A Manually Curated Novel Knowledge Management System for Genetic and Epigenetic Molecular Determinants of Colon Cancer

Ana Barat* and Heather J. Ruskin

Biocomputation Laboratory, Sci-Sym (Scientific Computing and Complex Systems Modelling) Research Centre, School of Computing, Dublin City University (DCU), Ireland

Abstract: *Background*: Disruption of the cell process by inappropriate modifications at the epigenetic level, together with chromatine rearrangements and mutations result in abnormal regulation of gene expression in cancer. Collection and integration of molecular data related to cancer onset and development should improve the understanding of oncogenesis as well as facilitate prevention and risk assessment. The aim of this work is to complement existing database and data mining provision on human pathology epigenetics, by supporting access to colon cancer-related knowledge from both phenotype and molecular phenomena association viewpoints.

Description: We developed StatEpigen: an integrated and highly specific web-based resource predominantly based on manual annotations, that provides users with information on associations between epigenetic phenomena and other crucial molecular events, such as abnormal gene expression and mutation, in colon cancer-related phenotypes. The main characteristic of the resource is access to frequencies of occurrence of molecular events, or associations of these, for a wide range of detailed colon cancer-specific phenotypes, with a particular interest on cancer initiation. The resource integrates data from published references and provides access to annotated information through a query, data visualisation and integration interface, with statistical features designed to highlight molecular event associations that may be significant at certain stages of oncogenesis.

Conclusion: The integrated resource presented here is a detailed and focused framework of curated data targeting colon cancer. Capturing knowledge on how epigenetic phenomena are associated with other important molecular features of colorectal cancer, the resource can be seen as a novel add-on to the large scale database provision in this area. It is available at the following URL: http://statepigen.sci-sym.dcu.ie.

Keywords: Colon cancer, data base, epigenetic, DNA hypermethylation, genotype-phenotype correlations, oncogenesis.

BACKGROUND

Links to epigenetic signatures, such as *DNA methylation*, *histone modifications* and *changes in chromatin*, have been established in cancer [1-7], in autoimmune disorders [8] and neuropsychiatric disorders [9, 10], in response to stress [11-14] and also in the ageing processes [1, 2, 8]. Epigenetic signatures match often to differential or abnormal gene expression profiles, with epigenetically triggered silencing or over-expression of genes described in recent work [15-19]. Of particular interest are epigenetic phenomena and related molecular features which occur during very early stages of disease initiation, since understanding these improves potential for developing strategies for early diagnosis [20, 21].

Colon cancer is one of the most common malignancies worldwide, particularly prevalent in the western world. As it is a pathology that usually develops very slowly, in a stepwise manner [3], data are available for a whole series of intermediary stages of the disease, providing insight on the developmental dynamics. These data are particularly valuable for risk assessment and early diagnosis, as there is no curative treatment for all stages of colorectal cancer [22]. Recent research [18, 19] shows that epigenetic abnormalities (such as DNA methylation) play a very important role in colon carcinogenesis. While large amounts of epigenetic data are becoming available on colon cancer, systematic synthesis of these is lacking. Clearly, with growing knowledge on molecular processes in disease epigenomes**,** complexity of manipulation and interpretation increases, and formal representation, straightforward computational access and improved ability to systematically manipulate and analyse this information is vital [23].

Efforts to meet these needs are ongoing and several cancer epigenetics-oriented resources already exist. MethCancerDB [24] focuses on gene-centred CpG island information and experimental design in order to assist researchers in assay planning. MethyCancer [25] integrates data on CpG island clones and global CpG island predictions, DNA methylation and gene expression data (from CGAP[1] and UniGene [26]). PubMeth [27] focuses on methylation frequency of cancer samples based on automatic, followed by manual, annotation without systematically distinguishing between cancer subphenotypes.

There is, to our knowledge at the current time, no resource, which incorporates information on epigenetic events,

*Address correspondence to this author at the Biocomputation Laboratory, Sci-Sym (Scientific Computing and Complex Systems Modelling) Research Centre, School of Computing, Dublin City University (DCU), Ireland; Tel: +35317005513; Fax: +35317005442: E-mail: abarat@computing.dcu.ie

[1]http://cgap.nci.nih.gov/

1876-8202/10

such as loss of imprinting and histone modifications. Additionally, current literature on genetic-epigenetic interactions is growing steadily, but remains largely qualitative, "patchy" and widely dispersed, lacking in systematic organisation. Relationships between epigenetic and genetic events are largely undefined within existing knowledge systems, but are crucial to understanding the dynamics and the cause/effect mechanisms of pathology initiation [28]. In evaluating markers for early detection and prognosis, it is important to know what factors may affect the accuracy of a marker and in which subpopulations the marker may be more accurate [20]. Thus, many references which describe frequencies of occurrence of molecular events, also describe frequencies of occurrence of an event in subpopulations, defined by the presence of another molecular event. Direct access to this kind of information *via* on-line resources is important. Furthermore, articles on sample epigenomes, reporting variability according to various subphenotypes (based on histologies, subhistologies and clinicopathological factors) are becoming increasingly common. Without this being captured in computational resources, these datasets remain fragmented and have limited potential for future exploitation.

Epigenetic models for pathologies such as colon cancer are essential to pinpoint underlying biological process mechanisms and predicting disease behaviour. The StatEpigen idea stemmed from the need for a clean set of data, needed to drive dynamic probabilistic models, at a molecular level, of cancer onset and progress. After investigation of data sources available, it became clear that published genetic / epigenetic quantitative interaction data existed mostly as primary literature and, consequently, was largely inaccessible for computational investigation. Thus, StatEpigen was designed to complement existing resources, such as PubMeth [27], MethCancerDB [24], MethyCancer [25], COSMIC [29], specifically with respect to gaps highlighted above. The system is based on a collection of epigenetic and genetic statistical information on colon cancer and associated phenotypes. Data on other cancers are also gradually being incorporated. The resource focuses predominantly on early stages of disease, investigating occurrence of abnormal molecular events in such early phenotypes as premalignant mucosa, aberrant crypt foci, adenomas, polyps and others. The interest is on how epigenetic events (such as CpG island hyper- and hypomethylation, various histone modifications, loss of heterozygocity etc.) are correlated with:

- each other.
- other molecular events such as gene expression, various types of mutations and polymorphisms.
- more complex molecular signatures such as MSI (Microsatellite Instability), CIMP (CpG Island Methylator Phenotype) and others.
- simultaneous molecular events; i.e. combinations of event types (as above), which occur simultaneously in the same samples.

For this reason, we have targeted the following types of datasets:

- molecular events and their frequency of occurrence, given the phenotype of the analysed samples.
- molecular events and their frequency of occurrence in sample subsets, characterised by another distinct molecular event which occurs in all analysed samples of a subset of given phenotype. These records are referred to as "Conditional Events" and are an important feature in this resource.

CONSTRUCTION AND CONTENT

Scientific Literature Curation and Annotation

Manual annotation is currently recognised as a major component of the gold standard in biological annotation. References containing data of interest for StatEpigen are first selected from PubMed using a keyword search. The query involves 4 categories of keywords: (i) related to the organs affected, (ii) related to colon cancer-associated phenotypes (histologies, subhistologies, type of cell abnormalities and possible pre-existent conditions), (iii) related to epigenetics (including aberrant methylation, histone modifications, chromatine abnormalities) and (iv) indicative that the reference may contain statistical information on both epigenetic alterations and associations between epigenetic alterations and other molecular features related to oncogenesis. The PubMed query is based on a preliminary colon cancer literature study and retrieves ~ 180 abstracts per year. This number has the tendency to increase slightly each year (e.g. 194 hits in 2008 versus 180 hits in 2007). The retrieved abstracts are manually examined and 60% on average are found relevant for this specific database. The selected papers are then assigned a curation priority index. Papers containing novel information or very rich in relevant information are given high curation priority indexes and assigned to a rapid curation queue (~ 35% of all abstract-selected papers). The papers with lower curation indexes are assigned to a slower curation queue. At present, StatEpigen contains information from about 270 papers belonging to the high interest group and having been published between 2000 and 2010. Normally, information is added to the database in a year-wise manner. For example, manually curated information from papers published in 2009 is included in the public database in 2010. Work on the slow curation group papers continues.

Manual data curation is performed using a specifically implemented curation editor (password protected[2]). The target curation unit is the molecular event with its associated frequency of occurrence. After examination of the information in full text, the phenotypic, epigenetic and genetic data are extracted and checked against the database *via* the editor to verify if currently present in the system. If the information is absent, the editor is used to allocate internal accession numbers for new objects such as genes, various types of molecular events and phenotypes. Genetic information is systematically mapped to stable identifiers from public databases UniProtKB [30] and Ensembl [31]. These data and sample-related information are then used to create the main records of the system, i.e. frequencies of Simple and Conditional Molecular Events.

The choice of manual annotation, as opposed to automatic annotation, is due to the complexity of data-types researched on the one hand (correlation of a molecular alteration of interest with other molecular alterations), as well as

[2]Request for access can be obtained on an application.

the current limitations of text-mining algorithms on the other. For example, the main information targeted here - conditional molecular events - are given in various formats, usually in tables, which differ from one reference to another. Also, a large amount of information is in display form, e.g. figures and photos of gels corresponding to assays. This information is only amenable to manual curation.

Database Provision

The curated data are incorporated in a purpose-designed database, implemented in MySQL, which follows a relational schema, generated according to an expert assessment of the data structures encountered in the literature. Efforts have been made to design a data model, which supports extension and refinement, in order to incorporate additional details as new data are published. The knowledge base is necessarily designed to be robust, to support the variable data classifications from the published literature, and to allow for evolution of others. For example, StatEpigen supports quantitative data, obtained using various methods of methylation quantifying.

Web-Based User Interface

A web-based user interface *http://statepigen.scisym.dcu.ie* provides fast data access and dissemination, as well as straightforward querying and result display facilities. The user interface for StatEpigen employs a design, implemented in PHP, Javascript, HTML and CSS.

Data Structures in StatEpigen

While the entity-relationship diagram of the database is given as supplementary material, two data structures of StatEpigen are looked at more closely here.

Frequencies of Molecular Events

At the core of the data model are the molecular events themselves (epigenetic and genetic) with their associated frequencies in various samples. Based on these, two main data structures are derived:

- Simple Molecular Event records: the incidence of a molecular event for a given phenotype.

- Conditional Molecular Event records: the incidence of a molecular event (the measured feature) in a subset of samples of specified phenotype, characterised by another molecular event (the condition feature).

A Simple Event record contains information about the molecular event itself, such as: name, name of the associated gene (where appropriate), event details (specification). Further, the record gives statistical information related to the event, including the number of samples analysed, event occurrence in the samples, any qualitative or quantitative information available on the event, and phenotype-related information. The status of the event refers to qualitative information about it and is determined by the assay used (as described in the specific reference). For example, a paper may describe only the detection or give semi-quantitative details about an event, such as whether its intensity of methylation is high or low. Other articles give extensive quantitative details on the analysed molecular events. This is dealt with by the fields "Quantified" and "Units", giving the quantification of the event and its respective units (which depend on the assay). Finally, the record fields also give the phenotype of the samples (see next section), and the PubMed ID of the reference. Fig. (**1**) shows an example of three Simple Event records, for three different sets of colon adenoma samples, which have been verified for adenomatous polyposis coli (APC) promoter hypermethylation.

A Conditional Molecular Event record is very similar to that for a Simple Molecular Event, but incorporates a further condition, based on another event, known to take place in all analysed samples. As for a Simple Event, a Conditional Event record gives qualitative and/or quantitative information on an event (measured feature), its frequency, the number of samples analysed, information about the phenotype of the analysed samples and the PubMed ID of the reference. The only difference in this case is that all the samples are characterised by another molecular event (condition feature), hence all information about this feature is also incorporated and displayed.

Fig. (**2**) gives an example of three Conditional Event records. Here, the frequency of APC hypermethylation (measured feature) is verified in three different subsets of adenomas, for which other molecular events, such as MGMT, CDKN2A p14 and p16 promoter hypermethylation (condition feature options) are known to take place in all analysed samples. Note how APC hypermethylation takes place more

	Pmid	Gene Id Gene	Name Event	Specif Event	Status Event	Frequency Event	Nb Tested Samples	Histology	Subhisto	Origin
☑	15526363	APC	hyperMeth_CpGprom		YES	0.250	8	adenoma	tubulovillous	sporadic -
☑	15389252	APC	hyperMeth_CpGprom		YES	0.276	29	adenoma	serrated	sporadic -
☑	15122305	APC	hyperMeth_CpGprom		YES	0.358	95	adenoma	dysplastic low grade	sporadic -

Fig. (1). Example of Simple Molecular Event records: APC promoter hypermethylation. All phenotypes are colon adenomas with various subhistologies. The annotation 'sporadic, -' for the origin field means that the adenomas are either sporadic, or have unspecified pre-existing conditions. The records have been obtained by performing an "Advanced Search", filtering by early tumours such as adenomas, and the APC gene. Only the first three rows of the search result have been illustrated here and the empty columns have been eliminated from the figure for convenience.

	Pmid	Gene1	Name Event1	Specif Evvent1	Status Evvent1	Nb Samples Event1	Gene2	Name Event2	Specif Event2	Status Event2	Ferequency Event2	Histology	Subhisto	Origin
		Molecular Feature Known to Be Present in All Analysed Samples:				Analysed samples:	Molecular Feature Measured in the Samples:				freq ev2 in samples where ev1 takes place	Phenotypic information:		
☑	16902913	MGMT	hyperMeth_CpG	prom	YES	39	APC	hyperMeth_CpG	prom	YES	0.769	adenoma	-	sporadic, -
☑	16902913	CDKN2A:p14	hyperMeth_CpG	prom	YES	31	APC	hyperMeth_CpG	prom	YES	0.806	adenoma	-	sporadic, -
☑	16902913	CDKN2A:p16	hyperMeth_CpG	prom	YES	25	APC	hyperMeth_CpG	prom	YES	0.880	adenoma	-	sporadic, -
		Condition Feature					Measured Feature							

Fig. (2). Example of Conditional Molecular Events: APC promoter hypermethylation given that another event takes place in all analysed samples. All-sample events are: MGMT, p14 and p16 promoter hypermethylation, respectively. The given records have been obtained from the search used to obtain the records from Fig. (**1**).

often in these samples than in those presented in Fig. (**1**), showing how the occurrence of APC promoter hypermethylation is biased by other molecular signatures in the cell.

Phenotypes

An important feature of this resource is visualisation of molecular events in the context of various intermediary stages of disease; hence, an appropriate detailed phenotype classification is specified. A phenotype is described here by the following components: *histology*, *subhistology*, *dysplasia*, *origin*, *type of cancer*. While most of these characteristics are straightforward, some clarification on what is meant by "origin" may be needed. Given the cancer context, it is useful to distinguish different ways that a tumour originates. For example, while most colon tumours are sporadic, others are associated with previously existing conditions, such as hereditary manifestations [32], Peutz-Jeghers Syndrome [33], Ulcerative Colitis [34, 35] and others. The "origin" field captures this information, when available. Both disease and normal phenotypes are included in StatEpigen. Related to normal phenotypes, the resource distinguishes samples from healthy organs and healthy-appearing tissues adjacent to tumours. This distinction is important since epigenetic profiles from healthy samples may differ from those in apparently healthy samples surrounding tumours [35, 36], a phenomenon known as "field effect" [37].

In addition, data on more than 100 cancer cell lines are available in StatEpigen. Each distinct cell line is considered as a separate phenotype. Work is also in progress to annotate StatEpigen with existing ontologies. Further details are given in the Discussion section.

UTILITY

Querying Interfaces

StatEpigen queries can be formed in a number of ways, which are schematically represented in Fig. (**3**).

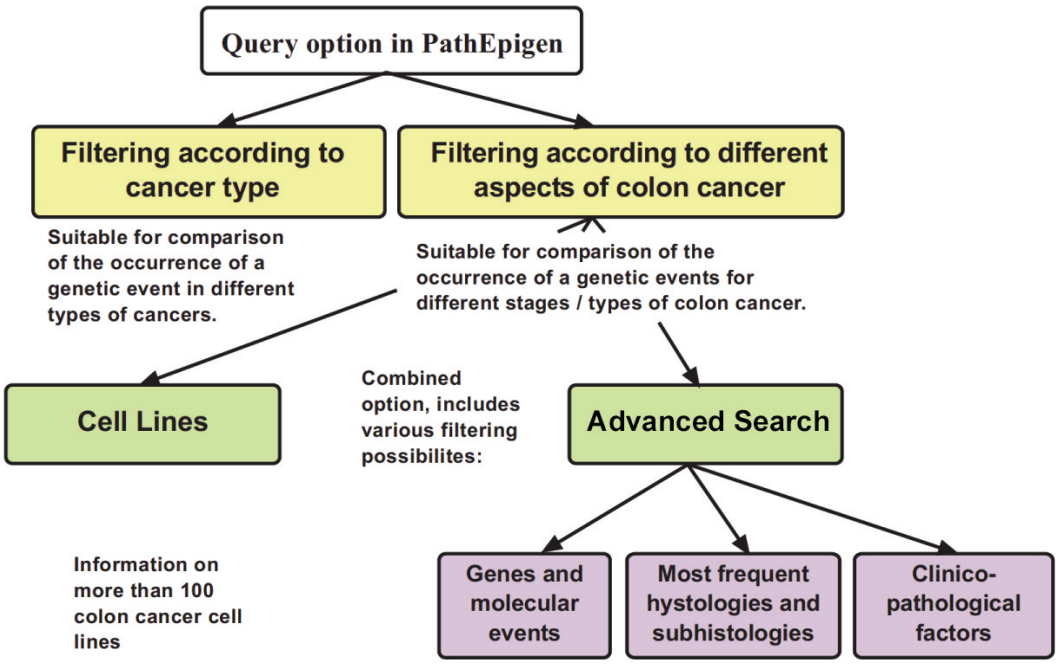

Fig. (3). Diagram representing the types of search available in StatEpigen.

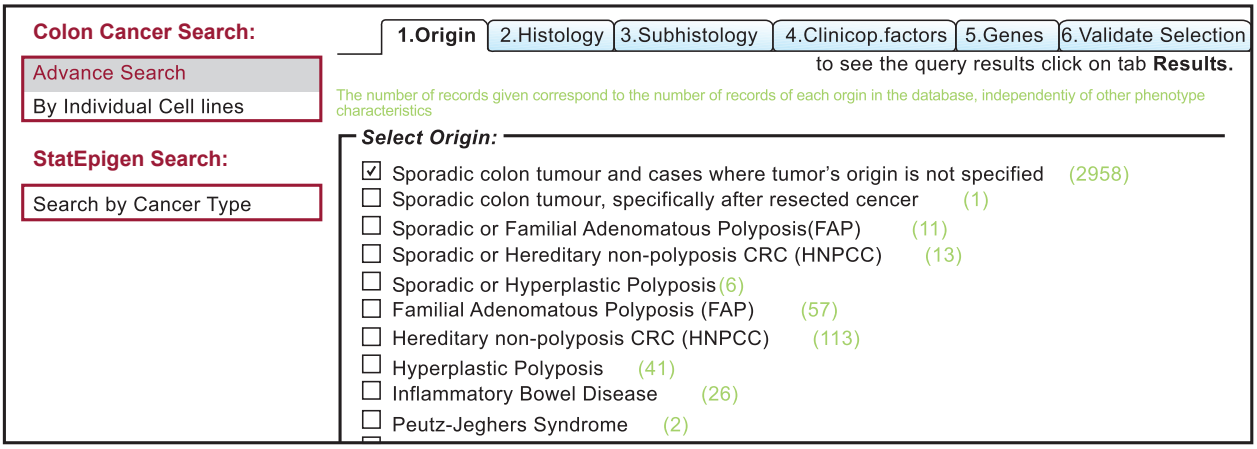

Fig. (4). Illustration of the StatEpigen menu displaying the available search options (left) and the "Advanced Search" tabular interface (right).

Advanced Search

The main querying feature of StatEpigen is "Advanced Search". This search allows filtering both by colon cancer phenotype and also by genes and molecular events (Fig. **4**). Browsing is performed by selecting from a number of tabs, including origin, histology, subhistology, dysplasia, clinicopathological factors and genes / molecular events. This allows filtering by tumour origin and by histology / subhistology: the interface here contains both common and less common histologies and subhistologies. In addition, this search permits verification of promoter methylation in stool, serum and other biological samples [5, 38-42]. This kind of analysis is valuable, allowing methylated genes to be explored as markers for improving screening for colorectal cancer [38]. In Fig. (**4**), filtering is carried out according to origin in the first tab of the "Advanced Search" interface, and can then be refined by histology / subhistology choices, (tabs 2 and 3). The user can also decide to filter only the records, which have been analysed according to clincopathological factors, (tab 4). Most studies on molecular determinants of cancer verify whether the clinicopathological factors separate a sample in subsets with different molecular event frequencies. For example, Lee *et al.* [43] found that separating samples, according to *colon location*, shows a significant difference in the frequencies of the methylation of DAP-kinase, E-cadherin and CDKN2A (p14) genes. For this reason, we have annotated data according to clinicopathological factors, and have implemented this as a separate querying option. When an "Advanced Search" query is performed and the "Clinicopathological Factors" option is checked, data not containing clinicopathological factor annotations are excluded, to prevent biased summary and statistics computing downstream.

Finally, one can filter according to a gene and/or molecular event of interest, (tab 5).

By clicking on "Validate Selection", the query is submitted and the number of available Simple and Conditional Events are displayed to the terminal. At this point, two options are possible, (Fig. **5**):

i. To Consider the Retrieved Data

The user can choose to visualise the data in details by clicking buttons "See Simple Events" and/or "See Conditional Events", (Fig. **5**). This allows tables of database records to be displayed, sorted by column (according to the gene, name of molecular event, histology etc.), selected / unselected for downloading and pipelined for subsequent analysis. The retrieved data format is as given in Fig. (**1**) for Simple and Fig. (**2**) for Conditional Events. The "Summary and Statistics" button provides access to a page allowing to summarise the retrieved results and to visualise the results of further statistical comparisons between molecular event occurrences in different sample types.

ii. To Refine the Query

Revert to the querying form(s). The previous filtering choices are memorised by the system and visible to the user.

For example, suppose that the user is interested in the behaviour of the APC gene in early pre-invasive colon lesions. It is widely known that APC mutation is an early event in colorectal tumorigenesis [44]. As the occurrence of mutations in various genes is known to be inter-related with epigenetic disregulations [45], all events and possible correlations will be searched for the APC gene. Thus, filtering by both phenotypes and genes is necessary. An advanced search is performed for adenomas and polyps (in samples with no associated pre-existing condition[3]) and the APC gene. Fig. (**1** and **2**) are table extracts from the retrieved results.

Search by Cell Lines

Data on more than 100 colon cancer cell lines are available in StatEpigen, permitting good integrated profiles of cell line cancer-related pathways to be visualised, using the option "By Individual Cell Lines".

Summary and Statistics Tools

Each time a colon cancer focused search is performed, data summarisation and statistics options are available, (Fig. **5**). For example, results on occurrence of a molecular event in a given phenotype are often available from more then one source. Of potential interest, especially when each individual set of samples is small, is to compute the occurrence of the

[3]These include the following: i) specifically sporadic samples, ii) datasets where samples associated with pre-existing conditions were explicitly excluded from the study by the authors, iii) samples for which the reference does not specifically mention any associated pre-existing conditions.

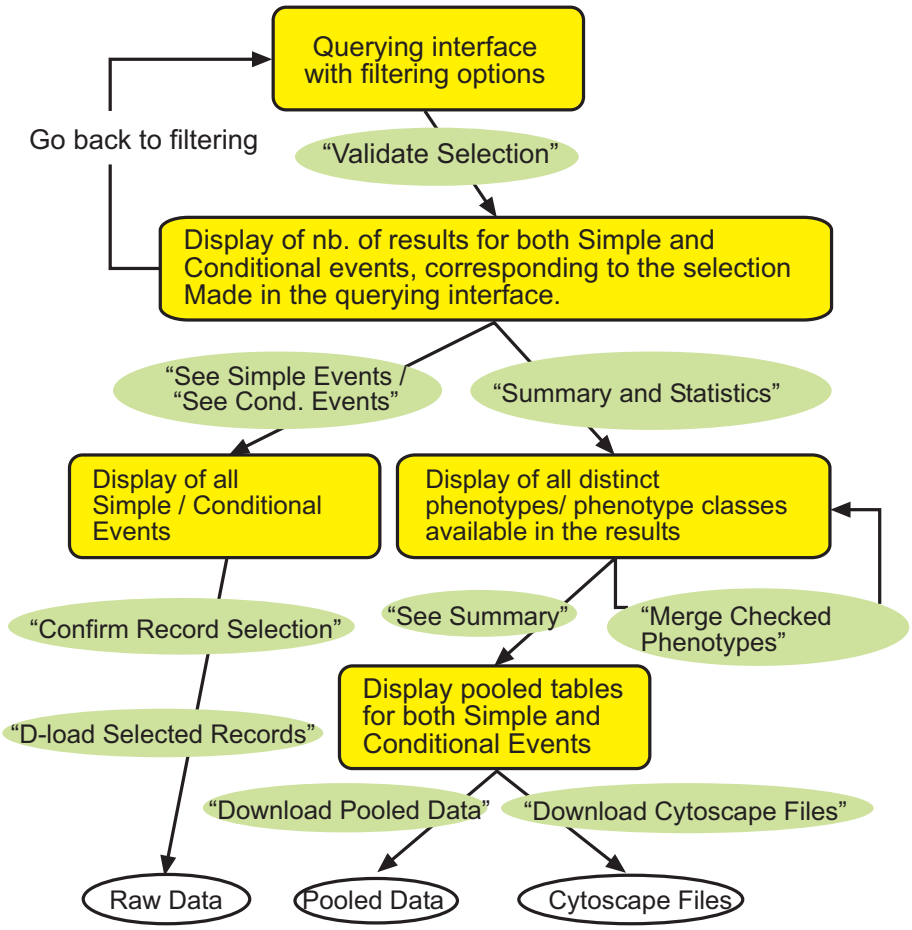

Fig. (5). Data visualisation and analysis flow diagram. Data can be downloaded in comma separated format and Cytoscape format (for summarisation data only). Cytoscape [46] is used to graphically represent static gene interaction, molecular pathways and protein interaction as networks. It can also be used to represent associations between molecular events. More details and an example are provided in the "Help" section of the website.

event, taking into consideration a number of available sets of samples. The "Summary and Statistics" option is intended to summarise and group together all records which give information on a *unique* event, in order to help the user with data visualisation. This brings together results from a list of sources, (which are easy to trace back), allowing for pooled frequencies of the molecular events to be computed for phenotypes of choice. The option also enables statistical tests to be carried out, to ascertain potentially correlated molecular events and significant differences in frequencies of event occurrence across phenotypes.

The options "Advanced Search" and "Summary and Statistics" are used to best advantage together, when focused on a particular molecular event/gene. For example, Simple and Conditional Events, related to the APC gene, are retrieved in both early non-invasive lesions (adenomas here) and carcinomas of colon. Clicking on "Summary and Statistics" will first give the listing of all phenotypes present in the retrieved results (Fig. **6**). At this point, the user can (i) first pool available phenotypes into classes and then "See Summary" and see the pooled records for each molecular event by each newly defined phenotype class, or (ii) click "See Summary" and display the pooled records for each molecular event by each phenotype in the table (Fig. **5**).

i. Visualise Pooled Results by Phenotype Classes

StatEpigen uniquely defines each phenotype by its components: origin, histology, subhistology, dysplasia and cancer type. Thus, e.g. *tubular dysplastic adenoma* is a different phenotype to *villous dysplastic adenoma*, because the "subhistology" component is different. However, clinical researchers may wish to define their own phenotype classes, e.g. to compare *all* adenoma types against *all* carcinomas. In Fig. (**6**), all carcinomas (both poorly and well differentiated) have already been merged to a class (id=10) and all early non-invasive tumours are selected for merging. A new phenotype class, which includes finer-grained phenotypes such as tubular, serrated, villous and other adenomas, is obtained by clicking "Merge Checked Phenotypes". The newly defined class is given an id, from the first row of the list of the merged phenotypes. Hence, in merging the early-invasive tumours from Fig. (**6**), the new phenotype class receives the id 39, resulting in two final classes: 10 (carcinomas) and 39 (adenomas).

Fig. (**7**) illustrates how Simple Events related to the APC gene have been pooled, both for adenomas and carcinomas. For example, APC abnormal methylation occurs slightly more often (46%) in early non-invasive tumours than in

Merge	ID	Histology	Subhistology	Dysplasia	Origin
☑	39	adenoma	-	-	sporadic,-
☑	15	adenoma	serrated	-	sporadic,-
☑	2	adenoma	tubulovillous	-	sporadic,-
☑	53	adenoma	tubular	dysplastic	sporadic,-
☑	54	adenoma	villous	dysplastic	sporadic,-
☑	60	adenoma	-	dysplastic low grade	sporadic,-
☐	10	carcinoma	-, poorly differentiated, well differentiated	-	sporadic,-

See Summary | Merge Checked Phenotypes | Reset

Fig. (6). List of distinct phenotypes for early non-invasive colon tumours (adenomas), found by advanced search, as described in example from the text. The rows are checked in order to merge into one composite phenotype. The un-checked row was obtained after merging three distinct phenotypes of carcinomas (poorly differentiated, well differentiated and with unspecified subhistology).

Nb	PMIDs	Gene	Name	Specif	Status	Average Freq	Total nb exper	p-val	Phen
Details - >		Molecular Feature Measured in all Samples:				Average Freq Event:	Analysed Samples:	Compared are the freq for Phen on lins a & b	Hover over phenotype ID to see details:
1	15526363 17260021 16902913 15389252 10969779 16336454 17167178 15386372 15122305	APC	hyperMeth_CpG	prom	YES	0.46	398	0.001, 1 & 3	39: adenomas
6	12782759 16902913	APC	LOH	5q	YES	0.111	117	0.0142, 6 & 2	39: adenomas
7	16902913 17167178	APC	mutation	any	YES	0.527	112	0.032, 7 & 5	39: adenomas
3	15760919 15386372 10969779 15122305 15326380 16336454 15502094 18977219 15375009 15526363 17167178 18451217 17260021	APC	hyperMeth_CpG	prom	YES	0.354	889	0.001, 3 & 1	10: carcinomas
2	12397640 12782759 10969779 15326380 10419592	APC	LOH	5q	YES	0.222	212	0.0142, 2 2 & 6	10: carcinomas
5	17167178 18451217 10969779 15034581	APC	mutation	any	YES	0.643	311	0.032, 5 & 7	10: carcinomas
4	15326380 10389752	APC	gene_expression		YES	0.233	150		10: carcinomas

Fig. (7). Pooled records, sorted by gene and phenotype. Shown are the 7 rows of the integrated records table, obtained with two composite phenotypes: id=39 – colon adenomas and id=10 – colon carcinomas. The p-values relate to comparison of frequencies of occurrence of molecular events in the two different phenotypes and are obtained applying the χ_1^2 test. To show for which pairs of rows, (hence couples of phenotypes), the p-values are calculated, comments such as "1 & 3" (row labelled "1" compared to row labelled "3") are given. As new data are continuously added to the database, the values for the pooled event frequencies are expected to slightly vary.

carcinomas (35.4%), (χ_1^2 test, p=0.001). APC mutation is an early event too (~52,7%), but it seems to occur even more often in carcinomas[4] (~64.3%), p=0.032. Loss of heterozygosity (LOH) on 5q shows a cumulative effect from adenoma to carcinoma (p=0.014). Fig. (**8**) shows the table for pooled Conditional Events. This table is of interest, because it provides statistical information on event associations. Rows 17 and 18 of this table clearly suggest that abnormal methylation constitutes a mechanism of APC gene inactivation. Further, APC abnormal methylation increases in frequency in subpopulations where other promoters are also methylated, (namely MGMT, p14, p16): rows 30-37. Also, APC promoter hypermethylation appears to be associated with APC loss of heterozygocity (LOH) in carcinomas, (rows 1-2, 11-12) and as APC hypermethylation occurs early in tumour progression, this association applies to early tu-

[4]Ideally, in order to perform statistical tests on pooled data, experiments using the same technology should be performed under the same conditions. The disadvantage of the present version is that it does not store assay methodology. However, here the data are pooled on the basis that most are obtained, using similar well-known methods, e.g. PCR-based techniques using bisulfite conversion, (predominantly MS-PCR), for methylation detection). These methods have good sensitivity and specificity; drawbacks are connected with the possibility of contamination of the analysed sample and obtaining false-positive results [21, 47]. As techniques to detect and quantify molecular events develop further, annotating records for assay methodology is likely to become important, in order to support refined data pooling in future. This upgrade to StatEpigen is now being considered.

Nb	Refs	Gene1	Name1	Spe-cif1	Sta-tus1	Nb samples event1	Gene2	Name2	Spe-cif2	Sta-tus2	Freq event2	p-val 1	p-val 2	Phen
Details - >		Molecular Feature Known to Be Present in All Analysed Samples:					Molecular Feature Measured in the Samples:				freq ev2 in samples where ev1 takes place	Compared are freq (ev2 given ev1) & freq (ev2)	Compared are freq (ev2 given ev1) on lines a & b	Hover over phen ID to see details:
1	15326380	APC	LOH	5q	YES	11	APC	hyperMeth_CpG	prom	YES	0.364	0.9733	0.496, 1&2	10
2	15326380	APC	LOH	5q	NO	105	APC	hyperMeth_CpG	prom	YES	0.267	0.0467	0.496, 1&2	10
11	10969779	APC	hyperMeth_CpG	prom	YES	14	APC	LOH	5q	YES	0.5	0.0096	0.124, 11&12	10
12	10969779	APC	hyperMeth_CpG	prom	NO	17	APC	LOH	5q	YES	0.235	0.893	0.124, 11&12	10
13	10969779	APC	mutation	any	YES	47	APC	hyperMeth_CpG	prom	YES	0.064	0.0009	0.0248, 13&14	10
14	10969779	APC	mutation	any	NO	19	APC	hyperMeth_CpG	prom	YES	0.263	0.4279	0.0248, 13&14	10
17	10969779	APC	hyperMeth_CpG	prom	NO	20	APC	gene_expression		YES	1	na	0.0001, 17&18	39
18	10969779	APC	hyperMeth_CpG	prom	YES	4	APC	gene_expression		YES	0	na	0.0001, 17&18	39
30	16902913	MGMT	hyperMeth_CpG	prom	YES	39	APC	hyperMeth_CpG	prom	YES	0.769	0.0009	0.0001, 30&35	39
35	16902913	MGMT	hyperMeth_CpG	prom	NO	39	APC	hyperMeth_CpG	prom	YES	0.333	0.0966	0.0001, 30&35	39
31	16902913	CDKN2A:p14	hyperMeth_CpG	prom	YES	31	APC	hyperMeth_CpG	prom	YES	0.806	0.0009	0.001, 31&36	39
36	16902913	CDKN2A:p14	hyperMeth_CpG	prom	NO	47	APC	hyperMeth_CpG	prom	YES	0.383	0.2664	0.001, 31&36	39
32	16902913	CDKN2A:p16	hyperMeth_CpG	prom	YES	25	APC	hyperMeth_CpG	prom	YES	0.88	0.0009	0.001, 32&37	39
37	16902913	CDKN2A:p16	hyperMeth_CpG	prom	NO	53	APC	hyperMeth_CpG	prom	YES	0.396	0.3446	0.001, 32&37	39

Fig. (8). Extract from the table of integrated records for Conditional Events pooled for 2 composite phenotypes – colon adenomas (id=39) and carcinomas (id=10). The figure has been edited for convenience and a selection only of all existing records and columns is shown. The p-value 1 relates to comparison of Conditional and Simple Event frequencies: e.g. frequency of APC hypermethylation in tumours with hypermethylated MGMT versus frequency of APC hypermethylation in general. A low p-value, highlighted in red, means that the frequency of the Conditional Event is significantly different to that for its Simple counterpart, (from Fig. **7**). The χ_1^2 test is used to compare proportions. The assumption of independence is respected on the basis that each Conditional Event sample is first removed from the corresponding pooled Simple Event record, and the test is thus performed on completely independent samples. To compare the frequencies in complementary subsets (e.g. rows 1 and 2 or 11 and 12), the two-sided exact Fisher's test is applied, giving p-value 2.

mours too. In Fig. (**8**) (rows 11-12), Fisher's exact test was used to compare presence of 5q LOH according to APC promoter methylation status. Clearly, LOH on 5q occurs more often in samples with hypermethylated APC promoter (50% - see Fig. **8**, row 11) than in cancer samples with unknown APC promoter status (22.2% - see Fig **7**, row labelled 2), p-value=0.0096. The other way round can also be tested: rows 1-2, Fig. **8** show presence APC promoter hypermethylation according to 5q LOH status. Here, APC hypermethylation occurs more rarely in samples with no 5q LOH (26,7% - see Fig. **8**, row 2), than in cancer samples with unknown 5q LOH status (35.4% - see Fig. **7**, row 3), p-value=0.0467. References [33, 48] acknowledge the association, suggesting that APC hypermethylation might be a second-hit mechanism to silence APC on one allele after LOH silences the gene on the other one. Conversely, the information from rows labelled 13 and 14, Fig. (**8**), shows that APC hypermethylation appears very rarely together with APC mutations, p-value=0.0248.

ii. Visualise Pooled Results by Initial Phenotype Definitions

Finally, we present an example, in which subphenotypes are of immediate interest, illustrating the value of defining fine-grained phenotypes. An "Advanced Search" on DAPK1 promoter abnormal methylation in adenomas was performed and the "Summary and Statistics" option was applied without merging. Fig. (**9**) shows values of DAPK1 promoter hypermethylation frequencies for carcinomas and 2 different subphenotypes of adenomas, according to clinicopathological factors (left and right colon). For each distinct phenotype, a χ^2 test is performed to compare the frequency of the given samples with the samples from all the other available phenotypes. All p-values, which are smaller than 0.1, are displayed. The χ^2 test confirms that the frequencies of DAPK1 promoter hypermethylation are not uniformly distributed across the sample populations. Both for carcinoma and dysplastic low grade adenoma, the promoter of DAPK1 is significantly more often hypermethylated in samples from right colon compared to samples from left colon. Also, there is no difference in the methylation frequencies in right colon dysplastic adenoma and carcinoma, suggesting that DAPK1 methylation is an early event in the right colon. Conversely, hypermethylation of DAPK1 was found significantly less often in left colon dysplastic adenoma compared to left colon carcinoma (lines 2 and 5, p=0.0287), suggesting that for left colon, the given phenomenon even may occur later then in the right colon. In the case of serrated adenoma, the findings are different, with a high frequency of methylation in the left colon (line 4 - no difference to the frequencies in the right colon).

DISCUSSION AND FUTURE DEVELOPMENT

StatEpigen, with its several thousand records from more than 270 references on colon cancer, together with its data querying, integration and visualisation interface, represents

Nb	PMIDs	Gene	Name	Specif	Status	Average Freq	Total nb exper	p-val	Clin. Path. F.	Phen
Details - >		Molecular Feature Measured in all Samples:				Average Freq Event:	Analysed Samples:	Compared are the freq for Phen on lines a & b	Clinico Pathological Factors:	Phenotype
1	15122305	DAPK1	hyperMeth_CpG	prom	YES	0.607	56	0.0146, 1 & 2 0.0009, 1 & 6	right	carcinoma
2	15122305	DAPK1	hyperMeth_CpG	prom	YES	0.398	93	0.0146, 2 & 1 0.0287, 2 & 5 0.0091, 2 & 6	left	carcinoma
3	15389252	DAPK1	hyperMeth_CpG	prom	YES	0.333	6	not signif.	right	serrated adenoma
4	15389252	DAPK1	hyperMeth_CpG	prom	YES	0.441	34	0.0121, 4 & 6	left	serrated adenoma
5	15122305	DAPK1	hyperMeth_CpG	prom	YES	0.618	34	0.0287, 5 & 2 0.0009, 5 & 6	right	dysplastic low grade adenoma
6	15122305	DAPK1	hyperMeth_CpG	prom	YES	0.197	61	0.0009, 6 & 1 0.0091, 6 & 2 0.0121, 6 & 4 0.0009, 6 & 5	left	dysplastic low grade adenoma

Fig. (9). Shows information on DAPK1 promoter hypermethylation in carcinomas and various adenomas. The search was done according to clinicopathological factors, with the choices on right colon and left colon.

an effort to apply informatics tools beyond CpG island methylation and gene expression studies, in order to integrate and comprehend molecular pathway steps in cancer initiation at a phenotype level. The new and unique features of the resource include a detailed description of phenotypes, (together with the option to integrate data by user-defined phenotype classes), the option of querying by clinicopathological factors, by stool and serum samples and by normal samples, including normal samples adjacent to tumours at various stages and of various origins. The ability to query and visualise frequencies of molecular events, conditional on occurrence of other events, is also new and represents a step towards defining genetic/epigenetic signatures of colon cancer subtypes. Downloadable data sets make this resource useful not only for clinicians and experimentalists, but also for those interested in building stochastic models and applying bioinformatics methods. One example of a similar such resource is COSMIC (Catalogue of Somatic Mutations in Cancer) [29]. This is acknowledged as critical database for cancer researchers. Consequently, complementing mutation information with that for epigenetic-genetic interactions may be expected to provide a valuable contribution to investigation of pathways in cancer development.

Although manually curated databases have the advantage of a lower error rate [49], no data extraction technique has perfect accuracy [50] and there have been contradictory reports about data quality in curated biological databases [51]. The error rate of StatEpigen was specifically calculated. To do this, two datasets were selected, the first containing simple events, the second conditional events, and each comprising of the order of 400 records. The records where checked against the original references. Records with any aberration from the original information were considered as errors, (and corrected). At present, the estimated error rate on both simple and conditional events is 5%, which is comparable to the error rates of other biomedical resources [52]. In order to improve the data quality and avoid error propagation in the future, existing records of StatEpigen also pass through a second check; the new records are edited by one curator initially and then verified, corrected and validated by another curator.

A future aim of the project is to annotate StatEpigen with a formal, well-principled cancer phenotype ontology based on Open Biomedical Ontologies (OBO) Foundry [53]. The specific way, in which the phenotypes are defined in the resource, will allow implementation of automatic mapping to an existing ontology. Candidates include currently available ontologies of disease phenotypes, (e.g. Human Disease Ontology, Human Phenotype Ontology). Although these do not encompass all phenotypes available in StatEpigen, additions are possible. Phenotype ontology annotation of the database will support integration and/or cross-referencing with other ontologies and databases. This representation of the specific phenotypes will help disease modelling and support subsequent project objectives, namely – optimisation of the existing data-mining interfaces of the resource and development of new ones.

The American Association for Cancer Research Human Epigenome Task Force and the EU Networks of Excellence Scientific Advisory Board, have outlined the computational challenges for a comprehensive human epigenome project [23]. According to these authors, such a project requires a strong bioinformatics platform, for which a priority is to establish a central large-scale relational database and web interface, assisted by analytical and statistical tools, to present data to the scientific community**.** On the other hand, the advantages of relatively small specific resources like the one presented here is that they can remain detailed and specific. The longer term destiny of this type of resource relies on work targeted to incorporate them to the larger scale provision. StatEpigen is designed in a flexible manner so that it

can be both optimised and developed further, but also integrated to a central large-scale platform.

CONCLUSIONS

The StatEpigen resource has been flexibly designed and implemented for multiple purpose usage. As such, it offers a quality integrated data source and can support a number of key requirements such as comparison between phenotypes, genes and epigenetic events, profile and signature search, as well as a basis for modelling. Its data-mining capability offers useful tools to both clinical and bioinformatics researchers, while the resource as a whole has the potential to help identify and support future research directions.

As new data are incorporated into StatEpigen, frequencies characterising molecular events and their correlations are continuously being refined and, as such, increase in value for statistical modelling, risk assessment and predictions. In the present version, data for cancers other than colon cancer provide a basis for comparison only, but future work to extend the resource to other pathologies is anticipated. Current work on the project is focused on completing the existing database with annotations on the assays used and on the dietary and environmental factors correlating to the molecular events, expanding StatEpigen to other gastrointestinal cancers and improving the data visualisation and data-mining interface.

AVAILABILITY AND REQUIREMENTS

Project name: StatEpigen database;

Project home page: http://statepigen.sci-sym.dcu.ie

ACKNOWLEDGEMENTS

The authors would like to thank IRCSET, (the Irish Research Council for Science Engineering) and Dublin City University for funding this work. We are grateful to Ubaldo Colmenar (Software Engineer, Universidad Politécnica de Madrid), for his extensive involvement in building the user interface, to Mathieu André (intern student from Institut National de Sciences Appliquées de Lyon, France), for implementing the interface for data annotation and collection, and to Dr. Andrey Pichugin (Institut de Cancerologie Gustave Roussy, Villejuif, France), for his feedback and suggestions on the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Ahuja N, Li Q, Mohan AL, Baylin SB, Issa JP. Aging and DNA methylation in colorectal mucosa and cancer. Cancer Res 1998; 58(23): 5489-94.

[2] Yuasa Y. DNA methylation in cancer and ageing. Mech Ageing Dev 2002; 123(12): 1649-54.

[3] Grady WM. Epigenetic events in the colorectum and in colon cancer. Biochem Soc Trans 2005; 33(Pt 4): 684-8.

[4] Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. Hum Mol Genet 2007; 16 Spec No 1: R50-9.

[5] Belshaw NJ, Elliott GO, Foxall RJ, *et al*. Profiling CpG island field methylation in both morphologically normal and neoplastic human colonic mucosa. Br J Cancer 2008; 99(1): 136-42.

[6] Irizarry RA, Ladd-Acosta C, Wen B, *et al*. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009; 41(2): 178-86.

[7] van Haaften G, Dalgliesh GL, Davies H. *et al*. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet 2009; 41(5):521-3.

[8] Lu Q, Qiu X, Hu N, Wen H, Su Y, Richardson BC. Epigenetics, disease, and therapeutic interventions. Ageing Res Rev 2006; 5(4): 449-67.

[9] Abel T and Zukin RS. Epigenetic targets of HDAC inhibition in neurodegenerative and psychiatric disorders. Curr Opin Pharmacol 2008; 8(1): 57-64.

[10] Allan AM, Liang X, LuoY, *et al*. The loss of methyl-CpG binding protein 1 leads to autism-like behavioral deficits. Hum Mol Genet 2008; 17(13): 2047-57.

[11] Meaney MJ. Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. Annu Rev Neurosci 2001; 24: 1161-92.

[12] Weaver IC, Cervoni N, Champagne FA, *et al*. Epigenetic programming by maternal behavior. Nat Neurosci 2004; 7(8): 847-54.

[13] Plagge A, Isles AR, Gordon E, *et al*. Imprinted Nesp55 influences behavioral reactivity to novel environments. Mol Cell Biol 2005; 25(8): 3019-26.

[14] Darnaudery M, Maccari S. Epigenetic programming of the stress response in male and female rats by prenatal restraint stress. Brain Res Rev 2008; 57(2): 571-85.

[15] Youssef EM, Estecio MR, Issa JP. Methylation and regulation of expression of different retinoic acid receptor beta isoforms in human colon cancer. Cancer Biol Ther 2004; 3(1): 82-6.

[16] Mori Y, Cai K, ChengY, *et al*. A genome-wide search identifies epigenetic silencing of somatostatin, tachykinin-1, and 5 other genes in colon cancer. Gastroenterology 2006; 131(3): 797-808.

[17] Ogino S, Brahmandam M, Kawasaki T, Kirkner GJ, Loda M, Fuchs CS. Combined analysis of COX-2 and p53 expressions reveals synergistic inverse correlations with microsatellite instability and CpG island methylator phenotype in colorectal cancer. Neoplasia 2006; 8(6): 458-64.

[18] Ogino S, Kawasaki T, Kirkner GJ, Dorfman I, Loda M, Fuchs CS. Down-regulation of p21 (CDKN1A/CIP1) is inversely associated with microsatellite instability and CpG island methylator phenotype (CIMP) in colorectal cancer. J Pathol 2006; 210(2): 147-54.

[19] Ogino S, Kawasaki T, Ogawa A, Kirkner GJ, Loda M, Fuchs CS. TGFBR2 mutation is correlated with CpG island methylator phenotype in microsatellite instability-high colorectal cancer. Hum Pathol 2007; 38(4): 614-20.

[20] Alonzo TA, Siegmund KD. Statistical methods for evaluating DNA methylation as a marker for early detection or prognosis. Dis Markers 2007; 23(1-2): 113-20.

[21] Sulewska A, Niklinska W, Kozlowski M, *et al*. Detection of DNA methylation in eucaryotic cells. Folia Histochem Cytobiol 2007; 45(4): 315-24.

[22] Papailiou J, Bramis KJ, Gazouli M, Theodoropoulos G. Stem cells in colon cancer. A new era in cancer theory begins. Int J Colorectal Dis 2010; [Epub ahead of print].

[23] Jones PA, Archer TK, Baylin SB, *et al*. Moving AHEAD with an international human epigenome project. Nature 2008; 454(7205): 711-5.

[24] Lauss M, Visne I, Weinhaeusel A, Vierlinger K, Noehammer C, Kriegner A. MethCancerDB-aberrant DNA methylation in human cancer. Br J Cancer 2008; 98(4): 816-7.

[25] He X, Chang S, Zhang J, *et al*. MethyCancer: The database of human DNA methylation and cancer. Nucleic Acids Res 2008; 36(Database issue): D836-41.

[26] Sayers EW, Barrett T, Benson DA, *et al*. Database resources of the national center for biotechnology information. Nucleic Acids Res 2009; 37(Database issue): D5-15.

[27] Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W. PubMeth: A cancer methylation database combining text-mining and expert annotation. Nucleic Acids Res 2008; 36(Database issue): D842-6.

[28] Herceg Z. Epigenetics and cancer: Towards an evaluation of the impact of environmental and dietary factors. Mutagenesis 2007; 22(2): 91-103.

[29] Forbes SA, Bhamra G, Bamford S, *et al*. The catalogue of somatic mutations in cancer (COSMIC). Curr Protoc Hum Genet 2008; Chapter 10, Unit 10.11.

[30] Bairoch A, Apweiler R, Wu CH, *et al*. The universal protein resource (UniProt). Nucleic Acids Res 2005; 33(Database issue):D154-9.

[31] Hubbard TJ, Aken BL, Ayling S, *et al.* Ensembl 2009. Nucleic Acids Res 2009; 37(Database issue): D690-7.
[32] Deng G, Bell I, Crawley S, *et al.* BRAF mutation is frequently present in sporadic colorectal cancer with methylated hMLH1, but not in hereditary nonpolyposis colorectal cancer. Clin Cancer Res 2004; 10(1 Pt 1): 191-5.
[33] Esteller M, Avizienyte E., Corn P, *et al.* Epigenetic inactivation of LKB1 in primary tumors associated with the peutz-jeghers syndrome. Oncogene 2000; 19(1): 164-8.
[34] Wheeler JM, Kim HC, Efstathiou JA, Ilyas M, Mortensen NJ, Bodmer WF. Hypermethylation of the promoter region of the E-cadherin gene (CDH1) in sporadic and ulcerative colitis associated colorectal cancer. Gut 2001; 48(3): 367-71.
[35] Kukitsu T, Takayama T, Miyanishi K, *et al*. Aberrant crypt foci as precursors of the dysplasia-carcinoma sequence in patients with ulcerative colitis. Clin Cancer Res 2008; 14(1): 48-54.
[36] Cui H, Horon IL, Ohlsson R, Hamilton SR, Feinberg AP. Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability. Nat Med 1998; 4(11): 1276-80.
[37] Giovannucci E, Ogino S. DNA methylation, field effects, and colorectal cancer. J Natl Cancer Inst 2005; 97(18): 1317-9.
[38] Chen WD, Han ZJ, Skoletsky J, *et al.* Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene. J Natl Cancer Inst 2005; 97(15):1124-32.
[39] Huang ZH, Li LH, Yang F, Wang JF. Detection of aberrant methylation in fecal DNA as a molecular screening tool for colorectal cancer and precancerous lesions. World J Gastroenterol 2007; 13(6): 950-4.
[40] Leung WK, To KF, Man EP, *et al.* Detection of hypermethylated DNA or cyclooxygenase-2 messenger RNA in fecal samples of patients with colorectal cancer or polyps. Am J Gastroenterol 2007; 102(5): 1070-6.
[41] Petko Z, Ghiassi M, Shuber A, *et al.* Aberrantly methylated CDKN2A, MGMT, and MLH1 in colon polyps and in fecal DNA from patients with colorectal polyps. Clin Cancer Res 2005; 11(3): 1203-9.
[42] Zou H, Harrington JJ, Shire AM, *et al.* Highly methylated genes in colorectal neoplasia: Implications for screening. Cancer Epidemiol Biomarkers Prev 2007; 16(12): 2686-96.
[43] Lee S, Hwang KS, Lee HJ, Kim JS, Kang GH. Aberrant CpG island hypermethylation of multiple genes in colorectal neoplasia. Lab Invest 2004; 84(7): 884-93.
[44] Judson H, Stewart A, Leslie A, *et al.* Relationship between point gene mutation, chromosomal abnormality, and tumour suppressor gene methylation status in colorectal adenomas. J Pathol 2006; 210(3): 344-50.
[45] Suehiro Y, Wong CW, Chirieac LR, *et al.* Epigenetic-genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in colorectal carcinoma. Clin Cancer Res 2008; 14(9): 2560-9.
[46] Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13(11): 2498-504.
[47] Reed K, Poulin ML, Yan L, *et al.* Comparison of bisulfite sequencing PCR with pyrosequencing for measuring differences in DNA methylation. Anal Biochem 2010; 397(1):96-106.
[48] Arnold CN, Goel A, Niedzwiecki D, *et al.* APC promoter hypermethylation contributes to the loss of APC expression in colorectal cancers with allelic loss on 5q. Cancer Biol Ther 2004; 3(10): 960-4.
[49] Popadin KY, Mamirova LA, Kondrashov FA. A manually curated database of tetrapod mitochondrially encoded tRNA sequences and secondary structures. BMC Bioinformatics 2007; 14(8): 441.
[50] Donaldson I, Martin J, de Bruijn B, *et al.* PreBIND and textomy-mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 2003; 27(4): 11.
[51] Cusick ME, Yu H, Smolyar A, *et al.* Literature-curated protein interaction datasets. Nat Methods 2009; 6(1): 39-46.
[52] Jones CE, Brown AL, Baumann U. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics 2007; 22(8): 170.
[53] Smith B, Ashburner M, Rosse C, *et al.* The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007; 25(11): 1251-5.

Received: August 18, 2010 Revised: October 15, 2010 Accepted: October 15, 2010