This information has not been peer-reviewed. Responsibility for the findings rests solely with the author(s).

Deposited research article

Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22

Karine Mégy, Stéphane Audic and Jean-Michel Claverie

Addresses: Genomic and Structural Information, UMR 1998 CNRS / Aventis, 31, chemin Joseph Aiguier, 13402 Marseille Cedex 20 , France.

Correspondence: Jean-Michel Claverie. E-mail: Jean-Michel.Claverie@igs.cnrs-mrs.fr

Posted: 10 January 2003

Genome ***Biology*** 2003, **4**:P1

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/2/P1

© 2003 BioMed Central Ltd

Received: 6 January 2003

This is the first version of this article to be made available publicly. This article was submitted to *Genome* ***Biology*** for peer review.

.deposited research

AS A SERVICE TO THE RESEARCH COMMUNITY, GENOME **BIOLOGY** PROVIDES A 'PREPRINT' DEPOSITORY TO WHICH ANY ORIGINAL RESEARCH CAN BE SUBMITTED AND WHICH ALL INDIVIDUALS CAN ACCESS FREE OF CHARGE. ANY ARTICLE CAN BE SUBMITTED BY AUTHORS, WHO HAVE SOLE RESPONSIBILITY FOR THE ARTICLE'S CONTENT. THE ONLY SCREENING IS TO ENSURE RELEVANCE OF THE PREPRINT TO GENOME **BIOLOGY**'S SCOPE AND TO AVOID ABUSIVE, LIBELLOUS OR INDECENT ARTICLES. ARTICLES IN THIS SECTION OF THE JOURNAL HAVE **NOT** BEEN PEER-REVIEWED. EACH PREPRINT HAS A PERMANENT URL, BY WHICH IT CAN BE CITED. RESEARCH SUBMITTED TO THE PREPRINT DEPOSITORY MAY BE SIMULTANEOUSLY OR SUBSEQUENTLY SUBMITTED TO GENOME **BIOLOGY** OR ANY OTHER PUBLICATION FOR PEER REVIEW; THE ONLY REQUIREMENT IS AN EXPLICIT CITATION OF, AND LINK TO, THE PREPRINT IN ANY VERSION OF THE ARTICLE THAT IS EVENTUALLY PUBLISHED. IF POSSIBLE, GENOME **BIOLOGY** WILL PROVIDE A RECIPROCAL LINK FROM THE PREPRINT TO THE PUBLISHED ARTICLE.

Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22

Karine Mégy, Stéphane Audic and Jean-Michel Claverie[1]

Karyn.Megy@igs.cnrs-mrs.fr, Stephane.Audic@igs.cnrs-mrs.fr, Jean-Michel.Claverie@igs.cnrs-mrs.fr
Genomic and Structural Information - UMR 1998 CNRS / Aventis - 31, chemin Joseph Aiguier - 13402 Marseille Cedex 20 - France - tel : +33(0)4.91.16.45.48 - fax : +33(0)4.91.16.45.49

[1] : corresponding author

Abstract

Background: Clusters of genes co-expressed are known in prokaryotes (operons) and were recently described in several eukaryote organisms, including Human. According to some studies, these clusters consist of housekeeping genes, whereas other studies suggest that these clustered genes exhibit similar tissue specificity. Here we further explore the relationship between co-expression and chromosomal co-localization in the human genome by analyzing the expression status of the genes along the best-annotated chromosomes 20, 21 and 22.

Methods: Gene expression levels were estimated according to their publicly available ESTs and gene differential expressions were assessed using a previously described and validated statistical test. Gene sequences for chromosomes 20, 21 and 22 were taken from the *Ensembl* annotation.

Results: We identified clusters of genes specifically expressed in similar tissues along chromosomes 20, 21 and 22. These co-expression clusters occurred more frequently than expected by chance and may thus be biologically significant.

Conclusion: The co-expression of co-localized genes might be due to higher chromatin structures influencing the gene availability for transcription in a given tissue or cell type .

Background

Since the publication of two "complete" first drafts of the human genome [1, 2], a huge continuing effort is being made to annotate the human genome. Whereas some regions remain poorly annotated, the exact positions of most - protein coding - genes are now defined. This allows the systematic analysis of the influence of the position of genes on various of their properties, such as their expression level and tissue distribution. The positional clustering of co-expressed genes is common in prokaryotes (operons) and was recently described in *Saccharomyces cerevisiae* [3], in *Caenorabditis elegans* [4, 5] and in *Drosophila melanogaster* [6, 7]. Throughout the human genome, it is often supposed that genes are randomly distributed, except for tandem duplicates. However, clusters of highly expressed genes were recently revealed in the Human genome [8, 9]. To date, no clear functional relationships between genes in these clusters have been identified and their biological meaning, if any, is yet to be determined.

Two studies were carried out on the expression level of sets of co-localized human genes. Caron *et al*. [8] analyzed the gene expression profiles for any chromosomal regions in various tissue types (Human Transcriptome Map). The genes studied corresponded to about 24,000 UniGene clusters and expression levels were estimated from 12 SAGE libraries made in different conditions. This study revealed about 50 large regions, called RIDGEs (*Region of IncreaseD Gene Expression*), showing a clustering of highly expressed genes. A similar study by Lercher *et al.* [9] (based on 11,000 UniGene clusters and 14 SAGE libraries) suggested that such RIDGEs might mostly consist of housekeeping genes and no clusters of genes with similar tissue expression profiles were identified.

In order to specifically analyze tissue specific expression, other studies were based on sets of genes expressed in a given tissue. Gabrielsson *et al.* [10] performed a micro-array analysis of genes expressed in the adipose tissue. Mapping these genes back on the human genome, revealed clusters

of adipose tissue specific genes on chromosomes 11, 19 and 22. Using ESTs, Dempsey *et al.* [11] focused on genes from chromosomes 21 and 22 expressed in the cardio-vascular system (CVS). They showed some chromosomal clustering of these genes. Bortoluzi *et al.* [12] performed a similar study on genes expressed in the skeletal muscle. They identified positional clusters of skeletal muscle genes on chromosomes 17, 19 and X. Finally, an EST analysis of the murine placenta by Ko *et al*. [13] identified clusters of placenta specific genes on chromosomes 2, 7, 9 and 17. Overall, these studies suggest that clusters of tissue specific genes do exist, and might be more frequent than initially thought.

Previous studies were based on the whole set of genes expressed in a particular tissue, irrespective of the behavior of these genes in other tissues. In order to evaluate the clustering of genes specifically expressed in any tissue - not specified in advance -, we performed a comprehensive analyzed of the expression profiles of all genes identified along human chromosomes 20, 21 and 22. These chromosomes were chosen as the most complete and best annotated available human chromosomes. For each gene, we first estimated the expression level in various tissues from the public EST database and then computed the probability of differential expression in each tissue. We then compared these probabilities with those calculated for the neighboring genes and looked for a succession of genes specifically over-expressed (SOGs) in a given tissue. This procedure revealed more of such clusters than expected at random.

Results

Relationship between Gene Expression and Tissue type

The following analyses were based on the number of specifically expressed genes (SEGs) in each tissue category and on chromosomes 20, 21 and 22 (using a *p-value* > 0.90). Tissue categories were pooled in three groups according to their origin: a diseased group (DIZ), a healthy and infant group (INF), and a healthy and adult group (ADLT).

Chromosome analysis. On each chromosome, 80% of the genes were found differentially expressed in at least one tissue category. The same proportion was found by Su *et al.* [14] in an analysis of the human transcriptome map. The remaining 20% represents genes ubiquitously expressed (i.e. housekeeping genes), or weakly expressed genes. The expression level of such genes - represented by low EST numbers - cannot be reliably estimated nor their differential expression status.

Genes with erratic expression levels. The number of tissue types associated with significant differential expression ($p > 0.90$) was estimated for each gene. We noticed that some genes were statistically identified as "differentially expressed" in more than 50% of the tissue types (Table 2). Our statistical test is performed by comparing the number of cognate ESTs found for each library type to the number found for all other library types aggregated as one virtual "average" tissue. With this procedure, genes exhibiting expression levels fluctuating highly above or below the average (over all the other tissues), may appear significantly differentially over- or under- expressed in numerous libraries. The genes we found exhibiting this erractic behavior were all highly expressed, corresponding to a large number of ESTs such as ribosomal proteins, known to be found in all tissues. This strongly suggests that the erratic EST counts (from almost none, to much higher than average) has an artifactual origin, e.g. an untold "normalization" procedure. Indeed, it is (and was) customary for a number of EST sequencing laboratories not to record (or even not to pick the clones corresponding to) the many instances of the most abundant transcripts (such as ribosomal proteins, elongation factor EF-Tu, and the like). This *ad hoc* -but not consistent- subtraction of the most abundant ESTs (even though the libraries are not normalized) is the most probable cause for the corresponding gene to appear either over- or under-expressed in many tissues. We thus removed them from our subsequent analyses.

Correlation Islands

We searched for correlation islands, defined as clusters of at least three successive SOGs in a common tissue (see Material and methods). Nine, 5 and 17 clusters of SOGs were found for chromosomes 20, 21 and 22, respectively. To assess the statistical significance of these results, we computed the probability of finding such a number of clusters under a random permutation of the gene order along the chromosomes. This probability was found to be very low (Table 3). We can thus confidently conclude that there are more clusters that expected by chance, and further explore their potential biological meaning.

The functional annotation of these gene clusters is shown in Table 4. No functional correlation was identified within the clusters, but such a correlation would be hard to establish given the lack of a defined function for many of the genes.

Two clusters (III and IV on chromosome 22) were each composed of three genes, two of them being annotated as having exactly the same function. These genes are from computer prediction and may correspond to a single gene erroneously interpreted as two different genes. As this particularity concern only two clusters, we did not consider them further.

Discussion

The analysis of all genes along chromosomes 20, 21 and 22, identified clusters of co-expressed genes (e.g. the known immunoglobulin cluster), and genes expressed in every tissue (e.g. some ribosomal proteins). The visualization of SEGs in various conditions allowed expression variations to be detected in diseased *vs.* healthy or infant *vs.* adult tissues. For instance, we noticed that a small cluster of immunoglobulin apparently specific of ovary diseased tissues. These immunoglobulins may be involved in an immune response specific to this pathology.

ESTs were grouped according to the tissue type: organ, developmental and pathological states. While comparing the gene expression across adult healthy tissues is biologically meaningful, comparing gene expression across pathological states (DIZ group) is more problematic as it

involves treating different pathological conditions as one. For instance, different cancer types - each with its specific expression patterns- may arise in the same organ [15]. In principle, only diseased tissues corresponding exactly to the same disorder should be pooled. When dealing with diseased tissues, our protocol was thus expected to provide a distorted view of their gene expression patterns.

As in all statistical studies, sample size is important. As less fetal/infant libraries were available, less fetal or infant tissue specific gene clusters were detected.

In a study of Drosophila gene clusters, Spellman *et al.* [6] found that no functional relationship could be detected between the genes within a cluster. Our study again failed to reveal any relationships between the gene forming co-expressed/co-localized clusters. However, the large proportion of genes with no defined function is not allowing any final conclusion to be drawn.

Chromatin is usually described as been divided into "open" domains, where genes have the potential to be expressed, and domains of "closed" regions, where gene expression is shut down. The existence of co-expressed/co-localized gene clusters is consistent with a model where large chromatin regions would change their activity (openness) status in a tissue specific manner, allowing neighboring genes to be transcribed or shut down in a coordinated way. Such a model, confirmed by our study, has been around for quite sometimes, although experimental evidence have been obtained for only a few tissues and cell types [16, 17].

Materials and methods

EST and Libraries

Human ESTs were obtained from dbEST (release Oct.2001) [18]. Pooled, subtracted or normalized libraries were removed from the study. The remaining 1270 libraries were classified in three groups: 489 libraries from diseased tissues, whatever their developmental stage (DIZ), 194 libraries from healthy fetal or infant tissues (INF) and 587 libraries from healthy adult tissues (ADLT). The classification was made with the data extracted from the 'keywords' and 'developmental stage' fields of the library description. A similar analysis was performed on the three groups. ESTs were then

masked for vector, common repeats and low complexity sequences using RepeatMasker (URL: repeatmasker.genome.washington.edu/.) and Repbase [19]. After these steps, 2,251,840 ESTs remained: 1,147,369 in the diseased libraries group, 478,320 in the infant libraries group and 626,151 in the adult libraries group. In each group, libraries were categorized into 40 organs as described in Table 1. Each library was classified in a tissue category if at least one of its keywords characterized the tissue category. Libraries were individually characterized by keywords extracted from the library description, in the 'lib', 'keywords', 'tissue description', 'tissue type', and ‘cell type‘ and ’organ’ fields. Tissue categories were individually characterized by representative keywords, such as the name of the category or its synonyms. A library could only belong to a single category. Finally, the classification was visually verified. Categories with less than 1,000 ESTs were removed. The numbers of EST for every tissue category in the groups DIZ, INF and ADLT are shown in Table 1. The list of the libraries composing each tissue category is given as supplementary data.

Genes on Chromosomes 20, 21 and 22

Gene sequences were downloaded from *Ensembl* (release Nov.2001). As we analyzed the gene expression along the chromosomes, the various transcripts of a single gene were not considered. We used both known and novel genes predicted by *Ensembl.* Respectively, 694, 243 and 595 gene sequences were found for chromosomes 20, 21 and 22. As the *Ensembl s*equence identifier might change from one release to the next, a correspondence between the *Ensembl* sequences and the NCBI sequences is given in Table 4.

Gene Expression Profiles

Every gene was compared to the total EST set of the corresponding group at high stringency (%identity> 95 and match length > 66% of the query sequence) with BLAST 2.2.1. The expression profile was derived from the cognate ESTs in each tissue category relative to the total number of

ESTs in the tissue category of this group. All expression profiles were stored in a matrix with rows corresponding to genes and columns corresponding to tissue categories. The M*ij* element thus correspond to the relative frequency of cognate ESTs for gene *i* in tissue category *j*.

Differential Gene Expression

To assess its differential expression in a tissue category and for a given group (DIZ, INF or ADLT), every contig was compared to the total EST set of this group at high stringency (previously described matrix). The hit list of cognate matches was then separated in two groups: ESTs from the corresponding tissue category *vs*. any other tissue categories. The statistical significance of the difference in frequencies between these two groups was computed according to a previously published formula [20]. The groups of diseased, infant or adult tissues were treated independently.

Correlation Island

Correlation islands were considered as clusters of at least three successive SOGs (p-value > 0.90) in the same tissue category. To assess the biological meaning of these clusters, we estimated the probability of finding such a number of clusters under a randomization of the gene position along the chromosomes (5,000 randomizations). The probabilities are presented in Table 3.

Acknowledgements

KM was supported by a grant from the Region Provence-Alpes Cote d'Azur and AVENTIS Pharma. We thank Deborah Byrne for reading the manuscript.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W*, et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA*, et al.*: **The sequence of the human genome**. *Science* 2001, **291**:1304-1351.

3. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression**. *Nat Genet* 2000, **26**:183-186.

4. Blumenthal T: **Gene clusters and polycistronic transcription in eukaryotes**. *Bioessays* 1998, **20**:480-487.

5. Roy PJ, Stuart JM, Lund J, Kim SK: **Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans**. *Nature* 2002, **418**:975-979.

6. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome**. *J Biol* 2002, **1**:5.

7. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI: **Large clusters of co-expressed genes in the Drosophila genome**. *Nature* 2002, **420**:666-669

8. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA*, et al.*: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains**. *Science* 2001, **291**:1289-1292.

9. Lercher MJ, Urrutia A, OHurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome**. *Nat Genet* 2002, **31**:180-183.

10. Gabrielsson BL, Carlsson B, Carlsson LM: **Partial genome scale analysis of gene expression in human adipose tissue using DNA array**. *Obes Res* 2000, **8**:374-384.

11. Dempsey AA, Pabalan N, Tang HC, Liew CC: **Organization of human cardiovascular-expressed genes on chromosomes 21 and 22**. *J Mol Cell Cardiol* 2001, **33**:587-591.

12. Bortoluzzi S, Rampoldi L, Simionati B, Zimbello R, Barbon A, d'Alessi F, Tiso N, Pallavicini A, Toppo S, Cannata N*, et al.*: **A comprehensive, high-resolution genomic transcript map of human skeletal muscle**. *Genome Res* 1998, **8**:817-825.

13. Ko MS, Threat TA, Wang X, Horton JH, Cui Y, Pryor E, Paris J, Wells-Smith J, Kitchen JR, Rowe LB*, et al.*: **Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome**. *Hum Mol Genet* 1998, **7**:1967-1978.

14. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A*, et al.*: **Large-scale analysis of the human and mouse transcriptomes**. *Proc Natl Acad Sci U S A* 2002, **99**:4465-4470.

15. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT*, et al.*: **A gene expression database for the molecular pharmacology of cancer**. *Nat Genet* 2000, **24**:236-244.

16. Armstrong JA, Emerson BM. **Transcription of chromatin: these are complex times.** *Curr*

Opin Genet Dev 1998, **8**:165-172.

17. Akashi K, He X, Chen J, Iwasaki H, Niu C, Steenhard B, Zhang J, Haug J, Li L. **Transcriptional accessibility for multi-tissue and multi-hematopoietic lineage genes is hierarchically controlled during early hematopoiesis.** *Blood* 2002 (in press).

18. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags"**. *Nat Genet* 1993, **4**:332-333.

19. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements**. *Trends Genet* 2000, **16**:418-420.

20. Audic S, Claverie J.-M.: **The significance of digital gene expression profiles**. *Genome Res* 1997, **7**:986-995.

Tables

Table 1. Number of ESTs and keywords characterizing the tissue types.

Tissues		Keywords	INF	DIZ	ADLT
Stem_cell		stem hematopoietic	-	-	1,592
Embryo		whole body trophoblast	10,346	-	-
Placenta		placent	110,502	43,395	-
Bone		bone osteo ewings	-	16,551	7,407
Cartilage		cartilage	-	5,017	5,224
Ear		cochlea ear	14,966	-	-
Eye		eye retina retino cornea ocular	6,370	28,235	30,036
Skin		skin keratinocyt melanoma melanocytic psoriasis derm	-	93,511	10,457
Cardio-vascular	Heart	myocardium valve heart cardiac	20,979	-	8,263
	Vascular	artery vein aorta endotheli blood vascul hemopoietic venous venae huvec platelet	-	-	39,737
Head_neck		mouth tongue tonsil head oral cavity gingiva	5,817	2,628	5,787
Muscle_skeletal		skeletal rhabdomyosarcom	-	5,442	40,819
Soft_tissue		adipose peritoneum omentum synovium fibroblast connective synovial epithelioid fibrosarcoma liposarcoma	3,721	4,182	5,692
Endocrine	adrenal	adrenal	-	16,620	10,483
	thyroid	thyroid parathyroid	-	4,294	5,125
	pineal	pineal pituitary	-	12,155	17,458
Exocrines_breast		breast mammary nipple areola	-	31,204	22,379
Respiratory		lung pleura trachea bronchi larynx pharynx nasal nasopharyn laryngal bronchi laryngeal olfact	4,406	82,164	38,108
Digestive	liver	liver hepato bile gallbladder	90,480	38,103	43,846
	saliv	salivary parotid paratid	-	11,193	-
	pancreas	pancrea langerhans	-	61,178	17,042
	stomach	stomach gastric gastro	-	23,580	-
	esophagus	esophag buccal	-	3,807	-
	bowel	bowel intestine appendix cecum colon duodenu ileum jejunum colit	-	70,282	9,933
Genitourinary	femal_ovary	ovary oviduct ovarian	-	65,196	6,106
	femal_uterus	uterus endometrium pregnant exocervical	-	10,5749	985
	femal_others	cervix fallopian	-	29,380	-
	testis	testis epididym seminal vesicle	-	25,484	6,034
	prostate	prostate	-	55,609	22,785
	Urinaire kidney	kidney wilms	1,427	57,891	13,776
	Urinaire others	bladder ureter urethra	-	20,736	-
	others	gonad genitourina wolffian germ germinal mesonephros Mullerian paramesonephros urogenital	-	11,930	4,341
Lymphoreticular	immuno	leukocyte monocyt macrophag t-cell lymph b-cell bone marrow leukemi leukaemia mononuclear myelo lymphoblastoid T-lymphoc B-lymphoc	-	75,877	72,618
	spleen	spleen	93,645	-	4,460
	thymus	thymus	4,680	-	-

Table 1. Number of ESTs and keywords characterizing the tissue types.

Tissues		Keywords	INF	DIZ	ADLT
Central Nervous System	brain	brain amygdala medulla oblongata cerebrum cortex frontal occipital hippocampus cerebellum corpus callosum basal ganglia striatum globus pallidus putamen caudate substantia nigra subthalamic tectum prosencephalon diencephalon thalamus corpora mesencephalon mesencephali quadrigemina glioblastoma astrocyt neuron cranial dura mater	108,992	99,576	61,843
	dorsal root ganglion	dorsal root	-	-	1,520
Peripheric Nervous System		sympathetic nerve nervous spinal neuronal dendrit	-	18,360	22,456
Neural System		neuro	-	26,463	-
NO_TISSU_Sp			68	662	88,540

'DIZ' is for diseases tissues, 'INF' for infant and foetal healthy tissues and 'ADLT' for adult healthy tissues. '-' represents categories with less than 1,000 ESTs.

Table 2. Genes from chromosomes 20, 21 and 22 expressed in most of the tissues.

Chromosomes	*Ensembl* gene identifiant	Function of the gene
chromosome 20	ENSG00000132668	ribosomal protein
chromosome 21	ENSG00000128093	cyclophilin A
chromosome 22	ENSG00000128327	ribosomal protein
	ENSG00000128360	ribosomal protein
	ENSG00000100316	ribosomal protein

Table 3. Number of clusters for chromosomes 20, 21 and 22

	Real Nc	Random Nc	Probability
chr.20	9	5.2	3.8×10^{-2}
chr.21	5	1.7	2.8×10^{-2}
chr.22	17	6.8	8×10^{-4}

Number of clusters at real position (Real Nc), mean number of clusters at random position (Random Nc) and probability to find the actual number of clusters by chance.

Table 4a. Clusters of successive genes from chromosome 20.

Chr.20	Tissue Category	*Ensembl* gene identifiant (chromosomic order)	Gene function	NCBI gene identifiant
I.	PNS_ADLT	ENSG00000132646	proliferating cell nuclear Ag (PCNA) - cyclin	NM_002592
		ENSG00000101290	CDP-diacylglycerol synthase 2 (CDS2)	Y16521
		ENSG00000149345	ubiquitin-conjugating enzyme E2D 3	NM_003340
II.	Breast_DIZ	ENSG00000101339	N-acetyltransferase 5 (ARD1 homolog, yeast)	NM_016100
		ENSG00000101343	Crn, crooked neck-like 1 (CGI-201)	NM_016652
		ENSG00000089101	no defined function	HSJ1178H5
III.	Eye_ADLT	ENSG00000125966	matrix metalloproteinase 24 (MMP24)	XM_047216
		ENSG00000126005	integrin beta 4 binding protein	BC019305
		ENSG00000125965	groADLTh differentiation factor 5 (GDF5)	NM_000557
		ENSG00000126001	centrosomal protein 2 (CEP2 = C-Nap1)	NM_007186
IV.	Genito_	ENSG00000125995	no defined function	AK000548
	urinair_	ENSG00000131051	splicing factor	NM_004902
	other_DIZ	ENSG00000126002	no defined function	XM_087888
V.	Testis_ADLT	ENSG00000124177	KIAA protein	XM_029763
		ENSG00000149593	no defined function	NM_032221
		ENSG00000149598	no defined function	AY034072
VI.	Pineal_DIZ	ENSG00000100982	no defined function	XM_053387
		ENSG00000124137	hypothetical C2H2 zinc finger protein	NM_022095
		ENSG00000100985	matrix metalloproteinase 9 (gelatinas collagenas)	NM_004994
VII.	Respiratory	ENSG00000130706	cell membrane glycoprotein (surface antigen)	BC003059
	INF	ENSG00000130702	laminin alpha5 chain precursor (forte expr. lung)	NM_005560
		ENSG00000130705	ribosomal protein	NM_001024
VIII.	Eye_DIZ	ENSG00000088876	No defined function	HSJ734P14
		ENSG00000101361	Nucleolar protein	XM_044915
		ENSG00000101365	isocitrate dehydrogenase 3 (NAD+) beta	NM_006899
IX.	Immuno	ENSG00000101146	RAE1 - mRNA export protein	NM_003610
	ADLT	ENSG00000132819	Seb4D(CLL-associated antigen KW-5)	AF432218
		ENSG00000124097	chromosomal protein	AF076674

The function and the *Ensembl* and NCBI identification numbers are given for each gene.

Table 4b. Clusters of successive genes from chromosome 21.

Chr.21	Tissue Category	*Ensembl* gene identifiant (chromosomic order)	Gene function	NCBI gene identifiant
I.	Pancreas ADLT	ENSG00000099582	adenovirus receptor	NM_001338
		ENSG00000099583	BTG family - mb 3 (antiproliferative protein)	NM_000606
		ENSG00000099585	no defined function	NM_017447
II.	Genital femal	ENSG00000099580	no defined function (similar to a rat kinase)	NM_017447
	others DIZ	ENSG00000023067	heat shock transcription factor 2 binding ptn	NM_007031
		ENSG00000128150	H2B histone family	NM_080593
		ENSG00000099597	no defined function	XM_035973
III.	NS DIZ	ENSG00000099522	no defined function	NM_032261
		ENSG00000099524	lanosterol synthase	NM_002340
		ENSG00000074707	germinal associated nuclear protein	AJ01009
IV.	Respiratory	ENSG00000023120	Phosphofructokinase, liver (PFKL)	XM_036042
	ADLT	ENSG00000099439	candidate gene for APECED	HSY11392
		ENSG00000099440	transient recept potential cation channel TRPM2	XM_009803
V.	Pineal_DIZ	ENSG00000099500	collagen, type VI, alpha 1 (COL6A1)	NM_001848
		ENSG00000099505	collagen, type VI, alpha 2 (COL6A2)	NM_001849
		ENSG00000139071	collagen, type VI, alpha 2 (COL6A2)	XM_086775

The function and the *Ensembl* and NCBI identification numbers are given for each gene.

Table 4c. Clusters of successive genes from chromosome 22.

Chr.22	Tissue Category	*Ensembl* gene identifiant (chromosomic order)	Gene function	NCBI gene identifiant
I.	Lympho-reticular ADLT	ENSG00000128256	immunoglobulin lambda gene	HSIGLV
		ENSG00000128275	immunoglobulin lambda gene	D87016
		ENSG00000128280	immunoglobulin light chain Vl1	HSU03902
		ENSG00000100089	immunoglobulin lambda light chain	HSZ85009
		ENSG00000128273	immunoglobulin lambda light chain	HUMIGLVF
II.	Lympho-reticular ADLT	ENSG00000128299	immunoglobulin light chain	HUMIGLZI
		ENSG00000128291	immunoglobulin lambda gene	D86994
		ENSG00000128265	immunoglobulin lambda light chain	HSZ85032
III.	Kidney_DIZ Pineal_ADLT	ENSG0000099964	macrophage migration inhibitory factor	HSMMIHFA
		ENSG0000099974	D-dopachrome tautomerase	HSU84143
		ENSG0000099977	D-dopachrome tautomerase	HSU49785
IV.	Skin_DIZ	ENSG00000100099	no defined function	HS1048E94
		ENSG00000100104	similar to mouse tuftelin-interacting protein 10	HS1048E9A
		ENSG00000100109	similar to mouse tuftelin-interacting protein 10	HS1048E9A
		ENSG00000100118	high-mobility group (nonhistone K.al) ptn 1	NM_002128
V.	Eye_ADLT	ENSG00000100284	target of myb1 (chicken) (TOM1)	NM_005488
		ENSG00000100292	heme oxygenase (decycling) 1 (HMOX1)	NM_002133
		ENSG00000100297	P1-Cdc46	HSP1CDC46
VI.	Spleen_ADLT	ENSG00000128340	no defined function	HS151B14
		ENSG00000100051	Rac3 (small G protein)	AF008591
		ENSG00000100055	cytohesin-4 (CYT4)	AF075458
VII.	Exocrine_breast_ DIZ	ENSG00000100097	galectin	HSLEC14K
		ENSG00000100101	no defined function	HS37E167
		ENSG00000100106	no defined function	HS37E16
VIII.	CNS_brain_DIZ	ENSG00000100106	no defined function	HS37E16
		ENSG00000138967	H1 histone family,member 0 (H1F0)	NM_005318
		ENSG00000100116	glycine C-acetyltransferase (GCAT)	XM_009974
IX.	CNS_brain ADLT	ENSG00000100311	platelet-derived groADLTh factor beta plptd	XM_009997
		ENSG00000100316	ribosomal protein	NM_000967
		ENSG00000100321	Synaptogyrin 1 (SYNGR1)	XM_009999
X.	Immuno ADLT	ENSG00000100389	ribosomal protein	XM_050589
		ENSG00000100393	E1A binding protein p300 (EP300)	XM_010013
		ENSG00000100395	H-l(3)mbt-like protein	HSA305227
XI.	Bone_DIZ	ENSG00000100387	Ring-box 1	BC017370
		ENSG00000100389	ribosomal protein	XM_050589
		ENSG00000100393	E1A binding protein p300 (EP300)	XM_010013
XII.	CNS_brain DIZ	ENSG00000100399	no defined function	HS756G23
		ENSG00000100401	RanGTPase activating ptn	HsRanGAP1
		ENSG00000100403	ubiquitous tetratricopptd containg ptn RoXaN	AF188530
XIII.	Skin_DIZ	ENSG00000100401	RanGTPase activating ptn	HsRanGAP1
		ENSG00000100403	ubiquit. tetratricopptd containg ptn RoXaN	AF188530
		ENSG00000100410	no defined function	HS223H9
		ENSG00000100412	nuclear aconitase	HSU80040
		ENSG00000100413	no defined function	XM_039448
		ENSG00000100417	phosphomannomutase	HSU86070
		ENSG00000100419	Thyroid autoantigen 70kD (Ku antigen)	BC018259
		ENSG00000100138	RNA binding protein (OTK27)	AF155235
XIV.	Breast_DIZ Bowel_ADLT	ENSG00000100294	similar to fatty acid synthase	XM_040148
		ENSG00000100300	peripheral benzodiazepine receptor related	HumBenza
		ENSG00000100304	no defined function	NM_015140
XV.	Skin_DIZ	ENSG00000100344	no defined function	HS796I171
		ENSG00000100347	no defined function	XM_043614
		ENSG00000022364	beta parvin (PARVB = CLINT = affixin)	NM_013327
XVI.	Skin_ADLT	ENSG00000100201	DEA-box protein p72	HSU59321
		ENSG00000100211	no defined function	HS508I15

		ENSG00000100216	Tom22 – mitochondrial import receptor	AB041906
XVII.	Brain_INF	ENSG00000100423	no defined function	AF131851
		ENSG00000100425	bromodomain containing protein 1	NM_014577
		ENSG00000100426	no defined function	XM_010055

The function and the *Ensembl* and NCBI identification numbers are given for each gene.