

**Searching for Ancient Gene Expression Signatures in Tumours**

Anna Sofia Trigos Gomez

Submitted in partial fulfilment of the requirements of the degree of Master of Science (Bioinformatics) (with coursework component)

November 2014

Faculty of Science

The University of Melbourne

**Abstract**

Many of the main features of multicellularity, such as an effective cell-to-cell adhesion, communication and differentiation, are lost in cancer. One explanation is that cancer cells undergo a process of atavism, where they reactivate ancient transcriptional programs that evolved during the emergence of primitive organisms. However, to our knowledge, there have not been studies that directly address empirical validation of this hypothesis. This study aimed to search for evidence of a possible atavism process in cancer cells by assessing the activation of ancient genes and functions in four tumour types, taking both gene-level and systems-level approaches.

Our results suggest that lung, breast and head and neck tumour cells are expressing an older transcriptome than their normal counterparts due to the combined effect of two opposing trends: the overexpression of ancient genes that date back to the emergence of cellular and eukaryotic life, and the decreased expression of post-metazoan genes. Additionally, our gene-set analysis of functionally related genes points towards the reactivation of ancient transcriptional programs. Finally, we discuss the relationship between candidate biological mechanisms that are activated in tumours and their role in the atavism.

**Declaration**

*This is to certify that:*

1. *The thesis comprises only my original work towards the masters except where indicated in the Preface,*
2. *Due acknowledgement has been made in the text to all other material used,*
3. *The thesis is less than 15,000 words in length, exclusive of text in images, tables, bibliographies and appendices.*

**Preface**

This research project was carried out at the Peter MacCallum Cancer Centre under the supervision of Dr David Goode.

**Acknowledgements**

I would like to thank Dr David Good for giving me the opportunity to work on this research topic and for his unwavering support and guidance during this journey.

I thank my mother, grandfather and late father for encouraging me to embark on postgraduate studies and all their continuous support during my degree.

I also want to thank all the lecturers I have had during my life. Each one of you have imparted a bit of the wisdom I needed to pursue this project.

Finally, I would like to thank my classmates for their comradery.

Table of Contents

[List of tables and figures viii](#_Toc404262869)

[Introduction 1](#_Toc404262870)

[Literature review 2](#_Toc404262871)

[Accumulation of mutations leads to cancer initiation and progression 2](#_Toc404262872)

[Link between cancer and unicellular life 2](#_Toc404262873)

[The role of stress in cancer and the acquisition of similarities with unicellular organisms 3](#_Toc404262874)

[Atavism hypothesis of cancer 4](#_Toc404262875)

[Cells can lose features of multicellularity 5](#_Toc404262876)

[Research Question and Objectives 6](#_Toc404262877)

[Significance 7](#_Toc404262878)

[Methodology 8](#_Toc404262879)

[1. Classification of human genes by their point of emergence 8](#_Toc404262880)

[1.1 Determining the founders of human genes 8](#_Toc404262881)

[1.2 Comparison of the assignment of phylostrata to human genes by other authors 9](#_Toc404262882)

[1.3 Verification of the phylostrata of emergence of genes by enrichment of functional categories 9](#_Toc404262883)

[2. Age of the cancer transcriptome 11](#_Toc404262884)

[2.1 Gene expression data 11](#_Toc404262885)

[2.2 Transcriptome age index 11](#_Toc404262886)

[2.3 Identifying trends in the shifts of expression of phylostrata in tumours 12](#_Toc404262887)

[2.4 Gene-level interpretation of the change of the TAI in tumour using non-parametric tests 12](#_Toc404262888)

[2.5 Discerning the age of the overlapping differently expressed genes (DEG) in LUAD, BRCA, HNSC and KIRC 13](#_Toc404262889)

[3. Evidence of atavism in gene set analysis 13](#_Toc404262890)

[3.1 Identifying the functional gene sets that change their expression in the phylostrata 13](#_Toc404262891)

[3.2 Association between the ancientness of a gene function and its expression level 15](#_Toc404262892)

[3.3 Change in expression of genes of functional sets across phylostrata 15](#_Toc404262893)

[4. Programming languages 15](#_Toc404262894)

[Results 16](#_Toc404262895)

[1. Phylostratigraphic analysis of human genes 16](#_Toc404262896)

[2. The age of the cancer transcriptome 19](#_Toc404262897)

[2.1 Possible confounders of the TAI 20](#_Toc404262898)

[2.2 Interpreting the shift of the TAI 25](#_Toc404262899)

[2.3 Ancientness of the core set of DEG in tumours 33](#_Toc404262900)

[3. Functional studies 35](#_Toc404262901)

[3.1 Differentially expressed gene functions in each phylostratum 35](#_Toc404262902)

[3.2 Dating GO terms 36](#_Toc404262903)

[3.3 Enrichment of ancient functions in tumours 38](#_Toc404262904)

[3.4 Identifying the gene sets that undergo a process of atavism 39](#_Toc404262905)

[Discussion 40](#_Toc404262906)

[1. Phylostratigraphy captures the major evolutionary changes that led to the current human genome composition 40](#_Toc404262907)

[2. Patterns in the up- and downregulation of genes in cancer can be traced back to their ancientness 41](#_Toc404262908)

[2.1 Tumours have an older transcriptome than normal cells 41](#_Toc404262909)

[2.2 Pre-multicellularity genes are upregulated in tumours, whereas later genes are downregulated 42](#_Toc404262910)

[2.3 Multiple confounders led to the lack of strong and clear signals from HNSC and KIRC samples 42](#_Toc404262911)

[2.4 A core set of DEG detected in the four cancer types is enriched with pre-multicellularity genes 43](#_Toc404262912)

[3. Ancient functions are upregulated in tumours 43](#_Toc404262913)

[3.1 The age of the GO terms parallels widespread evolutionary innovations 43](#_Toc404262914)

[3.2 There is a signature set of gene functions differentially expressed in the phylostrata 44](#_Toc404262915)

[3.3 Ancient gene functions are upregulated in tumours 44](#_Toc404262916)

[3.4 Evidence of atavism can also be found at the functional-set level 45](#_Toc404262917)

[4. Activation of ancient mechanisms in cancer 45](#_Toc404262918)

[4.1 The Warburg effect 45](#_Toc404262919)

[4.2 An ancient stress response 46](#_Toc404262920)

[4.3 Selection pressures and atavism 47](#_Toc404262921)

[Conclusions 48](#_Toc404262922)

[Further work 49](#_Toc404262923)

[List of References 50](#_Toc404262924)

[Supplementary information 58](#_Toc404262925)

[Supplementary methods and results 58](#_Toc404262926)

[S1. RNAseq expression data 58](#_Toc404262927)

[S2. Estimation of expression fold changes and identification of differentially expressed genes (DEG) 60](#_Toc404262928)

[S3. Comparing the age of the tumour and normal transcriptomes using microarray data 65](#_Toc404262929)

[Supplementary tables 70](#_Toc404262930)

# List of figures and tables

[Figure 1. Phylogenetic tree used for the classification of human genes into phylostrata 16](#_Toc404262931)

[Figure 2. Histogram of the number of OrthoMCL species allocated to each phylostratum 17](#_Toc404262932)

[Figure 3. Distribution of the number of genes assigned to the phylostrata with two different phylogenetic frameworks 18](#_Toc404262933)

[Figure 4. Heatmap of the enriched functional categories in the phylostrata 19](#_Toc404262934)

[Figure 5. Transcriptome Age Index (TAI) of normal and tumour LUAD, BRCA, HNSC and KIRC samples 20](#_Toc404262935)

[Figure 6. Correlation between library size and TAI of LUAD, BRCA, HNSC and KIRC normal and tumour samples 21](#_Toc404262936)

[Figure 7. Comparison of the TAI of BRCA tumour and normal samples after different phylostratigraphy techniques 23](#_Toc404262937)

[Figure 8. Correlation between the TAI and purity of KIRC tumour samples 24](#_Toc404262938)

[Figure 9. TAI of paired normal-tumour samples in LUAD, BRCA, HNSC and KIRC 25](#_Toc404262939)

[Figure 10. Distribution of the tumour-normal expression proportion differences across phylostrata 27](#_Toc404262940)

[Figure 11. Boxplots of the ratio of tumour/normal expression proportions of paired samples 29](#_Toc404262941)

[Figure 12. Overlap of DEG between cancer types 33](#_Toc404262942)

[Figure 13. Up- and downregulated functions in the phylostrata 36](#_Toc404262943)

[Figure 14. Distribution of sixth-generation GO terms by their phylostratum of emergence 37](#_Toc404262944)

[Figure 15. Grouping of the point of emergence of GO-terms into umbrella phylostrata 38](#_Toc404262945)

[Table I Number of samples by tumour/normal type after filtering 11](#_Toc404263025)

[Table II Number of species with GO annotations assigned to each phylostratum 14](#_Toc404263026)

[Table III. Wilcoxon two-sided p-values for the difference of the means of the tumour and normal expression proportions of the four cancer types 28](#_Toc404263027)

[Table IV. Two-sided p-values of the ratio of the tumour/normal expression proportions 30](#_Toc404263028)

[Table V. Two-sided p-values for positive and negative enrichment of phylostrata 31](#_Toc404263029)

[32](#_Toc404263030)

[Table VII. Two- and one-sided p-values of the Fisher Exact tests for enrichment of phylostrata in the genes of the core set of overlapping DEG 34](#_Toc404263031)

[Table VIII. The number of up- and downregulated functions in the umbrella phylostratum and the two-sided p-values of enrichment tests 39](#_Toc404263032)

[Table IX. Summary of functional sets that underwent atavism 39](#_Toc404263033)

# Introduction

Cancer is one of the leading diseases worldwide, with 8.2 million deaths occurred in 2012 due to cancer (World Health Organization, 2014). Although great research efforts in the last decades have led to considerable improvements in the diagnosis, treatment and prognosis of patients, much is still to be understood about the biology of cancer.

The development of next-generation sequencing has allowed viewing the larger picture of cancer, and has opened doors to analysing cancer as a complex system. Furthermore, a greater multidisciplinary approach to researching cancer has allowed novel interpretations to come forth.

Parallelisms have been made between the main features of cancer and important transitions in the evolution of life. The atavism hypothesis of cancer suggests that cancer cells are regressing to an ancient state of unicellularity through the re-activation of mechanisms of survival that had been suppressed with the emergence of multicellularity. However, validation of the association of a cancerous state and the expression of primitive genes has not been previously explored.

Therefore, this study aimed to search for evidence of atavism by focusing on changes in gene expression between tumour and normal cells in the context of the ancientness of the genes. Under this framework, we expect that the re-activation of ancient mechanisms should be evidenced by the increase of expression of ancient genes and the underexpression of newer genes. Various complementary approaches were taken to verify this hypothesis, using systems-, gene- and functional-level analyses, and relying on four distinct tumour types as models.

# Literature review

## Accumulation of mutations leads to cancer initiation and progression

Cancer is characterized by the abnormal growth of cells caused by the dysregulation of various pathways leading to a sustained growth, evading growth suppressors and resisting cell death (Hanahan & Weinberg, 2011). Being primarily a genetic disease, the identification of driver mutations and other genomic alterations has been set as a priority in cancer research (Hahn et al., 2009).

Theories such as the classical somatic evolution of cancer explain that the accumulation of mutation events throughout the lifetime of a cell may allow the cell to break free from the constraints of normal cell proliferation, leading to an uncontrolled autonomous proliferation (Podlaha, Riester, De, & Michor, 2012).

Cancers have been traditionally viewed as undergoing a Darwinian process, where natural selection of the cells with the best fitness (proliferative and survival advantage) drives cancer progression to ever more malignant phenotypes (Greaves, 2002). The mutation rate is increased during cancer progression in comparison to the spontaneous mutation events in normal cells, which has led to the proposal of a mutator phenotype in cancer (Loeb, Bielas, & Beckman, 2008; Loeb, 2001), which would account for their rapid increase in fitness in very short time frames. The mutator phenotype is thought to arise in response to environmental carcinogens, errors during replication and repair mechanisms, and endogenous damage from reactive species (Loeb, 2011). Basically, the acquisition and progression of cancer has been attributed to random events.

## Link between cancer and unicellular life

As early as 1996, the idea that such a resilient disease was caused by mutations occurring in random order seemed counter-intuitive (Israel, 1996). More recently, this view has been retaken by many other authors (Davies & Lineweaver, 2011; Vincent, 2010; Vincent, 2012), whose main argument is that random sequential mutational events cannot satisfactorily explain the diverse set of traits expressed by cancer cells.

A striking characteristic of cancer is its similarity with simpler, unicellular or colonial organisms. The intricate network of proliferation control exhibited by multicellular organisms is lost in cancer, and cells start behaving as rogue and selfish cells. Main features of complex multicellularity, such as effective cell to cell adhesion, communication, differentiation (Knoll, 2011; Rokas, 2008) are lost in cancer (Hanahan & Weinberg, 2011). Actually, it has been found that the emergence of cancer genes can be traced back to the origin of cellular organisms as well as multicellularity (Domazet-Loso & Tautz, 2010), and tumours have been identified in a primitive multicellular organism, hydra (Domazet-Lošo et al., 2014).

Unicellular organisms also use a ‘mutator phenotype’ as a defence mechanism, because these tend to favour DNA replication under adverse conditions, even if this genetic information is aberrant (Israel, 1996), given that more genomic variability prepares them for any adversities they might encounter (Cahill, Kinzler, Vogelstein, & Lengauer, 1999). Therefore, parallelisms can be drawn between the ‘mutator phenotype’ and the massive unicellular cell response where gene expression is highly noisy and altered (López-Maury, Marguerat, & Bähler, 2008).

Cancer has also been described as an ecosystem of individual cells, where individual members are subjected to evolutionary and ecological processes, such as natural selection, genetic drift, artificial selection by medical treatment, competition, dispersal and colonization (Merlo, Pepper, Reid, & Maley, 2006). Processes such as metastasis can be viewed as a colonization mechanism to escape stress. Finally, immortality and asexual reproduction are also characteristics of both cancer cells and unicellular organisms (Vincent, 2012).

## The role of stress in cancer and the acquisition of similarities with unicellular organisms

The relationship between cellular stress and cancer is that of a positive feedback loop: cancer is both driven by and causes stress (Bartek, Mistrik, & Bartkova, 2012; Halazonetis et al., 2008). Each step of tumourigenesis can be related to specific stress factors: cell damage caused by stress (UV, reactive oxygen species) at the initiation process; inflammation, hostile pH, nutrient deprivation and hypoxia during progression; and loss of adhesion and attack of the immune system during metastasis (Murakami, Noguchi, Takeda, & Ichijo, 2007). A stress phenotype has even been described as the common ground of many tumours, which includes DNA damage/replication, mitotic, proteotoxic, oxidative, metabolic stresses (Luo, Solimini, & Elledge, 2009).

These stress factors are not exclusive to cancer cells. Stress conditions that ancient unicellular organisms encountered, such as radiation, nutrient deprivation, free radicals, low oxygen and pH, are also some of the challenges that cancer cells must face (Vincent, 2012). Considering that stress responses in unicellular organisms are predictable when the stressor has been previously encountered in the evolutionary history of the species (López-Maury et al., 2008), finding similarities in the stress response of cancer and unicellular species would suggest that cancer cells might be resorting to these primitive machineries to deal with these adverse conditions.

Under abrupt stress conditions, yeast cells resort to large genome rearrangements, such as selective chromosome duplications, that confer them a transient adaptive advantage (Yona et al., 2012). Aneuploidy is also commonly found in cancer cells, and is thought to enhance tumourgenesis by causing an increase in the levels of cancer-promoting genes in the cells (Gordon, Resio, & Pellman, 2012), or by causing further genetic karyotypic instability (Nicholson & Duesberg, 2009). Studies have shown that aneuploidy can diminish the negative effects of environmental stress conditions on cell growth rate, suggesting an increased adaptability of aneuploid yeast cells relative to euploid cells (Pavelka et al., 2010).

## Atavism hypothesis of cancer

Considering the behavioural similarities of cancer cells and primitive organisms, a hypothesis that explains the resilience and ability of cancer cells to survive and proliferate, is that of an atavism process. Cancer cells might be activating a highly efficient set of ancient survival genes that were honed by unicellular organisms through millions of years of evolution to survive the harsh environmental conditions encountered during that time (Davies & Lineweaver, 2011): in the course of evolution, the survival genes of these rudimentary eukaryotes were modified or suppressed to give rise to cell differentiation and more sophisticated modes of cell organization. However, these genes are still latent, and become active (proto-oncogenes are activated to oncogenes) when cell survival is compromised due to stress, inflammation or infection, and all other defence mechanisms fail (DNA repair, apoptosis, immune system, etc). This causes an irreversible atavism that shifts the focus from survival of the organism to the survival of individual cells, which will ultimately progress into malignancy. However, this theory does not rule out that randomness plays an important role, especially because the initial inactivation of the constraints of multicellularity is thought to operate by stochastic mutations (Vincent & Van Seuningen, 2012).

The degree of atavism, how far back the atavism takes cancer cells, has also been addressed. A regression to unicellular organisms which are facultatively colonial has been suggested, given that cancer can show cooperativity, competition and adherence when growing in colonies, but are also able to originate a new line from a single cell (Vincent, 2010; Vincent, 2012). In this view, cancer cells are seen as exhibiting attributes of holozoan opisthokonts, a clade which includes metazoans but excludes fungi (Vincent, 2010). A reversion to rudimentary colonial organisms, in the range of polyps, sponges and choanoflaggelates has also been suggested (Davies & Lineweaver, 2011). On the other hand, some authors have gone as far as to define cancer as another species belonging to another phylum, due to parallelisms with speciation mechanisms seen in bacteria and asexual species (Duesberg & Rasnick, 2000; Vincent, 2010).

### Cells can lose features of multicellularity

Although the association of cancer cells with unicellular life has been discussed mainly in theory, there are actually well-studied examples of cancer-related cells that lose features of multicellularity, such as a differentiated state.

In the stem cell model of cancer, tumours originate from a reduced number of cells with stem-cell like properties, named cancer stem cells, which have a large proliferative potential and are able to initiate tumour growth by generating progeny at varying degrees of differentiation (Reya, Morrison, Clarke, & Weissman, 2001; Vincent & Van Seuningen, 2012). The exact origin of cancer stem cells has not been elucidated, but one hypothesis suggests that differentiated cells dedifferentiate to a stem-like state, with oncogenetic mutations being the drives of the dedifferentiation (Shah, Patel, Pathak, Swain, & Kumar, 2014).

The epithelial–mesenchymal (E-M) transition is a fundamental step during the embryonic development of metazoans (Thiery, 2002). During this transition, loss of the highly structural epithelial tissue and the acquisition of the loosely organized phenotype of mesenchymal cells are driven by the loss of cell polarity and cell-cell adhesion, together with new migratory properties (Thiery, Acloque, Huang, & Nieto, 2009), suggesting the loss of cell multicellular properties. The E-M transition has been found to generate cancer stem cells (Mani et al., 2008), as mesenchyma cells are able to generate multiple cell types (Thiery et al., 2009). Furthermore, pathways related to the E-M transition have been identified in tumour models (Huber, Kraut, & Beug, 2005), and molecules essential for cell adhesion that are lost during this process are also an integral part of cancer progression and metastasis (Thiery, 2002).

Finally, retainment of the pluripotent capacity of differentiated cells is exemplified by induced pluripotent stem cells (iPSC) in adult somatic cells. The transduction of certain transcription factors is enough for de-differentiation and to acquire features of human embryonic stem cells, as well as proliferating extensively and being inducible to differentiate into other cell types (Takahashi et al., 2007). Additionally, iPSC have been shown to have tumorogenic properties, creating a link between iPSC and cancer stem cells (Chen et al., 2012; Nishi et al., 2014).

# Research Question and Objectives

Although the atavism hypothesis of cancer has gained momentum in the last few years, there have been no studies directly evaluating the validity of its claims. Therefore, we aimed to search for experimental evidence of the reversal of cancer cells to a primitive state.

Under the atavism hypothesis, tumour cells re-activate an ancient toolkit for survival used by primitive cells. Signatures of this re-activation are expected to be present in the transcriptome of cancer cells, namely by the increased expression of ancient genes and functions. In this project, the composition of the cancer transcriptome will be evaluated from an evolutionary perspective, in order to determine the ancientness of the expressed genes.

The following goals were set:

1. **Determine the age of human genes using phylostratigraphy.**
   1. Validate the distribution of gene ages with published results obtained by other phylostratigraphy methods.
   2. Validate the ages of the genes by investigating their functional roles in an evolutionary context.
2. **Assess the difference between the age of the transcriptome of cancer and normal cells, using lung, breast, head and neck and kidney as sample cancer types.**
   1. Compare the age of the transcriptome in tumour and normal samples using Transcriptome Age Index (TAI).
   2. Evaluate the possible confounders that might bias the above comparison.
   3. Investigate the differences in the age of the transcriptome both at the systems- and gene-levels.
   4. Search for an association between the age of a gene and its status as a core differentially expressed gene in the tumours.
3. **Search for evidence of atavism using gene set analysis.**
   1. Detect functional gene sets that drive the change of the age of the transcriptome of tumours.
   2. Find evidence of the reactivation of ancient functional mechanisms in tumours.
   3. Identify the functional gene sets that undergo a process of atavism.

# Significance

Despite major efforts, cancer initiation and progression are still not entirely understood and are under intensive research. Putting cancer in the context of a loss of an advanced state of multicellularity and the re-activation of primitive cellular mechanisms adds another dimension to the biology behind cancer. Studies in this area of research might ultimately lead to further elucidation of the pathways and molecular mechanisms that regulate cancer behind the scenes, which might shed light on possible therapeutic targets.

# Methodology

## Classification of human genes by their point of emergence

### 1.1 Determining the founders of human genes

Phylostratigraphy was employed to classify human genes according to their age by mapping them to a phylogenetic tree (Domazet-Loso & Tautz, 2008, 2010). This method relies on the concept of founder genes, which are used to determine the point of emergence of the genes (Domazet-Loso & Tautz, 2003, 2008, 2010). The founder of a gene was said to be the oldest ortholog of the given human gene, as defined in the OrthoMCL database (L. Li, Stoeckert, & Roos, 2003). The OrthoMCL database lists 15782 groups of orthologous, paralogous and co-orthologous proteins of 150 species, defined by the similarity of their amino acid sequences as well as their clustering patterns after being processed by a Markov Cluster algorithm. Protein sequences are preferred in orthology algorithms because selective pressures generally occur at the protein level, leading to more stable protein sequences than that of DNA sequences. There were a total of 21632 human proteins in the OrthoMCL database, and only the groups of orthologous proteins with at least one human protein were processed.

In the OrthoMCL algorithm, proteomes of the 150 species in the database underwent an all-against-all BLASTP, reciprocal best similarity pairs were defined as being orthologous, whereas paralogs were the sequences that were reciprocally more similar to the base orthologous pair than to any other sequence from another genome. Orthologs and paralogs were subsequently represented in a graph weighted by their similarity and an unsupervised cluster algorithm (Markov Cluster algorithm - MCL) clustered protein sequences into families.

Species with orthologs of human proteins were mapped to a phylogenetic tree using the NCBI Taxonomy Common Tree browser (Benson, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2009; NCBi Resource Coordinators, 2014; Sayers et al., 2009) (<http://www.ncbi.nlm.nih.gov/taxonomy>). Taxons with at least one species present in OrthoMCL were included in the final phylogenetic framework, which consisted of 16 taxons (phylostrata), ranging from including all cellular organisms (phylostratum 1) to only humans (phylostratum 16).

The most ancient phylostratum represented in a group of orthologs was considered as the point of emergence of the human protein. However, manual verification of the assignment of genes revealed that there were spurious sequences similarities between the human proteins and proteins of species of early phylostrata due to their large number and diversity. These false positives caused bias towards the assignment of genes to earlier phylostrata. Therefore, human proteins that had been originally classified into phylostrata 1, 2 or 3 by orthology to a single sequence from a species of either of these phylostrata were reclassified into the phylostrata of the next oldest ortholog. Additionally, preliminary analysis showed that phylostratum 3 appeared to be enriched with genes with functions related to the nervous system (*data not shown*), even when there is agreement that primitive nervous systems appeared in basal metazoans (Jacobs et al., 2007), but application of the stricter requirement correctly shifted the terms to that of the metazoan phylostrata.

The protein IDs were translated to their gene IDs using Ensembl BioMart (Kasprzyk, 2011) with the Ensembl Genome Browser 73 (EMBL-EBI & The Wellcome Trust Sanger Institute, 2013) and the *Homo sapiens* genes (GRCh37.p12) database (Genome Reference Consortium, 2013). Genes that coded for multiple proteins which belonged to different orthologous groups were excluded, given that an accurate classification of proteins into orthologous groups should contain all proteins of a given gene. Using this filtering, 49 genes were excluded from the analysis. Finally, only the human genes that had been assigned a phylostratum and that were present in the platforms used for RNAseq (see below) were kept, resulting in a total of 16736 human genes included in the analysis.

### 1.2 Comparison of the assignment of phylostrata to human genes by other authors

Other authors (Domazet-Loso & Tautz, 2008) have previously classified human genes by their age, using protein sequence similarity searches with BlastP to search for matches with species in the NR (non-redundant database) of NCBI, as well as TBlasTN searches for trace and EST archives. The 19 phylostrata defined by these authors were collapsed into the framework of 16 phylostrata used in this study for direct comparison.

### 1.3 Verification of the phylostrata of emergence of genes by enrichment of functional categories

As a way to verify the correct dating of genes, functional enrichment tests were used to find the overrepresented gene functions in the phylostrata. The functions of a gene should agree with the functional capacities of the organisms belonging to the phylostrata where the gene has been assigned.

#### 1.3.1 Generation of a curated set of Gene Ontology annotations

Gene Ontology (GO) annotations (The Gene Ontology Consortium, 2000) provided from the UniProt Consortium (GOA) (Dimmer et al., 2012) were selected to define the functions of genes. The Gene Association File for *Homo sapiens* with a GOC Validation Date 07/10/2014 and a submission date of 7/10/2014 was used. Given that our aim was to focus on the biological functions of the genes, only GO annotations from the Biological Process branch were used. Gene annotations that have conflicting functions in the literature and that were tagged with the qualifier ‘NOT’ were excluded from the analysis. Annotations inferred from electronic annotations (IEA) were also excluded, due to their low reliability (Skunca, Altenhoff, & Dessimoz, 2012).

The hierarchical nature of the GO-terms annotation tree allowed the selection of a subset of broad GO terms that provided enough functional information to compare the enriched terms between phylostrata. GO term nodes that were separated by three edges from the highest parent node were found to provide a good trade-off between specificity and generality, so the GO terms were collapsed up to this level. This set of annotations will hereafter be referred to as ‘third-generation GO terms’. Viewing of the hierarchical tree was done on the Quick-GO browser (Binns et al., 2009).

The following method was used for the collapse:

1. All ancestors of all GO terms as well as the list of GO terms that were separated by 3 edges from the highest parent node of Biological process (GO: 0008150) were obtained with the GO.db R package (Carlson, 2014), version 2.14.0, with a packaged date of 2014-03-18. Only pairs of GO terms that were connected by an edge of type ‘is a’ were considered.
2. The ancestors of each GO term were scanned, and those that were 3 edges from the highest parent node were kept.
3. An equivalency table of GO terms was created, where the corresponding third-generation GO term for any other GO terms was listed.
4. The original GO term annotations of the genes were replaced with the GO terms of the equivalency table, creating the list of third-generation annotations.

In other words, all GO annotations were mapped to the third-generation GO terms, which subsequently replaced the original GO annotations of the genes.

#### 1.3.2 Hypergeometric tests for enrichment analysis of functional terms using BiNGO

The hypergeometric tests performed by the BiNGO plug-in (version 3.0.2) (Maere, Heymans, & Kuiper, 2005) of Cytoscape (version 3.1.1) (Shannon et al., 2003) were used to test for enrichment of third generation GO-term categories. The p-values obtained from hypergeometric tests represent the probability that a random sub-sample of genes of a test set belong to a certain category, given the total number of genes in the reference set that belong to that category.

The input files were: the go.obo ontology file downloaded from the Gene Ontology website (released on the 2014-07-23, format version 1.2, csv Revision: 19061), the list of third-generation gene annotations, and the entire set of 16736 human genes as the reference. The test was performed only to assess for overrepresentation of terms from the Biological Process branch, and was performed simultaneously for all phylostrata by inputting the genes of each separate phylostrata as independent clusters of genes. Only the overrepresented categories after Benjamin & Hochberg False Discovery Rate correction (p < 0.05) were chosen to be outputted. Out of the resulting enriched terms, only third-generation GO terms were considered.

#### 1.3.3 Further classification of the enriched functional terms and their distribution across phylostrata

Comparison of enriched terms across phylostrata was achieved by further classifying the significant categories into umbrella terms. The number of enriched terms of each umbrella term found in the phylostrata was counted, and their proportion in each phylostrata was calculated by dividing the number of terms in the phylostrata by the total number of terms of the umbrella term. A heatmap was generated to visualize the results, using the heatmap.2 function from the gplots R package.

## Age of the cancer transcriptome

### 2.1 Gene expression data

Gene expression data from Breast Invasive Carcinoma (BRCA), Lung Adenocarcinoma (LUAD), Head and Neck Squamous Cell Carcinoma (HNSC) and Kidney Renal Clear Cell Carcinoma (KIRC) tumour samples and their normal counterparts were used to explore the difference of the age of the transcriptome between normal and cancer cells. The raw RSEM RNAseq expression values were obtained from The Cancer Genome Atlas (TCGA) (https://tcga-data.nci.nih.gov/tcga/). The normal samples corresponded to distal tissue (>2 cm) from the site of the tumour. Metastatic samples and samples with low median correlation were excluded from the analysis and the raw RSEM values were median-scaled (see Supplementary methods). Tumour/normal paired samples were obtained by matching patient IDs. The final number of samples is shown is Table I.

**Table I Number of samples by tumour/normal type after filtering**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **Tumour** | **Normal** | **Total** | **Paired\*** |
| **LUAD** | 393 | 57 | 450 | 55 |
| **BRCA** | 1041 | 111 | 1152 | 109 |
| **HNSC** | 495 | 42 | 537 | 40 |
| **KIRC** | 518 | 72 | 590 | 72 |

\*Paired samples correspond tumour/normal samples from the same patient.

### 2.2 Transcriptome age index

The transcriptome ages of tumour and normal samples were calculated using the Transcriptome Age Index (TAI) method (Domazet-Lošo & Tautz, 2010). The calculation was as follows:

*Equation 1*

where *psi* is the phylostratum of each gene and *ei* is the median-scaled RSEM gene expression value of the gene *i*. Under this scheme, a lower TAI would suggest an older transcriptome due to the overexpression of genes of lower phylostrata and/or the underexpression of modern genes.

Wilcoxon two-sided and one-sided tests were used to find statistical differences between the TAI of cancer and normal samples. The median TAIs were used as a summary statistic to define each population.

### 2.3 Identifying trends in the shifts of expression of phylostrata in tumours

The TAI provides an overall numeric value of the age of the cancer transcriptome, but it is not possible to distinguish which phylostrata drive the changes in the TAI. To pinpoint the differences in the total expression between phylostrata in tumour and normal cells, the proportion of the total library size that each phylostratum represented in each sample was calculated:

*Equation 2*

where *Ppsi*is the proportion of expression abundance corresponding to genes of phylostratum *i*, *eij* is the median-scaled RSEM expression value of gene *j* in the phylostratum *i*, *mi* is the number of genes in phylostratum *i* and *n* is the total number of phylostrata. The differences in mean expression proportions of each phylostratum in tumour and normal samples were determined by subtraction (tumour-normal). Two- sided Wilcoxon tests were performed to determine significant differences between the mean proportion of expression of a phylostratum in tumour and normal samples, and one-sided tests were used to determine the direction of the difference.

A similar analysis was performed using only the paired samples. Here, the ratio of phylostratum library sizes of cancer and normal samples was determined for each pair. Two-sided and one-sided Wilcoxon tests were used to determine if the ratios were different to 1, and if the ratios had a tendency to be over or below 1.

### 2.4 Gene-level interpretation of the change of the TAI in tumour using non-parametric tests

#### 2.4.1 Testing for the up- and downregulation of genes in phylostrata

Shifts in the TAI were also viewed from the gene-level perspective, using two complementary approaches. First, the R package limma (Smyth, 2005) and the voom function (Law, Chen, Shi, & Smyth, 2014) were used to calculate point estimates of log fold changes of genes in the tumour samples compared to their normal counterparts. Only genes with counts per million (cpm) greater than 1 in all samples were included in the analysis (see Supplementary methods). Two- and one-sided Wilcoxon tests were implemented for each phylostratum to test for enrichment, where the log fold changes of a specific phylostratum were deemed to be either greater (more positive) or less (more negative) than the rest.

#### 2.4.2 Association between the upregulated or downregulated status of a differentially expressed (DE) gene and its age

The second method consisted in testing for association between the number of up- and downregulated genes and the phylostrata. A gene was considered to be DE if its adjusted p-value was less than 0.05. If the logFC of a gene was greater than 0, then it was considered to be upregulated in tumours, and vice versa. Two- and one-sided Fisher Exact tests were implemented to test whether there was a larger number of upregulated or downregualted genes in each phylostratum.

### 2.5 Discerning the age of the overlapping differently expressed genes (DEG) in LUAD, BRCA, HNSC and KIRC

The DEG in common between the four cancer types were identified, and the overlap was visualized with a Venn Diagram, obtained with the VennDiagram R package. This subset of genes was considered to be a core set of genes associated with a cancerous state. The enrichment of genes of a certain phylostrata in this core set, in comparison with the DEG that were not part of this set, were detected by means of two- and one-sided Fisher tests.

## 3. Evidence of atavism in gene set analysis

### 3.1 Identifying the functional gene sets that change their expression in the phylostrata

We aimed to detect the gene functions that had the greatest change in expression in each phylostratum in the four cancer types. GO-terms were used to define functional groups.

#### 3.1.1 Generation of a subset of sixth-generation GO terms

A curated set of human Gene Ontology annotations were obtained following the method used to subset the GO term annotations in section 1.3. However, given that a higher degree of specificity was needed for the following analysis, sixth-generation GO term annotations were used instead of the third-generation annotations. Analogously, this second subset of nodes were separated by 6 edges from the highest parent node (GO: 0008150) in at least one path.

#### 3.1.2 Dating GO-terms to their phylostratum of emergence

GO-term annotations in human genes were dated using as reference the annotations of other species published by the Gene Ontology Consortium (The Gene Ontology Consortium, 2000), including the annotations offered by the Protein Data Bank, UniProt and the Comprehensive Microbial Resource. Model species and/or species that had been analysed by the OrthoMCL database were included (Table S *5*), giving a total of 54 species used as reference. These species were assigned to the phylogenetic tree constructed previously (Table II), and their GO term annotations were collapsed to sixth-generation GO terms. The phylostratum of emergence of a GO term was defined as the phylostratum of the most primitive species where the annotation appeared. Filamentous species or species that were multicellular at any stage of their life cycle were excluded from phylostrata 2 and 3, since these species might be annotated with GO-terms related to multicellularity.

**Table II Number of species with GO annotations assigned to each phylostratum**

|  |  |
| --- | --- |
| **Phylostratum** | **Number of species** |
| **1** | 24 |
| **2** | 3 |
| **3** | 2 |
| **4** | 1 |
| **5** | 2 |
| **6** | 3 |
| **7** | 1 |
| **8** | 2 |
| **9** | 1 |
| **10** | 1 |
| **11** | 1 |
| **12** | 4 |
| **13** | 3 |
| **14** | 1 |
| **15** | 1 |
| **16** | 1 |

#### 3.1.3 Curation of the sixth-generation GO term annotations

Before performing tests for enrichment of GO terms, it was necessary to distinguish that the functions of a gene could have either emerged together with the gene, or could have been acquired later in evolution. The former would represent the initial or original functions of the gene, and therefore these functions should have preceded or emerged simultaneously with the gene. In contrast, later acquired gene functions would succeed the time of emergence of the gene and would confound the results because they would not represent the initial capabilities of the gene.

The equivalency of the phylostrata used to date both human genes and GO terms allowed us to determine which gene functions were gained after the emergence of the gene founder. The GO-term annotations of a gene that originated after the emergence of the gene were excluded: if a gene was annotated with a GO term that emerged in a later phylostratum (e.g., 6) than the one the gene belonged to (e.g., 2), that GO annotation was disregarded.

#### 3.1.4 Identification of differentially expressed functional gene sets in the phylostrata

Functional sets were generated by grouping genes with at least one GO annotation in common. For each cancer type, functional gene sets up- or downregulated in each phylostratum were determined by two- and one-sided Wilcoxon test. A functional gene set was said to be significantly either up- or downregulated in a phylostratum if the ranks of the point estimated logFC of the genes were consistently greater or lower than the rest of the genes in the phylostratum. The cut-off for the nominal p-value was set at 0.01. No correction for multiple testing was performed given that the assumptions of independence of the sets were violated: there is large overlap between the genes contained in the functional sets, as well as redundancy in the GO terms due to the hierarchical and interconnected nature of the ontology. Instead, the agreement between significant functional gene sets between tumour types was explored.

In order to summarize the results, umbrella terms were used to group the sixth-generation GO terms into broad categories. The number of up- and downregulated GO terms sets within each umbrella group were counted and the results were represented via a heatmap.

### 3.2 Association between the ancientness of a gene function and its expression level

Wilcoxon tests were used to determine the up- and downregulated status of a functional gene set in each of the four cancer types, without taking into account the phylostrata. Subsequently, the phylostrata of the GO-terms were grouped into umbrella phylostrata: Unicellular (Phylostrata 1-3), Early Metazoan (Phylostrata 4-9) and Mammalian (Phylostrata 10-16) groups. Fisher Exact tests were used to determine whether the GO-terms that emerged at each umbrella phylostratum were over-represented in the up- or downregulated gene sets.

### 3.3 Change in expression of genes of functional sets across phylostrata

We aimed to identify gene sets whose older genes (unicellular phylostrata 1-3) were upregulated and newer genes (phylostrata 4-16) downregulated, as these gene sets were considered to be undergoing atavism. Only gene subsets with gene members that belonged to both of these groups were considered. Two- and one-sided Wilcoxon tests using the logFC of the genes were used to determine the up- and downregulated status of the older vs. newer genes using a p-value cutoff of 0.05. Only gene subsets that had significant p-values with the two-sided and either of the one-sided tests were considered significant. The sets that met the above mentioned criteria were singled out. This was repeated for the four cancer types, and sets that showed evidence of atavism in at least two cancer types were listed.

## 4. Programming languages

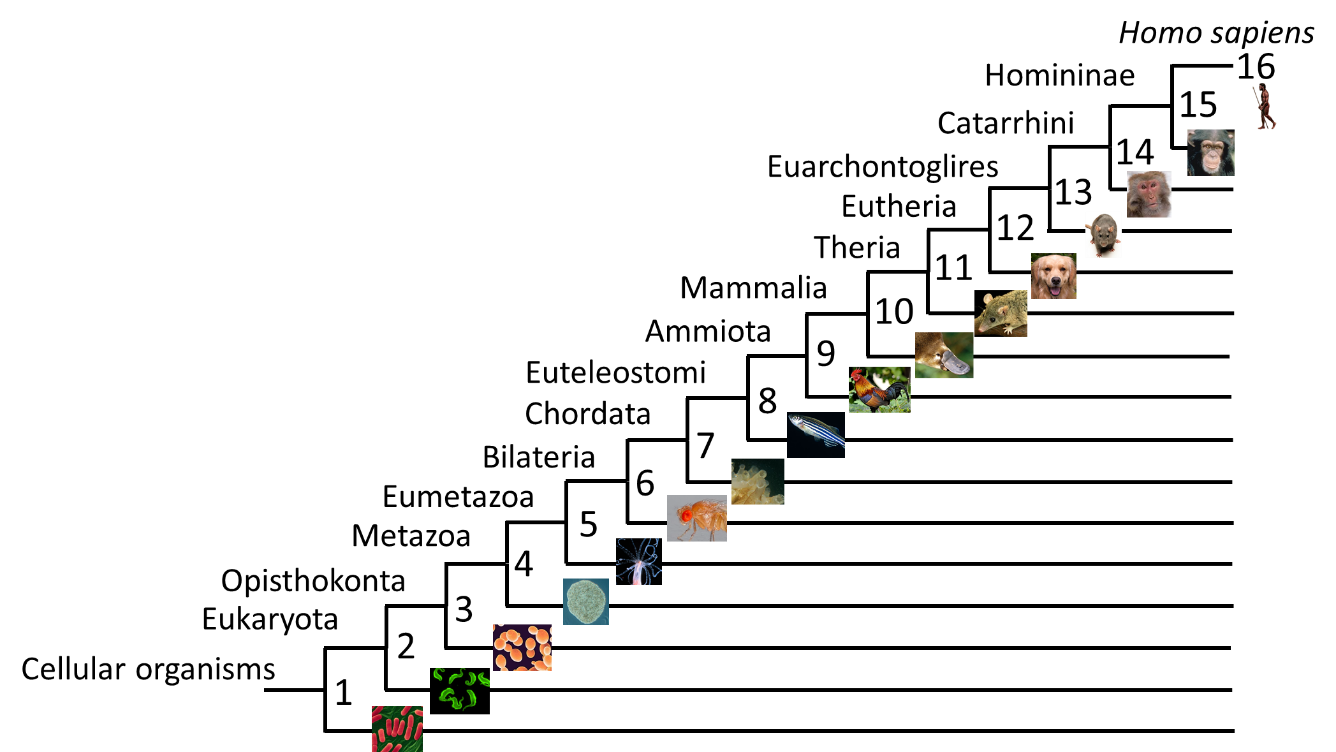
Statistical analyses were performed with R version 3.1.1 ([R Core Team, 2014](#_ENREF_17)). Parsing and scripting were done with Perl version 5.

# Results

## Phylostratigraphic analysis of human genes

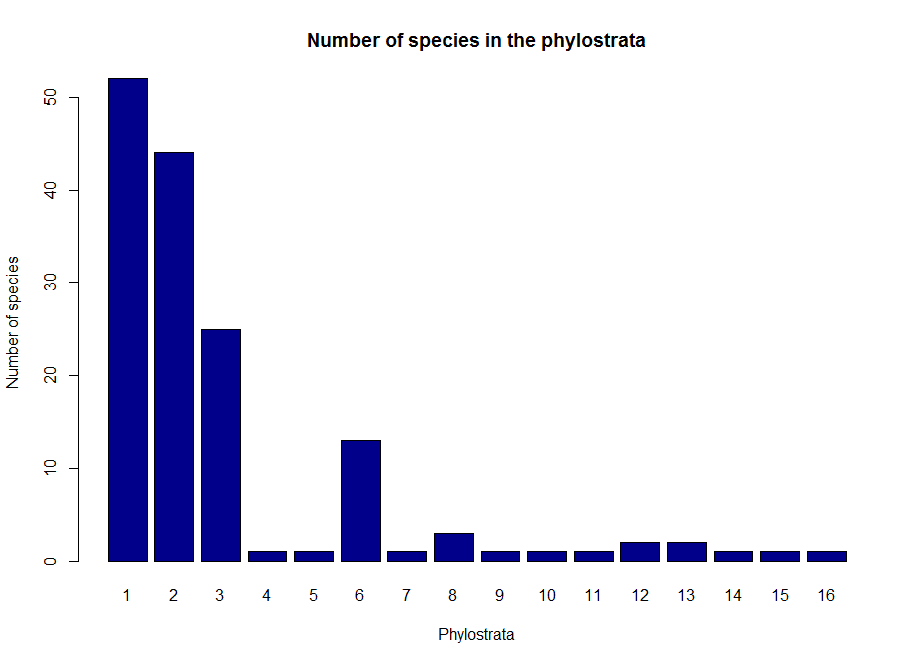
A phylogenetic tree was built with the species present in OrthoMCL as the first step for assigning a point of emergence to the human genes (***Figure 1***). The 16 phylostrata represent evolutionary milestones from the emergence of cellular life to humans, and therefore hold all possible points where the founder of a human gene could have emerged.

The tree shown in ***Figure 1*** can be viewed as a hierarchical bifurcating tree, where earlier phylostrata are the most inclusive, and higher phylostratum numbers correspond to species that are more closely related to humans. Of utmost importance is the establishment of multicellular life in animals, which can be pinpointed to phylostratum 4 (Metazoa). Phylostratum 1 and 2 represent unicellular organisms, whereas Phylostratum 3 includes both unicellular and primitive multicellular organisms, such as filamentous fungi.



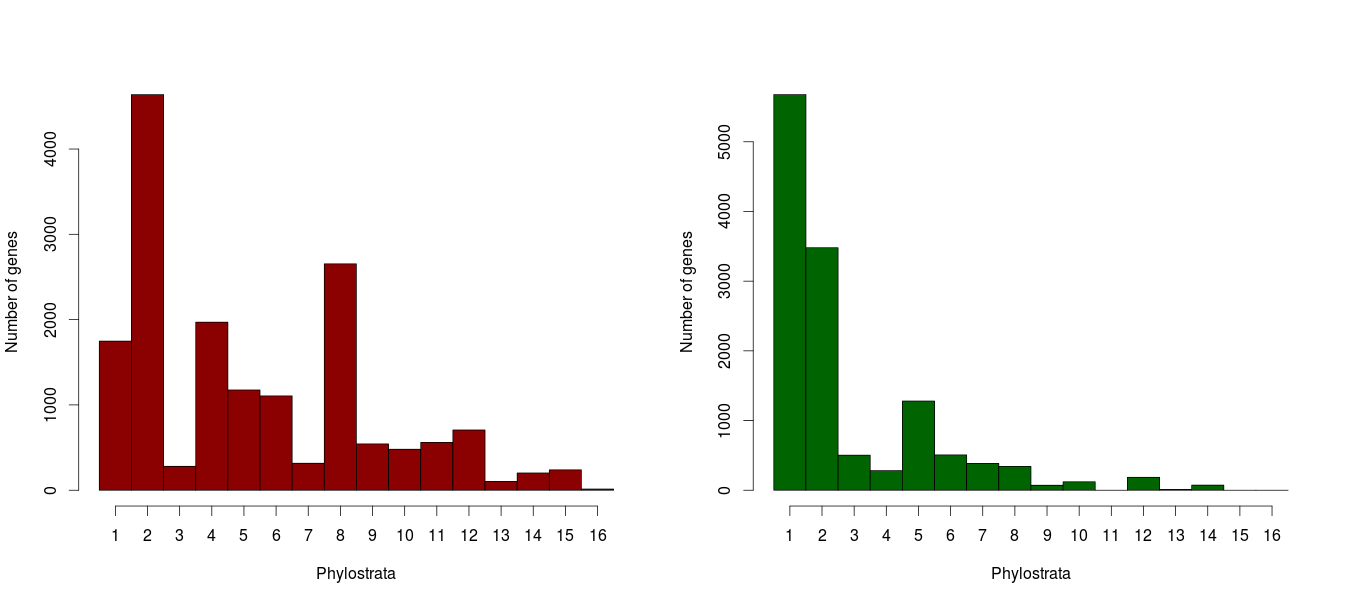
**Figure 1. Phylogenetic tree used for the classification of human genes into phylostrata.** The tree represents the phylogenetic classification of Homo sapiens, starting from the emergence of cellular life. Classification of the OrthoMCL species into taxons was carried out with the NCBI Taxonomy Common tree browser. The taxons with at least one species were included in the framework, and are defined as phylostrata. Species were assigned to the most recent phylostratum to which they belonged. The phylostratum rank its degree of ancientness, where more primitive phylostrata have smaller ranks.

Classification of the species analysed by OrthoMCL revealed that a large percentage of species belonged to early phylostrata (***Figure 2***). The most populated phylostrata correspond to those that include species with small genome sizes, or are interesting as human pathogens, model organisms, or commercially (phylostrata 1-3, 6).



**Figure 2. Histogram of the number of OrthoMCL species allocated to each phylostratum.** The number of species allocated varies across the phylostrata, where 80% of the species found in OrthoMCL are concentrated in phylostrata 1 to 3. There is also a spike in the number of species that belong to phylostratum 6.

The oldest phylostratum of the orthologs of a gene was used as its phylostratum of emergence. ***Figure 3a*** shows the distribution of the resulting human gene ages. Comparison of this distribution with the one obtained from the data of Domazet-Loso & Tautz (2008) (***Figure 3b***) reveals similar trends. There is a high number of genes assigned to phylostratum 2 in both distributions, and this number rapidly decreases in the following phylostrata. However, there were considerable peaks in the number of genes assigned to phylostrata 4 and 8 in the distribution obtained with the method used in this study that were not detected by the other authors.

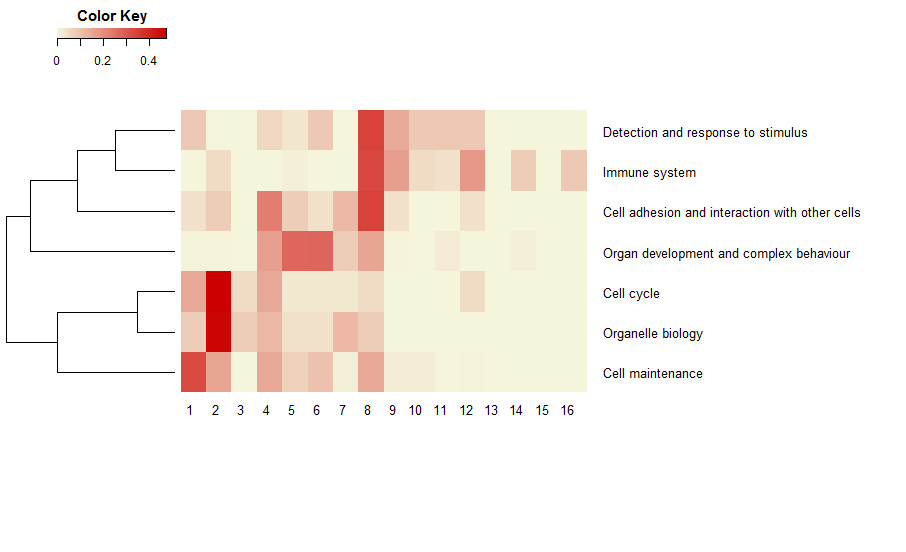


**b.**

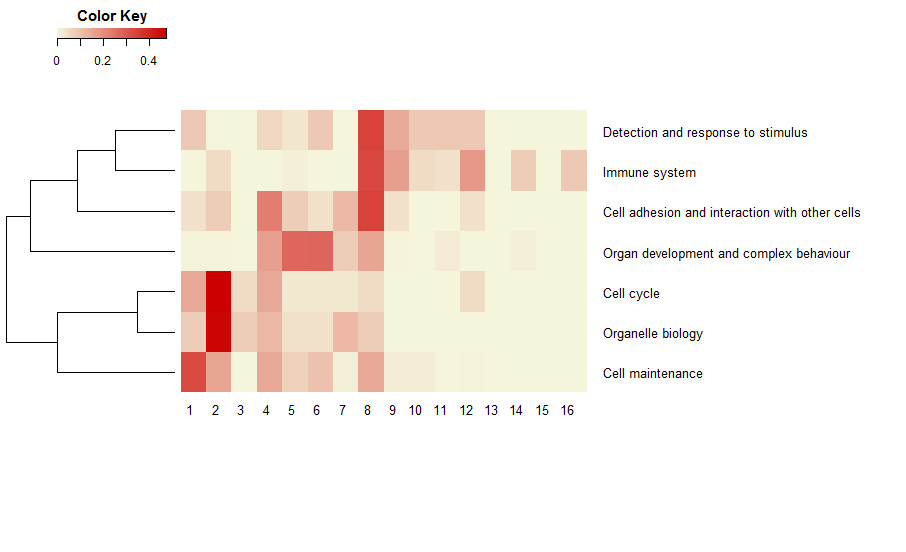
**a.**

**Figure 3. Distribution of the number of genes assigned to the phylostrata with two different phylogenetic frameworks.** (a) Distribution obtained with the phylogenetic tree obtained from the orthologous groups of OrthoMCL (**Figure 1**). A total of 16736 genes were mapped. (b) Distribution obtained by the classification carried out by another author (Domazet-Loso & Tautz, 2008). A total of 12916 genes were mapped. The distribution in (b) is highly skewed towards Phylostrata 1 and 2, assigning most of the human genes to these phylostrata. Distribution (a) is more evenly distributed, with spikes in the number of genes assigned to phylostratum 2, 4 and 8. A total of 6667 genes were assigned to non-metazoan phylostrata in (a), which represents 39.84% of the genes, whereas these numbers are 9661 and 74.80% in (b).

The GO terms enriched in the phylostrata were determined to ensure that the assignment of genes to phylostrata reflected the progressive acquisition of increasingly complex functions by genes. **Figure 4** shows that gene functions related to cell maintenance, organelle biology and the cell cycle are enriched in phylostrata 1 and 2, whereas mid-range phylostrata (4-8) are enriched for functions related to organ development and cell adhesion and interaction. Finally, the genes assigned to later phylostrata (8-16) are involved in the immune system and detection and response to stimulus. The hierarchical clustering dendogram shown in the figure further emphasizes these observations, as there are two main clusters that separate the basic and complex functions on the basis of their distribution across phylostrata.



Phylostrata



**Figure 4. Heatmap of the enriched functional categories in the phylostrata.** Enriched third-generation GO-terms found with BiNGO were classified into seven broad categories (umbrella terms). The intensity of the colours represents the proportion of the GO-terms of a category found in each phylostrata, where the colour intensities of each row sum to one. The dendogram represents the hierarchical clustering of the umbrella terms, where there are two main groups: terms involved in basal cell functions that are enriched in early phylostrata, and those related to more complex functions associated with multicellular life. The diagonal pattern followed by the proportion of enriched terms suggests a tendency for enriched terms to be related to increasingly complex and specific functions as the phylostrata increases.

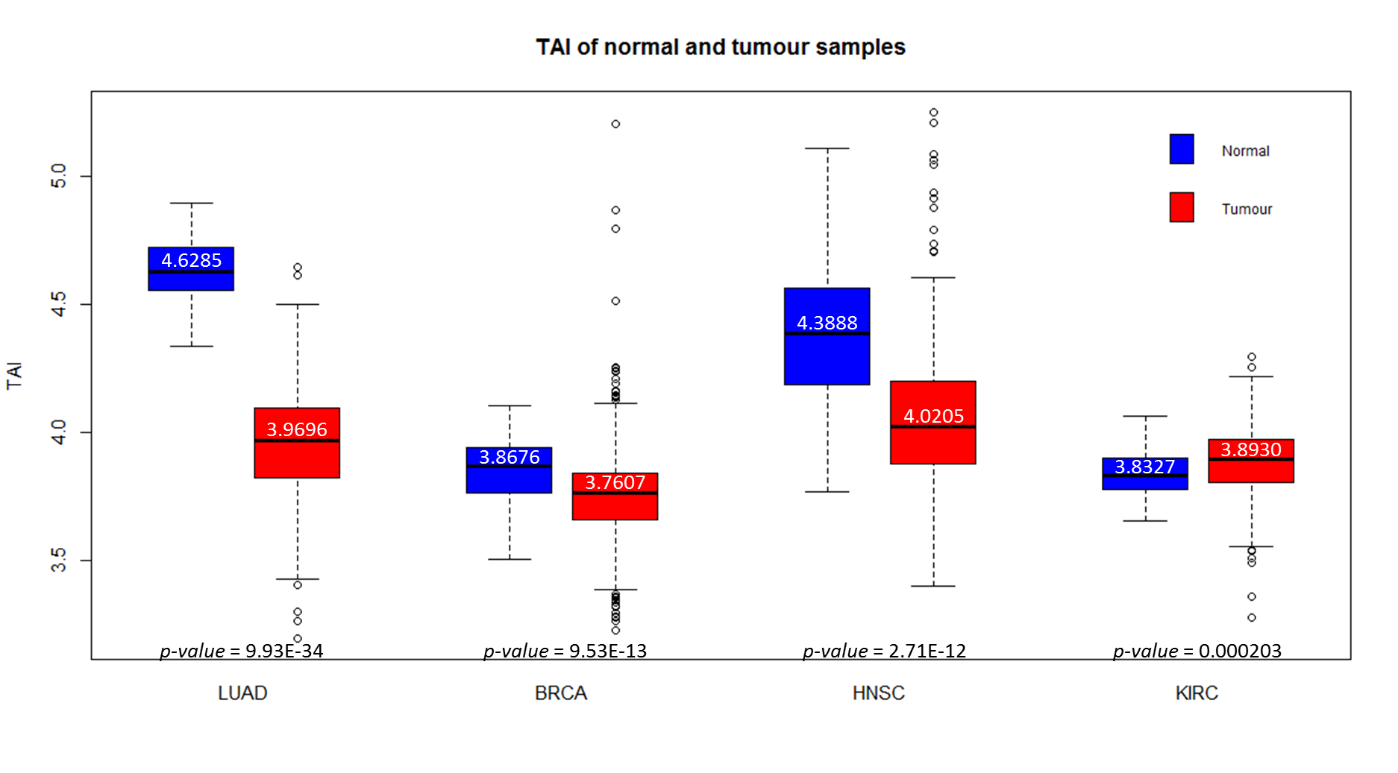
## The age of the cancer transcriptome

Calculation of the TAI of normal and tumour samples revealed a significantly lower TAI in the tumour samples in LUAD, BRCA and HNSC than their normal counterparts, whereas the opposite trend was seen for the KIRC samples (***Figure 5***). This suggests that for the first three tissue types, the age of the transcriptome of tumour cells is older than the age of the transcriptome of normal cells.

The degree of difference between the ages of the tumour and normal transcriptomes varies across tissue types and can be visualized by the vertical ranges of the tumour and normal boxplots. There are marked differences between the TAI of normal and tumour LUAD samples, where the boxplot boxes did not overlap, whereas the ranges largely overlapped in the BRCA, HNSC and KIRC samples.

It is also important to point out that the spread of TAI across samples varies between tissue types. Even when the TAI of normal samples is higher than in tumour samples in BRCA, the TAI of normal BRCA samples is lower than the TAI of tumour LUAD samples. A similar conclusion can be drawn when comparing BRCA and HNSC normal and tumour samples. This observation suggests that the age of the transcriptome depends not only on the normal or tumour state of the tissue, but also on the tissue type of origin, making comparison of TAI between tissue types not straightforward.

Finally, it is important to point out the limited spread of TAI values for the normal samples in LUAD, BRCA and KIRC, in contrast to the large spread of the TAI in tumour samples. This trend was not observed in normal HNSC samples.



**Figure 5. Transcriptome Age Index (TAI) of normal and tumour LUAD, BRCA, HNSC and KIRC samples.** In the case of LUAD, BRCA and HNSC, the TAI of tumour samples is lower than that of normal samples of the same tissue type, suggesting an older transcriptome in the tumour samples. The opposite trend was observed for KIRC samples. Boxplots reflect the spread of TAI across samples.The spread and range of the TAI of normal and tumour samples varies greatly between tissue types. Normal and tumour samples were median-scaled normalized simultaneously for each cancer type before calculation of the TAI. The white numbers are the median TAI. Two-sided p-values obtained from two-sided Wilcoxon tests (shown) were used to determine if the tumour samples had a different mean TAI than normal samples. Complementary one-sided tests were used to detect the direction of the difference (data not shown).

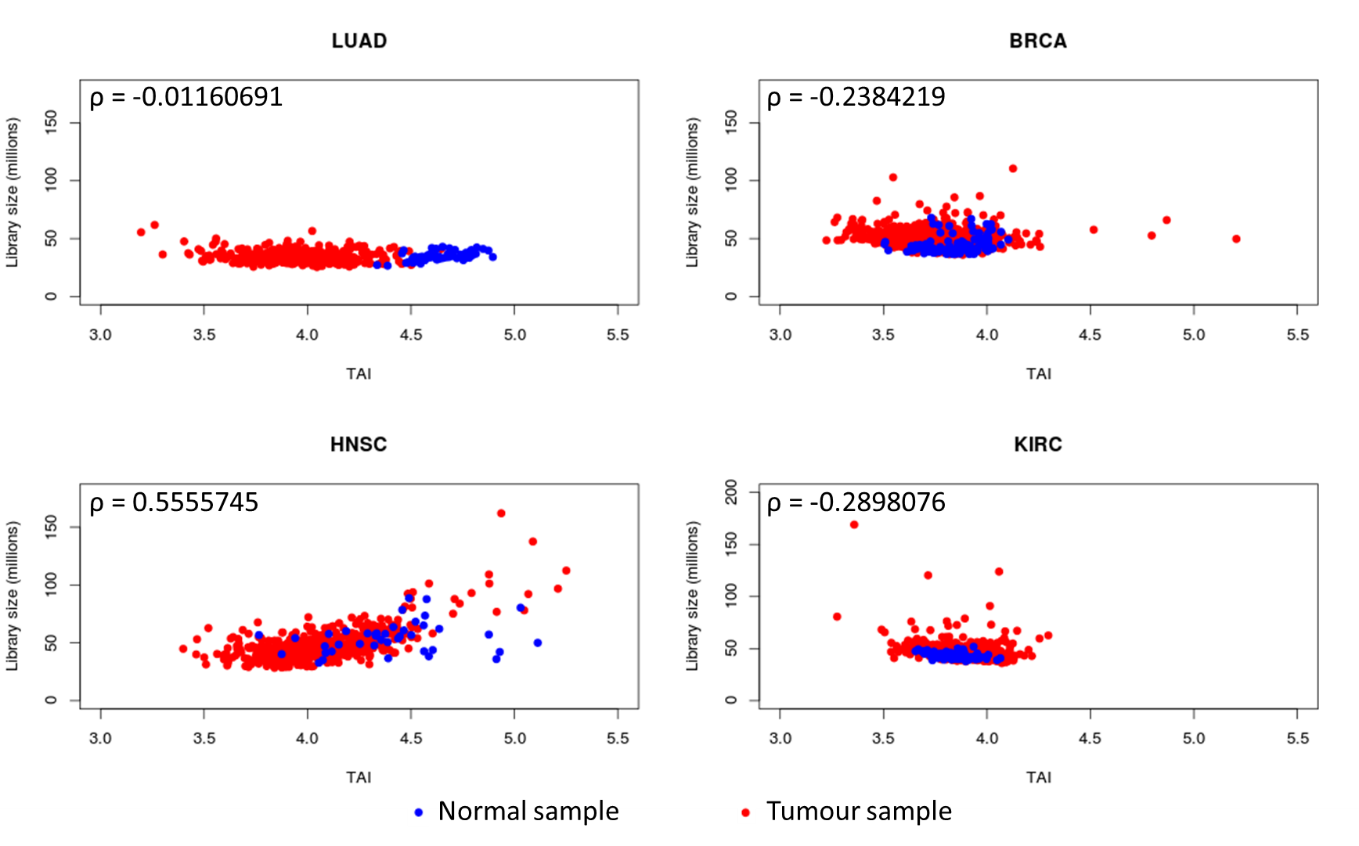
### 2.1 Possible confounders of the TAI

We identified four possible vulnerabilities of the TAI technique: (1) a correlation between the sample library size (the sum of the expression of all genes in a sample) and the normalization method of the expression data, (2) the specific phylostratigraphic technique used, (3) the sample purities and (4) systematic differences between the patients from which the normal and tumour samples were taken that would lead to artificial differences between the TAI of tumour and normal samples.

##### 2.1.1 Correlation between sample library size and TAI

We suspected that for some normalization techniques, some genes would be down-weighted or up-weighted differently in different samples, and that the shrinkage factor of highly expressed genes could have been less or more severe than for lowly expressed genes. The TAI and the proportion of abundance of each phylostratum is expected to be robust to the sample library sizes, given that their calculation involves normalizing (dividing) for library size. However, to rule out the sample library size as a confounder of the TAI, the Spearman correlation between the TAI of normal and tumour samples, and the library size of the median-scaled samples was calculated (***Figure 6***). The TAI of LUAD samples had the least correlation with the library sizes, whereas the TAI of BRCA and KIRC samples had a medium-level negative correlation, and the HNSC had a positive correlation.

It is important to mention that the median-scaling of the raw RSEM samples involves the multiplication of the expression values by a constant factor, but this constant term is nullified by the method used to calculate the TAI, which involves normalization by the library size. Thus, the normalization method only scales the library sizes.



**Figure 6. Correlation between library size and TAI of LUAD, BRCA, HNSC and KIRC normal and tumour samples.** The correlation obtained for LUAD samples was close to zero, whereas BRCA and KIRC show a negative correlation and HNSC samples show a positive correlation. Normal samples of LUAD, BRCA and KIRC appear agglomerated in specific areas of the plots due to similar library sizes and TAI, suggesting homogeneous samples. This does not occur in HNSC normal samples. The dispersed cancer samples suggest sample heterogeneity. The library sizes of the median-scaled samples were calculated as the sum of expression of the 16,736 genes in a sample. Correlation values were obtained using Spearman’s rank correlation coefficient.

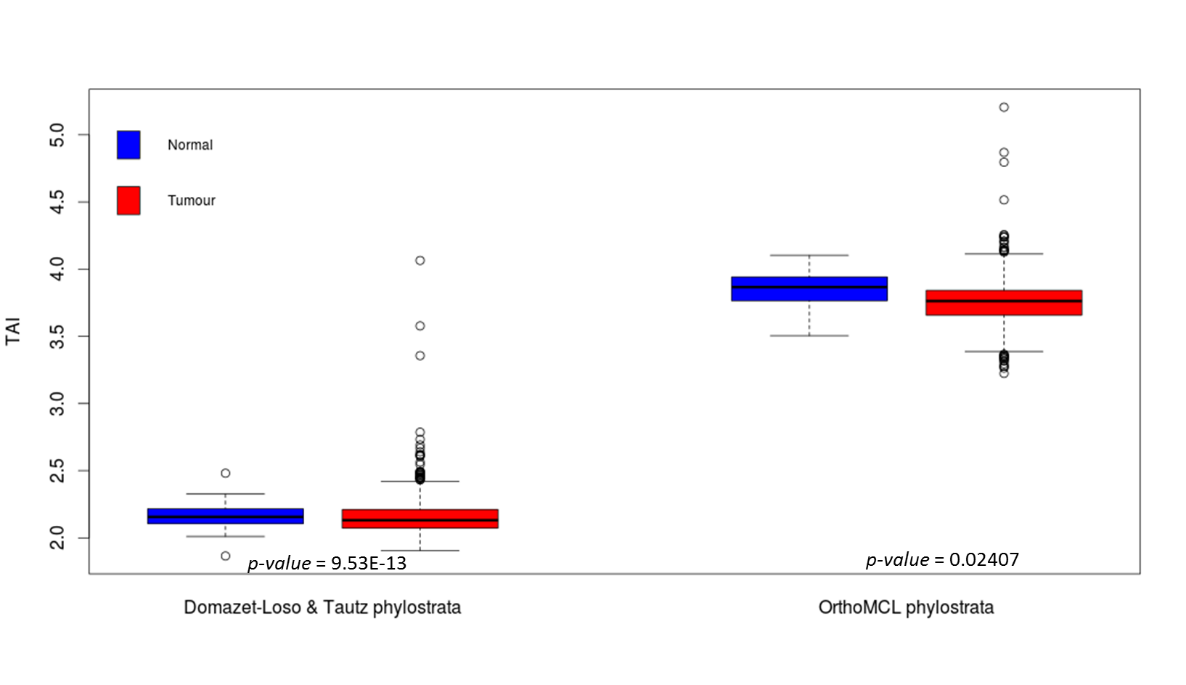
Similar analyses were made after calculation of the TAI with library-size-scaled expression samples, which generated the same correlation values as the one presented here, and samples normalized by TCGA, which showed a consistent positive correlation (0.01-0.68). However, we observed similar differences between the TAI of tumour and normal samples calculated with these expression levels to the ones discussed here (*data not shown)*. The TAI calculated from the median-scaled library sizes were chosen to be presented here and used hereafter given that it provided the least consistent correlation between library size and TAI across cancer types.

The correlation of the TAI of the BRCA, HNSC and KIRC samples with the library sizes has numerous explanations. Interpretation of the correlation between library size and TAI as a confounder of the TAI implies that the TAI and the transcriptome sizes are random variables independent of each other. However, under the assumption that the atavism hypothesis, tumour samples would be constrained to have a lower TAI than normal samples, so these samples are expected to be concentrated towards the left side of the plots. Additionally, considering that cancer is a disease of the dysregulation of expression levels of genes, the total transcriptome size of tumour samples would be expected to be spread over a different range than the normal samples. This is evident in ***Figure 6***, where the normal samples of LUAD, BRCA and KIRC have a very small vertical spread compared to the tumour samples. With these two implicit constraints in place, a correlation between library size and TAI is not entirely surprising: the negative correlation obtained for BRCA and KIRC implies that the library sizes of tumour samples tend to be larger, and the positive correlation of the HNSC means that the transcriptome of tumour samples is smaller. However, we cannot entirely rule out that the library sizes are playing a role in the differences in the TAI of tumours and normal samples.

##### 2.1.2 Comparison of the TAI under different phylostratigraphic schemes

The strong reliance of the TAI on the dating of genes makes the method very vulnerable to the phylostratigraphic method used and its possible biases. The general trends observed in the TAI should be independent of the method used to assign the gene ages. We calculated the TAI of median-normalized BRCA normal and tumour samples with the phylostratigraphic results obtained by another author (Domazet-Loso & Tautz, 2008). The 19 phylostrata obtained by this author were collapsed into the framework of 16 phylostrata used in this study.

The TAI of normal and tumour BRCA samples obtained with both phylostratigraphy methods is shown in ***Figure 7*** for comparison. Similar to our results, the TAI of tumour samples was also found to be significantly lower with this alternative phylostratigraphy approach, further validating our results. However, the initial TAI obtained for the BRCA samples ranged between 3.0 to 5.5, but the TAI using the approach of Domazet-Loso & Tautz (2008) ranged between 1.5 and 3.5. These results are expected considering the bias towards phylostrata 1 and 2 in the gene ages determined by these authors (***Figure 3***).

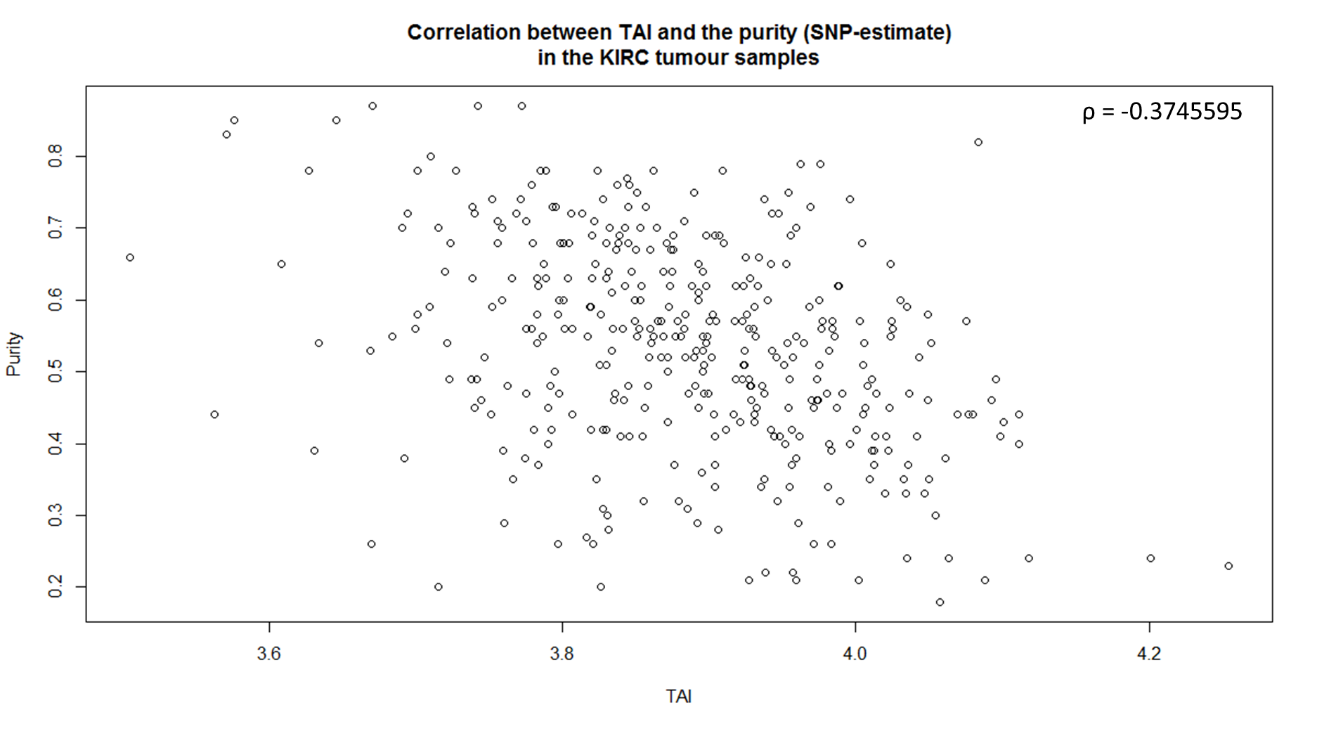


**Figure 7. Comparison of the TAI of BRCA tumour and normal samples after different phylostratigraphy techniques.** The TAI of BRCA median-scaled samples was calculated with the phylostrata of genes obtained by Domazet-Loso & Tautz (2008). The TAI obtained with the phylostratigraphic technique used in this study is shown for comparison. In both cases the tumour samples had a significantly lower TAI than their normal counterparts. However, the scale and range of the TAI values is non-overlapping and non-comparable. P-values shown are two-sided p-values obtained from two-sided Wilcoxon tests to determine if the tumour samples had a different mean TAI than normal samples. Complementary one-sided tests were used to determine the direction of the difference (data not shown).

##### 2.1.3 Correlation between the sample purity and TAI

A corollary of the atavism hypothesis is the presence of a high correlation between the TAI and sample purity. As the contamination of tumour samples with normal tissue types increases, the TAI should progressively resemble that of the normal tissue, which translates as an increase in the TAI. We determined the Spearman correlation between the TAI and the sample purities of KIRC tumour samples, as measured by somatic copy alterations by other authors (The Cancer Genome Atlas Research Network, 2013) (***Figure 8***). The strong negative correlation of -0.37 suggest that the TAI results are strongly dependent on the purity of the samples.

It is also worth mentioning that other purity metrics were also explored, such as the percent of tumour cells and nuclei provided directly by TCGA. We found that these estimates overestimated the sample purites when we compared them to those determined by somatic copy alterations. Therefore, it was not surprising that we found no correlation between the TAI and the percentage of tumour cells (*data not shown)*.



**Figure 8. Correlation between the TAI and purity of KIRC tumour samples**. There is a strong negative correlation between tumour sample purity and TAI, indicating that a lower TAI is associated with a higher number of tumour cells in the sample. Sample purity estimates correspond to those obtained with somatic copy alteration analyses (The Cancer Genome Atlas Research Network, 2013). The correlation value was obtained with Spearman’s rank correlation coefficient.

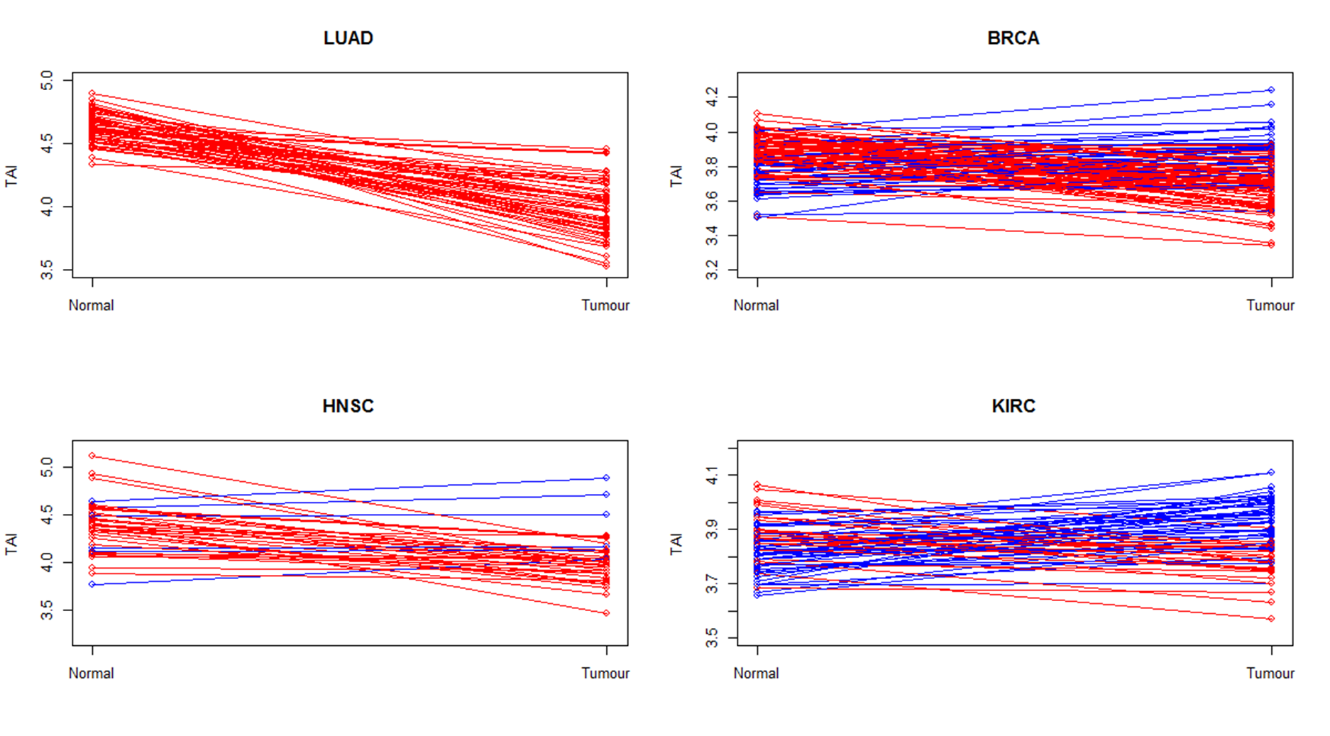
#### 2.1.3 Inter-patient differences

We wanted to rule out that the lower TAI of tumour samples was due to systematic differences between the patients from which the normal and tumour samples were taken. Namely, there were a large number of tumour samples without normal controls from the same patient that could lead to biases in the tumour samples not accounted for in the normal samples.

We compared the TAI of normal and tumour samples taken from the same patient to see if lower TAI in the tumour samples were detected (***Figure 9***). In the case of the LUAD patients, 100% of the tumour samples had a lower TAI than their normal counterparts from the same patient. A similar trend was observed in BRCA and HNSC, although this percentage decreased to 68.8% and 85%. Finally, in less than half of KIRC samples, the lower TAI belonged to the tumour samples (37.5%).

The results obtained in the paired analysis of the TAI reflects the results obtained in ***Figure 5***. The large difference in the TAI of normal and tumour LUAD samples is corroborated by a lower TAI in all tumour paired samples. In the case of the BRCA, HNSC and KIRC samples, the differences between tumour and normal TAI were found to be less evident, so some paired samples with a lower TAI in normal samples were expected.

In summary, inter-patient variation does not seem to be affecting the TAI results obtained with the entire dataset.



**Figure 9. TAI of paired normal-tumour samples in LUAD, BRCA, HNSC and KIRC.** The TAI of normal-tumour samples of the same patient are plotted. Red lines were drawn between paired samples that have a lower TAI in tumour samples than in normal. Blue lines were plotted in the opposite cases. In LUAD, BRCA and HNSC paired samples, the TAI of the tumour tissue was mostly lower than in the normal tissue (100%, 68.8% and 85%, respectively). The opposite trend was evident for KIRC patients (37.5%).

### 2.2 Interpreting the shift of the TAI

We sought to break down the distribution of expression levels across phylostrata to pinpoint the phylostrata that underwent the most significant changes in expression in tumour tissues. We took a systems-level approach to detect gross changes in the overall expression level of the phylostrata, as well as gene-level analysis to test for up- and downregulation of individual genes in the phylostrata.

#### 2.2.1 Systems-level changes in the overall expression level of the phylostrata

Comparison of the expression levels of phylostrata in tumour and normal samples was done using two complementary analyses, which exploited both the large number of samples available as well as the power offered by paired analyses.

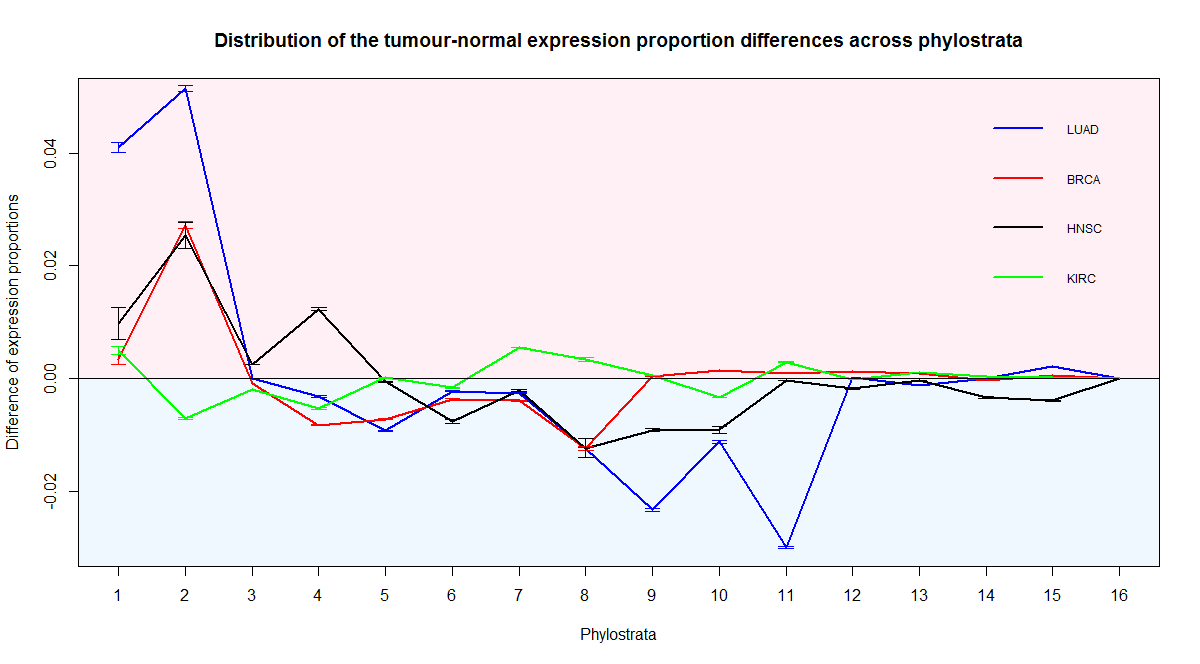
For the first approach, the difference in the mean proportion of gene expression between tumour and normal samples was determined for each phylostratum (**Figure 10** & ***Table III***). The overall expression level of phylostrata 1 and 2 was greater in tumour than in normal for LUAD, BRCA and HNSC. Interestingly, this increase in expression is more marked in phylostrata 2 than in phylostrata 1, but are significant in LUAD and HNSC samples in both cases. The change is only significant in phylostratum 2 of BRCA samples.

The expression of phylostrata 3 did not show a significant change in tumours, which was probably algebraically due to the low number of human genes assigned to this phylostratum. However, it can be also interpreted in the LUAD and BRCA samples as the intersection between the positive difference in expression in phylostrata 1 and 2, and the negative difference seen in later phylostrata.

Contrary to the results obtained for phylostrata 1 and 2, the expression levels of phylostrata 4 to 8 are significantly lower in tumour LUAD and BRCA samples, and this observation can be extended to phylostratum 11 in LUAD samples. HNSC also showed a similar pattern, but from phylostrata 5 to 10. However, the signal is not consistently significant across all these phylostrata, with non-significant differences in phylostrata 5 and 7. The small number of human genes assigned to phylostrata 13 onwards (***Figure 3***) is probably responsible for the limited change in expression of these phylostrata, and the significant differences obtained are interpreted as stochastic noise.

KIRC samples were not found to follow the patterns of changes in expression observed in LUAD, BRCA and HNSC samples. The differences in proportions in KIRC samples are small, and fluctuate from a positive to a negative difference without an evident trend.

In summary, we identified an increase in the expression of genes of early phylostrata (1 and 2) in tumours, and a decrease in expression of genes of later phylostrata (4 to 8-11). The rest of the phylostrata showed a minimal difference in the magnitude of the proportions. This trend was evident in LUAD, BRCA and HNSC samples, but not in KIRC samples.



**Figure 10. Distribution of the tumour-normal expression proportion differences across phylostrata.** The proportion differences were calculated by subtracting the mean proportion of phylostrata expressions of tumour and normal samples. A positive difference in a phylostrata indicates that the proportion of expression was greater in tumour samples than in normal samples. The proportion of the transcriptome corresponding to genes that emerged in phylostrata 1 and 2 was greater in tumour samples than in normal samples for LUAD, BRCA and HNSC. This was only true for the genes of phylostratum 1 in KIRC. For all cancer types, the genes of Phylostratum 3 did not show significant differences between tumour and normal samples. On the other hand, in LUAD, BRCA and HNSC samples, the expression of genes from phylostrata 4 (LUAD and BRCA) or 5 (HNSC) to phylostrata 8 (BRCA), 10 (HNSC) or 11 (in LUAD) was decreased in tumours. This trend was not seen in KIRC samples. The overall expression of genes of phylostratum 12 onwards changed little in magnitude in all cancer types. Error bars were calculated as mean ± total variance, where the total variance is the sum of the variances of the expression proportions of the normal and tumour samples. Considering that the entire set of samples was used for each cancer type, which included at least 450 samples, it was not surprising that the variance of the difference of the means was consistently small.

**Table III. Wilcoxon two-sided p-values for the difference of the means of the tumour and normal expression proportions of the four cancer types.** Wilcoxon one-sided p-values were used to corroborate the direction of the change. P-values were considered to be significant if p < 0.05.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phylostrata** | **LUAD** | **BRCA** | **HNSC** | **KIRC** |
| **1** | 6.20E-28 | 0.132535 | 0.001636 | 0.022986 |
| **2** | 7.57E-33 | 7.40E-37 | 4.25E-07 | 8.35E-08 |
| **3** | 0.851304 | 8.90E-09 | 2.03E-11 | 1.34E-20 |
| **4** | 0.0018 | 1.25E-21 | 4.54E-09 | 1.64E-06 |
| **5** | 1.67E-15 | 4.37E-21 | 0.697321 | 0.837189 |
| **6** | 0.00011 | 6.16E-24 | 6.41E-06 | 6.29E-05 |
| **7** | 9.39E-14 | 4.32E-29 | 0.555964 | 1.03E-21 |
| **8** | 6.17E-10 | 8.81E-19 | 0.00312 | 0.008705 |
| **9** | 6.72E-26 | 0.834541 | 3.39E-08 | 0.005463 |
| **10** | 1.21E-14 | 0.000455 | 2.54E-06 | 1.07E-24 |
| **11** | 2.65E-32 | 0.024854 | 0.446783 | 5.83E-12 |
| **12** | 0.021225 | 4.90E-21 | 0.072238 | 0.000106 |
| **13** | 1.61E-21 | 0.44481 | 0.231969 | 3.91E-28 |
| **14** | 0.707727 | 3.88E-12 | 0.297169 | 5.85E-16 |
| **15** | 4.61E-17 | 2.65E-36 | 1.98E-09 | 1.97E-07 |
| **16** | 5.87E-15 | 0.001411 | 2.08E-07 | 2.53E-19 |

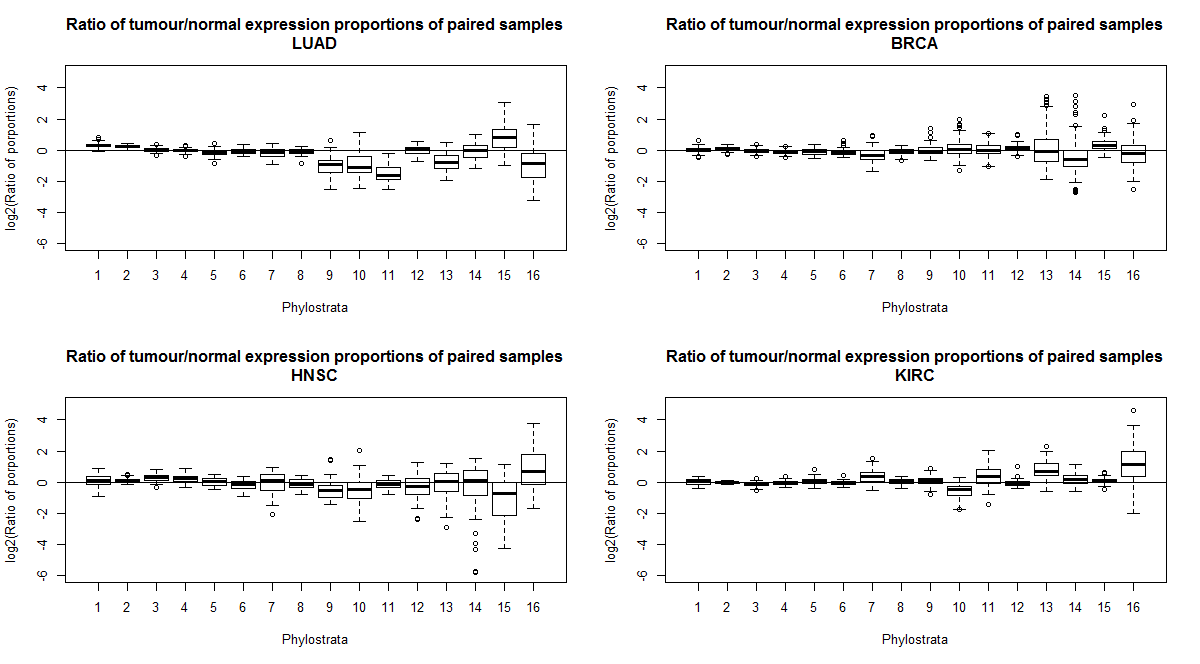
\*Pink =higher expression proportion in tumour samples

\*Blue=lower expression proportion in tumour samples

\*White=no significant difference between normal and tumour samples

Next, we wanted to ensure that the patterns of up- and downregulated phylostrata observed in the previous sections could be reproducible using paired samples. The tumour/normal ratio of the expression proportions of paired samples was calculated for each phylostratum (**Figure 11** & **Table *IV***).

The results obtained were consistent with those of the previous section, but the power to detect significant differences was greater, most likely due to the use of paired sample data, which removes the noise of generated by between-patient differences. The expression proportion of Phylostrata 1 and 2 was consistently higher in LUAD, BRCA and HNSC tumours, whereas post-Opisthokonta (LUAD and BRCA) or post-Metazoan (HNSC) phylostrata were lower. Similar to the previous observations of KIRC tumour samples, there is not a consistent pattern in the increase or decrease of expression of the phylostrata.



**Figure 11. Boxplots of the ratio of tumour/normal expression proportions of paired samples.** The log2 ratio of expression proportions of paired samples was used as a measure of the overall change of expression of the phylostrata. Boxplots represent the spread of log2 ratios across pairs of tumour-normal samples. For each sample pair, a log2 ratio of greater than 0 in a phylostratum indicates an increase in the total expression of the genes of that phylostratum in the tumour sample. The trends observed in Figure 10 are reproduced here. In LUAD, BRCA and HNSC samples, the overall expression of genes of tumour samples that emerged in phylostrata 1 and 2 was increased, whereas for the expression of later phylostrata was decreased.

**Table IV. Two-sided p-values of the ratio of the tumour/normal expression proportions.** Wilcoxon two-sided tests were used to determine if the ratio of expression proportions was significantly different to 1. One-sided Wilcoxon tests confirmed the direction of the change. P-values < 0.05 were considered significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phylostrata** | **LUAD** | **BRCA** | **HNSC** | **KIRC** |
| **1** | 1.50E-10 | 0.040093 | 0.043626 | 0.022873 |
| **2** | 1.20E-10 | 2.44E-13 | 7.71E-05 | 4.14E-05 |
| **3** | 0.15074 | 0.000854 | 1.81E-08 | 4.91E-09 |
| **4** | 0.051413 | 1.57E-09 | 9.58E-07 | 0.000498 |
| **5** | 1.06E-06 | 9.55E-07 | 0.404938 | 0.041928 |
| **6** | 0.017534 | 2.42E-12 | 0.001863 | 0.004558 |
| **7** | 0.000484 | 1.30E-10 | 0.412534 | 2.50E-10 |
| **8** | 0.002741 | 1.18E-07 | 0.170088 | 0.065265 |
| **9** | 7.16E-10 | 0.545412 | 5.45E-05 | 0.011469 |
| **10** | 1.07E-05 | 0.007874 | 0.000357 | 3.65E-12 |
| **11** | 1.14E-10 | 0.426547 | 0.079329 | 2.62E-07 |
| **12** | 0.337386 | 6.19E-10 | 0.270692 | 0.000225 |
| **13** | 1.35E-09 | 0.291373 | 0.589976 | 1.34E-12 |
| **14** | 0.635936 | 1.72E-05 | 0.527213 | 8.26E-05 |
| **15** | 2.71E-08 | 2.93E-15 | 0.001367 | 0.000519 |
| **16** | 0.00011 | 0.110427 | 0.000281 | 1.50E-10 |

\*Pink =ratio is greater than 1

\*Blue=ratio is less than 1

\*White=ratio is not significantly different to 1

#### 2.2.2 Gene-level changes of expression in the phylostrata

A limitation of the previous analyses is their vulnerability to genes that have a great change in expression levels between tumours and normal tissues which will dominate the signal of the phylostrata. We employed non-parametric gene-level tests to corroborate the association of the expression fold changes of genes and their ancientness.

Wilcoxon tests were used to rank all genes by their fold change, and determine whether the genes of a certain phylostrata ranked significantly higher (were more upregulated) or lower than the rest of the genes. This method can be interpreted as test for enrichment of phylostrata with more positive or more negative fold changes, where the ranking metric to test for enrichment is the fold change. The use of Wilcoxon tests as a means of testing for enrichment of sets has been previously used by other authors, but using the t-statistic as the ranking metric (Michaud et al., 2008).

The results for LUAD and BRCA samples broadly support the predictions made by the atavism hypothesis (***Table V***), as phylostrata 1 and 2 are enriched for genes with positive fold-changes, whereas post-metazoan phylostrata are enriched for genes with less positive or negative log fold-changes.

Contrary to the strong patterns obtained for LUAD and BRCA samples, only phylostratum 2 was enriched with upregulated genes and phylostratum 6 was enriched with genes with lower fold changes in HNSC. The KIRC samples did not follow the patterns observed in LUAD and BRCA, and suggested by HNSC samples.

**Table V. Two-sided p-values for positive and negative enrichment of phylostrata.** The tumour/normal log fold change estimates of the genes and two- and one-sided Wilcoxon tests were used as a comparative metric to find if genes of a certain phylostrata were more up- or downregulated than the rest of the phylostrata. The genes of pre-metazoan phylostrata in LUAD and BRCA samples tended to be more upregulated in tumours than the rest of the genes. The opposite occurs for post-metazoan phylostrata. This pattern was modestly observed in HNSC, but did not appear to be followed by the KIRC samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phylostrata** | **LUAD** | **BRCA** | **HNSC** | **KIRC** |
| **1** | 3.55E-36 | 3.51E-11 | 1.52E-06 | 5.88E-09 |
| **2** | 1.27E-67 | 2.58E-47 | 3.39E-06 | 0.147956 |
| **3** | 0.470427287 | 0.415629 | 0.428323 | 0.729988 |
| **4** | 3.40E-06 | 2.50E-07 | 0.098182 | 0.072978 |
| **5** | 4.88E-08 | 7.75E-07 | 0.475209 | 0.929242 |
| **6** | 1.45E-10 | 6.29E-09 | 0.017005 | 0.083791 |
| **7** | 2.55E-14 | 3.29E-15 | 0.41762 | 0.786649 |
| **8** | 2.13E-43 | 2.39E-21 | 0.504799 | 0.00907 |
| **9** | 8.13E-10 | 0.003432 | 0.487501 | 0.009491 |
| **10** | 0.005409531 | 0.735881 | 0.638057 | 0.004454 |
| **11** | 0.000894581 | 0.384727 | 0.769199 | 8.07E-05 |
| **12** | 1.42E-05 | 0.017948 | 0.521307 | 0.16122 |
| **13** | 0.024793793 | 0.325681 | 0.952126 | 0.774208 |
| **14** | 0.252656027 | 0.102176 | 0.787749 | 0.100591 |
| **15** | 0.757143268 | 0.875994 | 0.113983 | 0.26576 |
| **16** | 0.004423579 | 0.442146 | NA | NA |

\*Pink = the log fold changes of the genes of the phylostratum are greater (more positive) than the rest of the genes.

\*Blue = the log fold changes of the genes of the phylostratum are lesser (more negative) than the rest of the genes.

\*White = the log fold changes of the genes of the phylostratum are not significantly different to the rest of the genes.

The previous method was taken a step forward by counting the number of upregulated and downregulated differentially expressed genes in each phylostratum, followed by Fisher tests to determine if the number of genes in each category was higher or lower than expected in each phylostratum (***Table VI***).

The results obtained with this method mimic the results previously obtained. The number of upregulated genes is higher than expected in phylostrata 1 and 2 of LUAD and BRCA samples, whereas the number of downregulated genes is higher than expected in later phylostrata. The HNSC results resemble those obtained with LUAD and BRCA samples, but to a lesser degree. Finally, the KIRC samples do not seem to follow the pattern observed in the other cancer types.

**Table VI. Number of upregulated and downregulated DEG, and the p-values of overrepresentation.** Fisher Exact tests were used to determine if the number of up or downregulated genes was greater than expected, considering the number of DEG in a phylostrata and the total number of DEG. The number of upregulated genes in phylostrata 1 and 2 is greater than expected, but less than expected in later phylostrata. HNSC is only in agreement with the LUAD and BRCA samples in phylostrata 2 and 6. KIRC samples follow the opposite trend.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **LUAD** | | | **BRCA** | | | **HNSC** | | | **KIRC** | | |
| **Phylostrata** | **Number upregulated** | **Number downregulated** | **Two-sided p-value** | **Number upregulated** | **Number downregulated** | **Two-sided p-value** | **Number upregulated** | **Number downregulated** | **Two-sided p-value** | **Number upregulated** | **Number downregulated** | **Two-sided p-value** |
| **1** | 730 | 343 | 1.19E-32 | 619 | 510 | 2.97E-09 | 352 | 508 | 3.33E-05 | 482 | 685 | 7.21E-07 |
| **2** | 1821 | 1026 | 5.96E-64 | 1703 | 1373 | 4.26E-34 | 1083 | 1033 | 1.75E-05 | 1474 | 1569 | 0.587253 |
| **3** | 81 | 74 | 0.808344 | 77 | 89 | 1 | 54 | 71 | 0.365546 | 84 | 89 | 0.93876 |
| **4** | 486 | 619 | 2.14E-07 | 482 | 716 | 2.01E-06 | 349 | 433 | 0.090209 | 533 | 639 | 0.063275 |
| **5** | 230 | 342 | 4.19E-08 | 245 | 381 | 0.000106 | 200 | 196 | 0.211326 | 293 | 323 | 0.834232 |
| **6** | 198 | 299 | 1.51E-07 | 196 | 326 | 1.99E-05 | 134 | 192 | 0.019161 | 233 | 276 | 0.314237 |
| **7** | 31 | 114 | 1.12E-13 | 30 | 113 | 1.70E-10 | 37 | 48 | 0.512044 | 64 | 69 | 1 |
| **8** | 378 | 762 | 3.97E-40 | 406 | 723 | 9.93E-15 | 369 | 388 | 0.458359 | 592 | 539 | 0.001935 |
| **9** | 57 | 115 | 1.43E-06 | 50 | 93 | 0.005166 | 45 | 50 | 1 | 87 | 67 | 0.034585 |
| **10** | 40 | 57 | 0.05205 | 36 | 37 | 0.639942 | 26 | 26 | 0.780797 | 52 | 33 | 0.016208 |
| **11** | 40 | 69 | 0.002637 | 49 | 48 | 0.474165 | 32 | 35 | 1 | 66 | 33 | 0.000234 |
| **12** | 44 | 94 | 4.24E-06 | 49 | 98 | 0.001107 | 38 | 38 | 0.728966 | 83 | 72 | 0.168793 |
| **13** | 4 | 15 | 0.010137 | 9 | 15 | 0.417598 | 3 | 4 | 1 | 10 | 10 | 1 |
| **14** | 12 | 16 | 0.450001 | 11 | 21 | 0.213684 | 7 | 6 | 0.783045 | 20 | 12 | 0.11231 |
| **15** | 16 | 14 | 0.856647 | 20 | 21 | 0.875592 | 10 | 3 | 0.048462 | 25 | 17 | 0.16335 |
| **16** | 0 | 4 | 0.056388 | 1 | 0 | 0.466011 | 0 | 0 | 1 | 0 | 0 | 1 |

\*Pink = the number of upregulated genes in the phylostratum is greater than expected.

\*Blue = the number of downregulated genes in the phylostratum is greater than expected.

\*White = neither the number of upregulated or downregulated genes is different to what is expected.

\*Note: significant p-values means that the number of upregulated and downregulated genes in a phylostrata is not associated with the total number of up and downregulated genes

### 2.3 Ancientness of the core set of DEG in tumours

We found a core set of 2815 overlapping DEG in the four cancer types studied (***Figure 12***). Although this number increases when only 3 or 2 tumour types are considered, there is still a considerable number of DEG that are unique to a single tissue type.

C:\Users\Anna\Documents\Unimelb\Thesis\TAI all types\Overlap_of_DEG_cancer_types.jpg

**Figure 12. Overlap of DEG between cancer types.** There is a core set of 2815 genes that are differentially expressed in the four cancer types. A large set of 2038 DEG are shared by the LUAD, BRCA, KIRC, but not HNSC. The number of DEG unique to each cancer type is close to 600 for LUAD, and KIRC, but only 375 for BRCA and 137 for HNSC. Although the KIRC samples have not followed the phylostrata-associated patterns of overall gene expressions, their DEG are largely consistent with the other cancer types.

We aimed to discern whether the genes of the core set were enriched for genes of early phylostrata and poor in newer genes using Fisher Exact tests. To avoid having the overrepresentation of early genes in all DEG (**Table VI**) as a confounder of the test, the Fisher tests were performed with only the total number DEG as the reference, rather than the entire set of genes. The core gene set was found to be enriched with genes of phylostrata 1 and 2, whereas post-bilaterian genes (phylostrata 6, 8-10, 12 and 13) were underrepresented (**Table *VII***).

**Table VII. Two- and one-sided p-values of the Fisher Exact tests for enrichment of phylostrata in the genes of the core set of overlapping DEG.** Fisher Exact tests were used to determine if the number of genes of a certain phylostrata in the core set was greater or lower than expected given the total number of DEG of that phylostrata. The core set of DEG is enriched for genes that emerged in phylostrata 1 and 2, but is poor in genes of phylostrata 6, 8, 9, 10, 12 and 13.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Phylostrata** | **DEG in core set** | **DEG not in core set** | **Two-sided p-value** | **Greater p-value** | **Less p-value** |
| **1** | 416 | 1055 | 0.0007465 | 0.0003847 | 0.9996946 |
| **2** | 1087 | 2837 | 6.79E-08 | 3.84E-08 | 1 |
| **3** | 62 | 155 | 0.1772853 | 0.1042649 | 0.9209691 |
| **4** | 375 | 1173 | 0.6802255 | 0.6812652 | 0.3417806 |
| **5** | 199 | 614 | 0.9327301 | 0.5684518 | 0.4649814 |
| **6** | 154 | 565 | 0.0357773 | 0.984727 | 0.019187 |
| **7** | 44 | 148 | 0.6130405 | 0.7410228 | 0.3174002 |
| **8** | 344 | 1283 | 0.0003142 | 0.9998684 | 0.0001684 |
| **9** | 40 | 198 | 0.0038059 | 0.9987954 | 0.0020617 |
| **10** | 15 | 131 | 1.66E-05 | 0.9999973 | 8.30E-06 |
| **11** | 32 | 128 | 0.1956862 | 0.9324519 | 0.0966762 |
| **12** | 31 | 186 | 0.0002332 | 0.9999463 | 0.000111 |
| **13** | 2 | 30 | 0.0125767 | 0.9986902 | 0.0074026 |
| **14** | 5 | 39 | 0.0517598 | 0.9918362 | 0.0238192 |
| **15** | 9 | 47 | 0.1618082 | 0.9564513 | 0.0856452 |
| **16** | 0 | 4 | 0.578098 | 1 | 0.3218595 |

\*Pink = the number of genes of the phylostratum in the core set is greater than expected.

\*Blue = the number of genes of the phylostratum in the core set is less than expected.

\*White = the number of genes of the phylostrata in the core is not statistically different to what is expected.

Note: The DEG not in the core set are those that were DE in at least one cancer type, but not in all four

## Functional studies

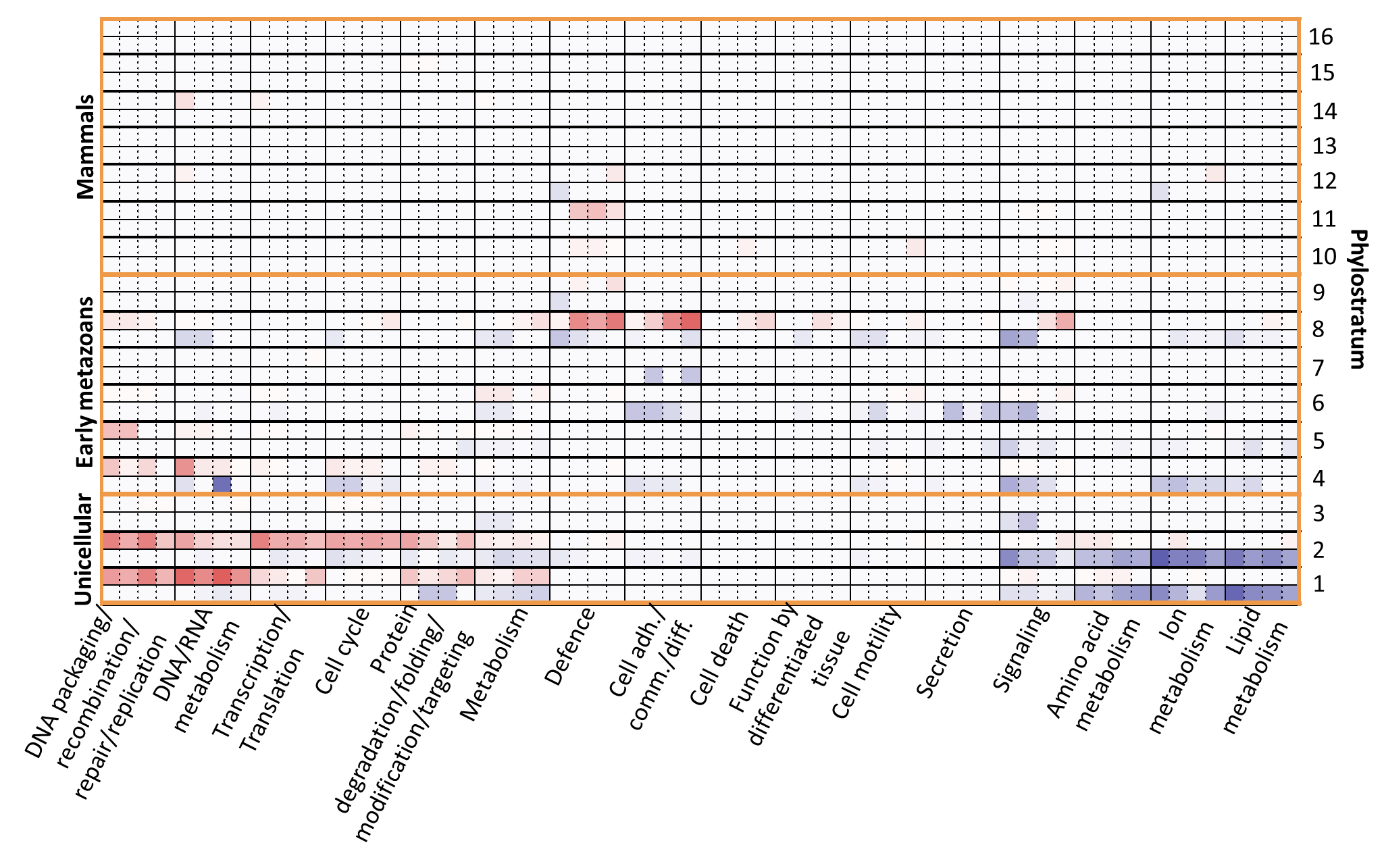
Until now our analysis of the atavism hypothesis has been from the point of view of the association between the age of genes and their expression levels. Another dimension of the atavism hypothesis can be addressed by also taking into account the biological functions of the genes.

### 3.1 Differentially expressed gene functions in each phylostratum

The collapsed set of gene annotations was used to determine the functional groups that were up- and downregulated in each phylostratum for the four tumour types (***Figure 13***). There was large consistency of the up- and downregulated functions between cancer types in unicellular and early metazoan phylostrata. In the case of mammalian phylostrata, there was overall consistency in the nearly non-existent differentially expressed functions, which agrees with the lack of gene expression differences between normal and tumours previously found.

The functional sets that have the largest number of upregulated GO-term sets were those related to DNA and RNA processes and metabolism, the cell cycle and protein processes, and were mainly found in unicellular phylostrata. However, gene sets involved in signalling and amino acid, ion and lipid metabolism were strongly downregulated in these phylostrata. However, strong signals of these functions faded in newer phylostrata, and were not present in mammalian phylostrata.

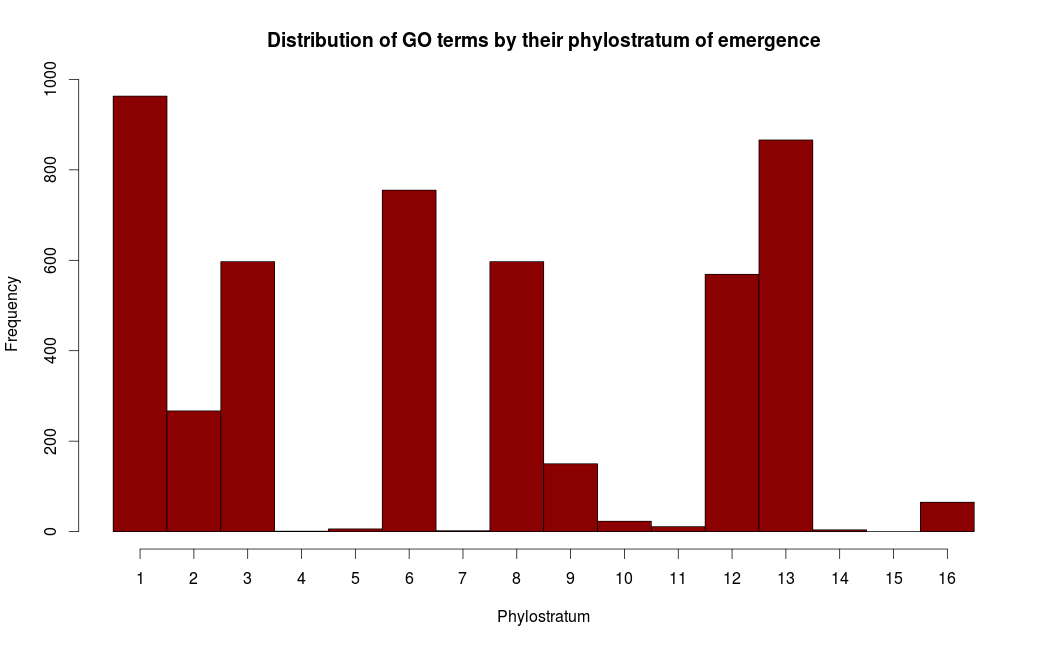
Contrarily, gene sets involved in defence and cell adhesion, communication and differentiation that show differential expression are concentrated in genes emerged in early metazoan genes, but not in unicellular phylostrata.

******

**Figure 13. Up- and downregulated functions in the phylostrata.** The phylostrata correspond to the time of emergence of genes.Thenumber of up- (red) and downregulated (blue) specific GO-term functions were counted for every umbrella term for each cancer type. Dotted vertical lines within each umbrella term correspond to results for each cancer type presented in the following order: LUAD, BRCA, HNSC and KIRC. There is an overall consistency between the number of up- and downregulated GO-terms across cancer types. The strongest signals are found in unicellular phylostrata, which are mainly upregulated in DNA- and RNA-related processes, cell cycle and protein metabolism, and downregulated in amino acid, ion and lipid metabolism. Early metazoans are characterized by an upregulation of defence- and cell-adhesion, communication and differentiation- related GO-terms. There are no evident strong signals in mammalian phylostrata.

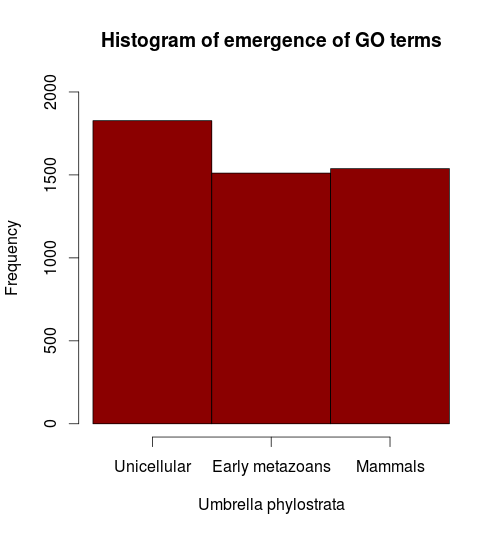
### Dating GO terms

The sixth-generation GO terms were dated by their phylostratum of emergence in order to pinpoint the ancientness of the functions (***Figure 14***). The distribution of functional innovations is uneven through the phylostrata, likely due to the limited number of species with annotations in certain phylostrata and the incomplete annotation of species that are not routinely used as model organisms. However, it is also possible that some phylostrata do not represent large evolutionary jumps that would involve the generation of numerous new functions and GO terms.



**Figure 14. Distribution of sixth-generation GO terms by their phylostratum of emergence.** Assignment ofGO terms to a phylostratum was determined by the oldest species annotated with the term. The bars represent the number of functional innovations originated in a phylostratum. The small number of functional innovations in phylostratum 4, 5, 7, 10, 11, 14 and 15 are probably due to a combination of the low-quality annotation of species of these phylostrata and a real biological lack of functional innovation.

The uneven distribution and large gaps in the number of GO-terms across the phylostrata could cause biased results towards better characterized phylostrata, due to a loss of power when checking for enrichment of a certain phylostrata in the upregulated gene sets and to detect patterns. Therefore, the phylostrata of emergence were collapsed into three groups, covering the Unicellular, Early Metazoan or Mammalian origins of GO terms (***Figure 15***). This new grouping revealed that a roughly similar number of GO terms emerged in each group.



**Figure 15. Grouping of the point of emergence of GO-terms into umbrella phylostrata.** The phylostrata of emergence of GO terms shown in **Figure 14** were collapsed into three major groups: Unicellular (Phylostrata 1-3), Early metazoans (Phylostrata 4-9) and Mammals (Phylostrata 10-16), in order to reduce the bias due to the low quality of annotations of non-commonly used organisms. This grouping led to a roughly uniform distribution of the number of evolutionary innovations across the umbrella phylostrata.

### 3.3 Enrichment of ancient functions in tumours

Enrichment tests were performed to detect whether upregulated functional sets mostly represented ancient functions (***Table VIII***). This trend was found to be true for LUAD and BRCA tumour samples, whose upregulated gene sets were enriched with functions that dated back to unicellularity. Additionally, we found that downregulated gene sets were mostly related with functions that emerged with early metazoans. Finally, the most recent mammalian functions were not enriched in either the upregulated or downregulated genes, suggesting that these newer functional sets did not have an important change in expression in tumours. However, an opposite trend was observed for the HNSC and KIRC samples, with the upregulation of early metazoan functions.

**Table VIII. The number of up- and downregulated functions in the umbrella phylostratum and the two-sided p-values of enrichment tests.** The up- and downregulated status of all gene sets in the four tumour types was determined with Wilcoxon tests using the tumour/normal log fold change estimates of the genes. Fisher Exact tests were subsequently used to test for enrichment of each umbrella phylostratum in the up- and downregulated gene sets. Upregulated gene sets in LUAD and BRCA tumour samples are enriched for GO terms that emerged in the unicellular umbrella phylostratum, whereas downregulated gene sets are enriched for those emerged in early metazoans. The opposite was observed for HNSC and KIRC samples.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **LUAD** | | | **BRCA** | | | **HNSC** | | | **KIRC** | | |
| **Umbrella phylostratum** | **Number upregulated** | **Number downregulated** | **p-value** | **Number upregulated** | **Number downregulated** | **p-value** | **Number upregulated** | **Number downregulated** | **p-value** | **Number upregulated** | **Number downregulated** | **p-value** |
| **Unicellular** | **166** | **235** | 1.03E-23 | 151 | 235 | 2.49E-20 | 170 | 185 | 4.18E-09 | 193 | 166 | 5.80E-08 |
| **Early metazoans** | **21** | **261** | 3.64E-25 | 16 | 213 | 2.49E-20 | 152 | 55 | 6.82E-09 | 157 | 52 | 1.29E-06 |
| **Mammals** | **3** | **4** | 0.401 | 0 | 0 | 1 | 1 | 0 | 1 | 9 | 0 | 0.01576 |

\*Pink: Upregulated gene sets are enriched with sets originated at the umbrella phylostratum.

\*Blue: Downregulated gene sets are enriched with sets originated at the umbrella phylostratum.

\*White: Neither up- or downregulated gene sets are enriched with sets of the umbrella phylostratum.

### 3.4 Identifying the gene sets that undergo a process of atavism

The atavism hypothesis can also be studied at the level of functional gene sets. A functional gene set was said to be undergoing atavism if its genes that emerged during unicellularity were upregulated, whereas its multicellularity genes were downregulated. Here, we focused on making a descriptive analysis by identifying the sets that met the above criteria.

We found a total of 159 functional gene sets that seemingly underwent a process of atavism in at least two cancer types. The functional gene sets were involved in a broad range of cellular functions (***Table IX***, **Table S 6**), suggesting that the atavism of functional groups is widespread.

**Table IX. Summary of functional sets that underwent atavism.** For each functional set, the logFC of genes were used as a comparative metric to determine the up- and downregulated status of genes belonging to unicellular and multicellular phylostrata with Wilcoxon tests. The groups of the functional sets with up regulated pre-multicellularity genes and downregulated post-multicellularity genes in at least two cancer types are shown here.

|  |  |
| --- | --- |
| **Process** | **Functional GO-term sets** |
| **Cell process** | Cell cycle, cell death, cell morphology, cell motility, cell response to stress, membrane biogenesis, nuclear transport |
| **DNA process** | DNA damage |
| **DNA/RNA process** | Transcription, translation, nucleotide/nucleoside metabolism |
| **Signaling & Metabolism** | Amino acid, carbohydrate, ion and lipid metabolism, energy, cell signaling pathways, secretion |
| **Protein process** | Modification, degradation, transport, targeting |

# Discussion

## Phylostratigraphy captures the major evolutionary changes that led to the current human genome composition

Phylostratigraphy has been used to assess the phylogenetic origin of genes involved in the embryonic development in *Drosophila melanogaster* (Domazet-Lošo, Brajković, & Tautz, 2007), human genetic diseases (Domazet-Loso & Tautz, 2008), cancer genes (Domazet-Loso & Tautz, 2010), as well as the vertebrate head sensory system (Sestak, Božičević, Bakarić, Dunjko, & Domazet-Lošo, 2013), and here we have employed it to find the age distribution of human genes.

An important observation is the limited number of human genes that were assigned to the base of the Opisthokonta, Phylostratum 3. This result can be explained by the near exclusivity of fungal species in this phylostratum (24 fungi and 1 Choanoflagellida) and the composition of their genome. The genomes of some fungi are considered to be small in comparison to other Eukaryotes (Tunlid & Talbot, 2002). For those species with large genomes, these are thought to have expanded through large duplication events with important gene losses (Dujon et al., 2004), or significant gene losses followed by the acquisition of new genes (Martin et al., 2010; Spanu et al., 2010; Tunlid & Talbot, 2002). In both scenarios, these newly acquired genes occurred after the diversification of the opisthokonts into fungi and metazoans, and would therefore not be present in their last common ancestor. Genes in common between humans and these fungi probably represent those that were necessary for the jump from primitive Eukaryota to Opisthokonta, and thus most of the genes shared between humans and opisthokonts (i.e., phylostratum 3) would have already been present in phylostratum 2.

The peaks in the number of genes assigned to phylostratum 4 and 8 can be attributed to large duplication events. A large-scale duplication event has been pinpointed to the onset of metazoans (Huerta-Cepas, Dopazo, Dopazo, & Gabaldón, 2007; Wenger & Galliot, 2013). These same authors found evidence of possibly whole-genome duplication events after the split of the chordata (phylostratum 7), but right before the radiation of vertebrates (phylostratum 8).

One important confounder of phylostratigraphy is the presence of genes that originated from duplication events within a single genome, known as paralogous genes {FormattingCitation}(Sonnhammer & Koonin, 2002). The distribution of gene ages constructed from the OrthoMCL algorithm suggest that this method was able to distinguish between orthologs and paralogs of the species, and correctly assigned the human genes to its founder gene, whether it had been the original ortholog or a paralog generated by a previous duplication event. For example, we found that 39.84% of the human genes mapped to Phylostrata 1-3, whereas the method of Domazet-Loso & Tautz (2008) led to 74.80% of the genes being assigned to these early phylostrata, suggesting that their method favours mapping to the gene before the duplication event. Therefore, the differences in the peaks observed with the distribution obtained with the phylostrata assignments of Domazet-Loso & Tautz (2008) probably indicate differences in the degree sensitivity of the detection thresholds and subsequent curation methods when differentiating orthologs and paralogs. However, other authors have obtained similar results to those obtained in this study, where 33% of human proteins were found to map to pre-metazoan phylostrata (Wenger & Galliot, 2013), greatly supporting our results.

We further verified the gene assignments to phylostrata by enrichment of the functional GO-terms, and found that basal cell functions were confined to early pre-multicellularity phylostrata (1 and 2), whereas functions related to a multicellular state were enriched in early post-multicellularity phylostrata (4-8). Finally, complex functions related to the immune system and the detection and response to stimulus are mostly enriched in late phylostrata (phylostratum 8 onwards). Overall, there was an evident association between the phylostrata of the genes and the complexity of their functions, supporting our assignments of genes to phylostrata.

## Patterns in the up- and downregulation of genes in cancer can be traced back to their ancientness

### 2.1 Tumours have an older transcriptome than normal cells

The transcriptome age index (TAI) provided information about changes in the age of the transcriptome of normal and tumour samples. This technique has proven useful in discerning the changes of transcriptome ages in ontogeny (Domazet-Lošo & Tautz, 2010), stem-cell types (Hemmrich et al., 2012) and in structures involved in the vertebrate sensory system (Sestak et al., 2013).

Importantly, we found that the tumour transcriptomes of LUAD, BRCA and HNSC were significantly older than their normal samples, suggesting that ancient genes are increasing their transcription levels in tumours, which is consistent with the atavism hypothesis.

On the other hand, the degree of TAI differences between tumour and normal samples varied with tumour type. The larger spread of tumour TAI values in comparison to the normal types is likely due to the presence of multiple cancer sub-types and stages in the samples with a variety of mutational profiles, aberrant RNA transcripts, pathway activations and epigenetic changes (Cancer & Atlas, 2012; The Cancer Genome Atlas Research Network, 2013, 2014). Interestingly, the TAI of the HNSC normal samples was found to be highly variable, which can be attributed to the diversity of the locations where the samples were taken from, which include the larynx, floor of mouth, tongue, oral cavity, buccal mucosa and hard palate. Furthermore, the TAI was also found to be highly dependent on the identity of the normal tissue (lung, breast, head and neck or kidney), suggesting that the age of the transcriptome is also dependent on the tissue type.

### 2.2 Pre-multicellularity genes are upregulated in tumours, whereas later genes are downregulated

Having already established that tumour samples have an older transcriptome, we studied the changes in expression levels of the phylostrata as well as their enrichment of over- and underexpressed genes. We found that the expression levels and the genes of phylostrata 1 and 2 were consistently upregulated in LUAD, BRCA and HNSC tumours, whereas later phylostrata (4 to 8-11) were underexpressed, indicating that these two complementary trends were responsible for the older tumour transcriptome. These results were consistent with the atavism hypothesis, where older genes are expected to have an increased expression in tumours.

Importantly, genes of Phylostrata 1, 2 and 3 emerged in unicellular ancestors, whereas genes of Phylostrata 4 onwards date back to a multicellular ancestor. Multicellularity arose multiple times in the course of evolution, but only once in metazoans, and there is general agreement that multicellularity in fungi, metazoans and plants arose as independent events considering that genes involved in multicellularity in metazoans are not present in fungi (Knoll, 2011; Niklas, 2014; Ruiz-Trillo et al., 2007). Hence, the oldest ancestor of multicellularity in humans is represented in Phylostratum 4, and the regulatory mechanisms of multicellularity in metazoans are highly conserved across the descendant lineages (Ruiz-Trillo et al., 2007). Therefore, although some of the modern-day species used as reference to map human genes to phylostratum 3 are currently multicellular, the human common ancestor would have been unicellular.

A link between multicellularity and cancer has been previously suggested, given that cancer genes have been associated with the emergence of metazoans (Domazet-Loso & Tautz, 2010) and many of the hallmarks of cancer are the basis for a unicellular state (Hanahan & Weinberg, 2011). Our results further these observations by suggesting that tumour cells are employing mechanisms of ancient unicellular organisms, and are switching off genes associated with a multicellular state.

### 2.3 Multiple confounders led to the lack of strong and clear signals from HNSC and KIRC samples

Unlike the lower TAI found in LUAD, BRCA and HNSC tumour samples and the distribution of tumour-normal expression difference in the phylostrata, normal KIRC samples were found to have an older transcriptome than their tumour counterparts and ancient genes were found to be upregulated.

We suspect that low sample purity was acting as a major confounder that dissipated any strong biological signal that could have arisen from the data. Somatic copy number alteration studies revealed that the sample purity of the tumour KIRC samples was only 54±14%, where stromal or endothelial contaminations, and tumour heterogeneity are thought to be the culprit (Gerlinger et al., 2012; The Cancer Genome Atlas Research Network, 2013). Considering the high correlation between the TAI and tumour sample purity (***Figure 8***), the results obtained with the KIRC samples cannot be faithfully attributed to tumour and normal sample differences. However, we cannot rule out that the aetiology of KIRC samples does not involve the activation of ancient mechanisms of survival.

Unlike in the system-level analysis where the HNSC samples had mimicked the results obtained with LUAD and BRCA (Results sections 2-2.2.1), there were overall less differences between tumour and normal samples in the gene-level analyses (2.2.2). The previously discussed high heterogeneity of the tumour and normal samples could have avoided the detection of significant differences. Nevertheless, we cannot exclude the presence of a limited subset of genes with large changes in expression levels that were mainly responsible for the changes of the TAI and expression differences of the phylostrata, but whose effect was not detected when using non-parametric approaches.

### 2.4 A core set of DEG detected in the four cancer types is enriched with pre-multicellularity genes

We found that an overlapping set of 2815 genes that were differentially expressed in the four cancer types was more enriched with pre-multicellularity genes than the DEG that were not part of this set. This core set of genes can be preliminarily interpreted as a ubiquitous set responsible for the general biological features associated with a cancerous state, and the rest of the DEG are those involved with tissue-specific features of the tumour. Further work would provide additional clarification of this point, but an association between possibly basal or key genes in cancer and unicellularity functions can be preliminarily suggested.

## Ancient functions are upregulated in tumours

### 3.1 The age of the GO terms parallels widespread evolutionary innovations

Dating of the GO terms revealed an uneven distribution of the number of GO-terms assigned to each phylostrata. The large number of functional innovations in phylostrata 1 to 3 and 6, 8, 12 and 13 reflect the large evolutionary steps given at these stages: the emergence and radiation of cellular and eukaryotic life, the emergence of bilateral symmetry, bones and nourishing via a placenta. Phylostratum 4, 5, 7, 10 and 11 had a very low number of GO terms, probably due to limited annotation of species of these phylostrata. Although the reference species used for phylostrata 14 and 15 are considerably well-studied (*Macaca mulatta* and *Pan troglodytes*), there are very few GO-terms that emerged in these phylostrata, probably because of the relatedness of the species involved. Finally, the peak of GO-terms annotated to phylostratum 16 probably represent a combination of innovations in humans, as well as a more thorough annotation of human genes.

### 3.2 There is a signature set of gene functions differentially expressed in the phylostrata

Functional gene sets of unicellular phylostrata showed the greatest changes in expression between tumour and normal samples, which were overall consistent between tumour types. Functions related to DNA and RNA processes and metabolism, the cell cycle and protein processes being upregulated, and those involved in signalling and amino acid, ion and lipid metabolism were strongly downregulated. These results suggest that the overexpression of genes belonging to unicellular phylostrata previously found (**Figure 10**, ***Table V***, ***Table VI***), was due to the interplay between the aforementioned functions.

On the other hand, gene sets involved in defence and cell adhesion, communication and differentiation showed differential expression in genes emerged in early metazoan genes. This result is in agreement with the emergence of multicellularity in early metazoans. The strongly upregulated functions of the immune system in phylostratum 8 probably refer to the adaptive instead of the innate immune system, given that the former dates back to early vertebrates (phylostratum 8), whereas the latter dates back to early eukaryotes (Danilova, 2006; Janeway, Travers, Walport, & Shlomchik, 2001).

Interestingly, we had previously found that genes of early metazoan phylostrata were strongly downregulated, but such a strong downregulation was not observed in any particular gene functions. It is possible that the accumulation of the slight downregulation of multiple gene functions accounted for the previously observed results.

### 3.3 Ancient gene functions are upregulated in tumours

Upregulated functional sets in BRCA and LUAD were mostly composed of ancient functions that can be traced back to the origin of unicellularity, whereas early metazoan functions were likely to be downregulated. Mammalian functions were mostly unchanged. These results support an important corollary of the atavism hypothesis: if cancer cells are regressing to an ancient state, then we would expect that most upregulated functions would date back to primitive organisms.

The absence of this pattern in HNSC and KIRC samples might be partially due to low purity and sample heterogeneity, which has been found to have a large negative effect on previous results. It is also possible that a complete process of atavism is not occurring in these tumour types, and that the atavism is evidenced by the reactivation of early metazoan functions.

### 3.4 Evidence of atavism can also be found at the functional-set level

Finally, we aimed to perform a descriptive analysis to identify the gene set that underwent a process of atavism. Namely, the regression of cancer cells to a primitive state need not be limited to a phenomenon that occurs globally in the cell, but it might also occur within individual gene sets: the older genes of the set would be overexpressed in tumours, whereas newer genes would be downregulated. Other authors have addressed this corollary by suggesting that newly acquired genes and functions are the least embedded in cancer and would thus be prone to dysregulation and loss-of-function, which could be used as therapeutic drug targets (Lineweaver, Davies, & Vincent, 2014). A range of specific gene sets involved in basal cell processes, defence, DNA, RNA and protein processes and metabolism were found to comply with this criteria. Additionally, there was large overlap between these gene sets and those that were found to be differentially expressed in tumours, suggesting an important role of these functions in the process of atavism.

## Activation of ancient mechanisms in cancer

The results obtained from the functional analysis can be used as starting points to identify the ancient mechanisms that are activated in cancer.

### 4.1 The Warburg effect

Unlike normal cells that rely on aerobic respiration and the mitochondrial tricarboxylic acid (TCA) cycle to produce NADH and ATP, cancer cells exhibit the Warburg effect, where there is a switch to a less efficient mechanism for energy production known as aerobic glycolysis, with the additional downside of the production of lactate (Vander Heiden, Cantley, & Thompson, 2009; Warburg, 1956). The use of lactic acid fermentation in many bacterial species has led many authors to propose that the Warburg effect is a core example of a process of atavism (Davies & Lineweaver, 2011; Lineweaver et al., 2014; Vincent, 2012).

Considering that lactic acid fermentation is a trademark of anaerobic bacteria and early protozoan eukaryotes (Upcroft & Upcroft, 2001), the genes involved in this pathway would most likely belong to Phylostrata 1 and 2. The widespread upregulation of genes of these phylostrata would suggest that these Warburg effect genes are being activated. However, we did not find direct evidence of the activation of genes involved in aerobic glycolysis pathways. It is possible that shifting to a fermentation metabolism only requires slight upregulation of the involved genes, which would be missed by our methods. Additionally, GO-term annotations might not adequately reflect the involvement of genes in these pathways.

Conversely, we did find similarities between a state of anaerobic metabolism and cancer cells, primarily related to the downregulated ancient genes involved in amino acid metabolism. Amino acid synthesis is known to be very limited in lactic acid bacteria (Ljungh & Wadstrom, 2009), suggesting that our results are consistent with a Warburg effect managed by genes from unicellular phylostrata. Additionally, there is further evidence of the Warburg effect in our results, since glycolytic cells show a decreased level of lipid content in comparison to their oxidative counterparts (Alasnier, Remignon, & Gandemer, 1996), and we found a decreased lipid metabolism in tumour samples.

### 4.2 An ancient stress response

The Cellular Stress Response (CSR) involves the activation of certain functional gene sets that were identified as significant in this study, such as DNA damage sensing/repair, cell cycle checkpoints, protein degradation and repair, chromosome maintenance, nucleic acid and chromatic stabilization and repair (Kultz, 2003; Kültz, 2005). The CSR is a minimal response of a cell to stress, making it highly conserved in archaea, bacteria, yeasts and humans (Gasch et al., 2000; Kültz, 2005). Actually, the activation of ancient stress mechanisms by cancer had been hypothesized many years ago (Israel, 1996).

The stress response proteins are among the most conserved, so their involvement in basal functions of cell maintenance is expected. Actually, the stress response of unicellular cells is enriched with genes related to a high growth rate (Brauer et al., 2008), and the stress response mechanism disrupts DNA repair mechanisms and promotes mutagenesis in yeast and mammalian cells under stress (Shor, Fox, & Broach, 2013; Yuan, Narayanan, Rockwell, & Glazer, 2000). Yeast under nutrient stress have been found to exhibit characteristics of the Warburg effect (Brauer et al., 2008). These features suggest that ancient stress response mechanisms might be drivers of the increased growth rate of tumours.

Further parallelisms can be drawn between this minimal stress response and the results found in this study. For example, cancer cells are known to exhibit an increased lipid metabolism (Zhao, Butler, & Tan, 2013), which is thought to aid in membrane synthesis, energy production and lipid protein modification (Mashima, Seimiya, & Tsuruo, 2009). However, separating genes by their ancientness allowed us to determine that unicellular genes involved in lipid metabolism are mostly downregulated. Downregulation of lipid metabolism has been identified in bacterial species under certain types of stress, together with the underexpression of genes involved in amino acid metabolism (Mols et al., 2010). Furthermore, some multicellular species are known to downregulate their lipid metabolism under stress (Yampolsky et al., 2014), and endoplasmatic reticulum stress has been associated with diminished lipid metabolism (Hotamisligil, 2010).

On the other hand, highly conserved stress proteins can modulate and are related to the expression of molecules of the immune system (Borghesi & Lynes, 1996; Moseley, 2000), which might relate the upregulation of immune system-related gene sets in early metazoans.

The activation of genes involved in the CSR is consistent with one of the pillars of the atavism hypothesis, as it involves the re-activation of genes evolved millions of years ago to protect the cell from the hostile environment. In summary, the functional studies performed in this study point towards the re-activation of ancient stress mechanisms.

### 4.3 Selection pressures and atavism

Validation of the atavism hypothesis *per se* is challenging. This work aimed to derive evidence of atavism from the analysis of the gene expression data, and we have interpreted the functional results obtained under the light of atavism. However, experiments specifically designed to test the hypothesis can also be applied. For example, altering in normal differentiated cells multiple hub genes of ancient processes and genes intimately linked to multicellularity to determine if it would lead to the acquisition of cancer properties; or inactivating these hub genes in cancer to see its malignant features are lost.

Nevertheless, even with these two complementary approaches and obtaining results that are in agreement with atavism hypothesis, it would be difficult to differentiate selection for a primitive state and selection for cells with a survival advantage. For instance, we have found that the upregulation of DNA replication/repair and DNA/RNA metabolism executed by ancient genes contributes significantly to the lower TAI in tumours, but a valid argument is that this upregulation is due to a selection for faster dividing cells, rather than selection for cells with a more primitive expression phenotype.

By definition, selection in cancer occurs at the cellular level, where cells that are the best adapted to survive under the conditions will outnumber those that are most vulnerable. However, in our point of view, an increased fitness is brought about by the reversal to a primitive state. Using the previous example, selection pressures would favour faster dividing cells, but these cells would have an increased rate of division *because of* their ancient expression phenotype. In other words, an ancient expression phenotype will be indirectly selected for.

The question of whether cancer is being driven by atavism, or if atavism is a consequence of cancer progression cannot be definitely answered with the available information. However, the process of atavism can still be seen as an implicit mechanism occurring behind the scenes of tumour progression.

# Conclusions

The main aim of this study was to find evidence of atavism in the gene expression levels of tumours. We hypothesised that if cancer cells were reactivating an ancient toolkit for survival, then increased levels of expression should be evidenced in primitive genes and gene functions.

Human genes were dated in order to obtain a global picture of the ancientness of the human genome, and our results were widely consistent with known evolutionary events in the history of the human species.

The dating of human genes allowed the use of the Transcriptome Age Index technique to compare the ancientness of tumour and normal transcriptomes. Lung, breast and head and neck tumours were shown to have an older transcriptome than normal tissues, and that this was mainly due to the combined action of two opposing forces: the upregulation of genes that date back to unicellularity and the downregulation of post-multicellularity genes.

Functional set analyses revealed that basal gene functions that emerged during unicellularity or the early stages of multicellularity are upregulated in tumours. Importantly, similarities between the activated ancient functions and those related to a general cellular stress response reinforced the hypothesis that tumour cells are reactivating ancient mechanisms of survival.

Finally, the results obtained in this study provide evidence of the reversal of cancer cells to a primitive state by means of the overexpression of ancient genes and functions and the detachment from newly acquired cellular mechanisms, making our results consistent with the atavism hypothesis.

# Further work

The pioneer nature of this project made it limited to searching for evidence of atavism in only four cancer types by focusing on signatures in the cancer transcriptome and broad cellular functions. Further validation of the atavism hypothesis, its extent and universality, as well as characterization of the pathways involved should be explored.

* Directly comparing the gene expression levels in cancer cells to those of cells of primitive species would shed light on the processes that are conserved. For example, unicellular organisms and basal metazoans can be used as a reference of primitive survival mechanisms, and gene-network level studies can serve as a basis for finding conserved large-scale pathways. Furthermore, taking advantage of the presence of human cells that exhibit features of a primitive and de-differentiated state, such as normal and cancer stem cells, and mesenchymal cells, would allow the detection of subtle atavism processes that do not necessarily cross over the species-specific or multicellular-organism barriers. Another approach could directly assess if genes and functions involved in multicellularity are switched off in cancer.
* The link between atavism in cancer and the use of conserved stress mechanisms should be studied in more depth by drawing parallelisms between the stress response of primitive cells and cancer. Specifically, it would be of interest to determine if there is greater similarity in the stress response of unicellular organisms and cancer cells, rather than with normal cells.
* Other tumour types and sub-types can be used to assess the ubiquity of atavism in cancer. Their genotype and degree of malignancy can also be taken into account to search for relationships between the degree of atavism and the aetiology of cancer.
* Finally, characterization of the atavism in cancer need not be limited to gene expression. A broad exploration can involve searching for atavism signatures in different -omics data, including genomic, proteomic, epigenetic and metabolomics data which will further delimit the extent and implications of the atavism.

# List of References

Alasnier, C., Remignon, H., & Gandemer, G. (1996). Lipid Characteristics Associated with Oxidative and Glycolytic Fibres in Rabbit Muscles. *Meat Science*, *43*(96), 213–224. doi:10.1016/S0309-1740(96)00015-0

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Bartek, J., Mistrik, M., & Bartkova, J. (2012). Thresholds of replication stress signaling in cancer development and treatment. *Nature Structural & Molecular Biology*, *19*(1), 5–7. doi:10.1038/nsmb.2220

Benson, D. a, Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2009). GenBank. *Nucleic Acids Research*, *37*(Database issue), D26–31. doi:10.1093/nar/gkn723

Binns, D., Dimmer, E. C., Huntley, R. P., Barrell, D. G., O’Donovan, C., & Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Proteomics*, *25*(22), 3045–3046. doi:10.1002/pmic.200800002

Borghesi, L. A., & Lynes, M. A. (1996). Stress proteins as agents of immunological change: some lessons from metallothionein. *Cell Stress & Chaperones*, *1*(2), 99–108. doi:10.1379/1466-1268(1996)001<0099:SPAAOI>2.3.CO;2

Brauer, M. J., Huttenhower, C., Airoldi, E. M., Rosenstein, R., Matese, J. C., Gresham, D., … Botstein, D. (2008). Coordination of Growth Rat , Cell Cycle, Stress Response, and Metabolic Activity in Yeast, *19*(January), 352–367. doi:10.1091/mbc.E07

Bullard, J., Purdom, E., Hansen, K., & Dudoit, S. (2010). Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiment. *BMC Bioinformatics*, *11*(94). doi:10.1186/1471-2105-11-94

Cahill, D. P., Kinzler, K. W., Vogelstein, B., & Lengauer, C. (1999). Genetic instability and darwinian selection in tumours. *Trends in Cell Biology*, *9*(12), M57–60. doi:10.1016/S0968-0004(99)01466-8

Cancer, T., & Atlas, G. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. doi:10.1038/nature11412

Carlson, M. (2014). GO.db: A set of annotation maps describing the entire Gene Ontology. R package.

Chen, L., Kasai, T., Li, Y., Sugii, Y., Jin, G., Okada, M., … Seno, M. (2012). A model of cancer stem cells derived from mouse induced pluripotent stem cells. *PloS One*, *7*(4), e33544. doi:10.1371/journal.pone.0033544

Danilova, N. (2006). The Evolution of Immune Mechanisms. *Journal of Experimental Zoology*, *306B*, 496–520. doi:10.1002/jez.b

Davies, P. C. W., & Lineweaver, C. H. (2011). Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. *Physical Biology*, *8*(1), 015001. doi:10.1088/1478-3975/8/1/015001.Cancer

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., … Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14*(6), 671–683. doi:10.1093/bib/bbs046

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O’Donovan, C., Martin, M. J., … Apweiler, R. (2012). The UniProt-GO Annotation database in 2011. *Nucleic Acids Research*, *40*(Database issue), D565–70. doi:10.1093/nar/gkr1048

Domazet-Lošo, T., Brajković, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics*, *23*(11), 533–539. doi:10.1016/j.tig.2007.08.014

Domazet-Lošo, T., Klimovich, A., Anokhin, B., Anton-Erxleben, F., Hamm, M. J., Lange, C., & Bosch, T. C. G. (2014). Naturally occurring tumours in the basal metazoan Hydra. *Nature Communications*, *5*(May), 4222. doi:10.1038/ncomms5222

Domazet-Loso, T., & Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Research*, *13*(10), 2213–9. doi:10.1101/gr.1311003

Domazet-Loso, T., & Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular Biology and Evolution*, *25*(12), 2699–707. doi:10.1093/molbev/msn214

Domazet-Loso, T., & Tautz, D. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biology*, *8*, 66. doi:10.1186/1741-7007-8-66

Domazet-Lošo, T., & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, *468*(7325), 815–8. doi:10.1038/nature09632

Duesberg, P., & Rasnick, D. (2000). Aneuploidy, the somatic mutation that makes cancer a species of its own. *Cell Motility and the Cytoskeleton*, *47*(2), 81–107. doi:10.1002/1097-0169(200010)47:2<81::AID-CM1>3.0.CO;2-#

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., … Souciet, J.-L. (2004). Genome evolution in yeasts. *Nature*, *430*(6995), 35–44. doi:10.1038/nature02579

EMBL-EBI, & The Wellcome Trust Sanger Institute. (2013). Ensembl Genome Browser 73.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., … Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, *11*(12), 4241–57.

Genome Reference Consortium. (2013). Human Build 37 patch release 12.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Math, D., … Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *The New England Jounral of Medicine*, *366*(10), 883–892.

Gordon, D. J., Resio, B., & Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. *Nature Reviews. Genetics*, *13*(3), 189–203. doi:10.1038/nrg3123

Greaves, M. (2002). Cancer causation: the Darwinian downside of past success? *The Lancet Oncology*, *3*(4), 244–251. doi:10.1016/S1470-2045(02)00716-7

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. C., & Shyr, Y. (2013). Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PloS One*, *8*(8), e71462. doi:10.1371/journal.pone.0071462

Hahn, W. C., Dunn, I. F., Kim, S. Y., Schinzel, A. C., Firestein, R., Guney, I., & Boehm, J. S. (2009). Integrative genomic approaches to understanding cancer. *Biochim Biophys Acta*, *1790*(6), 617–632. doi:10.1016/j.bbagen.2009.02.003.Integrative

Halazonetis, T. D., Gorgoulis, V. G., Bartek, J., Science, S., Series, N., & Mar, N. (2008). An Oncogene-lnduced DNA Damage Model for Cancer Development. *Science*, *319*(5868), 1352–1355. doi:10.1126/science.1140735

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–74. doi:10.1016/j.cell.2011.02.013

Hemmrich, G., Khalturin, K., Boehm, A.-M., Puchert, M., Anton-Erxleben, F., Wittlieb, J., … Bosch, T. C. G. (2012). Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity. *Molecular Biology and Evolution*, *29*(11), 3267–80. doi:10.1093/molbev/mss134

Hotamisligil, G. S. (2010). Endoplasmic reticulum stress and the inflammatory basis of metabolic disease. *Cell*, *140*(6), 900–17. doi:10.1016/j.cell.2010.02.034

Huber, M. a, Kraut, N., & Beug, H. (2005). Molecular requirements for epithelial-mesenchymal transition during tumor progression. *Current Opinion in Cell Biology*, *17*(5), 548–58. doi:10.1016/j.ceb.2005.08.001

Huerta-Cepas, J., Dopazo, H., Dopazo, J., & Gabaldón, T. (2007). The human phylome. *Genome Biology*, *8*(6), R109. doi:10.1186/gb-2007-8-6-r109

Israel, L. (1996). Tumour progression: random mutations or an integrated survival response to cellular stress conserved from unicellular organisms? *Journal of Theoretical Biology*, *178*(4), 375–80. doi:10.1006/jtbi.1996.0033

Jacobs, D. K., Nakanishi, N., Yuan, D., Camara, A., Nichols, S. a, & Hartenstein, V. (2007). Evolution of sensory structures in basal metazoa. *Integrative and Comparative Biology*, *47*(5), 712–23. doi:10.1093/icb/icm094

Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. J. (2001). *Immunobiology* (5th editio.). New York: Garland Science.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database*, *2011*, bar049. doi:10.1093/database/bar049

Knoll, A. H. (2011). The Multiple Origins of Complex Multicellularity. *Annual Review of Earth and Planetary Sciences*, *39*(1), 217–239. doi:10.1146/annurev.earth.031208.100209

Kultz, D. (2003). Evolution of the cellular stress proteome: from monophyletic origin to ubiquitous function. *Journal of Experimental Biology*, *206*(18), 3119–3124. doi:10.1242/jeb.00549

Kültz, D. (2005). Molecular and evolutionary basis of the cellular stress response. *Annual Review of Physiology*, *67*(1), 225–57. doi:10.1146/annurev.physiol.67.040403.103635

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. doi:10.1186/gb-2014-15-2-r29

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. doi:10.1186/1471-2105-12-323

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500. doi:10.1093/bioinformatics/btp692

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–89. doi:10.1101/gr.1224503

Lineweaver, C. H., Davies, P. C. W., & Vincent, M. D. (2014). Targeting cancer’s weaknesses (not its strengths): Therapeutic strategies suggested by the atavistic model. *BioEssays*, *36*(9), 827–35. doi:10.1002/bies.201400070

Ljungh, A., & Wadstrom, T. (2009). *Lactobacillus Molecular Biology: From Genomics to Probiotics*. Lund, Sweden: Horizon Scientific Press. Retrieved from http://books.google.com.au/books?id=d7GUXvQhxnoC

Loeb, L. A. (2001). A Mutator Phenotype in Cancer. *Cancer Research*, *61*(8), 3230–3239. doi:61:3230-3239

Loeb, L. A. (2011). Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Reviews Cancer*, *11*(6), 450–457. doi:10.1038/nrc3063

Loeb, L. A., Bielas, J. H., & Beckman, R. A. (2008). Cancers exhibit a mutator phenotype: clinical implications. *Cancer Research*, *68*(10), 3551–7. doi:10.1158/0008-5472.CAN-07-5835

López-Maury, L., Marguerat, S., & Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews. Genetics*, *9*(8), 583–93. doi:10.1038/nrg2398

Luo, J., Solimini, N. L., & Elledge, S. J. (2009). Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. *Cell*, *136*(5), 823–837. doi:10.1016/j.cell.2009.02.024.Principles

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, *21*(16), 3448–9. doi:10.1093/bioinformatics/bti551

Mani, S. a, Guo, W., Liao, M.-J., Eaton, E. N., Ayyanan, A., Zhou, A. Y., … Weinberg, R. a. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell*, *133*(4), 704–15. doi:10.1016/j.cell.2008.03.027

Martin, F., Kohler, A., Murat, C., Balestrini, R., Coutinho, P. M., Jaillon, O., … Wincker, P. (2010). Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature*, *464*(7291), 1033–8. doi:10.1038/nature08867

Mashima, T., Seimiya, H., & Tsuruo, T. (2009). De novo fatty-acid synthesis and related pathways as molecular targets for cancer therapy. *British Journal of Cancer*, *100*(9), 1369–72. doi:10.1038/sj.bjc.6605007

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–97. doi:10.1093/nar/gks042

Merlo, L. M. F., Pepper, J. W., Reid, B. J., & Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nature Reviews. Cancer*, *6*(12), 924–35. doi:10.1038/nrc2013

Michaud, J., Simpson, K. M., Escher, R., Buchet-Poyau, K., Beissbarth, T., Carmichael, C., … Scott, H. S. (2008). Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics*, *9*, 363. doi:10.1186/1471-2164-9-363

Mols, M., van Kranenburg, R., Tempelaars, M. H., van Schaik, W., Moezelaar, R., & Abee, T. (2010). Comparative analysis of transcriptional and physiological responses of Bacillus cereus to organic and inorganic acid shocks. *International Journal of Food Microbiology*, *137*(1), 13–21. doi:10.1016/j.ijfoodmicro.2009.09.027

Moseley, P. (2000). Stress proteins and the immune response. *Immunopharmacology*, *48*(3), 299–302. doi:10.1016/S0162-3109(00)00227-7

Murakami, S., Noguchi, T., Takeda, K., & Ichijo, H. (2007). Stress signaling in cancer. *Cancer Science*, *98*(10), 1521–7. doi:10.1111/j.1349-7006.2007.00551.x

NCBi Resource Coordinators. (2014). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *42*(Database issue), D7–17. doi:10.1093/nar/gkt1146

Nicholson, J. M., & Duesberg, P. (2009). On the karyotypic origin and evolution of cancer cells. *Cancer Genetics and Cytogenetics*, *194*(2), 96–110. doi:10.1016/j.cancergencyto.2009.06.008

Niklas, K. J. (2014). The evolutionary-developmental origins of multicellularity. *American Journal of Botany*, *101*(1), 6–25. doi:10.3732/ajb.1300314

Nishi, M., Sakai, Y., Akutsu, H., Nagashima, Y., Quinn, G., Masui, S., … Ryo, a. (2014). Induction of cells with cancer stem cell properties from nontumorigenic human mammary epithelial cells by defined reprogramming factors. *Oncogene*, *33*(5), 643–52. doi:10.1038/onc.2012.614

O’Connor, B. D., Merriman, B., & Nelson, S. F. (2010). SeqWare Query Engine: storing and searching sequence data in the cloud. *BMC Bioinformatics*, *11*(Suppl 12), S2. doi:10.1186/1471-2105-11-S12-S2

Pavelka, N., Rancati, G., Zhu, J., Bradford, W. D., Saraf, A., Florens, L., … Li, R. (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature*, *468*(7321), 321–5. doi:10.1038/nature09529

Podlaha, O., Riester, M., De, S., & Michor, F. (2012). Evolution of the cancer genome. *Trends in Genetics*, *28*(4), 155–63. doi:10.1016/j.tig.2012.01.003

Reya, T., Morrison, S. J., Clarke, M. F., & Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature*, *414*(6859), 105–11. doi:10.1038/35102167

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–40. doi:10.1093/bioinformatics/btp616

Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, *23*(21), 2881–7. doi:10.1093/bioinformatics/btm453

Robinson, M. D., & Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, *9*(2), 321–32. doi:10.1093/biostatistics/kxm030

Rokas, A. (2008). The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annual Review of Genetics*, *42*, 235–51. doi:10.1146/annurev.genet.42.110807.091513

Ruiz-Trillo, I., Burger, G., Holland, P. W. H., King, N., Lang, B. F., Roger, A. J., & Gray, M. W. (2007). The origins of multicellularity: a multi-taxon genome initiative. *Trends in Genetics*, *23*(3), 113–8. doi:10.1016/j.tig.2007.01.005

Sayers, E. W., Barrett, T., Benson, D. a, Bryant, S. H., Canese, K., Chetvernin, V., … Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *37*(Database issue), D5–15. doi:10.1093/nar/gkn741

Sestak, M. S., Božičević, V., Bakarić, R., Dunjko, V., & Domazet-Lošo, T. (2013). Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Frontiers in Zoology*, *10*(1), 18. doi:10.1186/1742-9994-10-18

Shah, A., Patel, S., Pathak, J., Swain, N., & Kumar, S. (2014). The evolving concepts of cancer stem cells in head and neck squamous cell carcinoma. *The Scientific World Journal*, *2014*. doi:10.1155/2014/842491

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–504. doi:10.1101/gr.1239303

Shor, E., Fox, C. a, & Broach, J. R. (2013). The yeast environmental stress response regulates mutagenesis induced by proteotoxic stress. *PLoS Genetics*, *9*(8), e1003680. doi:10.1371/journal.pgen.1003680

Skunca, N., Altenhoff, A., & Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology*, *8*(5), e1002533. doi:10.1371/journal.pcbi.1002533

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

Sonnhammer, E. L. L., & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, *18*(12), 619–620. doi:10.1016/S0168-9525(02)02793-2

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. a, Soanes, D. M., Stüber, K., … Panstruga, R. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, *330*(6010), 1543–6. doi:10.1126/science.1194573

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, *131*(5), 861–72. doi:10.1016/j.cell.2007.11.019

The Cancer Genome Atlas Research Network. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, *499*(7456), 43–9. doi:10.1038/nature12222

The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, *511*(7511), 543–550. doi:10.1038/nature13385

The Gene Ontology Consortium. (2000). Gene Ontology : tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556

Thiery, J. P. (2002). Epithelial-mesenchymal transitions in tumour progression. *Nature Reviews. Cancer*, *2*(6), 442–54. doi:10.1038/nrc822

Thiery, J. P., Acloque, H., Huang, R. Y. J., & Nieto, M. A. (2009). Epithelial-mesenchymal transitions in development and disease. *Cell*, *139*(5), 871–90. doi:10.1016/j.cell.2009.11.007

Tunlid, A., & Talbot, N. J. (2002). Genomics of parasitic and symbiotic fungi. *Current Opinion in Microbiology*, *5*(5), 513–9. doi:10.1016/S1369-5274(02)00355-7

Upcroft, P., & Upcroft, J. a. (2001). Drug targets and mechanisms of resistance in the anaerobic protozoa. *Clinical Microbiology Reviews*, *14*(1), 150–64. doi:10.1128/CMR.14.1.150-164.2001

Vander Heiden, M. G., Cantley, L. C., & Thompson, C. B. (2009). Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. *Science*, *324*(5930), 1029–1033. doi:10.1126/science.1160809

Vincent, A., & Van Seuningen, I. (2012). On the epigenetic origin of cancer stem cells. *Biochimica et Biophysica Acta*, *1826*(1), 83–8. doi:10.1016/j.bbcan.2012.03.009

Vincent, M. (2012). Cancer: A de-repression of a default survival program common to all cells?: A life-history perspective on the nature of cancer. *BioEssays*, *34*(1), 72–82. doi:10.1002/bies.201100049

Vincent, M. D. (2010). The Animal Within: Carcinogenesis and the Clonal Evolution of Cancer Cells Are Speciation Events Sensu Stricto. *Evolution*, *64*(4), 1173–1183. doi:10.1111/j.1558-5646.2009.00942.x

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., … Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, *38*(18), e178. doi:10.1093/nar/gkq622

Warburg, O. (1956). On the Origin of Cancer Cells. *Science*, *123*(3191), 309–314.

Wenger, Y., & Galliot, B. (2013). Punctuated emergences of genetic and phenotypic innovations in eumetazoan, bilaterian, euteleostome, and hominidae ancestors. *Genome Biology and Evolution*, *5*(10), 1949–68. doi:10.1093/gbe/evt142

World Health Organization. (2014). Cancer. Fact sheet N°297. Retrieved October 23, 2014, from http://www.who.int/mediacentre/factsheets/fs297/en/

Yampolsky, L. Y., Zeng, E., Lopez, J., Williams, P. J., Dick, K. B., Colbourne, J. K., & Pfrender, M. E. (2014). Functional genomics of acclimation and adaptation in response to thermal stress in Daphnia. *BMC Genomics*, *15*(1), 859. doi:10.1186/1471-2164-15-859

Yona, A. H., Manor, Y. S., Herbst, R. H., Romano, G. H., Mitchell, A., Kupiec, M., … Dahan, O. (2012). Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(51), 21010–5. doi:10.1073/pnas.1211150109

Yuan, J., Narayanan, L., Rockwell, S., & Glazer, P. M. (2000). Diminished DNA Repair and Elevated Mutagenesis in Mammalian Cells Exposed to Hypoxia and Low pH. *Cancer Research*, *60*(16), 4372–4376.

Zhao, Y., Butler, E. B., & Tan, M. (2013). Targeting cellular metabolism to improve cancer therapeutics. *Cell Death & Disease*, *4*(3), e532. doi:10.1038/cddis.2013.60

# Supplementary information

## Supplementary methods and results

### S1. RNAseq expression data

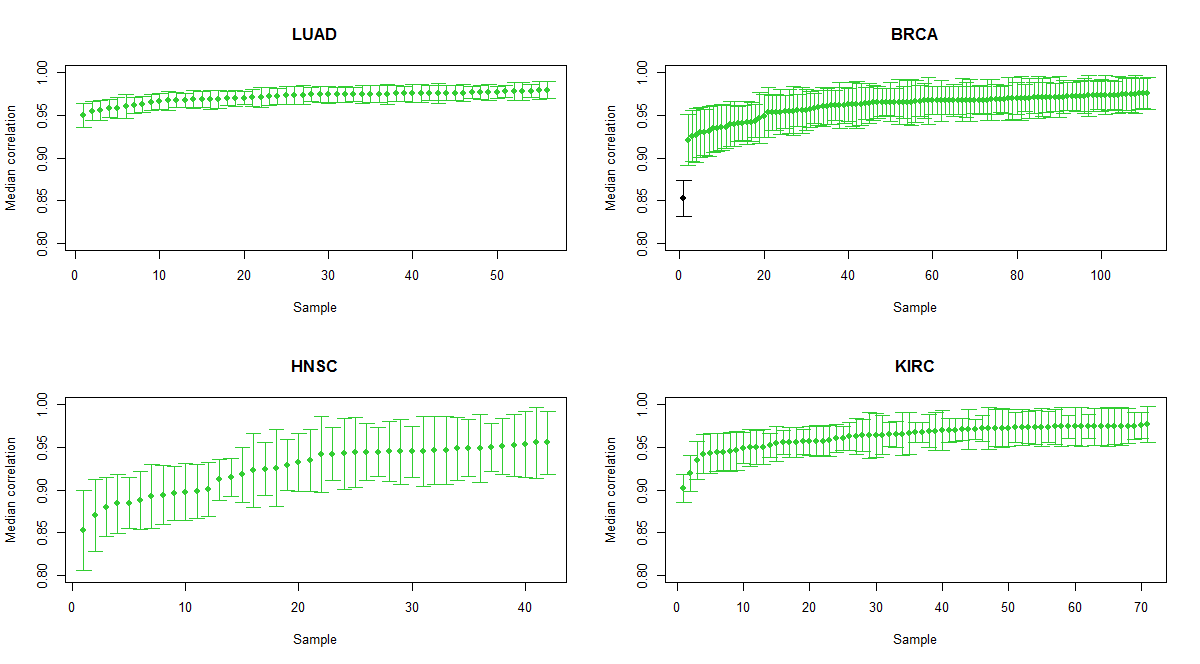
#### S1.1 Data collection

RNAseq data of LUAD, BRCA, HNSC and KIRC samples were downloaded from The Cancer Genome Atlas (TCGA). Metastatic samples were excluded from the dataset. The expression profiles were generated by the TCGA Research Network using their standard protocols. Briefly, the sequencing reads were obtained by an Illumina HiSeq sequencer, gene-abundance quantifications were generated by aligning the reads with MapSplice (Wang et al., 2010) and applying the RSEM algorithm (RNA-Seq by Expectation-Maximization) (Li & Dewey, 2011). Both normalized and raw RSEM values were initially downloaded. The normalized samples had been normalized with the upper quartile method on SeqWare (O’Connor, Merriman, & Nelson, 2010).

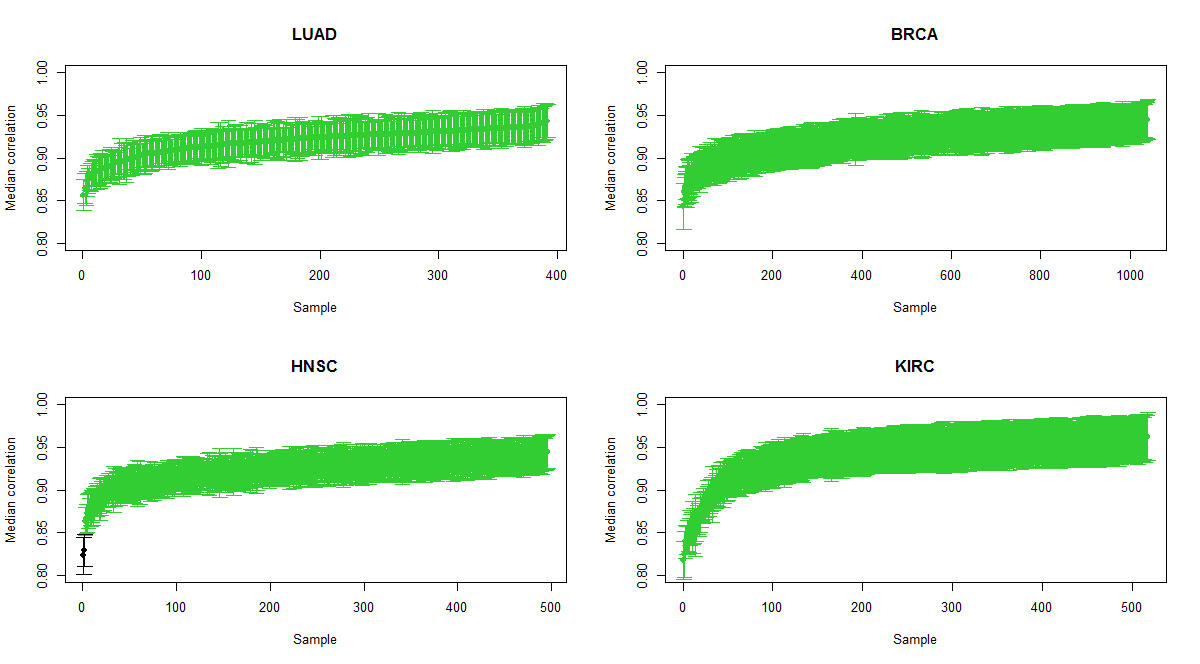
The expression values resulting from the RSEM algorithm were preferred over RPKM values because they account for multi-mapping reads, the values are comparable between samples and they take into account the effective read length (Li & Dewey, 2011; B. Li, Ruotti, Stewart, Thomson, & Dewey, 2010). However, both RSEM and RKPM have shown to produce similar and concordant results, both at the gene and exon levels (Guo et al., 2013). Additionally, the RSEM algorithm does not require a reference genome for mapping, which reduces biases that might occur when mapping to the human genome reference.

#### S1.2 Quality control

The quality of the cancer and control data sets was assessed by calculating the median Spearman correlation of each sample with all other samples of the same normal or tumour tissue type (***Figure S 1***, ***Figure S 2***). Raw unprocessed RSEM values were used for this step. One normal BRCA sample, and two tumour HNSC samples were removed. An additional HNSC normal sample was removed due to suspected contamination with another cell type. The TAI of this sample was 8.42, whereas the TAI of the normal and tumour samples of the four tumour samples ranged from 3 to 6, making the sample an evident outlier.



**Figure S 1. Median Spearman correlation of normal raw LUAD, BRCA, HNSC and KIRC samples.** Error bars represent the variance of the correlations between samples. The median correlations of all normal samples were above 0.8, suggesting homogeneous samples. However, the BRCA sample (black) whose median correlation was not within the error bars of the next closest sample was removed due to the evident dissimilarity with the other BRCA samples.



**Figure S 2. Median Spearman correlation of tumour LUAD, BRCA, HNSC and KIRC samples.** Error bars represent the variance of the correlations between samples. Similar to the case of normal samples, the median correlations were above 0.8. However, two HNSC samples (black) whose median correlations were not within the error bars of the next closest sample were removed due to the evident dissimilarity with the other HNSC samples.

### S2. Estimation of expression fold changes and identification of differentially expressed genes (DEG)

limma-voom version 3.20.5 (Smyth, 2005) was used to estimate the log fold changes and find DEG of the BRCA, LUAD, HNSC and KIRC samples. limma uses linear models and empirical Bayes models to find DEGs. Additionally, the voom function (Law et al., 2014) was used to estimate the mean-variance relationship from the data, which was incorporated as a weight for each observation.

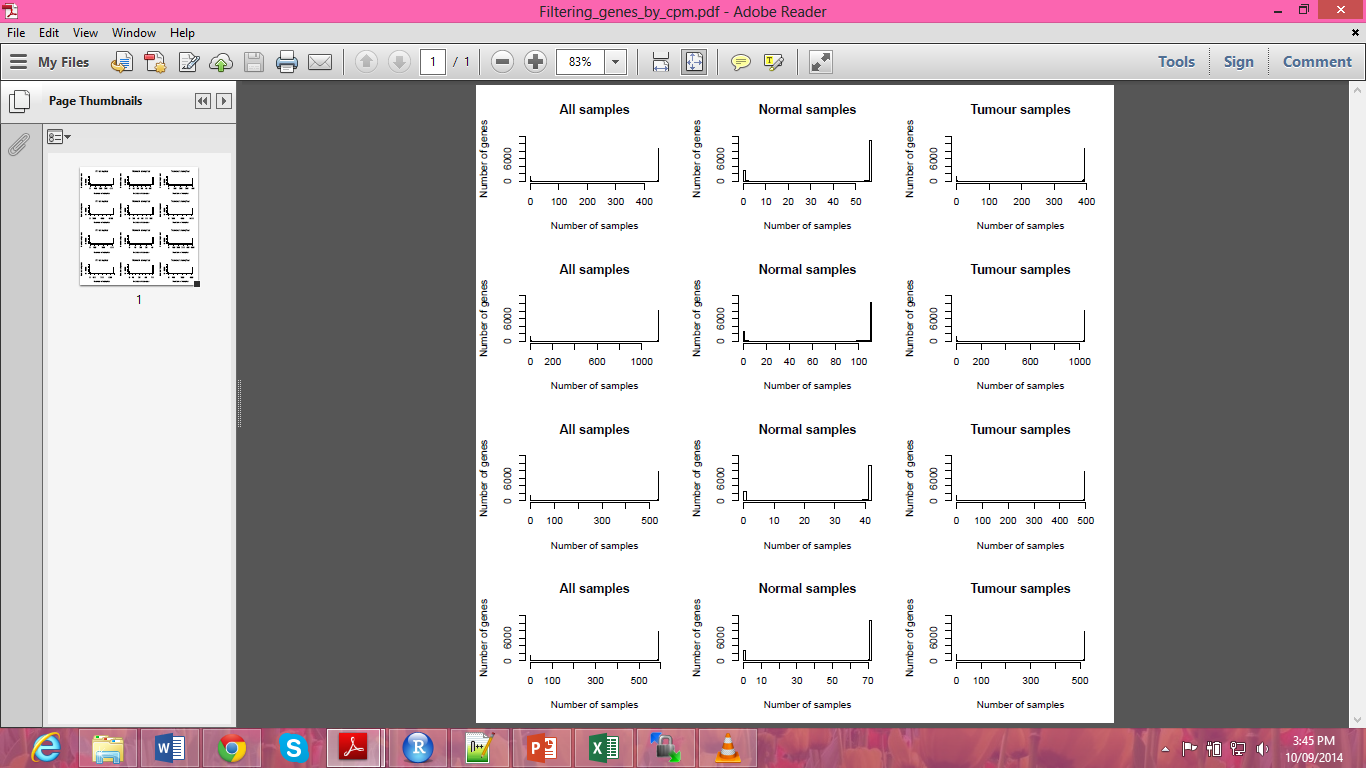
#### S2.1 Choice of input data

Raw RSEM values were preferred as input to limma, followed by a built-in normalization in limma to ensure compatibility of the data with the statistical assumptions of limma and voom. Even so, preliminary tests with the TCGA-normalized RSEM values resulted in very similar results (*data not shown*).

#### S2.2 Filtering out genes with low expression values

Genes with counts per million (cpm) less than 1 were considered to not be sufficiently highly expressed, as suggested by the limma manual. However, it was necessary to determine a threshold number of samples in which a gene had to fail this criteria for it to be removed. The distribution of the number of genes with a cpm greater than 1 against the number of samples was determined for tumour and normal samples, as well as the total sum of samples, in order to find appropriate thresholds (***Figure S 3***).

It is evident from the distributions taking into account the combined number of tumour and normal samples that most genes have a cpm>1 in all samples. However, equivalent distributions taking into account normal and tumour samples separately were investigated due to possibility of marked differences between the levels of expression of genes in normal and tumour samples. There might have been highly expressed genes in normal samples that were switched off in tumour cells, or vice versa, and would consequently be removed from the data set if only their overall expression across all samples (tumour and normal) was considered. These genes are of high importance because they would have a large variability between cancer and normal samples. Most genes have an expression level of cpm>1 in all tumour or normal samples. Finally, genes that had a cpm>1 in all samples, in all normal or in all tumour samples were kept (***Table* S *1***). Around 60% of the genes of the four tumour types survived filtering.



**a.**

**b.**

**c.**

**d.**

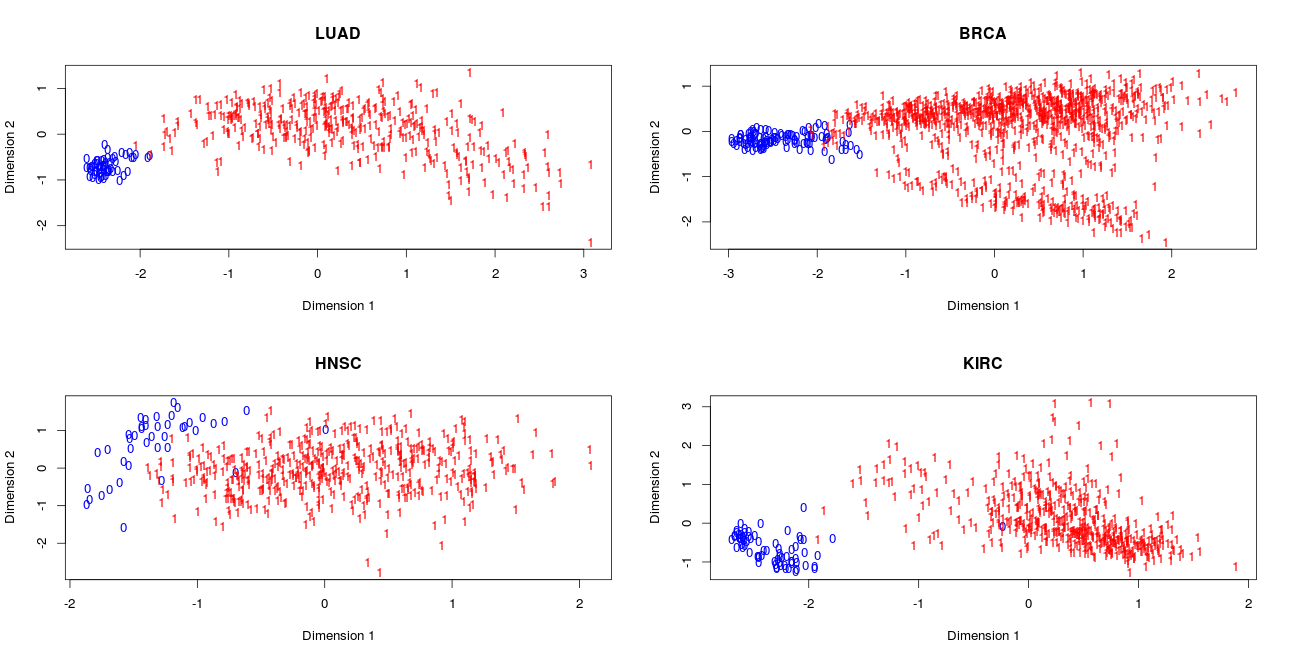
**Figure S 3. Distribution of the number of genes with cpm greater than 1 in all, normal or tumour samples for LUAD, BRCA, HNSC and KIRC samples.** a. LUAD. b. BRCA. c. HNSC. d. KIRC. Most genes have a cpm>1 in all samples, as evidenced by the righter-most peak in all plots.

**Table S 1. Number of genes that survived filtering.** Genes with cpm>1 in all tumour or normal samples were kept. The initial number of genes was 16736.

|  |  |  |
| --- | --- | --- |
| **Cancer type** | **Number of genes after filtering** | **Percentage of genes surviving filtering (%)** |
| **LUAD** | 10974 | 65.57 |
| **BRCA** | 10362 | 61.91 |
| **HNSC** | 9398 | 56.15 |
| **KIRC** | 10639 | 63.57 |

#### S2.3 MDS diagnostic plots

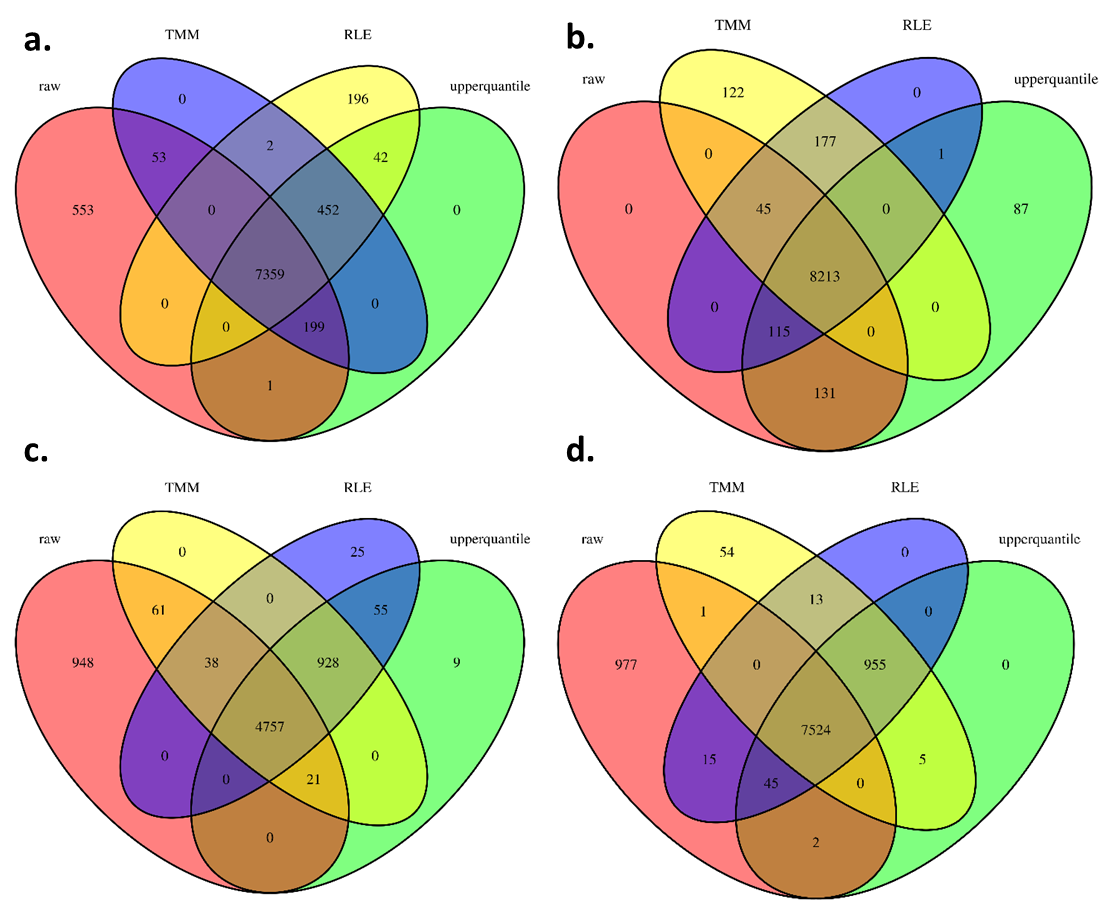
MDS plots were used to rule out strong batch effects and ensure the clustering of normal and tumour samples (***Figure S 4***). Normal and tumour samples were clearly separated in the first dimension, suggesting that most of the variation of the data was due to the normal/tumour state of the tissue, rather than batch effects. The tight clustering in a small area of the normal samples is a validation of their homogeneity, whereas tumour samples present more biological variation.



**Figure S 4. MDS plots of LUAD, BRCA, HNSC and KIRC samples.** Blue zeroes represent normal samples, red ones represent tumour samples. For each tissue type, the normal and tumour samples are roughly clustered together and are separated in the first dimension, discarding strong batch effects. The heterogeneity of tumour samples is evidenced by the larger horizontal spread of the samples in comparison to the normal counterparts. The normal HNSC samples are the least tightly clustered.

#### S2.4 Comparison of normalization methods

Limma offers three different normalization methods: TMM (Trimmed Mean of M values), RLE (Relative Log Expression values) and upperquantile , which were separately applied to the data. The overlap between the DEG detected with the three methods, as well as with no normalization, was determined for the four cancer types (**Figure S *5***). A gene was said to be differentially expressed if its adjusted p-value was less than 0.05. We found a great overlap between the DEG obtained.



**Figure S 5. Overlap of the DEG obtained after different normalization methods.** TMM, RLE and upperquantile normalization methods were applied to the data. No normalization was also tested. a. LUAD, b. BRCA, c. HNSC and d. KIRC samples. There is large overlap between the DEG identified with the different normalization methods in the four cancer types (between 4757 and 8213 genes). There was a larger number of uniquely identified DEG when no normalization was used, which probably represents false positives. The largest DEG overlap occurred between the upperquantile and RLE normalization methods, suggesting that both methods give nearly equivalent results.

TMM assumes that most genes are not differentially expressed (Dillies et al., 2013) and RLE assumes that the number of replicate samples is small (Anders & Huber, 2010), but the upperquantile method does not make these assumptions (Bullard, Purdom, Hansen, & Dudoit, 2010), making it the preferred choice considering that a large number of genes could be detected as being differentially expressed due to the large sample sizes.

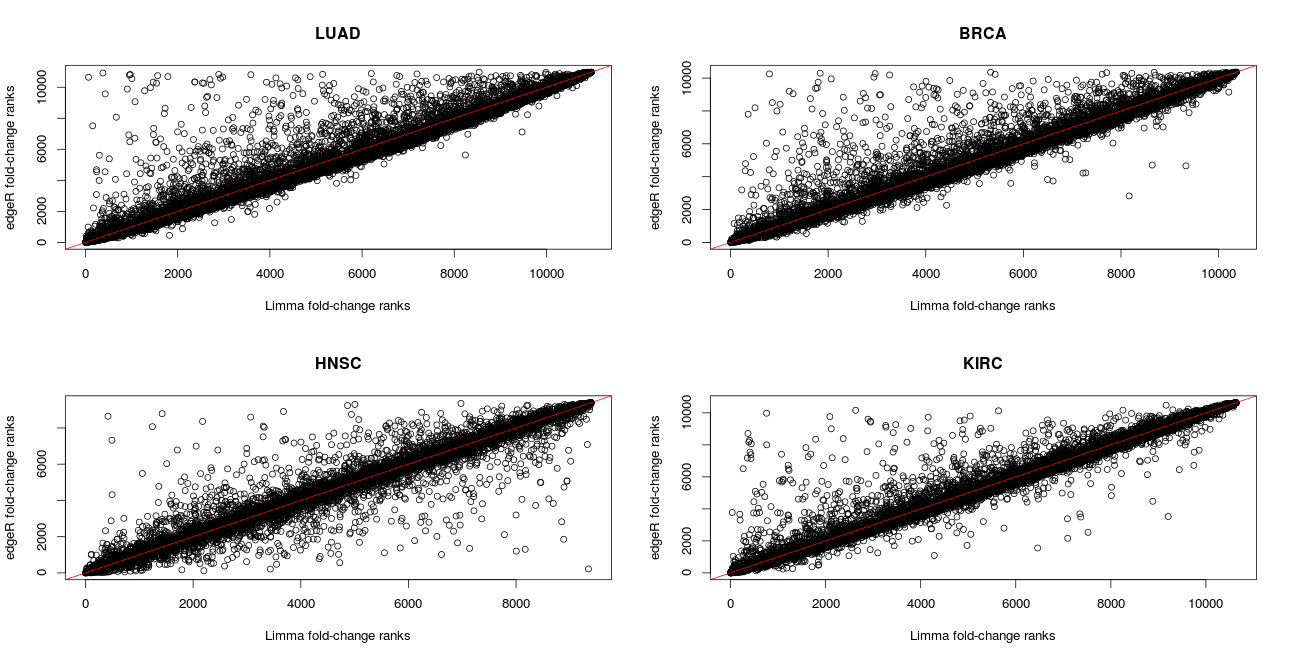
Finally, the number of DEG detected in each tumour type is in Table S *6*. A high percentage of the genes that survived cpm filtering were DE (61-82%), which represents 34-51% of all genes.

**Table S 2. Number of DEG obtained for each tumour type and the percentage that they represent from the total number of genes that survived the cpm filtering and the initial number of genes.** The TCGA raw RSEM values were upperquantile normalized after filtering according to their cpm. A gene was considered to be differentially expressed if its adjusted p-value was lower than 0.05.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DEG** | **Percentage of filtered genes that are DE (%)** | **Percentage of total genes that are DE (%)** |
| **LUAD** | 8131 | 74.09 | 48.58 |
| **BRCA** | 8547 | 82.48 | 51.07 |
| **HNSC** | 5770 | 61.40 | 34.48 |
| **KIRC** | 8531 | 80.19 | 50.97 |

#### S2.5 Correlation between the DEG obtained with limma and edgeR

We aimed to compare the limma results with those obtained with another R package to obtain DEG, edgeR, version 3.5.29, which analyses count-based expression data by assuming that the data follows a negative binomial distribution rather than the linear models used by limma *(McCarthy, Chen, & Smyth, 2012; Robinson, McCarthy, & Smyth, 2010; Robinson & Smyth, 2007, 2008)*. The same cpm filtering criteria and upperquantile normalization were used. The genes were ranked by their fold change estimated by either package, and the ranks were compared (***Figure* S *6***). There was high correlation (>0.96) between the ranks across the four tumour types, suggesting that the results obtained with limma are in agreement with those obtained with edgeR. The limma package was used hereafter, given that the processing time was smaller than for edgeR.



ρ = 0.963

ρ = 0.966

ρ = 0.969

ρ = 0.976

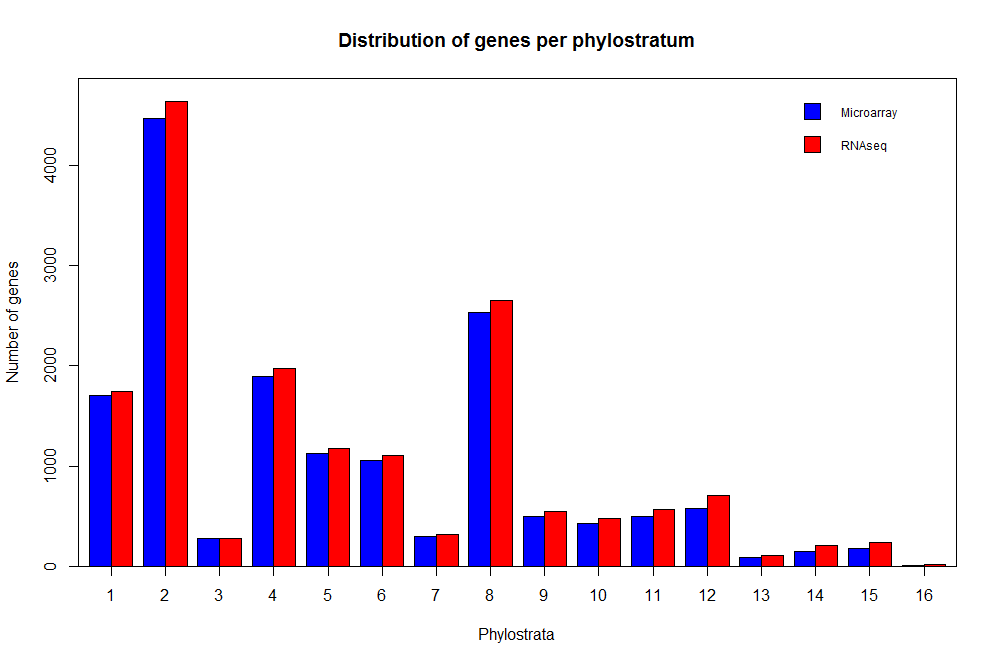
**Figure S 6. Comparison of the fold-change ranks of genes obtained with limma and edgeR.** For each tumour type, the genes were ranked by the fold changes estimated by limma and edgeR. The Spearman correlation coefficient between the ranks was calculated, which shows a consistently high correlation between the ranks.

### S3. Comparing the age of the tumour and normal transcriptomes using microarray data

The TAI has been traditionally calculated using microarray, rather than RNAseq data (Domazet-Lošo & Tautz, 2010). Therefore, we aimed to reproduce our results using TCGA microarray data of BRCA samples.

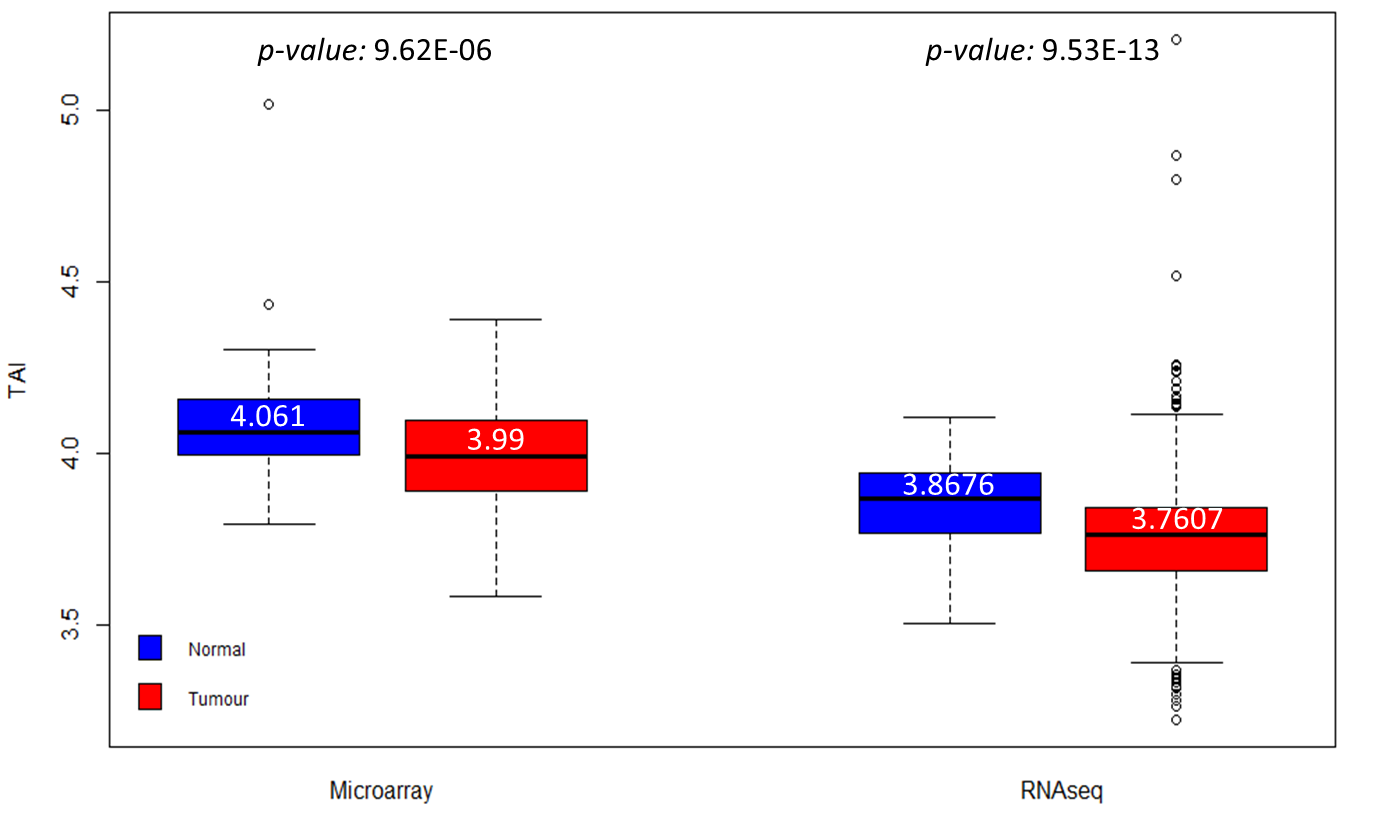
Agilent level 1 raw microarray intensities of 530 tumour and 63 normal samples were downloaded from TCGA. Limma was used to read in the entire raw data simultaneously, the “normexp” background correction method was applied, with an offset of 16. The red channel was subsequently quantile normalized between arrays, and used hereafter. The median intensity of the probes mapping to each gene was used to represent the expression value of the gene. Probes that belong to multiple genes were excluded from the analysis.

The assignment of genes to phylostrata was carried out as previously explained in the Methods section. A total of 15764 genes overlapped between the microarray platform and the genes available in OrthoMCL. The distribution of genes assigned to the phylostrata closely follows the distribution obtained for the RNAseq platform (**Figure S 7**).



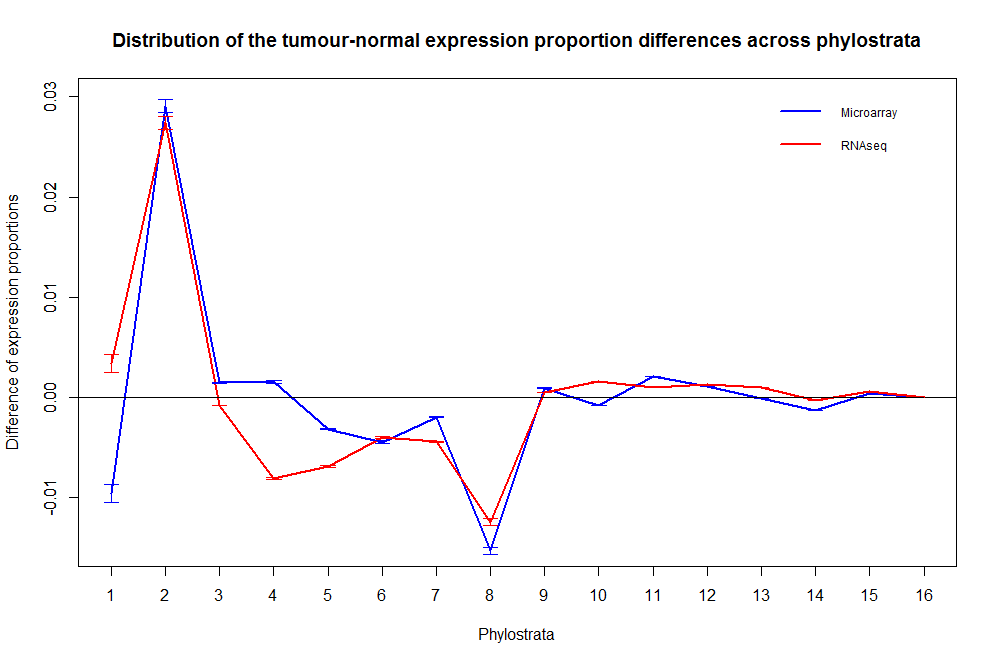
**Figure S 7. Distribution of the number of genes assigned to each phylostratum using the microarray and RNAseq TCGA data.** The distribution obtained with the genes available on the microarray platform closely resembles the one previously found for the RNAseq analysis, suggesting that there are not new biases in the number of genes assigned to the phylostrata.

Next, the TAI of the normal and tumour BRCA samples were calculated using the microarray expression values. The TAI of tumour samples was lower than the TAI of normal samples, and the ranges of the TAI were similar to those obtained with the RNAseq data (**Figure S *8***), suggesting that the results obtained with the TAI are independent of the technology used to measure the expression levels.



**Figure S 8. TAI of normal and tumour BRCA samples calculated from microarray and RNAseq data.** The microarray TAI of tumour samples is lower than the normal samples, once again suggesting an older transcriptome. Overall, the results obtained with the microarray expression values are analogous to those obtained with the RNAseq data. The white numbers are the median TAI. Two-sided p-values obtained from two-sided Wilcoxon tests (shown) were used to determine if the tumour samples had a different mean TAI than normal samples.

We calculated the difference in expression proportions of the phylostrata with the same method used for the RNAseq data, and plotted both results for comparison (**Figure S *9*** & **Table S *3***). The overall distribution generated with the microarray data closely resembles the one generated with RNAseq data, but with alterations in the expression proportion differences of Phylostratum 1, 4 and 5.



**Figure S 9. Difference in expression proportions of phylostrata from BRCA samples, using RNAseq and microarray data.** The expression proportions obtained with both technologies follow a similar distribution, except for Phylostratum 1, 4 and 5.

**Table S 3. Wilcoxon two-sided p-values for the difference of the means of the BRCA tumour and normal expression proportions obtained with microarray data.** Wilcoxon one-sided p-values were used to corroborate the direction of the change. P-values were considered to be significant if p < 0.05.

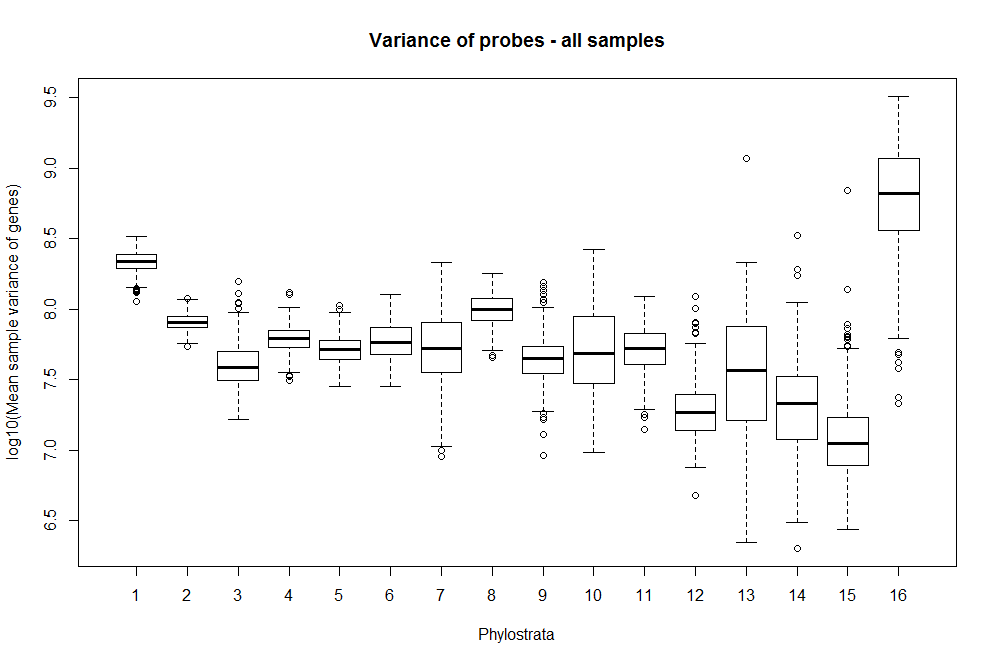
|  |  |
| --- | --- |
| **Phylostratum** | **BRCA-microarray p-value** |
| **1** | 1.72E-09 |
| **2** | 4.95E-32 |
| **3** | 4.40E-17 |
| **4** | 0.000357134 |
| **5** | 2.62E-13 |
| **6** | 1.96E-15 |
| **7** | 2.84E-18 |
| **8** | 2.20E-18 |
| **9** | 0.00180192 |
| **10** | 0.09561027 |
| **11** | 1.82E-08 |
| **12** | 4.06E-05 |
| **13** | 0.193281 |
| **14** | 2.60E-08 |
| **15** | 5.19E-13 |
| **16** | 0.04431486 |

\*Pink =higher expression proportion in tumour samples

\*Blue=lower expression proportion in tumour samples

\*White=no significant difference between normal and tumour samples

We suspected that an elevated variance between the probes of the genes of phylostrata 1, 4 and 5 might explain the previous discrepancies. The tumour and normal samples were pooled, and the variance between the probes of each gene was calculated for each sample. The mean variance across genes was subsequently determined for each phylostratum (**Figure S *10***). The variances of the genes of phylostratum 1 were much higher than the rest, which might have contributed to the discrepancy previously observed for this phylostratum. However, the genes of phylostrata 4 and 5 did not have evidently higher variances than the rest of the phylostrata, suggesting that there are other factors that affected the expression proportions of these phylostrata.



**Figure S 10. Mean expression variance of the genes across phylostrata.** Boxplots were built with the mean variance of the gene probes for each sample. Normal and tumour samples were pooled for this representation. The variance of the probes of Phylostratum 1 is higher than the variance of the rest of the phylostrata. Evident increased variances were not observed for Phylostrata 4 and 5. The large variance present in genes of phylostrata 16 is most likely due to the very small number of genes in this phylostratum.

## Supplementary tables

Table S 4. Phylostrata of OrthoMCL species

|  |  |  |
| --- | --- | --- |
| **ABBREV** | **Species** | **Phylostratum** |
| ecol | Escherichia coli str. K-12 substr. W3110 | 1 |
| ftul | Francisella tularensis subsp. tularensis SCHU S4 | 1 |
| gsul | Geobacter sulfurreducens PCA | 1 |
| halo | Halobacterium sp. NRC-1 | 1 |
| hwal | Haloquadratum walsbyi DSM 16790 | 1 |
| hbut | Hyperthermus butylicus DSM 5456 | 1 |
| ihos | Ignicoccus hospitalis KIN4/I | 1 |
| lmon | Listeria monocytogenes EGD-e | 1 |
| msed | Metallosphaera sedula DSM 5348 | 1 |
| msmi | Methanobrevibacter smithii ATCC 35061 | 1 |
| mjan | Methanocaldococcus jannaschii DSM 2661 | 1 |
| mmar | Methanococcus maripaludis S2 | 1 |
| tvol | Thermoplasma volcanium GSS1 | 1 |
| tmar | Thermotoga maritima MSB8 | 1 |
| tpal | Treponema pallidum subsp. pallidum str. Nichols | 1 |
| vcho | Vibrio cholerae O1 biovar El Tor str. N16961 | 1 |
| wend | Wolbachia endosymbiont of Culex quinquefasciatus Pel | 1 |
| wsuc | Wolinella succinogenes DSM 1740 | 1 |
| yent | Yersinia enterocolitica subsp. enterocolitica 8081 | 1 |
| ypes | Yersinia pestis CO92 | 1 |
| mlep | Mycobacterium leprae Br4923 | 1 |
| mtub | Mycobacterium tuberculosis H37Rv | 1 |
| nequ | Nanoarchaeum equitans Kin4-M | 1 |
| nmar | Nitrosopumilus maritimus SCM1 | 1 |
| rsol | Ralstonia solanacearum GMI1000 | 1 |
| rbal | Rhodopirellula baltica SH 1 | 1 |
| rpro | Rickettsia prowazekii str. Madrid E | 1 |
| rtyp | Rickettsia typhi str. Wilmington | 1 |
| sent | Salmonella enterica subsp. enterica serovar Typhi str. CT18 | 1 |
| sfle | Shigella flexneri 2a str. 301 | 1 |
| saur | Staphylococcus aureus subsp. aureus Mu50 | 1 |
| smar | Staphylothermus marinus F1 | 1 |
| spne | Streptococcus pneumoniae TIGR4 | 1 |
| ssol | Sulfolobus solfataricus P2 | 1 |
| syne | Synechococcus sp. WH 8102 | 1 |
| atum | Agrobacterium tumefaciens str. C58 | 1 |
| aaeo | Aquifex aeolicus VF5 | 1 |
| aful | Archaeoglobus fulgidus DSM 4304 | 1 |
| bant | Bacillus anthracis str. 'Ames Ancestor' | 1 |
| bsui | Brucella suis 1330 | 1 |
| bmal | Burkholderia mallei ATCC 23344 | 1 |
| bpse | Burkholderia pseudomallei 1710b | 1 |
| cmaq | Caldivirga maquilingensis IC-167 | 1 |
| cjej | Campylobacter jejuni subsp. jejuni NCTC 11168 | 1 |
| ckor | Candidatus Korarchaeum cryptofilum OPF8 | 1 |
| cpne | Chlamydophila pneumoniae CWL029 | 1 |
| ctep | Chlorobium tepidum TLS | 1 |
| cbot | Clostridium botulinum A3 str. Loch Maree | 1 |
| cper | Clostridium perfringens str. 13 | 1 |
| cbur | Coxiella burnetii RSA 493 | 1 |
| deth | Dehalococcoides ethenogenes 195 | 1 |
| drad | Deinococcus radiodurans R1 | 1 |
| ddis | Dictyostelium discoideum AX4 | 2 |
| edis | Entamoeba dispar SAW760 | 2 |
| ehis | Entamoeba histolytica HM-1:IMSS | 2 |
| einv | Entamoeba invadens IP1 | 2 |
| glab | Giardia intestinalis ATCC 50581 | 2 |
| glam | Giardia lamblia ATCC 50803 | 2 |
| glae | Giardia lamblia P15 | 2 |
| gthe | Guillardia theta | 2 |
| lbra | Leishmania braziliensis | 2 |
| linf | Leishmania infantum | 2 |
| lmaj | Leishmania major strain Friedlin | 2 |
| lmex | Leishmania Mexicana | 2 |
| micr | Micromonas sp. RCC299 | 2 |
| tpse | Thalassiosira pseudonana CCMP1335 | 2 |
| tann | Theileria annulata strain Ankara | 2 |
| tpar | Theileria parva strain Muguga | 2 |
| tgon | Toxoplasma gondii | 2 |
| tvag | Trichomonas vaginalis G3 | 2 |
| tbru | Trypanosoma brucei | 2 |
| tbrg | Trypanosoma brucei gambiense | 2 |
| tcon | Trypanosoma congolense | 2 |
| tcru | Trypanosoma cruzi strain CL Brener | 2 |
| tviv | Trypanosoma vivax | 2 |
| vcar | Volvox carteri f. nagariensis | 2 |
| ncan | Neospora caninum | 2 |
| osat | Oryza sativa Japonica Group | 2 |
| otau | Ostreococcus tauri | 2 |
| ppat | Physcomitrella patens subsp. patens | 2 |
| pram | Phytophthora ramorum | 2 |
| pber | Plasmodium berghei str. ANKA | 2 |
| pcha | Plasmodium chabaudi chabaudi | 2 |
| pfal | Plasmodium falciparum 3D7 | 2 |
| pkno | Plasmodium knowlesi strain H | 2 |
| pviv | Plasmodium vivax SaI-1 | 2 |
| pyoe | Plasmodium yoelii yoelii str. 17XNL | 2 |
| rcom | Ricinus communis | 2 |
| tthe | Tetrahymena thermophila SB210 | 2 |
| atha | Arabidopsis thaliana | 2 |
| bbov | Babesia bovis T2Bo | 2 |
| crei | Chlamydomonas reinhardtii | 2 |
| chom | Cryptosporidium hominis TU502 | 2 |
| cmur | Cryptosporidium muris RN66 | 2 |
| cpar | Cryptosporidium parvum Iowa II | 2 |
| cmer | Cyanidioschyzon merolae strain 10D | 2 |
| anid | Emericella nidulans | 3 |
| ecun | Encephalitozoon cuniculi GB-M1 | 3 |
| eint | Encephalitozoon intestinalis | 3 |
| ebie | Enterocytozoon bieneusi | 3 |
| egos | Eremothecium gossypii | 3 |
| gzea | Gibberella zeae PH-1 | 3 |
| klac | Kluyveromyces lactis NRRL Y-1140 | 3 |
| lbic | Laccaria bicolor S238N-H82 | 3 |
| mgri | Magnaporthe oryzae 70-15 | 3 |
| ylip | Yarrowia lipolytica CLIB122 | 3 |
| mbre | Monosiga brevicollis MX1 | 3 |
| ncra | Neurospora crassa OR74A | 3 |
| pchr | Phanerochaete chrysosporium | 3 |
| scer | Saccharomyces cerevisiae S288c | 3 |
| psti | Scheffersomyces stipitis CBS 6054 | 3 |
| spom | Schizosaccharomyces pombe | 3 |
| afum | Aspergillus fumigatus Af293 | 3 |
| aory | Aspergillus oryzae RIB40 | 3 |
| calb | Candida albicans | 3 |
| cgla | Candida glabrata CBS 138 | 3 |
| cimm | Coccidioides immitis RS | 3 |
| cpos | Coccidioides posadasii RMSCC 3488 | 3 |
| cneo | Cryptococcus bacillisporus | 3 |
| cneg | Cryptococcus neoformans var. grubii H99 | 3 |
| dhan | Debaryomyces hansenii CBS767 | 3 |
| tadh | Trichoplax adhaerens | 4 |
| nvec | Nematostella vectensis | 5 |
| dmel | Drosophila melanogaster | 6 |
| isca | Ixodes scapularis | 6 |
| sman | Schistosoma mansoni | 6 |
| apis | Acyrthosiphon pisum | 6 |
| aaeg | Aedes aegypti | 6 |
| agam | Anopheles gambiae str. PEST | 6 |
| amel | Apis mellifera | 6 |
| bmor | Bombyx mori | 6 |
| bmaa | Brugia malayi | 6 |
| cbri | Caenorhabditis briggsae AF16 | 6 |
| cele | Caenorhabditis elegans | 6 |
| cpip | Culex pipiens | 6 |
| phum | Pediculus humanus | 6 |
| cint | Ciona intestinalis | 7 |
| trub | Takifugu rubripes | 8 |
| tnig | Tetraodon nigroviridis | 8 |
| drer | Danio rerio | 8 |
| ggal | Gallus gallus | 9 |
| oana | Ornithorhynchus anatinus | 10 |
| mdom | Monodelphis domestica | 11 |
| ecab | Equus caballus | 12 |
| clup | Canis lupus familiaris | 12 |
| mmus | Mus musculus | 13 |
| rnor | Rattus norvegicus | 13 |
| mmul | Macaca mulatta | 14 |
| ptro | Pan troglodytes | 15 |
| hsap | Homo sapiens | 16 |

Table S 5. Species used to date GO-terms.

|  |  |
| --- | --- |
| **Phylostratum** | **Species** |
| 1 | Escherichia coli |
| 1 | Pseudomonas aeruginosa PAO1 |
| 1 | Dickeya dadantii |
| 1 | Agrobacterium tumefaciens str C58 |
| 1 | Anaplasma phagocytophilum HZ |
| 1 | Bacillus anthracis Ames |
| 1 | Campylobacter jejuni RM1221 |
| 1 | Carboxydothermus hydrogenoformans Z-2901 |
| 1 | Clostridium perfringens ATCC13124 |
| 1 | Colwellia psychrerythraea 34H |
| 1 | Coxiella burnetii RSA 493 |
| 1 | Dehalococcoides ethenogenes 195 |
| 1 | Ehrlichia chaffeensis Arkansas |
| 1 | Geobacter sulfurreducens PCA |
| 1 | Hyphomonas neptunium ATCC 15444 |
| 1 | Listeria monocytogenes 4b F2365 |
| 1 | Methylococcus capsulatus Bath |
| 1 | Neorickettsia sennetsu Miyayama |
| 1 | Pseudomonas fluorescens Pf-5 |
| 1 | Pseudomonas syringae DC3000 |
| 1 | Pseudomonas syringae pv. phaseolicola 1448A |
| 1 | Shewanella oneidensis MR-1 |
| 1 | Silicibacter pomeroyi DSS-3 |
| 1 | Vibrio cholerae El Tor N16961 |
| 2 | Trypanosoma brucei |
| 2 | Plasmodium falciparum |
| 2 | Leishmania major |
| 3 | Schizosaccharomyces pombe |
| 3 | Saccharomyces cerevisiae |
| 4 | Trichoplax adhaerens |
| 5 | Nematostella vectensis |
| 5 | Hydra vulgaris |
| 6 | Caenorhabditis elegans |
| 6 | Drosophila melanogaster |
| 6 | Aedes aegypti |
| 7 | Ciona intestinalis |
| 8 | Danio rerio |
| 8 | Xenopus laevis |
| 9 | Gallus gallus |
| 10 | Ornithorhynchus anatinus |
| 11 | Monodelphis domestica |
| 12 | Equus caballus |
| 12 | Sus scrofa |
| 12 | Canis lupus familiaris |
| 12 | Bos Taurus |
| 13 | Cricetulus griseus |
| 13 | Rattus novergicus |
| 13 | Mus musculus |
| 14 | Macaca mulatta |
| 15 | Pan troglodytes |
| 16 | Homo sapiens |

Table S 6. Gene sets that were found to undergo a process of atavism in at least two cancer types

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Umbrella** | **Set** | **LUAD** | **BRCA** | **HNSC** | **KIRC** |
| **Amino acid metabolism** | carboxylic acid metabolic process | Yes | Yes | No | No |
| monocarboxylic acid metabolic process | Yes | Yes | No | No |
| peptidyl-amino acid modification | Yes | Yes | No | No |
| **Carbohydrate metabolism** | cellular carbohydrate biosynthetic process | Yes | Yes | No | No |
| regulation of carbohydrate biosynthetic process | Yes | Yes | No | No |
| regulation of polysaccharide biosynthetic process | Yes | Yes | No | No |
| regulation of polysaccharide metabolic process | Yes | Yes | No | No |
| **Cell cycle** | negative regulation of cytoskeleton organization | Yes | Yes | Yes | No |
| cell cycle arrest | Yes | Yes | No | No |
| G1/S transition of mitotic cell cycle | Yes | Yes | No | No |
| G2/M transition of mitotic cell cycle | Yes | Yes | No | No |
| positive regulation of cell cycle process | Yes | Yes | No | No |
| positive regulation of cytoskeleton organization | Yes | Yes | No | No |
| regulation of actin cytoskeleton organization | Yes | Yes | No | No |
| **Cell death** | intrinsic apoptotic signaling pathway | Yes | Yes | No | No |
| negative regulation of programmed cell death | Yes | Yes | No | No |
| positive regulation of programmed cell death | Yes | Yes | No | No |
| regulation of apoptotic process | Yes | Yes | No | No |
| **Cell morphology** | cell projection morphogenesis | Yes | Yes | No | No |
| **Cell motility** | cell chemotaxis | Yes | Yes | No | No |
| **Cell response to stress** | cellular response to hydrogen peroxide | Yes | Yes | No | No |
| **DNA damage** | DNA damage checkpoint | Yes | Yes | Yes | No |
| regulation of response to DNA damage stimulus | Yes | Yes | No | No |
| **DNA metabolism** | DNA metabolic process | Yes | Yes | No | No |
| regulation of DNA metabolic process | Yes | Yes | No | No |
| **Energy** | glucose transport | Yes | Yes | Yes | No |
| glucose metabolic process | Yes | Yes | No | No |
| **Ion metabolism** | cation homeostasis | Yes | Yes | No | No |
| cellular cation homeostasis | Yes | Yes | No | No |
| cellular ion homeostasis | Yes | Yes | No | No |
| **Lipid metabolism** | glycolipid biosynthetic process | Yes | Yes | No | No |
| glycolipid metabolic process | Yes | Yes | No | No |
| sphingolipid metabolic process | Yes | Yes | No | No |
| **Membrane biogenesis** | membrane lipid biosynthetic process | Yes | Yes | No | No |
| **Metabolism** | hexose transport | Yes | Yes | Yes | No |
| cellular protein catabolic process | Yes | Yes | No | No |
| cellular protein complex assembly | Yes | Yes | No | No |
| glycosaminoglycan catabolic process | Yes | Yes | No | No |
| hexose biosynthetic process | Yes | Yes | No | No |
| hexose metabolic process | Yes | Yes | No | No |
| modification-dependent macromolecule catabolic process | Yes | Yes | No | No |
| modification-dependent protein catabolic process | Yes | Yes | No | No |
| negative regulation of cellular biosynthetic process | Yes | Yes | No | No |
| negative regulation of cellular catabolic process | Yes | Yes | No | No |
| negative regulation of macromolecule biosynthetic process | Yes | Yes | No | No |
| positive regulation of cellular biosynthetic process | Yes | Yes | No | No |
| positive regulation of cellular protein metabolic process | Yes | Yes | No | No |
| positive regulation of hydrolase activity | Yes | Yes | No | No |
| positive regulation of macromolecule biosynthetic process | Yes | Yes | No | No |
| positive regulation of phosphorus metabolic process | Yes | Yes | No | No |
| regulation of cellular macromolecule biosynthetic process | Yes | Yes | No | No |
| regulation of phosphate metabolic process | Yes | Yes | No | No |
| response to hexose | Yes | Yes | No | No |
| positive regulation of transferase activity | Yes | Yes | No | No |
| cellular protein complex disassembly | Yes | No | No | Yes |
| cyclic nucleotide catabolic process | No | Yes | No | Yes |
| **Nuclear transport** | nuclear export | Yes | Yes | No | No |
| nuclear import | Yes | Yes | No | No |
| **Nucleotide/nucleoside metabolism** | nucleoside triphosphate catabolic process | Yes | Yes | Yes | No |
| nucleoside catabolic process | Yes | Yes | No | No |
| nucleoside metabolic process | Yes | Yes | No | No |
| nucleoside phosphate biosynthetic process | Yes | Yes | No | No |
| nucleoside phosphate catabolic process | Yes | Yes | No | No |
| nucleoside phosphate metabolic process | Yes | Yes | No | No |
| nucleoside triphosphate metabolic process | Yes | Yes | No | No |
| nucleotide biosynthetic process | Yes | Yes | No | No |
| nucleotide catabolic process | Yes | Yes | No | No |
| nucleotide metabolic process | Yes | Yes | No | No |
| positive regulation of nucleobase-containing compound metabolic process | Yes | Yes | No | No |
| purine nucleoside catabolic process | Yes | Yes | No | No |
| purine nucleoside metabolic process | Yes | Yes | No | No |
| purine nucleoside triphosphate catabolic process | Yes | Yes | No | No |
| purine nucleoside triphosphate metabolic process | Yes | Yes | No | No |
| purine nucleotide biosynthetic process | Yes | Yes | No | No |
| purine nucleotide catabolic process | Yes | Yes | No | No |
| purine nucleotide metabolic process | Yes | Yes | No | No |
| purine ribonucleoside metabolic process | Yes | Yes | No | No |
| purine ribonucleotide catabolic process | Yes | Yes | No | No |
| purine ribonucleotide metabolic process | Yes | Yes | No | No |
| regulation of nucleoside metabolic process | Yes | Yes | No | No |
| regulation of nucleotide catabolic process | Yes | Yes | No | No |
| regulation of nucleotide metabolic process | Yes | Yes | No | No |
| negative regulation of nucleobase-containing compound metabolic process | Yes | Yes | No | No |
| **Protein degradation** | proteasome-mediated ubiquitin-dependent protein catabolic process | Yes | Yes | No | No |
| proteolysis | Yes | Yes | No | No |
| regulation of peptidase activity | Yes | Yes | No | No |
| proteasomal protein catabolic process | Yes | Yes | No | No |
| regulation of proteasomal protein catabolic process | Yes | Yes | No | No |
| **Protein metabolism** | negative regulation of cellular protein metabolic process | Yes | Yes | No | No |
| negative regulation of protein metabolic process | Yes | Yes | No | No |
| positive regulation of protein metabolic process | Yes | Yes | No | No |
| regulation of cellular protein metabolic process | Yes | Yes | No | No |
| regulation of protein catabolic process | Yes | Yes | No | No |
| **Protein modification** | cellular protein modification process | Yes | Yes | No | No |
| glycoprotein biosynthetic process | Yes | Yes | No | No |
| post-translational protein modification | Yes | Yes | No | No |
| protein modification by small protein conjugation or removal | Yes | Yes | No | No |
| protein processing | Yes | Yes | No | No |
| regulation of protein modification process | Yes | Yes | No | No |
| regulation of protein processing | Yes | Yes | No | No |
| **Protein transport/targeting** | protein targeting | Yes | No | No | Yes |
| protein targeting to membrane | Yes | No | No | Yes |
| regulation of protein transport | Yes | Yes | No | No |
| **RNA metabolism** | RNA metabolic process | Yes | Yes | No | Yes |
| RNA processing | Yes | Yes | No | Yes |
| negative regulation of RNA metabolic process | Yes | Yes | No | No |
| positive regulation of RNA metabolic process | Yes | Yes | No | No |
| regulation of RNA biosynthetic process | Yes | Yes | No | No |
| regulation of RNA metabolic process | Yes | Yes | No | No |
| ribonucleoside catabolic process | Yes | Yes | No | No |
| ribonucleoside metabolic process | Yes | Yes | No | No |
| ribonucleoside triphosphate catabolic process | Yes | Yes | No | No |
| ribonucleoside triphosphate metabolic process | Yes | Yes | No | No |
| ribonucleotide biosynthetic process | Yes | Yes | No | No |
| ribonucleotide catabolic process | Yes | Yes | No | No |
| ribonucleotide metabolic process | Yes | Yes | No | No |
| ribose phosphate biosynthetic process | Yes | Yes | No | No |
| RNA biosynthetic process | Yes | Yes | No | No |
| RNA phosphodiester bond hydrolysis | Yes | Yes | No | No |
| RNA phosphodiester bond hydrolysis, endonucleolytic | Yes | Yes | No | No |
| RNA transport | Yes | Yes | No | No |
| RNA catabolic process | Yes | No | No | Yes |
| **Secretion** | exocytosis | Yes | Yes | No | No |
| **Signaling** | enzyme linked receptor protein signaling pathway | Yes | Yes | No | No |
| G-protein coupled receptor signaling pathway | Yes | Yes | No | No |
| inositol lipid-mediated signaling | Yes | Yes | No | No |
| inositol phosphate metabolic process | Yes | Yes | No | No |
| MAPK cascade | Yes | Yes | No | No |
| negative regulation of signal transduction | Yes | Yes | No | No |
| phosphatidylinositol-mediated signaling | Yes | Yes | No | No |
| positive regulation of GTPase activity | Yes | Yes | No | No |
| positive regulation of intracellular signal transduction | Yes | Yes | No | No |
| positive regulation of kinase activity | Yes | Yes | No | No |
| positive regulation of signal transduction | Yes | Yes | No | No |
| protein phosphorylation | Yes | Yes | No | No |
| Ras protein signal transduction | Yes | Yes | No | No |
| regulation of GTPase activity | Yes | Yes | No | No |
| regulation of intracellular signal transduction | Yes | Yes | No | No |
| regulation of kinase activity | Yes | Yes | No | No |
| regulation of MAPK cascade | Yes | Yes | No | No |
| regulation of protein kinase activity | Yes | Yes | No | No |
| signal transduction by phosphorylation | Yes | Yes | No | No |
| small GTPase mediated signal transduction | Yes | Yes | No | No |
| transmembrane receptor protein tyrosine kinase signaling pathway | Yes | Yes | No | No |
| **Transcription** | RNA 3'-end processing | Yes | Yes | Yes | Yes |
| DNA-templated transcription, initiation | Yes | Yes | No | Yes |
| mRNA splice site selection | Yes | Yes | No | No |
| negative regulation of gene expression | Yes | Yes | No | No |
| positive regulation of gene expression | Yes | Yes | No | No |
| regulation of transcription, DNA-templated | Yes | Yes | No | No |
| spliceosomal complex assembly | Yes | Yes | No | No |
| transcription from RNA polymerase II promoter | Yes | Yes | No | No |
| transcription, DNA-templated | Yes | Yes | No | No |
| mRNA catabolic process | Yes | No | No | Yes |
| mRNA metabolic process | Yes | No | No | Yes |
| mRNA processing | No | No | Yes | Yes |
| regulation of mRNA processing | No | No | Yes | Yes |
| **Translation** | translation | Yes | No | No | Yes |
| translational elongation | Yes | No | No | Yes |

**Table S 7. Classification of third-generation GO terms used for Figure 4**

|  |  |
| --- | --- |
| **Umbrella term** | **Third-generation GO-terms** |
| **Cell maintenance** | regulation of plasma lipoprotein particle levels |
| amino-acid betaine metabolic process |
| cellular component assembly |
| cellular extravasation |
| prenylation |
| cellular hormone metabolic process |
| cellular macromolecule localization |
| demethylation |
| endomembrane system organization |
| flavonoid metabolic process |
| hormone transport |
| membrane raft organization |
| mineralocorticoid metabolic process |
| nitric oxide metabolic process |
| protein glycosylation |
| regulation of binding |
| regulation of cell shape |
| regulation of cellular component size |
| regulation of growth |
| response to fibroblast growth factor |
| response to hormone |
| response to nitrogen compound |
| response to nutrient |
| response to organonitrogen compound |
| response to oxygen-containing compound |
| response to temperature stimulus |
| response to topologically incorrect protein |
| response to transforming growth factor beta |
| single-organism behavior |
| secondary metabolic process |
| sulfide oxidation |
| vitamin B6 metabolic process |
| maintenance of location in cell |
| establishment of tissue polarity |
| establishment or maintenance of cell polarity |
| iron-sulfur cluster assembly |
| macromolecular complex subunit organization |
| alditol metabolic process |
| carbohydrate derivative metabolic process |
| carbohydrate derivative transport |
| carbohydrate metabolic process |
| cellular aldehyde metabolic process |
| cellular amino acid metabolic process |
| cellular aromatic compound metabolic process |
| cellular biosynthetic process |
| cellular carbohydrate metabolic process |
| cellular catabolic process |
| cellular ketone body metabolic process |
| cellular ketone metabolic process |
| cellular lipid metabolic process |
| cellular macromolecule metabolic process |
| cellular nitrogen compound metabolic process |
| cofactor metabolic process |
| daunorubicin metabolic process |
| doxorubicin metabolic process |
| drug metabolic process |
| glycosyl compound metabolic process |
| heme metabolic process |
| heterocycle metabolic process |
| lipid metabolic process |
| macromolecule metabolic process |
| macromolecule methylation |
| multicellular organismal metabolic process |
| nitrogen cycle metabolic process |
| one-carbon metabolic process |
| organic acid metabolic process |
| organic cyclic compound metabolic process |
| organic hydroxy compound metabolic process |
| organic substance biosynthetic process |
| organic substance catabolic process |
| organic substance transport |
| organonitrogen compound metabolic process |
| organophosphate ester transport |
| organophosphate metabolic process |
| phosphorus metabolic process |
| pigment metabolic process |
| protein metabolic process |
| regulation of metabolic process |
| single-organism biosynthetic process |
| single-organism carbohydrate metabolic process |
| single-organism catabolic process |
| small molecule metabolic process |
| sulfur compound metabolic process |
| thioester metabolic process |
| urea metabolic process |
| xenobiotic metabolic process |
| fluid transport |
| generation of precursor metabolites and energy |
| amide transport |
| establishment of protein localization to membrane |
| establishment of protein localization to organelle |
| hydrogen transport |
| intracellular transport |
| multi-organism intracellular transport |
| multi-organism transport |
| organic acid transport |
| oxidation-reduction process |
| polyketide metabolic process |
| positive regulation of biological process |
| positive regulation of molecular function |
| protein activation cascade |
| protein demethylation |
| protein localization to membrane |
| receptor internalization |
| regulation of catalytic activity |
| regulation of cellular component organization |
| regulation of cellular process |
| regulation of localization |
| regulation of homeostatic process |
| regulation of hormone levels |
| regulation of signaling |
| secretion |
| secretion by cell |
| signal transduction |
| signal transduction by phosphorylation |
| single-organism cellular localization |
| single-organism intracellular transport |
| single-organism transport |
| transmembrane transport |
| tricarboxylic acid cycle |
| cellular response to stress |
| energy derivation by oxidation of organic compounds |
| negative regulation of biological process |
| neurotransmitter metabolic process |
| regulation of developmental process |
| regulation of membrane lipid distribution |
| phagosome maturation |
| cell migration |
| cell projection organization |
| regulation of locomotion |
| taxis |
| chemotaxis |
| cellular developmental process |
| cellular homeostasis |
| electron transport chain |
| **Organelle biology** | cellular component assembly involved in morphogenesis |
| cellular component disassembly |
| cellular component morphogenesis |
| cellular component movement |
| establishment of vesicle localization |
| extracellular structure organization |
| mitochondrion organization |
| nuclear envelope organization |
| nucleus organization |
| organelle fission |
| organelle localization |
| organelle organization |
| peroxisome organization |
| single-organism organelle organization |
| vesicle organization |
| vesicle targeting |
| vesicle-mediated transport |
| **Immune system** | lymphocyte migration |
| T cell mediated cytotoxicity |
| T cell tolerance induction |
| immune response to tumor cell |
| immune response-regulating cell surface receptor signaling pathway involved in phagocytosis |
| inflammatory response to antigenic stimulus |
| leukocyte degranulation |
| positive T cell selection |
| T cell costimulation |
| adaptive immune response |
| antigen processing and presentation of exogenous antigen |
| antigen processing and presentation of peptide antigen |
| antigen processing and presentation of peptide or polysaccharide antigen via MHC class II |
| cytokine production |
| cytokine production involved in immune response |
| defense response |
| defense response to other organism |
| defense response to virus |
| humoral immune response |
| innate immune response |
| leukocyte activation involved in immune response |
| leukocyte chemotaxis |
| leukocyte mediated immunity |
| leukocyte proliferation |
| lymphocyte activation |
| myeloid leukocyte activation |
| myeloid leukocyte migration |
| natural killer cell mediated cytotoxicity |
| natural killer cell mediated cytotoxicity directed against tumor cell target |
| regulation of immune system process |
| response to bacterium |
| response to virus |
| cell activation |
| cell activation involved in immune response |
| complement activation |
| **Organ development and complex behaviour** | anatomical structure arrangement |
| response to BMP |
| sclerotome development |
| appendage development |
| appendage morphogenesis |
| cardiac conduction system development |
| cerebellar granular layer development |
| cerebellar granular layer formation |
| cerebellar granular layer morphogenesis |
| cerebral cortex development |
| circadian regulation of gene expression |
| circadian regulation of translation |
| circadian sleep/wake cycle, sleep |
| cornea development in camera-type eye |
| dendrite development |
| dendritic cell antigen processing and presentation |
| digestive tract development |
| digestive tract morphogenesis |
| endocrine pancreas development |
| entrainment of circadian clock |
| female pregnancy |
| fertilization |
| floor plate formation |
| glomerular capillary formation |
| glomerulus development |
| glomerulus morphogenesis |
| kidney rudiment formation |
| kidney vasculature morphogenesis |
| learning or memory |
| lung development |
| mammary gland formation |
| mating |
| adult behavior |
| multi-organism reproductive behavior |
| multicellular organismal movement |
| myoblast proliferation |
| nephron morphogenesis |
| neural precursor cell proliferation |
| neural tube development |
| optic cup formation involved in camera-type eye development |
| organ growth |
| osteoblast proliferation |
| pancreas development |
| photoreceptor cell maintenance |
| regulation of circadian rhythm |
| regulation of neurotransmitter levels |
| renal system vasculature morphogenesis |
| reproductive behavior |
| respiratory tube development |
| sex differentiation |
| spinal cord development |
| stem cell proliferation |
| synapse organization |
| synaptic vesicle transport |
| tissue remodeling |
| ureter development |
| angiogenesis |
| blood coagulation, fibrin clot formation |
| blood vessel development |
| blood vessel morphogenesis |
| brain development |
| cardiac chamber development |
| cardiac chamber morphogenesis |
| dendritic cell migration |
| forebrain development |
| gland development |
| gonad development |
| heart development |
| hematopoietic or lymphoid organ development |
| inner ear development |
| kidney development |
| morphogenesis of a branching structure |
| muscle cell proliferation |
| muscle structure development |
| nephron development |
| nerve development |
| neural retina development |
| organ development |
| organ morphogenesis |
| outflow tract morphogenesis |
| palate development |
| pattern specification process |
| regulation of anatomical structure size |
| regulation of blood pressure |
| regulation of body fluid levels |
| regulation of multicellular organismal process |
| retina development in camera-type eye |
| retina morphogenesis in camera-type eye |
| sensory organ development |
| system development |
| system process |
| telencephalon development |
| tissue development |
| tissue morphogenesis |
| tube development |
| tube formation |
| tube morphogenesis |
| cochlea development |
| coagulation |
| Sertoli cell differentiation |
| Spemann organizer formation |
| luteolysis |
| male pronucleus assembly |
| pallium development |
| development of primary sexual characteristics |
| multicellular organismal development |
| ventral midline development |
| homeostatic process |
| insemination |
| feeding behavior |
| molting cycle |
| plasma lipoprotein particle organization |
| reproductive structure development |
| semicircular canal development |
| epithelial cell proliferation |
| dermatome development |
| somite development |
| sperm-egg recognition |
| embryo development |
| embryonic morphogenesis |
| embryonic organ development |
| **Detection and response to stimulus** | detection of abiotic stimulus |
| detection of chemical stimulus |
| detection of external stimulus |
| detection of stimulus involved in sensory perception |
| regulation of response to stimulus |
| response to external biotic stimulus |
| response to mechanical stimulus |
| cellular response to abiotic stimulus |
| cellular response to biotic stimulus |
| cellular response to chemical stimulus |
| cellular response to endogenous stimulus |
| response to acid |
| response to organic substance |
| response to oxidative stress |
| response to toxic substance |
| response to tumor cell |
| response to wounding |
| response to xenobiotic stimulus |
| detection of biotic stimulus |
| detection of other organism |
| **Cell cycle** | DNA packaging |
| DNA repair |
| RNA localization |
| RNA transport |
| actin filament-based process |
| cell cycle |
| cell cycle process |
| cell development |
| cell division |
| chromosome organization |
| chromosome segregation |
| cytoskeleton organization |
| histone modification |
| membrane disassembly |
| membrane organization |
| microtubule organizing center organization |
| microtubule-based process |
| nucleobase-containing compound metabolic process |
| nucleobase-containing small molecule metabolic process |
| ribonucleoprotein complex biogenesis |
| telomere maintenance |
| translational initiation |
| single-organism membrane budding |
| transposition |
| cell death |
| execution phase of apoptosis |
| regulation of sequence-specific DNA binding transcription factor activity |
| viral RNA genome replication |
| viral gene expression |
| viral process |
| **Cell adhesion and interaction with other cells** | cell communication |
| cell junction organization |
| cell adhesion |
| cell-cell signaling |
| cell-substrate adhesion |
| entry into other organism involved in symbiotic interaction |
| symbiosis, encompassing mutualism through parasitism |
| movement in host environment |
| transport of virus |
| cell recognition |
| multi-organism membrane fusion |
| regulation of cell killing |

**Table S 8. Classification of sixth-generation GO terms used for Figure 13**

|  |  |
| --- | --- |
| **Umbrella term** | **Sixth-generation GO-term** |
| **Amino acid metabolism** | cellular amino acid catabolic process |
| monocarboxylic acid metabolic process |
| monocarboxylic acid catabolic process |
| alpha-amino acid catabolic process |
| monocarboxylic acid biosynthetic process |
| aspartate family amino acid catabolic process |
| carboxylic acid biosynthetic process |
| lysine metabolic process |
| cellular amino acid metabolic process |
| cellular response to amino acid stimulus |
| L-amino acid import |
| amino acid import |
| peptidyl-amino acid modification |
| cellular biogenic amine metabolic process |
| carnitine biosynthetic process |
| dicarboxylic acid catabolic process |
| tricarboxylic acid metabolic process |
| phosphatidylserine metabolic process |
| glutamine family amino acid catabolic process |
| branched-chain amino acid catabolic process |
| carboxylic acid transport |
| neutral amino acid transport |
| regulation of cellular amino acid metabolic process |
| amino acid transmembrane transport |
| amino acid transport |
| aromatic amino acid family catabolic process |
| carboxylic acid catabolic process |
| monocarboxylic acid transport |
| dicarboxylic acid metabolic process |
| carnitine transport |
| **Cell adhesion/communication/differentiation** | positive regulation of cell-substrate adhesion |
| collagen biosynthetic process |
| collagen catabolic process |
| hyaluronan biosynthetic process |
| hyaluronan metabolic process |
| regulation of collagen biosynthetic process |
| apical junction assembly |
| tight junction assembly |
| regulation of cell shape |
| columnar/cuboidal epithelial cell differentiation |
| generation of neurons |
| positive regulation of fat cell differentiation |
| positive regulation of myeloid cell differentiation |
| positive regulation of myeloid leukocyte differentiation |
| regulation of myeloid leukocyte differentiation |
| striated muscle cell differentiation |
| axonogenesis |
| cardiac muscle tissue development |
| neuron projection morphogenesis |
| hemidesmosome assembly |
| in utero embryonic development |
| regulation of gastrulation |
| regulation of striated muscle tissue development |
| regulation of metanephros development |
| cilium morphogenesis |
| branching morphogenesis of an epithelial tube |
| cell projection morphogenesis |
| negative regulation of cell development |
| canonical Wnt signaling pathway |
| central nervous system neuron differentiation |
| mesenchymal cell development |
| mesenchymal cell differentiation |
| positive regulation of Wnt signaling pathway |
| regulation of canonical Wnt signaling pathway |
| cell morphogenesis involved in differentiation |
| cell morphogenesis involved in neuron differentiation |
| axon extension involved in axon guidance |
| myeloid leukocyte differentiation |
| negative regulation of leukocyte differentiation |
| negative regulation of myeloid leukocyte differentiation |
| neuron projection extension involved in neuron projection guidance |
| positive regulation of cell development |
| positive regulation of cell morphogenesis involved in differentiation |
| ureteric bud morphogenesis |
| **Cell cycle** | regulation of actin filament depolymerization |
| single-organism membrane budding |
| actin polymerization or depolymerization |
| mitotic interphase |
| mitotic chromosome condensation |
| regulation of microtubule cytoskeleton organization |
| mitosis |
| mitotic recombination |
| G1/S transition of mitotic cell cycle |
| spindle organization |
| negative regulation of cell cycle phase transition |
| regulation of mitotic cell cycle phase transition |
| regulation of cell cycle phase transition |
| positive regulation of cell cycle process |
| negative regulation of cell cycle process |
| actin cytoskeleton reorganization |
| actin filament bundle organization |
| mitotic sister chromatid segregation |
| positive regulation of cytoskeleton organization |
| positive regulation of epithelial cell proliferation |
| regulation of actin cytoskeleton organization |
| regulation of actin filament bundle assembly |
| regulation of mesenchymal cell proliferation |
| microtubule bundle formation |
| regulation of actin filament length |
| regulation of actin polymerization or depolymerization |
| negative regulation of epithelial cell proliferation |
| positive regulation of smooth muscle cell proliferation |
| regulation of endothelial cell proliferation |
| **Cell death** | regulation of apoptotic process |
| negative regulation of programmed cell death |
| intrinsic apoptotic signaling pathway by p53 class mediator |
| intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator |
| negative regulation of apoptotic signaling pathway |
| regulation of extrinsic apoptotic signaling pathway |
| extrinsic apoptotic signaling pathway via death domain receptors |
| extrinsic apoptotic signaling pathway |
| **Cell motility** | negative regulation of cell motility |
| positive regulation of cell motility |
| ameboidal cell migration |
| cell chemotaxis |
| endothelial cell migration |
| muscle cell migration |
| positive regulation of cell migration |
| positive regulation of chemotaxis |
| negative regulation of chemotaxis |
| neuron migration |
| regulation of cell migration |
| regulation of ruffle assembly |
| **Defense** | positive regulation of defense response |
| regulation of innate immune response |
| defense response to virus |
| T cell proliferation |
| antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway |
| cellular response to interferon-gamma |
| cellular response to type I interferon |
| complement activation, lectin pathway |
| immune response-activating cell surface receptor signaling pathway |
| immune response-regulating cell surface receptor signaling pathway |
| immunoglobulin mediated immune response |
| interferon-gamma-mediated signaling pathway |
| leukocyte chemotaxis |
| lymphocyte proliferation |
| negative regulation of complement activation |
| negative regulation of humoral immune response |
| positive regulation of adaptive immune response |
| positive regulation of leukocyte mediated immunity |
| regulation of complement activation |
| regulation of defense response to virus |
| regulation of leukocyte mediated cytotoxicity |
| regulation of lymphocyte activation |
| regulation of lymphocyte mediated immunity |
| regulation of lymphocyte proliferation |
| type I interferon signaling pathway |
| negative regulation of leukocyte proliferation |
| negative regulation of lymphocyte activation |
| negative regulation of leukocyte activation |
| regulation of mononuclear cell proliferation |
| regulation of acute inflammatory response |
| regulation of inflammatory response |
| negative regulation of viral genome replication |
| negative regulation of viral process |
| regulation of viral genome replication |
| detection of bacterium |
| activation of immune response |
| B cell activation |
| complement activation |
| immune response-activating signal transduction |
| leukocyte migration |
| regulation of T cell activation |
| regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains |
| regulation of interferon-beta production |
| **DNA packaging/recombination/repair/replication** | DNA conformation change |
| nucleosome organization |
| nucleosome assembly |
| chromatin assembly |
| chromatin assembly or disassembly |
| DNA packaging |
| DNA replication-independent nucleosome organization |
| centromere complex assembly |
| chromatin modification |
| chromatin remodeling |
| covalent chromatin modification |
| histone exchange |
| histone modification |
| regulation of DNA recombination |
| DNA recombination |
| base-excision repair |
| nucleotide-excision repair, DNA gap filling |
| recombinational repair |
| double-strand break repair |
| DNA repair |
| nucleotide-excision repair |
| mismatch repair |
| telomere maintenance via recombination |
| DNA-dependent DNA replication |
| DNA replication |
| regulation of DNA replication |
| DNA strand elongation |
| DNA strand elongation involved in DNA replication |
| **DNA/RNA metabolism** | DNA catabolic process, endonucleolytic |
| regulation of DNA metabolic process |
| DNA catabolic process |
| nucleoside phosphate biosynthetic process |
| regulation of nucleotide catabolic process |
| regulation of nucleoside metabolic process |
| regulation of nucleotide metabolic process |
| nucleoside monophosphate catabolic process |
| purine nucleoside monophosphate catabolic process |
| ribonucleoside monophosphate catabolic process |
| nucleobase metabolic process |
| deoxyribonucleotide metabolic process |
| pyrimidine nucleoside biosynthetic process |
| nucleic acid phosphodiester bond hydrolysis |
| pyridine nucleotide metabolic process |
| nucleoside triphosphate biosynthetic process |
| purine nucleoside triphosphate biosynthetic process |
| purine nucleoside monophosphate biosynthetic process |
| nicotinamide nucleotide metabolic process |
| regulation of nucleotide biosynthetic process |
| positive regulation of nucleobase-containing compound metabolic process |
| nucleotide-sugar biosynthetic process |
| nucleotide-sugar metabolic process |
| protein-DNA complex assembly |
| RNA catabolic process |
| ncRNA metabolic process |
| RNA metabolic process |
| mitochondrial RNA metabolic process |
| ncRNA processing |
| ribonucleoside triphosphate biosynthetic process |
| ribonucleoside monophosphate biosynthetic process |
| RNA biosynthetic process |
| DNA metabolic process |
| nucleoside metabolic process |
| nucleoside monophosphate biosynthetic process |
| nucleoside monophosphate metabolic process |
| nucleoside phosphate metabolic process |
| nucleoside salvage |
| nucleotide biosynthetic process |
| nucleotide metabolic process |
| pyrimidine nucleobase metabolic process |
| pyrimidine nucleoside metabolic process |
| pyrimidine nucleoside salvage |
| pyrimidine nucleotide metabolic process |
| pyrimidine-containing compound salvage |
| regulation of RNA biosynthetic process |
| regulation of RNA metabolic process |
| purine nucleoside monophosphate metabolic process |
| ribonucleoside monophosphate metabolic process |
| purine nucleoside biosynthetic process |
| purine nucleotide biosynthetic process |
| ribonucleoside biosynthetic process |
| ribonucleotide biosynthetic process |
| ribose phosphate biosynthetic process |
| negative regulation of RNA metabolic process |
| **Endocytosis/Secretion** | retrograde vesicle-mediated transport, Golgi to ER |
| regulation of endocytosis |
| negative regulation of secretion |
| positive regulation of protein secretion |
| exocytosis |
| positive regulation of secretion |
| peptide hormone secretion |
| peptide secretion |
| regulation of insulin secretion |
| insulin secretion |
| regulation of peptide hormone secretion |
| regulation of peptide secretion |
| **Function by differentated tissue** | detection of visible light |
| blood coagulation, fibrin clot formation |
| actin-myosin filament sliding |
| heart contraction |
| learning or memory |
| negative regulation of blood coagulation |
| phototransduction |
| regulation of blood coagulation |
| regulation of heart rate |
| regulation of muscle contraction |
| regulation of muscle hypertrophy |
| striated muscle contraction |
| **Ion metabolism** | hydrogen ion transmembrane transport |
| potassium ion transmembrane transport |
| metal ion transport |
| regulation of metal ion transport |
| regulation of ion transmembrane transporter activity |
| regulation of transmembrane transporter activity |
| positive regulation of ion transport |
| positive regulation of transporter activity |
| divalent inorganic cation transport |
| divalent metal ion transport |
| positive regulation of ion transmembrane transporter activity |
| organic anion transport |
| negative regulation of ion transmembrane transporter activity |
| sodium ion transport |
| cellular cation homeostasis |
| cellular ion homeostasis |
| regulation of anion transport |
| calcium ion transmembrane transport |
| cation homeostasis |
| monovalent inorganic cation transport |
| positive regulation of calcium ion transport |
| regulation of ion transmembrane transport |
| regulation of potassium ion transmembrane transport |
| potassium ion transport |
| ferric iron transport |
| transition metal ion transport |
| trivalent inorganic cation transport |
| sodium ion transmembrane transport |
| **Lipid metabolism** | lipid oxidation |
| fatty acid oxidation |
| fatty acid catabolic process |
| fatty acid beta-oxidation |
| short-chain fatty acid metabolic process |
| glycerolipid catabolic process |
| triglyceride metabolic process |
| glycerolipid biosynthetic process |
| long-chain fatty acid metabolic process |
| glycerophospholipid metabolic process |
| fatty acid elongation |
| short-chain fatty acid catabolic process |
| fatty acid transport |
| long-chain fatty acid transport |
| unsaturated fatty acid metabolic process |
| glycerophospholipid biosynthetic process |
| phospholipid biosynthetic process |
| regulation of fatty acid metabolic process |
| regulation of lipase activity |
| positive regulation of lipase activity |
| phospholipid transport |
| sphingolipid biosynthetic process |
| sterol biosynthetic process |
| regulation of phospholipase activity |
| lipid translocation |
| glycosphingolipid metabolic process |
| acyl-CoA biosynthetic process |
| acylglycerol metabolic process |
| terpenoid metabolic process |
| negative regulation of lipid biosynthetic process |
| negative regulation of lipid metabolic process |
| neutral lipid catabolic process |
| positive regulation of lipid kinase activity |
| positive regulation of lipid metabolic process |
| regulation of lipid catabolic process |
| acylglycerol catabolic process |
| sphingolipid metabolic process |
| unsaturated fatty acid biosynthetic process |
| positive regulation of steroid metabolic process |
| regulation of steroid biosynthetic process |
| regulation of lipid biosynthetic process |
| glycerol metabolic process |
| glycolipid metabolic process |
| **Metabolism** | glucan biosynthetic process |
| glycogen biosynthetic process |
| acetyl-CoA metabolic process |
| glucose metabolic process |
| ethanol metabolic process |
| diterpenoid metabolic process |
| ethanolamine-containing compound metabolic process |
| monosaccharide biosynthetic process |
| hexose biosynthetic process |
| positive regulation of phosphorus metabolic process |
| regulation of cellular macromolecule biosynthetic process |
| tetrahydrofolate metabolic process |
| glucose transport |
| hexose transport |
| glycosaminoglycan biosynthetic process |
| glycosaminoglycan catabolic process |
| positive regulation of transferase activity |
| bicarbonate transport |
| macromolecule glycosylation |
| positive regulation of glucose transport |
| positive regulation of monooxygenase activity |
| modification-dependent macromolecule catabolic process |
| modification-dependent protein catabolic process |
| cellular protein catabolic process |
| positive regulation of cellular biosynthetic process |
| negative regulation of cellular protein metabolic process |
| regulation of glucose import |
| cellular response to hydrogen peroxide |
| hydrogen peroxide catabolic process |
| acetyl-CoA biosynthetic process |
| alditol metabolic process |
| biotin metabolic process |
| cellular glucan metabolic process |
| glycogen metabolic process |
| negative regulation of nucleobase-containing compound metabolic process |
| positive regulation of cellular protein metabolic process |
| positive regulation of protein metabolic process |
| regulation of cellular protein metabolic process |
| positive regulation of cellular catabolic process |
| positive regulation of hydrolase activity |
| regulation of ketone biosynthetic process |
| regulation of endopeptidase activity |
| regulation of peptidase activity |
| negative regulation of macromolecule biosynthetic process |
| proton transport |
| ATP synthesis coupled electron transport |
| mitochondrial ATP synthesis coupled electron transport |
| mitochondrial electron transport, NADH to ubiquinone |
| energy coupled proton transport, down electrochemical gradient |
| positive regulation of protein catabolic process |
| positive regulation of mitochondrion organization |
| regulation of mitochondrial membrane permeability |
| **Protein degradation/folding/modification/targeting** | protein insertion into mitochondrial membrane |
| proteasomal protein catabolic process |
| proteasome-mediated ubiquitin-dependent protein catabolic process |
| positive regulation of ubiquitin-protein ligase activity |
| cellular response to unfolded protein |
| endoplasmic reticulum unfolded protein response |
| 'de novo' protein folding |
| protein folding |
| Golgi transport vesicle coating |
| Golgi vesicle budding |
| post-translational protein modification |
| regulation of protein modification process |
| cellular protein complex disassembly |
| protein oligomerization |
| protein glycosylation |
| protein modification by small protein conjugation or removal |
| cellular protein modification process |
| ER to Golgi vesicle-mediated transport |
| post-Golgi vesicle-mediated transport |
| protein processing |
| cotranslational protein targeting to membrane |
| protein targeting to membrane |
| protein targeting |
| protein targeting to mitochondrion |
| regulation of protein transport |
| protein localization to mitochondrion |
| SRP-dependent cotranslational protein targeting to membrane |
| protein localization to endoplasmic reticulum |
| protein targeting to ER |
| intracellular protein transmembrane transport |
| **Signaling** | cyclic nucleotide biosynthetic process |
| oxidative phosphorylation |
| cyclic-nucleotide-mediated signaling |
| protein dephosphorylation |
| second-messenger-mediated signaling |
| negative regulation of kinase activity |
| signal transduction by phosphorylation |
| MAPK cascade |
| calcium-mediated signaling |
| small GTPase mediated signal transduction |
| G-protein coupled receptor signaling pathway |
| positive regulation of signal transduction |
| cyclic nucleotide catabolic process |
| inositol phosphate metabolic process |
| negative regulation of phosphorus metabolic process |
| regulation of MAPK cascade |
| transmembrane receptor protein tyrosine kinase signaling pathway |
| positive regulation of kinase activity |
| enzyme linked receptor protein signaling pathway |
| ERK1 and ERK2 cascade |
| G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger |
| cyclic nucleotide metabolic process |
| cytokine-mediated signaling pathway |
| inositol lipid-mediated signaling |
| negative regulation of intracellular signal transduction |
| negative regulation of signal transduction |
| negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway |
| phosphatidylinositol-mediated signaling |
| positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway |
| protein kinase B signaling |
| protein phosphorylation |
| receptor internalization |
| regulation of BMP signaling pathway |
| regulation of G-protein coupled receptor protein signaling pathway |
| regulation of intracellular signal transduction |
| regulation of lipid kinase activity |
| regulation of pathway-restricted SMAD protein phosphorylation |
| regulation of phosphate metabolic process |
| regulation of phosphatidylinositol 3-kinase signaling |
| regulation of protein activation cascade |
| regulation of protein kinase B signaling |
| regulation of small GTPase mediated signal transduction |
| regulation of transmembrane receptor protein serine/threonine kinase signaling pathway |
| response to cAMP |
| signal transduction by p53 class mediator |
| icosanoid metabolic process |
| phosphatidylinositol metabolic process |
| cellular response to transforming growth factor beta stimulus |
| positive regulation of oxidoreductase activity |
| nucleobase-containing small molecule interconversion |
| positive regulation of GTPase activity |
| regulation of GTPase activity |
| regulation of Ras GTPase activity |
| I-kappaB kinase/NF-kappaB signaling |
| regulation of I-kappaB kinase/NF-kappaB signaling |
| negative regulation of G-protein coupled receptor protein signaling pathway |
| regulation of kinase activity |
| regulation of protein kinase activity |
| positive regulation of intracellular signal transduction |
| **Transcription/Translation** | regulation of mRNA catabolic process |
| mRNA catabolic process |
| regulation of transcription, DNA-templated |
| DNA-templated transcription, elongation |
| RNA 3'-end processing |
| mRNA metabolic process |
| spliceosomal snRNP assembly |
| DNA-templated transcription, termination |
| RNA processing |
| RNA splicing |
| mRNA processing |
| 3'-UTR-mediated mRNA stabilization |
| positive regulation of gene expression |
| regulation of gene expression, epigenetic |
| transcription from RNA polymerase III promoter |
| transcription from mitochondrial promoter |
| negative regulation of gene expression |
| posttranscriptional regulation of gene expression |
| transcription, DNA-templated |
| transcription from RNA polymerase II promoter |
| translational elongation |
| tRNA aminoacylation |
| tRNA aminoacylation for protein translation |
| tRNA metabolic process |
| rRNA metabolic process |
| ribonucleoprotein complex assembly |
| RNA transport |
| nuclear export |
| regulation of translation |
| translation |