Supplementary Information

Supplementary Table 1. Mutated gene data. For each of the mutated genes, the associated tumour type and the source of information (reference database or published article) are given, as well as the gene annotations in Kegg (Kanehisa et al, 2008), Reactome (Matthews et al, 2009), Biocarta (http://www.biocarta.com/), Gene Ontology Biological Process (Ashburner et al, 2000) and Interpro (Hunter et al, 2009).

Supplementary Table 2. Pathways, processes and protein domains containing a significant number of mutated genes in the different tumour types. All results with Q-values lower than 0.1 are shown.

Supplementary Methods. Complementary information for the statistical test and online pathways and processes resources.

Supplementary Methods

Introduction

The identification of pathways and processes containing more mutated genes than expected by chance in each tumour type has been done with a Fisher statistical test. This test implies the definition of a statistical background to identify significant pathways/processes while taking into account how many pathway/process genes have been screened for mutations in each tumour type. However, the dataset of 5,272 mutated genes have been created by combining many information sources (e.g., high-throughput resequencing data, literature data catalogued in databases), and is very heterogeneous. While analyzing the significance of genome-wide tumour resequencing studies is easy since all of them explore the full genome, low-scale or individual resequencing experiments are very different : These experiments are designed to evaluate a few genes and they report only positive results. In this case, the size of the dataset that was assessed is undefined. To complicate things further a number of datasets are of intermediate nature (for example the resequencing studies of gene families, such as kinases (Greenman et al, 2007).

To gain confidence in the significance of the statistical enrichment results, we designed the following experiment to compare the results obtained while analysing 2 different large-scale resequencing studies in the same tumour type.

Parallel pathway/process statistical enrichment for 2 large-scale colorectal resequencing studies

The following experiment was designed to estimate the differences in pathways/processes associated to a given tumour-type while computing the enrichment analysis in parallel for 2 large-scale resequencing studies, each associated to its proper statistical background (complete set of genes that have been screened for mutations). To our knowledge, no tumour type have been screen twice for the same dataset of genes (neither screened twice genome-wide). Hence, we chose to compare mutated genes identified in colorectal tumours 1) by a kinase resequencing study (Greenman et al, 2007) and 2) by a genome-wide resequencing study (Wood et al, 2007). Mutated gene data extracted from literature and small-scale studies were added to these 2 datasets identically (Wang et al, 2004; Bardelli et al, 2003; Forbes et al, 2008; Thomas et al, 2007; Futreal et al, 2004; Hamosh et al, 2005). We challenged pathways and processes extracted from Biocarta (http://www.biocarta.com/), Kegg (Kanehisa et al, 2008) and Reactome (Matthews et al, 2009) (an integrated display of these pathways/processes can be found in the HPD database (Chowbina et al, 2009)). Few significant results were obtained with the previously defined Q-value threshold of 0.01 (see supplementary Method Tables below). However, comparing the ranked list of obtained pathways, we observed that the first bin (all results were divided in 5 bins of equal sizes, the first bins of each database are presented in supplementary Method Tables A, B and C) encompassing the lowest Q-values, contain more common pathways/processes than expected by chance (P-value < 0.003) for the 3 pathway databases (P-value score estimated as the number of common pathways/processes obtained after 10,000 bin randomisation).

Furthermore, the real analysis, obtained after combining the large-scale datasets, and defining the union of screened genes as a background, mainly identify pathway/process that are also retrieved in the separated parallel analyses (supplementary Method Tables).

Our conclusion is that the pathways/processes obtained for the parallel analyses of 2 large-scale colorectal cancer resequencing studies, would not give many significant results when considered separately, even if they are comparable and point to very similar results. Merging the mutated genes data with different backgrounds does not bias the results (the observations are essentially the sames in the different sources), and positively contributes to give them an additional statistical significance. Furthermore, this experiment also indicated that the mutated genes data coming from literature databases and low-scale studies, that were added identically to the 2 datasets for the parallel analysis, and that are less prone to false-positives, contribute greatly to the discovery of tumour-associated pathways and processes.

Beyond this specific issue we would like to point out that the system and the type of analysis presented here will evolve towards the inclusion of larger data sets with all the cancer genome projects in course: Time will make the system increasingly robust and supported by homogeneous large-scale studies. The statistical significance assessment will be not only easier, but also closer to what is anticipated in this initial publication. Furthermore, we can also imagine in the future integrating different -omics data to study cancer-related alterations of cellular pathways (Balestrieri et al, 2009).

Supplementary Method Table

Colorectal cancer mutated genes
Biocarta, Kegg and Reactome pathway/process enrichment analyses

For each database, the pathway/processes in the first bin are presented (all results were divided in 5 bins of equal sizes). Pathways/process retrieved in the first bins in the 2 parallel experiements, as well as in the real experiment presented in the paper, are coloured in blue

A) Biocarta

Common Pathways/Processes	
21/31	P<0.0001

Genome-wide resequencing (Wood et al.)	Wood q-value	Kinase-family resequencing (Greenman et al.)	Greenman q-value
h_p53hypoxiaPathway	0.020	h_p53hypoxiaPathway	0.330
h_tgfbPathway	0.020	h_ctcfPathway	0.330
h_mTORPathway	0.050	h_trkaPathway	0.330
h_ctcfPathway	0.070	h_chemicalPathway	0.380
h_eif4Pathway	0.090	h_cblPathway	0.380
h_her2Pathway	0.090	h_cardiacegfPathway	0.420
h_igf1mtorpathway	0.100	h_tffPathway	0.580
h_trkaPathway	0.120	h_vegfPathway	0.600
h_crebPathway	0.120	h_ps1Pathway	0.600
h_pitx2Pathway	0.120	h_cdc42racPathway	0.600
h_cblPathway	0.120	h_mTORPathway	0.600
h_egfPathway	0.150	h_egfPathway	0.600
h_alkPathway	0.180	h_ifnaPathway	0.600
h_gsk3Pathway	0.190	h_plcPathway	0.600
h_ps1Pathway	0.220	h_erbB4pathway	0.600
h_HBxPathway	0.220	h_telPathway	0.600
h_tffPathway	0.250	h_RELAPathway	0.600
h_vegfPathway	0.250	h_edg1Pathway	0.600
h_shhPathway	0.250	h_tgfbPathway	0.600
h_cdc42racPathway	0.250	h_bcellsurvivalPathway	0.600
h_no1Pathway	0.290	h_alkPathway	0.650
h_igf1rPathway	0.290	h_cell2cellPathway	0.680
h_cell2cellPathway	0.290	h_eif4Pathway	0.680
h_ptdinsPathway	0.310	h_arfPathway	0.680
h_arfPathway	0.310	h_no1Pathway	0.680
h_ptenPathway	0.310	h_pitx2Pathway	0.700
h_telPathway	0.310	h_igf1mtorpathway	0.850
h_il4Pathway	0.310	h_hesPathway	0.850
h_erkPathway	0.360	h_mef2dPathway	0.850
h_edg1Pathway	0.410	h_tob1Pathway	0.850
h_akapCentrosomePathw	0.410	h_her2Pathway	0.870

Statistical significant results for the merged datasets (results presented in the manuscript)

h_trkaPathway	Trka Receptor Signaling Pathway	0.01
h_p53hypoxiaPathway	Hypoxia and p53 in the Cardiovascular system	0.01

B) Kegg

Common Pathways/Processes	
12/15	P<0.0001

Genome-wide resequencing (Wood et al.)	Wood q-value	Kinase-family resequencing (Greenman et al.)	Greenman q-value
hsa05210	0.000	hsa05210	0.000
hsa05213	0.000	hsa05213	0.002
hsa05218	0.000	hsa05214	0.003
hsa05212	0.000	hsa05218	0.005
hsa05215	0.000	hsa05212	0.005
hsa04520	0.000	hsa04012	0.006
hsa05216	0.000	hsa05215	0.014
hsa05214	0.000	hsa05211	0.023
hsa05223	0.000	hsa05219	0.031
hsa05219	0.000	hsa04320	0.032
hsa04510	0.000	hsa04510	0.040
hsa04012	0.000	hsa05223	0.043
hsa04320	0.000	hsa04070	0.064
hsa05220	0.000	hsa04530	0.135
hsa05221	0.000	hsa05220	0.140

Statistical significant results for the merged datasets (results presented in the manuscript)

hsa05210	Colorectal cancer	6.7E-013
hsa05213	Endometrial cancer	5.4E-010
hsa04012	ErbB signaling pathway	4.1E-009
hsa05214	Glioma	1.5E-008
hsa05212	Pancreatic cancer	1.4E-007
hsa05215	Prostate cancer	3.7E-007
hsa05218	Melanoma	6.5E-007
hsa04510	Focal adhesion	1.9E-006
hsa05223	Non-small cell lung cancer	2.0E-006
hsa05221	Acute myeloid leukemia	3.2E-006
hsa05219	Bladder cancer	1.1E-005
hsa05216	Thyroid cancer	2.5E-005
hsa04520	Adherens junction	2.5E-005
hsa05211	Renal cell carcinoma	1.6E-004
hsa05220	Chronic myeloid leukemia	1.7E-004
hsa04360	Axon guidance	2.9E-004
hsa04912	GnRH signaling pathway	2.9E-004
hsa04320	Dorso-ventral axis formation	2.9E-004
hsa04730	Long-term depression	4.4E-004
hsa04370	VEGF signaling pathway	5.6E-004
hsa04010	MAPK signaling pathway	6.7E-004
hsa04540	Gap junction	9.6E-004
hsa04930	Type II diabetes mellitus	9.6E-004
hsa04916	Melanogenesis	1.0E-003
hsa04720	Long-term potentiation	1.0E-003
hsa04664	Fc epsilon RI signaling pathway	1.0E-003
hsa05217	Basal cell carcinoma	1.7E-003
hsa04920	Adipocytokine signaling pathway	2.1E-003
hsa04150	mTOR signaling pathway	2.2E-003
hsa04910	Insulin signaling pathway	4.1E-003
hsa04662	B cell receptor signaling pathway	4.9E-003
hsa05222	Small cell lung cancer	5.0E-003
hsa04310	Wnt signaling pathway	5.8E-003

C) Reactome

Common Pathways/Processes	
04/07	P<0.003

Genome-wide resequencing (Wood et al.)	Wood q-value	Kinase-family resequencing (Greenman et al.)	Greenman q-value
REACT_9417	0.01	REACT_13685	0.08
REACT_498	0.01	REACT_16888	0.37
REACT_16888	0.03	REACT_9417	0.37
REACT_11061	0.03	REACT_604	0.39
REACT_14797	0.05	REACT_1505	0.46
REACT_6844	0.06	REACT_152	0.46
REACT_13685	0.25	REACT_498	0.46

Statistical significant results for the merged datasets (results presented in the manuscript)

REACT_498	Signaling by Insulin receptor	0
REACT_11061	Signalling by NGF	0.01

Pathways, processes and other nework-related resources

Pathways and Process databases

Kegg (Kanehisa et al, 2008)
http://www.genome.jp/kegg/

Reactome (Matthews et al, 2009)
http://www.reactome.org/

Biocarta
http://www.biocarta.com/genes/index.asp

Gene Ontology (Ashburner et al, 2000)
http://www.geneontology.org/

BioPax
http://www.biopax.org/

HPD (Chowbina et al, 2009)
http://discern.uits.iu.edu:8340/HPD/help.php

Protein interaction databases

HPRD (Peri et al, 2003)
http://www.hprd.org/

Intact (Hermjakob et al, 2004)
http://www.ebi.ac.uk/intact/main.xhtml

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM & Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet* **25:** 25-29

Balestrieri C, Alberghina L, Vanoni M & Chiaradonna F (2009) Data recovery and integration from public databases uncovers transformation-specific transcriptional downregulation of cAMP-PKA pathway-encoding genes. *BMC Bioinformatics* **10 Suppl 12:** S1

Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, Saha S, Markowitz S, Willson JKV, Parmigiani G, Kinzler KW, Vogelstein B & Velculescu VE (2003) Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300:** 949

Chowbina SR, Wu X, Zhang F, Li PM, Pandey R, Kasamsetty HN & Chen JY (2009) HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics* **10 Suppl 11:** S5

Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA & Stratton MR (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10:** Unit 10.11

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N & Stratton MR (2004) A census of human cancer genes. *Nat Rev Cancer* **4:** 177-83

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J et al (2007) Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153-8

Hamosh A, Scott AF, Amberger JS, Bocchini CA & McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:** D514-7

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D & Apweiler R (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32:** D452-455

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T & Yamanishi Y (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36:** D480-4. http://www.genome.jp/kegg/

Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L & D'Eustachio P (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37:** D619-22. http://www.reactome.org/

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC et al (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13:** 2363-2371

Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, Wang M, Feng W, Zander T, MacConaill L, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, Ziaugra L, Wong K, Gabriel S, Beroukhim R et al (2007) High-throughput oncogene mutation profiling in human cancer. *Nat Genet* **39:** 347-51

Wang Z, Shen D, Parsons DW, Bardelli A, Sager J, Szabo S, Ptak J, Silliman N, Peters BA, van der Heijden MS, Parmigiani G, Yan H, Wang T, Riggins G, Powell SM, Willson JKV, Markowitz S, Kinzler KW, Vogelstein B & Velculescu VE (2004) Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* **304:** 1164-6

Wood LD, Parsons DW, Jones S, Lin J, SjÃ¶blom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z et al (2007) The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108-13