BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 23 2011, pages 3300–3305
doi:10.1093/bioinformatics/btr559

Gene expression

Advance Access publication October 11, 2011

Gathering insights on disease etiology from gene expression profiles of healthy tissues

A. Sofia Silva[1,†], Shona H. Wood[1], Sipko van Dam[1], Sven Berres[1], Anne McArdle[2] and João Pedro de Magalhães[1,*]

[1]Integrative Genomics of Ageing Group, Institute of Integrative Biology and [2]Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, UK

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Gene expression profiles have been widely used to study disease states. It may be possible, however, to gather insights into human diseases by comparing gene expression profiles of healthy organs with different disease incidence or severity. We tested this hypothesis and developed an approach to identify candidate genes associated with disease development by focusing on cancer incidence since it varies greatly across human organs.

Results: We normalized organ-specific cancer incidence by organ weight and found that reproductive organs tend to have a higher mass-normalized cancer incidence, which could be due to evolutionary trade-offs. Next, we performed a genome-wide scan to identify genes whose expression across healthy organs correlates with organ-specific cancer incidence. We identified a large number of genes, including genes previously associated with tumorigenesis and new candidate genes. Most genes exhibiting a positive correlation with cancer incidence were related to ribosomal and transcriptional activity, translation and protein synthesis. Organs with enhanced transcriptional and translational activation may have higher cell proliferation and therefore be more likely to develop cancer. Furthermore, we found that organs with lower cancer incidence tend to express lower levels of known cancer-associated genes. Overall, these results demonstrate how genes and processes that predispose organs to specific diseases can be identified using gene expression profiles from healthy tissues. Our approach can be applied to other diseases and serve as foundation for further oncogenomic analyses.

Contact: jp@senescence.info

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 11, 2011; revised on August 10, 2011; accepted on September 29, 2011

1 INTRODUCTION

Large-scale gene expression analyses employing microarrays have been widely used to study human diseases. The majority of such studies compare disease and non-disease states or different pathological conditions to identify disease biomarkers and gain insights into pathophysiological processes. Few studies, however, have compared data between healthy tissues to identify normal tissue-specific pathways that predispose or contribute to disease. One study focused on the expression of disease-associated genes in healthy tissues and found that they tend to be overexpressed in tissues where their disruption causes pathology (Lage *et al.*, 2008), but this remains a largely unexplored area. Herein, we wanted to develop a method that takes advantage of gene expression profiles from healthy organs to determine new candidate genes and processes associated with complex diseases.

Although many diseases are tissue specific, for others disease incidence and severity varies widely across organs. It is plausible that variation in particular organ features affects their predisposition to pathophysiological mechanisms and therefore identifying such features will provide clues on disease etiology. Given that large-scale gene expression data is available for the major human organs (Su *et al.*, 2004), the basic premise behind this work is that it may be possible to gather insights into human disease by comparing gene expression profiles of healthy organs with differences in disease incidence or severity. Our aim in this work was to test this hypothesis and develop a method for identifying genes and processes associated with organ-specific disease incidence using gene expression data from healthy tissues. Because cancer can originate in multiple organs, we focused our proof of concept analysis on cancer incidence.

Cancer is caused primarily by alterations in the genome of the affected cells (Hanahan and Weinberg, 2000; Stratton *et al.*, 2009), yet organ characteristics may predispose or protect from tumorigenesis. In fact, there is a great variation in the incidence of cancer across organs. Differences in cell proliferation and turnover have been put forward as an explanation, though the topic remains controversial (Ward *et al.*, 1993). A previous study found that while gain-of-function cancer genes (i.e. oncogenes) are overexpressed in tissues more associated with cancer, loss-of-function cancer genes (i.e. tumor suppressor and caretaker genes) are underexpressed in such tissues (Lage *et al.*, 2008).

In order to study the differences in cancer incidence between different human organs, we conducted a genome-wide correlation study between cancer incidence normalized for organ weight and gene expression patterns using microarray data from healthy tissues. Our results reveal a large number of genes and processes whose expression is associated with cancer incidence and show it is possible to employ gene expression profiles across healthy organs to

*To whom correspondence should be addressed.
†Present address: Centro de Investigação em Ciências da Saúde, Faculdade de Ciências da Saúde, Universidade da Beira Interior, Covilhã, Portugal.

identify mechanisms predisposing organs to disease development. This methodological framework may be useful for further studies on cancer as well as other diseases.

2 METHODS

The main aim behind this method is to detect genes whose expression correlates with organ-specific disease incidence. Groups of genes correlating, positively or negatively, with disease incidence can then be analyzed with standard functional enrichment tools to detect processes and functions that can be related to the disease process and gain new insights. Analyses also focus on identifying trends among known disease-associated genes such as determining whether they tend to be overexpressed in the organs with high disease incidence, as done previously (Lage *et al.*, 2008), or the proportion of tissue specific genes correlating with disease incidence.

2.1 Data sources and processing

Epidemiological data on cancer incidence rates for multiple organs were obtained from the United States Cancer Statistics (United States Department of Health and Human Services, 2009), referring to the 1999–2005 period and using data from all ages and ethnic groups. For both men and women, age-standardized cancer incidence rates per 100 000 person-years were used. This cancer incidence data presented heart and skeletal muscle together as 'Soft Tissue including Heart'. In order to estimate the incidence for heart and skeletal muscle separately, data belonging to the Northern Ireland Cancer Registry was used to estimate the percentage of cancers in the heart and skeletal muscle belonging to the soft tissue including heart cancer registries. Based on data from the Northern Ireland Cancer Registry (http://www.qub.ac.uk/research-centres/nicr/), it was estimated that 4.2% of soft tissue including heart cancer is due to cancer in skeletal muscle and that 0.68% is due to heart cancer. These percentages were applied to the US data for incidence in soft tissue to estimate the incidence of cancers in the heart and skeletal muscle.

A number of environmental factors, including smoking, infections, alcohol consumption and diet, are known to impact on the incidence of specific cancer types. Even though it is impossible to control for all these factors and focus only on intrinsic biological determinants of cancer in different organs, and because smoking is such a major, specific and artificial risk factor for lung cancer, only data from never-smokers was used for lung cancer incidence (Thun *et al.*, 2008). Moreover, data from Arab Gulf States (Al-Madouj and Al-Zahrani, 2005) was used as a second population to validate results from the US population, since an Arab population is expected to have a much lower alcohol consumption.

Because larger organs will have more cells, cancer incidence rates were normalized to the size of the organ. Although there are differences in cell density between organs, these are difficult to quantify and thus organ weight was used as an approximation to correct for any bias due to size. Typical or average organ weights were obtained from standard sources (Crichton-Browne, 1879; de la Grandmaison *et al.*, 2001; Ludwig, 2002; Nagaoka *et al.*, 2004; White *et al.*, 1987). To standardize the results, weight of tissues were normalized by the mean weight of women (58 kg) and men (70 kg), when applicable. Mass-normalized cancer incidence (CI_{organ}) was obtained by dividing the age-adjusted cancer incidence rate (CI_{age}) by typical organ weight (M_{organ}) divided by the gender weight (M_{gender}):

$$CI_{organ} = \left(\frac{CI_{age}}{M_{organ} / M_{gender}} \right)$$

When applicable, the average of CI_{organ} for males and females was used. Data employed in our calculations plus CI_{organ} results for all organs are available as Supplementary Material and on our website (http://genomics.senescence.info/cancer/tissues.html).

Tissue-specific gene expression data from the Genomics Institute of the Novartis Research Foundation (GNF data) was used in this study. GNF data was generated using both custom-designed and Affymetrix arrays that interrogate the expression of the vast majority of protein-encoding human genes and were used to profile a panel of 79 human tissues (Su *et al.*, 2004). Normalized gcRMA-condensed data was downloaded from the BioGPS portal (http://biogps.gnf.org/downloads/). Cancerous tissues in the GNF data were excluded and only healthy tissues/organs were used. Probe sets with a maximum expression <150 were filtered, resulting in 7737 probes. Mappings between GNF and Affymetrix arrays' probes and genes were obtained from Gene Expression Omnibus and from the BioGPS portal.

Gene expression GNF data were matched to mass-normalized cancer incidence. Where possible, suborgans were used; however, limitations in either the GNF dataset or cancer incidence meant that in most cases (e.g. uterus and brain) whole organs were used. Data on colon and rectum were combined. Gene expression data on lymph nodes was mapped to lymphoma cancer incidence and data on bone marrow was mapped to leukemia cancer incidence, as these are the tissues in which these cancer types most often originate. In total, mass-normalized cancer incidence was obtained for 22 tissues in the GNF dataset (Supplementary Table S1).

Data on cancer mutations in germline and somatic tissues were obtained from the COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic/) (Bamford *et al.*, 2004).

2.2 Identifying genes and processes associated with cancer incidence

Both gene expression and mass-normalized cancer incidence data were log-transformed and a regression analysis was performed between these two variables for the 7737 probes in the GNF dataset passing our filters using custom R scripts. Pearson's correlation coefficients and *P*-values were obtained that reflect the correlation between a probe's gene expression signal across tissues and tissue-specific cancer incidence. To assess the impact of outliers, results with the Pearson's correlation were compared with those obtained using the Jackknife correlation. Correlation coefficients correlated strongly between the two ($r^2 = 0.84$), with the most significant genes from the Pearson's correlation also highly significant when using the Jackknife correlation (Supplementary Fig. S1). As such, this shows that the results obtained are not due to individual outliers.

Using US data for all organs, 161 probes were significant at $P < 0.05$ with Benjamini–Hochberg correction (Benjamini and Hochberg, 1995); *P*-values from the correlation analysis below 1.05×10^{-3} were deemed significant. When excluding muscle and heart for the validation analysis, no probes were significant with Benjamini–Hochberg correction and therefore the top 1% probes (i.e. 77 probes) were used for downstream functional enrichment analysis.

To identify functions and processes correlated with cancer incidence in healthy tissues, significant probes were split into probes with a positive correlation and those with a negative correlation with cancer incidence. The genes corresponding to these probes were then analyzed in DAVID, a web-accessible set of tools that allow researchers to infer the biological meaning behind large lists of genes (Huang da *et al.*, 2009).

To detect genes with tissue-specific expression, gene expression signals for healthy tissues were averaged (*S*) and standard deviation (SD) for all probes was calculated. Probes with tissue-specific expression in a given tissue were defined as probes with a signal intensity $> S + \mathrm{SD} \times 2$. The number of tissue-specific probes for each organ was counted and log-transformed. The correlation between the number of probes with tissue-specific expression patterns and cancer incidence was determined using standard regression analysis.

Genes in which mutations have been associated with cancer were analyzed for correlations with cancer incidence. Since very few of these genes exhibited significant associations with cancer incidence from the above analysis, the number of all probes targeting cancer-associated genes with a positive and negative correlation with cancer incidence were counted. Significant deviations from the expected ratio (59.4% genes with positive

and 40.6% with negative correlations, based on the average for all probes) were determined using a cumulative binomial test.

3 RESULTS

In this study, we sought to determine whether it is possible to gain insights into disease development processes by comparing gene expression profiles of healthy organs with organ-specific disease incidence and, if so, develop a method that performs such analyses. Organ-specific disease incidence can be obtained from epidemiological studies, but because larger organs will have more cells we think it is crucial to normalize disease incidence by organ weight. Organ-specific gene expression profiles are publicly available in databases such as the GNF dataset (Su *et al.*, 2004). Genes can then be scanned for those whose expression correlates with mass-normalized disease incidence and common pathways detected by functional enrichment analysis methods (de Magalhaes *et al.*, 2010). Known disease-associated genes can also be analyzed for statistically significant patterns across organs of varying disease incidence.

To test our approach, we focused on the large observed differences in cancer incidence between human organs. Data on cancer incidence rates by site was obtained from the USA (United States Department of Health and Human Services, 2009). Since our focus was on intrinsic biological features of each organ and its cells, we tried to minimize effects of environmental factors and as such used lung cancer incidence data from never-smokers. To account for the differing numbers of cells between organs, tissue-specific cancer incidence rates were normalized by organ weight (see Section 2). Nonetheless, it is interesting to note that cancer incidence did not correlate with organ weight (Supplementary Fig. S2); yet, because our focus is on cellular properties that may predispose to cancer the use of mass-normalized cancer incidence will only accentuate those properties.

After normalizing, we obtained mass-normalized organ-specific cancer incidences for multiple organs (Fig. 1). The prostate had the highest cancer incidence rate, whereas skeletal muscle had the lowest rate. There was a large variance in mass-normalized cancer incidence, spanning six orders of magnitude from prostate to skeletal muscle. Interestingly, five of the eight tissues with the highest mass-normalized cancer incidence are related to reproduction (Fig. 1).

To identify genes and processes associated with cancer incidence, we mapped mass-normalized cancer incidence with organs in the microarray GNF tissue-specific gene expression dataset (see Section 2). We then correlated the gene expression signal of each microarray probe across organs with the mass-normalized organ cancer incidence. Overall, we found 161 significant probes (Benjamini–Hochberg test; see Section 2). The genes corresponding to the top 20 probes are displayed in Table 1 and include several ribosomal genes as well as proteins that are part of the transcriptional

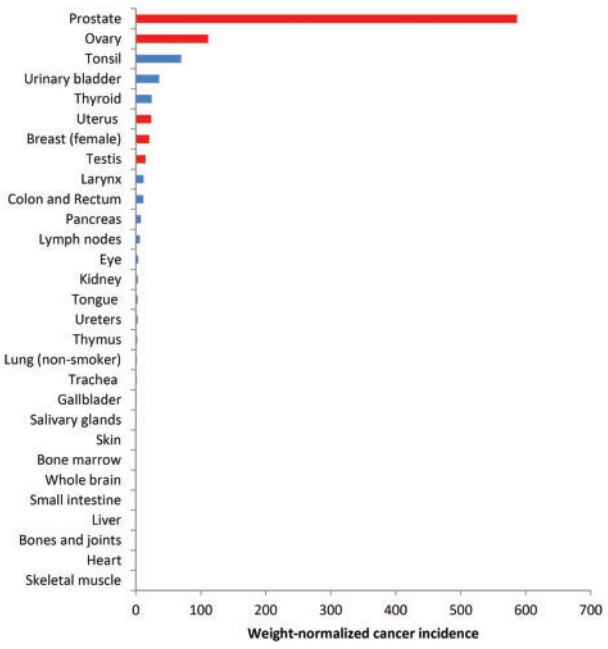

Fig. 1. Weight-normalized cancer incidence per organ. Organs associated with reproduction are highlighted in red. Data for all organs is provided in Supplementary Table S2.

Table 1. Top 20 genes most strongly correlated with cancer incidence

Gene symbol	Gene name	*P*-value	r^2	*n* significant probes
RPL3P7	Ribosomal protein L3 pseudogene 7	4.6×10^{-8}	0.78	1
RPL3	Ribosomal protein L3	8.4×10^{-8}	0.77	4
EEF1A1	Eukaryotic translation elongation factor 1 alpha 1	7.3×10^{-6}	0.64	4
C6orf48	Chromosome 6 open reading frame 48	1.4×10^{-5}	0.62	1
RPS3A	Ribosomal protein S3A	1.6×10^{-5}	0.61	3
HMGN1	High-mobility group nucleosome binding domain 1	1.9×10^{-5}	0.61	1
RPS27	Ribosomal protein S27 (metallopanstimulin 1)	2.6×10^{-5}	0.60	1
C6orf106	Chromosome 6 open reading frame 106	2.6×10^{-5}	0.59	1
PABPC1	Poly(A) binding protein, cytoplasmic 1	3.0×10^{-5}	0.59	1
RPL41	Ribosomal protein L41	3.5×10^{-5}	0.58	1
RPL5	Ribosomal protein L5	3.6×10^{-5}	0.58	2
BTF3	Basic transcription factor 3	3.7×10^{-5}	0.58	2
RPL15	Ribosomal protein L15	3.8×10^{-5}	0.58	1
RPS7	Ribosomal protein S7	4.8×10^{-5}	0.57	1
RPL22	Ribosomal protein L22	5.6×10^{-5}	0.56	3
RPN2	Ribophorin II	5.7×10^{-5}	0.56	3
TACC1	Transforming, acidic coiled-coil containing protein 1	5.7×10^{-5}	0.56	1
RPL9	Ribosomal protein L9	5.7×10^{-5}	0.56	1
PABPC3	Poly(A) binding protein, cytoplasmic 3	5.9×10^{-5}	0.56	1
SON	SON DNA binding protein	6.4×10^{-5}	0.56	1

For genes with multiple microarray probes, values shown refer for the most significant probe. All genes exhibited a positive correlation with cancer incidence, except for C6orf106 which exhibited a negative correlation.

Table 2. Clusters from DAVID with an enrichment score >2.5 are displayed

Cluster	Enrichment score	Categories	Benjamini
1	11.9	Translational elongation	8.7×10^{-28}
		Ribosome	5.6×10^{-22}
		Protein biosynthesis	7.2×10^{-19}
		Translation	4.9×10^{-18}
		rRNA processing	2.6×10^{-4}
		ncRNA processing	0.0095
2	3.6	Nuclear lumen	1.2×10^{-4}
3	2.8	RNA-binding	5.7×10^{-9}
		mRNA splicing	0.0024
		Methylation	0.0047

Categories within clusters were chosen based on the most informative annotations among those in the cluster with Benjamini–Hochberg <0.05.

machinery. In addition, there were some poorly studied genes, like *C6orf48*, *C6orf106* and *TACC1* (Table 1). Two other significant genes of note were taurine upregulated 1 (*TUG1*) and WD repeat and SOCS box-containing 1 (*WSB1*), both of which positively correlated with cancer incidence ($P=1.1\times10^{-4}$ and 1.8×10^{-4}, respectively). The vast majority of top probes (132 out of 160) correlated positively with cancer incidence. In the top 20, only *C6orf106* showed a negative correlation with cancer incidence. Our full results are available as Supplementary Material and on our website (http://genomics.senescence.info/cancer/tissues.html).

The significant probes were matched with their gene symbols and were then analyzed in DAVID to identify common functions and processes. Probes with positive and negative correlations with cancer incidence were analyzed separately. Among genes with a positive correlation with cancer incidence, we identified four significant clusters using DAVID, all of which containing categories with a significant enrichment even after Benjamini–Hochberg correction (Table 2). By far, the most significant cluster from DAVID included ribosomal genes, genes involved in transcription and protein synthesis. Among genes that correlated negatively with cancer incidence, only categories related to muscle were significant and thus assumed to be a bias caused by the low cancer incidence of skeletal muscle and heart (data not shown).

Given the above potential bias due to skeletal muscle, and because muscle has been reported to have unique ribosomal biogenesis (Thorrez *et al.*, 2008), the analysis described above was repeated after eliminating skeletal muscle and heart. However, ribosomal activity still appeared in DAVID analyses as the top, highly significant cluster (enrichment score of 8.1) with ribosome (Benjamini $=9.9\times10^{-13}$), translational elongation (1.3×10^{-10}) and protein biosynthesis (7.6×10^{-11}) having highly significant enrichments. Similarly, we repeated our analysis using cancer incidence data from an Arab population (see Section 2), in order to validate our results in a population under different environment and in particular one in which alcohol consumption is low, and results were largely confirmatory with the top cluster (enrichment score of 15.1) encompassing highly significant categories involving transcription, translation and ribosomes (data not shown).

We then investigated whether the expression of genes in which mutations have been associated with cancer may be related to organ-specific cancer incidence. Several genes in which mutations have been associated with cancer were among our significant genes, including *RPL22* ($P=5.6\times10^{-5}$) and *SFPQ* (7.8×10^{-5}). Although there were slightly more probes with a positive correlation with organ-specific cancer incidence than a negative correlation (59.4 and 40.6%, respectively), among probes ($n=218$) targeting genes in which mutations have been associated with cancer about three-fourth had a positive correlation with organ-specific cancer incidence while one-fourth had a negative correlation. This difference was statistically significant ($P<10^{-5}$).

Lastly, for each tissue we counted the number of genes expressed in a tissue-specific fashion (see Section 2) and determined whether this in turn could also be related to mass-normalized cancer incidence, but found no significant correlation (data not shown).

4 DISCUSSION

Gene expression analyses of pathological conditions have been widely used, yet few studies have focused on comparisons across healthy tissues to identify normal tissue-specific pathways that contribute to disease. One can consider human organs as different experimental samples with varying disease incidences and understanding these differences could provide new biological insights into disease etiology. As a proof of concept, we focused on cancer incidence and developed a simple method to determine genes and processes associated with organ-specific cancer incidence. Elucidating the biological reasons for differences in cancer incidence among organs could help our understanding, diagnosis and treatment of cancer.

Although to our knowledge no systematic studies have been conducted to date, it is expected that the number of cells within a tissue is proportional to cancer risk (Albanes and Winick, 1988). We were hence surprised to find that organ cancer incidence does not correlate with organ weight (Supplementary Fig. S2). On the other hand, the two heaviest organs in our analysis (muscle and bone) have relatively low cancer incidence, whereas the organ with the highest incidence (prostate) is a relatively small organ. Our results provide no evidence then that organ size determinants, like progenitor cell number (Stanger *et al.*, 2007) and p53-mediated stress and apoptosis (Mesquita *et al.*, 2010), contribute to cancer. Since no doubt a lower number of cells will make it less likely to develop cancer, however, there must be strong features in cells from tissues with high cancer incidence that predispose them to cancer development. To find these features, we used mass-normalized cancer incidence to control for number of cells in organs and emphasize cellular properties that may predispose to cancer.

To our knowledge, our work is the first to calculate cancer incidence across organs controlling for the organ's weight. This normalization provides a measure of organ-specific cancer incidence that can serve as basis for further studies, in particular since a large variation in mass-normalized cancer incidence was observed for healthy human organs. Strikingly, we observed that reproductive organs were overrepresented among those with the higher rates of cancer incidence (Fig. 1). This may reflect evolutionary trade-offs involving selective pressures related to reproduction. Reproductive organs may be under stronger evolutionary selection, because reproduction is more important than late-life survival and thus alleles that favor reproduction early in life will be selected for even if they are deleterious later in life, as predicted by evolutionary theory (Kirkwood and Austad, 2000). Indeed, although cancer is an age-related disease, it is interesting to note that testicular cancer is

the most prevalent cancer type in men aged 15–34 years (Bosl and Motzer, 1997). From a physiological perspective, it is also possible that the fact that reproductive tissues are more responsive to hormones plays a role in their increased cancer incidence.

By employing microarray data across healthy human tissues, we identified multiple genes associated, most positively but many negatively too, with organ-specific cancer incidence. Some of these genes, like *C6orf48*, *C6orf106*, *RPN2*, *TACC1*, *TUG1* and *WSB1*, may merit further study and thus our work provides candidate genes for future experiments. In particular, *RPN2* expression has already been associated with drug resistance in breast cancer (Honma *et al.*, 2008), while *TACC1* could play a role in translation and cell division and is a candidate cancer gene (Conte *et al.*, 2003) and *WSB1* has been associated with pancreatic cancer progression (Archange *et al.*, 2008). Another gene that may merit further study is *TUG1*, a non-coding RNA associated with development and shown to repress p53-dependant cell-cycle regulation (Khalil *et al.*, 2009), but to our knowledge not previously associated with cancer.

The results from our genome-wide scan showed a strong correlation between organ-specific cancer incidence normalized for weight and expression of genes associated with transcription and protein synthesis, most notably ribosomal proteins (e.g. RPL3, RPS3A and RPS27) and proteins associated with transcriptional activity like EEF1A1, HMGN1, PABPC1 and BTF3 (Tables 1 and 2). Our results were highly statistically significant and were consistent when muscle and heart tissues were excluded from the correlation analysis, as these have been reported to have unusual patterns of ribosomal activity (Thorrez *et al.*, 2008), and when using an Arab population which is exposed to different environmental cancer risks.

Translational control and ribosome biogenesis are associated with cell growth and proliferation and the loss of key points during protein synthesis might contribute to the initiation and progression of cancer (Clemens and Bommer, 1999; Holland *et al.*, 2004; Ruggero and Pandolfi, 2003). Overexpression of ribosomal proteins, in fact, has been consistently associated with tumorigenesis (Ruggero and Pandolfi, 2003), and changes in proto-oncogenes and tumor suppressor genes that occur in cancer often cause an upregulation of ribosome biogenesis (Montanaro *et al.*, 2008). Similarly, translation components have been found overexpressed in some cancers (Dua *et al.*, 2001), including EEF1A1 whose altered expression has been linked to transformation (Clemens and Bommer, 1999; Lamberti *et al.*, 2004). It is plausible that organs with a higher cancer incidence contain a larger fraction of proliferating cells and this is reflected in higher ribosomal biogenesis and transcription. That said, one caveat of our approach is that for most organs we employ cancer incidence and gene expression values that represent average values for the organ, yet differences between specific cell populations (like epithelial cells or stem cells) could be important in cancer development and may to some degree influence our results.

Since increased ribosomal biogenesis, transcription and protein synthesis are associated with cell growth and proliferation, which in turn is a hallmark of cancer, our results are not surprising even if they validate our rationale. One hypothesis is that a higher activation of the transcriptional machinery decreases the number of steps necessary for cancer to evolve since activation of ribosomal biogenesis and transcription are frequent hallmarks of cancer. In other words, higher protein synthesis may 'prime' cells for neoplastic development, perhaps by predisposing cells to high proliferation. Therefore, our results fit in with the notion that while mutations leading to the development of cancer might affect any tissue, those organs with more active cells are much more likely to develop cancer since cancer cells themselves show increased transcription. It seems that normal cells from tissues more prone to develop cancer have characteristics typical of cancer.

We observed that genes in which mutations have been associated with cancer tend to have expression patterns across tissues that correlate positively with cancer incidence more often than they correlate negatively. As such, it seems that organs with lower cancer incidence tend to express lower levels of known cancer-associated genes, which is in line with previous results for loss-of-function cancer-associated genes, but in contrast with results for gain-of-function cancer genes (Lage *et al.*, 2008). This discrepancy may be related to the mass normalizing of cancer incidence that we employ. We also did not observe any correlation between the number of tissue-specific genes and cancer incidence, which is in line with previous results showing that cancers express tissue-specific genes with selective expression in tissues different from the tissue the cancers' originate (Axelsen *et al.*, 2007).

Although many gene expression analyses have been performed to study human disease, comparing data on healthy tissues to gather mechanistic insights is a largely unexplored area. Since it can affect so many different tissues, cancer is particularly suited for such approach. Further studies may take advantage of more powerful analytical techniques, for example by employing next-generation sequencing technologies (de Magalhaes *et al.*, 2010). Similar approaches may also be employed to study tumor progression, invasion and metastasis organ preference of various cancer types. This may be useful for studying the tissue microenvironment in which cancer develops and in particular for studying metastasis development since many cancer types follow specific metastatic patterns. As we enter the era of personalized medicine and large-scale sequencing (de Magalhaes *et al.*, 2010), including in cancer genomics (Stratton *et al.*, 2009), analyses of healthy tissues could become a powerful paradigm in oncogenomics that complement standard analyses contrasting cancer and healthy tissues. Our approach employing data from healthy tissues can also serve as foundation for analyses of other diseases and systemic diseases in particular. Healthy tissue gene expression data can thus inform about pathologies associated with the tissue and contribute to disease systems biology analyses.

5 CONCLUSIONS

Our work provides a new paradigm to study disease etiology that may be particularly suitable to study cancer. Taken together, our results demonstrate that organ gene expression background is important in cancer development and in particular suggest that tissues with higher transcriptional and translational activation are more likely to develop cancer. Although our results are somewhat confirmatory, they demonstrate how employing data from healthy tissues can provide insights into disease development. We also identified numerous genes associated with cancer incidence in healthy human tissues, including new candidate genes for further studies. Our work thus provides a framework for future research to understand variation in disease incidence across human organs using gene expression profiles from healthy organs.

ACKNOWLEDGEMENTS

The authors wish to thank Jessica Lingley and Shraddha Sankhe for assistance in collecting cancer statistical data, Venkatesh Patel for helping with functional enrichment analyses, Ilídio Correia for useful suggestions and to all group members for support and fruitful discussions. Further thanks to Roger Barraclough, Nuria Lopez-Bigas and Philipp Bucher for comments on previous drafts of the manuscript.

Funding: Biotechnology and Biological Sciences Research Council grant (BB/H008497/1) (to J.P.dM.); Ellison Medical Foundation (to J.P.dM.); Marie Curie International Reintegration Grant within European Commission-FP7 (to J.P.dM.); Erasmus programme (to A.S.S. and S.B.).

Conflict of Interest: none declared.

REFERENCES

Al-Madouj,A.N. and Al-Zahrani,A.S. (2005) *Eight-year cancer incidence among nationals of the GCC states: 1998—2005*. Gulf Center for Cancer Registration, Riyadh, Saudi Arabia. Available at http://www.sgh.org.sa/PDF/cancer%201998-2005.pdf

Albanes,D. and Winick,M. (1988) Are cell number and cell proliferation risk factors for cancer? *J. Natl Cancer Inst.*, **80**, 772–774.

Archange,C. *et al.* (2008) The WSB1 gene is involved in pancreatic cancer progression. *PLoS One*, **3**, e2475.

Axelsen,J.B. *et al.* (2007) Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles. *Proc. Natl Acad. Sci. USA*, **104**, 13122–13127.

Bamford,S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355–358.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Bosl,G.J. and Motzer,R.J. (1997) Testicular germ-cell cancer. *N. Engl. J. Med.*, **337**, 242–253.

Clemens,M.J. and Bommer,U.A. (1999) Translational control: the cancer connection. *Int. J. Biochem. Cell Biol.*, **31**, 1–23.

Conte,N. *et al.* (2003) TACC1-chTOG-Aurora A protein complex in breast cancer. *Oncogene*, **22**, 8102–8116.

Crichton-Browne,J. (1879) On the weight of the brain and its component parts in the insane. *Brain*, **2**, 42.

de la Grandmaison,G.L. *et al.* (2001) Organ weight in 684 adult autopsies: new tables for a Caucasoid population. *Forensic Sci. Int.*, **119**, 149–154.

de Magalhaes,J.P. *et al.* (2010) Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res. Rev.*, **9**, 315–323.

Dua,K. *et al.* (2001) Translational control of the proteome: relevance to cancer. *Proteomics*, **1**, 1191–1199.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Holland,E.C. *et al.* (2004) Signaling control of mRNA translation in cancer pathogenesis. *Oncogene*, **23**, 3138–3144.

Honma,K. *et al.* (2008) RPN2 gene confers docetaxel resistance in breast cancer. *Nat. Med.*, **14**, 939–948.

Huang da,W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Khalil,A.M. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 11667–11672.

Kirkwood,T.B. and Austad,S.N. (2000) Why do we age?, *Nature*, **408**, 233–238.

Lage,K. *et al.* (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA*, **105**, 20870–20875.

Lamberti,A. *et al.* (2004) The translation elongation factor 1A in tumorigenesis, signal transduction and apoptosis: review article. *Amino Acids*, **26**, 443–448.

Ludwig,J. (2002) *Handbook of Autopsy Practice*. Humana Press, Totowa, NJ.

Mesquita,D. *et al.* (2010) A dp53-dependent mechanism involved in coordinating tissue growth in Drosophila. *PLoS Biol.*, **8**, e1000566.

Montanaro,L. *et al.* (2008) Nucleolus, ribosomes, and cancer. *Am. J. Pathol.*, **173**, 301.

Nagaoka,T. *et al.* (2004) Development of realistic high-resolution whole-body voxel models of Japanese adult males and females of average height and weight, and application of models to radio-frequency electromagnetic-field dosimetry. *Phys. Med. Biol.*, **49**, 1.

Ruggero,D. and Pandolfi,P.P. (2003) Does the ribosome translate cancer? *Nat. Rev. Cancer*, **3**, 179–192.

United States Department of Health and Human Services. (2009) *United States Cancer Statistics: 1999-2005, WONDER On-line Database*. Centers for Disease Control and Prevention and National Cancer Institute.

Stanger,B.Z. *et al.* (2007) Organ size is limited by the number of embryonic progenitor cells in the pancreas but not the liver. *Nature*, **445**, 886–891.

Stratton,M.R. *et al.* (2009) The cancer genome. *Nature*, **458**, 719–724.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062.

Thorrez,L. *et al.* (2008) Using ribosomal protein genes as reference: a tale of caution. *PLoS One*, **3**, e1854.

Thun,M.J. *et al.* (2008) Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. *PLoS Med.*, **5**, e185.

Ward,J.M. *et al.* (1993) Cell proliferation not associated with carcinogenesis in rodents and humans. *Environ. Health Perspect.*, **101** (Suppl 5), 125–135.

White,D.R. *et al.* (1987) Average soft-tissue and bone models for use in radiation dosimetry. *Br. J. Radiol.*, **60**, 907.