Genome **Biology**

The mutational landscape of chromatin regulatory factors across 4,623 tumor samples

Abel Gonzalez-Perez[1†], Alba Jene-Sanz[1†] and Nuria Lopez-Bigas[1,2*]

Abstract

Background: Chromatin regulatory factors are emerging as important genes in cancer development and are regarded as interesting candidates for novel targets for cancer treatment. However, we lack a comprehensive understanding of the role of this group of genes in different cancer types.

Results: We have analyzed 4,623 tumor samples from thirteen anatomical sites to determine which chromatin regulatory factors are candidate drivers in these different sites. We identify 34 chromatin regulatory factors that are likely drivers in tumors from at least one site, all with relatively low mutational frequency. We also analyze the relative importance of mutations in this group of genes for the development of tumorigenesis in each site, and in different tumor types from the same site.

Conclusions: We find that, although tumors from all thirteen sites show mutations in likely driver chromatin regulatory factors, these are more prevalent in tumors arising from certain tissues. With the exception of hematopoietic, liver and kidney tumors, as a median, the mutated factors are less than one fifth of all mutated drivers across all sites analyzed. We also show that mutations in two of these genes, *MLL* and *EP300*, correlate with broad expression changes across cancer cell lines, thus presenting at least one mechanism through which these mutations could contribute to tumorigenesis in cells of the corresponding tissues.

Background

Highly conserved molecular mechanisms are responsible for maintaining genome integrity and tightly regulated gene expression, which is essential for cell survival. Those include the fine regulation of chromatin structure, mainly maintained through three distinct processes: the post-translational modification of histone tails, the replacement of core histones by histone variants, and the direct structural remodeling by ATP-dependent chromatin-remodeling enzymes [1]. The proteins that control this system, broadly referred to as chromatin regulatory factors (CRFs), contribute to the establishment of chromatin structures that modulate the expression of large gene sets, either by establishing more inaccessible regions or by placing histone marks that open the chromatin and allow the binding of other factors. These CRFs help to maintain cellular identity, and mutations in them

(commonly called epimutations) often lead to a deregulation of gene expression that may contribute to tumorigenesis [2]. CRFs are broadly classified in three main groups: histone tail modifiers (including histone acetyltransferases, histone deacetylases (HDACs), histone methyltransferases and histone demethylases, that deposit or remove acetyl or methyl groups, respectively); DNA methyltransferases (DNMTs) and putative demethylases (that affect cytosines at CpG islands); and ATP-dependent chromatin remodeling complexes (responsible for the repositioning of nucleosomes).

Until recently, DNMT proteins had not been found mutated in cancer [3], but DNMT3A, and later DNMT1 and DNMT3B, were reported as altered in patients with myelodysplastic syndromes and in acute monocytic leukemia, where their mutation status also predicted prognosis [4,5]. Mutations in ATP-dependent chromatin-remodeling complexes are recurrent in, amongst others, ovarian and clear cell renal cancers [2]. The regulation of the trimethylation of histone H3 at K27 mark (H3K27me3) by the Polycomb complex, a key component to maintain stem cell identity, is also

* Correspondence: nuria.lopez@upf.edu
†Equal contributors
[1]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Dr. Aiguader 88, Barcelona, Spain
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

BioMed Central

frequently compromised in a variety of cancer types, including those from breast, bladder, pancreas, prostate and lymphomas [6]. Histone demethylases have also been implicated in the development of a wide variety of tumors. Moreover, recent whole exome sequencing studies in large cohorts of tumor samples have highlighted as main findings the inactivating mutations on proteins that regulate the epigenomic state of cells [7]. Alterations in KAT6B [8], SMARCC1 [9] and NSD1 [10] have been described in uterine, cervical and skin pre-malignant lesions, respectively. This presents these proteins as potential biomarkers, thus adding early cancer detection to the possible uses of CRFs in the clinic.

This current accumulation of evidence for the role of CRFs in cancer has attracted the attention of the scientific community towards CRFs as novel targets for cancer treatment. In 2006, the first HDAC inhibitor (HDACi), Vorinostat, was approved by the US Food and Drugs Administration to treat a specific type of lymphoma, and more than 20 molecules of this type are currently under preclinical and clinical investigation [11]. Some DNMT inhibitors have been recently approved by the US Food and Drugs Administration to treat myelodysplastic syndromes, and their combination with HDACi is a subject of intense study in clinical trials [12]. Some studies raise hopes for the possible use of HDACis to overcome drug resistance [13]. Interestingly, an in-depth review by Patel *et al.* on 46 potentially druggable yet chemically unexplored proteins in the Cancer Gene Census (CGC) identified six CRFs: ATRX, KAT6A, KDM6A, NSD3, PBRM1 and SMARCA4 [14].

Even though CRFs are emerging as important players in cancer development [15-20], to our knowledge no systematic analysis on the alterations of a comprehensive catalog of CRFs in different tumors has been performed to date. Moreover, most studies have focused their efforts in the in-depth characterization of specific genes that appear mutated at high frequencies, underestimating the impact of lowly recurrent drivers (those genes whose mutation is likely to be functional, but occurs in few samples) on tumorigenesis. For instance, a very recent report [21] focused only on the SWI/SNF family took into account the frequency of mutations of their members rather than their likelihood of driving tumorigenesis.

In this paper, we carry out a systematic exploration of the role of CRFs in tumorigenesis in different tissues. To that end, we first compiled and manually curated a comprehensive list of CRFs, for which we annotated any previously known implications in cancer. Secondly, we analyzed 4,623 tumor samples from 13 anatomical sites to identify which of the CRFs are driver candidates in these different sites, employing two approaches recently introduced by us [22,23]. Finally, we took advantage of the profiles of genomic and transcriptomic alterations revealed by the Cancer Cell Line Encyclopedia (CCLE) [24] to explore the effects of mutations in two likely driver CRFs on the expression of broad gene modules across 905 cancer cell lines.

Results
Analysis of chromatin regulatory factor tumor somatic mutations identifies 34 likely drivers in 13 cancer sites

To determine which CRFs may be involved in cancer emergence and development in primary tumors from 13 anatomical sites upon mutation, we first collected and manually curated a list of CRFs from the literature. This catalog contained 183 proteins grouped into eleven major functional classes, the most populated of which were the HDACs, the histone acetyltransferases and the histone methyltransferases. (The detailed list of CRFs in all functional classes is presented in Additional file 1: Table S1). Only 26 of them are included in the CGC. However, we found that many of these CRFs (115 out of 183) have some evidence, mainly in scattered reports from the past two years, of genomic or transcriptomic alterations in human tumors (Table 1 and Additional file 1: Table S2).

In IntOGen-mutations [83], during the past year, we have collected and analyzed datasets of cancer somatic mutations produced by several research groups across the world. Some of them have been generated within the framework of large international initiatives like The Cancer Genome Atlas (TCGA) [84] and the International Cancer Genomes Consortium [85], while others are the fruit of independent laboratories. Taken together, these datasets [86] contain somatic mutations detected in 4,623 primary tumor samples obtained from 13 anatomical sites (Table 2). Each dataset has been analyzed separately, to compensate for differences between tumor histologies and subtypes, and between sequencing analysis pipelines. First, we used an approach recently developed by us, OncodriveFM [22], to detect genes that, across the cohort of tumor samples, tend to accumulate functional mutations. We give the name 'FM bias' to this significant trend towards the accumulation of functional mutations. The FM bias is a signal of positive selection during cancer development and therefore FM-biased genes are likely candidates to drivers. Second, we identified genes whose mutations tend to significantly cluster in certain regions of their protein sequence (CLUST bias) also via an approach recently developed in our group, OncodriveCLUST [23]. Both FM-biased and CLUST-biased genes constitute sound candidates to cancer drivers [87] in these 13 anatomical sites. We have also combined the *P* values of FM bias and CLUST bias of individual genes across the datasets of tumor samples obtained from the same anatomical site. In summary, we have obtained a measurement of FM bias

Table 1 Described oncogenic alterations in chromatin regulatory factors that are candidate drivers in at least one tissue

Gene	Literature evidence
ARID1A	Mutated in cc ovarian carcinoma and RCC (CGC), bladder [25], HCC [26], endometrium [27], colorectal [28], gastric adenocarcinoma [29], pancreatic cancer [30], lung adenocarcinoma [31], Burkitt lymphoma [32] and aggressive neuroblastoma [33].
	Down-regulated in aggressive breast cancer [34],
KMT2C	Mutated in medulloblastoma (CGC), HCC [26], bladder [25], prostate cancer [35], colorectal cancer [36], gastric adenocarcinoma [29], NSCLC [37], breast cancer [38] and pancreatic cancer [30].
	Deleted in leukemia [39].
DNMT3A	Mutated in AML (CGC), ALL and lung cancer [40].
	Over-expressed in ovarian aggressive tumors [41].
KDM6A	Mutated in kidney, esophageal squamous cell carcinoma, multiple myeloma (CGC), lung cancer [42], medulloblastoma [43], ccRCC [44], bladder [25] and prostate [35].
	Over-expressed in breast tumors with poor prognosis [45].
	Deleted in lung cancer [46].
PBRM1	Mutated in ccRCC, breast (CGC) and pancreatic cancer [47].
NSD1	Mutated in AML (CGC) and NMSC [10].
	Gained in lung adenocarcinoma of never-smokers [48].
TET2	Mutated in MDS (CGC), CMML and AML [49].
SETD2	Mutated in ccRCC (CGC).
	Down-regulated in breast tumors [50].
SMARCA4	Mutated in NSCLC (CGC), lung adenocarcinoma [31], medulloblastoma [43] and Burkitt lymphoma [32].
	Over-expressed in glioma [51] and in melanoma progression [52].
	Gained in lung [42].
KMT2D	Mutated in medulloblastoma, bladder [25], renal cancer (CGC), DLBCL [53].
	Over-expressed in breast and colon tumors [54].
CHD4	Mutated in high MSI gastric and colorectal cancers [55].
	Down-regulated in gastric and colorectal cancers [55].
NCOR1	Mutated in breast [56] and bladder cancer [25].
	Down-regulated in aggressive breast tumors [57].
EP300	Mutated in colorectal, breast and pancreatic cancers, ALL, AML, DLBCL (CGC), bladder [25], SCLC [58] and endometrium [27].
	Up-regulated in esophageal squamous cell carcinoma [59] and advanced HCC [60].
	Loss of heterozygosity in glioblastoma [61].
KDM5C	Mutated in ccRCC (CGC).
ARID2	Mutated in hepatocellular carcinoma (CGC), melanoma [62], NSCLC [63] and pancreatic cancer [30].
	Deleted in NSCLC [63].
ATF7IP	-
ASXL1	Mutated in MDS and CMML (CGC), myeloproliferative neoplasm; [64], AML with myelodysplasia-related changes [65] and castration-resistant prostate cancer [66].
MLL	Mutated in AML, ALL (CGC), bladder [25], SCLC [58], HCC [26] and gastric tumors [29].

Table 1 Described oncogenic alterations in chromatin regulatory factors that are candidate drivers in at least one tissue *(Continued)*

BAZ2A	Over-expressed in CLL [67].
CHD3	Mutated in high MSI gastric and colorectal cancers [55].
ATRX	Mutated in pediatric glioblastoma, neuroendocrine pancreatic tumors (CGC) and high grade adult gliomas [68].
ARID1B	Mutated in breast tumors [56].
MBD1	Over-expressed in pancreatic cancer [69].
BAP1	Mutated in uveal melanoma, breast, NSCLC and RCC (CGC).
	Over-expressed in NSCLC with good prognosis [70].
INO80	-
CHD2	Mutated in high MSI gastric and colorectal cancers [55] and CLL [71].
	Down-regulated in relapsed colon cancer [72].
ARID4A	-
DOT1L	-
ASH1L	Mutated in lung cancer cell lines [42].
	Gained in hepatocellular carcinoma [73].
BPTF	Gained in neuroblastoma and lung cancer [74].
RTF1	-
PHC3	Mutated and lost in osteosarcoma [75].
SMARCA2	Mutated in NMSC [76] and CLL [77].
	Down-regulated in lung adenocarcinoma [78] and gastric cancer [79].
	Amplified in AML [80].
SETDB1	Recurrently amplified and over-expressed in melanoma [81].

This is an exhaustive compilation of alterations ([a]) reported in CRFs showing FM bias and CLUST bias in at least one tissue (Figure 1). Gene names correspond to HUGO Gene Nomenclature Committee-approved symbols. In bold typeface, genes included in the CGC [82]. ALL, acute lymphocytic leukemia; AML, acute myeloid leukemia; cc, clear cell; CGC, Cancer Gene Census; CLL, chronic lymphocytic leukemia; RCC, renal cell carcinoma; CMML, chronic myelomonocytic leukemia; CRPC, castration-resistant prostate cancer; ESCC, esophageal squamous cell carcinoma; HCC, hepatocellular carcinoma; HL, Hodgkin lymphoma; MCL, mantle cell lymphoma; MDS, myelodysplastic syndrome; MSI, microsatellite instability; MPN, myeloproliferative neoplasm; NMSC, non-melanoma skin cancer; NSCLC, non-small cell lung carcinoma; RCC, renal cell carcinoma.
[a]Evidence based solely on cancer cell lines is excluded from this table. Only evidence in human samples have been used. Effects of pharmacological inhibition are not included. Germline polymorphisms are also excluded.

and CLUST bias for each mutated gene at the level of one dataset of tumor samples (or project), and also at the level of each anatomical site (or tissue). This catalog of likely driver genes has allowed us, for the first time, to systematically explore the involvement of epigenetic mechanisms (via mutations in CRFs) in tumorigenesis in 4,623 tumor samples from 13 anatomical sites.

After an exhaustive search within the list of likely driver genes, we found that 34 CRFs from our manually curated list are FM biased and/or CLUST biased in at least one site (Figure 1, upper panel). Sixteen of them appear as likely drivers in more than one project, and only liver carcinomas appear free of likely driver CRFs -

Table 2 Description of the datasets of tumor somatic mutations collected and analyzed to detect candidate cancer driver genes

Site	Dataset name	Description	Authors	Obtained from	Tumor samples	References
Bladder	BLADDER UROTHELIAL TCGA	Bladder urothelial carcinoma	TCGA	Synapse	98	-
Brain	BRAIN GLIOBASTOMA TCGA	Glioblastoma multiforme	TCGA	Synapse	290	[84]
	BRAIN GLIOBASTOMA JHU	Glioblastoma multiforme	John Hopkins University	ICGC DCC	88	[88]
	BRAIN PEDIATRIC DKFZ	Pediatric brain tumors	DKFZ	ICGC DCC	113	[89,90]
Breast	BREAST JHU	Breast cancer	Johns Hopkins University	ICGC DCC	42	[91]
	BREAST WTSI	Breast cancer	Welcome Trust/ Sanger Institute	ICGC DCC	100	[56]
	BREAST TN UBC	Triple negative breast cancer	University of British Columbia	PubMed	65	[92]
	BREAST TCGA	Breast invasive carcinoma	TCGA	Synapse	762	[93]
	BREAST BROAD	Breast cancer	BROAD Institute	PubMed	103	[94]
	BREAST ER + WU	ER + breast cancer	Washington University	PubMed	77	[38]
Colorectal	COLORECTAL ADENO JHU	Colorectal adenocarcinoma	Johns Hopkins University	ICGC DCC	36	[91]
	COLORECTAL ADENO TCGA	Colorectal adenocarcinoma	TCGA	Synapse	193	[28]
Head and neck	HEAD/NECK SQUAMOUS BROAD	Head and neck squamous cell carcinoma	Broad Institute	SM	74	[95]
	HEAD/NECK SQUAMOUS TCGA	Head and neck squamous cell carcinoma	TCGA	Synapse	301	-
Hematopoietic	CLL SPAIN	Chronic lymphocytic leukemia	Spanish Ministry of Science	ICGC DCC	109	[71,96]
	CLL DFCI	Chronic lymphocytic leukemia	Dana Farber Cancer Institute	SM	90	[97]
	AML TCGA	Acute myeloid leukemia	TCGA	Synapse	196	[98]
Kidney	KIDNEY CLEAR CELL TCGA	Kidney clear cell carcinoma	TCGA	Synapse	417	[99]
Liver	LIVER IARC	Liver cancer	IACR	ICGC DCC	24	[100]
Lung	LUNG ADENO WU	Lung adenocarcinoma	Washington University School of Medicine	ICGC DCC	162	[101]
	LUNG NON SMALL CELL MCW	Non small cell lung cancer	Medical College of Wisconsin	SM	31	[37]
	LUNG SQUAMOUS TCGA	Lung squamous cell carcinoma	TCGA	Synapse	174	[102]
	LUNG ADENO TCGA	Lung adenocarcinoma	TCGA	Synapse	228	-
	LUNG SMALL CELL UCOLOGNE	Small cell lung cancer	University Cologne	SM	27	[58]
	LUNG SMALL CELL JHU	Small cell lung cancer	Johns Hopkins University	SM	42	[103]
Ovary	OVARY TCGA	Ovarian serous cystadenocarcinoma	TCGA	Synapse	316	[104]
Pancreas	PANCREAS JHU	Pancreatic cancer	Johns Hopkins University	ICGC DCC	114	[105]
	PANCREAS OICR	Pancreatic cancer	Ontario Institute for Cancer Research	ICGC DCC	33	[106]
	PANCREAS QCMG	Pancreatic cancer	Queensland Centre for Medical Genomics	ICGC DCC	67	[106]
Stomach	GASTRIC PFIZER	Gastric cancer	Pfizer Worldwide Research and Development	SM	22	[107]
Uterus	UTERI TCGA	Uterine corpus endometrioid carcinoma	TCGA	Synapse	230	-

The results of all the analyses may be browsed and retrieved through IntOGen-mutations. TCGA, The Cancer Genome Atlas; ICGC, International Cancer Genomes Consortium; DCC, ICGC Data Coordination Center; DKFZ, German Cancer Research Center; IACR, International Agency for Research on Cancer; SM, Supplementary Material of articles.
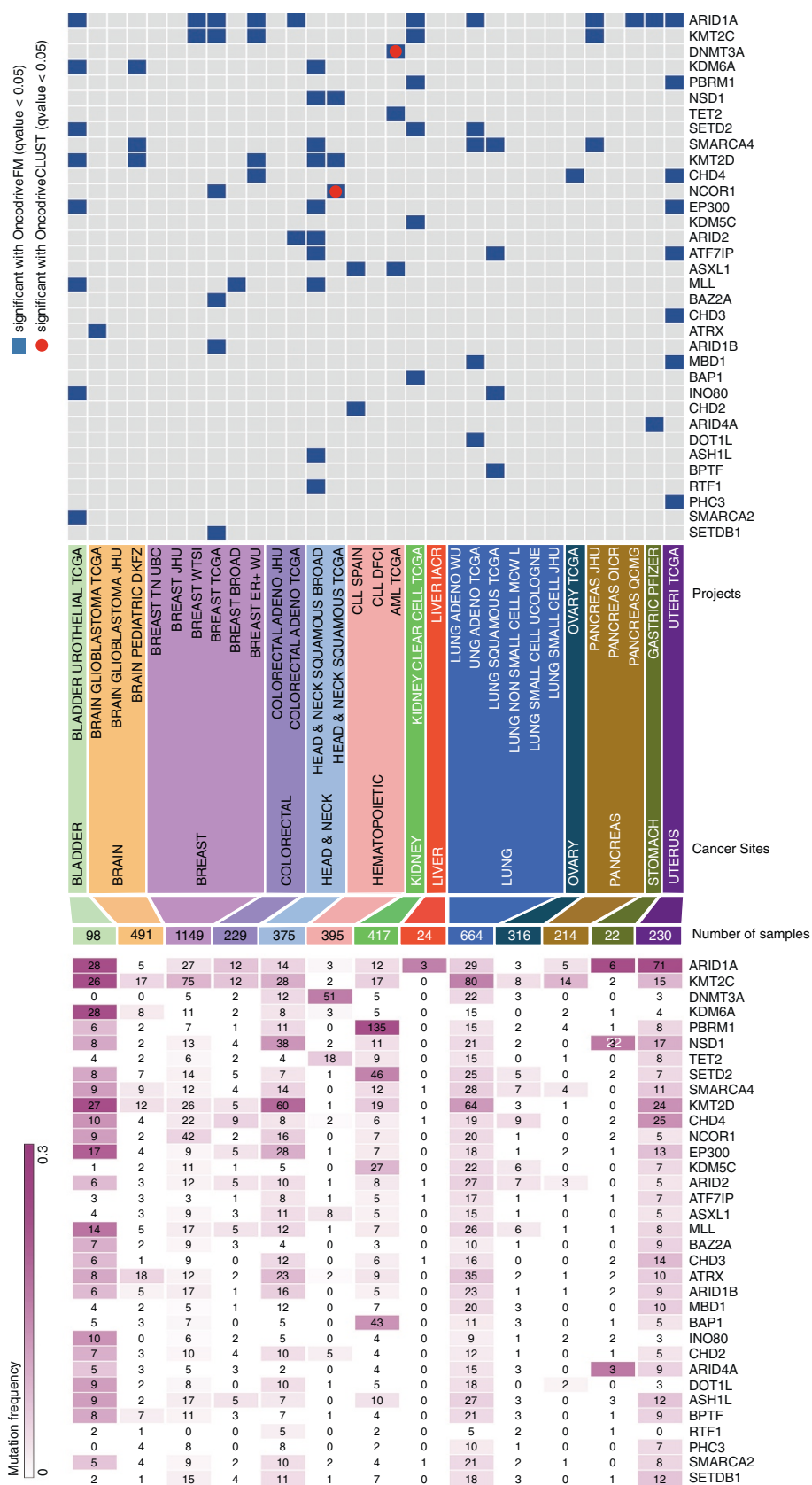
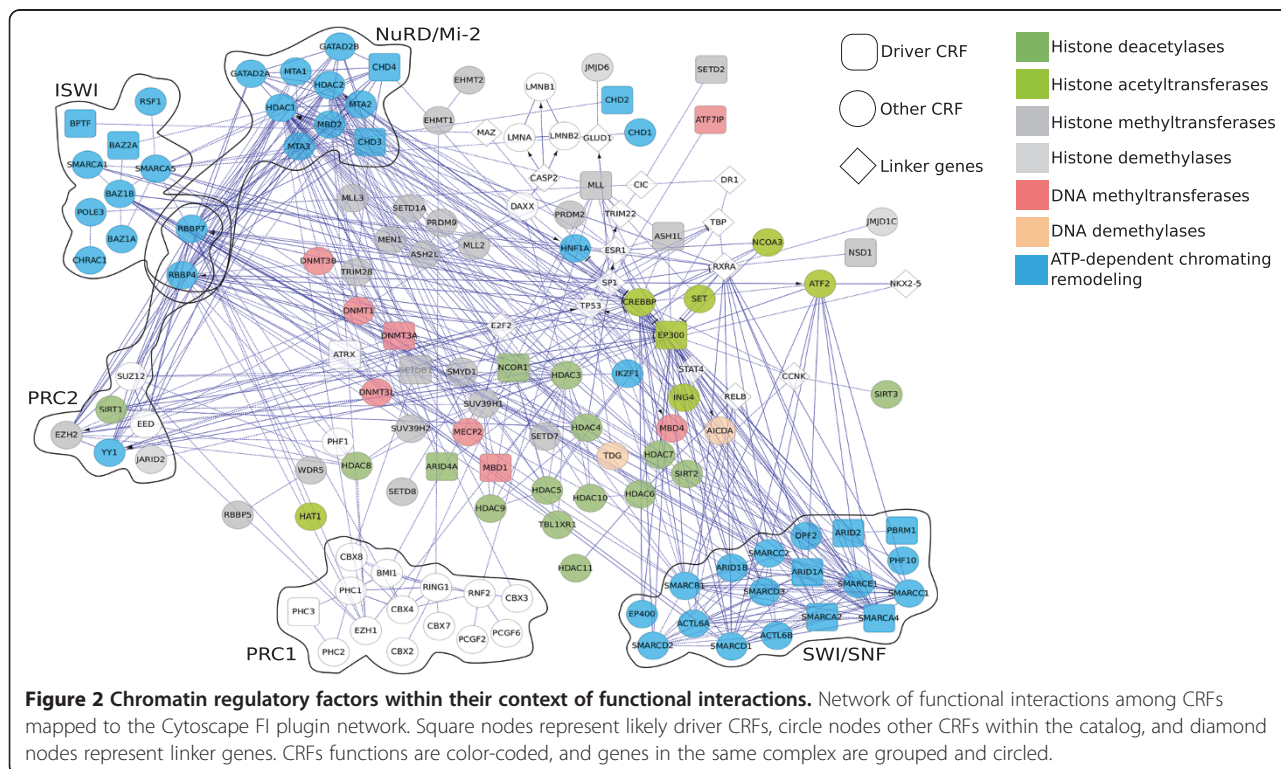**Figure 1** (See legend on next page.)

(See figure on previous page.)
Figure 1 Likely driver chromatin regulatory factors across the datasets of somatic mutations in IntOGen-mutations. The heat-map in the top panel identifies FM-biased and CLUST-biased CRFs across the 31 datasets from 13 sites in IntOGen-mutations, whose original projects are detailed in the middle panel. The heat-map in the bottom panel contains the number of samples with mutations in each likely driver CRF in each site. Cells in the heat-map are colored following mutational frequency.

although this may be attributed to the small sample size of the dataset. Several driver CRFs are mutated at frequencies above 10% in at least one site (Figure 1, lower panel). Functional relationships among many of them - 124 CRF genes can be mapped onto a pre-compiled [108] functional interaction network (Figure 2) - suggest the possibility that mutations in different genes produce similar malignancies (see below). We can therefore make the general observation that CRFs - 34 in the dataset collected by us - potentially act as mutational drivers in most of the cancer sites studied.

Driver candidates are significantly overrepresented within our catalog of CRFs (34 driver CRFs from 183 human CRFs in our list versus 348 total drivers from 22,696 human genes; Fisher's P value 1.26×10^{-25}). In addition, when analyzed as a group, the 183 CRFs in our catalog appear FM biased in all sites except liver (Figure 3A), which indicates that collectively they tend to accumulate mutations that on average possess higher functional impact than the background of the corresponding tumors. Taken together, these two observations

suggest that CRFs as a group may have an important role in tumorigenesis in the 13 sites with data in IntOGen.

Because CRFs usually act as multiprotein complexes, we also determined whether the best established of these complexes exhibit discernible signals of positive selection as a group across tumor samples. We did this in two ways. First, we computed the FM bias of six complexes described in Additional file 1: Table S1 and whose components appear illustrated in the network of functional interactions in Figure 2. We established that five of the complexes - ISWI being the exception - significantly accumulate highly impacting mutations in at least one site (Figure 3A). Second, we observed that pairs of proteins of the same complex tend to be mutated following a pattern of mutual exclusivity within cancer sites (Figure 3B and Additional file 1: Table S3). For example, the exploration of the SWI/SNF complex in breast tumors revealed that *ARID1A* tends to be mutated in samples where neither *SMARCA4*, *ARID2* nor *SMARCA2* are mutated. These two observations imply that multi-
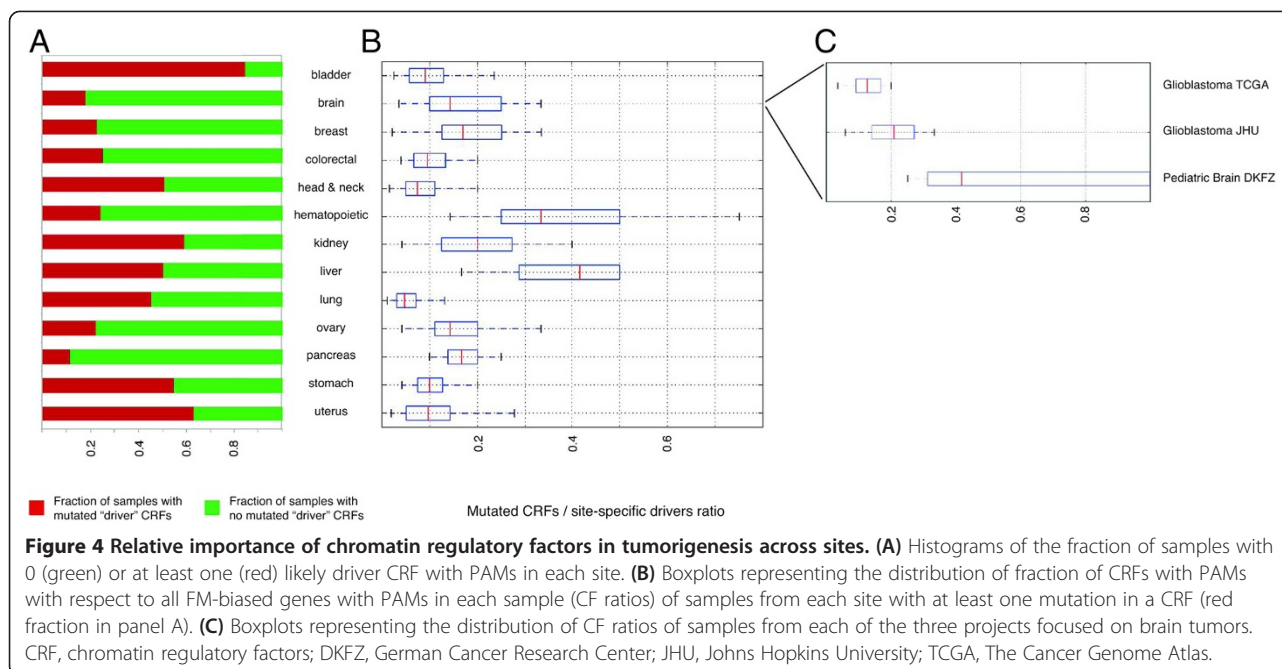

Figure 2 Chromatin regulatory factors within their context of functional interactions. Network of functional interactions among CRFs mapped to the Cytoscape FI plugin network. Square nodes represent likely driver CRFs, circle nodes other CRFs within the catalog, and diamond nodes represent linker genes. CRFs functions are color-coded, and genes in the same complex are grouped and circled.

Figure 3 FM bias, mutation frequencies and mutually exclusivity of chromatin regulatory factors as part of complexes. (A) Left heat-map shows the *P* value of FM bias analysis for all CRFs and for each complex. Right heat-map shows the number of samples with PAMs in the complex and the color indicates the mutation frequency (number of samples with PAMs divided by number of samples of this cancer type analyzed). **(B)** Heat-map of samples and genes of each complex, PAMs are represented as green cells in the heat-map. Tumor samples from each site have headers of the corresponding colors. Samples and genes in the heat-map are ordered based on mutually exclusive alterations within each site using Gitools built-in function for this purpose. Number of samples with PAMs in the gene (N) and the mutation frequency (Freq) of the gene in whole dataset are shown at the right of each heat-map. Gene names in bold indicated that the gene is one of the 34 detected as candidate drivers. PAM, protein-affecting mutations.

protein complexes, rather that individual genes, are the subjects of positive selection during tumorigenesis in the cancer sites under study.

The implication of chromatin regulatory factors in tumorigenesis strongly depends on the anatomical site and the tumor type

To determine whether there are differences in the implication of CRFs in tumorigenesis across the anatomical sites in IntOGen, we first computed the number of likely driver genes in general, and likely driver CRFs in particular, that bear protein sequence-affecting mutations, or PAMs (non-synonymous, stop, frameshift-causing insertions or deletions (indels)) in each tumor sample. From these data, the simplest way of representing the relative importance of mutations in CRFs in tumorigenesis across sites consists of counting the number of samples with at least one FM-biased CRF bearing a PAM (Figure 4A). In this metric, bladder urothelial carcinomas

and endometrial carcinomas stand out, with more than 80% and 60%, respectively, of the samples with at least one mutated CRF. On the opposite extreme, less than 10% of brain and hematopoietic tumor samples contain mutated likely driver CRFs.

We then computed the fraction of CRFs with PAMs with respect to all FM-biased genes with PAMs in each sample (CF ratio) (Figure 4B). This measure gives an indication of the relative importance of CRFs in the tumorigenesis process in each sample. Although liver or hematopoietic are not among the sites with the highest proportion of tumor samples with mutated CRFs (Figure 4A), these appear to be very important in the development of tumors in these sites (see the corresponding boxplots of Figure 4B). A closer look at the repertoire of mutated drivers in the samples of the three brain tumor datasets currently in IntOGen reveals that whereas mutations in classic tumor suppressors and oncogenes dominate the landscape of glioblastomas,

Figure 4 Relative importance of chromatin regulatory factors in tumorigenesis across sites. (A) Histograms of the fraction of samples with 0 (green) or at least one (red) likely driver CRF with PAMs in each site. **(B)** Boxplots representing the distribution of fraction of CRFs with PAMs with respect to all FM-biased genes with PAMs in each sample (CF ratios) of samples from each site with at least one mutation in a CRF (red fraction in panel A). **(C)** Boxplots representing the distribution of CF ratios of samples from each of the three projects focused on brain tumors. CRF, chromatin regulatory factors; DKFZ, German Cancer Research Center; JHU, Johns Hopkins University; TCGA, The Cancer Genome Atlas.

mutations in CRFs are more predominant in pediatric medulloblastomas. The median of the ratio of mutations in CRFs to mutations in all drivers across medulloblastoma samples is 0.4, compared to 0.21 and 0.1 in glioblastoma JHU (Johns Hopkins University; see Table 2) and glioblastoma TCGA (The Cancer Genome Atlas; see Table 2), respectively (Figure 4C). The samples of these two glioblastoma datasets exhibit a repertoire of mutated 'classical' tumor suppressors and oncogenes, such as *TP53*, *PTEN* and *EGFR* (Figure 5). As observed in the previous section, mutations in CRFs are likely drivers in tumors from most cancer types. Nevertheless, the latter result suggests that these mutations are circumscribed to a relatively small number of tumor samples, although future reviews of the catalogs of CRFs may increase the proportions calculated here.

Mutations in chromatin regulatory factors correlate with transcriptomic alterations of gene modules in cancer cell lines

To further understand the possible implication of CRFs in tumorigenesis, we explored the effects of CRF mutations on changes in the transcriptional levels of broad gene sets in cancer cell lines. To this end, we employed the data produced by The Cancer Cell Line Encyclopedia project, which has sequenced 1,651 protein-coding genes, of which 43 are CRFs according to our curated list (see Additional file 1: Table S1 for a detailed classification). First, to check whether cancer cell lines behave comparably to primary tumors in the patterns of altered gene modules, we carried out a sample level enrichment analysis (SLEA) [110] over cancer cell lines using Gene Ontology Biological Process

terms that are altered in specific cancer tissues. We found that cancer cell lines repeated the transcriptional profiles typical of their corresponding primary tumors (Additional file 2: Figure S1).

We then assessed the transcriptional impact of PAMs on *EP300* and *MLL3* (the only CRFs sustaining PAMs in sufficient cell lines: 115 and 191, respectively) to determine whether the impact of these PAMs on epigenetic regulation could translate into changes of the transcriptional levels of broad gene sets. The underlying hypothesis was that genes whose transcription was modulated by specific histone marks that became affected by PAMs on these two genes would present expression changes detectable when analyzed as a group. We collected regulatory modules of histone modifications in three cell types (Additional file 1: Table S4) and performed SLEA separately on cell lines originated from blood and solid tissues (Figure 6). As a result of the SLEA, we obtained a value of significance of the over-expression or under-expression (as a z-score) of each module in each cell line. We then compared the z-scores of cell lines that bear mutations in the gene in question (*EP300* or *MLL3*) to those cell lines where it does not, using the Wilcoxon-Mann–Whitney test. The *P*-values of the right-tail and left-tail comparisons were then adjusted using the Benjamini-Hochberg approach. Figure 6 presents the modules that rendered either significant right-tail or left-tail *P* values for any of the two genes. It shows that, in general, cell lines from solid tissues with mutations in either *EP300* or *MLL3* exhibited lower expression of repressed chromatin gene modules (H3K27 me3 and late replicating genes), and higher expression of gene modules with activating histone marks (marked by
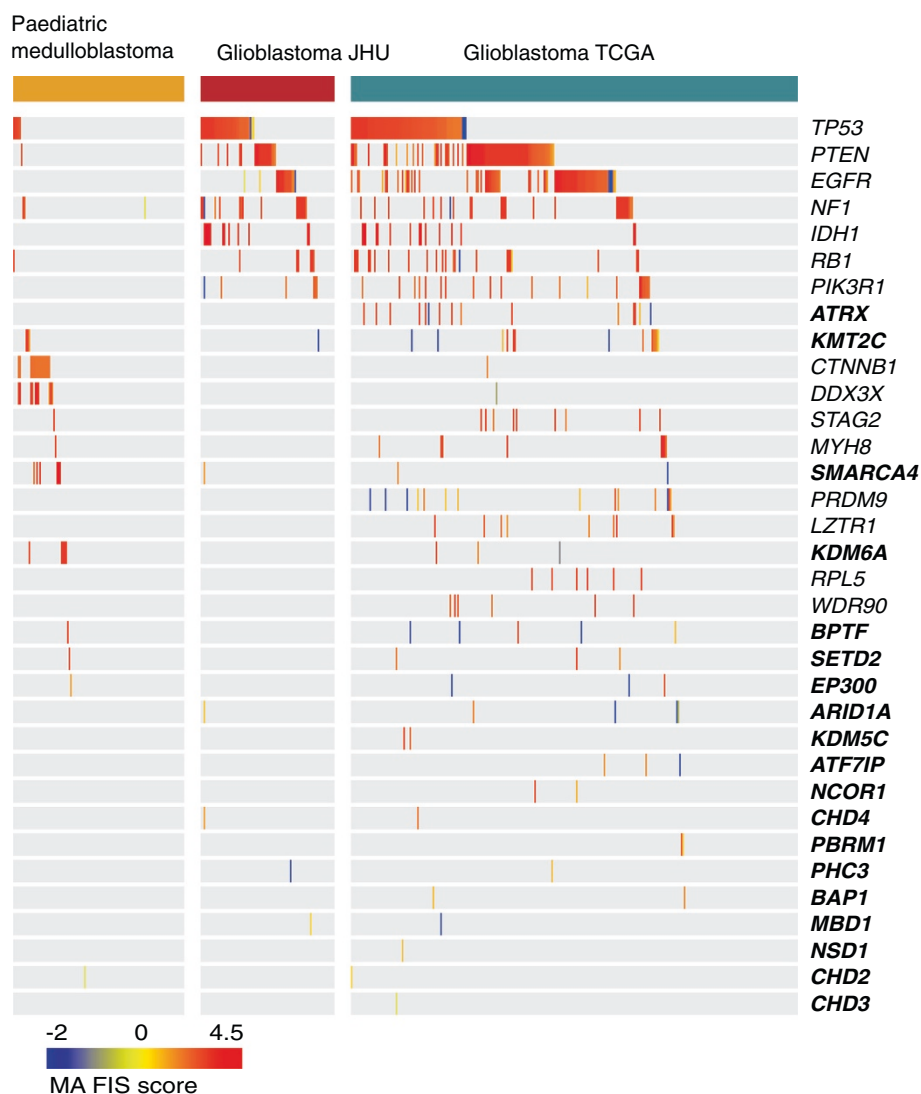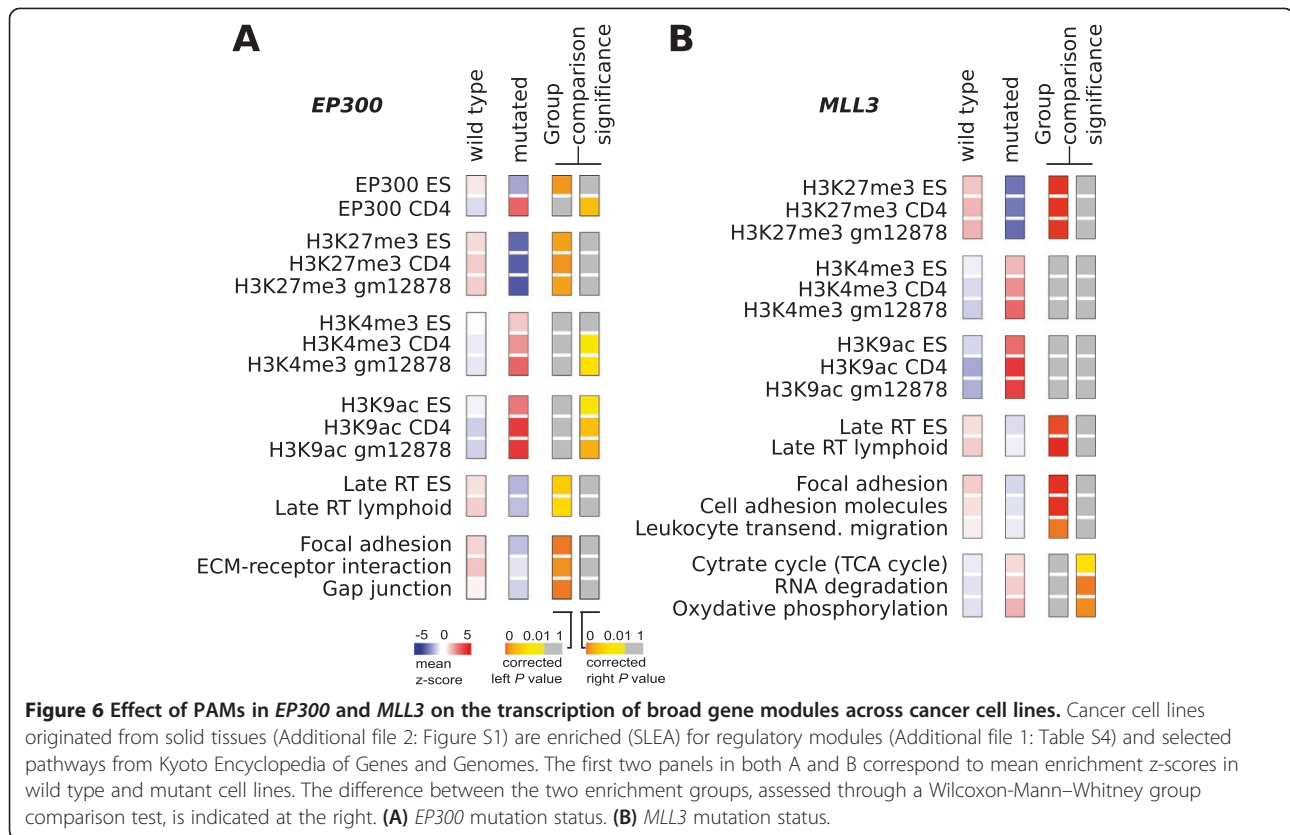
Figure 5 Mutational status of tumor samples from the three brain datasets included in IntOGen. The genes represented in the heat-map comprise all FM-biased CRFs that bear one mutation in at least one brain tumor sample (in bold typeface) plus the top 15 FM-biased genes in brain obtained from IntOGen. Mutations are represented by their MutationAssessor [109] functional impact scores (FIS). Samples and genes in the heat-map are ordered based on mutually exclusive alterations within dataset. FIS, functional impact score; MA, MutationAssessor score. JHU, Johns Hopkins University; TCGA, The Cancer Genome Atlas.

H3K4me3 and H3K9ac; Table 1). The under-expression of the H3K27me3 module, regulated by Polycomb, has been associated to a stem cell-like signature and more aggressive tumors [86]. Moreover, cell lines with mutations in *MLL3* showed higher expression of cell cycle-related modules. Taken together, these results suggest that mutations in CRFs may affect the transcriptional levels of gene sets bearing histone marks related to these CRFs.

Discussion

In this study, we found that several CRFs are likely involved in tumorigenesis in cancers from 13 anatomical sites. We uncovered these genes as putative causes of

the studied malignancies through the use of the FM bias and CLUST bias analyses, rather than the mere recurrence of mutations in genes across tumor samples. Moreover, by focusing on multiprotein complexes formed by several CRFs, we found evidence that suggest that these, rather than individual genes, are the subjects of positive selection during tumorigenesis. These two approaches constitute novelties with respect to the most recent and comprehensive analysis [21], which found recurrent mutations in SWI/SNF proteins across more than 650 tumor samples of 10 anatomical sites. Another important methodological novelty of our work consists in the use of CF ratios to assess the relevance of

Figure 6 Effect of PAMs in *EP300* and *MLL3* on the transcription of broad gene modules across cancer cell lines. Cancer cell lines originated from solid tissues (Additional file 2: Figure S1) are enriched (SLEA) for regulatory modules (Additional file 1: Table S4) and selected pathways from Kyoto Encyclopedia of Genes and Genomes. The first two panels in both A and B correspond to mean enrichment z-scores in wild type and mutant cell lines. The difference between the two enrichment groups, assessed through a Wilcoxon-Mann–Whitney group comparison test, is indicated at the right. **(A)** *EP300* mutation status. **(B)** *MLL3* mutation status.

mutations in CRFs in tumorigenesis in cancers from different sites. The employment of this ratio normalizes the number of CRFs with PAMs in the samples of a site by its intrinsic burden of mutations in driver genes. It is thus possible to observe that PAMs in CRFs, although highly prevalent in carcinomas of the uterus, probably play a relatively small role in their tumorigenesis because these tumors bear mutations in many other driver genes. However, mutations in CRFs appear to play a bigger role in tumorigenesis in hematopoietic malignancies than they do in tumors from other sites, although only few hematopoietic tumors bear PAMs in CRFs (Figure 4).

A group of pediatric medulloblastomas also possess abnormally high CF ratios, which implies that a high proportion of their mutated drivers are actually CRFs. It has been suggested that both pediatric and hematopoietic malignancies have very low mutational rates and therefore fewer drivers take part in their emergence than in solid adult tumors [111]. One could hypothesize from our results that alteration of either the transcriptional control or the chromatin maintenance of broad gene modules - as we observed in cell lines - via mutations in CRFs may be the crucial step of tumorigenesis in at least some of these tumors. This hypothesis, which could be experimentally tested, is another important contribution of the present work.

A third important contribution is the list of putative driver CRFs, which is available at IntOGen [112]. In particular, two of them were uncovered as putative drivers in more than one site (*CHD4* and *ATF7IP*) and are not annotated in the CGC [82]. They therefore constitute interesting candidates for novel epigenetic drivers (Figure 1). These additions to the list of driver CRFs might contribute to the research for anticancer drugs that takes CRFs as suitable targets.

Conclusions

We present the first systematic approach to characterize the repertoire of CRFs that could constitute mutational cancer drivers in tumors from 13 anatomical sites. We found that likely driver CRFs appear across tumor samples from most of these 13 sites, although the number of affected samples is in general low, except in the case of tumors from several sites, such as bladder, kidney and uterus. Mutations in CRFs appear to be in general only one of several contributing mechanisms towards tumorigenesis in most cancer samples. Finally, we have proved that mutations in two CRFs correlate with broad expression changes across cancer cell lines, thus presenting at least one mechanism through which these mutations could contribute to tumorigenesis in cells of the corresponding tissues. Our results expand the current knowledge

on the involvement of CRFs in tumorigenesis in several malignancies. Furthermore, they can contribute to formulate hypotheses on the mechanistic basis for this association. All the results presented here are available for browsing through the IntOGen-mutations platform [83,112] and using Gitools interactive heat-maps [113].

Materials and methods

Chromatin regulatory factors

We manually compiled a list of 183 genes coding for CRF proteins from the literature, based on protein function and known essential association to complexes important for the regulation of chromatin structure. A detailed classification of these CRFs is presented in Additional file 1: Table S1; the information was obtained from the Uniprot database [114] and the manuscripts referenced within the additional file. The relevant proteins for the purpose of this analysis are described in Table 1 and Additional file 1: Table S2.

FM-biased genes in primary tumors

FM-biased genes exhibit a bias towards the accumulation of functional mutations across a cohort of tumor samples and are therefore candidate cancer drivers. We have compiled 31 datasets of tumors from 13 anatomical sites and detected the FM-biased genes in each of them with the approach described in [22]. Genes that were not expressed across the major (TCGA) datasets included in IntOGen (obtained from syn1734155) were eliminated from the OncodriveFM analysis at this point. The overlap of drivers obtained from different datasets of mutations detected in tumors from the same anatomical site is shown in Additional file 2: Figure S2. Finally, we combined the gene-wise P values obtained for datasets of the same anatomical site to obtain a single P value that measures the bias of the gene towards the accumulation of functional mutations in different tumors from the same site. The corrected genes FM bias P values in these 13 tissues are stored in the IntOGen knowledgebase [87]. The collection of the datasets of tumor somatic mutations, their processing and browsing through IntOGen are thoroughly described in [83]. Details of the 31 tumor somatic mutations datasets are presented in Table 2.

CLUST-biased genes in primary tumors

PAMs in CLUST-biased genes tend to be grouped in regions of the proteins in a higher degree than synonymous mutations across the same dataset. This grouping constitutes another signal of positive selection that points to likely cancer drivers. The method to compute the CLUST bias in genes across datasets of tumor somatic mutations (OncodriveCLUST) is described in Tamborero et al., [23]. We computed the CLUST bias of all genes with PAMs across the 31 datasets compiled

and stored in IntOGen-mutations [83]. Genes that were not expressed across the major (TCGA) datasets included in IntOGen (obtained from syn1734155) were eliminated from the OncodriveCLUST analysis at this point. As with the FM bias, we combined the gene-wise P values obtained for datasets of tumor samples from the same anatomical site. The corrected genes' CLUST bias P values in these 13 tissues are stored in the IntOGen knowledgebase [87]. The collection of the datasets of tumor somatic mutations, their processing and browsing through IntOGen are thoroughly described in [83]. Details of the 31 tumor somatic mutations datasets are presented in Table 2.

Analysis of mutational frequencies of tumor samples

We defined a group of broad consequence types as corresponding to PAMs for all analyses of the mutational frequencies of tumor samples. All non-synonymous, stop and frameshift indels were included in this group. We recorded two numbers in the 4,623 tumor samples included in the study: the number of PAMs in any of the 34 likely driver CRFs detected across the 13 sites; and the number of PAMs in any of the 382 likely driver genes detected across the 13 sites. We then computed the ratio (CRFs-to-drivers ratio, or CF ratio) between these two numbers to assess the relevance of mutations in CRFs in tumorigenesis in every tumor sample. Note that because the 34 likely driver CRFs were included within the catalog of 382 likely drivers, the CF ratio takes values between 0 (no mutations in CRFs) and 1 (all mutated drivers in the sample are CRFs). Finally, we computed the number of tumor samples from each site with at least one PAM in a CRF and the distribution of their CF ratios.

Functional network analysis

We mapped the 183 CRFs in our catalog to the functional interactions network within the Cytoscape FI plugin [108,115], allowing the presence of linker genes to maximize the number of connected CRFs. Using Cytoscape, we then grouped genes in the same multiprotein complex (from the ones shown in Additional file 1: Table S2). We also mapped the biological functions of CRFs in the network using nodes colors, and whether they appeared as likely drivers through nodes shapes.

Cancer cell lines data processing

Expression arrays from the CCLE were downloaded from the Gene Expression Omnibus [GEO:GSE36133] as raw CEL files, and pre-processed as previously described [110]. The input data for enrichment analysis was obtained by median centering the expression value of each gene across cancer cell lines and dividing this value by the standard deviation. The obtained value is the

measure of expression level for the gene in a sample as compared to its expression level in all other samples in the dataset. We built separate expression matrices for cancer cell lines obtained from hematological system or solid primary cells, since the expression profiles of these two groups were shown to clearly differ in the original publication [24].

SLEA was performed using Gitools version 1.6.0 [116]. We used the z-score method as described previously [117]. This method compares the mean (or median) expression value of genes in each module to a distribution of mean (or median) of 10,000 random modules of the same size. Such enrichment analysis is run for each sample and the result is a z-score, which is a measure of the difference between the observed and expected mean (or median) expression values for genes in a module. We applied the mean z-score enrichment values, which are the arithmetic means of z-scores for individual samples, separately in cell lines obtained from hematological system or in those obtained from solid primary cells. To test for significant differences between the z-score means between groups of cell lines we used the Mann–Whitney test [118] implemented in Gitools. All heat-maps were generated with Gitools [119].

To detect potential PAMs in genes within the list of CRFs (Additional file 1: Table S1), we downloaded processed mutations data (single nucleotide variants and small indels) for 1,651 protein-coding genes (7 May 2012 version, excluding common polymorphisms and single nucleotide variants with an allelic fraction >10%) from the CCLE website [120]. We computed the consequence types of these variants using the Ensembl (v69) Variant Effect Predictor wrapped within the IntOGen-mutations pipeline [83].

Public gene regulation datasets

We collected lists of genes occupied by a specific histone mark or bound by a regulatory factor, and computationally predicted chromatin states, from available sources (Additional file 1: Table S4). These included human genome-wide occupancy datasets from ChIP-seq experiments in several cell types [121-125] that we processed using Bowtie [126] (version 0.12.5, hg19 genome assembly, unique alignments, allowing two mismatches) for short read aligning. For peak detection of transcription factors we used MACS [127] (version 1.4.1, settings: --nomodel and --bw parameter set to twice the shift size whenever a control immunoprecipitation was not available). For broad histone modifications (that is, H3K27me3), we used SICER [128] (version 1.1, setting gap size to 600). Regions were assigned to protein-coding genes (Ensembl v69) if they overlapped either to the gene body or up to 5 kb upstream from the transcription start site, using BedTools [129]. Overall peak calling performance

was evaluated with CEAS [130]. Other gene sets were obtained from KEGG [131] and Gene Ontology [132]. The list and mappings (in Ensembl v67 IDs) of KEGG and Gene Ontology Biological Process terms were downloaded through the Gitools importer [116].

Additional files

Additional file 1: Supplementary Tables S1, S2, S3 and S4 with titles and descriptions, and supplementary references.

Additional file 2: Supplementary Figure S1 and S2.

Additional file 3: Supplementary Table S5.

References

1. Papamichos-Chronakis M, Peterson CL: **Chromatin and the genome integrity network.** *Nat Rev Genet* 2013, **14**:62–75.
2. Elsässer SJ, Allis CD, Lewis PW: **New epigenetic drivers of cancers.** *Science* 2011, **331**:1145–1146.
3. Bestor TH: **Unanswered questions about the role of promoter methylation in carcinogenesis.** *Ann N Y Acad Sci* 2003, **983**:22–27.
4. Yan X-J, Xu J, Gu Z-H, Pan C-M, Lu G, Shen Y, Shi JY, Zhu YM, Tang L, Zhang XW, Liang WX, Mi JQ, Song HD, Li KQ, Chen Z, Chen SJ: **Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia.** *Nat Genet* 2011, **43**:309–315.
5. Walter MJ, Ding L, Shen D, Shao J, Grillot M, McLellan M, Fulton R, Schmidt H, Kalicki-Veizer J, O'Laughlin M, Kandoth C, Baty J, Westervelt P, DiPersio JF, Mardis ER, Wilson RK, Ley TJ, Graubert TA: **Recurrent DNMT3A mutations in patients with myelodysplastic syndromes.** *Leukemia* 2011, **25**:1153–1158.
6. Jene-Sanz A, Váraljai R, Vilkova AV, Khramtsova GF, Khramtsov AI, Olopade OI, Lopez-Bigas N, Benevolenskaya EV: **Expression of polycomb targets predicts breast cancer prognosis.** *Mol Cell Biol* 2013, **33**:3951–3961.

7. You JS, Jones PA: **Cancer genetics and epigenetics: two sides of the same coin?** *Cancer Cell* 2012, **22**:9–20.

8. Moore SDP, Herrick SR, Ince TA, Kleinman MS, Cin PD, Morton CC, Quade BJ: **Uterine leiomyomata with t(10;17) disrupt the histone acetyltransferase MORF.** *Cancer Res* 2004, **64**:5570–5577.

9. Shadeo A, Chari R, Lonergan KM, Pusic A, Miller D, Ehlen T, Van Niekerk D, Matisic J, Richards-Kortum R, Follen M, Guillaud M, Lam WL, MacAulay C: **Up regulation in gene expression of chromatin remodelling factors in cervical intraepithelial neoplasia.** *BMC Genomics* 2008, **9**:64.

10. Quintana RM, Dupuy AJ, Bravo A, Casanova ML, Alameda JP, Page A, Sánchez-Viera M, Ramírez A, Navarro M: **A transposon-based analysis of gene mutations related to skin cancer development.** *J Invest Dermatol* 2013, **133**:239–248.

11. Giannini G, Cabri W, Fattorusso C, Rodriquez M: **Histone deacetylase inhibitors in the treatment of cancer: overview and perspectives.** *Future Med Chem* 2012, **4**:1439–1460.

12. Baylin SB, Jones PA: **A decade of exploring the cancer epigenome — biological and translational implications.** *Nat Rev Cancer* 2011, **11**:726–734.

13. Sharma SV, Haber DA, Settleman J: **Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents.** *Nat Rev Cancer* 2010, **10**:241–253.

14. Patel MN, Halling-Brown MD, Tym JE, Workman P, Al-Lazikani B: **Objective assessment of cancer genes for drug discovery.** *Nat Rev Drug Discov* 2013, **12**:35–50.

15. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, Kandoth C, Payton JE, Baty J, Welch J, Harris CC, Lichti CF, Townsend RR, Fulton RS, Dooling DJ, Koboldt DC, Schmidt H, Zhang Q, Osborne JR, Lin L, O'Laughlin M, McMichael JF, Delehaunty KD, McGrath SD, Fulton LA, Magrini VJ, Vickery TL, Hundal J, Cook LL, Conyers JJ, *et al*: **DNMT3A mutations in acute myeloid leukemia.** *New Engl J Med* 2010, **363**:2424–2433.

16. Yamashita Y, Yuan J, Suetake I, Suzuki H, Ishikawa Y, Choi YL, Ueno T, Soda M, Hamada T, Haruta H, Takada S, Miyazaki Y, Kiyoi H, Ito E, Naoe T, Tomonaga M, Toyota M, Tajima S, Iwama A, Mano H: **Array-based genomic resequencing of human leukemia.** *Oncogene* 2010, **29**:3723–3731.

17. Uno K, Takita J, Yokomori K, Tanaka Y, Ohta S, Shimada H, Gilles FH, Sugita K, Abe S, Sako M, Hashizume K, Hayashi Y: **Aberrations of the hSNF5/INI1 gene are restricted to malignant rhabdoid tumors or atypical teratoid/rhabdoid tumors in pediatric solid tumors.** *Genes Chromosomes Cancer* 2002, **34**:33–41.

18. Jiao Y, Shi C, Edil BH, de Wilde RF, Klimstra DS, Maitra A, Schulick RD, Tang LH, Wolfgang CL, Choti MA, Velculescu VE, Diaz LA Jr, Vogelstein B, Kinzler KW, Hruban RH, Papadopoulos N: **DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors.** *Science* 2011, **331**:1199–1203.

19. Banine F, Bartlett C, Gunawardena R, Muchardt C, Yaniv M, Knudsen ES, Weissman BE, Sherman LS: **SWI/SNF chromatin-remodeling factors induce changes in DNA methylation to promote transcriptional activation.** *Cancer Res* 2005, **65**:3542–3547.

20. Jones S, Wang T-L, Shih I-M, Mao T-L, Nakayama K, Roden R, Glas R, Slamon D, Diaz LA Jr, Vogelstein B, Kinzler KW, Velculescu VE, Papadopoulos N: **Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma.** *Science* 2010, **330**:228–231.

21. Shain AH, Pollack JR: **The Spectrum of SWI/SNF mutations. Ubiquitous in human cancers.** *PLoS One* 2013, **8**:e55119.

22. Gonzalez-Perez A, Lopez-Bigas N: **Functional impact bias reveals cancer drivers.** *Nucl Acids Res* 2012, **40**:e169.

23. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N: **Oncodrive CLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.** *Bioinformatics* 2013, **29**:2238–2244.

24. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, *et al*: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.** *Nature* 2012, **483**:307–603.

25. Gui Y, Guo G, Huang Y, Hu X, Tang A, Gao S, Wu R, Chen C, Li X, Zhou L, He M, Li Z, Sun X, Jia W, Chen J, Yang S, Zhou F, Zhao X, Wan S, Ye R, Liang C, Liu Z, Huang P, Liu C, Jiang H, Wang Y, Zheng H, Sun L, Liu X, Jiang Z, *et al*: **Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder.** *Nat Genet* 2011, **43**:875–878.

26. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, *et al*: **Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators.** *Nat Genet* 2012, **44**:760–764.

27. Gallo ML, O'Hara AJ, Rudd ML, Urick ME, Hansen NF, O'Neil NJ, Price JC, Zhang S, England BM, Godwin AK, Sgroi DC, NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program, Hieter P, Mullikin JC, Merino MJ, Bell DW: **Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes.** *Nat Genet* 2012, **44**:1310–1315.

28. The Cancer Genome Atlas Research Network: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487**:330–337.

29. Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, Rajasegaran V, Heng HL, Deng N, Gan A, Lim KH, Ong CK, Huang D, Chin SY, Tan IB, Ng CC, Yu W, Wu Y, Lee M, Wu J, Poh D, Wan WK, Rha SY, So J, Salto-Tellez M, Yeoh KG, Wong WK, Zhu YJ, Futreal PA, Pang B, *et al*: **Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes.** *Nat Genet* 2012, **44**:570–574.

30. Biankin AV, Waddell N, Kassahn KS, Gingras M-C, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, Patch AM, Wu J, Chang DK, Cowley MJ, Gardiner BB, Song S, Harliwong I, Idrisoglu S, Nourse C, Nourbakhsh E, Manning S, Wani S, Gongora M, Pajic M, Scarlett CJ, Gill AJ, Pinho AV, Rooman I, Anderson M, Holmes O, Leonard C, Taylor D, *et al*: **Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes.** *Nature* 2012, **491**:399–405.

31. Seo J-S, Ju YS, Lee W-C, Shin J-Y, Lee JK, Bleazard T, Lee J, Jung YJ, Kim JO, Shin JY, Yu SB, Kim J, Lee ER, Kang CH, Park IK, Rhee H, Lee SH, Kim JI, Kang JH, Kim YT: **The transcriptional landscape and mutational profile of lung adenocarcinoma.** *Genome Res* 2012, **22**:2109–2119.

32. Love C, Sun Z, Jima D, Li G, Zhang J, Miles R, *et al*: **The genetic landscape of mutations in Burkitt lymphoma.** *Nat Genet* 2012, **44**:1321–1325.

33. Sausen M, Leary RJ, Jones S, Wu J, Reynolds CP, Liu X, *et al*: **Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma.** *Nat Genet* 2013, **45**:12–17.

34. Mamo A, Cavallone L, Tuzmen S, Chabot C, Ferrario C, Hassan S, *et al*: **An integrated genomic approach identifies ARID1A as a candidate tumor-suppressor gene in breast cancer.** *Oncogene* 2012, **31**:2090–2100.

35. Lindberg J, Mills IG, Klevebring D, Liu W, Neiman M, Xu J, *et al*: **The mitochondrial and autosomal mutation landscapes of prostate cancer.** *Eur Urol* 2012, **63**:702–708.

36. Watanabe Y, Castoro RJ, Kim HS, North B, Oikawa R, Hiraishi T, *et al*: **Frequent alteration of MLL3 Frameshift mutations in microsatellite deficient colorectal cancer.** *PLoS One* 2011, **6**:e23320.

37. Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P, *et al*: **Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing.** *Carcinogenesis* 2012, **33**:1270–1276.

38. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, *et al*: **Whole-genome analysis informs breast cancer response to aromatase inhibition.** *Nature* 2012, **486**:353–360.

39. Ruault M, Brun ME, Ventura M, Roizès G, De Sario A: **MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently deleted in myeloid leukaemia.** *Gene* 2002, **284**:73–81.

40. Kim MS, Kim YR, Yoo NJ, Lee SH: **Mutational analysis of DNMT3A gene in acute leukemias and common solid cancers.** *APMIS* 2012, **12**:85–94.

41. Bai X, Song Z, Fu Y, Yu Z, Zhao L, Zhao H, *et al*: **Clinicopathological significance and prognostic value of DNA methyltransferase 1, 3a, and 3b expressions in sporadic epithelial ovarian cancer.** *PLoS One* 2012, **7**:e40024.

42. Liu J, Lee W, Jiang Z, Chen Z, Jhunjhunwala S, Haverty PM, *et al*: **Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events.** *Genome Res* 2012, **22**:2315–2327.

43. Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, *et al*: **Novel mutations target distinct subgroups of medulloblastoma.** *Nature* 2012, **488**:43–48.

44. van Haaften G, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, *et al*: **Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer.** *Nat Genet* 2009, **41**:521–523.

45. Paolicchi E, Crea F, Farrar WL, Green JE, Danesi R: **Histone lysine demethylases in breast cancer.** *Cr Rev Oncol-Hem* 2012, **86**:97–103.

46. Liu L, Zhao E, Li C, Huang L, Xiao L, Cheng L, et al: **TRIM28, a new molecular marker predicting metastasis and survival in early-stage non-small cell lung cancer.** *Cancer Epidem* 2013, **37**:71–78.

47. Shain AH, Giacomini CP, Matsukuma K, Karikari CA, Bashyam MD, Hidalgo M, et al: **Convergent structural alterations define SWItch/Sucrose NonFermentable (SWI/SNF) chromatin remodeler as a central tumor suppressive complex in pancreatic cancer.** *Proc Natl Acad Sci USA* 2012, **109**:E252–E259.

48. Job B, Bernheim A, Beau-Faller M, Camilleri-Broët S, Girard P, Hofman P, et al: **Genomic aberrations in lung adenocarcinoma in never smokers.** *PLoS One* 2010, **5**:e15145.

49. Abdel-Wahab O, Mullally A, Hedvat C, Garcia-Manero G, Patel J, Wadleigh M, et al: **Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies.** *Blood* 2009, **114**:144–147.

50. Al Sarakbi W, Sasi W, Jiang WG, Roberts T, Newbold RF, Mokbel K: **The mRNA expression of SETD2 in human breast cancer: correlation with clinico-pathological parameters.** *BMC Cancer* 2009, **9**:290.

51. Bai J, Mei P-J, Liu H, Li C, Li W, Wu Y-P, et al: **BRG1 expression is increased in human glioma and controls glioma cell proliferation, migration and invasion in vitro.** *J Cancer Res Clin Oncol* 2012, **138**:991–998.

52. Saladi SV, Keenen B, Marathe HG, Qi H, Chin K-V, de la Serna IL: **Modulation of extracellular matrix/adhesion molecule expression by BRG1 is associated with increased melanoma invasiveness.** *Mol Cancer* 2010, **9**:280.

53. Pasqualucci L, Trifonov V, Fabbri G, Ma J, Rossi D, Chiarenza A, et al: **Analysis of the coding genome of diffuse large B-cell lymphoma.** *Nat Genet* 2011, **43**:830–837.

54. Natarajan TG, Kallakury BV, Sheehan CE, Bartlett MB, Ganesan N, Preet A, et al: **Epigenetic regulator MLL2 shows altered expression in cancer cell lines and tumors from human breast and colon.** *Cancer Cell Int* 2010, **10**:13.

55. Kim MS, Chung NG, Kang MR, Yoo NJ, Lee SH: **Genetic and expressional alterations of CHD genes in gastric and colorectal cancers.** *Histopathology* 2011, **58**:660–668.

56. Stephens PJ, Tarpey PS, Davies H, Loo PV, Greenman C, Wedge DC, et al: **The landscape of cancer genes and mutational processes in breast cancer.** *Nature* 2012, **486**:400–404.

57. Zhang Z, Yamashita H, Toyama T, Sugiura H, Ando Y, Mita K, et al: **NCOR1 mRNA is an independent prognostic factor for breast cancer.** *Cancer Lett* 2006, **237**:123–129.

58. Peifer M, Fernández-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, et al: **Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer.** *Nat Genet* 2012, **44**:1104–1110.

59. Li Y, Yang H-X, Luo R-Z, Zhang Y, Li M, Wang X, et al: **High expression of p300 has an unfavorable impact on survival in resectable esophageal squamous cell carcinoma.** *Ann Thorac Surg* 2011, **91**:1531–1538.

60. Yokomizo C, Yamaguchi K, Itoh Y, Nishimura T, Umemura A, Minami M, et al: **High expression of p300 in HCC predicts shortened overall survival in association with enhanced epithelial mesenchymal transition of HCC cells.** *Cancer Lett* 2011, **310**:140–147.

61. Dokmanovic M, Clarke C, Marks PA: **Histone deacetylase inhibitors: overview and perspectives.** *Mol Cancer Res* 2007, **5**:981–989.

62. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat J-P, et al: **A landscape of driver mutations in melanoma.** *Cell* 2012, **150**:251–263.

63. Manceau G, Letouzé E, Guichard C, Didelot A, Cazes A, Corté H, et al: **Recurrent inactivating mutations of ARID2 in non-small cell lung carcinoma.** *Int J Cancer* 2012, **132**:2217–2221.

64. Vainchenker W, Delhommeau F, Constantinescu SN, Bernard OA: **New mutations and pathogenesis of myeloproliferative neoplasms.** *Blood* 2011, **118**:1723–1735.

65. Devillier R, Gelsi-Boyer V, Brecqueville M, Carbuccia N, Murati A, Vey N, et al: **Acute myeloid leukemia with myelodysplasia-related changes are characterized by a specific molecular pattern with high frequency of ASXL1 mutations.** *Am J Hematol* 2012, **87**:659–662.

66. Grasso CS, Wu Y-M, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al: **The mutational landscape of lethal castrate resistant prostate cancer.** *Nature* 2012, **487**:239–243.

67. Hanlon K, Rudin CE, Harries LW: **Investigating the targets of MIR-15a and MIR-16-1 in patients with Chronic Lymphocytic Leukemia (CLL).** *PLoS One* 2009, **4**:e7169.

68. Jiao Y, Killela PJ, Reitman ZJ, Rasheed BA, Heaphy CM, de Wilde RF, et al: **Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas.** *Oncotarget* 2012, **3**:709–722.

69. Xu J, Zhu W, Xu W, Yao W, Zhang B, Xu Y, et al: **Up-regulation of MBD1 promotes pancreatic cancer cell epithelial-mesenchymal transition and invasion by epigenetic down-regulation of E-cadherin.** *Curr Mol Med* 2013, **13**:387–400.

70. Fan L-H, Tang L-N, Yue L, Yang Y, Gao Z-L, Shen Z: **BAP1 is a good prognostic factor in advanced non-small cell lung cancer.** *Clin Invest Med* 2012, **35**:E182–E189.

71. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al: **Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia.** *Nat Genet* 2012, **44**:47–52.

72. Bandrés E, Malumbres R, Cubedo E, Honorato B, Zarate R, Labarga A, et al: **A gene signature of 8 genes could identify the risk of recurrence and progression in Dukes' B colon cancer patients.** *Oncol Rep* 2007, **17**:1089–1094.

73. Skawran B, Steinemann D, Weigmann A, Flemming P, Becker T, Flik J, et al: **Gene expression profiling in hepatocellular carcinoma: upregulation of genes in amplified chromosome regions.** *Mod Pathol* 2008, **21**:505–516.

74. Buganim Y, Goldstein I, Lipson D, Milyavsky M, Polak-Charcon S, Mardoukh C, et al: **A novel translocation breakpoint within the BPTF gene is associated with a pre-malignant phenotype.** *PLoS One* 2010, **5**:e9657.

75. Deshpande AM, Akunowicz JD, Reveles XT, Patel BB, Saria EA, Gorlick RG, et al: **PHC3, a component of the hPRC-H complex, associates with E2F6 during G0 and is lost in osteosarcoma tumors.** *Oncogene* 2007, **26**:1714–1722.

76. Moloney FJ, Lyons JG, Bock VL, Huang XX, Bugeja MJ, Halliday GM: **Hotspot mutation of Brahma in non-melanoma skin cancer.** *J Invest Dermatol* 2009, **129**:1012–1015.

77. Doménech E, Gómez-López G, Gzlez-Peña D, López M, Herreros B, Menezes J, et al: **New mutations in chronic lymphocytic leukemia identified by target enrichment and deep sequencing.** *PLoS One* 2012, **7**:e38158.

78. Reisman DN, Sciarrotta J, Wang W, Funkhouser WK, Weissman BE: **Loss of BRG1/BRM in human lung cancer cell lines and primary lung cancers: correlation with poor prognosis.** *Cancer Res* 2003, **63**:560–566.

79. Yamamichi N, Inada K, Ichinose M, Yamamichi-Nishina M, Mizutani T, Watanabe H, et al: **Frequent loss of Brm expression in gastric cancer correlates with histologic features and differentiation state.** *Cancer Res* 2007, **67**:10727–10735.

80. Hélias C, Struski S, Gervais C, Leymarie V, Mauvieux L, Herbrecht R, et al: **Polycythemia vera transforming to acute myeloid leukemia and complex abnormalities including 9p homogeneously staining region with amplification of MLLT3, JMJD2C, JAK2, and SMARCA2.** *Cancer Genet Cytogenet* 2008, **180**:51–55.

81. Ceol CJ, Houvras Y, Jane-Valbuena J, Bilodeau S, Orlando DA, Battisti V, et al: **The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset.** *Nature* 2011, **471**:513–517.

82. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4**:177–183.

83. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al: **IntOGen-mutations identifies cancer drivers across tumor types.** *Nat Meth* 2013, **5**:5.

84. McLendon R, Friedman A, Bigner D, Meir EGV, Brat DJ, Mastrogianakis GM, et al: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**:1061–1068.

85. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al: **International network of cancer genome projects.** *Nature* 2010, **464**:993–998.

86. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, et al: **An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors.** *Nat Genet* 2008, **40**:499–507.

87. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, et al: **IntOGen: Integration and data-mining of multidimensional oncogenomic data.** *Nat Meth* 2010, **7**:92–93.

88. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al: **An integrated genomic analysis of human glioblastoma multiforme.** *Science* 2008, **321**:1807–1812.

89. Rausch T, Jones DTW, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, et al: **Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations.** *Cell* 2012, **148**:59–71.

90. Schwartzentruber J, Korshunov A, Liu X-Y, Jones DTW, Pfaff E, Jacob K, et al: **Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma.** *Nature* 2012, **482**:226–231.

91. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318**:1108–1113.

92. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al: **The clonal and mutational evolution spectrum of primary triple-negative breast cancers.** Nature 2012, **486**:395–399.
93. The Cancer Genome Atlas Network: **Comprehensive molecular portraits of human breast tumours.** Nature 2012, **490**:61–70.
94. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, et al: **Sequence analysis of mutations and translocations across breast cancer subtypes.** Nature 2012, **486**:405–409.
95. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al: **The mutational landscape of head and neck squamous cell carcinoma.** Science 2011, **333**:1157–1160.
96. Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, et al: **Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia.** Nature 2011, **475**:101–105.
97. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, et al: **SF3B1 and other novel cancer genes in chronic lymphocytic leukemia.** New Engl J Med 2011, **365**:2497–2506.
98. The Cancer Genome Atlas Research Network: **Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia.** New Engl J Med 2013, **368**:2059–2074.
99. The Cancer Genome Atlas Research Network: **Comprehensive molecular characterization of clear cell renal cell carcinoma.** Nature 2013, **499**:43–49.
100. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, et al: **Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma.** Nat Genet 2012, **44**:694–698.
101. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al: **Somatic mutations affect key pathways in lung adenocarcinoma.** Nature 2008, **455**:1069–1075.
102. The Cancer Genome Atlas Research Network: **Comprehensive genomic characterization of squamous cell lung cancers.** Nature 2012, **489**:519–525.
103. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al: **Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer.** Nat Genet 2012, **44**:1111–1116.
104. The Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.** Nature 2011, **474**:609–615.
105. Jones S, Zhang X, Parsons DW, Lin JC-H, Leary RJ, Angenendt P, et al: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** Science 2008, **321**:1801–1806.
106. ICGC data portal. [http://dcc.icgc.org]
107. Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, et al: **Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer.** Nat Genet 2011, **43**:1219–1223.
108. Wu G, Feng X, Stein L: **A human functional protein interaction network and its application to cancer data analysis.** Genome Biol 2010, **11**:R53.
109. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** Nucl Acids Res 2011, **39**:e118–e118.
110. Gundem G, Lopez-Bigas N: **Sample level enrichment analysis unravels shared stress phenotypes among multiple cancer types.** Genome Med 2012, **4**:28.
111. Schinzel AC, Hahn WC: **Oncogenic transformation and experimental models of human cancer.** Front Biosci 2008, **13**:71–84.
112. Integrative Onco Genomics: mutations. [http://www.intogen.org/mutations]
113. Gitools IntOgen datasets. [www.gitools.org/datasets#IntOGen]
114. The UniProt Consortium: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** Nucleic Acids Res 2011, **40**:D71–D75.
115. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** Genome Res 2003, **13**:2498–2504.
116. Perez-Llamas C, Lopez-Bigas N: **Gitools: analysis and visualisation of genomic data using interactive heat-maps.** PLoS ONE 2011, **6**:e19541.
117. Lopez-Bigas N, De S, Teichmann SA: **Functional protein divergence in the evolution of Homo sapiens.** Genome Biol 2008, **9**:R33.
118. Mann HB, Whitney DR: **On a test of whether one of two random variables is stochastically larger than the other.** Ann Math Statist 1947, **18**:50–60.
119. Gitools. [http://www.gitools.org]
120. Broad-novartis cancer cell line encyclopedia. [http://www.broadinstitute.org/ccle].

121. The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** Nature 2007, **447**:799–816.
122. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al: **Human DNA methylomes at base resolution show widespread epigenomic differences.** Nature 2009, **462**:315–322.
123. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al: **High-resolution profiling of histone methylations in the human genome.** Cell 2007, **129**:823–837.
124. Viré E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, et al: **The Polycomb group protein EZH2 directly controls DNA methylation.** Nature 2005, **439**:871–874.
125. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al: **Sequencing newly replicated DNA reveals widespread plasticity in human replication timing.** Proc Natl Acad Sci 2010, **107**:139–144.
126. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** Genome Biol 2009, **10**:R25.
127. Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, et al: **Model-based Analysis of ChIP-Seq (MACS).** Genome Biol 2008, **9**:R137.
128. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: **A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.** Bioinformatics 2009, **25**:1952–1958.
129. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** Bioinformatics 2010, **26**:841–842.
130. Ji X, Li W, Song J, Wei L, Liu XS: **CEAS: cis-regulatory element annotation system.** Nucl Acids Res 2006, **34**:W551–W554.
131. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** Nucleic Acids Res 2012, **40**:D109–D114.
132. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al: **Gene Ontology: tool for the unification of biology.** Nat Genet 2000, **25**:25–29.