METHODS

Datasets. Tumour sample data are from the TCGA. Details about the cohorts and analysed samples can be found in the Supplementary Information.

Resampling methods for ranked list generation in imbalanced datasets. Preliminary testing and the ee-MWW values are reported in the Supplementary Information.

Gene Ontology networks. Gene Ontology (GO) enrichment was computed using MWW test statistics for the genes positively regulated in tumours with *FGFR3-TACC3* or other genetic alterations of interest (for example, *RAS* and *EGFR-SEPT14*). The significant GO terms from MWW-gene set test (GST) analysis (Supplementary Information) were further analysed using the Enrichment Map[28] application of Cytoscape[29]. In the network, nodes represent the terms and edges represent known term interactions and are defined by the number of shared genes between the pair of terms. Size of the nodes is proportional to statistical significance of the enrichment (Fig. 1b and Extended Data Fig. 1c) or the number of genes in the category (Fig. 3b and Extended Data Figs 5c, 6c, f, h). The overlap between gene sets is computed according to the overlap coefficient (OC), defined as:

$$\mathrm{OC} = \frac{|A \cap B|}{\min(|A|, |B|)}$$

where A and B are two gene sets, and $|X|$ equals to the number of elements within set X[30]. We set a cutoff of $\mathrm{OC} > 0.5$ to select the overlapping gene sets.

Correlation analysis between GO NES and the expression of F3–T3. We selected 19 human samples with F3–T3 fusions from ref. 31 and the TCGA fusion gene Data Portal[32]. Starting from fastq data, we applied the ChimeraScan pipeline[33] to compute the total number of reads supporting the fusion (Supplementary Table 6m). From TCGA, we obtained the legacy level 3 RNA sequencing by expectation maximization (RSEM) counts of the samples. By using the EDASeq methodology[34], we corrected the counts for GC content and applied full-quantile normalization. We transformed the normalized counts in the transcripts per million abundance quantification, applied MWW-GST to each sample and collected the NES. We used the MDSigDB collections c5.bp, c5.mf, c5.cc and hallmark collections of gene sets. We compared each gene set with the number of reads supporting the F3–T3 fusions by using the Spearman's rank correlation index (Supplementary Table 6n). To test the correlation, we assumed the alternative hypothesis of the correlation greater than zero.

Assembly of the transcriptional interactomes. To identify master regulators of the gene expression signature activated in the F3–T3-positive glioma subgroup, we first assembled independent transcriptional networks from gene expression profiles of GBM and pan-glioma datasets using the regularized gradient boosting machine algorithm (RGBM)[17] (package available from CRAN at https://cran.r-project.org/web/packages/RGBM/index.html). RGBM was used to identify regulators of the molecular subtypes of brain tumours[17,35]. We used gene expression profiles and a predefined list of 2,137 gene regulators or transcription factors (master regulators) as input. This process was independently applied to obtain GBM and pan-glioma transcriptional interactomes comprising 430,104 (median regulon size: 203) and 300,969 (median regulon size: 141) transcriptional interactions, respectively, of which 188,238 were overlapping.

Master regulator activity. To identify the master regulators of the gene expression signature activated in F3–T3-positive glioma, we modified a method that we had previously described[16]. In brief, the activity of a master regulator MR, defined as the index that quantifies the activation of the transcriptional program of that specific master regulator in each sample S_i, is calculated as follows:

$$\mathrm{Act}(S_i, \mathrm{MR}) = \frac{1}{N}\sum_{k=1}^{N} t_{ki}^{+} - \frac{1}{M}\sum_{j=1}^{M} t_{ji}^{-}$$

where t_{ki}^{+} is the expression level of the k-th positive target of the master regulator in the i-th sample, t_{ji}^{-} is the expression level of the j-th negative target of the master regulator in the i-th sample, N (or M) the number of positive (or negative) targets present in the regulon of the considered master regulator. If $\mathrm{Act}(S_i, \mathrm{MR}) > 0$, the master regulator is activated in that particular sample, if $\mathrm{Act}(S_i, \mathrm{MR}) < 0$, the master regulator is inversely activated, if $\mathrm{Act}(S_i, \mathrm{MR}) \approx 0$, it is deactivated. We used the MWW test to select master regulators that showed a significant difference between the F3–T3-positive samples and all the other samples. In Supplementary Table 7a, b, we present the list of master regulators obtained by applying master regulators analysis $\left(\left|\log_2\left(\frac{\mathrm{NES}}{1-\mathrm{NES}}\right)\right| > 2.0\right)$ and significance of differential activity <0.01.

Topological data analysis. Topological data analysis[14,15] (TDA) of the pan-glioma dataset was based on the Mapper algorithm[36]. The topological network was built using the Ayasdi platform (http://www.ayasdi.com). Several open-source implementations of Mapper are available (https://github.com/MLWave/kepler-mapper, http://danifold.net/mapper/, https://github.com/RabadanLab/sakmapper, https://github.com/paultpearson/TDAmapper). TDA was performed using the expression matrix of the top 100 genes differentially expressed between F3–T3-positive tumours and the remaining tumours as shown in Extended Data Fig. 6a. Mapper uses a dimentionality reduction algorithm and produces a topological representation of the data that preserves locality. The projection space of the dimentional-reduction algorithm is covered with overlapping bins. The data points that fall in each bin are then clustered in the original high-dimentional space. A network is constructed by assigning a node to each cluster, and clusters that share one or more samples are connected by an edge. The result is a low-dimensional network representation of the data in which nodes represent sets of samples with similar global transcriptional profiles, and edges connect nodes that have at least one sample in common. For our analysis we used 2D Locally Linear Embedding[37] as dimentional-reduction algorithm and variance normalized Euclidean metric[38] as distance. Single-linkage clustering was performed in each of the pre-images of the bins using a previously described algorithm[39]. The number of bins (resolution) for each dimension was 20 and the degree of overlap (gain) between neighbouring bins was 66%. The size of the bin was chosen such that the number of samples in each row or column of bins was the same. The open-source implementations of Mapper produce results consistent with those obtained from the Ayasdi platform[40].

Transcriptomic analysis of human astrocytes. We performed comparative analysis of gene expression of human astrocytes transduced with a lentivirus expressing F3–T3 treated with vehicle (F3–T3 and DMSO, $n = 5$ replicates), F3–T3 treated with the FGFR inhibitor PD173074 for 12 h (F3–T3 and PD173074, $n = 5$ replicates), F3–T3(K508M) treated with vehicle (F3–T3(K508M) and DMSO, $n = 3$ replicates) and empty vector treated with vehicle (vector DMSO, $n = 3$ replicates). Expression data were obtained using the Illumina human HT12v4 gene expression array. The list of 4,034 differentially expressed genes between the F3–T3 and DMSO and F3–T3 and PD173074 groups (t-test $P < 0.01$ and MWW test $P < 0.01$) was used to construct a heat map comprising the whole dataset in which vector and DMSO and F3–T3(K508M) and DMSO are control groups. Samples were clustered using the hierarchical clustering algorithm based on the Ward linkage method and Euclidean distance as implemented in R. Finally, the GO enrichment analysis was performed using the ranked list obtained from three independent comparisons: F3–T3 versus F3–T3 treated with PD173074; F3–T3 versus F3–T3(K508M); F3–T3 versus vector using the Java version of GSEA. For each comparison, statistically significant GO terms with $Q < 10^{-6}$ were selected. The statistically significant pathways common to all three comparisons were included in the construction of the visual network using the Enrichment Map application[28] of Cytoscape[29]. The microarray data have been deposited in ArrayExpress with accession number E-MTAB-6037.

Identification of proteins phosphorylated by the F3–T3 gene fusion using mass spectrometry. Cells were lysed in buffer containing 9 M urea, 20 mM HEPES pH 8.0, 0.1% SDS and a cocktail of phosphatase inhibitors. Six milligrams of protein were reduced with 4.5 mM DTT, alkylated with 10 mM iodoacetamide and digested with trypsin overnight at 37 °C. Samples were desalted on a C18 cartridge (Sep-Pak plus C18 cartridge, Waters). Each sample was prepared in triplicate. Phosphopeptide enrichments were performed as previously described[41]. An LTQ Orbitrap XL (ThermoFisher) in-line with a Paradigm MS2 HPLC (Michrom bioresources) was used to acquire high-resolution mass spectrometry and tandem mass spectrometry data. Technical duplicate data for each of the metal-oxide affinity chromatography elutions and triplicate data for the phosphotyrosine immunoprecipitation samples were acquired.

RAW files were converted to mzXML using msconvert[42] and searched against the Swissprot Human protein database (9 January 2013 release) appended with common proteomics contaminants and reverse sequences as decoys. Searches were performed with X!Tandem (version 2010.10.01.1) using the k-score plugin[43,44]. For all searches the following search parameters were used: parent monoisotopic mass error of 50 parts per million (p.p.m.); fragment ion error of 0.8 daltons; allowing for up to two missed tryptic cleavages. Variable modifications were oxidation of methionine (+15.9949@M), carbamidomethylation of cysteine (+57.0214@C), and phosphorylation of serine, threonine, and tyrosine (+79.9663@[STY]). The search results were then post-processed using PeptideProphet and ProteinProphet[45–47]. Spectral counts were obtained for each cell line using ABACUS[48]. Immunoprecipitation data of phospho-tyrosine enrichment were processed through ABACUS separately from the MOAC enrichment data. ABACUS results were filtered to only retain proteins with a ProteinProphet probability >0.7. Only phosphorylated peptides with a probability >0.8 were considered for spectral counting. For tyrosine enrichment these ABACUS parameters resulted in a protein false discovery rate (FDR) of 0.0045. This ABACUS output was used for all subsequent analysis to quantify the relative abundance of phosphorylated peptides or proteins. Phospho-site localization was performed with an in-house reimplementation of the Ascore algorithm as previously described[49]. Ascore values represent the probability of detection owing to chance, with scores