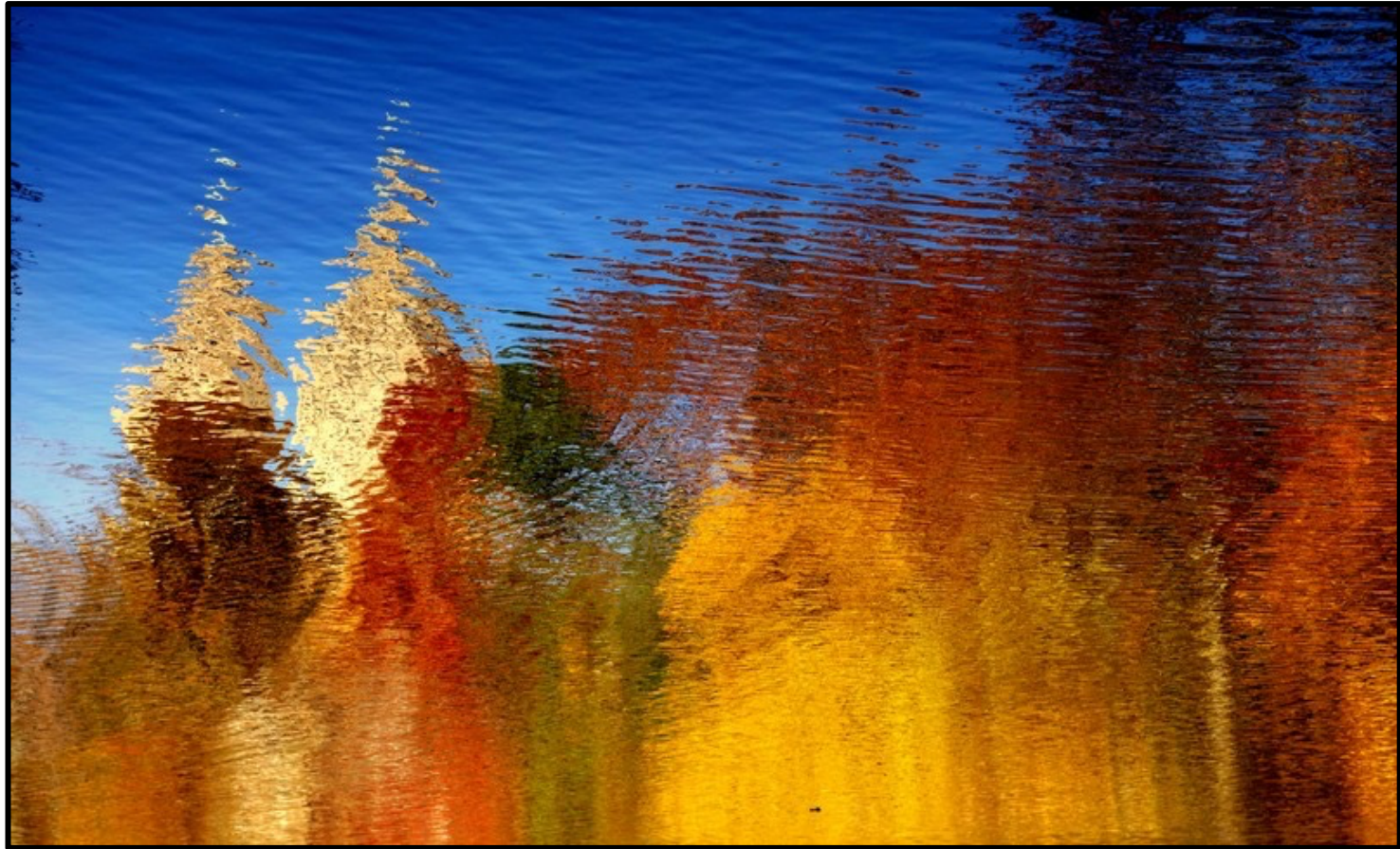


LARVA: An integrative framework for Large-scale Analysis of Recurrent Variants in noncoding Annotations



M Gerstein, Yale

Slides freely downloadable from **Lectures.GersteinLab.org**
& “tweetable” (via @markgerstein). See last slide for references & more info.

Finding Key Variants in Cancer Genomes: the Needle in the Haystack

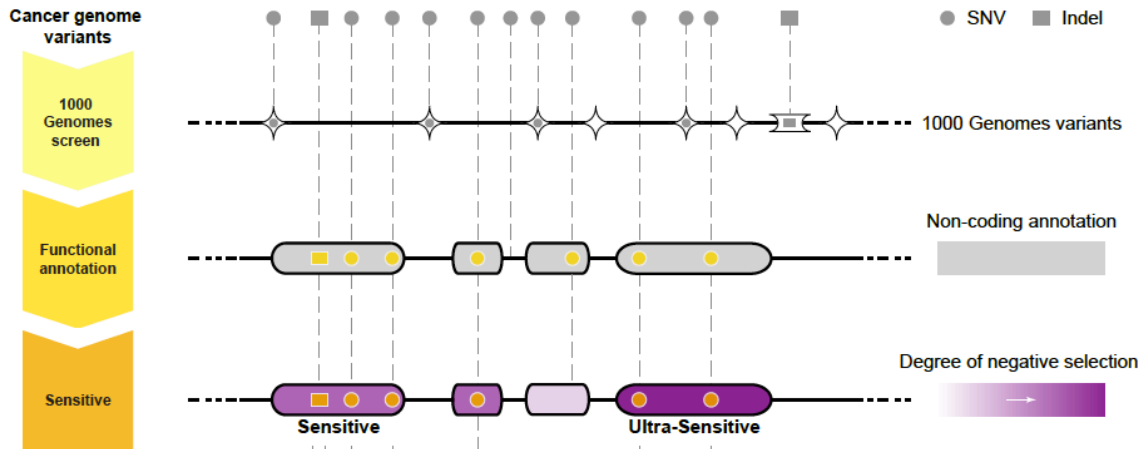
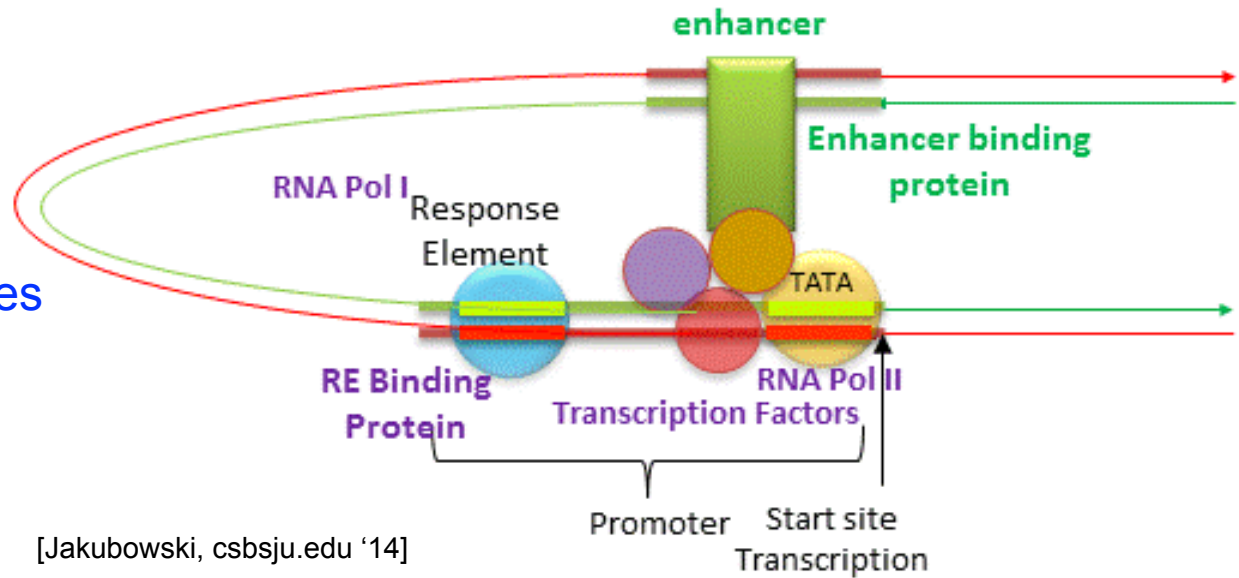


[Image credit: www.yourgrantauthority.com '15]

- Increasing number of whole genome sequenced for tumor/normal pairs
 - Eg >2500 for PCAWG
- Lots of somatic mutations in an average tumor (~5K/sample), particularly in non-coding regions
- A focus is distinguishing drivers & passengers
 - Canonical **Drivers** are mutations driving cancer progression
 - Thought to be under positive selection
 - Recur in the same position, gene or functional element across tumors in different individuals
 - **Passengers** are thought not be significant to driving cancer progression
 - Collateral damage
 - Could result from impaired DNA repair processes
- Most driver work has focused on genes
 - eg Youn & Simon ('11). *Bioinformatics*; Lawrence et al. ('13). *Nature*

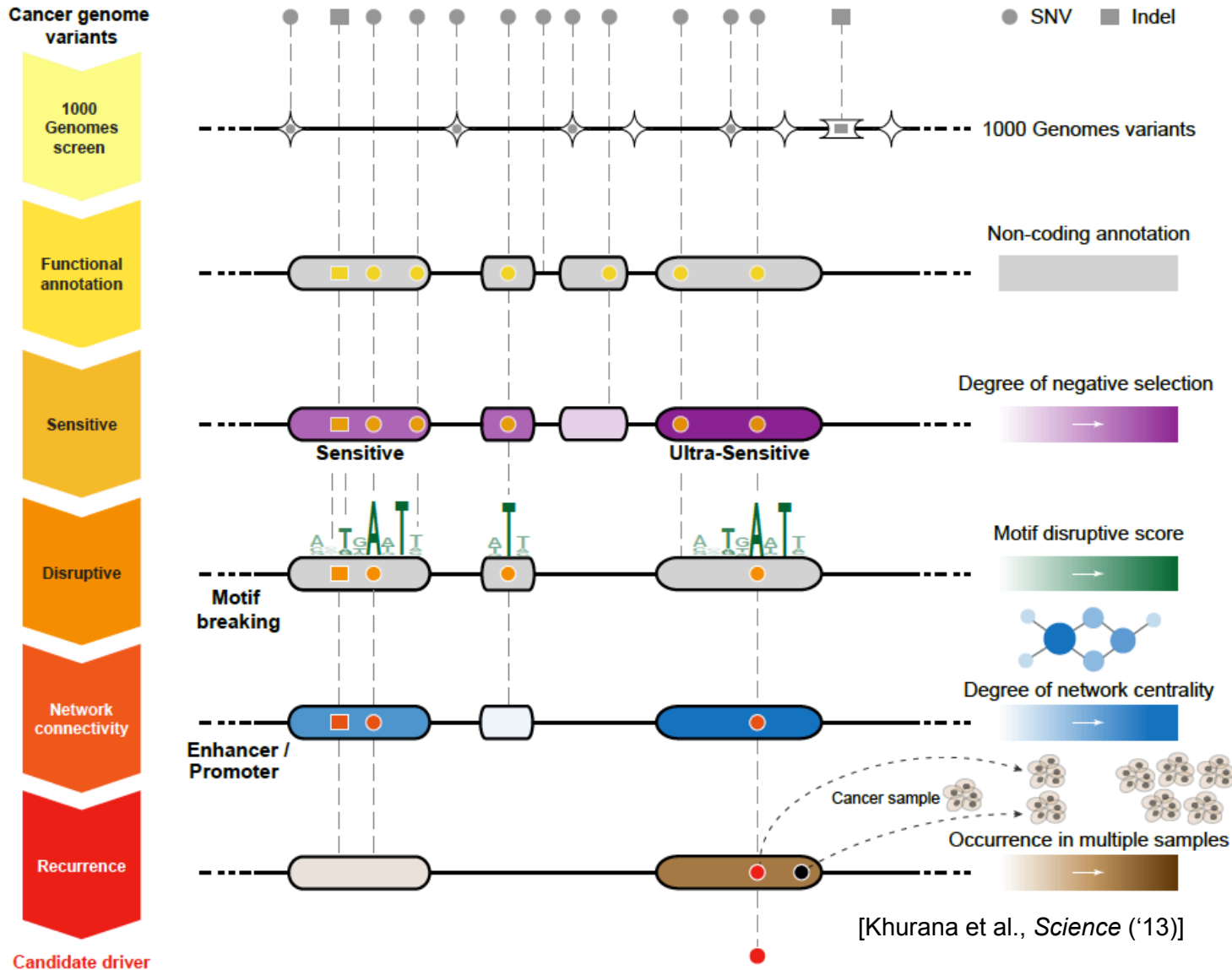
Noncoding Annotations

- Promoters
- TF binding sites
- Transcription start sites
- DHS sites
- Enhancers

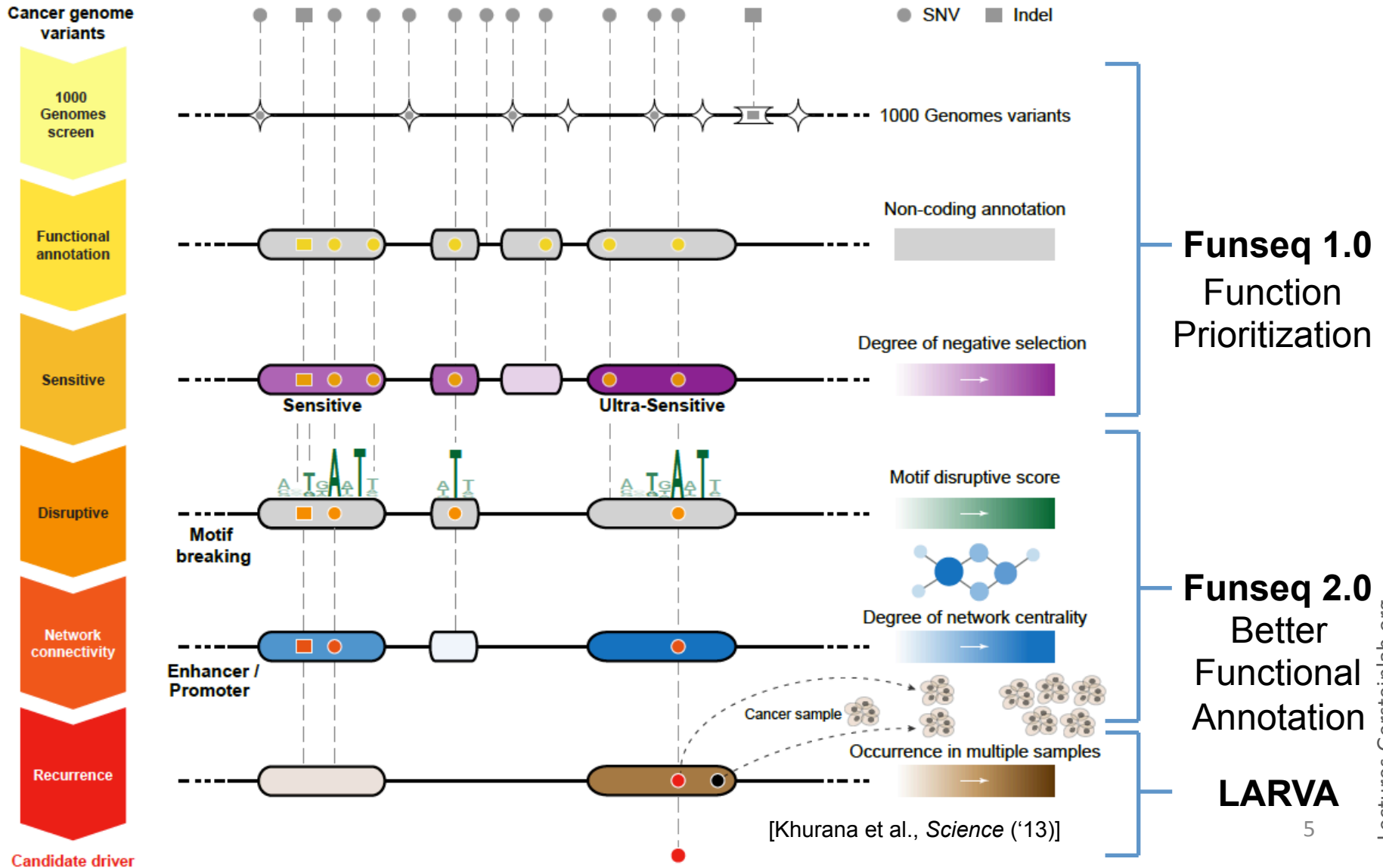


Ultra-sensitive & Ultra-conserved elements Non-coding regions more conserved than expectation across the human population & between species [Bejerano et al. ('04). Science; Khurana et al., Science ('13)]

Identification of non-coding candidate drivers amongst somatic variants: FunSeq

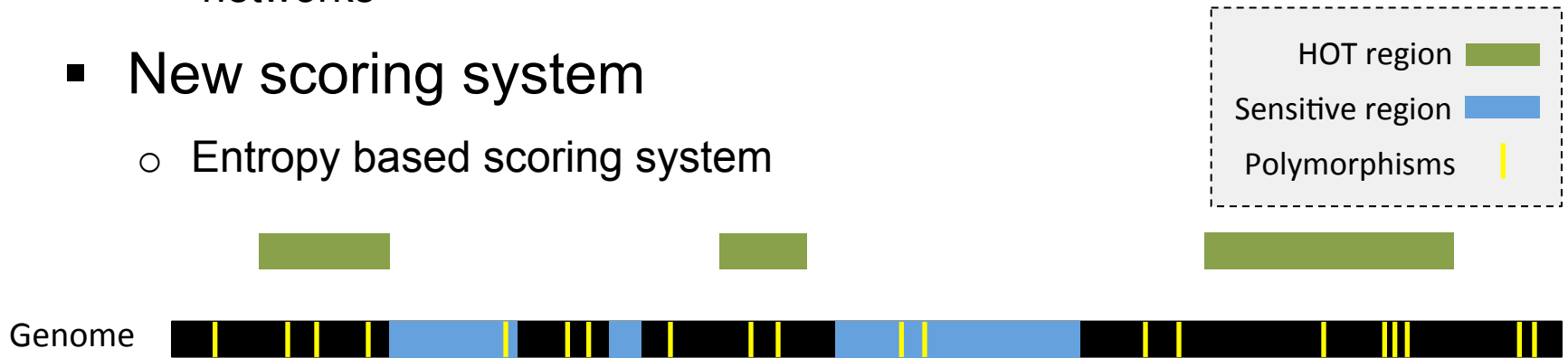


Identification of non-coding candidate drivers amongst somatic variants: FunSeq



From Funseq 1.0 to Funseq 2.0

- Elaborated features
 - Motif disruption score: changes in PWMs
 - Network centrality analysis: PPI, regulatory, and phosphorylation networks
- New scoring system
 - Entropy based scoring system



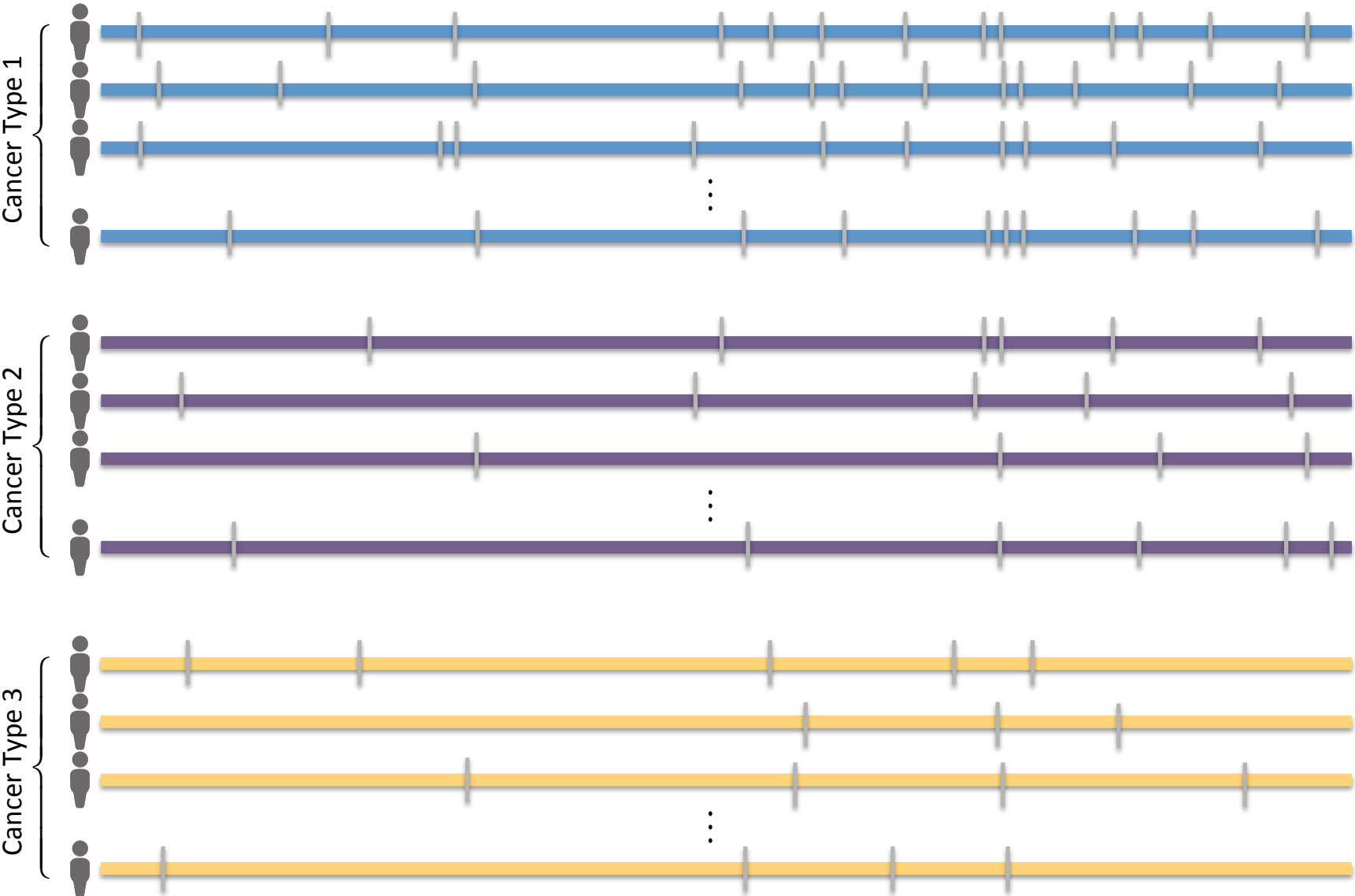
$$p = \frac{3}{20}$$

Feature weight: $w_d = 1 + p_d \log_2 p_d + (1 - p_d) \log_2 (1 - p_d)$

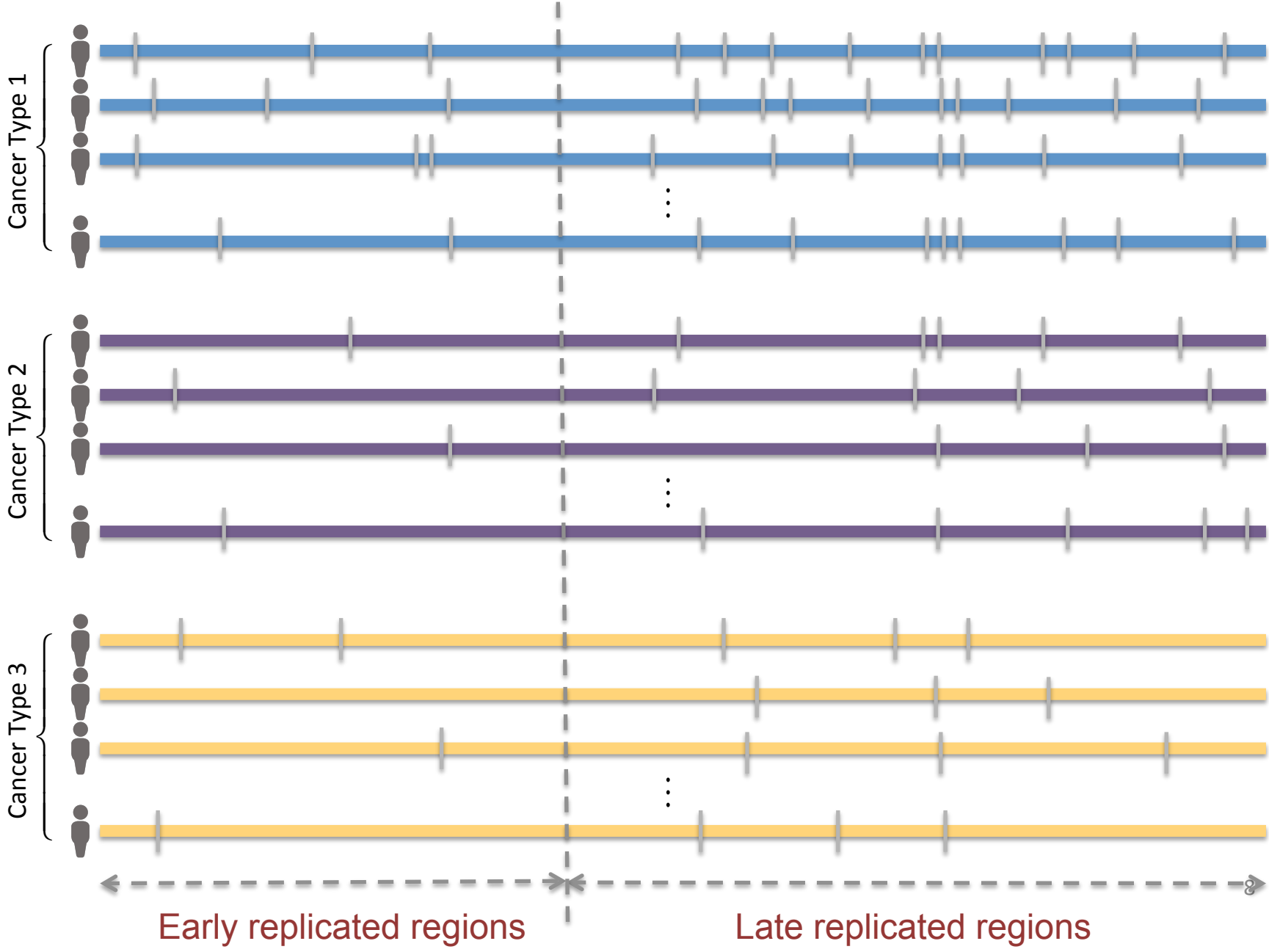
$p \uparrow$ $w_d \downarrow$ $p = \text{probability of the feature overlapping natural polymorphisms}$

For a variant: $\text{Score} = \sum w_d$ of observed features

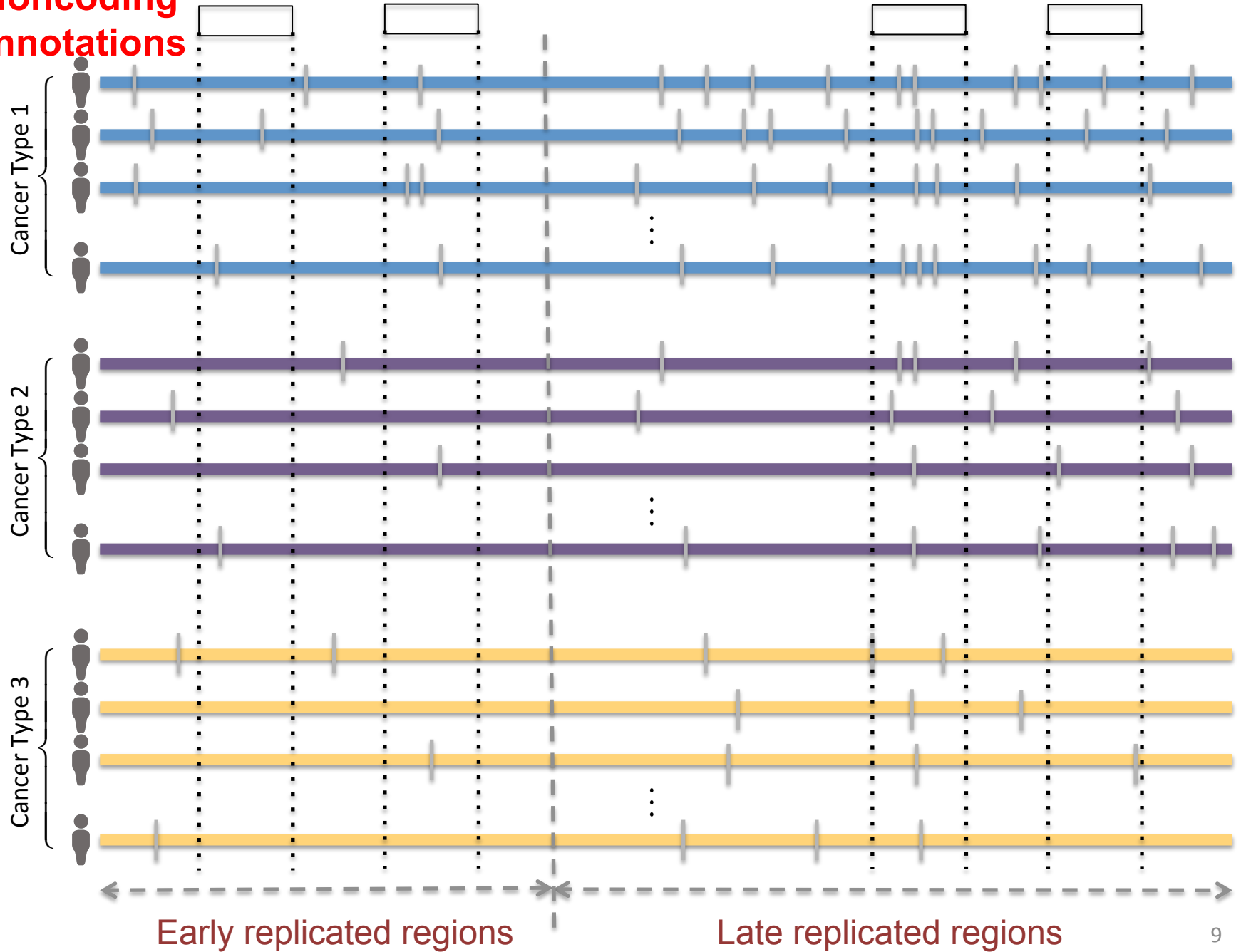
Mutation recurrence



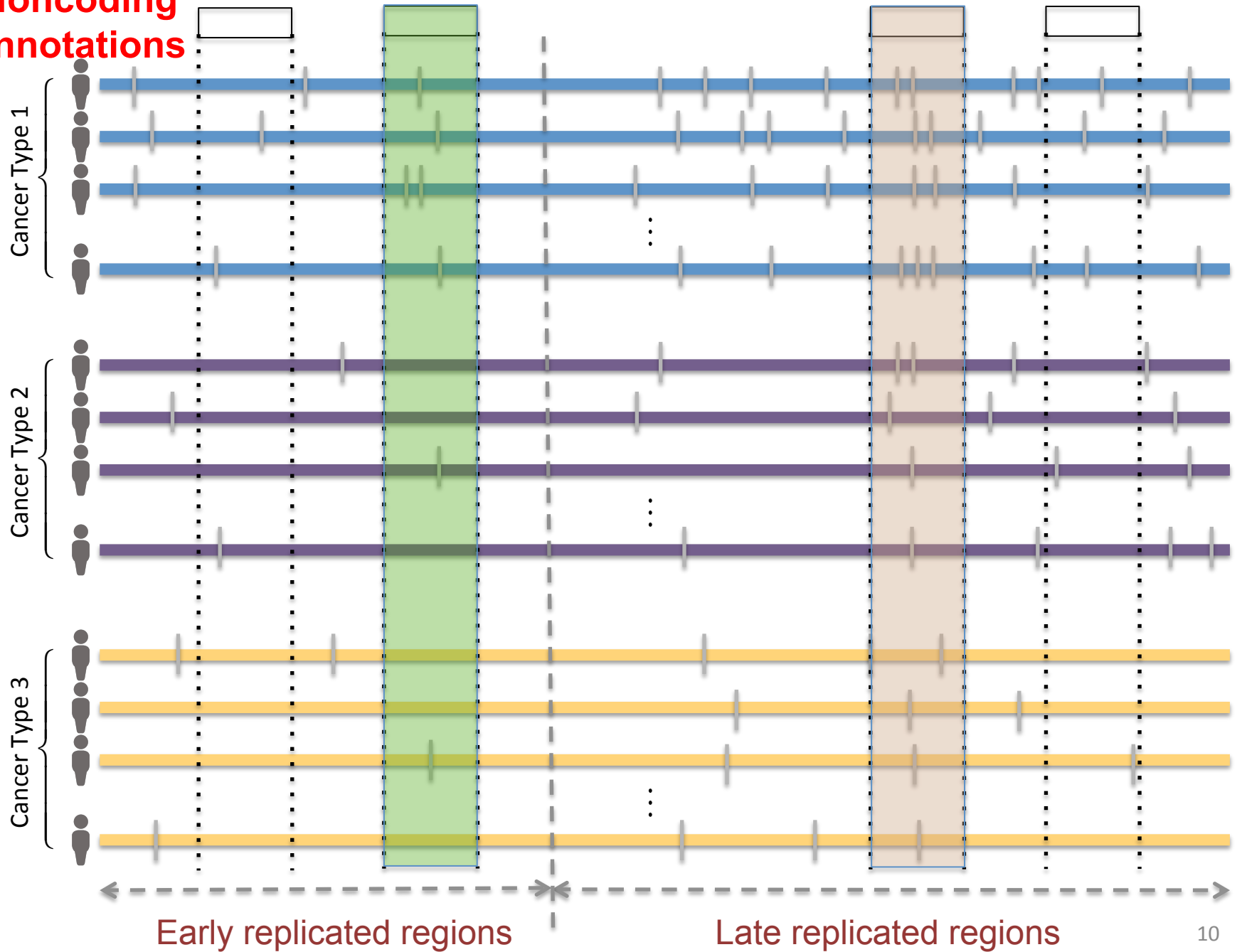
Mutation recurrence



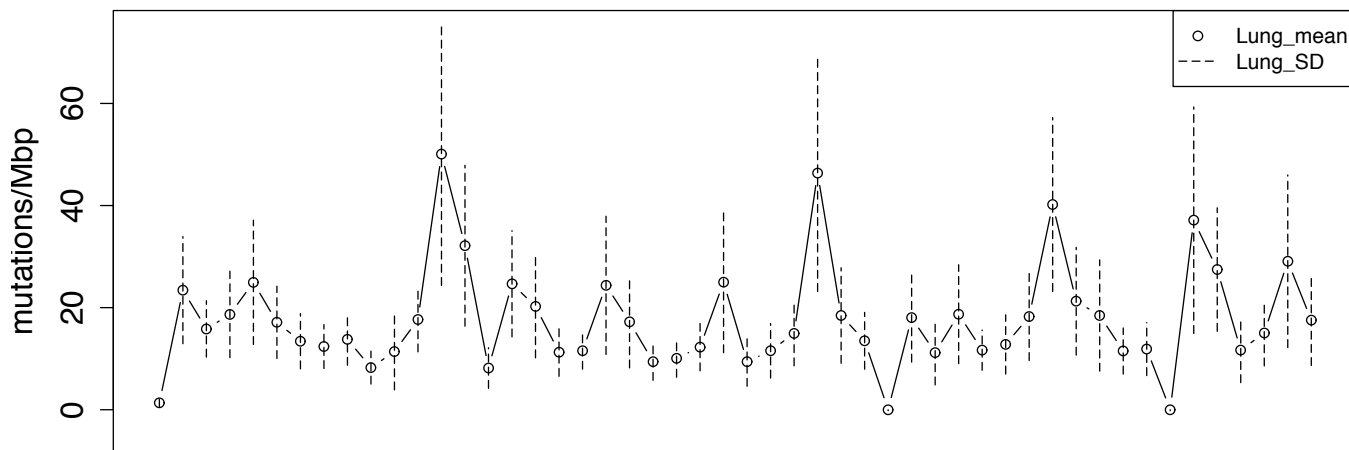
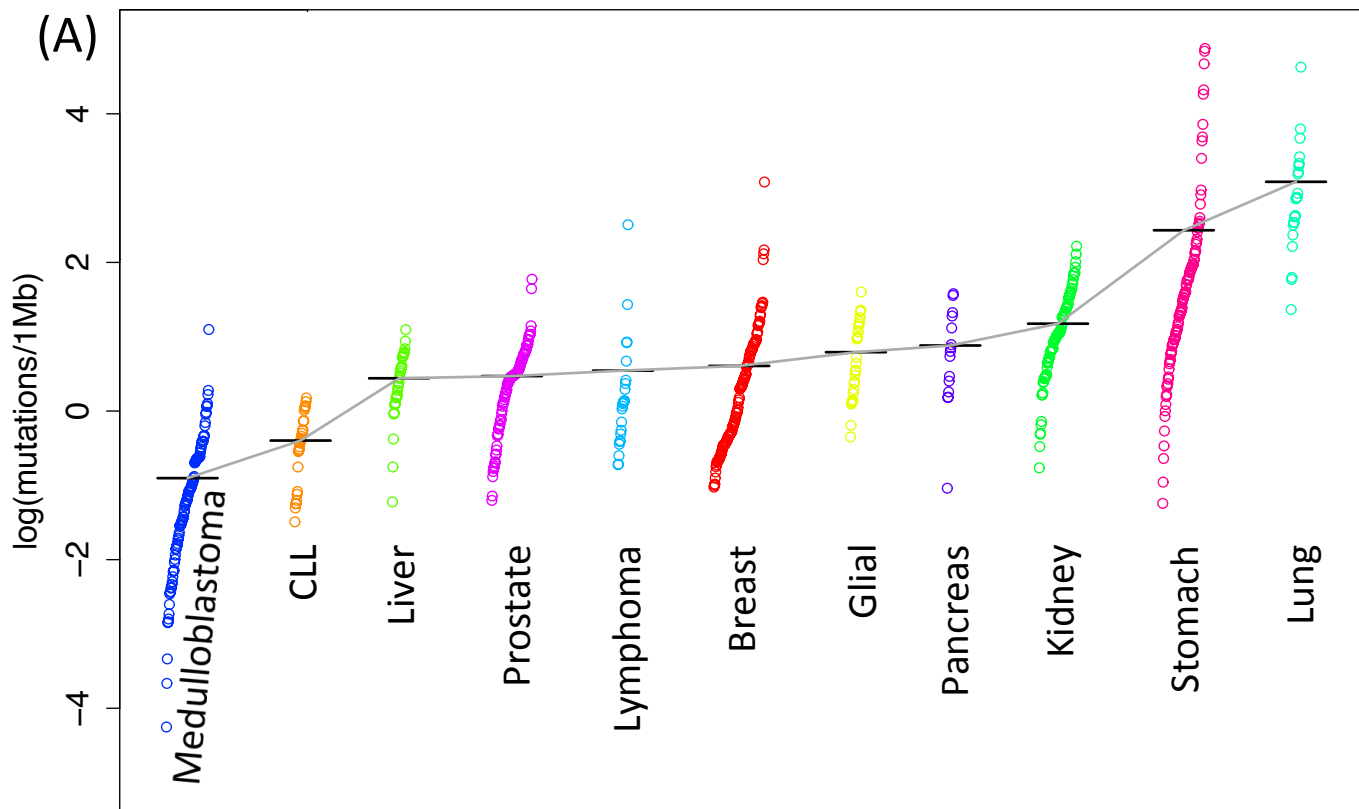
Noncoding annotations



Noncoding annotations



Cancer Somatic Mutational Heterogeneity, across cancer types, samples & regions



1 Mbp genome regions (locations chosen at random)

Cancer Somatic Mutation Modeling

- 3 models to evaluate the significance of mutation burden
- Suppose there are k genome elements. For element i , define:
 - n_i : total number of nucleotides
 - x_i : the number of mutations within the element
 - p_i : the mutation rate
 - R : the replication timing bin of the element

Model 1: Constant Background Mutation Rate (Model from Previous Work)

$$x_i : \text{Binomial}(n_i, p)$$

Model 2: Varying Mutation Rate

$$x_i | p_i : \text{Binomial}(n_i, p_i)$$

$$p_i : \text{Beta}(\mu, \sigma)$$

Model 3: Varying Mutation Rate with Replication Timing Correction

$$x_i | p_i : \text{Binomial}(n_i, p_i)$$

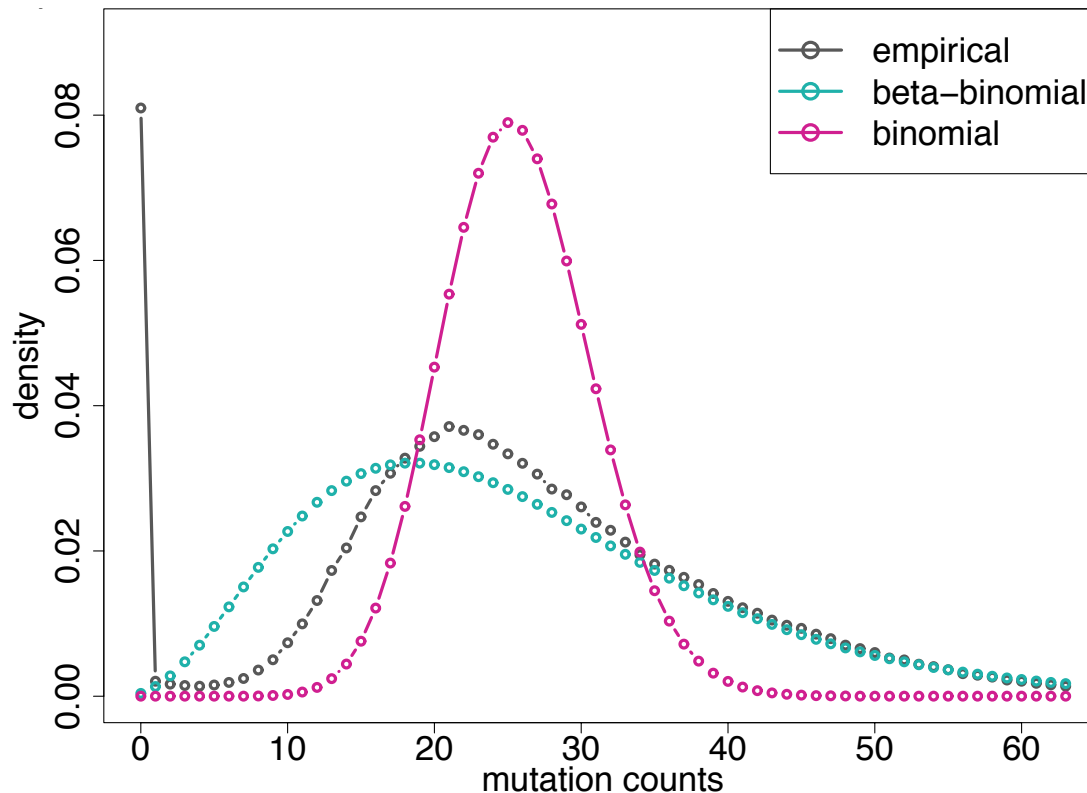
$$p_i : \text{Beta}(\mu | \mathbf{R}, \sigma | \mathbf{R})$$

$$\mu | \mathbf{R}, \sigma | \mathbf{R} : \text{constant within the same } \mathbf{R} \text{ bin}$$

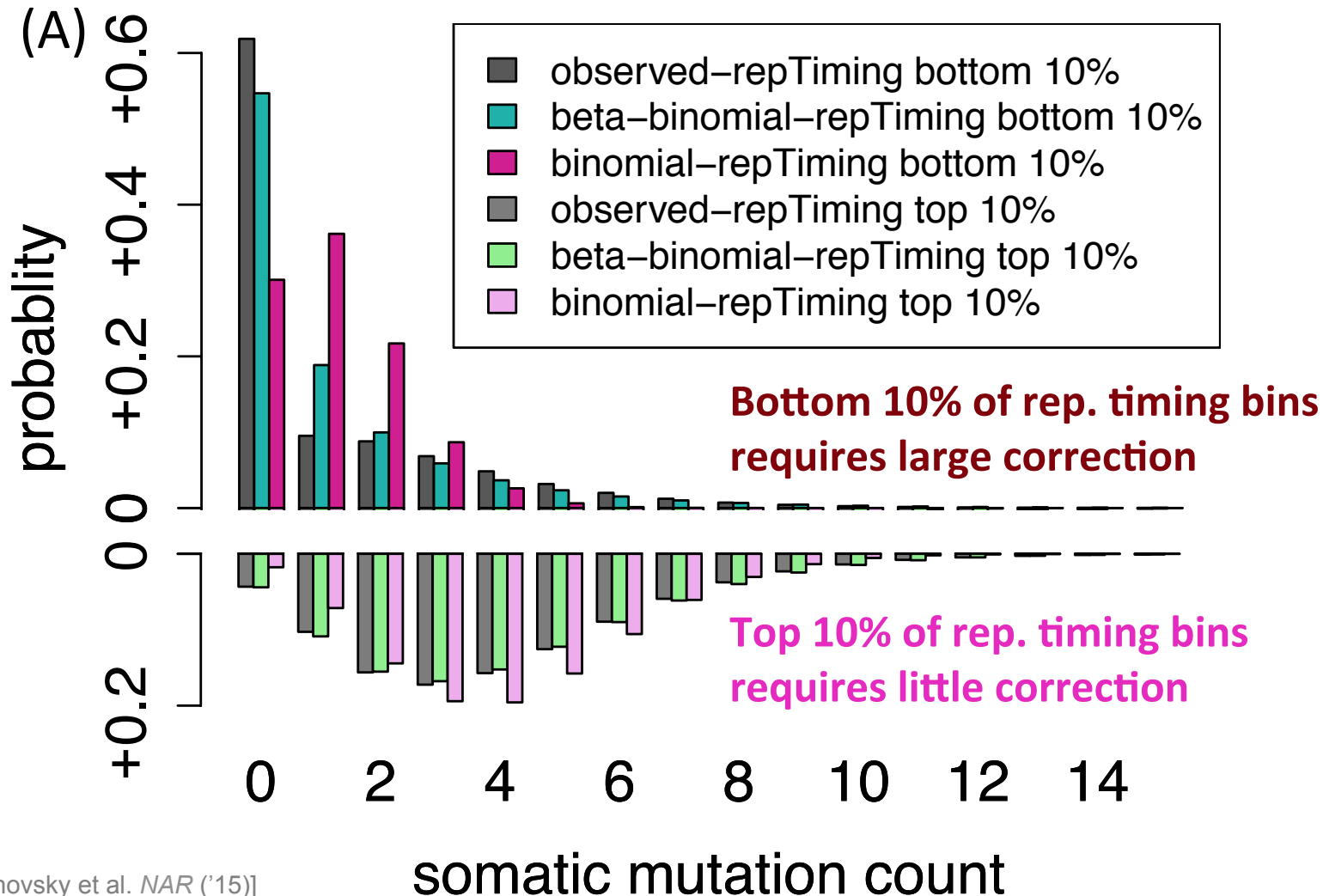
[Lochovsky et al. *NAR* ('15)]

LARVA Model Comparison

- Comparison of mutation count frequency implied by the binomial model (model 1) and the beta-binomial model (model 2) relative to the empirical distribution
- The beta-binomial distribution is significantly better, especially for accurately modeling the over-dispersion of the empirical distribution

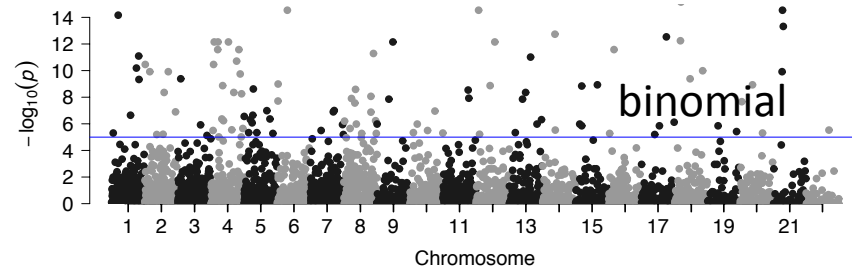
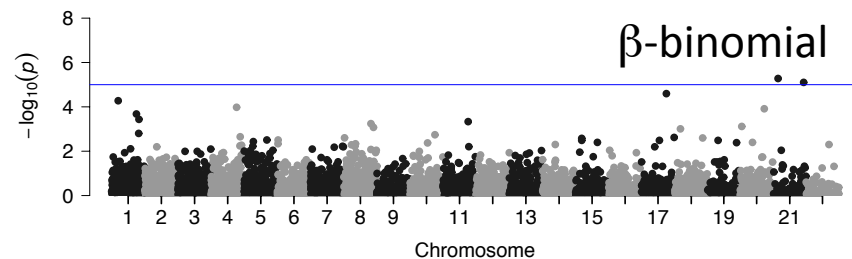
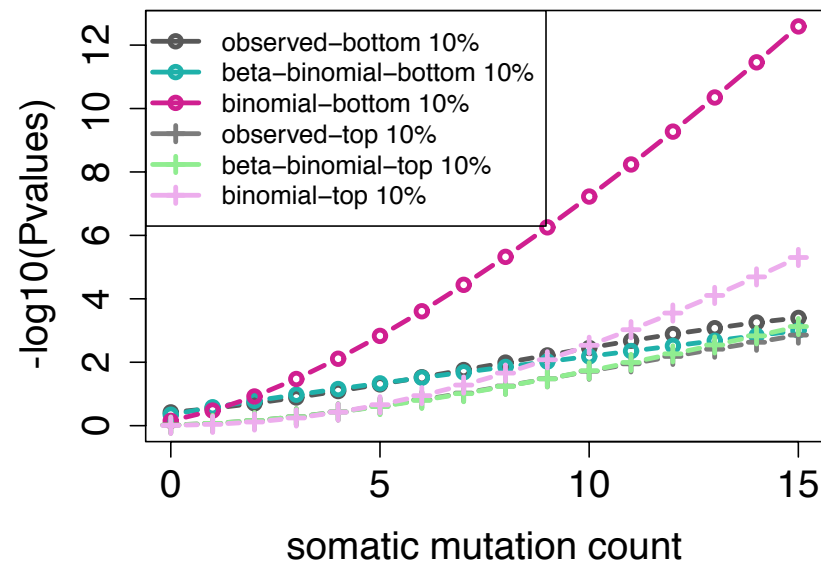
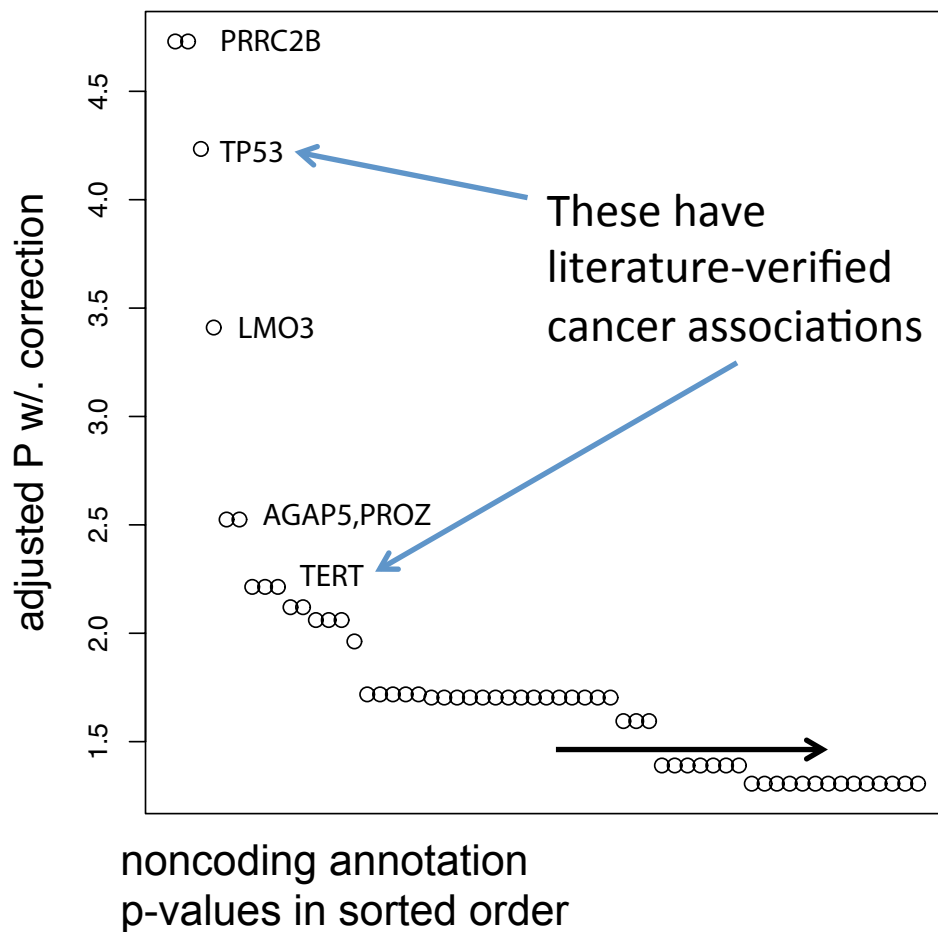


Adding DNA replication timing correction further improves the beta-binomial model



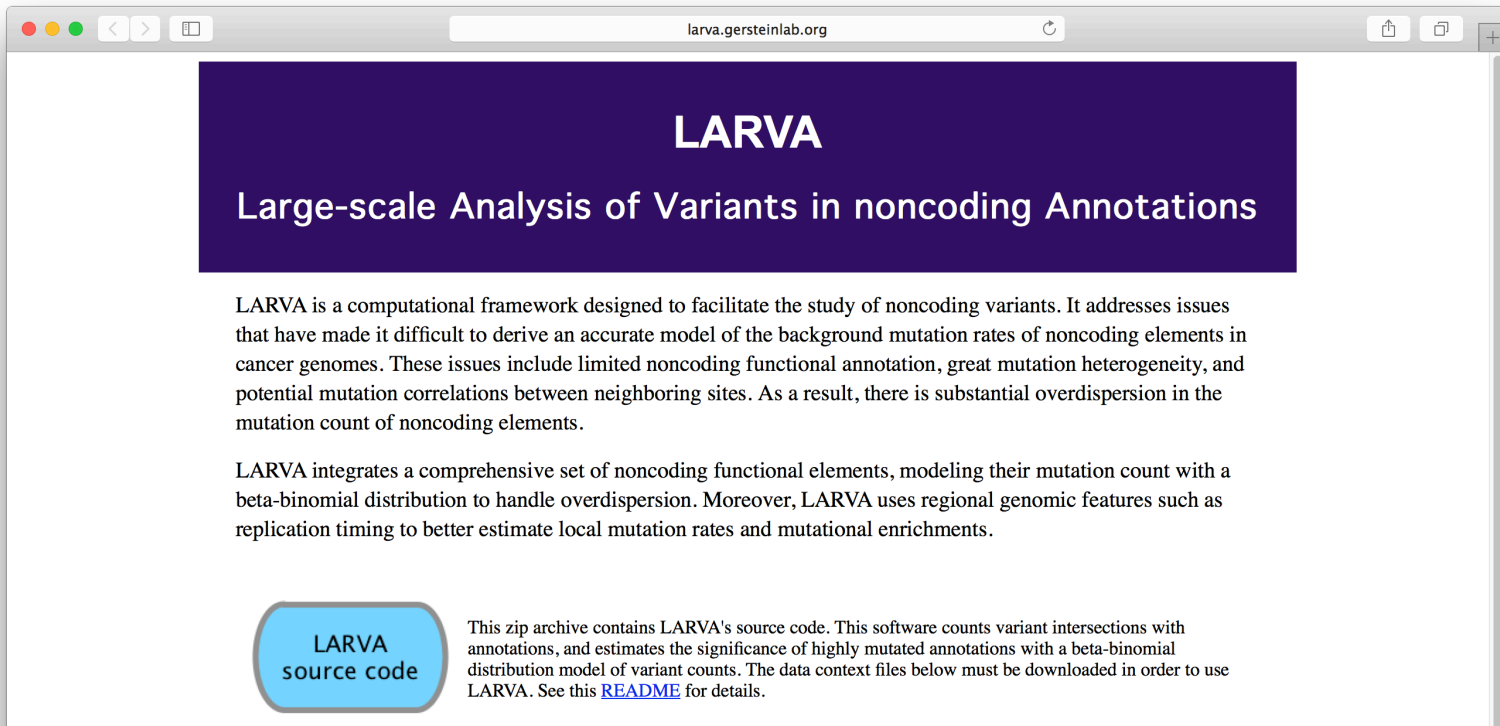
LARVA Results

TSS LARVA results



LARVA Implementation

- <http://larva.gersteinlab.org/>
- Freely downloadable C++ program
 - Verified compilation and correct execution on Linux
- A Docker image is also available to download
 - Runs on any operating system supported by Docker
- Running time on transcription factor binding sites (a worst case input size) is ~80 min
 - Running time scales linearly with the number of annotations in the input



The screenshot shows a web browser window with the URL larva.gersteinlab.org. The page features a dark purple header with the text "LARVA" in white, followed by the subtitle "Large-scale Analysis of Variants in noncoding Annotations" in white. Below the header, there is a paragraph of text describing the framework's purpose and challenges. A second paragraph explains the computational model used. At the bottom, there is a blue button labeled "LARVA source code" and a link to the README file.

LARVA
Large-scale Analysis of Variants in noncoding Annotations

LARVA is a computational framework designed to facilitate the study of noncoding variants. It addresses issues that have made it difficult to derive an accurate model of the background mutation rates of noncoding elements in cancer genomes. These issues include limited noncoding functional annotation, great mutation heterogeneity, and potential mutation correlations between neighboring sites. As a result, there is substantial overdispersion in the mutation count of noncoding elements.

LARVA integrates a comprehensive set of noncoding functional elements, modeling their mutation count with a beta-binomial distribution to handle overdispersion. Moreover, LARVA uses regional genomic features such as replication timing to better estimate local mutation rates and mutational enrichments.

LARVA source code This zip archive contains LARVA's source code. This software counts variant intersections with annotations, and estimates the significance of highly mutated annotations with a beta-binomial distribution model of variant counts. The data context files below must be downloaded in order to use LARVA. See this [README](#) for details.

Acknowledgements



- **LARVA**.gersteinlab.org
 - L **Lochovsky***, J **Zhang***, Y Fu, E Khurana
- **FunSeq2**.gersteinlab.org
 - Y **Fu**, Z Liu, S Lou, J Bedford, X Mu, K Yip, E Khurana

Info about content in this slide pack

- General PERMISSIONS
 - This Presentation is copyright Mark Gerstein, Yale University, 2015.
 - Please read permissions statement at **www.gersteinlab.org/misc/permissions.html** .
 - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
 - Paper references in the talk were mostly from Papers.GersteinLab.org.
- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see <http://streams.gerstein.info> .
 - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: <http://www.flickr.com/photos/mbgmbg/tags/kwpotppt>