Problem

We have a huge amount of reads from RNA transcripts + potentially some rubbish sequences

We decided to classify given reads into the one of following three groups:

a) Reads of good quality not interrupted by introns

on the genome level - Exon reads

a) Reads of good quality interrupted by introns - Splicing reads

b) Badly sequenced reads + No hits sequences - Bad reads

1. Exon reads can give us information on exons location and expression level of genes

2. Splicing reads can provide information on splice sites and introns positions

1. and 2. could be used to improve gene prediction accuracy

Realization in TRANSOMICS pipeline

Using our ***SCAN2*** program we mapped (with parameters for very fast and uninterrupted mapping) each read to contigs (or chromosomes) and compute three values: number of good hits, number of bad hits and quality of the best mapping (alignment), which were used for initial read sorting to

Group 1: EXON READS

Reads that have a number of good hits to some contig higher in a certain number times the number of hits for any other contigs

Group 2: BAD READS

Reads that have good mappings in several contigs or many bad mappings in various contigs

Group 3: No uninterrupted hits (No significant hits +SPLICING READS)

The last group ~ 10% of the total read number and can be studied more thoroughly

By our ***EST_MAP*** program we mapped reads of Group 3 to chromosomes and then selected SPLICING READS

Example of mapping by EST_MAP

mapped perfectly to a chromosome as 2 fragments, with an intron between them

```
[DR] Sequence:       4(      1)  L:       36
Blocks of alignment: 2
    1 E:  4679323     26 [ag GT] P:   4679323         1  L:      26, G: 100.00, W:    520, S:7.99124
    2 E:  4679397     10 [AG ga] P:   4679397        27 L:      10, G: 100.00, W:    200, S:4.69493
```

Data used

Organism	***Genome version***	***Genome size***	***RNAseq data size (example)***
H. sapiens*	GRCh37 (hg19)	~3 Gb	~24 Gb (GM12878_2x75)
C. elegans	WS200	~98 Mb	~6.8 Gb (SRX001873)
D. melanogaster	version 5 (dmel_r5.20_FB2009_07)	~165 Mb	~13.1 Gb (cell line Kc167)

* repeats masking:

for Human genome, repeats found by RepeatMasker were masked

(simple repeats and low complexity regions were not masked since they can be parts of protein coding regions);

for Drosophila and C.elegans genomes repeats were not masked.

Reads data

C. elegans

experiment : polyA+ RNAseq random fragment library (Illumina)
lab : UWGS-RW

1. SRX004863 & SRX004864: early embryo	(7.5 Gb + 11 Gb)
2. SRX004865 & SRX004866: late embryo	(7.6 Gb + 7.3 Gb)
3. SRX004867: mid-L1	(16 Gb)
4. SRX001872: mid-L2	(13 Gb)
5. SRX001875: mid L3	(7.7 Gb)
6. SRX001874: mid L4	(5.1 Gb)
7. SRX001873: young adult (pre-gravid)	(6.8 Gb)

+ combined set of reads from all stages

Drosophila

lab : Celniker modENCODE supergroup

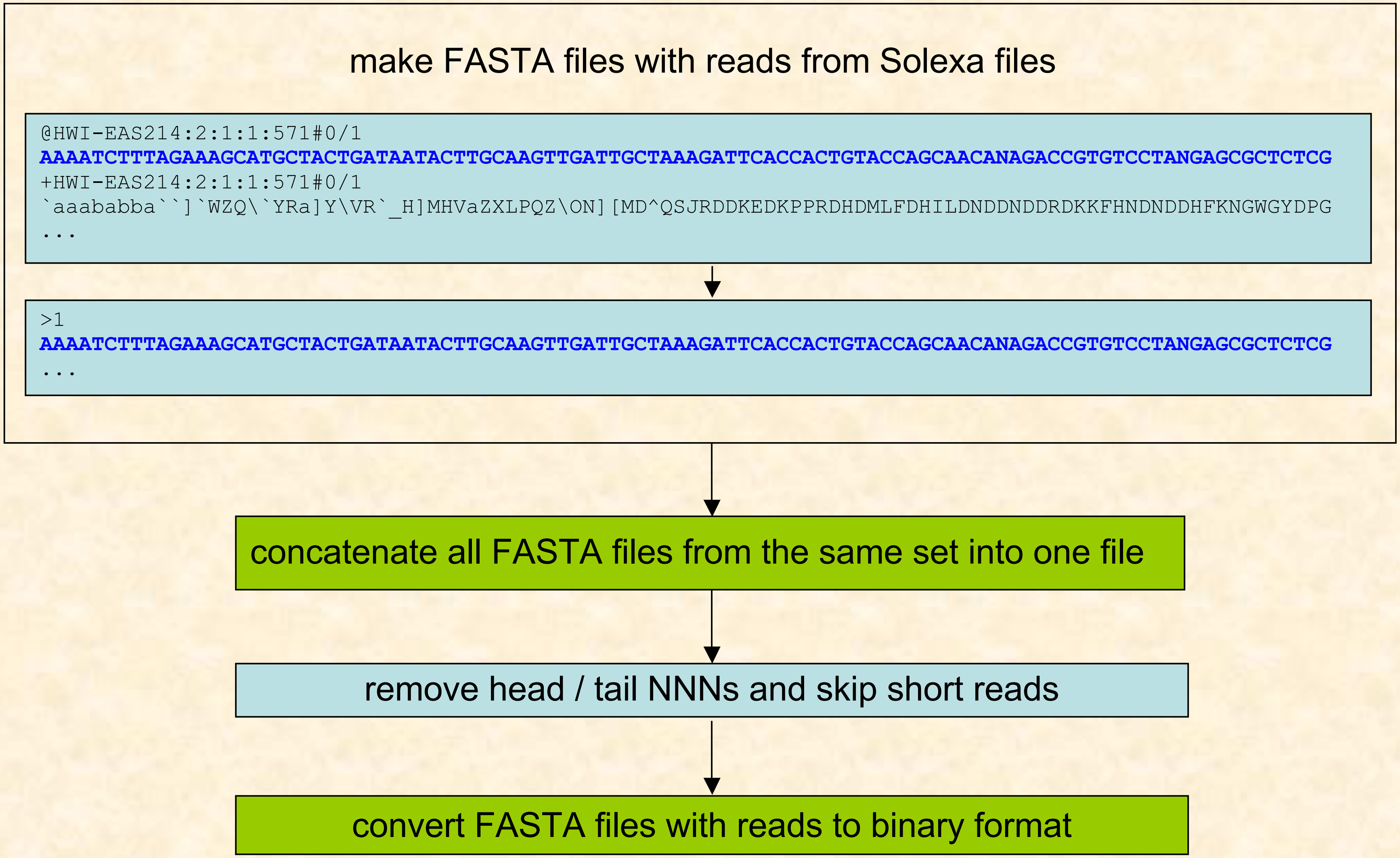
experiment : cell line S2-DRSC	this set was split into Untreated and treated	(25.8 Gb)
experiment : cell line CME_W1_CI		(7.4 Gb)
experiment : cell line Kc167		(13.1 Gb)
experiment : cell line ML-DmBG3-c2		(6.9 Gb)

+ combined set of reads from all cell lines

Human

experiment : Solexa Human polyA+ total RNA, paired reads, GM12878 (~24 Gb)
lab : Wold lab, Caltech

Steps of TRANSOMICS pipeline: Preparing reads data



make FASTA files with reads from Solexa files

```
@HWI-EAS214:2:1:1:571#0/1
AAAATCTTTAGAAAGCATGCTACTGATAATACTTGCAAGTTGATTGCTAAAGATTCACCACTGTACCAGCAACANAGACCGTGTCCTANGAGCGCTCTCG
+HWI-EAS214:2:1:1:571#0/1
`aaababba``]`WZQ\`YRa]Y\VR`_H]MHVaZXLPQZ\ON][MD^QSJRDDKEDKPPRDHDMLFDHILDNDDNDDRDKKFHNDNDDHFKNGWGYDPG
...
```

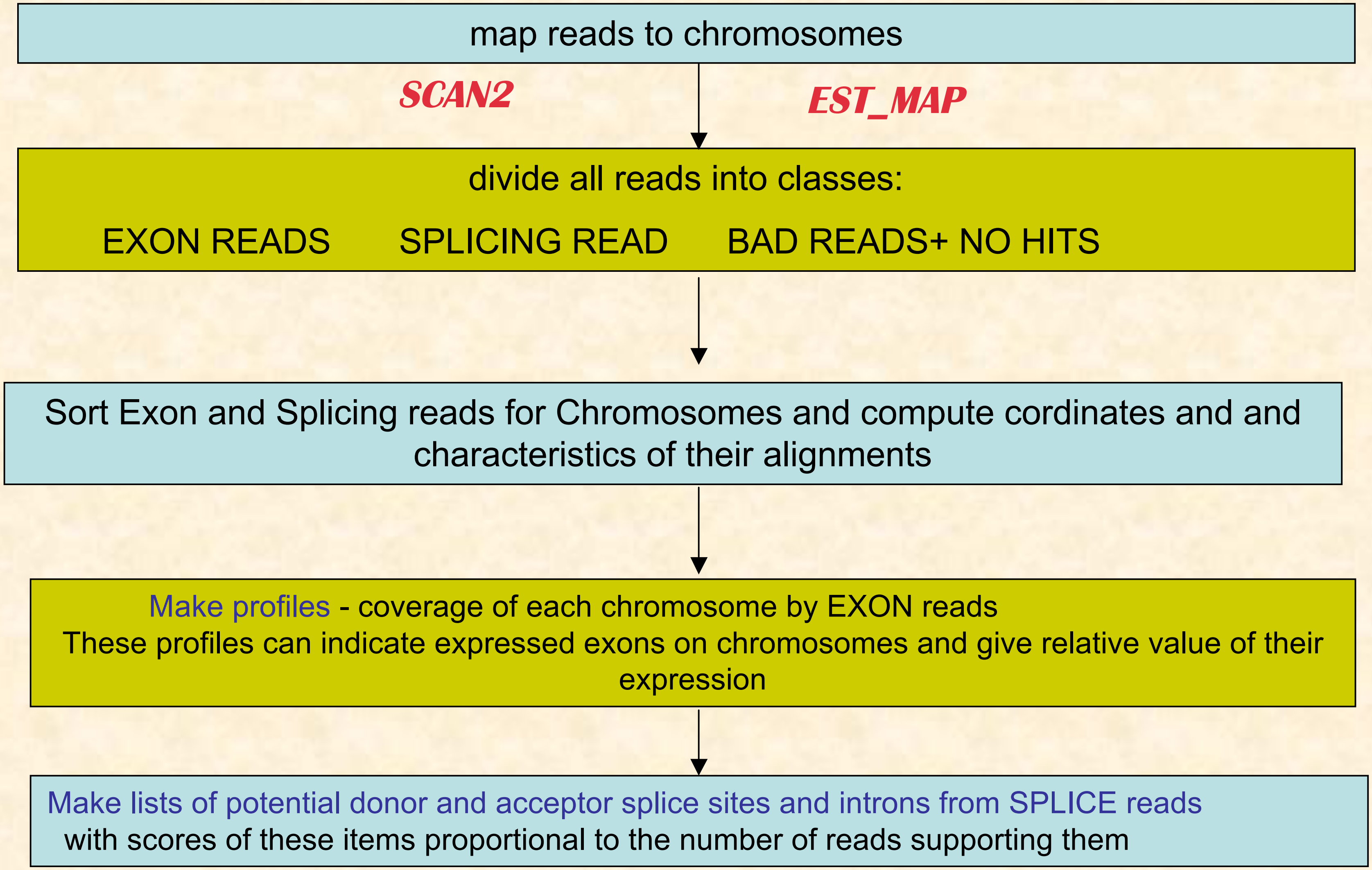
```
>1
AAAATCTTTAGAAAGCATGCTACTGATAATACTTGCAAGTTGATTGCTAAAGATTCACCACTGTACCAGCAACANAGACCGTGTCCTANGAGCGCTCTCG
...
```

concatenate all FASTA files from the same set into one file

remove head / tail NNNs and skip short reads

convert FASTA files with reads to binary format

TRANSOMICS pipeline flow



map reads to chromosomes

SCAN2

EST_MAP

divide all reads into classes:

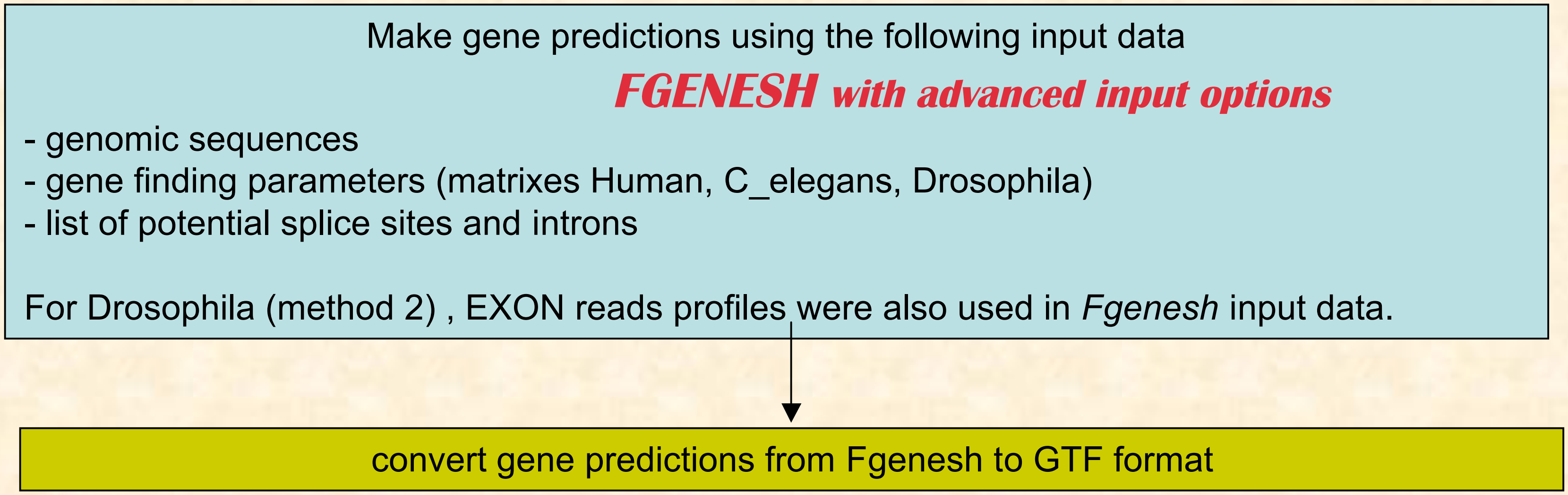
EXON READS SPLICING READ BAD READS+ NO HITS

Sort Exon and Splicing reads for Chromosomes and compute cordinates and and characteristics of their alignments

Make profiles - coverage of each chromosome by EXON reads
These profiles can indicate expressed exons on chromosomes and give relative value of their expression

Make lists of potential donor and acceptor splice sites and introns from SPLICE reads
with scores of these items proportional to the number of reads supporting them

Transomics pipeline flow (continued)



Make gene predictions using the following input data

FGENESH with advanced input options

- genomic sequences
- gene finding parameters (matrixes Human, C_elegans, Drosophila)
- list of potential splice sites and introns

For Drosophila (method 2) , EXON reads profiles were also used in *Fgenesh* input data.

convert gene predictions from Fgenesh to GTF format

Gene predictions have been done for each experiment and with combined set of reads from all cell lines

Calculating expression levels

Profiles (coverage of each chromosome by EXON reads) were used for calculating expression data

For each gene (exon), RPKM was calculated as follows:

RPKM = 1000000000 * (profile_sum_locus / profiles_sum_all) / length (in bp),

where
profile_sum_locus - sum of profile coverage of gene (exon) by mapped reads;
profiles_sum_all - sum of profile coverage of chromosome by mapped reads, and sum over all chromosomes for a given organism;
length - length of gene (exon) in base pairs (bp).

In our modified RPKM formulae we worked with profiles rather than reads themselves, and used the multiplier (profile_sum_locus / profiles_sum_all)

instead of the multiplier
(number of reads mapped to gene (exon) / overall number of mapped reads).

Results reported

For each experiment, only genes with RPKM > 0.01 were reported.

For structure predictions using reads from all sets/stages for a given organism all genes were reported.

*C.elegans, SRX001873: young adult **example:***

~6.8 Gb (all Solexa files in fastq format)
60 903 898 reads

after removing head / tail NNNs and skipping short reads (and converting to FASTA format):

~2.7 Gb (FASTA files)
59 547 560 reads

conversion to binary format
~4.8 Gb (binary files)

mapping reads

59547560 - 100% - all reads
41150605 - 69.1% - EXON reads mapped as uniterrapted fragment
1002486 - 1.7% - SPLICE READS mapped to chromosomes as 2 fragments
(alignment with potential internal intron)
17394469 - 29.2% - BAD reads (mapped not so well or mapped to multiple chromosomes)

Times of processing data

TIME FOR :chr2 vs. reads of SRX001873

C.e. Chr2 ~ 15 Mb SRX001873: young adult (pre-gravid) ~ 6.8 Gb

Pipeline steps:

Data preparation step: make FASTA files with reads from Solexa files, remove head / tail NNNs and skip short reads, convert to binary format:	20 min
Map reads to the chromosome:	2 h 30 min
Sort reads by chromosomes (perl script):	1 h 30 min
Make EST_MAP alignments for splice sites discovery (to all chromosomes):	1 h
Analysing alignments, list of potential splice sites and introns:	10 sec
Fgenesh gene predictions:	6 min
Make profile (coverage of chromosome by reads):	8 min
Calculating expression data (perl scripts):	30 sec
Conversion to GTF format:	2 sec

Effect on gene predictions

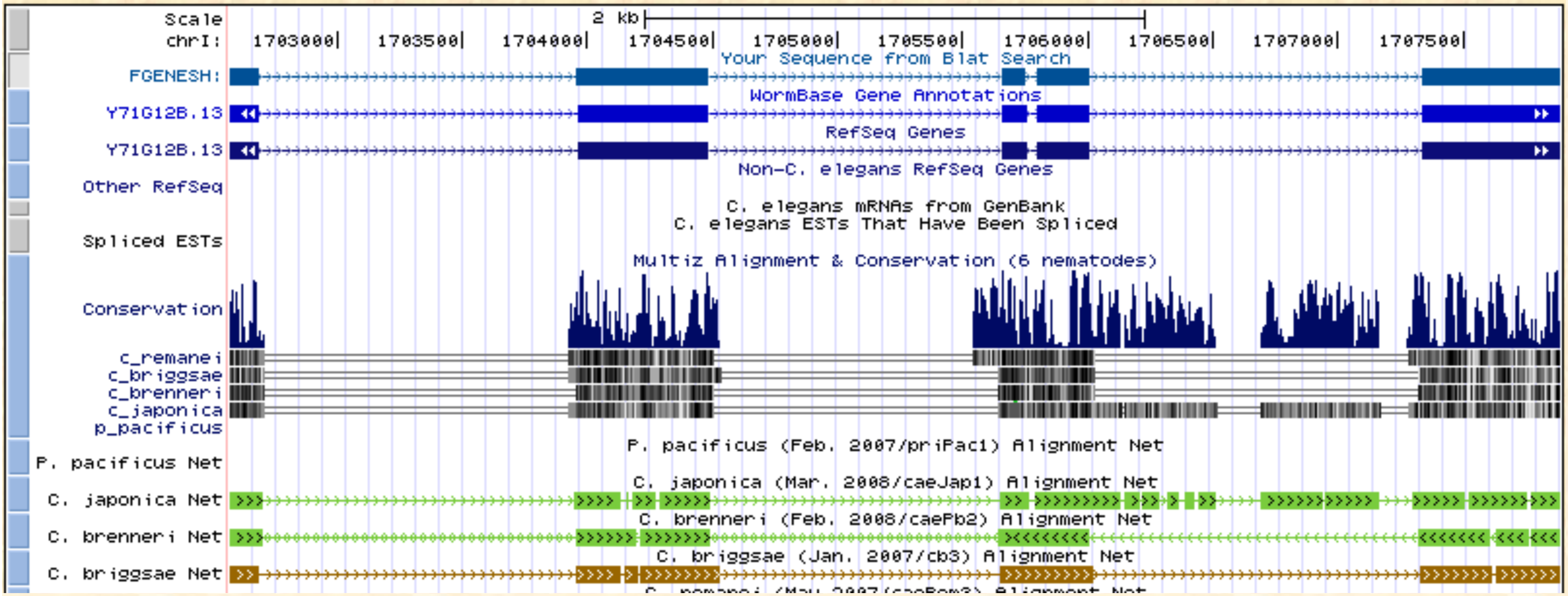
TEST of Fgenesh gene prediction accuracy: for 10 Ngasp sequences of C.elegans

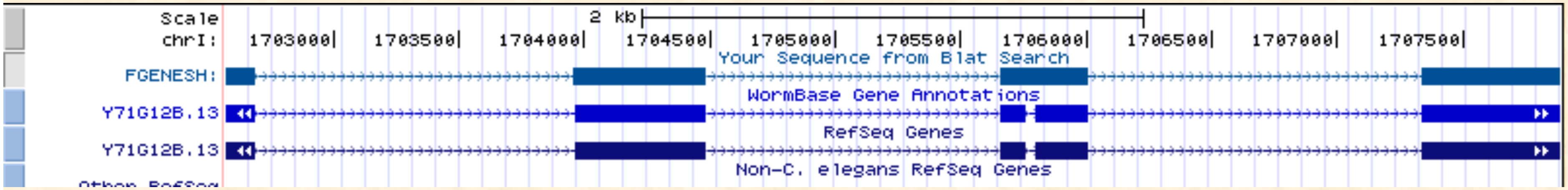
FGENESH +	SN nuc %	SP nuc %	Exon exact Sn	Exon exact Sp	Exon ovr Sn	Exon ovr Sp	Exactly predicted genes
Ab initio	95.2	86.0	79.1	72.7	94.0	84.4	794
Ngasp EST	94.8	87.3	81.5	76.0	94.4	86.0	967
With bigger EST set	94.9	87.6	82.4	77.0	94.5	86.4	1035
With EST and ALT SPLICING predictions	95.3	87.5	83.6	76.5	94.9	86.5	1130
WITH SPLICE READS	95.9	86.7	82.2	74.5	95.4	84.9	944

High accuracy of ab initio predictions on nucleotide level on these data leaves a small room to increase it.

Using reads from different experiments for Alternative splicing variants discovery

C.e., chr1, EXPERIMENT 7 transcript RPKM "7.86"


Scale
chrI:
2 kb
1703000|
1703500|
1704000|
1704500|
1705000|
1705500|
1706000|
1706500|
1707000|
1707500|
Your Sequence from Blat Search
FGENESH:
WormBase Gene Annotations
Y71G12B.13
RefSeq Genes
Y71G12B.13
Non-C. elegans RefSeq Genes
Other RefSeq
C. elegans mRNAs from GenBank
C. elegans ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (6 nematodes)
Conservation
c_remanei
c_briggsae
c_brenneri
c_japonica
p_pacificus
P. pacificus (Feb. 2007/priPac1) Alignment Net
P. pacificus Net
C. japonica (Mar. 2008/caeJap1) Alignment Net
C. japonica Net
C. brenneri (Feb. 2008/caePb2) Alignment Net
C. brenneri Net
C. briggsae (Jan. 2007/cb3) Alignment Net
C. briggsae Net



C.e., chr1, EXPERIMENT 5 transcript RPKM "7.10"

Alternative splicing of Drosophila copia-specific 2.1-kb mRNA

cell line Kc167 transcript RPKM "9077.06”

```
  G  Str   Feature   Start         End     Score            ORF            Len     rpkm
 337 +    1 CDSf    3074367 -   3074691   24.41   3074367 -   3074690    324    11151.92
 337 +    2 CDSl    3077641 -   3077747    5.29   3077643 -   3077747    105     2774.91
 337 +      PolA    3078701                1.25
>FGENESH: 337   2 exon (s) 3074367  - 3077747   143 aa, chain +
MDNCGFVLDSGASDHLINDESLYTDSVEVVPPLKIAVAKQGEFIYATKRGIVRLRNDHEI
TLEDVLFCKEAAGNLMSVKRLQEAGMSIEFDKSGVTISKNGLMVVKNSENQLADIFTKPL
PAARFVELRDKLGLLQDDQSNAE
```

cell line CME_W1_CI transcript RPKM "11390.86”

```
  G  Str   Feature   Start         End     Score            ORF            Len     rpkm
 312 +    1 CDSf    3073518 -   3073845   19.90   3073518 -   3073844    327    1845.05
 312 +    2 CDSi    3073909 -   3074691    5.17   3073911 -   3074690    780   15851.23
 312 +    3 CDSl    3077641 -   3077747    5.49   3077643 -   3077747    105    8012.87
 312 +      PolA    3078701                1.25
>FGENESH: 312   3 exon (s) 3073518  - 3077747   405 aa, chain +
MDKAKRNIKPFDGEKYAIWKFRIRALLAEQDVLKVVDGLMPNEVDDSWKKAERCAKSTII
EYLSDSFLNFATSDITARQILENLDAVYERKSLASQLALRKRLLSLKLSTGAKIEEMDKI
SHLLITLPSCYDGIITAIETLSEENLTLAFVKNRLLDQEIKIKNDHNDTSKKVMNAIVHN
NNNTYKNNLFKNRVTKPKKIFKGNSKYKVKCHHCGREGHIKKDCFHYKRILNNKNKENEK
QVQTATSHGIAFMVKEVNNTSVMDNCGFVLDSGASDHLINDESLYTDSVEVVPPLKIAVA
KQGEFIYATKRGIVRLRNDHEITLEDVLFCKEAAGNLMSVKRLQEAGMSIEFDKSGVTIS
KNGLMVVKNSENQLADIFTKPLPAARFVELRDKLGLLQDDQSNAE
```

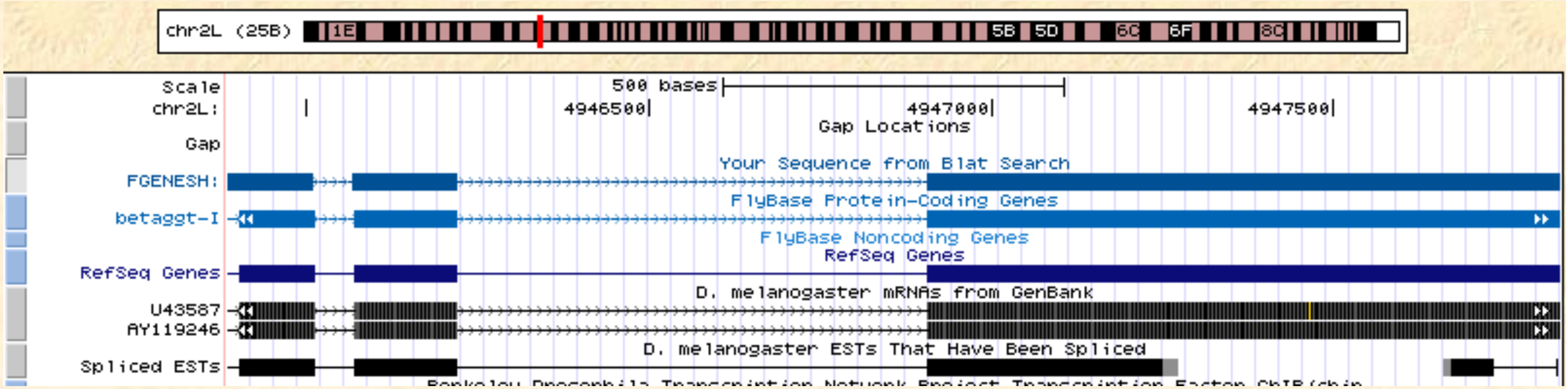
cell line Kc167 transcript RPKM "5403.70"

```
G  Str   Feature   Start         End     Score            ORF            Len     rpkm
395 +    1 CDSf    3650274 -   3650601   11.29   3650274 -   3650600    327     959.76
395 +    2 CDSi    3651046 -   3651447   14.29   3651048 -   3651446    399    9729.30
395 +    3 CDSl    3654397 -   3654503    1.16   3654399 -   3654503    105    2774.91
395 +      PolA    3655193                1.25
>FGENESH: 395   3 exon (s) 3650274  - 3654503   278 aa, chain MDKAKRNIKPFDGEKYAIWKFRIRALLAEQDVLKVVDGLMPNEVDDSWKKAERCAKSTII
EYLSDSFLNFATSDITARQILENLDAVYERKSLASQLALRKRLLSLKLSKNEKQVQTATT
HGIAFMVKEVNNTSVMDNCGFVLDSGASDHLINDESLYTDSVEVVPPLKIAVAKQGEFIY
ATKRGIVRLRNDHEITLEDVLFCKEAAGNLMSVKRLQEAGMSIEFDKSGVTISKNGLMVV
KNSENQLADIFTKPLPAARFVELRDKLGLLQDDQSNAE
```

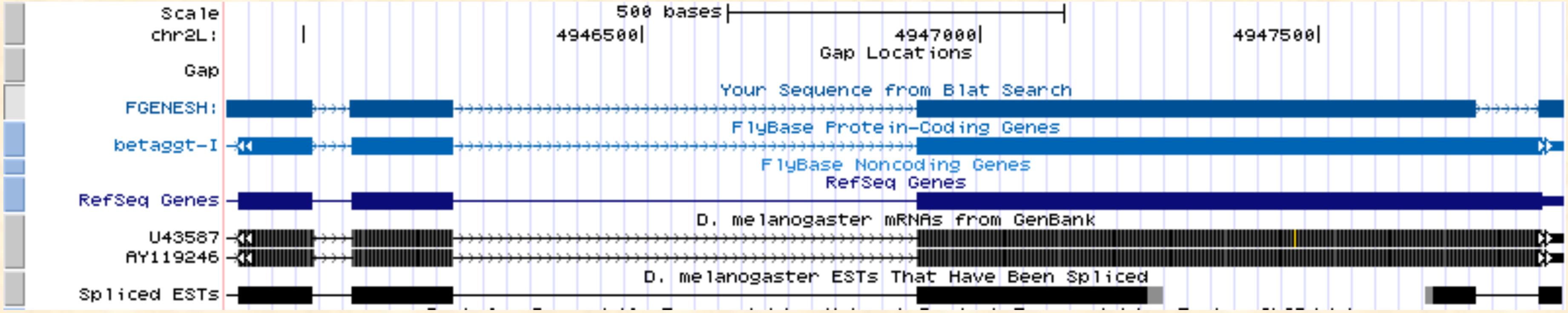
cell line CME_W1_CI transcript RPKM "21308.42”

```
 G  Str   Feature   Start         End     Score            ORF            Len     rpkm
364 +    1 CDSf    3651123 -   3651447   21.51   3651123 -   3651446    324    25685.72
364 +    2 CDSl    3654397 -   3654503    1.81   3654399 -   3654503    105     8012.87
364 +      PolA    3655193                1.25
>FGENESH: 364   2 exon (s) 3651123  - 3654503   143 aa, chain +
MDNCGFVLDSGASDHLINDESLYTDSVEVVPPLKIAVAKQGEFIYATKRGIVRLRNDHEI
TLEDVLFCKEAAGNLMSVKRLQEAGMSIEFDKSGVTISKNGLMVVKNSENQLADIFTKPL
PAARFVELRDKLGLLQDDQSNAE
```

Alternative splicing in Drosophila genes

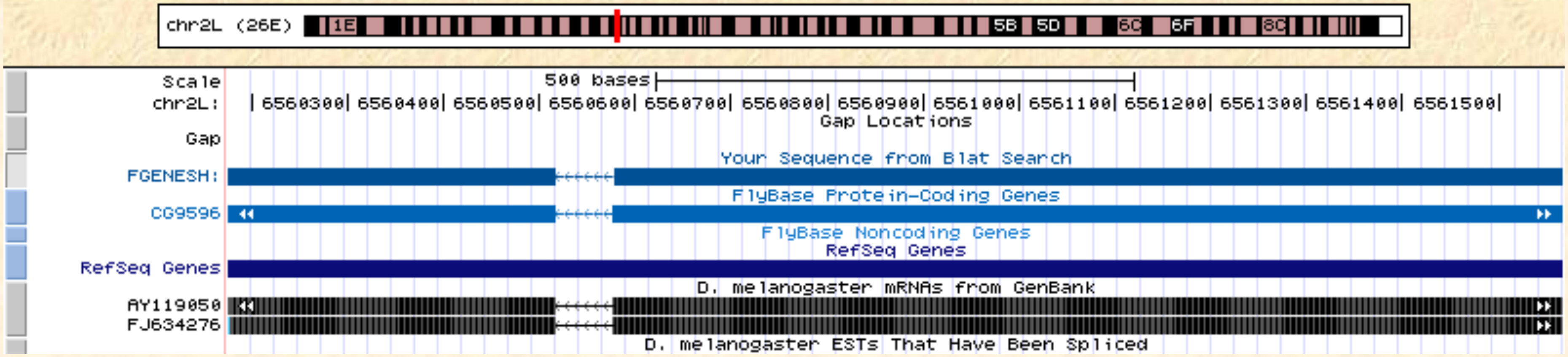


cell line CME_W1_CI transcript RPKM "12.01"

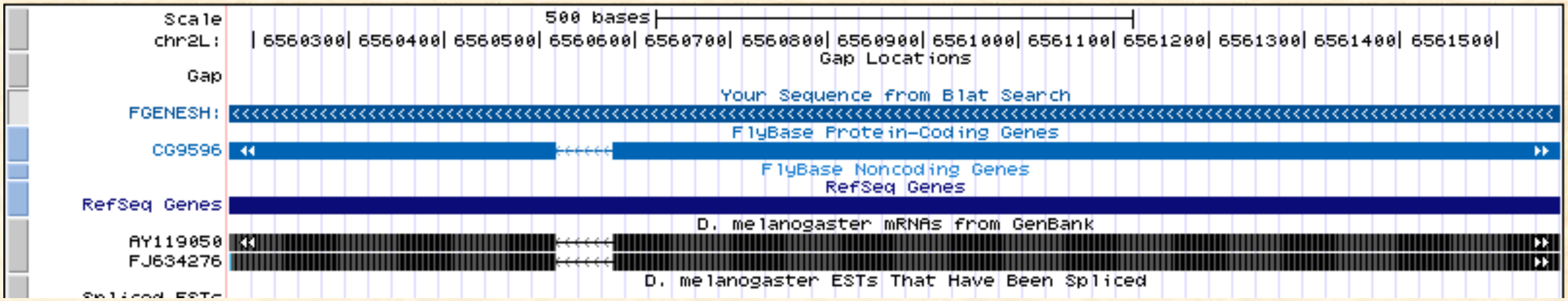


cell line Kc167 transcript RPKM "11.88"

Alternative splicing in Drosophila genes



cell line Kc167 transcript RKPM 5.95



cell line CME_W1_CI transcript RKPM 8.7

Participants

Victor Solovyev Igor Seledtsov

Peter Kosarev Vladimir Molodtsov

Department of Computer Science, Royal Holloway, University of London, UK;

Softberry Inc., USA

2 quad core processor computers

Note: it is a first version of TRANSOMICS pipeline with methods developed or adjust to treat read data without availability of proper learning data.

Further progress certainly can be done having available training sets data (to experiment with methods), accounting paired reads, quality and other reads information.