Gene Set Enrichment Analysis and Ingenuity Pathway Analysis of Metastatic Clear Cell Renal Cell Carcinoma Cell Line

Authors: Mohammed I. Khan¹, Konrad J. Dębski², Michał Dabrowski², Anna M. Czarnecka¹, Cezary Szczylik¹

Authors’ information:
1. Molecular Oncology Laboratory, Department of Oncology, Military Institute of Medicine, ul. Szaserów 128, 04-141 Warsaw, Poland
2. Bioinformatics Laboratory, Center of Neurobiology, Nencki Institute of Experimental Biology PAS, ul. Pasteura 3, 02-093 Warsaw, Poland

Address for correspondence:
Mohammed I. Khan, e-mail: imrankhanbioinfo@gmail.com, Phone: +48-261817172

Keywords: Metastatic renal cell carcinoma; gene set enrichment analysis (GSEA); Ingenuity Pathway Analysis (IPA)

Abbreviations: mccRCC: metastatic clear cell renal cell carcinoma; GSEA: gene set enrichment analysis; mRCC: metastatic renal cell carcinoma; IPA: Ingenuity Pathways Analysis; IPKB: Ingenuity Pathway Knowledge Base; FC: fold change; GO: Gene ontology;

Running head: GSEA and IPA of metastatic ccRCC Caki-1
Abstract:

In recent years, genome-wide RNA expression analysis has become a routine tool that offers a great opportunity to study and understand the key role of genes that contribute to carcinogenesis. Various microarray platforms and statistical approaches can be used to identify genes that might serve as prognostic biomarkers and be developed as antitumor therapies in the future. Metastatic renal cell carcinoma (mRCC) is a serious, life-threatening disease, and there are few treatment options for patients. In this study, we performed one-color microarray gene expression (4×44K) analysis of the mRCC cell line Caki-1 and the healthy kidney cell line ASE-5063. A total of 1921 genes were differentially expressed in the Caki-1 cell line (1023 upregulated and 898 downregulated). Gene set enrichment analysis (GSEA) and Ingenuity Pathway Analysis (IPA) approaches were used to analyze the differential-expression data. The objective of this research was to identify complex biological changes that occur during metastatic development using Caki-1 as a model mRCC cell line. Our data suggest that there are multiple deregulated pathways associated with metastatic clear cell renal cell carcinoma (mccRCC), including integrin-linked kinase (ILK) signaling, leukocyte extravasation signaling, IGF-I signaling, CXCR4 signaling, and PI3K/AKT/mTOR signaling. The IPA upstream analysis predicted top transcriptional regulators that are either activated or inhibited, such as estrogen receptors, TP53, KDM5B, SPDEF, and CDKN1A. The GSEA approach was used to further confirm enriched pathway data following IPA.

Keywords: Metastatic renal cell carcinoma, gene set enrichment analysis (GSEA), Ingenuity Pathway Analysis (IPA)

Introduction:

Metastatic renal cell carcinoma (RCC) is a serious, life-threatening disease. Approximately 30% of patients with RCC have advanced disease at their initial diagnosis, and 60% of these patients have greater mortality because of metastases and the aggressiveness of their disease (41, 45, 53). Metastasis of RCC has been reported to almost all organs in the human body, including the brain, skin, bones, pancreas, thyroid, lungs, and liver (4, 6). Treatment is difficult, because metastatic renal cell carcinoma (mRCC) shows limited or no responsiveness to conventional anticancer therapies (31). Targeted therapies are currently considered the standard of care for mRCC patients. Five VEGF pathway inhibitors—including tyrosine kinase inhibitors (TKIs), such as sorafenib (18), sunitinib (52), pazopanib (69), and axitinib (61); monoclonal antibodies directed against VEGF, such as bevacizumab, used in
combination with interferon-alpha (IFNα) (19); and mTOR inhibitors, such as everolimus (51) and temsirolimus (2)—are now approved by the FDA to treat mRCC. Most recently, nivolumab—an anti-PD-1 monoclonal antibody—has been shown to be effective against RCC (50). Clear cell renal cell carcinoma (ccRCC), the most frequent RCC subtype, is characterized by very high mortality rate of 40% because of the frequency of distant metastasis (57, 75). Understanding the specific alternations in metastatic ccRCC (mccRCC) is the first step toward development of novel therapeutics. Therefore, there is an urgent need to perform gene expression studies on mccRCC in order to find new biomarkers. In addition, there is a need to develop novel prognostic and therapeutic biomarkers that could allow early detection of ccRCC before onset of metastasis based on gene expression data.

Recently, Subramanian et al. suggested a novel approach for gene expression data analysis. Their approach, called gene set enrichment analysis (GSEA) (70), requires knowledge-based databases accumulating large-scale expression data sets. The GSEA approach is then applied to gene expression data to select whole sets of functionally related groups of genes by comparison with selected gene sets from available databases. In this way, GSEA reveals many biological interpretations in common.

We used GSEA to analyze microarray data from the mccRCC cell line Caki-1, which displays epithelial morphology and expresses wild-type von Hippel-Lindau (VHL) (25, 29). Caki-1 cells represent a suitable in vitro model for studying the human proximal tubule epithelium and mRCC (26, 29). This cell line was compared with human kidney epithelial cells (ASE-5063), which is an ideal cell line to study the development and biochemical function of the kidneys (3, 14). Differentially expressed gene profiling data of the above cell lines were further analyzed using Ingenuity Pathway Analysis (IPA) software to interpret the biological changes, altered canonical pathways, upstream transcriptional regulators, and gene networks of the metastatic Caki-1 cell line. The goal of this study was to use GSEA and IPA approaches to determine: 1) complex biological changes that occur in mccRCC; 2) how many and which genes are differentially expressed—either upregulated or downregulated—in mRCC; 3) significantly changed biological processes and pathways in mRCC; and 4) an interaction network of differentially expressed genes.

Materials and methods:

1. Cell lines and RNA extraction:

Human renal carcinoma cell line Caki-1 (wild-type VHL) (ATCC, HTB-46™) was bought from American Type Culture Collection (Virginia, USA), and human healthy kidney
epithelial cell line ASE-5063 was bought from Applied StemCell Inc. (California, USA). Both cell lines were maintained in RPMI-1640+GlutaMAX™-I medium (LifeTechnologies, California, USA) with 10% FBS (Biochrom GmbH, Cambridge, UK) and 1% PenStrep (100 U/mL penicillin and 100μg/mL Streptomycin) from AMRESCO LLC (Ohio, USA). Cells were expanded in T-75 flasks (Orange Scientific, Braine-l’Alleud, Belgium) before isolation of total RNA. Cells were dissociated using Accutase™ cell detachment solution (BD Biosciences, California, USA). Total RNA was isolated using Total RNA Mini Plus from A&A Biotechnology (Gdynia, Poland) as described in the manufacturer’s protocol. Isolated RNA from the healthy kidney cell line (ASE) was used as a biological reference control. The isolated RNA was stored at −80ºC until the microarray experiment was carried out.

2. Microarray procedure:

RNA quality and integrity was measured by 2100 Bioanalyzer (Agilent, California, USA). Amplification, labeling, generation of cRNA, and hybridization were done by PERLAN Technologies (Warsaw, Poland) on Agilent’s human GE 4×44K v2 (G4845A) microarrays, as described previously by Stankiewicz et al. (67). Limma R/Bioconductor package version 3.22.7 (62) was used to perform microarray data background correction, quantile normalization, filtering of probes with low intensity in less than half of samples (3 out of 6), probe summarization at the gene level, quality control, principal component analysis (PCA), and statistical analysis of differentially expressed genes. Low-intensity probes were defined as those whose intensity was below a threshold set at 10% above the third quartile of negative probes. PCA was performed with the prcomp R function based on summarized expression data.

3. Bioinformatics methods:

3.1 Differentially expressed genes

Differential gene expression levels between healthy control kidney cell line ASE and mccRCC cell line Caki-1 were estimated with a moderated T-test using the limma package (62). P-values were corrected with the Benjamini-Hochberg algorithm (false discovery rate: FDR). Genes were considered as differentially expressed if the adjusted p-value was less than 0.05 and absolute fold change was greater than or equal to 2.

3.2 Enriched Gene Ontology (GO) terms
The list of differentially expressed genes from mccRCC cell line Caki-1 compared with healthy kidney cell line ASE was uploaded into the Protein Analysis Through Evolutionary Relationships (PANTHER) classification system to infer the functions of genes based on their evolutionary relationship (46). PANTHER’s Gene List Analysis option was used to identify overrepresentation of GO terms in the gene list data. The most significantly enriched ontologies were presented in a pie chart based on the list of up- or downregulated genes involved in each of the terms.

3.3 Pathway and network analysis by IPA
The list of differentially expressed genes in the Caki-1 cell line, containing gene identifiers and corresponding expression values, was uploaded into the Ingenuity Pathway Analysis (IPA) software (Qiagen, California, USA). The “core analysis” function included in the software was used to interpret the differentially expressed data, which included biological processes, canonical pathways, upstream transcriptional regulators, and gene networks. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathway Knowledge Base (IPKB).

3.4 GSEA analysis of differentially expressed genes
The GSEA analysis was done using GSEA software version 2.2.2.0 (49, 70), which uses predefined gene sets from the Molecular Signatures Database (MSigDB v5.0) (70). A gene set is a group of genes that share pathways, functions, chromosomal localization, or other features. For the present study, we used all the C collection sets for GSEA analysis (i.e., C1–C7 collection in MsigDB) and list of ranked genes based on a score calculated as \(-\log_{10}\) of p-value multiplied by sign of fold change. The minimum and maximum criteria for selection of gene sets from the collection were 10 and 500 genes, respectively.

RESULTS:
Differentially expressed genes of mccRCC (Caki-1) cell line.
We compared the expression profiles of mccRCC cell line Caki-1 with healthy human kidney epithelial cell line ASE. As described in the materials and methods section, differentially expressed genes were obtained from the Caki-1 cell line. Using a moderated T-test with p-value corrected using the Benjamini-Hochberg algorithm, we found 1921 genes that were differentially expressed with a significance level less than 0.05 (FDR) and absolute fold change greater than or equal to 2 (Figure 1). Of these, 1023 genes were significantly
upregulated and 898 genes were significantly downregulated. The complete list of deregulated genes was submitted as supplementary table S1. The data discussed in this publication have been deposited in NCBI’s Gene Expression Omnibus (16) and are accessible through GEO Series accession number GSE78179.

**Enriched biological pathways by IPA analysis**

Using IPA, we examined the relation between these highly significant genes (1921 genes with FDR [adjusted p-value] < 0.05 and absolute fold change ≥ 2) to determine the most significant canonical pathways and biological networks involved in mccRCC. Our analysis revealed highly significant overlap of 198 canonical pathways (p < 0.05) connected with apoptosis, cancer, cell cycle regulation, cellular immune response, and cellular growth, proliferation, and development. The most significant of these are presented in Table 1.
**Table 1:** Top significantly enriched canonical pathways in mccRCC (Caki-1) by IPA.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>-log(p-value)</th>
<th>Ratio</th>
<th>Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILK Signaling</td>
<td>1.1E01</td>
<td>2.6E-01</td>
<td>FLNB,FN1,MYH9,MYL6,BMP2,ILK,MYL6B,ITGB8,PPP1R4B,PDGFC,ITGB3,FLNA,ATF4,IRS2,VCL,ITGB4,ACTG2,CTNNB1,DSP,ACTN1,ITGB1,FBLIM1,LIMS2, CFL1,CASP3,ROH,PPP2R5D,ACTN3,CREBBP,VEGFC,MYL9,DOCK1,ITGB2,FO,S,CDH1,ROHQ,RND3,LIMS1,CDC42,ROHA,SH2B2,RSU1,PIK3CD,ACTN2,ITGB6, ACTG1,MMP9</td>
</tr>
<tr>
<td>Leukocyte Extravasation Signaling</td>
<td>7.93E00</td>
<td>2.23E-01</td>
<td>CD99,MMP7,ICAM1,MYL6,PTK2B,MMP14,MMP15,WASL,ABL1,CLDN6,MAPK11,ITGB3,CLDN4,PRKCE,VCL,ACTG2,CTNNB1,MMP1,VASP,ACTN1,PRKCA,ITGB1,CXCR4,ACTN3,ITGA5,THY1,RAPGEF3,GNAl2,ITGB2,PRKCI,CLDN1,CDC42, VAV3,CLDN16,ROHA,NCF2,PIK3CD,CLDN2,ACTN4,ACTG1,MMP9,CLDN3,MSN</td>
</tr>
<tr>
<td>IGF-1 Signaling</td>
<td>4.01E00</td>
<td>2.16E-01</td>
<td>IGFBP6,SOCS3,CTGF,YWHAH,RRAS,YWHAZ,HRAS,IGFBP5,IGFBP7,IGFBP2,Y WHA,FOS,PRKCINOV,RPS6KB2,IGFBP3,SOCS2,IRS2,PIK3CD,SNF,MAP2K1</td>
</tr>
<tr>
<td>Molecular Mechanisms of Cancer</td>
<td>3.54E00</td>
<td>1.45E-01</td>
<td>CDKN2A,WNT3,SMAD3,NCSTN,HRAS,BBC3,BIRC3,FZD2,CASP3,RRAS,CREBBP,CDK6,ITGA5,RAPGEF3,AURKA,RA3,CDH1,ROHQ,RND3,CDC42,ROHA,ARH GEF18,PIK3CD,BMP2,NFKBIE,CDK4,ABL1,FZD1,MAPK11,SMO,PRKCE,BID,CT NNB1,MAP2K1,BMP1,PRKCA,ITGB1,ROHC,ADCY3,ADCY6,BAK1,GNAl2,FAD D,FOS,PLCB4,PRKCI,BMP7,RPB1,CDK11B,GLI1,FZD7,WNT5A</td>
</tr>
<tr>
<td>CXCR4 Signaling</td>
<td>3.48E00</td>
<td>1.79E-01</td>
<td>MYL6,RRAS,ROHC,CXCR4,ERG1,ADCY3,ADCY6,HRAS,MYL6B,GNAl2,MYL9,GNB1,FOS,DOCK1,GNB4,PLCB4,ROHQ,PRKCI,RND3,ROHA,ITPR3,PRKCE,PIK3 CD,GNG5,MAP2K1,PRKCA,MYL12A</td>
</tr>
<tr>
<td>Bladder Cancer Signaling</td>
<td>3.34E00</td>
<td>2.09E-01</td>
<td>CDKN2A,CXCL8,MMP7,RRAS,FGF2,MMP14,MMP15,ABL1,CDK4,HRAS,VEGFC,PDGFC,FGF1,CDH1,MAP2K1,MMP9,PIK3CD,EGF</td>
</tr>
<tr>
<td>ERK5 Signaling</td>
<td>2.99E00</td>
<td>2.22E-01</td>
<td>YWHAH,RRAS,SGK1,CREBBP,YWHAZ,HRAS,YWHAQ,CTF1,FOS,RPS6KB2,AT F4,SNF,WNK1,EGF</td>
</tr>
<tr>
<td>PI3K/AKT Signaling</td>
<td>2.67E00</td>
<td>1.74E-01</td>
<td>ITGB1,CDC37,YWHAH,RRAS,NFKBIE,PPP2R5D,GDF15,ILK,YWHAZ,HRAS,ITG A5,YWHAQ,SYNJ2,GYS1,LIMS1,RPS6KB2,PIK3CD,CDK11B,SNF,CTNNB1,MA P2K1</td>
</tr>
<tr>
<td>JAK/Stat Signaling</td>
<td>2.41E00</td>
<td>1.94E-01</td>
<td>STAT4,FOS,SOCS3,STAT6,RRAS,CISH,SOCS2,HRAS,PIK3CD,STAT2,IL6,CEBPB ,STAT1,MAP2K1</td>
</tr>
<tr>
<td>VEGF Signaling</td>
<td>2.32E00</td>
<td>1.8E-01</td>
<td>PTK2B,RRAS,ACTN3,CREBBP,HRAS,PDGFC,PIK3CD,ACTN4,VCL,ACTG2,SNF,A CTG1,MAP2K1,EIF1AY,ACTN1,PRKCA</td>
</tr>
<tr>
<td>HIF1α Signaling</td>
<td>2.18E00</td>
<td>1.7E-01</td>
<td>MMP7,SLC2A1,RRAS,MMP14,CREBBP,MMP15,VEGFC,HRAS,PDGFC,MAPK11,</td>
</tr>
<tr>
<td>Pathway</td>
<td>p-value</td>
<td>q-value</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>PDGF Signaling</td>
<td>2.14E00</td>
<td>1.82E-01</td>
<td></td>
</tr>
<tr>
<td>VDR/RXR Activation</td>
<td>1.76E00</td>
<td>1.69E-01</td>
<td></td>
</tr>
<tr>
<td>Renal Cell Carcinoma Signaling</td>
<td>1.76E00</td>
<td>1.74E-01</td>
<td></td>
</tr>
<tr>
<td>EGF Signaling</td>
<td>1.62E00</td>
<td>1.79E-01</td>
<td></td>
</tr>
<tr>
<td>PPAR Signaling</td>
<td>1.57E00</td>
<td>1.56E-01</td>
<td></td>
</tr>
<tr>
<td>mTOR Signaling</td>
<td>1.50E00</td>
<td>1.32E-01</td>
<td></td>
</tr>
<tr>
<td>STAT3 Pathway</td>
<td>1.25E00</td>
<td>1.51E-01</td>
<td></td>
</tr>
<tr>
<td>TGF-β Signaling</td>
<td>1.09E00</td>
<td>1.38E-01</td>
<td></td>
</tr>
<tr>
<td>TREM1 Signaling</td>
<td>1.05E00</td>
<td>1.43E-01</td>
<td></td>
</tr>
<tr>
<td>Prostate Cancer Signaling</td>
<td>1.02E00</td>
<td>1.38E-01</td>
<td></td>
</tr>
<tr>
<td>Insulin Receptor Signaling</td>
<td>9.95E-01</td>
<td>1.25E-01</td>
<td></td>
</tr>
<tr>
<td>IL-6 Signaling</td>
<td>8.35E-01</td>
<td>1.21E-01</td>
<td></td>
</tr>
<tr>
<td>AMPK Signaling</td>
<td>8.09E-01</td>
<td>1.13E-01</td>
<td></td>
</tr>
<tr>
<td>Epithelial Adherens Junction Signaling</td>
<td>8.21E00</td>
<td>2.52E-01</td>
<td></td>
</tr>
</tbody>
</table>

**PDGF Signaling**
- **RRAS, MAP3K1, ABL1, HRAS, PDGFC, PDGFB, SYNJ2, FOS, CAV1, SPHK1, PIK3CD, STAT1, MAP2K1, PRKCA**

**VDR/RXR Activation**
- **IGFBP6, PRKCI, SPP1, GADD45A, IGFBP3, PRKCE, IGFBP5, CDKN1B, SEMA3B, NCO2, CEBPB, THBD, PRKCA**

**Renal Cell Carcinoma Signaling**
- **MET, FOS, SL2CA1, RRAS, CDC42, RBX1, CREBBP, HRAS, PIK3CD, MAP2K1, TCEB1, PDGFB**

**EGF Signaling**
- **FOS, MAP3K1, ITPR3, HRAS, PIK3CD, STAT1, MAP2K1, MAPK11, PRKCA, EGFR**

**PPAR Signaling**
- **RRAS, NFKBIE, CREBBP, NR1H3, HRAS, PDGFC, PDGFB, NR2F1, FOS, IL18, NCO2, NRIP1, MAP2K1, CITED2**

**mTOR Signaling**
- **PLD2, DDIT4, RHOOC, RRAS, PPP2R5D, HRAS, EIF4G3, VEGFC, FKBP1A, EIF4A2, PDGFC, HMOX1, RPS4Y1, PRKCI, RHOQ, RND3, ROHA, RPS27L, RPS6KB2, PRKCE, RPS4Y2, PIK3CD, PRKCA, EIF4B**

**STAT3 Pathway**
- **MAP3K12, SOCS3, RRAS, FGFR1, CISH, FGFR2, HRAS, SOCS2, MAP2K1, MAPK11, EGFR**

**TGF-β Signaling**
- **FOS, RRAS, CDC42, SMAD3, BMP2, CREBBP, HRAS, BMP7, SERPINE1, MAP2K1, MAPK11, PMEPA1**

**TREM1 Signaling**
- **ITGB1, CXCL8, IL18, IJCAM1, CCL2, CASP1, ITGA5, CD86, IL6, CASP5**

**Prostate Cancer Signaling**
- **RRAS, NFKBIE, CREBBP, ABL1, ATF4, HRAS, PIK3CD, CDKN1B, CTNNB1, MAP2K1, GSTP1**

**Insulin Receptor Signaling**
- **PPP1R14C, SOCS3, SGK1, RRAS, HRAS, PPP1R14B, SYNJ2, PRKCI, GYS1, RHOQ, SH2B2, RPS6KB2, ASIC1, IRS2, PIK3CD, MAP2K1**

**IL-6 Signaling**
- **CXCL8, SOCS3, TNFAIP6, RRAS, NFKBIE, HRAS, CEBPB, IL6, MAPK11, COL1A1, FOS, IL18, PIK3CD, MAP2K1**

**AMPK Signaling**
- **PFKFB3, SL2CA1, CPT1B, PPP2R5D, CREBBP, PFKP, MAPK11, SMARCA4, SMARCD3, PFKM, CCNA2, GYS1, FASN, ATF4, IRS2, SMARCC2, PIK3CD, SMARCC1, HMGR, ADRB2**

**Epithelial Adherens Junction Signaling**
- **MYH9, ARP1C1B, MYL6, WASL, HRAS, MYL6B, TUBB, CLIP1, ACTR3, TUBA1C, ACTG2, VCL, CTNNB1, ACTN1, EGFR, TUBB3, NOTCH3, RRAS, TUBB4B, FGFR1, TUBB2A, ACTN3, TUBA4A, FGFL, MYL9, MET, CDH2, CDH1, TUBA1A, TUBB6, TUBB8, CDC42, RHOA, ACTN4, PARD3, ACTG1**
Interaction network of differentially expressed genes (IPA)

IPA identified significant networks associated with the differentially expressed genes in the Caki-1 cell line. These networks were scored based on the number of genes participating in any particular network. We identified 17 gene networks with scores from 25 to 86 genes. In Figure 2, we present the top four gene networks and their associated functions (A) cellular growth and proliferation, organismal survival, and cellular movement; (B) inflammatory response, cell death and survival, and cellular movement; (C) cellular movement, immune cell trafficking, and hematological system development and function; and (D) cell cycle, cellular growth and proliferation, and cell death and survival.

Top upstream transcriptional regulators

IPA upstream functional analysis was used to predict the top upstream transcriptional regulators from differentially expressed genes of the mccRCC cell line. These predictions are based on the literature compiled in the IPKB. The analysis examined how many known targets of the upstream regulators were present in the mccRCC (Caki-1) cells. An overlap p-value was computed based on significant overlap between genes in the dataset and known targets regulated by the transcriptional regulator. The activation z-score algorithm was used to make predictions. IPA predicted top transcriptional regulators that were activated in our dataset, such as estrogen receptor (ER) (z-score = 2.552; overlap p = 3.36E-23), TP53 (z-score = 4.667; overlap p = 1.03E-22), KDM5B (z-score = 2.233; overlap p = 1.38E-08), SPDEF (z-score = 2.673; overlap p = 1.00E-07), and CDKN1A (z-score = 2.108; overlap p = 1.12E-07). Furthermore, TGFB1 (z-score = −3.909; overlap p = 3.99E-22), ERBB2 (z-score = −2.872; overlap p = 1.20E-18), IL13 (z-score = −2.682; overlap p = 9.54E-10), and ESR1 (z-score = −2.05; overlap p = 8.28E-12) were predicted to be inhibited in our dataset.

Correlation of mccRCC with other diseases

To evaluate the positive correlation between mccRCC and other diseases, we compared the differentially expressed genes of mccRCC (Caki-1) with other disease gene sets. The comparison was performed with the existing genomic data in IPKB. The differentially expressed genes found in mccRCC (Caki-1) were highly associated with cancers such as urogenital cancer (p = 8.78E-07; 656 molecules), renal cancer and tumors (p = 4.08E-07; 156 molecules), and prostate cancer and tumors (p = 6.93E-05; 150 molecules); and functions
such as migration of cells ($p = 2.18 \times 10^{-21}$; 236 molecules), expression of RNA ($p = 7.94 \times 10^{-14}$; 268 molecules), and invasion of tumor cell lines ($p = 4.59 \times 10^{-20}$; 151 molecules) (Figure 3).

**GSEA analysis**

We also used GSEA, a pathway enrichment method that evaluates microarray data at the level of gene sets (70), to analyze differentially expressed genes. We found 237 gene sets that were significantly enriched (FDR $< 0.25$) for differentially expressed genes in mccRCC (Caki-1) as compared with healthy kidney cells (ASE). The full list of significantly enriched gene sets can be found in supplementary table S2. Of these, 224 gene sets were significantly upregulated and 13 gene sets were downregulated in the Caki-1 cell line. An overview of deregulated enriched gene sets is presented in Table 2 and Figure 4. Furthermore, we performed leading-edge analysis with the 237 significantly enriched gene sets (cutoff FDR $< 0.25$) in GSEA. The most significant oncogenic signature (C6) pathways are presented as a heat map in Figure 5.

**Enriched gene ontology (GO) analysis**

The main enriched GO terms were categorized on the basis of molecular function and biological process to determine which were enriched in the differentially expressed genes found in mccRCC (Caki-1) (Figure 6). All the genes were placed into broad functional categories on the basis of the PANTHER database. A total of 1921 (up– and downregulated) genes were grouped into several classes according to their major molecular functions and biological processes. The major molecular functions associated with these differentially expressed genes were catalytic activity (28.80%) (GO:0003824), ligands binding (31.70%) (GO:0005488), transporter activity (4.60%) (GO:0005215), receptor activity (6.30%) (GO:0004872), and nucleic acid–binding transcriptional factors (7.00%) (GO:0001071). Some biological processes that seemed to be enriched in the differentially expressed genes of mccRCC were cellular process (37.80%) (GO:0009987), apoptotic process (3.50%) (GO:0006915), biological regulation (19.50%) (GO:0065007), metabolic process (43.50%) (GO:0008152), and immune system process (7.50%) (GO:0002376).
Table 2. Top deregulated gene sets from GSEA analysis: ES: enrichment score; p-value < 0.05; FDR < 0.25; NES: normalized enrichment score; C2: curated gene sets; C5: GO gene sets; C6: oncogenic signatures gene sets; C7: immunologic signatures.

<table>
<thead>
<tr>
<th>GSEA set name</th>
<th>Biological features</th>
<th>MSigDB</th>
<th>Number of genes in set</th>
<th>Gene count</th>
<th>ES (ASE vs Caki-1)</th>
<th>NES (ASE vs Caki-1)</th>
<th>Nominal p-value</th>
<th>FDR q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>REACTOME_HEMOSTASIS</td>
<td>Genes involved in hemostasis</td>
<td>C2</td>
<td>466</td>
<td>80</td>
<td>0.21</td>
<td>1.96</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>REACTOME_CLASS_A1_RHODOPSIN LIKE_RECEPTORS</td>
<td>Genes involved in Class A/1 (Rhodopsin-like receptors)</td>
<td>C2</td>
<td>305</td>
<td>40</td>
<td>-0.24</td>
<td>-1.71</td>
<td>0.04</td>
<td>0.23</td>
</tr>
<tr>
<td>NEGATIVE_REGULATION_OF_BIOLOGICAL_PROCESS</td>
<td>Any process that stops, prevents, or reduces the frequency, rate, or extent of a biological process</td>
<td>C5</td>
<td>677</td>
<td>120</td>
<td>0.18</td>
<td>1.98</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>AKT_UP_MTOR_DN.V1_DN</td>
<td>Genes downregulated by everolimus in mouse prostate tissue transgenically expressing human AKT1 gene</td>
<td>C6</td>
<td>183</td>
<td>37</td>
<td>0.29</td>
<td>1.95</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>MEK_UP.V1_UP</td>
<td>Genes upregulated in MCF-7 cells (breast cancer) positive for ESR1</td>
<td>C6</td>
<td>196</td>
<td>49</td>
<td>0.25</td>
<td>1.94</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>P53_DN.V1_DN</td>
<td>Genes downregulated in NCI-60 panel of cell lines with mutated TP53</td>
<td>C6</td>
<td>192</td>
<td>66</td>
<td>-0.21</td>
<td>-1.86</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>ESC_V6.5_UP_LATE.V1_UP</td>
<td>Genes upregulated during late stages</td>
<td>C6</td>
<td>190</td>
<td>28</td>
<td>-0.31</td>
<td>-1.87</td>
<td>0.01</td>
<td>0.17</td>
</tr>
<tr>
<td>Dataset ID</td>
<td>Description</td>
<td>Genes downregulated</td>
<td>Genes upregulated</td>
<td>p-value</td>
<td>fold change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>-------------------</td>
<td>---------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL2_UP.V1_UP</td>
<td>Genes downregulated in Sez-4 cells (T lymphocyte)</td>
<td>C6</td>
<td>192</td>
<td>30</td>
<td>0.28</td>
<td>1.72</td>
<td>0.03</td>
<td>0.10</td>
</tr>
<tr>
<td>KRAS.KIDNEY_UP.V1_UP</td>
<td>Genes upregulated in epithelial kidney cancer cell lines overexpressing an oncogenic form of KRAS</td>
<td>C6</td>
<td>195</td>
<td>45</td>
<td>-0.22</td>
<td>-1.63</td>
<td>0.04</td>
<td>0.22</td>
</tr>
<tr>
<td>GSE7460_CD8_TCELL_VS_CD4_TCELL_ACT_DN</td>
<td>Genes downregulated in comparison of ActCD8 versus ActCD4</td>
<td>C7</td>
<td>200</td>
<td>29</td>
<td>0.31</td>
<td>1.86</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>GSE18791_CTRL_VS_NEWCASTLE_VIRUS_DC_8H_DN</td>
<td>Genes downregulated in comparison of control conventional dendritic cells (cDC)</td>
<td>C7</td>
<td>200</td>
<td>37</td>
<td>-0.30</td>
<td>-2.04</td>
<td>0.01</td>
<td>0.24</td>
</tr>
<tr>
<td>GSE1460_NAIVE_CD4_TCELL_ADULT_BLOOD_VS_THYMIC_STROMAL_CELL_DN</td>
<td>Genes downregulated in naive CD4 T cells from adult blood versus thymic stromal cells</td>
<td>C7</td>
<td>200</td>
<td>45</td>
<td>0.25</td>
<td>1.82</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>GSE30962_PRIMARY_VS_SECONDARY_CHRONIC_LCMV_INF_CD8_TCELL_UP</td>
<td>Genes downregulated in comparison of virus specific (gp33) exhausted CD8 T cells</td>
<td>C7</td>
<td>199</td>
<td>26</td>
<td>0.33</td>
<td>1.93</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>GSE13229_IMM_VS_MATURE_NKCELL_UP</td>
<td>Genes upregulated in comparison of immature NK cells</td>
<td>C7</td>
<td>200</td>
<td>27</td>
<td>0.31</td>
<td>1.76</td>
<td>0.02</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>versus mature NK cells</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>222</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>223</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>224</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>225</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>226</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>227</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>228</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>229</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>230</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>231</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>232</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>233</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>234</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>235</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>236</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>237</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>238</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>239</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

The evolution from primary tumor to metastatic disease is a multistep process that requires primary tumor cells to detach, survive during circulation, attach to a distant organ, adapt to the new environment, and outgrow the secondary lesion to develop metastasis (10, 41). Patients with mRCC are difficult to manage clinically because the available therapies lack significant efficacy (23). Clear cell renal cell carcinomas (ccRCC) are the predominant type of RCC characterized by chemotherapy or radiation-therapy resistance. In the present study, we used Caki-1, a ccRCC cell line derived from a metastatic site in the skin (24), for a gene expression study in comparison with a healthy kidney epithelial cell line (ASE) (14) derived from purified epithelial cells from whole kidneys. We chose the mccRCC Caki-1 cell line, which has wild-type VHL status (21), while other cell lines, such as 786-O and Caki-2, have a mutated VHL gene and represent cell lines derived from primary ccRCC. The other reason for selecting Caki-1 cells as a good model in this research is the cell line’s capability to develop tumors in many xenograft studies (15, 26, 58). These tumors, moreover, have been found to be highly resistant to sunitinib treatment (35). A number of differentially expressed genes in the Caki-1 cell line can be found in supplementary table S1. For example, chemokine receptor CXCR4 was reported to be 2.01-fold upregulated in the Caki-1 cell line; this was previously reported as a cancer stem cell marker, and CXCR4-positive cells exhibit great resistance to tyrosine kinase inhibitors (27, 42, 47). In addition, CXCR4 expression has significant prognostic value and therapeutic importance in RCC patients (13). IL6 expression, meanwhile, was 19.88-fold upregulated in the Caki-1 cell line. Similar findings were reported by Walther et al. (74) and Angelo et al. (1), who found frequent secretion of IL6 by renal cancer cell lines and elevated serum IL6 levels in patients with mRCC (40, 74). STATs are a family of cytoplasmic proteins that take part in transcription of genes involved in diverse cellular activities in diseased states (8, 44), and the expression of STAT1, STAT2, STAT4, and STAT6 genes was observed to be elevated (4.86, 4.25, 3.48, and 3.03, respectively) in Caki-1 cells. Our findings suggest that activated STATs in mccRCC could promote tumorigenesis by preventing apoptosis and enhancing proliferation, angiogenesis, invasiveness, and immune response evasion. Although many genes were differentially expressed in the mccRCC Caki-1 cell line in this study, it must be recalled that in many tumors, metastatic genes are principally tumor type– and organ site–specific (48, 63, 72).

Disruption of multiple biological pathways is a hallmark of many tumors, including mccRCC. Our data illustrated in this study show the complexity of mccRCC and the extent to
which genes are significantly involved in different pathways and functions. Many of these
gen genes are associated with survival and provide insight into the broader carcinogenic process.
Our data from IPA illustrate a number of key pathways and their involvement in renal
carcinogenic process. Many of these pathways comprise genes involved in the cell cycle, cell
growth, apoptosis, and differentiation (Table 1). One of the top pathways that was
significantly expressed by genes from Caki-1 cells was ILK signaling. ILK is a serine-
threonine protein kinase involved in the cell cycle, metastasis, migration, angiogenesis, and
tumor cell proliferation (36). Knockdown of ILK was shown to inhibit cell migration and
invasion in the Caki-1 cell line and downregulation of epithelial–mesenchymal transition
(EMT) markers in an in vivo study (34). Another recent study showed that in ccRCC patients,
ILK expression was significantly increased in high-grade tumors compared with low-grade
tumors (17). Furthermore, tumors with positive ILK expression showed a higher proliferation
index than tumors with negative ILK expression. Upregulation of ILK was reported to have a
role in regulating resistance to gemcitabine in lung cancer by promoting EMT and cellular
drug efflux (39). Since mRCC is a drug-resistant tumor, these findings could make ILK an
attractive future target for resistant-mRCC therapeutics. Other canonical pathways revealed
by IPA analysis in Caki-1 cells were leukocyte extravasation signaling, IGF-I signaling,
CXCR4 signaling, bladder cancer signaling, ERK5 signaling, PI3K/AKT/mTOR signaling,
JAK-STAT signaling, and VEGF signaling. An in vitro study of the IGF signaling system
confirmed that mRCC cells strongly expressed IGF-I and IGFBP-3 (11). Furthermore, in
mRCC, autocrine IGF-I and IGFBP-3 act to stimulate and inhibit growth, respectively. Thus
far, insulin and IGF-I signaling have not been used as biomarkers for RCC screening or for
clinical trials that could be proven to be beneficial for treatment of mRCC patients (65).
Another significant pathway that was reported with mRCC in our dataset using both GSEA
and IPA enrichment was PI3K/AKT/mTOR. This pathway has been studied with regard to
mRCC in several studies (7, 22, 60).

An important gene network was identified involving the AKT, CCND1, and HIF1A1
genes (Figure 2A). Activated AKT has been reported in conjunction with decreased PTEN
expression, which is specific for ccRCC (37). AKT inhibitor–induced apoptosis in ccRCC
(37) could be used as a therapeutic option for a subset of mRCC patients with elevated
AKT activity. CCND1 is an oncogene that is commonly upregulated in different cancers and
act as a HIF-regulated gene in RCC cell lines (55, 64). The unusual network between CCND1
and HIF1A1 in Caki-1 cells with wild-type VHL might suggest an intrinsic property of cells
in the renal tubular epithelium. We found a second network (Figure 2B) around SMARCA4,
a gene in the SWI/SNF family. SMARCA4 has a tumor-suppressive function in many solid tumors, and its inhibition leads to apoptosis, cell cycle changes, and myeloid differentiation in leukemic cells (12). However, the relationship between SMARCA4 and mccRCC is still unknown, so further investigation is needed. A third important network (Figure 2C) involved tumor necrosis factor-α (TNF-α), extracellular signal–regulated kinase (ERK), and chemokine CXCL8. TNF-α and CXCL8 and their legends promote tumorigenesis by facilitating tumor proliferation, angiogenesis, and metastasis. In ccRCC, TNF-α is responsible for inducing epithelial-mesenchymal transition via the activation of nuclear factor kappa B (NF-κB) (76). The ERK signaling pathway is often upregulated in human tumors, and blockage of ERK pathway is an attractive target for developers of anticancer drugs (43). Anti-metastatic activity in RCC cells can be halted by inhibiting the ERK pathway (20). The fourth important network we found was mainly concentrated around tumor suppressor gene TP53 (Figure 2D).

We used IPA upstream analysis to identify the cascade of upstream transcriptional regulators in observed gene expression changes in our data sets. This analysis predicted which transcriptional regulators were involved and whether they were likely activated or inhibited. The differentially expressed genes were associated with upstream regulators that had previously been associated with other tumors, such as estrogen receptors (ER), TP53, KDM5B, SPDEF, and CDKN1A (32, 38, 59, 66, 68). In our dataset, these transcriptional regulators were activated and influenced the downstream genes and pathways they regulate. Expression of ER-β has been found in RCC cell lines (77). Moreover, proliferation of RCC cells has been seen to decrease significantly after treatment with estrogen (17-β-estradiol, E2) via ER-β activation. Consequently, ER activation in Caki-1 could suggest its importance for mccRCC prognosis and novel therapeutics in treatment of RCC. P53 is one of the most widely studied tumor suppressor proteins in human cancer (54). The Caki-1 mccRCC cell line has wild-type TP53 status, and TP53 is rarely mutated in RCC (33). However, an increase in TP53 expression—without mutation—has been linked with reduced overall survival and more rapid disease progression in RCC (56). Upregulated expression of KDM5B (JARID1B) has been found in various cancers, including RCC (38). Inhibition of KDM5B affects apoptosis and reduces growth of bladder cancer cells through the E2F-RB pathway (38). More specific research on RCC would be needed before KDM5B could be identified as a targeting molecule, however.

We used the GSEA approach to further confirm enriched pathway data following IPA. When comparing the mccRCC Caki-1 cell line with the healthy kidney ASE cell line, many gene sets were significantly upregulated in Caki-1. This phenomenon seems to be more
obvious in metastatic RCC. This might suggest that many upregulated pathways are responsible for invasion, proliferation, invasiveness, and resistance against various cancer therapies (Table 2 and supplementary table S2). For example, activating transcription factor 2 (ATF2), which is included in the gene set ATF2_S_UP.V1_UP, has been reported to be involved in DNA damage and skin tumor development (5). The CAMP_UP.V1_DN gene set regulates the cAMP signaling pathway, and mutations in members of the cAMP signaling cascade are responsible for tumor formation (73). The RB_P130_DN.V1_UP gene set was also found to be upregulated in mccRCC cells. Rb is a tumor suppressor gene that plays an important role in negative regulation of the cell cycle and tumor progression (28) and might have therapeutic potential for mccRCC in the future.

We also looked for similarities and differences between our Caki-1 cell line data and the molecular data of ccRCC patients obtained by The Cancer Genome Atlas Research Network (TCGA) (9), including genes, pathways, and networks. Among the top 50 most significantly mutated genes in ccRCC patients, we found five (TXNIP, TCEB1, NPNT, CCNB2, FBF2) that also appeared as differentially expressed genes in our Caki-1 cell line data. The gene for tumor protein p53 was found to be mutated in the TCGA data, while in the Caki-1 cell line, this gene was shown to be an important upstream transcriptional regulator. Furthermore, cyclin-dependent kinase inhibitor 2A (CDKN2A), a tumor suppressor gene, was downregulated (−3.08) in Caki-1 cells; this gene was somatically altered in ccRCC patients, according to TCGA data. A difference was observed in two histone demethylase gene isoforms (KDM5B and KDM5C) from the JARID family: KDM5B (JARID1B) was reported as an activated upstream transcriptional regulator in Caki-1 cells, while KDM5C (JARID1C) was among the mutated genes in ccRCC patients. Members of the ARID family of DNA binding proteins (ARID3A, ARID3B, ARID5A, ARID5B) and the SMARCA4 gene were downregulated in Caki-1 cells, while isoforms ARID1A and SMARCA4 were included in a frequently mutated subnetwork (9) that participates in the PBAF SWI/SNF chromatin remodeling complex (30). To gain insights into common enriched pathways reported in both TCGA ccRCC patients and Caki-1 cells, we examined the most significantly enriched pathways that we found by IPA and GSEA. TCGA reported the top 25 enriched pathways among patients, which matched up well with seven important pathways reported in our study: p53, HIF-1 and 2 alpha, ATF2, TGF-β–mediated signaling via SMAD2-3 pathway, PDGF-beta signaling, and IL-6–mediated signaling. Additionally, we found that the PI3K/AKT/mTOR pathway reported in Caki-1 cells was also reported as altered in 28% of tumors from patients with ccRCC (9). Reduced AMP-activated kinase (AMPK) and increased
acetyl-CoA carboxylase (ACC) were both correlated with poor prognosis and worse survival in ccRCC patients in the TCGA data. We also found significant alterations to the AMPK pathway in the Caki-1 cell line. AMPK, as discussed previously, is a cellular sensor that negatively regulates mTOR signaling pathways. Together, decreased levels of AMPK and increased levels of ACC could result in a metabolic shift toward increased fatty acid synthesis (71). Major differences between our data and the TCGA data were observed in genes involved in poor prognosis correlated with upregulation of genes involved in the pentose phosphate pathway (TALDO1) and fatty acid synthesis (FSN); both genes were downregulated in the Caki-1 cell line (−2.05 and −2.70, respectively). Taking all of this into account, the data presented in this research shows that mccRCC manipulates more than one molecular mechanism and that multiple pathways are linked with the aggressiveness of the disease. To better understand mccRCC, we must take into consideration different altered mechanisms that could provide the foundation for novel biomarkers and development of new targeted therapies.

Acknowledgments
This study was supported by the Foundation for Polish Science TEAM project TEAM/2010-6/8. The authors acknowledge the support of the Scribendi, Inc. for professional editing and proofreading of this manuscript.

Disclosures
All authors declare no competing commercial interests in relation to the submitted work.

References


**Figure 1:** (A) Volcano plot showing relation between magnitude of gene expression change (log2 of fold change; x-axis) and statistical significance of this change (−log10 of FDR; y-axis) when comparing Caki-1 and ASE cell lines. Colored points represent differentially expressed genes (cutoff FDR < 0.05) with magnitude of change equal or greater than 2 that are either overexpressed (red) or underexpressed (blue) in Caki-1 compared with ASE. (B) Principal component analysis (PCA) of microarray data derived from Caki-1 (red triangle) and ASE (blue circle) cell lines. Each mark represents an individual sample. Note that Caki-1 samples are clearly separate from the ASE samples according to the first and second principal components.

**Figure 2:** IPA-identified top four most significant gene networks with score ≥ 25. (A) Cellular growth and proliferation, organismal survival, cellular movement. (B) Inflammatory response, cell death and survival, cellular movement. (C) Cellular movement, immune cell trafficking, hematological system development and function. (D) Cell cycle, cellular growth and proliferation, cell death and survival.

**Figure 3:** Correlation of mccRCC with other human diseases.

**Figure 4:** Gene set enrichment analysis (GSEA)-enrichment plots of representative gene sets from Table 2: (A) reactome hemostasis, (B) negative regulation of biological process, (C) oncogenic pathway AKT/PI3K, (D) oncogenic KRAS signaling.

**Figure 5:** Heat map of leading oncogenic pathways and their corresponding genes. The heat map presented here shows only the upregulated genes (red). Genes are shown on the vertical bars colored from deep red (top rank) to pink (lowest rank).

**Figure 6:** Pie chart of most significant enriched gene ontologies of differentially expressed genes of mccRCC (Caki-1) based on Figure 6(A), molecular function (MF), and Figure 6(B), biological process. The number indicates a count of altered genes that fall into a certain category together with the corresponding gene ontology ID number.