BMC Genomics

Research article

Analysis of a human brain transcriptome map

Ping Qiu[1], Lawrence Benbow[1], Suxing Liu[2], Jonathan R Greene[1] and Luquan Wang*[1]

Address: [1]Bioinformatics Group and Human Genomic Research Department, Schering-Plough Research Institute, 2015 Galloping Hill Road, Kenilworth, New Jersey 07033, USA and [2]Tumor Biology Department, Schering-Plough Research Institute, 2015 Galloping Hill Road, Kenilworth, New Jersey 07033, USA

E-mail: Ping Qiu - ping.qiu@spcorp.com; Lawrence Benbow - lawrence.benbow@spcorp.com; Suxing Liu - suxing.liu@spcorp.com; Jonathan R Greene - jonathan.greene@spcorp.com; Luquan Wang* - luquan.wang@spcorp.com

*Corresponding author

Published: 16 April 2002

BMC Genomics 2002, **3**:10

Received: 21 December 2001
Accepted: 16 April 2002

This article is available from: http://www.biomedcentral.com/1471-2164/3/10

Abstract

Background: Genome wide transcriptome maps can provide tools to identify candidate genes that are over-expressed or silenced in certain disease tissue and increase our understanding of the structure and organization of the genome. Expressed Sequence Tags (ESTs) from the public dbEST and proprietary Incyte LifeSeq databases were used to derive a transcript map in conjunction with the working draft assembly of the human genome sequence.

Results: Examination of ESTs derived from brain tissues (excluding brain tumor tissues) suggests that these genes are distributed on chromosomes in a non-random fashion. Some regions on the genome are dense with brain-enriched genes while some regions lack brain-enriched genes, suggesting a significant correlation between distribution of genes along the chromosome and tissue type. ESTs from brain tumor tissues have also been mapped to the human genome working draft. We reveal that some regions enriched in brain genes show a significant decrease in gene expression in brain tumors, and, conversely that some regions lacking in brain genes show an increased level of gene expression in brain tumors.

Conclusions: This report demonstrates a novel approach for tissue specific transcriptome mapping using EST-based quantitative assessment.

Background

Sequencing of Expressed Sequence Tags (ESTs) has resulted in the rapid identification of expressed genes [1]. ESTs are single-pass, partial sequences of cDNA clones from a large number of disease and normal tissue libraries. ESTs have been used extensively for gene discovery and for transcript mapping of genes from a wide number of organisms [2–4]. Even with the finished working draft of the human genome, the generation of a complete and non-redundant catalog of human genes is still a big challenge facing the genome research community. Full-length cDNA data are currently available for only 10,000 human genes [5], less than one-third of the total using the most conservative recent estimates of human gene numbers [6,7]. Evidence of differential expression is one of the most important criteria in prioritizing the exploitation of genes in both academic and pharmaceutical research [8–10].

While identifying individual differentially expressed genes attracts most of the interest, a genome wide transcriptome map may not only provide a tool to identify candidate genes that are over-expressed or silenced in certain disease tissue, but may also help to understand the structure and organization of the genome. Genomes are the blueprints of life and they should not be considered as a simple collection of genes. In fact, the organization of genes into operons, complex regulons [11], or pathogenicity islands [12] suggests that related functions usually share physical proximity. Different types of transcriptome maps can help to identify different types of transcription domains. Those domains can now be analyzed as to how they relate to known nuclear substructures, such as nuclear speckles, PML bodies and coiled bodies [13–15].

Two strategies have been commonly used to evaluate large-scale gene expression: experimental and computational. The former is represented by DNA microarray technology [16]. Computational methods consist of generating a large number of random ESTs from non-normalized cDNA libraries. The variation in the relative frequency of those tags, stored in databases, are then used to point out the differential expression of the corresponding genes: this is the so called "digital Northern" comparison. Digital Northern data can be used to provide quantitative assessment of differential expression within a certain limit [17]. Velculescu et al. [18] introduced another digital method called serial analysis of gene expression (SAGE). The SAGE method requires only nine nucleotides, therefore allowing a larger throughput. In both protocols, the number of tags is reported to be proportional to the abundance of cognate transcripts in the tissue or cell type used to make the cDNA library.

The recently announced first draft of the human genome [19,20] holds in it an unprecedented wealth of information, available for public study and scrutiny. How are genes organized in the human genome? Is there any distribution pattern of tissue specific genes in terms of chromosomal location? In this study, we combined the concept of digital Northern and transcript mapping for all public and Incyte LifeSeq ESTs to evaluate the tissue specific transcriptome. The goal of this paper is not to evaluate the digital expression of individual genes; instead we are looking at the tissue enriched digital expression level for a given chromosomal region. Particularly, we looked at the distribution pattern of brain-enriched genes in the genome and how that pattern changes in brain tumor tissues. We are well aware of the fact that this method and associated approaches are quite primitive. However, the tissue specific transcriptome data strongly suggest that human genome organization is correlated to the tissue type and its dynamics.

Results

Distribution of brain-enriched genes along the chromosomes

With the unavailability of the complete annotated human gene catalog, it is not practical to document each individual gene that is expressed in brain within one chromosome region. Since the number of sequence tags is reported to be proportional to the abundance of cognate transcripts in the tissue or cell type used to make a given cDNA library, the number of ESTs within a chromosome region should reflect the abundance of the cognate transcripts in that region. Therefore comparison of the abundance of brain tissue derived transcripts relative to those from other tissues within the same chromosome region can highlight regions that have more brain-enriched gene expression.

We performed digital expression analysis of brain-enriched genes across the human genome with a window size of 5 Mbp and an interval of 1 Mbp. The transcript density factor for normal (non-tumor) brain libraries (TDF_{NB}) was calculated as described in Methods. Figure 1 is an example of the distribution of TDF_{NB} over chromosome 1 using publicly available EST sequences from dbEST and reveals a number of "peaks" that represent transcripts that appear to be preferentially expressed in brain tissues. To check the validity of these peak regions and make sure that the difference is not due to random picking or partial sequencing of cDNA libraries (which is the common random fluctuation caused by digital Northern approach) [17], the analysis was repeated using ESTs and the associated library information from the Incyte Genomics LifeSeq database. The distribution pattern of TDF_{NB} shows an overall correlation coefficient of 0.658 for the whole genome between these two data sources. If we only analyze the region with Z-score >= 2 (i.e. peak regions), the correlation coefficient is 0.935 which suggests that the peak regions resulting from the analysis of public data are most likely not artifacts. Figure 2 shows the comparison of the distributions of TDF on chromosome 1 calculated from ESTs derived from brain tissue libraries vs. ESTs derived from breast tissue libraries. The overall pearson correlation coefficient for these two tissues is 0.113 which suggests that the peak regions observed in Figure 1 are brain specific.

There are 16 high TDF_{NB} regions (enriched with brain specific expression) with Z-score >= 2 (Table 1). To assess the validity of our finding using public data, similar analysis using Incyte LifeSeq data shows that the same peaks can be derived (data not shown). Table 2 summarizes all the low TDF_{NB} regions (lack of brain specific expression) over the whole genome with Z-score >= 2. It's interesting to note that the majority of high TDF_{NB} or low TDF_{NB} regions have close to average gene density indicating that

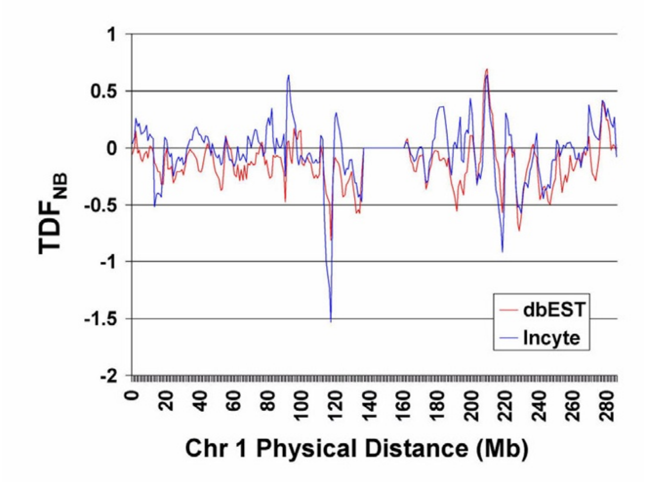

Figure 1
Comparison of distribution of TDF_{NB} calculated from data derived from dbEST and Incyte Genomics LifeSeq for chromosome 1. Pearson Correlation Coefficient = 0.658.

those regions are not biased toward extremely high or low gene density.

Expression profile change of some chromosome regions with extreme TDF_{NB} in brain tumor

Our analysis strongly suggests that brain-enriched genes are distributed throughout the genome in a non-random fashion. Some regions are dense with brain-enriched genes or brain specific expression. It would be interesting to know if any of these patterns change during tumorgenesis. A similar analysis was performed using ESTs generated from brain tumor libraries and their digital expression profile relative to the pooled tissue was plotted against the genome. The chromosomal distributions of these putative brain tumor enriched transcripts and the normal brain enriched transcripts are quite different. Table 3 lists all the chromosome regions with high TDF in non-tumor brain libraries (TDF_{NB}) which become low TDF or neutral TDF in brain tumor (TDF_{TB}). Chr15, 21–25 Mbp, Chr12, 85–89 Mbp, and Chr18, 45–52 Mbp (Figure 3) are some of the examples. While most of the low TDF regions in normal brain remain low in brain tumor, a few regions did become high TDF regions in brain tumor tissues (Table 3). Chr2, 93–99 Mbp and Chr19, 53–58 Mbp (Figure 4) are two examples. The digital expression profile in those regions was further confirmed by using data from Incyte LifeSeq (data not shown).

Discussion

A genome is not a simple collection of genes. It has been reported that significant correlation exists between the distribution of genes along the chromosome and the physical architecture of the cell in bacteria [21]. The human genome is much more complex and a complete understanding of its organization awaits completion of the finished sequence as well as a definitive annotation of the human gene catalog. Considerable evidence has already shown that related genes tend to exist as clusters in the genome. For example, 80% of the over 900 olfactory genes

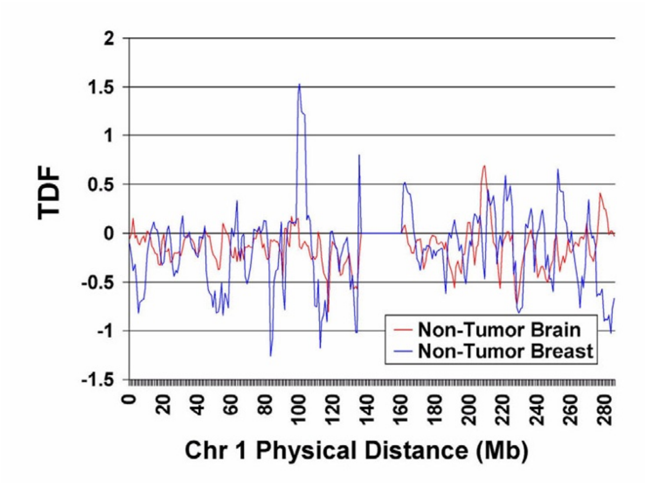

Figure 2
Comparison of distribution of TDF on chromosome 1 calculated from ESTs derived from brain tissue libraries vs. ESTs derived from breast tissue libraries. Pearson Correlation Coefficient = 0.113.

Table 1: Summary of chromosome regions with significant brain-enriched gene expression (high TDF_{NB} region) (window size 5 Mbp, Z-score >= 2.0). All regions are confirmed by separate analysis using Incyte LifeSeq ESTs. Pearson Correlation Coefficient = 0.935.

Chromosome	Region(Mbp)	TDF_{NB}	Gene Density Ratio
1	208–212	0.68	0.33
2	14–18	0.46	0.76
3	90–96	0.39	0.58
5	184–188	0.30	1.0
6	82–86	0.32	0.79
6	138–142	0.28	0.67
7	73–77	0.56	0.59
10	145–149	0.43	0.1
12	85–89	0.36	0.75
13	65–69	0.35	0.71
15	21–26	0.69	0.78
18	80–87	0.88	0.83
20	8–15	0.54	0.52
22	46–50	0.30	0.17
X	103–109	0.43	0.58
X	150–155	0.57	0.25

Table 2: Summary of chromosome regions lacking in brain-enriched gene expression (low TDF_{NB} regions) (window size 5 Mbp, Z-score >= 2.0). All regions are confirmed by separate analysis using Incyte LifeSeq ESTs. Pearson Correlation Coefficient = 0.935.

Chromosome	Region (Mbp)	TDF_{NB}	Gene Density Ratio
1	117–111	-0.80	1.11
1	218–222	-0.56	0.50
1	229–233	-0.62	0.79
2	92–99	-0.99	1.79
3	175–179	-0.76	1.0
3	217–211	-0.97	0.76
4	75–81	-1.31	2.78
7	152–159	-0.80	1.26
9	117–122	-0.76	0.99
10	57–61	-0.82	0.62
11	55–59	-1.24	0.25
11	111–117	-0.92	0.94
12	55–60	-0.71	4.15
14	111–116	-1.08	0.81
19	30–35	-0.60	0.41
19	53–57	-0.62	1.58
21	7–11	-1.83	0.03
22	10–14	-0.43	0.10

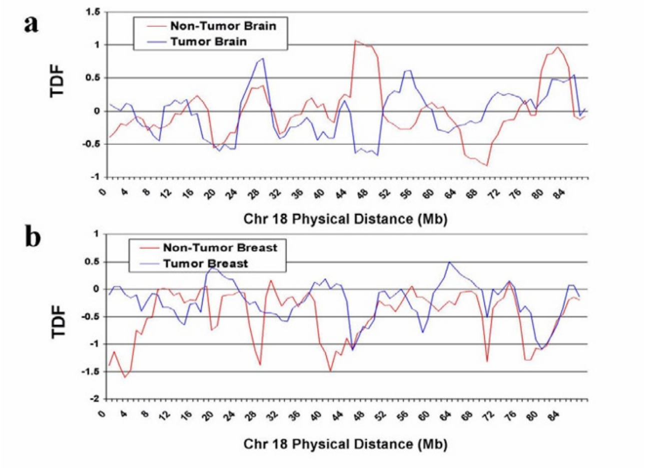

Figure 3
Differential expression between tumor tissues vs. non-tumor tissues (chromosome 18). a) An example of transition from a brain high TDF region on chromosome 18 (45 Mbp–52 Mbp) in normal tissue libraries to low TDF region in brain tumor libraries; b) Corresponding non-tumor breast tissue vs. tumor breast tissue transcriptome map on the same chromosome. The brain high TDF region observed that changed to low TDF region in a) was not observed in b).

are found in clusters of 6–138 genes [22]. The 3.6 Mbp human major histocompatibility complex (MHC) on chromosome 6p21.3 is a critical repository for the immune response genes [23]. Extensive analysis of the genomic organization of the MHC region has revealed that at least 27 of its resident genes possess duplicated copies in at least one of the three other restricted chromosomal regions 1q21-q25, 9q33-q34 and 19q13.1-p13.4. For another example, ABC transporter gene family members are located on 6p21.3, 1q25 and 9q34 as clusters.

The development of distinct tissue and cell types is a fundamental characteristic of growth in higher organisms. Tissue and cellular differentiation, in turn, is highly dependent on specific patterns of gene expression and transcript accumulation. Many studies have been successfully used to pinpoint genes exhibiting tissue or disease specific expression. This study suggests another approach that focuses on the tissue specific transcriptome map study and attempts to study the genomic proximity of tissue specific genes. We show here that some regions on specific chromosomes are enriched with brain-enriched gene expression. Given the chromosome window size (5 Mbp) used to calculate the TDF and the average size of the gene (~30 Kbp), each 5 Mbp region should contain on average about 170 genes. In addition, we report here only regions that have a normal level of gene density. Therefore the very high TDF_{NB} regions are most likely contributed by the high level expression of brain-enriched gene clusters or the regions containing a high density of brain-enriched genes.

We are aware of the importance of those brain-enriched genes that scatter across the genome whose brain specific or differential expression is diluted by the large volume of neighbor genes within the 5 Mbp window region. Another limitation of this approach is the inability to reveal some regions with equal up-down differential expression. The goal of this study is to pinpoint the chromosomal regions with the most significant brain-enriched gene expression and to elucidate the non-randomness of the tissue specific expression over the chromosome. Naturally, gene proximity within chromosomes is already known to be significant. Finding neighbors of a given gene can shed light on that gene, especially when the neighbors contain objects with similar features. Therefore, this tissue specific transcriptome map study may not only help us to understand the genome organization in the future, but may also provide means to gain leads to the functions of many genes for which this information is not currently available.

Methods

Resources for databases and computer programs

GenBank release 120 was downloaded from [ftp://ncbi.nlm.nih.gov] . ESTs and their associated tissue library information and clone information were extracted and or-

Table 3: High TDF_{NB} and low TDF_{NB} (Z-score >= 2) regions which show significant differential expression between tumor brain tissues and non-tumor brain tissues.

Chromosome	Region (Mbp)	TDF_{NB}	TDF_{TB}
2	14–18	0.45	0.06
2	91–99	-0.93	0.11
10	58–62	-0.74	0.92
12	85–89	0.36	-0.56
13	54–61	-0.78	0.19
15	21–25	0.74	0.21
18	45–52	1.0	-1.63
19	53–58	-0.64	1.8
20	10–15	0.64	-0.62
22	30–34	-0.48	0.15
X	103–109	0.43	0.09
X	24–28	-0.72	-0.11

ganized in a relational database (Sybase, SQL Server Release 11.0, CA, Sybase Inc.). The EST cDNA libraries were manually curated and catalogued into non-tumor brain libraries and brain tumor libraries. We obtained 208604 dbEST ESTs from 369 non-tumor brain cDNA libraries, 67351 ESTs from 148 brain tumor cDNA libraries. We also extracted 363473 ESTs from 100 Incyte LifeSeq non-tumor brain cDNA libraries and 126538 ESTs from their 23 brain tumor libraries. All non-commercial software used in this study was written in PERL 5.0.

Transcript mapping

Transcript mapping was done based on the October 2000 Freeze of the University of California at Santa Cruz's working draft sequence [http://genome.ucsc.edu] , which presents a tentative assembly of the finished and draft human genomic sequence based on the Washington University-Saint Louis clone map [http://genome.wustl.edu/gsc] . We mapped all the public ESTs (2.5 millions ESTs) from dbEST as well as the ESTs from Incyte Genomics' LifeSeq database (5.1 millions ESTs) using a local alignment software package AAT. AAT is a local alignment software which extended the BLAST algorithm by assigning fixed penalty to long gaps [24]. To reduce the number of undesirable matches due to interspersed repeats, the DNA sequence is screened for interspersed repeats using the RepeatMasker program (Smit, AFA & Green, P et al [http://ftp.genome.washington.edu/RM/RepeatMasker.html]). Only those ESTs that have over 95% identity to the genomic counterpart over half the length of ESTs' length or 50 bp whichever is longer are included.

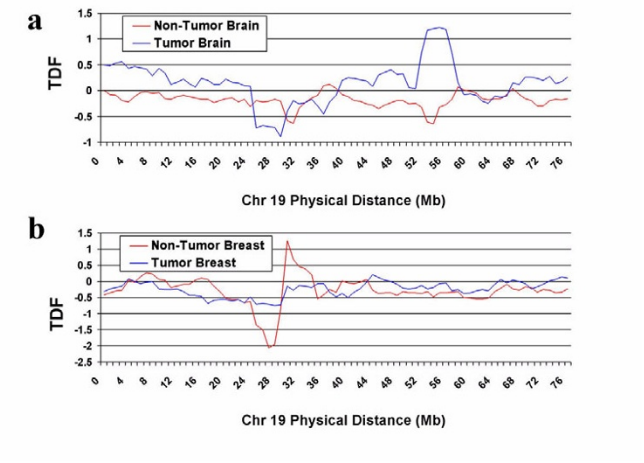

Figure 4
Differential expression between tumor tissues vs. non-tumor tissues (chromosome 19). a) An example of transition from a brain low TDF region on chromosome 19 (53 Mbp–58 Mbp) in normal tissue libraries to high TDF region in brain tumor libraries; b) Corresponding non-tumor breast tissue vs. tumor breast tissue transcriptome map on the same chromosome. The brain low TDF region observed that changed to high TDF region in a) was not observed in b).

Calculation of the transcript density factor (TDF)

The TDF for normal brain-enriched gene expression (TDF_{NB}) is calculated using 5 Mbp window moving along the chromosome with 1 Mbp interval and defined as:

$$TDF_{NB} = \ln(R_{NB}/R)/(T_{NB}/T)$$

Where

R_{NB} is the number of EST clones derived from non-tumor brain tissue within a window of 5 Mbp.

R is the number of EST clones derived from all non-tumor tissue pooled libraries within a window of 5 Mbp.

T_{NB} is the number of total mapped EST clones derived from non-tumor brain tissues.

T is the number of total mapped EST clones derived from all non-tumor tissue pooled libraries.

The distribution of TDF_{NB} of all the 5 Mbp regions should approximate a Gaussian distribution. Theoretically, TDF_{NB} should approach 0 if the expression level of genes in brain tissues has no difference from that in pooled tissues within a 5 Mbp chromosomal region. We define all the regions with high TDF_{NB} (Z-score >= 2) as brain-enriched regions. Those chromosome regions that have high brain-enriched gene expression are referred as high TDF_{NB} region. Those chromosome regions that have low brain-enriched gene expression are referred as low TDF_{NB} region.

The calculation of transcript density factor for brain tumor (TDF_{TB}) is similar to the calculation of TDF_{NB}.

$TDF_{TB} = \ln(R_{TB}/R)/(T_{TB}/T)$

Where

R_{TB} is the number of EST clones derived from tumor brain tissue within a window of 5 Mbp.

R is the number of EST clones derived from all non-tumor tissue pooled libraries within a window of 5 Mbp.

T_{TB} is the number of total mapped EST clones derived from tumor brain tissue libraries.

T is the number of total mapped EST clones derived from all non-tumor tissue libraries.

Correlation analysis of distribution pattern of TDF using public and Incyte data

The Pearson correlation coefficient (r) represents the degree of similarity (strength of correlation) between two sets of data. The correlation coefficient is calculated as follows:

$$r = \frac{n\,(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left[n \sum X^2 - (\sum X)^2\right]\left[n \sum Y^2 - (\sum Y)^2\right]}}$$

where X_i is TDF values derived from public data and Y_i is the TDF values derived from the Incyte data. Values for the Pearson correlation coefficient range from -1 to 1 where zero indicates no correlation, -1 indicates a perfect negative correlation and 1 indicates a perfect positive correlation.

The calculation of gene density

Using 5 Mbp window moving along the chromosome with 1 Mbp interval, the gene density ratio is defined as UniGene (Version 5.002) [4,25] count divided by the average UniGene count in each 5 Mbp region. Average gene density ratio is equal to 1.0.

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MHM, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al*: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-1656
2. Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, *et al*: **Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence.** *Nature* 1995, **377(suppl.):**3-174
3. Polymeropoulos MH, Xiao H, Sikela JM, Adams MD, Venter JC: **Chromosomal distribution of 320 genes from a brain cDNA library.** *Nature Genet* 1993, **4:**381-386
4. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, *et al*: **A gene map of the human genome.** *Science* 1996, **274:**540-546
5. Maglott DR, Katz KS, Sicotte H, Pruitt KD: **NCBI's LocusLink and Refseq.** *Nucleic Acids Res* 2000, **28:**126-128
6. Ewing B, Green P: **Analysis of expressed sequence tags indicates 35000 human genes.** *Nature Genet.* 2000, **25:**232-234
7. Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier F, Saurin W, Weissenbach J: **Estimate of human gene number provided by genome-wide analysis using DNA Tetraodon nigroviridis DNA sequence.** *Nat. Genet* 2000, **25:**235-238
8. Nowak R: **Entering the postgenome era.** *Science* 1995, **270:**368-369
9. Adams MD: **Progress towards a complete set of human genes.** *In Genomes, molecular biology and drug discovery* 1996
10. Bains W: **Virtually sequenced: the next genomic generation.** *Nature Biotechnol.* 1996, **14:**711-713
11. Collado-Vides J: **A syntactic representation of units of genetic information – a syntax of units of genetic information.** *J. Theor. Biol.* 1991, **148:**401-429
12. Finlay BB, Falkow S: **Common themes in microbial pathogenicity revisited.** *Microbiol. Mol. Biol. Rev.* 1997, **61:**136-169
13. Wansink DG, Schul W, van der Kraan I, van Steensel B, van Driel R, de Jong L: **Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus.** *J. Cell Biol.* 1993, **122:**283-293
14. Wei X, Somanathan S, Samarabandu J, Berezney R: **Three-dimensional visualization of transcription sites and their association with splicing factor-rich nuclear speckles.** *J. Cell Biology* 1999, **146:**543-58
15. Jackson DA, Iborra FJ, Manders EM, Cook PR: **Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei.** *Mol. Biol. Cell* 1998, **9:**1523-1536
16. Ekins R, Chu FW: **Microarrays: their origins and applications.** *Trends in Biotechnology* 1999, **17:**217-218
17. Audic S: **The significance of Digital Gene Expression Profiles.** *Genome Res.* 1997, **7:**986-995
18. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487
19. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921
20. 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
21. Danchin A, Guerdoux_Jamet P, Moszer I, Nitschk P: **Mapping the bacterial cell architecture into the chromosome.** *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 2000, **355:**179-190
22. Glusman G, Yanai I, Rubin I, Lancet D: **The Complete Human Olfactory Subgenome.** *Genome Res* 2001, **11:**685-702
23. Campbell RD, Trowsdale J: **Map of the human MHC.** *Immunol. Today* 1993, **14:**349-352
24. Huang X, Adams MD, Zhou H, Kerlavage AR: **A Tool for Analyzing and Annotating Genomic Sequences.** *Genomics* 1997, **46:**37-45
25. Schuler GD: **Pieces of the puzzle: expressed sequence tags and the catalog of human genes.** *J. Mol. Med.* 1997, **75:**694-698