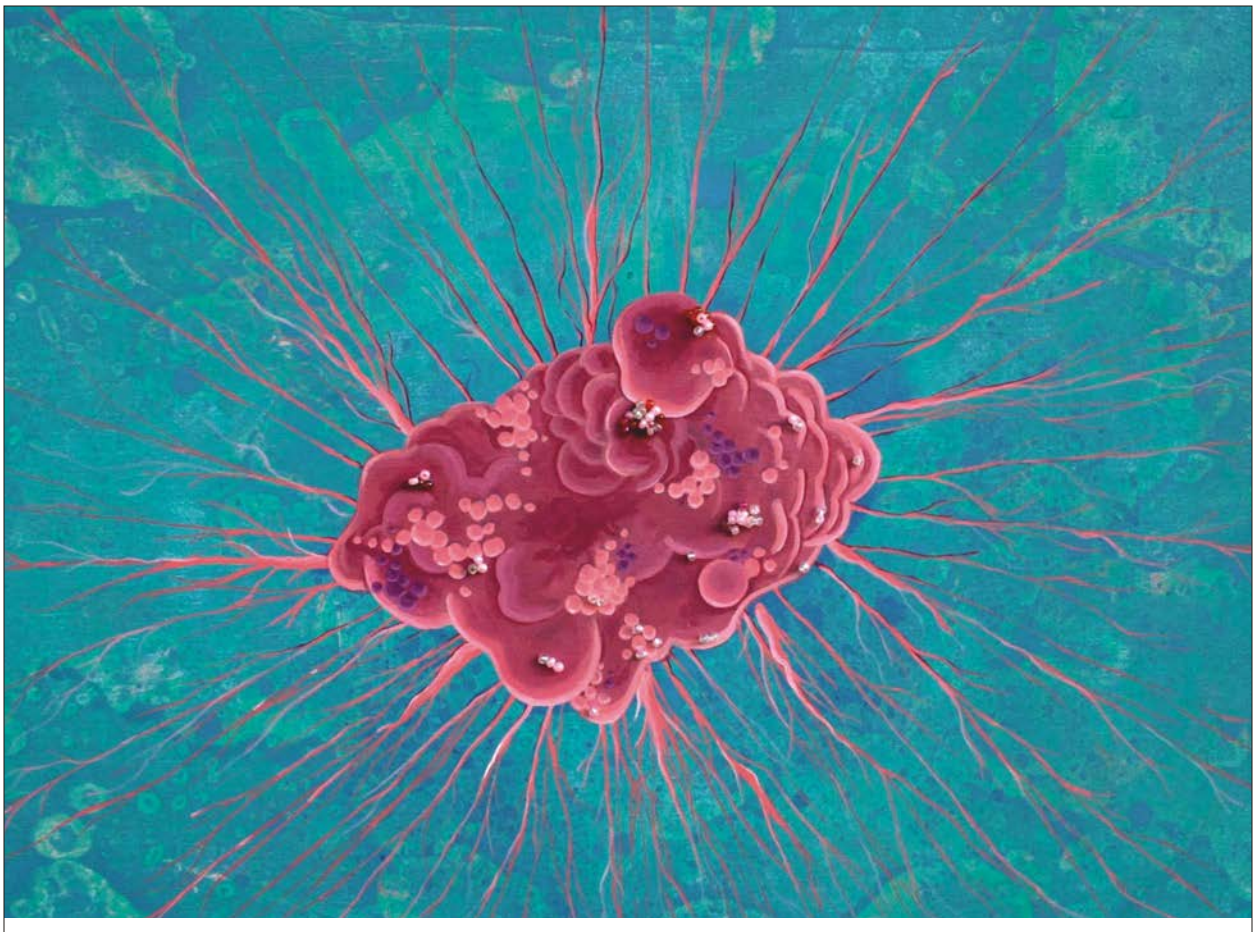

International Journal of Cancer Research

ISSN 1811-9727

Academic Journals Inc.

www.academicjournals.com

International Journal of Cancer Research 4 (4): 137-145, 2008
ISSN 1811-9727

Genomic Distribution, Expression and Pathways of Cancer Metasignature Genes Through Knowledge Based Data Mining

[1,2]Muhummadh Khan and [1]Kaiser Jamil
[1]Genomics and Bioinformatics Lab, Indo American Cancer Institute and Research Centre, Banjara Hills, Hyderabad, 500034, AP India
[2]MGNIRSA, Gagan Mahal, Hyderabad, 500029, AP India

Abstract: In the present study, we tried to integrate information from various sources to understand the role of cancer metasignature genes in initiation and progression of neoplasia. We analyzed these cancer metasignature genes for their chromosomes distribution, expression profiles in normal human tissues and the cellular pathways in which they are involved using the relevant data from the biological databases. It is concluded that cancer metasignature genes are needed for proper functioning and maintenance of normal cellular physiology. We report that multiple numbers of these genes were involved in three cellular processes; cell division cycle, antigen processing and presentation and proteasome dependent proteolysis. They are also involved in a myriad of metabolic and genetic processes. We propose that it is possible that almost all cell types of the human body contain a dormant genetic network which enables them to overcome senescence or evade apoptosis. This network when activated due to some effective trigger may lead to cancer generation and progression.

Key words: Cancer, metasignature genes, expression profiles, chromosomal location, cellular pathways, neoplasm, undifferentiation

INTRODUCTION

Cancer is a generic term for a group of more than 100 diseases that can affect any part of the body (Stewart and Kleihues, 2003). Basically it is caused due to over expression or mutation of proto-oncogene or deactivation of tumor suppressors or failure of DNA repair. Cancer can result from accumulation of mutations and other heritable changes in susceptible cells. So far, abnormalities in about 350 genes have been implicated in human cancers, but the true number of cancer genes is unknown (Higgins *et al*., 2007; Forbes *et al*., 2006). One defining feature of cancer is the rapid creation of abnormal cells which grow beyond their normal tissue boundaries. These cancerous cells can invade adjoining parts of the body and spread to other organs; this process is referred to as metastasis. Metastasis are the major cause of death from cancer.

The battle against cancer has many fronts. But the main focus amongst the researchers has been to understand the origin and progress of this disease at the molecular level. Cancer may arise from mutation in a single cell. The transformation from a normal cell into a tumor cell is a multistage process, typically a progression from a pre-cancerous lesion to malignant tumors. During the progress of the disease, hundreds of different genes are seen activated or deactivated at different times (Bucca *et al*., 2004; Kaiser, 2005; Greenman *et al*., 2007). In 1990s, DNA microarray or DNA chip technology was developed, which enabled researchers to measure the expression levels of hundreds of genes simultaneously. Using this technique, a comprehensive understanding of the cell can be

Corresponding Author: Kaiser Jamil, Genomics and Bioniformatics Lab, Indo American Cancer Institute and Research Centre, Banjara Hills, Hyderabad, 500034, AP India Tel: +91-40-23540348 Fax: +91-40-23542120

achieved (Hanai *et al*., 2006). Studies have been conducted to identify the gene expression pattern in specific cancers like the lung cancer, colon cancer and leukemia etc. (Bucca *et al*., 2004; Kim *et al*., 2005; Halvorsen *et al*., 2005; Talbot *et al*., 2005; De Pitta *et al*., 2005). Meta-analysis of such microarray datasets was done to identify group of genes which were universally activated in all cancers (Rhodes *et al*., 2004). This means that all cancer types share the common features of unregulated cell proliferation and invasion. In the present study, we analyzed the chromosomal distribution, expression profile in normal body tissues and the cellular pathways of the genes of cancer metasignature. Since the gene sets were derived through statistical procedures, we wanted to investigate whether these signature genes yield biological relevant information regarding fundamental aspects of cancer.

MATERIALS AND METHODS

This study involves investigation of chromosomal location, cellular pathways and expression profile of cancer metasignature genes. Using bioinformatics tools, as described in the methodology below, we have carried out this study in the Bioinformatics Lab of our institute.

The metasignature gene lists were obtained from the Rhodes *et al*. (2004). There are two datasets of cancer metasiganature genes; the neoplastic and the undifferentiated. The neoplastic cancer metasignature genes dataset contains a list of 69 genes where as the undifferentiated cancer metasignature genes dataset consists of 67 genes. The gene symbols used in the list were from HUGO Gene Nomenclature Committee (HGNC) database, which is the accepted standard for gene names (Bruford *et al*., 2008). The two metasignature genes datasets were combined to generate a single dataset with no repeating genes. This combined gene list was submitted to web server WEBGESTALT for analyzing their chromosomal distribution in the human genome and their expression in normal body tissues (Zhang *et al*., 2005). These genes were clustered based on the information from KEGG pathway database.

Web Gestalt (WEB-based Gene Set Analysis Toolkit) is a web-based integrated data mining system to help biologists in exploring large sets of genes. It is composed of four modules: gene set management, information retrieval, organization/visualization and statistics. The management module uploads, saves, retrieves and deletes gene sets, as well as performs boolean operations to generate the unions, intersections or differences between different gene sets. The information retrieval module retrieves information for up to 20 attributes for all genes in a gene set. The organization/visualization module organizes and visualizes gene sets in various biological contexts, including Gene Ontology, tissue expression pattern, chromosome distribution, metabolic and signaling pathways, protein domain information and publications. Web Gestalt can be accessed at http://bioinfo.vanderbilt.edu/webgestalt.

RESULTS AND DISCUSSION

Chromosomal Distribution

The existence of a general cancer metasignature may not be entirely surprising, because all cancer types share the common features of unregulated cell proliferation and invasion and it would follow that the genes that are essential to these processes would be highly expressed in multiple cancer types. It is interesting that a small number of genes are almost universally activated, given the vast array of transforming mechanisms that are known to initiate cancer and the variety of tissue types represented. Since the neoplastic and the undifferentiated metasignatures contained an overlap of 16 genes. We combined the metasignatures by removing the repeating 16 genes and analysed their chromosomal distribution (Fig. 1). About 35% of metasignature genes were located on chromosomes 1, 7 and 2, whereas chromosomes 22 and Y did not contain any genes. It is to be noted that about 9 chromosomes (x, 15, 16, 10, 13, 18, 21, 22, Y) contained only 5% of the metasignature genes. According to Fig. 1, the chromosomes 1, 7 and 2 contain high number of genes. But any conclusions drawn from this could

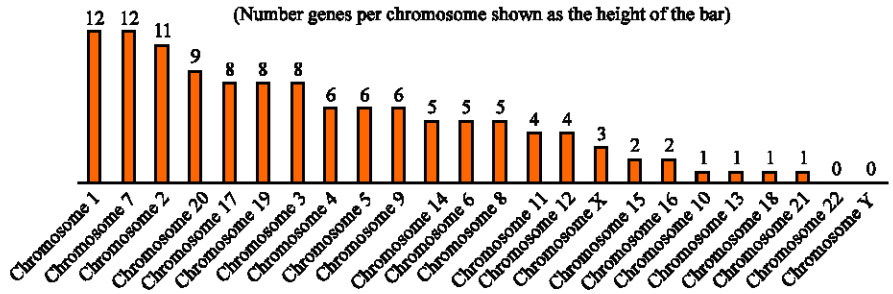

Fig. 1: Chromosomal distribution of the cancer metasignature genes in the human genome

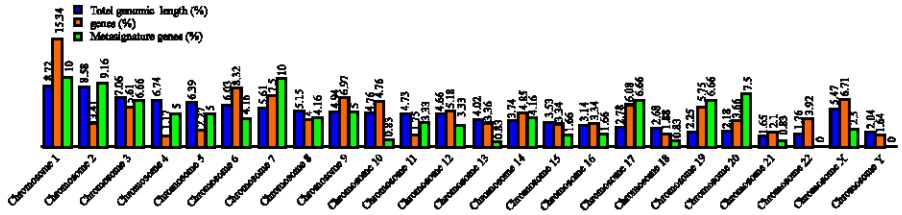

Fig. 2: Percentage of cancer metasignature genes per chromosome compared with the percentages of chromosome length and its gene density. The blue bars represent the percentage of genomic length of the chromosome, the red bars represent the percentage of total number of genes present in the given chromosome and the green bars represent the percentage of cancer metasignature genes in the given chromosomes. All the numbers on the top of the bars are given in percentages

be erroneous since the chromosomes vary widely amongst themselves in content and size. Hence we took into consideration the chromosome length, number of genes it contained and the number of metasignature genes that were located on it for a given chromosome. These numbers were converted into percentages from the total length of human genome i.e., number of base pairs according to the human genome database at NCBI (Wheeler *et al*., 2008) and overall number of genes located on that chromosome from the Vega human genome browser (Wilming *et al*., 2008). These percentages were plotted as graph (Fig. 2).

It is interesting to note that about 30% of the metasignature genes are located on chromosomes 7, 17, 19 and 20 which contribute only 12.5% to the total genomic length and contain about 22% of the total genes. The chromosomes 2, 4, 5 and 11 contain less number of total genes when compared to their genomic lengths. These chromosomes contain about 22% of the metasignature genes. Overall eight chromosomes, irrespective of their genomic length and gene density, contain more than 50% of the metasignature genes.

Many genes exist as gene families which are defined as groups of genes with sequence homology and related overlapping functions. Such genes known to exhibit pattern in their location on the chromosomes like the HOX genes, human α-globin gene cluster, multiple members of the MDR-ADH (MDR: medium-chain dehydrogenases/reductases; ADH: alcohol dehydrogenase) family, flavin-containing monooxygenase genes (Abbasi and Grzeschik, 2007; Hernandez *et al*., 2004; Lai *et al*., 2005: Lim and Bowles, 2004; Tang *et al*., 2006; Gonzalez-Duarte and Albalat, 2005). In the case of the cancer metasignature genes, their chromosomal distribution is biased but it is evident that these genes do not exhibit any related pattern in their chromosomal location. We assume that in cancer progression genes from varied regions of the human genome are expressed and the genomic loci of these genes are certainly not related to the genomic length or the gene density of the chromosomes.

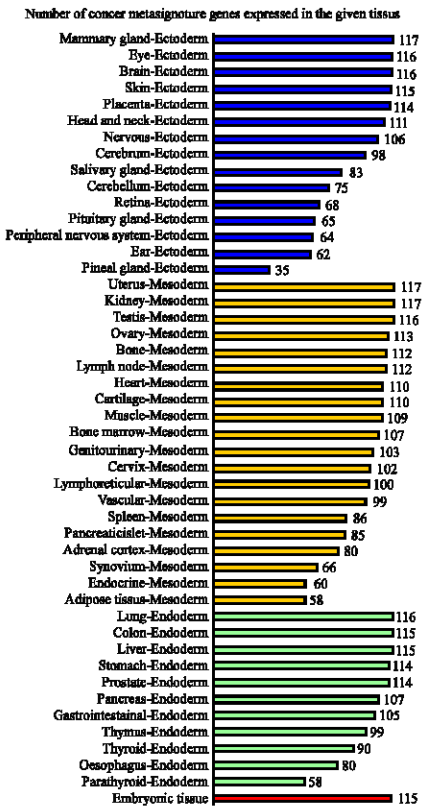

Fig. 3: Expression of the cancer metasignature genes in the normal human body tissues. The tissues were broadly classified based on their embryonic germ layer. The ectoderm derived tissues are colored blue, mesoderm derived tissues are colored orange and the endoderm derived tissues are colored green. The embryonic tissue is colored red. The numbers at the end of the bars represent the number of metasignature genes

Expression in Normal Human Tissues

We analysed the expression of the metasignature genes in the normal tissues. These tissues were classified based on the embryonic germ layers that were derived from; ectoderm, mesoderm and endoderm. We carried out this exercise in order to check any bias in the expression of the metasignature genes in any germ layer derived tissues. There are many different types of cancer. All cancers, however, fall into one of four broad categories based on the tissue of origin. Carcinomas are tumors that arise in the tissues that line the body's organs. These tissues are derived from the ectoderm and the endoderm. About 80% of all cancer cases are carcinomas. Hence we expected biased over expression of the metasignature genes in these tissues. The sarcomas, leukemia's and lymphomas are cancers of the tissues derived from the mesoderm. Sarcomas are tumors that originate in bone, muscle, cartilage, fibrous tissue or fat. Leukemias are cancers of the blood or blood-forming organs and the lymphomas affect the lymphatic system.

We found that the metasignature genes which depict the common transcriptional profile of cancer are expressed in most of the normal tissues (Fig. 3). About 80% of the metasignature genes are expressed in 30 out of 47 tissues. Only 4 tissues showed 50% or less of metasignature genes. Basing on the germ layers it was seen that the ectoderm, mesoderm and endoderm contained similar expression

profile of these genes. Only the ectoderm derived tissue showed a drop in the expression when compared to the other two. Even with this reduced expression levels the most of the tissues showed more than 50% of genes expressed. By this we infer that the cancer metasignature are universal in most of the tissues of the normal human body. It is plausible that almost every cell in the human body has a dormant genetic network which enables it to overcome senescence or evade apoptosis. This may lead to cancer generation and progression through uncontrolled cell division.

KEGG Pathways Based Clustering

KEGG is a database of biological systems, consisting of genetic building blocks of genes and proteins (KEGG GENES), chemical building blocks of both endogenous and exogenous substances (KEGG LIGAND) and molecular wiring diagrams of interaction and reaction networks (KEGG PATHWAY) (Kanehisa *et al*., 2008). KEGG PATHWAY is a collection of manually drawn pathway maps representing the knowledge on the molecular interaction and reaction networks for Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Human Diseases and the structure relationships in Drug Development. We used the information of cellular pathways from this database to cluster the metasignature genes.

The analysis showed that multiple numbers of genes were involved in three cellular processes; cell division cycle, antigen processing and presentation and proteosome dependent proteolysis. There were about 13 genes relating to cell cycle (Table 1). It is known that the cancer cells undergo continuous and uncontrolled cell division. This analysis further confirms this fact since the metasignatures contain prominent genes which are essential for proper functioning of the cell cycle like the Cyclins, Cyclin dependent Kinases, Histone deacetylases, Mitotic checkpoint protein coding gene, Mini-chromosome maintenance protein coding genes and Polo-like Kinases.

The cyclins and cyclin dependent kinases are required for progression through the various stages of the mitotic cell cycle (Malumbres, 2007). Histone deacetylase deacetylates the histones which is important process in chromosomal remodeling (Kim *et al*., 2006). It also interacts with retinoblastoma tumor-suppressor protein and this complex is a key element in the control of cell proliferation and differentiation. Together with metastasis-associated protein-2, it deacetylates p53 and modulates its effect on cell growth and apoptosis (Wolffe, 1996). The mitotic checkpoint protein coding gene (MAD2L1) is essential for preventing the onset of anaphase until all chromosomes are properly aligned at the metaphase plate (Taylor *et al*., 2004). The Mini-Chromosome Maintenance (MCM) proteins are required for establishing pre-replication complexes which trigger the activation of cyclin-dependent kinases for progression of the cell cycle (Forsburg, 2008). Polo-Like Kinases (PLK1) is an important regulator of several events during mitosis. Recent reports show that PLK1 is involved in both G2 and mitotic DNA damage checkpoints (Ferrari, 2006). It can be seen from the above brief description of the genes that the metasignature of cancer contains an essential set of cell cycle related genes.

Table 1: Cellular processes in which the metasignature genes are involved with the No. of genes and their entrez gene id

KEGG pathways of cellular processes

Pathways	No. of genes	Entrez gene IDs
Cell cycle	13	1875, 3065, 4085, 4171, 4172, 4175, 5111, 5347, 890, 891, 983, 990, 991
Antigen processing and presentation	5	10437, 1514, 5721, 6890, 821
Gap junction	3	10382, 203068, 983
Cell communication	1	1278
Focal adhesion	1	1278
Toll-like receptor signaling pathway	1	4283
Natural killer cell mediated cytotoxicity	1	3460
Leukocyte transendothelial migration	1	4318

There are about five genes in the metasignature whose protein products are involved in major Histocompatibility Component (MHC) pathway one and two; where the antigens are presented to the CD8 and CD4 T-cells and the natural killer cell (Goodsell, 2005). It is becoming increasingly clear that MHC may also play a role in the natural control of cancer cells. Cancer cells contain many mutated proteins that may be displayed by MHC to alert the immune system. Tumor cells may also express normal proteins but in unusual places or in abnormal amounts, providing a potential signal to mobilize an immune response. The possibility of enhancing this response with vaccines is an exciting goal of current research (Goodsell, 2005).

The ubiquitin mediated protein modification or degradation through proteasome is a versatile system employed by cell for regulation of various processes like transcription, histone modification and degradation of unwanted or misfolded proteins (Hershko, 2005). There are about seven genes in the metasignature which code for proteins required for this system. This pathway comes under the genetic information processing KEGG pathways (Table 2). The metasignature genes were involved in a numerous metabolic pathways ranging from the energy generation, nucleotide metabolism and biosynthesis of amino acids, glycans and folate biosynthesis (Table 3). Eight metasignature genes were involved in the environmental information processing like the Notch, Jak stat, WNT and TGF Beta signalling pathways and about 4 genes were involved in pathways related to other diseases like the prion disease, type 1 diabetes mellitus and the *E. coli* infection (Table 4, 5).

Table 2: Genetic Information processing pathways in which the metasignature genes are involved with the No. of genes and their entrez gene id

KEGG pathways of genetic information processing		
Pathways	No. of genes	Entrez gene IDs
Proteasome	5	10213, 5682, 5695, 5704, 5708
Aminoacyl-tRNA biosynthesis	3	2617, 3376, 6897
Ubiquitin mediated proteolysis	2	11065, 991
RNA polymerase	1	5440

Table 3: Metabolic pathways in which the metasignature genes are involved with the No. of genes and their entrez gene id

KEGG pathways of metabolism		
Pathways	No. of genes	Entrez gene IDs
Purine metabolism	3	10606, 4830, 5440
Glycine, serine and threonine metabolism	2	2617, 6897
Pyrimidine metabolism	2	4830, 5440
Fructose and mannose metabolism	2	7264, 8473
Citrate cycle (TCA cycle)	2	47, 6391
Pyruvate metabolism	1	3939
Cysteine metabolism	1	3939
Propanoate metabolism	1	3939
One carbon pool by folate	1	10797
Reductive carboxylate cycle (CO_2 fixation)	1	47
Folate biosynthesis	1	8836
Methionine metabolism	1	191
Glycan structures-biosynthesis 2	1	8473
Glutamate metabolism	1	2730
Oxidative phosphorylation	1	6391
Glycolysis/Gluconeogenesis	1	3939
Ether lipid metabolism	1	5050
Glycerolipid metabolism	1	8473
N-Glycan biosynthesis	1	8813
Glutathione metabolism	1	2730
Selenoamino acid metabolism	1	191
Valine, leucine and isoleucine biosynthesis	1	3376
Glyoxylate and dicarboxylate metabolism	1	10797

Table 4: Environmental information processing pathways in which the metasignature genes are involved with the No. of genes and their entrez gene id

KEGG pathways of environmental information processing		
Pathways	No. of genes	Entrez gene IDs
Notch signaling pathway	2	1857, 3065
Cytokine-cytokine receptor interaction	2	3460, 4283
ABC transporters-General	1	6890
Wnt signaling pathway	1	1857
TGF-beta signaling pathway	1	1875
ECM-receptor interaction	1	1278
Jak-STAT signaling pathway	1	3460

Table 5: Pathways relating to human diseases in which the metasignature genes are involved with the No. of genes and their entrez gene id

KEGG pathways of human diseases		
Pathways	No. of genes	Entrez gene IDs
Colorectal cancer	2	1857, 332
Pathogenic *Escherichia coli* infection	2	10382, 203068
Type 1 diabetes mellitus	1	3329
Prion disease	1	3329
Cholera-Infection	1	10952
Chronic myeloid leukemia	1	3065

From this exercise, it is evident that the metasignature genes play key roles in a multitude of cellular pathways. In cancer, the whole cellular machinery is thrown out of control into chaos leading to uncontrolled growth and division of the cell. We believe, from the KEGG pathway based analysis, that the sustained proliferation of the cancer cell is only possible due to involvement of various genes which are important in many fundamental cellular pathways needed for the cell to survive.

CONCLUSION

We found that cancer metasignature genes are scattered throughout the human genome and their chromosomal distribution does not show any relation to the genomic length and the gene density of the chromosomes. Most of the genes are generally expressed in almost all normal body tissues derived from the three embryonic germ layers. The KEGG pathway based analysis showed that these genes are involved in a myriad of metabolic and genetic processes in the cell. From the present investigation, we conclude the cancer metasignature genes are vital for proper functioning and maintenance of the normal cellular physiology. It is possible that almost all cell types of the human body contain a dormant genetic network which enables them to overcome senescence or evade apoptosis. The sustained proliferation of the cancer cell is only possible due to involvement of various genes which are important in many fundamental cellular pathways needed for the cell to survive. This may lead to cancer generation and progression through uncontrolled cell division.

REFERENCES

Abbasi, A.A. and K.H. Grzeschik, 2007. An insight into the phylogenetic history of HOX linked gene families in vertebrates. BMC Evol. Biol., 7: 239-239.

Bruford, E.A., M.J. Lush, M.W. Wright, T.P. Sneddon, S. Povey and E. Birney, 2008. The HGNC Database in 2008: A resource for the human genome. Nucleic Acids Res., 36: 445-448.

Bucca, G., G. Carruba, A. Saetta, P. Muti, L. Castagnetta and C.P. Smith, 2004. Gene expression profiling of human cancers. Ann. N. Y. Acad. Sci., 1028: 28-37.

De Pitta, C., L. Tombolan, M. Campo Dell'Orto, B. Accordi and G. Te-Kronnie *et al*., 2005. A leukemia-enriched cDNA microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia. Haematologica, 90: 890-898.

Ferrari, S., 2006. Protein kinases controlling the onset of mitosis. Cell Mol. Life Sci., 63: 781-795.

Forbes, S., J. Clements, E. Dawson, S. Bamford and T. Webb *et al*., 2006. Cosmic 2005. Br. J. Cancer, 94: 318-322.

Forsburg, S.L., 2008. The MCM helicase: Linking checkpoints to the replication fork. Biochem. Soc. Trans., 36: 114-119.

Gonzalez-Duarte, R. and R. Albalat, 2005. Merging protein, gene and genomic data: The evolution of the MDR-ADH family. Heredity, 95: 184-197.

Goodsell, D.S., 2005. The molecular perspective: Major histocompatibility complex. Stem Cells, 23: 454-455.

Greenman, C., P. Stephens, R. Smith, G.L. Dalgliesh and C. Hunter *et al*., 2007. Patterns of somatic mutation in human cancer genomes. Nature, 446: 153-158.

Halvorsen, O.J., A.M. Oyan, T.H. Bo, S. Olsen and K. Rostad *et al*., 2005. Gene expression profiles in prostate cancer: Association with patient subgroups and tumour differentiation. Int. J. Oncol., 26: 329-336.

Hanai, T., H. Hamada and M. Okamoto, 2006. Application of bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields. J. Biosci. Bioeng., 101: 377-384.

Hernandez, D., A. Janmohamed, P. Chandan, I.R. Phillips and E.A. Shephard, 2004. Organization and evolution of the flavin-containing monooxygenase genes of human and mouse: Identification of novel gene and pseudogene clusters. Pharmacogenetics, 14: 117-130.

Hershko, A., 2005. The ubiquitin system for protein degradation and some of its roles in the control of the cell division cycle. Cell Death Differ., 12: 1191-1197.

Higgins, M.E., M. Claremont, J.E. Major, C. Sander and A.E. Lash, 2007. Cancer Genes: A gene selection resource for cancer genome projects. Nucleic Acids Res., 35: 721-726.

Kaiser, J., 2005. Genomics. Tackling the cancer genome. Science, 309: 693-693.

Kanehisa, M., M. Araki, S. Goto, M. Hattori and M. Hirakawa *et al*., 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res., 36: 480-484.

Kim, T.M., H.J. Jeong, M.Y. Seo, S.C. Kim and G. Cho *et al*., 2005. Determination of genes related to gastrointestinal tract origin cancer cells using a cDNA microarray. Clin. Cancer Res., 11: 79-86.

Kim, T.Y., Y.J. Bang and K.D. Robertson, 2006. Histone deacetylase inhibitors for cancer therapy. Epigenetics, 1: 14-23.

Lai, C.Q., L.D. Parnell and J.M. Ordovas, 2005. The APOA1/C3/A4/A5 gene cluster, lipid metabolism and cardiovascular disease risk. Curr. Opin. Lipidol., 16: 153-166.

Lim, E.K. and D.J. Bowles, 2004. A class of plant glycosyltransferases involved in cellular homeostasis. EMBO J., 23: 2915-2922.

Malumbres, M., 2007. Cyclins and related kinases in cancer cells. J. BUON., 12: 45-52.

Rhodes, D.R., J. Yu, K. Shanker, N. Deshpande and R. Varambally *et al*., 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc. Natl. Acad. Sci. USA., 101: 9309-9314.

Stewart, B.W. and P. Kleihues, 2003. World cancer report, international agency for research on cancer (IARC), World Health Organisation (WHO). http://193.51.164.194/IARCPress/wcr/index.php.

Talbot, S.G., C. Estilo, E. Maghami, I.S. Sarkaria and D.K. Pham *et al*., 2005. Gene expression profiling allows distinction between primary and metastatic squamous cell carcinomas in the lung. Cancer Res., 65: 3063-3071.

Tang, Y., Z. Wang, Y. Huang, D.P. Liu, G. Liu, W. Shen, X. Tang, D. Feng and C.C. Liang, 2006. Gene order in human alpha-globin locus is required for their temporal specific expressions. Genes Cells, 11: 123-131.

Taylor, S.S., M.I. Scott and A.J. Holland, 2004. The spindle checkpoint: A quality control mechanism which ensures accurate chromosome segregation. Chromosome Res., 12: 599-616.

Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant and K. Canese *et al*., 2008. Database resources of the national center for biotechnology information. Nucleic Acids Res., 36: D13-21.

Wilming, L.G., J.G. Gilbert, K. Howe, S. Trevanion, T. Hubbard and J. L. Harrow, 2008. The vertebrate genome annotation (Vega) database. Nucleic Acids Res., 36: D753-D760.

Wolffe, A.P., 1996. Histone deacetylase: A regulator of transcription. Science, 272: 371-372.

Zhang, B., S. Kirov and J. Snoddy, 2005. WebGestalt: An integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res., 33: W741-W748.