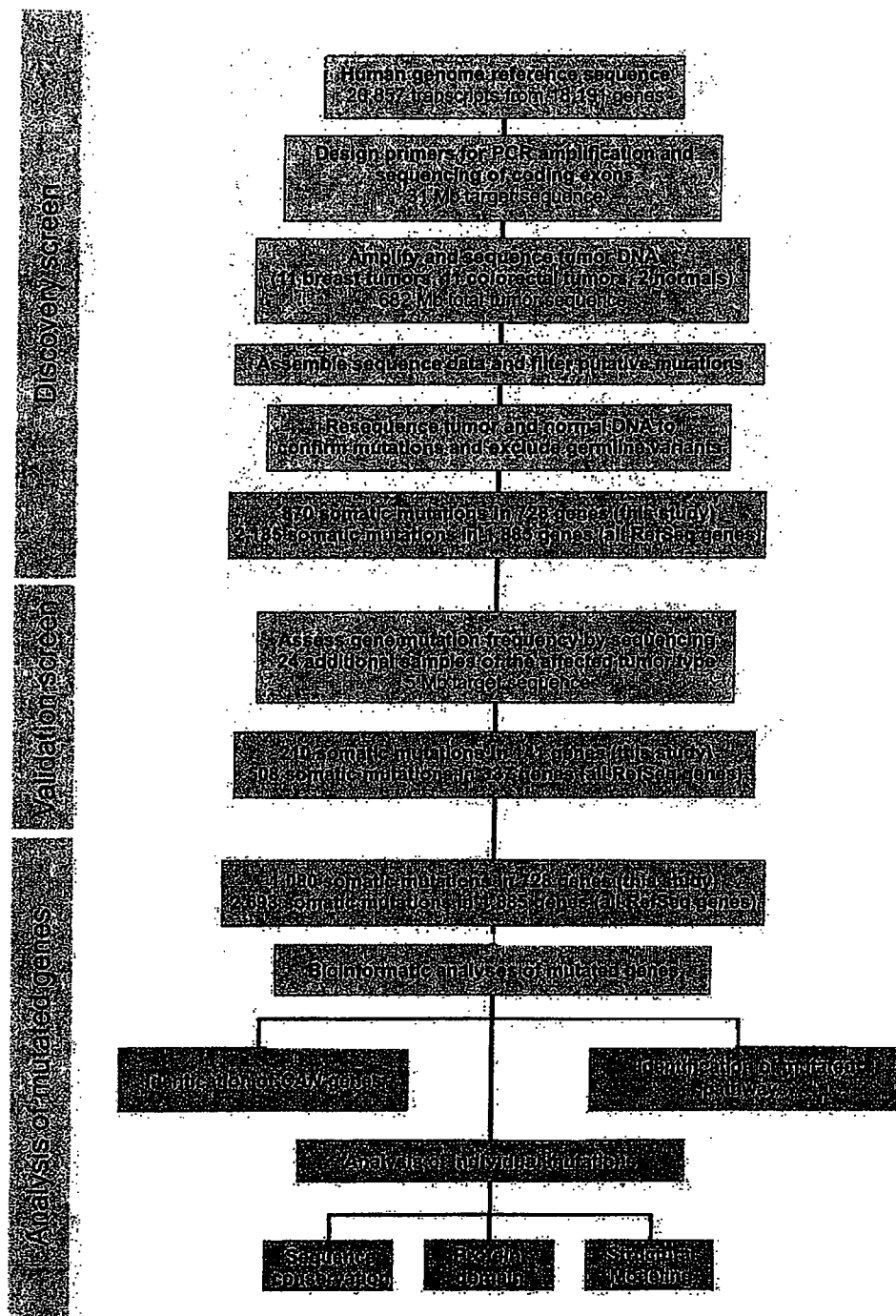


Fig. 4

Table 1. Summary of somatic mutations

Tumor Type	Screen	Gene Set	Mutated genes	Coding changes					Non-coding 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

ND = Not determined because synonymous mutations were not evaluated in the RefSeq genes analyzed in Sjöblom *et al.* (5).

Fig. 5

table S1. Primers used for PCR amplification and sequencing*

Gene	RefSeq accession #	Coding Exon #	Genomic position	Forward primer - Relative position	Forward primer sequence	Reverse primer - Relative position	Reverse primer sequence
A1B0	NM_130786.2	1	chr18:8355581-8355581	-172	M13-CTGAGAGAGTTGACCGCTTTGCC	80	TATGCTGCTTCTGTGACTGC
A1B0	NM_130786.2	2	chr18:8355581-8355581	-227	M13-GAGAGACCAACAGGAGAGAG	86	TGCGCTCTATCACTGACCAAG
A1B0	NM_130786.2	3	chr18:8355581-8355581	-232	GTAGGCTCTGGAGATGAGTGG	93	M13-GTGAGCTTGAGGAGAGGAG
A1B0	NM_130786.2	4	chr18:8355581-8355581	-109	M13-CTGACCTCGCTGTCTCTTAC	91	TCAAGATGATGATGCTGTATC
A1B0	NM_130786.2	5	chr18:8355581-8355581	-103	M13-GAGGCTGTGTCTGTGATG	85	TTAGCTTGATGATGACCTGCT
A1B0	NM_130786.2	6	chr18:8355581-8355581	-94	AGACTGACGACAGGAGAGAG	67	M13-GAGGCTGAGGAGAGAGAG
A1B0	NM_130786.2	7	chr18:8355581-8355581	-123	ACAGGAGAGAGAGAGAGG	103	M13-GAGGCTGAGGCTGACAGG
A1B0	NM_130786.2	8	chr18:8355581-8355581	-88	M13-GATTTCGCGGAGAGAGG	206	ACCCATGACGATTGAGGACAC
A2B1	NM_145891.1	1	chr18:7323303-7323303	-107	M13-ATTTGGGCTCGAGGCTTTGAG	106	CTTCTCACTGATGACGAGAG
A2B1	NM_145891.1	2	chr18:7323303-7323303	-75	GATCTGGGAGAGAGGCTTTGCTG	78	M13-AGAGTGGCTGGATGATGAG
A2B1	NM_145891.1	3	chr18:7323303-7323303	-183	CACGAGATGACAGCTTACG	53	M13-CCGAGACTGATGAGAGAGG
A2B1	NM_145891.1	4	chr18:7323303-7323303	-88	M13-ATGAGAGCTGGGAAACAATAC	154	AGATGAGCCGATGCTTTCTG
A2B1	NM_145891.1	5	chr18:7323303-7323303	-120	M13-GATGTTAGAGAGCTGAGCGAG	73	CATCAAGACCCGATGATGCT
A2B1	NM_145891.1	6	chr18:7323303-7323303	-153	M13-TTTGTTATGACCTGAGCGCTTT	94	AAAGCATGCTGATGATGCTGAG
A2B1	NM_145891.1	7	chr18:7323303-7323303	-133	CCGAGAGAGAGCTGCTTAC	119	M13-TTGGCCAGAGAAATCTTAC
A2B1	NM_145891.1	8	chr18:7323303-7323303	-35	TGATTTCTGTGCTGTTGCTG	18	M13-CCGATGCTGAGGATGATAC
A2B1	NM_145891.1	9	chr18:7323303-7323303	-190	TGATCTGCTGATGCTGCTG	72	M13-ATGAGAGCTGATGATGCTG
A2B1	NM_145891.1	10	chr18:7323303-7323303	-161	AAATGTTCTGCTGCTGCGGAC	161	M13-TACAGGATGAGGACCTCAC
A2B1	NM_145891.1	11	chr18:7323303-7323303	-87	M13-GATTTGCTGATGCTGCTGCTG	121	ACGTTGAGGATGAGGACCTAAC
A2B1	NM_145891.1	12	chr18:7323303-7323303	-125	M13-ATTTGCTGCTGCTGAGAGAG	79	TAAAGCGAGGATGAGGACCTC
A2B1	NM_145891.1	13	chr18:7323303-7323303	-137	M13-ATGAGCTGCTGCTGAGTGG	220	AAAGCTTCACTGAGGAGAGG
A2B1	NM_145891.1	14	chr18:7323303-7323303	-71	M13-ATCAGAGCAATGCGAGAGAG	159	TGCTCTCTTATATGCGAGAGAG
A2B1	NM_145891.1	14	chr18:7323303-7323303	-137	AATGCACTGCGATGAGGTGATG	153	M13-GTCTGCGAGGCTGCTTCTG
A2M	NM_000014	1	chr12:3159625-3159625	-73	M13-TTGGGAGCAGATGATGATG	195	CGTTAGAGAGGCTTGCATATAG
A2M	NM_000014	2	chr12:3159625-3159625	-169	TCTTATGATGAGCTGCTGAC	83	M13-TGCAATCTTACGAGGCTCC
A2M	NM_000014	3	chr12:3159625-3159625	-81	M13-TCCAGGAGAGGCTGACAG	71	CCAAAGTAGGCTCTGAGGCTG
A2M	NM_000014	4	chr12:3159625-3159625	-224	M13-GGAGAGAGCTGCTGCTTGTG	122	TCTTGGCTGCTGCTGCTGCTG
A2M	NM_000014	5	chr12:3159625-3159625	-221	CATGCTGCTGATGATGATGATG	103	M13-GGAGAGTTGAGAGCTTCTCT
A2M	NM_000014	6	chr12:3159625-3159625	-169	M13-GCTTCAAGCTGATGAGG	78	CATGAGCAATGAGAGGCTCTG
A2M	NM_000014	7	chr12:3159625-3159625	-115	M13-TGAAATCTTCAATGAGAGAG	83	TTTGTGTTGCCATGCTGCTG
A2M	NM_000014	8	chr12:3159625-3159625	-165	M13-AGCAGAGAGAGGCTGAGAG	180	TGTGAAAGCTCAAGTGGGCTG
A2M	NM_000014	9	chr12:3159625-3159625	-78	M13-TGACAGAGCTTATGCTGAC	150	GACTGCGCTTCAAGTCAATTC
A2M	NM_000014	10	chr12:3159625-3159625	-199	CCGAGATGCAAGAGAGAGAG	282	M13-GTGAATCTGATGATGCTGCTG
A2M	NM_000014	11	chr12:3159625-3159625	-120	M13-GTCTGATGCTGCTGAGAG	114	CGTTAGATCGCTTGTGATATTC
A2M	NM_000014	12	chr12:3159625-3159625	-178	M13-GATGCTGATGATGATGAG	54	CTTAAAGAGCTTCAATGATGCTG
A2M	NM_000014	13	chr12:3159625-3159625	-181	TGTGAGGCTGATGATGCTG	123	M13-GGAGAGCTTATGATGATGCTG
A2M	NM_000014	14	chr12:3159625-3159625	-119	ATG	117	M13-GTCCGACTGCTGCTGATG
A2M	NM_000014	15	chr12:3159625-3159625	-82	M13-TGGAATGCTTGTGATTAATCA	138	TCTGTACGATGAGAGGCCCA
A2M	NM_000014	16	chr12:3159625-3159625	-88	AGAGGAGAGAGCTGCTGCA	79	M13-CTGCAATGAGAGAGAGG
A2M	NM_000014	17	chr12:3159625-3159625	-92	M13-TGGAATGCTTGTGATTAATCA	160	CAGCGCTAACCATGTACCTCC
A2M	NM_000014	18	chr12:3159625-3159625	-159	AGAAAGCTTCTGTGATGAGCT	175	M13-CAATGCTGAGGATGATGATG
A2M	NM_000014	19	chr12:3159625-3159625	-153	TTCGATCTGCTGCTGCTGATCA	149	M13-TGCTTATGCTGATGATGCTG
A2M	NM_000014	20	chr12:3159625-3159625	-105	TGAGGAGAGCTGATGATGAC	92	M13-AAAGTGCAGCTGTAGTGCT
A2M	NM_000014	21	chr12:3159625-3159625	-127	M13-TGCTGCTGCTGCTGCTGCTG	104	CAGAAAGCAATGAGAGAGAGAGAG
A2M	NM_000014	22	chr12:3159625-3159625	-166	M13-TGCTGCTGAGGCTGCTGCTG	107	ATGCTCATGATGCTGCTGCA
A2M	NM_000014	23	chr12:3159625-3159625	-80	M13-GGAGAGAGCTGCTGCTGCTG	102	AGCCGCAAGAGAGAGAGAGAG
A2M	NM_000014	24	chr12:3159625-3159625	-146	TTCGCTTCAATCATGCTTATCTAC	136	M13-GGAGAGTTGAGAGAGAGAGAG
A2M	NM_000014	25	chr12:3159625-3159625	-175	M13-AGGAGAGCTGCTGCTGCTG	132	TGTGATGATCTGCTGATGCTGCTG
A2M	NM_000014	26	chr12:3159625-3159625	-75	GAGGCTGCTGCTGCTGCTGCTG	83	M13-TGCTGCTGAGGCTTCTGCTG
A2M	NM_000014	27	chr12:3159625-3159625	-106	GGGATGATGATGATGATGATG	105	M13-GAGGCTGCTGCTGCTGCTG
A2M	NM_000014	28	chr12:3159625-3159625	-79	M13-TGTGATGCTGCTGATGCTGCTG	86	CAGAGAGAGAGAGAGAGAGAG
A2M	NM_000014	29	chr12:3159625-3159625	-73	M13-GGCTGCTGCTGCTGCTGCTG	78	TGTGCGCTGCTTGTGATGCTG
A2M	NM_000014	30	chr12:3159625-3159625	-84	M13-AGGCTGCTGCTGCTGCTGCTG	87	GTCGACCTGCTGCTGCTGCTG
A2M	NM_000014	31	chr12:3159625-3159625	-77	M13-TGAGAGAGAGCTGCTGCTGCTG	157	M13-CCGAGAGAGCTGCTGCTGCTG
A2M	NM_000014	32	chr12:3159625-3159625	-86	M13-GGCTGCTGCTGCTGCTGCTG	209	CGAGAGAGCTGCTGCTGCTG
A2M	NM_000014	33	chr12:3159625-3159625	-117	AAAGCATGATGATGATGATGCTG	90	M13-TTCTGAAATCTGATGATGAGAG
A2M	NM_000014	34	chr12:3159625-3159625	-141	TGTGCAATTTAATGAGAGGCTG	54	M13-CCGAGAGAGCTGCTGCTGCTG
A2M	NM_000014	35	chr12:3159625-3159625	-86	CTTCTGCTGCTGCTGCTGCTG	156	M13-TGAGGATGATGAGAGAGAGAG
A2M	NM_000014	36	chr12:3159625-3159625	-155	GATGCTGCTGCTGCTGCTGCTG	119	M13-TTGAATGATGAGAGAGAGAGAG
AAGALT	NM_017436.4	1	chr22:4143330-4144455	-281	M13-TGTGCTGCGAGAGAGAGAG	258	AGCAGATGAGAGAGAGTGGT
AAGALT	NM_017436.4	1	chr22:4143330-4144455	111	ATGCTAGCTGAGAGCTGCTGCTG	570	M13-AAGCAATGAGCTGCTGCTGCTG
AAGALT	NM_017436.4	1	chr22:4143330-4144455	157	ATAGCTGCTGAGAGAGAGAGAG	714	M13-GATGAGAGCTGCTGCTGCTGCTG
AAGALT	NM_017436.4	1	chr22:4143330-4144455	584	CATTGTGCTGAGAGAGAGAGAG	928	M13-GAGAGCTGAGAGAGAGAGAG
AAGALT	NM_017436.4	1	chr22:4143330-4144455	706	ACTGCTGCTGAGAGAGAGAGAG	120	M13-GCTCCAGAGCTGCTGAGAG
AGMT	NM_016161.1	1	chr13:138325800-138326176	-35	TGATTTAAGAGAGAGAGCTGCTG	56	M13-TCTGAGAGCTGCTGCTGCTGCTG
AGMT	NM_016161.1	2	chr13:138325800-138326176	-113	M13-TGAGAGAGAGCTGCTGCTGCTG	4	TGCGAGGAGAGAGAGAGAGAG
AGMT	NM_016161.1	2	chr13:138325800-138326176	80	ATCATGTGAGAGAGAGAGAGAG	2	M13-GCAGAGAGAGAGAGAGAGAG
AGMT	NM_016161.1	2	chr13:138325800-138326176	306	GAGAGCTTCCAGAGAGAGAGAG	100	M13-TGCAATGATGAGAGAGAGAG
AAA1	NM_207288	1	chr7:34664350-34664405	-148	M13-TGTGCTGCTGCTGCTGCTG	132	CAGGAGAGAGAGAGAGAGAG
AAA1	NM_207288	1	chr7:34664350-34664405	-110	M13-TGCTGCTGCTGCTGCTGCTG	164	GCTGTAATCTGCTGAGAGAGAG
AAA1	NM_207288	1	chr7:34664350-34664405	-82	M13-TGTGCTGCTGCTGCTGCTG	208	TAGAGAGAGAGAGAGAGAGAG
AAA1	NM_207288	2	chr7:34664350-34664405	-145	TGCTGCTGCTGCTGCTGCTGCTG	110	M13-CCGAGAGAGAGAGAGAGAG
AAA1	NM_207288	3	chr7:34664350-34664405	-285	TCCAAAGAGAGAGAGAGAGAG	84	M13-GAGAGAGAGAGAGAGAGAG
AAA1	NM_207288	3	chr7:34664350-34664405	-182	CCCTGCTGCTGCTGCTGCTGCTG	94	M13-GCTGCTGCTGCTGCTGCTG
AAA1	NM_207288	3	chr7:34664350-34664405	-144	M13-TGCTGAGAGAGAGAGAGAG	245	GAGATGCTGCTGCTGCTGCTG
AAA1	NM_207288	3	chr7:34664350-34664405	-144	M13-TGCTGAGAGAGAGAGAGAG	79	TGTGATGAGAGAGAGAGAGAG
AAA1	NM_207288	3	chr7:34664350-34664405	-115	M13-CCGAGAGAGAGAGAGAGAG	107	AATGAGAGAGAGAGAGAGAGAG

When a gene gives rise to multiple transcripts, the primers for altered exons are listed under only one transcript name. Coding exons larger than 550 bp were amplified and sequenced with multiple primer pairs. (Position of the 5' and forward PCR primer relative to the first base of the exon, 5'UTR and/or introns indicated by position of the first base of the exon, signal peptide position of the 3' end of the reverse PCR primer relative to the last base of the exon, SH13 denotes the universal sequencing primer 5'-GTGTAAGCGACCGGCGT-3', S indicates that the exons resulting from the primer pair did not meet our quality criteria, and that a 30% increase in the target region have a 3-fold quality score of 20 to 2.75% of the 22 major samples analyzed in the Discovery screen.

Note that this is only the first page of a very large file

Fig. 6

table S2. Distribution of somatic mutations in individual tumors

	C:G→T:A	C:G→G:C	C:G→A:T	T:A→C:G	T:A→G:C	T:A→A:T	5'-CpG-3'	5'-TpC-3'	Insertion	Deletion	Duplication	Total Mutations	Nonsynonymous Mutations	Synonymous Mutations												
Colorectal Cancers																										
Co74	This Study	19	(42.2)	6	(13.3)	8	(17.8)	1	(2.2)	4	(8.9)	5	(11.1)	16	(35.6)	7	(15.6)	0	(0.0)	2	(4.4)	0	(0.0)	45	33	12
	Total RefSeq	43	(35.0)	16	(13.0)	24	(19.5)	9	(7.3)	14	(11.4)	11	(8.9)	32	(26.0)	19	(15.4)	1	(0.8)	3	(2.4)	2	(1.6)	123	111	12
Co92	This Study	18	(60.0)	1	(3.3)	3	(10.0)	2	(6.7)	4	(13.3)	2	(6.7)	14	(46.7)	3	(10.0)	0	(0.0)	0	(0.0)	0	(0.0)	30	21	9
	Total RefSeq	54	(60.0)	5	(5.6)	11	(12.2)	7	(7.8)	8	(8.9)	3	(3.3)	43	(47.8)	8	(8.9)	0	(0.0)	0	(0.0)	1	(1.1)	90	81	9
Co108	This Study	26	(66.7)	0	(0.0)	6	(15.4)	3	(7.7)	1	(2.6)	2	(5.1)	23	(59.0)	2	(5.1)	0	(0.0)	1	(2.6)	0	(0.0)	39	32	7
	Total RefSeq	56	(68.3)	1	(1.2)	6	(7.3)	7	(8.5)	5	(6.1)	5	(6.1)	48	(58.5)	5	(6.1)	1	(1.2)	1	(1.2)	0	(0.0)	82	75	7
Mx22	This Study	15	(68.2)	1	(4.5)	5	(22.7)	1	(4.5)	0	(0.0)	0	(0.0)	10	(45.5)	2	(9.1)	0	(0.0)	0	(0.0)	0	(0.0)	22	18	4
	Total RefSeq	45	(65.2)	5	(7.2)	8	(11.6)	2	(2.9)	2	(2.9)	0	(0.0)	35	(50.7)	4	(5.8)	0	(0.0)	3	(4.3)	4	(5.8)	69	65	4
Mx27	This Study	28	(73.7)	1	(2.6)	8	(21.1)	1	(2.6)	0	(0.0)	0	(0.0)	24	(63.2)	4	(10.5)	0	(0.0)	0	(0.0)	0	(0.0)	38	29	9
	Total RefSeq	56	(65.1)	3	(3.5)	15	(17.4)	4	(4.7)	3	(3.5)	2	(2.3)	50	(58.1)	7	(8.1)	1	(1.2)	0	(0.0)	2	(2.3)	86	77	9
Mx30	This Study	17	(65.4)	1	(3.8)	2	(7.7)	3	(11.5)	0	(0.0)	2	(7.7)	13	(50.0)	2	(7.7)	0	(0.0)	1	(3.8)	0	(0.0)	26	18	8
	Total RefSeq	35	(61.4)	3	(5.3)	5	(8.8)	3	(5.3)	2	(3.5)	3	(5.3)	22	(38.6)	7	(12.3)	0	(0.0)	3	(5.3)	3	(5.3)	57	49	8
Mx32	This Study	17	(58.6)	0	(0.0)	6	(20.7)	3	(10.3)	2	(6.9)	1	(3.4)	14	(48.3)	3	(10.3)	0	(0.0)	0	(0.0)	0	(0.0)	29	23	6
	Total RefSeq	45	(59.2)	2	(2.6)	10	(13.2)	8	(10.5)	7	(9.2)	1	(1.3)	38	(50.0)	5	(6.6)	0	(0.0)	2	(2.6)	1	(1.3)	76	70	6
Mx38	This Study	9	(34.6)	5	(19.2)	4	(15.4)	2	(7.7)	4	(5.2)	6	(7.8)	22	(28.6)	14	(18.2)	1	(1.3)	3	(3.9)	3	(3.9)	77	73	4
	Total RefSeq	31	(40.3)	9	(11.7)	11	(14.3)	9	(11.7)	4	(5.2)	5	(7.8)	22	(28.6)	14	(18.2)	1	(1.3)	3	(3.9)	3	(3.9)	99	91	8
Mx41	This Study	12	(41.4)	2	(6.9)	5	(17.2)	5	(17.2)	4	(13.8)	0	(0.0)	11	(37.9)	4	(13.8)	0	(0.0)	3	(3.0)	0	(0.0)	29	21	8
	Total RefSeq	54	(54.5)	6	(6.1)	16	(16.2)	10	(10.1)	7	(7.1)	3	(3.0)	45	(45.5)	14	(14.1)	0	(0.0)	3	(3.0)	0	(0.0)	99	91	8
Mx42	This Study	11	(37.9)	6	(20.7)	5	(17.2)	5	(17.2)	2	(6.9)	0	(0.0)	11	(37.9)	3	(10.3)	0	(0.0)	0	(0.0)	0	(0.0)	29	24	5
	Total RefSeq	43	(53.1)	8	(9.9)	10	(12.3)	7	(8.6)	3	(6.9)	0	(0.0)	35	(43.2)	6	(7.4)	0	(0.0)	5	(6.2)	1	(1.2)	81	76	5
Mx43	This Study	35	(68.6)	2	(3.9)	9	(17.6)	1	(2.0)	2	(3.9)	0	(0.0)	28	(54.9)	7	(13.7)	0	(0.0)	2	(3.9)	0	(0.0)	51	30	21
	Total RefSeq	72	(70.6)	3	(2.9)	14	(13.7)	3	(5.3)	4	(3.9)	2	(2.0)	57	(55.9)	10	(9.8)	0	(0.0)	3	(2.9)	1	(1.0)	102	81	21
Validation	This Study	75	(57.3)	8	(6.1)	27	(20.6)	7	(5.3)	2	(1.5)	7	(5.3)	58	(44.3)	18	(13.7)	1	(0.8)	2	(1.5)	2	(1.5)	131	101	30
	Total RefSeq	160	(53.5)	22	(7.4)	44	(14.7)	19	(6.4)	13	(4.3)	18	(6.0)	115	(38.5)	42	(14.0)	4	(1.3)	14	(4.7)	5	(1.7)	299	269	30
Breast Cancers																										
B1C	This Study	3	(21.4)	5	(35.7)	2	(14.3)	1	(7.1)	0	(0.0)	2	(14.3)	4	(28.6)	0	(0.0)	0	(0.0)	1	(7.1)	0	(0.0)	14	11	3
	Total RefSeq	9	(22.0)	6	(14.6)	5	(12.2)	6	(14.6)	6	(14.6)	4	(9.8)	6	(14.6)	1	(2.4)	0	(0.0)	4	(9.8)	1	(2.4)	41	38	3
B2C	This Study	28	(35.9)	19	(24.4)	9	(11.5)	4	(5.1)	11	(10.3)	4	(5.1)	25	(14.2)	58	(33.0)	1	(1.3)	5	(6.4)	0	(0.0)	78	56	22
	Total RefSeq	58	(33.0)	49	(27.8)	20	(11.4)	8	(4.5)	11	(6.3)	15	(8.5)	25	(14.2)	58	(33.0)	1	(1.3)	5	(6.4)	0	(0.0)	176	154	22
B3C	This Study	21	(56.8)	4	(10.8)	4	(10.8)	5	(13.5)	1	(2.7)	1	(2.7)	12	(32.4)	4	(10.8)	0	(0.0)	0	(0.0)	0	(0.0)	37	23	14
	Total RefSeq	48	(57.8)	7	(8.4)	6	(7.2)	13	(15.7)	3	(3.6)	2	(2.4)	25	(30.1)	8	(9.6)	0	(0.0)	3	(3.6)	1	(1.2)	83	69	14
B4C	This Study	10	(31.3)	4	(12.5)	5	(15.6)	5	(15.6)	2	(6.3)	2	(6.3)	4	(12.5)	7	(21.9)	0	(0.0)	4	(12.5)	0	(0.0)	32	24	8
	Total RefSeq	22	(28.9)	17	(22.4)	10	(13.2)	8	(10.5)	4	(5.3)	5	(6.6)	8	(10.5)	22	(28.9)	0	(0.0)	10	(13.2)	0	(0.0)	76	68	8
B5C	This Study	17	(42.5)	7	(17.5)	6	(15.0)	5	(12.5)	2	(5.0)	1	(2.5)	7	(17.5)	11	(27.5)	0	(0.0)	2	(5.0)	0	(0.0)	40	29	11
	Total RefSeq	38	(40.0)	20	(21.1)	11	(11.6)	12	(12.6)	6	(6.3)	6	(6.3)	17	(17.9)	27	(28.4)	0	(0.0)	2	(2.1)	0	(0.0)	95	84	11
B6C	This Study	8	(36.4)	5	(22.7)	2	(9.1)	3	(13.6)	1	(4.5)	3	(13.6)	6	(27.3)	2	(9.1)	0	(0.0)	0	(0.0)	0	(0.0)	22	19	3
	Total RefSeq	24	(30.4)	17	(21.5)	16	(20.3)	5	(6.3)	4	(5.1)	4	(5.1)	24	(30.4)	9	(11.4)	2	(2.5)	7	(8.9)	0	(0.0)	79	76	3
B7C	This Study	43	(45.7)	25	(26.6)	6	(6.4)	7	(7.4)	2	(2.1)	3	(3.2)	13	(13.8)	37	(39.4)	0	(0.0)	8	(8.5)	0	(0.0)	94	66	28
	Total RefSeq	80	(36.2)	73	(33.0)	23	(10.4)	17	(7.7)	4	(1.8)	10	(4.5)	30	(13.6)	86	(38.9)	0	(0.0)	14	(6.3)	0	(0.0)	221	193	28
B8C	This Study	14	(34.1)	9	(22.0)	5	(12.2)	0	(0.0)	6	(14.6)	3	(7.3)	5	(12.2)	11	(26.8)	1	(2.4)	2	(4.9)	1	(2.4)	41	33	8
	Total RefSeq	25	(26.3)	20	(21.1)	21	(22.1)	2	(2.1)	8	(8.4)	8	(10.8)	15	(15.8)	23	(24.2)	2	(2.1)	8	(8.4)	1	(1.1)	95	87	8
B9C	This Study	9	(24.3)	7	(18.9)	8	(21.6)	6	(16.2)	2	(5.4)	4	(10.8)	6	(16.2)	10	(27.0)	0	(0.0)	0	(0.0)	0	(0.0)	37	31	6
	Total RefSeq	22	(25.9)	17	(20.0)	18	(21.2)	9	(10.6)	5	(5.9)	8	(9.4)	10	(11.8)	22	(25.9)	0	(0.0)	6	(7.1)	0	(0.0)	85	79	6
B10C	This Study	17	(28.8)	16	(27.1)	7	(11.9)	8	(13.6)	3	(5.1)	5	(8.5)	8	(13.6)	17	(28.8)	0	(0.0)	3	(5.1)	0	(0.0)	59	45	14
	Total RefSeq	37	(28.5)	34	(26.2)	21	(16.2)	17	(13.1)	6	(4.6)	7	(5.4)	16	(12.3)	35	(26.9)	0	(0.0)	8	(6.2)	0	(0.0)	130	116	14
B11C	This Study	19	(36.5)	17	(32.7)	11	(21.2)	3	(5.8)	0	(0.0)	1	(1.9)	5	(9.6)	33	(63.5)	0	(0.0)	1	(1.9)	0	(0.0)	52	38	14
	Total RefSeq	59	(36.4)	65	(40.1)	24	(14.8)	5	(3.1)	0	(0.0)	7	(4.3)	19	(11.7)	104	(64.2)	0	(0.0)	2	(1.2)	0	(0.0)	162	148	14
Validation	This Study	24	(30.4)	8	(10.1)	27	(34.2)	10	(12.7)	1	(1.3)	6	(7.6)	11	(13.9)	9	(11.4)	2	(1.5)	3	(3.8)	0	(0.0)	79	60	19
	Total RefSeq	69	(33.0)	29	(13.9)	53	(25.4)	22	(10.5)	5	(2.4)	12	(5.7)	30	(14.4)	27	(12.9)	2	(1.0)	15	(7.2)	2	(1.0)	209	190	19
ND = Not determined because synonymous mutations were not evaluated in the RefSeq genes analyzed in Spillman et al. (57)																										

ND = Not determined because synonymous mutations were not evaluated in the RefSeq genes analyzed in Sjblom et al. (57)

Fig. 7

8
Fin

Gene	Accession	Tumor	Tumor Type	Screen	Nucleotide (genomic) ¹	Amino acid (protein) ²	Mutation Type	SIPT Score	LogRE Value	LS-SNP
ABCA1	NM_005502.2	Mx29	Colorectal	Validation	g.chr5:146525109>A	p.A2108T	Misense	0.02	0.43	-4.48
ABCA1	NM_005502.2	Mx42	Colorectal	Discovery	g.chr5:146581357>G	p.A2108T	Misense	0.34		
ABCA12	NM_173076	H17C2	Colorectal	Validation	g.chr22:151727280>A (homozygous)	p.E210D	Misense	0.55		
ABCA2	NM_001088.1	B810T	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Synonymous	0.78		
ABCA3	NM_001088.1	B81T	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.39		
ABCA4	NM_000350.1	B1C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.06	0.02	-0.60
ABCA6	NM_060264.2	Mx49	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		
ABCB1	NM_009927.3	Mx41	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01		
ABCB10	NM_012089.1	B4C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	INDEL	0.01	0.48	0.04
ABCB11	NM_003742	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCB11	NM_003742	Mx20	Colorectal	Validation	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCB5	NM_005688.1	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCB8	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCB8	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC5	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC8	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC8	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9	NM_005586	B4C	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.58	0.12	0.36
ABCC9	NM_005586	C674	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.19	-0.23	
ABCC9	NM_005586	B11C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.03	0.68	
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.33		
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	INDEL			
ABCC9	NM_007184.2	B9C	Breast	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.01	0.48	0.04
ABCC9	NM_007184.2	Mx43	Colorectal	Discovery	g.chr16:2237671>C	p.Y266V	Misense	0.05		-0.89
ABCC9										

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

GA17	NIL_005390.2	B5C	B5C22T	Breast	Discovery	g.chr1:3256979A-G	c.239A>G	p.E60G	Misense	0.04	0.18	-1.31
GA81	NIL_002038.2	B5C	B5C22T	Breast	Validation	g.chr1:14489410A-G	c.240A>G	p.V65C	Misense	0.02		
GA81	NIL_002038.2	B5C	B5C22T	Breast	Discovery	g.chr1:14471732C>A	c.1160C>A	p.T37N	Misense	0.01		
GA82	NIL_001078.4	Mx2	Mx2	Colorectal	Discovery	g.chr2:18413460A>C	c.100C>G	p.A210A	Synonymous	0.46		
GABR4S	NIL_000811.1	Mx2	Mx2	Breast	Discovery	g.chr16:43389187C>G (homozygous)	c.1594A>C	p.P31A	Misense	0	0.08	1.22
GABRP	NIL_014211.1	B5C	Mx2	Colorectal	Discovery	g.chr4:4677128A>A	c.630C>A	p.P10H	Misense	0	-0.02	0.68
GALGT2	NIL_03446.1	Mx2	Mx2	Breast	Discovery	g.chr5:16104220C>A	c.1247A>G	p.H46H	Misense	0.52	-0.38	0.18
GALG2	NIL_00101556	B7C	Hs5	Colorectal	Discovery	g.chr17:7071764A>G	c.1378C>A	p.A48H	Misense	0.21	-0.16	-0.51
GALNS	NIL_000512.2	Cv74	Cv74	Colorectal	Discovery	g.chr16:2717460111C2>A	c.556G>T	p.A18S	Misense	0.44	1.41	-0.51
GALNT17	NIL_01034645	B5C	B5C22T	Breast	Discovery	g.chr16:2717460111C2>A	c.192C>A	p.P510T	Misense	0.01	1.88	-0.33
GALNT2	NIL_01034645	B5C	B5C22T	Breast	Discovery	g.chr16:2717460111C2>A	c.191C>T	p.R81W	Indel			
GALNT5	NIL_014584.1	B5C	B5C22T	Breast	Validation	g.chr1:1743358A>T	c.1708A>T	p.A57V	Misense	0.56	-0.01	
GALNTL2	NIL_054110	B7C	B5C	Breast	Discovery	g.chr2:15796588A>T	c.1521A>T	p.L602E	Misense	0.21	1.30	-1.59
GAPL1L1	NIL_024034.3	Mx2	Mx2	Breast	Discovery	g.chr1:1628124G>A	c.1242G>A	p.A14R	Synonymous	0.79	-0.84	
GDF9	NIL_001001557	B11C	Mx2	Colorectal	Discovery	g.chr16:3570826C>G	c.592A>G	p.T197S	Misense	0	-0.46	-0.47
GFI1	NIL_005265	Mx2	Mx2	Breast	Discovery	g.chr16:3570826C>G	c.421C>T	p.R141W	Misense	0		
GFI1B	NIL_004188.2	Mx2	Mx2	Colorectal	Discovery	g.chr1:13289472G>A	c.1081C>C	p.R287R	Synonymous	1		
GGA1	NIL_013385.2	B7C	B5C22T	Breast	Discovery	g.chr1:13289472G>A	c.1450A>G	p.A231H	Misense	0.09	0.63	-0.66
GGA3	NIL_014463.1	B7C	B5C22T	Breast	Discovery	g.chr22:3635152C>G	c.1450A>G	p.P49A	Misense	0.51	0.10	0.60
GGRHR	NIL_000823.1	Mx4	Mx4	Breast	Discovery	g.chr17:70747110C>T (homozygous)	c.1823C>A	p.A44E	Misense	0.42	-0.13	
GMAP9	NIL_176571	B5C	B5C	Colorectal	Discovery	g.chr7:3078176A>G	c.135C>A	p.A19E	Synonymous	0	2.41	1.32
GNOT-1	NIL_163257	B5C	B5C	Breast	Discovery	g.chr7:148653277A	c.4177A	p.A19E	Misense	0		
GPI3	NIL_133281	B5C	B5C	Breast	Discovery	g.chr7:148653277A	c.1806C>T	p.A19E	Misense	0.3	0.03	1.01
GPI4	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.Q48E	Misense	0.04	-0.35	
GPI5	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI6	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI7	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI8	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI9	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI10	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI11	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI12	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI13	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI14	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI15	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI16	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI17	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI18	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI19	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI20	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI21	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI22	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI23	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI24	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI25	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI26	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI27	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI28	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI29	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI30	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI31	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI32	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI33	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI34	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI35	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI36	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI37	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI38	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI39	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI40	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI41	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI42	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI43	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI44	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI45	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI46	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI47	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI48	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI49	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI50	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI51	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI52	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI53	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI54	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI55	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI56	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI57	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI58	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI59	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI60	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI61	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI62	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI63	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI64	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI65	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI66	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI67	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI68	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI69	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI70	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI71	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI72	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI73	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI74	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI75	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI76	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI77	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous	1		
GPI78	NIL_005267	Cv108	Mx4	Colorectal	Discovery	g.chr16:1811611C>G	c.1450C>G	p.P71P	Synonymous			

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

STAT4	NM_001351512	B10C	Discovery	g.chr2:191764870G>C g.chr12:31894843_31894944delCT	c.334G>C c.1789_1790delCT	p.E172Q fs	0	2.57	Misense INDEL
STAT1	NM_019255.1	B6C	Discovery	g.chr10:133850963delT	c.1710delT	p.L710delT	1	0.11	Misense
STAT2	NM_030766.2	MA27	Discovery	g.chr2:27155120G>T	c.1710G>T	p.L710G>T	0	1.35	Misense
STAT3	NM_014387.2	B1C	Discovery	g.chr12:28000000G>C	c.283G>C	p.R283G	0.11	0.18	Misense
STAT4	NM_171424.1	B7C	Discovery	g.chr12:278104280C>G	c.316A>G	p.P316A	1	-0.02	Misense
STAT5A	NM_001019161	MA28	Discovery	g.chr6:598194206A (homozygous)	c.519G>A	p.E519A	0.08	-0.19	Misense
STAT5B	NM_001020640	C07A	Discovery	g.chr13:2351698G>T	c.165G>T	p.Q165H	1	0.08	Synonymous
STAT6	NM_018837.2	B10C	Validation	g.chr15:94737981A>C g.chr15:94737981A>C g.chr20:45172665T>C	c.181A>C c.1717A>C c.189T>C	p.R181H p.L171H p.L189T>C	1	-0.02	Misense
STAT7	NM_018837.2	B10C	Discovery	g.chr2:3721834A>G	c.283G>A	p.E283A	0.08	-0.19	Misense
STAT8	NM_018837.2	B10C	Discovery	g.chr6:4502931C>G	c.287C>G	p.C287G	0.08	-0.19	Misense
STAT9	NM_018837.2	B10C	Discovery	g.chr6:133251194C>G	c.287C>G	p.C287G	0.08	-0.19	Misense
STAT10	NM_018837.2	B10C	Discovery	g.chr10:148444181D>C	c.487C>C	p.E487C	0.08	-0.19	Misense
STAT11	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT12	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT13	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT14	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT15	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT16	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT17	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT18	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT19	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT20	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT21	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT22	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT23	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT24	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT25	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT26	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT27	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT28	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT29	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT30	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT31	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT32	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT33	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT34	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT35	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT36	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT37	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT38	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT39	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT40	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT41	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT42	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT43	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT44	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT45	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT46	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT47	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT48	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT49	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT50	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT51	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT52	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT53	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT54	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT55	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT56	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT57	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT58	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT59	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT60	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT61	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT62	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT63	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT64	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT65	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT66	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT67	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT68	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT69	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT70	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT71	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT72	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT73	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT74	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT75	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT76	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT77	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT78	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT79	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT80	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT81	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT82	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT83	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT84	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT85	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT86	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT87	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT88	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT89	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT90	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT91	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT92	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT93	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT94	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT95	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT96	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT97	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT98	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT99	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT100	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT101	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT102	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT103	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT104	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT105	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT106	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT107	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT108	NM_018837.2	B10C	Discovery	g.chr6:152713327T>G	c.1650G>A	p.L1650H	1	0.08	Misense
STAT109	NM_018837.2	B10C	Discovery	g.chr					

g.chr20:22383335>T (homozygous)	g.chr20:22418400>T g
---------------------------------	---

[illegible]

[illegible]

Accession	Gene	Protein	Position	Variant	Frequency	Population	Phylogenetic	Conservation	Impact	Annotation
ZNF538	NA1_014717	Mad2	101	G>A	100%	CEU	High	0.95	Missense	p.L45V
ZNF540	NM_152093.2	Ccr2	102	C>T	100%	CEU	High	0.95	Missense	p.M276I
ZNF541	NM_032255.1	B2C	103	A>G	100%	CEU	High	0.95	Missense	p.S158L
ZNF548	NM_175544.2	B11C	104	G>T	100%	CEU	High	0.95	Missense	p.L15V
ZNF549	NM_152389	B7C	105	A>G	100%	CEU	High	0.95	Missense	p.D12Y
ZNF560	NM_152476.1	A6a8	106	G>T	100%	CEU	High	0.95	Missense	p.G106E
ZNF568	NM_158539	A6a8	107	A>G	100%	CEU	High	0.95	Missense	p.S173C
ZNF569	NM_152484.2	B8B1	108	G>T	100%	CEU	High	0.95	Missense	p.S173C
ZNF570	NM_152484.2	B8B1	109	G>T	100%	CEU	High	0.95	Missense	p.S173C
ZNF572	NM_152484.2	B8B1	110	G>T	100%	CEU	High	0.95	Missense	p.S173C
ZNF582	NM_154480	A6a8	111	G>T	100%	CEU	High	0.95	Missense	p.C512F
ZNF624	NM_202787.1	A6a1	112	G>T	100%	CEU	High	0.95	Missense	p.A477C
ZNF644	NM_201283.1	B11C	113	G>T	100%	CEU	High	0.95	Missense	p.Y243H
ZNF646	NM_014689.2	B9Z4T	114	G>T	100%	CEU	High	0.95	Missense	p.S432C
ZNF648	NM_014689.2	B11C	115	G>T	100%	CEU	High	0.95	Missense	p.S432C
ZNF659	NM_024697.1	A6a2	116	G>T	100%	CEU	High	0.95	Missense	p.E593Q
ZNF671	NM_024697.1	A6a2	117	G>T	100%	CEU	High	0.95	Missense	p.E593Q
ZNF674	NM_001631.1	A6a2	118	G>T	100%	CEU	High	0.95	Missense	p.A1137I
ZNF674	NM_001631.1	A6a2	119	G>T	100%	CEU	High	0.95	Missense	p.A1137I
ZNF684	NM_001631.1	A6a2	120	G>T	100%	CEU	High	0.95	Missense	p.A1137I
ZNF704	NM_173831	B8C	121	G>T	100%	CEU	High	0.95	Missense	p.H142Q
ZNF714	NM_152515	A6a8	122	G>T	100%	CEU	High	0.95	Missense	p.A384T
ZNF754	NM_153022.1	B5C	123	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	124	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	125	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	126	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	127	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	128	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	129	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	130	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	131	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	132	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	133	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	134	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	135	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	136	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	137	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	138	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	139	G>T	100%	CEU	High	0.95	Missense	p.S568I
ZNF754	NM_153022.1	B5C	140	G>T	100%	CEU	High	0.95	Missense	p.S568I</

table S4A. Colorectal CAN-genes

Gene	Accession	CaMP RANK*	CaMP Score External [†]	Passenger Probability External [†]	CaMP Score SNP [†]	Passenger Probability SNP [†]	CaMP Score NS/S [†]	Passenger Probability NS/S [†]	Number of Mutations [†]	Mutations/ Mb [†]	Number of STOPS, INDELS, splice site mutations [†]	Number of SIFT- significant mutations [†]	Number of LogRE-value- significant mutations [†]	Number of LS-SNP- significant mutations [†]
ABCA1	NM_005902	46	3.76	0.04	3.53	0.17	2.74	0.93	3	12.4	0	1	0	1
ABCF1	NM_003742	90	2.75	0.04	2.75	0.17	2.09	0.93	2	15.2	1	0	0	0
ACSL5	NM_016234	73	3.10	0.04	3.10	0.17	2.41	0.93	2	27.7	0	0	0	0
ADAM19	NM_033274	125	2.22	0.04	1.98	0.17	2.71	0.78	2	21.0	0	0	1	0
ADAM29	NM_014289	58	3.43	0.04	3.43	0.17	2.71	0.88	2	36.9	0	0	0	0
ADAMTS18	NM_199355	21	5.18	0.01	5.00	0.08	3.99	0.89	3	43.7	1	0	0	1
ADAMTS20	NM_175951	100	2.47	0.04	2.47	0.17	1.88	0.91	2	11.1	0	0	0	0
ADAMTSL3	NM_207517	14	5.54	0.01	5.94	0.03	4.80	0.44	4	22.4	0	0	0	0
ADARB2	NM_018702	135	1.68	0.04	1.74	0.17	1.22	0.77	2	30.2	0	1	2	0
AGC1	NM_001135	30	4.40	0.02	4.24	0.12	3.32	0.79	3	16.3	1	2	1	0
AKAP12	NM_005100	115	2.23	0.04	2.24	0.17	1.64	0.85	2	11.5	0	0	0	0
AKAP9	NM_004274	75	3.26	0.04	3.07	0.17	2.45	0.93	2	8.5	0	0	0	0
ALK	NM_004304	58	3.82	0.04	3.45	0.17	2.53	0.93	3	17.8	0	0	0	0
APC	NM_000038	1	65.08	0.00	78.19	0.00	66.76	0.00	33	143.7	32	1	1	1
ARHGEF10	NM_014929	112	2.28	0.04	2.31	0.17	1.68	0.88	2	16.4	0	0	0	0
ATP1A1	NM_032189	78	2.97	0.04	3.00	0.17	2.35	0.93	2	17.1	0	0	0	0
ATP12A1	NM_020410	118	2.14	0.04	2.20	0.17	1.83	0.74	2	15.4	1	1	0	0
ATP13A5	NM_188505	38	3.89	0.04	3.90	0.13	3.20	0.74	2	14.4	0	0	0	0
BCL9	NM_004326	87	2.76	0.04	2.81	0.17	2.20	0.93	2	26.2	2	0	0	0
C10orf137	NM_015608	19	5.46	0.00	5.27	0.06	4.24	0.60	3	28.3	0	0	0	0
C12orf7	NM_024546	57	3.43	0.04	3.45	0.17	2.75	0.88	2	20.0	0	0	0	0
C14orf15	NM_010228	102	2.43	0.04	2.46	0.17	1.85	0.91	2	28.3	0	0	0	0
C16orf2	NM_019558	117	2.43	0.04	2.22	0.17	1.59	0.82	2	20.0	0	0	0	0
C17orf1	NM_012072	69	3.13	0.04	3.13	0.17	2.45	0.93	2	30.4	1	0	0	0
CACNA2D3	NM_016388	33	4.47	0.02	4.00	0.14	3.05	0.85	3	25.8	0	2	2	3
CD109	NM_133483	68	3.18	0.04	3.19	0.17	2.51	0.85	2	14.8	0	0	0	0
CHL1	NM_006814	83	2.89	0.04	2.91	0.17	2.25	0.93	2	14.8	0	0	0	0
CHRF153T	NM_001014372	39	3.99	0.02	3.88	0.10	3.13	0.74	2	20.2	0	0	0	0
CLSTN2	NM_022131	107	2.37	0.04	2.38	0.17	1.75	0.88	2	28.0	0	2	1	0
COL1A1	NM_175813	60	3.38	0.04	3.42	0.17	2.72	0.88	2	13.9	0	0	0	0
COL1A2	NM_000910	118	2.22	0.04	2.24	0.17	1.62	0.85	2	10.3	0	2	2	0
GPAMD9	NM_015692	121	2.91	0.04	2.96	0.17	1.54	0.82	2	19.6	0	0	0	0
GSM3	NM_188123	6	9.93	0.00	8.71	0.00	6.72	0.11	6	12.6	1	0	0	0
CUTL1	NM_001815	97	2.60	0.04	2.51	0.17	1.92	0.83	2	53.9	0	1	1	0
CA4A1	NM_153388	128	2.23	0.04	1.98	0.17	1.38	0.78	2	25.9	0	0	0	0
DPP10	NM_020888	70	3.11	0.04	3.12	0.17	2.46	0.93	2	9.3	0	0	0	0
DSCAML1	NM_020883	140	3.11	0.04	3.12	0.17	2.46	0.93	2	25.9	0	0	0	0
DTNB	NM_153361	113	2.28	0.04	2.30	0.17	1.84	0.85	2	28.2	0	0	0	0
EDD1	NM_015902	48	3.75	0.03	3.62	0.17	2.75	0.80	3	10.3	0	1	1	0
EPHA3	NM_005233	10	7.15	0.00	7.07	0.01	5.70	0.18	4	41.4	0	0	0	0
EPH8	NM_004445	12	6.80	0.00	6.22	0.03	4.83	0.44	4	41.4	0	0	0	0
ERCC8	NM_001124	48	3.61	0.03	3.60	0.16	2.81	0.82	2	17.4	0	2	2	0
EYA4	NM_004100	63	3.24	0.04	3.24	0.17	2.54	0.93	2	26.5	0	1	1	1
FBX1	NM_000132	103	2.42	0.04	2.48	0.17	1.88	0.91	2	6.1	0	0	0	0
FBX2	NM_001969	22	5.18	0.01	4.79	0.10	3.81	0.74	4	13.1	1	0	0	0
FBXW7	NM_033832	5	10.26	0.00	9.64	0.00	7.89	0.02	5	76.7	2	2	3	0
FLJ1306	NM_019957	83	2.58	0.04	2.51	0.17	1.93	0.85	2	46.0	0	0	0	0
FLNC	NM_032180	59	3.39	0.04	3.42	0.17	2.72	0.88	2	43.4	0	0	0	0
FN1	NM_001458	89	2.70	0.04	2.68	0.17	1.99	0.82	2	7.2	1	0	0	0
GALNS	NM_000512	85	2.84	0.04	2.75	0.17	2.17	0.89	2	8.4	0	0	0	0
GLJ3	NM_000168	106	2.36	0.04	2.38	0.17	2.16	0.89	2	34.8	0	1	1	0
GNAS	NM_000516	28	4.33	0.04	4.31	0.11	3.27	0.74	3	13.5	0	0	0	0
GPR112	NM_030752	91	2.83	0.04	2.68	0.17	2.10	0.93	2	76.3	0	3	3	0
GPR168	NM_017551	76	3.02	0.04	3.04	0.17	2.38	0.83	2	8.5	0	0	0	0
GRII1	NM_000638	96	2.52	0.04	2.54	0.17	1.91	0.83	2	17.7	0	0	0	0
GUCY42	NM_000855	132	2.08	0.04	1.85	0.17	1.30	0.77	2	19.7	0	0	0	0
HAPLN1	NM_003847	109	6.06	0.00	5.84	0.03	4.88	0.37	3	52.7	1	2	2	0
HIST1H1B	NM_001864	74	3.06	0.04	3.08	0.17	1.70	0.88	2	11.0	0	0	0	0
HIST1H1B	NM_005322	43	3.77	0.03	3.77	0.14	3.05	0.79	2	55.2	1	1	1	0
IGFBP3	NM_000588	114	2.56	0.04	2.30	0.17	1.61	0.85	2	89.4	0	0	0	0
IGSF22	NM_173588	129	2.15	0.04	1.93	0.17	1.35	0.78	2	27.5	0	1	1	0
IRS4	NM_003804	80	2.93	0.04	2.85	0.17	2.32	0.93	2	21.4	0	0	0	0

Fig. 9

[illegible]

UQCRC2	NM_003366	35	3.97	0.02	3.95	0.12	3.21	0.74	2	39.6	1	0	0	0	0
ZNF282	NM_005095	86	2.50	0.04	2.53	0.17	1.52	0.93	2	15.5	1	1	0	0	1
ZNF442	NM_030324	41	3.80	0.03	3.80	0.13	3.06	0.76	2	32.4	0	0	0	0	0
ZNF221	NM_015481	72	3.11	0.04	3.10	0.17	2.44	0.93	2	17.1	0	0	0	0	0

* Ranking determined with the SNP-based passenger rate, though the rankings were similar regardless of the assumed passenger mutation rate. † Three different CIMP scores and passenger probabilities are listed for each gene, as these values depended on the assumed passenger mutation rate (see text of Supplementary Online Material for details). The five values listed correspond to passenger rates based on external data ("External"), SNPs ("SNP") or NSG ratios ("NSG"). As noted in the text, the External and NSG-based rates are likely to be underestimates and overestimates, respectively, of the true passenger rates. ‡ Amino acid changing mutations per megabase of DNA successfully sequenced in the Discovery and Validation screens. A nucleotide was considered to be successfully sequenced if it had a Phred quality score ≥ 20 . § A mutation was considered SIFT-significant if its score was ≤ 0.05 . ¶ A mutation was considered PolyPhen-significant if its score was ≥ 0.6 . * A mutation was considered LoF-significant if its score was ≤ -0.2 .

table S4B. Breast CAN-genes

Gene	Accession	CaMP RANK*	CaMP Score External†	Passenger Probability External†	CaMP Score SNP†	Passenger Probability SNP†	CaMP Score NS/‡	Passenger Probability NS/‡	Number of Mutations/ Mb†	Number of splice site mutations	Number of SIFT- significant mutations†	Number of LogR-E-value- significant mutations†	Number of LS-SNP- significant mutations†
ABCA3	NM_001089	20	5.16	0.00	4.67	0.04	3.55	0.49	3	18.0	1	0	1
ABCB8	NM_007188	49	3.78	0.01	3.47	0.10	2.70	0.56	2	25.7	1	1	0
ABPI1	NM_001091	44	3.84	0.01	3.58	0.10	2.81	0.56	2	27.3	0	1	0
ACADM	NM_000016	37	4.02	0.01	3.89	0.10	2.81	0.56	2	27.3	0	1	0
ADAM12	NM_003474	11	5.45	0.00	4.98	0.03	3.81	0.46	3	98.1	1	1	1
AEGP	NM_006920	105	3.24	0.02	2.94	0.10	2.21	0.55	2	32.4	0	2	0
AH1	NM_001824	4	7.86	0.00	7.23	0.00	5.73	0.12	2	18.8	0	0	0
AMFR	NM_001144	58	3.86	0.01	3.35	0.10	2.60	0.56	2	24.0	0	1	1
APC2	NM_005883	132	2.81	0.02	2.54	0.11	1.87	0.67	2	32.5	0	1	0
ARFGAP2	NM_004430	121	2.89	0.02	2.72	0.11	2.04	0.56	2	11.1	1	0	1
ARRHGAP25	NM_001007231	94	3.28	0.02	3.03	0.10	2.23	0.56	2	10.7	1	0	1
ASL	NM_001024943	50	3.76	0.01	3.47	0.10	2.23	0.56	2	28.0	1	0	1
ATM1	NM_001840	103	3.28	0.02	2.97	0.10	2.68	0.56	2	40.4	1	0	0
ATP9B1	NM_005603	3	7.93	0.00	7.28	0.00	5.73	0.11	4	22.4	0	2	1
BAT2	NM_006886	136	2.67	0.02	2.41	0.11	1.77	0.57	2	31.6	2	1	0
BCL11A	NM_018014	107	3.21	0.02	2.94	0.10	2.15	0.56	2	9.5	0	0	0
BGN	NM_001711	34	4.11	0.01	3.79	0.09	3.00	0.55	2	31.7	0	0	0
BRCAR1	NM_007286	112	3.10	0.02	2.83	0.11	2.18	0.56	2	55.6	0	0	0
C14orf100	NM_016475	40	3.82	0.01	3.63	0.10	2.80	0.56	2	10.0	0	0	1
C14orf135	NM_023135	35	4.09	0.01	3.76	0.09	2.99	0.55	2	79.7	1	1	0
C14orf165	NM_023135	63	3.41	0.02	3.12	0.10	2.37	0.56	2	31.5	0	0	0
C14orf161	NM_024784	83	3.41	0.01	3.16	0.10	2.37	0.56	2	19.0	0	0	0
C22orf15	NM_005678	81	3.46	0.01	3.16	0.10	2.37	0.56	2	26.3	0	0	0
CACNA1F	NM_005183	96	3.28	0.02	3.00	0.10	2.40	0.56	2	26.3	0	0	0
CDH20	NM_031891	39	3.97	0.01	3.67	0.10	2.32	0.56	2	9.5	0	2	1
CEN7B1	NM_014718	67	3.55	0.01	3.27	0.10	2.91	0.56	2	23.3	0	1	0
CEN7G1	NM_014770	55	3.72	0.01	3.27	0.10	2.51	0.56	2	24.7	0	0	0
CHD5	NM_015557	113	3.08	0.02	2.78	0.11	2.14	0.56	2	24.0	0	0	0
CIC	NM_015125	110	3.15	0.02	2.85	0.11	2.14	0.56	2	10.5	0	0	0
CLCH3	NM_001628	89	3.38	0.02	3.07	0.10	2.32	0.56	2	23.2	1	0	0
CMYA1	NM_194283	91	3.38	0.02	3.07	0.10	2.32	0.56	2	13.5	0	0	0
CNNM4	NM_020184	86	3.41	0.02	3.11	0.10	2.34	0.56	2	28.4	0	1	0
CNTN3	NM_020872	84	3.43	0.01	3.12	0.10	2.38	0.56	2	18.6	0	0	0
CNTN8	NM_014681	85	3.31	0.02	3.00	0.10	2.25	0.56	2	21.0	0	0	0
COL1A1	NM_001654	19	5.15	0.00	4.68	0.04	3.54	0.48	3	18.4	0	0	0
COL8A1	NM_001856	120	3.04	0.02	2.73	0.11	2.01	0.58	2	18.4	0	0	0
COL7A1	NM_000094	27	4.89	0.01	4.30	0.08	3.32	0.54	3	9.0	1	0	0
CPAMD8	NM_015882	6	7.25	0.00	6.30	0.01	4.89	0.23	4	20.9	0	1	1
CSPF1	NM_024790	70	3.52	0.01	3.22	0.10	2.46	0.56	2	24.9	0	0	0
CYP11A1	NM_004395	45	3.84	0.01	3.54	0.10	2.77	0.56	2	36.8	0	0	0
DBH1	NM_004395	9	6.15	0.00	5.62	0.02	4.43	0.24	3	56.9	1	1	1
DPAGT1	NM_001382	52	3.76	0.01	3.48	0.10	2.68	0.56	2	38.8	0	0	0
DPYD	NM_000110	43	3.91	0.01	3.59	0.10	2.83	0.56	2	42.7	0	0	0
FLJ13251	NM_023073	109	3.22	0.02	2.93	0.10	2.23	0.56	2	21.7	0	0	0
FLJ13479	NM_024706	5	7.89	0.00	6.70	0.01	5.14	0.21	4	9.8	0	1	2
FLJ40889	NM_182625	106	3.40	0.02	2.94	0.10	2.13	0.56	2	70.8	0	1	0
FLNA	NM_001456	28	4.58	0.01	4.20	0.10	3.20	0.56	2	24.1	2	0	0
FLNB	NM_001457	134	2.66	0.02	2.46	0.11	1.78	0.57	2	10.9	0	0	0
FLNC	NM_001458	128	2.84	0.02	2.56	0.11	1.96	0.56	2	7.5	1	1	3
FREB1	NM_144986	111	3.12	0.02	2.83	0.11	2.11	0.56	2	7.3	0	0	0
GABY	NM_002039	78	3.48	0.01	3.17	0.10	2.41	0.56	2	14.1	0	0	0
GALNT17	NM_001034445	71	3.50	0.01	3.20	0.10	2.43	0.56	2	26.6	0	0	0
GALNT5	NM_001458	74	3.50	0.01	3.19	0.10	2.44	0.56	2	30.5	0	0	0
GGA1	NM_013365	72	3.50	0.01	3.19	0.10	2.44	0.56	2	22.0	0	1	0
GLI1	NM_005289	87	3.41	0.02	3.09	0.10	2.33	0.56	2	29.5	0	0	0
GRIN2D	NM_000836	15	5.39	0.00	4.89	0.03	3.89	0.46	3	27.9	0	0	0
GSN	NM_001777	101	3.28	0.02	2.98	0.10	2.23	0.56	2	35.2	0	0	0
HBOC4	NM_006037	98	3.28	0.02	2.98	0.10	2.23	0.56	2	22.8	0	0	0
HDLBP	NM_203346	21	5.03	0.00	4.61	0.05	3.47	0.52	3	16.7	1	1	1
HSD3	NM_158631	62	3.67	0.01	3.33	0.10	2.57	0.56	2	23.5	1	1	1
HSPC99	NM_014149	51	3.78	0.01	3.46	0.10	2.67	0.56	2	60.9	0	0	0

Fig. 10

ICAM5	NNL_003259	57	3.66	0.01	3.37	0.10	2.61	0.56	2	25.2	0	0	1	1
INHBE	NNL_031479	48	3.78	0.01	3.47	0.10	2.68	0.56	2	53.7	0	0	0	0
IRTA2	NNL_031281	97	3.29	0.02	2.99	0.10	2.99	0.96	2	19.3	1	0	1	0
JARDYB	NNL_006816	85	3.40	0.02	3.12	0.10	2.40	0.56	2	12.7	0	0	0	0
KEAP1	NNL_203500	61	3.67	0.01	3.34	0.10	2.57	0.56	2	40.2	0	0	0	0
KIAA0427	NNL_014772	41	3.95	0.01	3.63	0.10	2.87	0.56	2	31.5	0	0	1	0
KIAA0467	NNL_015229	131	2.79	0.02	2.54	0.11	1.91	0.56	2	7.8	0	0	0	0
KIAA0664	NNL_01033502	115	3.08	0.02	2.78	0.11	2.07	0.56	2	15.0	0	0	0	0
KIAA0934	NNL_014874	8	6.61	0.00	6.12	0.01	4.55	0.24	2	14.6	0	0	0	0
KIAA0939	NNL_025164	98	3.28	0.02	2.98	0.10	2.24	0.56	2	26.8	2	1	0	0
KIAA1632	NNL_020964	125	3.21	0.02	2.67	0.11	1.98	0.56	2	16.3	0	0	0	0
KIAA1946	NNL_177454	14	5.43	0.00	4.95	0.03	3.81	0.46	3	14.5	0	0	0	0
KIF44	NNL_014875	13	5.45	0.00	4.86	0.03	3.79	0.46	3	34.4	0	0	0	0
KPNA6	NNL_002289	64	3.84	0.01	3.33	0.10	2.56	0.56	2	21.8	0	0	0	0
LAMA1	NNL_005559	69	3.50	0.01	3.22	0.10	2.57	0.56	2	45.7	0	1	1	0
LOC240165	NNL_00102418	7	6.60	0.00	6.16	0.01	4.84	0.20	2	6.3	0	1	0	0
LRBA	NNL_009726	28	4.71	0.01	4.31	0.06	3.31	0.55	3	69.9	0	1	0	0
LRRG7	NNL_020784	114	3.04	0.02	2.76	0.11	2.03	0.56	2	12.8	0	3	2	2
LRRP1P	NNL_004735	117	5.28	0.00	4.76	0.03	3.82	0.46	3	14.4	1	1	0	1
MAGEE1	NNL_020932	65	3.59	0.01	3.29	0.10	2.54	0.56	2	43.4	0	0	0	0
MAIPKBP1	NNL_014984	100	3.27	0.02	2.89	0.10	2.24	0.56	2	27.0	0	0	0	0
MGC24047	NNL_176840	22	4.72	0.00	4.38	0.04	3.54	0.46	2	13.1	0	1	1	0
MYH1A	NNL_005580	54	3.76	0.01	3.45	0.10	2.68	0.56	2	106.3	0	0	0	0
MYH1	NNL_005963	16	5.27	0.00	4.81	0.03	3.72	0.46	3	35.2	0	2	1	1
MYH9	NNL_002473	128	2.91	0.02	2.65	0.11	1.98	0.56	2	14.6	0	1	0	0
MYO1B	NNL_032808	123	2.95	0.02	2.68	0.11	2.04	0.56	2	10.3	0	1	0	0
MYO1G	NNL_033054	60	3.65	0.01	3.35	0.10	2.60	0.56	2	8.0	0	2	0	2
NCB50R	NNL_016230	33	4.20	0.01	3.87	0.08	3.08	0.54	2	18.6	0	0	1	0
NCDA6	NNL_014071	119	3.07	0.02	2.74	0.11	2.04	0.56	2	41.1	0	0	0	0
NOTCH1	NNL_017617	23	4.82	0.01	4.37	0.06	3.35	0.54	3	12.4	0	2	0	2
NUP133	NNL_018230	82	3.45	0.01	3.16	0.10	2.42	0.56	2	18.9	0	2	0	0
NUP214	NNL_003065	127	2.91	0.02	2.65	0.11	1.98	0.56	2	9.1	0	1	0	1
OTOF	NNL_184323	30	4.58	0.01	4.15	0.09	3.13	0.56	3	18.6	0	1	0	0
PDCD11	NNL_018935	78	3.48	0.01	3.17	0.10	2.41	0.56	2	25.0	0	1	0	0
PFC	NNL_004878	138	2.62	0.02	2.40	0.11	1.96	0.56	2	10.1	0	0	0	0
PIK3CA	NNL_002821	63	3.63	0.01	3.33	0.10	2.98	0.56	2	39.1	0	0	0	0
PLD2	NNL_002663	2	10.98	0.00	10.11	0.00	6.18	0.00	5	55.0	0	4	0	0
PLEKHA8	NNL_002659	118	3.04	0.02	2.77	0.11	2.03	0.56	2	20.6	0	0	0	0
PRHIE	NNL_014606	36	3.96	0.01	3.68	0.10	2.85	0.56	2	52.4	0	0	0	0
PRP4B	NNL_000493	42	4.01	0.01	3.62	0.10	2.85	0.56	2	36.2	0	1	0	1
PRP81	NNL_002764	63	3.38	0.02	3.06	0.10	2.30	0.56	2	30.1	0	0	2	0
RAPH1	NNL_215569	92	4.21	0.01	3.88	0.06	3.05	0.54	2	57.8	0	2	1	0
RASAL2	NNL_170982	80	3.38	0.02	3.06	0.10	2.31	0.56	2	20.8	0	0	0	0
RPX2	NNL_008835	90	3.48	0.01	3.17	0.10	2.42	0.56	2	18.2	0	0	0	0
RGL1	NNL_015149	75	3.49	0.01	3.18	0.10	2.32	0.56	2	25.4	0	1	0	0
RHS2	NNL_014677	108	3.46	0.01	2.93	0.10	2.43	0.56	2	22.7	0	1	2	2
RP1L1	NNL_176857	135	2.70	0.02	2.43	0.11	1.76	0.56	2	18.3	0	0	0	0
RPGRIP1	NNL_020388	130	2.87	0.02	2.58	0.11	1.94	0.57	2	11.0	0	0	0	0
SEMA5B	NNL_001031702	117	3.07	0.02	2.76	0.11	2.05	0.56	2	17.1	0	0	0	0
SEPHS2	NNL_012248	56	3.71	0.01	3.44	0.10	2.85	0.56	2	17.7	0	0	0	0
SIX4	NNL_017420	12	5.41	0.00	4.97	0.03	3.76	0.46	3	48.8	0	0	0	1
SLC4A3	NNL_001044	59	3.66	0.01	3.35	0.10	2.58	0.56	2	44.0	0	0	0	0
SLC9A3	NNL_182932	36	4.01	0.01	3.69	0.10	2.83	0.56	2	30.8	0	1	0	0
SLCSA10	NNL_183061	68	3.52	0.01	3.24	0.10	2.50	0.56	2	50.4	0	0	1	0
SORL1	NNL_003105	28	4.64	0.01	4.30	0.06	3.25	0.53	3	13.2	2	1	0	0
SP110	NNL_004509	47	3.91	0.01	3.52	0.10	2.78	0.56	2	26.0	0	0	0	0
SPTAN1	NNL_003127	25	4.72	0.01	4.32	0.07	3.33	0.54	3	11.4	0	0	0	0
STARDB	NNL_014725	73	3.50	0.01	3.19	0.10	2.43	0.56	2	22.5	0	3	1	1
SULF2	NNL_018337	46	3.84	0.01	3.54	0.10	2.77	0.56	2	21.2	0	0	0	0
TACC2	NNL_200862	137	2.65	0.02	2.40	0.11	1.80	0.57	2	9.2	0	0	0	0
TAF1	NNL_004606	140	2.79	0.02	2.34	0.11	1.63	0.59	2	9.8	0	0	0	0
TCF1	NNL_000545	79	3.48	0.01	3.17	0.10	2.41	0.56	2	29.9	0	1	0	0
TDRD8	NNL_001010870	139	2.89	0.02	2.40	0.11	1.71	0.56	2	11.6	0	0	0	0
TECTA	NNL_005422	31	4.68	0.01	4.15	0.03	3.07	0.56	3	14.2	0	1	1	0
TG	NNL_003235	10	5.84	0.00	5.24	0.08	3.85	0.46	4	13.8	1	0	0	0
TIBS3	NNL_007112	77	3.50	0.01	3.17	0.10	2.41	0.56	2	36.2	1	0	0	0

TIMELESS	NM_003520	124	2.86	0.02	2.88	0.11	1.86	0.58	2	15.5	0	1	1	0
TLN1	NM_006289	133	2.98	0.02	2.47	0.11	1.83	0.57	2	7.3	0	1	0	0
TMEM123	NM_052932	24	4.70	0.00	4.33	0.04	3.49	0.46	2	98.0	0	0	0	0
TPRS35	NM_163609	102	3.28	0.02	2.97	0.10	2.23	0.56	2	23.0	0	0	0	0
TPS3	NM_000546	1	55.19	0.00	51.59	0.00	44.15	0.00	16	410.2	8	10	10	11
TRIOBP	NM_001039141	68	3.54	0.01	3.29	0.10	2.57	0.56	2	11.8	0	0	0	0
VEPH1	NM_024821	18	5.34	0.00	4.70	0.04	3.54	0.50	3	33.8	0	0	0	0
XDH	NM_000379	53	3.78	0.01	3.48	0.10	2.71	0.59	2	14.4	0	2	0	0
ZFP14	NM_199427	88	3.39	0.02	3.07	0.10	2.32	0.58	2	28.1	0	2	0	0
ZFP1528	NM_015346	122	3.08	0.02	2.71	0.11	2.04	0.56	2	10.0	0	0	0	0
ZNF669	NM_182484	104	3.28	0.02	2.95	0.10	2.19	0.56	2	34.8	0	0	0	0
ZNF648	NM_014699	116	3.06	0.02	2.78	0.11	2.08	0.56	2	11.5	1	1	1	1

* Ranking determined with the SNP-based passenger rate, though the rankings were similar regardless of the assumed passenger mutation rate. † Three different GAMP scores and passenger probabilities are listed for each gene, as shown values displayed on the assumed passenger mutation rate (see text of Supplementary Online Material for details). The three values listed correspond to passenger rates based on external data ("Edman"), SNP ("SNP"), and NGS ("NGS"). As noted in the text, the Edman and NGS-based rates are likely to be underestimates and overestimates, respectively, of the true passenger rates. ‡ Amino acid changing mutations per megabase of DNA successfully sequenced in the Discovery and Validation screens. A nucleotide was considered to be successfully sequenced if it had a Phred quality score ≥ 20 . § A mutation was considered BFT-significant if its score was ≤ 0.05 . ¶ A mutation was considered logP-E-significant if its score was ≤ 0.05 . ** A mutation was considered LB-significant if its score was ≤ 0.2 .

table S5. Summary of mutation prevalence study

Gene name	Accession number	Total nonsynonymous mutations [*]	Nonsynonymous mutations in 96 additional tumors	Passenger Probability External [†]	Passenger Probability SNP [†]	Passenger Probability NS/S [†]
<i>ABCA1</i>	NM_005502	4	1	0.75	0.96	0.96
<i>ACSL5</i>	NM_016234	2	0	0.39	0.74	0.95
<i>ADAM19</i>	NM_033274	2	0	0.62	0.87	0.95
<i>ADAM29</i>	NM_014269	2	0	0.30	0.69	0.94
<i>ADAMTS18</i>	NM_199355	5	2	0.00	0.07	0.82
<i>ADAMTSL3</i>	NM_207517	9	5	0.00	0.00	0.03
<i>APC</i>	NM_000038	171	138	0.00	0.00	0.00
<i>C10orf137</i>	NM_015608	3	0	0.20	0.73	0.94
<i>CNTN4</i>	NM_175613	3	1	0.02	0.17	0.88
<i>EDD1</i>	NM_015902	4	1	0.99	1.00	0.96
<i>EPHA3</i>	NM_005233	10	6	0.00	0.00	0.00
<i>EPHB6</i>	NM_004445	4	0	0.08	0.58	0.93
<i>EYA4</i>	NM_004100	2	0	0.27	0.62	0.95
<i>FBXW7</i>	NM_033632	14	9	0.00	0.00	0.00
<i>GNAS</i>	NM_000516	5	2	0.00	0.04	0.67
<i>GUCY1A2</i>	NM_000855	7	4	0.00	0.00	0.01
<i>HIST1H1B</i>	NM_005322	2	0	0.07	0.31	0.86
<i>KRAS</i>	NM_004985	79	62	0.00	0.00	0.00
<i>LGR6</i>	NM_021636	3	1	0.05	0.33	0.94
<i>MAP2K7</i>	NM_145185	6	3	0.00	0.00	0.02
<i>MMP2</i>	NM_004530	5	2	0.00	0.02	0.61
<i>NAV3</i>	NM_014903	8	3	0.00	0.01	0.64
<i>OR51E1</i>	NM_152430	4	1	0.00	0.01	0.28
<i>PIK3CA</i>	NM_006218	28	23	0.00	0.00	0.00
<i>PRKD1</i>	NM_002742	5	3	0.00	0.00	0.39
<i>PTEN</i>	NM_000314	4	1	0.00	0.07	0.48
<i>PTPRU</i>	NM_005704	3	0	0.93	0.98	0.96
<i>RET</i>	NM_020975	5	2	0.01	0.17	0.89
<i>RUNX1T1</i>	NM_175634	4	1	0.01	0.16	0.86
<i>SCN3B</i>	NM_018400	2	0	0.09	0.34	0.90
<i>SEC8L1</i>	NM_021807	5	2	0.00	0.03	0.63
<i>SLC29A1</i>	NM_004955	2	0	0.17	0.53	0.93
<i>TBX22</i>	NM_016954	2	0	0.23	0.58	0.95
<i>TCF7L2</i>	NM_030756	10	7	0.00	0.00	0.01
<i>TGM3</i>	NM_003245	4	1	0.02	0.18	0.91
<i>TNN</i>	NM_022093	5	0	0.00	0.11	0.81
<i>TP53</i>	NM_000546	79	61	0.00	0.00	0.00
<i>UHRF2</i>	NM_152896	2	0	0.41	0.77	0.95
<i>UQCRC2</i>	NM_003366	2	0	0.11	0.38	0.87
<i>ZNF442</i>	NM_030824	2	0	0.17	0.51	0.90

^{*}Number of mutations found in the entire dataset, i.e., those identified in the Discovery and Validation Screens plus those found in the study of 96 additional tumors. In the study of the 96 additional tumors, eight synonymous mutations in seven genes were identified in addition to the nonsynonymous mutations listed. [†]Passenger probabilities were determined for the entire dataset. Three different passenger probabilities are listed for each gene, as these values depended on the assumed passenger mutation rate (see text of Supplementary Online Material for details). The three values correspond to passenger rates based on external data ("External"), SNPs ("SNP") or NS/S ratios ("NS/S"). As noted in the text, the External and NS/S-based rates are likely to be underestimates and overestimates, respectively, of the true passenger rates.

Fig. 11

table S6A. Gene groups and pathways preferentially mutated in colorectal cancers

Group or Pathway Name	Mutation enrichment			Gene enrichment			Composite Score*
	Observed Number of Mutations	Basepairs Sequenced (Mb)	Expected Number of Mutations	Group P value	Observed Number of Genes Mutated	Expected Number of Genes Mutated	
Receptor-mediated HIF regulation	14	1,042,383	1.6	2.37E-07	7	36	27.66
Plasmin signaling	19	2,260,162	3.5	6.35E-07	8	46	23.14
Insulin receptor signaling pathway	13	1,104,510	1.7	2.88E-06	6	40	17.84
Cytoskeletal remodeling	35	6,020,047	9.3	1.29E-08	17	184	17.56
TGF- β WNT and cytoskeletal remodeling	27	4,548,793	7.0	7.63E-07	16	173	12.15
ECM remodeling	16	2,774,082	4.3	4.48E-04	9	59	10.97
Insulin regulation of protein synthesis	11	1,383,630	2.1	5.60E-04	7	51	9.58
EPO-induced PI3K/AKT pathway and Ca(2+) influx	16	2,615,695	4.0	2.40E-04	8	66	9.42
Role of HDAC and CaMK in control of skeletal myogenesis	16	2,808,381	4.3	4.89E-04	10	85	8.36
Endothelial cell contacts by non-junctional mechanisms	11	2,258,524	3.5	2.16E-02	7	38	6.59
CoA biosynthesis	5	555,753	0.9	3.61E-02	4	21	5.90
Arachidonic acid production	6	839,235	1.3	3.98E-02	5	35	4.29
anterior/posterior pattern formation	40	807,602	1.2	3.68E-42	5	22	202.17
positive regulation of bone mineralization	4	218,310	0.3	1.07E-02	3	4	31.72
transmembrane receptor protein tyrosine phosphatase signalling pathway	7	634,916	1.0	2.32E-03	3	9	18.86
cell migration	18	1,686,003	2.6	1.73E-06	5	34	18.19
axogenesis	13	1,500,241	2.3	6.56E-05	7	44	14.29
ureteric bud development	5	421,256	0.6	1.43E-02	3	13	9.14
gastrulation (sensu Mammalia)	6	666,436	1.0	1.64E-02	3	13	8.84
response to DNA damage stimulus	7	784,201	1.2	7.76E-03	5	28	8.71
transport	24	6,768,802	10.4	6.10E-03	19	206	4.39
synaptic transmission	20	5,076,526	7.8	5.34E-03	16	179	4.38
Cell adhesion_Cadherins	69	7,197,510	11.1	4.28E-29	23	231	60.65
Cell adhesion_Amyloid proteins	63	8,196,003	12.6	1.23E-21	23	257	40.18
Signal transduction_Androgen receptor signaling cross-talk	21	2,249,733	3.5	2.14E-08	14	96	24.01
Development_Heart development	33	5,701,993	8.8	4.29E-08	12	120	15.82
Proteolysis_Connective tissue degradation	27	4,856,498	7.5	2.46E-06	17	134	15.28
Reproduction_FSH-beta signaling pathway	33	5,466,540	8.4	1.64E-08	19	214	14.84
Cell adhesion_Cell-matrix interactions	36	7,837,765	12.1	1.87E-06	20	175	14.08
Proteolysis_ECM remodeling	22	3,853,738	5.9	2.39E-05	13	95	13.58
Cell adhesion_Attractive and repulsive receptors	38	7,382,208	11.3	4.79E-07	19	201	12.83
Development_NeurogenesisAxonal guidance	39	8,325,958	12.8	3.70E-07	22	240	12.66
Cell adhesion_Integrin printing	27	4,851,724	7.5	2.48E-06	13	130	12.04
Cell adhesion_Platelet-endothelium-leucocyte interactions	28	5,860,391	9.1	2.46E-05	16	170	9.31
Development_Neurogenesis:Synaptogenesis	30	6,604,793	10.2	2.43E-05	21	239	8.70
Reproduction_GnRH signaling pathway	30	6,568,381	10.1	2.20E-05	19	231	8.23
Transmission of nerve impulse	28	6,741,842	10.4	9.27E-05	22	245	7.78
Development_Skeletal muscle development	38	10,287,613	15.8	9.42E-05	18	210	7.41
Proteins that interact with TCF7L2	37	2,336,240	3.6	2.90E-10	11	108	20.85
Proteins that interact with CDS1	29	1,715,524	2.8	1.07E-08	8	43	31.84
Proteins that interact with PKNOX1	14	418,999	0.6	2.52E-07	5	24	28.52
Proteins that interact with PTPRU	6	249,636	0.4	7.23E-02	4	9	10.89

* The composite score indicates the product of the negative log of the CaMP P-value and the fold gene enrichment and is used to rank the pathways or groups within each category.

Fig. 12

table S6B. Gene groups and pathways preferentially mutated in breast cancers

Group or Pathway Name	Mutation enrichment			Gene enrichment			Composite Score*
	Observed Number of Mutations	Basepairs Sequenced (Mb)	Expected Number of Mutations	Group P-value	Observed Number of Genes Mutated	Expected Number of Genes Mutated	
IL9 signaling pathway	12	967,590	1.5	1.98E-05	7	1.7	18.98
VEGF signaling pathway	11	818,253	1.3	2.67E-05	5	1.3	18.11
VEGF signaling via VEGFR2 - generic cascades	15	1,931,557	3.1	6.80E-05	10	2.3	18.00
VEGF signaling and activation	15	1,702,536	2.7	3.01E-05	9	2.3	17.57
NGF activation of NF- κ B	13	1,137,718	1.8	1.59E-05	7	2.1	16.38
Angiopoietin - Tie2 signaling	13	1,329,449	2.1	3.92E-05	8	2.5	16.38
PDGF signaling via STATs and NF- κ B	12	1,212,647	1.9	2.40E-05	7	2.1	13.84
Leptin signaling via PI3K-dependent pathway	14	1,411,982	2.2	6.90E-05	8	2.7	13.77
GDNF family signaling	14	1,528,309	2.4	3.68E-05	9	3.1	13.08
EPO-induced PI3K/AKT pathway and Ca(2+) influx	19	3,247,230	5.1	1.42E-04	11	3.5	12.19
Angiotensin activation of Akt	14	1,640,930	2.8	5.70E-05	8	2.8	12.17
G-Protein alpha-q signaling cascades	15	2,012,064	3.2	1.01E-04	9	3.1	11.78
FAK signaling	18	3,062,547	4.8	2.21E-04	11	3.5	11.41
Role of HDAC and CaMK in control of skeletal myogenesis	20	3,288,134	5.2	5.60E-05	12	4.5	11.41
Fc epsilon RI pathway	16	2,263,461	3.6	9.24E-05	11	3.9	11.40
AKT signaling	13	1,287,030	2.0	3.89E-05	7	2.7	11.28
Arginine metabolism	9	960,285	1.5	1.09E-03	7	1.8	11.25
RAN regulation pathway	12	1,558,128	2.5	5.02E-04	7	2.1	11.25
A2A receptor signaling	13	1,351,207	2.1	4.54E-05	8	2.6	11.19
Cytoskeleton remodeling	32	6,840,728	10.8	2.64E-05	21	8.6	11.14
TOR and CD28 co-stimulation in activation of NF- κ B	12	1,190,956	1.9	6.17E-05	7	2.7	10.77
IP3 signaling	24	4,337,777	6.9	4.02E-05	15	6.2	10.71
FGF-family signaling	11	1,440,891	2.3	9.84E-04	6	1.8	10.08
ERBB-family signaling	12	1,402,983	2.2	2.35E-04	7	2.6	9.85
HGF signaling pathway	11	1,221,093	1.9	3.32E-04	6	2.2	9.44
IL4 signaling pathway	5	1,189,351	1.9	2.65E-04	3	2.3	9.27
Regulation of acetyl-CoA carboxylase 1 activity in keratinocytes	13	455,934	0.7	1.92E-02	8	0.6	8.90
Plasmin signaling	13	2,202,874	3.5	2.27E-03	8	2.4	8.74
PLAU signaling	11	1,222,704	1.9	3.31E-04	6	2.4	8.63
GFGR signaling pathway	13	1,954,350	3.1	8.30E-04	8	2.5	8.52
PIP3 signaling in B lymphocytes	14	2,074,652	3.3	4.47E-04	8	3.2	8.35
Leukocyte chemotaxis	18	3,018,555	4.8	1.85E-04	11	4.9	8.30
Calcium signaling	20	4,086,047	6.5	6.17E-04	13	5.2	8.09
CD28 signaling	13	1,923,208	3.0	7.48E-04	8	2.5	7.31
Chemokines and adhesion	23	5,053,728	8.0	5.05E-04	17	7.8	7.15
Integrin outside-in signaling	18	3,494,871	5.5	7.89E-04	9	4.1	6.80
NFAT in immune response	13	1,945,915	3.1	8.03E-04	8	3.7	6.72
Immunological synapse formation	16	2,746,097	4.3	5.76E-04	10	5.3	6.16
CREB pathway	17	3,143,906	5.0	7.53E-04	10	5.1	6.12
MIF-mediated glucocorticoid regulation	6	691,618	1.1	1.99E-02	5	1.4	5.99
Aspartate and asparagine metabolism	5	514,753	0.8	2.74E-02	4	1.1	5.94
MIF in innate immunity response	9	1,334,099	2.1	8.38E-03	4	2.9	5.84
Notch activating pathway for NF- κ B	8	1,134,672	1.8	1.27E-02	5	1.7	5.46
Anti-apoptotic TNF α /NF- κ B/IAP pathway	6	694,543	1.1	2.00E-02	5	1.6	5.02
Activation of PKC via G-Protein coupled receptor	12	2,246,686	3.5	8.16E-03	10	4.2	4.48
Integrin-mediated cell adhesion	17	4,447,603	7.0	2.05E-02	12	4.5	4.29
ATM/ATR regulation of G1/S checkpoint	8	1,346,431	2.1	2.88E-02	6	2.2	4.08
CXCR4 signaling via second messenger	9	1,561,411	2.5	2.10E-02	7	2.8	3.89
Glucocorticoid receptor signaling	6	881,600	1.4	4.74E-02	6	1.9	3.84
CCR4-induced leukocyte adhesion	9	1,751,872	2.8	3.69E-02	6	2.2	3.69
TLR ligands and common TLR signalling pathway	7	1,020,802	1.6	2.61E-02	6	2.5	3.63
Function MEF2 in T lymphocytes	15	3,787,605	6.0	2.57E-02	11	4.7	3.03
CCR3 signaling in eosinophils	13	3,043,671	4.8	2.64E-02	11	5.7	2.78
TGF β , WNT and cytoskeletal remodeling	18	5,005,986	7.9	2.62E-02	16	9.1	

Fig. 13

Gene Ontology	8	708,014	1.1	8.59E-04	5	23	1.2	4.1	1.18E-02	12.67
B cell differentiation	11	1,339,957	2.1	6.05E-04	6	30	1.6	3.8	8.44E-03	12.23
extracellular matrix organization and biogenesis	5	541,530	0.9	3.22E-02	3	7	0.4	8.1	1.16E-02	12.16
myoblast fusion	72	18,441,949	28.1	5.44E-08	55	807	42.5	1.3	4.52E-02	9.41
signal transduction	31	7,297,477	11.5	1.12E-04	25	205	10.8	2.3	2.97E-04	9.11
transport	13	2,349,971	3.7	3.88E-03	9	48	2.4	3.7	1.57E-03	8.97
cell-cell adhesion	8	863,788	1.4	2.73E-03	4	22	1.2	3.5	3.89E-02	8.86
striated muscle development	21	3,543,870	5.6	4.95E-05	15	141	7.4	2.0	1.23E-02	8.70
positive regulation of transcription from RNA polymerase II promoter	15	2,184,014	3.5	2.24E-04	9	76	4.0	2.3	2.66E-02	8.21
negative regulation of transcription	5	510,816	0.8	2.67E-02	5	20	1.1	4.8	7.23E-03	7.48
spermatid development	32	8,318,158	13.1	4.08E-04	27	234	12.3	2.2	3.63E-04	7.43
cell adhesion	11	1,594,879	2.5	2.15E-03	7	56	2.9	2.4	3.78E-02	6.34
regulation of transcription	6	751,085	1.2	2.61E-02	4	20	1.1	3.8	2.98E-02	6.02
actin filament organization	13	2,824,711	4.1	9.13E-03	9	81	4.3	2.1	3.67E-02	4.31
DNA repair	47	9,612,283	15.2	7.96E-06	29	205	10.8	2.7	8.74E-06	19.09
Cytoskeleton_Actin filaments	30	4,834,208	7.8	9.69E-07	21	144	7.8	2.8	9.37E-05	16.67
Inflammation_IgE signaling	37	7,101,319	11.2	1.07E-06	25	201	10.6	2.4	2.13E-04	14.11
Immune_Phagocytosis	22	3,288,115	5.2	1.05E-05	16	109	5.7	2.8	5.99E-04	13.89
Inflammation_IFN-gamma signaling	32	5,888,712	9.5	4.00E-06	23	181	9.5	2.4	2.81E-04	13.03
Inflammation_Kallicrein-kinin system	16	1,884,095	2.7	9.12E-06	10	76	4.0	2.5	1.10E-02	12.60
Inflammation_IL-10 anti-inflammatory response	44	11,372,301	18.0	2.82E-05	29	210	11.1	2.6	1.31E-05	11.94
Development_Skeletal muscle development	28	4,878,880	7.7	5.62E-06	19	160	8.4	2.3	1.82E-03	11.85
Development_Hemopoiesis, Erythropoietin pathway	33	6,146,432	8.7	2.43E-06	24	228	12.0	2.0	2.24E-03	11.23
Immune_Phagosome in antigen presentation	41	9,606,351	15.2	1.05E-05	27	236	12.4	2.2	4.09E-04	10.82
Cell adhesion_Integrin-mediated cell-matrix adhesion	18	2,702,922	4.3	5.72E-05	12	95	5.0	2.4	7.54E-03	10.18
Inflammation_IL-2 signaling	19	2,829,237	4.5	3.76E-05	13	108	5.7	2.3	7.94E-03	10.12
Inflammation_TREM1 signaling	18	2,568,081	4.1	3.71E-05	12	105	5.5	2.2	1.48E-02	9.82
Signal transduction_Leptin signaling	33	3,168,254	10.9	1.53E-05	23	231	12.2	1.9	5.04E-03	9.11
Reproduction_GnRH signaling pathway	20	3,681,860	5.0	3.95E-05	13	121	6.4	2.0	1.76E-02	8.98
Inflammation_Amphoterin signaling	27	5,541,861	8.8	5.44E-05	21	191	10.1	2.1	2.54E-03	8.91
Cell adhesion_Synaptic contact	32	7,119,304	11.2	3.78E-05	20	206	10.8	1.8	1.07E-02	8.16
Cell adhesion_Platelet aggregation	21	4,751,871	7.5	1.40E-03	16	114	6.0	2.7	8.60E-04	7.61
Development_Neuromuscular junction	39	9,551,315	15.1	3.35E-05	23	262	13.8	1.7	1.82E-02	7.47
Cytoskeleton_Regulation of cytoskeleton rearrangement	34	8,498,041	13.4	1.29E-04	24	240	12.6	1.9	4.05E-03	7.39
Development_NeurogenesisAxonal guidance	21	3,664,808	5.8	6.74E-05	15	162	8.5	1.8	3.38E-02	7.34
Immune_TCR signaling	25	5,437,437	8.6	2.38E-04	20	188	9.9	2.0	4.41E-03	7.32
Inflammation_Innate inflammatory response	30	6,519,243	10.3	4.71E-05	21	239	12.6	1.7	2.29E-02	7.23
Development_NeurogenesisSynaptogenesis	25	5,607,944	8.9	1.53E-04	16	166	8.7	1.8	2.17E-02	6.99
Development_Angiogenesis	35	8,435,822	13.3	5.59E-05	20	170	8.9	2.2	1.55E-03	6.97
Cell adhesion_Platelet-endothelium-leucocyte interactions	31	8,198,102	12.9	6.19E-04	30	380	20.0	1.5	2.71E-02	6.38
Reproduction_Spermatogenesis, motility and copulation	20	4,454,813	7.0	1.66E-03	21	201	10.8	2.0	4.35E-03	6.37
Cell adhesion_Attractive and repulsive receptors	18	3,247,888	5.1	4.13E-04	12	134	7.1	2.3	3.77E-03	6.31
Proteolysis_Connective tissue degradation	19	3,896,649	6.2	9.28E-04	14	124	6.5	1.8	4.13E-02	6.22
Immune_BCR pathway	23	5,027,462	7.9	4.93E-04	16	171	9.0	1.8	2.70E-02	5.98
DNA damage_Checkpoint	14	2,554,852	4.0	2.63E-03	11	95	5.0	2.2	1.76E-02	5.88
Neurophysiological process_Visual perception	38	11,249,850	17.8	8.03E-04	28	302	15.9	1.8	6.34E-03	5.45
Muscle contraction	25	5,769,709	9.1	5.10E-04	18	214	11.3	1.6	4.58E-02	5.28
Reproduction_FSH-beta signaling pathway	22	5,430,633	8.6	2.73E-03	14	130	8.8	2.0	1.39E-02	5.25
Cell adhesion_Integrin priming	19	4,100,873	6.5	1.66E-03	16	180	9.5	1.7	3.91E-02	4.69
Inflammation_Histamine signaling	12	2,141,358	3.4	5.63E-03	11	103	5.4	2.0	2.84E-02	4.57
Inflammation_Inflammasome	29	8,118,458	12.8	2.28E-03	22	257	13.5	1.6	2.57E-02	4.30
Cell adhesion_Amyloid proteins	25	6,651,916	10.5	2.93E-03	20	224	11.8	1.7	2.25E-02	4.30
Cell cycle_G2-M	8	1,235,946	2.0	1.99E-02	7	57	3.0	2.3	4.07E-02	3.97
Inflammation_IL-12,15,18 signaling	20	129,862	0.2	7.44E-23	2	3	0.2	12.7	1.23E-02	280.36
Proteins that interact with PRODH	25	779,898	1.2	4.00E-13	6	12	0.6	9.5	5.13E-05	117.80
Proteins that interact with ERCC6	48	5,202,381	8.2	4.34E-06	21	167	8.8	2.4	3.04E-03	12.82
Proteins that interact with PARP1	32	2,872,800	4.5	1.48E-05	12	101	5.3	2.3	3.24E-02	10.90
Proteins that interact with CSNK1D	16	1,634,469	2.6	9.42E-02	8	39	2.1	3.9	3.30E-03	4.00
Proteins that interact with GRIK3										

* The composite score indicates the product of the negative log of the CaMP P-value and the fold gene enrichment and is used to rank the pathways or groups within each category.