Challenge 4 –
Moving beyond coding regions

Mark Gerstein

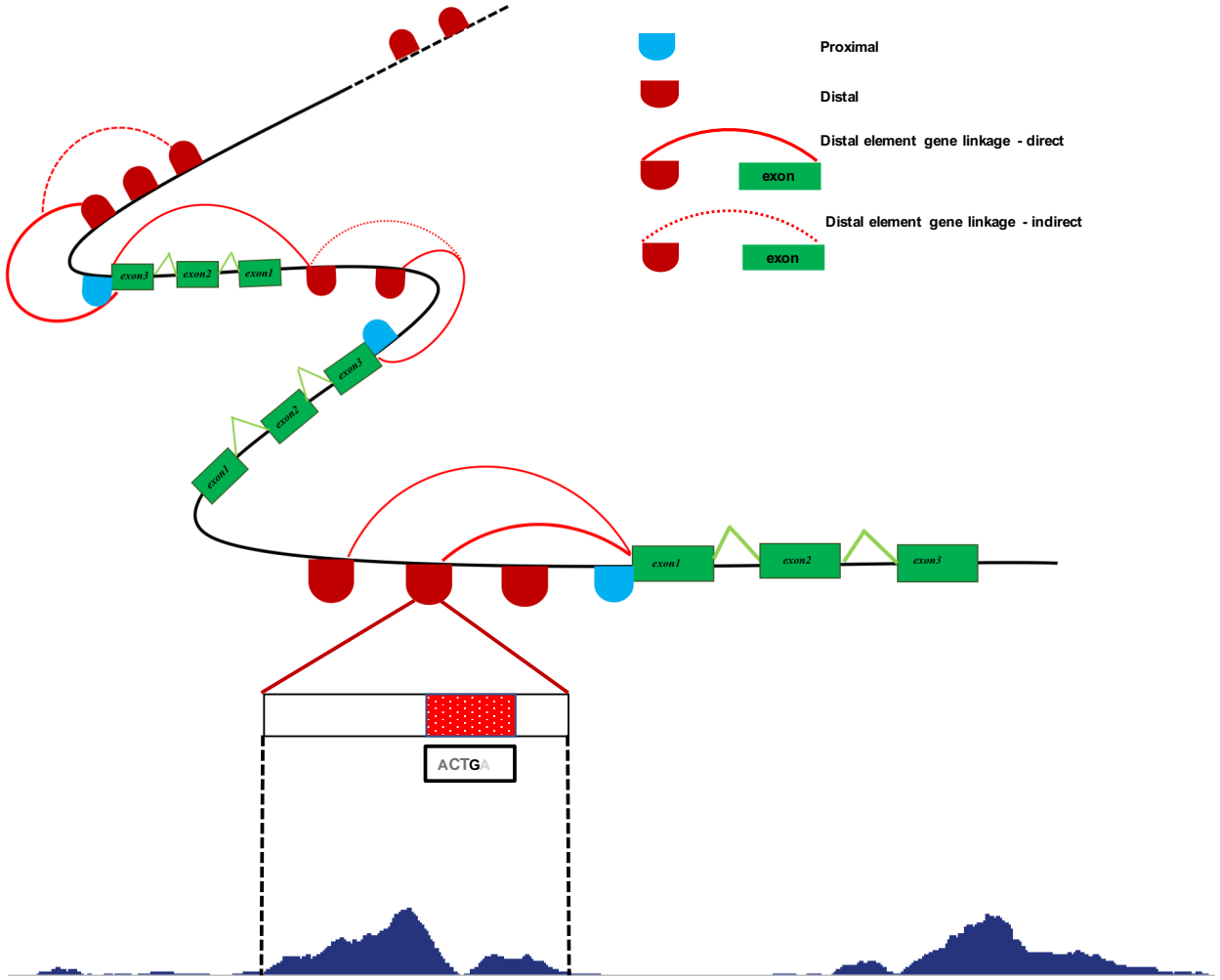
Yale CMG

Moving beyond coding

- CMG projects have been mostly WES, with a few incorporating WGS (incl. Dubowitz & unsolved cases).

- Compared to WES, interpreting variants in non-coding regions is challenging.
  3 things to consider in this regard
  (annotation qual., differential impact, variant qual.) …

Things to consider in moving beyond coding #1:
Quality & scale of coding v. non-coding annotation & the impact of this on statistical power

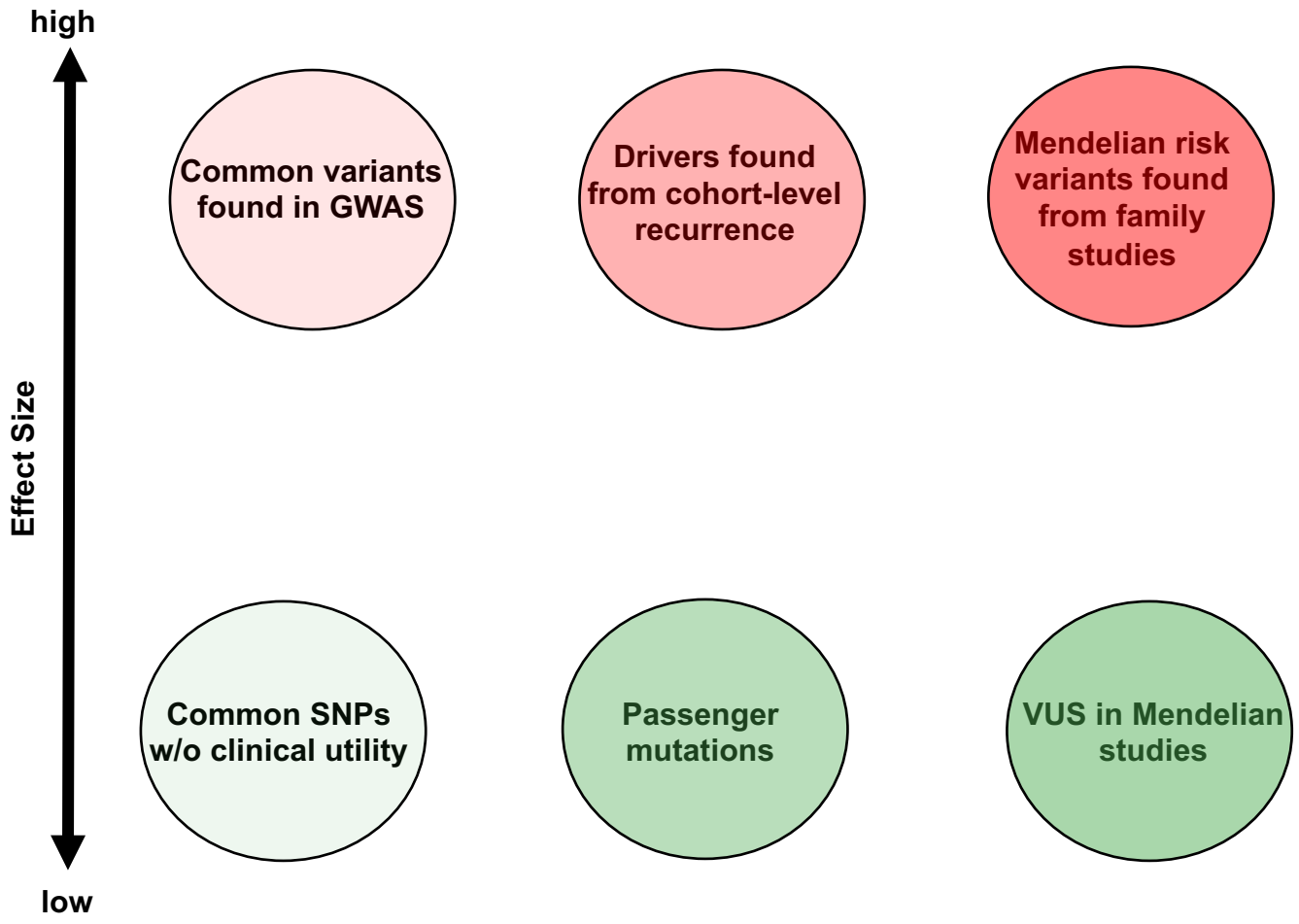


[*Nature* 547: 40]

- **ENCODE** has developed non-coding annotations & a number of tools have been developed to synthesize these (eg HaploReg, FunSeq, &c)
- Compared to coding regions, the underlying functional **territory of non-coding regions is not as well defined** nor is the differential effect of different mutations
- This creates **power issues** in non-coding variant prioritization. More precise (ie more compact) annotation may be useful.
- Also, integration of **tissue-specific** annotations & epigenetic data is important for deciphering impact of non-coding variants

Things we need to consider in moving from coding to non-coding #2:

Most of the high-impact variants found so far tend to occur in coding regions (lessons from cancer genomics)

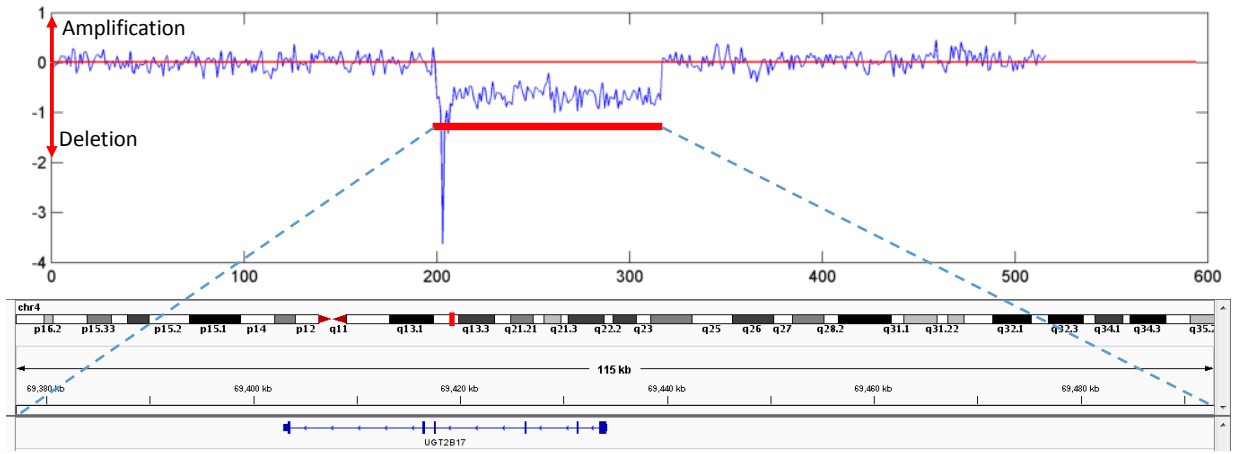


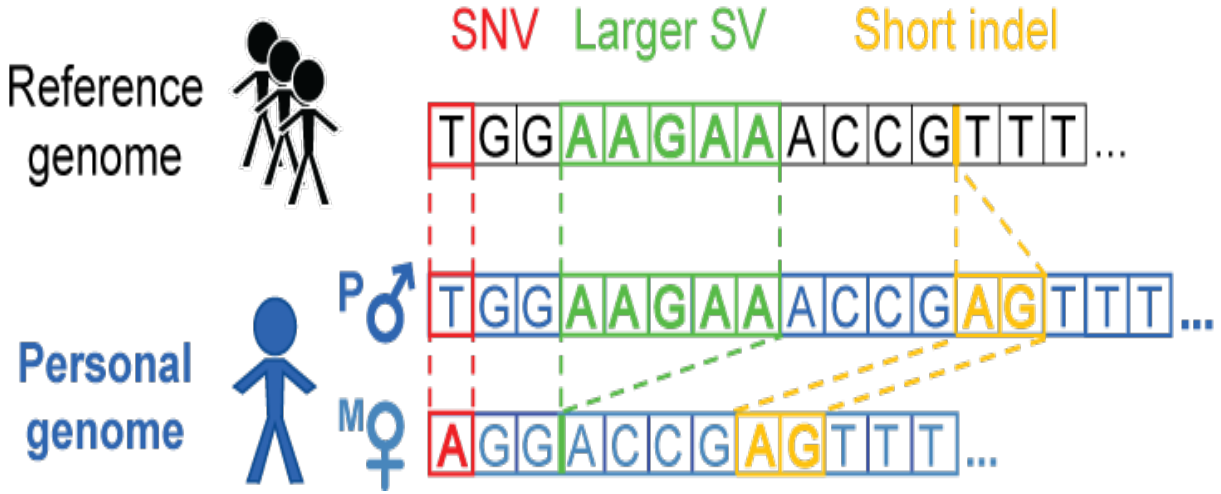
[Adapted from Thomas et al., Lancet ('15)]

- **Somatic coding driver vs non-coding passenger** as an example of extreme dichotomy. Or is this a function of ascertainment ?
- Despite 1000s of WGS call sets, very **few non-coding drivers** have been found in cancer genomics [Rhienbay et al bioRxiv ’17; Khurana et al NRG ’16]
- In general (ie for CMG), do high-impact variants tend to occur in coding regions & “softer” regulatory ones, in non-coding regions?

Things we need to consider in moving from coding to non-coding #3:

Variant calls (even coding ones) from WGS maybe more informative & accurate





- WGS can detect **full spectrum of variants** including SNPs, INDELs, & SVs. **SVs, in particular**, are harder to interpret just in terms of exomes [Yang et al. AJHG '15].
- Accuracy of mapping can be better (even to coding), esp. w/ regard to **repeats** & pseudogenes [Zhang et al. PLOS Comp. Bio. '17].
- Potentially better uniformity in coverage may lead to better accuracy in coding variants (& handling of mosaicism) [Belkadi et al. PNAS. '15].
- WGS also makes possible **more precise references for mapping** – ie individual-specific, personal dipoloid genomes & population specific references [Chaisson et al. NRG ' 15; Rozowsky et al. MSB '11] .