Supplemental Material for FasterDB and Elexir

Contents

1 Datasets used to build FasterDB/Elexir **2**
- 1.1 FasterDB transcript database . . . 2
- 1.2 Elexir experimental datasets . . . 2
- 1.3 FasterDB splicing factor datasets . . . 5
  - 1.3.1 siRNA datasets . . . 5
  - 1.3.2 CLIPseq datasets . . . 5

2 Algorithms and methods **7**
- 2.1 FasterDB . . . 7
  - 2.1.1 Splicing event detection . . . 7
  - 2.1.2 Complementary features . . . 8
- 2.2 Exon array analysis and Elexir . . . 9
- 2.3 Splicing factors . . . 10
  - 2.3.1 siRNA dataset analysis . . . 10
  - 2.3.2 CLIPseq dataset analysis . . . 10

3 Web interfaces **12**
- 3.1 FasterDB tutorial . . . 12
- 3.2 Elexir tutorial . . . 21
- 3.3 Splicing factor tutorial . . . 24

Part 1
Datasets used to build FasterDB/Elexir

This chapter focuses on the datasets we have used. In particular, are mentioned in this section from where the datasets were obtained and how they were preprocessed. When applicable, the URLs of the tools are mentioned and corresponding publications are indicated.

1.1 FasterDB transcript database

Human and mouse exons have been collected from EnSEMBL[1] (release 60, assemblies GRCh37 and NCBIM37) and aligned against the NCBI[2] transcript database using Megablast (v2.2.25).

The human and mouse exons extracted from the transcripts (hereinafter 'transcript exons') have been aligned against genomic sequences (using sim4[3]) as to define their chromosomal coordinates and have been clustered by genomic position in order to define their genomic counterparts (hereinafter 'genomic exons').

1.2 Elexir experimental datasets

Elexir proposes the visualization of the gene expression profiles at the exon level using public expression datasets. For each gene, two and three different profiles are generated respectively for mouse and human (i) expression over various tissues/organs, (ii) expression in various cell lines, and (iii) expression in various cancer cell lines.

The tissue expression profiles are derived from a sample dataset provided by Affymetrix[4]. It consists in 11 normal human triplicate tissues.

The cell line expression dataset was obtained from public repositories such as the ENCODE consortium project and represents a large collection of human and mouse cell lines from different origin (both in term of tissues and cell types). In addition, some datasets were extracted from GEO[5] using the following projects: GSE15998 (mouse), GSE19090 (human), and GSE15805 (human). The table 1.1 contains detailed information about the selected experiments for human, grouped by cell type of interest (endothelial, epithelial, and fibroblast). Other cell lines that have not been used in the original publication are presented in Table 1.2. The mouse cell lines

[1] http://www.ensembl.org
[2] http://www.ncbi.nlm.nih.gov/
[3] http://pbil.univ-lyon1.fr/members/duret/cours/inserm210604/exercise4/sim4.html
[4] http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx
[5] www.ncbi.nlm.nih.gov/geo/

are displayed in Table 1.3. For human, expression datasets from cancer cell lines have been gathered, details are displayed in Table 1.4.

Endothelial cell lines			
Cell line	**Definition**	**Sex**	**Karyotype**
HBMEC	Brain microvascular endothelial cells	U	Normal
HMVEC-dAd	Adult dermal microvascular endothelial cells	F	Normal
HMVEC-dBl-Ad	Adult blood microvascular endothelial cells	F	Normal
HMVEC-dBl-Neo	Neonatal blood microvascular endothelial cells	M	Normal
HMVEC-dNeo	Neonatal microvascular endothelial cells	M	Normal
HPAEC	Pulmonary artery endothelial cells	F	Normal
HRGEC	Renal glomerular endothelial cells	U	Normal
HUVEC	Umbilical vein endothelial cells	U	Normal
Epithelial cell lines			
Cell line	**Definition**	**Sex**	**Karyotype**
HAEpiC	Amniotic epithelial cells	U	Normal
HCPEpiC	Choroid plexus epithelial cells	U	Normal
HEEpiC	Esophageal epithelial cells	U	Normal
HIPEpiC	Iris pigment epithelial cells	U	Normal
HMEC	Mammary epithelial cells	U	Normal
HNPCEpiC	Non-pigment ciliary epithelial cells	U	Normal
HRCEpiC	Renal cortical epithelial cells	U	Normal
HRPEpiC	Retinal pigment epithelial cells	U	Normal
HRE	Renal epithelial cells	U	Normal
NHBE	Bronchial epithelial cells	F	Normal
PrEC	Prostate epithelial cell line	U	Normal
SAEC	Small airway epithelial cells	U	Normal
Fibroblast cell lines			
Cell line	**Definition**	**Sex**	**Karyotype**
AG04449	Fetal buttock/thigh fibroblast	M	Normal
AG04450	Fetal lung fibroblast	M	Normal
AG09319	Gum tissue fibroblasts	F	Normal
AoAF	Aortic adventitial fibroblast cells	F	Normal
BJ	Skin fibroblast	M	Normal
Fibrobl	Child fibroblast	F	Normal
HCF	Cardiac fibroblasts	U	Normal
HCFaa	Cardiac fibroblasts (adult atrial)	F	Normal
HFF	Foreskin fibroblast	M	Normal
HGF	Gingival fibroblasts	U	Normal
HMF	Mammary fibroblasts	F	Normal
HPAF	Pulmonary artery fibroblasts	U	Normal
HPdLF	Periodontal ligament fibroblasts	M	Normal
HPF	Pulmonary fibroblasts	U	Normal
HVMF	Villous mesenchymal fibroblast cells	U	Normal
NHDF-Ad	Adult dermal fibroblasts	F	Normal
NHDF-neo	Neonatal dermal fibroblasts	U	Normal
NHLF	Lung fibroblasts	U	Normal

Table 1.1: Human cell lines for which exon expression profiles are included in FasterDB, grouped by cell type of interest. (M male, F female, U unknown)

Cell line	Definition	Sex	Karyotype
AG09309	Adult toe fibroblast	F	-
AG10803	Abdominal skin fibroblasts	M	-
GM12878	B-lymphocyte	F	Normal
GM06990	B-lymphocyte	F	-
HAc	Astrocytes-cerebellar	U	Normal
HA-sp	Astrocytes spinal cord	U	Normal
HCM	Cardiac myocytes	U	Normal
HConF	Conjunctival fibroblast	U	-
HEK293	Embryonic kidney	U	-
hESC	Human embryonic stem cell	-	-
HL-60	Promyelocytic leukemia cells	F	Cancer
HSMM	Skeletal muscle myoblasts	U	Normal
Jurkat	T lymphoblastoid	M	Cancer
K562	Leukemia	F	Cancer
NB4	Leukemia	U	Cancer
NH-A	Astrocytes	U	Normal
RPTEC	Renal	U	Normal
SKMC	Skeletal muscle	U	Normal
SK-N-SH	Neuroblastoma	F	Cancer
Th1	Primary Th1 T cells	U	-
Th2	Primary Th2 T cells	U	-
WI-38	Embryonic lung	F	Normal

Table 1.2: Other human cell lines for which exon expression profiles are included in FasterDB. (M male, F female, U unknown)

Cell line	Tissue	Cell type	Disease
Raw264.7	Abelson murine leukemia virus-induced tumor; ascites	Macrophage	Abelson murine leukemia virus-induced tumor
Neuro-2a	Brain	Neuroblast	Neuroblastoma
C3H/10T1/2	Embryo	-	Sarcoma
BAF3	Bone marrow	pro-B-cell	-
MIMCD	Kidney, medulla/collecting duct	SV40 transformed	-
NIH/3T3	Embryo	Fibroblast	-
3T3-L1	Embryo	Fibroblast	-
C2C12	Muscle	Myoblast	-
Min6	Pancreatic	Beta cell	-

Table 1.3: Mouse cell lines for which exon expression profiles are included in FasterDB.

Cell line	Definition	Sex	Karyotype
A549	Lung carcinoma	M	Cancer
BE2_C	Neuroblastoma	M	Cancer
Caco-2	Colorectal adenocarcinoma	M	Cancer
Gliobla	Glioblastoma	U	Cancer
HCT-116	Colorectal carcinoma	M	Cancer
HeLa-S3	Cervical carcinoma	F	Cancer
HepG2	Hepatocellular carcinoma	M	Cancer
LNCaP	Prostate adenocarcinoma	M	Cancer
MCF-7	Mammary gland, adenocarcinoma	F	Cancer
Medullo	Medulloblastoma	U	Cancer
PANC-1	Pancreatic carcinoma	M	Cancer
SK-N-MC	Supra-orbital human brain tumor	F	Cancer
WERI-Rb-1	Retinoblastoma	F	Cancer

Table 1.4: Cancer cell lines for which exon expression profiles are included in FasterDB. (M male, F female, U unknown)

1.3 FasterDB splicing factor datasets

Additional experimental datasets have been collected for splicing factors. The Table 1.5 contains the splicing factors for which datasets have been collected as well as their canonical binding motifs.

Splicing factor	Array datasets	CLIP datasets	Motifs
9G8			acgagagay, wggacra
CUG-BP			ygctyk, stgt
DAZAP1			aaatag
ESRP	✓		tgg
FOX2		✓	gcatg
FUS		✓	ggtg
HNRNPA1		✓	tagrsw
HNRNPA2B1		✓	gtagtag, aggatng
HNRNPAB			atagca
HNRNPC		✓	tttt
HNRNPF		✓	gtngtng, gtggat
HNRNPH1	✓	✓	dggg, ggygg
HNRNPL	✓		cacaca
HNRNPL_2			caca
HNRNPM		✓	ggttggtt
HNRNPU		✓	gtgtg
HuR	✓	✓	tttdttt
MBNL			ygcy
NOVA1			ycab
PTB	✓	✓	ctctct, tctt
QKI			actaa
RBM4	✓		gggg
SC35			tgcygyy, gryymsyr
SF2ASF			crsmsgw, tgrwgvh
SRP20			wcwwc, ctcktcy
SRP40			yywcwss
SRP55			yrcrkm
TIA		✓	ttttt
TRA2A			gaagaggaag
TRA2B			aagtgtt,gaagaa, ghvvganr
YB1			caaccacaa

R = A or G, Y = C or T, K = G or T, M = A or C, S = C or G, W = A or T, B = C or G or T, D = A or G or T, H = A or C or T, V = A or C or G, N = A or C or G or T

Table 1.5: Summary of the available datasets for various splicing factors.

1.3.1 siRNA datasets

The Table 1.5 contains the list of splicing factors for which exon arrays have been performed upon their depletion through siRNA and have been uploaded in our database. The Table 1.6 contains more details about the origin of the datasets. More details about the way these datasets have been processed can be found in section 2.2.

1.3.2 CLIPseq datasets

FasterDB also permits the visualization of CLIPseq datasets of several splicing factors. These datasets were collected from public databases and processed when necessary. The Table 1.5

contains the list of splicing factors for which such datasets have been uploaded in our database. The Table 1.7 contains details about the origin of the datasets for both human and mouse.

siRNA target	**GEO**	**Contributors**	**PubMed**
ESRP	GSE17468	Warzecha CC *et al.*	19829082
HNRNP H	GSE12386	Xiao X	19749754
HNRNP L	GSE8945	Hung LH and Bindereif A	18073345
HuR	GSE29780	Mukherjee N	21723170
PTB	GSE23514	Llorian M *et al.*	20711188
RBM4	GSE32933	Markus MA *et al.*	-

Table 1.6: Origin of the siRNA datasets integrated into FasterDB.

Human			
Splicing factor	**GEO**	**Contributors**	**PubMed**
FOX2	UCSC Track	Yeo GW *et al.*	19136955
FUS	GSE40651	Yeo GW and Cleveland DW	2302393
hnRNP A1	GSE34993	Huelga SC and Yeo GW	22574288
hnRNP A2B1	GSE34993	Huelga SC and Yeo GW	22574288
hnRNP C	E-MTAB-1371	Zarnack *et al.*	23374342
hnRNP F	GSE34993	Huelga SC and Yeo GW	22574288
hnRNP H1	GSE23694	Katz Y *et al.*	21057496
hnRNP M	GSE34993	Huelga SC and Yeo GW	22574288
hnRNP U	GSE34993	Huelga SC and Yeo GW	22574288
hnRNP U	GSE34491	Xiao R *et al.*	22325991
HuR	GSE29943	Lebedeva S *et al.*	21723171
HuR	GSE29780	Mukherjee N	21723170
PTB	GSE19323	Xue Y *et al.*	20064465
TIA1	E-MTAB-432	Wang Z *et al.*	21048981
TIAL1	E-MTAB-432	Wang Z *et al.*	21048981
Mouse			
Splicing factor	**GEO**	**Contributors**	**PubMed**
FUS	GSE40651	Yeo GW and Cleveland DW	2302393
TDP43	GSE40651	Yeo GW and Cleveland DW	2302393
Tra2	UCSC Track	Grellscheid S *et al.*	22194695

Table 1.7: Provenance of the CLIPseq datasets integrated into FasterDB for human and mouse.

Part 2

Algorithms and methods

2.1 FasterDB

2.1.1 Splicing event detection

The alignments between the human and mouse transcripts and their respective genomes described in 1.1 allow the definition of seven major splicing events (alternative first exon, alternative last exon, alternative 3' splice site, alternative 5' splice site, intron retention, exon deletion, and exon skipping) whose definitions can be found below:

Alternative first exon. To be defined as an alternative first exon, a transcript exon has to be the first exon of at least one transcript. In addition, if there are other internal exons at this genomic position, it has to start at least 10 nt upstream of the first position of the corresponding genomic exon.

Alternative last exon. To be defined as an alternative last exon, a transcript exon has to be the last exon of at least one transcript and, if there were other internal exons at this genomic position, it has to end at least 10 nt downstream of the last position of the corresponding genomic exon.

Alternative 3' splice site. It corresponds to a transcript exon whose starting position is different from the starting position of its corresponding genomic exon.

Alternative 5' splice site. It corresponds to a transcript exon whose ending position is different from the ending position of its corresponding genomic exon.

Intron retention. An intron retention corresponds to a complete intronic sequence being included into at least one transcript.

Exon deletion. There exists an exon deletion when a transcript exon presents an internal deletion when aligned against its corresponding genomic exon.

Exon skipping. An exon skipping event is detected when an exon is absent in at least one transcript and if it is not an alternative first or last exon.

In addition, in order to identify even more alternative last exons, the dbEST database[1] was obtained from the NCBI. Sequences annotated as 3' EST and/or where presence of a polyatail was reported have been retrieved. Polyadenylation tails have been extracted and sequences

[1] http://www.ncbi.nlm.nih.gov/dbEST/

were then aligned to the genome using Blat[2]. These novel alignments allowed us to identify additional alternative last exons following the criteria described above.

These splicing events can then be used to guide the expression data analysis (in order to check only the known splicing events). Alternatively, the expression data analysis can be made with no *a priori* (as to be able to detect novel splicing events).

2.1.2 Complementary features

Additional characteristics provided by other databases and/or processed by a series of tools have been added to FasterDB. The following paragraphs describes these features, their origin and, if necessary, how the datasets were processed.

miRNA binding sites

miRNA binding sites were predicted using Pita[3], Miranda[4] and Pictar[5] algorithms. All predictions were linked to the MiR database[6].

Splice sites

For each predicted alternative splice site, scores were computed using MaxEntScan[7]. This score is calculated using a sequence that covers both sides of the splicing site (3 bases in the exon and 6 bases in the intron for 5' splice sites; 20 bases in the intron and 3 bases in the exon for 3' splice sites). MaxEntScan uses Maximum Entropy Models (MEMs) to compute log odds ratios that represent the likelihood of being a real splicing site (the higher a score, the higher the probability that the associated sequence is a true splice site).

UTR and motifs in UTRs

For each gene, and using their associated transcripts, a non redundant repertory of untranslated regions was established. For each transcript, coding regions were defined as being the longest sequence between a start and a stop codons. Thus, 5' untranslated transcript regions were defined as regions upstream of the start codon and 3' untranslated transcript regions were defined as regions downstream of the stop codon. Then transcript UTRs were aligned and clustered similarly to what is described in section 1.1. For each group, the longest untranslated region was defined as the genomic UTR reference. If the sequence was small, we only kept it when a transcription start site was present at the beginning of the region or a polyadenlyation site was present at the end for 5' UTR and 3' UTR respectively. Furthermore, genomic UTR references were analyzed using PatSearch in order to determine if putative motifs were present.

Conserved exons

Conserved exons between human and mouse organisms have been identified. In a first step, human genes and their orthologous mouse genes were retrieved from EnsEMBL using Biomart[8]. Each human exon was then aligned against each exon of its orthologous mouse gene. Conserved

[2]http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
[3]http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html
[4]http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/
[5]http://pictar.mdc-berlin.de/
[6]www.mirbase.org/
[7]genes.mit.edu/burgelab/
[8]http://www.ensembl.org/biomart/martview/

exons were defined if the percentage identity and the DNA coverage criterias validated our thresholds.

Splicing factor binding sites

Splicing factor binding sites were sought using PatSearch on the genomic sequences. Patterns shorter than 10 nucleotides, as defined in the IUPAC code, were searched. Each exon sequence and the first 200 nucleotides of both flanking introns were mined for motifs.

in silico PCR

A multi-alignment of the transcript exons was performed using ClustalW. For a given gene, the sequences of the transcript exons corresponding to one genomic exon were aligned. The results for each genomic exon position were then assembled to present the multi-alignment of the different cDNAs. This is used later to help users to design PCR primers.

2.2 Exon array analysis and Elexir

Elexir is a tool designed to analyze and visualize the gene differential expression profile between two conditions. The exon arrays are first analyzed using the APT affymetrix software[9] (ignoring the mismatch probes, using RMA and quantile normalization), this produces the DABG and the raw intensities for each probe.

Then, and contrary to most approaches, the intensities are not summarized per exon or per gene. Instead, expression ratios between conditions A and B are computed for each probe and only then, the signal (*i.e.*, the ratios) is summarized per exon or per gene. Notice that only probes with no cross hybridization, low GC percentage ($< 18\%$), and detection above the background (DABG p-value < 0.05) are selected. When there are not enough probes or if detection above the background is not consistent, the gene is considered as not expressed (see the two algorithms below).

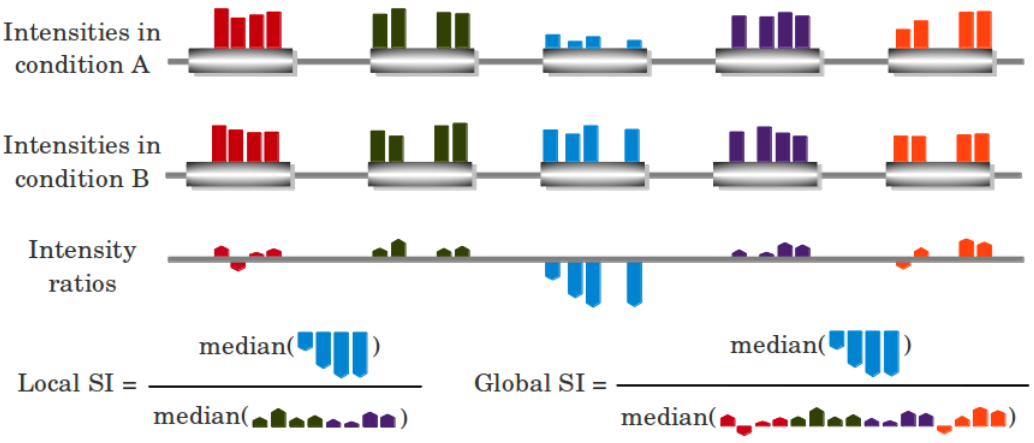

Figure 2.1: Illustration of the algorithms used to compute the Local SI and Global SI.

[9]http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx

In order to detect splicing events, inclusion/exclusion rates of each exon are computed, using two distinct algorithms that are summarized below and illustrated in figure 2.1. Both methods start from the same data, that is the ratio of the intensities between the two conditions for all probes. The objective is then to compute a splicing index for a given exon as illustrated in figure 2.1. When possible, the Local SI values are accompanied by a p-value obtained with a Student t-test to assess their significance.

Local Splicing index (Local SI) With this method, the raw differential expression of the exon n is summarized by computing the median of the intensity ratios ($IntA$, $IntB$) for all probes that target this exon (see equation (2.1)). This value is then adjusted by the differential expression of the flanking exons, which is computed exactly like the raw differential expression of the exon n.

$$Local\ SI = \frac{\underset{probe\ i \in exon\ n}{median}\left(\frac{IntA_i}{IntB_i}\right)}{\underset{probe\ j \in exon\ n-1,n+1}{median}\left(\frac{IntA_j}{IntB_j}\right)} \tag{2.1}$$

Global Splicing Index (Global SI) Similarly to the first method, the raw differential expression of the exon is summarized by computing the median of the intensity ratios for all probes that target this exon (see equation (2.2)). Then, this value is adjusted in a different way, by taking the raw differential expression of the whole gene (instead of the flanking exons).

$$Global\ SI = \frac{\underset{probe\ i \in exon\ n}{median}\left(\frac{IntA_i}{IntB_i}\right)}{\underset{probe\ j \notin exon\ n}{median}\left(\frac{IntA_j}{IntB_j}\right)} \tag{2.2}$$

2.3 Splicing factors

2.3.1 siRNA dataset analysis

Exon array after splicing factor depletion (see Table 1.6) have been analyzed as described above (section 2.2).

2.3.2 CLIPseq dataset analysis

When available, direct integration of processed datasets was performed. When necessary, the positions were converted to hg19 using liftover[10]. If only the raw files were available, the CLIP datasets have been mapped using Bowtie[11]. The Table 2.1 summarizes what has been done for each dataset.

[10] http://genome.ucsc.edu/cgi-bin/hgLiftOver
[11] http://bowtie-bio.sourceforge.net/index.shtml

Human		
Dataset	**Splicing factor(s)**	**Method**
GSE40651	FUS	Data lifted to hg19, integration
GSE34993	hnRNP A1/A2B1/F/M/U	Data lifted to hg19, integration
EMTAB1371	hnRNP C	Simple integration
GSE23694	hnRNP H1	SAM to BED conversion, data lifted to hg19, integration
GSE34491	hnRNP U	Data lifted to hg19, integration
GSE29943	HuR	Adapter removal, Bowtie mapping, integration
GSE29780	HuR	Simple integration
GSE19323	PTB	Data lifted to hg19, integration
EMTAB432	TIA1/TIAL1	Adapter removal, Bowtie mapping, integration
UCSC Track	FOX2	Data lifted to hg19, integration
Mouse		
Dataset	**Splicing factor(s)**	**Method**
GSE40651	FUS/TDP43	Simple integration
UCSC Track	Tra2	Simple integration

Table 2.1: Analysis of the CLIPseq datasets for FasterDB

Part 3
Web interfaces

This chapter describes the different web interfaces to our tools.

3.1 FasterDB tutorial

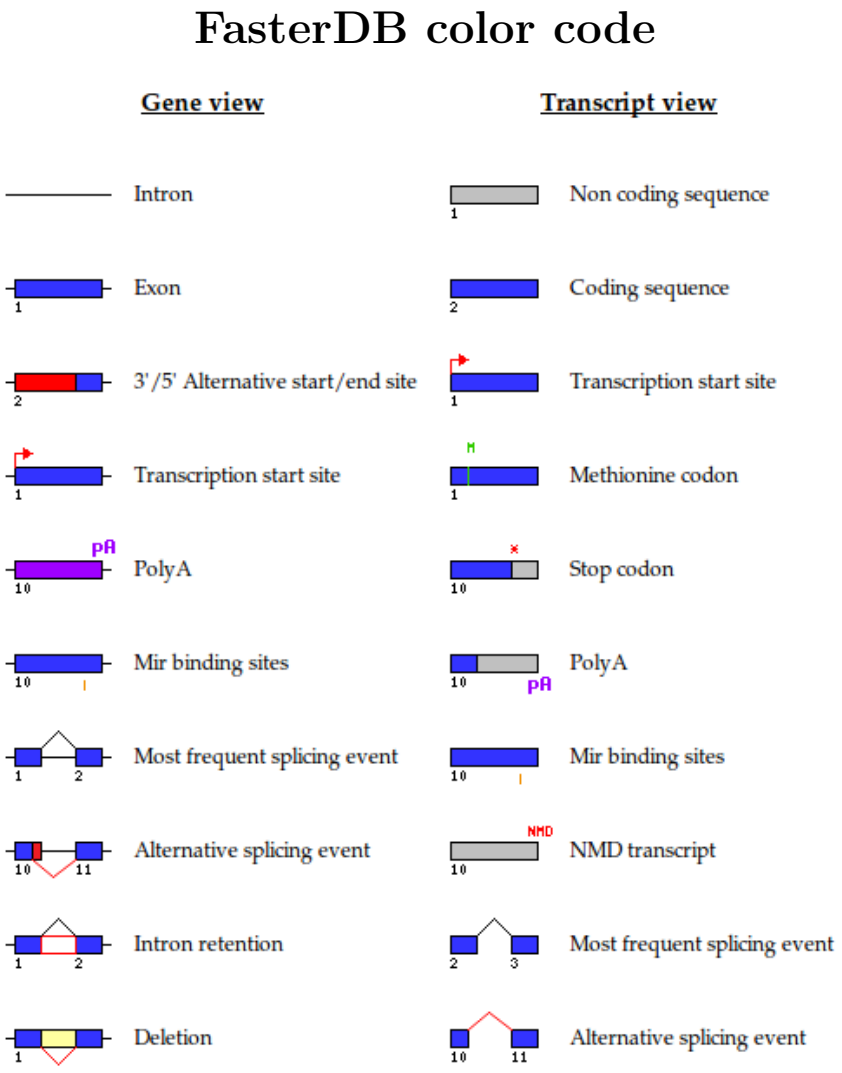

Figure 3.1: FasterDB and Elexir are using a specific color code to draw genes, transcripts and exons.

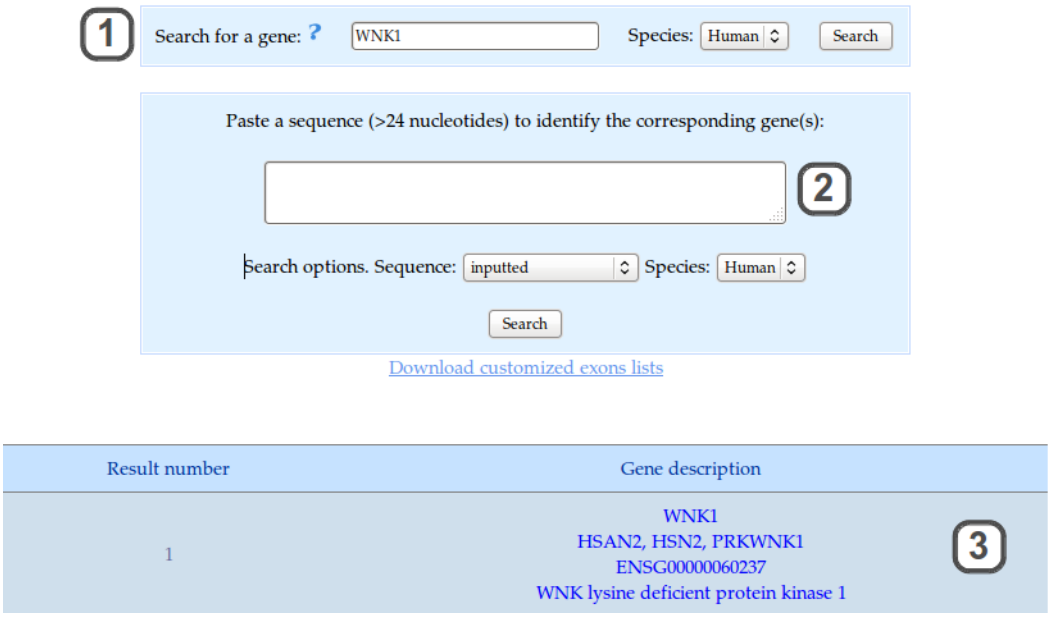

Figure 3.2: FasterDB home page is the search page. It can be used to access the gene pages (see above the search corresponding to the human WNK1 gene). The search engine can be used by providing keywords, gene identifiers (ENSEMBL, HUGO gene name, synonyms or gene description), chromosomal coordinates (square 1), or sequence (perfect match sequence against FasterDB genes - square 2) for mouse or human. The results of the request are displayed below the search parameters (square 3). Clicking on the help icon displays the search criteria for users as well as clickable examples.

FasterDB gene page for human WNK1

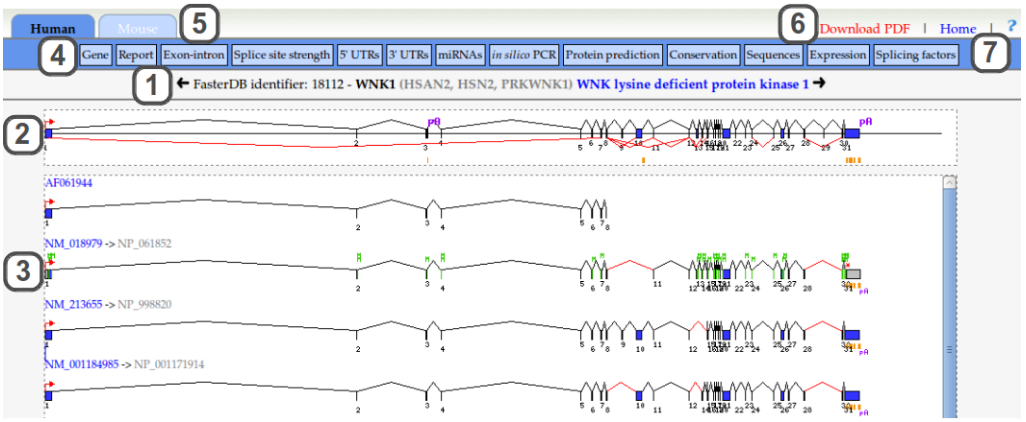

Figure 3.3: The gene page contains all the information about a given gene. In particular, it contains information about the gene (FasterDB identifier, HUGO name and synonyms) (square 1), a schematic representation of the gene (see section 3.1 for the color code) (square 2), a schematic representation of the transcripts associated to the gene (square 3), a toolbar with several buttons, each displaying additional information about the gene (see the following figures for more information) (square 4), a link to the FasterDB human's orthologous gene in the mouse genome (or vice-versa) (square 5), and a link to download the graphical representation of this gene page - square 6).

It is possible to browse the neighboring genes downstream or upstream by using the left or right arrows located around to the basic gene information (square 1). Alternatively, it is also possible to go back to the home page for another search (square 7). The help icon toggles the display of the figure legend (*i.e.*, color code as in Figure 3.2). Graphical elements are clickable to either obtain the corresponding fasta sequences (*e.g.*, exons, introns) or to link out to external references, such as the NCBI, where more information is available (*e.g.*, transcripts).

FasterDB gene table for human WNK1

	Gene
Symbol	WNK1
Synonyms	HSAN2, HSN2, PRKWNK1
Description	WNK lysine deficient protein kinase 1
EnsEMBL link	ENSG00000060237
Chromosome	12 - forward strand
Chromosomal location (UCSC link)	862089 - 1020618 (158530bp)
Sequence	sequence
Associated transcripts	AF061944 NM_018979 NM_213655 NM_001184985 NM_014823 BC021121 BC071959 BC094862 AJ296290 BC141881 AB002342 DQ925669 BC013629

Figure 3.4: The gene table can be displayed by clicking on the 'Gene' button of the main toolbar (see Figure 3.3). This table shows general information about the gene of interest (official symbol, synonyms, description, chromosomal location and associated transcripts). Links to external resources are also available: clicking on the Ensembl ID of the gene leads to its Ensembl page while clicking on chromosomal coordinates directs to the UCSC browser page of the gene. Finally, clicking on 'sequence' displays the fasta sequences of all the exons and introns of the gene.

FasterDB report table for human WNK1

Report	
Exon	Number of supporting transcripts
Transcription initiation & first exon(s)	
1	1
Transcription termination & last exon(s)	
3	3
31	4
Exon skipping	
2	1
3	1
4	1
5	1
6	1
9	6
10	5
11	1
13	5
14	4
29	7
Alternative 3' splice sites	
7	1
12	1
Alternative 5' splice sites	
1	1
24	1

Polyadenylation							
Identifier	Position	Gene coordinates of cleavage site	Chromosomal coordinates of cleavage site	Pattern	PolyA tail	Accession number	Type
PA1	3	74575	936663	AATAAA	yes	BC021121	Cdna
PA2	3	74575	936663	AATAAA	yes	BC071959	Cdna
PA3	3	74575	936663	AATAAA	yes	BC094862	Cdna
PA4	31	158530	1020618	AATAAA	yes	NM_018979	Cdna
PA5	31	158530	1020618	AATAAA	no	NM_213655	Cdna
PA6	31	158530	1020618	AATAAA	no	NM_001184985	Cdna
PA7	31	158530	1020618	AATAAA	yes	NM_014823	Cdna

Alternative 3' splice site (acceptor)			
Identifier	Position	Distance (bp)	Transcript
ASS1	7	111	BC141881
ASS2	12	3	NM_014823 AB002342

Figure 3.5: Features relative to alternative splicing events are summarized in the report table that can be displayed by clicking on the 'Report' button of the main toolbar (see Figure 3.3). For each type of alternative splicing event, the exon concerned and the number of transcripts having this alternative splicing event are displayed. Clicking on a given alternative splicing event displays further information about this event. Two examples of transcription end and exon skipping events tables are shown above. The 'Transcription Termination & last Exons' link shows polyadenylation sites. For each polyadenylation site, its exonic position, gene and chromosomal coordinates are shown. Moreover information about the signal pattern, the accession number and the type of sequence (cDNa/Est) of the transcript used to define the polyadenylation site are displayed. Finally, the presence of a polyA tail is tested and notified when detected in the sequence. Lastly, the 'Exon Skipping' table shows information about skipped exons by displaying their position and the transcripts that have been used to perform this annotation.

FasterDB exon-intron table for human WNK1

Exon-Intron

Position	Type	Gene coordinates	Chromosomal coordinates	Length	Sequence	Features
1	Exon	1 - 1402	862089 - 863490	1402	sequence	download
1	Intron	1403 - 60719	863491 - 922807	59317	sequence	-
2	Exon	60720 - 60892	922808 - 922980	173	sequence	download
2	Intron	60893 - 74119	922981 - 936207	13227	sequence	-
3	Exon	74120 - 74340	936208 - 936428	221	sequence	download
3	Intron	74341 - 77080	936429 - 939168	2740	sequence	-
4	Exon	77081 - 77238	939169 - 939326	158	sequence	download
4	Intron	77239 - 104238	939327 - 966326	27000	sequence	-
5	Exon	104239 - 104327	966327 - 966415	89	sequence	download
5	Intron	104328 - 106322	966416 - 968410	1995	sequence	-
6	Exon	106323 - 106542	968411 - 968630	220	sequence	download
6	Intron	106543 - 108090	968631 - 970178	1548	sequence	-
7	Exon	108091 - 108421	970179 - 970509	331	sequence	download
7	Intron	108422 - 109160	970510 - 971248	739	sequence	-
8	Exon	109161 - 109348	971249 - 971436	188	sequence	download
8	Intron	109349 - 112187	971437 - 974275	2839	sequence	-
9	Exon	112188 - 112442	974276 - 974530	255	sequence	download

Figure 3.6: Next comes the summary table of the exon/intron structure of the gene that can be displayed by clicking on the 'Exon-Intron' button of the main toolbar (see Figure 3.3). For each element, coordinates on the chromosome and on the gene are displayed as well as the element length. Clicking on 'sequence' displays the nucleotidic sequence of the element. For exons only, clicking on the 'download' link, in the 'features' column, produces a file containing further information about the exon of interest.

FasterDB splice site strength table for human WNK1

Splice site strength

Identifier	Site	Exon	Strength	Alternative	Sequence
S1	5' (donor)	1	9.65	no	CAGgtaaag
S2	5' (donor)	1	-19.55	yes	GCAgagcag
S3	5' (donor)	2	10.74	no	AACgtaagt
S4	3' (acceptor)	2	9.34	no	gttttggtttttgattttagGAT
S5	5' (donor)	3	7.76	no	TAGgtatgt
S6	3' (acceptor)	3	9.33	no	gtatactttgctttttctagGTA
S7	5' (donor)	4	8.46	no	AGTgtaagt
S8	3' (acceptor)	4	7.73	no	aatcgttcttttaatttaagGTA
S9	5' (donor)	5	11	no	AAGgtaagt
S10	3' (acceptor)	5	8.5	no	tttctgtttatggttttcagGGG
S11	5' (donor)	6	6.41	no	ATGgtaaat
S12	3' (acceptor)	6	5.68	no	cttttccctctgtttggaagATA
S13	5' (donor)	7	3.41	no	TATgtacgt
S14	3' (acceptor)	7	11.24	no	tttacccttttattctgtagGTA
S15	3' (acceptor)	7	-8.49	yes	agcggcagttggtacgggagGAG
S16	5' (donor)	8	10.36	no	GTGgtaagt
S17	3' (acceptor)	8	7.51	no	gcgattcatttttccttcagCTG
S18	5' (donor)	9	5.19	no	ACTgtatgt

Figure 3.7: The splice site strength table can be accessed by clicking on the 'Splice site strength' button of the main toolbar (see Figure 3.3). For each splice site, its type (donor or acceptor), its strength according to the MaxEntScan method and whether the site is alternative or not are indicated. Furthermore, the sequence (9 and 23 nucleotides for donor and acceptor sites respectively) is shown in upper case for the exonic part and in lower case for the intronic part.

FasterDB 5‘UTR and 3‘UTR tables for human WNK1

	5' untranslated regions	
	UTR 1	UTR 2
Supporting GenBank accession identifers	NM_018979 BC021121 BC071959 BC094862	BC141881
Length	643	701
Number of ATGs in frame	0	0
Total number of ATGs	4	5
Number of GTGs in frame	6	0
Total number of GTGs	15	12
Number of CTGs in frame	5	4
Total number of CTGs	13	9
Number of micro ORFs in frame	0	0
Total number of micro ORFs	2	3
GC percentage	77	50
Number of pyrimidine tracks	6	3

	3' untranslated regions	
	UTR 1	UTR 2
Supporting GenBank accession identifers	BC021121 BC071959 BC094862	NM_018979 BC013629
Length	206	2660
Number of ATGs in frame	0	15
Total number of ATGs	1	44
Number of GTGs in frame	1	15
Total number of GTGs	4	52
Number of CTGs in frame	1	17
Total number of CTGs	4	44
GC percentage	33	39
Number of pyrimidine tracks	1	22

Figure 3.8: This table can be displayed by clicking on the ‘5‘UTR’ and ‘3‘UTR’ buttons of the main toolbar (see Figure 3.3). Several information about each of the 5 untranslated regions are provided. For each UTR, the subset of transcripts having this UTR is displayed as well as the number of ATG, GTG and CTG motifs found either globally in the UTR sequence or in frame. The length and the GC content are also computed for each UTR. Finally, the number of pyrimidine tracks as well as the number of microORFs is displayed.

FasterDB miRNA table for human WNK1

miRNAs				
Identifier	**Name**	**Gene coordinates**	**Chromosomal coordinates**	**Prediction algorithm**
Group 1				
M1	hsa-miR-613	74370 - 74388	936458 - 936476	miranda
M2	hsa-miR-206	74370 - 74388	936458 - 936476	miranda
M3	hsa-miR-1	74370 - 74388	936458 - 936476	miranda
M4	hsa-miR-146b-5p	74475 - 74498	936563 - 936586	miranda
Group 2				
M1	hsa-miR-19b	116213 - 116230	978301 - 978318	miranda
M2	hsa-miR-19a	116213 - 116230	978301 - 978318	miranda
M3	hsa-miR-448	116328 - 116348	978416 - 978436	miranda
M4	hsa-miR-136	116344 - 116367	978432 - 978455	miranda
M5	hsa-miR-23a	116346 - 116366	978434 - 978454	miranda
M6	hsa-miR-23b	116346 - 116366	978434 - 978454	miranda
M7	hsa-miR-153	116439 - 116463	978527 - 978551	miranda
M8	hsa-miR-217	116442 - 116466	978530 - 978554	miranda
M9	hsa-miR-374a	116631 - 116653	978719 - 978741	miranda
M10	hsa-miR-374b	116632 - 116653	978720 - 978741	miranda

Figure 3.9: This table summarizes information about micro RNA binding sites found in the gene UTR sequence and is displayed by clicking on the 'miRNA' button of the main toolbar (see Figure 3.3). For each miRNA, its identifier, gene and chromosomal coordinates, together with the algorithm that predicted this miRNA binding site (among Miranda, pita, targetscan) are indicated. When a polyadenylation site is present, a blue bar is displayed between the miRNA binding sites.

FasterDB protein prediction table for human WNK1

Protein prediction	
Transcript name	**Sequence**
AF061944	sequence
NM_018979	sequence
NM_213655	sequence
NM_001184985	sequence
NM_014823	sequence
BC021121	sequence
BC071959	sequence
BC094862	sequence
AJ296290	sequence
BC141881	sequence
AB002342	sequence
DQ925669	sequence
BC013629	sequence

Figure 3.10: The protein predictions are available upon clicking on the 'Protein prediction' button of the main toolbar (see Figure 3.3). This table contains the protein predictions for the transcripts associated to the gene of interest when these transcripts are known or predicted protein coding transcripts. The fasta sequences of these proteins can be accessed by using the 'sequence' link in the table.

FasterDB conserved exon table for human WNK1

Conserved exons		
Gene identifier mouse	**Position human**	**Position mouse**
4851	1	1
4851	2	2
4851	3	3
4851	4	4
4851	5	7
4851	6	8
4851	7	9
4851	8	10
4851	9	11
4851	10	12
4851	11	13
4851	12	14
4851	13	15
4851	14	16
4851	15	17

Figure 3.11: This table can be displayed by using the 'Conservation' button of the main toolbar (see Figure 3.3). First, the mouse orthologous of the human gene is indicated (or vice-versa). Then, for each human exon, its orthologous mouse exon, if it exists, is identified.

FasterDB sequence tool


```
>Exon1
GGCCTCGGGGAAGGGGGGGCCCGCTCCTCAGGCGCCGAGGCTCCGAGGCT
>Exon2
TTGGTGTTTTATCATTTTAAACCACATTCTGTTTTGGTTTTTGATTTTAG
>Exon3
AACAAACTCCTATCATTGATAACTTGTTTGGTATACTTTGCTTTTTCTAG
>Exon4
CGTCTGAAGCTTCTGCTCGCATTGAGTCTGAATCGTTCTTTTAATTTAAG
>Exon5
AAGTTCAAAGGCCCTGCTTTTATTAATGTATTTCTGTTTATGGTTTTCAG
>Exon6
AACTAATGGTGTTTTATTTTTGTTCCTTTTCTTTTCCCTCTGTTTGGAAG
>Exon7
GTTGTTTTCTTTCAATATACTACTGCTTAATTTACCCTTTTATTCTGTAG
>Exon8
TGATTTGTCTTTCTCTCTCTTTTTTTTGGCGATTCATTTTTCCTTCAG
>Exon9
TTGTTTTCACAGTACTGTGTTTTTCATGTGTGTGTTTGTTTTGTGTTGAG
>Exon10
AATTTCACATATGAATGTATGAATTACTTGTCTTATTCATGTTGATACAG
>Exon11
CACTTACTTTAGGCCTTCTCTAATTTGTTGTGTTCACTTCTTTCCTTCAG
>Exon12
ATGCTGTTATTGATTTGAAATAAAACTGAATCATTGTATTTTATTCTTAG
```

Figure 3.12: Intronic or exonic sequences can be downloaded by clicking on the 'Exon-Intron' button of the main toolbar or by clicking on any exon or intron on the graphical representation of the gene. Alternatively, the 'Sequences' button of the main toolbar allows the retrieval of the flanking genetic sequences of all exons or introns of the displayed gene. Users have to select the length of the requested sequences. This functionality is useful for instance, to retrieve the exon 3' splice sites, by retrieving, for all exons, the 50 nucleotides upstream as shown above.

FasterDB expression table for human WNK1

Expression			
Exon position	Tissues	Cell lines	Cancer cell lines
1	Expression	Expression	Expression
2	Expression	Expression	Expression
3	Expression	Expression	Expression
4	Expression	Expression	Expression
5	Expression	Expression	Expression
6	Expression	Expression	Expression
7	Expression	Expression	Expression
8	Expression	Expression	Expression
9	Expression	Expression	Expression
10	Expression	Expression	Expression
11	Expression	Expression	Expression
12	Expression	Expression	Expression
13	Expression	Expression	Expression
14	Expression	Expression	Expression
15	Expression	Expression	Expression

Experiments	Expressed	Gene expression level	NI	Global SI	Local SI	p-value
Heart	Yes	415	1.67	3.26	3.08	0.0000
Thyroid	Yes	244	-1.10	1.79	1.50	0.0000
Muscle	Yes	262	-1.03	1.91	1.39	0.0004
Testis	Yes	215	-1.11	1.76	1.21	0.0007
Cerebellum	Yes	188	-1.69	1.16	-1.33	0.0000
Pancreas	No	120	-3.33	-1.96	-1.39	0.0040
Kidney	Yes	457	-1.75	1.10	-1.69	0.0000
Prostate	Yes	174	-2.22	-1.30	-2.04	0.0000
Liver	Yes	189	-2.70	-1.56	-2.22	0.0000
Breast	Yes	207	-3.45	-2.04	-2.27	0.0000
Spleen	Yes	204	-4.00	-2.38	-3.23	0.0000

Figure 3.13: The expression table contains links to the expression profile of the exons in tissues, in cell lines and in cancer cell lines (human only). It can be accessed by using the 'Expression' button of the main toolbar (see Figure 3.3). For each group (tissues/cell lines/cancer cell lines), the dedicated link redirects the user to the expression profile page that details the expression profile of the selected exon. For each tissue/cell line, the gene expression level is estimated through several indexes (see section 2.2). The table can be sorted by any column. Finally, the p-value of the student test based on the local SI values is displayed when available. Clicking on the name of the tissues/cell lines links out to Elexir where the entire gene profile can be investigated (see section 3.2).

FasterDB splicing factor tool

Splicing factors

Splicing factor	Exon array	CLIP data	Exon position
CELF1 (CUG-BP)			1 Go
DAZAP1			1 Go
ELAVL1 (HuR)	Yes	Yes	1 Go
ESRP1/2 (ESRP)	Yes		1 Go
FUS		Yes	1 Go
HNRNPA1		Yes	1 Go
HNRNPA2B1		Yes	1 Go
HNRNPAB			1 Go
HNRNPC		Yes	1 Go
HNRNPF/HNRNPH1	Yes	Yes	1 Go
HNRNPL	Yes		1 Go
HNRNPL_2			1 Go
HNRNPM		Yes	1 Go
HNRNPU		Yes	1 Go
MBNL			1 Go
NOVA1			1 Go
PTBP1/2 (PTB)	Yes	Yes	1 Go
QKI			1 Go
RBFOX2 (FOX2)		Yes	1 Go
RBM4	Yes		1 Go
SFRS3 (SRP20)			1 Go
SFRS6 (SRP55)			1 Go
SFRS7 (9G8)			1 Go
SRSF1 (SF2ASF)			1 Go
SRSF2 (SC35)			1 Go
SRSF5 (SRP40)			1 Go
TIA1/TIAL1		Yes	1 Go
TRA2A			1 Go
TRA2B			1 Go
YBX1 (YB1)			1 Go

Figure 3.14: The splicing factor tool, accessible through the 'Splicing factor' button of the main toolbar (see Figure 3.3), allows one to check for a selected exon and splicing factor whether there exist motifs corresponding to the splicing factors around the exon starts/ends, and whether experimental datasets (siRNA, CLIPseq) are available. See for instance above the example for the WNK1 gene. More details about this functionality are provided in section 3.3.

3.2 Elexir tutorial

Elexir is a web application which allows user to choose an experiment within a different set of stored exon array experiments. This interface has been developed in order to allow the end-user to browse and to easily query one or more gene expression between conditions with possible replicates for each condition. Different tested conditions can be chosen for each experiment. A

condition can be defined as a cell line, tissue or treatment. Paired or unpaired analysis can be done depending on samples relationship used for test and control conditions.

The page contains the structure of the queried gene, the same structure as in FasterDB using the same color code (see section 3.1). Below is the intensity report that is centered around a schematic graph of the gene with corresponding probes for each exon. The scheme represents both the exons in light blue and the introns in white (when intronic probes are displayed - see the control panel). Above and below the plots are bars, whose heights represent the normalized probe intensities (in log2 scale) respectively for the test and the control conditions. Above the scheme of the gene, the colors reflect the ratio between probe intensity in test and control conditions. Green or red colors mean that the probe intensity in the test condition is lower or higher, respectively, compared to the control condition. Black bars point to no difference between both conditions. Below the scheme of the gene, the colors indicate the quality of the probe. Dark grey probes are considered as not expressed (DABG p-value > 0.05), strikethrough bars represent probes that align multiple times on the genome (and that are therefore non specific), red enclosed bars indicate a too low GC content for the associated probes. This color code is also summarized in Figure 3.15.

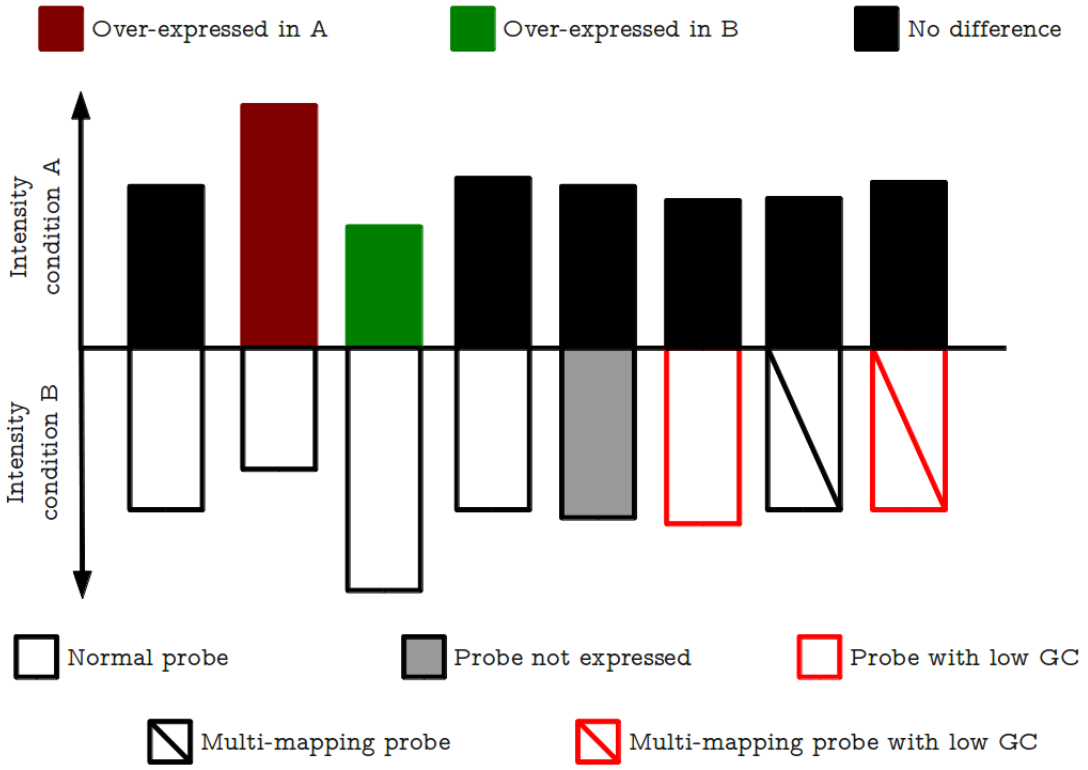

Figure 3.15: The Elexir color code

Information for each probe can be easily accessed either by a simple click on the bar or by the underlying descriptive table. In addition, a control panel is located on top to adapt the parameters and thus ease the vizualisation process.

Elexir search engine

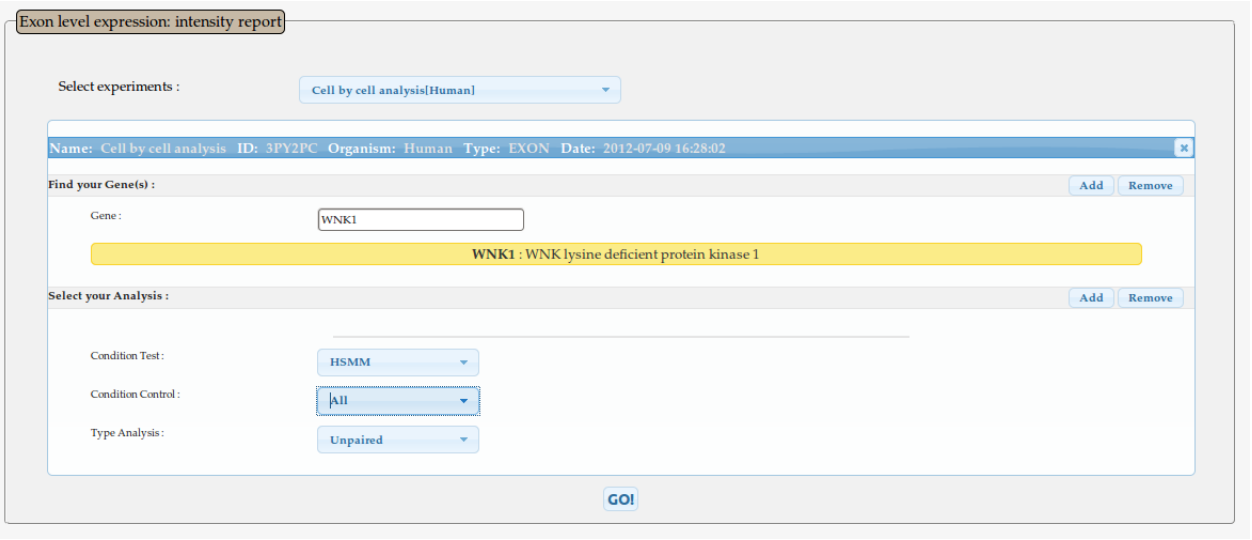

Figure 3.16: The home page of Elexir is the search page that can be used to access the gene page when all parameters have been inputed (see for instance the search corresponding to the human WNK1 gene above). On this search page, a user first needs to select the dataset to investigate (*e.g.*, human tissues, mouse cell lines), then the gene to display and the parameters of the analysis (which conditions are used as test, which conditions are used as control). Notice that several genes can be inputed one by one and the gene page will then contain the results for each gene.

Elexir gene page for human WNK1

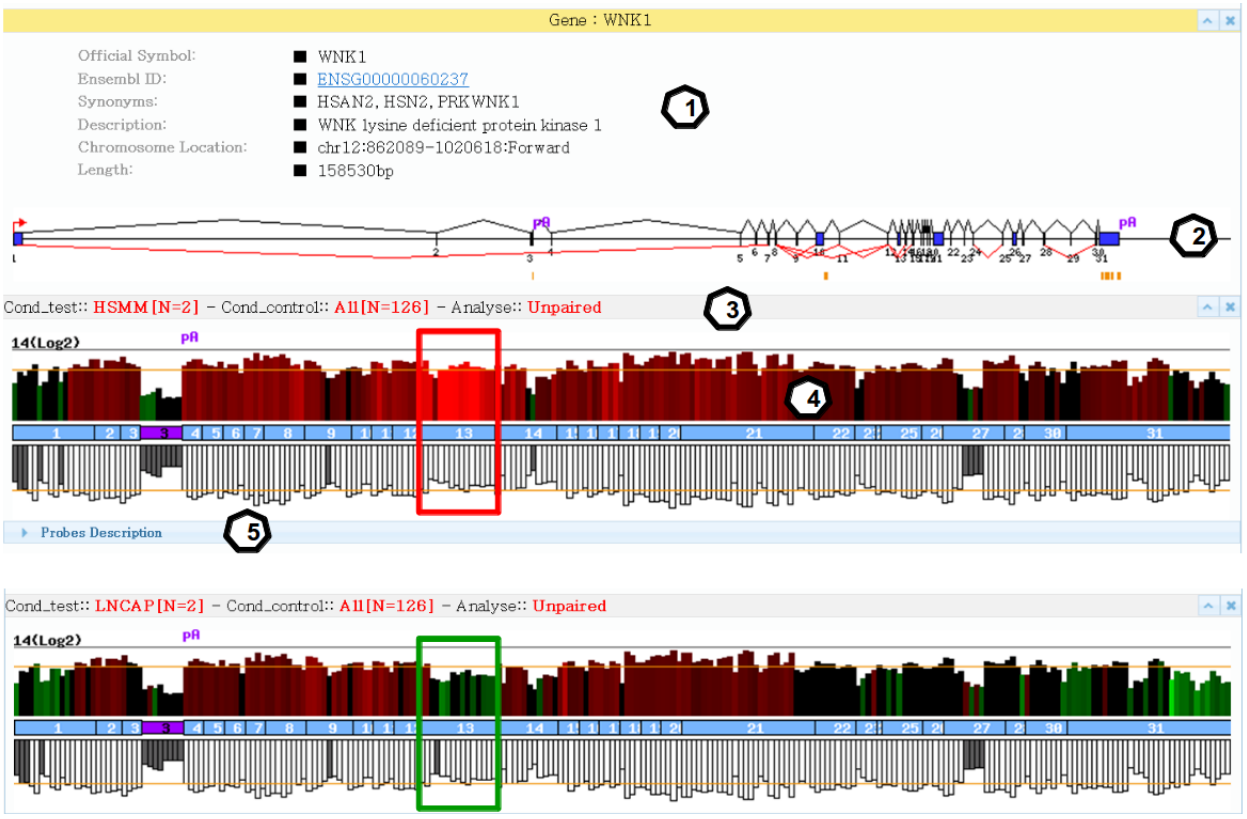

Figure 3.17: The WNK1 gene page in Elexir. First, information about the gene and its structure is depicted (same image as in FasterDB) (heptagons 1 and 2). Then, details of the analysis are summarized (heptagon 3) and the results are displayed per probe along the gene (heptagon 4). The height of each probe corresponds to its intensity level. Red probes indicate inclusion in test condition while green probes indicate its exclusion (inclusion in control condition). Black probes point no difference between both conditions. In addition, a description of the probes can be found below the results (heptagon 5). For an example, see Figure 3.17 that exhibits the WNK1 expression profile in two cell lines (HSMM and LNCAP) versus all the other cell lines.

3.3 Splicing factor tutorial

The motifs identified around splicing sites are displayed on a special page for a single exon. On this page, the exon is represented by a blue box and the flanking introns by solid black lines. Light yellow boxes indicate the positions where the motif associated to the splicing factor has been identified. Notice that the orthologous mouse exon is also displayed when available. Sequences of exon or introns (upstream and downstream the exon) can be displayed simply by clicking on exon or intron in the graphical presentation.

The siRNA datasets are visualized using Elexir (see section 3.2).

The CLIPseq datasets are represented by histograms covering the whole gene (whose FasterDB structure is also displayed) that indicate the number of reads that have been mapped to each chromosomal location. When several experiments exist, several histograms are displayed,

one for each experiment.

Splicing factor: motif page

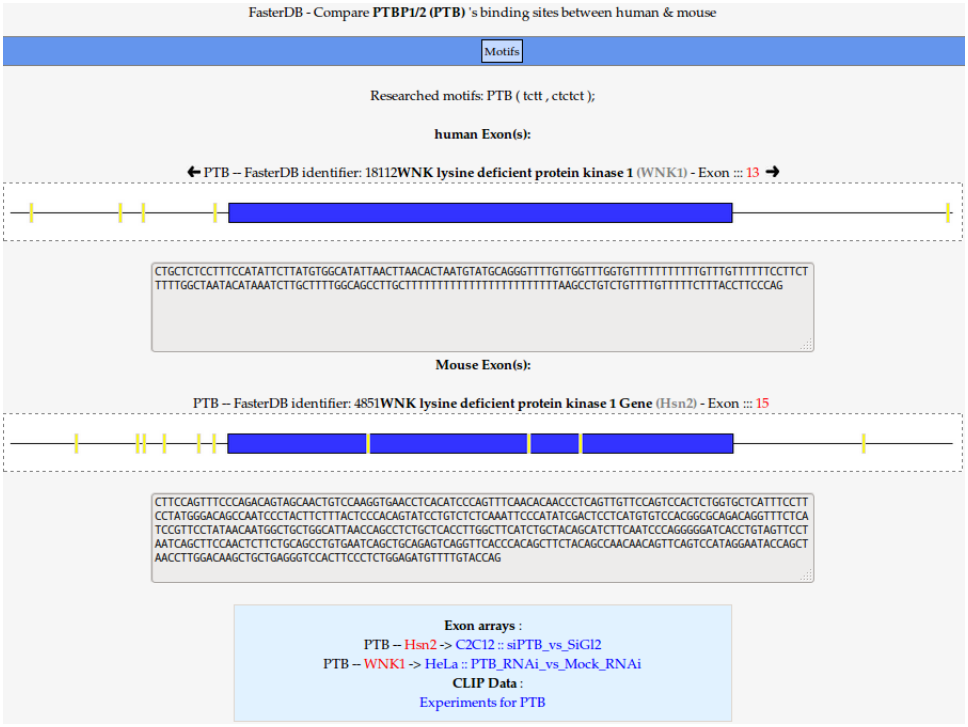

Figure 3.18: The motif page displays for a selected exon and motif, the motifs found in and around the exon (200 nucleotides in the bordering introns). It can be accessed from the main toolbar of the gene page (see Figure 3.3) through the 'Splicing factor' button). The example above shows the PTB motifs for exon 13 of the WNK1 gene and for its orthologous exon (exon 15 of the mouse WNK gene).

Splicing factor: CLIPseq datasets

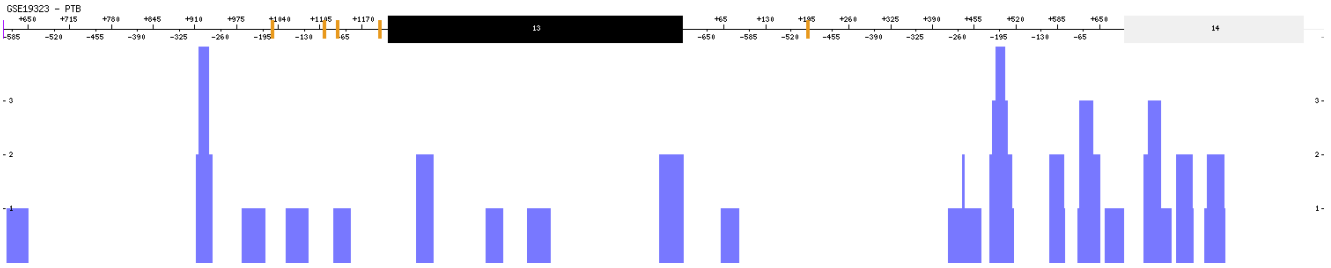

Figure 3.19: This page can be accessed from the links at the bottom of the motif page (see Figure 3.18). Reads from CLIPseq experiment corresponding to binding sites of a selected splicing factor are represented as a raw signal below exons schematic presentation. When available, multiple CLIPSeq profiles are displayed one on top of another. The above example illustrates the data available for the exon 13 of WNK1 and the PTB splicing factor.