Supplemental Material for "Pairagon: A Highly Accurate, HMM-based cDNA-to-Genome Aligner"

David V. Lu, et al.

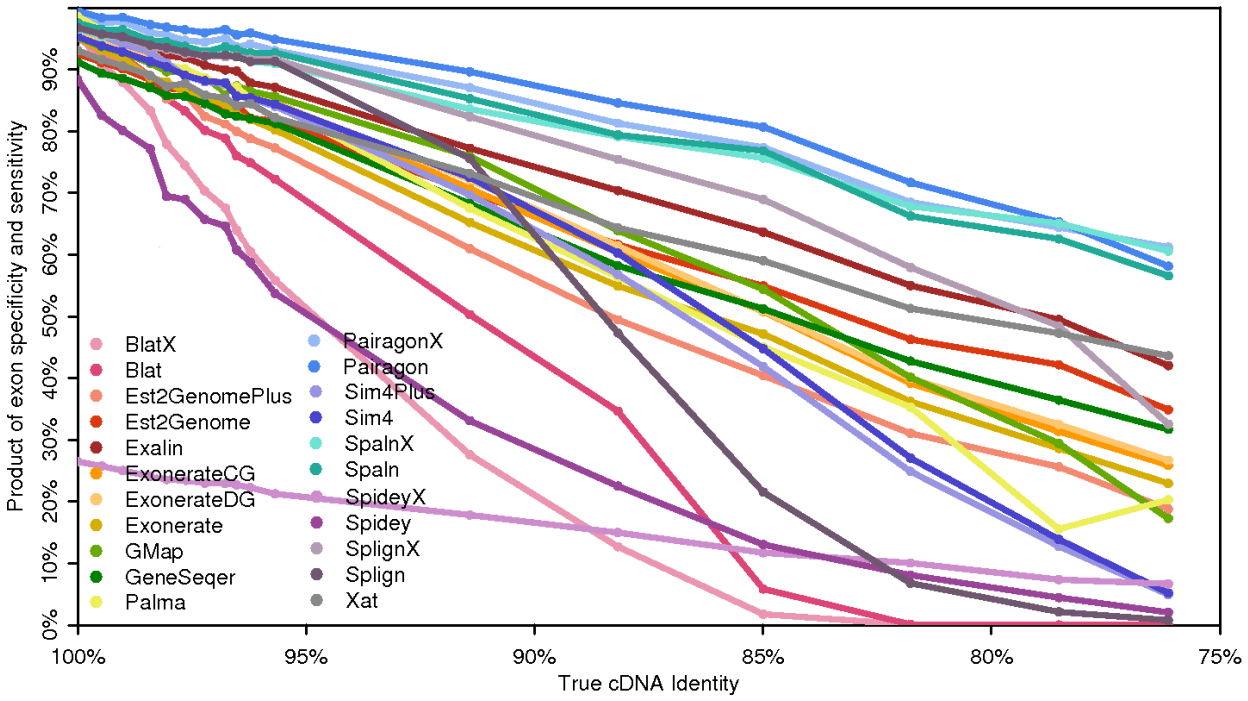
April 13, 2009

Note: All alignments for these experiments are available online at
`http://mblab.wustl.edu/software/pairagon`

1 Experiment 1 Results for All Aligners

D	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
Identity	1.000	0.995	0.990	0.984	0.981	0.976	0.972	0.968	0.965	0.962
Blat	97.5	94.2	92.6	89.1	85.5	83.4	80.4	79.0	76.2	75.0
BlatX	96.7	91.3	88.3	83.7	78.3	74.6	70.5	67.7	64.0	60.6
Est2Genome	93.5	92.1	91.1	89.2	87.9	87.5	85.9	84.8	85.0	82.8
Est2Genome+	95.7	93.1	91.9	88.8	85.6	86.4	82.7	81.4	80.3	79.0
Exalin	97.5	96.1	95.4	93.4	92.5	92.3	91.0	90.3	90.1	88.0
Exonerate	95.9	93.9	92.7	89.4	88.6	86.9	86.1	85.0	83.5	82.3
ExonerateCG	97.2	95.4	94.7	92.3	90.0	89.9	88.7	87.9	86.8	86.0
ExonerateDG	96.9	95.2	94.2	92.1	90.3	90.3	88.6	88.0	86.8	86.2
GMap	95.9	94.4	93.5	91.7	90.3	90.3	89.3	87.1	88.0	87.1
GeneSeqer	92.5	90.5	89.8	87.9	86.0	86.2	85.2	83.5	83.0	82.6
Pairagon	99.6	98.4	98.5	97.3	96.9	96.5	95.9	96.5	95.6	95.9
PairagonX	98.8	97.7	97.9	96.2	95.9	95.1	94.8	95.4	94.1	94.5
Palma	98.7	96.2	95.2	91.3	89.2	87.5	86.1	84.2	83.7	81.7
Sim4	96.0	94.5	93.5	91.8	90.9	89.4	88.6	88.2	86.1	85.7
Sim4+	97.4	95.6	94.3	92.8	90.9	88.9	87.8	87.3	85.6	84.8
Spaln	97.8	96.8	96.7	94.6	94.6	93.6	92.5	93.6	93.1	92.5
SpalnX	97.0	95.7	95.8	94.1	93.7	93.1	92.4	92.7	92.5	91.4
Spidey	89.2	83.2	80.7	77.6	69.9	69.3	66.1	64.9	61.1	58.9
SpideyX	29.7	28.9	28.1	27.0	26.4	26.2	25.8	25.7	25.4	24.8
Splign	97.2	96.1	95.7	94.3	93.9	93.2	92.6	92.6	92.5	91.7
SplignX	97.1	96.0	95.7	94.0	93.7	92.9	92.7	92.4	92.6	91.9
Xat	94.0	92.4	91.5	89.9	88.2	88.5	86.8	86.2	85.1	85.3

D	0.00	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Identity	1.000	0.957	0.914	0.882	0.850	0.818	0.785	0.761
Blat	97.5	72.3	50.2	34.8	8.4	0.2	0.0	0.0
BlatX	96.7	55.9	27.4	12.7	2.4	0.1	0.0	0.0
Est2Genome	93.5	82.4	70.7	62.1	55.3	46.4	42.5	35.1
Est2Genome+	95.7	77.5	60.9	49.5	40.4	31.0	25.7	19.0
Exalin	97.5	87.4	77.2	70.3	63.6	54.8	49.4	42.1
Exonerate	95.9	81.0	65.8	55.7	48.1	37.5	29.9	24.4
ExonerateCG	97.2	84.9	70.8	60.6	51.0	39.6	32.3	27.0
ExonerateDG	96.9	84.6	71.8	60.9	50.5	39.8	32.6	27.0
GMap	95.9	86.3	76.4	64.8	55.9	42.1	32.2	20.1
GeneSeqer	92.5	81.7	67.7	57.4	50.1	41.7	35.4	30.5
Pairagon	99.6	95.0	89.7	84.8	81.0	72.1	65.7	58.6
PairagonX	98.8	93.5	87.7	82.4	78.7	70.4	66.4	63.4
Palma	98.7	80.5	60.2	46.9	35.0	25.7	10.0	12.0
Sim4	96.0	84.7	72.6	60.1	44.2	25.9	12.7	4.5
Sim4+	97.4	83.6	68.8	55.7	40.5	23.5	11.5	4.2
Spaln	97.8	92.3	84.4	78.2	75.3	64.1	60.2	54.9
SpalnX	97.0	91.0	83.2	78.4	74.9	66.6	63.9	59.3
Spidey	89.2	53.9	33.0	22.3	12.8	7.8	4.2	1.8
SpideyX	29.7	23.8	19.8	16.7	12.9	11.0	8.1	7.3
Splign	97.2	91.7	78.2	55.4	30.6	11.5	4.0	1.5
SplignX	97.1	91.9	82.8	76.8	71.0	60.5	52.4	38.8
Xat	94.0	83.2	73.9	65.1	59.9	51.7	47.8	43.8



2 Experiment 2 Results for All Aligners

D	0	0.1	0.2	0.3	0.4	0.5	0.6
Identity	0.998	0.946	0.901	0.859	0.829	0.789	0.761
Blat	97.0	68.0	45.9	30.5	10.3	0.4	0.0
BlatX	95.1	46.9	20.8	8.1	1.8	0.1	0.0
Est2Genome	94.4	82.0	71.9	64.4	57.7	50.8	45.1
Est2Genome+	95.0	75.3	60.9	50.2	42.5	35.8	28.7
Exalin	97.6	89.2	82.1	75.5	68.7	63.6	58.3
Exonerate	95.3	83.6	73.4	63.6	53.1	44.4	31.7
ExonerateCG	96.1	81.3	70.3	57.1	47.1	36.3	24.3
ExonerateDG	95.9	82.4	71.1	58.4	49.1	38.1	25.9
GMap	91.6	83.4	75.4	68.1	58.2	45.2	28.4
GeneSeqer	72.7	64.0	56.7	49.2	42.5	35.7	29.5
Pairagon	98.9	92.9	87.7	83.6	79.3	75.2	68.7
PairagonX	98.9	92.9	87.7	84.5	81.1	77.7	73.9
Palma	98.4	79.2	60.9	46.8	33.1	22.9	14.6
Sim4	93.3	81.0	69.2	58.2	44.2	30.4	15.7
Sim4+	93.9	80.5	67.8	56.1	42.0	28.8	15.0
Spaln	97.6	88.5	81.0	75.5	71.7	65.9	62.9
SpalnX	97.1	88.0	81.0	76.2	71.3	66.1	62.0
Spidey	92.4	55.1	31.8	19.2	10.1	5.2	2.7
SpideyX	72.2	59.0	48.4	39.5	33.1	27.6	20.5
Splign	97.4	86.4	58.8	15.8	3.6	1.3	0.3
SplignX	97.2	88.3	80.6	74.4	67.1	55.6	42.0
Xat	93.1	83.0	75.4	69.5	63.5	58.6	52.5

3 Error Classifications

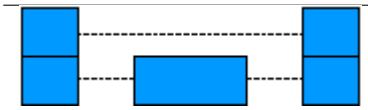
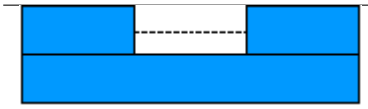
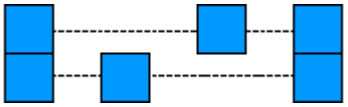
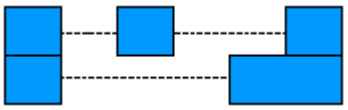
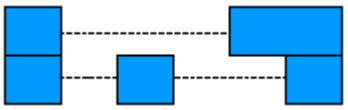
3.1 Method

The method we used to classify errors is similar to the way that alternate splicing patterns are classified (Cite Guigo), except instead of comparing two plausible gene structures, we are comparing the reference gene annotation to the gene annotation produced by the alignment. The nucleotides of each exon are marked with '1' and everywhere else is marked with '0'. The binary representation of the annotation is placed above the representation for the alignment. Repeat alignment columns are then elided.

A finite state machine then processes the columns to determine what sorts of errors occur. Any place where the numbers within the column do not match indicate the presence of some sort of error. The patterns that the finite state machine recognizes are enumerated below.

3.2 Gene Structure Errors

Gene structure errors are noted by either an exon or intron that is completely wrong, i.e. no part of it maps to a correct exon or intron.

Pattern	Name	Notes
	Missing Exon	Similar patterns exist for missing 5' and 3' exons. If the missing exon is less than 30 bases long, it is reclassified as a Missing Microexon.
	Extra Exon	This occurs when the alignment contains an exon that is not found in the annotation, and is surrounded by valid exons.
	Missing Intron	This occurs when the aligner merges two exons and marks the intron as a large gap.
	Extra Intron	This occurs when an intron is placed in the middle of an exon, usually resulting from a large gap.
	Shifted Exon	This occurs when an exon only found in the reference annotation is adjacent to an exon that only occurs in the alignment. This only counts as one error.
	Assimilated Exon	This occurs when an exon is 'assimilated' into another exon, such that the one exon is missing and the other has an incorrect boundary. This also just counts as one error.
	Add and Shift	This is the reverse of an assimilated exon, occurring when an exon is split into two separate pieces, one of which does not overlap the original exon.

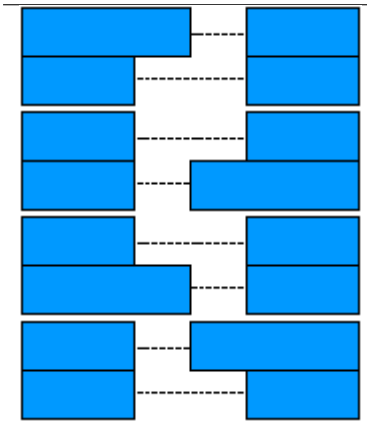
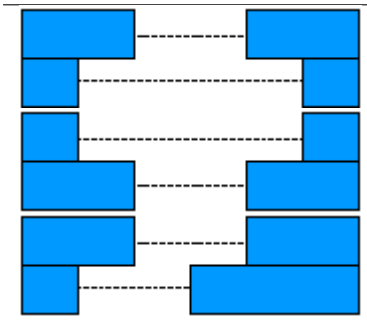
3.3 Transcript Boundary Errors

Transcript boundary errors encapsulate errors where the first or last exon is partially correct and the outermost boundary (i.e. the transcript start or transcript finish) is incorrect.

Pattern	Name	Notes
	Missing Start/Missing Finish	One end of the transcript is unaligned.
	Extra Start/Extra Finish	The alignment stretches beyond the boundary of the reference annotation.

3.4 Splicing Errors

Splicing errors occur when an intron is correctly marked, but the exact boundaries are not.

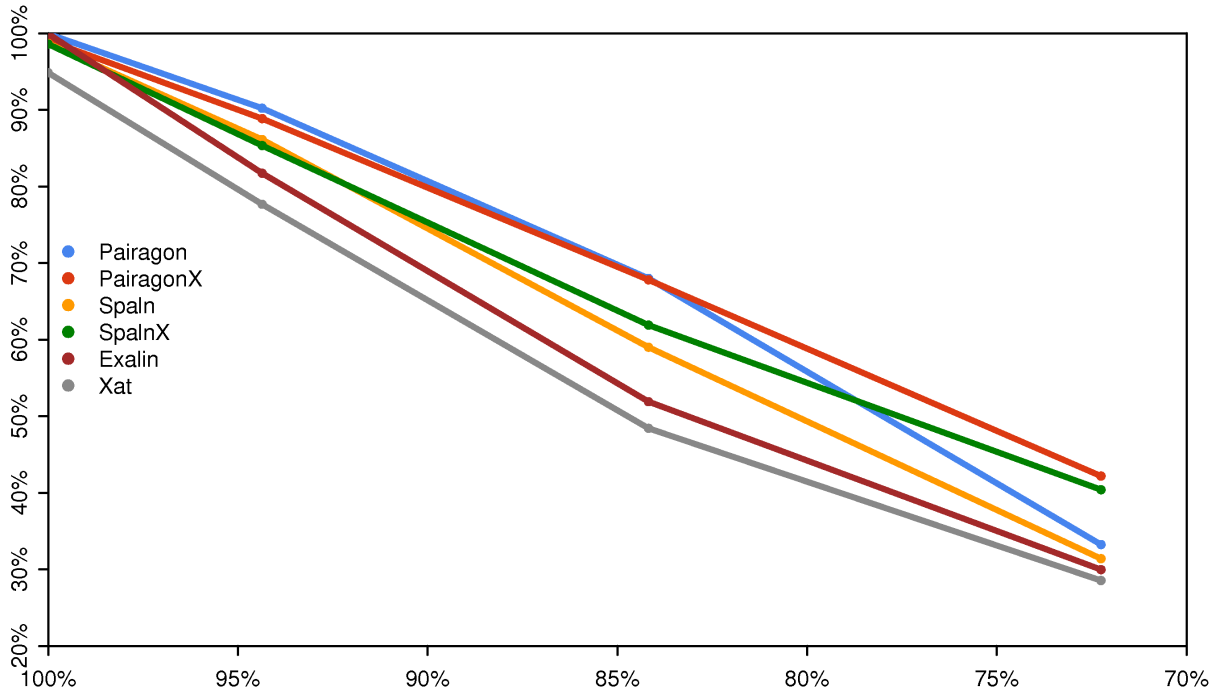
Pattern	Name	Notes
	One-Site Errors	This occurs when only one of boundaries of the intron is incorrect.
	Two-Site Errors	This occurs when both boundaries of the intron are incorrect.

3.5 Other Errors

Other errors include producing no alignment and producing an alignment on the wrong strand (not pictured).

4 Full Transcript Results

D	0.0	0.1	0.3	0.6
Identity	1.00	0.94	0.84	0.72
Blat	98.61	65.17	20.51	0.00
BlatX	98.80	51.56	7.10	0.00
Est2Genome	98.17	77.40	44.68	24.31
Est2Genome+	98.25	69.37	33.16	14.26
Exalin	99.94	81.01	50.77	29.39
Exonerate	99.51	69.15	34.27	13.13
ExonerateCG	57.82	42.37	23.23	9.13
ExonerateDG	95.01	66.62	32.55	11.84
GMap	99.82	82.38	48.55	19.73
GeneSeqer	93.79	70.75	37.19	17.54
HiSim4	97.51	77.10	35.17	4.40
Pairagon	99.92	90.12	67.76	32.50
PairagonX	99.20	89.21	68.50	43.30
Sim4	99.78	81.80	39.48	4.59
Sim4+	99.84	78.76	35.85	4.46
Spaln	98.66	84.56	55.27	27.67
SpalnX	98.60	84.40	59.59	37.67
Spidey	92.57	49.17	13.75	0.72
SpideyX	31.95	21.76	11.90	3.92
Splign	97.52	80.23	35.33	0.71
SplignX	97.44	80.20	49.59	13.11
Xat	95.01	77.60	47.92	27.81



5 Timing Data

Average runtime over all alignments in Experiment 1, as reported by the UNIX time command.

Program	Time [s]
Sim4	0.00
Splign	0.01
GMap	0.02
Spaln	0.02
Spidey	0.02
Blat	0.10
Xat	0.10
Exalin	1.55
Exonerate	1.57
GeneSeqer	1.97
Est2Genome	2.77
Pairagon	89.52
Palma	236.30