LSHG-CT-2003-503329

ATD

The Alternate Transcript Diversity Project

Specific Targeted Research or Innovation Project
Thematic Priority 1
Life sciences, genomics and biotechnology for health

Publishable Final Activity Report

Period covered: 01. March 2004 to 28. February 2007
Date of preparation:

Start date of project: 01. March 2004 Duration: 3 years

Project coordinator: Prof. Daniel GAUTHERET
Project coordinator organisation: Inserm

Table of Contents

1. Project Execution 3
1.1 Summary description of project objectives 3
1.2 Contractors involved 3
1.3 Work performed and results achieved 4
2. Dissemination and use 8
Annex 1 – Final plan for using and disseminating the knowledge 9
Section 1 - Exploitable knowledge and its use 9
Section 2 – Dissemination of knowledge 9
Section 3 - Publishable results 12

1. Project Execution

Alternate Transcript Diversity

ATD

Alternate Transcript Diversity

www.atdproject.org

1.1 Summary description of project objectives

Production of mature transcripts in vertebrates is regulated at three stages: Transcription initiation, splicing and polyadenylation. The combinatorial arrangement of variations at each stage generates, from a single gene, a variety of mRNA isoforms with different start sites, exons or 3' UTRs. Expression of these Alternative Transcripts (AT) has been observed to be specific to tissue-type or developmental stage. Disruptions in expression patterns have serious consequences for an organism and are associated with numerous diseases, including cancer, multiple sclerosis, heart failure and neurodegenerative disorders. Identifying disease-specific ATs can lead to development of novel drug targets or markers.

The ATD project is a collaborative multi-disciplinary effort that intends to characterize alternative transcripts (AT) comprehensively throughout the human genome and assess the differential expression of these isoforms in time and space, in normal and disease-related tissues. Based on bioinformatics approaches, the project integrates the various processes affecting transcript structures (initiation, splicing and polyadenylation), while implementing strict quality control procedures, such as seeking evolutionary proof through comparative sequence data analysis between human and mouse; and seeking evidences from protein domain information. Further characterizations of ATs are carried out through activities such as identification of regulatory patterns or the study of expression specificity in terms of association with diseases, developmental stages or tissues. Standard vocabularies and models are developed in order to represent gene structures and their expression patterns. The validity of bioinformatics prediction of disease-specific ATs will be examined through RT-PCR experiments in selected tissues. Primers and probes that uniquely represent each of the ATs in the database will be developed and made available to the community. The AT discovery effort will be accompanied by database integration, and dissemination to the scientific community.

1.2 Contractors involved

INSERM (CO1, France): coordinator; EMBL (CR2a, **EBI** UK. CR2b, **EMBL** Germany); University of Western Cape - **UWC** (CR5, South Africa); Fundacio IMIM - **FIMIM** (CR6, Spain); Estonian Biocenter - **EBC** (CR7, Estonia); University of Heidelberg - **UH** (CR8, Germany); **INSERM-Transfert** (CR9, France): project management; Max Delbruck Center – **MDC** (CR12, Germany); Centre de Regulació Genòmica - **CRG** (CR13, Spain).

Coordinator contact details: Daniel Gautheret, INSERM TAGC, Université de la Méditerranée, Luminy case 928, 13288 MARSEILLE cedex 09, France. E-mail : gautheret @ esil.univ-mrs.fr

1.3 Work performed and results achieved

The eight partners of the ATD project have now spent three years working together. We have organized six scientific meetings and many inter-group visits, and exchanged hundreds of E-mails. The results are gratifying in many respects.

First, we have produced one of the best resources on alternative transcription worldwide. This resource is called ASTD, the Alternative Splicing and Transcript Diversity database, and is hosted at the EBI (http://www.ebi.ac.uk/astd/). The database is the result of a computational pipeline developed by ATD partners [1]. Alternative transcript information is provided for human, mouse and rat. Table 1 presents statistics for the human section. This compares favorably with the main alternative splicing database in the USA (ASAP II, Nucl. Acids. Res. 35:D93-D98, 2007), which contains alternative transcript data for only ~12000 human genes. In addition, the ASTD database is now fully integrated with other EBI databases such as Ensembl, which guarantees seamless analysis sessions for users and a high visibility for ATD research.

Table1. Statistics for Release 1 of the human ASTD, based on Ensembl Human v36

⊟ **Gene information**

Number of genes :	16560
Number of genes with an ASTD transcript :	16560
Number of genes with an ASTD transcription_start_site :	7694
Number of genes with an ASTD polyA_site :	7824
Number of genes with an ASTD splicing event :	11018
Number of genes with multiple ASTD transcripts :	13867
Number of genes showing splice events and multiple transcription_start_sites :	6656
Number of genes showing splice events and multiple polyA_sites :	4296
Proportion of genes undergoing alternative splicing:	67 %
Proportion of genes undergoing alternative polyadenylation:	47 %
Proportion of genes undergoing alternative transcription_start_sites:	46 %

The ASTD query interface (Figure 2A) is another major deliverable of the ATD project and the key to a widespread dissemination of our work. This interface enables researchers to query the immense complexity of the alternative transcriptome along many different lines. Users can locate alternative transcripts in the complete genome (Figure 1B) and track them down to the gene level visualizing synthetically all known isoforms for a gene (Figure 1C), to the transcript level visualizing detailed exon information (Figure 1D), or to the “event” level visualizing how alternative splicing events occur (Figure 1E). Furthermore, for each transcript isoform, users have access to expression information presented using a normalized vocabulary (Figure 1F).

The ASTD web server also provides various analysis tools that enable researchers to identify alternative transcripts of special medical or biological interest (Figure 2). Users may analyse specific genes or genome positions (Figure 2A), or select a larger group of genes on the basis of their function (Figure 2B) or their expression in specific tissues or diseases (Figure 2C). In one powerful analysis tool, users can select two complex pools of tissues or diseases, and the ASTD server will display all alternative transcripts that are specifically expressed in one of the pools and not in the other (Figure 2D).

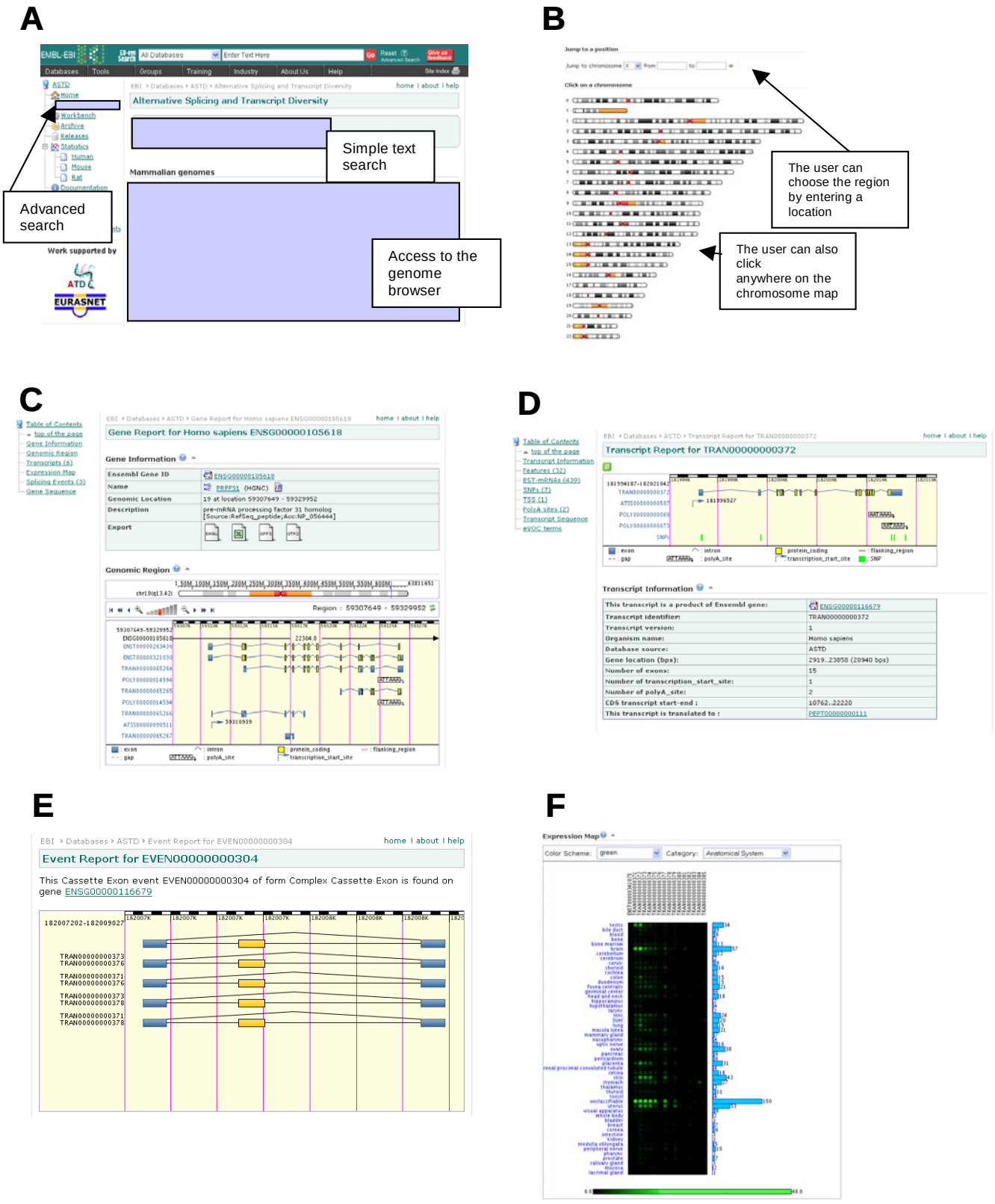

Figure 1. Screenshots of the ASTD database access pages. A: the home page, giving access to human, mouse and rat alternative transcript datasets. B: the genome browser, allowing access to genes directly from the chromosome map. C: the gene view, showing all transcript variants for a single gene, together with coding/non-coding information. D: the transcript view, showing positions of exons, sequence, etc. for a single transcript. E: the Event view, showing proofs for a specific alternative splicing event. F: the expression map, showing tissue-, disease- or development-specific expression levels of each isoform. This exemple shows expression levels per tissue

Converting alternative transcript data into functional discovery was another major objective of our project. For this purpose, we created tools for regulatory motif identification. We now have a complete system that is able to identify over-represented sequence motifs in very long lists of differentially expressed transcripts (as provided by the database analysis tool) more efficiently than any other software [art. in prep.]. Using this and other tools developed by ATD partners, we derived new lists of motifs potentially involved in post-transcriptional regulation. Some of these motifs are being further scrutinized, while others have been published [2].

Figure 2. Screenshots of the ASTD database query pages. A: advanced gene selection; B: advanced Gene Ontology-based selection; C: advanced expression state-based selection; D: differential expression analysis from two selected gene pools.

The ATD project had a significant emphasis on the experimental validation of predicted transcripts. In the third year, we completed the experimental validation of over 500 different transcripts in human and mouse. Mouse transcripts were selected for their particularly long 3' end structure and inter-species conservation. About 40 previously unknown transcripts with long 3' extension were discovered and published [3]. Human transcripts were selected for their putative expression in cancer cells. Starting with a selection of over 400 genes from ASTD queries, successive screening procedures produced a short list of 73 alternative transcripts with proven cancer-specific expression. This is probably one of the most exciting results of the ATD project, and it is also very rewarding because it involved

collaborative inputs from many ATD partners who supported this effort all the way from the initial computational predictions to the final, high impact biomedical findings. These experimental validations of tumor-specific isoforms are being published [4 and art. in prep. by Univ Heidelberg].

En route to these final achievements, we developed several tools and methods that will serve the general alternative transcription community. This includes a new benchmark to evaluate alternative splicing annotation methods [5], new software for the identification of tissue biases in expression data [6,7] and a new annotation model for the representation of alternative transcripts [article in prep. and Deliverable 2.2]. We also published several computational analyses that exploited ATD data to focus on diverse biological aspects of alternative splicing [8] and polyadenylation [9,10].

References

1. Le Texier V, Riethoven JJ, Kumanduri V, Gopalakrishnan C, Lopez F, Gautheret D, Thanaraj TA. (2006) AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. **BMC Bioinformatics** 7:169.
2. Legendre M, Ritchie W, Lopez F, Gautheret D. (2006). Differential repression of alternative transcripts: a screeen for miRNA targets. **PloS Comput. Biol.** 2(5): e43.
3. Moucadel V, Lopez F, Ara T, Benech P, Gautheret D (2007). Beyond the 3' end: experimental validation of extended transcript isoforms. **Nucl. Acids Res.** *In press.*
4. Pospisil H, Herrmann A, Butherus K, Pirson S, Reich JG, Kemmner W (2006). Verification of predicted alternatively spliced Wnt genes reveals two new splice variants (CTNNB1 and LRP5) and altered Axin-1 expression during tumour progression. **BMC Genomics** 13;7:148.
5. Shah PK, Jensen LJ, Boue S, Bork P (2005). Extraction of transcript diversity from scientific literature. **PLoS Comput Biol.** 1(1):e10.
6. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. **Nucleic Acids Res.** 14;33(5):1544-52.
7. Reimand J, Kull M, Peterson H, Hansen J and Vilo J (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. **Nucleic Acids Research**, 1–8
8. Plass M and Eyras E (2006) Differentiated Evolutionary Rates in Alternative Exons and the Implications for Splicing Regulation. **BMC Evolutionary Biology** 6(1):131.
9. Ara T, Lopez F, Ritchie W, Benech P, Gautheret D. (2006). Conservation of alternative polyadenylation patterns in mammals. **BMC Genomics** 7:189.
10. Lopez F, Granjeaud S, Ara T, Ghattas B, Gautheret D. (2006). The disparate nature of “intergenic” polyadenylation sites. **RNA** 12:1794-1801.

2. Dissemination and use

The major publishable results generated to date are the following.

(i) ASTD database and web server.

(ii) Novel regulatory motifs associated to post-transcription gene regulation

(iii) Cancer-specific transcript isoforms.

Knowledge dissemination is carried out through the ATD consortium web site which contains information on all public deliverables of the ATD project, the organization of two "Alternate Transcript Diversity Symposium" (EBI in November 2004 and EMBL in March 2006) and, most importantly the ASTD database and webserver hosted at EBI (Partner CR2A).

The ASTD database and webserver will have very significant impact on the field of alternative transcript research. ASTD will be an invaluable tool for laboratories interested in the structure, function or expression of alternative transcripts from human and other model organisms. Its scope, data quality, integration with other genome databases should immediately position it as the major European resource on alternative transcription, on an equal standing with ASAP, the current leading alternative transcription resource in the USA. As both databases have their specific strengths, this high level competition will be a benefit for the whole community. The data-mining tools of ASTD are a particularly useful aspect, as they will assist clinicians in their quest for disease markers.

Although several ATD-related articles have been published in high ranking journals during the 3-year period (see Annex I), we expect that the works with highest impact are the ones resulting from our final collaborative efforts in year 3. The next year should thus witness several high-impact publications, the most important being the ASTD database itself, and the experimental validation of novel cancer markers.

Annex 1 – Final plan for using and disseminating the knowledge

Section 1 - Exploitable knowledge and its use

The primary goal of the ATD project was to create a freely-accessible integrated database of alternative transcripts, associated prediction algorithms and bioinformatics tools.

However, the project management team in collaboration with the partners concerned were also assessing on a regular basis the progress of the deliverables with potential for IP protection and exploitation specifically concerning results related to WP4:
- Disease specific ATs, co-expressed isoforms and associated regulatory patterns (Del 4.2, 3.2, 3.3); Potential disease-specific ATs, which can be further evaluated for their use as disease markers, prognosis tool or drug target as potential are being closely evaluated and followed pending their experimental confirmation. In addition, co-regulated transcript and associated regulatory motifs involved in transcription, splicing or polyadenylation control will be assessed as targets for nucleic acid based drug, or as hot spots for disease-causing mutations, with applications in risk patient screening.

At this point promising data have been generated but it is still too early to give any details on the strategy since further validation experiments are necessary. The project manager (Inserm Transfert) however will be staying in close contact with the partners who own the knowledge (i.e. Partner UH) to ensure that knowledge is protected as soon as sufficient validation, i.e. on cancer specific transcript isoforms, is available. Should IP protection not be appropriate, the results, i.e. identification of cancer-specific transcript isoforms, will be published as soon as possible.

Section 2 – Dissemination of knowledge

Overview table

Planned/ actual Dates	**Type**	**Type of audience**	**Countries addressed**	**Size of audience**	**Partner responsible /involved**
14/06/2004	*Project website*	*General public*	International		*CR9, CO1*
22-23/11/2005	*Conference Symposium on alternate transcript diversity, Hinxton, Cambridge*	*Research*	International	*200*	*CR2a, CO1, CR2b, CR5, CR6, CR7, CR12*
27/04–01/05/ 2005	*Conference, Regulation of pre-mRNA splicing, Aussois, France*	*Research*	International		*CO1*
09-12/06/2005	*International Conference of Bioinformatics, Tartu, Estonia*	*Research*	International		*CR7*
23-24/06/2005	*ISMB2005 Alternative Splicing Special Interest Group meeting, Detroit, USA*	*Research*	International		*CO1, CR2a*
28/09-1/10/2005	*Conference, ECCB 2005, Madrid, Spain*	*Research*	International	*900*	*CR6*
21-23/03/2006	*Conference, Symposium on Alternate Transcript Diversity II, EMBL Heidelberg, Germany*	*Research*	International	*200*	*CR2a, CO1, CR2b, CR5, CR7,CR8, CR13*
March 2006	*3rd Annual Conference MidSouth Computational Biology and*	*Research*	International		*CR5*

Planned/ actual Dates	Type	Type of audience	Countries addressed	Size of audience	Partner responsible /involved
	Bioinformatics Society				
22.-26. April 2006	*Conference, EURASNET kick-off meeting*	*Research*	International	*120*	*CR2b*
5.-10. August 2006	*Conference, Alternative Splicing "Special Interest Group Meeting, ISMB 2006*	*Research*	International		*CR5, CR13*
13.-17. September 2006	*Conference, Genome Informatics meeting. CSHL and Welcome Trust joint meeting*	*Research*	International		*CR13*
October 2006	*Workshop, Advanced Bioinformatics workshop, CSHL*	*Research*	International		*CR5*
October 2006	*BeNeLux Bioinformatics Conference, Bioinformatics for Food and Health*	*Research*	International		*CR5*
18.-20. December 2006	*Conference, 17th International Conference on Genome Informatics*	*Research*	International		*CR5*
December 2006	*Workshop, Genome Network Expression Cluster Workshop, Riken Institute*	*Research*	International		*CR5*
28.-30. January 2007	*Conference, First Southern African Bioinformatics Workshop 2007*	*Research*	International	*100*	*CR5*
21.-24. January 2007	*Conference, ECCB 2006*	*Research*	International		*CR13*
14.-18. April 2007	*Conference, EURASNET annual meeting*	*Research*	International	*120*	*CR2b*

ATD project website (www.atdproject.org): The website is used as the major tool for dissemination of generated knowledge to the general public. All public data and deliverables have been made available via the public website under the header "ATD data releases":
http://www.atdproject.org/QuickPlace/it-atdproject/Main.nsf/h_Toc/33EEC00CD44BF442C1257115003DB3E7/?OpenDocument

In addition, the beta version of the **ASTD Database Website** will be released end of April 2007 (http://www.ebi.ac.uk/astd/).

Conference: Symposium on Alternate Transcript Diversity – Data, Biology, and Therapeutics" 22.-23. November, 2004 in Hinxton, UK (http://www.ebi.ac.uk/Information/events/atd-sympo/). The conference was organized by partner CR2a. The majority of ATD members participated at the symposium. Data generated within the ATD project were presented by the consortium members Alphonse Thanaraj (CR2a, oral presentation), Daniel Gautheret (Inserm, oral presentation), Winston Hide (CR5, oral presentation), Eduardo Eyras (CR6, poster selected for oral presentation) and Takeshi Ara (Inserm, poster) .

Conference: Regulation of pre-mRNA splicing, 27.04.-01.05.2005, Aussois, France (http://www.cnrs.fr/SDV/Actions/cjmluhrmann_e.html).
Poster presentations by partner CO1.

Conference : 7th International Conference BIOINFORMATICS 2005, June 9-12, 2005 Tartu, Estonia (http://bioinfo.ebc.ee/bio2005/indexen.php).
Partner CR7 was part of the organizing committee.

Conference: ISMB 2005 - Alternative Splicing Special Interest Group meeting June 23-24, 2005, Detroit USA (http://biolinfo.org/as-sig/).
Oral presentations by partners CO1 and CR2a.

Conference: ECCB 2005 - 4th European Conference on Computational Biology September 28 –October 01, 2005, Madrid, Spain (http://www.eccb05.org/).
Oral presentations by partners CO1, CR6
Poster presentations by partner CR2a

Conference: Symposium on Alternate Transcript Diversity II, 21. – 23. March 2006, EMBL Heidelberg, Germany
Partner CR2A was part of the organizing committee. Partner CO1 and CR5 were part of the Programme committee and chaired two independent sessions.
Oral presentations by Partners CR2B, CR13
Poster presentations by Partners CR13, CO1 and CR5.
A poster on the ATD database was presented by the EBI and the ATD consortium.

Conference: 3rd Annual Conference MidSouth Computational Biology and Bioinformatics Society. Baton Rouge, Louisiana, March 2006
Oral presentation Partner CR5

Conference: EURASNET kick-off meeting, Sigtes, Spain - 22nd-26th April 2006
Oral presentation Partner CR2b

Conference: Alternative Splicing Special Interest Group Meeting, ISMB 2006, 5. – 10 August 2006, Fortaleza, Brazil
Oral presentations Partners CR5 and CR13

Conference: Genome Informatics meeting. CSHL & Wellcome Trust joint meeting. Cambridge, UK, 13-17 September 2006
Oral presentation Partner CR13

Conference: 17th international conference on Genome Informatics 2006, Yokahama, Japan, 18.- 20. December
Poster presentation Partner CR5

Workshop: Genome Network Expression Cluster Workshop, Riken Institute, Yokohama, Japan, December 2006
Oral presentation Partner CR5

Workshop: Advanced Bioinformatics workshop. Cold Spring Harbour Laboratories. New York, October 2006
Oral presentation Partner CR5

Conference: BeNeLux Bioinformatics Conference. Bioinformatics for Food and Health, Wageningen, October 2006
Oral presentation Partner CR5

Conference: ECCB2006 conference, Eilat, Israel, 21-24 January 2007.
Poster presentation Partner CR13

Conference: First Southern African Bioinformatics Workshop, Johannesburg, 28-30 January 2007:
Partner CR5 was part of the programme committee.
Poster presentation Partner CR5

Conference: EURASNET annual meeting, Bagnol, France - 14th-18th April 2007
Oral presentation Partner CR2b

Section 3 - Publishable results

Articles published since project start.

- Moucadel V, Lopez F, Ara T, Benech P, Gautheret D (2007). Beyond the 3' end: experimental validation of extended transcript isoforms. **Nucl. Acids Res.** *In press.*
- Reimand J, Kull M, Peterson H, Hansen J and Vilo J (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. **Nucleic Acids Research**, 1–8
- Lopez F, Granjeaud S, Ara T, Ghattas B, Gautheret D. (2006). The disparate nature of "intergenic" polyadenylation sites. **RNA** 12:1794-1801.
- Ara T, Lopez F, Ritchie W, Benech P, Gautheret D. (2006). Conservation of alternative polyadenylation patterns in mammals. **BMC Genomics** 7:189.
- Legendre M, Ritchie W, Lopez F, Gautheret D. (2006). Differential repression of alternative transcripts: a screeen for miRNA targets. **PloS Comput. Biol.** 2(5): e43.
- Le Texier V, Riethoven JJ, Kumanduri V, Gopalakrishnan C, Lopez F, Gautheret D, Thanaraj TA. (2006) AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. **BMC Bioinformatics** 7:169.
- Plass M and Eyras E (2006) Differentiated Evolutionary Rates in Alternative Exons and the Implications for Splicing Regulation. **BMC Evolutionary Biology** 6(1):131.
- Pospisil H, Herrmann A, Butherus K, Pirson S, Reich JG, Kemmner W (2006). Verification of predicted alternatively spliced Wnt genes reveals two new splice variants (CTNNB1 and LRP5) and altered Axin-1 expression during tumour progression. **BMC Genomics** 13;7:148.
- Shah PK, Jensen LJ, Boue S, Bork P (2005). Extraction of transcript diversity from scientific literature. **PLoS Comput Biol.** 1(1):e10.
- Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. **Nucleic Acids Res.** 14;33(5):1544-52.

Manuscripts submitted

- Ritchie W, Granjeaud S, Puthier & Gautheret. Measure of Transcript Isoform Entropy Shows Extent of Splicing Disruption in Cancer. Submitted.

Manuscripts in preparation:

- The ATD consortium. *The Alternative Splicing and Transcript Database.*
- Ch. Fallsehr; O. Hoffmann, M. Kull, V Jaak, W. Hide, M. v. Knebel Doeberitz. *Experimental validation of predicted feature specific alternative transcripts from the Altsplice database - comparison of experimental reality to database representation*
- M Suyama, E Harrington, S Vinokourova, M von Knebel Doeberitz and P Bork *Co-occurrence of intronic splicing regulators in mammalian genomes*
- E Harrington and P Bork Sircah. A *tool for the detection and visualisation of alternative transcripts*
- O Hofmann, M Sammath, C Mungall, K Eilbeck, S Lewis, E Eyras, W Hide. *Standards in Alternative Splicing.*
- M Kull, O Hofmann, C Fallsehr, M von Knebel Doeberitz, W Hide, J Vilo. *DIDAS: Identifying expression state specific splice variants.*
- C Fallsehr, M Kull, O Hofmann, J Vilo, W Hide, M von Knebel Doeberitz. *Cancer-specific splice-variants in colon and lung tumors.*