Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mechanisms of Ageing and Development 126 (2005) 193–208
mechanisms of ageing and development
www.elsevier.com/locate/mechagedev

Text-based analysis of genes, proteins, aging, and cancer

Jeremy R. Semeiks, L.R. Grate, I.S. Mian*

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Available online 26 October 2004

Abstract

The diverse nature of cancer- and aging-related genes presents a challenge for large-scale studies based on molecular sequence and profiling data. An underexplored source of data for modeling and analysis is the textual descriptions and annotations present in curated gene-centered biomedical corpora. Here, 450 genes designated by surveys of the scientific literature as being associated with cancer and aging were analyzed using two complementary approaches. The first, ensemble attribute profile clustering, is a recently formulated, text-based, semi-automated data interpretation strategy that exploits ideas from statistical information retrieval to discover and characterize groups of genes with common structural and functional properties. Groups of genes with shared and unique Gene Ontology terms and protein domains were defined and examined. Human homologs of a group of known *Drosphila* aging-related genes are candidates for genes that may influence lifespan (*hep*/MAPK2K7, *bsk*/MAPK8, *puc*/LOC285193). These JNK pathway-associated proteins may specify a molecular hub that coordinates and integrates multiple intra- and extracellular processes via space- and time-dependent interactions with proteins in other pathways. The second approach, a qualitative examination of the chromosomal locations of 311 human cancer- and aging-related genes, provides anecdotal evidence for a ''phenotype position effect'': genes that are proximal in the linear genome often encode proteins involved in the same phenomenon. Comparative genomics was employed to enhance understanding of several genes, including open reading frames, identified as new candidates for genes with roles in aging or cancer. Overall, the results highlight fundamental molecular and mechanistic connections between progenitor/stem cell lineage determination, embryonic morphogenesis, cancer, and aging. Despite diversity in the nature of the molecular and cellular processes associated with these phenomena, they seem related to the architectural hub of tissue polarity and a need to generate and control this property in a timely manner.

Keywords: Statistical information retrieval; Progenitor/stem cell lineage determination; Embryonic morphogenesis; Cancer; Aging; Phenotype position effect

1. Introduction

Genomic and genetic approaches to the study of cancer and aging are widespread. An underutilized resource in investigations of these and other biological phenomena is the biomedical literature. Efforts to exploit data in the form of text seek to convert information into knowledge in a systematic and quantitative manner, most often using tools and techniques developed for modeling text documents in other scientific disciplines (reviewed in Shatkay and Feldman, 2003). For example, probabilistic graphical models have proved invaluable in the analysis of molecular sequence and profiling data, and recently this statistical framework has been applied to a corpus of documents about *C. elegans* in order to gain insights into aging (Blei et al., 2004).

A common outcome of biomedical research is the generation of a collection of ''interesting'' genes and/or proteins, for example, genes involved in response to stress, differentially expressed between normal and aged tissue samples, and so on. Here, the focus is two extant collections of genes that reviews of the scientific literature designated as being involved in cancer and in aging. This work illustrates how analysis of these collections using a combination of two general approaches currently underexploited in computational biology yields new and enhanced insights into these genes and phenomena. The first approach, ensemble attribute profile clustering, is the semi-automated functional grouping of genes and/or gene products based on their shared annotations from a set of terms specified in curated textual corpora. The second approach considers the order of

* Corresponding author.
E-mail address: smian@lbl.gov (I.S. Mian).

0047-6374/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.mad.2004.09.028

genes in a genome (as opposed to their actual sequences) and is based on the hypothesis that genes that are close in terms of linear order are involved in the same phenomenon (''phenotype position effect''). The integrated text-based analysis described here both recapitulates known properties of genes involved in the generalized phenotypes of cancer and aging, and suggests new candidates for genes associated with these phenomena.

Ensemble attribute profile clustering is a recently formulated strategy designed to assist in the analysis of a set of genes (Semeiks et al., 2004). Using data in the form of textual descriptions rather than molecular sequences or profiles, it addresses the task of discovering and characterizing groups of genes with common functional, structural, and other properties (''attributes''). The approach exploits ideas from statistical information retrieval, a field which, in contrast to Natural Language Processing, emphasizes vector space and probabilistic models for document representation, retrieval and analysis (Salton, 1988). Vector space models of documents ignore syntax and semantics when recasting a document as a bag of words. Similarly, gene attribute profiles, the vector space model representation of genes employed by ensemble attribute profile clustering, neglect the order in which domains occur in a protein, discard the manner in which terms in a biological ontology are organized, and so on when recasting a gene as a set of attributes. Next, simple probabilistic graphical models are estimated from a set of gene attribute profiles, the models used to define groups of genes with similar patterns of attributes, and the biology of genes assigned to groups examined by a user. Previously, this type of graphical model was employed to cluster transcript profiles and thus determine groups of genes (or experiments) with common patterns of expression (Moler et al., 2000a, b; Bhattacharjee et al., 2001). Here, analysis of 291 cancer- and 159 aging-related genes using attributes derived from the gene-centered LocusLink corpus and ensemble attribute profile clustering demonstrates the practical utility of this semi-automated data interpretation strategy.

Recent studies indicate that the distribution of genes along eukaryotic chromosomes is not random, that is, nearby genes may be co-expressed, be involved in the same metabolic pathway, interact with each other, and share regulatory regions, histone modification states, or regulatory elements (reviewed in Hurst et al., 2004). Most investigations pertaining to the biological consequences of gene order have examined evidence regarding the effect of a gene's genomic location on the molecular endpoint of transcription, that is, an expression position effect. This work postulates the existence of a phenotype position effect, the interplay between genome organization and cellular/tissue/organismal endpoints. Inspection of genes in human genomic regions containing known cancer- and aging-related genes suggests relationships between progenitor/stem cell lineage determination, embryonic morphogenesis, cancer, and aging. The roles of tissue polarity and cryptic genetic variation are discussed.

2. Methods

2.1. Known cancer- and aging-related genes

The two published collections of cancer- and aging-related genes reexamined here were derived from surveys of the scientific literature (cancer, Futreal et al., 2004), http://www.sanger.ac.uk/genetics/CGP/Census/; aging, March 2004 SAGEKE database (K. LaMarco, personal communication), http://sageke.sciencemag.org). The 291 cancer genes were exclusively human, whereas the 167 aging genes were from a range of species (*Homo sapiens*, 20 genes; *Mus musculus*, 18; *Drosophila melanogaster*, 30; *Caenorhabditis elegans*, 97; *Rattus norvegicus*, 2).

2.2. Ensemble attribute profile clustering

In order to discover and characterize groups of cancer- and aging-related genes with similar patterns of structural and functional properties, the two published collections were analyzed using ensemble attribute profile clustering. A comprehensive discussion of this general-purpose approach can be found elsewhere (Semeiks et al., 2004) so only a summary of the tasks involved is given below.

Data conversion maps gene identifiers in a source list to entries in a curated biomedical corpus. Here, the corpus employed was LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink). For the July 2004 build of LocusLink, 291 of the cancer genes and 159 of the aging genes could be equated with distinct genetic loci, that is, LocusIDs. These two collections will be referred to as the Cancer and Aging data sets.

Feature generation represents each gene as a vector of (exchangeable) attributes where an attribute corresponds to a property of the gene and/or its protein product(s). Each element or ''feature'' of the vector signifies the significance or weight of the attribute in question in the gene of interest. Here, only protein-related attributes in general, and LocusLink-assigned Gene Ontology (GO; http://www.geneontology.org) (Ashburner et al., 2000) and Conserved Domain Database (CDD) (Marchler-Bauer et al., 2003) controlled vocabulary terms in particular, were considered. Examples of two attributes are the GO terms `transcription, DNA-dependent` and CDD term `Tyrosine kinase, catalytic domain. Phosphotransferases`.

Feature selection seeks to avoid overfitting the data during subsequent clustering by discarding uninformative attributes. Here, the problem of overfitting was reduced by retaining an attribute only if it was assigned to at least three of the N total genes in a set. If P is the number of attributes satisfying a selection criterion, the N P-dimensional vectors

can be aggregated into an $N \times P$ gene attribute profile matrix. The Cancer and Aging data sets comprised 291 251-dimensional and 159 126-dimensional gene attribute profile vectors, respectively. All the attributes considered here were treated as binary random variables so each element of the matrix specified whether an attribute was (''1'') or was not (''0'') assigned to a gene.

Model-based clustering estimates a simple probabilistic (graphical) model from feature vectors that partitions the vectors into groups with similar patterns of features. Here, every GO/CDD term attribute in a gene attribute profile vector was a binary random variable so each variable in the finite mixture (naïve Bayes) model was modeled using a Bernouilli distribution. AutoClass (Cheeseman and Stutz, 1996), the specific finite mixture model implementation used here, employs a Bayesian approach to determine automatically the number of clusters K which best fit the data. An AutoClass model was trained using a gene attribute profile matrix and the following non-default parameter settings: `max_duration = 10800`, `max_cycles = 500`, `start_j_list = 40,45`, and `force_new_search_p = true`. A trained model was used to assign each gene attribute profile vector to one of the K clusters and thus produce a partitioning of the N genes into groups with similar patterns of attributes.

Cluster combination finds a robust partitioning of the data (consensus clusters) by integrating results from multiple independent runs of one or more clustering algorithm(s). Here, since different intializations of the program AutoClass can yield distinct numbers of clusters and assignments of gene attribute profiles to clusters, four independent AutoClass models were estimated for a given gene attribute profile matrix. The ensuing four partitionings of genes in a data set were combined as follows. For each (AutoClass) model, genes assigned to the same cluster were identified. A unanimous voting scheme was used to fuse these results: genes that co-occurred in all four models were equated with belonging to the same consensus cluster. This conservative approach can result in some of the N genes not being equated with a consensus cluster and hence not appearing in the final results.

Cluster interpretation uses consensus clusters of genes, attributes important in defining consensus clusters, external knowledge, and human reasoning to elucidate the shared functional, structural and biological properties of genes in a collection. Here, two reports were produced for a data set. The first focused on consensus clusters, displaying the genes assigned to consensus clusters along with additional information found in LocusLink: gene symbol, GO and CDD terms, cytogenetic location, and physiological pathways from the Kyoto Encyclopedia of Genes and Genomes or KEGG database (http://www.genome.jp/kegg). The second report focused on the common features of a consensus cluster, displaying the count and relative frequencies of all attributes associated with genes in a particular consensus cluster. For a consensus cluster containing two or more genes, an ''influential attribute'' was defined as one that occured with a frequency ≥ 0.5 in the genes assigned to that consensus cluster.

2.3. Phenotype position effects

In order to examine the interplay between gene order, cancer, and aging, genomic regions containing the Cancer and Aging genes were examined. Here, every LocusID for a human gene in each of these two data sets was mapped to its corresponding accession in the NCBI Reference Sequence database (Refseq; http://www.ncbi.nlm.nih.gov/RefSeq). The Refseqs were mapped to chromosomal coordinates in the May 2004 UCSC build of the Golden Path database (specifically, the maximal transcription start and end coordinates in the table `refFlat`; http://genome.ucsc.edu). These coordinates were used to calculate the spacing between the Cancer and Aging genes in terms of both number of basepairs and number of intervening RefSeq accessions.

The homologs of a given gene in other species were taken to be those specified by HomoloGene (http://www.ncbi.nlm.nih.gov/HomoloGene/). Information on the biology of human and mouse genes was taken from their respective LocusLink entries. For *D. melanogaster* genes, Flybase (http://www.flybase.org) was used to access GO terms and other information, including FlyGRID defined genetic and physical protein-protein interactions.

3. Results

3.1. Cancer and Aging consensus clusters and influential attributes recapitulate known global features of cancer and aging

Given a collection of genes, ensemble attribute clustering provides a systematic, semi-automated method for discovering groups of genes with similar patterns of attributes (consensus clusters), characterizing groups in terms of their influential attributes, and ascertaining the general structural, functional and other properties common to the collection. Tables 1 and 2 summarize the Cancer and Aging consensus clusters and their GO/CDD term influential attributes (information on individual genes, consensus clusters, and full attribute lists can be found in Supplementary Material). A consensus cluster may have no associated influential attributes if most of its genes have few or no LocusLink-assigned attributes (for example, the *C. elegans* genes in Aging consensus cluster 0), or if no attribute frequencies meet the criterion of ≥ 0.5 for such a designation (attribute frequencies in Cancer consensus cluster 0 are ≤ 0.22).

The Cancer consensus clusters and influential attributes are consistent with the observation that ''the most commonly represented domain encoded by cancer genes is the protein kinase, followed by domains involved in DNA binding and

Table 1
Consensus clusters and influential attributes for the 291 genes in the Cancer data set

Consensus cluster		LocusLink gene symbol (LocusID) and GO/CDD attributes
0	53 Genes	ABI1 (10006); AF15Q14 (57082); AF1Q (10962); ASPSCR1 (79058); ATIC (471); BCL5 (603); BCL7A (605); BHD (201163); BRD4 (23476); BTG1 (694); CCDC6 (8030); CCNB1IP1 (57820); CEP1 (11064); CHIC2 (26511); CXXC6 (80312); ELKS (23085); FANCD2 (2177); FGFR1OP (11116); FIP1L1 (81608); FNBP1 (23048); FSTL3 (10272); GMPS (8833); HIST1H4I (8294); HRPT2 (79577); IGH@ (3492); IGK@ (50802); IGL@ (3535); IRTA1 (83417); KIAA1618 (57714); LASP1 (3927); LHFP (10186); LMO1 (4004); LMO2 (4005); LPP (4026); MAML2 (84441); MDS1 (4197); MECT1 (23373); MLLT2 (4299); MUTYH (4595); NPM1 (4869); NUT (256646); PRCC (5546); PRO1073 (29005); PSIP1 (11168); RAP1GDS1 (5910); SBDS (51119); SS18 (6760); SS18L1 (26039); STL (7955); TCL1A (8115); TFPT (29844); TRD@ (6964)
	Attributes	None
1	39 Genes	CBFA2T1 (862); COPEB (1316); ELL (8178); ERG (2078); ETV1 (2115); ETV6 (2120); EWSR1 (2130); FLI1 (2313); FOXO1A (2308); FOXO3A (2309); HLF (3131); HMGA2 (8091); HOXA9 (3205); LYL1 (4066); MAF (4094); MLL (4297); MLLT1 (4298); MLLT10 (8028); MLLT3 (4300); MLLT6 (4302); MLLT7 (4303); MYC (4609); MYCL1 (4610); MYCN (4613); MYST4 (23522); OLIG2 (10215); PAX3 (5077); PAX7 (5081); RB1 (5925); REL (5966); RUNX1 (861); SMARCB1 (6598); TAL1 (6886); TAL2 (6887); TCF3 (6929); TFE3 (7030); TLX1 (3195); WT1 (7490); ZNF278 (23598)
	Attributes	CC: nucleus; intracellular; cell MF: nucleic acid binding; binding; DNA binding; transcription regulator activity; transcription factor activity BP: transcription, DNA-dependent †; transcription †; regulation of transcription, DNA-dependent †; regulation of transcription †; nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; cellular process †; cellular physiological process †; cell growth and/or maintenance †
2	30 Genes	AF5Q31 (27125); ATF1 (466); BCL11A (53335); BCL11B (64919); CBFA2T3 (863); CDX2 (1045); CEBPA (1050); ETV4 (2118); GATA1 (2623); HOXA11 (3207); HOXA13 (3209); HOXC13 (3229); HOXD11 (3237); HOXD13 (3239); LAF4 (3899); MEN1 (4221); NOTCH1 (4851); NR4A3 (8013); PAX5 (5079); PAX8 (7849); PBX1 (5087); PPARG (5468); SSX4 (6759); TCF1 (6927); TCF12 (6938); TFEB (7942); TLX3 (30012); ZNF198 (7750); ZNF384 (171017); ZNFN1A1 (10320)
	Attributes	CC: cell; intracellular; nucleus MF: binding †; nucleic acid binding; DNA binding; transcription regulator activity; transcription factor activity BP: transcription, DNA-dependent †; transcription †; nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; regulation of transcription, DNA-dependent; regulation of transcription; development
3	16 Genes	ALK (238); BMPR1A (657); EGFR (1956); ERBB2 (2064); FGFR1 (2260); FGFR2 (2263); FGFR3 (2261); FLT3 (2322); FLT4 (2324); KIT (3815); MET (4233); NTRK1 (4914); NTRK3 (4916); PDGFRA (5156); PDGFRB (5159); RET (5979)
	Attributes	CC: membrane †; integral to membrane †; cell †; plasma membrane; integral to plasma membrane MF: transmembrane receptor protein kinase activity †; transmembrane receptor activity †; transferase activity †; signal transducer activity †; receptor activity †; protein kinase activity †; kinase activity †; catalytic activity †; binding †; ATP binding †; adenyl nucleotide binding †; transmembrane receptor protein tyrosine kinase activity; protein-tyrosine kinase activity BP: protein modification †; protein metabolism †; protein amino acid phosphorylation †; phosphorylation †; metabolism †; signal transduction; cellular process; cell surface receptor linked signal transduction; cell communication; enzyme linked receptor protein signaling pathway; transmembrane receptor protein tyrosine kinase signaling pathway; cellular physiological process; cell growth and/or maintenance CDD: Tyrosine kinase, catalytic domain. Phosphotransferases
4	16 Genes	CBFB (865); CREBBP (1387); CTNNB1 (1499); FEV (54738); IRF4 (3662); MHC2TA (4261); NCOA2 (10499); NCOA4 (8031); POU2AF1 (5450); PRRX1 (5396); RARA (5914); SMAD4 (4089); SSX1 (6756); SUFU (51684); TRIP11 (9321); ZNF145 (7704)
	Attributes	CC: intracellular †; cell †; nucleus MF: protein binding †; binding †; transcription regulator activity; transcription factor binding; transcription cofactor activity; nucleic acid binding; transcription coactivator activity; DNA binding; transcription factor activity BP: transcription, DNA-dependent †; transcription †; nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; regulation of transcription, DNA-dependent; regulation of transcription; cellular process

Table 1 (*Continued*)

Consensus cluster		LocusLink gene symbol (LocusID) and GO/CDD attributes
5	16 Genes	ATM (472); BRCA2 (675); DDB2 (1643); ERCC4 (2072); ERCC5 (2073); FANCG (2189); MLH1 (4292); MSH2 (4436); MSH6 (2956); NBS1 (4683); PMS1 (5378); PMS2 (5395); RAD51L1 (5890); SET (6418); XPA (7507); XPC (7508)
	Attributes	CC: nucleus †; intracellular †; cell † MF: binding †; nucleic acid binding; DNA binding; damaged DNA binding BP: nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; DNA metabolism †; response to stress; response to DNA damage stimulus; DNA repair; cellular process; cellular physiological process; cell proliferation; cell growth and/or maintenance; cell cycle; regulation of cell cycle
6	15 Genes	APC (324); BCL2 (596); FH (2271); GPC3 (2719); IL21R (50615); MLLT4 (4301); NF1 (4763); NF2 (4771); PDGFB (5155); PTCH (5727); PTEN (5728); PTPN11 (5781); TRA@ (6955); TSC1 (7248); TSHR (7253)
	Attributes	CC: cell; membrane MF: binding BP: cellular process †; cellular physiological process; cell growth and/or maintenance; cell proliferation; regulation of cell cycle; cell cycle
7	11 Genes	ABL2 (27); AKT2 (208); BCR (613); BRAF (673); CDK4 (1019); JAK2 (3717); LCK (3932); MAP2K4 (6416); PIM1 (5292); STK11 (6794); TEC (7006)
	Attributes	MF: transferase activity †; protein kinase activity †; kinase activity †; catalytic activity †; binding; ATP binding; adenyl nucleotide binding; protein serine/threonine kinase activity BP: protein modification †; protein metabolism †; protein amino acid phosphorylation †; phosphorylation †; metabolism †; cellular process; signal transduction; intracellular signaling cascade; cellular physiological process; cell growth and/or maintenance; cell communication
8	11 Genes	BCL10 (8915); CALCR (799); EXT1 (2131); EXT2 (2132); MALT1 (10892); RABEP1 (9135); SH3GL1 (6455); SMO (6608); TFRC (7037); TNFRSF6 (355); TSC2 (7249)
	Attributes	CC: cell †; intracellular; membrane; cytoplasm; integral to membrane MF: signal transducer activity BP: cellular process †; cellular physiological process; signal transduction; cell growth and/or maintenance; cell communication; metabolism
9	9 Genes	BIRC3 (330); BRCA1 (672); CBL (867); NSD1 (64324); PML (5371); TIF1 (8805); TRIM33 (51592); WHSC1 (7468); WHSC1L1 (54904)
	Attributes	CC: ubiquitin ligase complex †; intracellular †; cell †; nucleus MF: zinc ion binding †; ubiquitin-protein ligase activity †; transition metal ion binding †; metal ion binding †; ligase activity †; catalytic activity †; binding †; acid-D-amino acid ligase activity †; nucleic acid binding; DNA binding; protein binding; transcription regulator activity BP: ubiquitin cycle †; protein ubiquitination †; protein modification †; protein metabolism †; metabolism †; transcription, DNA-dependent; transcription; regulation of transcription, DNA-dependent; regulation of transcription; nucleobase, nucleoside, nucleotide and nucleic acid metabolism; cellular process; cellular physiological process; cell growth and/or maintenance
10	6 Genes	EVI1 (2122); FUS (2521); GOLGA5 (9950); GPHN (10243); LCP1 (3936); SFPQ (6421)
	Attributes	CC: intracellular †; cell †; nucleus; cytoplasm MF: binding; nucleic acid binding; metal ion binding; DNA binding
11	5 Genes	HSPCA (3320); NCKIPSD (51517); NUP214 (8021); NUP98 (4928); TPR (7175)
	Attributes	CC: intracellular †; cell †; nucleus; cytoplasm; pore complex; nuclear pore; nuclear membrane; membrane; integral to membrane BP: transport †; protein metabolism †; metabolism †; cellular process †; cellular physiological process †; cell growth and/or maintenance †; protein transport; protein-nucleus import
12	5 Genes	CYLD (1540); HSPCB (3326); NACA (4666); PCM1 (5108); RPL22 (6146)
	Attributes	CC: intracellular †; cytoplasm †; cell † MF: binding BP: metabolism †; protein metabolism
13	5 Genes	ARNT (405); BCL6 (604); DDIT3 (1649); DEK (7913); EP300 (2033)
	Attributes	CC: nucleus †; intracellular †; cell †

Table 1 (*Continued*)

Consensus cluster		LocusLink gene symbol (LocusID) and GO/CDD attributes
		MF: transcription regulator activity †; protein binding †; nucleic acid binding †; DNA binding †; binding †; transcription factor activity; transcription factor binding; transcription cofactor activity BP: transcription, DNA-dependent †; transcription †; regulation of transcription, DNA-dependent †; regulation of transcription †; nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; cellular process †; cellular physiological process †; cell growth and/or maintenance †; transcription from Pol II promoter; signal transduction; cell proliferation; cell communication
14	5 Genes	ACSL6 (23305); COX6C (1345); SDHB (6390); SDHC (6391); SDHD (6392)
	Attributes	CC: mitochondrion †; intracellular †; cytoplasm †; cell †; membrane; mitochondrial membrane; integral to membrane MF: transporter activity; catalytic activity; electron transporter activity; binding BP: metabolism †; energy pathways; electron transport; tricarboxylic acid cycle; main pathways of carbohydrate metabolism
15	5 Genes	CCND1 (595); EPS15 (2060); GAS7 (8522); MN1 (4330); SEPT6 (23157)
	Attributes	MF: binding BP: cellular process †; cellular physiological process †; cell proliferation †; cell growth and/or maintenance †; cell cycle; regulation of cell cycle
16	5 Genes	HRAS (3265); KRAS2 (3845); MSF (10801); NRAS (4893); PNUTL1 (5413)
	Attributes	MF: hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides †; hydrolase activity †; guanyl nucleotide binding †; GTP binding †; GTPase activity †; catalytic activity †; binding † BP: cellular process †; cellular physiological process †; cell proliferation †; cell growth and/or maintenance †; cell cycle †; small GTPase mediated signal transduction; signal transduction; regulation of cell cycle; intracellular signaling cascade; cell communication CDD: Ras subfamily of RAS small GTPases
17	4 Genes	FANCA (2175); FANCC (2176); FANCE (2178); FANCF (2188)
	Attributes	CC: nucleus †; intracellular †; cell †; cytoplasm BP: response to stress †; response to DNA damage stimulus †; nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; DNA repair †; DNA metabolism †; protein metabolism; protein complex assembly
18	4 Genes	MYH11 (4629); MYH9 (4627); TPM3 (7170); TPM4 (7171)
	Attributes	CC: intracellular †; cytoskeleton †; cytoplasm †; cell †; actin cytoskeleton † MF: protein binding †; cytoskeletal protein binding †; binding †; actin binding †; calmodulin binding; ATP binding; adenyl nucleotide binding BP: morphogenesis †; development †; organogenesis; muscle development; cellular process; cellular physiological process; cell growth and/or maintenance; cell motility CDD: Intermediate filament proteins
19	4 Genes	ARHGAP26 (23092); ARHGEF12 (23365); BCL9 (607); CHN1 (1123)
	Attributes	MF: GTPase regulator activity; GTPase activator activity; enzyme regulator activity; enzyme activator activity; signal transducer activity; binding BP: cellular process †; cellular physiological process; cell growth and/or maintenance; signal transduction; cell communication CDD: GTPase-activator protein for Rho-like GTPases
20	3 Genes	ERCC2 (2068); ERCC3 (2071); TP53 (7157)
	Attributes	CC: nucleus †; intracellular †; cell †; transcription factor complex; nucleoplasm; DNA-directed RNA polymerase II, holoenzyme MF: protein binding †; nucleic acid binding †; hydrolase activity †; DNA binding †; catalytic activity †; binding †; ATP binding †; adenyl nucleotide binding †; metal ion binding; hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides; helicase activity; DNA helicase activity; DNA-dependent ATPase activity; ATP-dependent helicase activity; ATP-dependent DNA helicase activity; ATPase activity, coupled

Table 1 (*Continued*)

Consensus cluster		LocusLink gene symbol (LocusID) and GO/CDD attributes
		BP: transcription, DNA-dependent †; transcription †; response to stress †; response to DNA damage stimulus †; regulation of transcription, DNA-dependent †; regulation of transcription †; regulation of programmed cell death †; regulation of apoptosis †; programmed cell death †; positive regulation of apoptosis †; nucleotide-excision repair †; nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; induction of programmed cell death †; induction of apoptosis †; DNA repair †; DNA metabolism †; cellular process †; cellular physiological process †; apoptosis †; transcription from Pol II promoter; transcription-coupled nucleotide-excision repair; sensory perception of mechanical stimulus; perception of sound; organismal physiological process; cell proliferation; cell growth and/or maintenance; cell cycle
21	3 Genes	BLM (641); RECQL4 (9401); WRN (7486)
	Attributes	CC: nucleus †; intracellular †; cell † MF: nucleic acid binding †; hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides †; hydrolase activity †; helicase activity †; DNA helicase activity †; catalytic activity †; binding †; ATP-dependent helicase activity †; ATP binding †; ATPase activity, coupled †; adenyl nucleotide binding †; DNA binding BP: nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; DNA metabolism †; response to stress; response to DNA damage stimulus; DNA repair; development CDD: Helicase superfamily c-terminal domain †; DEAD/DEAH box helicase. Members of this family include the DEAD and DEAH box helicases. Helicases are involved in unwinding nucleic acids. The DEAD box helicases are involved in various aspects of RNA metabolism, including nuclear transcription, pre mRNA †; ATP-dependent DNA helicase[Replication, recombination and repair] †
22	3 Genes	CLTC (1213); COL1A1 (1277); PICALM (8301)
	Attributes	CC: intracellular †; cytoplasm †; cell †; plasma membrane; membrane MF: structural molecule activity BP: transport †; cellular process †; cellular physiological process †; cell growth and/or maintenance †

"Genes" specifies the number of genes in a consensus cluster. "Attributes" lists influential attributes, attributes that occur with a frequency ≥ 0.5 in a consensus cluster containing ≥ 2 genes and 1.0 otherwise († signifies an attribute with frequency 1.0). The abbreviations are as follows: CC, GO Cellular Component term; MF, GO Molecular Function term; BP, GO Biological Process term; and CDD, Conserved Domain Database protein domain.

Table 2
Consensus clusters and influential attributes for the 159 genes in the Aging data set

Consensus cluster		LocusLink gene symbol (LocusID) and GO/CDD attributes
0	100 Genes	1F813 (172164); 1I156 (172513); 3L221 (176427); 4G426 (177345); age-1 (174762); ced-3 (178272); che-11 (179666); che-13 (186607); che-2 (180401); che-3 (172593); clk-1 (175729); clk-2 (176065); clk-3 (267067); ctl-1 (259738); daf-1 (176829); daf-10 (184883); daf-12 (181263); daf-16 (172981); daf-18 (176869); daf-19 (174577); daf-2 (175410); daf-28 (267111); daf-4 (175781); daf-6 (181584); daf-9 (180889); daz-1 (173931); drd (45844); eat-1 (177701); eat-10 (267157); eat-13 (267158); eat-18 (190914); eat-2 (175072); eat-3 (174476); eat-5 (172480); eat-6 (179796); eat-7 (267154); egl-4 (176991); fem-3 (177734); fer-15 (267268); fog-1 (171895); fog-2 (267275); fog-3 (172863); gcy-18 (178237); gcy-6 (191644); GH1 (2688); glp-1 (176286); gro-1 (175725); hsp-6 (178873); ins-1 (177936); ins-7 (191688); isp-1 (177609); lbp-7 (191701); let-363 (172167); lrs-2 (172499); mec-1 (188259); mes-1 (181362); mev-1ANDced-9 (260040); nuc-1 (181174); osm-1 (181715); osm-3 (177141); osm-5 (180585); osm-6 (179631); pdk-1 (180475); pgl-1 (177461); rad-8 (268088); Rdh (38449); scl-1 (186053); sir-2.1 (177924); spe-10 (268146); spe-26 (177929); tax-2 (172723); tax-4 (176297); tkr-1 (192055); unc-1 (180458); unc-13 (172497); unc-15 (172491); unc-2 (180570); unc-20 (268221); unc-24 (177594); unc-25 (176713); unc-26 (178284); unc-29 (172703); unc-30 (178265); unc-31 (178233); unc-32 (176257); unc-35 (268223); unc-4 (174544); unc-46 (268226); unc-47 (176431); unc-54 (259839); unc-6 (180961); unc-64 (176743); unc-7 (181608); unc-76 (180014); unc-78 (180631); unc-79 (175523); unc-80 (268236); vit-1 (181034); vit-2 (180781); vit-5 (180630)
	Attributes	None
1	22 Genes	Beach1 (33819); Cct1 (117353); CETP (1071); cher (42066); Ef1alpha100E (43736); fwd (45374); HSPA9B (3313); Indy (40049); Kl (16591); KL (9365); l(3)DTS3 (249462); LMNA (4000); MTP (4547); PARK2 (5071); Pcmt (40668); Plau (18792); POSH (36990); Rgn (19733); Sug (39314); Terc (21748); UCHL1 (7345); VhaSFD (34997)
	Attributes	CC: cell MF: catalytic activity BP: metabolism; development

Table 2 (*Continued*)

Consensus cluster		LocusLink gene symbol (LocusID) and GO/CDD attributes
2	13 Genes	ATM (472); BLM (641); CKN1 (1161); EcR (35540); Ercc2 (13871); ovo (31429); p53 (42722); Pit1 (18736); Prop1 (19127); Rpd3 (38565); TERT (7015); Trp53 (22059); WRN (7486)
	Attributes	CC: nucleus; intracellular; cell MF: binding; nucleic acid binding; DNA binding; transcription regulator activity BP: nucleobase, nucleoside, nucleotide and nucleic acid metabolism †; metabolism †; transcription, DNA-dependent; transcription; regulation of transcription; response to stimulus; regulation of transcription, DNA-dependent; DNA metabolism; response to stress; response to DNA damage stimulus
3	5 Genes	Eip71CD (39675); Hsp68 (42852); Msra (110265); Prdx1 (18477); Sod2 (20656)
	Attributes	MF: oxidoreductase activity; catalytic activity BP: metabolism †; response to stress; response to stimulus; response to biotic stimulus; protein metabolism; oxygen and reactive oxygen species metabolism
4	4 Genes	Akt1 (41957); chico (64880); InR (42549); Pi3K92E (42446)
	Attributes	CC: cell; plasma membrane; membrane MF: transferase activity; signal transducer activity; phosphotransferase activity, alcohol group as acceptor; kinase activity; catalytic activity; binding; protein kinase activity BP: transmembrane receptor protein tyrosine kinase signaling pathway †; signal transduction †; regulation of organ size †; regulation of growth †; regulation of cellular process †; regulation of cell size †; regulation of cell growth †; regulation of body size †; positive regulation of organ size †; positive regulation of growth †; positive regulation of cell size †; positive regulation of cell growth †; morphogenesis †; metabolism †; insulin receptor signaling pathway †; development †; cellular morphogenesis †; cell surface receptor linked signal transduction †; cell growth and/or maintenance †; cell growth †; cell communication †; response to stimulus; positive regulation of body size; phosphorylation; organogenesis; death; response to biotic stimulus; protein modification; protein metabolism; protein amino acid phosphorylation; organismal physiological process; determination of adult life span; aging
5	4 Genes	APOE (348); Cat (40048); Sod (39251); Sod2 (36878)
	Attributes	CC: intracellular †; cytoplasm †; cell † MF: antioxidant activity †; oxidoreductase activity; catalytic activity; superoxide dismutase activity; binding BP: response to stimulus †; response to biotic stimulus †; oxygen and reactive oxygen species metabolism †; organismal physiological process †; metabolism †; development †; death †; determination of adult life span; defense response; aging; superoxide metabolism; response to stress; response to oxidative stress; cell communication
6	3 Genes	APP (351); PSEN1 (5663); PSEN2 (5664)
	Attributes	CC: plasma membrane †; membrane †; intracellular †; integral to plasma membrane †; integral to membrane †; Golgi apparatus †; endoplasmic reticulum †; cytoplasm †; cell †; nucleus; nuclear inner membrane; chromosome MF: protein binding; binding BP: signal transduction †; programmed cell death †; death †; cell growth and/or maintenance †; cell communication †; apoptosis †; intracellular signaling cascade; DNA replication and chromosome cycle
7	3 Genes	Ghr (14600); Ghrhr (14602); Npy (24604)
	Attributes	CC: extracellular space †; extracellular †; membrane; integral to membrane; cell MF: transmembrane receptor activity †; signal transducer activity †; receptor activity † BP: signal transduction; G-protein coupled receptor protein signaling pathway; cell surface receptor linked signal transduction; cell communication
8	3 Genes	bsk (44801); hep (32256); puc (40958)
	Attributes	MF: catalytic activity †; transferase activity; signal transducer activity; receptor signaling protein serine/threonine kinase activity; receptor signaling protein activity; protein serine/threonine kinase activity; protein kinase activity; phosphotransferase activity, alcohol group as acceptor; kinase activity

Table 2 (*Continued*)

Consensus cluster		LocusLink gene symbol (LocusID) and GO/CDD attributes
		BP: signal transduction †; protein modification †; protein metabolism †; protein kinase cascade †; ovarian follicle cell development (sensu Insecta) †; oogenesis (sensu Insecta) †; oogenesis †; morphogenesis of embryonic epithelium †; morphogenesis of a polarized epithelium †; morphogenesis †; micropyle formation †; metabolism †; MAPKKK cascade †; JNK cascade †; intracellular signaling cascade †; insect chorion formation †; establishment of tissue polarity †; establishment of planar polarity †; embryonic development (sensu Insecta) †; dorsal closure †; dorsal appendage formation †; development †; cell growth and/or maintenance †; cell communication †; response to stress; response to stimulus; protein amino acid phosphorylation; phosphorylation; organogenesis; cellular morphogenesis
9	2 Genes	Igf1r (16001); Insr (16337)
	Attributes	CC: membrane †; integral to membrane †; cell † MF: transmembrane receptor protein tyrosine kinase activity †; transmembrane receptor activity †; transferase activity †; signal transducer activity †; receptor activity †; protein-tyrosine kinase activity †; protein kinase activity †; phosphotransferase activity, alcohol group as acceptor †; kinase activity †; catalytic activity †; binding †; ATP binding †; adenyl nucleotide binding † BP: transmembrane receptor protein tyrosine kinase signaling pathway †; signal transduction †; protein modification †; protein metabolism †; protein amino acid phosphorylation †; phosphorylation †; organogenesis †; morphogenesis †; metabolism †; development †; cell surface receptor linked signal transduction †; cell communication † CDD: Receptor L domain. The L domains from these receptors make up the bilobal ligand binding site. Each L domain consists of a single-stranded right hand beta-helix. This Pfam entry is missing the first 50 residues of the domain †; Furin-like cysteine rich region †; Fibronectin type 3 domain †

See Table 1 for a description of the format.

transcriptional regulation" (Futreal et al., 2004). The automatically-generated consensus clusters and influential attributes in Table 1 were inspected to create an overview of the cancer-related genes (an overview of the aging-related genes in Table 2 is presented later). Since this subjective process can highlight different facets of the data, the summaries given below should be viewed as one set of possible synopses. Unless indicated otherwise, the attributes listed (phrases in Courier font) are GO terms.

- Transcription, DNA-dependent is associated with six consensus clusters (102 genes) that are distinguished by their binding partner and biological role: consensus cluster 1, DNA binding and cell growth and/or maintenance; 2, DNA binding and development; 4, protein binding and transcription coactivator activity; 9, ubiquitin ligase complex and protein modification; 13, signal transduction and cell growth and/or maintenance; and 20, response to DNA damage stimulus and regulation of apoptosis.
- Signal transducer activity or signal transduction is associated with six consensus clusters (52 genes) that are distinguished by their catalytic activity, location and role: 3, protein-tyrosine activity, CDD tyrosine kinase, catalytic domain. Phosphotransferases, cell surface receptor linked signal transduction, and cell growth and/or maintenance; 7, protein serine/threonine kinase activity, and intracellular signaling cascade; 8, membrane; 13, transcription; 16, GTPase activity and CDD Ras subfamily of RAS small GTPases; and 19, GTPase activator activity. Note that two of these clusters (3 and 7) are associated also with the more specific function of protein kinase activity.
- Cell cycle is associated with five consensus clusters (44 genes) that are distinguished by their ocation and role: 5, nucleus; 6, membrane; 15, binding; 16, GTPase activity; and 20, transcription.
- Response to stress or DNA repair is associated with four consensus clusters (27 genes) that are distinguished by their binding partner and role: 5, DNA binding and cell cycle; 17, protein complex assembly; 20, regulation of transcription; and 21, DNA helicase activity, CDD helicase superfamily c-terminal domain.
- Transport or transporter activity is associated with three consensus clusters (12 genes) that are distinguished by their location and transported molecule: 11, protein metabolism and nuclear pore; 14, mitochondrion and electron transporter activity; and 22, cytoplasm.
- Miscellaneous (13 genes). 9, ubiquitin ligase complex; 18, cytoskeleton and morphogenesis.

The Aging consensus clusters and influential attributes shown in Table 2 are consistent with the generally accepted view that aging-related proteins are components of genomic instability, oxidative damage, developmental, immune, and neuroendocrine pathways. One set of synopses is as follows.

- `Response to stress` is associated with three consensus clusters (22 genes) that are distinguished by their activity and role: 2, `response to DNA damage stimulus`; 3, `oxygen and reactive oxygen species metabolism`; and 5, `oxygen and reactive oxygen species metabolism`, `development`, and `death`.
- `Signal transducer activity` or `signal transduction` is associated with five consensus clusters (15 genes) that are distinguished by their activity, location, and role: 4, `transmembrane receptor protein tyrosine kinase signaling pathway` and `regulation of cell size`; 6, `golgi apparatus` and `apoptosis`; 7, `G-protein coupled receptor protein signaling pathway` and `extracellular`; 8, `receptor signaling protein serine/threonine kinase activity` and `morphogenesis of embryonic epithelium`; and 9, `transmembrane receptor protein tyrosine kinase signaling pathway`, `development`, and CDD `Fibronectin type 3 domain`. Note that three of the clusters (4, 8, and 9) are associated with the more specific process of `protein kinase activity`.
- `Transcription, DNA-dependent` and `DNA binding` is associated with consensus cluster 2 (13 genes).

KEGG pathways annotated to the Cancer and Aging genes but not used during construction of attribute profiles provide additional insights (complete information can be found in Supplementary Material). Both data sets include genes linked to `MAPK signaling` and `Cytokine–cytokine receptor interaction`. A pathway unique to Aging is `neurodegenerative disorders` (UCHL1 and PARK2 in consensus cluster 1; APOE in 5; APP, PSEN1, and PSEN2 in 6).

3.2. Cancer and Aging genes involved in protein (mis)folding and turnover

Inspection of individual genes and their annotations allows features associated with a small fraction of genes in a data set but overlooked by ensemble attribute clustering to be identified. As illustration, attention here is restricted to the emerging areas of protein misfolding and turnover in cancer and aging (reviewed in Sangster et al., 2004; Söti and Csermely, 2003). For example, Hsp90 possesses the capacity to uncover a highly pleiotropic range of phenotypes (Sangster et al., 2004) and *D. melanogaster* molecular chaperones such as Hsp70 and its JNK-inducible relative Hsp68 are believed to have repair functions downstream of JNK signaling. Aging consensus cluster 8 (Table 2) consists of three genes in the JNK cascade that are believed to extend lifespan by confering tolerance to oxidative stress (Wang et al., 2003).

Both the Cancer and Aging data sets contain molecular chaperones, molecules that assist in the folding of proteins, prevent the aggregation of proteins and nucleotides, and participate in the elimination of ubiquitinated molecules. The Cancer data set contains two heat shock protein (Hsp) genes: HSPCA (heat shock 90 kDa protein 1, α) and HSPCB (heat shock 90 kDa protein 1, β). The Aging data set contains three Hsp genes: *C. elegans* hsp-6 (heat shock protein, 70.8 kD), *D. melanogaster* Hsp68 (heat shock protein 68), and HSPA9B (heat shock 70 kDa protein 9B, PBP74, mortalin-2). Whether cancer and aging are, respectively, most closely linked to the Hsp90 and Hsp70 families remains to be determined.

More Cancer genes than Aging genes appear to be associated with the ubiquitin cycle and protein catabolism. Thirteen Cancer consensus genes are annotated with GO terms containing the word "ubiquitin" or CDD domains with the phrase "protein turnover" (nine are in consensus cluster 9): ASPSCR1, BIRC3, BRCA1, CBL, CYLD, GAS7, NSD1, PML, TFRC, TIF1, TRIM33, WHSC1, and WHSC1L1. Only four Aging consensus genes meet these same criteria: Msra, PARK2, Prdx1, and UCHL1.

3.3. Cell and tissue polarity: new candidates for genes involved in aging and age-related diseases

In addition to being components of the JNK cascade, the three *D. melanogaster* genes in Aging consensus cluster 8 are involved in planar cell polarity, that is, the creation of certain polarized epithelial tissue (*bsk*, basket; *hep*, hemipterous; *puc*, puckered). Tissue morphogenesis and function require the coupling of individual epithelial cell apical-basolateral polarity to the extracellular environment, resulting in an additional axis of polarity within the epithelium. In an epithelium, the apical surfaces of cells face the lumen whereas the basolateral surfaces contact adjacent cells and the underlying connective tissue. One mechanism for establishing and maintaining cell polarity is protein and lipid trafficking in biosynthetic, recycling and transcytosis pathways involving the endoplasmic reticulum (ER), golgi apparatus, and vesicles (reviewed in Mostov and ter Beest, 2003). Disruption of cell polarity has been implicated in neoplastic transformation of neuroepithelial cells (Klezovitch et al., 2004). The creation of planar cell polarity in *D. melanogaster* and the process of vertebrate gastrulation and neurulation both involve the Frizzled pathway (Mlodzik, 2002; Van Aelst and Symons, 2002). Extracellular polarity signals interact with the membrane receptor Frizzled, and a key downstream effector of Frizzled-mediated signaling is the JNK cascade. In breast tissue, the precise organization and

arrangement of epithelial cells, myoepithelial cells, stromal cells (fibroblasts, adipocytes, immune effector cells, cells of the vascular system), the basement membrane and extracellular matrix is necessary for functional integrity and disruption of such tissue polarity, especially the spatial relationship of epithelial and myoepithelial cells, is intimately linked to cancer (reviewed in Bissell et al., 2003). These observations and those discussed below suggest that the JNK cascade and tissue polarity may be molecular and architectural hubs with key roles not only in cancer and aging, but also in progenitor/stem cell lineage determination and embryonic morphogenesis.

Mammalian equivalents of three components of the *D. melanogaster* JNK cascade are new candidates for genes that may influence lifespan. The human homologs of *hep, bsk*, and *puc* are, respectively, MAP2K7 (mitogen-activated protein kinase kinase 7; alias MKK7), MAPK8 (mitogen-activated protein kinase 8; alias JNK), and LOC285193 (similar to RIKEN cDNA 0710001B24). The first two, MAP2K7 and MAPK8, are well studied members of the JNK cascade. Apart from the LOC285193 gene encoding a protein with a dual specificity phosphatase domain, this human homolog of *puc* remains largely uncharacterized. The *puc* product has protein–protein interactions with components of many signaling pathways, including Frizzled, Wnt receptor, TGF-β receptor, BMP receptor, EGF receptor, transmembrane receptor protein tyrosine kinase, smoothened, and torso. This suggests that *hep*/MAP2K7, *bsk*/MAPK8, and *puc*/LOC285193 may coordinate and integrate multiple intra- and extracellular processes via space- and time-dependent interactions with proteins in other pathways.

The LOC285193 product may have a role in age-related diseases and tissue regeneration. *puc* encodes a JUN kinase phosphatase and negative regulator of JNK cascade that is involved in actin cytoskeleton organization and biogenesis, embryonic epidermal differentiation, peripheral nervous system development, ovarian follicle cell development, and wound healing. The latter function may involve Notch signaling because the *puc* product has protein-protein interactions with *wg*, a morphogen that is a ligand for the Notch and Frizzled-2 surface receptors. In mice, insufficient activation of the receptor Notch-1 by its ligand Delta has been linked to the loss of regenerative potential of old skeletal muscle (Conboy et al., 2003). The Jagged1/Notch pathway is believed to be a mediator of TGF-β induced epithelial-to-mesenchymal transition (EMT) in mammary gland, kidney tubule and epidermal epithelia (Zavadil et al., 2004). EMTs underlie the cell plasticity required in embryonic development and are observed frequently in advanced carcinogenesis. Alagille syndrome is a multisystem, autosomal dominant disorder with highly variable expressivity caused by mutations in the Jagged1 (JAG1) gene. Investigation of the tissue and ligand-specific relationship(s) between LOC285193 and Notch signaling may yield insights into the biology of normal, malignant and aged tissues.

3.4. Phenotype position effects and comparative genomics: new candidates for cancer- and aging-related genes

The phenomenon of phenotype position effects was investigated by using the May 2004 UCSC build of the Golden Path to determine the genomic coodinates of all genes in the human genome, identifying the locations of the 291 human Cancer and 20 human Aging genes, and examining genes in the intervening regions (interlopers). Although at least two Cancer genes were found on every chromosome, only 13 of the 23 chromosomes contained one or more Aging genes (complete information on the 311 genes can be found in Supplementary Material). The Cancer gene IRF4 (interferon regulatory factor 4) is the penultimate gene on the p arm of chromosome 6 and thus the closest gene in the data sets to a telomere. Two genes in a data set that were separated by 0 and 1 interlopers were defined as ‘‘contiguous’’ and ‘‘consecutive’’ pairs respectively. A pair was deemed ‘‘isolated’’ if, on both sides, it was flanked by five or more genes not in the data set. The number of interlopers for the Aging and Cancer data sets were in the ranges 26–699 and 0–273, respectively. For the Aging data set, the different distribution across chromosomes and the absence of contiguous pairs may be due to its smaller size.

Examination of contiguous and consecutive pairs yields a variety of anecdotal evidence for both the existence of phenotype position effects and the functions of neighboring genes. The notation ‘‘X–Z’’ refers to two genes in the human genome that are next to each other in terms of linear order and where ‘‘X’’ and ‘‘Z’’ are genes in the Cancer or Aging data sets. Similarly, ‘‘X–[Y]–Z’’ refers to a situation where there is one interloper, the gene ‘‘Y’’. Table 3 is a synthesis of information on four isolated contiguous pairs of Cancer genes. One hypothesis regarding CDX2–FLT3 and FANCC–PTCH is that CDX2 and FANCC are effectors of the signaling initiated by their respective adjacent receptors, FLT3 and PTCH. In Table 1, cancer consensus cluster 2 (CDX2) may include genes that are ligands (small molecules) or downstream effectors (transcription factors) of the transmembrane receptor protein kinases in consensus cluster 3 (FLT3). Similarly, consensus cluster 17 (FANCC) may include genes that respond to signals received by membrane proteins in consensus cluster 6 (PTCH).

Table 4 describes isolated consecutive pairs and includes predictions for interlopers. Similarly, Table 5 discusses interleaved contiguous and consecutive pairs. Combining phenotype position effects with comparative genomics not only provides enhanced understanding of proteins that have been characterized previously, but also generates predictions about open reading frames (FANCD2–[C3ORF10]–VHL; BCL6–[FLJ42393]–LPP).

Aging and Cancer genes form clusters in the same genomic region. Table 6 gives information on the pair of genes with the fewest interlopers, ATM–[KDELC2]–[SLAC2-B]–DDX10. The Aging genes UCHL1 and KL link some of the contiguous and consecutive pairs discussed

Table 3
A systematic examination of isolated contiguous pairs in the Cancer data set

Location	LocusLink gene symbol (description) and user-generated summary
	EVI1 (ecotropic viral integration site 1)–MDS1 (myelodysplasia syndrome 1).
3q26.2	EVI1 binds DNA, inhibits c-Jun N-terminal kinase (JNK), and is involved in megakaryocytic differentiation.
3q26.2	MDS1 is a transcription factor associated with Myelodysplasia syndrome-1 (a chromosomal translocation breakpoint in the region between these genes is associated with leukemia).
	TLX3 (T-cell leukemia, homeobox 3)–NPM1 (nucleophosmin, nucleolar phosphoprotein B23, numatrin).
5q35.1	TLX3 is an orphan homeobox gene associated with T-cell acute lymphoblastic leukaemia.
5q35.1	NPM1 is an NF-κB co-activator for the induction of the human SOD2 gene, regulates p53, is expressed at low levels in lymphoblasts from patients with Fanconi anemia, and is associated with acute promyelocytic leukaemia.
	FANCC (Fanconi anemia, complementation group C)–PTCH (patched homolog, Drosophila).
9q22.32	FANCC regulates expression of genes involved in myeloid differentiation and inflammation, cooperates with Hsp70 to support survival of hematopoietic cells, and FANCC-deficient hematopoietic stem/progenitor cells exhibit aberrant cell cycle control. Fanconi anemia is characterized by bone marrow failure, cancer predisposition and DNA damage hypersensitivity.
9q22.32	PTCH is a receptor for sonic hedgehog (a secreted molecule implicated in the formation of embryonic structures) and is associated with various carcinomas.
	CDX2 (caudal type homeobox transcription factor 2)–FLT3 (fms-related tyrosine kinase 3).
13q12.2	CDX2 is a transcription factor, has a role in integrating pathways controlling embryonic axial elongation and anterior/posterior patterning, and is associated with gastric cancer.
13q12.2	FLT3 is vascular endothelial growth factor receptor involved in cell surface receptor linked signal transduction, and is associated with acute myeloid leukaemia.

Genes are ordered by chromosomal location (the symbol and description are taken from the July 2004 release of LocusLink). The gene-specific summary was created for this work by synthesizing information available in the corresponding LocusLink entry (a shortage of space precludes the inclusion of all citatations to published work, but these and other information can be accessed from the WWW page for a gene whose symbol is listed). Unless otherwise indicated, the genes discussed are human genes.

Table 4
A systematic examination of isolated consecutive pairs in the Cancer data set

Location	LocusLink gene symbol (description) and user-generated summary
	MSH2 (mutS homolog 2, colon cancer, nonpolyposis type 1 E. coli)–[KCNK12 (potassium channel, subfamily K, member 12)]–MSH6 (mutS homolog 6 *E. coli*)).
2p22-p21	MSH2 is a damaged DNA binding protein involved in mismatch repair. It is associated with hereditary nonpolyposis colorectal cancer type 1, ovarian cancer, T-cell lymphoma, and other disorders.
2p16.3	MSH6 is a damaged DNA binding protein involved in mismatch repair. It is associated with hereditary nonpolyposis colorectal cancer type 5, endometrial type ovarian cancer, and other disorders.
2p21	[KCNK12] is a tandem-pore K^+ channel. It may be associated with colorectal or ovarian cancer. Another member of this protein sequence family, KCNK9, is amplified and overexpressed in several types of human carcinomas.
	BCL11A (B-cell CLL/lymphoma 11A, zinc finger protein)–[PAPOLG (poly(A) polymerase gamma)]–REL (v-rel reticuloendotheliosis viral oncogene homolog, avian).
2p16.1	BCL11A is a zinc-finger transcription factor involved in hematopoietic cell differentiation. It is associated with B-cell lymphoma and leukaemia.
2p16.1	REL is a transcription factor in the NF-κB cascade that may be involved in keratinocyte senescence via the induction of MnSOD and hence oxidative stress. It is associated with Hodgkin's lymphoma.
2p16.1	[PAPOLG] may be involved in Notch signaling and morphogenesis, and may be associated with lymphoma or leukemia. The *D. melanogaster* homolog, *hrg* (hiiragi), has protein-protein interactions with members of the Notch signaling pathway, and regulation of hrg level is essential for controlled cytoplasmic polyadenylation and early development.
	HOXD13 (homeobox D13)–[HOXD12 (homeobox D12)]–HOXD11 (homeobox D11).
2q31.1	HOXD13 is transcription factor involved in morphogenesis, and is associated with brachydactyly and synpolydactyly.
2q31.1	HOXD11 is a transcription factor involved in morphogenesis, and is associated with acute myeloid leukaemia.
2q31.1	[HOXD12] is a transcription factor. It may be involved in embryonic patterning. Deletions that remove the HOXD cluster are associated with limb and genital abnormalities.
	FANCD2 (Fanconi anemia, complementation group D2)–[C3ORF10 (chromosome 3 open reading frame 10)]–VHL (von Hippel-Lindau syndrome).
3p25.3	FANCD2 may be involved in DNA damage response and is essential during embryogenesis to prevent inappropriate apoptosis in cells and tissues undergoing high levels of proliferative expansion. Fanconi anemia is associated with congenital abnormalities and a predisposition to cancer.
3p25.3	VHL is a component of a protein complex involved in the ubiquitination and degradation of hypoxia-inducible-factor (HIF), a transcription factor that plays a central role in the regulation of gene expression by oxygen. It is associated with a variety of tumors and congenital polycythemia.

Table 4 (*Continued*)

Location	LocusLink gene symbol (description) and user-generated summary
3p25.3	[C3ORF10] is uncharacterized hematopoietic stem/progenitor cell protein MDS027. It may be involved in intracellular trafficking, be a determinant of cell and tissue polarity, interact with syntaxins, and be associated with embryonic morphogenesis and cancer. The maize homolog, BRK1, promotes polarized growth and division of leaf epidermal cells (Frank and Smith, 2002). The *D. melanogaster* homolog, CG30173/SIP1, has protein-protein interactions with the t-SNARE syntaxin 1A, a protein essential for both synaptic transmission and Golgi-plasma membrane vesicle trafficking. The extracellular morphogen epimorphin/syntaxin-2, a member the syntaxin family of vesicle fusion proteins, directs distinct morphogenic pathways in mammary epithelial cells depending upon the context in which the protein is presented (polar versus apolar) (Radisky et al., 2003).
	BCL6 (B-cell CLL/lymphoma 6, zinc finger protein 51))–[FLJ42393 (FLJ42393 protein)]–LPP (LIM domain containing preferred translocation partner in lipoma).
3q27.3	BCL6 is a transcription factor involved in the inflammatory response of B-cells. It is associated with diffuse large-cell lymphoma.
3q27.3-q28	LPP contains a LIM domain and has a role as a focal adhesion adaptor protein. It is associated with lipoma and myeloid leukemia (in benign and malignant tumors, mutant LPP is permanently localized in the nucleus).
3q27.3	[FLJ42393] may have a role in signal transduction and inflammatory response and may be associated with hemopoietic disorders.
	(TACC3)–FGFR3 (fibroblast growth factor receptor 3, achondroplasia, thanatophoric dwarfism)–[LETM1 (leucine zipper-EF-hand containing transmembrane protein 1)]–WHSC1 (Wolf-Hirschhorn syndrome candidate 1).
4p16.3	FGFR3 is a membrane receptor that binds to acidic and basic fibroblast growth hormones and is involved in bone development and maintenance. It is associated with multiple types of skeletal dysplasia.
4p16.3	WHSC1 is a DNA binding protein involved in transcription regulation, is a component of the ubiquitin ligase complex, and may have a role in early development. It is associated with multiple myelomas, severe mental and developmental defects.
4p16.3	[LETM1] is a membrane protein that binds calcium ions. It may be a component of a signal transduction pathways related to early differentiation of normal tissues, and be associated with hematopoietic stem cell, musculoskeletal, cancer and other disorders. TACC3 (transforming, acidic coiled-coil containing protein 3) encodes a member of the TACC family of microtubule-associated proteins that have roles in transcriptional control, are involved in embryonic development, cell growth and differentiation, and are associated with multiple myeloma, breast and gastric cancer. TACC3 is upregulated when differentiating erythroid progenitor cells are treated with erythropoietin.
	WHSC1L1 (Wolf-Hirschhorn syndrome candidate 1-like 1)–[LETM2 (leucine zipper-EF-hand containing transmembrane protein 2)]–FGFR1 (fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)–(FLJ43582)–(TACC1).
8p12	WHSC1L1 is a DNA binding protein involved in transcription regulation, is a component of the ubiquitin ligase complex, and may have a role in early development.
8p12	FGFR1 is a membrane receptor that binds both acidic and basic fibroblast growth factors and is involved in limb induction. It is associated with Pfeiffer and Jackson-Weiss syndromes.
8p12	[LETM2] may be similar to LETM1 (this 8p12 region is a replication of the 4p16 region discussed above).
	LHFP (lipoma HMGIC fusion partner)–[COG6 (component of oligomeric golgi complex 6)]–FOXO1A (forkhead box O1A, rhabdomyosarcoma).
13q13.3-q14.11	LHFP binds DNA and is associated with lipoma.
13q14.11	FOXO1A is a forkhead transcription factor that may have a role in myogenic growth and differentiation. It is associated with alveolar rhabdomyosarcoma and skeletal muscle atrophy.
13q14.11	[COG6] is a component of a conserved oligomeric Golgi (COG) complex involved in intracellular protein trafficking and glycoprotein modification. It may be involved in the repair and refolding of stress-damaged proteins, have a role in early differentiation, and be associated lifespan and age-related diseases. The *D. melanogaster* homolog, CG1968, has protein-protein interactions with proteins involved in heat shock (Hsp70Ba, Hsp70Bc, Hsp70Bb, Hsp70Aa) and transcription regulation (NC2alpha, lbl, Bap60). Mutation of the COG7 component of the COG complex is linked to type IIe Congenital Disorders of Glycosylation and fibroblasts from such patients exhibit abnormal distribution of COG6.
	WAS (Wiskott-Aldrich syndrome, eczema-thrombocytopenia)–[SUV39H1 (suppressor of variegation 3-9 homolog 1, Drosophila)]–GATA1 (GATA binding protein 1, globin transcription factor 1)–(HDAC6).
Xp11.23	WAS is a small GTPase regulatory/interacting protein expressed in hematopoietic cells and is involved in the transduction of signals from cell surface receptors to the actin cytoskeleton. It is associated with immune dysregulation and microthrombocytopenia.
Xp11.23	GATA1 is a transcription factor involved in erythroid development (switching of fetal to adult hemoglobin). It is associated with X-linked dyserythropoietic anemia with thrombocytopenia and megakaryotic leukaemia.
Xp11.23	[SUV39H1] is a heterochromatic protein that accumulates transiently at centromeric positions during mitosis and is involved in transcriptional regulation via chromatin modification. It may have a role in the survival and proliferation of hematopoietic stem cells and X-lined immunodeficency thrombocytopenia or cancer. The mouse homolog, Suv39h1, cooperates with Smads in BMP-induced repression, has been linked to the epigenetic regulation of telomere length, and forms a complex with a DNA binding protein that acts to regulate the expression of several hematopoietic genes. HDAC6 (histone deaceytlate 6) is a microtubule-associated deacetylase that is linked to the ubiquitin network, and is involved in growth-factor-induced chemotaxis, cell cycle progression and development; the Xp11.22-23 region is characterized by instability in several cancers and neurological disorders.

A gene symbol in square parenthesis denotes a gene not in the Cancer or Aging data set (an interloper), and for which the concept of phenotype position effect is used to make predictions about its function (evidence in support of the hypothesis is provided). Round parentheses denote additional genes of interest. See Table 3 for a description of the format.

Table 5
A systematic examination of interleaved contiguous and consecutive pairs in the Cancer data set

Location	LocusLink gene symbol (description) and user-generated summary
	FIP1L1 (FIP1 like 1, S. cerevisiae)–CHIC2 (cysteine-rich hydrophobic domain 2)–[GSH-2 (homeobox protein GSH-2)]–PDGFRA (platelet-derived growth factor receptor, alpha polypeptide)–KIT (v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog).
4q12	FIP1L1 is a subunit of cleavage and polyadenylation specificity factor that binds to U-rich RNA elements and stimulates poly(A) polymerase. It is associated with hypereosinophilia.
4q12	CHIC2 is a member of the cysteine-rich hydrophobic (CHIC) protein family and is localized to vesicles and plasma membranes. It is associated with acute myeloid leukaemia.
4q12	PDGFRA is a cell surface tyrosine kinase receptor for growth factors that are mitogens for cells of mesenchymal origin. It is associated with somatic gastrointestinal stromal tumors and leukaemia.
4q12	KIT is a transmembrane tyrosine kinase involved in signal transduction. It is associated with gastrointestinal stromal tumors, germ cell tumors, leukemia, and piebaldism.
4q12	[GSH-2] is a transcription factor. It may have a role in defining the identity of progenitor cells and be associated with cancer and neural disorder(s). GSH-2 is ectopically expressed in leukemic cells, and fusion of FIP1L1 and PDGFRA via deletion of the interstitial region is associated with hypereosinophilia. The mouse homolog, Gsh2, is involved in regional patterning of lateral ganglionic eminence progenitor cells and brain development.
	HOXA9 (homeobox A9)–[HOXA10 (homeobox A10)]–HOXA11 (homeobox A11)–HOXA13 (homeobox A13).
7p15.2	HOXA9 is homeobox transcription cluster involved in embryonic development. It is associated with myeloid leukaemia.
7p15.2	HOXA11 is a homeobox transcription factor involved in embryonic and uterine development. It is associated with radioulnar synostosis with amegakaryocytic thrombocytopenia.
7p15.2	HOXA13 is a homeobox transcription factor involved in embryonic and skeletal development. It is associated with hand-foot-genital syndrome.
7p15.2	[HOXA10] is a homeobox transcription factor involved in fertility, embryo viability, and regulation of hematopoietic lineage commitment.

See Table 4 for a description of the format.

Table 6
A systematic examination of the two genes in the Cancer and Aging data sets, DDX10 and ATM, respectively, that are separated by the fewest genes

Location	LocusLink gene symbol (description) and user-generated summary
	ATM (ataxia telangiectasia mutated, includes complementation groups A, C and D)–[KDELC2 (KDEL, Lys-Asp-Glu-Leu, containing 2)]–[SLAC2-B]–DDX10 (DEAD, Asp-Glu-Ala-Asp, box polypeptide 10).
11q22.3	ATM is a PI3/PI4-kinase involved in DNA repair, telomere length regulation, and cell cycle regulation. It is associated with aging, breast cancer, lymphomas, and leukemia.
11q22.3	DDX10 is an ATP-dependent RNA helicase involved in ribosome assembly. It is associated with myeloid malignancies.
11q22.3	[KDEL2] is an uncharacterized protein with a KDEL sequence that retains soluble proteins in the ER. [SLAC2-B] binds the small GTP-binding protein Rab27A, a protein involved in membrane trafficking. KDELC2 and SLAC2-B may be involved in intracellular transport and associated with cancer or aging. The Aging gene ATM may have a role in intracellular vesicle and/or protein transport because it binds to β-adaptin in cytoplasmic vesicles. Ribosomes attached to the surface of the rough ER insert newly synthesized proteins directly into the ER, which then processes them and passes them to the Golgi apparatus.

See Table 4 for a description of the format.

above. UCHL1 is part of a block on 4p, (TACC3)–FGFR3–[LETM1]–WHSC1–[122]–RHOH–[4]–UCHL1. Although WHSC1 is 127 genes upstream of UCHL1, the equivalent Cancer gene in the duplicated and inverted 8p region, WHSC1L1, is 22 genes downstream of the Aging gene WRN (UCHL1 has a role in protein deubiquitination; WHSC1 and WHSC1L1 are components of the ubiquitin ligase complex). KL is part of a block on 13q, CDX2–FLT3–[17]–BRCA2–[4]–KL–[20]–LHFP–[COG6]–FOXO1A (the mouse homolog of KL, Kl, is a type I membrane protein with β-glucuronidase activity that can hydrolyze steroid β-glucuronides; COG6 is involved in glycoprotein modification).

4. Discussion

Ensemble attribute profile clustering-based analysis of 291 cancer- and 159 aging-related genes highlights factors both shared by and unique to these phenotypes. Common processes and pathways include those that maintain the integrity of the genome and cellular milieu (response to stress), cell surface receptor linked and intracellular signal transduction, transcription, cell growth and maintenance, cell communication, and apoptosis. Protein metabolism involving ubiquitin and GTPase related activities seem more prevalent in proteins implicated in cancer whilst response to oxidative damage appears to be more common in proteins

linked to aging. The results suggest that studies of the human homologs of a group of three *D. melanogaster* genes known to influence lifespan may be informative (*hep*/MAP2K7, *bsk*/MAPK8, *puc*/LOC285193). Circumstantial evidence supports the potential of such studies in enhancing knowledge of aging and age-related diseases. In particular, LOC285193 is in a genomic region that has been associated with susceptibility to type II diabetes mellitus (NIDDM1 region; 2q37.3) and its peritelomeric location suggests investigations of a role in cellular senescence may be warranted. The two genes 3′ of LOC285193 are involved in proteolysis and peptidolysis and these three contiguous genes are separated by short intergenic regions of $\leq$10,000 base pairs: LOC285193–[RNPEPL1 (arginyl aminopeptidase, aminopeptidase B,-like 1)]–[CAPN10 (calpain 10)].

A qualitative assessment of phenotype position effects suggests that a feature of genomic regions containing Cancer and Aging genes is the close physical and functional relationship between cell fate determination, pattern formation in an embryo, tumor initiation and progression, lifespan modification, and age-related diseases. The genes and proteins linking these processes include ones with roles in the JNK cascade; Notch signaling; heat shock response; protein, lipid and nucleic acid trafficking; chromatin organization; and mechanical integrity of the nucleus. Despite the panoply of intra- and extracellular mechanisms involved in these phenomena, one feature that connects them is the task of generating and controling tissue polarity.

Investigations of the genes and proteins shown in Tables 3–6, especially the interlopers (genes in square parentheses), may yield insights into the molecular and cellular mechanisms linking progenitor/stem cell lineage determination, embryonic morphogenesis, cancer, and aging. Frequently, extant studies have examined their role(s) in the context of one of these phenomena. For example, the interloper GSH-2 (Table 5) is in a genomic region associated with human leukemia and other cancers (4q12), but the mouse homolog Gsh2 has been studied as a homebox protein involved in brain and central nervous system development, and pattern specification. Candidates for further experimental studies include (i) PAPOLG (poly(A) polymerase γ) which may be involved in the Notch pathway (2p16.1); (ii) C3ORF10 (uncharacterized hematopoietic progenitor cell protein) which may be involved in intracellular trafficking and cell and tissue polarity (3p25.3); (iii) FLJ42393 which may be involved in signal transduction and inflammatory response (3q27.3); (iv) LETM1/LETM2 (membrane calcium binding proteins) which may be involved in differentiation of normal tissues (4p16.3/8p12); (v) COG6 (a component of a conserved oligomeric golgi complex involved in intracellular protein trafficking and glycoprotein modification) which may be involved in the repair and refolding of stress-damaged proteins (13q14.11); and (vi) SUV39H1 (histone methylase and repressor of transcription) which may be involved in hematopoiesis and differentiation (Xp11.23).

A key question raised by this work is how the same palette of genes can contribute to and specify a seemingly disparate plethora of phenotypes. One hypothesis is related to cryptic genetic variation, variation that is not observed in a population under normal conditions but is uncovered by environmental or genetic perturbations (Gibson and Dworkin, 2004). Genomic regions containing the genes in question may be especially responsive to intra- and extracellular signals such as small molecules, serum, xenobiotic agents, heat shock, irradiation, reactive oxygen species, the circadian clock and so on. The encoded proteins may themselves unveil variation that is generally neutral but becomes adaptive or deleterious after specific or non-specific environmental change and/or the introduction of novel alleles. Some cryptic genetic variation may be manifested in the unperturbed state and thus contribute to the maintenance of variation needed for visible traits. Elucidating the pertinent internal and external factors, and ascertaining their spatial and temporal integration should assist in understanding the similarities and differences between progenitor/stem cell lineage determination, embryonic morphogenesis, cancer, and aging.

Acknowledgements

This work was supported by the National Institute on Aging, National Institute of Envirnomental Health Sciences, U.S. Department of Energy (OBER) and California Breast Cancer Research Program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mad.2004.09.028.

References

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. 25, 25–29.

Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E., Lander, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D., Meyerson, M., 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. U.S.A. 98, 13790–13795.

Bissell, M., Rizki, A., Mian, I., 2003. Tissue architecture: the ultimate regulator of breast function. Curr. Opin. Cell Biol. 15, 753–762.

Blei, D., Franks, K., Jordan, M., Mian, I., 2004. Statistical modelling of biomedical corpora: mining the Caenorhabditis Genetic Center Bibliography. Mech. Age. Dev., in press.

Cheeseman, P., Stutz, J., 1996. Bayesian classification (AutoClass): theory and results. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press, pp. 153–180.

Conboy, I., Conboy, M., Smythe, G., Rando, T., 2003. Notch-mediated restoration of regenerative potential to aged muscle. Science 302, 1575–1577.

Frank, M., Smith, L., 2002. A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells. Curr. Biol. 12, 849–853.

Futreal, P., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M., 2004. A census of human cancer genes. Nature Rev. Cancer 4, 177–183.

Gibson, G., Dworkin, I., 2004. Uncovering cryptic genetic variation. Nature Rev. Genet. 5, 681–690.

Hurst, L., Pál, C., Lercher, M., 2004. The evolutionary dynamics of eukaryotic gene order. Nature Rev. Genet. 5, 299–310.

Klezovitch, O., Fernandez, T., Tapscott, S., Vasioukhin, V., 2004. Loss of cell polarity causes severe brain dysplasia in lgl1 knockout mice. Genes Dev. 18, 559–571.

Marchler-Bauer, A., Anderson, J., DeWeese-Scott, C., Fedorova, N., Geer, L., He, S., Hurwitz, D., Jackson, J., Jacobs, A., Lanczycki, C., Liebert, C., Liu, C., Madej, T., Marchler, G., Mazumder, R., Nikolskaya, A., Panchenko, A., Rao, B., Shoemaker, B., Simonyan, V., Song, J., Thiessen, P., Vasudevan, S., Wang, Y., Yamashita, R., Yin, J., Bryant, S., 2003. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. 31, 383–387.

Mlodzik, M., 2002. Planar cell polarization: do the same mechanisms regulate Drosophila tissue polarity and vertebrate gastrulation? TRENDS Genet. 18, 564–571.

Moler, E., Chow, M., Mian, I., 2000a. Analysis of molecular profile data using generative and discriminative methods. Physiol. Genom. 4, 109–126.

Moler, E., Radisky, D., Mian, I., 2000b. Integrating naïve Bayes models and external knowledge to examine copper and iron homeostasis in Saccharomyces cerevisiae. Physiol. Genom. 4, 127–135.

Mostov, K.T.S., ter Beest, M., 2003. Polarized epithelial membrane traffic: conservation and plasticity. Nature Cell Biol. 5, 287–293.

Radisky, D., Hirai, Y., Bissell, M., 2003. Delivering the message: epimorphin and mammary epithelial morphogenesis. Trends Cell Biol. 13, 426–434.

Salton, G., 1988. Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer. Addison-Wesley.

Sangster, T., Lindquist, S., Queitsch, C., 2004. Under cover: causes, effects and implications of Hsp90-mediated genetic capacitance. BioEssays 26, 348–362.

Semeiks, J., Rizki, A., Bissell, M., Mian, I., 2004. Ensembl attribute clustering: discovering and characterizing groups of genes with similar patterns of functional and structural features, in preparation.

Shatkay, H., Feldman, R., 2003. Mining the biomedical literature in the genomic era: an overview. J. Comput. Biol. 10, 821–855.

Söti, C., Csermely, P., 2003. Aging and molecular chaperones. Exp. Gerontol. 38, 1037–1040.

Van Aelst, L., Symons, M., 2002. Role of Rho family GTPases in epithelial morphogensis. Genes Dev. 16, 1032–1054.

Wang, M., Bohmann, D., Jasper, H., 2003. JNK signaling confers tolerance to oxidative stress and extends lifespan in Drosophila. Dev. Cell 5, 811–816.

Zavadil, J., Cermak, L., Soto-Nieves, N., Bottinger, E., 2004. Integration of TGF-beta/Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition. EMBO J. 23, 1155–1165.