CSE 549: RNA-Seq aided gene finding

Stony Brook University

Finding Genes

We'll break gene finding methods into 3 main categories.

ab initio

latin — "from the beginning"
w/o experimental evidence

based on predictive modeling

how well do genomic sequences score under our "gene model"?

comparative

make use of knowledge across species

a known human gene is strong evidence for a chimp gene

many "housekeeping" genes are incredibly similar across highly divergent species

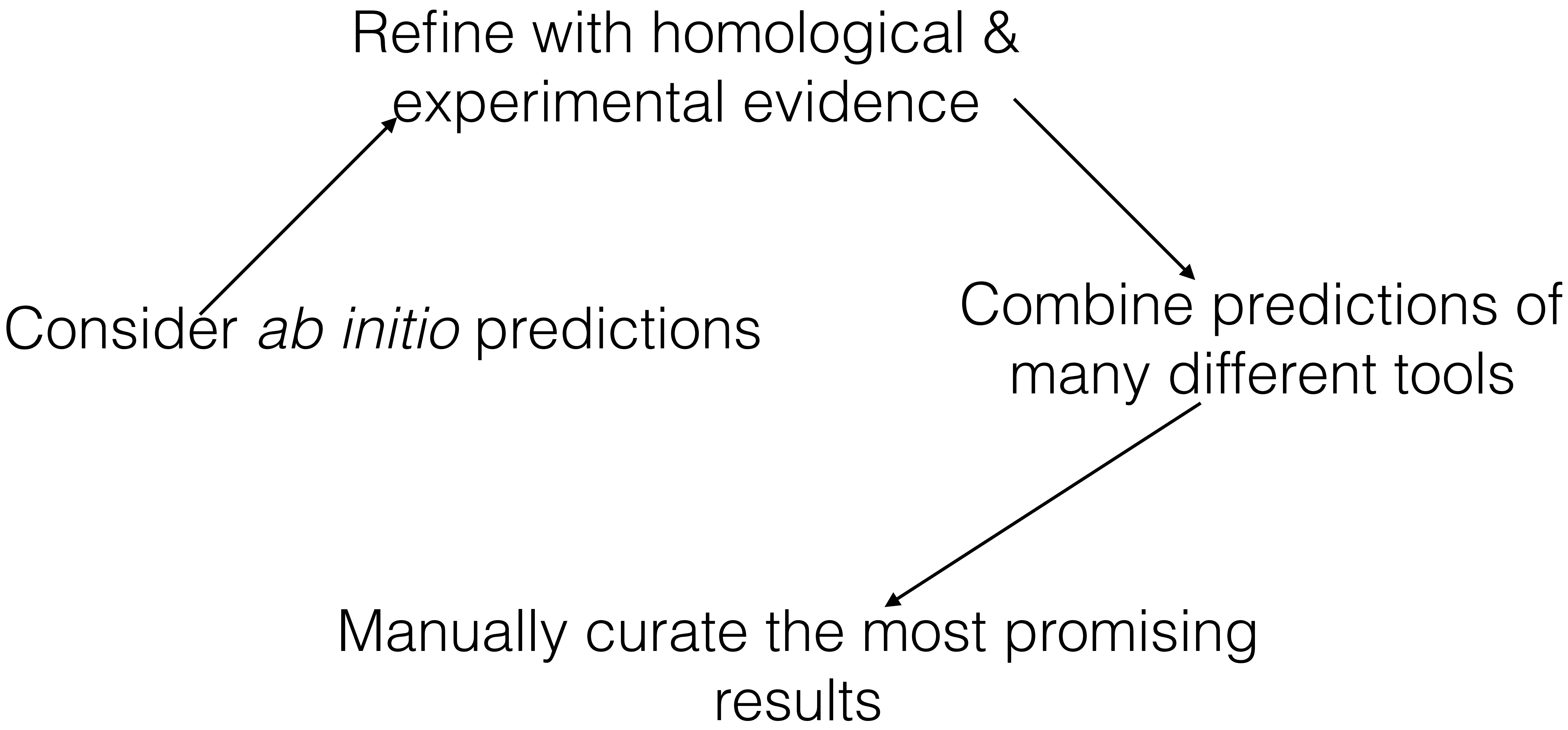
combined / extrinsic

Make use of experimental evidence (e.g. RNA-seq)

Evidence highlights transcribed regions

Gene structure extracted from evidence (potentially combined with model predictions)

Typical Approaches to Annotation are “Hybrid” Methods



Consider *ab initio* predictions

Refine with homological & experimental evidence

Combine predictions of many different tools

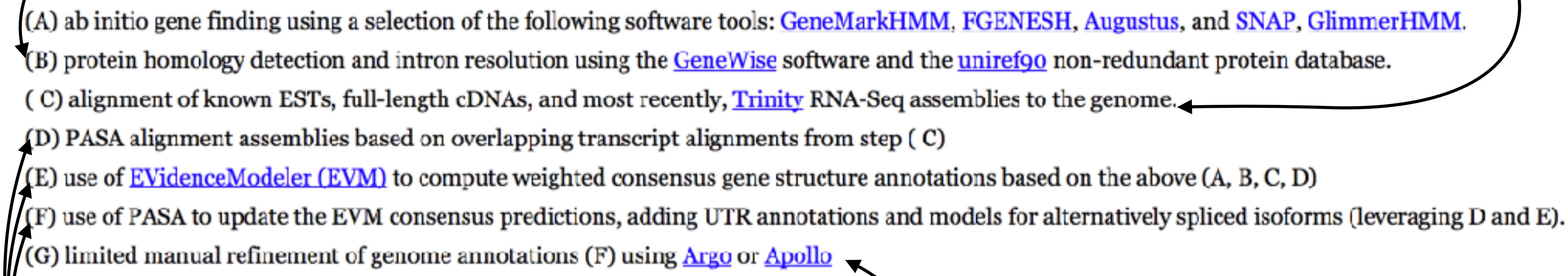
Manually curate the most promising results

Hybrid Gene Finding "Pipelines"

ab initio

comparative

de novo (experimental)



(A) ab initio gene finding using a selection of the following software tools: GeneMarkHMM, FGENESH, Augustus, and SNAP, GlimmerHMM.

(B) protein homology detection and intron resolution using the GeneWise software and the uniref90 non-redundant protein database.

(C) alignment of known ESTs, full-length cDNAs, and most recently, Trinity RNA-Seq assemblies to the genome.

(D) PASA alignment assemblies based on overlapping transcript alignments from step (C)

(E) use of EVidenceModeler (EVM) to compute weighted consensus gene structure annotations based on the above (A, B, C, D)

(F) use of PASA to update the EVM consensus predictions, adding UTR annotations and models for alternatively spliced isoforms (leveraging D and E).

(G) limited manual refinement of genome annotations (F) using Argo or Apollo

evidence combiners

manual curation

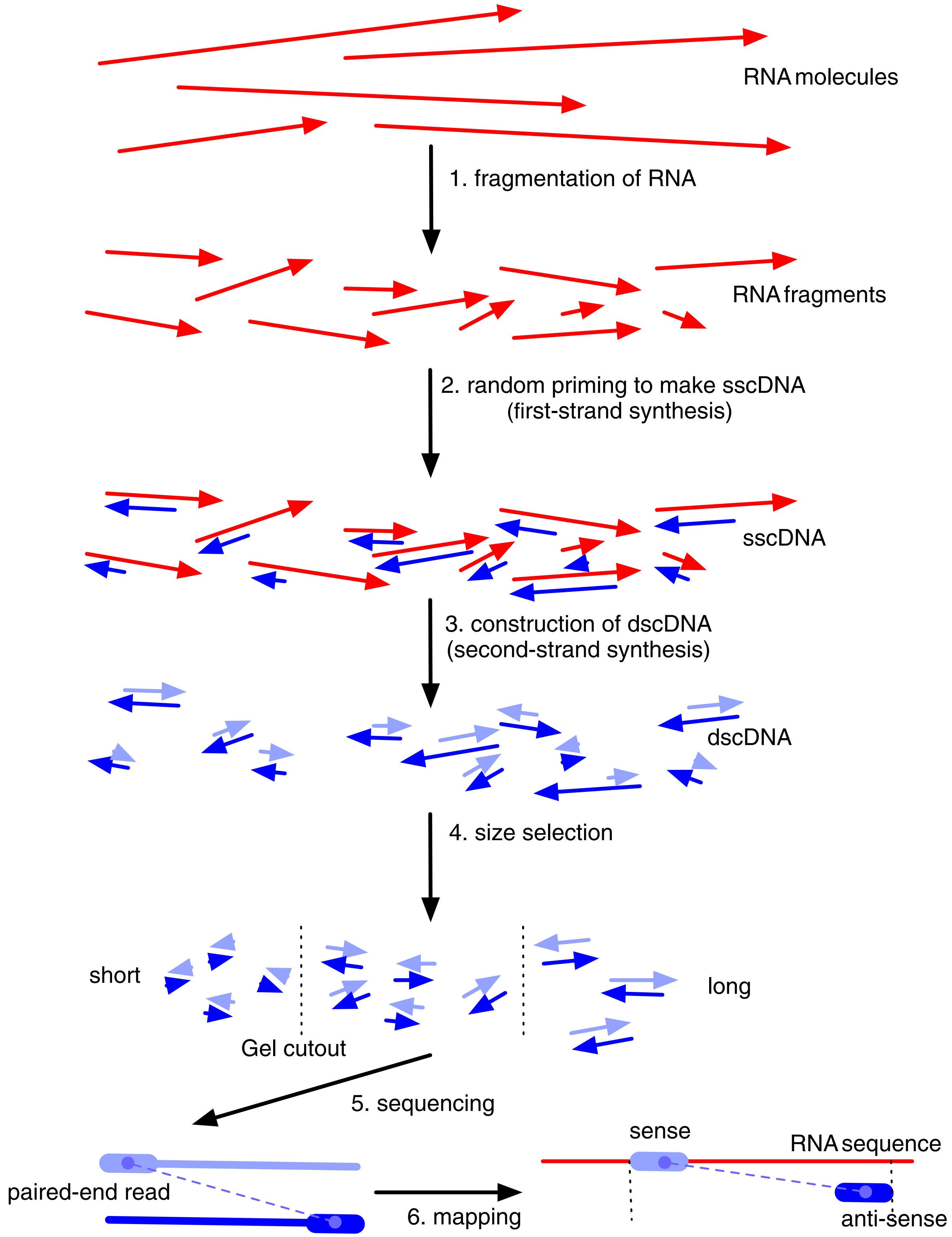
http://pasa.sourceforge.net

What is “experimental” data?

There are many ways we can obtain “experimental” evidence of a gene.

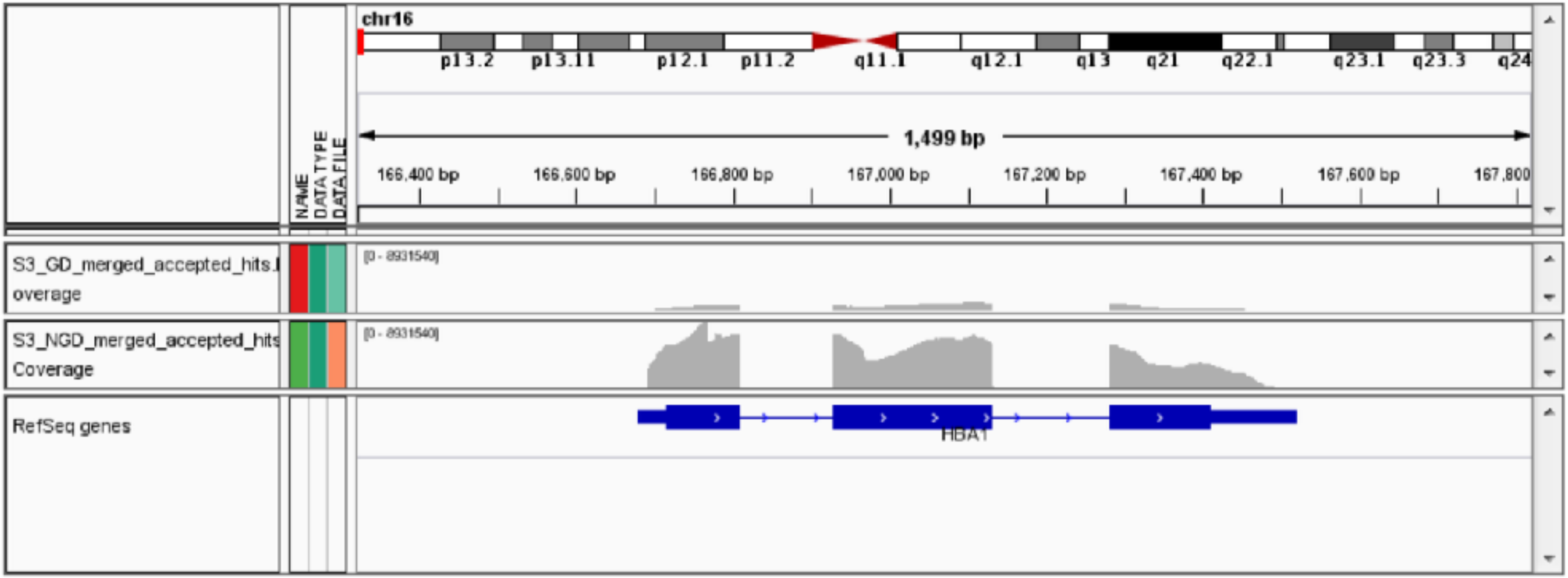
- Sequencing of protein product (mass spec.)
    - Expensive & slow, but provides direct evidence of protein coding genes (reverse translation)
- Expressed Sequence Tags (EST)
    - Targeted sequencing (typically Sanger sequencing) of expressed transcripts
- RNA-Seq
    - High throughput sequencing of the “transcriptome”

RNA-Seq — Sequencing Transcripts



Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): R22.

How is RNA-seq Useful?

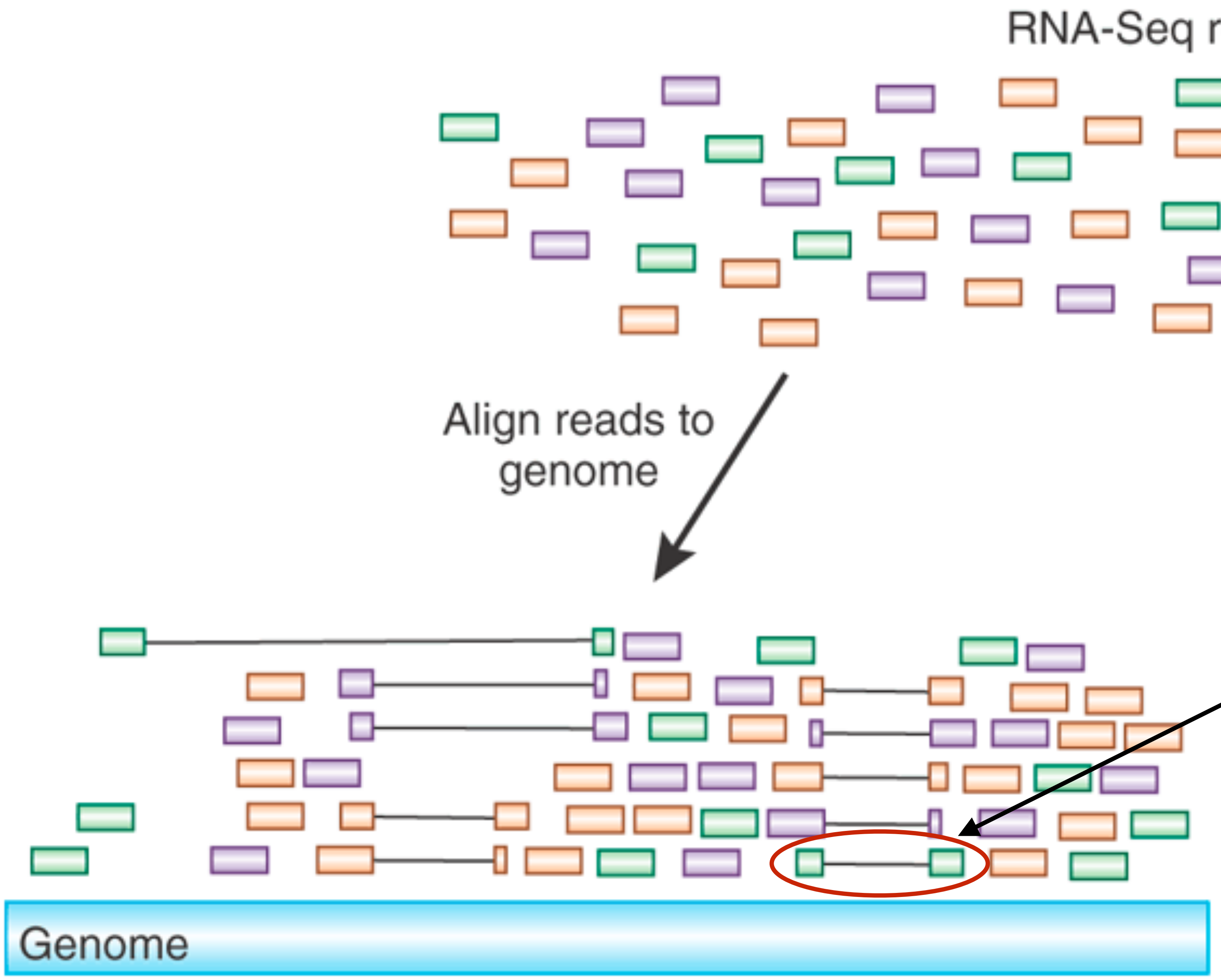


RNA-Seq reads come from a spliced transcript — if we can map them back to the genome, they give us evidence of **transcribed** regions.

Human genome contains > 14,000 pseudogenes [Pei et al. Genome Biology 2012]

Image from: Shin, Heesun, et al. "Variation in RNA-Seq Transcriptome Profiles of Peripheral Whole Blood from Healthy Individuals with and without Globin Depletion." PloS one 9.3 (2014): e91041.

RNA-seq Alignment is Hard

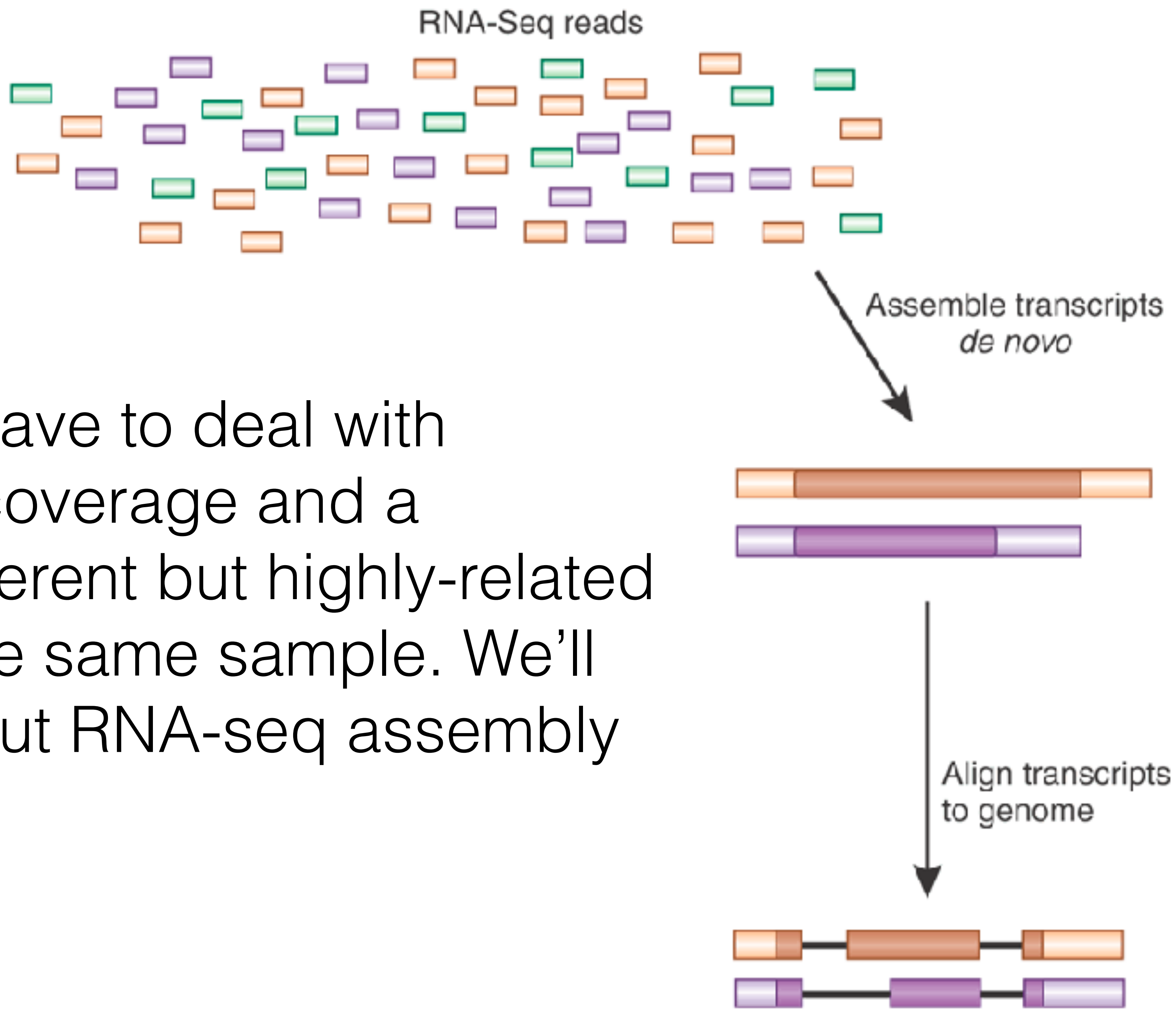


RNA-seq reads can span exon boundaries. Thus, contiguous regions of the read may map kilobases apart when aligning to the genome.

Improving the quality, sensitivity and speed of “spliced” alignment is still an active area of research. How can we be confident in a spliced-alignment when only a small portion of a read maps to an exon?

Haas, Brian J., and Michael C. Zody. "Advancing RNA-seq analysis." Nature biotechnology 28.5 (2010): 421.

RNA-seq Assembly is Harder



Assemblers have to deal with non-uniform coverage and a mixture of different but highly-related isoforms in the same sample. We'll talk more about RNA-seq assembly later.

Does Experimental Evidence Help?

There are many uses of RNA-seq apart from helping *ab initio* gene prediction.

Nonetheless, such evidence may be a powerful tool in helping us predict the existence of new genes.

Unsupervised Gene Finding w/o Experimental Evidence

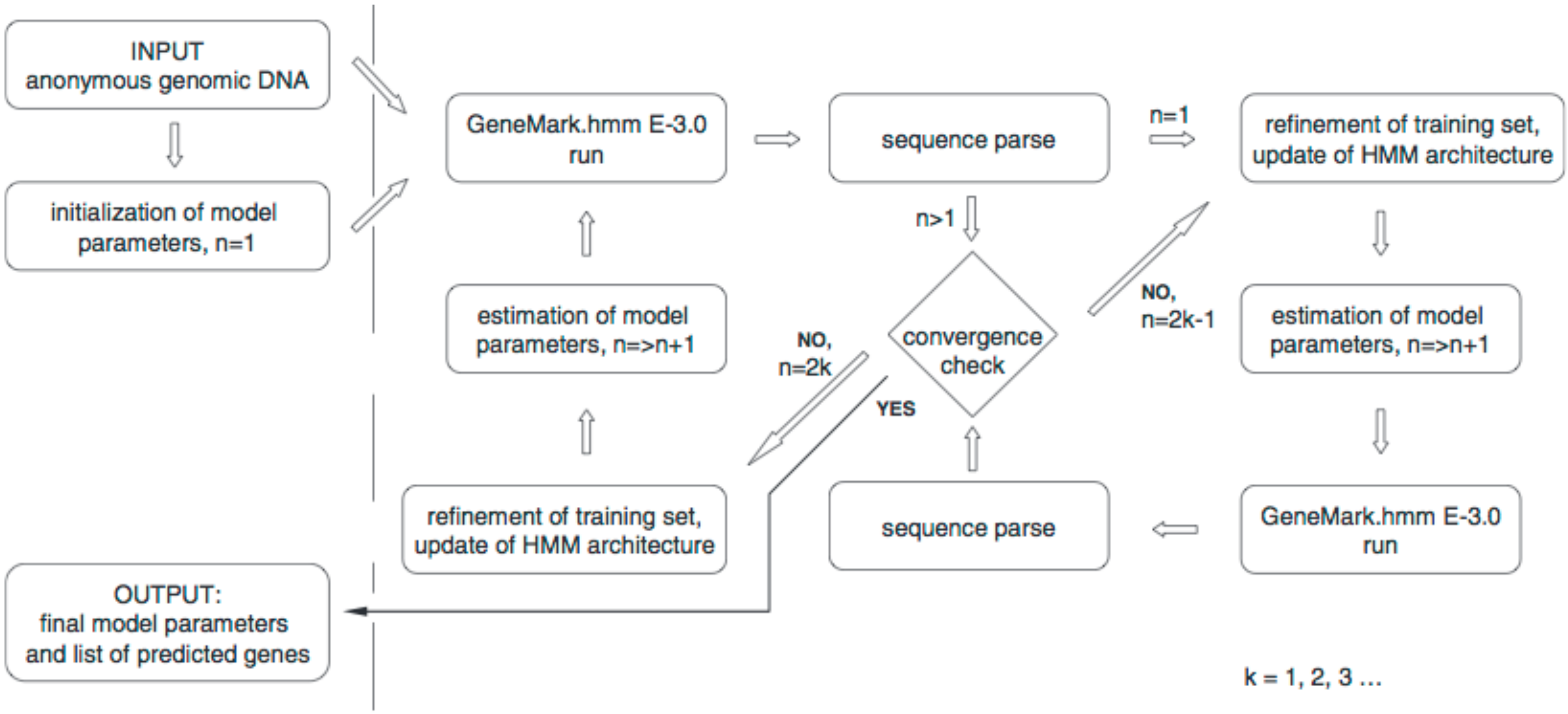

INPUT
anonymous genomic DNA
initialization of model parameters, n=1
GeneMark.hmm E-3.0 run
sequence parse
n=1
refinement of training set, update of HMM architecture
n>1
convergence check
NO, n=2k-1
estimation of model parameters, n=>n+1
NO, n=2k
YES
estimation of model parameters, n=>n+1
refinement of training set, update of HMM architecture
sequence parse
GeneMark.hmm E-3.0 run
OUTPUT:
final model parameters
and list of predicted genes
k = 1, 2, 3 ...

Figure 2. The step-wise diagram of the iterative unsupervised parameterization of HSMM implemented in GeneMark.hmm ES-3.0.

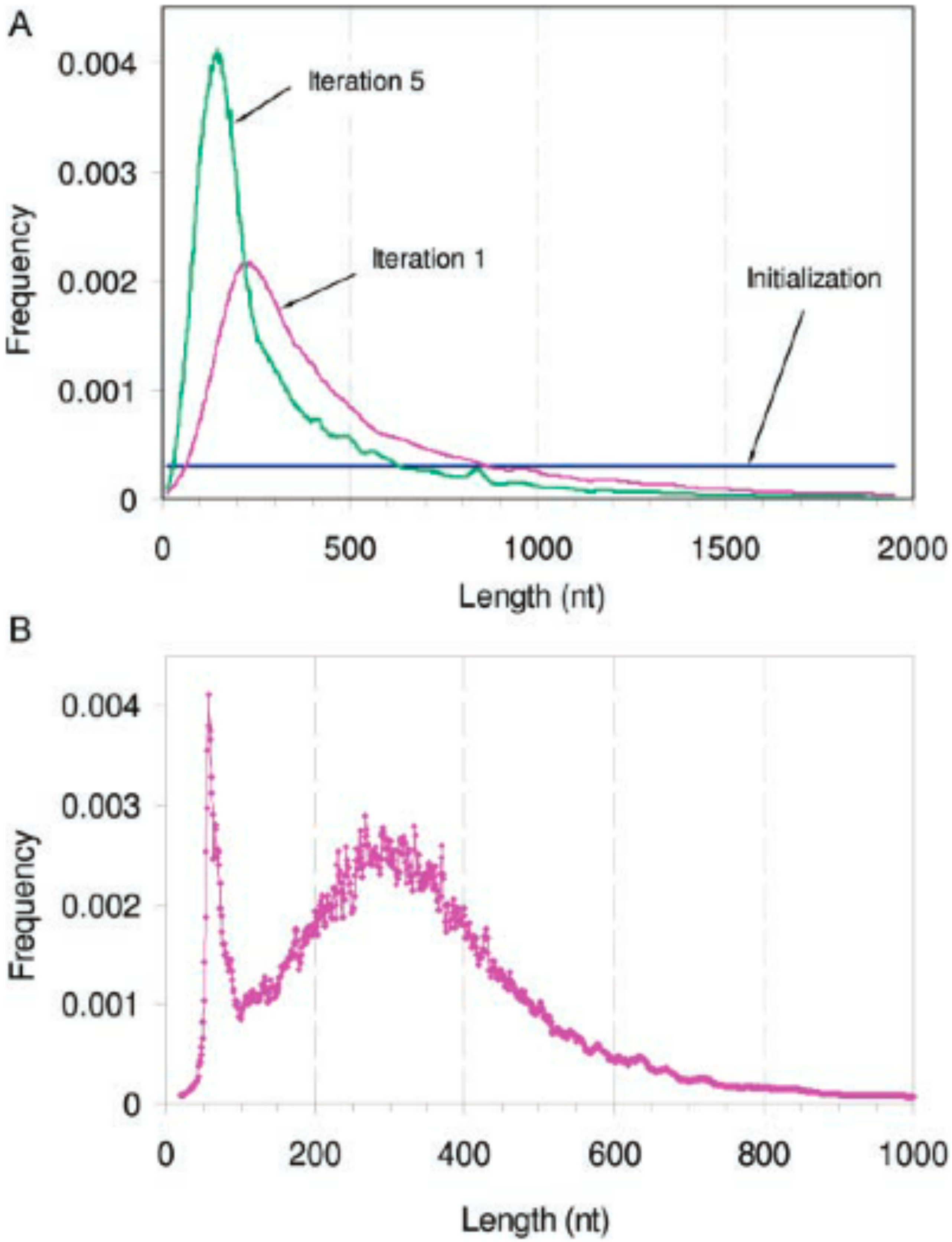
Start with heuristic / uninformative parameters and train via the Viterbi training algorithm we discussed previously.

Lomsadze, Alexandre, et al. "Gene identification in novel eukaryotic genomes by self-training algorithm." Nucleic Acids Research 33.20 (2005): 6494-6506.

Learning Feature Length Distributions

D. mel exon lengths

C. intestinalis intron lengths



Lomsadze et al. 2005

Unsupervised Training Competes With Supervised Training

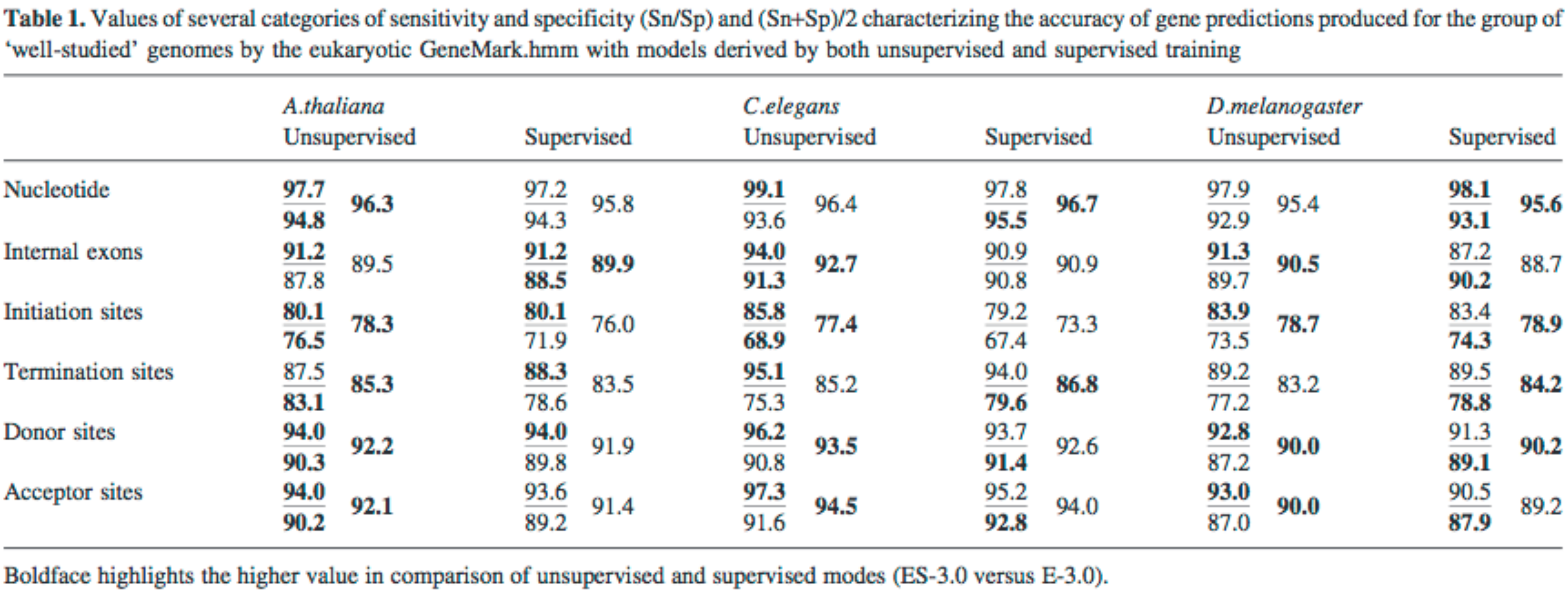


Table 1. Values of several categories of sensitivity and specificity (Sn/Sp) and (Sn+Sp)/2 characterizing the accuracy of gene predictions produced for the group of 'well-studied' genomes by the eukaryotic GeneMark.hmm with models derived by both unsupervised and supervised training

	A.thaliana				*C.elegans*				*D.melanogaster*			
	Unsupervised		Supervised		Unsupervised		Supervised		Unsupervised		Supervised	
Nucleotide	**97.7** **94.8**	**96.3**	97.2 94.3	95.8	**99.1** 93.6	96.4	97.8 **95.5**	**96.7**	97.9 92.9	95.4	**98.1** **93.1**	**95.6**
Internal exons	**91.2** 87.8	89.5	**91.2** **88.5**	**89.9**	**94.0** **91.3**	**92.7**	90.9 90.8	90.9	**91.3** 89.7	**90.5**	87.2 **90.2**	88.7
Initiation sites	**80.1** **76.5**	**78.3**	**80.1** 71.9	76.0	**85.8** **68.9**	**77.4**	79.2 67.4	73.3	**83.9** 73.5	**78.7**	83.4 **74.3**	**78.9**
Termination sites	87.5 **83.1**	**85.3**	**88.3** 78.6	83.5	**95.1** 75.3	85.2	94.0 **79.6**	**86.8**	89.2 77.2	83.2	89.5 **78.8**	**84.2**
Donor sites	**94.0** **90.3**	**92.2**	**94.0** 89.8	91.9	**96.2** 90.8	**93.5**	93.7 **91.4**	92.6	**92.8** 87.2	**90.0**	91.3 **89.1**	**90.2**
Acceptor sites	**94.0** **90.2**	**92.1**	93.6 89.2	91.4	**97.3** 91.6	**94.5**	95.2 **92.8**	94.0	**93.0** 87.0	**90.0**	90.5 **87.9**	89.2

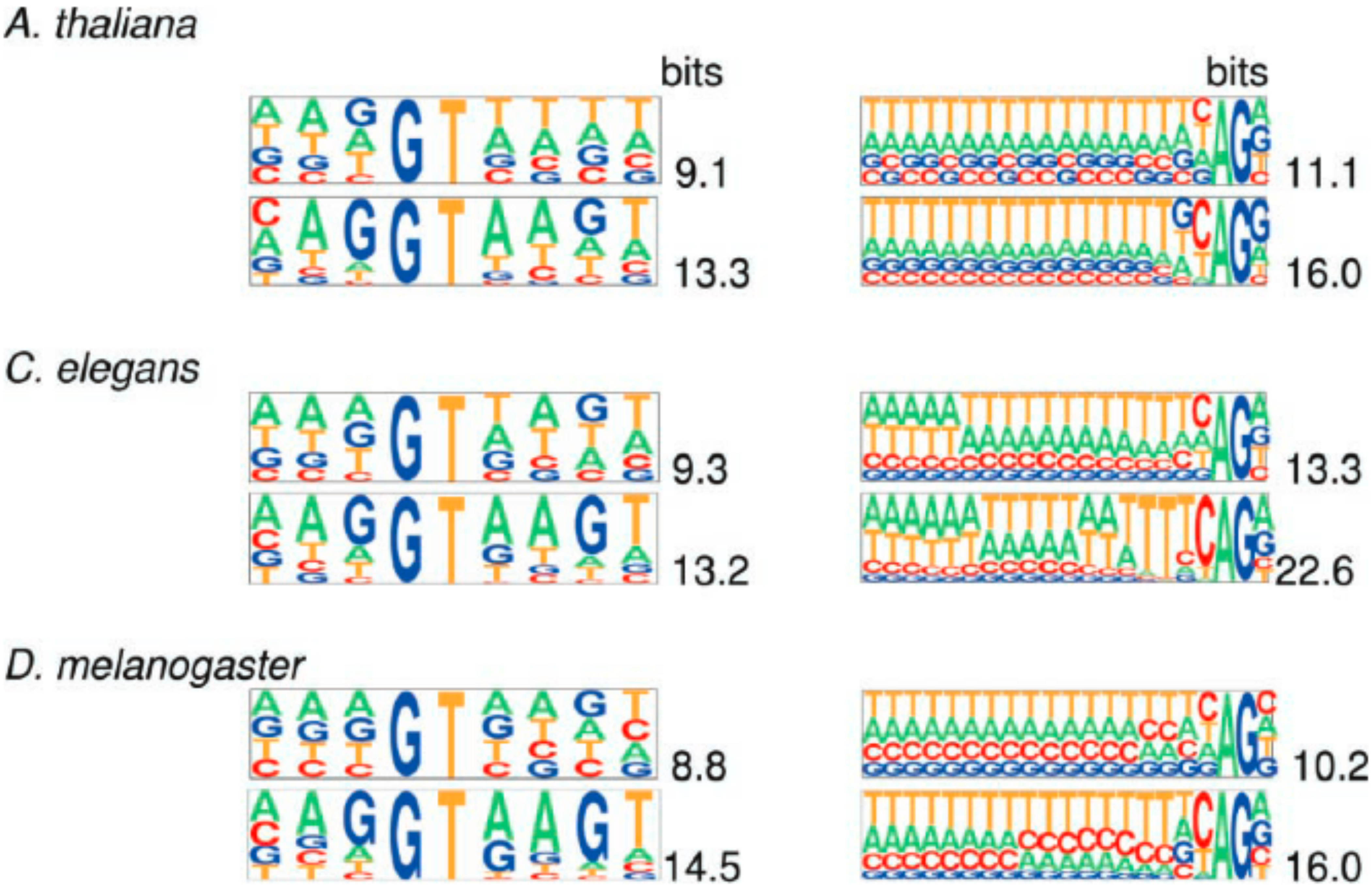
Boldface highlights the higher value in comparison of unsupervised and supervised modes (ES-3.0 versus E-3.0).

Lomsadze et al. 2005

Learning Site-Specific Features

Donor

Acceptor



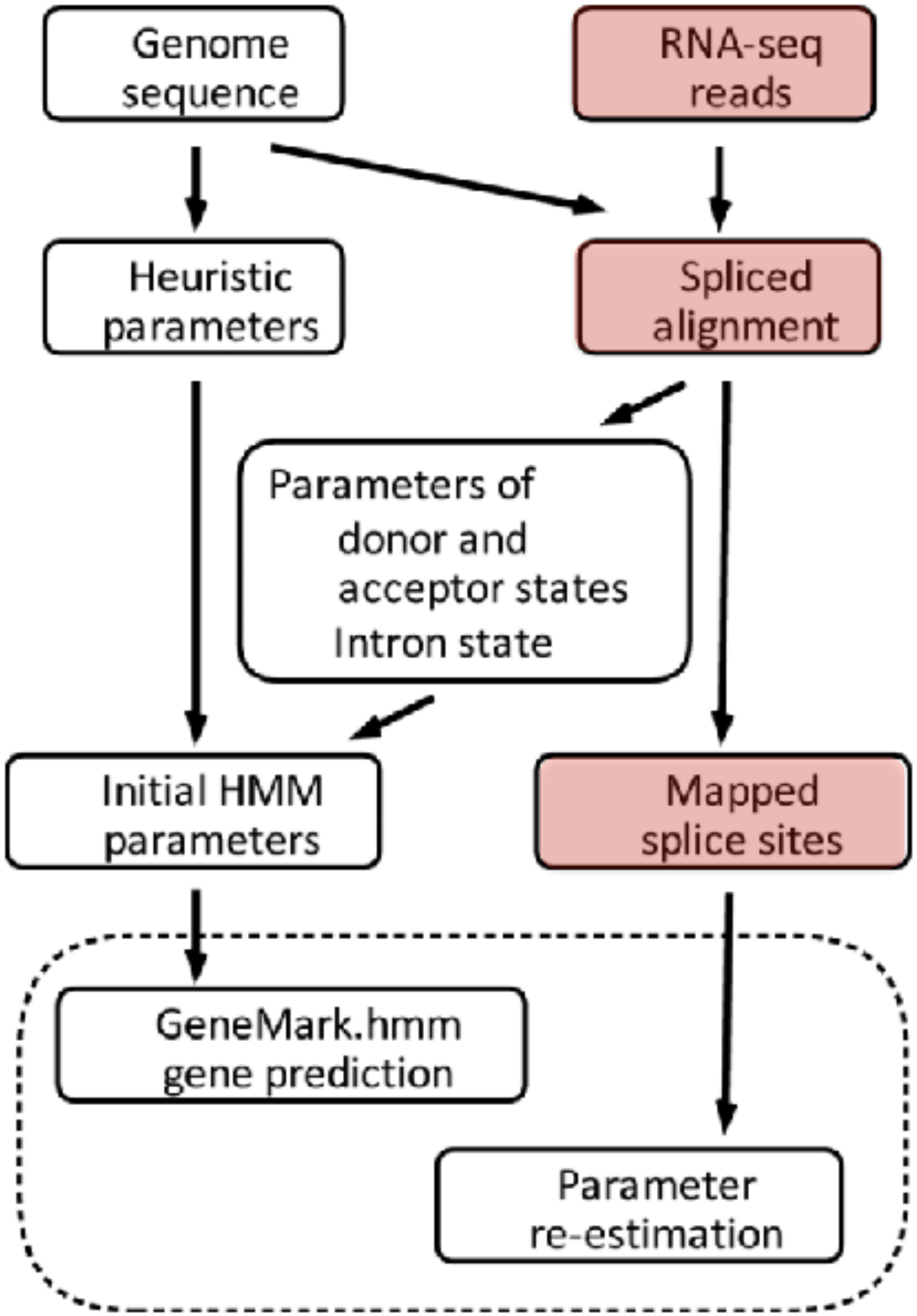
← 1st iter

← convergence

Lomsadze et al. 2005

Incorporating Experimental Evidence

GeneMark ET Procedure



Lomsadze, Alexandre, Paul D. Burns, and Mark Borodovsky. "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm." Nucleic acids research (2014): gku557.

Incorporating Experimental Evidence

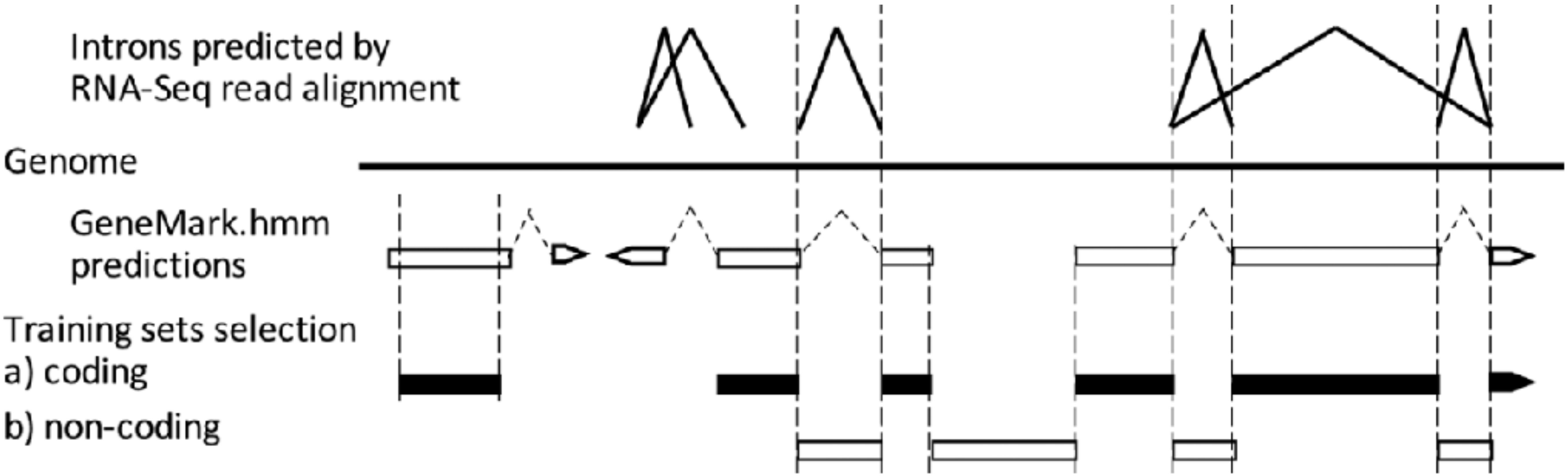


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

Lomsadze et al. 2014

Effect of Using Spliced-Alignments

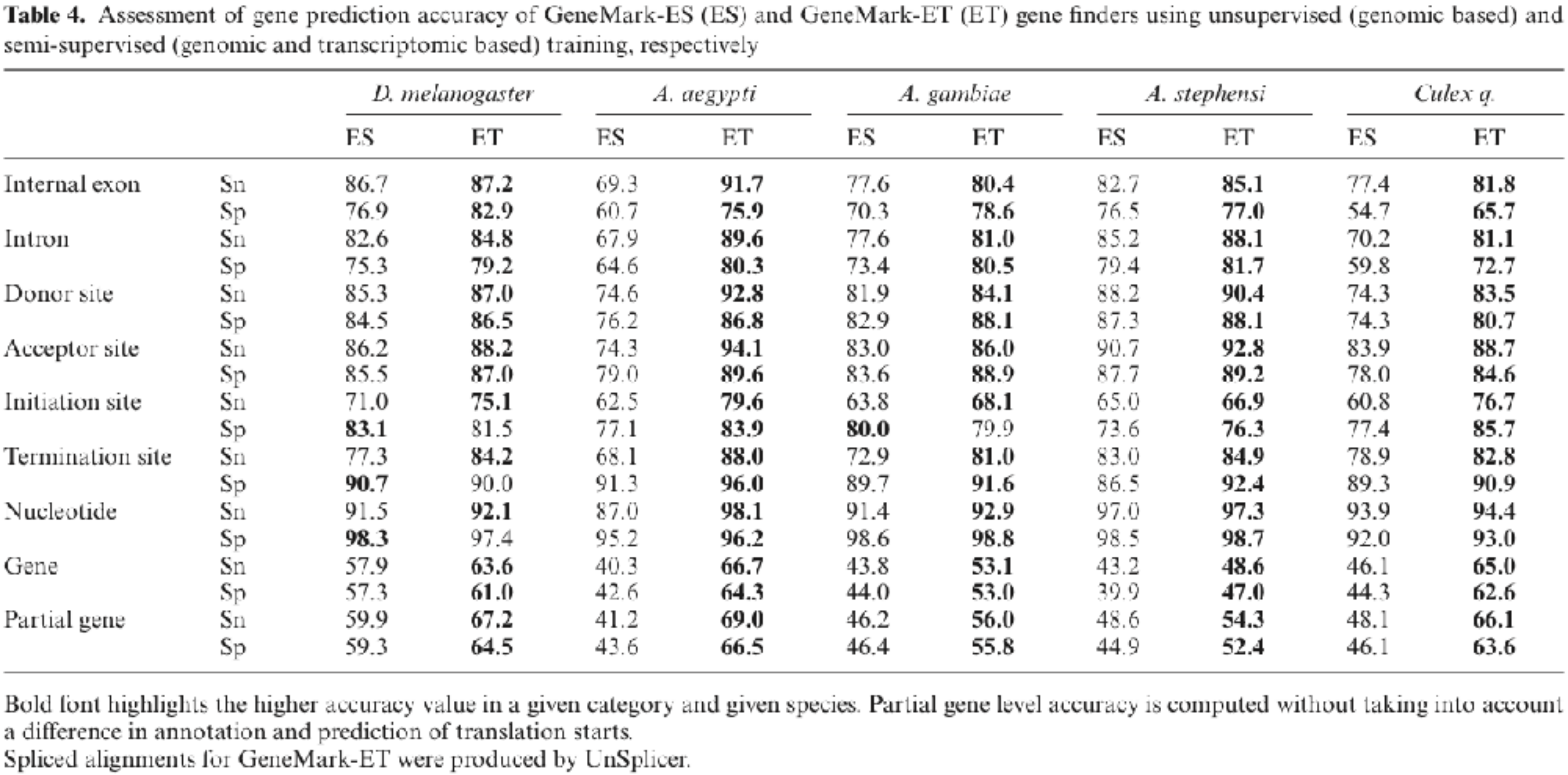


Table 4. Assessment of gene prediction accuracy of GeneMark-ES (ES) and GeneMark-ET (ET) gene finders using unsupervised (genomic based) and semi-supervised (genomic and transcriptomic based) training, respectively

		D. melanogaster		*A. aegypti*		*A. gambiae*		*A. stephensi*		*Culex q.*	
		ES	ET	ES	ET	ES	ET	ES	ET	ES	ET
Internal exon	Sn	86.7	**87.2**	69.3	**91.7**	77.6	**80.4**	82.7	**85.1**	77.4	**81.8**
	Sp	76.9	**82.9**	60.7	**75.9**	70.3	**78.6**	76.5	**77.0**	54.7	**65.7**
Intron	Sn	82.6	**84.8**	67.9	**89.6**	77.6	**81.0**	85.2	**88.1**	70.2	**81.1**
	Sp	75.3	**79.2**	64.6	**80.3**	73.4	**80.5**	79.4	**81.7**	59.8	**72.7**
Donor site	Sn	85.3	**87.0**	74.6	**92.8**	81.9	**84.1**	88.2	**90.4**	74.3	**83.5**
	Sp	84.5	**86.5**	76.2	**86.8**	82.9	**88.1**	87.3	**88.1**	74.3	**80.7**
Acceptor site	Sn	86.2	**88.2**	74.3	**94.1**	83.0	**86.0**	90.7	**92.8**	83.9	**88.7**
	Sp	85.5	**87.0**	79.0	**89.6**	83.6	**88.9**	87.7	**89.2**	78.0	**84.6**
Initiation site	Sn	71.0	**75.1**	62.5	**79.6**	63.8	**68.1**	65.0	**66.9**	60.8	**76.7**
	Sp	**83.1**	81.5	77.1	**83.9**	**80.0**	79.9	73.6	**76.3**	77.4	**85.7**
Termination site	Sn	77.3	**84.2**	68.1	**88.0**	72.9	**81.0**	83.0	**84.9**	78.9	**82.8**
	Sp	**90.7**	90.0	91.3	**96.0**	89.7	**91.6**	86.5	**92.4**	89.3	**90.9**
Nucleotide	Sn	91.5	**92.1**	87.0	**98.1**	91.4	**92.9**	97.0	**97.3**	93.9	**94.4**
	Sp	**98.3**	97.4	95.2	**96.2**	98.6	**98.8**	98.5	**98.7**	92.0	**93.0**
Gene	Sn	57.9	**63.6**	40.3	**66.7**	43.8	**53.1**	43.2	**48.6**	46.1	**65.0**
	Sp	57.3	**61.0**	42.6	**64.3**	44.0	**53.0**	39.9	**47.0**	44.3	**62.6**
Partial gene	Sn	59.9	**67.2**	41.2	**69.0**	46.2	**56.0**	48.6	**54.3**	48.1	**66.1**
	Sp	59.3	**64.5**	43.6	**66.5**	46.4	**55.8**	44.9	**52.4**	46.1	**63.6**

Bold font highlights the higher accuracy value in a given category and given species. Partial gene level accuracy is computed without taking into account a difference in annotation and prediction of translation starts.

Spliced alignments for GeneMark-ET were produced by UnSplicer.

Lomsadze et al. 2014

Effect of Using Spliced-Alignments

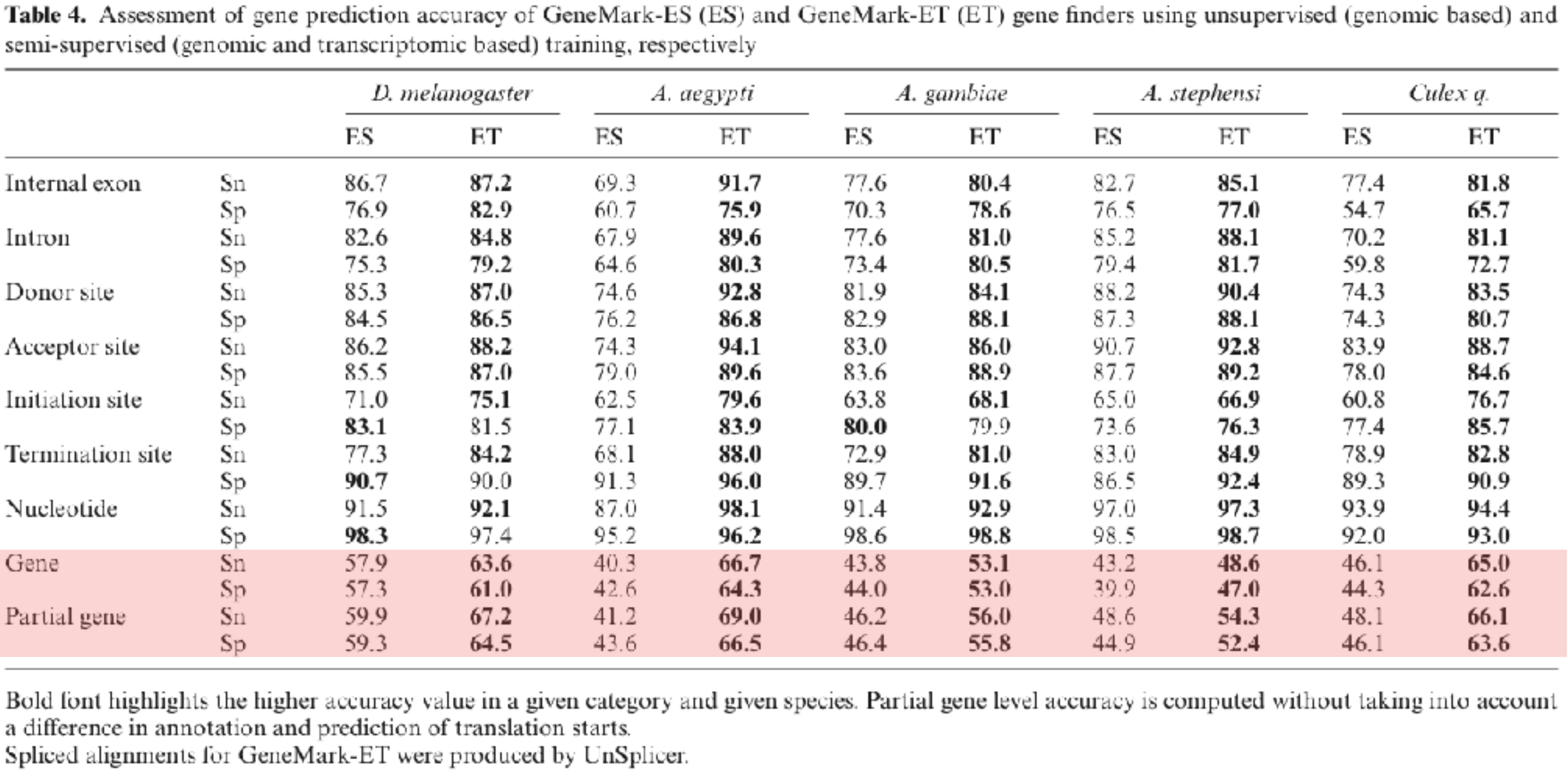


Table 4. Assessment of gene prediction accuracy of GeneMark-ES (ES) and GeneMark-ET (ET) gene finders using unsupervised (genomic based) and semi-supervised (genomic and transcriptomic based) training, respectively

		D. melanogaster		*A. aegypti*		*A. gambiae*		*A. stephensi*		*Culex q.*	
		ES	ET	ES	ET	ES	ET	ES	ET	ES	ET
Internal exon	Sn	86.7	**87.2**	69.3	**91.7**	77.6	**80.4**	82.7	**85.1**	77.4	**81.8**
	Sp	76.9	**82.9**	60.7	**75.9**	70.3	**78.6**	76.5	**77.0**	54.7	**65.7**
Intron	Sn	82.6	**84.8**	67.9	**89.6**	77.6	**81.0**	85.2	**88.1**	70.2	**81.1**
	Sp	75.3	**79.2**	64.6	**80.3**	73.4	**80.5**	79.4	**81.7**	59.8	**72.7**
Donor site	Sn	85.3	**87.0**	74.6	**92.8**	81.9	**84.1**	88.2	**90.4**	74.3	**83.5**
	Sp	84.5	**86.5**	76.2	**86.8**	82.9	**88.1**	87.3	**88.1**	74.3	**80.7**
Acceptor site	Sn	86.2	**88.2**	74.3	**94.1**	83.0	**86.0**	90.7	**92.8**	83.9	**88.7**
	Sp	85.5	**87.0**	79.0	**89.6**	83.6	**88.9**	87.7	**89.2**	78.0	**84.6**
Initiation site	Sn	71.0	**75.1**	62.5	**79.6**	63.8	**68.1**	65.0	**66.9**	60.8	**76.7**
	Sp	**83.1**	81.5	77.1	**83.9**	**80.0**	79.9	73.6	**76.3**	77.4	**85.7**
Termination site	Sn	77.3	**84.2**	68.1	**88.0**	72.9	**81.0**	83.0	**84.9**	78.9	**82.8**
	Sp	**90.7**	90.0	91.3	**96.0**	89.7	**91.6**	86.5	**92.4**	89.3	**90.9**
Nucleotide	Sn	91.5	**92.1**	87.0	**98.1**	91.4	**92.9**	97.0	**97.3**	93.9	**94.4**
	Sp	**98.3**	97.4	95.2	**96.2**	98.6	**98.8**	98.5	**98.7**	92.0	**93.0**
Gene	Sn	57.9	**63.6**	40.3	**66.7**	43.8	**53.1**	43.2	**48.6**	46.1	**65.0**
	Sp	57.3	**61.0**	42.6	**64.3**	44.0	**53.0**	39.9	**47.0**	44.3	**62.6**
Partial gene	Sn	59.9	**67.2**	41.2	**69.0**	46.2	**56.0**	48.6	**54.3**	48.1	**66.1**
	Sp	59.3	**64.5**	43.6	**66.5**	46.4	**55.8**	44.9	**52.4**	46.1	**63.6**

Bold font highlights the higher accuracy value in a given category and given species. Partial gene level accuracy is computed without taking into account a difference in annotation and prediction of translation starts.
Spliced alignments for GeneMark-ET were produced by UnSplicer.

Lomsadze et al. 2014

Effect of Using Spliced-Alignments

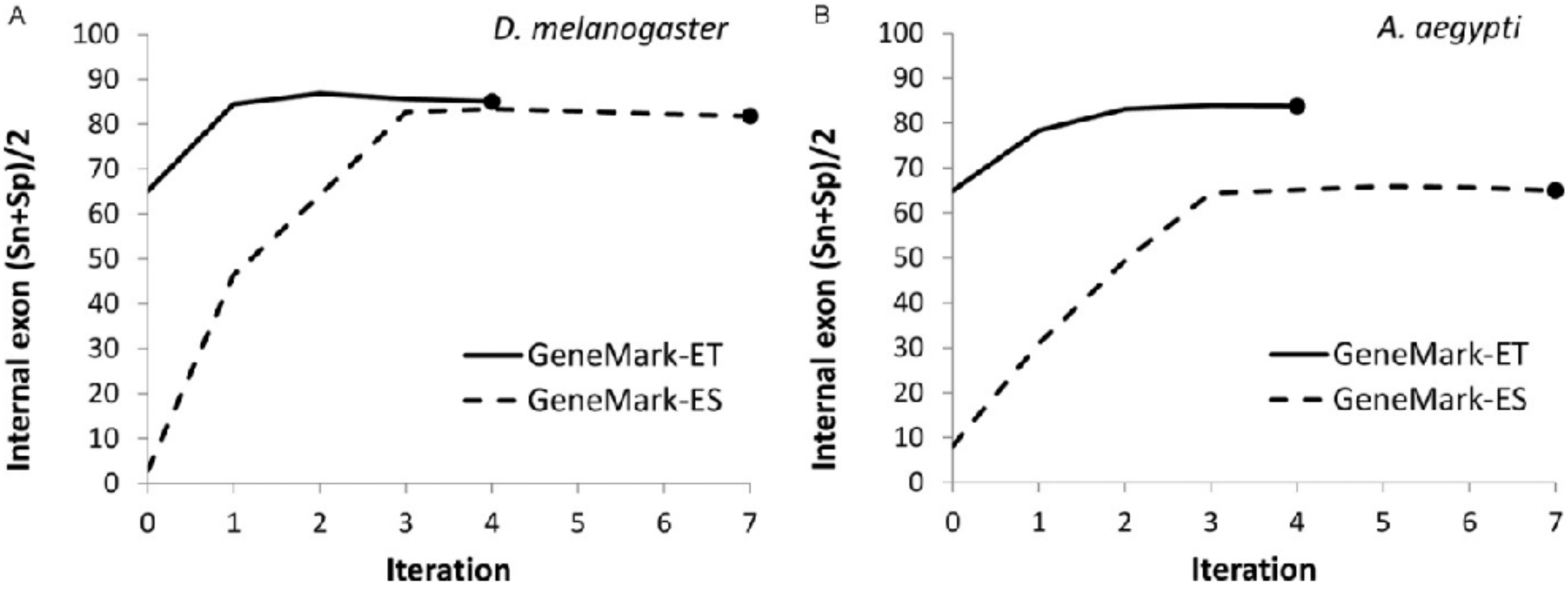


Figure 4. Observed dynamics of change in iterations of the mean of Sn and Sp internal exon prediction values for the GeneMark-ET and GeneMark-ES algorithms in cases of *Drosophila melanogaster* (A) and *Anopheles aegypti* (B) genomes.

Lomsadze et al. 2014

Interesting Observation About Exon Lengths

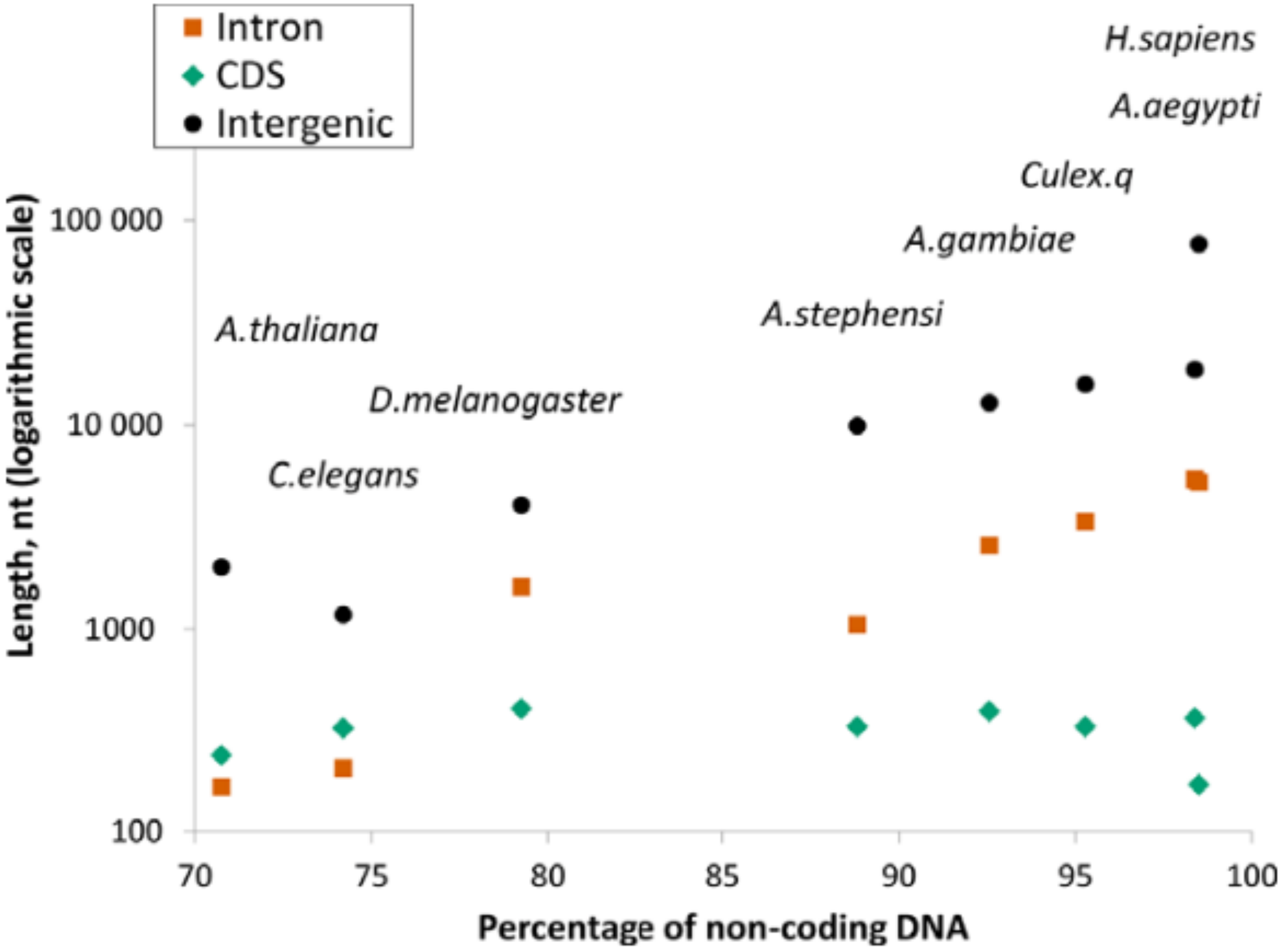


Figure 1. The dot plot graph depicting average lengths of exons, introns and intergenic regions against the value of percentage of non-coding DNA in a given genome was made for the five insect genomes used in the GeneMark-ET tests as well as for several other eukaryotic species. The average lengths of intron and intergenic regions correlate with the genome length while the average length of protein-coding exons (CDS) does not show dependence on the genome size.

Lomsadze et al. 2014

Combining Evidence from Multiple Predictions

Haas, Brian J., et al. "Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments." Genome biology 9.1 (2008): R7.

Next 6 Slides

Other Ways of Using Experimental Evidence

Experimental evidence (RNA-seq, in particular) is a great help in improving gene prediction. However, its uses stretch **far** beyond assisting *ab initio* gene prediction.

Transcript quantification

Differential expression, alternative splicing analysis

Fusion/chimera detection

Variant (SNP, SV, CNV) detection

Transcript assembly

Genome guided & de novo

Build higher-level models of transcription

co-expression networks -> regulatory networks

Other Ways of Using Experimental Evidence

Transcript quantification

How much of each gene (or transcript / isoform) is present in a particular experiment?

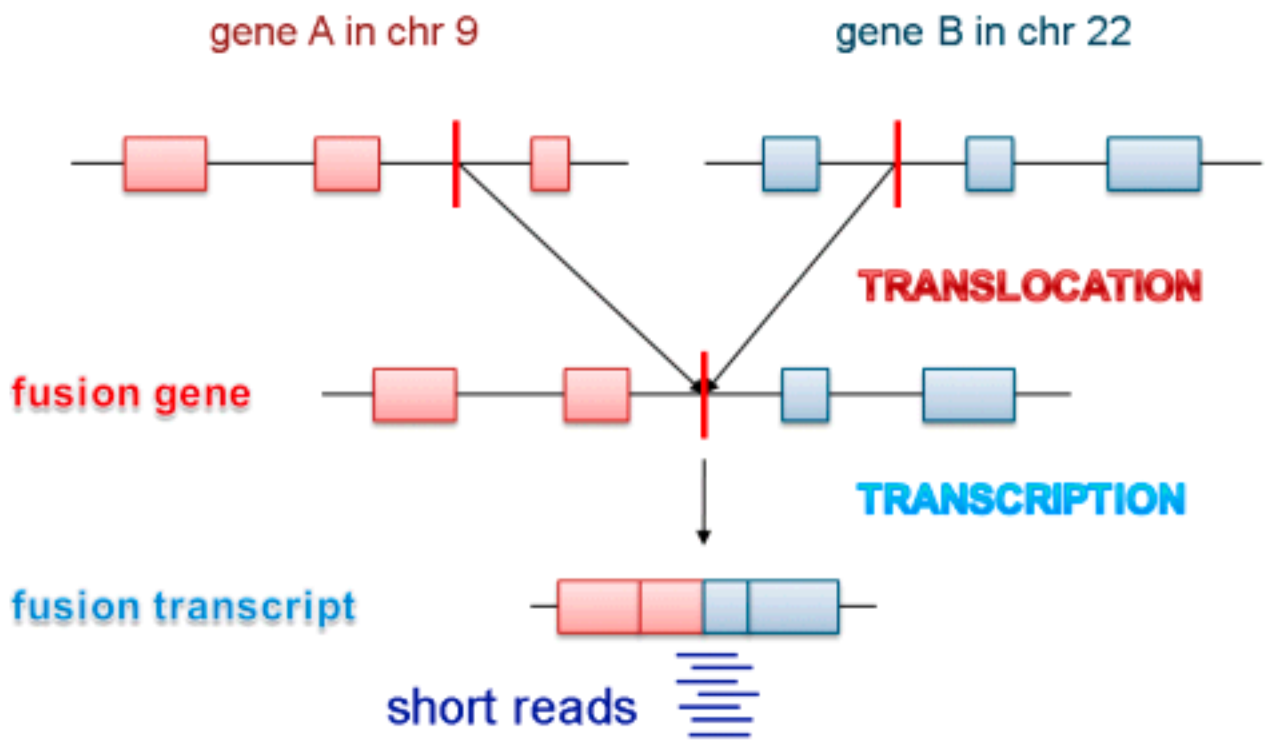
Differential expression, alternative splicing analysis

What are the *statistically significant* differences in expression or splicing between experimental conditions?

These tools can be used to study e.g. differences between healthy / diseased tissue, or how gene expression differs across tissue types.

Other Ways of Using Experimental Evidence

Fusion/chimera detection



Variant (SNP, SV, CNV) detection

Find small (SNP) or large (SV) variation in how read map back to their genes of origin

Find differences in the number of copies of a gene in the DNA (CNV)

image from: http://biome.ewha.ac.kr:8080/FusionGene/

Other Ways of Using Experimental Evidence

Transcript assembly

With sufficiently deep sequencing, we can hope to assemble transcripts present in an experiment in a manner similar to how we assemble DNA.

This often lets us find previously undiscovered genes, as well as novel splice variants (combination of exons that make up an isoform of the gene).

Genome guided and *de novo*

Assembly can either rely on knowing the reference genome (making the problem much easier), or can be done directly from the RNA-seq reads without the reference (or via hybrid approaches).

Other Ways of Using Experimental Evidence

Build higher-level models of transcription

Co-expression networks -> regulatory networks

By looking at how the expression of different genes covaries across many different experimental conditions and tissue types, we can begin to view the set of genes as a network.

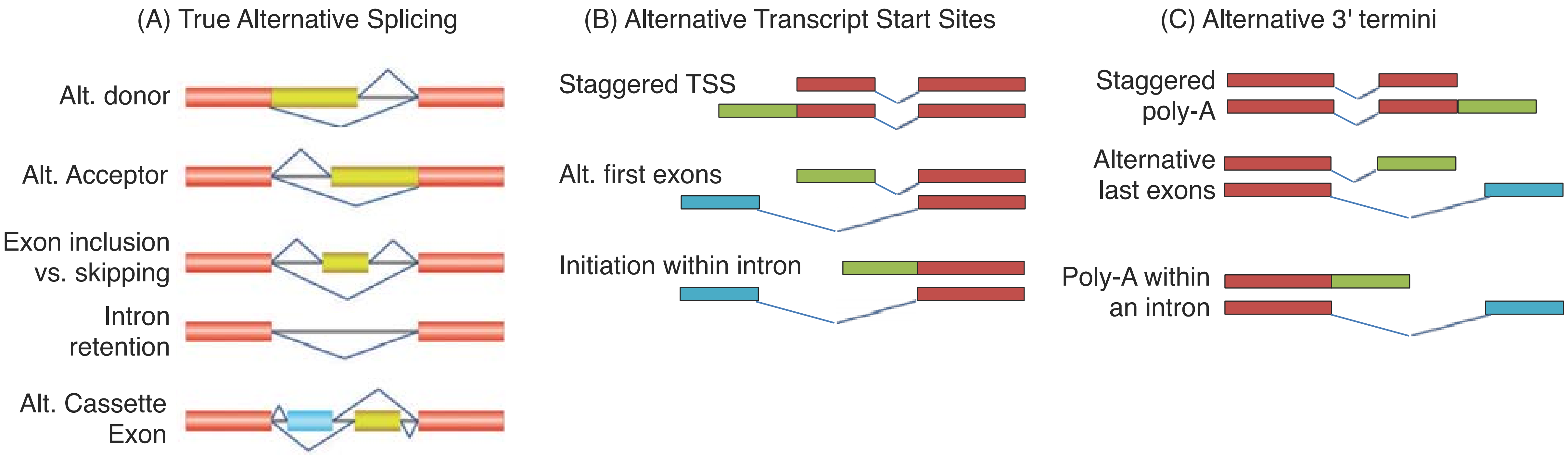
Which genes tend to be “turned on” when others are “turned off”.

Use such information to try and determine regulatory relationships between genes — which genes control others, and how.

Genome-guided Assembly, an Example

We'll take a look at how Cufflinks, one of the most popular tools for RNA-seq-based transcript discovery & quantification assembles transcripts from data.

Transcript (rather than just "gene") discovery lets us explore the different variants of a gene that are actually expressed in a cell.



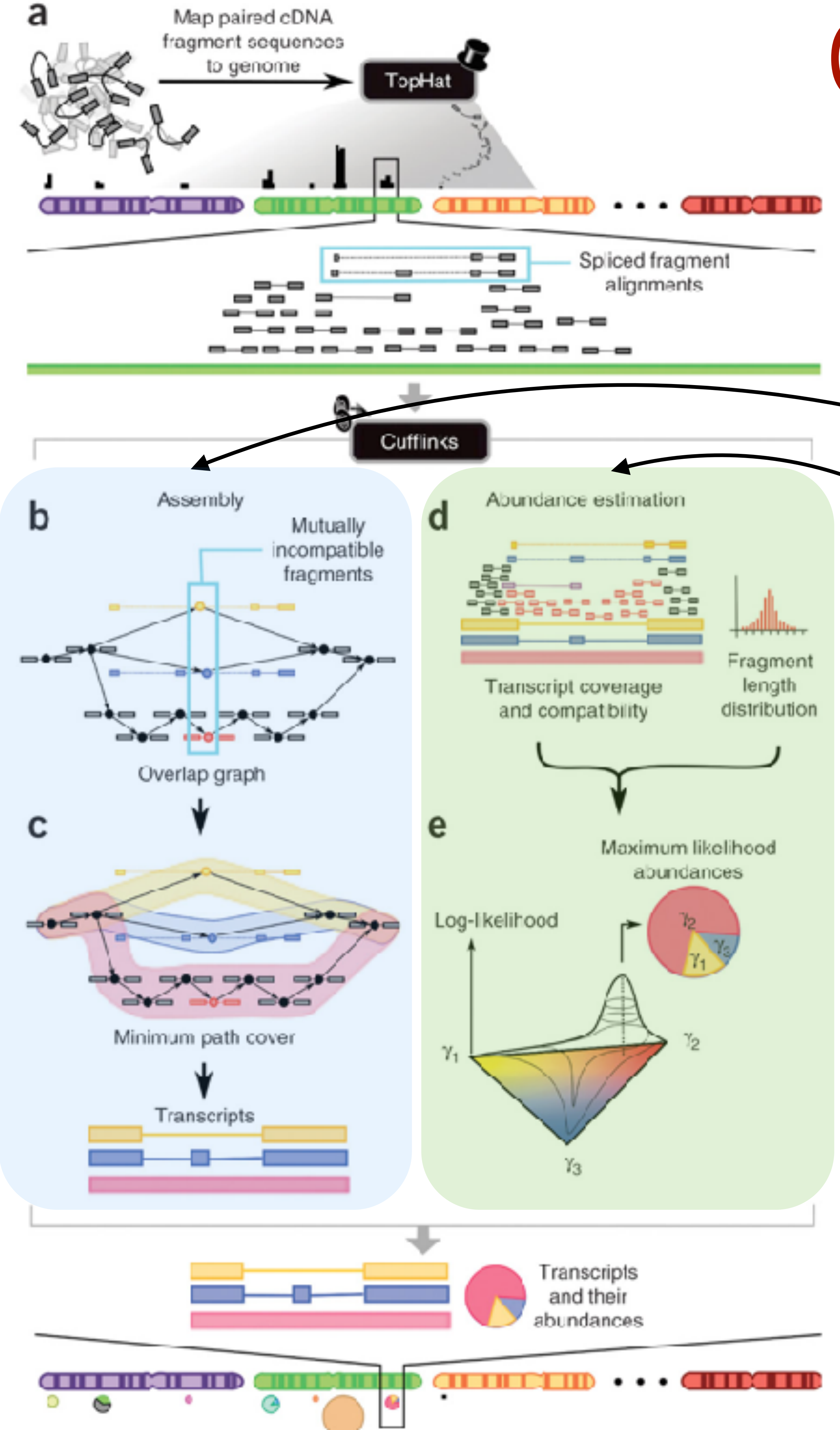
Caveat: Cufflinks solves the identification and quantification in largely separate phases. This turns out to discard a lot of useful information. Newer approaches attempt to combine these two phases, which results in both better identification and better quantification. Still, Cufflinks is an incredibly useful (and widely used) tool (cited ~4,500 times).

Cufflinks Transcript Assembly

Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter.· Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat Biotechnol,* 28(5): 511–515 (2010)

* Slide from Carl Kingsford

Cufflinks Pipeline


a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ_1
γ_2
γ_3
Transcripts and their abundances

Alignment

Assembly

Quantification

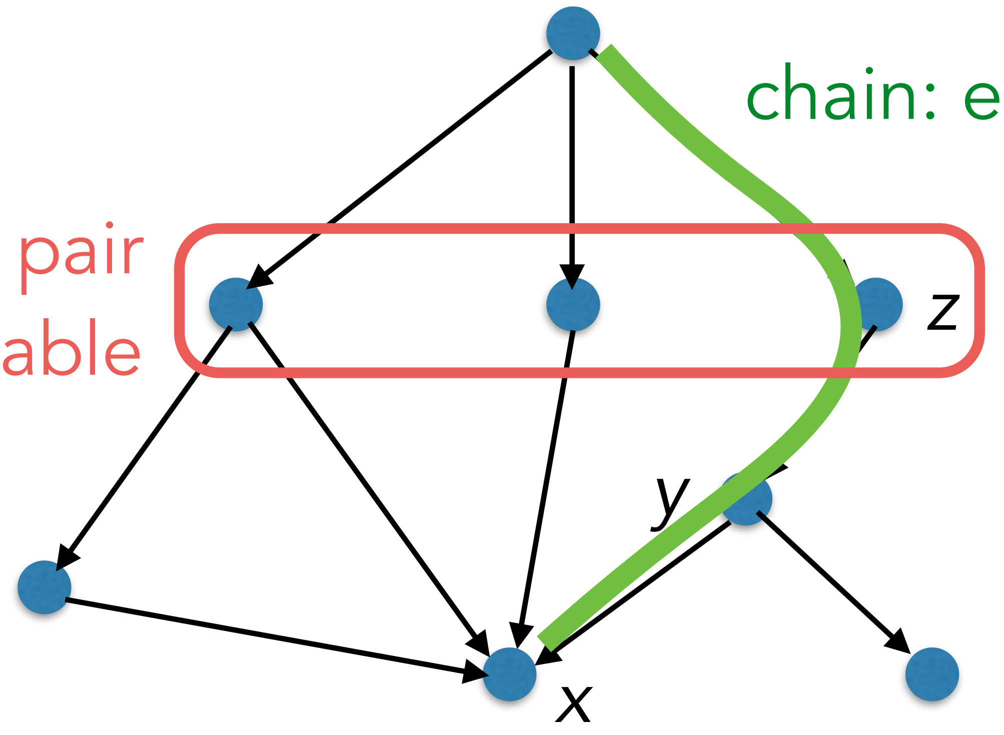
Partially Ordered Sets

Def. A pair (S, ≤) is a partial order if, for all x, y ∈ S:

(transitivity) $x \leq y$ and $y \leq z \Rightarrow x \leq z$

(reflexivity) $x \leq x$

(antisymmetry) $x \leq y$ and $y \leq x \Rightarrow x = y$



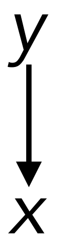
chain: every pair is comparable

antichain: every pair is incomparable

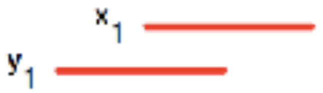
* Slide from Carl Kingsford

Cufflink's Partial Order

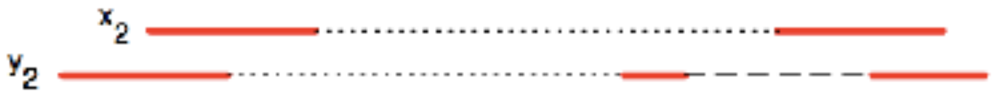
● = sequenced fragment:



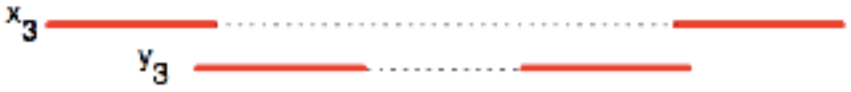
= x aligns to the left of y and x and y have compatible intron structure



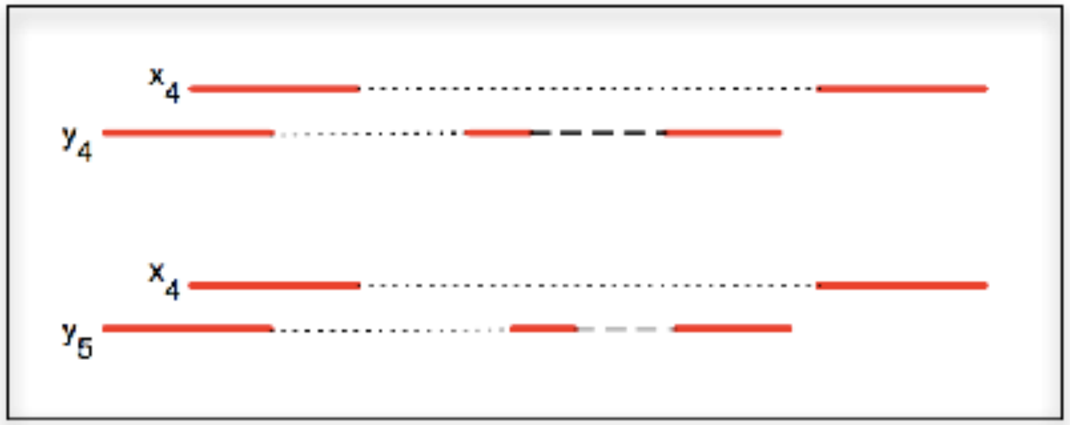
$y_1 \leq x_1$



incompatible b/c the right end of y_2 is split-mapped, implying an intron where there is no intron in x_2.



x_3 and y_3 are nested, and so are merged into a single fragment.



x_4 is uncertain because it could be compatible with either y_4 or y_5; x_4 is therefore thrown away.

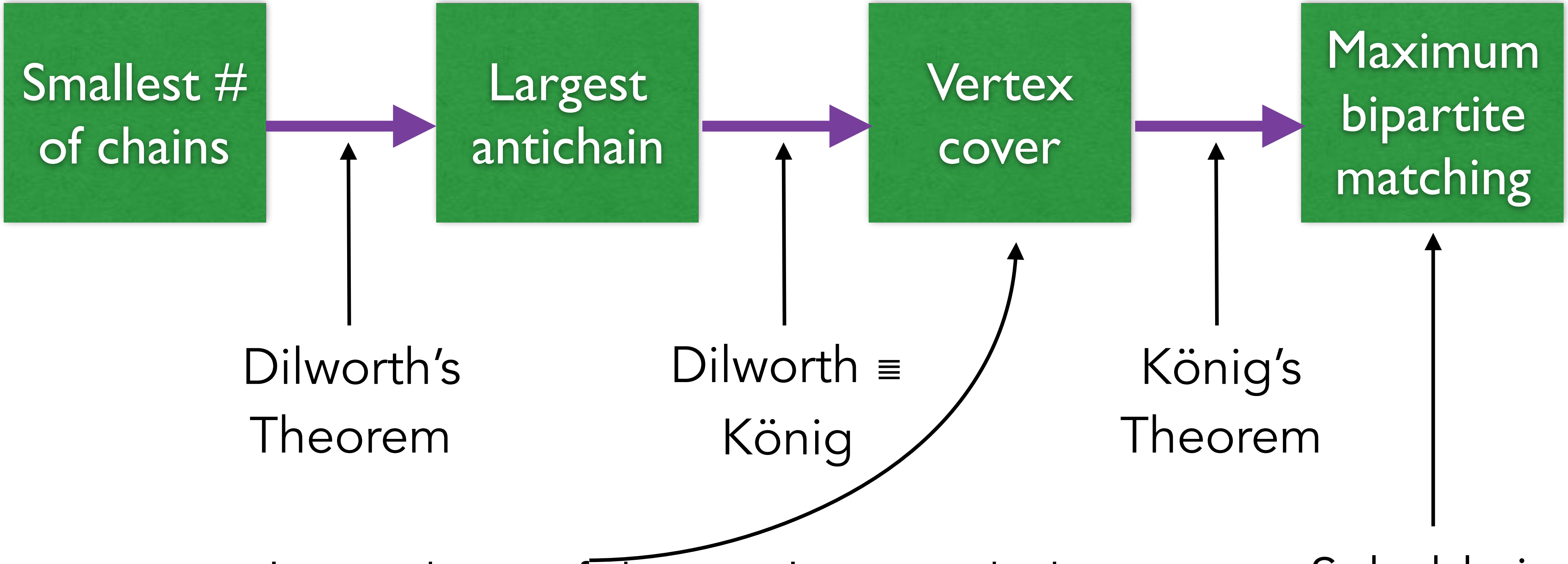
(Trapnell et al., 2010)

* Slide from Carl Kingsford

Cufflinks' Assembly Algorithm

(covering)

Partitioning partial order into smallest # of chains → "parsimonious" set of transcripts that explains the observed reads



Smallest # of chains

Largest antichain

Vertex cover

Maximum bipartite matching

Dilworth's Theorem

Dilworth ≡ König

König's Theorem

A vertex cover is a subset of the vertices such that each edge is adjacent to at least one vertex from the subset.

Solvable in $O(E\sqrt{V})$

* Slide from Carl Kingsford

Dilworth's Theorem

Thm (Dilworth). In a poset, the size of the largest antichain = the size of the minimum cover by chains.

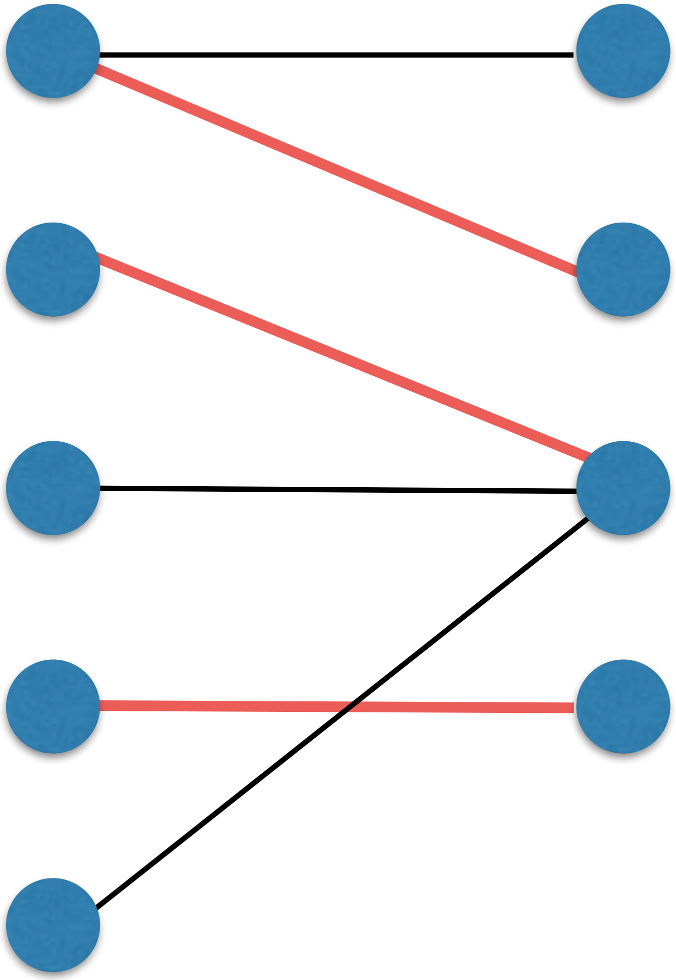
Proof intuition.

- The largest antichain must hit every chain (otherwise it could be made larger).
- It can't hit any chain twice, otherwise it would contain two comparable items.

* Slide from Carl Kingsford

König's Theorem

Thm (König). In a bipartite graph, the # of edges in a maximum matching = # of vertices in the smallest vertex cover.

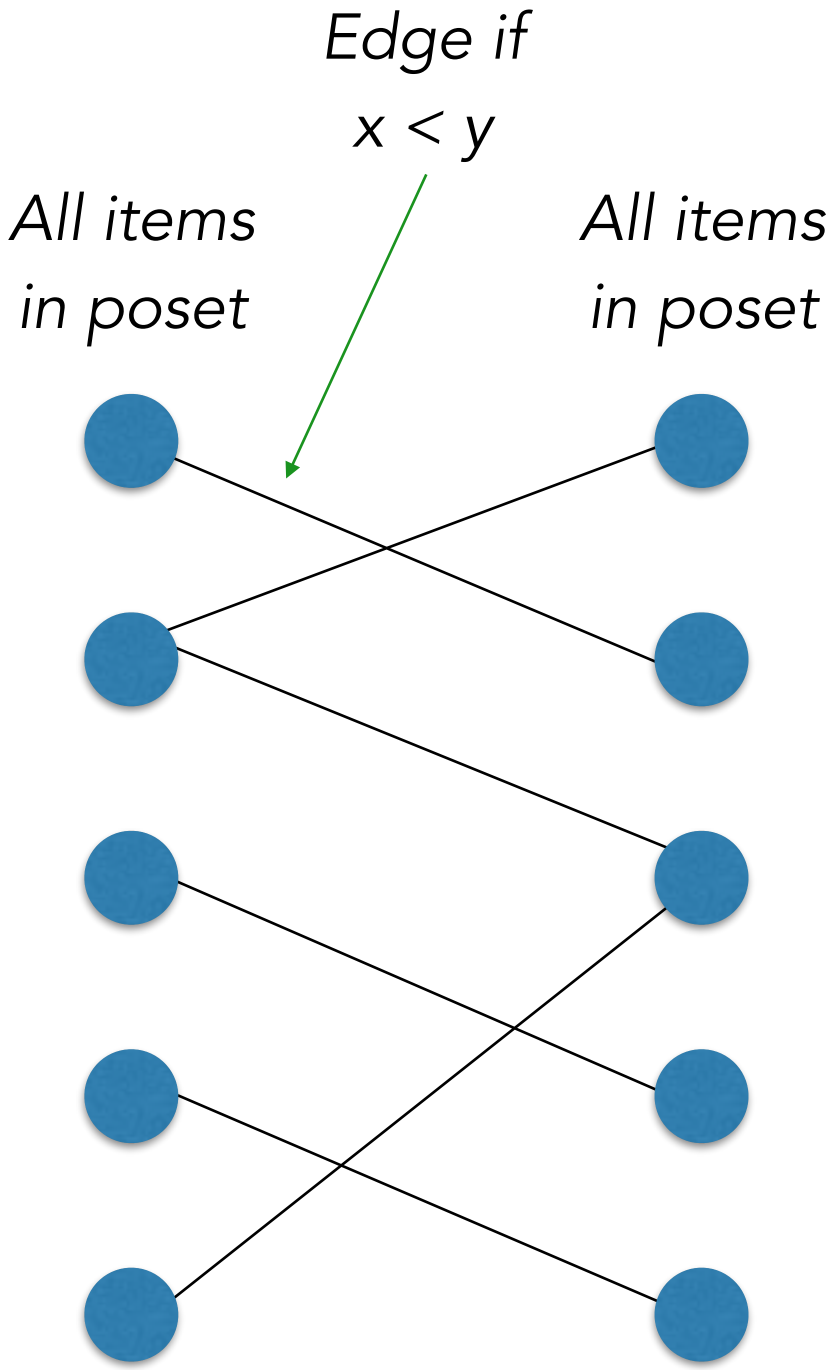


Proof intuition.

- In a maximum matching, every edge must be covered.
- Otherwise, if both endpoints are not matched, we could add that edge to the matching and increase its size.

* Slide from Carl Kingsford

Using Matching to Find a Minimal Chain Cover



Let M be the maximal matching.

By König's theorem, there is a (minimal) vertex cover C of the same size as M.

Let T be the elements of the poset that are not in C.

T is an antichain. Why?

If u and v were comparable, there would be an edge between them, and since neither u or v was in M, we could add that edge to M.

Make a set W of chains by u ≡ v if (u,v) ∈ M.

These equivalence classes are chains. Why?

Every pair of items in each equivalence class had an edge between them, meaning they were comparable.

* Slide from Carl Kingsford

|W| = |T|

M = maximal matching.

C = vertex cover of the same size as M.

T = antichain elements of poset that are not in C.

W = set of chains formed from edges of M.

n = # elements in poset

m = # of edges in matching

Size of *T* is n - m. Why?

Every edge uses up exactly one element on the LHS of the bipartite graph.

Size of *W* is n - m. Why?

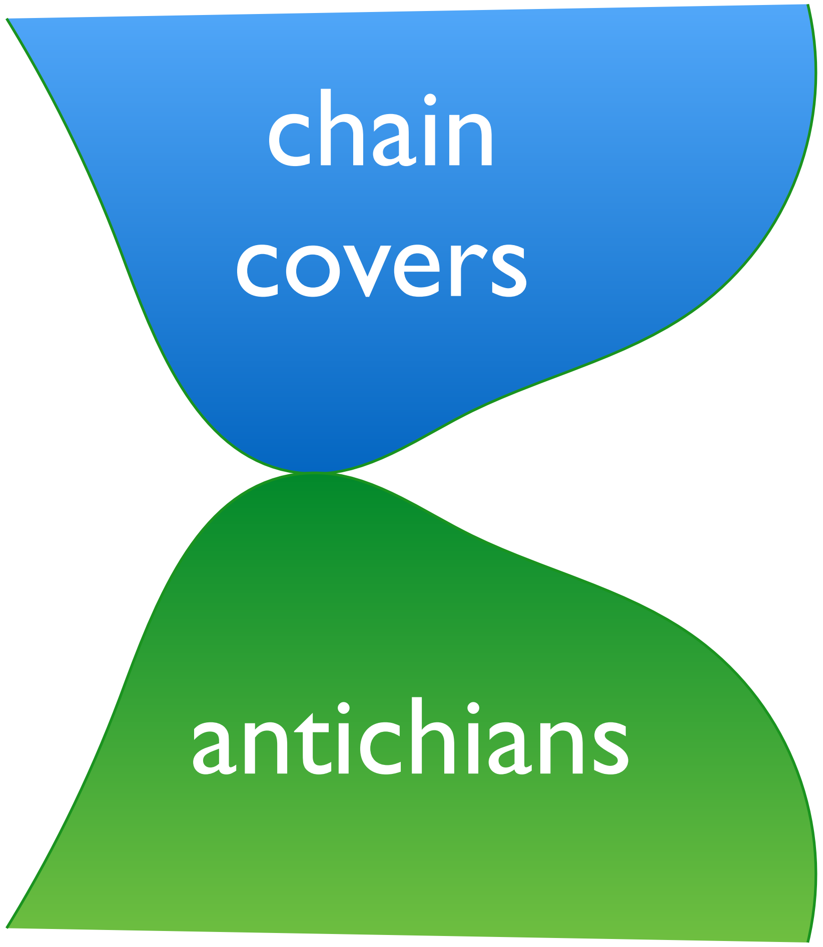
Consider set of n “chains” each consisting of a single element of poset.

Each edge (u,v) that we use to put v into the same poset as u reduces the number of chains by 1.

⇒ Number of equivalence-class chains = n - m

* Slide from Carl Kingsford

Why is W Minimum Size?



All antichains must be of size ≤ all chain covers.

Suppose not, and let A be an antichain bigger than cover Q.

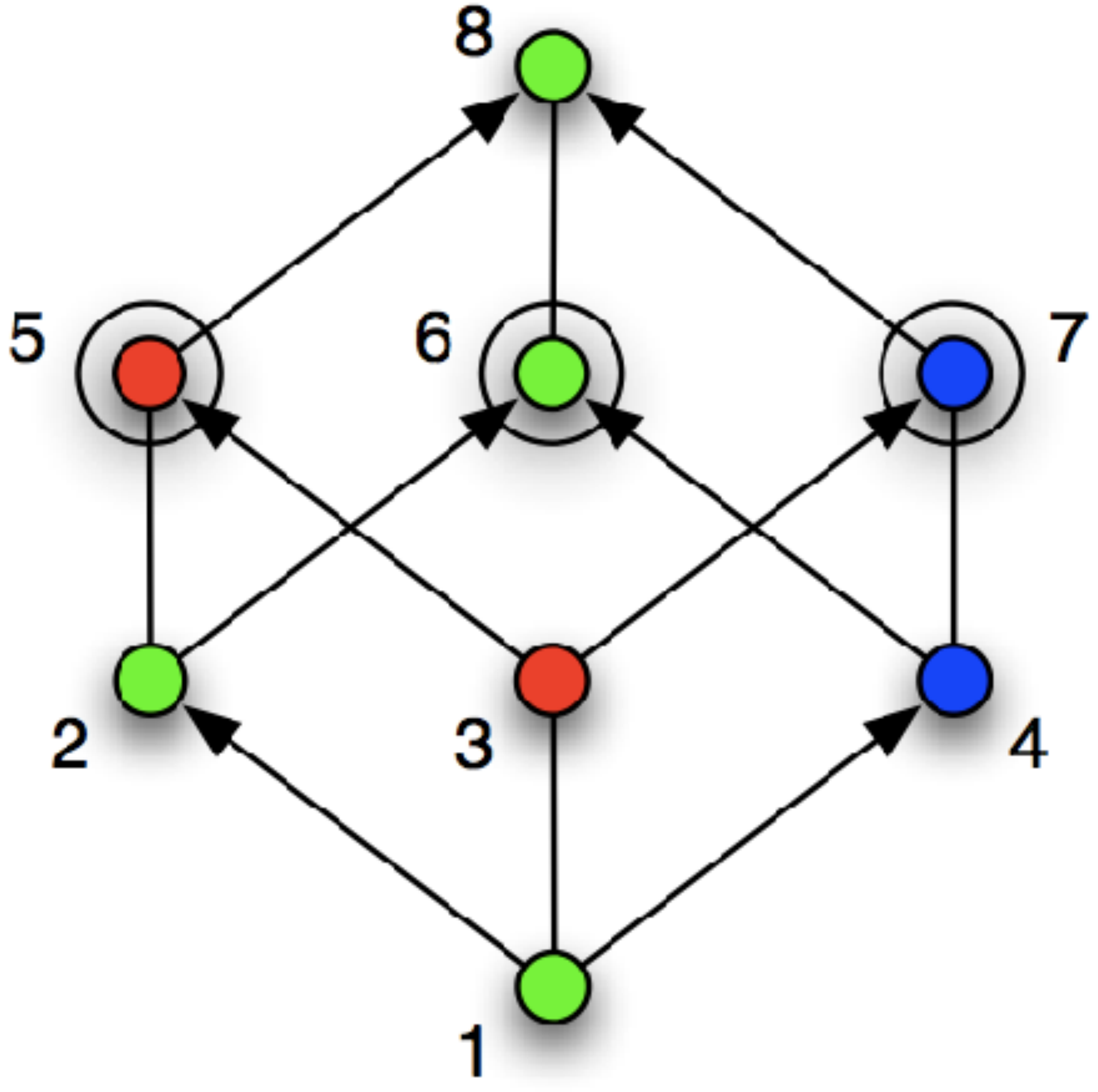
Then, by pigeonhole, A must contain at least 2 elements x, y from the same chain in Q.

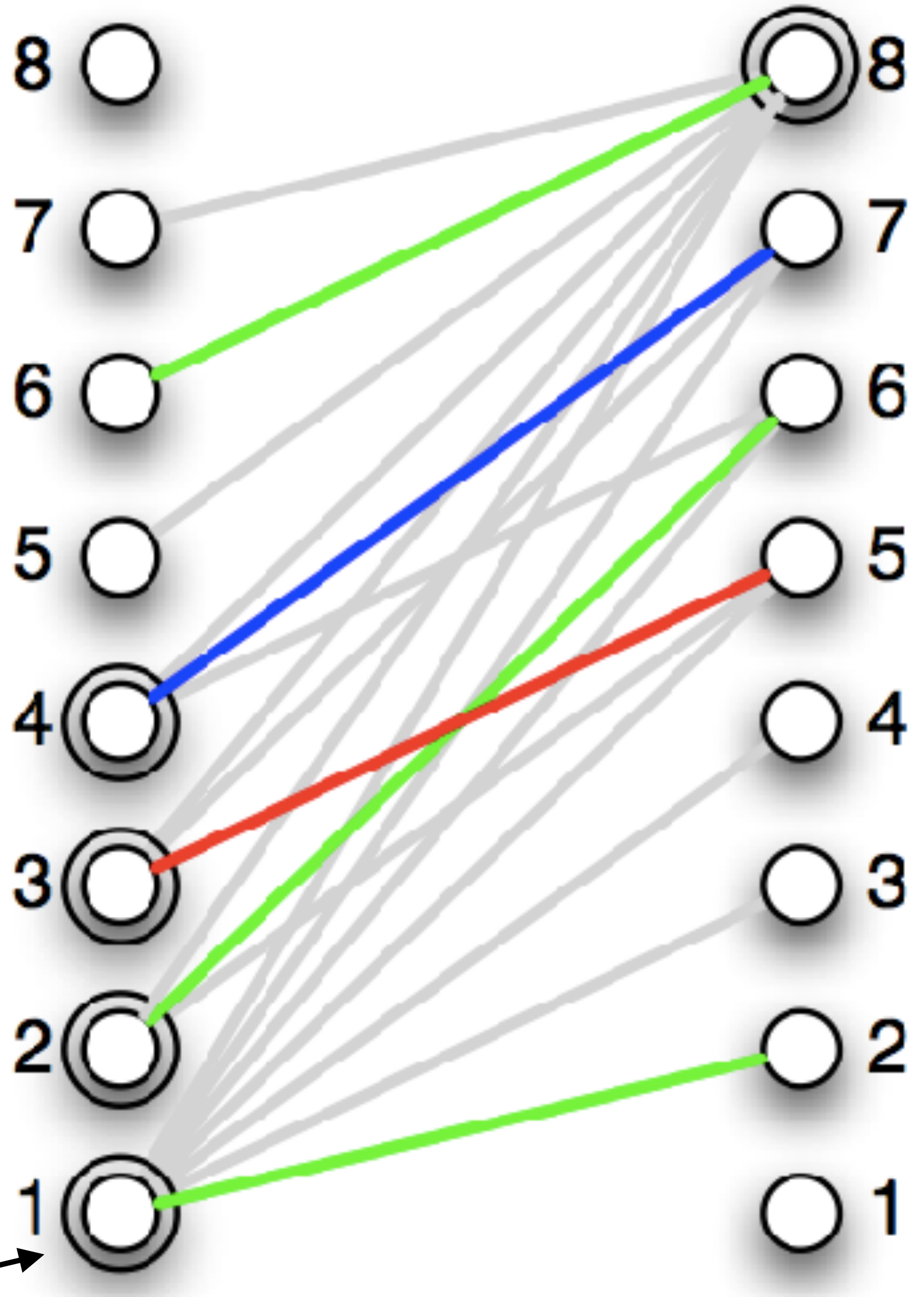
But x, y are comparable because they are in the same chain.

⇒ the pair (T,W) must be a largest antichain and a smallest W because they are the same size.

* Slide from Carl Kingsford

A Matching-Covering Example



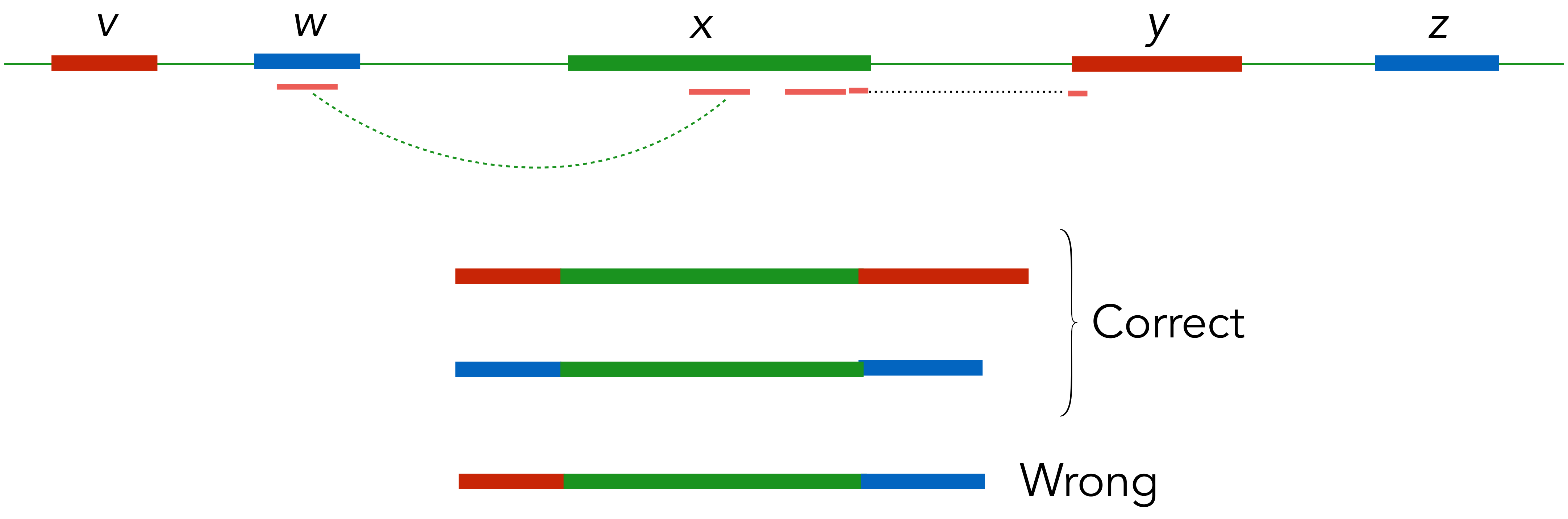


Double circles denote vertex cover

(Trapnell et al., 2010)

* Slide from Carl Kingsford

Selecting From Among Many Minimum Solutions



Idea: exons included in same transcript should have similar expression

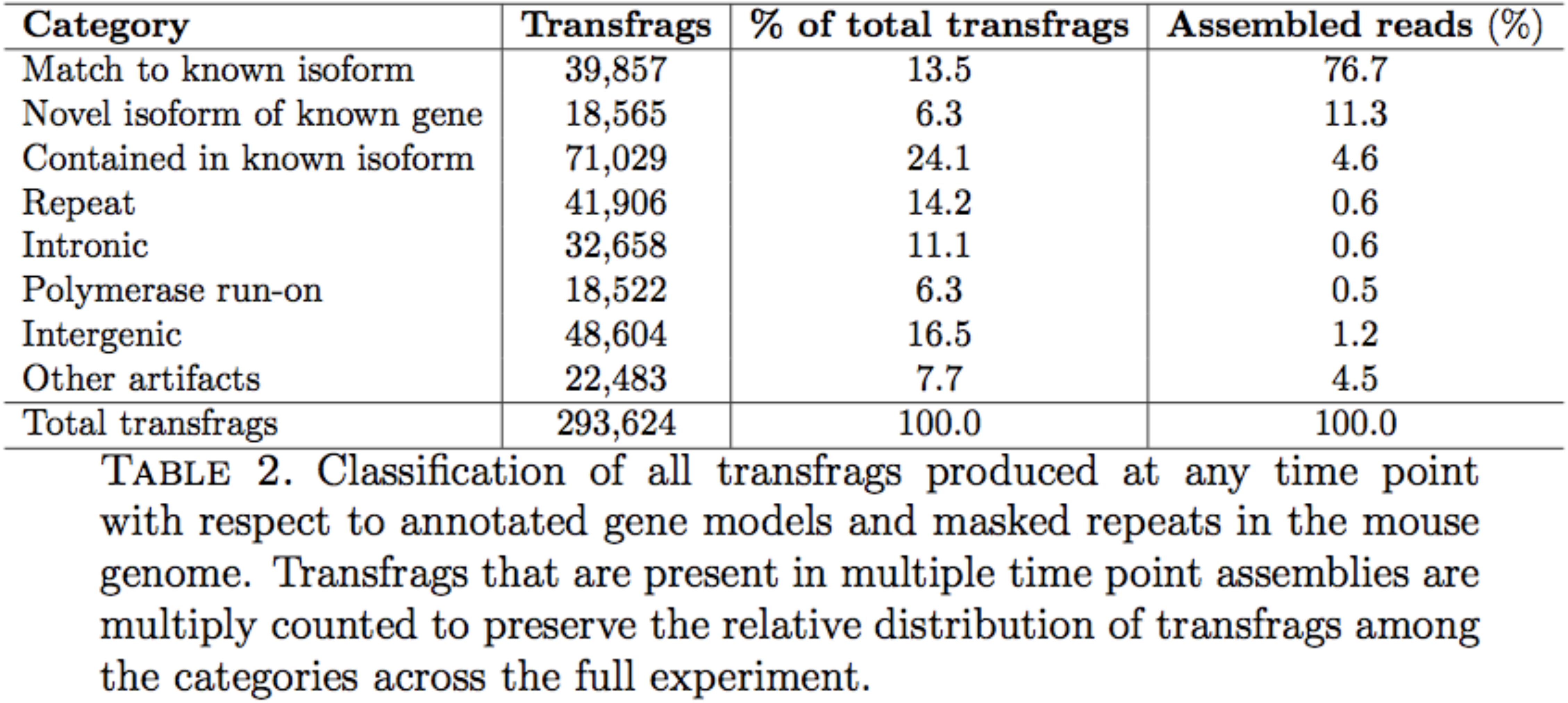
Estimate … that are compatib… by length of x).

measures how similar the exons' PSI values are

$$\text{weight}(x, y) = -\log(1 - |\psi_x - \psi_y|)$$

* Slide from Carl Kingsford

Discovery of Novel Isoforms



Category	Transfrags	% of total transfrags	Assembled reads (%)
Match to known isoform	39,857	13.5	76.7
Novel isoform of known gene	18,565	6.3	11.3
Contained in known isoform	71,029	24.1	4.6
Repeat	41,906	14.2	0.6
Intronic	32,658	11.1	0.6
Polymerase run-on	18,522	6.3	0.5
Intergenic	48,604	16.5	1.2
Other artifacts	22,483	7.7	4.5
Total transfrags	293,624	100.0	100.0

TABLE 2. Classification of all transfrags produced at any time point with respect to annotated gene models and masked repeats in the mouse genome. Transfrags that are present in multiple time point assemblies are multiply counted to preserve the relative distribution of transfrags among the categories across the full experiment.

(Trapnell et al., 2010)

* Slide from Carl Kingsford