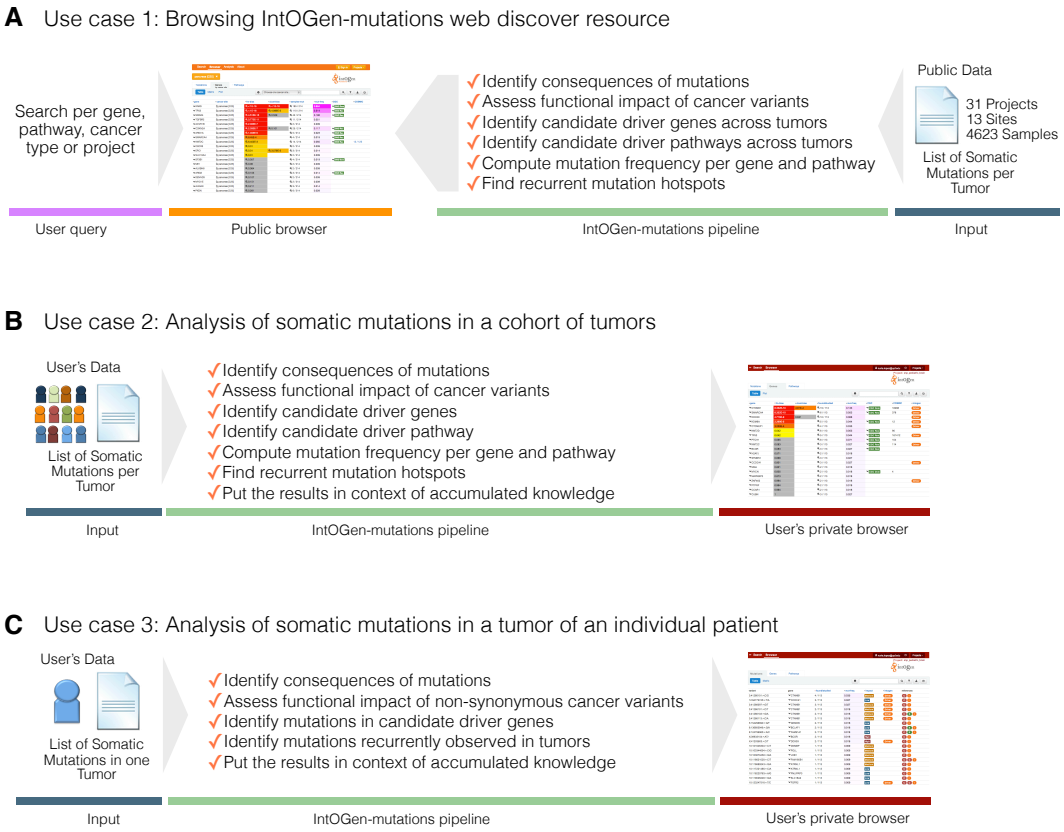
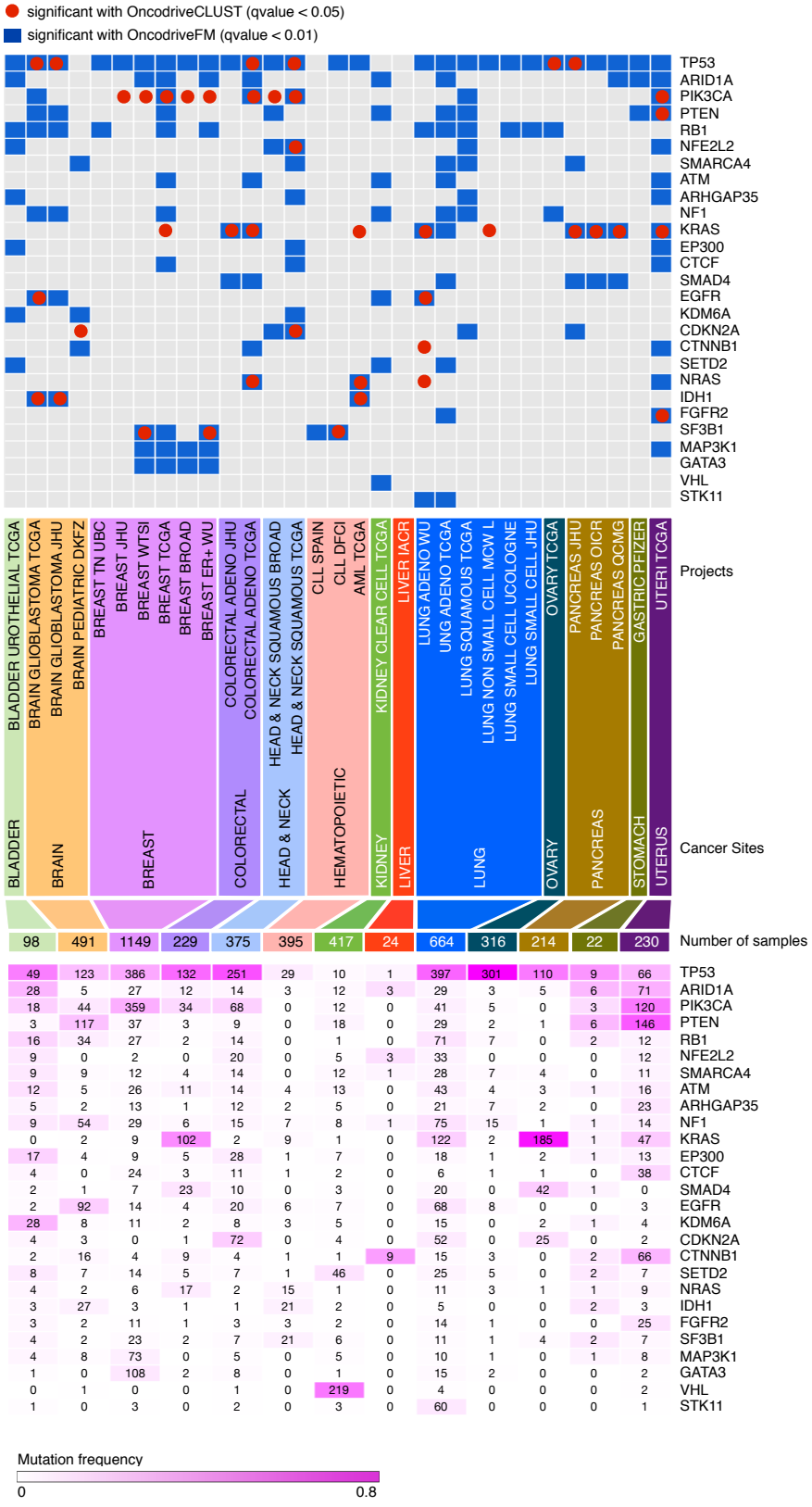


Supplementary Figure 1. Diagram representing three different use cases of the IntOGen-mutations resource. A) The IntOGen-mutations pipeline is used to compute the results of periodically obtained public data and populate the IntOGen web discovery tool. Users can browse the results of the pipeline through the IntOGen-mutations public browser searching per gene, pathway, cancer type or tumor sequencing project.

B) Users employ the IntOGen-mutations pipeline, either through the web server or locally to analyze their own data, thus obtaining a private browser with their results. In the use case 2, users input a list of somatic mutations from a cohort of tumor samples to identify putative driver genes. The results can be browsed within the context of accumulated knowledge in IntOGen.

C) The pipeline can also be used to rank somatic mutations identified in the tumor of an individual patient to assess their potential implication in cancer development.

Nature Methods: doi:10.1038/nmeth.2642



Supplementary Figure 2. Results in IntOGen for a list of selected genes. The upper panel indicates if the genes are detected as drivers by OncodriveFM (blue squares) or OncodriveCLUST (red circles) in each project. In the lower panel the projects are aggregated by cancer site. Numbers indicate the total of samples with mutations in each gene and cancer site; the frequency is shown in a color scale from white to purple.

Nature Methods: doi:10.1038/nmeth.2642

Supplementary Table 1

Somatic mutations datasets included within the IntOGen-mutations web discovery tool

Project name	Description	Site	Institution	Obtained from	Pubmed	Tumor samples
BLADDER UROTHELIAL TCGA	Bladder urothelial carcinoma	bladder	TCGA	Synapse	-	98
BRAIN GLIOBASTOMA TCGA	Glioblastoma multiforme	brain	TCGA	Synapse	18772890	290
BRAIN GLIOBASTOMA JHU	Glioblastoma multiforme		Johns Hopkins University	ICGC DCC	18772396	88
BRAIN PEDIATRIC DKFZ	Pediatric brain tumors		DKFZ	ICGC DCC	22265402/ 22286061	113
BREAST JHU	Breast cancer	breast	Johns Hopkins University	ICGC DCC	17932254	42
BREAST WTSI	Breast cancer		Welcome Trust/ Sanger Institute	ICGC DCC	22722201	100
BREAST TN UBC	Triple negative breast cancer		University of British Columbia	SM	22495314	65
BREAST TCGA	Breast invasive carcinoma		TCGA	Synapse	23000897	762
BREAST BROAD	Breast cancer		BROAD Institute	SM	22722202	103
BREAST ER+ WU	ER+ breast cancer		Washington University	SM	22722193	77
COLORECTAL ADENO JHU	Colorectal adenocarcinoma	colorectal	Johns Hopkins University	ICGC DCC	17932254	36
COLORECTAL ADENO TCGA	Colorectal adenocarcinoma		TCGA	Synapse	22810696	193
HEAD/NECK SQUAMOUS BROAD	Head and neck squamous cell carcinoma	Head & neck	Broad Institute	SM	21798893	74
HEAD/NECK SQUAMOUS TCGA	Head and neck squamous cell carcinoma		TCGA	Synapse	-	301
CLL SPAIN	Chronic lymphocytic leukemia	hematopoietic	Spanish Ministry of Science	ICGC DCC	21642962/ 22158541	109
CLL DFCI	Chronic lymphocytic leukemia		Dana Farber Cancer Institute	SM	22150006	90
AML TCGA	Acute myeloid leukemia		TCGA	Synapse	-	196
KIDNEY CLEAR CELL	Kidney clear cell carcinoma	kidney	TCGA	Synapse	-	417

TCGA						
LIVER IARC	Liver cancer	liver	IACR	ICGC DCC	22561517	24
LUNG ADENO WU	Lung adenocarcinoma	lung	Washington University School of Medicine	ICGC DCC	18948947	162
LUNG NON SMALL CELL MCW	Non small cell lung cancer		Medical College of Wisconsin	SM	22510280	31
LUNG SQUAMOUS TCGA	Lung squamous cell carcinoma		TCGA	Synapse	22960745	174
LUNG ADENO TCGA	Lung adenocarcinoma		TCGA	Synapse	-	228
LUNG SMALL CELL UCOLOGNE	Small cell lung cancer		University Cologne	SM	22941188	27
LUNG SMALL CELL JHU	Small cell lung cancer		Johns Hopkins University	SM	22941189	42
OVARY TCGA	Ovarian serous cystadenocarcinoma	ovary	TCGA	Synapse	21720365	316
PANCREAS JHU	Pancreatic cancer	pancreas	Johns Hopkins University	ICGC DCC	18772397	114
PANCREAS OICR	Pancreatic cancer		Ontario Institute for Cancer Research	ICGC DCC	-	33
PANCREAS QCMG	Pancreatic cancer		Queensland Centre for Medical Genomics	ICGC DCC	-	67
GASTRIC PFIZER	Gastric cancer	stomach	Pfizer Worldwide Research and Development	SM	22037554	22
UTERI TCGA	Uterine corpus endometrioid carcinoma	uterus	TCGA	Synapse	-	230

Supplementary Table 2

Number of likely driver genes identified by OncodriveFM and OncodriveCLUST in each tumors genome re-sequencing project analyzed.

Legend
CLUST biased genes: genes with significant mutations clustering
FM biased genes: genes with significant FM bias
Total: number of genes with one positive selection mark
Known: number of genes already known to be involved in cancer (annotated in the Cancer Gene Census)
r: ratio of known cancer genes-to-total genes within all genes exhibiting one positive selection mark

Project name	Cancer site	Samples	FM biased genes		CLUST biased genes	
			Total	Known (r)	Total	Known (r)
BLADDER UROTHELIAL TCGA	bladder	98	24	9 (0.375)	2	0 (0)
BRAIN GLIOBASTOMA TCGA	brain	290	26	11 (0.42)	12	4 (0.33)
BRAIN GLIOBASTOMA JHU		88	6	6 (1)	3	2 (0.67)
BRAIN PEDIATRIC DKFZ		113	5	3 (0.6)	2	1 (0.5)
BREAST JHU	breast	42	2	1 (0.5)	0	0
BREAST WTSI		100	8	6 (0.75)	3	3 (1)
BREAST TN UBC		65	1	1 (1)	2	1 (0.5)
BREAST TCGA		762	39	18 (0.46)	5	3 (0.6)
BREAST BROAD		103	7	5 (0.71)	7	3 (0.43)
BREAST ER+ WU		77	10	7 (0.7)	2	2 (1)
COLORECTAL ADENO JHU	colorectal	36	4	4 (1)	2	2 (1)
COLORECTAL ADENO TCGA		193	25	12 (0.48)	7	6 (0.86)
HEAD/NECK SQUAMOUS BROAD	Head and neck	74	7	3 (0.43)	2	2 (1)
HEAD/NECK SQUAMOUS TCGA		301	58	16 (0.28)	45	7 (0.15)
CLL SPAIN	hematopoietic	109	1	1 (1)	1	1 (1)
CLL DFCI		90	3	3 (1)	2	2 (1)
AML TCGA		196	15	15 (1)	9	9 (1)
KIDNEY CLEAR CELL TCGA	kidney	417	21	11 (0.52)	4	0 (0)
LIVER IARC	liver	24	0	0	1	1 (1)
LUNG ADENO WU	lung	162	4	4 (1)	4	3 (0.75)
LUNG NON SMALL CELL MCW		31	3	1 (0.33)	0	0
LUNG SQUAMOUS TCGA		174	50	9 (0.18)	0	0
LUNG ADENO TCGA		228	77	10 (0.13)	1	1 (1)
LUNG SMALL CELL UCOLOGNE		27	2	2 (1)	0	0
LUNG SMALL CELL JHU		42	6	2 (0.33)	0	0
OVARY TCGA	ovary	316	10	6 (0.6)	1	1 (1)
PANCREAS JHU	pancreas	114	7	6 (0.86)	3	2 (0.67)
PANCREAS OICR		33	4	3 (0.75)	1	1 (1)
PANCREAS QCMG		67	7	4 (0.57)	2	1 (0.5)
GASTRIC PFIZER	stomach	22	19	4 (0.21)	0	0
UTERI TCGA	uterus	230	75	19 (0.25)	5	5 (1)

Nature Methods: doi:10.1038/nmeth.2642

Supplementary Table 3

Number of likely driver genes identified by OncodriveFM and OncodriveCLUST in each cancer site analyzed.

Legend
CLUST biased genes: genes with significant mutations clustering
FM biased genes: genes with significant FM bias
Total: number of genes with one positive selection mark
Known: number of genes already known to be involved in cancer (annotated in the Cancer Gene Census)
r: ratio of known cancer genes-to-total genes within all genes exhibiting one positive selection mark

Site	Number of projects	Samples	FM biased genes		CLUST biased genes	
			Total	Known (r)	Total	Known (r)
Bladder	1	98	24	9 (0.375)	2	0 (0)
Brain	3	491	31	14 (0.45)	15	5 (0.33)
Breast	6	1148	36	16 (0.44)	13	6 (0.46)
Colorectal	2	229	23	13 (0.56)	7	6 (0.86)
Head and neck	2	375	58	16 (0.28)	45	7 (0.16)
Hematopoietic	3	395	18	17 (0.94)	12	12 (1)
Kidney	1	417	21	11 (0.52)	4	0 (0)
Liver	1	24	0	0	1	1 (1)
Lung	6	664	126	19 (0.15)	3	3 (1)
Ovary	1	316	10	6 (0.6)	1	1(1)
Pancreas	3	214	9	6 (0.67)	4	2 (0.5)
Stomach	1	22	19	4 (0.21)	1	0
Uterus	1	230	75	19 (0.25)	5	5 (1)
GLOBAL	31	4623	347	76 (0.22)	89	29 (0.33)

Supplementary Note 1
Comparison of candidate driver genes detected by IntOGen-mutations pipeline and those reported in the original publication

In order to further validate the results obtained with IntOGen-mutations pipeline we compared the results (lists of driver genes) reported in original publications with those obtained with the pipeline for some of the projects. We choose four projects that have used well-known methods to identify significantly mutated genes and for which we have used exactly the same list of somatic mutations as input for the pipeline. Note that for some TCGA projects included in IntOGen we have used the most up-to-date list of mutations, which may not correspond to the one reported in the publication, as more tumors sequences may have been made available after publication, for this reason those were not used for comparison. As a result of the comparison, we found that the methods to identify driver genes included in IntOGen-mutations (OncodriveFM and OncodriveCLUST) are able to identify most of the *bona fide* cancer drivers identified in the original manuscripts and often also identify as drivers some known cancer genes not reported in the original manuscript. Next, we discuss these comparisons in detail.

Comparison with Barneji et al (BREAST BROAD)

Banerji et al[1] report the identification of 10 significantly mutated genes using MutSig FDR < 0.1. They discarded 4 of them after manual review of reads and subsequent orthogonal confirmation of somatic events (see Supplementary Table 6A from Banerji et al). The IntOGen-mutations pipeline identifies the 6 significantly mutated genes detected by MutSig which are kept after the manual review and orthogonal confirmation (namely TP53, CBFB, GATA3, MAP3K1, PIK3CA and AKT1), while it does not detect as significant any of the discarded genes (namely PCGF2, ZBED4, WEE1, BZRAP1). In addition, OncodriveFM detects 3 other genes (NOTCH2, MLL and ARK1C3) and OncodriveCLUST identifies 3 more (ERBB2, RSBN1L and PRKCZ) (Figure S1). Note that both OncodriveFM and OncodriveCLUST identify AKR1C3. Three of the genes detected by the IntOGen-mutations pipeline and not MutSig are well-known cancer genes (NOTCH2, MLL and ERBB2), which reinforce the validity of the results. The other genes are novel candidates -and always we have to consider that some may be false positives- until their oncogenic role has been established.

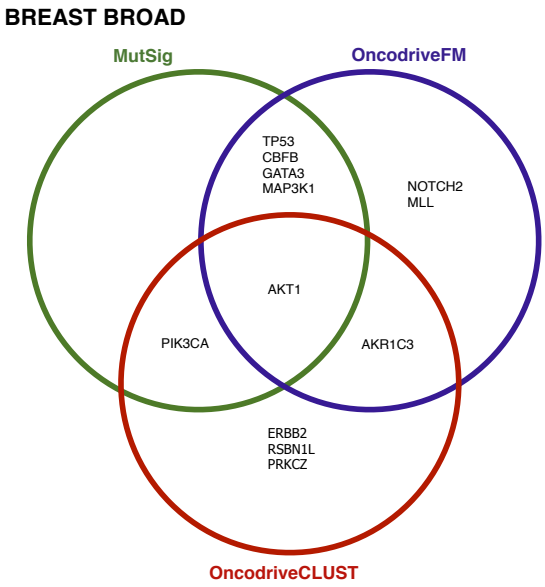


Figure S1. Venn diagram showing the overlap between genes detected by MutSig in Barneji et al and those identified by OncodriveFM and OncodriveCLUST using IntOGen-mutations pipeline.

Comparison with Stransky et al (HNSCC BROAD)

Stransky et al[2] describe the identification of 38 significantly mutated genes (false discovery rate q < 0.1) according to MutSig (see Supplementary Table 7 in Stransky et al). Nine of those are also detected by OncodriveFM and/or OncodriveCLUST (qvalue<0.05, see Figure S2). Among the other 29 some may be *bona fide* driver genes (eg. PRDM2), however some clear false positives are also recognized (eg. OR5L2, OR4C15). On the other hand OncodriveFM also detects 8 other genes, among which are MACF1, NFE2L2, NCOR1, NSD1 and PIK3CG. MACF1 (microtubule actin cross-linking factor 1) appears to be involved in the Wnt signaling pathway and functions as a positive regulator in the translocation of Axin and its associated complex from the cytoplasm to the cell membrane, an indispensable step to transduce signaling upon Wnt stimulation[3]. NFE2L2 (also named NRF2) is a transcription factor known to be involved in lung cancer[4,5]. NCOR1 is a protein that mediates repression of thyroid-hormone and retinoic-acid receptors by promoting chromatin condensation and preventing access of the transcription machinery[6], and also takes part in ligand-dependent transcriptional repression by oestrogen receptor alpha. Mutations in this gene are also reported in breast cancer samples in Stephens et al[7]. NSD1 is a nuclear receptor binding with a SET domain: it is frequently translocated with NUP98 in acute myeloid leukemia (AML)[8] and mutations in this gene have been identified as the major cause of a childhood overgrowth syndrome (Sotos sydrome)[9]. Again, among the genes detected by OncodriveFM and OncodriveCLUST we always have to consider that some may be false positives, until their oncogenic role has been clearly established.

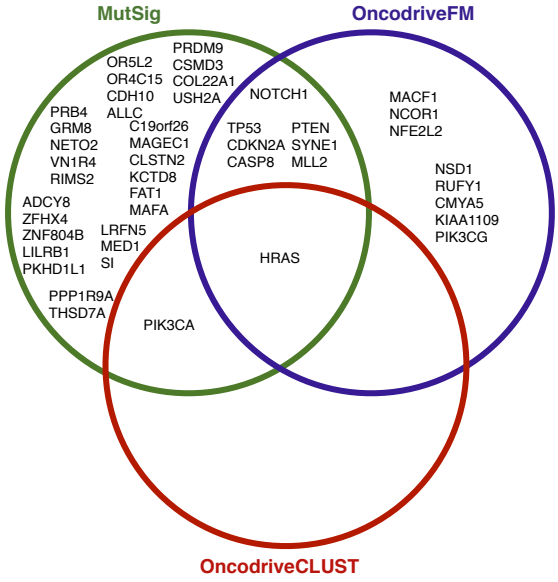


Figure S2. Venn diagram showing the overlap between genes detected by MutSig in Stransky et al and those identified by OncodriveFM and OncodriveCLUST in the IntOGen-mutations pipeline.

Comparison with Ellis et al (BREAST ER WU)

Ellis et al[10] identify 18 significantly mutated genes with a convolution false discovery rate (FDR) < 0.26 (Table 1 and Supplementary Table 6 in Ellis et al). Twelve of those are also detected by OncodriveFM and/or OncodriveCLUST (qvalue < 0.05, see Figure S3). OncodriveFM detects 8 other genes among which there are some well known cancer genes (eg. MLL2, CHD4, ARID1A) and some good candidates, for example RREB1, a transcription factor that binds specifically to the RAS-responsive elements (RRE) of gene promoters.

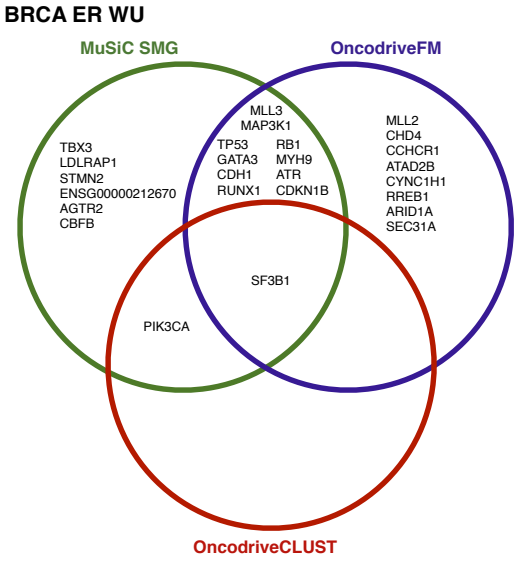


Figure S3. Venn diagram showing the overlap between genes detected by MuSiC (convolution test) and those identified by OncodriveFM and OncodriveCLUST in the IntOGen-mutations pipeline.

Comparison with TCGA LAML paper (AML TCGA)

The manuscript describing the TCGA work on de novo acute myeloid leukemia[11] identifies 23 significantly mutated genes using MuSiC[12] (false discovery rate < 0.05, Table S7 of TCGA manuscript). Most of those are detected by OncodriveFM and/or OncodriveCLUST (Figure S4). OncodriveFM also identifies ASXL1, GATA2 and NF1. ASXL1 is a protein associated to polycomb repressive complex 2 (PRC2)[13] that is recurrently mutated in patients with myelodysplastic syndrome, myeloproliferative neoplasms, and acute myeloid leukemia. GATA2 is a transcription factor involved in myeloid malignancies. Heritable mutations in GATA2 are associated with familial myelodysplastic syndrome and acute myeloid leukemia[14–18].

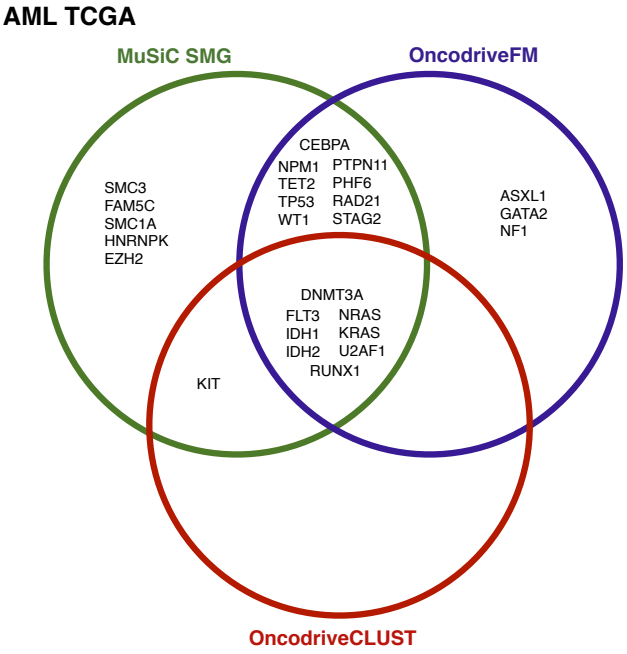


Figure S4. Venn diagram showing the overlap between genes detected by MuSiC and those identified by OncodriveFM and OncodriveCLUST in the IntOGen-mutations pipeline.

Comparison of the lists of candidate driver genes detected by OncodriveFM with known cancer genes

The absence of curated gold-standard datasets of driver and passenger genes limits efforts to validate the results of bioinformatics method to identify novel driver candidates. We have therefore taken an indirect approach to validate the results of our methods included in IntOGen. Briefly, we computed the rate of known cancer genes (from the Cancer Gene Census, or CGC[19]) amongst increasing lists of top candidate genes identified by OncodriveFM (Figure S5). (Because OncodriveCLUST detects only few genes in all sites, the same analysis would not add to the information already shown in Supplementary Table 3.) While high-ranking genes (25 top in blue) identified in all cancer sites are highly enriched for known cancer genes, as lower-ranking genes are incorporated to the analysis this rate steadily decreases. This indicates that top-ranking genes in the lists of candidates are all probably *bona fide* drivers. Since the lists of driver candidates reported in each site (see Supp. Table 3 for the actual size and known cancer genes rate of each list) fall include mostly or exclusively very high-ranking genes, these are enriched for likely cancer genes rather than spurious predictions.

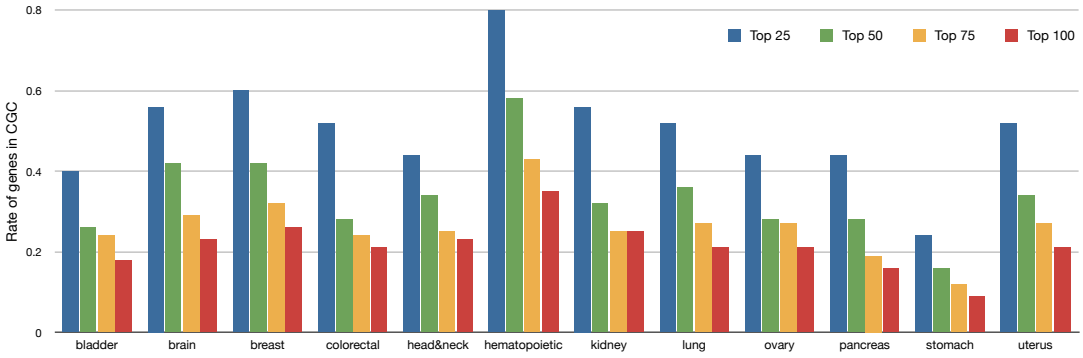


Figure S5. Rate of known cancer genes amongst the 25 top-ranking, the 50 top-ranking, the 75 top-ranking, and the 100 top-ranking OncodriveFM driver candidates in all cancer sites with data in IntOGen, except liver.

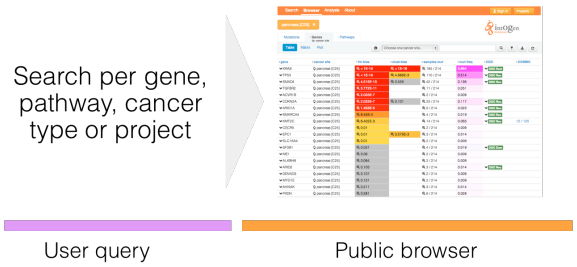
References

1. Banerji, S. *et al*. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
2. Stransky, N. *et al*. The mutational landscape of head and neck squamous cell carcinoma. *Science (New York, N.Y.)* **333**, 1157–60 (2011).
3. Chen, H.-J. *et al*. The role of microtubule actin cross-linking factor 1 (MACF1) in the Wnt signaling pathway. *Genes & development* **20**, 1933–45 (2006).
4. Shibata, T. *et al*. Cancer related mutations in NRF2 impair its recognition by Keap1-Cul3 E3 ligase and promote malignancy. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 13568–73 (2008).
5. Hammerman, P. S. *et al*. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* (2012).doi:10.1038/nature11404
6. Hörlein, A. J. *et al*. Ligand-independent repression by the thyroid hormone receptor mediated by a nuclear receptor co-repressor. *Nature* **377**, 397–404 (1995).
7. Stephens, P. J. *et al*. The landscape of cancer genes and mutational processes in breast cancer. *Nature* (2012).doi:10.1038/nature11017
8. Hollink, I. H. I. M. *et al*. NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. *Blood* **118**, 3645–56 (2011).
9. Douglas, J. *et al*. NSD1 mutations are the major cause of Sotos syndrome and occur in some cases of Weaver syndrome but are rare in other overgrowth phenotypes. *American journal of human genetics* **72**, 132–43 (2003).

10. Ellis, M. J. *et al*. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* (2012).doi:10.1038/nature11143
11. Cancer, T. & Atlas, G. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine* 130501100016000 (2013).doi:10.1056/NEJMoa1301689
12. Dees, N. D. *et al*. MuSiC: Identifying mutational significance in cancer genomes. *Genome Research* (2012).doi:10.1101/gr.134635.111
13. Abdel-Wahab, O. *et al*. ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. *Cancer cell* **22**, 180–93 (2012).
14. Hahn, C. N. *et al*. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nature genetics* **43**, 1012–7 (2011).
15. Pasquet, M. *et al*. High frequency of GATA2 mutations in patients with mild chronic neutropenia evolving to MonoMac syndrome, myelodysplasia, and acute myeloid leukemia. *Blood* **121**, 822–9 (2013).
16. Bödör, C. *et al*. Germ-line GATA2 p.THR354MET mutation in familial myelodysplastic syndrome with acquired monosomy 7 and ASXL1 mutation demonstrating rapid onset and poor survival. *Haematologica* **97**, 890–4 (2012).
17. Ostergaard, P. *et al*. Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nature genetics* **43**, 929–31 (2011).
18. Kazenwadel, J. *et al*. Loss-of-function germline GATA2 mutations in patients with MDS/AML or MonoMAC syndrome and primary lymphedema reveal a key role for GATA2 in the lymphatic vasculature. *Blood* **119**, 1283–91 (2012).
19. Futreal, P. A. *et al*. A census of human cancer genes. *Nature Reviews. Cancer* **4**, 177–183 (2004).

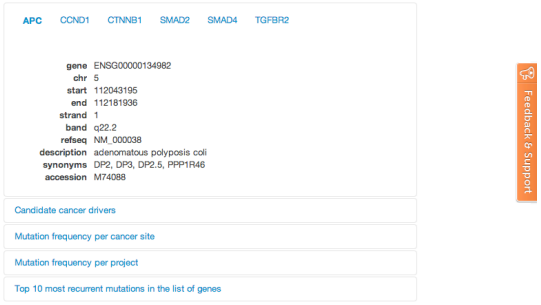
]Supplementary Note 2

Use case 1: Browsing IntOGen-mutations web discovery tool



The systematic analysis of more than 4500 tumors across projects and tumor sites allows researchers to have a wide view of genes and pathways involved in tumorigenesis. Cancer researchers can search IntOGen-mutations to find out which genes are candidate drivers for a given tumor site or the likelihood that a given gene (or geneset) is a driver across different malignancies. This use case entails the employment of the IntOGen-mutations web discovery tool as an aid to researchers in surveying genes or pathways that are related to tumorigenesis when mutated. We exemplify this use case with a survey of selected genes involved in the Hippo pathway. The goal is to find out all the ways this group of genes may be altered by mutations across the 31 projects currently in IntOGen.

One can start by obtaining the list of selected genes of the Hippo pathway from the KEGG HIPPO pathway[1] (hsa04390). They are: SMAD2, SMAD4, CTNNB1, APC, TGFBR2, CCND1. This list must be introduced in the google-like advanced search form of the IntOGen-mutations web discovery tool (http://www.intogen.org/mutations, see image).



In the boxes below you can have an overview of the mutational-pattern of these genes in cancer.

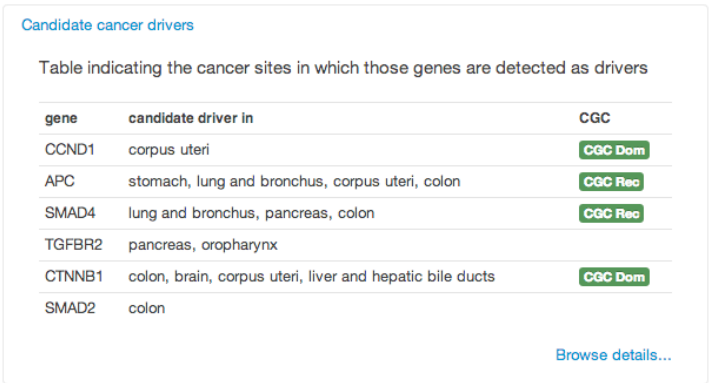

Candidate cancer drivers

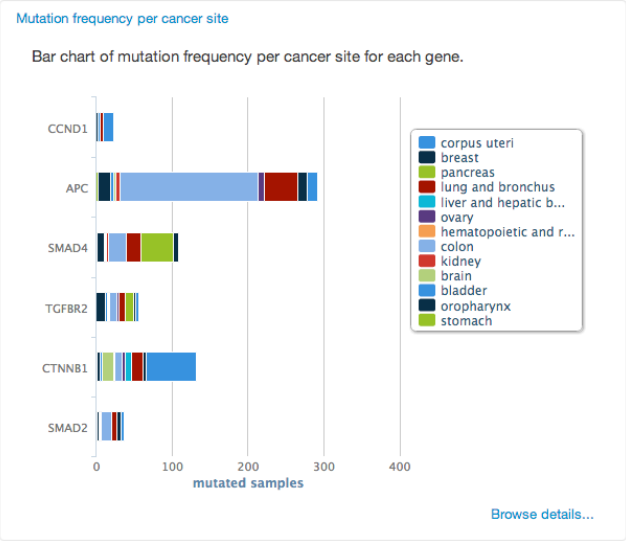
Table indicating the cancer sites in which those genes are detected as drivers

gene	candidate driver in	CGC
CCND1	corpus uteri	CGC Dom
APC	stomach, lung and bronchus, corpus uteri, colon	CGC Rec
SMAD4	lung and bronchus, pancreas, colon	CGC Rec
TGFBR2	pancreas, oropharynx	
CTNNB1	colon, brain, corpus uteri, liver and hepatic bile ducts	CGC Dom
SMAD2	colon	

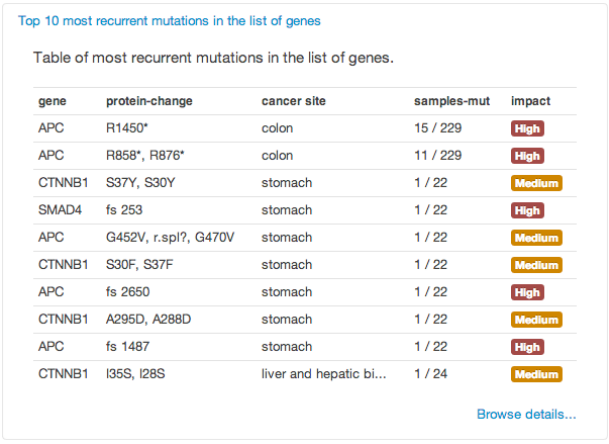
Browse details...

Clicking on “Candidate cancer drivers” you will have an overview of which of these genes are detected as drivers in different cancer sites.

Nature Methods: doi:10.1038/nmeth.2642



Clicking on “Mutation frequency per cancer site” you will see a bar chart showing the frequency of mutations of each gene in each cancer site. Similarly clicking on “Mutation frequency per project” you will see a similar bar chart showing the frequency of mutations of each gene in each project.



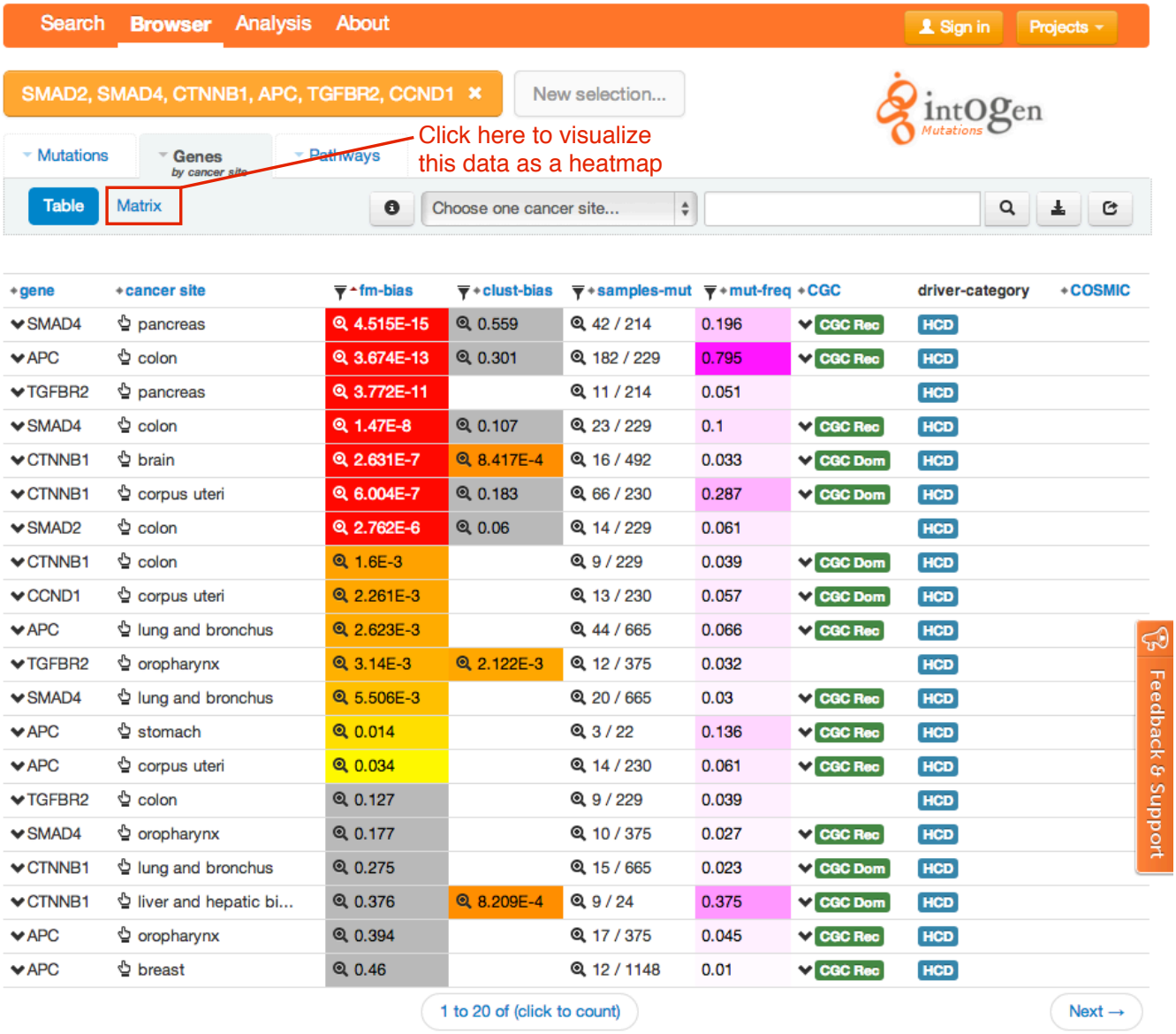
Clicking on “Top 10 most frequent mutations in the list of genes” you will see details of the impact and frequency of these mutations.

Following the links in each of the boxes you are directed to the browser where you can see further details.

For example, by clicking on “Browse details…” link in the box of “Candidate cancer drivers” you will see a table of FM bias and CLUST bias integrated q-values of these genes in the 12 cancer sites with data in IntOGen-mutations.

The filter of the 6 Hippo pathway genes has been applied to the data and appears at the top left corner, as an orange box. This table view presents also the frequency of mutations of each gene across tumors of each site, and links to the Cosmic database[2], if somatic mutations have been previously identified in the gene in cancer samples, or if it is annotated in the Cancer Gene Census[3]. With the mouse over the title of a column a tooltip with details of the data shown in that column is shown.

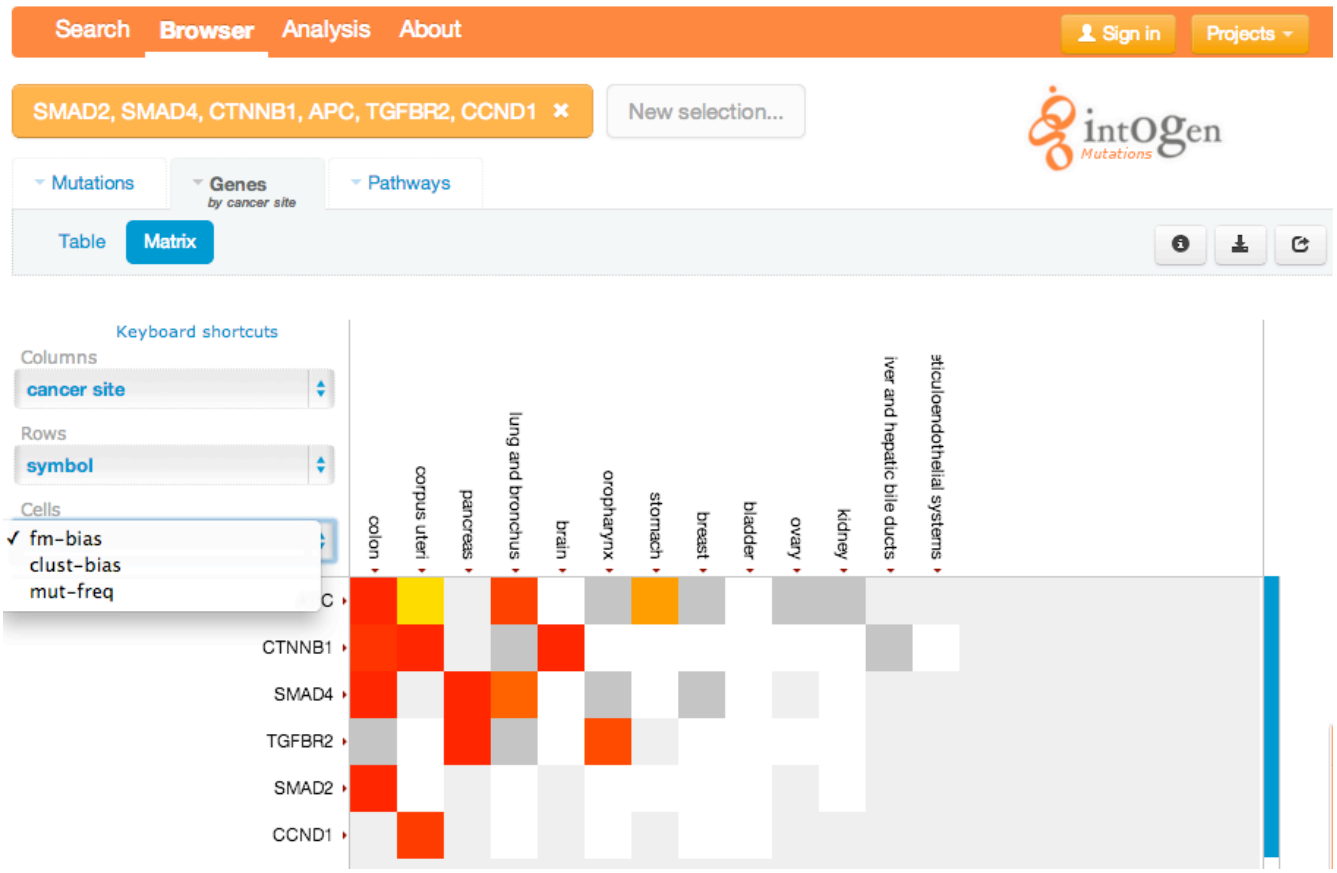
Nature Methods: doi:10.1038/nmeth.2642



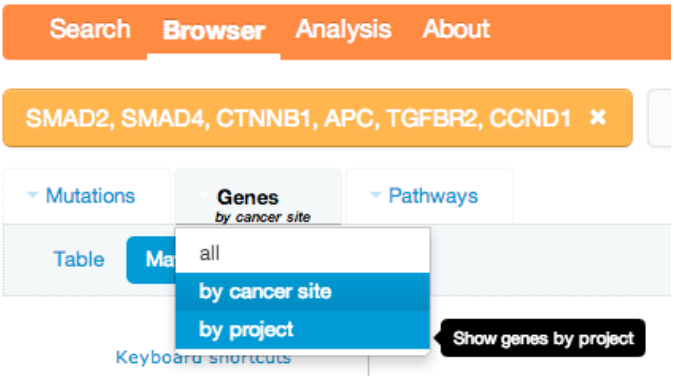
gene	cancer site	fm-bias	clust-bias	samples-mut	mut-freq	CGC	driver-category	COSMIC
SMAD4	pancreas	4.515E-15	0.559	42 / 214	0.196	CGC Rec	HCD	
APC	colon	3.674E-13	0.301	182 / 229	0.795	CGC Rec	HCD	
TGFBR2	pancreas	3.772E-11		11 / 214	0.051		HCD	
SMAD4	colon	1.47E-8	0.107	23 / 229	0.1	CGC Rec	HCD	
CTNNB1	brain	2.631E-7	8.417E-4	16 / 492	0.033	CGC Dom	HCD	
CTNNB1	corpus uteri	6.004E-7	0.183	66 / 230	0.287	CGC Dom	HCD	
SMAD2	colon	2.762E-6	0.06	14 / 229	0.061		HCD	
CTNNB1	colon	1.6E-3		9 / 229	0.039	CGC Dom	HCD	
CCND1	corpus uteri	2.261E-3		13 / 230	0.057	CGC Dom	HCD	
APC	lung and bronchus	2.623E-3		44 / 665	0.066	CGC Rec	HCD	
TGFBR2	oropharynx	3.14E-3	2.122E-3	12 / 375	0.032		HCD	
SMAD4	lung and bronchus	5.506E-3		20 / 665	0.03	CGC Rec	HCD	
APC	stomach	0.014		3 / 22	0.136	CGC Rec	HCD	
APC	corpus uteri	0.034		14 / 230	0.061	CGC Rec	HCD	
TGFBR2	colon	0.127		9 / 229	0.039		HCD	
SMAD4	oropharynx	0.177		10 / 375	0.027	CGC Rec	HCD	
CTNNB1	lung and bronchus	0.275		15 / 665	0.023	CGC Dom	HCD	
CTNNB1	liver and hepatic bi...	0.376	8.209E-4	9 / 24	0.375	CGC Dom	HCD	
APC	oropharynx	0.394		17 / 375	0.045	CGC Rec	HCD	
APC	breast	0.46		12 / 1148	0.01	CGC Rec	HCD	

The same information may be visualized as a heatmap following the 'Matrix' link (see image above). Matrix visualization allows to rapidly grasping that these genes are FM biased mutational drivers across 6 cancer sites, pointing to a likely involvement of this pathway in tumorigenesis in at least tumors from these 6 sites. This heatmap, built on the jheatmap technology (http://bg.upf.edu/jheatmap/) developed in our group is interactive: rows and columns can be moved and sorted at will. The values shown at the heatmap can be changed to OncodriveCLUST pvalues, as can be changed the names of rows and columns, using the options in the left panel. This action completes the picture described above showing that CTNNB1 is also a putative driver (by CLUST bias) in liver.

Nature Methods: doi:10.1038/nmeth.2642

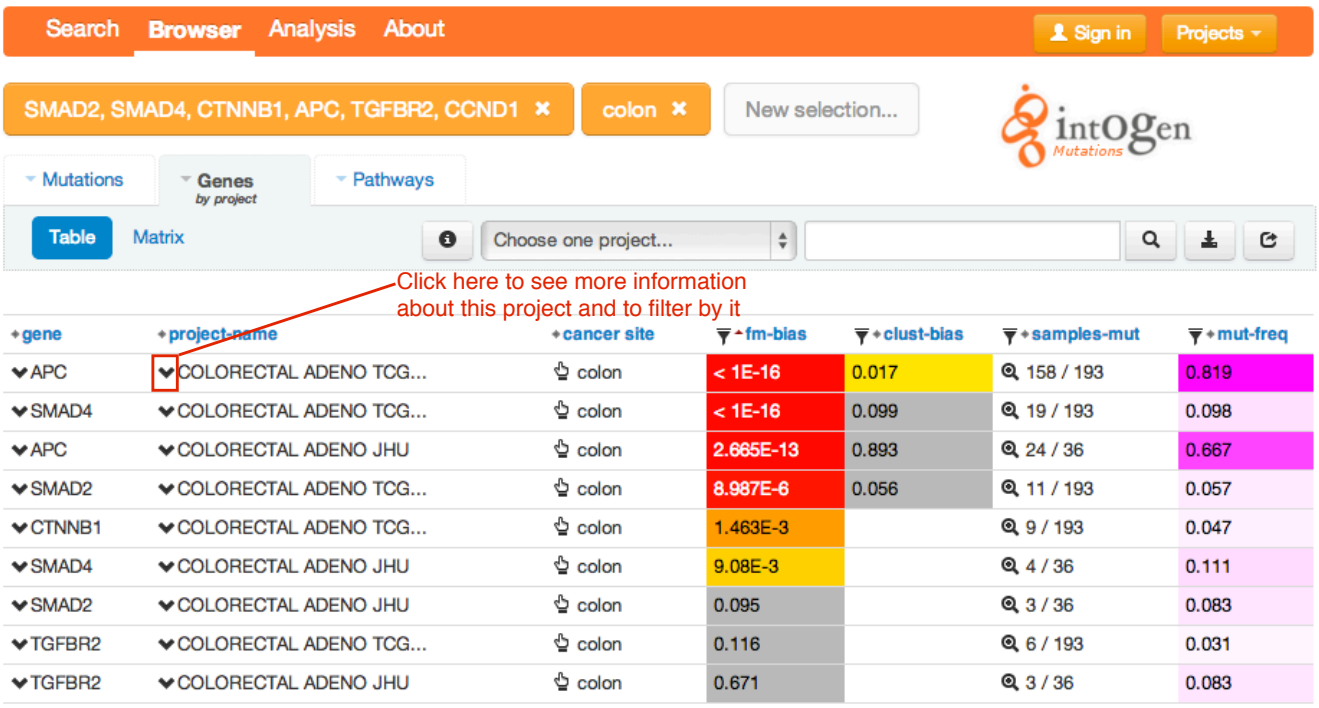


One question to ask from this data is the actual contributions of the projects to the integrated q-values shown in the heatmap. These can be retrieved by pulling down the menu at the Genes tab and selecting the 'by-project' information.



Then, for example by filtering the results for colon only it becomes apparent that whereas APC is significantly FM biased in both colorectal adenocarcinoma projects[4,5], SMAD2, SMAD4 and CTNNB1 are so only in the TCGA colorectal dataset. The combination, nevertheless highlights them as significant in this site because their FM bias pvalues in the Johns Hopkins colorectal adenocarcinoma dataset are sufficiently low.

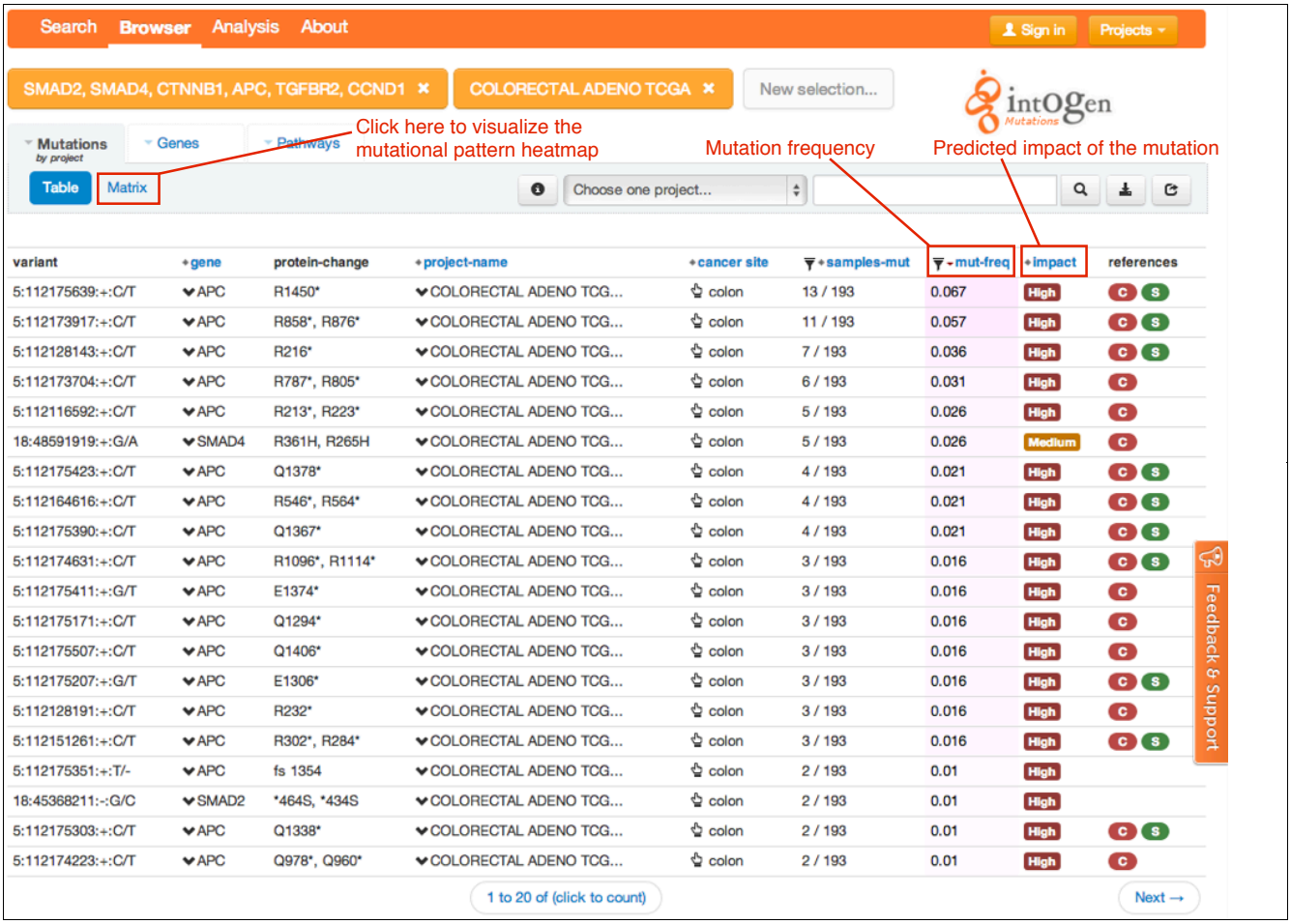
Nature Methods: doi:10.1038/nmeth.2642

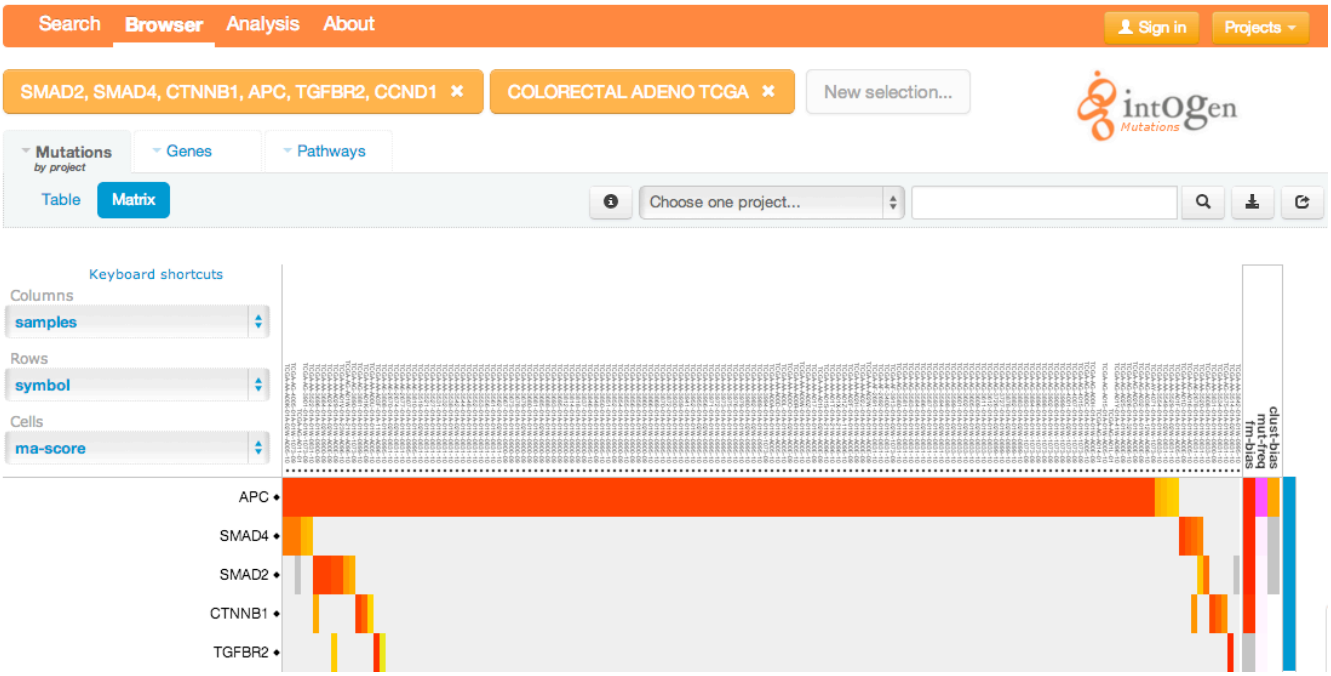


In IntOGen-mutations it is also possible to view details of individual mutations, such as its frequency and its predicted impact. To do so, one can first filter the results of only one dataset, say the TCGA colorectal adenocarcinoma dataset (see above), and then go to the Mutations 'by project' tab. A table with mutations is shown (see first image in next page). By clicking at a column header the table is sorted by the values of this column. For instance, to find the most frequency mutations, click the header of samples-mut. The mutation in the top is the most frequent one in this dataset, observed in 13 samples out of 193 samples analyzed. The predicted impact of the mutation is high. By clicking at the impact box, the user obtains details of the consequence of the mutation and its impact in protein function.

One can also visualize the mutational pattern per gene in a heatmap, which also allows checking whether mutations in these genes that act as part of the same pathway tend to be mutually exclusive across tumor samples of the same project. To do so, click on the ‘Matrix’ view button. Mutations in these genes across all tumor types in the dataset are then visualized as in interactive heatmap (see second image in next page), the color indicating the MutationAssessor scores (colors toward red a high impact). By default, these mutations appear sorted across the genes in a mutually exclusive manner. In this case, the heatmap shows that mutations in SMAD2, SMAD4, CTNNB1 and TGFBR2 tend to appear in the same samples where APC is mutated, underlining the likely character of founders of somatic mutations in APC in colorectal carcinomas.

Nature Methods: doi:10.1038/nmeth.2642





Nature Methods: doi:10.1038/nmeth.2642

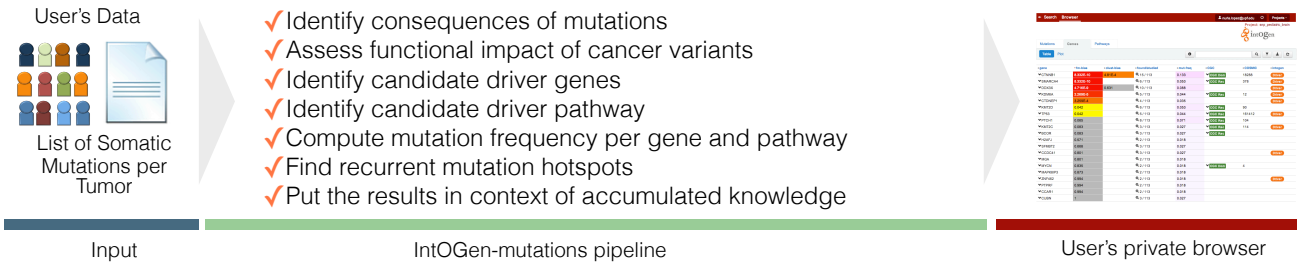
References

1. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
2. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, D945–950 (2010).
3. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Reviews. Cancer* **4**, 177–183 (2004).
4. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
5. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)* **318**, 1108–1113 (2007).

Nature Methods: doi:10.1038/nmeth.2642

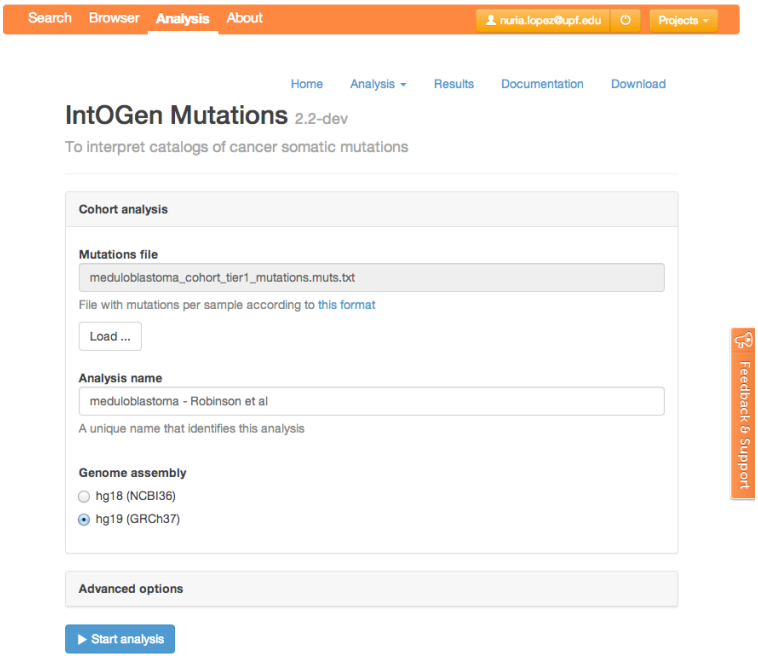
Supplementary Note 3

Use case 2. Analysis of somatic mutations in a cohort of tumors

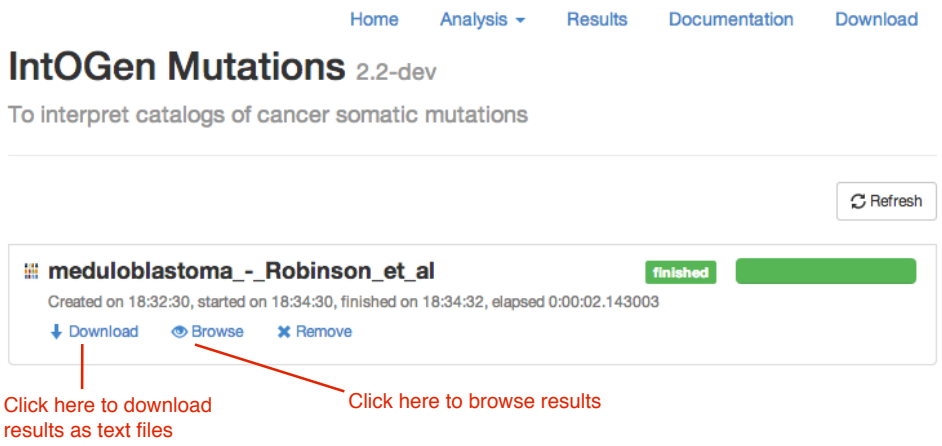


The IntOGen-mutations platform is the first tool that unites a pipeline to analyze the somatic mutations identified across a cohort of tumor samples with a web discovery tool containing accumulated knowledge on the role of somatic mutations in tumors obtained from systematic equivalent analysis of datasets of resequenced tumor genomes. Therefore, one important use of IntOGen-mutations is to identify likely driver genes across a cohort of tumors and compare them with the list of previously detected likely drivers in the same cancer site or in general that is provided by the IntOGen-mutations web discovery tool.

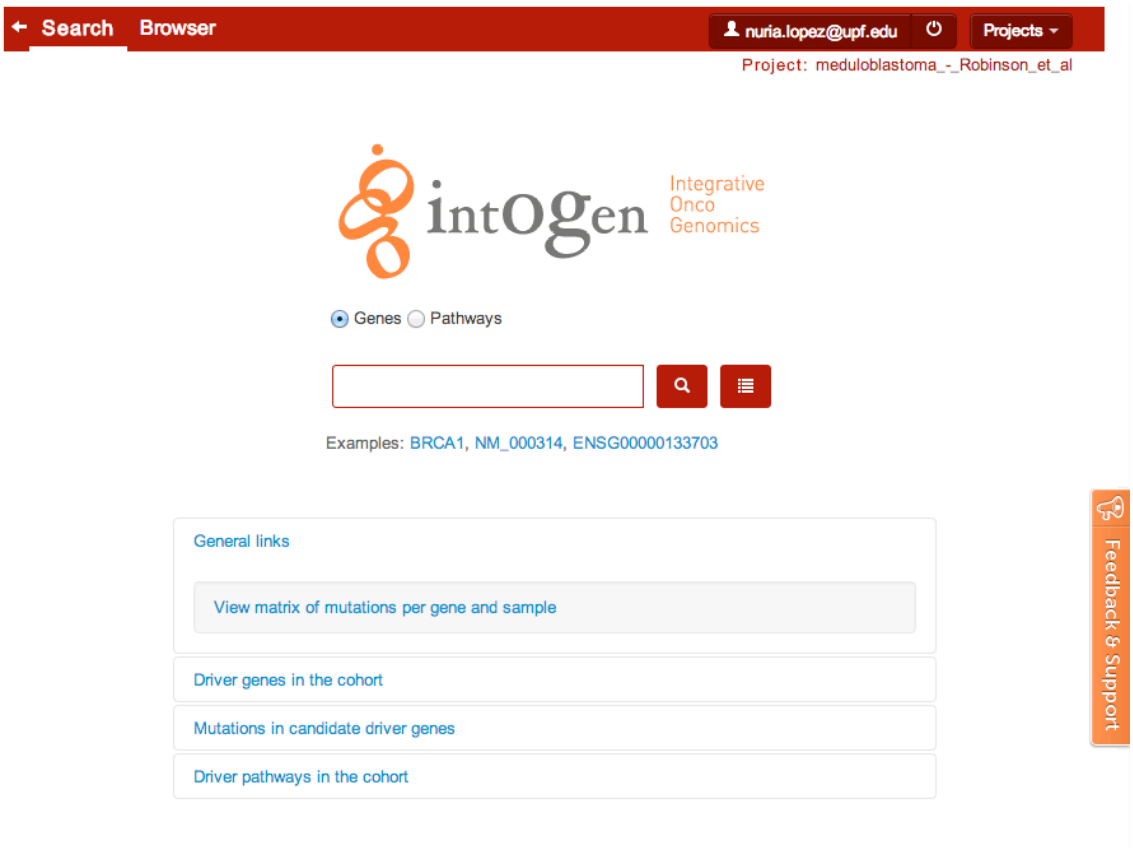
This use case consists in the analysis of the catalog of somatic mutations detected across a dataset of tumor samples aimed at identifying putative driver genes. To illustrate this use case, we employed a dataset of 931 high-confidence somatic mutations identified across 37 medulloblastoma tumor samples (Robinson *et al. Nature* **488**, 43–8 (2012)). The list of mutations was submitted to the online version of the IntOGen-mutations cohort analysis pipeline (http://www.intogen.org/mutations/analysis).

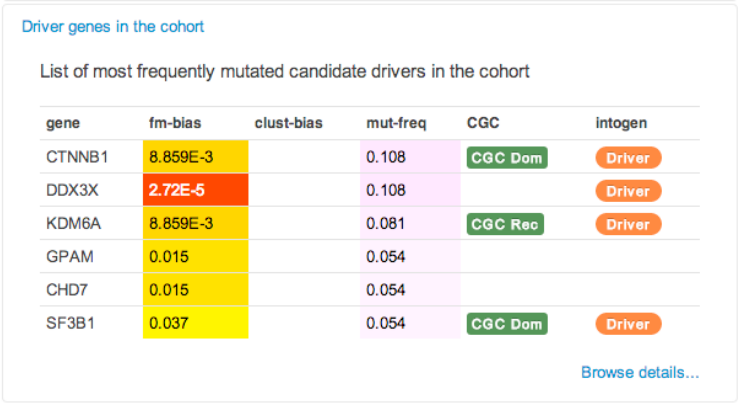


The pipeline then runs on the mutations in the dataset and upon completion, an IntOGen-mutations-like website with the results is generated which can be accessed through the 'Browse' button at the right of the project *md_tier1* progress bar.



The website containing the results of the analysis launched by the user is built on top of the Onexus system (http://www.onexus.org/) and has exactly the same configuration and design as the IntOGen-mutations web discovery tool. The only differences are the colors in the design of the page and the name of the website visible at the top right corner of the pages. A series of three reports added to the main page of the website generated by the pipeline guide the user in the interpretation of the main results of the analysis.





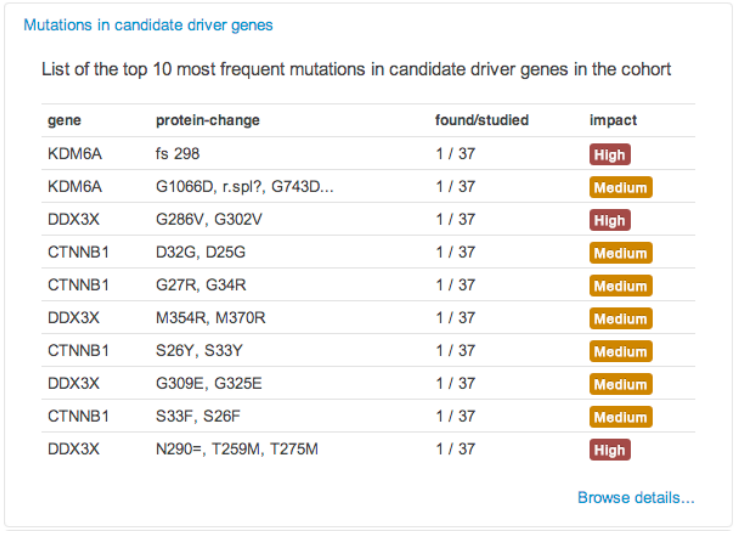
Driver genes in the cohort

List of most frequently mutated candidate drivers in the cohort

gene	fm-bias	clust-bias	mut-freq	CGC	intogen
CTNNB1	8.859E-3		0.108	CGC Dom	Driver
DDX3X	2.72E-5		0.108		Driver
KDM6A	8.859E-3		0.081	CGC Rec	Driver
GPAM	0.015		0.054		
CHD7	0.015		0.054		
SF3B1	0.037		0.054	CGC Dom	Driver

Browse details...

The box “Driver genes in the cohort” presents a list of candidate driver genes in the cohort of tumors under analysis. The report directly highlights the genes that have been previously found to be drivers in IntOGen or which are annotated in the CGC.



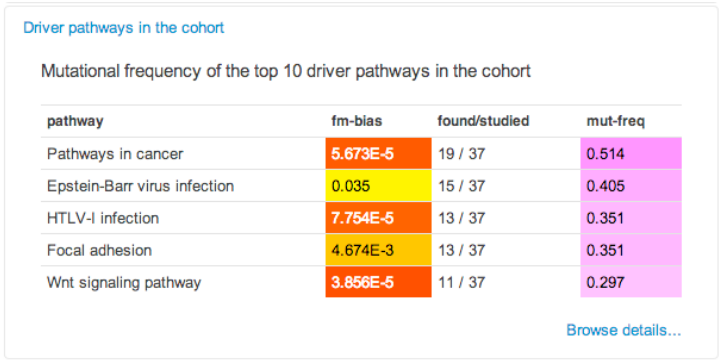
Mutations in candidate driver genes

List of the top 10 most frequent mutations in candidate driver genes in the cohort

gene	protein-change	found/studied	impact
KDM6A	fs 298	1 / 37	High
KDM6A	G1066D, r.spl?, G743D...	1 / 37	Medium
DDX3X	G286V, G302V	1 / 37	High
CTNNB1	D32G, D25G	1 / 37	Medium
CTNNB1	G27R, G34R	1 / 37	Medium
DDX3X	M354R, M370R	1 / 37	Medium
CTNNB1	S26Y, S33Y	1 / 37	Medium
DDX3X	G309E, G325E	1 / 37	Medium
CTNNB1	S33F, S26F	1 / 37	Medium
DDX3X	N290=, T259M, T275M	1 / 37	High

Browse details...

The box “Mutatios in candidate cancer driver genes” summarizes the most frequent mutations in candidate driver genes in the cohort.



Driver pathways in the cohort

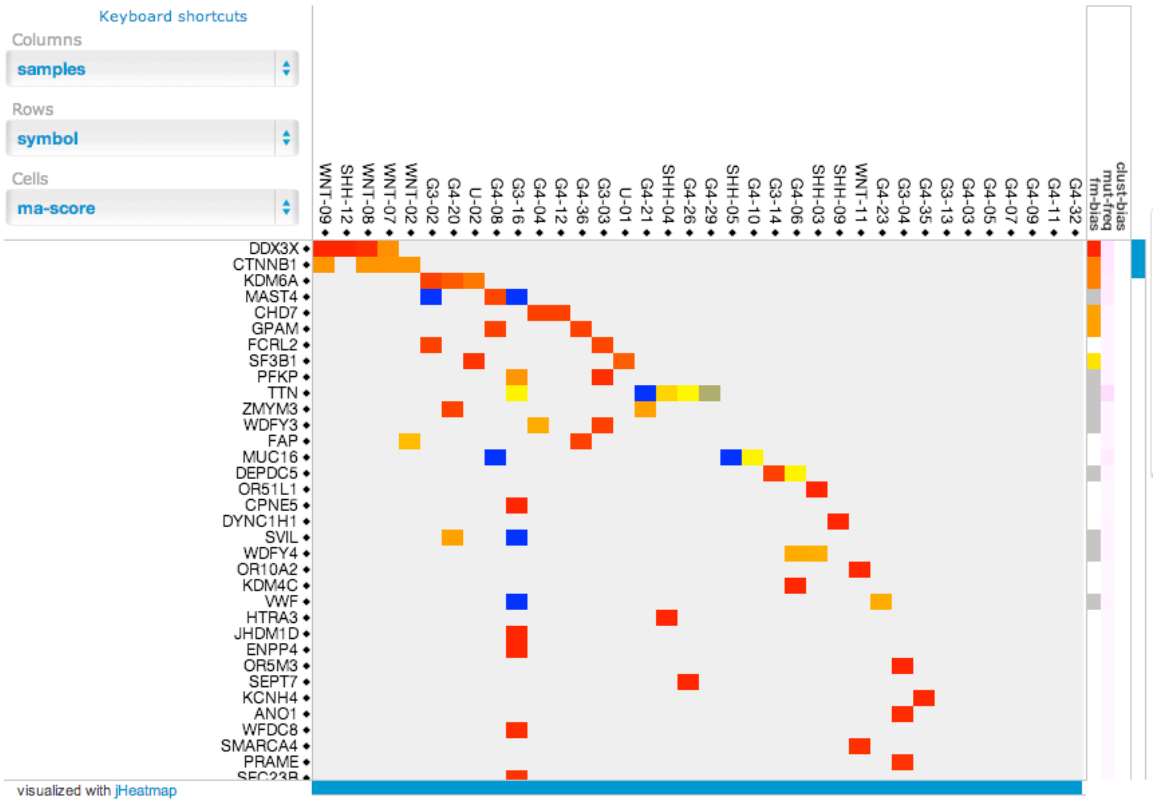
Mutational frequency of the top 10 driver pathways in the cohort

pathway	fm-bias	found/studied	mut-freq
Pathways in cancer	5.673E-5	19 / 37	0.514
Epstein-Barr virus infection	0.035	15 / 37	0.405
HTLV-I infection	7.754E-5	13 / 37	0.351
Focal adhesion	4.674E-3	13 / 37	0.351
Wnt signaling pathway	3.856E-5	11 / 37	0.297

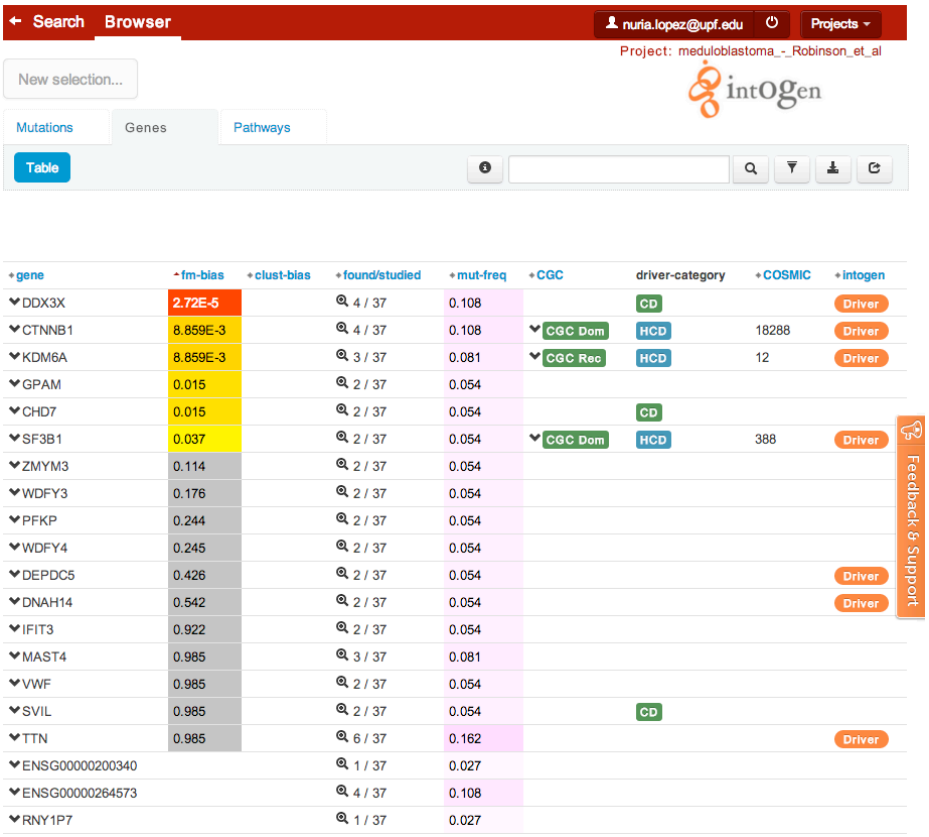
Browse details...

The box “Driver pathways in the cohort” focuses on the top FM biased pathways indicating also the number of samples with somatic mutations in genes of the pathway.

By clicking on “View matrix of mutations per gene and sample” the user is taken directly to the heatmap of all the mutations detected across tumor samples, which was described in the Use Case 1.

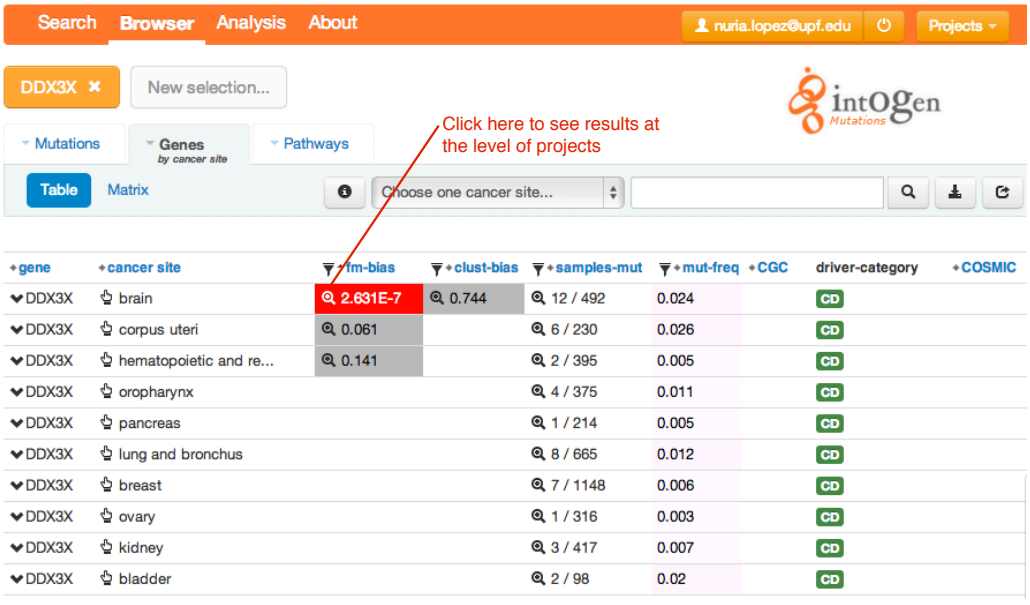


It is then possible to navigate, for instance to the Genes tab, where the user obtains information on the genes that have identified as putative drivers by OncodriveFM and/or OncodriveCLUST in the dataset under analysis. The rightmost column in this visualization marks genes that are putative drivers in at least one cancer site currently included within the IntOGen-mutations web discovery tool. From this visualization user learns that DDX3X, CTNNB1, KDM6A, GPAM, CHD7 and SF3B1 are top ranking FM biased genes in this dataset across the whole range of mutation frequency.



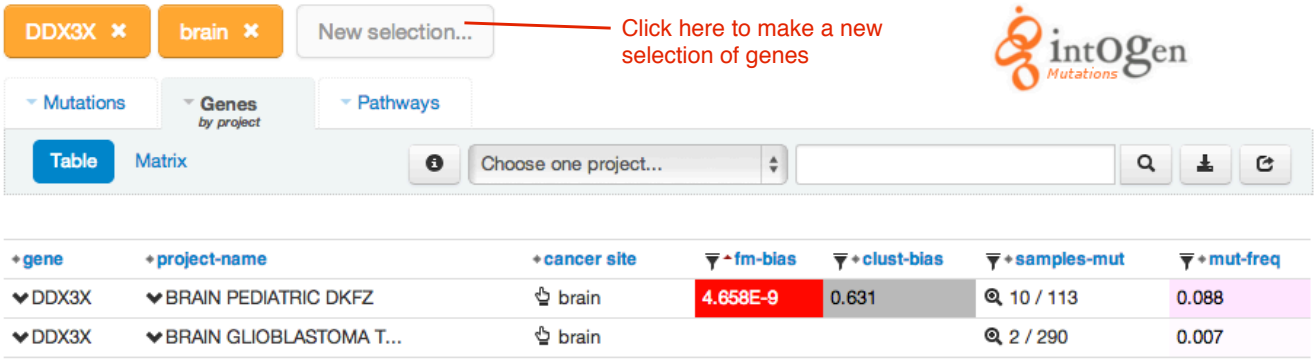
gene	fm-bias	clust-bias	found/studied	mut-freq	CGC	driver-category	COSMIC	intogen
DDX3X	2.72E-5		4 / 37	0.108		CD		Driver
CTNNB1	8.859E-3		4 / 37	0.108	CGC Dom	HCD	18288	Driver
KDM6A	8.859E-3		3 / 37	0.081	CGC Rec	HCD	12	Driver
GPAM	0.015		2 / 37	0.054				
CHD7	0.015		2 / 37	0.054		CD		
SF3B1	0.037		2 / 37	0.054	CGC Dom	HCD	388	Driver
ZMYM3	0.114		2 / 37	0.054				
WDFY3	0.176		2 / 37	0.054				
PFKP	0.244		2 / 37	0.054				
WDFY4	0.245		2 / 37	0.054				
DEPDC5	0.426		2 / 37	0.054				Driver
DNAH14	0.542		2 / 37	0.054				Driver
IFIT3	0.922		2 / 37	0.054				
MAST4	0.985		3 / 37	0.081				
VWF	0.985		2 / 37	0.054				
SVIL	0.985		2 / 37	0.054		CD		
TTN	0.985		6 / 37	0.162				Driver
ENSG00000200340			1 / 37	0.027				
ENSG00000264573			4 / 37	0.108				
RNY1P7			1 / 37	0.027				

The 'Driver' links in the 'intogen' column in the Table visualization connects the user directly with the information on each gene in the IntOGen-mutations web discovery tool. For instance, following the link that corresponds to DDX3X in this dataset, the user learns that it is indeed significantly FM biased in brain tumors, despite its relatively low mutational frequency.



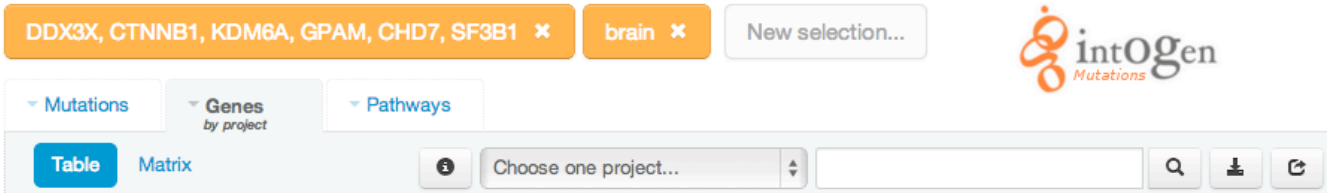
gene	cancer site	fm-bias	clust-bias	samples-mut	mut-freq	CGC	driver-category	COSMIC
DDX3X	brain	2.631E-7	0.744	12 / 492	0.024		CD	
DDX3X	corpus uteri	0.061		6 / 230	0.026		CD	
DDX3X	hematopoietic and re...	0.141		2 / 395	0.005		CD	
DDX3X	oropharynx			4 / 375	0.011		CD	
DDX3X	pancreas			1 / 214	0.005		CD	
DDX3X	lung and bronchus			8 / 665	0.012		CD	
DDX3X	breast			7 / 1148	0.006		CD	
DDX3X	ovary			1 / 316	0.003		CD	
DDX3X	kidney			3 / 417	0.007		CD	
DDX3X	bladder			2 / 98	0.02		CD	

Going deeper to the per-project information reveals that the significance of DDX3X is due to its accumulation of functional mutations in the BRAIN PEDIATRIC DKFZ dataset, which comprises pediatric medulloblastoma cases.



gene	project-name	cancer site	fm-bias	clust-bias	samples-mut	mut-freq
DDX3X	BRAIN PEDIATRIC DKFZ	brain	4.658E-9	0.631	10 / 113	0.088
DDX3X	BRAIN GLIOBLASTOMA T...	brain			2 / 290	0.007

If the candidate driver genes detected in this cohort (DDX3X, CTNNB1, KDM6A, GPAM, CHD7, SF3B1) are introduced as a new selection of genes in IntOGen-mutations web discovery tool we can see that three of these genes (DDX3X, CTNNB1 and KDM6A) pop up as likely candidate drivers of pediatric medulloblastomas.

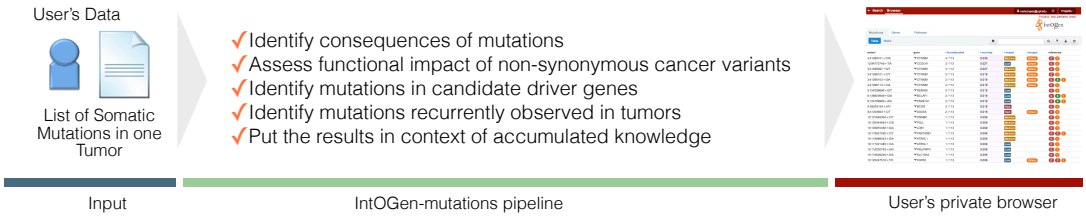


gene	project-name	cancer site	fm-bias	clust-bias	samples-mut	mut-freq
CTNNB1	BRAIN PEDIATRIC DKFZ	brain	2.4E-13	3.607E-4	15 / 113	0.133
DDX3X	BRAIN PEDIATRIC DKFZ	brain	4.658E-9	0.631	10 / 113	0.088
KDM6A	BRAIN PEDIATRIC DKFZ	brain	5.105E-5		5 / 113	0.044
CHD7	BRAIN GLIOBLASTOMA T...	brain			3 / 290	0.01
KDM6A	BRAIN GLIOBLASTOMA T...	brain			3 / 290	0.01
DDX3X	BRAIN GLIOBLASTOMA T...	brain			2 / 290	0.007
SF3B1	BRAIN GLIOBLASTOMA T...	brain			2 / 290	0.007
CTNNB1	BRAIN GLIOBLASTOMA T...	brain			1 / 290	0.003

Nature Methods: doi:10.1038/nmeth.2642

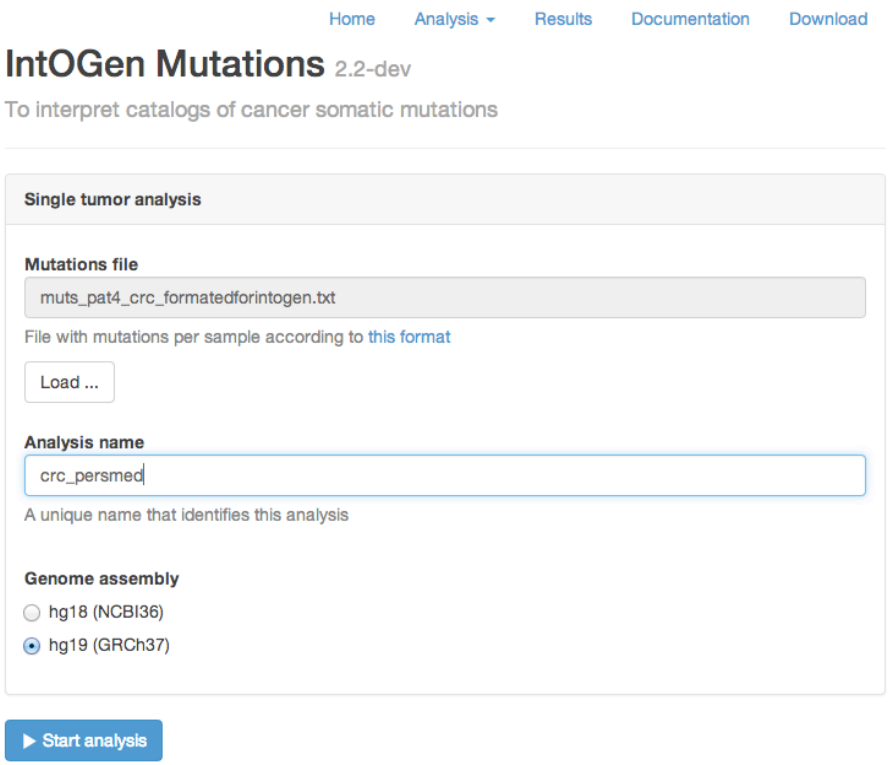
Supplementary Note 4

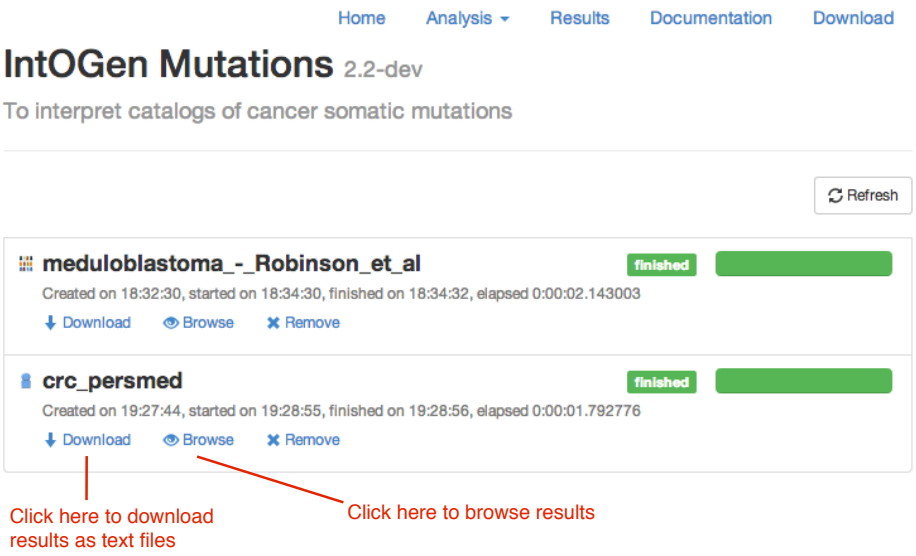
Use case 3. Analysis of somatic mutations in a tumor of an individual patient



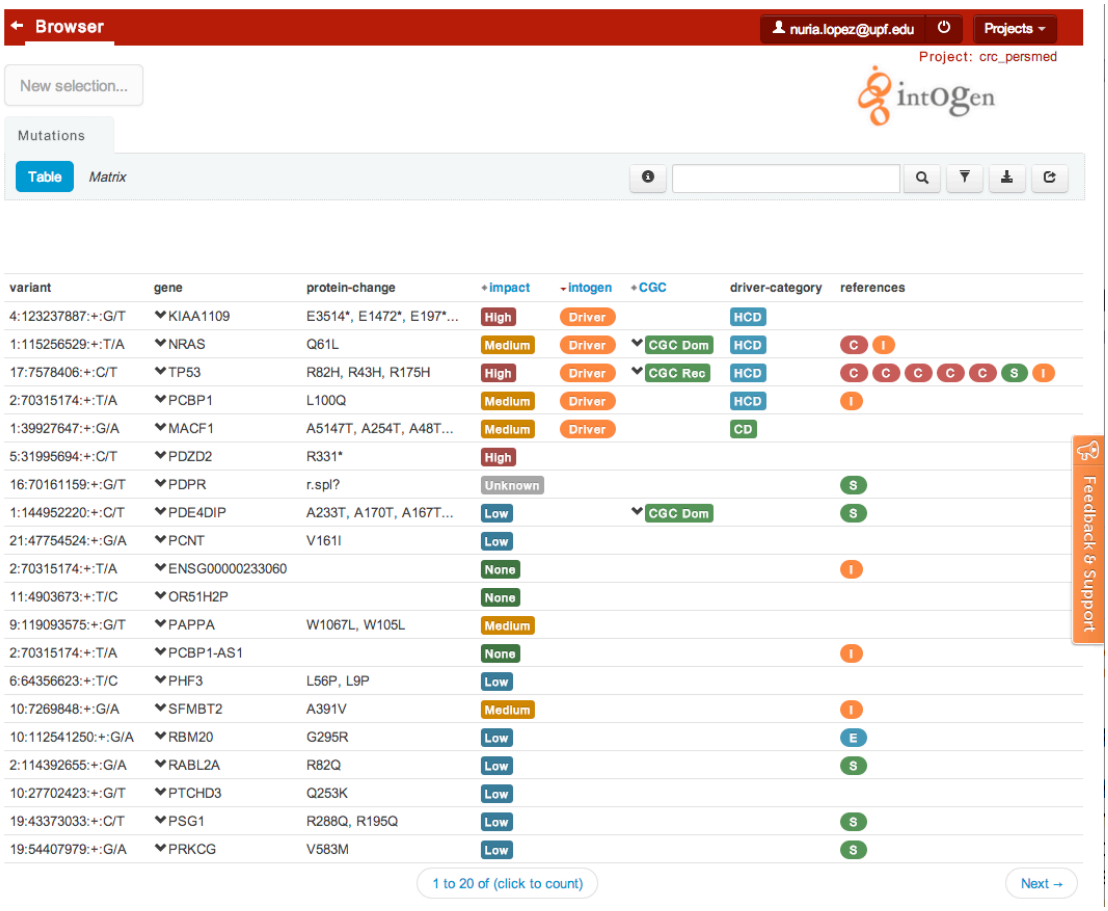
The IntOGen-mutations pipeline can be used to rank the somatic mutations identified in the tumor of an individual patient. Researchers with a list of mutations detected in a tumor can identify functionally impacting mutations, find mutations affecting cancer driver genes, and identify any mutations in the patient that have been previously observed in tumors. All this information can help to suggest which genes might have driven tumorigenesis in the patient, with the final aim of informing a personalized approach to treatment.

We exemplify this use case with the list of somatic mutations detected in the exome of a metastatic colon cancer[1]. The user starts by submitting these mutations to the online version of the IntOGen-mutations single tumor analysis pipeline at http://www.intogen.org/mutations/analysis and launching it. (For exemplary purposes, let's name this project *crc_persmed.*)





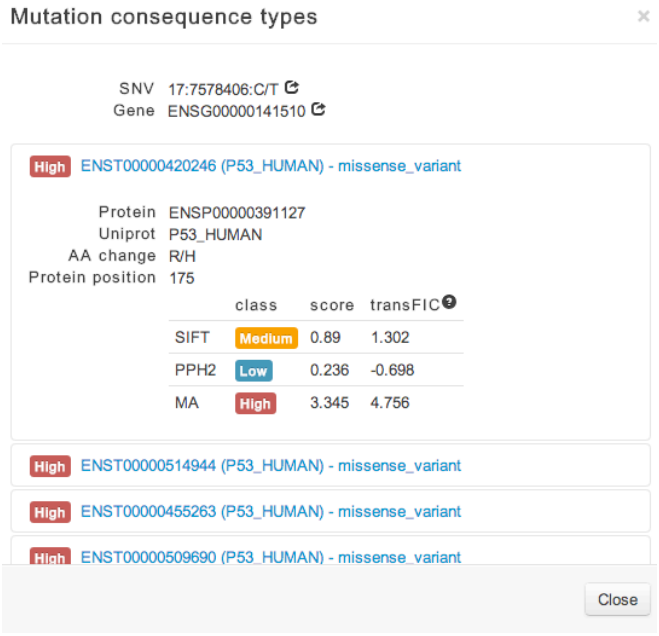
When the analysis is completed the user can download the results in the form of text files or browse the results. Upon entrance in the browser, the user is directed to the only available tab in this case, Mutations.



variant	gene	protein-change	impact	intogen	CGC	driver-category	references
4:123237887:+:G/T	KIAA1109	E3514*, E1472*, E197*...	High	Driver		HCD	
1:115256529:+:T/A	NRAS	Q61L	Medium	Driver	CGC Dom	HCD	C I
17:7578406:+:C/T	TP53	R82H, R43H, R175H	High	Driver	CGC Rec	HCD	C C C C C S I
2:70315174:+:T/A	PCBP1	L100Q	Medium	Driver		HCD	I
1:39927647:+:G/A	MACF1	A5147T, A254T, A48T...	Medium	Driver		CD	
5:31995694:+:C/T	PDZD2	R331*	High				
16:70161159:+:G/T	PDPR	r.spl?	Unknown				S
1:144952220:+:C/T	PDE4DIP	A233T, A170T, A167T...	Low		CGC Dom		S
21:47754524:+:G/A	PCNT	V161I	Low				
2:70315174:+:T/A	ENSG00000233060		None				I
11:4903673:+:T/C	OR51H2P		None				
9:119093575:+:G/T	PAPPA	W1067L, W105L	Medium				
2:70315174:+:T/A	PCBP1-AS1		None				I
6:64356623:+:T/C	PHF3	L56P, L9P	Low				
10:7269848:+:G/A	SFMBT2	A391V	Medium				I
10:112541250:+:G/A	RBM20	G295R	Low				E
2:114392655:+:G/A	RABL2A	R82Q	Low				S
10:27702423:+:G/T	PTCHD3	Q253K	Low				
19:43373033:+:C/T	PSG1	R288Q, R195Q	Low				S
19:54407979:+:G/A	PRKCG	V583M	Low				S

This table gives information on the effect of each mutation over the encoded protein, information about the evidences of the affected gene to be involved in cancer, and reports if the mutations have already been reported in IntOGen or other databases (references columns).

The table is sorted to show on top those that occur in IntOGen-mutations drivers. To gain further insight on the possible involvement of these mutations in driver genes in the process tumorigenesis, the user may follow the links in the Impact column. These classify the likely impact of mutations on protein function into four categories ranging from None to High, based on their consequence types and the transFIC[2] MutationAssessor[3] scores of non-synonymous mutations. This way, the user learns that the mutation detected on TP53 causes an aminoacid change in several of its transcripts with a putative high effect on the protein product.



The actionable mutation detected by the original report falls on the NRAS known driver possibly with an important effect on the activity of the protein. Following the 'Driver' link in the IntOGen column that corresponds to this mutation, the user learns that in IntOGen NRAS has been deemed significantly FM biased and CLUST biased in colorectal adenocarcinoma, as well as in other sites.