Published in final edited form as:
Nat Methods. 2013 October ; 10(10): 965–971. doi:10.1038/nmeth.2609.

A reversible gene trap collection empowers haploid genetics in human cells

Tilmann Bürckstümmer[1], **Carina Banning**[1], **Philipp Hainzl**[1,2], **Richard Schobesberger**[1], **Claudia Kerzendorfer**[2], **Florian M Pauler**[2], **Doris Chen**[2], **Nicole Them**[2], **Fiorella Schischlik**[2], **Manuele Rebsamen**[2], **Michal Smida**[2], **Ferran Fece de la Cruz**[2], **Ana Lapao**[1,2], **Melissa Liszt**[1,2], **Benjamin Eizinger**[1], **Philipp M Guenzl**[2], **Vincent A Blomen**[3], **Tomasz Konopka**[2], **Bianca Gapp**[2], **Katja Parapatics**[2], **Barbara Maier**[2,4], **Johannes Stöckl**[5], **Wolfgang Fischl**[1], **Sejla Salic**[1], **M Rita Taba Casari**[1], **Sylvia Knapp**[2,4], **Keiryn L Bennett**[2], **Christoph Bock**[2], **Jacques Colinge**[2], **Robert Kralovics**[2], **Gustav Ammerer**[6], **Georg Casari**[1], **Thijn R Brummelkamp**[2,3], **Giulio Superti-Furga**[2], and **Sebastian M B Nijman**[2]

[1]Haplogen GmbH, Vienna, Austria [2]Research Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM), Vienna, Austria [3]The Netherlands Cancer Institute, Amsterdam, The Netherlands [4]Department of Medicine 1, Laboratory of Infection Biology, Medical University of Vienna, Vienna, Austria [5]Institute of Immunology, Medical University of Vienna, Austria [6]Max F. Perutz Laboratories, University of Vienna, Vienna, Austria

Abstract

Knockout collections are invaluable tools for studying model organisms such as yeast. However, there are no large-scale knockout collections of human cells. Using gene-trap mutagenesis in near-haploid human cells, we established a platform to generate and isolate individual ‘gene-trapped cells’ and used it to prepare a collection of human cell lines carrying single gene-trap insertions. In most cases, the insertion can be reversed. This growing library covers 3,396 genes, one-third of the expressed genome, is DNA-barcoded and allows systematic screens for a wide variety of cellular phenotypes. We examined cellular responses to TNF-α, TGF-β, TNF-γ and TNF-related

Correspondence should be addressed to T.B. (tibu@haplogen.com), T.R.B. (t.brummelkamp@nki.nl), G.S.-F. (gsuperti@cemm.oeaw.ac.at) or S.M.B.N. (snijman@cemm.oeaw.ac.at).

Accession codes. GenBank: KF179301 (sequence encoding SRGAP1-PPM1H). PeptideAtlas: PASS00240. Gene Expression Omnibus: GSE48848.

Author Contributions

S.M.B.N., T.R.B. and G.S.-F. conceived the haploid gene-trap mutant collection and provided overall guidance. S.M.B.N. and T.B. analyzed data and, together with G.S.-F. and T.R.B., wrote the paper. S.M.B.N., T.R.B., T.B. and G.C. conceived the gene-trap vector design including barcodes and *loxP* sites and the clone-mapping pipeline. T.B. and C. Banning supervised the establishment of the mutant collection and performed validation experiments. P.H., A.L., M.L., W.F., S.S. and M.R.T.C. assisted in the establishment of the mutant collection and validation experiments. F.M.P., D.C., N.T., F.S., B.E., P.M.G., V.A.B., T.K., B.G., C. Bock and R.K. generated the samples for DNA and RNA sequencing and SNP arrays, and analyzed the data. R.S., B.E. and G.C. established the clone-mapping bioinformatics pipeline and databases. C.K., M.R., M.S. and F.F.d.l.C. performed clone-validation experiments. K.P., K.L.B. and J.C. generated samples for mass spectrometry and analyzed the data. B.M., J.S. and S.K. performed and analyzed the leukocyte typing experiments. G.C. and G.A. assisted in platform design.

Competing Financial Interests

The authors declare competing financial interests: details are available in the online version of the paper.

Europe PMC Funders Author Manuscripts

apoptosis-inducing ligand (TRAIL), to illustrate the value of this unique collection of isogenic human cell lines.

Over a decade after the complete human genome has been sequenced, functional annotation of the ~20,000 protein-coding genes remains incomplete. Thus, systematic and scalable methods for the interrogation of the biological functions of gene products are needed. In model organisms, the elucidation of protein function by genetic inactivation has been an extremely valuable approach. But the suitability of these organisms is limited for studying human pathology as many human disease genes lack orthologs in lower eukaryotes.

A major obstacle for experimental genetics in most higher organisms is that their genomes are diploid, masking the inactivation of single alleles. However, there is no fundamental biological reason dictating that a haploid genome precludes normal cellular behavior. For instance, haploid fish and amphibians have been generated, and experiments with mosaic chickens indicate that haploid cells can contribute to multiple lineages1–3. Furthermore, haploid embryonic stem cells from Medaka fish and mice have recently been obtained and shown to maintain pluripotency, underlining the notion that haploid cells can behave like their diploid counterparts4–6.

In humans, sub-diploidy is regularly observed in leukemias, and a stable near-haploid cell line (KBM7) has been subcloned from a chronic myeloid leukemia (CML) patient sample containing the *BCR-ABL1* gene fusion7–9. In contrast to many other established human cell lines, KBM7 cells can be reprogrammed to induced pluripotent stem cells, showing that they maintain the potential to differentiate into all three germ layers10.

Mutagenesis of near-haploid cells with a gene-trap retrovirus has recently been used to inactivate human genes and screen for phenotypes such as proliferative defects or sensitivity to pathogen infection11–19. However, as these screens must be performed in large pools of ~100 million cells, they have been thus far limited to positive selection for mutants resistant to a toxic agent (for example, virus, bacterial toxin or drug). An arrayed collection would allow detailed investigation of single clones or focused subsets of clones, although culturing large numbers of clones simultaneously would be challenging. We initiated a large-scale effort to subclone individual gene trap–containing cells with the aim to create a library of gene mutant cell lines. We make this unique collection of human cell clones with individual loss-of-function mutations available to the scientific community via http://clones.haplogen.org/, which will empower genetics in human cells.

Results

Genomic and proteomic characterization of KBM7 cells

We analyzed in detail the genetic makeup of KBM7 cells and the repertoire of expressed mRNAs and proteins. Previously, spectral karyotyping had revealed that most KBM7 subclones contain 25 chromosomes and are diploid for chromosome 8 (karyotype 25, XY, disomic for chromosome 8, containing the Philadelphia chromosomal translocation)7,11. Indeed, FACS analysis showed that interphase cells contained ~1*N* chromosomes (Fig. 1a).

Europe PMC Funders Author Manuscripts

Although this near-haploid karyotype was stable over several months of culture, diploid cells occasionally emerged (Fig. 1a), presumably through mitotic nondisjunction20.

To obtain a high-resolution karyotype of KBM7 cells, we performed high-density single-nucleotide polymorphism (SNP) array genotyping (Supplementary Table 1). Complete loss of the Y chromosome frequently occurs in KBM7 cells, and we observed this in some of the KBM7 clones that we investigated (Fig. 1b). In agreement with a largely haploid genome, we observed single copies for >95% of all genes. In contrast, we detected SNP heterozygosity and high signal intensities for the entire chromosome 8 and a small part of the long arm of chromosome 15, indicating that genes on these chromosomes are present in two copies. This confirmed that diploidy of these genes did not arise through a duplication event but that they were retained during clonal evolution8.

KBM7 cells are derived from a CML patient with a Philadelphia chromosome and display hallmarks of the myelomonocytic lineage as determined by leukocyte typing (Supplementary Table 2). The CML origin raises the question of which additional driver mutations may have accumulated in the KBM7 genome. Furthermore, the haploidy of KBM7 cells may phenotypically unmask germ-line–encoded recessive SNPs. To address this, we sequenced the exome and whole genome at ~21× and 34× coverage, respectively, and sequenced the messenger RNAs using 100-base-pair (bp) paired-end reads (Supplementary Table 3). Besides the *BCR-ABL1* translocation and potentially damaging point mutations in *TP53* and *NOTCH1*, we observed none of the recurrent aberrations in myeloid malignancies. Next we analyzed the proteome using mass spectrometry (Supplementary Table 4). In concordance with previous studies, protein abundance correlated with mRNA levels21 (Fig. 1c and Supplementary Table 5). As for other frequently used cell lines, ~75% of all Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were represented by at least 50% of their annotated proteins (Fig. 1d and Supplementary Table 6). For instance, we detected the majority of proteins annotated in diverse signaling pathways (for example, insulin, VEGF, NOD and PPAR) by protein mass spectrometry, suggesting that these pathways could be studied in KBM7 cells.

We noticed increased mRNA expression of genes from the diploid chromosome 8, likely as a consequence of the increased copy number (Fig. 1e). However, this differential expression was less pronounced at the protein level (Fig. 1f). This suggests that post-transcriptional proteostatic mechanisms may compensate for the increased gene dosage. To allow easy and up-to-date access to the above described data sets, we created a University of California Santa Cruz (UCSC) Genome Browser hub that additionally contains information about previously mapped insertion sites (http://kbm7.genomebrowser.cemm.at/)12.

A platform for the generation of haploid gene-trap mutants

Gene trap–based insertional mutagenesis operates by random insertion of a splice acceptor followed by a GFP marker and termination sequence into the genome, thus disrupting gene expression. The insertion site must be determined to identify the disrupted gene. The additional introduction of short unique DNA sequences as barcodes into the genome of cells allows easy tracing of individual clones in complex mixtures, enabling efficient pooled screens. In addition, barcodes greatly facilitate the retrieval of single-cell clones from

multiwell plates using massive parallel sequencing. Therefore, we introduced random 22-bp DNA sequences into the retroviral gene-trap vector to obtain a high-complexity barcoded vector library (Fig. 2a). We infected KBM7 cells with a multiplicity of infection <0.1 and subjected them to FACS for GFP expression. Even though the *GFP* transgene is promoterless, intergenic or antisense insertions can result in low GFP expression, presumably because of long terminal repeat activity or (cryptic) genomic promoter activity. As a consequence, not all GFP^+ clones have disrupting insertions in genes. Subsequently, we seeded cells using a limiting dilution strategy, expanded them in 96-well plates and stored them in liquid nitrogen (Fig. 2b). We mapped insertion sites in individual clones (Fig. 2c and Online Methods); examples of insertion site and DNA barcode sequences are shown in Supplementary Table 7.

Using this pipeline we mapped 23,468 clones with an equal number of unique barcodes (Fig. 3a). Of these, 321 clones contained gene-trap insertions in a coding exon, directly disrupting the respective open reading frame. Insertions in introns are predicted to be mutagenic if the gene-trap cassette is inserted in the sense orientation. Of 11,766 intronic insertion events, ~50% (6,352) were predicted to affect expression of the respective gene. Of the 3,396 trapped genes, 67% (2,289 genes) were expressed at FPKM (fragments per kilobase of exon per million fragments mapped) >3, and 81% (2,755 genes) were expressed at FPKM >1. As expected, essential genes such as those involved in ribosome biogenesis, splicing and amino acid metabolism were under-represented among the trapped genes (Supplementary Fig. 1 and Supplementary Table 8).

As retroviral vectors display an integration bias, the proportion of newly trapped genes decreases with the size of the collection. The bias we observed using our vector was very similar to that previously reported for murine leukemia virus (MLV)-based vectors22 (Supplementary Fig. 2). We modeled the relationship between trapped genes and mapped clones to investigate the proportion of expressed genes that can be recovered using our strategy (Supplementary Figs. 3 and 4). From this analysis we estimate that to recover 8,000 trapped genes (>75% of the expressed genome) ~35,000 mutant clones are required, which is within reach using the current platform.

FACS analysis of 171 clones revealed that 143 clones (~84%) remained haploid after 4–6 weeks of culture. Furthermore, high-density SNP analysis of eight mutant clones revealed only few minor genetic alterations, suggesting limited genetic drift (Supplementary Table 9). Information on all clones is available at http://clones.haplogen.org/ and in Supplementary Table 10.

Haploid gene trap mutants resemble gene knockouts

We generated the majority of the mutant clones using a vector containing a gene-trap cassette flanked by *loxP* sites. For the 4,512 clones with intronic insertions and *loxP* sites (62%), this allows reversible gene inactivation (Fig. 3b). Indeed, upon infection with a retrovirus expressing Cre recombinase, the gene-trap cassette was readily excised. Furthermore, protein expression was restored, albeit not fully for all examined clones (Fig. 3c and Supplementary Fig. 5). This suggests that in some instances the remaining retroviral and *loxP* sequences impinge on gene transcription.

To illustrate that the splice acceptor of the gene-trap cassette is efficiently used by the cellular splicing machinery and thereby disrupts the associated gene transcript, we first used a reverse trancriptase (RT)-PCR–based validation approach. We designed primers flanking the gene-trap insertion site and used them to amplify cDNA from selected clones. As expected, efficient splicing of the gene-trap into the endogenous transcript resulted in the absence of a PCR product (Fig. 3d). However, not all clones displayed a strong reduction in the amount of PCR product (data not shown). Although in some cases this may be explained due to the technical limitations of the assay, it indicates that in a fraction of the clones the gene-trap insertion does not efficiently disrupt expression.

For a second set of clones, we analyzed both protein expression and performed quantitative (q)RT-PCR analysis (Fig. 3e and Supplementary Table 10). Low mRNA levels resulted in undetectable expression at the protein level (Fig. 3e and Supplementary Table 11). We conclude that gene-trapping of haploid KBM7 cells can result in near-complete gene inactivation akin to that in conventional knockouts.

Reverse genetic analysis of selected mutants

We analyzed the phenotypes of selected gene-trapped clones to illustrate their value as tools for studying gene function. Signaling through tumor necrosis factor receptor 1 (TNFR1, encoded by the *TNFRSF1A* gene) contributes to inflammatory diseases, and interfering with its function has revolutionized the treatment of rheumatoid arthritis23. A clone containing an effective disruption of *TNFRSF1A* would be expected to no longer respond to stimulation with TNF-α. Indeed, cells with mutant *TNFRSF1A* did not degrade the downstream target IκB-α upon stimulation with TNF-α (Fig. 4a; qRT-PCR data for this clone and others used in Fig. 4 are in Supplementary Table 12). Consequently, these cells were almost completely impaired in their transcriptional response to TNF-α (Fig. 4b) and were resistant to TNF-α–induced apoptosis (Fig. 4c). Thus, cells with mutant *TNFRSF1A* are refractory to TNF-α–induced signaling and provide a tractable tool to study TNF-mediated effects.

The cytokine TGF-β transmits signals into the cell by binding the TGF-β receptor I (TGFBR1). Subsequent heterodimerization with the TGF-β receptor II (TGFBR2) results in the downstream phosphorylation of second messengers, including the receptor-activated SMAD proteins, which instruct cellular programs involved in cell cycle, apoptosis, differentiation and immune regulation24. Cells with mutant *TGFBR1* or *TGFBR2* could not phosphorylate SMAD2 upon treatment with TGF-β (Fig. 4d). Future studies, for instance, with gene-trapped cells reconstituted with TGF-β receptor point mutants, could be valuable for the unraveling of signaling events triggered by TGF-β.

Inactivating mutations in the gene encoding Neurofibromin 1 (NF1) cause hereditary and sporadic cancers. It is known that NF1 acts as a negative regulator of Ras by stimulating its intrinsic GTPase activity. We used the cells with mutation in *NF1* from the collection to test the impact of NF1 loss on Ras signaling upon serum starvation. We detected elevated phosphorylation of the downstream effector ERK (Fig. 4e), confirming the negative role of NF1 in the Ras-RAF-ERK pathway and illustrating that isogenic mutant KBM7 cells can be used to systematically study oncogenic signaling pathways.

Interferon gamma (IFN-γ) is a critical mediator of innate and adaptive immunity and signals by activating the JAK-STAT pathway through its cognate receptor. Access to isogenic cell lines bearing mutations in IFN-γ signaling components will allow a systematic dissection of the pathway. The collection already contains mutants for several components, including the genes *IFNGR2, JAK1, JAK2, SOCS7, GRB2, SOS1, SOS2, STAT3, STAT4, PTPN6* (which encodes SHP-1) and *CBLC* (which encodes c-CBL). Here we used cells with mutant *JAK2* to begin such an analysis. These cells displayed a severe but not complete reduction of phosphorylated STAT1 levels upon stimulation (Fig. 4f). This suggests alternative routes of activation, most likely via the close homolog JAK1. Accordingly, the transcriptional response to IFN-γ was severely blunted in KBM7 cells with mutant *JAK2* (Fig. 4g). STAT1 activation could be restored upon Cre recombinase–mediated excision of the gene trap, confirming that the blunted signaling was caused by impaired JAK2 expression (Supplementary Fig. 6).

Caspase-8 is an essential effector of extrinsic apoptotic stimuli, including Fas ligand and TRAIL25,26. We tested the *CASP8* mutant cells in the collection and found that they were indeed completely resistant to TRAIL-induced cleavage of the effectors caspase-3 and RIP1 (Fig. 4h), demonstrating its nonredundant role. Although cells with mutant *CASP8* were resistant to TRAIL-induced apoptosis, they retained sensitivity to another apoptosis-inducing agent, doxorubicin (Fig. 4i). A complete block of the caspase-8–mediated cell death pathway may allow future screens for genes circumventing this pathway.

ARID2 (also called BAF200) is a component of the PBAF SWI-SNF chromatin remodeling complex, and loss-of-function mutations have been found in a variety of tumors27–30. However, the molecular function of ARID2 and its role as a tumor suppressor are poorly understood. ARID2 has been found to be required for the induction of selected interferon-response genes, including *IFITM1*, suggesting a role in the antiviral response31. Supporting this essential role, KBM7 cells with mutant *ARID2* could not induce *IFITM1* mRNA levels beyond basal expression in response to treatment with IFN-γ (Fig. 4j).

Approximately 1% of eukaryotic proteins are modified with a glycosyl-phosphatidyl-inositol (GPI) moiety that anchors them to the outer leaflet of the plasma membrane32. Germ-line mutations in the genes that mediate GPI attachment result in a form of hemolytic anemia called Paroxysmal nocturnal hemoglobinuria. KBM7 cells express the GPI-anchored proteins CD55 and CD59, and our collection includes mutants for the GPI attachment factor genes *PIGS* and *PIGX*. When either gene was mutated, surface expression of both CD55 and CD59 was abrogated (Fig. 4k and Supplementary Fig. 7). This illustrates that KBM7 mutant cells can be used to produce proteins lacking specific post-translational modifications, which may be useful for the production of biologicals. Together, these data illustrate that the collection of mutant KBM7 cells can be used to interrogate a wide variety of cellular processes.

Discussion

We generated a collection of thousands of human isogenic cell lines with mutations in individual genes. Key features of the collection are: cells are of human origin, allowing

Europe PMC Funders Author Manuscripts

genotype to phenotype relationships in a relevant model organism; cells are comprehensively characterized in terms of their molecular makeup; complete gene inactivation can be achieved; cell lines are isogenic; many insertions (in introns) are reversible via Cre recombinase–*loxP* technology, thus avoiding the need for reconstitution as required for RNA interference experiments or nuclease-based knockouts; and individual mutants are marked with unique DNA barcodes, enabling additional functional genetics applications.

As the mRNA transcripts of these trapped genes are dysfunctional in many clones, they can resemble conventional knockout alleles. 'Gene-trapped' mouse embryonic stem cells have been widely used for the generation of 'knockout mice'33,34. However, as for mouse embryonic stem cell mutants, not every trapped gene in KBM7 cells results in the full inactivation of its associated transcripts. In some cases, the genomic context may prevent the efficient use of the adenovirus splice acceptor site, or alternative transcripts may mask the inactivation of a minor transcript. In other cases, a trapped gene may still yield a truncated transcript resulting in a gain of function or partially functional (i.e., hypomorphic) protein. Although these and potentially other mechanisms would yield mutant cells that do not represent full loss-of-function alleles, they may still prove valuable for the analysis of protein function. Moreover, hypomorphic alleles will be useful for enhancer or suppressor screens and thus allow the study of essential genes.

It has been well documented and we have also observed that retroviruses have a strong genomic integration bias. Indeed, their preference for the 5′ region of expressed genes has made them a preferred agent for gene trap–based mutagenesis. Although we intend to expand this mutant collection, the strongly skewed mutagenesis may stand in the way of the generation of a complete genome-wide collection. Therefore, KBM7 mutants made with other mutagens such as lentiviruses, transposons, chemicals or nucleases may aid in the construction of a complete genome-wide collection. The haploid genome of KBM7 cells makes them well suited for targeted nuclease–based technologies such as transcription activator–like effector nucleases as only one allele needs to be targeted, thus greatly increasing efficiency and simplifying genotyping (data not shown).

Although the gene trap–based approach for generating mutations in human cells can be applied in diploid cells, in most cases it would only result in full gene inactivation in haploid cells, which in humans is currently restricted to KBM7 or HAP1 cells13. No single cell line can be expected to recapitulate all aspects of human biology, and findings in KBM7 cells require validation in independent experimental systems. However, all available evidence suggests that KBM7 cells are not less valid than other popular cell lines as a model to study well-conserved processes that may be less dependent on a particular tissue context. Thus, we propose that the most appropriate use of this collection may be the systematic annotation of gene function for basic cellular processes, such as metabolism, secretion and glycosylation as well as for common signal transduction and transcriptional processes. Careful measurement of the transcriptional landscape across many mutant KBM7 cell lines, under basal conditions or after challenge with selected stimuli will allow modeling of the logic of genetic networks with unprecedented precision and reliability. The collection presented here paves the way for a systematic and rigorous assessment of the genetic requirements of many processes under many conditions in a human cell line.

Online Methods

Cell culture

KBM7 cells were grown in Iscove's modified Dulbecco's medium (IMDM) supplemented with 10% FCS and penicillin-streptomycin. A list of all mutant clones used in experiments is provided in Supplementary Table 13.

SNP arrays, and exome, genome and RNA sequencing

DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega). DNA was hybridized to Genome-Wide Human SNP 6.0 array (Affymetrix) and data were analyzed for chromosomal copy number changes using Genotyping Console version 4.1.1 software (Affymetrix). A deletion on chr.11p was detected encompassing *AMBRA1,* a gene involved in autophagy35.

Genomic DNA libraries were generated using the TruSeq DNA LT Sample Prep-Kit v2, and exome enrichment was performed using the TruSeq Exome Enrichment Kit (both from Illumina). The libraries were hybridized to Illumina flow cells V3 and sequenced on an Illumina HiSeq 2000 instrument (37 million 50-bp paired-end reads for whole-exome sequencing and 1,065 million 100-bp paired-end reads for whole-genome sequencing). Additionally, a second independent batch of wild-type cells was subjected to whole-genome sequencing (236 million 60-bp and 179 million 75-bp paired-end reads). Reads were aligned to the hg19 reference genome (Burrows Wheeler Aligner 0.5.9-r16; ref. 36) and potential PCR duplicates were removed by SAMtools (0.1.18)37. Coverage calculations and tracks were generated with the help of RseQC and SAMtools mpileup. In combination, 50% of the genome was covered at 35× or more (99% was covered at 14× or more). Variants were called with SAMtools mpileup + bcftools. The resulting SNV list was filtered for high quality and further annotated by ANNOVAR (version of November 2012)38 using Refseq gene annotation. The circos plot was created with the help of the Circos software (v0.63)39. Libraries for RNA-seq were prepared using 4 μg of *KBM7* RNA, RiboZero (Epicentre) and ScriptSeq v1 (Epicentre). Libraries were sequenced on a HiSeq2000 (50-bp single-end reads). 113 million reads were aligned with TopHat v2.0.3 using Bowtie 2.0.0.6 and the ENSEMBL gene annotation (University of California Santa Cruz Genome browser). RPKMs were calculated using the RSeQC package excluding multi-matching reads.

Besides the previously described *BCR-ABL1* translocation and potentially damaging mutations in *TP53* and *NOTCH1* (Supplementary Table 3), none of the recurrent aberrations in myeloid malignancies were observed in KBM7 cells, and major genes implicated in cancer, including *PTEN*, *KRAS*, *RB1*, *BRAF* and *EGFR* did not carry mutations. However, paired-end RNA sequencing indicated a small inversion on chromosome 12 that was confirmed by Sanger sequencing (Supplementary Fig. 8). The inversion disrupts two genes (*SRGAP1* and *PPM1H*) and generates a new, uncharacterized fusion transcript.

Mass spectrometry

PBS-washed KBM7 cell pellets were mixed with 300 μl of 50 mM HEPES pH 8.0, 2% SDS, 1 mM PMSF, protease inhibitor cocktail (Sigma-Aldrich) and incubated for 20 min at room

temperature. After heating the lysate for 5 min at 99 °C, samples were sonicated with a Covaris sonicator (Sonicator S2X, Covaris) and clarified by centrifugation at 16,000*g* for 10 min at room temperature. Protein concentration was measured with the BCA Protein Assay (Pierce Biotechnology). After reduction of disulfide bonds with dithiothreitol, 150 µg total protein were digested with trypsin according to the filter-aided sample preparation (FASP) protocol40. Fifty micrograms of the digest was concentrated and purified by solid phase extraction (SPE) and 50 off-line fractions were collected41. Peptides were then separated by liquid chromatography and analyzed by collision-induced dissociation (CID) on a hybrid LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) coupled to an Agilent 1200 HPLC nanoflow system (Agilent Biotechnologies). The acquired raw mass spectrometry (MS) data files were processed and searched against the human SwissProt database with the search engines MASCOT (v2.3.02, MatrixScience) and Phenyx (v2.5.14, GeneBio)42. A false positive detection rate (FDR) of <1% and <0.1% was determined for proteins and peptides, respectively, by applying the same procedure against a reversed database.

We observed substantial qualitative overlap (>50%) of the proteome of KBM7 cells with those of 11 other previously characterized human cell lines used for human genetics studies (Supplementary Table 14)43. Furthermore, KEGG pathway representation and coverage between the cell lines showed marked concordance (Supplementary Fig. 9). Thus, as judged by the proteins expressed high enough to be detected by mass spectrometry, there was concordance between KBM7 and several popular cell lines.

Pipeline for production and mapping of gene-trap mutant clones

The annotated sequence of the gene-trap cassette used for this study is available at http://clones.haplogen.org/. Gene-trap infected KBM7 cells12 were FACS-sorted for GFP expression and cloned by limiting dilution.

We generated orthogonal clone pools (rows, columns and plates) for each set of six 96-well plates to keep track of their location, and barcode sequences were PCR-amplified with indexed primers (i.e., ‘pool codes’; Fig. 2c). Barcodes and indices were read by paired-end sequencing, thus uniquely assigning a well and plate position to a single barcode (i.e., clone). This orthogonal pooling strategy, combined with indexed sequencing, allowed the unambiguous identification of the plate position for ~70% of the clones. In most cases, unassigned plate positions were due to slow-growing clones or empty wells. As each plate position is linked to a single barcode sequence, this procedure also flags wells containing multiple clones or multiple gene-trap integrations, both of which are discarded.

To map genomic insertion sites of individual clones, genomic DNA from identical pools was digested in parallel using NlaIII and MseI and processed by inverse PCR12. Barcodes and insertion sites were matched in a second paired-end sequencing run. The relatively short DNA sequence tags and strong bias in the inverse PCR limited the routine mapping efficiency to ~85%.

Genomic DNA was isolated from cell pools using the QIAamp DNA Mini Kit (Qiagen). Barcodes were amplified by PCR using GoTaq Polymerase (Promega). Primers were Fwd:

AATGATACGGCGACCACCGAGATCTACACAGAACTCGTCAGTTCCACCAC and Rev: CAAGCAGAAGACGGCATACGAGATXXXXXXXXXXXXXXXXGTGACTGGAGTTCAGACGTG, where X indicates the bases of pool-specific 15-bp index sequences.

PCR products were purified and subjected to paired-end sequencing using the following sequencing primers: first read (pool index): GATCGGAAGAGCACACGTCTGAACTCCAGTCAC and second read (barcode): GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT. Each barcode was linked to three pool codes to specify the row, the column and the plate that uniquely identify its position on a given 96-well plate.

To link barcodes with genomic insertion sites, we performed inverse PCRs of sets of 576 clones (six 96-well plates) that were all combined into one pool. Genomic DNA was isolated from cell pools and subjected to digestion with NlaIII or MseI. Digested samples were ligated and subjected to inverse PCR with the following primers: Fwd: AATGATACGGCGACCACCGAGATCTACACATCTGATGGTTCTCTAGCTTGCC and Rev: CAAGCAGAAGACGGCATACGAGATGTTTGTACAAAAAAGCAGGC.

PCR products were subjected to paired-end sequencing using the following sequencing primers: first read (genome): CTAGCTTGCCAAACCTACAGGTGGGGTCTTTCA and second read (barcode): GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT.

The first read identified the genomic integration site of the gene trap, and the second read identified the 22-bp barcode. Each barcode was thus linked to a single genomic integration site. We estimate that the costs for the generation, mapping and isolation of individual mutant clones using this platform are between 200 and 400 Euros ($250–500) per clone. Costs are expected to decrease over time.

RT-PCRs

To determine mRNA expression we isolated RNA and subjected it to reverse transcription. cDNA was analyzed by PCR using GoTaq Polymerase (Promega) according to manufacturer's instructions. PCR primers were custom-designed for each clone to flank the insertion site of the gene-trap cassette.

Reverse genetic analysis of selected mutant KBM7 cells

Cells stimulated with TNF-α (Peprotech) cells were lysed in Frackelton buffer (10 mM Tris/HCl pH 7.5, 50 mM NaCl, 30 mM sodium pyrophosphate, 1% Triton X-100, 50 mM NaF and protease inhibitors) and analyzed by western blotting using an IκB-α–specific antibody (C-21, dilution 1:500, Santa Cruz) and a tubulin-specific antibody (ab-7291, dilution 1:2,000, Abcam). To induce TNF-α–mediated apoptosis, cells were pretreated with 3 μg/ml cycloheximide (Sigma-Aldrich) for 1 h and subsequently stimulated with 30 ng/ml TNF-α for 24 h. Cell viability was assessed using CellTiterGlo (Promega).

To measure phospho-SMAD2 (Ser465 and Ser467, 1:500, 3108, Cell Signaling), cells were serum starved overnight and stimulated with 10 ng/ml TGF-β (Peprotech). Cell lysates were

analyzed by western blotting using SMAD1/2/3 (Santa Cruz sc-7960, 1:1,000) as a loading control.

To assess the response to TRAIL, cells were stimulated with TRAIL (Peprotech) for 4 h, harvested and lysed (50 mM HEPES pH 7.4, 250 mM NaCl, 5 mM EDTA, 1% NP40, 50 mM NaF and protease inhibitors). Lysates were analyzed by western blotting using cleaved caspase-3 (9661, Cell Signaling, 1:1,000), RIP1 (610458, BD Transduction Lab, 1:1,000) and tubulin (7291, Abcam, 1:1,000) antibodies. Cell viability was assessed using CellTiter-Glo (Promega) 16 h after addition of TRAIL.

Cellular response to IFN-γ was determined by stimulating with 100 ng/ml IFN-γ (Peprotech) for the indicated time periods. Cells were washed with cold PBS and lysed in Frackelton buffer. Lysates were analyzed by western blotting using a phospho-STAT1 (Tyr701)-specific antibody (9171, dilution 1:1,000, Cell Signaling) and a tubulin-specific antibody (ab-7291, dilution 1:2,000, Abcam).

Control KBM7 or *ARID2* clone was treated with IFN-γ for 24 h. Cells were then washed with PBS, and RNA was extracted using the RNeasy MinElute Cleanup Kit (Qiagen). *IFITM1* expression levels were measured by qRT-PCR using the KAPA SYBR FAST ABI Prism kit (Peqlab). *GAPDH* housekeeping gene was used as a control.

For microarray analysis, cells were collected and total RNA was extracted using the RNeasy Mini Kit (Qiagen). RNAs were analyzed using Primeview (Affymetrix) microarrays, and the data was robust multiarray average (RMA)-normalized.

Antibodies and FACS

Antibodies recognizing BTK (sc-1107, 1:1,000), CYLD (sc-74434, 1:1,000), ELF1 (sc-631, 1:1,000), HCK (sc-72, 1:1,000), NFATC2 (sc-7296, 1:1,000), P53 (DO-1, 1:1,000), IFI16 (sc-8023, 1:1,000) were obtained from Santa Cruz Biotechnology. Anti-Tec 06-561 (1:500) was purchased from Upstate (Millipore) and Anti-PTEN (138G6, 1:1,000), JAK2 (D2E12, 1:1,000) and RB1 (4H1, 1:1,000) from Cell Signaling Technologies.

For membrane staining, cells were incubated with below indicated monoclonal antibodies (mAb) for 30 min. Oregon Green–conjugated goat anti-mouse-Ig antibodies (Molecular Probes) was used as a second-step reagent. FACS analysis was performed on a FACSCalibur flow cytometer (BD Biosciences). The following mouse-anti-human mAbs were generated by O. Majdic (Institute of Immunology, Medical University of Vienna) and used for stainings (clone): MHC-I (W6/32), CD3 (VIT3b), CD11a (CD11a-5E6), CD11b (VIM12), CD13 (5-390), CD34 (9F2), CD40 (G28-5), CD44 (3F5), CD45 (VIT200), CD47 (AIV), CD63 (11C9 15), CD71 (VIP1) and CD147 (AAA1).

CD55 and CD59 surface expression was measured by washing cells with cold PBS, and staining with CD59-PE (p282) or CD55-FITC antibody (IA10) (BD Transduction Lab) for 20 min on ice. Samples were measured on a FACSCalibur.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank G. Winter and members of the Brummelkamp, Nijman and Superti-Furga laboratories for discussions and technical assistance, R. Martins for help with the illustrations in Figure 2, O. Majdic (Medical University of Vienna) for providing antibodies, J. Carette for advice on gene trap vector design, and H. Pickersgill for manuscript editing and suggestions. C. Banning was supported by a FemPower grant from Zentrum für Innovation und Technologie (Die Technologieagentur der Stadt Wien), and A.L. and M.L. were supported by a Zentrum für Innovation und Technologie Life Sciences 2011 grant. M.R. was supported by European Molecular Biology Organization fellowship (ALTF1346-2011). K.P. was supported by a European Research Council grant (ERC-2009-AdG-250179-i-FIVE).

References

1. Subtelny S. The development of haploid and homozygous diploid frog embryos obtained from transplantations of haploid nuclei. J Exp Zool. 1958; 139:263–305. [PubMed: 13654675]
2. Thorne MH, Collins RK, Sheldon BL. Live haploid-diploid and other unusual mosaic chickens (*Gallus domesticus*). Cytogenet Cell Genet. 1987; 45:21–25. [PubMed: 3474105]
3. Corley-Smith GE, Lim CJ, Brandhorst BP. Production of androgenetic zebrafish (*Danio rerio*). Genetics. 1996; 142:1265–1276. [PubMed: 8846903]
4. Yi M, Hong N, Hong Y. Generation of medaka fish haploid embryonic stem cells. Science. 2009; 326:430–433. [PubMed: 19833967]
5. Elling U, et al. Forward and reverse genetics through derivation of haploid mouse embryonic stem cells. Cell Stem Cell. 2011; 9:563–574. [PubMed: 22136931]
6. Leeb M, Wutz A. Derivation of haploid embryonic stem cells from mouse embryos. Nature. 2011; 479:131–134. [PubMed: 21900896]
7. Kotecki M, Reddy PS, Cochran BH. Isolation and characterization of a near-haploid human cell line. Exp Cell Res. 1999; 252:273–280. [PubMed: 10527618]
8. Andersson BS, et al. Ph-positive chronic myeloid leukemia with near-haploid conversion *in vivo* and establishment of a continuously growing cell line with similar cytogenetic pattern. Cancer Genet Cytogenet. 1987; 24:335–343. [PubMed: 3466682]
9. Holmfeldt K, Odic D, Sullivan MB, Middelboe M, Riemann L. Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains. Appl Environ Microbiol. 2012; 78:892–894. [PubMed: 22138992]
10. Carette JE, et al. Generation of iPSCs from cultured human malignant cells. Blood. 2010; 115:4039–4042. [PubMed: 20233975]
11. Carette JE, et al. Haploid genetic screens in human cells identify host factors used by pathogens. Science. 2009; 326:1231–1235. [PubMed: 19965467]
12. Carette JE, et al. Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. Nat Biotechnol. 2011; 29:542–546. [PubMed: 21623355]
13. Carette JE, et al. Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. Nature. 2011; 477:340–343. [PubMed: 21866103]
14. Guimaraes CP, et al. Identification of host cell factors required for intoxication through use of modified cholera toxin. J Cell Biol. 2011; 195:751–764. [PubMed: 22123862]
15. Papatheodorou P, et al. Lipolysis-stimulated lipoprotein receptor (LSR) is the host receptor for the binary toxin Clostridium difficile transferase (CDT). Proc Natl Acad Sci USA. 2011; 108:16422–16427. [PubMed: 21930894]
16. Reiling JH, et al. A haploid genetic screen identifies the major facilitator domain containing 2A (MFSD2A) transporter as a key mediator in the response to tunicamycin. Proc Natl Acad Sci USA. 2011; 108:11756–11765. [PubMed: 21677192]

Europe PMC Funders Author Manuscripts

17. Rosmarin DM, et al. Attachment of *Chlamydia trachomatis* L2 to host cells requires sulfation. Proc Natl Acad Sci USA. 2012; 109:10059–10064. [PubMed: 22675117]
18. Birsoy K, et al. MCT1-mediated transport of a toxic molecule is an effective strategy for targeting glycolytic tumors. Nat Genet. 2013; 45:104–108. [PubMed: 23202129]
19. Jacobson LS, et al. Cathepsin-mediated necrosis controls the adaptive immune response by Th2 (T helper type 2)-associated adjuvants. J Biol Chem. 2013; 288:7481–7491. [PubMed: 23297415]
20. Shi Q, King RW. Chromosome nondisjunction yields tetraploid rather than aneuploid cells in human cell lines. Nature. 2005; 437:1038–1042. [PubMed: 16222248]
21. Lundberg E, et al. Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol. 2010; 6:450. [PubMed: 21179022]
22. Lewinski MK, et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. 2006; 2:e60. [PubMed: 16789841]
23. Scheinecker C, Smolen JS. Rheumatoid arthritis in 2010: from the gut to the joint. Nat Rev Rheumatol. 2011; 7:73–75. [PubMed: 21289609]
24. Moustakas A, Heldin CH. The regulation of TGFbeta signal transduction. Development. 2009; 136:3699–3714. [PubMed: 19855013]
25. Varfolomeev EE, et al. Targeted disruption of the mouse Caspase 8 gene ablates cell death induction by the TNF receptors, Fas/Apo1, and DR3 and is lethal prenatally. Immunity. 1998; 9:267–276. [PubMed: 9729047]
26. Kischkel FC, et al. Apo2L/TRAIL-dependent recruitment of endogenous FADD and caspase-8 to death receptors 4 and 5. Immunity. 2000; 12:611–620. [PubMed: 10894161]
27. Biankin AV, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature. 2012; 491:399–405. [PubMed: 23103869]
28. Manceau G, et al. Recurrent inactivating mutations of ARID2 in non-small cell lung carcinoma. J Int Cancer. 2013; 132:2217–2221.
29. Hodis E, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150:251–263. [PubMed: 22817889]
30. Li M, et al. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet. 2011; 43:828–829. [PubMed: 21822264]
31. Yan Z, et al. PBAF chromatin-remodeling complex requires a novel specificity subunit, BAF200, to regulate expression of selective interferon-responsive genes. Genes Dev. 2005; 19:1662–1667. [PubMed: 15985610]
32. Orlean P, Menon AK. Thematic review series: lipid posttranslational modifications. GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycophospholipids. J Lipid Res. 2007; 48:993–1011. [PubMed: 17361015]
33. Skarnes WC, et al. A public gene trap resource for mouse functional genomics. Nat Genet. 2004; 36:543–544. [PubMed: 15167922]
34. Stanford WL, Cohn JB, Cordes SP. Gene-trap mutagenesis: past, present and beyond. Nat Rev Genet. 2001; 2:756–768. [PubMed: 11584292]
35. Fimia GM, et al. Ambra1 regulates autophagy and development of the nervous system. Nature. 2007; 447:1121–1125. [PubMed: 17589504]
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
37. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
38. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164. [PubMed: 20601685]
39. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]
40. Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods. 2009; 6:359–362. [PubMed: 19377485]

41. Bennett KL, et al. Proteomic analysis of human cataract aqueous humour: Comparison of one-dimensional gel LCMS with two-dimensional LCMS of unlabelled and iTRAQ(R)-labelled specimens. J Proteomics. 2011; 74:151–166. [PubMed: 20940065]

42. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics. 2003; 3:1454–1463. [PubMed: 12923771]

43. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics. 2012; 11

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

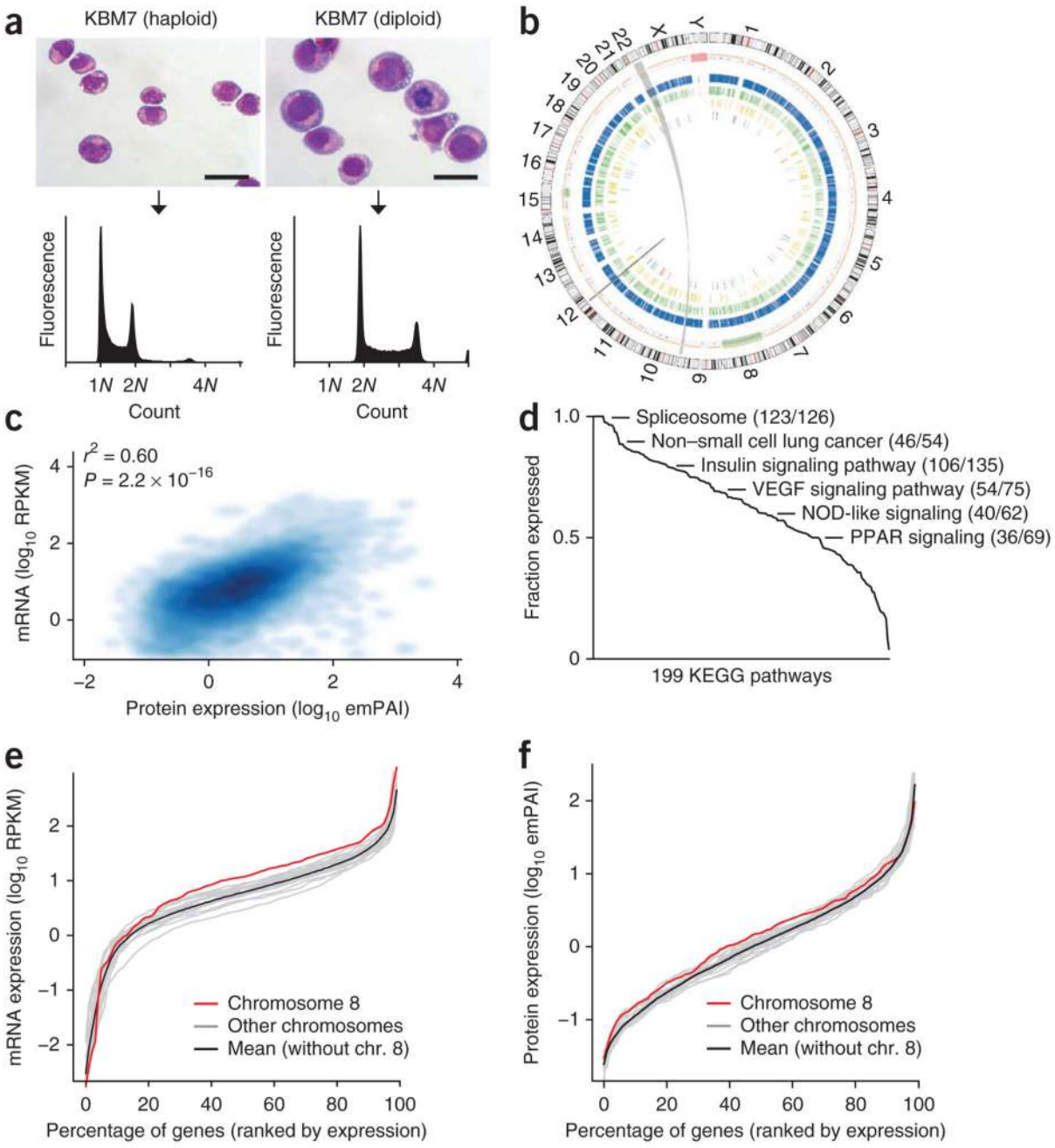

Figure 1. Genomic and proteomic characterization of KBM7 cells.

(**a**) Cytospin Giemsa staining and DNA content analysis by propidium iodide staining and FACS analysis of haploid and diploid KBM7 cells. Scale bar, 20 μm. (**b**) Circos plot of KBM7 cells. Outer to inner circles show chromosomes with cytogenetic bands and centromeres (red); homozygous (orange) and heterozygous (purple) variants, with disomic regions highlighted in green and complete losses highlighted in red; 'known' SNVs (listed in either dbSNP build 137, dbSNP build 129, 1000 Genomes version of April 2012, Exome Variant Server 5400 or 6500) are depicted as blue, 'unique' (not listed in any of the above

databases) as green, COSMIC (version 61) as yellow, frameshift as light blue and stop-gain variants as red ticks (the latter two in the same circle). The *BCR-ABL1* translocation and the *SRGAP1-PPM1H* inversion are indicated as gray ribbon. (**c**) Scatter density plot comparing expression of mRNAs (read per kilobase per million, RPKM) and proteins (exponentially modified protein abundance index, emPAI). Pearson's correlation coefficient P value (n = 9,835). (**d**) KEGG pathway coverage of selected pathways as assessed by protein expression. All KEGG pathways (199 pathways) were sorted based on coverage. (**e**,**f**) Per-chromosome analysis of mRNA (**e**) and protein (**f**) levels. mRNA expression of genes on the disomic chromosome 8 (**e**, $P = 4.5 \times 10^{-12}$ (Wilcoxon rank-sum) for comparing all autosomes (n = 9,453) with chromosome 8 (n = 385)) and protein expression (**f**, $P = 0.002$ (Wilcoxon rank-sum) for comparing all autosomes with chromosome 8).

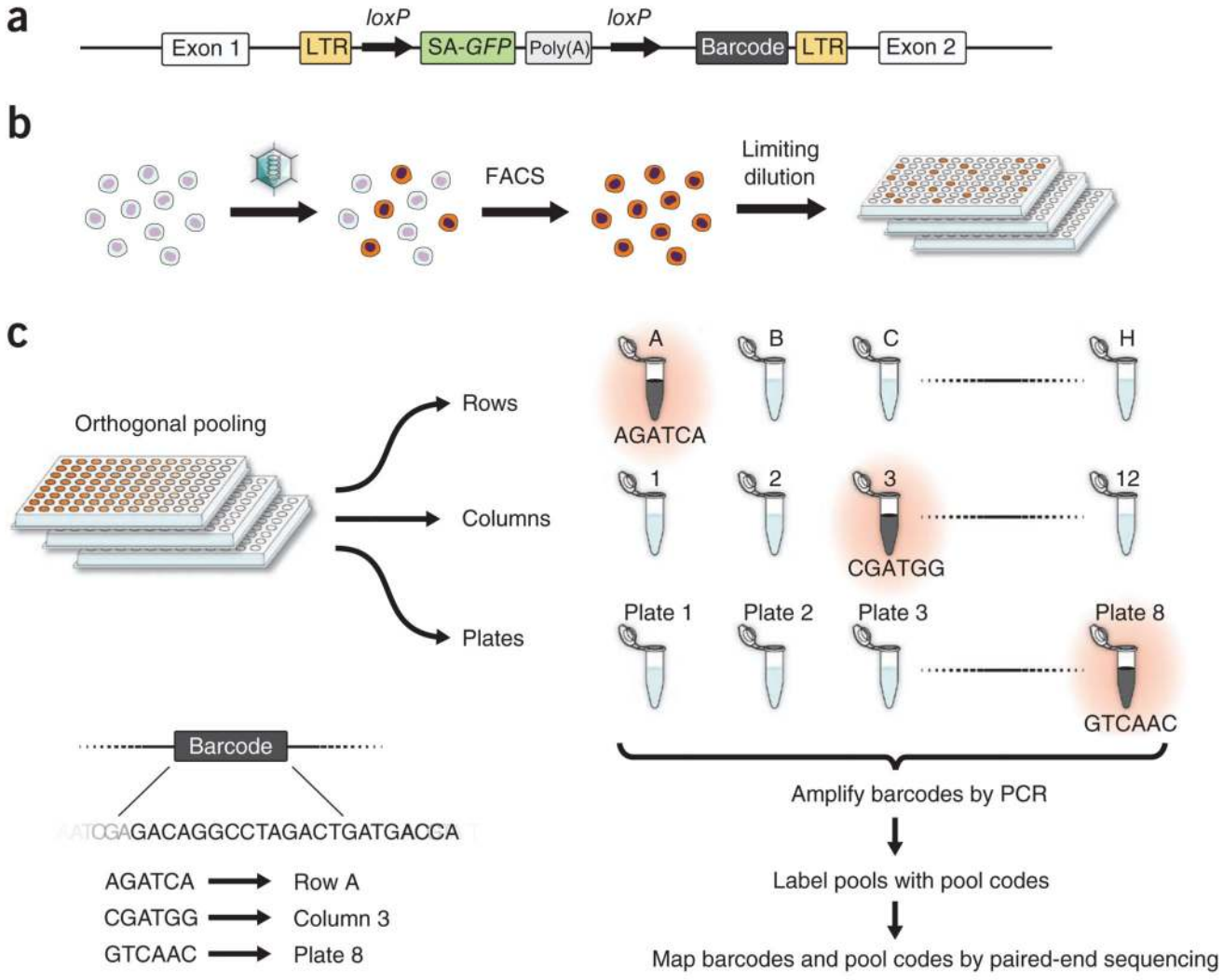

Figure 2. A pipeline for the generation of haploid 'gene-trapped' cells.

(**a**) Schematic of the retroviral gene-trap vector. SA, splice acceptor; LTR, long terminal repeat. Barcodes were introduced using a Gateway cloning strategy. (**b**,**c**) Overview of generation of clones (**b**) and mapping of barcode position (**c**).

Europe PMC Funders Author Manuscripts

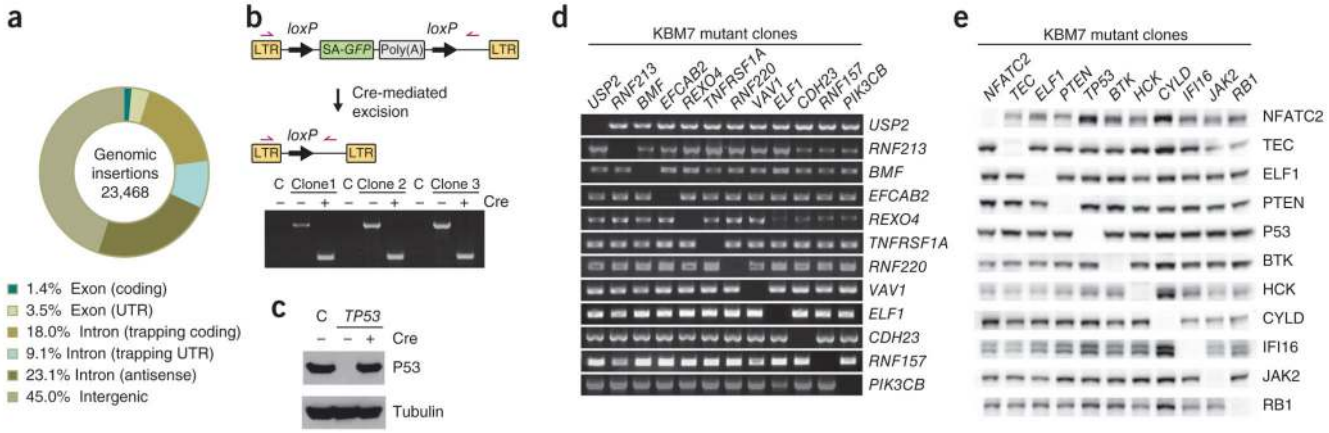

Figure 3. Gene-trapped mutants are reversible and can resemble gene knockouts.

(**a**) Distribution of all gene-trap integration sites mapped in the high-throughput sequencing pipeline. (**b**) Schematic of the retroviral gene-trap vector including parallel *loxP* recombination sites and magenta arrows indicating primer sites for PCR analysis (top). PCR analysis of genomic DNA from clones 4 d after infection with retrovirally encoded and tamoxifen-inducible Cre recombinase. Parental KBM7 cells were used as a negative control (C) for PCR specificity. SA, splice acceptor; LTR, long terminal repeat. (**c**) Western blot of KBM7 cells with a gene trap disrupting p53 expression, with and without treatment with Cre recombinase. Parental KBM7 cells were used as a control for p53 expression. (**d**) RT-PCR analysis of mRNA from indicated clones. Exonic PCR primers (indicated on the right) flanking the integration sites were designed and used to amplify cDNA from the mutants indicated at the top. (**e**) Western blot analysis of indicated clones. The antigens recognized by each antibody are indicated on the right.

Europe PMC Funders Author Manuscripts

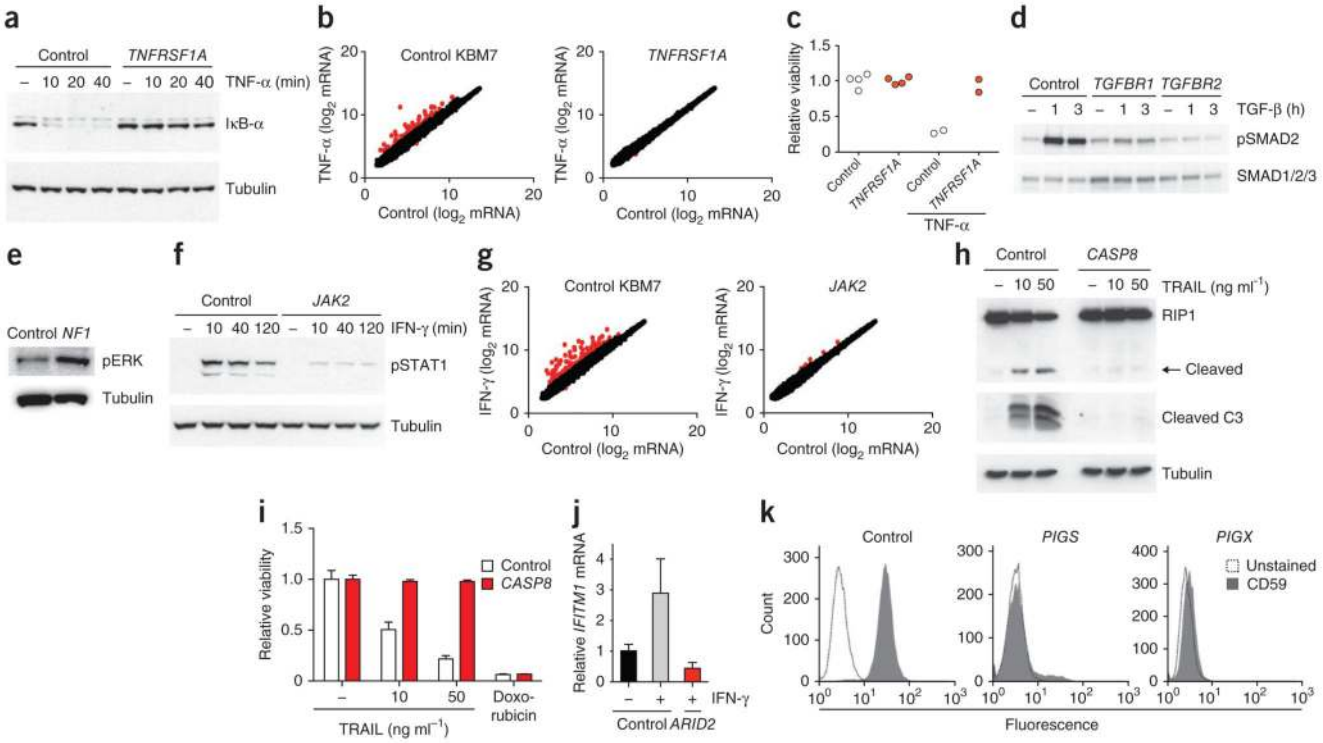

Figure 4. Molecular portraits of mutant KBM7 cells establish genotype-phenotype relationships.

(**a**) Western blot of IκB-α levels in *TNFRSF1A* mutant and control cells stimulated with TNF-α. (**b**) Transcriptome analysis of *TNFRSF1A* mutant and control cells stimulated with TNF-α for 6 h (scatter plots of Affymetrix probe sets). Fold change >2 is indicated in red. (**c**) Cell viability as measured by CellTiter-Glo of *TNFRSF1A* mutant ($n = 2$ biological replicates; two samples were treated and measured independently in the same experiment) and control cells ($n = 4$ biological replicates) treated with cycloheximide, with or without TNF-α for 12 h. Error bars, s.d. (**d**) Western blot analysis for phosphorylated SMAD2 (Ser465 and Ser467) in the indicated cells, stimulated with TGF-β or left untreated (–). (**e**) Western blot of *NF1* mutant and control cells for phosphorylated ERK (pERK). Cells were serum-starved overnight. (**f**) Western blot of phosphorylated STAT1 (pSTAT1) in *JAK2* mutant and control cells stimulated with IFN-γ. (**g**) Transcriptome analysis of *JAK2* mutant and control cells stimulated with IFN-γ for 6 h, as in **b**. (**h**) Western blot analysis for RIP1, cleaved caspase 3 (C3) and tubulin of *CASP8* mutant and control cells stimulated with TRAIL for 4 h. Arrow indicates cleaved RIP1. (**i**) Cell viability of *CASP8* mutant and control cells treated for 16 h with TRAIL. Doxorubicin served as a positive control. Error bars, s.d. ($n = 3$ independently treated replicates in the same experiment). (**j**) mRNA levels as measured by qRT-PCR of *IFITM1* in control or *ARID2* mutant cells treated with IFN-γ for 24 h or left untreated. Error bars, s.d. ($n = 3$ independent experiments). (**k**) FACS histograms of the *PIGS* and *PIGX* mutant cells stained with anti–CD59-PE antibody.

Europe PMC Funders Author Manuscripts