Supplementary Materials

TrueSight: a new algorithm for splice junction detection using RNA-seq

Yang Li, Hongmei Li-Byarlay, Paul Burns, Mark Borodovsky, Gene E. Robinson, and Jian Ma

Supplementary Methods

1. Expectation-Maximization with logistic regression

Initiation: The initial positive and negative sets $P^{(0)}$ and $N^{(0)}$ are selected as described in 'Initial spliced alignment datasets' section. Then, the logistic regression parameter $\beta^{(0)}$ is learned by the logistic regression package lr_trirls with TR-RILS algorithm implemented (1) with least square regularized weights. $x_i (i = 1, \dots, k)$ are initial *labeled* feature vectors that remain unchanged in the iteration process.

In *t*-th iteration, $t \geq 0$:

Expectation step: For $x_i, i = k+1, \dots, n$, a feature vector for junction *i*, we estimate the probability that x_i is in $P^{(t)}$ by

$$p\left(P^{(t)} \middle| x_i\right) = \frac{1}{1 + \exp(\beta_0^{(t)} + \sum_{j=1}^{10} \beta_j^{(t)} x_{ij})}$$

Classification step: $x_i (i = k+1, \dots, n)$ is assigned to $P^{(t+1)}$ if $p\left(P^{(t)} \middle| x_i\right) \geq 0.5$, otherwise it is assigned to $N^{(t+1)}$. Since we perform a binary classification, using 0.5 as threshold guarantees a choice of class for x_i $(i = k+1, \dots, n)$ with maximum posterior probability.

Maximization step: we use lr_trirls package to estimate new logistic regression parameters $\beta^{(t+1)}$ by maximizing logistic regression objective function (regularization term omitted):

$$\sum_{i=1}^{n} \left\{ w_i \log p\left(P^{(t+1)} \middle| x_i\right) + (1 - w_i) \log p\left(N^{(t+1)} \middle| x_i\right) \right\}$$

where $w_i = 1$ if $x_i \in P^{(t+1)}$; $w_i = 0$ if $x_i \in N^{(t+1)}$.

Iterations are terminated at convergence on *T*-th iteration, if $(P^{(T)}, N^{(T)})$ is the same as $(P^{(T-1)}, N^{(T-1)})$. By $\beta^{(T)}$ we designate a set of final logistic regression parameters. Thus, for junction *i*, we calculate the *SJ score* (SJS) as:

$$\frac{1}{1 + \exp(\beta_0^{(T)} + \sum_{j=1}^{10} \beta_j^{(T)} x_{ij})}$$

The score has the meaning of posterior probability.

2. Evaluating individual features

We used simulated datasets (see 'Simulated datasets' section) to evaluate the importance of each feature in the integrated model by calculating the area under the curve (AUC) values. We determined AUC values for the ten features in inferring true SJs in the three datasets (Table S1). We plotted the maximum AUC value for each feature and for the full model across all samples in Fig. 3. We also listed logistic regression parameters (β) for three datasets in Table S1.

3. Filters used in post-processing paired-end read

When aligning paired-end reads, several filters are used after assigning TrueSight score to all gapped alignment reads: (i) We retain the gapped alignment of a read with TrueSight score > 0.1, when its pair is not aligned. (ii) When both reads in a pair have alignments, their mapping orientations should be correct. Furthermore, the distance between pair-end reads should be within the user-defined maximum intron size. (iii) When both reads in a pair have canonical/semi-canonical gapped alignments, the transcript orientation inferred from their SJs should be consistent.

4. Generating honey bee RNA-seq data

Honey bees (*Apis mellifera*) were maintained at the University of Illinois Beekeeping Facility according to standard beekeeping procedures. Bees for RNA-seq were selected from colonies of single drone inseminated queens to reduce genetic variation for deep sequencing.

Total RNA was isolated from dissected fat body tissue using Qiagen RNeasy kits. cDNA was synthesized using 200 ng total RNA with the mix of Arrayscript reverse transcriptase (Ambion), random hexamer, Arrayscript buffer, dNTP, RNase Out in a 20 μl reaction.

Sequencing was done by Illumina HiSeq 2000. Ten RNA-seq samples have in total of 380 million paired-end reads (i.e. 190 million pairs) with each read being 100bp in length. The detailed coverage, alignment, and inferred AS numbers for each dataset are summarized in Table S6.

5. Modifying GLEAN gene model

5.1 Transcribed Islands

To obtain reliable 'transcribed islands' for the honey bee genome, we filtered TrueSight alignments for 10 samples by only retaining the best alignment (i.e. smallest number of mismatches for full alignment and highest TrueSight score for gapped alignment) for each RNA-seq read. After this filtering, we obtained read counts for each nucleotide position in the honey bee genome and searched for *transcribed islands* with the following criteria: (i) at least 5x coverage for each base of the island; and (ii) a transcribed island should be longer than 50bp.

Boundaries for 'transcribed islands' identified at this stage were determined independently from splicing signals or SJs inferred by TrueSight. The boundaries will be further refined by the subsequent gene model modification process.

All transcribed islands were compared with exons of GLEAN gene models; only islands non-overlapping with existing exons were retained for identifying novel exons.

5.2 TrueSight splice junction

SJs from independent TrueSight runs on ten RNA-seq samples were clustered together, and the highest TrueSight score was assigned for each SJ if it was detected in multiple samples. SJs with score higher than 0.5 were retained as TrueSight SJs (184,912, shown in Table S7) and were utilized to improve the current GLEAN models.

5.3 Improving GLEAN gene models with iterative algorithm

Initiation: By comparing TrueSight SJs with SJs from the GLEAN gene model ($model^0$, we define a set of exons and SJs as a primary gene $model$), TrueSight SJs are categorized into four subsets: (i) known SJs, which are already in $model^0$; (ii) novel SJs with both splice sites known ($novel_0^0$), which reflect exon skipping; (iii) novel SJs with only one known splice site ($novel_1^0$); and (iv) novel SJs with two novel splice sites ($novel_2^0$).

Iteration: In t^{th} ($t \geq 1$) iteration

Adding new link (SJ) to existing exons with $novel_0^{t-1}$

SJs in $novel_0^{t-1}$ provide novel links to existing exons in $model^{t-1}$, and are strong supports for cassette exons. $novel_0^{t-1}$ is added into $model^{t-1}$ to construct a new version of gene model $model^t$.

Modifying exon coordinates with $novel_1^{t-1}$

The original junction linking two exons ($a \sim b$; $c \sim d$) is $b \sim c$. If there is junction in $novel_1^{t-1}$: $b' \sim c$, such that $\|b' - b\| < 200, b' > a$, exon $a \sim b$ would have alternative boundary $a \sim b'$. If there is junction in $novel_1^{t-1}$: $b \sim c'$, such that $\|c' - c\| < 200, c' < d$, exon $c \sim d$ would have alternative boundary $c' \sim d$.

SJs used in modifying exon coordinates should be in the same strand as the exon. Both the SJs and modified exon boundaries are added into $model^t$.

All junctions in $novel_1^{t-1}$, which are not utilized in modifying exon boundaries and are not added into $model^t$, are treated as $novel^{t-1}$. Junctions in $novel_2^{t-1}$are also added into $novel^{t-1}$.

Junctions in the new set $novel^{t-1}$ are compared with $model^t$and categorized into $novel_0^t$, $novel_1^t$ and $novel_2^t$, which are used in the $t + 1^{th}$ iteration.

Termination: The modification process would terminate at T^{th}iteration when there is no junction in $novel_0^T$. SJs in $novel_1^{\mathrm{T}}$ and $novel_2^{\mathrm{T}}$ are utilized to add new exons and SJs to $model^T$.

Based on our RNA-seq data, after five iterations the number of SJs in original GLEAN gene models has increased from 53,884 to 66,847. The newly added junctions provided information for two types of AS: cassette exons (CE) and alternative exon boundaries (AEB).

5.4 Adding new exons and splice junctions

To be conservative in adding novel exons and SJs into the modified GLEAN model, we only use $novel_1^{\mathrm{T}}$ and $novel_2^{\mathrm{T}}$ with TrueSight score greater than 0.5.

Adding new exons and splice junctions with $novel_1^T$

For exon $a \sim b$, if there is a junction in $novel_1^{t-1}$: $b \sim p'$ such that we can find a transcribed island ($p \sim q$) satisfying $\|p' - p\| < 100, p' < q$, we can label the transcribed island ($p' \sim q$) as a new exon, with one boundary (p') fixed and the other (q) undetermined.

Symmetrically, for exon $a \sim b$, if there is a junction in $novel_1^{t-1}$: $q' \sim a$ such that we can find a transcribed island $(p \sim q)$ satisfying $\|q' - q\| < 100, q' > p$, we can label the transcribed island $(p \sim q')$ as a new exon, with one boundary (q') fixed and the other (p) undetermined.

The new exons identified in this process are in two subsets: (i) new exons with both boundaries fixed, since both ends of these exons are linked to known exons by junctions in $novel_1^T$. (ii) new exons with only one end defined (*Novel Terminal Exons*). These *Novel Terminal Exons* are either first/last exons of the whole transcripts, or linking to further novel exons by SJs in $novel_2^T$.

Adding new exons and splice junctions with $novel_2^T$

For a SJ in $novel_2^{\mathrm{T}}$: $q' \sim p'$, if there are two transcribed islands $(p_1 \sim q_1, p_2 \sim q_2)$ such that: $q_1 < p_2, \|q' - q_1\| < 100, q' > p_1$, $\|p' - p_2\| < 100, p' < q_2$, we can link these two transcribed islands together by the junction $q' \sim p'$.

There are two outcomes of this novel exon adding process: (i) novel multi-exon transcripts can be identified in inter-genic regions of the GLEAN models; (ii) novel exons might be connected to Novel Terminal Exons identified in the previous process, thus these novel exons serve as new *Novel Terminal Exons* for known genes.

Comparing with the original GLEAN models, 5,873 new exons were added and the number of SJs in the modified model has increased to 70,022.

Overall, we have identified 2,803 multi-exon transcripts in inter-genic regions of the GLEAN model, which cannot be connected to existing genes by known/predicted SJs.

6. Detecting Intron Retention

We define a honey bee intron as IR by two criteria: (i) each base of the intron has >5x coverage from TrueSight RNA-seq alignments in our dataset. (ii) IR inclusion ratio should be at least three times higher than the average honey bee intron inclusion ratio, which is 0.017. The first criterion guarantees the IR detectable by RNA-seq, and the second criterion will screen out potential false IRs caused by

RNA-seq artefacts mapped onto intron regions. Out of 59,674 honey bee introns in 9,355 multi-exon genes, we identified 5,258 (8.81%) intron retentions (IRs) in 2,846 (30.4%) genes.

References

1. Komarek, P. and Moore, A. (2003), Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs . *Artificial Intelligence and Statistics*, Vol. 83.

Supplementary Figures

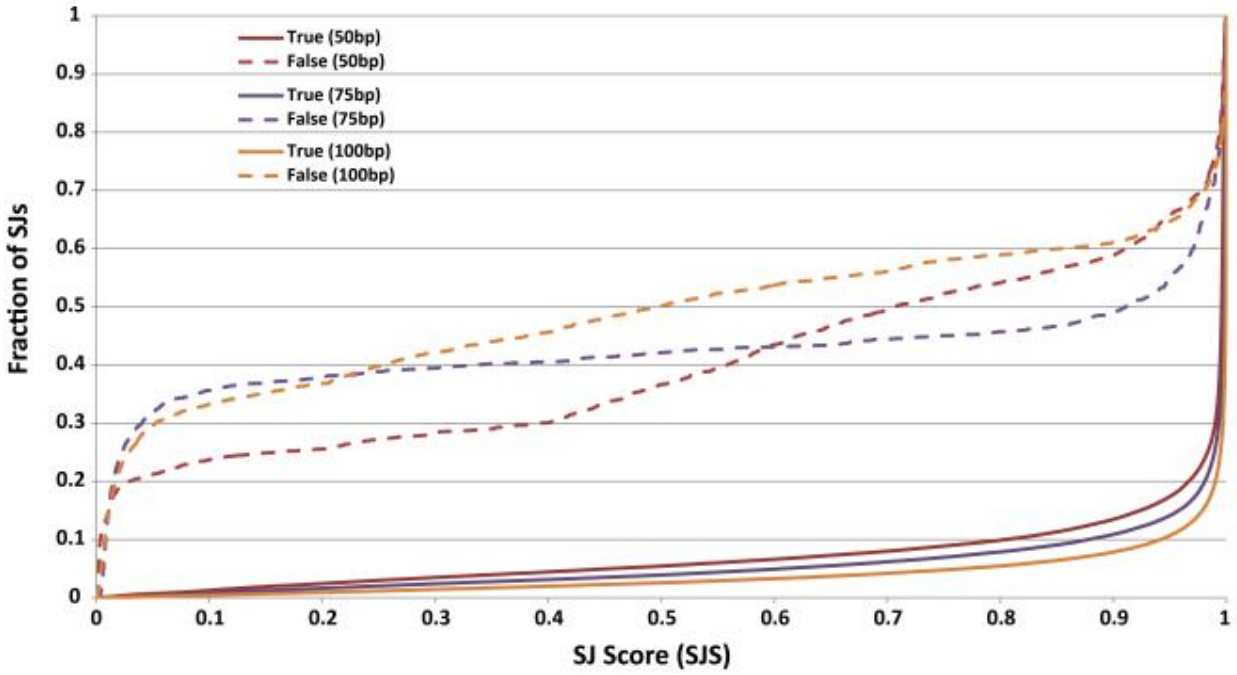

Figure S1: Distribution of TrueSight scores of true and false splice junctions. Y-axis is the fraction of true/false junctions under certain SJS.

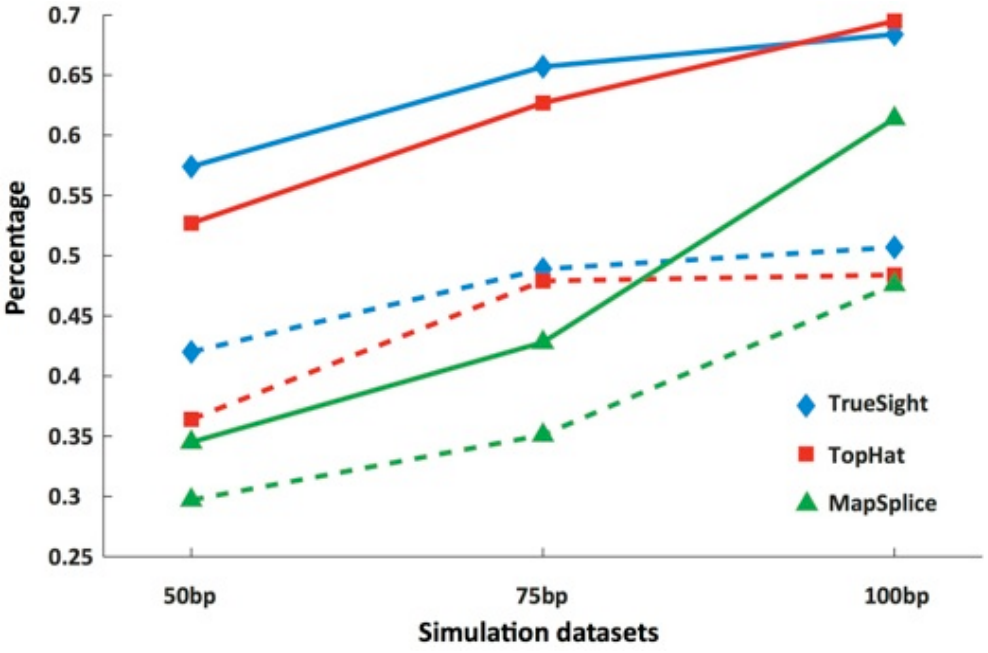

Figure S2: Sensitivity and specificity of transcripts assembled by Cufflinks from alignments generated by TrueSight, TopHat and MapSplice on 50, 75 and 100bp simulated datasets. Solid lines represent specificity and dashed lines represent sensitivity, comparing with UCSC Known Gene model (performed by Cuffcompare). All results are shown at intron-chain level, and a 'true' intron-chain has exactly the same intron set as a reference transcript, with possibly undetermined terminal exon boundaries, which are hard to correctly determine from RNA-seq data. *Sensitivity* is the fraction of recovered intron-chains in the simulated RNA-seq data and *specificity* is the ratio of the number of true intron-chains over the number of predicted ones.

Supplementary Tables

	50bp		75bp		100bp	
	AUC	**β**	**AUC**	**β**	**AUC**	**β**
Full model	0.985	6.713	**0.988**	6.353	0.983	4.962
Splicing signal (MC)	0.782	2.568	0.804	1.800	**0.810**	2.291
Splicing signal (PWM)	0.874	18.98	0.868	17.54	**0.883**	17.49
Coding potential	0.768	0.962	0.794	0.167	**0.797**	0.112
Coverage score	0.589	0.235	**0.610**	0.328	0.605	0.570
Intron size	0.831	0.212	**0.891**	0.368	0.888	0.164
Junction mapping #	0.634	0.375	0.691	2.248	**0.713**	1.096
Mapping length	**0.935**	15.65	0.912	15.43	0.908	17.36
Mapping entropy	0.906	2.834	**0.916**	3.022	0.905	3.436
Multiple mappings	0.705	3.023	0.690	3.680	**0.717**	3.188
Mismatches	**0.877**	0.202	0.866	0.832	0.862	1.650

Table S1: AUC value and final logistic regression parameters (β, with absolute value) for each of ten features and for full TrueSight model in inferring true SJs for the three simulated datasets. β value listed for the full model is $\boldsymbol{\beta_0}$. Highest AUC for each feature and full model are highlighted.

Dataset	Tools	Total	True	False
100bp	**TrueSight**	40,316	**39,705**	**611**
	TopHat	40,350	38,686	1,664
	MapSplice	40,630	38,934	1,696
	PASSion	22,245	19,937	2,308
75bp	**TrueSight**	15,996	**15,721**	**275**
	TopHat	16,199	15,536	3,242
	MapSplice	16,455	14,137	2,318
	PASSion	6,268	5,549	719
50bp	**TrueSight**	3,116	3,034	**82**
	TopHat	3,232	**3,079**	153
	MapSplice	3,174	2,683	491
	PASSion	969	736	233

Table S2: Results on multiple junctions covered by the same read. The number in the table represents the number of different multiple junctions identified, e.g. TrueSight has detected 40,316 different combinations of multiple junctions. Note that if there are three consecutive junctions: SJ_1, SJ_2 and SJ_3, and reads are found to cover the combination SJ_1-SJ_2, SJ_2-SJ_3 and SJ_1-SJ_2-SJ_3, three different combinations of multiple junctions are reported. True multiple junctions are combinations of annotated SJs, while false ones have at least one un-annotated SJ. Highest number of true multiple junctions and lowest number of false junctions are highlighted.

Species	Read length (bp)	Pair numbers (M)	SRA #	Reference genome	Gene annotations
Human	75	24.28	SRR065504	hg19	RefSeq, Ensembl, spliced EST, UCSC knownGene
Drosophila	76	13.60	SRR042297	dm3	FlyBase r5.42
Arabidopsis	75	20.90	SRR360205	TAIR10	TAIR10
C. elegans	102	12.24	SRR359066	ce10	RefSeq, Ensembl

Table S3: Real datasets used in performance evaluation. All datasets are paired-end RNA-seq reads, with length, coverage and SRA accession number listed. The genome reference used in alignment and the gene annotation references for SJs evaluation are also listed.

Dataset	Tools	Total	Both ends annotated		One SS novel	Both novel	SN (%)	SP (%)
			Known introns	Novel introns				
Human	**TrueSight**	209,447	174,543	9,386	19,269	6,249	**96.79**	**97.02**
	TopHat	192,939	164,291	8,613	14,494	5,541	91.10	**97.13**
	MapSplice	267,932	180,337	13,191	29,376	45,028	**100**	83.19
	PASSion	185,399	152,654	5,668	12,448	14,629	84.65	92.11
Drosophila	**TrueSight**	42,309	35,998	519	3,345	2,447	**94.30**	**91.84**
	TopHat	44,080	35,532	733	3,530	4,285	92.14	90.28
	MapSplice	96,204	38,565	777	6,626	50,236	**100**	47.78
	PASSion	40,566	34,617	443	2,162	3,344	89.76	**91.76**
Arabidopsis	**TrueSight**	113,825	99,070	701	8,479	5,575	**98.44**	**94.18**
	TopHat	127,310	98,339	2,655	9,218	17,098	97.84	86.57
	MapSplice	212,713	100,510	1,220	13,776	97,207	**100**	54.30
	PASSion	105,588	93,952	510	2,934	8,192	93.48	**92.24**
C. elegans	**TrueSight**	75,048	63,847	1,177	5,160	4,864	**100**	**93.52**
	TopHat	54,186	44,264	1,456	4,016	4,450	69.33	91.79
	MapSplice	92,105	63,807	1,192	6,018	21,088	**99.94**	77.10
	PASSion	66,252	60,247	525	2,372	3,108	94.36	**95.31**

Table S4: Evaluation results on real datasets. All mapped SJs from different tools are categorized into four classes as described in the main text. *Sensitivity* is the fraction of 'known introns' to the largest number of 'known introns' discovered by one method, thus the most exhaustive method is defined to have 100% sensitivity. *Specificity* is calculated by dividing the number of 'both novel' junctions over the 'total' number of splice junctions reported. Top two SN and SP for each dataset are highlighted.

Dataset	Tools	Semi-canonical		Non-canonical	
		SN (%)	SP (%)	SN (%)	SP (%)
50bp	TrueSight	71.92	89.40	12.08	19.34
	TopHat	74.23	61.15	0.67	30.00
	MapSplice	54.97	64.46	6.04	4.29
	PASSion	81.63	80.41	11.41	0.83
75bp	TrueSight	77.67	92.56	12.42	18.26
	TopHat	87.65	56.46	0.65	17.31
	MapSplice	75.16	63.56	13.72	6.85
	PASSion	82.05	92.85	11.76	3.04
100bp	TrueSight	81.89	94.84	13.21	35.65
	TopHat	89.99	50.60	0	16.98
	MapSplice	83.72	73.67	15.72	10.73
	PASSion	82.45	94.83	12.58	3.38

Table S5: Semi-canonical and non-canonical junctions identified by TrueSight, TopHat, MapSplice and PASSion on 50bp, 75bp and 100bp simulation datasets. TopHat and PASSion identified largest number of semi-canonical junctions, while TrueSight and PASSion achieved the highest specificity. When searching non-canonical junctions that have no specific splice site signal, TrueSight shows the highest SP, especially in 100bp reads. TrueSight, MapSplice and PASSion show approximately the same SN. Few correct non-canonical junctions were found by TopHat, while PASSion had lowest SP among all the methods tested. The SN and SP are defined the same as in Table 1: SN is the fraction of simulated junctions correctly detected by TrueSight; SP is the fraction of true junctions (comparing with RefSeq, Ensembl, spliced EST and UCSC Known Gene) among all predicted junctions.

Sample ID	Reads (M)	Aligned reads (M)	Uniquely aligned reads (M)	Gapped alignment (M)	Uniquely gapped alignment (M)	SJs	AS genes
1	38.23	30.29	28.98	7.88	7.69	97,267	5,012
2	40.96	32.37	31.18	8.46	8.26	101,380	5,154
3	32.34	24.97	23.93	6.35	6.19	91,516	5,037
4	37.03	29.80	28.68	7.73	7.53	103,607	5,056
5	36.32	29.15	27.95	7.57	7.37	95,343	5,079
6	37.85	29.70	28.65	7.74	7.58	94,194	5,067
7	38.45	30.87	29.81	8.03	7.85	99,798	5,105
8	36.70	29.46	28.13	7.71	7.49	96,898	5,016
9	47.49	11.86	11.34	3.16	3.05	78,087	4,600
10	34.69	27.57	26.68	7.30	7.15	95,822	4,989
Total	380.05	276.04	265.33	71.94	70.14	184,912	5,644

Table S6: RNA-seq dataset used in this study. All RNA-seq datasets are paired-end and each read pair is counted as two reads in the table. The total SJs are pooled from 10 samples, by assigning the highest SJS. Only SJs with SJS > 0.5 were reported. Note that the lower mapping rate in sample 9 was caused by deformed wing virus, which is prevalent in honey bees.

Table S7: List of AS genes based on ten RNA-seq datasets used in this study **(in a separate Excel file)**

Table S8: Modified honey bee GLEAN gene model **(in a separate Excel file)**