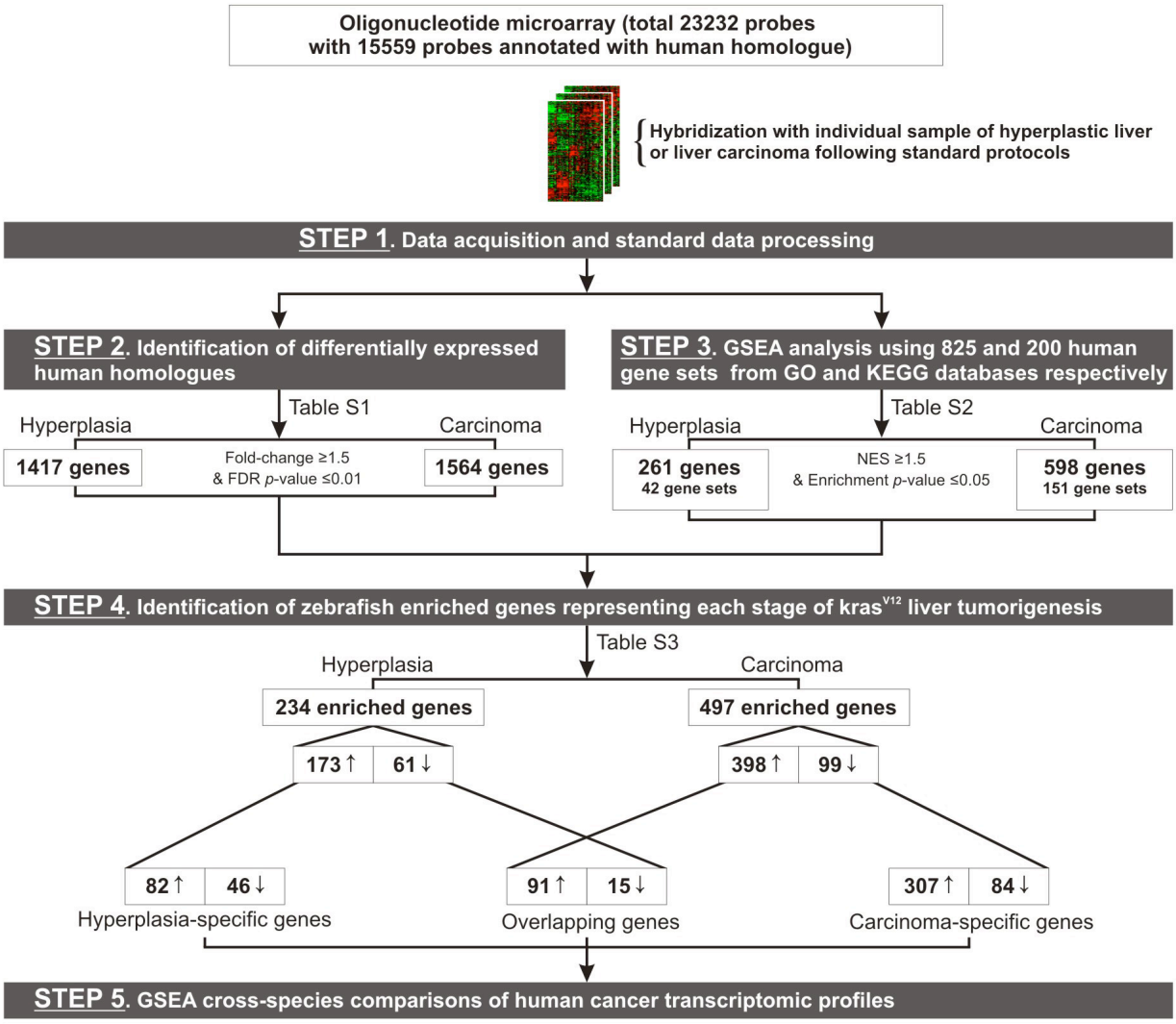

Fig. S3. Flowchart of microarray data analysis. **Step 1:** Data acquisition and standard data processing. Arrays were scanned for fluorescence signal detection using the GenePix 4000B Axon scanner (Molecular Devices, CA) and the generated images were analyzed using GenePix Pro® 4.0 image analysis software (Molecular Devices) to measure fluorescence signal intensities. All the arrays gave a mean signal-to-background ratio of more than 5 and >95% of the probes gave measurable signals. Microarray raw data were extracted and formatted to conform to the MIAME (Minimal Information About a Microarray Experiment) standards, and

further subjected to Lowess normalization and transformation to $\log_2$ ratio using the R system software (http://www.R-project.org) (Lam et al., 2009a). Differences between the $\log_2$ ratios of transgenic and normal livers were calculated to obtain the fold-change values which indicate up- or down-regulation in gene expression. Statistical significance of the expression level was performed using Student's t-test to yield the p-value of each gene. The resulting p-values were then adjusted for Benjamini and Hochberg false discovery rate (FDR) to minimize the number of false positive genes that could be identified by chance. **Step 2:** Identification of differentially-expressed genes having human homologues. Zebrafish genes having human homologues and differentially expressed in hyperplasia and carcinoma were filtered using the selection criteria of fold-change ≥1.5 and FDR-adjusted p-value ≤0.01 as statistical cut-offs (supplementary material Table S1). **Step 3:** GSEA analysis using 825 and 200 predefined human gene sets from the GO and KEGG databases respectively. GSEA version 2.0.1 from the Broad Institute was used with predefined Gene Ontology (GO) and KEGG gene sets retrieved from the GSEA Molecular Signature Database (MSigDB v2.5). Detailed description of each gene set can be found at the MSigDB website (www.broad.mit.edu/gsea/msigdb/index.jsp). GSEA analysis was completed using phenotypic permutation with a weighted enrichment statistic and the Ratio_of_Classes metric to rank genes. One thousand permutations of the data were completed to obtain a false discovery rate (FDR) q-value (Subramanian et al., 2005). Human gene sets with normalized enrichment score (NES) ≥1.5, nominal (NOM) enrichment p-value ≤0.05 and FDR q-value ≤0.25 were considered significant (supplementary material Table S2). The genes within each significant human gene set that contributed maximally to the GSEA score in zebrafish hyperplasia and HCC were also identified. **Step 4:** Identification of zebrafish enriched genes representing each stage of $kras^{V12}$ liver tumorigenesis. The zebrafish differentially-expressed

genes having human homologues which satisfied the statistical cut-offs in Step 2 and the genes obtained by GSEA analysis in Step 3 were then compared to identify the genes present in both steps. These genes were termed as zebrafish enriched genes. The enriched genes corresponding to each stage of $kras^{V12}$ liver tumorigenesis were then overlapped to obtain the hyperplasia-specific enriched genes, carcinoma-specific enriched genes as well as overlapping enriched genes between two stages (supplementary material Table S3). Arrows indicate up- and down-regulated genes. **Step 5:** GSEA cross-species comparisons of human cancer transcriptomic profiles. Raw data files of different published human cancer data sets were obtained from the Gene Expression Omnibus (GEO) database. All of the retrieved data sets were generated using the Affymetrix GeneChip® platform. Probes with *P*-call values of <80% were then discarded across all samples to minimize variations from biological samples and identify only reliably measured genes. The lists of up- and down-regulated zebrafish enriched genes from Step 4 were used for GSEA cross-species comparison with these human cancer data sets. Statistical significance of the cross-species comparison was determined by the selection criteria of false discovery rate q-value (FDR q-value) ≤0.05 and family-wise error rate p-value (FWER p-value) ≤0.05.

References

Lam, S.H., Krishna, Murthy, Karuturi, R. and Gong, Z. (2009a). Zebrafish spotted-microarray for genome-wide expression profiling experiments: data acquisition and analysis. *Methods Mol. Biol.* **546**, 197-226.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545-15550.