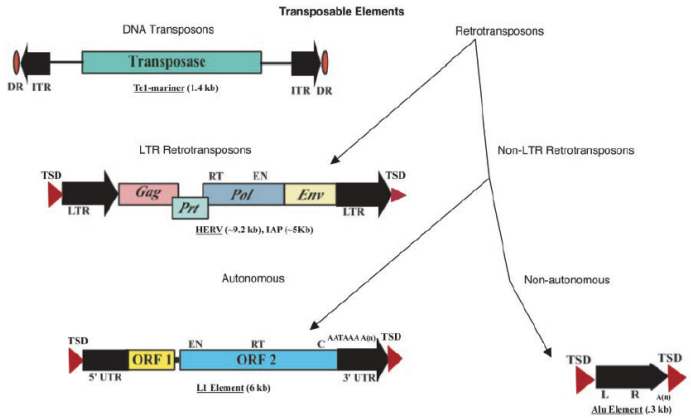
Genome Annotation

Stefan Prost [1]

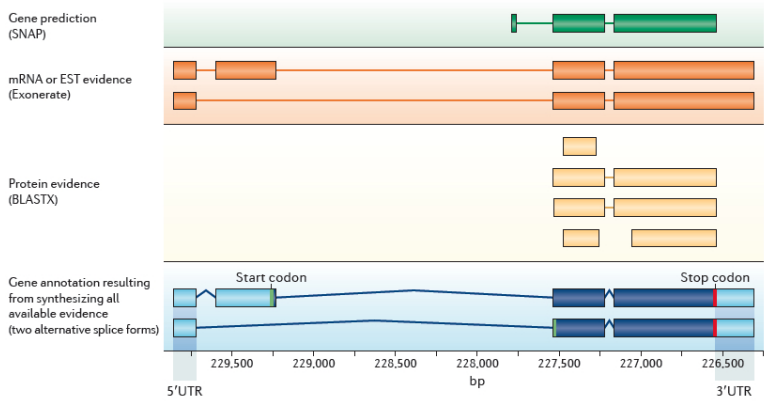
[1]Department of Integrative Biology, University of California, Berkeley, United States of America

May 27th, 2015

1 Repeat Annotation



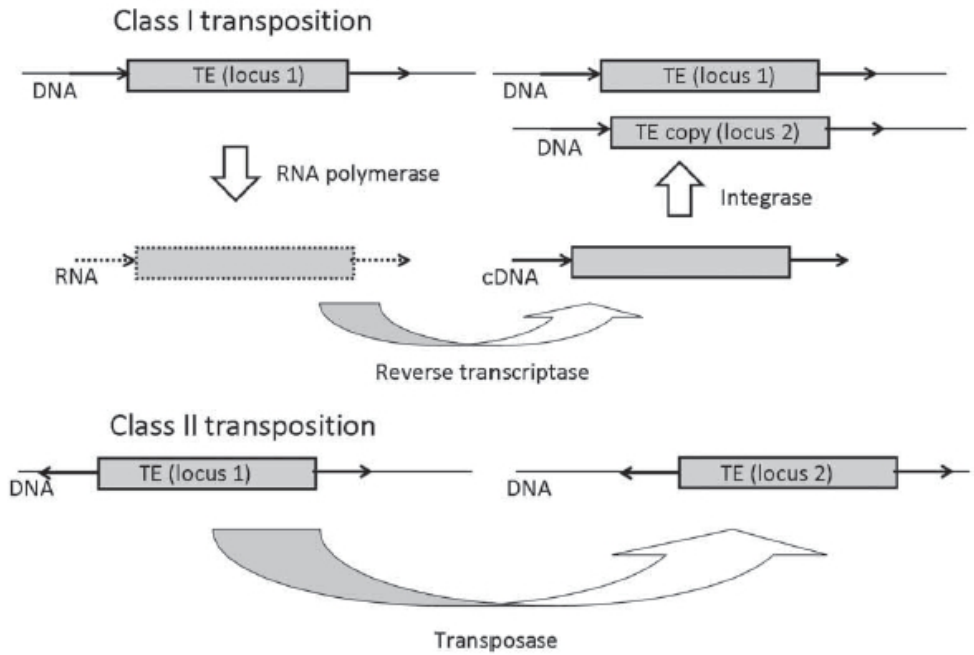
2 Gene Annotation



Repeat Annotation

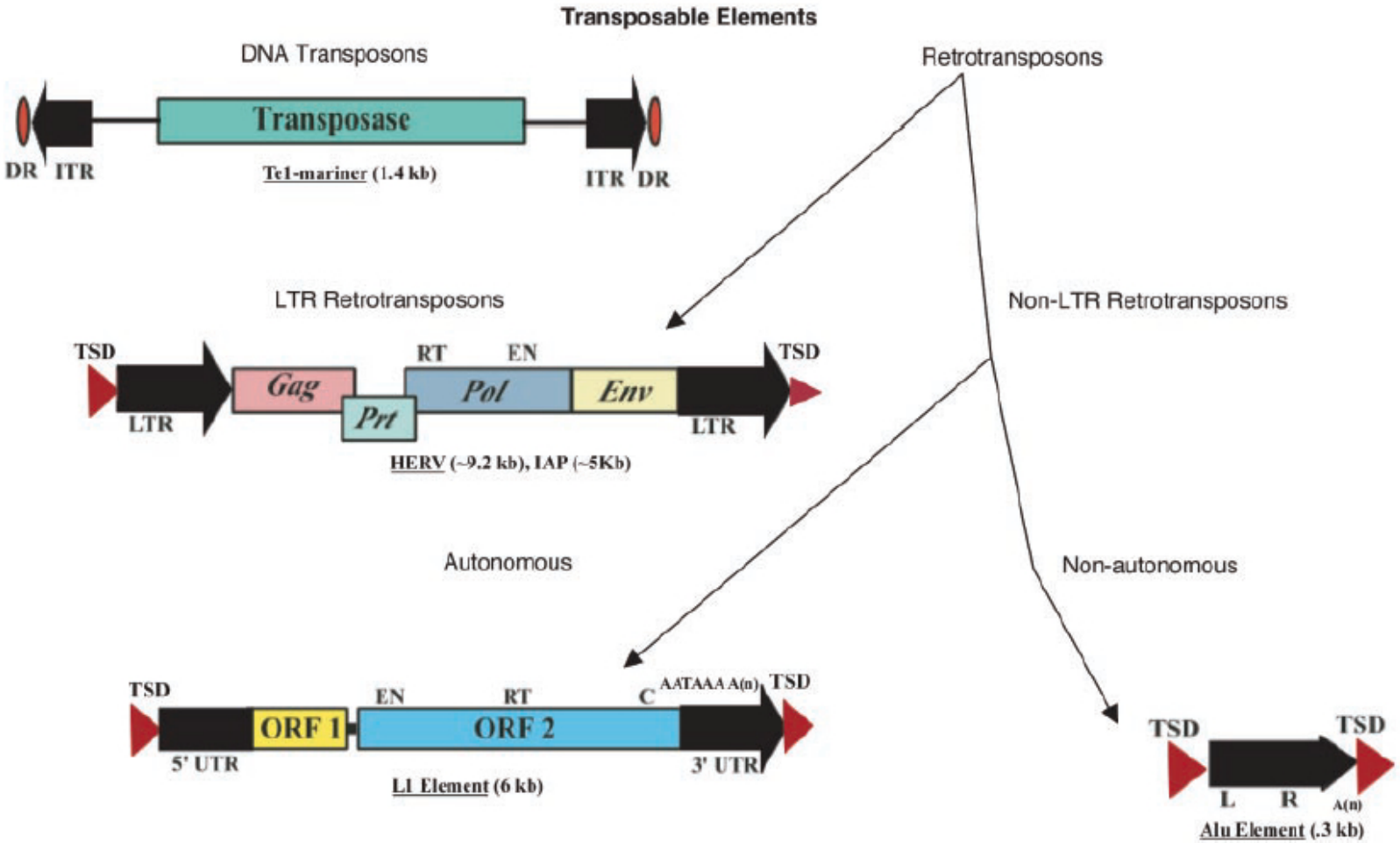
1. Mask Repeats for Gene Annotation
2. Study Repeat Content and Evolution

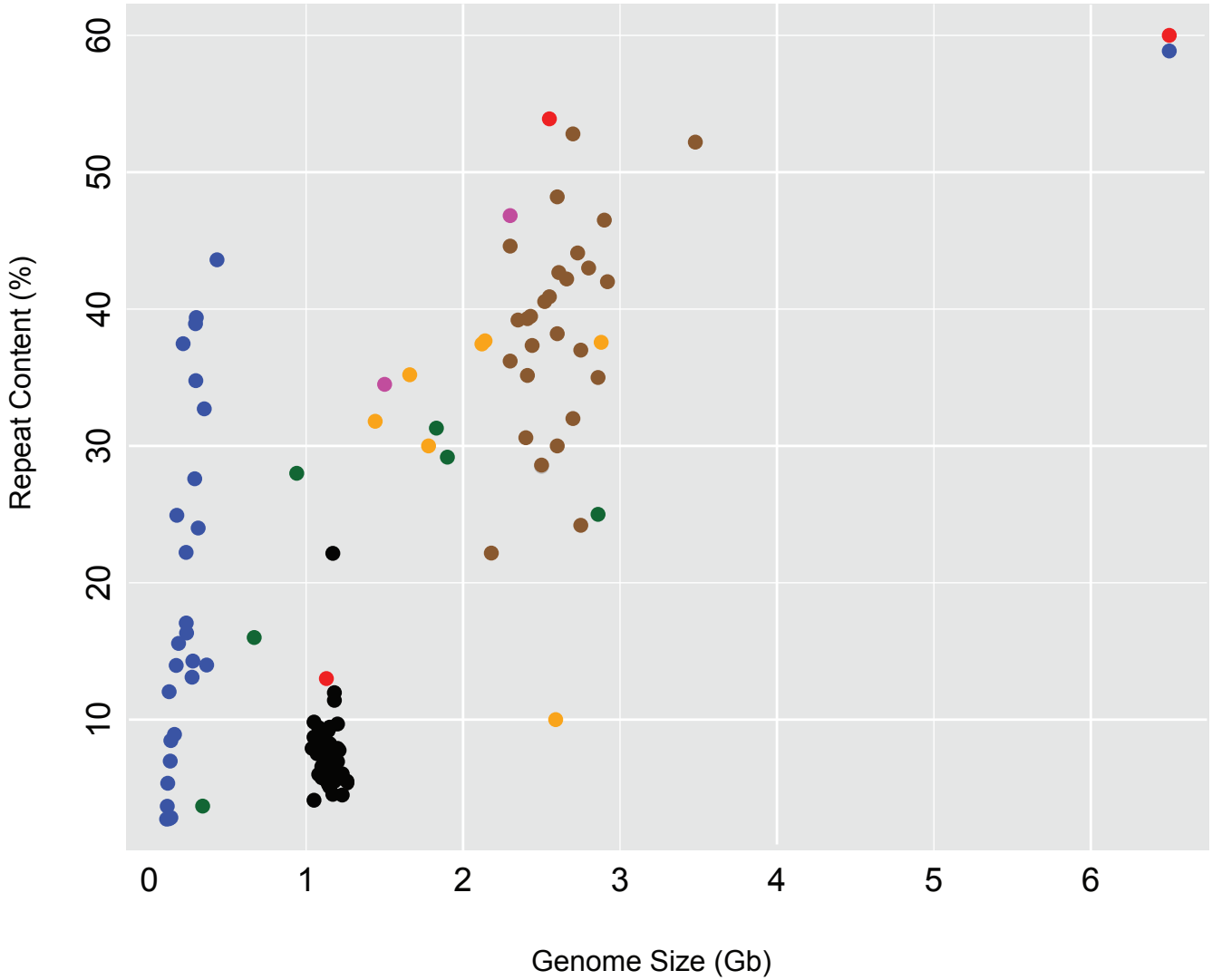
- **Low Complexity Sequences**
  - Simple Repeats and Satellites
- **Transposable Elements**
  - Class I: Retrotransposons *Copy and Paste*
  - Class II: DNA Transposons *Cut and Paste*



Hermann et al. 2013

- **Class I TEs**
  - Long Terminal Repeats (LTR)
    - Endogenous Retrovirus
  - Non-LTR
    - Long Interspersed Nuclear Elements (LINEs)
    - Short Interspersed Nuclear Elements (SINEs)
- **Class II TEs**
  - Subclass I (both DNA strands are cleaved)
  - Subclass II (Rolling-circle, self-synthesizing)


Transposable Elements
DNA Transposons
Transposase
DR ITR
Tc1-mariner (1.4 kb)
ITR DR
Retrotransposons
LTR Retrotransposons
TSD
LTR
Gag
Prt
RT
EN
Pol
Env
TSD
LTR
HERV (~9.2 kb), IAP (~5Kb)
Non-LTR Retrotransposons
Autonomous
Non-autonomous
TSD
5' UTR
ORF 1
EN
RT
ORF 2
C
AATAAA A(n)
TSD
3' UTR
L1 Element (6 kb)
TSD
L
R
A(n)
TSD
Alu Element (.3 kb)


Repeat Content (%)
10
20
30
40
50
60
0
1
2
3
4
5
6
Genome Size (Gb)

- **Homology**
  - RepeatMasker (http://www.repeatmasker.org/)
- **De novo**
  - RepeatModeler (http://www.repeatmasker.org/)
  - WindowMasker (Morgulis et al., 2005)
  - RepeatScout (Price et al., 2005)
  - Piler (Edgar and Myers, 2005)
- **De novo from reads**
  - REPdenovo (github)
  - Tedna (Zytnicki et al., 2014)

CAVE: De novo tools can wrongly identify highly conserved protein-coding genes

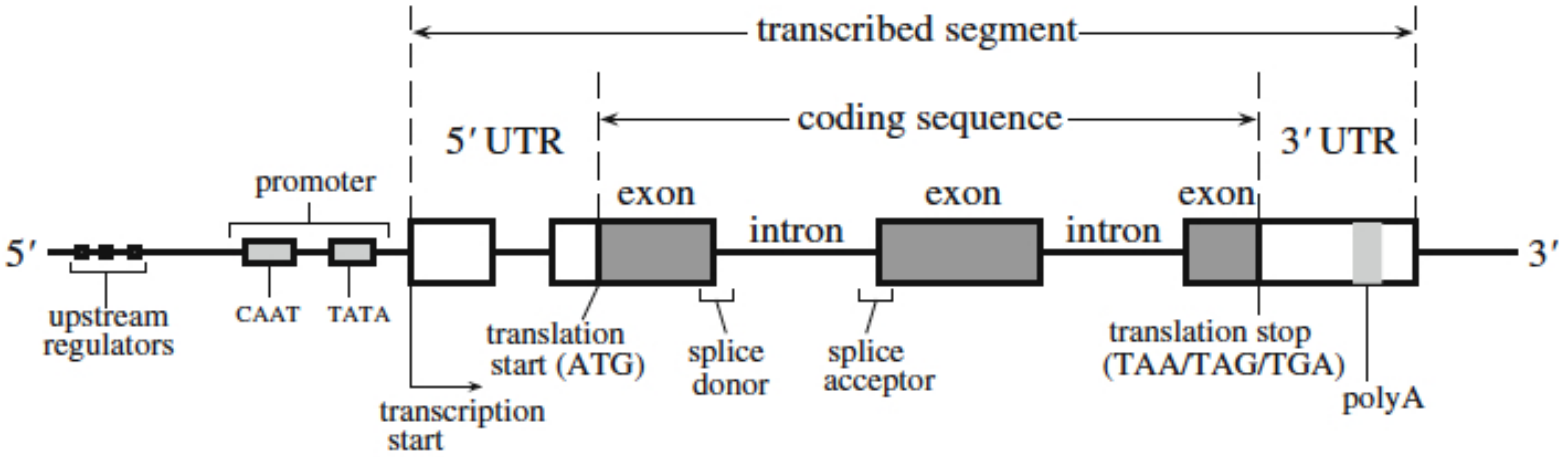
For Gene Annotation

1. RepeatMasker -pa xx -gccalc -nolow -species aves genome.fasta
2. BuildDatabase -name genome genome.fasta.masked
3. RepeatModeler -database genome
4. RepeatMasker -pa xx -gccalc -nolow -lib consensi.fa.classified genome.fasta[.masked]

For Repeat Annotation

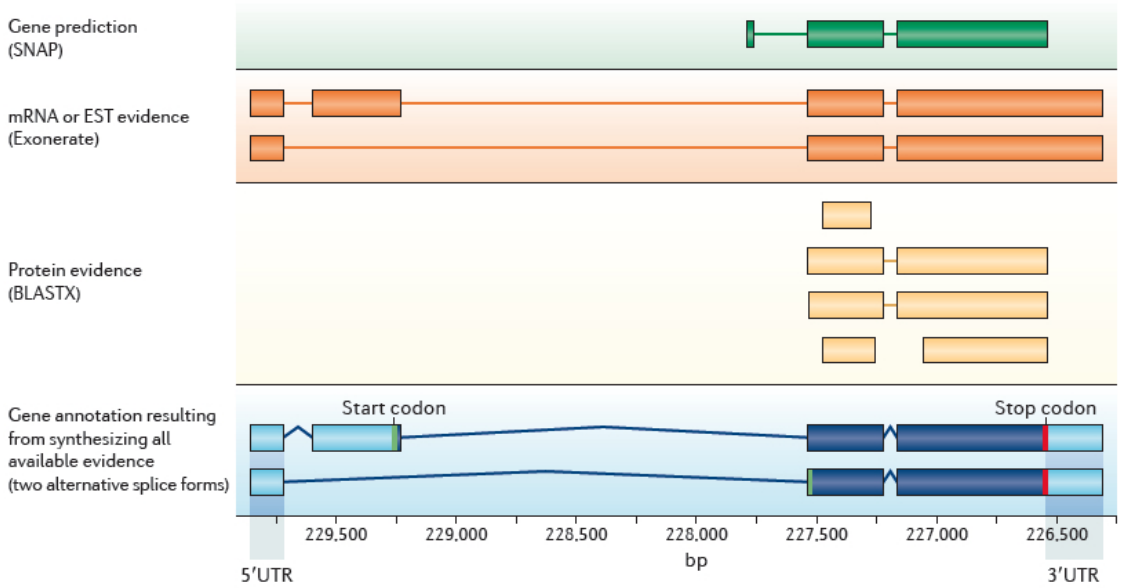
1. RepeatMasker -pa xx -a -gccalc -species aves genome.fasta
2. RepeatMasker -pa xx -a -gccalc -lib consensi.fa.classified genome.fasta[.masked]

Gene Structure



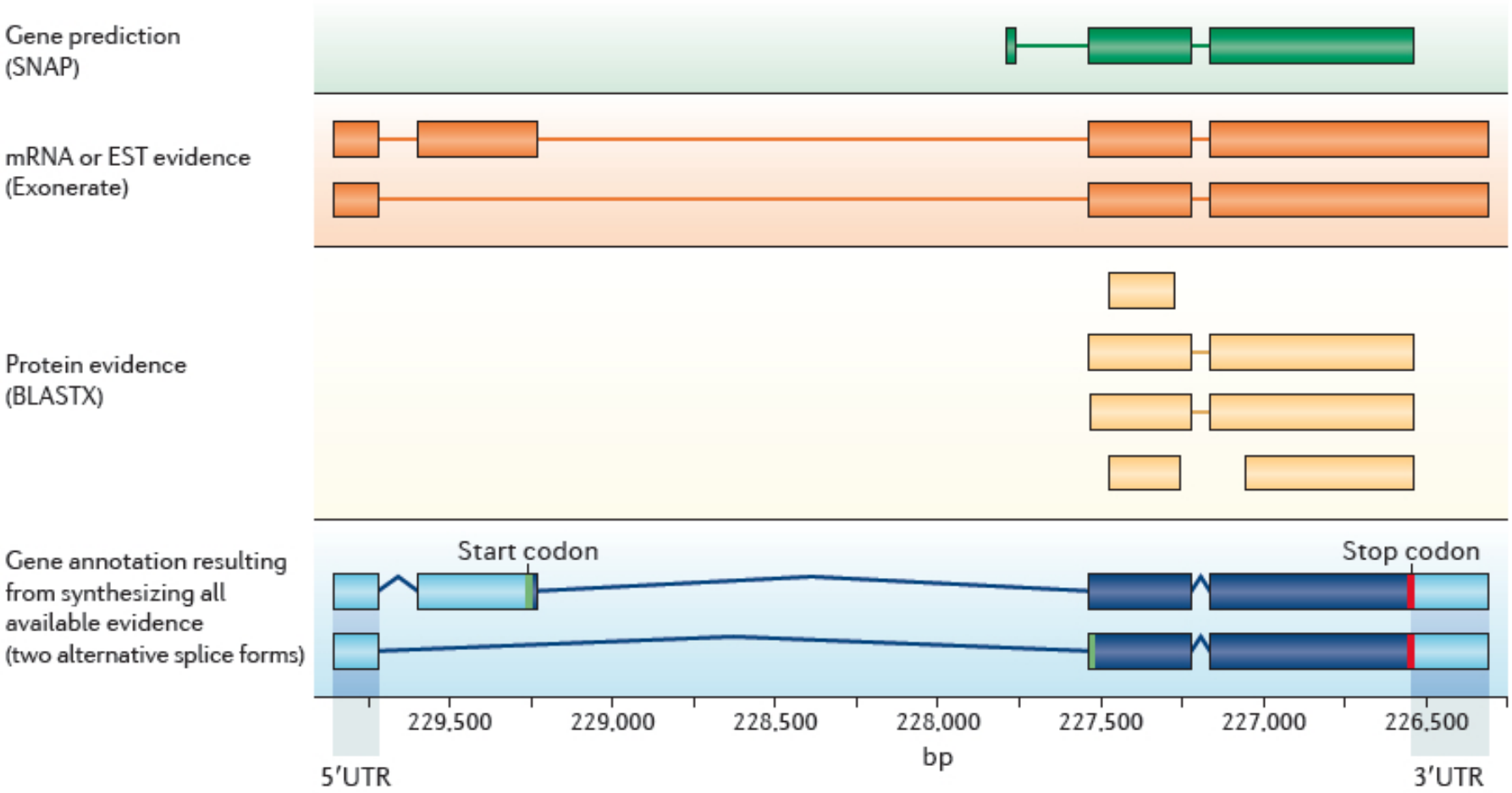
(Marina Axelson-Fisk, Springer-Verlag London, 2015)

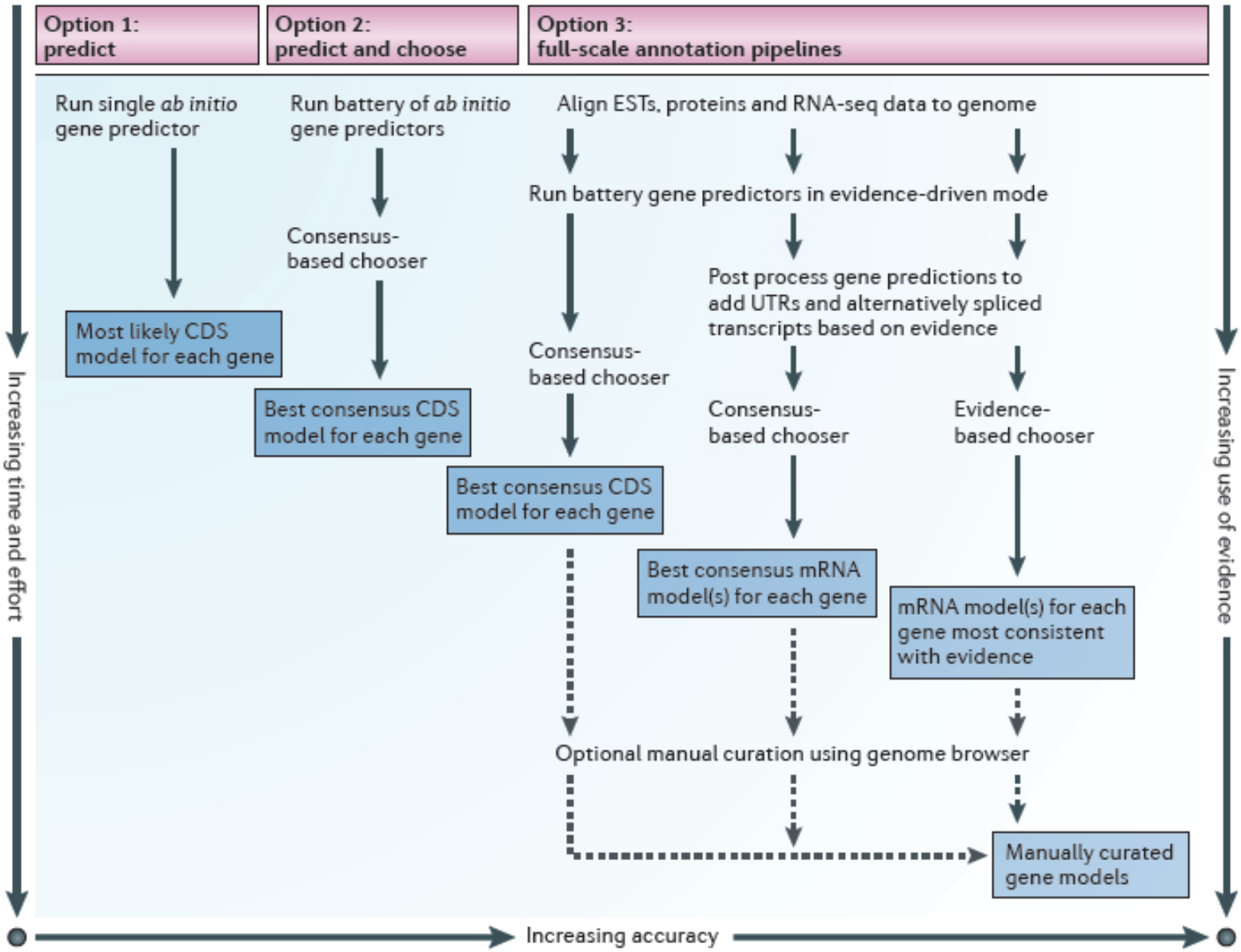
1. **Evidence Alignment**
   - Protein
   - EST
   - RNA-seq
2. **Ab initio Prediction**



Ab Initio Gene Prediction

- Do not need evidence data
- Need to be trained for the organism (codon frequencies, distribution of exon-intron length, etc.)
- Most find single most likely coding sequence (CDS)
- Do not report untranslated regions (UTRs)
- Cannot deal with alternative splicing
- Accuracy usually 60-70%
- With training can be up to 100%
- Need a very good genome assembly


Gene prediction
(SNAP)
mRNA or EST evidence
(Exonerate)
Protein evidence
(BLASTX)
Gene annotation resulting
from synthesizing all
available evidence
(two alternative splice forms)
Start codon
Stop codon
229,500
229,000
228,500
228,000
227,500
227,000
226,500
bp
5′UTR
3′UTR


Option 1: predict
Option 2: predict and choose
Option 3: full-scale annotation pipelines
Run single ab initio gene predictor
Run battery of ab initio gene predictors
Align ESTs, proteins and RNA-seq data to genome
Run battery gene predictors in evidence-driven mode
Consensus-based chooser
Post process gene predictions to add UTRs and alternatively spliced transcripts based on evidence
Most likely CDS model for each gene
Consensus-based chooser
Consensus-based chooser
Evidence-based chooser
Best consensus CDS model for each gene
Best consensus CDS model for each gene
Best consensus mRNA model(s) for each gene
mRNA model(s) for each gene most consistent with evidence
Optional manual curation using genome browser
Manually curated gene models
Increasing time and effort
Increasing use of evidence
Increasing accuracy

Definition: Sensitivity

Sensitivity (SN) is the fraction of the reference feature that is predicted by the gene predictor.

$$SN = TP/(TP + FN)$$

Definition: Specificity

Specificity (SP) is the fraction of the prediction overlapping the reference feature.

$$SP = TP/(TP + FP)$$

Definition: Accuracy

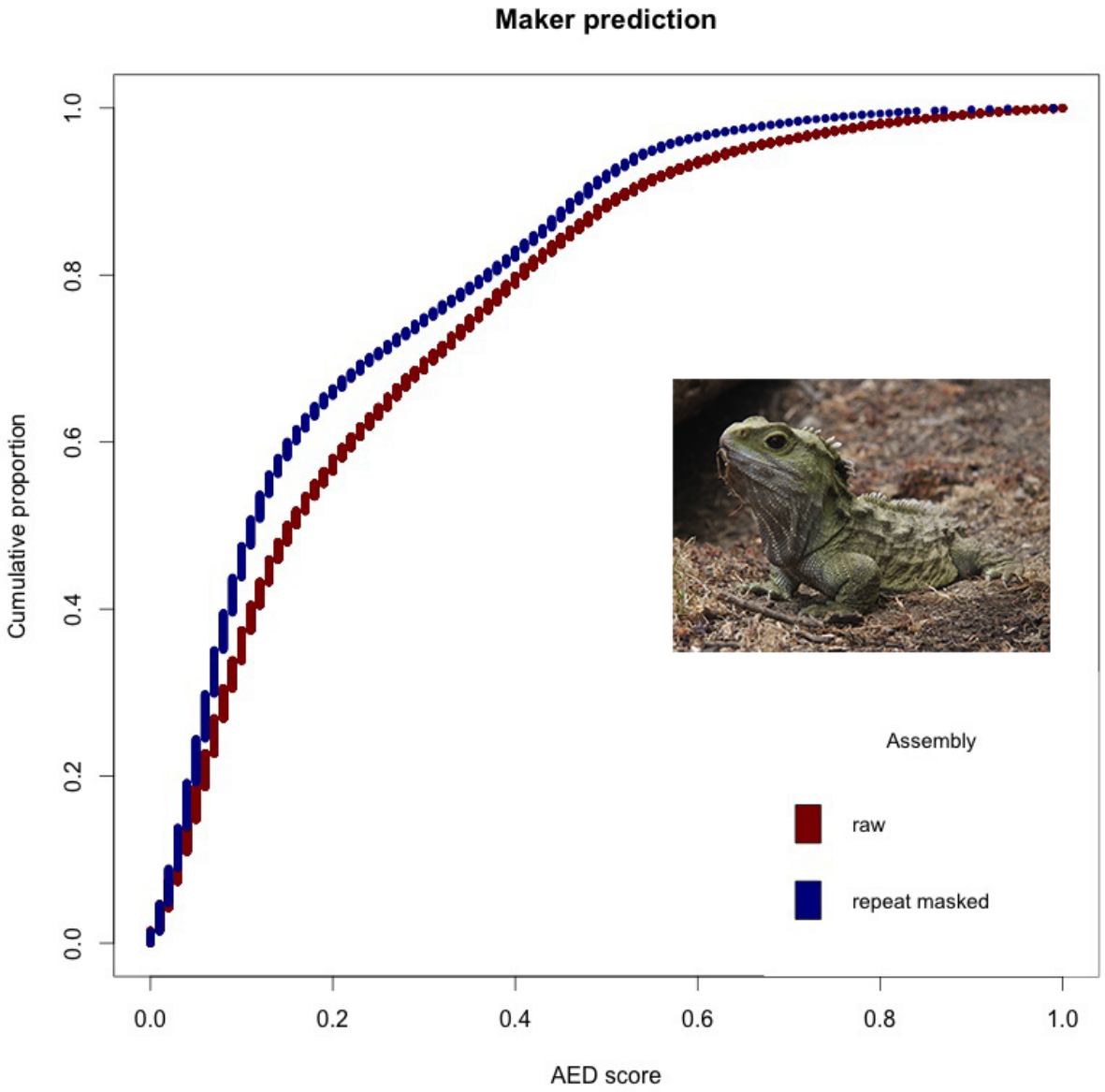
SN and SP are often combined into a single measure called accuracy (AC)

Definition: Annotation Edit Distance

Annotation Edit Distance (AED) is calculated in the same manner as SN and SP, but in place of a reference gene model, the coordinates of the union of the aligned evidence are used instead: $AED = 1 - AC$, where $AC = (SN + SP)/2$.

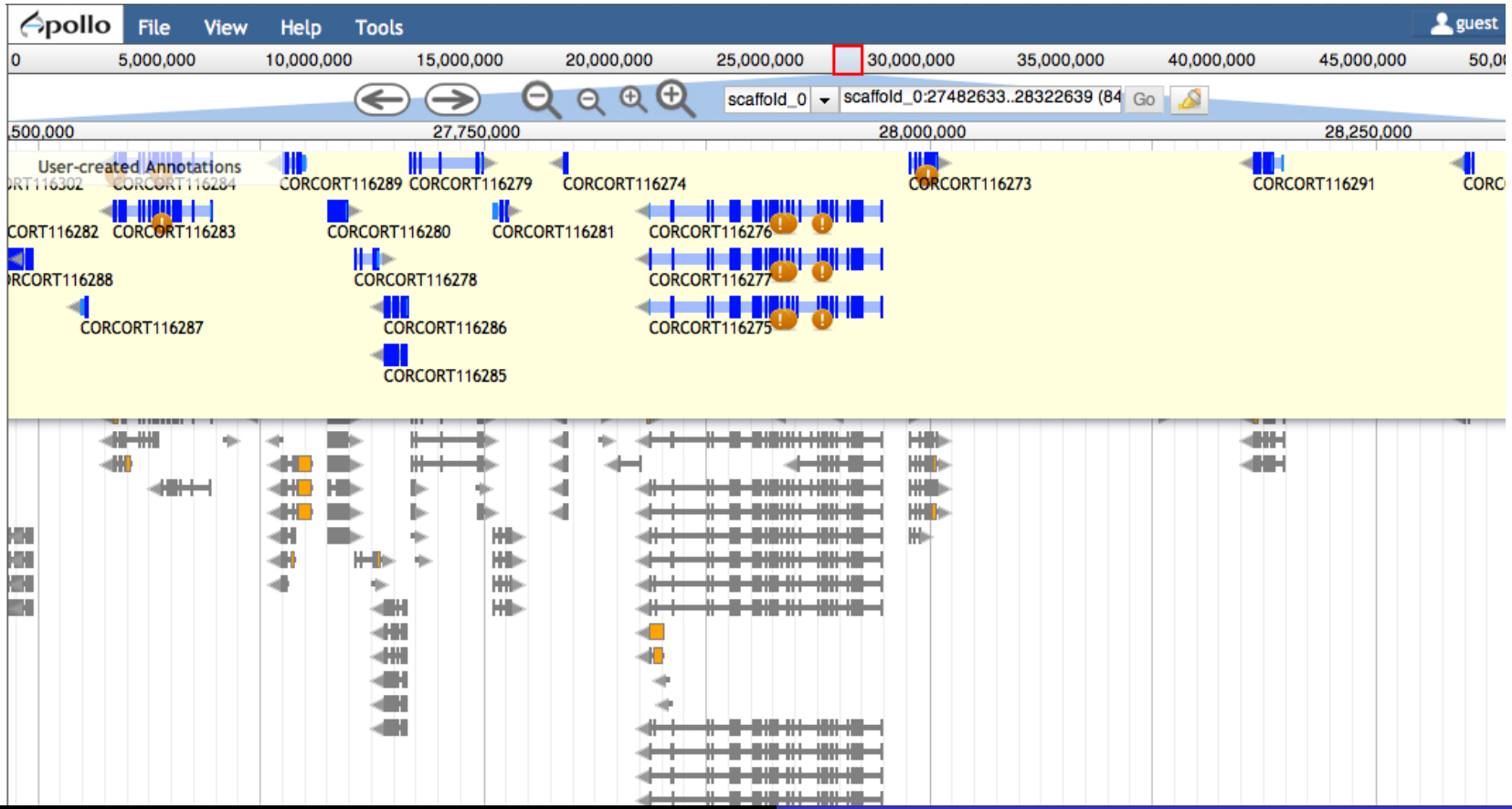
0... indicates perfect agreement with evidence

1... indicates no evidence support for annotation


Maker prediction
Cumulative proportion
0.0
0.2
0.4
0.6
0.8
1.0
AED score
0.0
0.2
0.4
0.6
0.8
1.0
Assembly
raw
repeat masked

- **Pipelines**
  - Maker2 (Holt and Yandell, 2011)
  - Pasa (Haas et al. 2003)
  - Ensembl (Curewen et al., 2004)
  - NCBI (Kitts, 2003)
- **Evidence Mapping**
  - BLAST / BLAT
  - Exonerate (Slater and Birey, 2005)
- **Ab Initio Gene Predictors**
  - Augustus (Stanke et al., 2006)
  - SNAP (Korf 2004)
  - GeneMark-ES (Zu et al., 2010)
- **Choosers and Combiners**
  - JIGSAW (Allen and Salzberg, 2005)
  - GLEAN (Elsik et al., 2007)

- **Visualization / Curation**
  - Artemis (Rutherford et al., 2000)
  - Apollo / WebApollo (Lewis et al., 2002)
  - JBROWSE (Skinner at al., 2009)
  - IGV (Robinson et al., 2011)



Maker 2