Supporting Information

Leary *et al.* 10.1073/pnas.0808041105

SI Methods

Clinical Samples and Cell Lines. DNA samples were obtained from xenografts and cell lines of ductal breast and colorectal carcinoma. Normal DNA samples were obtained from matched normal tissue or peripheral blood. Twenty two of the DNA samples included those used in the Discovery Screen of Sjöblom *et al.* and Wood *et al.* (1, 2). All tumor samples analyzed for copy number analyses are listed in Table S7. For the Illumina analyses, the colorectal cancer samples used were 10 cell lines and 26 xenografts, each developed from a liver metastasis of a different patient. The breast cancer samples used were 22 cell lines and 23 xenografts, each developed from a different patient. In addition, 11 colorectal cancer metastases [immunopurified using the BerEP4 antibody as described (3)], and seven cell lines were analyzed by Digital Karyotyping analyses. Available clinical information for samples that were analyzed by copy number and sequence analyses is available in table S2 of reference 2. All samples were obtained in accordance with the Health Insurance Portability and Accountability Act (HIPAA).

Digital Karyotyping. Digital Karyotyping libraries were constructed as described (4, 5). In brief, 17-bp tags of genomic DNA were generated using the NlaIII mapping and SacI fragmenting restriction enzymes. For each library, the experimental tags obtained were concatenated, cloned and sequenced. SAGE2002 software was used to extract the experimental tags from the sequencing data. The sequences of the experimental tags were compared to the predicted virtual tags extracted from the human genome reference sequence hg16 (NCBI Build 34, July 2003) and were visualized using the SageGenie DKView to identify potential alterations (http://cgap.nci.nih.gov/SAGE/DKViewHome). The coordinates of all identified alterations were translated to the human genome reference sequence hg17 (NCBI Build 35, May 2004) to allow comparison to Illumina data.

Homozygous deletions were identified using a sliding window size of 175 virtual tags (≈700 kb in size). Windows with a tag density ratio (observed tags in window/expected tags in window) <0.01 were considered to represent putative homozygous deletions and were further examined. Regions of homozygous deletions were defined as containing no experimental tags and the boundaries were determined as the outermost virtual tags with no matching experimental tags.

Amplifications were identified using sliding windows of variable sizes, as the most accurate window size for detection and quantification of amplifications is the exact size of the altered region. Windows with tag density ratios ≥ 6 were considered to represent amplified regions. Boundaries of the amplified region were determined by the outermost tag contained in a window with a tag density ratio >3 or by the virtual tag position after which there is sharp decline in the observed experimental tags.

High-Density SNP Arrays. The Illumina Infinium II Whole Genome Genotyping Assay employing the BeadChip platform was used to analyze tumor samples at 317,503 (317k), 555,351 (550k V1), or 561,466 (550k V3) SNP loci from the Human HapMap collection. All SNP positions were based on hg17 (NCBI Build 35, May 2004) version of the human genome reference sequence. The genotyping assay is a two step procedure that is based on hybridization to a 50 nucleotide oligo, followed by a two-color fluorescent single base extension. The image files of fluorescence intensities were processed using Illumina BeadStation software to provide intensity values for each SNP position. For each SNP, the normalized experimental intensity value (R) was compared to the intensity values for that SNP from a training set of normal samples and represented as a ratio (called the "Log R Ratio") of $\log_2(R_{\text{experimental}}/R_{\text{training set}})$.

Bioinformatic Analysis of High-Density SNP Array Data. Digital Karyotyping was used to inform and optimize the criteria for detection of focal homozygous deletions and high-copy amplifications using the Illumina arrays. Three colorectal cancer samples (Co44, Co82, and Co84) were assessed by Digital Karyotyping tag libraries as well as the Illumina arrays (Table S1). From these analyses criteria were developed to permit sensitive and specific detection of the Digital Karyotyping alterations using the Illumina platform as described below. These criteria were subsequently used to analyze an additional 46 breast and 33 colorectal cancers.

Detection of Homozygous Deletions. Homozygous deletions (HDs) were defined as two or more consecutive SNPs with a Log R Ratio value of ≤ -2. The first and last SNPs of the identified HD region were considered to be the boundaries of the alteration for subsequent analyses. The deletion breakpoint would be expected to be located between the boundary deleted SNPs and adjacent non-deleted SNPs; use of the inner deleted SNP boundaries provides the most conservative approach as use of the outer boundaries may include non-deleted regions. To eliminate chip artifacts and potential copy number polymorphisms, we removed all HDs that were included in copy number polymorphism databases (6, 7). As these analyses showed that copy number polymorphisms had conserved boundaries, we also removed all observed HDs with identical boundaries that occurred in multiple samples. Adjacent homozygous deletions separated by one or two SNPs were considered to be part of the same alteration. Adjacent HDs were evaluated separately for the purposes of determining affected genes, but were counted as single entries in Table 2 and Table S3. To identify genes affected by HDs, we compared the location of coding exons in the RefSeq and CCDS databases with the genomic coordinates of the observed HDs. Any gene with a portion of its coding region contained within a homozygous deletion was considered to be affected by the deletion.

Detection of Amplifications. High copy amplifications (i.e., >12 chromosomal copies as determined by Digital Karyotyping) were defined as regions having at least one SNP with a LogR ratio ≥ 1.4, at least one in ten SNPs with a LogR ratio ≥ 1, and an average LogR ratio of the entire region of ≥ 0.9. The boundaries of amplified regions were delimited by the outermost SNPs with LogR ratios ≥ 1. Similar to analyses of homozygous deletions, we removed all amplifications that had identical boundaries in multiple samples.

Because focal amplifications are more likely to be useful in identifying specific target genes, a second set of criteria were used to remove large chromosomal regions or entire chromosomes that showed copy number gains. These large alterations, called "complex amplifications," were thus distinguished from small focal alterations, called "simple amplifications." Based on observations from Digital Karyotyping, several steps were used to identify and remove complex amplifications. First, amplifications >3Mb in size and groups of nearby amplifications (within 1 Mb) that were also >3Mb in size were considered complex. Amplifications or groups of amplifications that occurred at a frequency of ≥4 amplifications in a 10Mb region, or ≥5 amplifications per chromosome were deemed to be complex. The amplifications remaining after these filtering steps were considered to be simple amplifications and were further examined. The complex regions were not included in subsequent statistical analyses but those containing candidate cancer genes are indicated in Table 1. To identify protein coding genes affected by amplifications, we compared the location of the start and stop positions of each gene within the RefSeq and CCDS databases with the genomic coordinates of the observed amplifications. As amplifications of a subgenic region (i.e., containing only a fraction of a gene) are less likely to have a functional consequence, we focused our analyses on genes whose entire coding regions were included in the observed amplifications.

A number of genes coamplified or codeleted with known oncogenes (CCND1, ERBB2, CCNE1, EGFR, MYC) or tumor suppressors (CDKN2A, PTEN, MAP2K4, TP53) were considered "known passengers" and eliminated from further statistical analysis. However, for completeness, these known passengers were listed along with their respective copy number alterations in Tables S3–S5, but these alterations were not used to calculate the passenger probabilities listed in Tables S4 and S5. Alterations of known passengers were also excluded from statistical analysis of pathways (Table S6).

Statistical Analysis of Deletions and Amplifications. For each of the genes involved in amplifications or deletions, we quantify the strength of the evidence that they may be drivers of carcinogenesis by reporting a driver probability, separately for amplifications and deletions. In each case, the passenger probability is an *a posteriori* probability that integrates information from the somatic mutation analysis of Wood *et al.* (2) with the data presented in this article. The passenger probabilities reported in Wood *et al.* (2) serve as *a priori* probabilities. These are available for three different scenarios of passenger mutation rates and results are presented separately for each. If a gene was not found to be mutated in Wood *et al.* (2) the prior passenger probability is set to the estimated proportion of passengers in the RefSeq set. Then, a likelihood ratio for "driver" versus "passenger" was evaluated using as evidence the number of samples in which a gene was found to be amplified (or deleted). Analysis is carried out separately by type of array, and then combined by multiplication of the relevant likelihood terms. The passenger term is the probability that the gene in question is amplified (deleted). For each sample, we begin by computing the probability that the observed amplifications (deletions) will include the gene in question by chance. Inclusion of all available SNPs is required for amplification, while any overlap of SNPs is sufficient for deletions. Specifically, if in a specific sample N SNPs are typed, and K amplifications are found, whose sizes, in terms of SNPs involved, are A_1 . A_K, a gene with G SNPs will be included at random with probability

$$(A_1\text{-}G + 1)/N + .. + (A_K\text{-}G + 1)/\mathrm{N}$$

for amplifications and

$$(A_1 + G\text{-}1)/N + .. + (A_K + G\text{-}1)/\mathrm{N}$$

for deletions

We then compute the probability of the observed number of amplifications (deletions) assuming that the samples are independent but not identically distributed Bernoulli random variables, using the Thomas and Taub algorithm (8), as implemented in R by M. Newton. Our approach to evaluating the passenger probabilities provides an upper bound, as it assumes that all of the deletions and amplifications observed only include passengers. The driver term of the likelihood ratio was approximated as for the passenger term, after multiplying the sample-specific passenger rates above by a gene-specific factor reflecting the increase (alternative hypothesis) of interest. This increase is estimated by the ratio between the empirical deletion rate of the gene and the expected deletion rate for that gene.

For each of the gene sets considered we quantify the strength of the evidence that they may include a higher-than-average proportion of driver genes. For each set, in a list of all of the RefSeq genes sorted by a score combining information on mutations, amplifications and deletions, we compared the ranking of the genes contained in the set with the ranking of those outside, using the rank-sum test, as implemented by the Limma package in Bioconductor (9). Scores were obtained by adding three log likelihood ratios for mutations, amplifications and deletions. This combination approach makes an approximating assumption of independence of amplifications and deletions. In general, amplified genes cannot be deleted, so independence is technically violated. However, because of the relatively small number of dramatic amplification and deletions, this assumption is tenable for the purposes of gene set analysis. Inspection of the log likelihoods suggest that they are roughly linear in the number of events, supporting the validity of this approximation as a scoring system. The statistical significance of deviation from the null hypothesis of a random distribution was calculated using Limma and then corrected for multiplicity by the q-value method (10) as implemented in version 1.1 of the package "q-value."

1. Sjoblom T, *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274.
2. Wood LD, *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
3. Saha S, *et al.* (2001) A phosphatase associated with metastasis of colorectal cancer. *Science* 294:1343–1346.
4. Wang TL, *et al.* (2002) Digital karyotyping. *Proc Natl Acad Sci USA* 99:16156–16161.
5. Leary RJ, Cummins J, Wang TL, Velculescu VE (2007) Digital karyotyping. *Nat Protoc* 2:1973–86.
6. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 38:75–81.
7. Sebat J, *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* 305:525–528.
8. Thomas MA, Taub AE (1982) Calculating binomial probabilities when the trial probabilities are unequal. *J Stat Comp Simul* 14:125–131.
9. Smyth GK (2005) in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds Gentleman V, Carey S, Dudoit R, Irizarry WH (Springer, New York), pp 397–420.
10. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100:9440–9445.

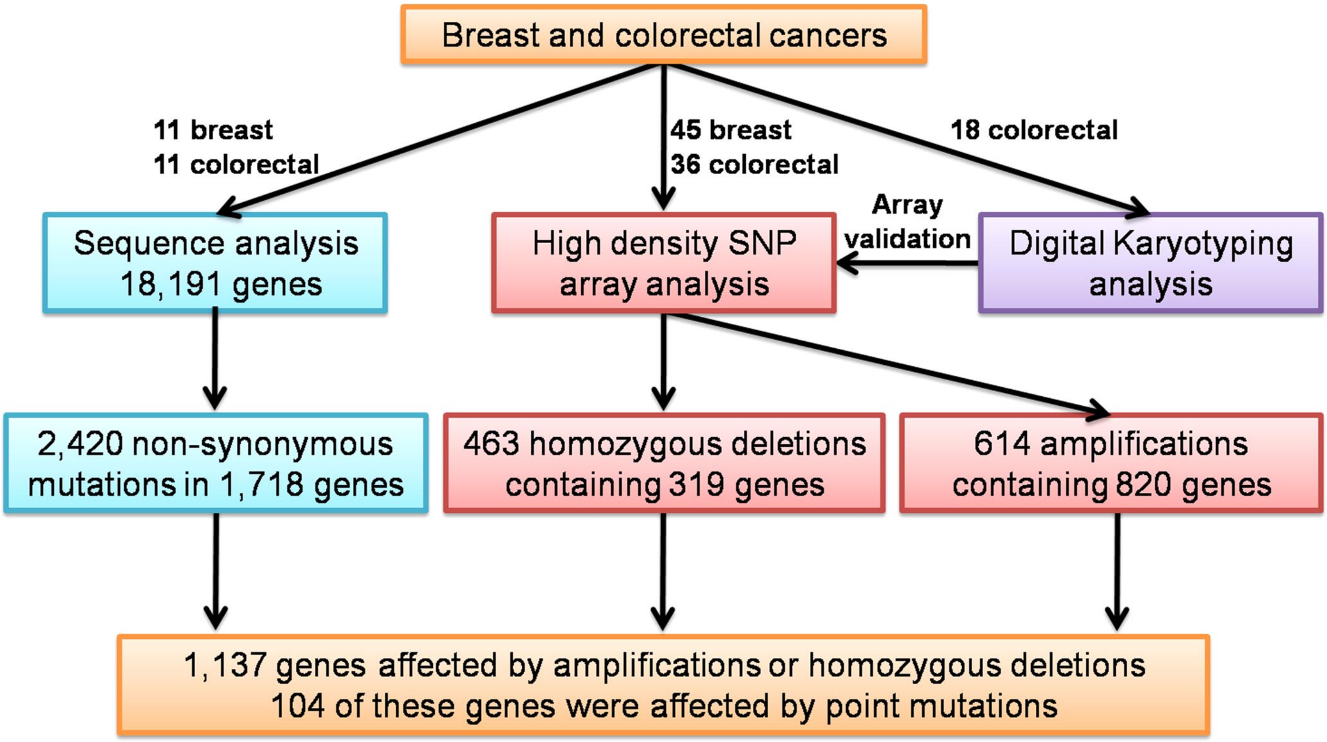

Fig. S1. Schematic of experimental approach for integration of copy number and sequence alterations in breast and colorectal cancers.

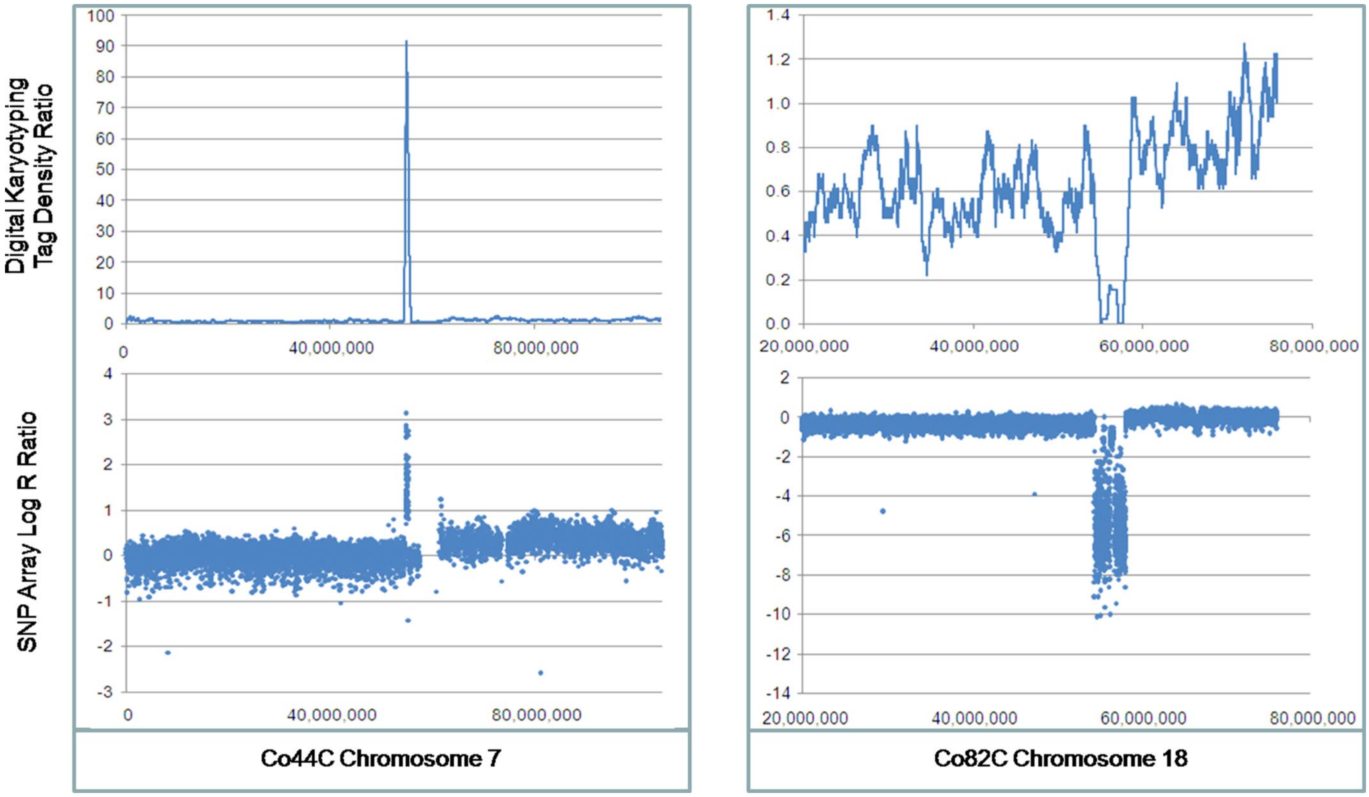

Fig. S2. Detection of amplifications and homozygous deletions using Illumina arrays and Digital Karyotyping. Digital Karyotyping results are shown in the top graphs, with the chromosomal coordinates indicated on the horizontal axis and the Digital Karyotyping tag density ratio indicated on the vertical axis. Illumina array results are shown in the bottom graphs, with the chromosomal coordinates indicated on the horizontal axis and the Log R Ratio indicated on the vertical axis. Digital Karyotyping data were used to validate the Illumina arrays and to develop approaches for sensitive and specific detection of focal amplifications and homozygous deletions.

Table S1. Comparison between Illumina array and Digital Karyotyping copy number analyses

Alteration	Tumor		Digital Karyotyping			Tag density	Illumina SNP arrays		Size,	
type	sample	Chr	Left boundary	Right boundary	Size, bp	ratio†	Left boundary	Right boundary	bp	Log R ratio†
HD	Co44C	5	59,059,409	59,807,807	748,399	0.0	59,109,232	59,522,525	413,294	−8.3
Amplification	Co84C	6	41,273,307	43,008,812	1,735,506	9.1	41,419,345	42,485,546	1,066,202	1.9
Amplification	Co44C	7	54,856,760	55,409,704	552,945	92.5	54,862,624	55,406,733	544,110	3.1
Amplification	Co84C	8	127,618,526	128,009,287	390,762	19.2	127,621,008	127,995,012	374,005	2.7
Amplification*	Co84C	8	128,750,189	128,857,861	107,673	8.3	128,750,181	128,848,183	98,003	2.0
Amplification	Co84C	11	34,337,207	35,266,401	929,195	33.0	34,359,268	35,265,359	906,092	3.0
Amplification*	Co82C	12	30,734,351	32,018,350	1,284,000	6.2	30,702,093	32,036,123	1,334,031	2.2
Amplification	Co84C	13	109,096,557	109,553,930	457,374	9.2	109,108,212	109,557,712	449,501	2.3
Amplification	Co84C	15	88,545,070	89,258,106	713,037	26.2	88,561,995	89,253,599	691,605	3.6
HD	Co82C	18	54,490,515	56,426,158	1,935,644	0.0	54,517,561	56,407,631	1,890,071	−10.1
HD	Co82C	18	56,629,615	58,250,812	1,621,198	0.0	56,875,085	58,225,845	1,350,761	−9.5
Amplification	Co84C	19	34,570,450	34,641,949	71,500	7.9	34,561,976	34,641,548	79,573	2.2
Amplification	Co84C	19	34,956,853	35,344,522	387,670	14.3	34,966,463	35,321,409	354,947	2.6
Amplification*	Co82C	19	43,386,048	45,698,030	2,311,983	8.4	43,834,169	45,620,784	1,786,616	2.4
Amplification*	Co84C	19	54,500,237	54,643,655	143,419	8.4	54,520,709	54,622,533	101,825	2.1

*Starred alterations indicate those that are identified by Digital karyotyping which are represented by multiple smaller amplifications on the Illumina arrays.

†Values for Tag Density Ratios and Log R Ratios represent observed maximum values for amplifications and minimum values for homozygous deletions.

Table S2. Copy number changes detected by Digital Karyotyping in colorectal cancer

Sample	Tags analyzed	Type of alteration	Chr	Left boundary, bp	Right boundary, bp	Size	Window size (tags)	Tag density ratio*	Number of affected genes
M10–23	113,163	Amplification	1	142,456,410	142,686,517	230,108	50	7.1	1
M12–05	60,438	HD	4	166,843,261	168,228,892	1,385,632	300	0.0	2
Co44C	114,462	HD	5	59,059,409	59,807,807	748,399	175	0.0	0
Co84C	441,113	Amplification	6	41,273,307	43,008,812	1,735,506	50	9.1	26
Co44C	114,462	Amplification	7	54,856,760	55,409,704	552,945	50	92.5	3
Co37C	74,314	Amplification	7	99,594,694	99,864,037	269,344	50	6.8	13
Co84C	441,113	Amplification	8	127,618,526	128,009,287	390,762	50	19.2	1
Co37C	74,314	Amplification	8	128,148,391	128,420,136	271,746	50	6.6	0
Co37C	74,314	Amplification	8	128,667,420	129,170,226	502,807	50	8.6	2
Co84C	441,113	Amplification	8	128,750,189	128,857,861	107,673	50	8.3	1
Co84C	441,113	Amplification	8	129,473,672	129,667,129	193,458	50	13.8	0
M11–1	110,884	Amplification	8	142,414,288	142,454,841	40,554	50	7.1	1
M11–1	110,884	Amplification	8	144,356,733	144,478,642	121,910	50	6.4	4
Co84C	441,113	Amplification	11	34,337,207	35,266,401	929,195	50	33.0	6
Co82C	128,368	Amplification	12	30,734,351	32,018,350	1,284,000	50	6.2	7
Co84C	441,113	Amplification	13	109,096,557	109,553,930	457,374	50	9.2	1
Co84C	441,113	Amplification	15	88,545,070	89,258,106	713,037	50	26.2	11
M12–02	132,232	Amplification	16	50,791,506	51,506,000	714,495	50	16.9	0
Co82C	128,368	HD	18	54,490,515	56,426,158	1,935,644	175	0.0	10
Co82C	128,368	HD	18	56,629,615	58,250,812	1,621,198	175	0.0	5
Co84C	441,113	Amplification	19	34,570,450	34,641,949	71,500	50	7.9	0
Co84C	441,113	Amplification	19	34,956,853	35,344,522	387,670	50	14.3	2
Co84C	441,113	Amplification	19	36,274,262	36,388,331	114,070	50	6.2	0
Co82C	128,368	Amplification	19	43,386,048	45,698,030	2,311,983	50	8.4	70
Co84C	441,113	Amplification	19	54,500,237	54,643,655	143,419	50	8.4	5

*Values for tag density ratios represent observed maximum values for amplifications and minimum values for homozygous deletions.

Other Supporting Information Files

Table S3 (XLS)
Table S4 (XLS)
Table S5 (XLS)
Table S6 (XLS)
Table S7 (XLS)