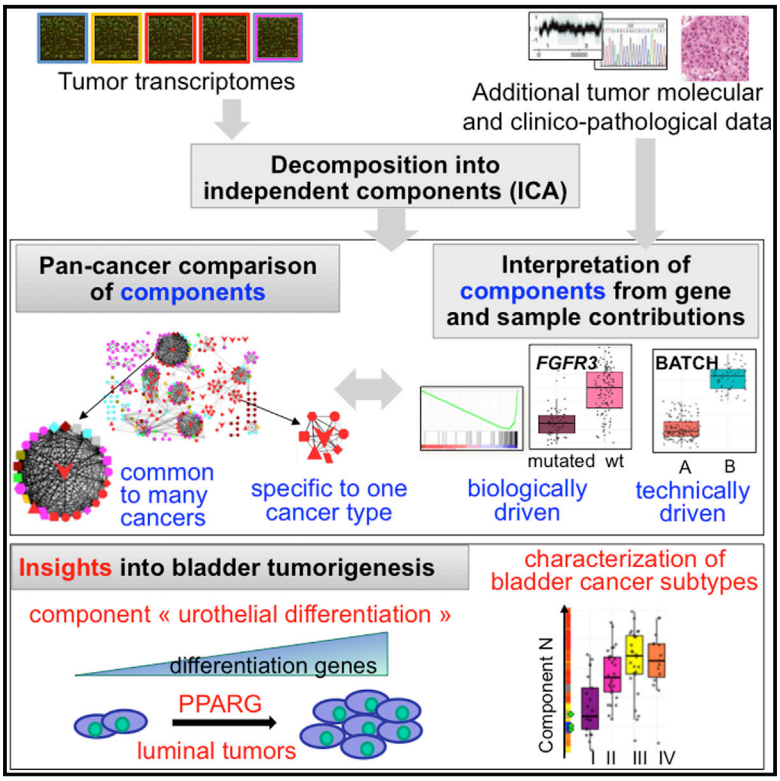
Report

Cell Reports

Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes

Graphical Abstract



Highlights

ICA analysis uncovers the various factors influencing transcriptomic data

ICA of multiple data sets reveals cancer-shared and cancer-type-specific signals

The components identified by ICA allow characterization of bladder tumor subtypes

PPARG is a protumorigenic gene associated with differentiation in bladder cancer

Authors

Anne Biton, Isabelle Bernard-Pierrot, ..., Andrei Zinovyev, François Radvanyi

Correspondence

francois.radvanyi@curie.fr

In Brief

Extracting biological insights from large-scale data is both a challenge and an opportunity. Biton et al. now analyze bladder tumor transcriptomes. An enrichment analysis of contributing genes combined with molecular and clinical annotations of tumor samples identifies biologically relevant components, some of which are shared with other carcinoma types.

Biton et al., 2014, Cell Reports 9, 1235–1245
November 20, 2014 ©2014 The Authors
http://dx.doi.org/10.1016/j.celrep.2014.10.035

Cell Reports
Report

Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes

Anne Biton,[1,2,3,4,16,18] Isabelle Bernard-Pierrot,[1,4,16] Yinjun Lou,[1,4,19] Clémentine Krucker,[1,4] Elodie Chapeaublanc,[1,4] Carlota Rubio-Pérez,[5] Nuria López-Bigas,[5,6] Aurélie Kamoun,[1,4] Yann Neuzillet,[1,4,7,8] Pierre Gestraud,[1,2,3] Luca Grieco,[1,2,3] Sandra Rebouissou,[1,4,20] Aurélien de Reyniès,[9] Simone Benhamou,[10,11] Thierry Lebret,[7,8] Jennifer Southgate,[12] Emmanuel Barillot,[1,2,3] Yves Allory,[13,14,15] Andrei Zinovyev,[1,2,3,17] and François Radvanyi[1,4,17,*]
[1]Institut Curie, Centre de Recherche, 75248 cedex 05 Paris, France
[2]INSERM, U900, 75248 cedex 05 Paris, France
[3]Ecole des Mines ParisTech, 77305 cedex Fontainebleau, France
[4]CNRS, UMR 144, Oncologie Moléculaire, Equipe Labellisée Ligue Contre le Cancer, Institut Curie, 75248 cedex 05 Paris, France
[5]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain
[6]Catalan Institution for Research and Advanced Studies (ICREA), 08003 Barcelona, Spain
[7]Service d'Urologie, Hôpital Foch, 92150 Suresnes, France
[8]Université de Versailles - Saint-Quentin-en-Yvelines, Faculté de Médecine Paris - Ile-de-France Ouest, 78280 Guyancourt, France
[9]Ligue Nationale Contre le Cancer, Cartes d'Identité des Tumeurs Program, 75013 Paris, France
[10]CNRS, UMR 8200, Institut de Cancérologie Gustave Roussy, 94805 Villejuif, France
[11]INSERM, U946, 75010 Paris, France
[12]Jack Birch Unit of Molecular Carcinogenesis, Department of Biology, University of York, York Y010 5DD, UK
[13]AP-HP, Département de Pathologie, Hôpitaux Universitaires Henri Mondor, 94000 Créteil, France
[14]INSERM, U955, 94000 Créteil, France
[15]Université Paris-Est, Faculté de Médecine, 94000 Créteil, France
[16]Co-first author
[17]Co-senior author
[18]Present address: Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, CA 94143-2922, USA
[19]Present address: ZheJiang University, Institute of Hematology, HangZhou, ZheJiang 310058, China
[20]Present address: INSERM, UMR-674, Génomique Fonctionnelle des Tumeurs Solides, IUH, 75010 Paris, France
*Correspondence: francois.radvanyi@curie.fr
http://dx.doi.org/10.1016/j.celrep.2014.10.035

SUMMARY

Extracting relevant information from large-scale data offers unprecedented opportunities in cancerology. We applied independent component analysis (ICA) to bladder cancer transcriptome data sets and interpreted the components using gene enrichment analysis and tumor-associated molecular, clinicopathological, and processing information. We identified components associated with biological processes of tumor cells or the tumor microenvironment, and other components revealed technical biases. Applying ICA to nine cancer types identified cancer-shared and bladder-cancer-specific components. We characterized the luminal and basal-like subtypes of muscle-invasive bladder cancers according to the components identified. The study of the urothelial differentiation component, specific to the luminal subtypes, showed that a molecular urothelial differentiation program was maintained even in those luminal tumors that had lost morphological differentiation. Study of the genomic alterations associated with this component coupled with functional studies revealed a protumorigenic role for PPARG in luminal tumors. Our results support the inclusion of ICA in the exploitation of multiscale data sets.

INTRODUCTION

Large-scale projects (e.g., Cancer Genome Project [CGP], The Cancer Genome Atlas [TCGA], or the International Cancer Genome Consortium [ICGC]), and the efforts of individual laboratories, are generating massive amounts of high-throughput molecular data often associated with clinicopathological characteristics. Transcriptome data for tumors are the most commonly available type of large-scale molecular data. These remain difficult to interpret because the transcriptome is influenced by various overlapping biological factors linked to the tumor cells or to the tumor microenvironment and nonbiological factors linked to sample processing or data generation. The need to deconvolute these factor effects led to the use of methods originally developed to solve the blind source separation problem (Jutten and Hérault, 1991), with the aim of recovering hidden signal

sources from the observed output mixture. Independent component analysis (ICA) is one of these methods. ICA models the level of expression of each gene in a given sample as a linear weighted sum of several independent components, where each component captures the effect of one of the factors/processes. The expression data matrix is thus decomposed into a number of components, each of which is characterized by an activation pattern both across genes and across samples. The genes with the largest projection onto a component (providing the greatest contribution) are the genes the most strongly influenced by the process associated with this component. The contribution value of a sample reflects the activity of the component in this sample. Most of the studies applying ICA to cancer transcriptome data have focused on the interpretation of components based on the contribution of the genes to the components (Liebermeister, 2002; Lee and Batzoglou, 2003; Carpentier et al., 2004; Chiappetta et al., 2004; Frigyesi et al., 2006; Teschendorff et al., 2007). No study, to our knowledge, has interpreted ICA results taking into account additional data associated with the samples, such as clinical and pathological data (including histopathological images), mutations and copy number alterations, and conditions of procurement and processing of the samples.

Bladder cancer is one of the most common cancers in North America and Europe (Ferlay et al., 2010). The stage classification differentiates between non-muscle-invasive (CIS [carcinoma in situ], Ta, T1) and muscle-invasive tumors (T2, T3, T4). Gene expression clustering is commonly used to identify subtypes in cancer with the aim of gaining new insights into the molecular heterogeneity of tumors and of improving the management of cancer patients. Recently, basal and luminal subtypes of muscle-invasive bladder cancer that shared molecular features with basal and luminal breast cancers have been identified (Sjödahl et al., 2013; Choi et al., 2014; Cancer Genome Atlas Research Network, 2014; Damrauer et al., 2014; Rebouissou et al., 2014).

We show here that by applying ICA to bladder cancer transcriptome data sets, exploiting additional molecular and clinicopathological data, and comparing to ICA applied to multiple cancer types, it is possible to obtain insights into bladder carcinogenesis.

RESULTS

ICA Applied to the CIT Bladder Cancer Transcriptome Data Set: Interpretation of Components Based on Gene Projections

ICA was applied to our reference data set, which was produced in the framework of the French national Cartes d'Identité des Tumeurs (CIT) program from a series of 198 tumors and three normal samples (CIT series). These 198 tumors comprised both non-muscle-invasive tumors (n = 106) (deposited in ArrayExpress, accession number E-MTAB-1940) and muscle-invasive tumors (n = 92) (E-MTAB-1803, Rebouissou et al., 2014). Twenty independent components (ICs) were computed.

The first interpretation of each component was based on analysis of the genes with the largest projections on the component (the contributing genes, the most strongly influenced by the process associated with this component; Table S1). We investigated the association of the component with a specific cell type, known pathways, or sets of coregulated genes, by studying the enrichment of these contributing genes in predefined gene sets (such as Gene Ontology, KEGG, or BioCarta, see Supplemental Experimental Procedures). We also investigated whether these genes were located in a particular chromosomal region or could be distinguished in terms of sequence features, such as GC content.

Eleven components were associated with a biological process on the basis of enrichment analysis (Tables 1, S2 and S3, sheet "Association_ICs_genomicLocation"). Among these components, we were able to distinguish between components associated with the tumor cells and components associated with the tumor microenvironment. Three of the tumor cell-driven components were linked to various types of differentiation (squamous and basal differentiation [CIT-6], urothelial differentiation [CIT-9], neural differentiation [CIT-18]), one was associated with the cell cycle (CIT-7), one was associated with mitochondria (CIT-4), and another (CIT-10) contained six neighboring genes located at 11p15.5, including the *IGF2*/*H19* imprinted locus (Figure S1A). Three components were associated with the stroma: CIT-3 was assigned to smooth muscle, CIT-8 to the immune response mediated by B and T lymphocytes, and CIT-12 to tumor-associated myofibroblasts. Two biologically interpretable components could not be attributed to either tumor cells or the stroma (CIT-5, CIT-14). CIT-5 was enriched in interferon response genes and CIT-14 was enriched in inflammatory and early response genes. One component, CIT-2, was not significantly enriched in any particular gene set, but gene projections onto this component were strongly correlated with GC content of the corresponding genes (Figure S1B).

ICA Applied to the CIT Bladder Cancer Transcriptome Data Set: Interpretation of the Components Based on Contributions of the Tumor Samples

We further interpreted the components by studying the association of each component with specific tumor sample features by analyzing the values of sample contributions to the components (Table S3, sheet "Sample_Contributions", and Figure S1).

The samples of the CIT series were comprehensively annotated (Table S3, sheet "Sample_Annotations") for clinical and pathological data, the mutational status of the genes most commonly mutated in bladder cancer, genomic alterations assessed by CGH array, transcriptomic signatures (the carcinoma in situ signature (Dyrskjøt et al., 2004), and the multiple regional epigenetic silencing phenotype (MRES phenotype) (Vallot et al., 2011)) and technical factors.

We investigated the distributions of the tumors of the CIT series on the components as a function of these parameters (Table S3, sheet "Association_Annotation_ICs"). This approach enabled us to interpret two additional components (CIT-1 and CIT-13) and deepened our understanding of two components previously assigned based on gene contributions (CIT-14 and CIT-18).

CIT-13 was associated with the two well-known pathways of bladder tumor progression (Figure 1A). Indeed, most of the tumors displaying a negative contribution to CIT-13 were Ta

Table 1. Interpretation of the Independent Components of the CIT Bladder Cancer Data Set Using the Contributing Genes

Source	Component ID	Component Name	Contributing Genes (n)	Interpretation of the Contributing Genes
Tumor cells	CIT-6	basal-like	151	enriched in gene sets "keratinization," "keratinocyte differentiation"; markers of epithelial basal cells: KRT5, KRT6A, KRT14, KRT16, KRT17, DSC3, etc.
	CIT-4	mitochondria	61	enriched in mitochondrial genes
	CIT-7	cell cycle	187	enriched in targets of E2F and genes involved in cell cycle
				markers of proliferation: TOP2A, CDC20, CCNB2, CDK1, BUB1B, AURKA, kinesins (KIF4A, KIF2C, KIF14, KIF15, KIF11, KIF20A, KIF23, KIFC1), etc.
	CIT-9	urothelial differentiation	181	markers of urothelial differentiation: UPK1A, UPK1B, UPK2, KRT20, PPARG, BAMBI. TFs involved in differentiation: FOXA1, GATA3, GRHL3
	CIT-10	11p15.5	163	five genes colocalized at 11p15.5 (*H19*, *SYT8*, *TNNT3*, *TNNI2*, *LSP1*)
	CIT-13	bladder cancer pathways[a]	175	
	CIT-18	neuroendocrine[b]	122	enriched in neuronal genes: *CELSR2*, *CELSR3*, *ENO2*, *EMX2*, *NMU*, etc.
Stroma	CIT-3	smooth muscle	197	enriched in targets of the serum response factor SRF, and in gene sets associated with smooth muscle
				markers of the smooth muscle: calponin 1 (CNN1), synemin (SYNM), actins (ACTG2, ACTC1, ACTA2), myosins (MYL9, MYLH11), desmin (DES), transgelin (TAGLN), etc.
	CIT-8	lymphocytes B&T (LB&T)	211	enriched in gene sets associated with immune reaction
				markers of lymphocytes (immunoglobulins, chemokines, members of the MHC class II, CD79A, POU2AF1, and markers of T cells CD2, CD53, CD3D, CD8A, GZMA, GZMB, LCP2, TNFRSF17, etc.)
	CIT-12	myofibroblasts	192	enriched in gene sets associated with extracellular matrix
				markers of myofibroblasts: periostin (POSTN,) lumican (LUM), decorin (DCN), collagens (COL1A1, COL1A2, COL5A1, COL5A2, COL6A2, COL6A3, COL15A1, COL3A1, COL12A1), etc.
Tumor cells or stroma	CIT-5	interferon response (IFN)	169	enriched in targets of interferon regulatory factors IRF1, IRF2, IRF7, IRF8, and the TF STAT1. Among the first 25 contributing genes, 11 encode interferon-induced proteins *IFIT1*, *IFIT3 IFI6*, *IFI27*, *IFI35*, *IFI44*, *IFI44L*, *MX1*, *MX2*, *IFI6*, *IFIH1*. *IRF9* is also a contributing gene. The presence of *MX1* and *CXCL10* in the contributing genes indicates a type I interferon response.
Involvement of technical factors	CIT-1	batch effect[a]	147	
	CIT-2	GC content	32	the projection values of all genes are highly correlated to their GC content (Pearson correlation = 0.66)
	CIT-14	inflammation/early response/type of surgery[b]	173	enriched in genes involved in inflammatory response and response to stress
				contributing genes including chemokine genes (*CXCL1*, *CXCL2*, *CXCL3*), early response to stress genes (*EGR1*, *EGR2*, *EGR3*, *IER3*), interleukin genes (*IL8*, *IL1A*), *FOS*, *FOSB*, *DUSP1*, and *DUSP5*

See also Figure S1 and Tables S1 and S2.
[a]Assignation of components using tumor contribution.
[b]Further assignation of components using tumor contribution.

papillary tumors, frequently mutated for *FGFR3* (Figure 1B), the marker of the Ta pathway. Conversely, the tumors displaying a positive contribution to CIT-13 were mostly T2-4 tumors, with high rates of *TP53* mutation, expressing markers preferentially associated with the CIS pathway (the CIS signature and the MRES phenotype) (Figure 1B). T1 tumors were found on both sides of the component (Figure 1B). The distribution of the tumors on CIT-13 was therefore entirely consistent with the definition of the two pathways of bladder tumor progression. CIT-1 was not associated with any biological information. The traceability of processing for the CIT samples made it possible to attribute this component to a batch effect (Figure S1C).

The strongest association for component CIT-14 was with the type of surgery (transurethral resection of the bladder tumor versus cystectomy) ($p = 1.89 \times 10^{-17}$, Table S3, sheet "Association_Annotation_ICs"; Figure S1D). The enrichment of the contributing genes of CIT-14 in early response genes (Table 1) may therefore in part reflect the different stresses induced by the two different surgical procedures. The enrichment of CIT-14 in genes previously reported as differentially expressed in

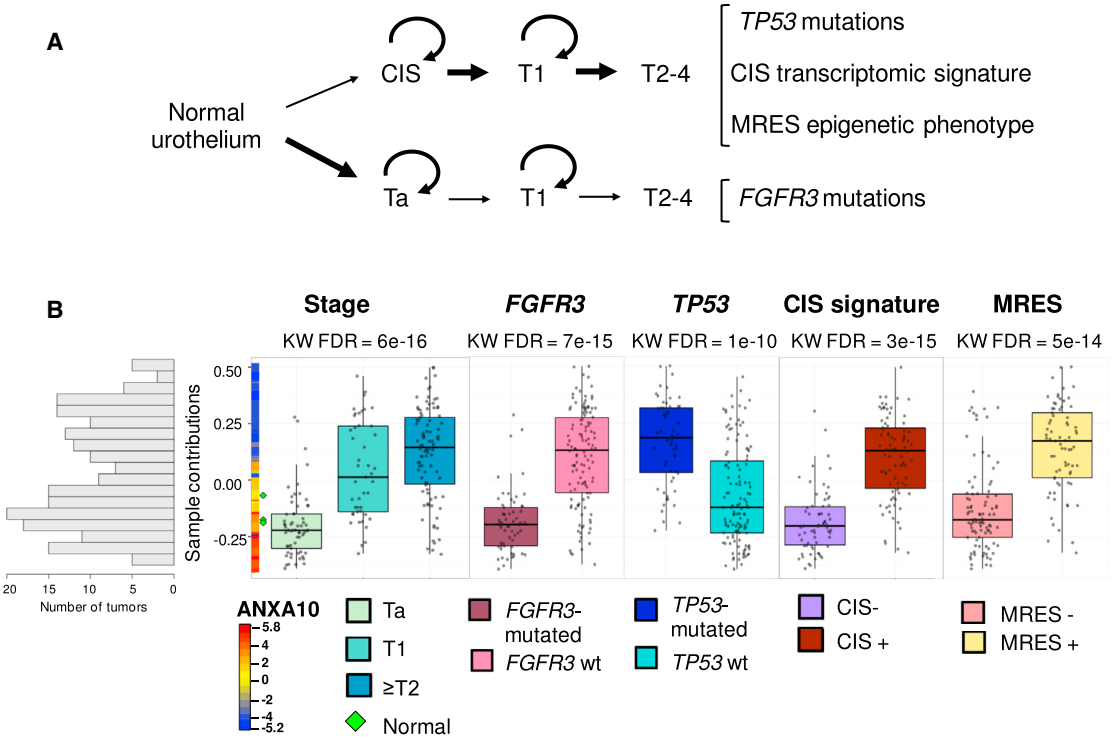


Figure 1. Association of the CIT-13 Component with the Two Pathways of Bladder Tumor Progression

(A) The two pathways of bladder tumor progression (Ta pathway and CIS pathway). Ta tumors constitute the largest group of bladder tumors (50%) at first diagnosis. Ta tumors often recur but only rarely progress via T1 tumors to muscle-invasive tumors (T2–T4 tumors). The Ta pathway is characterized by a high frequency of *FGFR3* mutations. The other pathway corresponds to the carcinoma in situ (CIS) pathway. CIS often progress (in about 50% of cases) to T1 and then to muscle-invasive tumors. This pathway is characterized by early *TP53* mutations (Spruck et al., 1994), a specific transcriptomic signature (Dyrskjøt et al., 2004), and an epigenetic phenotype (MRES, Vallot et al., 2011). For box plots, the bottom, top, and middle bands of the boxes indicate the 25th, 75th, and 50th percentiles, respectively. Whiskers extend to the most extreme data points no more than 1.5 interquartile range from the box.

(B) Distribution of the tumors on the CIT-13 component as a function of pathological and molecular characteristics. The histogram of the sample contributions to CIT-13 is shown on the left. The sample contribution values are given on the vertical axis. On this axis is also shown the expression level of one of the strongest contributing genes (*ANXA10*) of the IC, each point representing one tumor sample, and its color indicating the relative level of expression of the gene. The color scale for the expression level of *ANXA10* is shown below. The box plots and superimposed individual data points show the distribution of the samples on CIT-13 as a function of stage (Ta, T1, and tumors of higher stage: T2–T4), *FGFR3* and *TP53* mutation status, the CIS transcriptomic signature (Dyrskjøt et al., 2004), and the multiple regional epigenetic silencing (MRES) phenotype (Vallot et al., 2011). The normal samples are represented as green diamonds. False discovery rates (FDRs) for Kruskal-Wallis tests are provided at the top of each plot.

See also Table S3.

prostate cancers as a function of sample type (biopsy versus prostatectomy specimens) (Lin et al., 2006) supports this conclusion (Table S2, sheet 1). We cannot exclude the possibility that a biological process not directly related to the type of surgery may also contribute to this component, which we refer to as the inflammation/early response/type of surgery. The CIT-18 component, which was enriched in neural genes (Table 1), was due to the presence of neuroendocrine tumors in the CIT series (Figure S1E).

Representative pathological sections were available for the CIT tumor series. For each component, the sections were ranked according to the contribution of the corresponding tumor sample to the component (http://microarrays.curie.fr/publications/UMR144/LuminalBasalBladder/) and were reanalyzed in this context. This confirmed the assignment of components CIT-3, CIT-12, and CIT-7 to smooth muscle, myofibroblasts, and cell cycle, respectively (Figure S2).

ICA of Multiple Transcriptome Data Sets, Bladder, and Other Types of Cancer Allows Distinguishing Reproducible Cancer Type-Specific Components and Components Common to Different Cancer Types

We investigated the reproducibility of the components retrieved from our reference bladder tumor set (CIT series), by applying ICA to seven other bladder cancer transcriptome data sets, six series of breast carcinomas and eight series of carcinomas of various types (colon, rectum, endometrium, kidney, lung squamous cell and lung adenocarcinoma, ovary, prostate). The data sets, their accession numbers, and gene projections for all data sets are given in Table S4.

We used a correlation graph approach to compare the ICs across the different data sets (see Supplemental Experimental Procedures) (Figure 2). In the correlation graph, each node represents a component computed for a data set, and each edge connects two components if the corresponding gene projections

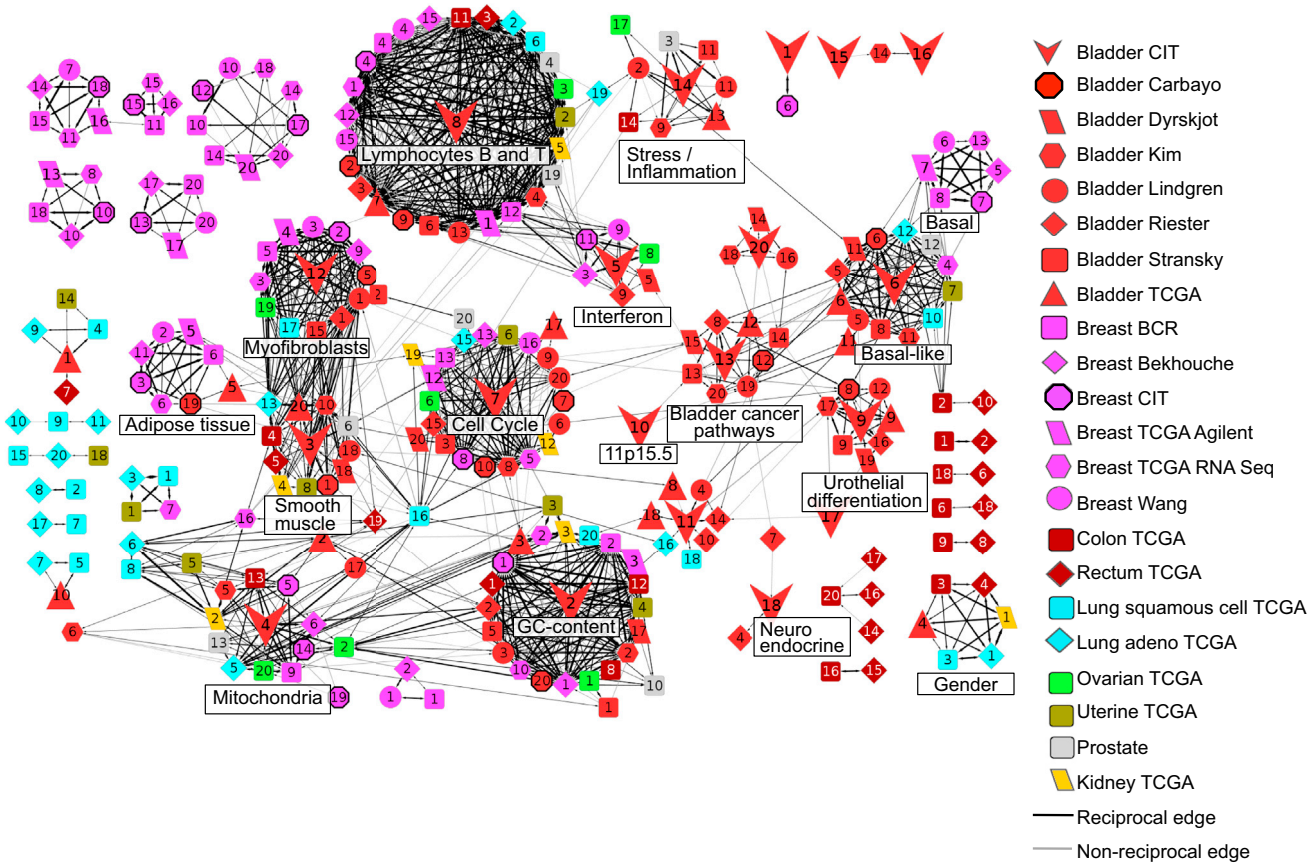


Figure 2. Comparison of the Independent Components across Data Sets of Various Cancer Types Using Correlation-Based Graphs

Correlation-based graph of the independent components (ICs) calculated for seven bladder cancer data sets plus 14 other gene expression data sets of different types of cancer, including six breast carcinoma data sets and one data set for each of the following cancer types: prostate, colon, and rectum adenocarcinomas, uterine corpus endometrioid carcinoma, kidney (clear cell carcinomas), lung (squamous cell carcinomas and adenocarcinomas), and ovarian serous adenocarcinomas. Each node denotes an IC; the node colors indicate the type of cancer, whereas their shapes distinguish between data sets of the same cancer type. The numerical IDs indicate the order in which the ICs were returned by the icasso method. The graphs were visualized with Cytoscape (Cline et al., 2007). The edge thickness indicates the degree of correlation between the two ICs linked by the edge. Only pairs of ICs with absolute correlation values exceeding 0.35 are shown. Reciprocal edges are colored black, whereas non-reciprocal edges are in gray. Reproducible components appear on the graph as a cluster of tightly interconnected nodes. See also Tables S4 and S5.

are correlated. Clusters of components (pseudocliques, see Supplemental Experimental Procedures) were observed in the structure of the correlation graph (the genes contributing to the components of each clique are given in Table S5). Some clusters were composed of densely interconnected nodes and included components common to most types of cancer (immune response mediated by B and T lymphocytes, cell cycle, GC content, myofibroblasts/smooth muscle). Other clusters were common to some, but not all types of cancer (including gender, basal-like, mitochondria, interferon response, inflammation/early response genes/type of surgery-associated clusters). Because we used several data sets for bladder and breast cancers, we were able to identify pseudocliques specific to these cancers. Two bladder-cancer-specific pseudocliques consisted of components found in all bladder cancer data sets: one was associated with urothelial differentiation, whereas the other was associated with the two pathways of bladder tumor progression.

Association of the Recently Identified Expression-Based Subtypes of Muscle-Invasive Bladder Cancer with Biological Processes

Muscle-invasive bladder cancers are clinically and biologically heterogeneous. Recently, expression-based clusters that resembled basal and luminal subtypes of breast cancer have been reported in muscle-invasive bladder cancer (Sjödahl et al., 2013; Cancer Genome Atlas Research Network, 2014; Choi et al., 2014; Damrauer et al., 2014; Rebouissou et al., 2014). To characterize these muscle-invasive bladder cancer subtypes, we compared their distribution on the different components, which had been interpreted in the CIT series. We studied the four subtypes from the TCGA classification (Cancer Genome Atlas Research Network, 2014): cluster I (papillary luminal), cluster II (luminal), cluster III (basal-like), and cluster IV. These four clusters were identified in the CIT series using a centroid-based predictor (see Supplemental Experimental Procedures). The six components displaying the most significant

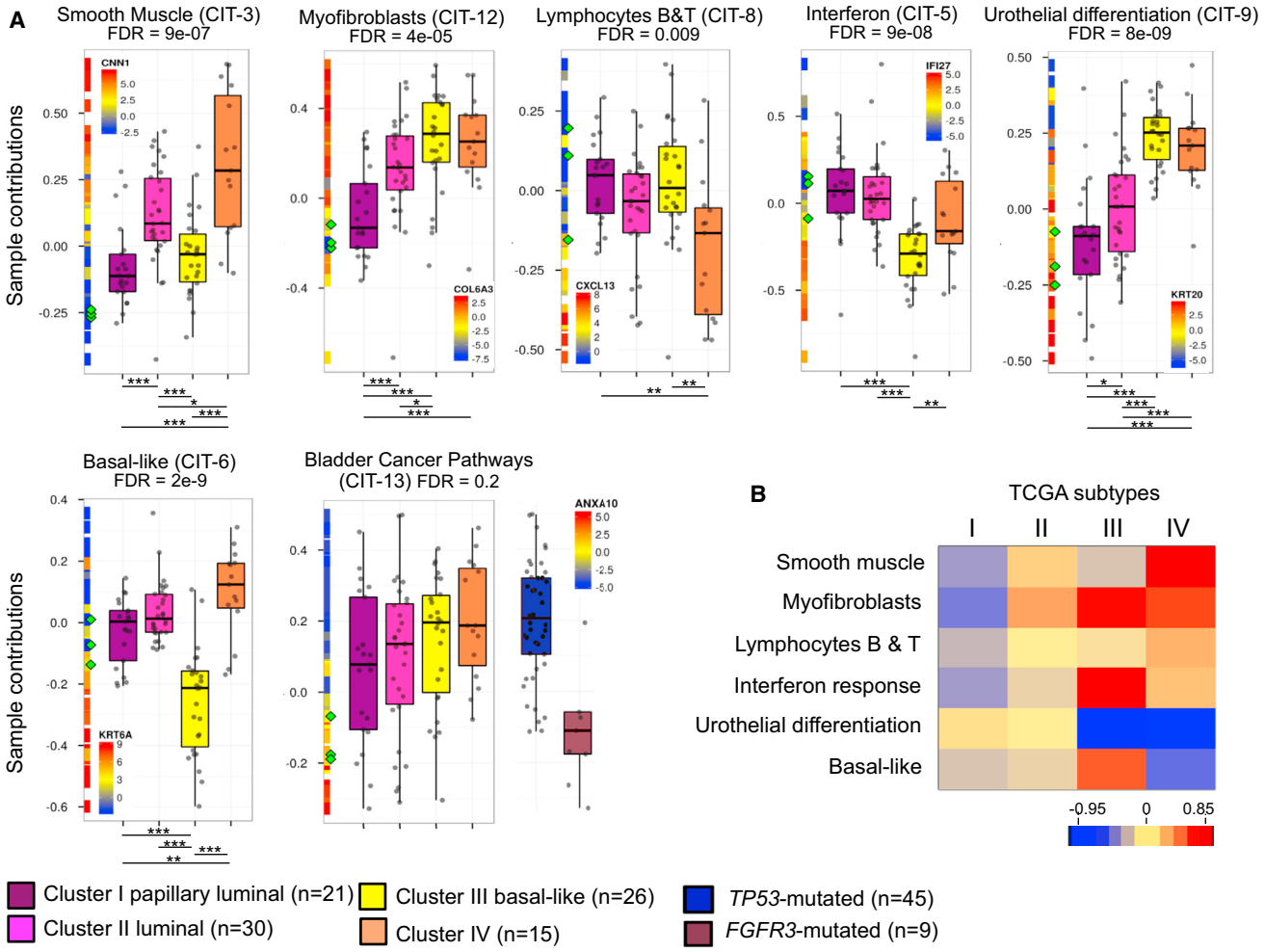


Figure 3. Characterization in the CIT Series of the Four Gene Expression Subtypes as Defined in the Bladder Cancer TCGA 2014 Classification

Cluster I (papillary luminal, in dark magenta), cluster II (luminal, in magenta), cluster III (basal-like, in yellow), and cluster IV (orange). The six components displaying the most significant differences between these subtypes as well as the progression pathway component are shown.

(A) Each box plot represents the distribution of a subgroup of tumors on the IC and is overlaid with the individual data points. The vertical lines at the left of the box plots indicate the expression level of one of the strongest contributing genes of the IC, each point representing one tumor sample, and its color indicating the relative level of expression of the gene. For box plots, the bottom, top, and middle bands of the boxes indicate the 25th, 75th, and 50th percentiles, respectively. Whiskers extend to the most extreme data points no more than 1.5 interquartile range from the box. The diamond-shaped green points indicate the positions of the normal samples. FDR values for Kruskal-Wallis tests are provided at the top of each plot. The p values (Wilcoxon test) for each pairwise comparison are shown at the bottom of each plot: *$0.01 \leq p < 0.05$; **$0.001 \leq p < 0.01$; ***$p < 0.001$.

(B) Heatmap representing the median of the sample contributions of each subtype to the components. For the purpose of this figure, the vector of sample contributions of some components (lymphocytes, interferon response, urothelial differentiation, basal-like) has been reversed such that the over- (respectively, under-) expression of the most contributing genes of the components are on the positive (respectively, negative) side of the component and thus appear in red (respectively, blue) color.

See also Figure S2 and Table S3.

differences between these clusters and the progression pathway component are shown in Figure 3A. As expected, basal-like tumors presented a basal differentiation (their distribution on the basal-like component [CIT-6] corresponded to high expression of basal markers, such as KRT6A), and a low level of urothelial differentiation (their distribution on the urothelial differentiation component (CIT-9) corresponded to low expression of urothelial differentiation markers, such as KRT20). We also showed that this subtype presented a strong interferon response (as shown by its distribution on the interferon-response component [CIT-5]). The distribution of basal-like tumors on the component associated with the bladder cancer pathways (CIT-13) was similar to that of *TP53*-mutated tumors and different from that of *FGFR3*-mutated tumors, suggesting that basal-like tumors belonged to the CIS pathway (see diagram in Figure 1A). Their distribution on the components associated with stroma (CIT-3 [smooth muscle component], CIT-8 [lymphocytes T and B component], and CIT-12 [myofibroblast component]) indicated

that basal-like tumors were relatively poor in smooth muscle cells and lymphocytes T and B and relatively rich in myofibroblasts (strong expression of myofibroblast markers). The two luminal subtypes (clusters I and II) presented a high level of differentiation. Indeed, their distribution on the urothelial differentiation component (CIT-9) corresponded to high expression of urothelial differentiation markers, such as KRT20. These two subtypes differed in terms of stroma, cluster II was richer in smooth muscle (CIT-3) and myofibroblasts (CIT-12). A schematic representation summarizing the characteristics of each subtype according to the components is shown in Figure 3B.

Genetic and Functional Evidence for a Role of PPARG in Urothelial Differentiation and in Bladder Carcinogenesis

The urothelial differentiation component (CIT-9), found in all bladder cancer data sets studied, was specifically associated with bladder luminal tumors. We studied this component in more detail.

Most luminal tumors (subtypes I and II) were distributed on the differentiated side of the component (high expression of differentiation markers such as *KRT20*, *GRHL3*, *UPK1A*, *BAMBI*, and *PPARG*) (Figure 4A). As expected, Ta low-grade tumors (G1/G2) were on the differentiated side of the component (Figure 4A) and were differentiated at the morphological level (Figure 4B; http://microarrays.curie.fr/publications/UMR144/LuminalBasalBladder/). Surprisingly the tumors with the highest levels of urothelial differentiation markers (most negative contributions to the component) were high-grade tumors (G3) (Figure 4A). By morphology, these tumors both lacked architectural organization and displayed unpolarized and atypical nuclei (Figure 4B; http://microarrays.curie.fr/publications/UMR144/LuminalBasalBladder/). These observations indicate that there was no relationship between molecular differentiation measured on this component and morphological grade assessed by architectural organization and nuclear atypia.

To identify protumorigenic genes potentially involved in the carcinogenesis of luminal tumors, we looked for altered genomic regions associated with the urothelial differentiation component. Among these regions, we selected the regions of gains that were found in tumors associated with the differentiated side of the component and that contained a contributing gene associated with differentiation. In the CIT series, a region of genomic alteration (chr3:1650731-12883379) encompassing *PPARG*, a contributing gene associated with urothelial differentiation (Table S6) was the most significantly associated with the component. The alterations observed in this region were mostly gains (in 51 of the 178 tumors, 29%; with losses observed in seven of the 178 tumors, 4%) (Figure 4C). Computation of the minimal region of gain (Rouveirol et al., 2006) (see Supplemental Experimental Procedures) identified a 774 kb region containing nine genes, including *PPARG*. These genomic alterations of *PPARG* were correlated with its expression in both tumors and bladder tumor-derived cell lines (Figure S3A). The use of high-resolution SNP arrays (Illumina Human 610) for 11 tumors presenting a gain of this region made it possible to narrow down the minimal region of gain to 273.9 kb including only *PPARG*. The distribution of the tumors of the TCGA series on the urothelial differentiation component according to genomic alterations confirmed that *PPARG* gains were associated with differentiated tumors (Figure S3B). These findings provide genetic evidence that PPARG was involved in tumorigenesis of these tumors. Consistent with this, pathway enrichment analysis of the contributing genes of the urothelial differentiation component showed enrichment in genes of the PPAR pathway: *PPARG*, *FABP4*, *HMGCS2*, *ACSL5*, *ACOX1*, *APOA2*, and *ACOX3* (Table S2, sheet "CIT9-GOstat"). To demonstrate the functional involvement of PPARG in bladder tumors, we studied the effect of small interfering RNA (siRNA)-mediated *PPARG* knockdown on the growth of nine bladder-cancer-derived cell lines with different levels of *PPARG* mRNA. Two different siRNAs targeting *PPARG* mRNA were used, and a downregulation of 65%–90% of *PPARG* mRNA was obtained (Figure S3C). The largest decrease in cell viability (50%) was observed for the two cell lines presenting a genomic alteration of *PPARG* (amplification for UMUC9 and gain for SD48) (Figure 4D, left panel). Overall, *PPARG* mRNA level was significantly correlated with the effect of siRNA-mediated *PPARG* knockdown, because higher levels of *PPARG* expression were associated with larger decreases in cell viability (Figure 4D, right panel). We also investigated the effect of *PPARG* knockdown on the cloning efficiency of three cell lines (SD48, UMUC9, and MGHU3). The inhibition of growth in the soft agar assay was similar (60% for cell lines with *PPARG* genomic alteration, SD48, and UMUC9, and 20% for MGHU3) to the decrease in cell viability in 2D culture (Figure 4E). These results indicated that PPARG was involved in tumor cell growth in PPARG expressing bladder cancer cells.

PPARG is a transcription factor of the nuclear receptor type 1 subfamily and has been implicated in urothelial differentiation (Varley et al., 2004). The presence of PPAR target genes among the contributing genes of the urothelial differentiation component suggested that PPARG could control the expression of several contributing genes of this component. We tested this hypothesis by comparing the transcriptome of the SD48 cell line treated with three different siRNAs targeting *PPARG* or with the transfection reagent (Lipofectamine). We found that 198 genes were significantly deregulated by siRNA treatment (107 genes significantly downregulated and 91 genes upregulated) (Table S7). This set of deregulated genes was significantly enriched (Fisher's test, $p < 0.0001$) in genes contributing to the CIT urothelial differentiation component (21 of 198; Figure 4F; Table S7). Most of the 21 genes of the urothelial differentiation component controlled by *PPARG* were upregulated by this transcription factor (downregulation observed upon *PPARG* knockdown for 18 of these 21 genes) and, as expected, were associated with differentiation (17 of these 18 genes). Conversely, two of the three genes downregulated by *PPARG* (upregulated following *PPARG* knockdown) were associated with undifferentiated tumors.

DISCUSSION

We applied independent component analysis (ICA) to 21 transcriptome data sets, including eight data sets for bladder cancer and five for breast cancer. For one of the bladder cancer data sets (the CIT data set that was generated by our group), clinical, pathological, molecular, and sample processing data were available. The use of these additional data made it possible to assign

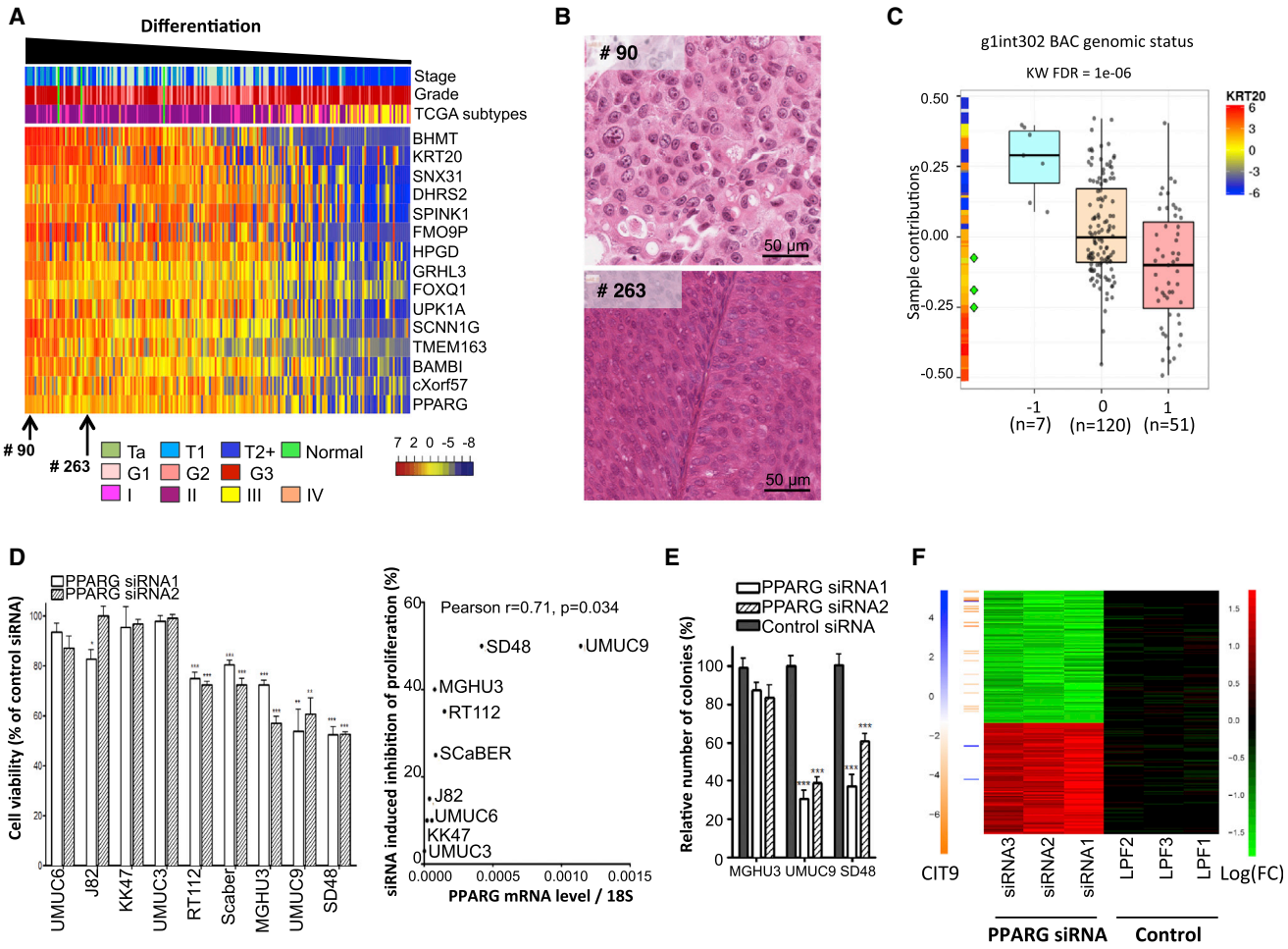


Figure 4. Identification of *PPARG* as an Oncogene Controlling Cell Viability and Differentiation in Bladder Tumors

(A) Heatmap of the top 15 contributing genes of CIT-9; the genes (rows) and samples (columns) are ranked according to their contribution to the component. The gene expression profiles are centered. These top 15 most contributing genes include several urothelial differentiation markers (*KRT20*, *GRHL3*, *UPK1A*, *BAMBI*, *PPARG*). The stage, grade, and TCGA tumor subtype classification (I, papillary luminal; II, luminal; III, basal-like; IV) are indicated.

(B) Histopathological sections of two tumors on the differentiated side of the component (high expression of urothelial differentiation markers). Case #90, a T1G3 tumor, despite having one of the highest levels of urothelial differentiation markers lacked any architectural organization and displayed unpolarized and atypical nuclei. Case #263, a TaG2 tumor, which was also on the differentiated side of the component, was differentiated at the morphological level.

(C) Distribution of the tumor samples according to their alteration status for the region on chromosome 3:1650731-12883379 encompassing *PPARG*, as measured by aCGH. FDR value for the Kruskal-Wallis test is provided at the top of the box plot. For box plots, the bottom, top, and middle bands of the boxes indicate the 25th, 75th, and 50th percentiles, respectively. Whiskers extend to the most extreme data points no more than 1.5 interquartile range from the box.

(D) Left: effect of PPARG knockdown on cell viability, as assessed by MTT assays, of nine bladder-cancer-derived cell lines. Two different siRNAs were used. Right: correlation between cell viability inhibition induced by *PPARG* knockdown (mean of the two different siRNAs) and *PPARG* expression.

(E) Effect of *PPARG* knockdown on anchorage-independent colony formation in soft agar. The values reported are the means ± SD of three independent experiments carried out in triplicate.

(F) Heatmap visualization of the 198 regulated transcripts in SD48 cells 96 hr after *PPARG* knockdown by three different siRNAs. The expression ratio (FC) was determined for each gene by comparison of the mean level of expression on lipofectamine control (LPF) chips with that on *PPARG* siRNA chips. The gene expression profiles are centered on the mean level of expression on control chips. For regulated genes contributing to the CIT-9 component with an absolute value for the contribution score exceeding 3 (n = 21), the contribution score is indicated according to the color scale on the left side of the heatmap.

See also Figure S3 and Tables S3, S6, and S7.

several components to a biological process or to technical factors and allowed us to characterize subtypes of tumors and to identify molecular mechanisms associated with a component.

Our analysis across different types of cancer identified both components common to some or all cancer types considered, and also components specific to a particular type of cancer. Three components common to all cancer types (B and T lymphocytes, cell cycle, myofibroblasts) probably correspond to the three multicancer gene signatures/attractors (lymphocyte-specific, mitotic chromosomal instability, and mesenchymal transition) identified very recently in breast, colon, and ovarian cancer by a different computational approach (Cheng et al., 2013).

Indeed, many of the genes associated with these components were also found to be markers for these signatures. The smooth muscle and myofibroblasts components were merged in some data sets (the TCGA and Kim et al., 2010 bladder data sets and in the lung adenocarcinoma data set), indicating that the expression of these two stroma components are closely related and are often present in the same samples for these data sets.

The components common to some but not all cancer types included the basal-like component. This component was found in bladder cancer, lung squamous carcinoma, lung adenocarcinoma, and endometrioid carcinoma and was related to a component specific to breast cancer, the basal component. The existence of a basal subgroup in breast cancer is widely recognized (Bertucci et al., 2012; Sørlie et al., 2001). It is becoming clear that basal-like subgroups also exist in other cancer types (Chung et al., 2004; Wilkerson et al., 2010; Di Palma et al., 2012), including, as shown very recently, bladder cancer (Sjödahl et al., 2013; Cancer Genome Atlas Research Network, 2014; Choi et al., 2014; Damrauer et al., 2014; Rebouissou et al., 2014).

TCGA results have shown that most colon and rectum adenocarcinomas cannot be distinguished at the genomic level (Cancer Genome Atlas Network, 2012). Consistent with this observation, adenocarcinomas of the colon and the rectum shared most of their components in our analysis.

The biologically meaningful ICs we identified included four related to tumor microenvironment. A detailed interpretation of these components should help to characterize the tumor microenvironment and the reciprocal interactions between cancer cells and stromal cells.

Luminal tumors differed from the other bladder tumors by their activities on different components: low interferon response and the presence of molecular urothelial differentiation. Surprisingly, molecular urothelial differentiation was maintained even in those luminal tumors that had lost morphological differentiation, as assessed by cytological atypia and the absence of architectural organization. This suggests a dependency on part of the differentiation program for tumor progression. We demonstrated that the transcription factor PPARG positively controls the expression of genes linked to molecular differentiation in tumor cells. In agreement with this, Choi et al. (2014) have recently shown a PPARG activation signature in luminal tumors. In addition, we provide here genetic and functional evidence that this gene is involved in tumorigenesis of molecularly differentiated bladder tumors. Interestingly, the use of pioglitazone, an antitype 2 diabetic drug that is an agonist of PPARG, has been associated with an increased risk of bladder cancer (Turner et al., 2014). This could be due to the protumorigenic role of PPARG, as shown here. A parallel can be drawn between PPARG and another nuclear receptor also associated with differentiation, the estrogen receptor. Indeed, the estrogen receptor is associated with differentiation in the breast, which is necessary for the growth of estrogen receptor-positive tumors, and the use of estrogen increases the risk of breast cancer (Collaborative Group on Hormonal Factors in Breast Cancer, 1997).

There are several advantages of searching for associations between expression and copy number through ICA, as we did here to find *PPARG*, rather than using a classical search for correlation between expression and copy number. The association between a genomic alteration and a component can be directly interpreted in the context of a biological process (represented by the component) and of multiple genes associated with this biological process (the most contributing genes on the component), rather than simply associating one genomic alteration with one gene. Moreover, it reduces the high number of candidates that can be obtained when analyzing the genomic alterations alone. We also found that genomic gains of *GATA3* were associated with the urothelial differentiation component (data not shown). GATA3 is important for the proliferation of luminal breast tumors (Kong et al., 2011) and is therefore likely to play the same role in bladder tumors presenting a molecular urothelial differentiation.

We have developed an R package (*MineICA*, available in Bioconductor: http://www.bioconductor.org/packages/release/bioc/html/MineICA.html) that implements most of the methods used in our study. We also provide a bladder tumor data set that contains not only molecular and anatomoclinical data but also histological data. Integrating image data into multiparametric analyses of tumors is an emerging domain of research. The resource we provide should be useful in this respect.

Our analysis shows that applying ICA to transcriptome data and exploiting clinical, pathological, and genetic alteration data obtained from ongoing large-scale projects will provide insights into cancer biology.

EXPERIMENTAL PROCEDURES

CIT Series: Tissues Samples and Data

Tissues samples of human bladder carcinomas were collected from patients at Henri Mondor Hospital (Créteil, France), Institut Gustave Roussy (Villejuif, France), and Foch Hospital (Suresnes, France). Details of samples and large-scale data are provided in Supplemental Experimental Procedures. All patients provided written informed consent, and the study was approved by the ethics committees of the different hospitals involved in the study (Comité de Protection des Personnes de l'hôpital Henri Mondor, Comité de Protection des Personnes de Boulogne - Ambroise Paré and Comité de Protection des Personnes de Bicêtre). All analyses were performed on the basis of anonymized patient data.

Bioinformatics Analysis

To compute the independent components, we applied the FastICA algorithm (Hyvärinen, 1999) to each data set using *icasso* (Himberg et al., 2004) for selecting stable components. We computed 20 independent components (ICs), using *pow3* nonlinearity and *deflationary* algorithm. We interpreted the genes associated with an IC ("most contributing genes") by performing enrichment analysis with GOstats (Falcon and Gentleman, 2007) and "GSEA Preranked" module (Subramanian et al., 2005). We compared the contributions of predefined groups of tumors to the ICs in Wilcoxon rank-sum and Kruskal-Wallis tests. The ICs were compared across data sets using correlation-based graphs.

Details of computational analyses are provided in Supplemental Experimental Procedures.

Experiments

To study the role of PPARG in bladder-cancer-derived cell lines, we evaluated the effect of PPARG knockdown using siRNAs on cell viability of nine cell lines (J82, KK47, MGHU3, SCaBER, SD48, RT112, UMUC3, UMUC6, UMUC9) and on colony formation in soft agar for three of these cell lines (MGHU3, SD48, UMUC9). PPARG mRNA expression level was determined using quantitative RT-PCR. Genes regulated after PPARG knockdown were determined using Affymetrix DNA microarray. Details of experimental procedures and statistical analysis of the results are provided in Supplemental Experimental Procedures.

ACCESSION NUMBERS

The ArrayExpress accession number for the CIT non-muscle-invasive tumor array data reported in this paper is E-MTAB-1940.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.10.035.

AUTHOR CONTRIBUTIONS

A.B., A.Z., and E.B. designed the computational methodology. A.B. assembled the data and performed the ICA analyses. A.B. implemented tools for data visualization and interpretation in cooperation with F.R. P.G. and E.B. advised on processing the data and tools implementation. A.B., F.R., and A.Z. were responsible for the interpretation of the independent components, helped by I.B.-P., S.R., and J.S. L.G. participated in TCGA data analysis and in the analysis of correlation graph. A.B., A.K., A.Z., and A.d.R. compared CIT and TCGA data. I.B.-P. designed and was responsible for the functional validation studies. C.K., Y.L., Y.N., and I.B.-P. performed the experimental validation studies, and A.K. was involved in their analysis. Y.A., E.C., and I.B.-P. prepared and analyzed histopathological data. Y.A. was responsible for the anatomopathological data and image interpretation, T.L. and S.B. for the clinical data, and C.R.-P., N.L.-P., Y.N., and I.B.-P. for discussing the therapeutic applications of the work. A.B., I.B.-P., F.R., and A.Z. wrote the manuscript. A.Z. supervised the computational work. F.R. coordinated the work. All authors discussed the results and implications and reviewed the manuscript.

ACKNOWLEDGMENTS

This work is part of the "Cartes d'Identité des Tumeurs" (CIT) program funded and developed by the "Ligue Nationale contre le Cancer" (LNCC) (http://cit.ligue-cancer.net). We thank E. Voirin, N. Servant, G. Lucotte, and P. Hupé for their help with bioinformatics data management and analysis. We thank members of the bladder cancer CIT consortium (P. Maillé and D. Vordos, Henri Mondor Hospital; M. Sibony, Cochin Hospital; A. Laplanche, IGR, INSERM; Y. Denoux and V. Molinié, Foch Hospital; E. Letouzé, LNCC) for their constant support. This work was supported by the LNCC (to "Oncologie Moléculaire" and "Computational Systems Biology of Cancer" accredited teams), the Institut Curie (to F.R., E.B., A.Z.), the "Centre National de la Recherche Scientifique" (CNRS) (to F.R.), the "Institut National de la Santé et de la Recherche Médicale" (INSERM) (to E.B., A.Z., S.B., and Y.A.), the INCa (INCa_2960 and 4382 to F.R. and Y.A.), ITMO cancer, systems biology program (to A.Z., E.B., and F.R.), the Labex (no. ANR-10-LBX-0038) part of the IDEX PSL (no. ANR-10-IDEX-0001-02 PSL) (to F.R.), and York Against Cancer (to J.S.). A.B. was supported by a grant from the INCa, from the LNCC, and by NIH grant 5U24 CA143799-04 as part of TCGA project, Y.L. by a grant from the "Fondation Franco-*Chinoise* pour la Science et ses Applications" (FFCSA), Y.N. by a grant from the "Fondation ARC pour la recherche sur le cancer," and A.K. by a grant from the LNCC. The results published here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Received: January 25, 2014
Revised: August 18, 2014
Accepted: October 13, 2014
Published: November 13, 2014

REFERENCES

Bertucci, F., Finetti, P., and Birnbaum, D. (2012). Basal breast cancer: a complex and deadly molecular subtype. Curr. Mol. Med. *12*, 96–110.

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315–322.

Carpentier, A.-S., Riva, A., Tisseur, P., Didier, G., and Hénaut, A. (2004). The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. Comput. Biol. Chem. *28*, 3–10.

Cheng, W.-Y., Ou Yang, T.-H., and Anastassiou, D. (2013). Biomolecular events in cancer revealed by attractor metagenes. PLoS Comput. Biol. *9*, e1002920.

Chiappetta, P., Roubaud, M.C., and Torrésani, B. (2004). Blind source separation and the analysis of microarray data. J. Comput. Biol. *11*, 1090–1109.

Choi, W., Porten, S., Kim, S., Willis, D., Plimack, E.R., Hoffman-Censits, J., Roth, B., Cheng, T., Tran, M., Lee, I.L., et al. (2014). Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer Cell *25*, 152–165.

Chung, C.H., Parker, J.S., Karaca, G., Wu, J., Funkhouser, W.K., Moore, D., Butterfoss, D., Xiang, D., Zanation, A., Yin, X., et al. (2004). Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell *5*, 489–500.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. *2*, 2366–2382.

Collaborative Group on Hormonal Factors in Breast Cancer (1997). Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. Lancet *350*, 1047–1059.

Damrauer, J.S., Hoadley, K.A., Chism, D.D., Fan, C., Tiganelli, C.J., Wobker, S.E., Yeh, J.J., Milowsky, M.I., Iyer, G., Parker, J.S., and Kim, W.Y. (2014). Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. Proc. Natl. Acad. Sci. USA *111*, 3110–3115.

Di Palma, S., Simpson, R.H.W., Marchiò, C., Skálová, A., Ungari, M., Sandison, A., Whitaker, S., Parry, S., and Reis-Filho, J.S. (2012). Salivary duct carcinomas can be classified into luminal androgen receptor-positive, HER2 and basal-like phenotypes. Histopathology *61*, 629–643.

Dyrskjøt, L., Kruhøffer, M., Thykjaer, T., Marcussen, N., Jensen, J.L., Møller, K., and Ørntoft, T.F. (2004). Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. Cancer Res. *64*, 4040–4048.

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics *23*, 257–258.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., and Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer *127*, 2893–2917.

Frigyesi, A., Veerla, S., Lindgren, D., and Höglund, M. (2006). Independent component analysis reveals new and biologically significant structures in micro array data. BMC Bioinformatics *7*, 290.

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. Neuroimage *22*, 1214–1222.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. *10*, 626–634.

Jutten, C., and Hérault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Process. *24*, 1–10.

Kim, W.J., Kim, E.J., Kim, S.K., Kim, Y.J., Ha, Y.S., Jeong, P., Kim, M.J., Yun, S.J., Lee, K.M., Moon, S.K., et al. (2010). Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. Mol. Cancer *9*, 3.

Kong, S.L., Li, G., Loh, S.L., Sung, W.K., and Liu, E.T. (2011). Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. Mol. Syst. Biol. *7*, 526.

Lee, S.-I., and Batzoglou, S. (2003). Application of independent component analysis to microarrays. Genome Biol. *4*, R76.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. Bioinformatics *18*, 51–60.

Lin, D.W., Coleman, I.M., Hawley, S., Huang, C.Y., Dumpit, R., Gifford, D., Kezele, P., Hung, H., Knudsen, B.S., Kristal, A.R., and Nelson, P.S. (2006). Influence of surgical manipulation on prostate gene expression: implications for molecular correlates of treatment effects and disease prognosis. J. Clin. Oncol. *24*, 3763–3770.

Rebouissou, S., Bernard-Pierrot, I., de Reyniès, A., Lepage, M.L., Krucker, C., Chapeaublanc, E., Hérault, A., Kamoun, A., Caillault, A., Letouzé, E., et al. (2014). EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype. Sci Transl Med *6*, 244ra91.

Rouveirol, C., Stransky, N., Hupé, P., Rosa, P.L., Viara, E., Barillot, E., and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. Bioinformatics *22*, 849–856.

Sjödahl, G., Lövgren, K., Lauss, M., Patschan, O., Gudjonsson, S., Chebil, G., Aine, M., Eriksson, P., Månsson, W., Lindgren, D., et al. (2013). Toward a molecular pathologic classification of urothelial carcinoma. Am. J. Pathol. *183*, 681–691.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*, 10869–10874.

Spruck, C.H., 3rd, Ohneseit, P.F., Gonzalez-Zulueta, M., Esrig, D., Miyao, N., Tsai, Y.C., Lerner, S.P., Schmütte, C., Yang, A.S., Cote, R., et al. (1994). Two molecular pathways to transitional cell carcinoma of the bladder. Cancer Res. *54*, 784–788.

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

Teschendorff, A.E., Journée, M., Absil, P.A., Sepulchre, R., and Caldas, C. (2007). Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput. Biol. *3*, e161.

Turner, R.M., Kwok, C.S., Chen-Turner, C., Maduakor, C.A., Singh, S., and Loke, Y.K. (2014). Thiazolidinediones and associated risk of bladder cancer: a systematic review and meta-analysis. Br. J. Clin. Pharmacol. *78*, 258–273.

Vallot, C., Stransky, N., Bernard-Pierrot, I., Hérault, A., Zucman-Rossi, J., Chapeaublanc, E., Vordos, D., Laplanche, A., Benhamou, S., Lebret, T., et al. (2011). A novel epigenetic phenotype associated with the most aggressive pathway of bladder tumor progression. J. Natl. Cancer Inst. *103*, 47–60.

Varley, C.L., Stahlschmidt, J., Lee, W.C., Holder, J., Diggle, C., Selby, P.J., Trejdosiewicz, L.K., and Southgate, J. (2004). Role of PPARgamma and EGFR signalling in the urothelial terminal differentiation programme. J. Cell Sci. *117*, 2029–2036.

Wilkerson, M.D., Yin, X., Hoadley, K.A., Liu, Y., Hayward, M.C., Cabanski, C.R., Muldrew, K., Miller, C.R., Randell, S.H., Socinski, M.A., et al. (2010). Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin. Cancer Res. *16*, 4864–4875.

Cell Reports, Volume *9*

Supplemental Information

Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes

Anne Biton, Isabelle Bernard-Pierrot, Yinjun Lou, Clémentine Krucker, Elodie Chapeaublanc, Carlota Rubio-Pérez, Nuria López-Bigas, Aurélie Kamoun, Yann Neuzillet, Pierre Gestraud, Luca Grieco, Sandra Rebouissou, Aurélien de Reyniès, Simone Benhamou, Thierry Lebret, Jennifer Southgate, Emmanuel Barillot, Yves Allory, Andrei Zinovyev, and François Radvanyi

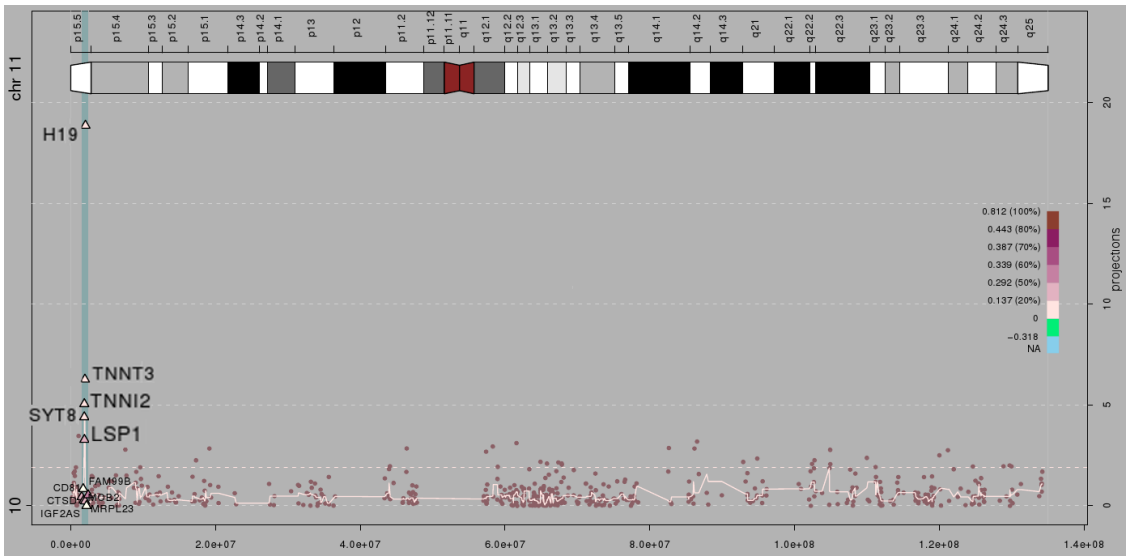
TABLE OF CONTENTS

SUPPLEMENTAL FIGURES

Figure S1, related to Table 1: Interpretation of five independent components from the CIT dataset.

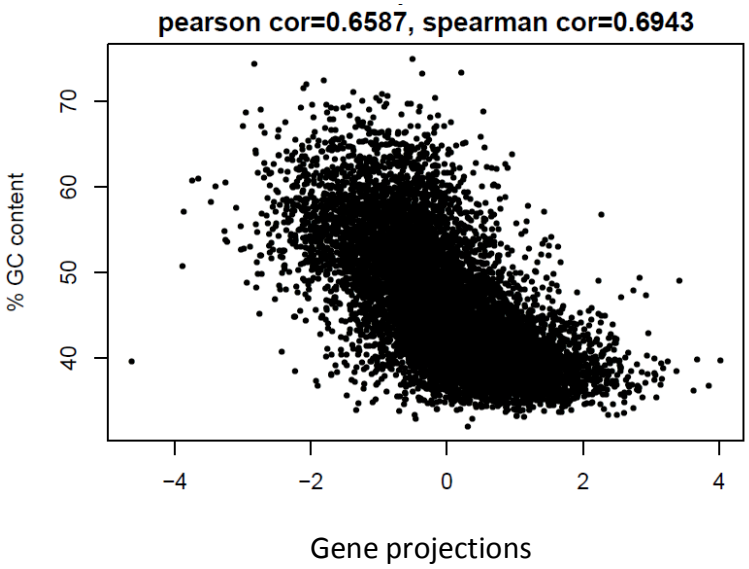
a) Cluster of genes contributing to the component CIT-10, co-localized at 11p15-5 genomic location

Each point represents a gene. The y-axis corresponds to the absolute projection value of the gene on CIT-10, whereas the x-axis corresponds to its genomic position. The solid line indicates the score (median of the projections) obtained for a window centered on each gene. Each window includes the three neighboring genes on each side of the gene. The pink dotted line (second dotted line from the bottom of the graph) indicates the threshold corresponding to the .999th percentile of the null distribution obtained by random gene position permutations. The color of the genes within the peak (triangles) reflects the correlation between their expression and their genome values (aCGH). The color scale is shown to the right of the plot. Here, the genes in the peak are not correlated with their copy numbers.



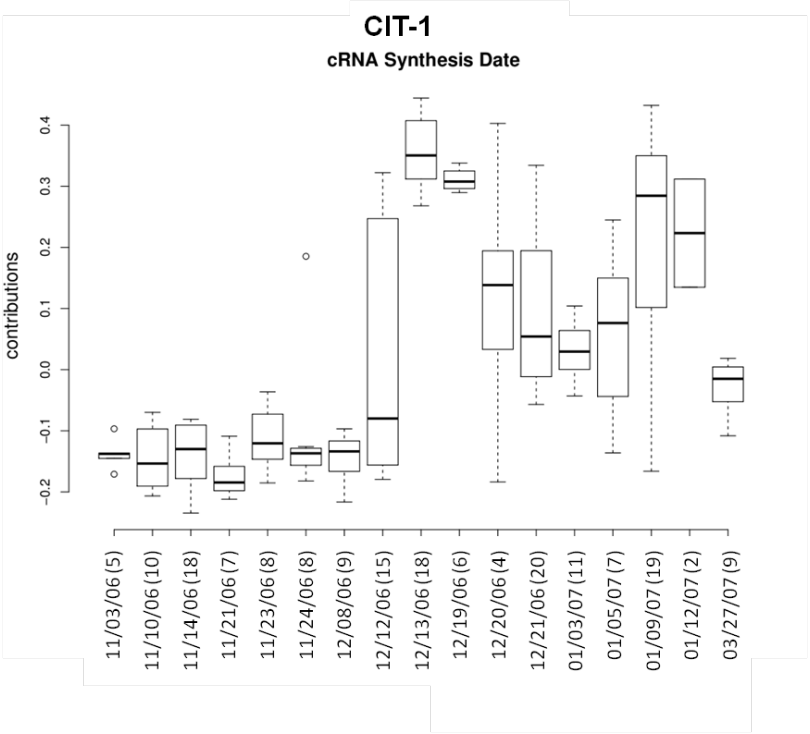
b) Projections of the genes onto the component CIT-2 are strongly correlated with their GC content

This scatter plot shows the proportion of G and C bases of the genes as a function of their gene projections on CIT-2, the Pearson correlation coefficient between the two vectors is 0.66. The fraction of G and C bases (%GC) in each gene was annotated with the Ensembl database through biomaRt (Durinck et al., 2005).



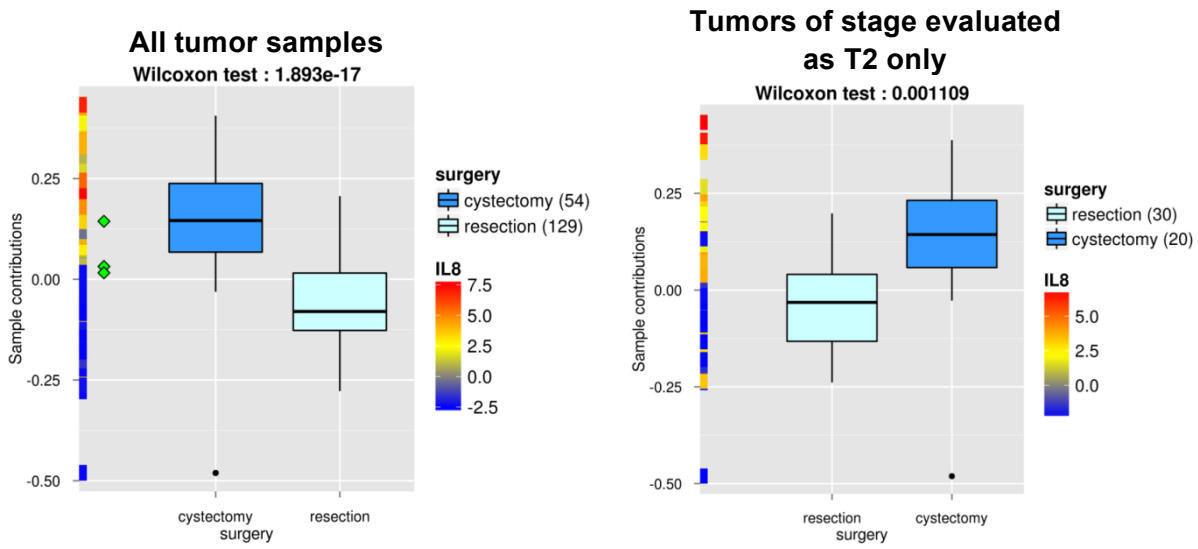
c) Association of the component CIT-1 with a batch effect

Distribution of the samples according to their date of cRNA synthesis on CIT-1. The distribution of samples in CIT-1 was clearly bimodal. We found no association of this component with biological information or tumor features, and CIT-1 was poorly reproducible across the datasets. This component separates two groups of samples, one processed before 12/12/2006, and one processed after 12/12/2006. The first group has negative contribution values and low levels of variation between cRNA synthesis dates, and the second group is distributed principally on the positive side of the component, with more variable contributions and poorer RNA quality. Questioning of the people who performed the microarray experiments revealed that the 16℃ water bath was out of service on 12/12/2006 and had to be changed during the synthesis step. The coincidence of the date of the water bath break and the date after which the samples were found to over-express the contributing genes of the component indicates that this incident was probably the source of the variability captured by this component. A change in the temperature of the water bath would indeed have affected the chemical reactions and therefore the efficiency of the cRNA synthesis.

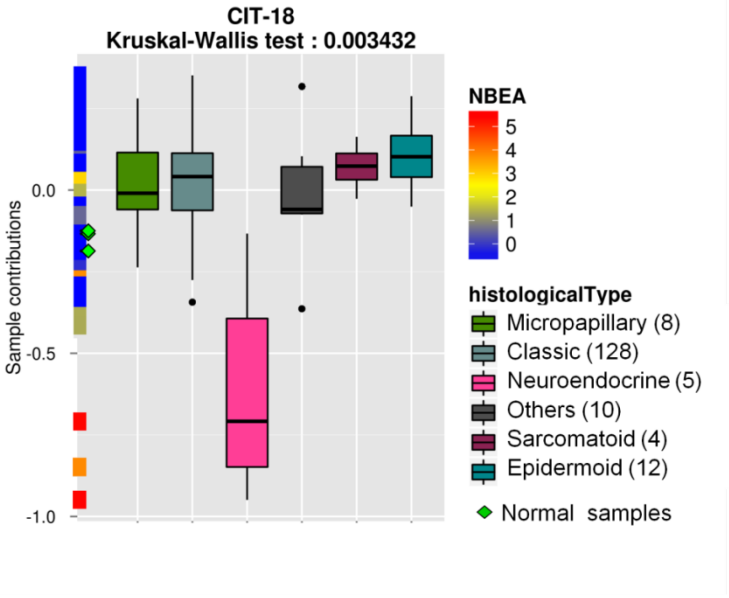


d) Distribution of the tumors according to the type of surgery (resection and cystectomy) on component CIT-14.

Tumors from the CIT series were obtained by either the radical removal of the bladder (cystectomy), or transurethral resection of the tumor (resection). The component CIT-14 indicates that cystectomy triggers a stronger stress response in the tumor cells in comparison with transurethral resection. CIT-14 indeed separates the tumor samples according to the type of surgery undergone by the patient, and its contributing genes are associated with inflammation and response to stress. On the left, all the tumor samples were considered, whereas on the right, only T2 tumors were considered. For this stage both types of surgery were used. The samples are differentially distributed on this component according to the type of surgery, and this separation persisted after controlling for stage. The x-axis indicates the type of surgery; the y-axis indicates the gene contributions to CIT-14. The vertical line at the left from each of the boxplot indicates the level of expression of one of the most contributing genes of CIT-14 (*IL8*), where each color point represents a tumor sample with its color showing the *IL8* gene expression. The diamond-shaped green points indicate the position of the normal samples. The BH FDR of the Wilcoxon tests are provided at the top of the plot.



e) Assignment of component CIT-18 as the neuroendocrine component



Distribution of the tumor samples on CIT-18 according to their histological class: The vector of tumor contributions to CIT-18 is determined by a small subgroup of tumors corresponding to bladder tumors histologically classified as neuroendocrine tumors. The *x*-axis indicates the tumor class; the *y*-axis indicates the contributions to CIT-18. The diamond-shaped green points indicate the position of the normal samples. The BH FDR of the Kruskal-Wallis test is provided at the top of the plot. The GO gene sets enriched in the contributing genes of this component ("GO BP: regulation of neurological system process, neurogenesis, neuron differentiation") were consistent with the association of this component with the neuroendocrine tumors. The vertical line at the left end of the boxplots indicates the level of expression of the neurobeachin gene (*NBEA)* one of the most contributing genes in CIT-18, each color point here represents a tumor sample with its color showing the *NBEA* gene expression.

Figure S2, related to Figure 3: Relationship between histological features and the molecular annotation of the components

The distributions of the tumor samples of the CIT series according to their contribution to the components smooth muscle, myofibroblast, and cell cycle are represented on the three upper panels. For each component, the sample contribution values are given on the vertical axis. The expression level of one of the strongest contributing genes of the component is also shown on this axis, each point representing one tumor sample, with its color indicating the relative level of expression of the gene (the color scale for gene expression is also shown). Normal samples are represented by green diamonds. For each component, representative pathological sections are shown for two tumors displaying opposite contributions to the component. For the smooth muscle component, the most positive values (corresponding to high expression of *CNN1*, a smooth muscle marker), were associated with the presence of smooth muscle as shown for tumor #252 (indicated by an arrow in the lower panel). No smooth muscle bundle is observed in tumors displaying the most negative contributions, as illustrated for tumor #93. For the myofibroblast component, tumors with the most positive contributions (corresponding to high expression of *COL6A3*, a myofibroblast marker) presented a fibrous stroma (indicated by an arrow) with small infiltrating tumor islets, as illustrated here with tumor #262. No stroma was observed in tumor #268 which displayed a negative contribution to the component. For the cell cycle component, tumors with the most negative contributions (corresponding to high expression of *TOP2A*, a cell cycle marker) presented frequent mitoses (indicated by arrows) as illustrated with tumor #262; mitoses were absent from tumors displaying a positive contribution (low expression of *TOP2A*) (as shown for tumor #274).

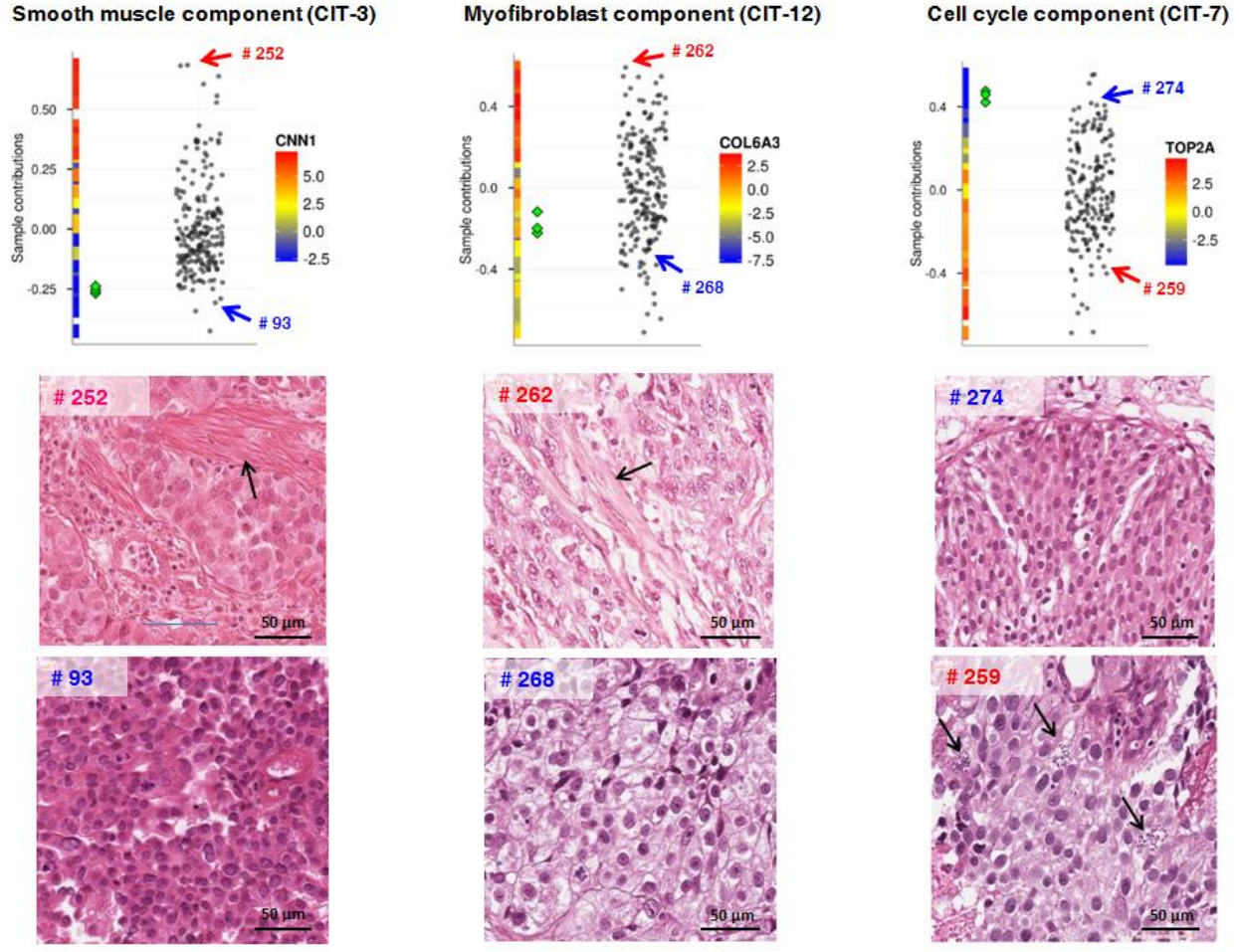
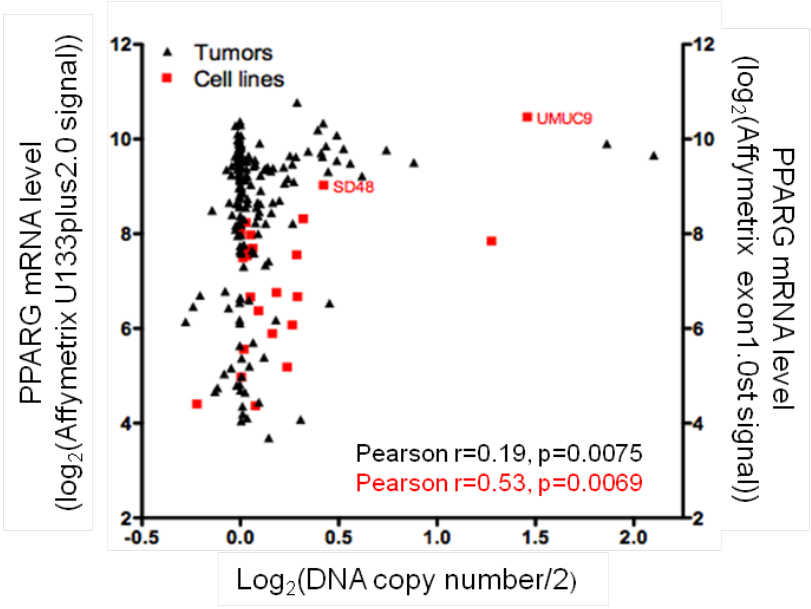
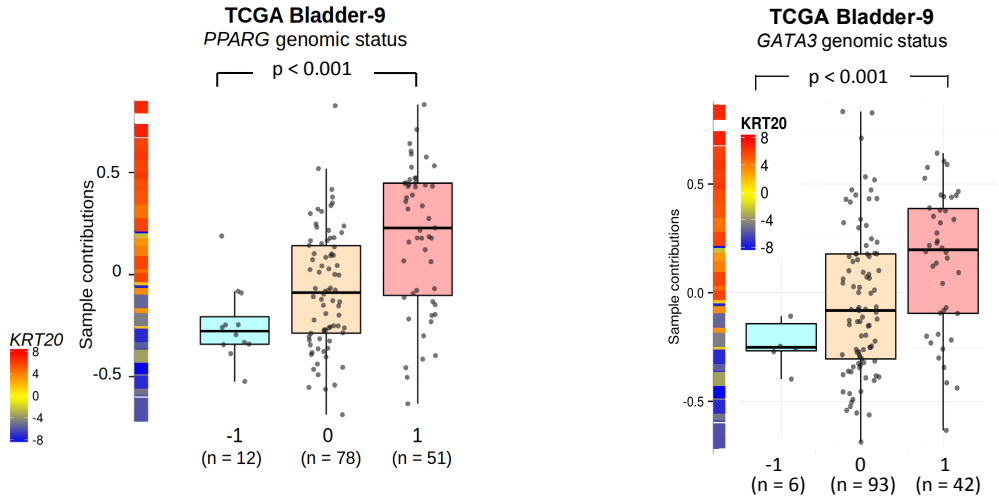


Figure S3, related to Figure 4: a) Correlation between *PPARG* mRNA and *PPARG* DNA copy number in bladder cell lines and tumors. b) Distribution of the tumor samples of the TCGA dataset according to their alteration status for *PPARG* on component TCGA Bladder-9. c) Efficiency of *PPARG* silencing by siRNAs specific for *PPARG*

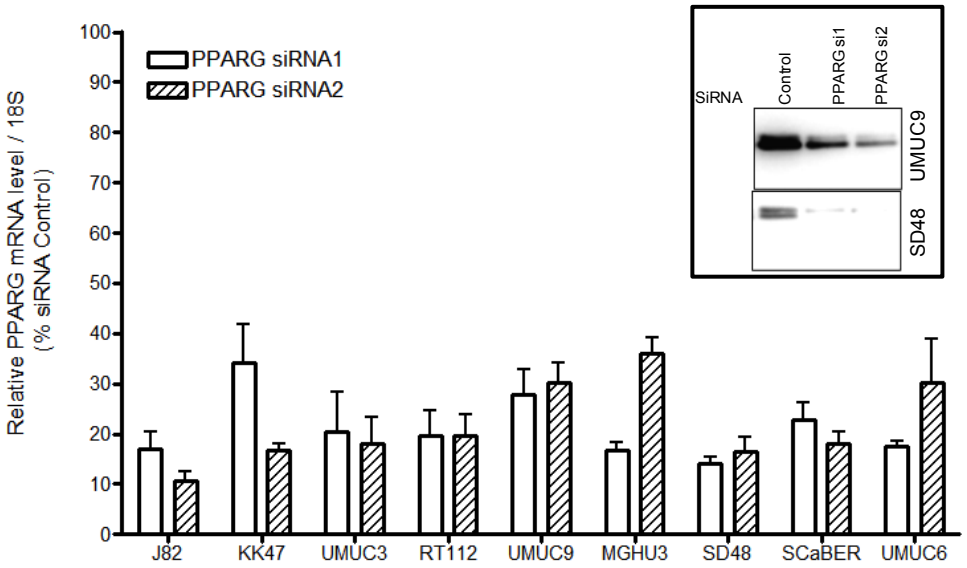
a) Linear regression analysis of the 178 bladder carcinomas (black) and of the 24 cell lines (red) for which we had both CGH and expression data. The Affymetrix signal (U133plus2.0 and Exon1.0st for tumors and cell lines respectively) and the CGH $\log_2$-ratio (aCGH and Affymetrix SNP6 for tumors and cell line respectively) were compared, using Pearson's correlation test. Expression levels were considered to be significantly modified with respect to DNA copy number if $p < 0.05$.



b) Distribution of the tumor samples of the TCGA Bladder cancer dataset on the component associated with urothelial differentiation as a function of their alteration status for the region encompassing *PPARG (left) and GATA3 (right)*, as measured by SNP 6.0 arrays. The *p*-value for the Wilcoxon test is provided at the top of the boxplot.



c) The level of *PPARG* mRNA was assessed and expressed relative to mRNA levels for *18S,* 72 hours after transfection with the specific or control siRNA. For the two cell lines presenting a *PPARG* genomic alteration, PPARG protein levels were also assessed by western-blotting.



SUPPLEMENTAL TABLES

Table S1, related to Table 1: Projections of the genes onto the different components for the CIT dataset

This table provides the scaled genes projections with absolute values exceeding 2.5 on the different components (CIT-1 to CIT-20). The genes are ranked according to their absolute scaled projection. The column definitions are listed below:

A. hgnc_symbol: HUGO gene symbol

B. scaled_proj: (scaled) projection of the gene onto the component

C. nbOcc_forThreshold:3: Number of components on which the gene has an absolute projection greater than 3.

D.comp_forThreshold:3: Indices of the components on which the gene has an absolute projection greater than 3.

The following columns (E, F, G, H, I, J, K) provide a description of the genes as available in biomaRt (Durinck et al., 2005), their genomic coordinates, and the last column gives the Ensembl IDs of the genes.

Table S2, related to Table 1: Association of the components of the CIT dataset with gene sets

Sheet 1: Enrichment of the component CIT-14 in genes identified as differentially expressed in prostate cancer according to the type of surgery
(Lin et al., 2006), identified 62 differentially expressed genes between prostate biopsy and prostatectomy specimens in a study analyzing the effects of surgical manipulation on overall gene expression. This table provides for the 41 genes of 62 that were available in the CIT dataset (column A), their projection across the 20 components of the CIT dataset (columns B to U), and the *p*-value (row 43) and FDR (row 44) of the statistical test for their enrichment in the contributing genes of the components. The projections with an absolute value larger than 3 are shown in bold.

Sheet 2: Association of the components of the CIT dataset with genomic locations

This table lists genomic regions including genes with high contributions on the components that were selected using a sliding window approach (see Supplemental Methods).

Column definitions are listed below:

A. Component index

B. Chromosome of the region

C. Chromosome band of the region

D. Start position of the region

E. End position of the region

F. Genes of the dataset included in the region; the score of the sliding window centered on the gene is given in parentheses.

G. Threshold used to select the windows

Following sheets:
This table contains the results of the enrichment analyses for each independent component of the CIT dataset. Two sheets are available per component, each sheet contains the results of the enrichment analyses performed with GOstats and GSEA, respectively.
The column definitions are listed below:

GOstats
A. DB: GO database.
B. ID: GO ID.
C. Term: GO Term.
D. P-value: probability of observing the number of genes annotated for the gene set in the selected gene list, given the total number of annotated genes in the reference population.
E. Odds ratio: odds ratio for each category term tested providing an indicator of the level of enrichment in genes within the list as against the universe.
F. Expected counts: expected number of genes in the selected gene list to be found at each tested category term/gene set.
G. Counts: number of genes in the selected gene list annotated for the gene set.
H. Size: number of genes from the universe annotated for the gene set.
I. In_geneSymbols: The Symbols of the genes in the selected gene list that are included in the gene set.

GSEA (see http://www.broadinstitute.org/gsea/doc/GSEAUserGuideTEXT.htm#_Interpreting_GSEA_Results for an explanation of the GSEA outputs)

A. Category: Category of the MSigDB database to which the gene set belongs.
B. Name: Name of the gene set.
C. NES: Normalized enrichment score (NES): reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes.
D. FDR.q.val: False discovery rate, estimated probability that a gene set with a given NES represents a false positive finding.
E. NOM.p.val: *P*-value: the statistical significance of the enrichment score.
F. Leading edge subset: the subset of members of the gene set that contribute most to the ES (only available for a subset of the gene sets).

Table S3, related to Figures 1, 3 and 4: Sample annotations and contributions

Sheet 1: Clinical, pathological and molecular annotations associated with the CIT samples

Each row refers to a sample and each column to a given annotation. The first column provides the sample ID. IDs of normal samples start with an "N".

age: age of the patient at the time of surgery.

gender: gender of the patient.

size: tumor size.

necrosis: necrosis status of the tumors.

grade98: 1998 WHO/ISUP classification of the grade of the bladder tumors.

grade73: 1973 WHO/ISUP classification of the grade of the bladder tumors.

stage: tumor stage (Ta, T1, T2, T3, T4).

stage_3levels: stages divided into three groups: Ta, T1, and stages of T2 or higher (T2+).

Invasiveness: stages divided into two groups: muscle-invasive tumors (T2+, noted as "+") and non-muscle-invasive tumors (Ta and T1, noted as "-").

FGFR3_mutation, RAS_mutation, TP53_mutation: mutation status for *FGFR3*, *RAS*, and *TP53* ("yes" if mutated, "no" if not mutated, NA if the information is not available).

EGFR_CN, E2F3_CN, MDM2_CN, RB_CN, PPARG_CN: copy number status for *EGFR*, *E2F3*, *MDM2*, *RB*, and *PPARG* as determined from aCGH data. -1 means loss, 0 no alteration, 1 gain.

CDKN2A_CN: copy number status for *CDKN2A* as determined by MLPA.

histologicalType: histological type of the tumor (six categories: micropapillary, sarcomatoid, classic, epidermoid, neuroendocrine, and others).

stage_evaluationPreSurgery: stage as evaluated before surgery.

surgery: type of surgery (resection or cystectomy).

vital_status: vital status of the patient at the end of the follow-up, alive or deceased.

recurrence: recurrence status of the patient.

MRES: multiple regional epigenetic silencing (MRES) status of the tumor sample according to the MRES transcriptomic signature (Vallot et al., 2011).

CIS_Orntoft: Classification of the tumor sample according to a CIS transcriptomic signature. (Dyrskjøt et al., 2004).

TCGA_subtypes: Classification of the tumor sample according to the four gene expression subtypes as defined in the TCGA classification.

Sheet 2: Contributions of the samples from the CIT dataset across the different components

Each row refers to a sample and each column to a component. The first column gives the sample IDs.

Sheet 3 and 4: Association of the components of the CIT dataset with clinical, pathological, and molecular features, using the contributions of the samples to the components

Each column refers to a component and each row to a given feature. The contributions of pre-established groups of samples were compared using Wilcoxon rank-sum tests for two-classes comparison and Kruskal-Wallis tests for the comparison of more than two classes. The first table provides the adjusted *p*-values (by variable, with method BH) of these tests, the second table contains the unadjusted *p*-values. For the variable "age", the correlation between the age of the patients and the contribution of their tumor sample to the component is given. The *p*-values and FDR values below 0.05 are shown in bold.

age: age of the patient at the time of surgery. as age is a continuous variable, we give in this row the Pearson correlation coefficient between age and sample contribution, and with the p-value of the correlation test.

gender: gender of the patient.

size: tumor size.

necrosis: necrosis status of the tumors.

grade98: 1998 WHO/ISUP classification of the grade of bladder tumors.

grade73: 1973 WHO/ISUP classification of the grade of bladder tumors.

stage: tumor stage (Ta, T1, T2, T3, T4).

stage_3levels: stages divided into three groups: Ta, T1, and stages T2 or higher (T2+).

Invasiveness: stages divided into two groups: the muscle-invasive tumors (T2+) and non-muscle invasive tumors (Ta and T1).

EGFR_CN, E2F3_CN, MDM2_CN, RB_CN, PPARG_CN: copy number status for the given genes as determined from aCGH.

CDKN2A_CN: copy number status for *CDKN2A* as determined by MLPA.

HistologicalType: histological type of the tumor (six categories: micropapillary, sarcomatoid, classic, epidermoid, neuroendocrine, and others).

stage_evaluationPreSurgery: stage as evaluated before surgery.

surgery: type of surgery (resection or cystectomy).

vital_status: vital status of the patient at the end of follow-up, alive or deceased.

recurrence: recurrence status of the patient.

MRES: multiple regional epigenetic silencing (MRES) status of the tumor sample according to the MRES transcriptomic signature (Vallot et al., 2011).

CIS_Orntoft: classification of the tumor sample according to a CIS transcriptomic signature (Dyrskjøt et al., 2004).

Sheet 5: Association of the components of the CIT dataset with genomic locations

This table lists genomic regions including genes with high contributions on the components that were selected using a sliding window approach.

Column definitions are listed below:

A. Component index

B. Chromosome of the region

C. Chromosome band of the region

D. Start position of the region

E. End position of the region

F. Genes of the dataset included in the region; the score of the sliding window centered on the gene is given in parentheses.

G. Threshold used to select the windows

Table S4, related to Figure 2: Description of the transcriptome datasets of bladder tumors and other tumor types and projections of the genes onto the components in different cancer transcriptome datasets

The column definitions are listed below:

First sheet (Datasets):

A. The cancer type of the dataset.

B. The ID of the dataset.

C. The number of tumor samples and the number of normal samples if any.

D. The technology used to generate the dataset.

E. The way the dataset has been normalized.

F. The gene annotation method.

G. The data access, the TCGA datasets have been downloaded from https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/.

Second sheet (Gene populations):

This table contains the number of common genes between each pair of dataset.

Following sheets:

For each dataset, the Table lists the genes with an absolute projection value of at least 3 on one or several components, together with their projections across the components. The genes (first column) are ranked according to the standard deviation (column “sd_expr”) of their expression profiles. The columns “nbOcc” and “components” give the number of components and the indices of the components on which the gene has an absolute projection larger than 3, respectively. The remaining columns give the projection values of the gene across the 20 components. The projections with absolute values higher than 3 are shown in bold.

Table S5, related to Figure 2: Contributing genes of the largest pseudo-cliques in different cancer data sets

This Table shows the genes that contribute to the components of pseudo-cliques. The genes are ranked by the median of their ranks in the components of the pseudo-clique. There is one sheet per pseudo-clique.

Column definitions are listed below:

A. Hgnc_Symbol: The gene Symbol.

B. Median rank: The median rank of the gene across the components of the pseudo-clique.

C. nbAn: The number of datasets/components of the pseudo-clique on which the gene has an absolute projection value greater than 3.

D. description: The description of the gene.

E. chr: The chromosome on which the gene is located.

F. band: The cytoband location of the gene.

G. Datasets: The datasets present in the pseudo-clique to which the gene contributes.

H. Min rank: The minimum rank of the gene across the components of the pseudo-clique.

I. Ranks: The distribution of gene ranks across all the components of the pseudo-clique.

J. scaled_proj: The projection values of the genes on the components of the pseudo-clique.

Table S6, related to Figure 4: Association of the CIT components with aCGH data

This Table lists the genomic alteration status of the bladder CIT samples and the components for which the sample contributions differed according to their genomic alterations, as determined from aCGH data (see Supplemental Methods and Rebouissou et al., 2014).

First sheet: copy number status for the available samples.

A. Bacterial artificial chromosome (BAC) ID.

B-D: Position of the BAC.

E-FX: copy number status, one column per sample.

Sheets 2-14: The association of each BAC with each component was assessed individually based on the GNL status (loss (-1), normal (0) and gain (1), amplifications were merged with gains). Contiguous BACs similarly associated with the component were merged. Each sheet lists the regions associated with a particular component.

Column definitions are listed below:

A. regionId: ID of the region, defined by the first and the last BAC in the BAC sequence defining the region.

B-F. The columns “begin, end, chr, band, and BACs” give the coordinates of the genomic region constituted by BACs with alteration profiles differentially distributed on the ICs.

G. The column “config” indicates the distribution of genomic status on the ICs, with two types of configurations:

- opposite genomic aberrations differently distributed on the component, with one of them distributed on the same side as the unaltered group (0,1 ≠ -1, or -1, 0 ≠ 1).

- opposite genomic aberrations both distributed differently from the unaltered group and on opposite sides of the component (-1>0>1, or -1<0<1).

H. nbGenes: number of genes included in the region.

I. genes: genes (available in the CIT dataset) included in the region.

Table S7, related to Figure 4: Differentially expressed genes upon *PPARG* knock-down in the SD48 cell line

This Table provides the results of the differential expression analysis, performed with *limma*, between SD48 cell line treated or not with siRNAs targeting *PPARG*.

Column definitions are listed below:

A. gene symbol: Gene symbols

B. chr: Chromosome on which the gene is located.

C. ID: Ensembl ID

D. logFC: estimate of the log2-fold change corresponding to the effect of siRNA treatment.

E. AveExpr: average log2-expression for the gene over all samples.

F. P.Value: raw *p*-value of the statistical tests.

G. adj.P.Val adjusted *p*-value (using BH FDR).

H. CIT9: projection of the gene on CIT-9 (if available in the CIT dataset).

I. description: gene description.

J. NbContribIC: number of components from the CIT dataset on which the projection of the gene has an absolute value higher than 3.

K-L: measured expression value of the gene in the various cell lines.

SUPPLEMENTAL METHODS

CIT series: tissues samples and data

A set of human bladder carcinomas was collected from patients treated surgically between 1988 and 2006 at Henri Mondor Hospital (Créteil, France), Institut Gustave Roussy (Villejuif, France) and Foch Hospital (Suresnes, France). All tumors were pathologically reviewed, staged according 2009 TNM. All patients provided written informed consent and the study was approved by ethics committees of various hospitals. Normal urothelial samples were obtained from organ donors (Diez de Medina et al, 1997). Expression of desmin, a specific smooth muscle cell marker, was measured by RT-qPCR, and tumor samples with expression higher than 25% of the median of 5 normal bladder smooth muscle were considered over-contaminated with stroma, and thus not included in the analysis. Human Bladder samples were analysed with Affymetrix HG-U133 Plus 2.0. DNA copy number was analyzed on the human genome-wide CIT-CGH array (V6) designed by the CIT-CGH consortium as described in Rebouissou et al., 2014.

The microarray Affymetrix and the array-CGH (comparative genomic hybridization) protocol and data analysis have been previously described in Rebouissou et al., 2014. The CIT data used here are available from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) under the accession numbers E-MTAB-1803 for the muscle invasive bladder tumours and E-MTAB-1940 for the non-muscle invasive tumours and normal samples.

Transcriptome data processing

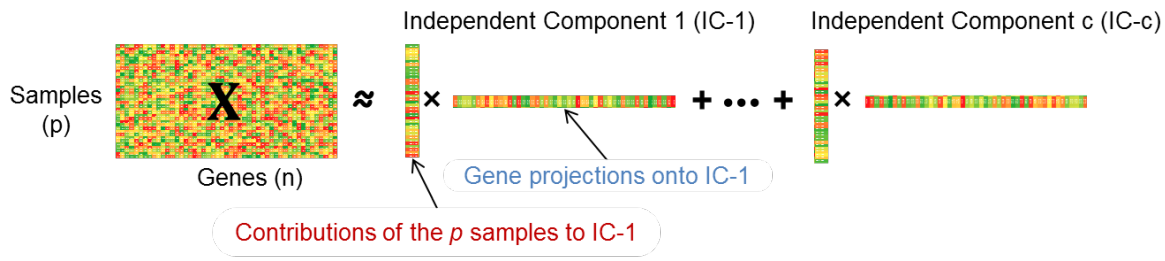
The datasets included in this analysis are described in Table S4. When raw data were available, we normalized the Affymetrix microarray-based datasets using GCRMA (Wu et al., 2004) or RMA methods (Irizarry et al., 2003) together with BrainArray custom chip description files (CDF) (Dai et al., 2005). For already processed datasets, we annotated the probe sets using the Bioconductor version 2.8.0 (Gentleman et al., 2004) annotation packages of the microarrays.

All RNASeq data were downloaded from the TCGA open-access HTTP directory (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/), and are Level 3 data summarizing read counts at the gene level. The IDs of the Level 3 datasets we used are available in Table S4. All RNASeq datasets were normalized with Upper Quartile (Bullard et al., 2010). Some of the tumors of the breast cancer TCGA series were also analyzed with microarrays. The two datasets were considered separately.

When expression measurements were available for more than 10,000 probe sets or genes, the 10,000 probes with the highest interquartile range (IQR) were selected before ICA. If more than one probe set mapped to a given gene, expression values at the gene level were defined as the median expression across probe sets for the gene. Official gene symbols were used as the gene identifiers. The overlap of gene populations between datasets is shown in Table S4, sheet 2.

Decomposition of transcriptome data using Independent Component Analysis

Independent Component Analysis (ICA) models gene expression as a weighted sum of independent signals, the independent components (ICs), each component capturing a different process influencing the expression of the genes. In this model, the gene expression matrix X of dimension p x n (p samples and n genes) is decomposed into a sum of c matrices (where c is the number of ICs), each of which is a matrix multiplication of two vectors of dimensions p and n. The vector of dimension n corresponds to the projections of the genes onto the IC and the vector of dimension p to the contributions of the p samples to the IC (or activities of the IC in the p samples). The genes having the largest projection onto a component (providing the greatest contribution) are the genes most strongly influenced by the process associated with this component. The contribution value of a sample reflects the activity of the component in the sample concerned.



For further details about ICA, the reader may refer to ((Comon, 1994)'(Hyvärinen, 1999a); (Cardoso, 1999); (Hyvärinen and Oja, 2000); (Hyvärinen et al., 2001); (Hyvärinen et al., 2009)).

Computation of the independent components.

We applied the FastICA algorithm (Hyvärinen, 1999b) to each dataset, and calculated 20 independent components, using *pow3* nonlinearity and *deflationary* algorithm. We ensured that only stable components were selected by running FastICA 500 times with *icasso* (Himberg et al., 2004). The number of components was chosen using the Cattell scree test on the CIT dataset (Cattell, 1966).

ICA was performed in the gene space, where each data vector corresponds to a gene, characterized by its expression in all samples. The expression of each gene was centered.

Analysis of the components based on gene projections.

We selected the genes associated with a component ("most contributing genes") by thresholding the absolute projections of expression at 3 or 4 standard deviations from the mean as in (Teschendorff et al., 2007). The functional gene set enrichment was assessed by the hypergeometric test (Bioconductor v2.8 package GOstats (Falcon and Gentleman, 2007)), using KEGG (Kanehisa et al., 2000) and Gene Ontology (Ashburner et al., 2000) categories, and "GSEA Preranked" module of GSEA-P v2.0 software (Subramanian et al., 2005), using c1-4 categories of MSigDB.

Association of the vector of gene projections with genomic location: we detected genomic regions enriched in contributing genes by shifting a sliding window gene by gene, including three neighboring genes on either side of the selected gene and computing the median of the

corresponding gene projection values. A null distribution of the scores was computed from 1000 permutations of the gene projection values. The 0.998th percentile of this distribution was used to select enriched windows.

Analysis of the components based on sample contributions

The vector of sample contributions reflects the activity of the factor represented by the component across the samples. Samples with opposite activities in this vector display opposite patterns of expression of the contributing genes of the component. By studying the distribution of groups of tumors on the component, we can therefore characterize them and associate a given tumor feature with the expression profile of the contributing genes. The contributions to a component of predefined groups of tumors were compared in Wilcoxon rank-sum tests and Kruskal-Wallis tests. The *p*-values were adjusted for multiple testing, for each variable, across the 20 ICs, by a Benjamini and Hochberg false discovery rate procedure. The vectors of sample contributions were associated to the components of the CIT dataset with the gene copy number profiles measured by BAC-based CGH arrays (see below).

Comparison of ICs across datasets using correlation-based graphs

The similarity of two ICs obtained in different datasets was assessed by calculating the absolute value of Pearson's correlation coefficient for the projection values of their common genes. The results of the correlation computation across various datasets were summarized in a correlation graph. Hereafter, the n^{th} component (out of N components) from dataset M is denoted $C_{M,n}$. In the correlation-based graph, edges are introduced in an asymmetric fashion: a directed edge from $C_{A,i}$ to $C_{B,j}$ signifies that $C_{B,j}$ is the IC most strongly correlated with $C_{A,i}$ among the $C_{B,1..N}$. Components reproducible across different datasets appear as pseudo-cliques in the graph. A pseudo-clique is an almost fully connected subgraph, missing relatively few edges. The layout of the graph was produced, using Cytoscape (Cline et al., 2007) in two steps. Firstly, for each CIT-component, all other components connected to it formed the initial CIT-centric node communities. If a component was correlated to several CIT-components then it was assigned to

the largest community. Communities defined in this fashion were extracted from the correlation graph. Secondly, for the rest of the graph, “Most Highly Connected Subgraph” Cytoscape plugin was applied to detect the remaining pseudo-cliques, which were also extracted from the correlation graph one by one. All these node communities were converted to meta-nodes using BiNoM (Bonnet et al., 2013), and used to compute the standard organic layout.

Definition of the TCGA subtypes

To build a classifier of the TCGA subtypes we used a centroid-based approach with cosine distance as similarity measure, and log2-scaled TCGA normalized data (RNASeqV2_Level3 matrix). For each subtype 100 marker genes were selected: the top 50 highest and top 50 lowest fold changes among genes showing a moderate t-test pvalue less than 1e-6 (comparison of expression values in the subtype /vs/ outside of the subtype). This selection yielded 319 distinct genes which were used to build the centroids. This classifier correctly classified 97% of the TCGA data (125/129 samples). To apply this predictor to the 201 normalized expression profiles from the CIT cohort, centroids were reduced to genes measured in both platforms, yielding 240 genes. The resulting classifier showed the same performance than in the TCGA data (95% of correct classification).

Image acquisition

FFPE Tissue microarrays (TMA) blocks were generated as previously described (Rebouissou et al., 2014), and Hematoxylin, Eosin and Safran stained TMA slides were scanned automatically using Aperio XT automate and analyzed using Calopix viewer (Tribvn).

Association of the components with genomic alterations

Our goal was to select components for which the sample contributions had differential distributions according to their genomic aberrations. The association of each BAC with each component was tested individually, based on the GNL status. Three levels of genomic status were considered, loss (-1), normal (0) and gain (1). We also calculated the correlation between

genome and transcriptome by the GTCA method (Neuvial et al., 2007) to assess the impact of genome alterations on gene expression.

For a component of interest and a given BAC b, let us denote by μ_{-1}, μ_0, and μ_1 the median contributions of the samples presenting a loss, no alteration, or a gain on b, respectively.

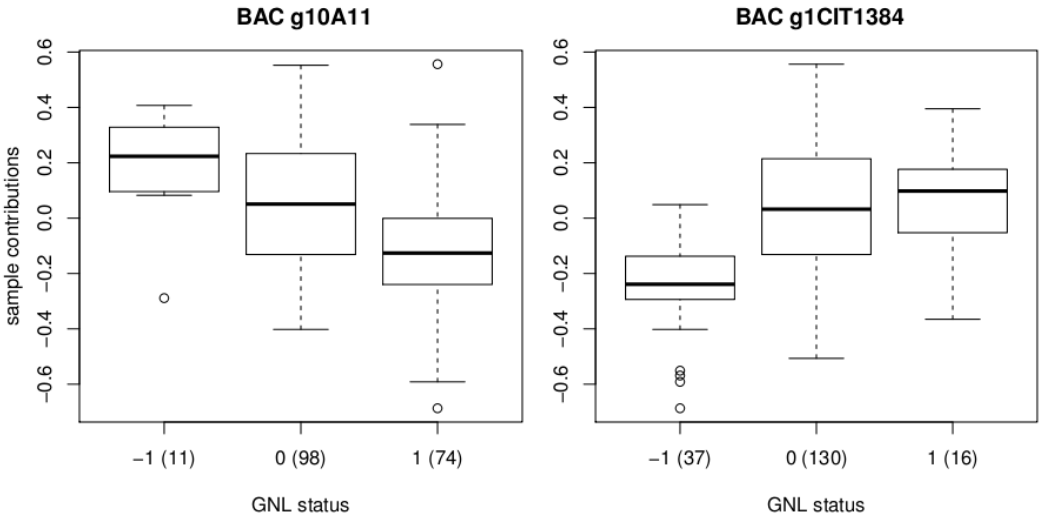
We considered two types of configuration:

a) Opposite genomic aberrations differentially distributed on the component, with one not different from the unaltered group (0): $\mu_{-1} \neq \mu_1 \wedge (\mu_{-1} = \mu_0 \vee \mu_1 = \mu_0)$. Two two-sided tests were used to select the corresponding configurations, the first testing the null hypothesis $H_0: \mu_{-1} = \mu_1$ and the second comparing the unaltered group with the closest group $g \in \{-1,1\}$ $H_0: \mu_{-1} = \mu_g$.

b) Opposite genomic aberrations both distributed differently from the unaltered group (0) and on opposite sides of the component: $\mu_{-1} < \mu_0 \wedge (\mu_1 > \mu_0 \vee ((\mu_{-1} > \mu_0) \wedge (\mu_1 < \mu_0))$. Two one-sided Wilcoxon tests were carried out to select this configuration. Each tested one of the following null hypotheses: $H_0: \mu_{-1} < \mu_0$, and $H_0: \mu_0 < \mu_1$ (the directions of the tests depended on the direction of the IC).

Illustration of the associations of interest between genomic status and component.

(left) The samples with a loss (-1) or a gain (1) of the BAC are distributed differently from the samples with no alteration, in opposite directions (-1>0>1); **(right)** The samples with a loss (-1) are distributed differently from the samples with a gain (1), which are not distributed differently from the samples with no alteration (-1<0,1).



We chose to ignore situations in which the losses (“-1”) and gains (“1”) were found on the same side of the component as the unaltered group (“0”). We also ignored the ambiguous case occuring when both gains and losses are distributed differently from each other and from “0” but are located on the same side of the IC.

If a particular type of alteration was represented by fewer than five samples, only one Wilcoxon test was performed, to compare the altered group with a sufficient number of samples and the unaltered group.

Corrections for multiple hypothesis testing based on the Benjamini and Hochberg FDR procedure were carried out for each type of test. A threshold was applied to the FDR on the appropriate tests for selecting BACs corresponding to the situations of interest. The BACs were ordered in the genome and the consecutive significant BACs were merged in genomic regions delimited by the start position of the first BAC and the end position of the last BAC. Regions on a same chromosome separated by only one BAC were merged.

The genes located within the selected regions were retrieved with the biomaRt (Durinck et al., 2005) R package and version 61 of the Human Ensembl database. As we considered only a subset of the available gene population, not all genes have projection values. One way of studying the list of genes included in a specific region is to select those with a high projection on

the studied IC. For these genes, we know that the genomic aberration has a potential impact on expression in the samples contributing to the IC.

Cell lines and cell culture

The bladder cancer-derived cell lines J82, KK47, SCaBER, SD48, RT112, UMUC3, UMUC6 and UMUC9 were obtained from the American Type Culture Collection or from the German Resource for biomedical material (DSMZ). MGH-U3 cells were kindly provided by Dr Yves Fradet. RT112 cells were cultured in RPMI medium, whereas all the other cells were cultured in Dulbecco's modified Eagle's medium (DMEM). Both media were supplemented with 2 mM glutamine and 10% fetal calf serum (FCS). All the reagents were purchased from Gibco-BRL (Cergy Pontoise, France). Cells were routinely grown at 37°C, under an atmosphere containing 5% CO_2.

RNA interference

Transient transfections were performed, using Lipofectamine RNAi according to the manufacturer's instructions for forward transfection (Invitrogen, Cergy Pontoise, France), with 20 nM siRNA, and cells were further cultured in DMEM containing 1% FCS. A negative control siRNA (luciferase GL2 siRNA, sense strand: CGUACGCGGAAUACUUCG) and three siRNAs specific for *PPAPRG* (sense strand for *PPARG* siRNA1:GCGACUUGGCAAUAUUUAUTT, *PPARG* siRNA2: CGGAGAACAAUCAGAUUGATT and *PPARG* siRNA3: GACAAAUCACCAUUCGUUATT) were purchased from Qiagen (Courtaboeuf, France). The three *PPARG* siRNAs were designed to knock down the expression of all known mRNA isoforms.

Quantitative real-time reverse transcription-PCR and DNA microarray analysis

Cells were transfected with 20 nM RNA in a six-well plate. RNA was isolated 72 h later, with RNeasy (Qiagen, Courtaboeuf, France) mini kits. Reverse transcription was performed with 0.5 µg of total RNA and a high-capacity reverse transcription kit (Invitrogen, Cergy Pontoise,

France). PCR was carried out in a Roche® light cycle 480 real-time thermal cycler, with Roche Taqman master mix (Roche®). The sequences of the primers used are available upon request. We analyzed 0.2 µg of total RNA with the Affymetrix human exon 1.0 ST DNA array, as previously described (Rebouissou et al., 2014). Raw data were analyzed as described above.

Cell viability assay

Cells were transfected in a 24-well plate, with 20 nM siRNA; 72 hours after transfection, the cells were incubated for 1 hour with 0.5 mg/ml MTT and then lysed in DMSO. Absorbance was measured at a wavelength of 550 nm.

Soft agar assay

We added 20,000 siRNA-transfected cells in DMEM supplemented with 10% FCS and 0.3% agar to each of triplicate wells containing medium and 0.8% agar, on 12-well plates. The plates were incubated for 14 to 21 days and colonies with diameters greater than 50 µm on observation under a phase-contrast microscope equipped with a measuring grid were scored as positive.

Statistical analysis

All functional experiments were carried out twice or three times, in triplicate. Data were analyzed as means ± SD. The control siRNA group was used as the reference. For *PPARG*, the Affymetrix signal was compared with the CGH $\log_2$-ratio and with cell growth inhibition after siRNA treatment in the cell lines, in Pearson's correlation tests. LIMMA (Smyth, 2005) was used to identify genes differentially expressed between siRNA-treated and Lipofectamine-treated cells from the SD48 cell line. The *p*-values were adjusted for multiple testing by BH FDR methods. Genes with a FDR below 5% were considered to be differentially expressed.

SUPPLEMENTAL REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. Nature genetics, *25*, 25-29.

Bonnet E., Calzone L., Rovera D., Stoll G., Barillot E., and Zinovyev A. (2013). BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. BMC Syst. Biol. *7*, 18.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Cardoso, J.F. (1999). High-order contrasts for independent component analysis. Neural Comput *11*, 157–192.

Cattell, R.B. (1966). The scree test for the number of factors. Multivar. Behav. Res. *1*, 245–276.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. *2*, 2366–2382.

Comon, P. (1994). Independent component analysis, a new concept? Signal Process. *36*, 287-314.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 33, e175.

Diez de Medina, S.G., Chopin, D., El Marjou, A., Delouvee, A., LaRochelle, W.J., Hoznez, A., Abbou, C., Aaronson, S.A., Thierry, J.P., Radvanyi, F.(1997). Decreased expression of keratinocyte growth factor receptor in a subset of human transitional cell bladder carcinomas. Oncogene *14*,323–330

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., Moor, B.D., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Dyrskjøt, L., Kruhøffer, M., Thykjaer, T., Marcussen, N., Jensen, J.L., Møller, K., and Ørntoft, T.F. (2004). Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. Cancer Res *64*, 4040–4048.

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics *23*, 257–258.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. Neuroimage *22*, 1214–1222.

Hyvärinen, A. (1999a). Survey on Independent Component Analysis. Neural Comput. Surv. *2*, 94–128.

Hyvärinen, A. (1999b). Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. Ieee Trans. Neural Networks *10*, 626–634.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. Neural Netw *13*, 411–430.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). Independent Component Analysis (John Wiley & Sons).

Hyvärinen, A., Hurri, J., and Hoyer, P.O. (2009). Natural image statistics: A Probabilistic Approach to Early Computational Vision. (Springer London).

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30.

Lin, D.W., Coleman, I.M., Hawley, S., Huang, C.Y., Dumpit, R., Gifford, D., Kezele, P., Hung, H., Knudsen, B.S., Kristal, A.R., et al. (2006). Influence of surgical manipulation on prostate gene expression: implications for molecular correlates of treatment effects and disease prognosis. J Clin Oncol *24*, 3763–3770.

Neuvial, Pi., Gestraud, P., Lucchesi, C., and Barillot, E. (2007). GTCA: Genome Transcriptome Correlation Analysis within R. In ISMB/ECCB, Vienna, Austria.,.

Rebouissou, S., Bernard-Pierrot, I., de Reyniès, A., Lepage, M.L., Krucker, C., Chapeaublanc, E., Hérault, A., Kamoun, A., Caillault, A., Letouzé, E., et al. (2014). EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype. Sci Transl Med. *6*, 244ra91.

Smyth GK (2005). “Limma: linear models for microarray data.” In Gentleman R, Carey V, Dudoit S, Irizarry R and Huber W (eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, pp. 397–420. Springer, New York.

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

Teschendorff, A.E., Journee, M., Absil, P.A., Sepulchre, R., and Caldas, C. (2007). Elucidating the altered transcriptional programs in breast cancer using independent component analysis. Plos Comput. Biol. *3*, e161.

Vallot, C., Stransky, N., Bernard-Pierrot, I., Hérault, A., Zucman-Rossi, J., Chapeaublanc, E., Vordos, D., Laplanche, A., Benhamou, S., Lebret, T., et al. (2011). A novel epigenetic phenotype associated with the most aggressive pathway of bladder tumor progression. J Natl Cancer Inst *103*, 47–60.

Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. J Am Stat Assoc, *99*, 909–917.