Supporting Information

Hart et al. 10.1073/pnas.1424012112

SI Materials and Methods

Exome Sequencing. Genomic DNA was prepared from MCF-10A and MCF-10A-H1047R cell lines using the Qiagen DNeasy Kit. DNA for each sample was sheared on an S2 Covaris instrument to a size range of 250–400 bp. One microgram of sheared DNA for each sample was then end-repaired, A-tailed with Taq polymerase, kinased, and ligated to standard Illumina TruSeq-barcoded adapters following Illumina-recommended protocols. The library was then PCR-amplified for six cycles. The amplified libraries were then hybridized to Agilent Human All Exon SureSelect Target Enrichment baits following the manufacturer's recommended protocols. After hybridization, the selected libraries were amplified for 12 more cycles and size-selected on a 2% agarose gel to recover library products with insert sizes in the 250- to 350-base size range. Size-selected libraries were loaded onto an Illumina sequencer for paired-end 2 × 100 base sequencing. Raw sequencing data were processed into fastq files using CASAVA 1.8 (Illumina). The raw reads used for variant calling are available from the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) as Project SRP050011. The resulting reads were aligned to HG19 using Burrows–Wheeler Aligner version 0.7.10 (1).

Copy Number Variation Analysis, Small Regions. Exome sequencing data were used to identify regions of the genome with significant copy number variations (CNVs). The aligned sequences were analyzed using ExomeDepth version 1.0.5 (2) in R. Parameters for transition probability and region size were varied and found not to influence the detected regions of CNV between H1047R and WT. The regions with a Bayes factor of >20 were selected as significant and annotated with the Conrad's frequent CNV regions (3) and overlapping genes.

Variant Discovery. Aligned reads were processed using the Genome Analysis Toolkit (GATK) DNAseq best practices procedure. Individual alignments were processed in parallel pipelines using GATK version 3.3 (4–6). Aligned reads were deduplicated of PCR duplicates and sorted using Picard version 1.92. Sequences were realigned around insertions and deletions (INDELs) using IndelRealigner in the GATK. Base quality scores were recalibrated using base quality score recalibration. Variants were called using HaplotypeCaller in genome variant call format (GVCF) mode. The GVCF files for each alignment, along with 30 whole-exome sequencing samples from the 1,000 Genomes Project (7), were merged and joint-genotyped using GVCFs. The 30 samples used were obtained from the NIH National Center for Biotechnology Information Sequence Read Archive with the following identifiers: SRR791615, SRR789354, SRR075808, SRR788987, SRR792222, SRR077191, SRR079641, SRR084777, ERR034516, SRR393083, SRR592040, ERR250449, SRR581046, SRR076859, SRR098821, SRR597254, SRR764704, SRR599986, SRR079346, SRR582150, SRR393048, SRR787987, SRR079957, SRR079981, SRR084780, SRR359476, SRR792006, SRR795398, SRR080086, SRR085883, SRR361976, and SRR796074. The called variants were further analyzed using VariantAnnotator and subjected to variant quality score recalibration of SNPs and INDELs separately. The variants that were called were analyzed using ANNOVAR version 08232013 (8).

CNV Analysis, Large Regions. Variant calls were used to determine CNV by changes in SNP allele frequency. The allele frequencies of 78,000 sites were determined from the exome sequencing data. The allele frequencies present in both MCF-10A and MCF-10A-H0147R were plotted against genomic coordinates in a way analogous to the methods used for CNV determination by SNP microarray (cshprotocols.cshlp.org/content/2008/6/pdb.top46.full). Regions of copy number gain are visible as changes in heterozygous allele frequency from 1/2–1/3, 2/3, 1/4, 3/4, etc. From this analysis, we found loss of heterozygosity across the entire X chromosome and alterations of the 5′ end of chromosome 5 and chromosome 22. Chromosome 5 is consistent with a high-level amplification (>four copies) of 5p13–15 in MCF-10A-H1047R. This region includes TERT and IL-7 receptor. Chromosome 22 has two copies near the 5′ end and a single copy throughout the 3′ portion; this is in contrast to the amplifications seen in MCF-10A. This region includes NF2 and EP300 tumor suppressors. However, loss or inactivation of both copies would be required to abolish their activity.

1. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589–595.
2. Plagnol V, et al. (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 28(21): 2747–2754.
3. Conrad DF, et al.; Wellcome Trust Case Control Consortium (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464(7289): 704–712.
4. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
5. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
6. Van der Auwera GA, et al. (2002) From FastQ Data to high-confidence variant calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics*, 10.1002/0471250953.bi1110s43.
7. Abecasis GR, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
8. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164.

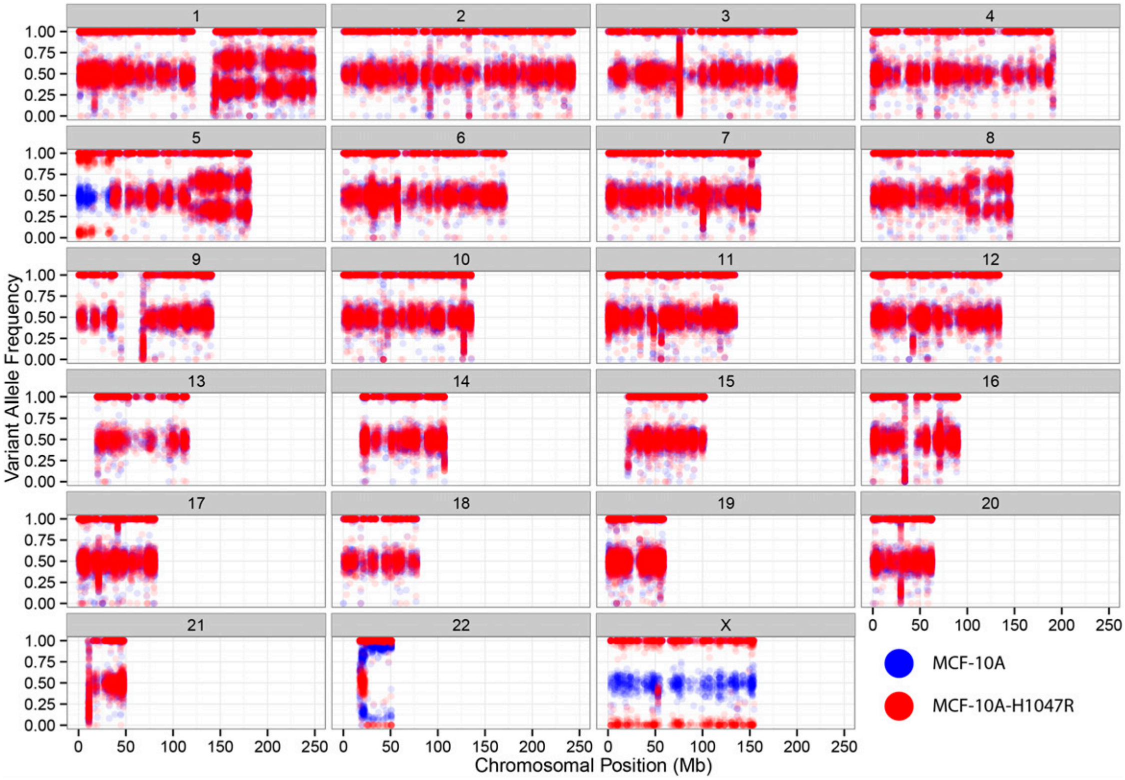

Fig. S1. Large CNV determination by SNP analysis. The variant allele frequency of 78,000 SNPs found in exome sequencing was determined and plotted against genomic position. Areas of genomic duplication can be seen as allele frequencies other than 0.5. An extra copy of a chromosomal region will cause allele frequencies of 0.33 and 0.67. Areas of unequal variant frequency can be seen when the blue and red points fail to overlap, as in chromosomes 5, 22, and X.

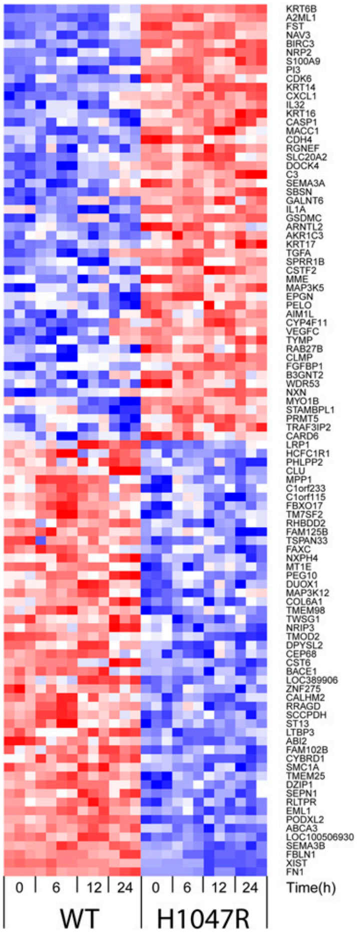

Fig. S2. Heat map of the top 50 consistently up- and down-regulated genes in the MCF-10A/MCF-10A-H1047R pair of cells.

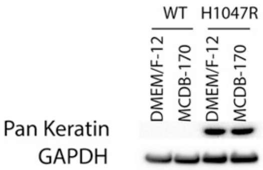

Fig. S3. Western blot using a pankeratin antibody shows up-regulation of keratins in MCF-10A-H1047R cells.

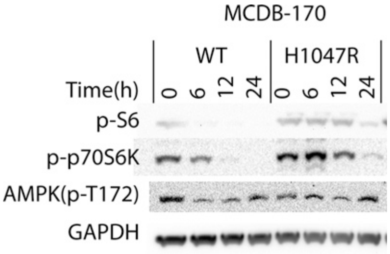

Fig. S4. Western blot of phosphorylated AMPK and TOR targets S6 and p70S6k. The phosphorylation of AMPK is up-regulated at the early and late time points in WT and mutant cells. This suggests a condition of energy starvation at these sampling times. The late time points also show a down-regulation of TOR activity, which could reflect activating phosphorylation of TSC2 by AMPK.

Table S1. Summary of exonic SNPs in MCF-10A and MCF-10A-H1047R cells

Table S1

The data are filtered to include only SNPs that result in a coding change between MCF-10A and MCF-10A-H1047R. If an SNP overlapped multiple coding transcripts, it is listed in the table multiple times. Significance (*P* values) was determined using Fisher's exact test.

Table S2. Summary of the focal amplifications and deletions between MCF-10A and MCF-10A-H1047R

Table S2

The data are filtered to regions with a Bayes factor, an indicator of significance, exceeding 20. In addition to genomic coordinates (HG19), the genes overlapping the region are listed.

Table S3. Summary of gene expression data from RNAseq

Table S3

Samples are identified with cell line, time point, and sample number. Expression values are expressed as counts per minute. Summary statistics, including $\log_2$ fold change (*log_2FC*) [$\log_2$(H1047R/WT)], log counts per minute (logCPM) ($\log_2$[mean(WT + H1047R)]), and *P* value (negative binomial exact test), are included to the right.

Table S4. Summary of gene expression data from SILAC

Table S4

Samples are expressed as $\log_2$(H1047R/WT). The *log_2FC*, confidence intervals, and *P* values are determined from bootstrapped medians of peptides uniquely assignable to a single gene.

Table S5. Summary of GSEA results

Table S5

Category indicates the MSigDB class for the signature. Data indicate the source of the dataset used: 0–24 h are RNAseq gene expression data from the indicated time points, and SILAC are from SILAC proteomics data.

Table S6. Metabolic changes in MCF-10A-H1047R cells

Table S6