Integrated analysis of the genome and the transcriptome by FANTOM

Shintaro Katayama, Mutsumi Kanamori and Yoshihide Hayashizaki

Date received (in revised form): 18th May 2004

Shintaro Katayama
is an engineer in bioinformatics, with research interest in the analysis of transcription regulatory network.

Mutsumi Kanamori
received her PhD from Himeji Institute of Technology. Her research interest is in the analysis of transcription regulatory networks.

Yoshihide Hayashizaki
has made great efforts to reveal the gene transcriptional network through organising the FANTOM consortium. The Genome Network Project started in 2004.

Keywords: *annotation, integrated analysis, genome, transcriptome*

Abstract

The key to reliable annotation of a mammalian genome is broad characterisation of the transcriptional output, the transcriptome. FANTOM, the functional annotation of mouse cDNA, is a large-scale analysis of both the genome and the transcriptome of the mouse. In the early days of this work, the transcripts were characterised using our sophisticated methods. After the timely release of the first draft of mouse genome sequences, interesting information was obtained by its integration with these one-by-one annotations. Moreover, each transcript included its expression profile. Here, the two integrated annotation methods used by FANTOM are reviewed: one-by-one and categorised. One-by-one annotation refers to naming carried out based on well-known transcripts or its fragments using the top-down-style pipeline developed mostly by the FANTOM project. Categorised annotation, which refers to transcript grouping, not only helps naming of unknown transcripts, but will be the most utilised method for integration of the genome and the transcriptome from now on.

INTRODUCTION

The FANTOM (the functional annotation of mouse cDNA) Consortium is a group of molecular biologists from the RIKEN Genomic Sciences Center in Yokohama and from elsewhere in Japan and the world. The first FANTOM meeting (FANTOM1) in 2001 was organised to annotate the initial output of the RIKEN pipeline, which consisted of 21,076 cDNA sequences.[1] Originally, annotation, that is, the naming of a gene on the basis of its function, was one of the fundamental tasks of biology and involved repeated experiments. However, because of the comprehensive and high-throughput studies,[2] there were a large number of transcripts to deal with and various annotation methods available. FANTOM1 was one attempt to develop a sophisticated annotation pipeline.

At the time of FANTOM1, the mouse genome project was in progress. In 1990, the Human Genome Project had suggested that the mouse was one of the model organisms that should have its genome sequenced. In 1999, the Mouse Genome Sequencing Consortium (MGSC), which at that time consisted of three major sequencing centres, launched a concerted effort to sequence the mouse genome. Although it is possible, through using various programs, to predict the approximate region of a gene from the genomic sequence, it is not yet easy to predict accurately the true transcriptional region.[3] Therefore, it is necessary to identify many gene transcripts and to analyse gene functions comprehensively.

For that reason, at the second FANTOM meeting (FANTOM2) in 2002,[4] the annotation pipeline was further developed on the basis of the various successes and problems encountered during FANTOM1 and demands that had arisen during the genome analysis. The number of target transcripts had increased to 60,770 sequences, and the number of well-known mouse genes had also increased. However, the most important advance was the release of the mouse draft genome[5] simultaneously with the FANTOM2 results. The participants of the FANTOM2 were able to use both the

S. Katayama,
Laboratory for Genome Exploration Research Group,
RIKEN Genomic Sciences Center (GSC),
RIKEN Yokohama Institute,
Suehiro-cho,
Tsurumi-ku,
Yokohama,
Kanagawa, 230–0045, Japan

Tel: +81 45 503 9222
Fax: +81 45 503 9216
E-mail: katayama@gsc.riken.jp

One-by-one annotation

genome and the transcripts. Moreover, large-scale expression analysis techniques, for example, microarray[6,7] technology, had become widely available. Thus, the field of biology was undergoing major changes. Fortunately, FANTOM2 was able to take advantage of those changes, and it became the first large-scale annotation project that integrated the genome with the transcripts and their expression profiles.

Ordinarily, all of the experiments and analyses were just for annotation. Here, the two annotation methods used by the FANTOM project are considered: one-by-one and categorised. One-by-one annotation refers to naming carried out based on well-known transcripts or its fragments using the top-down-style pipeline developed mostly by the FANTOM project. Categorised annotation, which refers to transcript grouping, not only helps naming of unknown transcripts, but will be the method most utilised to integrate the genome and the transcriptome from now on. Finally, some possible future developments in gene annotation are reviewed.

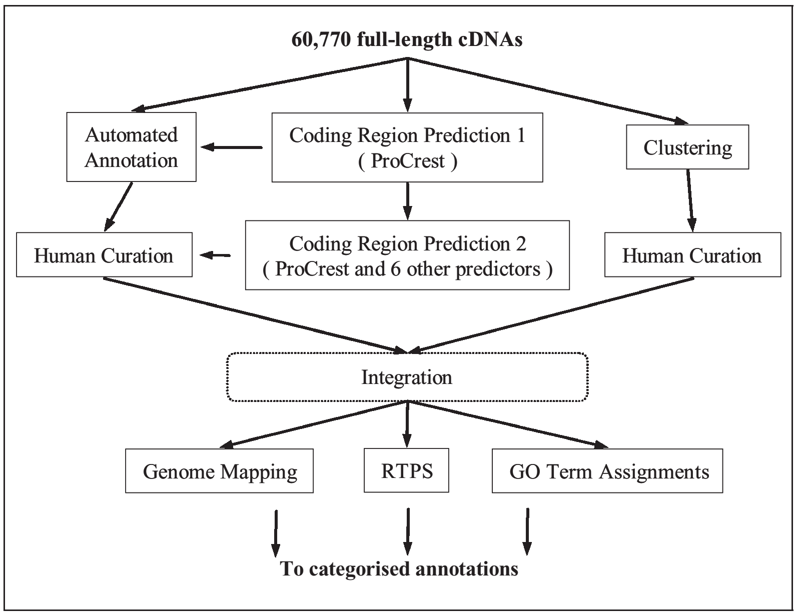

Figure 1: The FANTOM2 one-by-one annotation pipeline. RTPS, representative transcript and protein set; GO, gene ontology

ONE-BY-ONE ANNOTATION

This classic, known-base method relies on some key technologies, including coding region prediction and homology searches. However, for the annotation of a large cDNA set, a well-controlled pipeline is very important as well. Figure 1 shows the task flow of FANTOM2 one-by-one annotation. First, the coding regions are predicted, being annotated automatically on the basis of homologies with well-known genes, and clustered complete transcripts. Then, the original annotation is modified or added to, using annotations found by the web-based annotation system of the FANTOM Consortium. Finally, categorised annotations are prepared. Each step is now described.

Coding region prediction

General importance: Coding region prediction

Annotations of well-known genes are very easy because we can use homology searches. However, when we encounter unknown genes, we must first analyse them by identifying the functional region before we can annotate them. In the case of genes that code for proteins, the functional regions are those coding regions (CDS) that code for the protein domains. Therefore, if we cannot define the coding region first, annotation is difficult. Unfortunately, the problem is made more difficult because some cDNAs may be immature or truncated or may include a sequencing error. Fifteen per cent of all sequencing errors result in a frameshift. Since the genome had not yet been released to the public, whether the CDS included a nucleotide insertion or deletion had to be determined from only the sequence. Therefore, several different computational programs were used for CDS prediction: rsCDS, the NCBI CDS predictor, ProCrest, DECODER,[8] Longest-ORF and Truncated-ORF.

These methods have been compared by Furuno *et al.*,[9] and, here, ProCrest (Protein Coding Region Estimator) is mentioned in particular. ProCrest is an ideal CDS predictor that provides useful

Frameshift

Proper naming

information about the following points during cDNA annotation. First, it distinguishes the CDS from non-coding regions by using statistical techniques, so that the probability that a region is a CDS is quantified. Unlike DECODER, ProCrest does not require each transcript to have only one CDS. Second, ProCrest distinguishes between a full-length and a truncated CDS in the transcript. A transcript contains a full-length CDS if a high-potential region begins with an initiation codon and ends with a stop codon. We can be fairly certain that it is 5′-truncated if the 5′-end of the region is without an initiation codon. Conversely, it is 3′-truncated if the 3′-end does not have a stop codon. Moreover, ProCrest indicates the probability of a sequencing error causing a frameshift. A frameshift is considered likely when the reading frame of a region with high coding potential is switched, because ProCrest calculates the coding potential of each possible reading frame. Finally, ProCrest distinguishes between a mature transcript, in which the splicing is completed, and an immature one, which contains one or more introns. In a part of the sequence, a region showing a rapid drop in CDS coding potential may be an intron.

Uninformative name

Annotation pipeline

Candidates for comparison

Three kinds of evidence are generally used to evaluate coding potential. The first kind of evidence is hexanucleotide frequency among 4,096 possible hexanucleotide sequences (equals $4^3 \times 4^3$; for example, CCT-GTA). This classic method is used by Genscan,[10] ESTScan,[11] DECODER and other programs. The second kind of evidence is neighbouring amino acid frequencies, which vary according to the type of protein, among 400 possibilities (equals 20×20; for example, P-V). Degenerate codon frequencies, which depend on the number and expression frequency of tRNA anticodons, among 59 possibilities (equals $6 \times 3 + 4 \times 5 + 3 + 2 \times 9$; for example, CCn or GTn) is the last kind of evidence used. This last kind of evidence takes into account a bias toward nucleotide frequencies favouring high GC content. The learning data set comprised the 8,419 NCBI RefSeqs available at that time. The ProCrest coding viewer shows the coding-potential score, suggests one or more coding regions, and, if present, shows noteworthy facts as mentioned above in each transcript (Figure 2). Because of the high information content of its results, ProCrest was applied first to all cDNA, and the automated annotation was based on the ProCrest prediction. Then, during the human curation step, this prediction was compared with the results of other coding region predictors.

The first step: Automated annotation and clustering

During FANTOM1, it was apparent that accurate and consistent annotation by manual curation alone is very difficult. Proper naming of transcripts is difficult when only computational methods are used. There is no doubt that a lucid rule for accurate and consistent annotation is needed. Defining ‘proper naming’ is delicate, but it is reasonable to say that an uninformative name, such as a name that does not directly indicate function or one that consists only of an ID number, is unsuitable. For example, ‘HYPOTHETICAL 30-kDa PROTEIN’ and ‘RIKEN 0610005K03 gene’ are uninformative names because they do not directly indicate function. A set of approximately 50 regular expressions was developed that defined such ‘uninformative’ words and implemented an uninformative rule filter. This list can evolve over time as human curators identify additional information-poor terms. CAS[12] (cDNA annotation system), which is our automated annotation pipeline for preliminary high-quality annotation, was developed with these considerations in mind.

A lucid rule for accurate and consistent annotation is specifically a top-down-style flow chart. If two sequences match throughout, it can be regarded as the same gene. Of course, the candidates for comparison must be well-known transcripts because even if a gene were

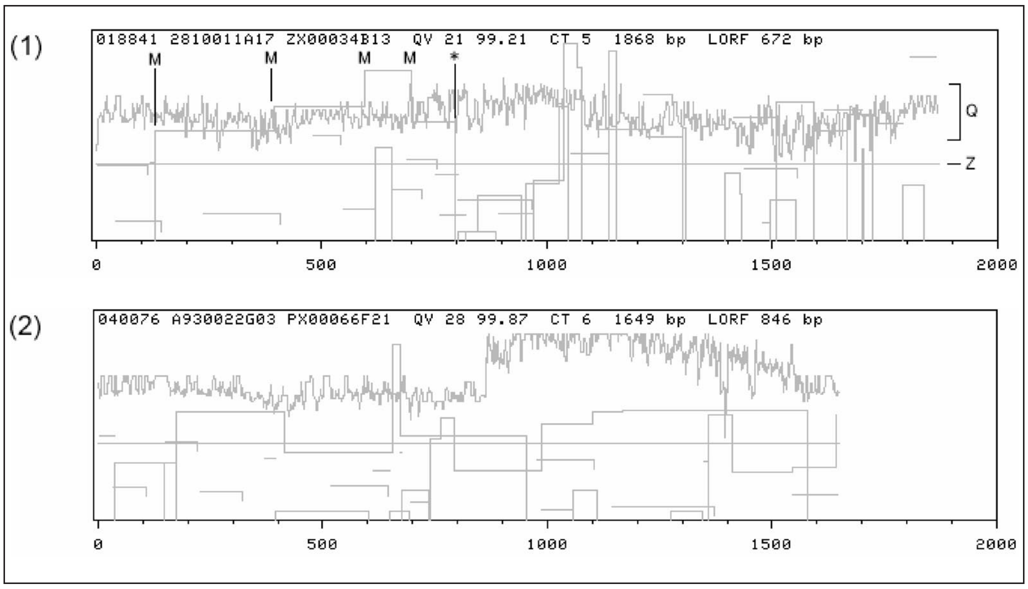

Figure 2: ProCrest coding viewer in FANTOM. (1) The coding-potential view for clone 2810011A17, which is similar to RAS-RELATED PROTEIN RAB-32 [*Homo sapiens*] (see Figure 3). The density plot marked 'Q' denotes sequence quality. Squares and angled lines show the open reading frames (ORF) in each possible reading frame. The height of each line is proportional to the coding-potential score in each ORF. Regions with coding potential are above line 'Z'; conversely, probable non-coding regions are below line 'Z'. M shows the location of ATG codons in one predicted coding region, and * marks the stop codon. (2) The coding-potential view for clone A930022G03, which is a FASCIN 2 (RETINAL FASCIN) homologue [*Homo sapiens*]. At about 900 bp is the boundary between two high-coding-potential regions. This boundary suggests a frameshift. In fact, there is also a gap in the alignment of the transcript with human FASCIN2 and a steep rise in sequence quality at around 900 bp

found to be similar to an unknown transcript, we would be no closer to understanding our gene. Therefore, candidates for comparison were identified among annotated transcripts of MGD,[13] RefSeq and LocusLink.[14] Also, a gene that is similar to another well-known gene only in the coding region might also be the same gene. For this reason, other candidates were identified for comparison with PIR[15] and in the Swiss-Prot[16] non-redundant database. Both the similarity and the length of the alignment block were considered. Then, one of the following four paths was followed, according to the degree of homology. If a similar well-known transcript could not be found, InterPro,[17] MDS[18] and SCOP[19] were used to find possible protein domains. If no fully wide coding region was identified, the well-known clusters that consist of expressed sequence tags (ESTs) were looked for. UniGene and TIGR[20] were used as the well-known clusters. If no well-known EST cluster was matched, each EST was looked for. If no protein domains, no EST clusters and no ESTs was identified, the name 'unclassifiable' was assigned. Thus, CAS used a top-down-style decision tree to place each transcript into one of 19 categories on the basis of its degree of similarity with known sequences. Human gene database, H-invitational Database (H-InvDB), has similar annotation methods, into five similarity categories.[21]

Clustering

The determination of a coding region is an indispensable task, but the clustering of transcripts simplifies naming. By clustering, we not only decrease the number of targets but also eliminate possible nucleotide insertions or deletions derived from cloning or sequencing problems. Moreover, a fragment may be shown to be a part of one longer transcript. On the other hand, a specific transcription start point or specific splicing may be hidden by clustering. If the genome with sufficient accuracy was released, probably cDNAs would be mapped on the genome and clustering realised as for other organisms.[21,22] However, since the genome was not sufficient at that time, an original cDNA-based clustering program, ClusTrans, was developed to cluster all of our transcripts, supplementary information 19 of Okazaki *et al.*[4]

The second step: Human curation

Human curation

Although a perfect decision tree was aimed for, mistakes deriving from data errors cannot be avoided. Human curation compensates for such mistakes. To review and appropriately modify the annotation, an integrated view is needed of the annotation and the evidence on which it is based. CAS integrates a human curation interface for reviewing and modifying the annotation (Figure 3). In addition to the annotation by the automated pipeline, the genome mapping coordinates, the results of CDS region prediction by other programs, the quality of sequencing, the complexity of the amino acids, the contig assembly and so on are considered by human curators. For example, the genome mapping coordinates are useful because they indicate whether the transcript is properly spliced. Sequence quality and contig assembly information is essential for assessing the quality of any annotation because sequence quality allows the curators to determine whether discrepancies between the cDNA sequences and possible matches in target databases reflect sequencing errors or genuine polymorphisms, mutations or closely related isoforms. The integrated approach makes a reliable judgment possible.

Representative transcript and protein set (RTPS)

Gene ontology (GO)

This step involves not only a review of the automated annotation but also its revision by an expert in genome nomenclature. Although we had great confidence in the automated annotation, we realised that additional human review and curation was still needed to evaluate and confirm the assignments. Thus, the large number of specialists in the FANTOM Consortium reviewed and refined the annotations. Therefore, CAS includes not only the automated annotation pipeline results but also the human curation results in each cDNA annotation. On the other hand, each cluster was annotated by Jackson Lab.[25]

The final step: Genome mapping, the representative transcript and protein set, and gene ontology terms

FANTOM attempted an integrated analysis of the genome and the transcriptome. cDNA mapping to the genome sequence gives better solutions for, for example, multigene families and complex alternative splicing than EST assemblies. After the human curation, the mouse draft genome was released and all transcripts were mapped to the genome. Although there are various tools for alignment, verification is still often required, but this problem will be considered later in this paper. The representative transcript and protein set (RTPS) derived from the RIKEN transcripts provides a substantial new discovery resource because of its non-redundancy. The RTPS pipeline, supplementary information 9 of Okazaki *et al.*,[4] clustered the transcripts and selected those that were representative. Gene ontology (GO) term assignments are also useful, but each GO term assignment must include the evidence codes. The genes in the MGD were assigned GO terms by MGI annotators.[26] In some cases, GO terms were also assigned by FANTOM Consortium members, or by using translation tables of Swiss-Prot keywords, InterPro domains,[27] EC assignments, and SCOP structural domains to GO terms. We must not forget, however, that there are differences

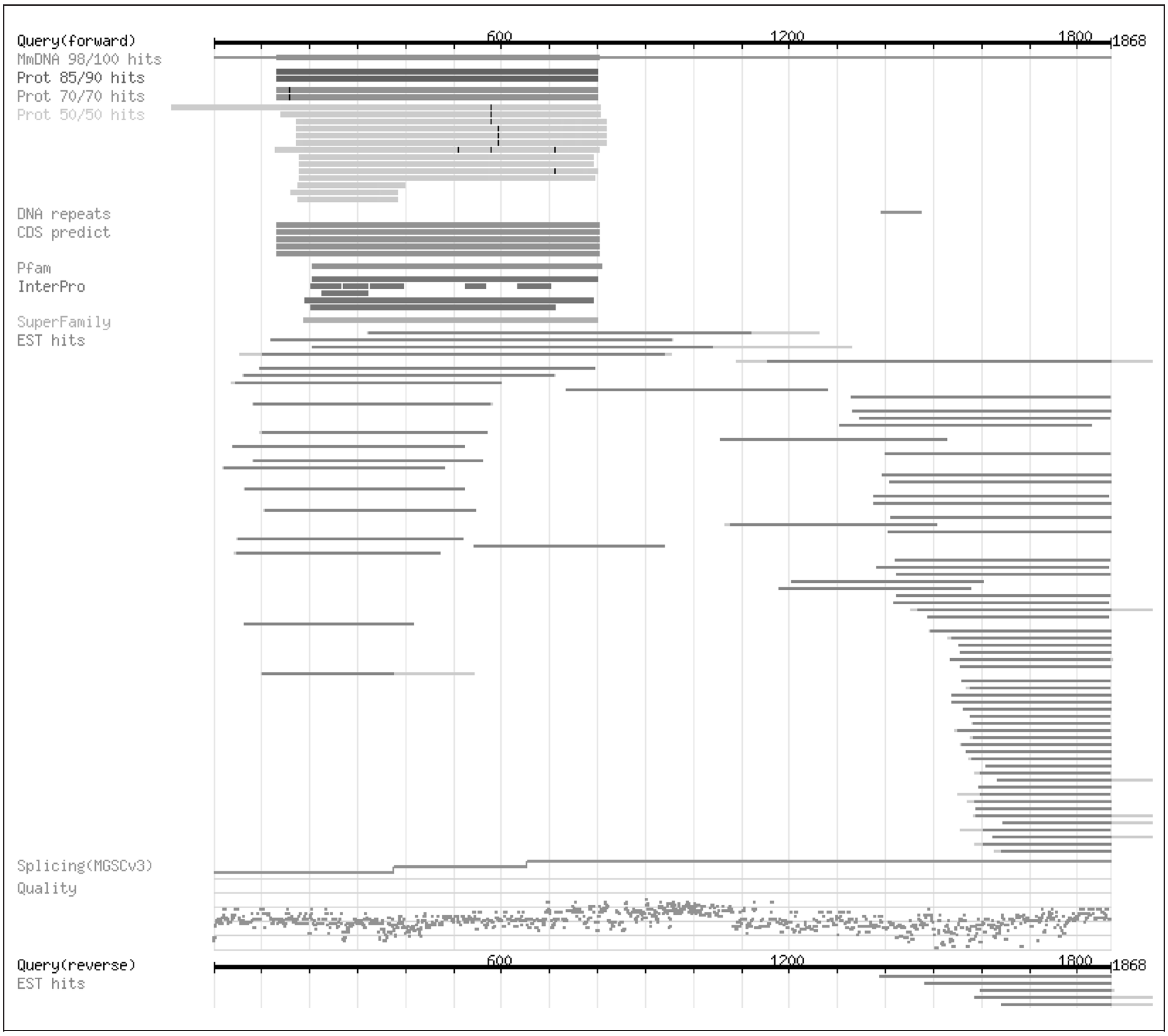

Figure 3: An integrated view of the annotation of the 1,868-nt clone 2810011A17, which is similar to the RAS-RELATED PROTEIN RAB-32 [*Homo sapiens*]. There are two alignments, forward and reverse of annotation target cDNA. 'MmDNA' track shows BLASTN[23] alignments to Mouse DNA sequence set from RefSeq+LocusLink+MGD. 'Prot' track shows FASTY[24] alignments to protein sequence set from Swiss-Prot + PIR. In these tracks, two numbers that are separated by slash mean ratios of identity and match length, respectively. If there is no similarity with well-known sequence, 'CDS predict' track and Protein motif tracks ('Pfam', 'InterPro' and 'Superfamily') help us to annotation. The steps in the 'Splicing (MGSCv3)' row show the exon boundaries

of the abstraction levels and evidence types among assignments of GO terms. Only a few genes were not assigned GO terms. Together, genome mapping, the RTPS and GO terms provide very important information. Many FANTOM collaborators use these resources, and they gave impressive results during categorised annotation.

Categorised annotation

CATEGORISED ANNOTATION

Categorised annotation refers to the discovery of functional properties and grouping. There are many groups of transcripts, for example, sense–antisense pairs and non-coding transcripts. When we annotate transcripts, if the target organism is closely related to a well-known organism,[22] one-by-one annotation is effective by similarity. If not, we have no choice but to use this categorised annotation. From now on, various categorisations may be developed. However, categorised annotation that integrates transcripts and the genome gives the knowledge of new kind. Therefore, we would like to focus attention on the integration of the genome and the transcriptome, which means the transcripts, their surroundings and their expression profiles.

Expression profiles

Antisense transcripts

Antisense transcripts

Because mRNA is single-stranded, a pair of mRNA transcripts might hybridise according to the Watson–Crick rule, thus possibly altering transcription, maturation, transport, stability or translation.[28] Intuitively, we see that transcripts from overlapping parts of the genome will sometimes hybridise. By genome mapping, Kiyosawa *et al.*[29] found 2,481 sense–antisense pairs and 899 further pairs exhibiting non-antisense bidirectional transcription. The major difference between these two groups was the expression pattern; many sense–antisense pairs were co-expressed. Moreover, they focused on CpG islands in the genome near these sense–antisense pairs. There are many mature tools for genome mapping and CpG-island prediction, but Kiyosawa *et al.* extended our knowledge of the reliability of such transcripts, because they integrated genome mapping and GpG-island prediction efficiently.

Secretome

Alternative splicing

Alternative splicing

Another approach to constructing the transcriptome is high-throughput sequencing of cDNA ends (ESTs), but full-length cDNA assemblies are better for complex alternative splicing analysis. Zavolan *et al.*[30] developed a new computational procedure to identify and classify the forms of splice variation present in a gene. BLAST is very useful for simply mapping to the genome, but Zavolan *et al.* achieved a more rigorous analysis by refining the alignment result using Sim4[31] or BLAT[32] and following the splicing consensus. More important, these strict alignment tools defined a maximum intron length. This is important because a transcript that is rare and has a very long intron might fail the mapping.

Imprinting

Expression profiles

There are many integrated viewers between gene expressions and cDNAs, for example, READ,[7] H-ANGEL in H-InvDB and expression tracks in UCSC Genome Browser Database.[33] However, further analysis and focusing are necessary to categorising. Two examples are given below.

The secretome comprises proteins secreted into the extracellular environment. Grimmond *et al.*[34] developed a computational strategy to identify the secretome derived from the RTPS. These proteins must include a signal peptide that is required for entry into the secretory pathway, and they lack any transmembrane domains or intracellular localisation signals. They identified 2,033 unique proteins, including more than 500 novel proteins and 92 proteins fewer than 100 amino acids in length. By expression profiling the secretome and performing a clustering analysis, they found that several groups of genes were highly expressed in a tissue-restricted fashion. This is important annotation information, but unfortunately it is not useful for naming. There is room for further investigation of the integration of gene naming conventions with gene expression profiles.

Nikaido *et al.*[35] also published an impressive study of transcripts that were differentially expressed by parthenogenote and androgenote embryos. They extracted natural antisense transcripts, non-coding RNAs and candidate transcripts from the imprinted region by a multiple database search. Clearly, imprinting cannot be predicted from the genome sequence and its annotation, but prediction becomes possible by including the transcriptome in the analysis. Genome mapping is also needed to extract the imprinting relevant to a disease locus. This is a typical example of how the integration of analyses of the genome, mRNA transcripts and gene expression profiles brings about new knowledge.

Cap analysis gene expresssion (CAGE)

CONCLUSION: BEYOND FANTOM

Genome network

Integrated analysis of the genome and the transcriptome creates a synergistic effect, providing extremely valuable insights into the nature of life. The transcriptome shows the exact locations of genes, whereas genome sequences provide information on promoters, exon–intron junctions, and so on. Recently, the goal is changing from a sequence-based to an expression-based analysis, that is, from the analysis of static cell status to the modelling of dynamic cell activity. However, the integration of the genome and the transcriptome should be the basis of all analyses, including dynamic analyses. An analysis of the mammalian circadian clock by Ueda *et al.*[36] is a notable example. First, they examined the expression profiles of many transcripts, and selected cycling genes. Then, they identified transcription factor response elements near the transcription start site on the genome.

Transcription factor response elements

One of the surprising results of FANTOM was the prominence of non-coding sequences. Out of 37,086 transcriptional units (TUs), 15,815 had a high probability of being non-coding. Although CAS defines transcripts coding for fewer than 100 amino acids as non-coding, as mentioned before, some natural, short coding genes also exist. Numata *et al.*[37] extracted 4,280 transcripts from 15,815 non-coding TUs as the candidate non-coding set. Moreover, further experimental and informatic analyses are being performed of such sequences. A top-down style decision tree for automated annotation of non-coding sequences is also needed. H-InvDB has annotation flow of non-coding, but it is a debatable point. What is the functional region for annotation in non-coding genes that is equivalent to the CDS in coding? Rfam[38] is one interesting example.

Non-coding

In 2003, we developed the cap analysis gene expression (CAGE) method.[39] The high-throughput nature of this technology offers a way of understanding gene networks by the correlation of promoter usage and gene transcription factor expression. This shifts the target of the analysis to the genome network, which means to the modelling of dynamic cell activity. We would be pleased if, as our research develops, it becomes useful to many other researchers.

Acknowledgments

Thanks are due to T. Kasukawa for valuable advice and to K. Yoshida for providing us with the samples. We gratefully acknowledge helpful discussions with J. Adachi on several points about ProCrest.

References

1. Kawai, J., Shinagawa, A., Shibata, K. *et al.* (2001), 'Functional annotation of a full-length mouse cDNA collection', *Nature*, Vol. 409(6821), pp. 685–690.
2. Carninci, P., Waki, K., Shiraki, T. *et al.* (2003), 'Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia', *Genome Res.*, Vol. 13(6B), pp. 1273–1289.
3. Wang, J., Li, S., Zhang, Y. *et al.*(2003), 'Vertebrate gene predictions and the problem of large genes', *Nat. Rev. Genet.*, Vol. 4(9), pp. 741–749.
4. Okazaki, Y., Furuno, M., Kasukawa, T. *et al.* (2002), 'Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs', *Nature*, Vol. 420(6915), pp. 563–573.
5. Waterston, R. H., Lindblad-Toh, K., Birney, E. *et al.* (2004), 'Initial sequencing and comparative analysis of the mouse genome', *Nature*, Vol. 420(6915), pp. 520–562.
6. Shoemaker, D. D., Schadt, E. E., Armour, C. D. *et al.* (2001), 'Experimental annotation of the human genome using microarray technology', *Nature*, Vol. 409(6822), pp. 922–927.
7. Bono, H., Yagi, K., Kasukawa, T. *et al.* (2003), 'Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays', *Genome Res.*, Vol. 13(6B), pp. 1318–1323.

8. Fukunishi, Y. and Hayashizaki, Y. (2001), 'Amino acid translation program for full-length cDNA sequences with frameshift errors', *Physiol. Genomics*, Vol. 5(2), pp. 81–87.

9. Furuno, M., Kasukawa, T., Saito, R. *et al.* (2003), ' CDS annotation in full-length cDNA sequence', *Genome Res.*, Vol. 13(6B), pp. 1478–1487.

10. Burge, C. and Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA', *J. Mol. Biol.*, Vol. 268(1), pp. 78–94.

11. Iseli, C., Jongeneel, C. V. and Bucher, P. (1999), 'ESTScan, a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences', *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pp. 138–148.

12. Kasukawa, T., Furuno, M., Nikaido, I. *et al.* (2003), 'Development and evaluation of an automated annotation pipeline and cDNA annotation system', *Genome Res.*, Vol. 13(6B), pp. 1542–1551.

13. Bult, C. J., Blake, J. A., Richardson, J. E. *et al.* (2004), 'The Mouse Genome Database (MGD): Integrating biology with the genome', *Nucleic Acids Res.*, Vol. 32 Database issue, pp. D476–481.

14. Pruitt, K. D. and Maglott, D. R. (2001), 'RefSeq and LocusLink: NCBI gene-centered resources', *Nucleic Acids Res.*, Vol. 29(1), pp. 137–140.

15. Barker, W. C., Garavelli, J.S., Huang, H. *et al.* (2000), 'The protein information resource (PIR)', *Nucleic Acids Res.*, Vol. 28(1), pp. 41–44.

16. Boeckmann, B., Bairoch, A., Apweiler, R. *et al.* (2003), 'The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003', *Nucleic Acids Res.*, Vol. 31(1), pp. 365–370.

17. Mulder, N. J., Apweiler, R., Attwood, T. K. *et al.* (2003), 'The InterPro Database, 2003 brings increased coverage and new features', *Nucleic Acids Res.*, Vol. 31(1), pp. 315–318.

18. Kawaji, H., Schonbach, C., Matsuo, Y. *et al.* (2002), 'Exploration of novel motifs derived from mouse cDNA sequences', *Genome Res.*, Vol. 12(3), pp. 367–378.

19. Gough, J. and Chothia, C. (2002), 'SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments', *Nucleic Acids Res.*, Vol. 30(1), pp. 268–272.

20. Pertea, G., Huang, X., Liang, F. *et al.* (2003), 'TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets', *Bioinformatics*, Vol. 19(5), pp. 651–652.

21. Imanishi, T., Itoh, T., Suzuki, Y. *et al.* (2004), 'Integrative annotation of 21,037 human genes validated by full-length cDNA clones', *PLoS Biol.*, Vol. 2(6), p. E162.

22. Kikuchi, S., Satoh, K., Nagata, T. *et al.* (2003), 'Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice', *Science*, Vol. 301(5631), pp. 376–379.

23. Altschul, S. F., Gish, W., Miller, W. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.*, Vol. 215(3), pp. 403–410.

24. Pearson, W. R., Wood, T., Zhang, Z. *et al.* (1997), 'Comparison of DNA sequences with protein sequences', *Genomics*, Vol. 46(1), pp. 24–36.

25. Baldarelli, R. M., Hill, D. P., Blake, J. A. *et al.* (2003), 'Connecting sequence and biology in the laboratory mouse', *Genome Res.*, Vol. 13(6B), pp. 1505–1519.

26. Hill, D. P., Davis, A. P., Richardson, J. E. *et al.* (2001), 'Program description: Strategies for biological annotation of mammalian systems: Implementing gene ontologies in mouse genome informatics', *Genomics*, Vol. 74(1), pp. 121–128.

27. Camon, E., Magrane, M., Barrell, D. *et al.* (2003), 'The Gene Ontology Annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro', *Genome Res.*, Vol. 13(4), pp. 662–672.

28. Vanhee-Brossollet, C. and Vaquero, C. (1998), 'Do natural antisense transcripts make sense in eukaryotes?', *Gene*, Vol. 211(1), pp. 1–9.

29. Kiyosawa, H., Yamanaka, I., Osato, N. *et al.* (2003), 'Antisense transcripts with FANTOM2 clone set and their implications for gene regulation', *Genome Res.*, Vol. 13(6B), pp. 1324–1334.

30. Zavolan, M., Kondo, S., Schonbach, C. *et al.* (2003), 'Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome', *Genome Res.*, Vol. 13(6B), pp. 1290–1300.

31. Florea, L., Hartzell, G., Zhang, Z. *et al.* (1998), 'A computer program for aligning a cDNA sequence with a genomic DNA sequence', *Genome Res.*, Vol. 8(9), pp. 967–974.

32. Kent, W. J. (2002), 'BLAT – the BLAST-like alignment tool', *Genome Res.*, Vol. 12(4), pp. 656–664.

33. Karolchik, D., Baertsch, R., Diekhans, M. *et al.* (2003), 'The UCSC Genome Browser Database', *Nucleic Acids Res.*, Vol. 31(1), pp. 51–54.

34. Grimmond, S. M., Miranda, K. C., Yuan, Z. *et al.* (2003), 'The mouse secretome: Functional classification of the proteins

secreted into the extracellular environment', *Genome Res*., Vol. 13(6B), pp. 1350–1359.

35. Nikaido, I., Saito, C., Mizuno, Y. *et al*. (2003), 'Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling', *Genome Res*., Vol. 13(6B), pp. 1402–1409.

36. Ueda, H. R., Chen, W., Adachi, A. *et al*. (2002), 'A transcription factor response element for gene expression during circadian night', *Nature*, Vol. 418(6897), pp. 534–539.

37. Numata, K., Kanai, A., Saito, R., *et al*. (2003), 'Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection', *Genome Res*., Vol. 13(6B), pp. 1301–1306.

38. Griffiths-Jones, S., Bateman, A., Marshall, M. *et al*. (2003), 'Rfam: An RNA family database', *Nucleic Acids Res*., Vol. 31(1), pp. 439–441.

39. Shiraki, T., Kondo, S., Katayama, S. *et al*. (2003), 'Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage', *Proc. Natl Acad. Sci. USA*, Vol. 100(26), pp. 15776–15781.