Consider the geneset: Why the transcripts used for variant annotation matter

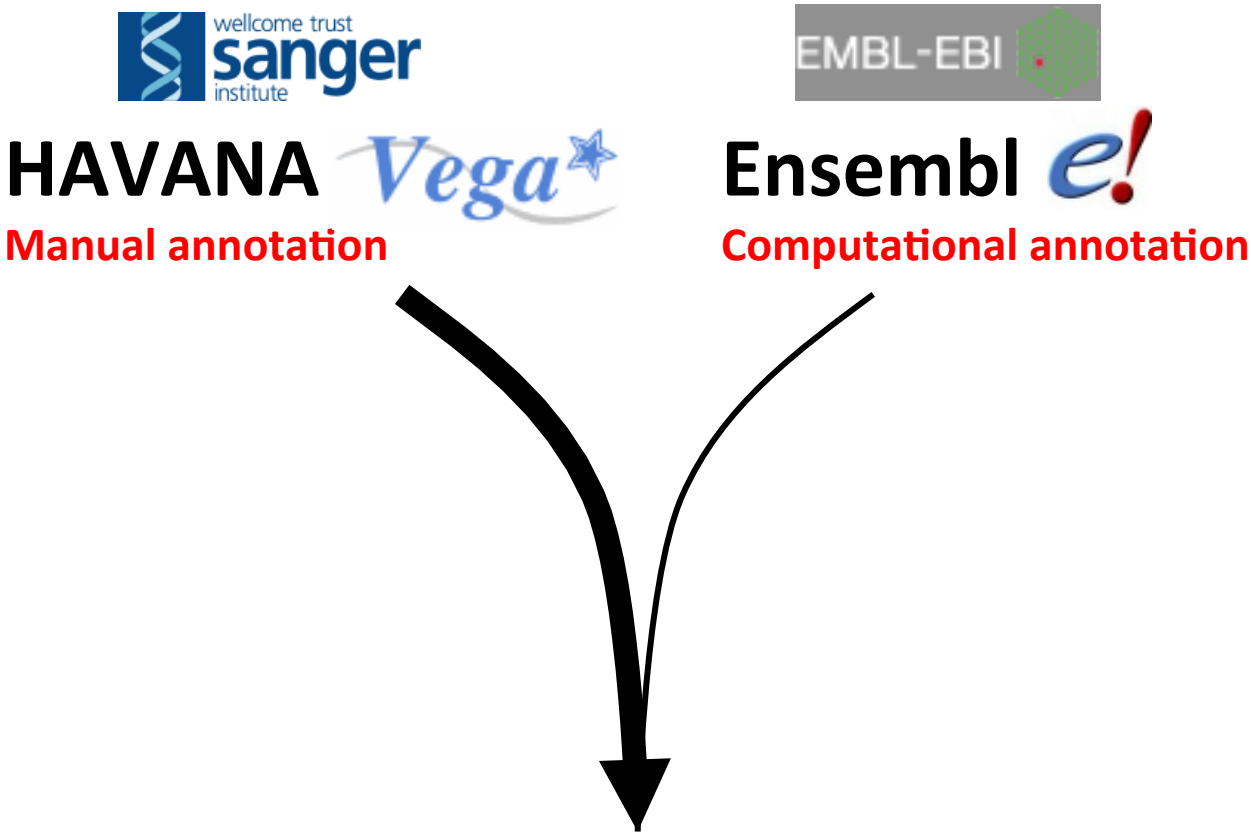
Mark Gerstein on behalf of

Adam Frankish

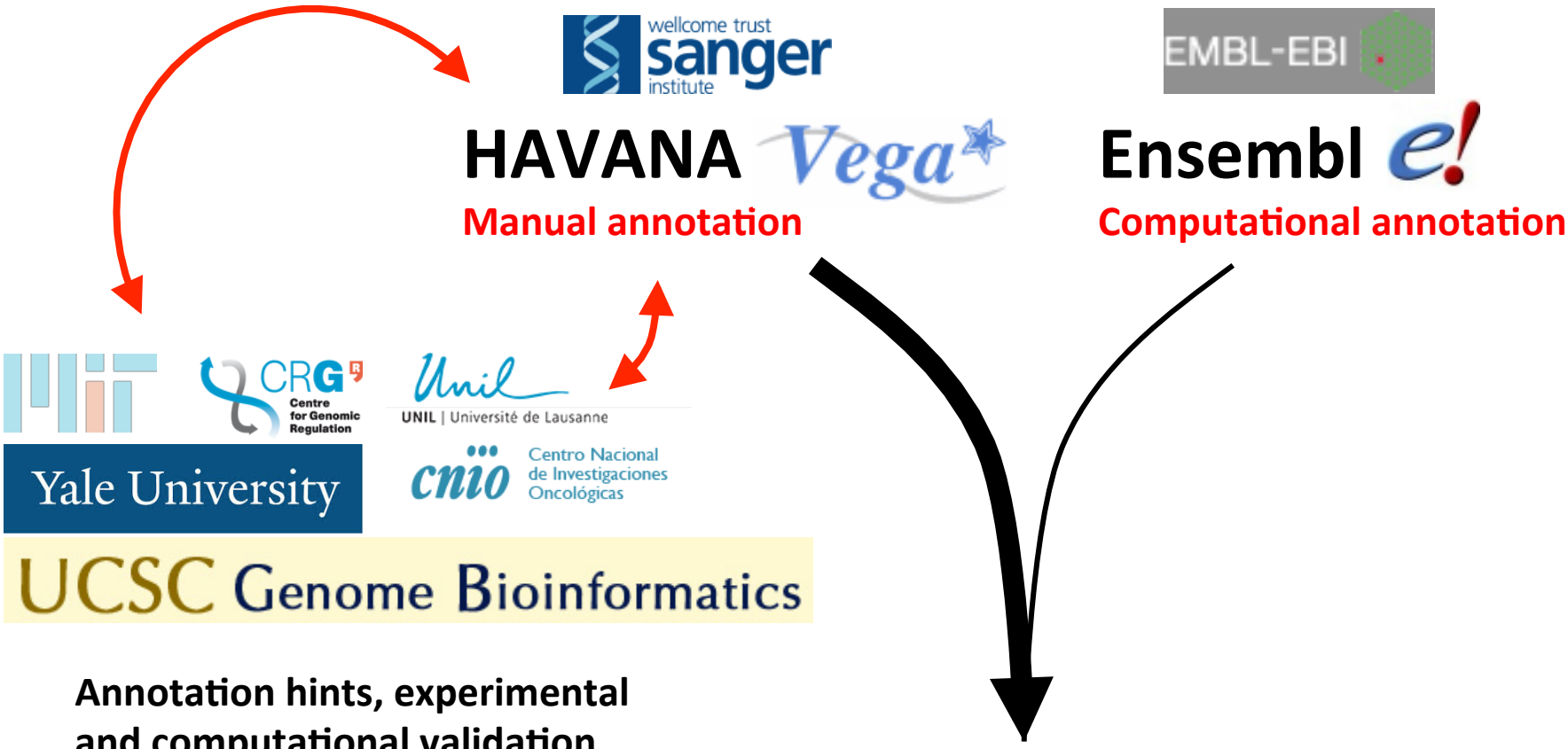
Computation Genomics, Wellcome Trust Sanger Institute

(both on behalf of the Gencode consortium)

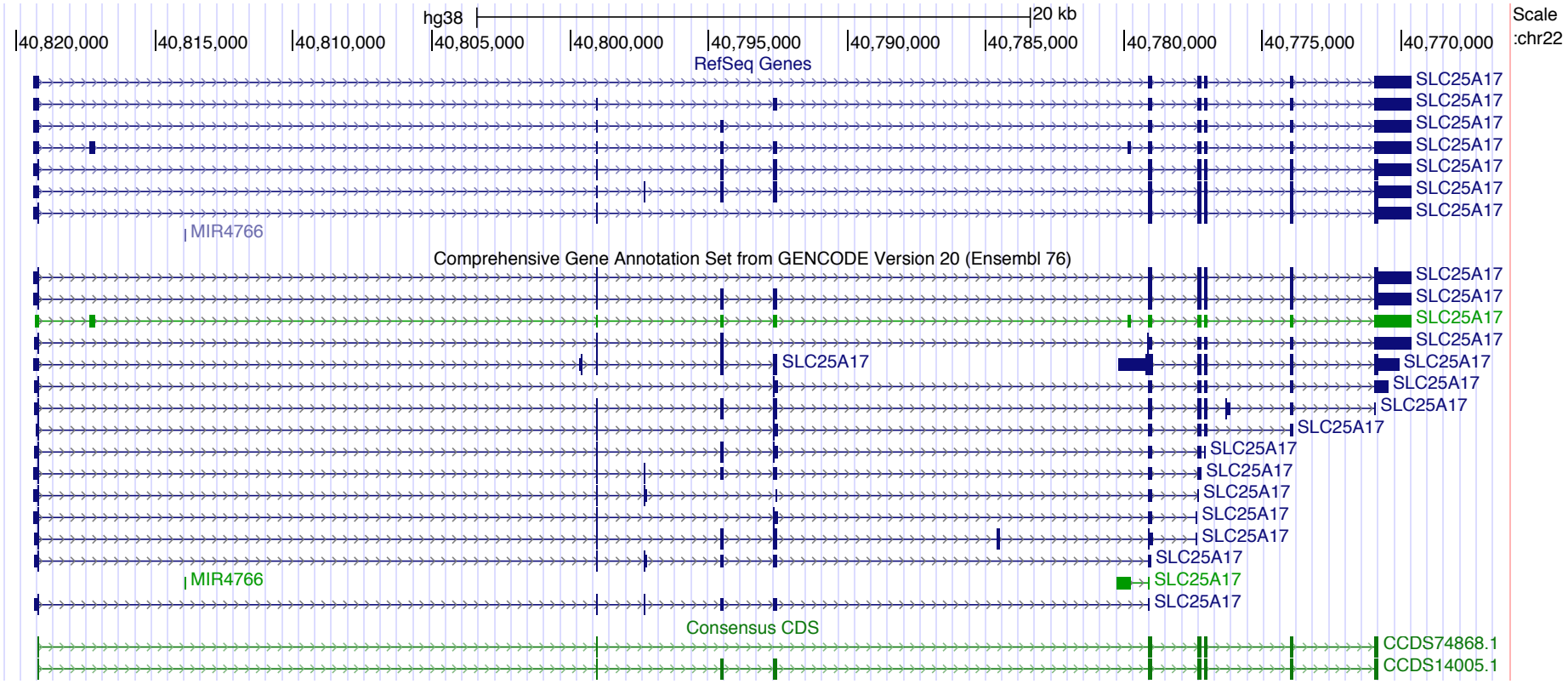
GENCODE gene annotation



The GENCODE consortium

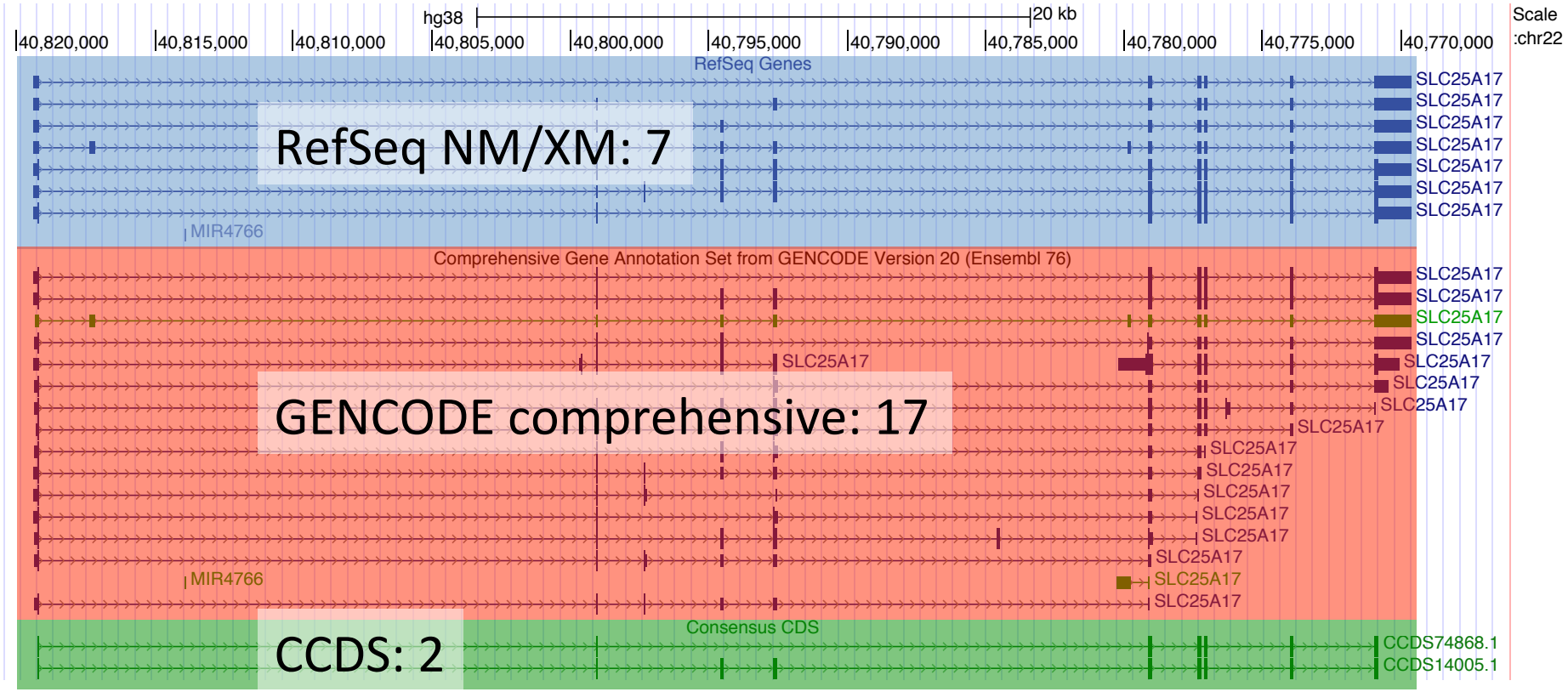


GENCODE and RefSeq are different

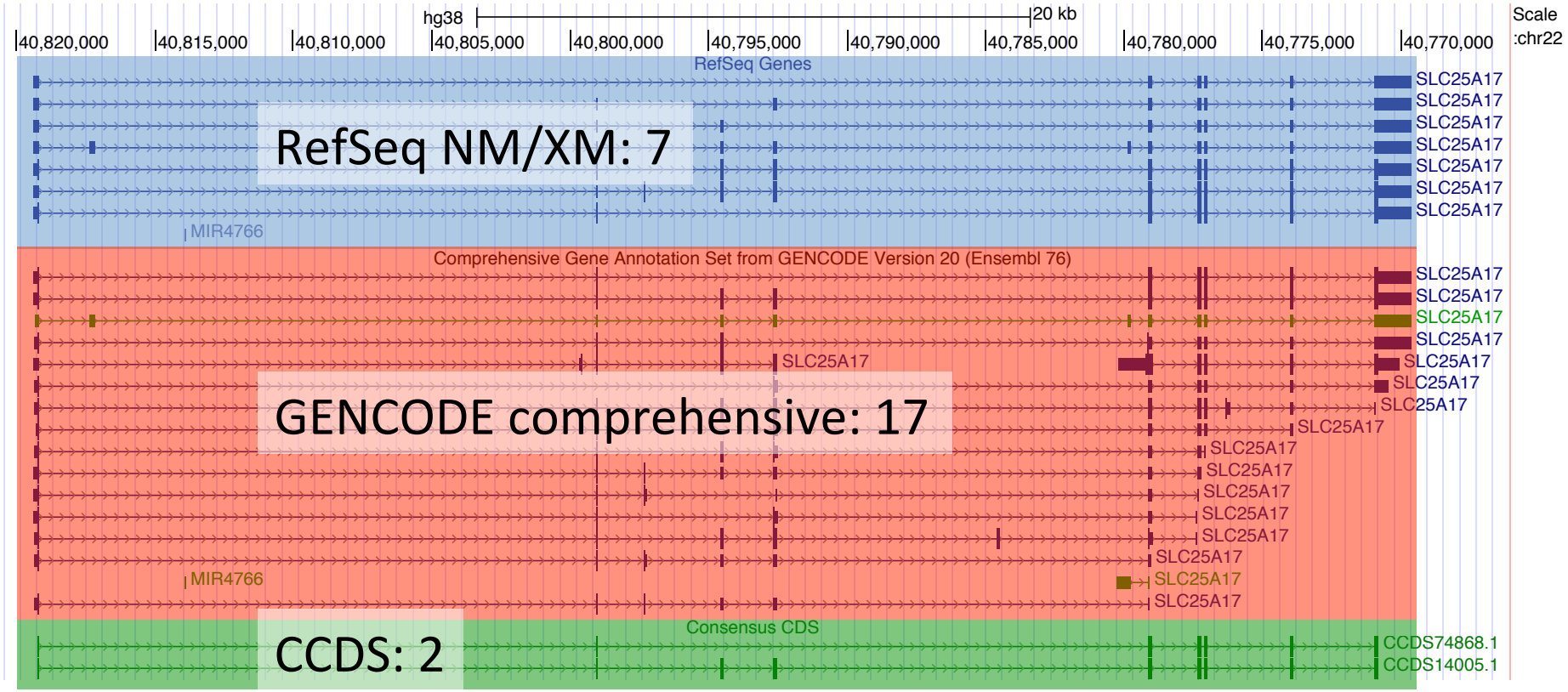


SLC25A17 on GRCh38

GENCODE has more alternative splicing



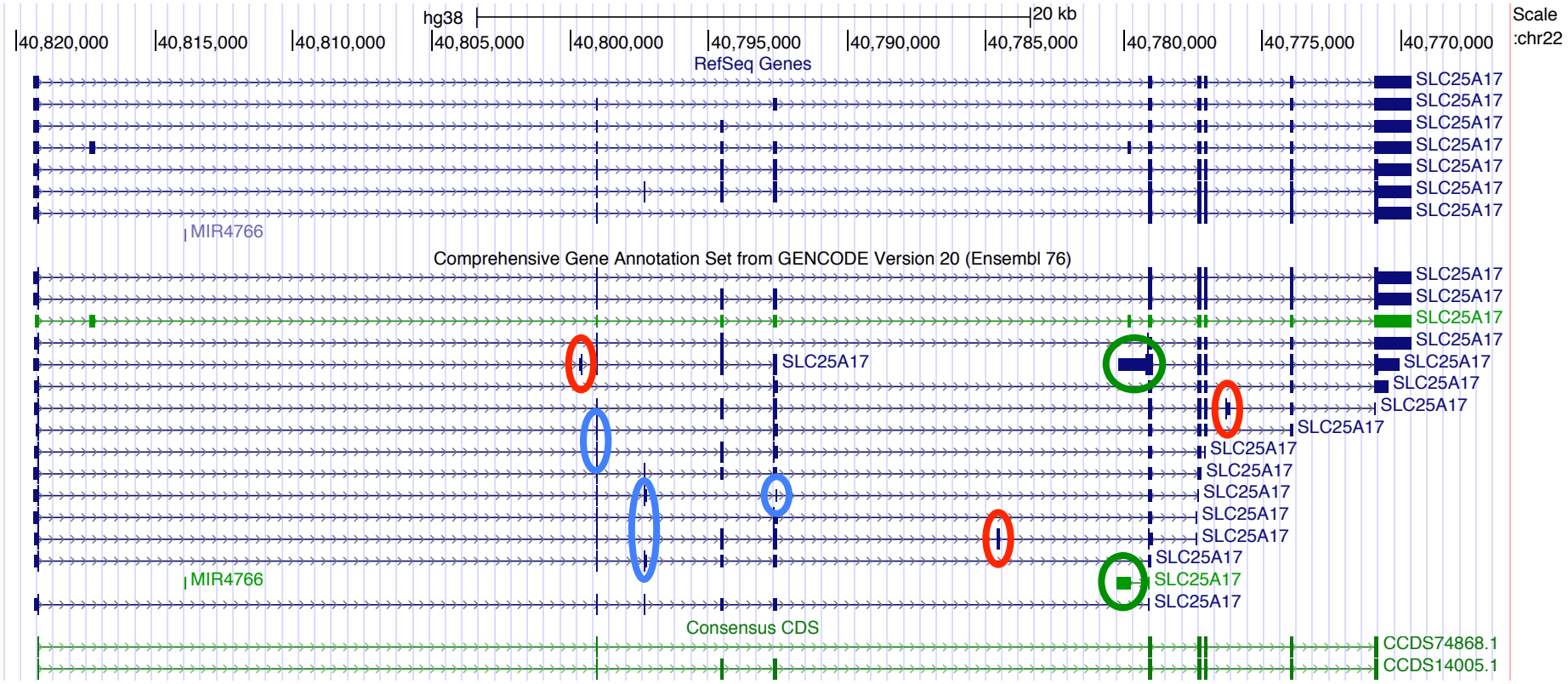
GENCODE has more alternative splicing



For each multi-exon protein-coding locus:

- GENCODE comprehensive – 7.6 transcripts, 4.3 unique translations
- RefSeq NM/XM – 3.4 transcripts, 2.7 unique translations
- RefSeq NM – 1.9 transcripts, 1.7 unique translations

GENCODE has more unique exonic features

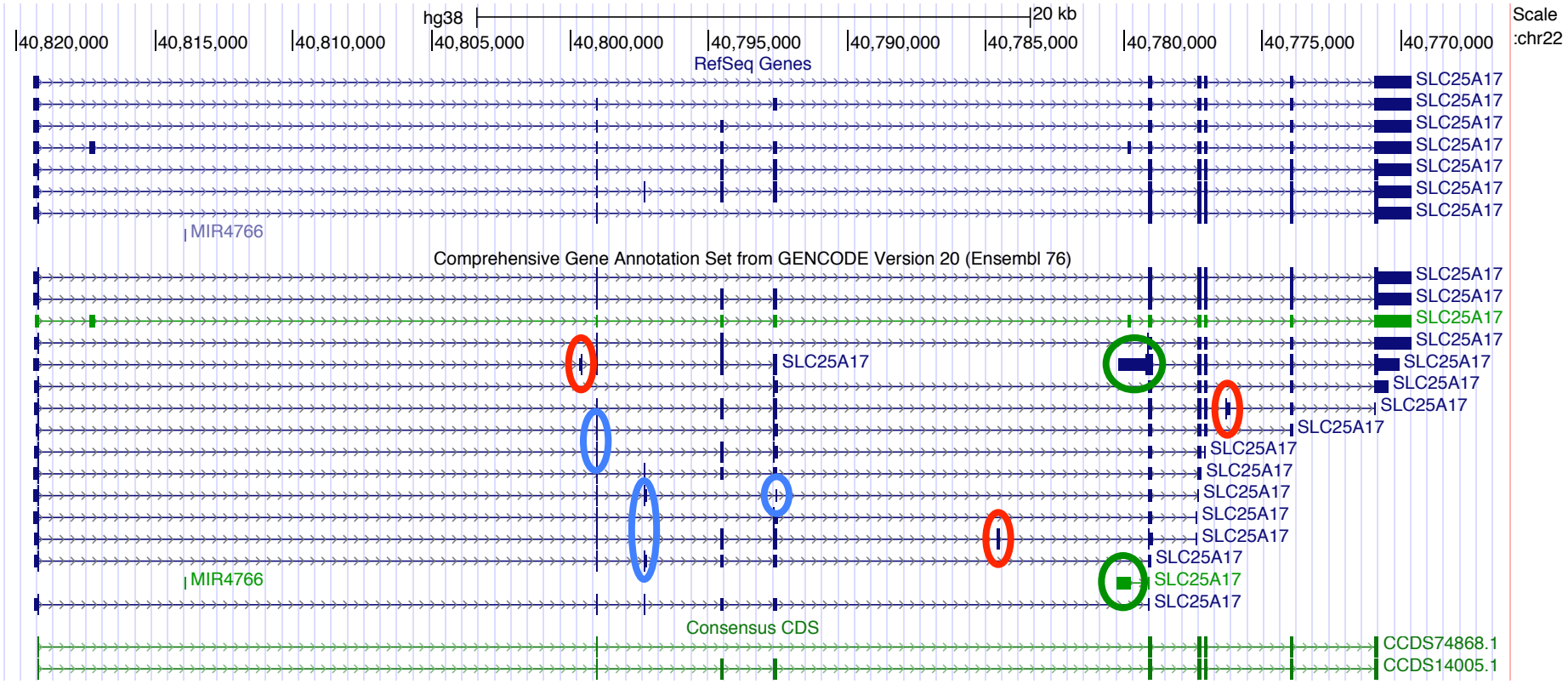


Novel cassette exon

Novel alternative splice site

Novel putative TSS, 5’ UTR

GENCODE has more unique exonic features



Novel cassette exon

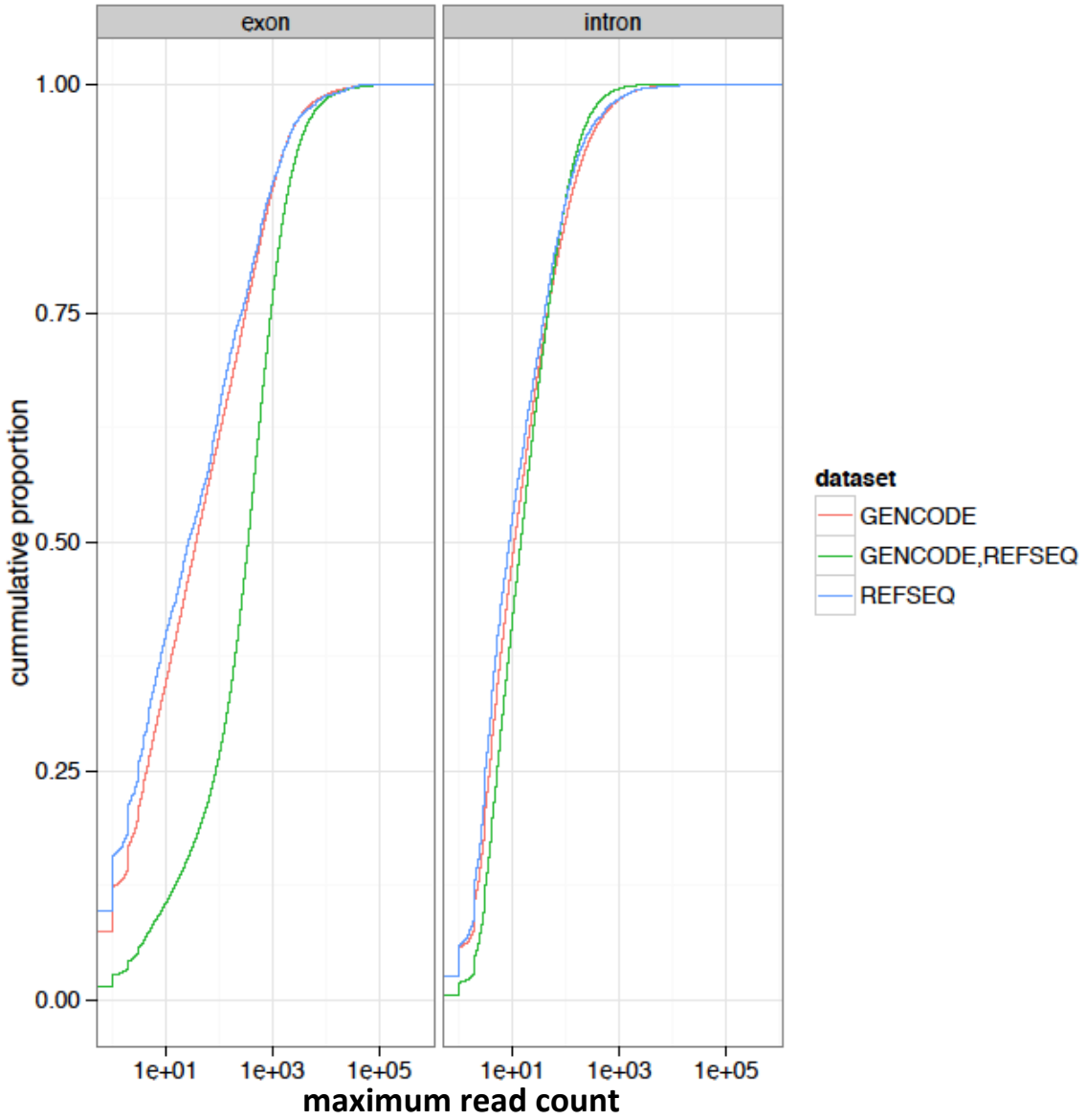
Novel alternative splice site

Novel putative TSS, 5' UTR

Genomic coverage of exons at protein-coding loci:

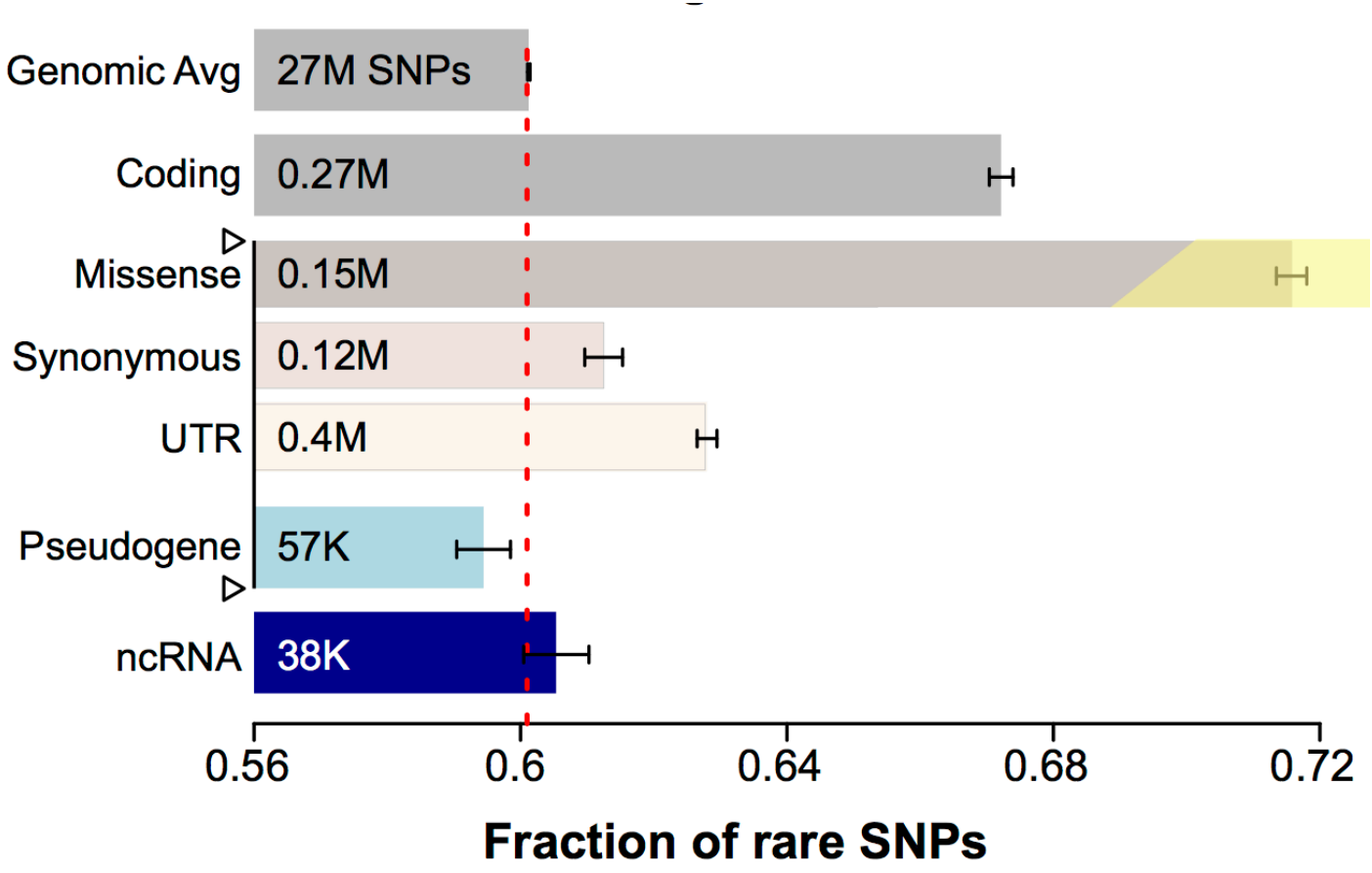
- GENCODE ~1.3%
- RefSeq NM/XM ~1%

Expression of unique exons and introns is similar for GENCODE and RefSeq



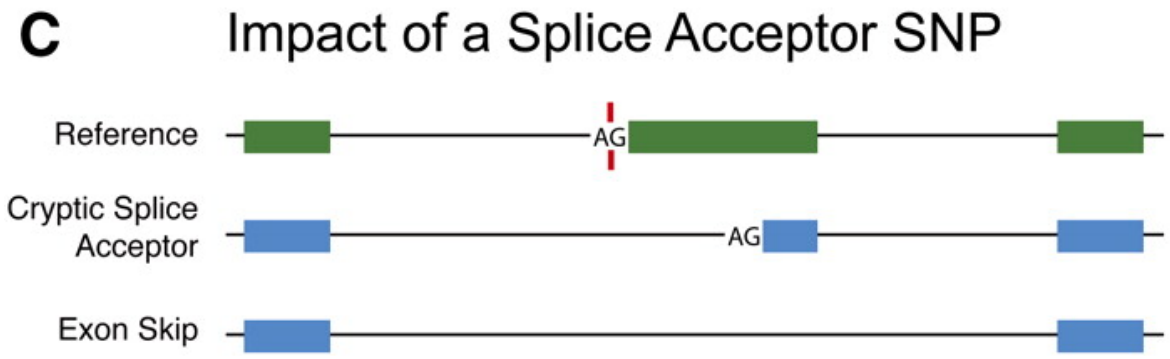
Gencode & Variants

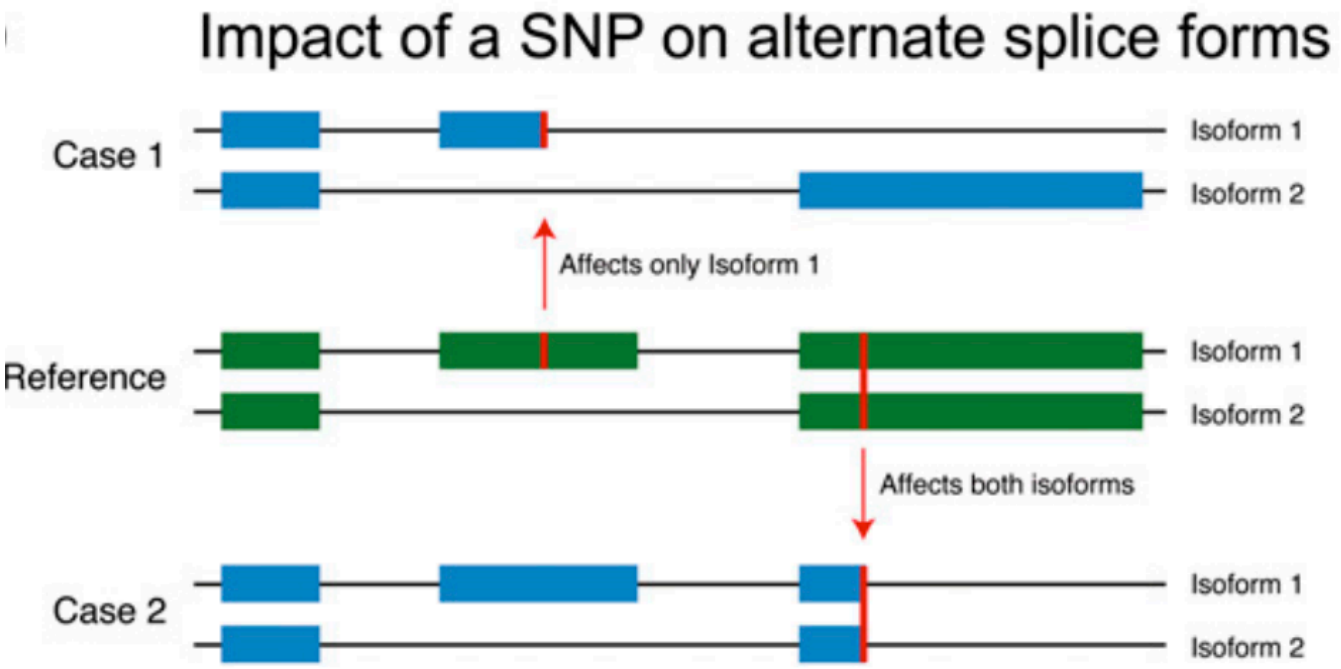
Gencode **consistently** annotates CDSes in the context of UTRs, Pseudogenes & ncRNAs, so can uniformly look at the effect of variants on these



[Khurana et al., *Science* ('13)]

One has to consider transcript structure to assay impact of variants on genes





[Balasubramanian et al., *Genes Dev.*, '11]

Reference transcript set affects consequence calling

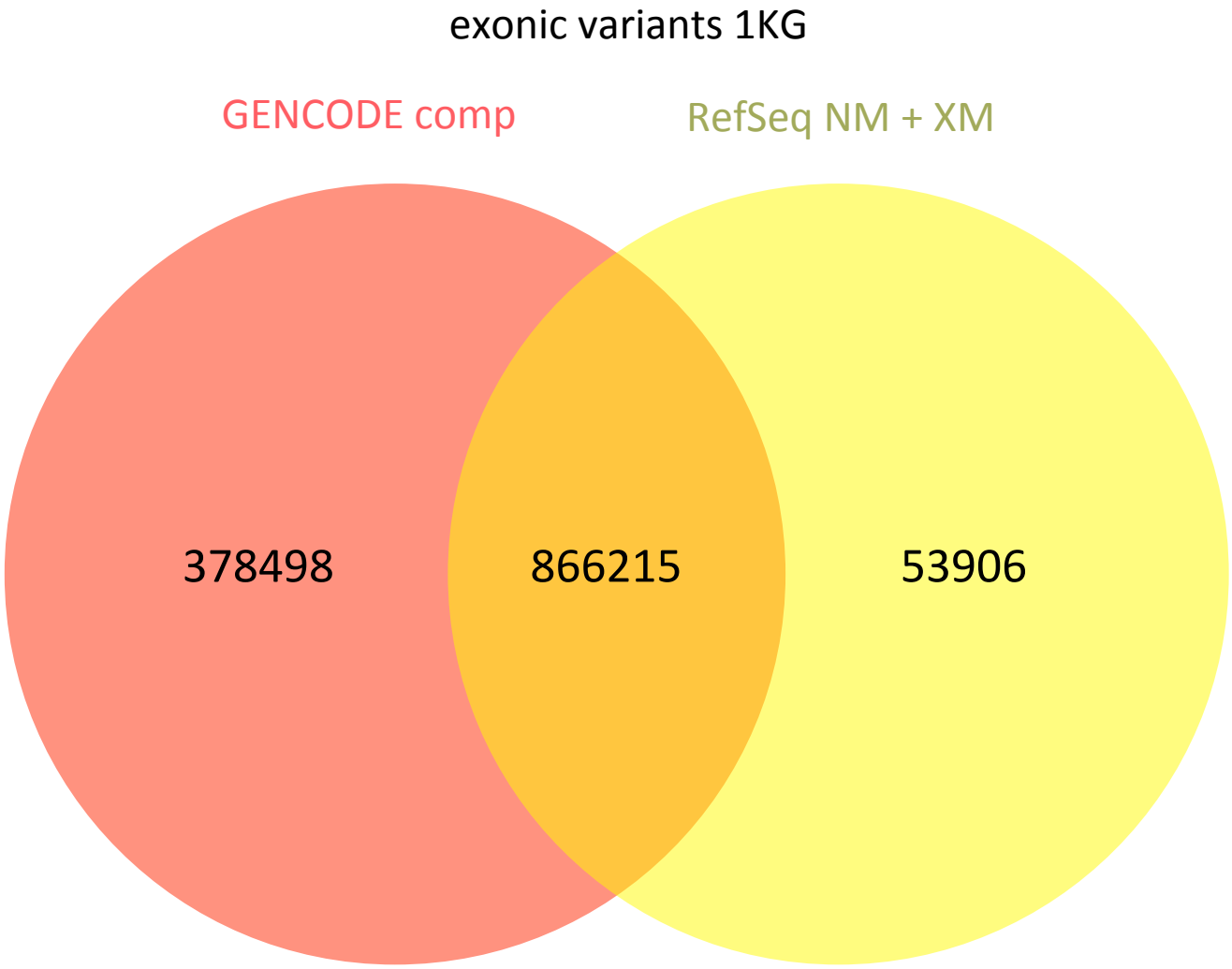
McCarthy *et al. Genome Medicine* 2014, **6**:26
http://genomemedicine.com/content/6/3/26

RESEARCH Open Access

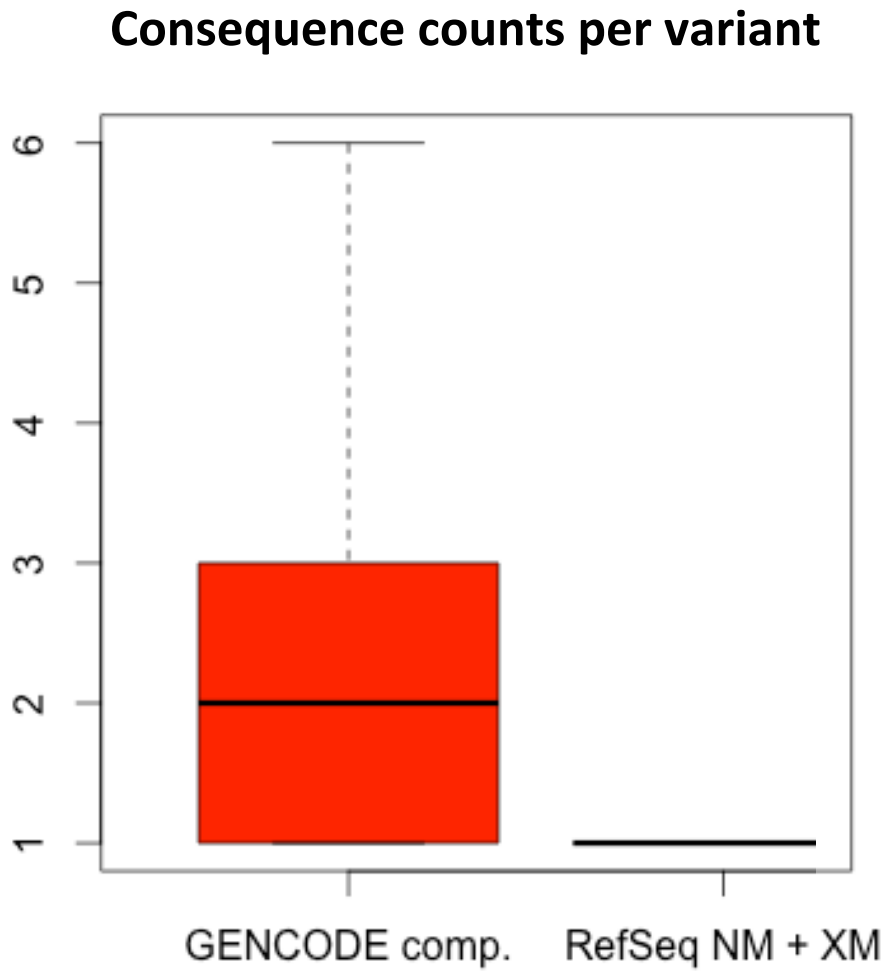
Choice of transcripts and software has a large effect on variant annotation

Davis J McCarthy[1,2*], Peter Humburg[2], Alexander Kanapin[2], Manuel A Rivas[2], Kyle Gaulton[2], The WGS500 Consortium, Jean-Baptiste Cazier[3] and Peter Donnelly[1,2]

More variants intersect with GENCODE transcripts

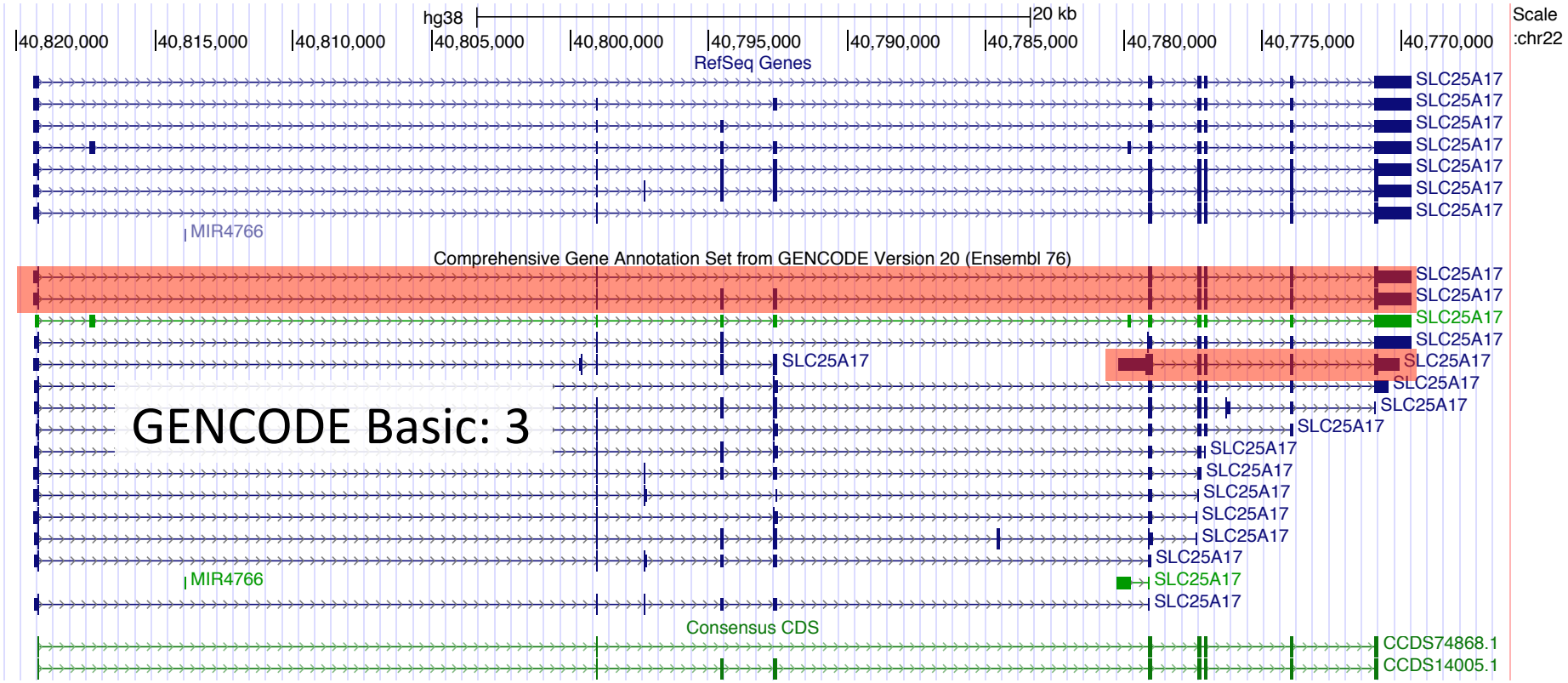


Transcript complexity makes calling variant consequence more difficult



Simplifying Gencode

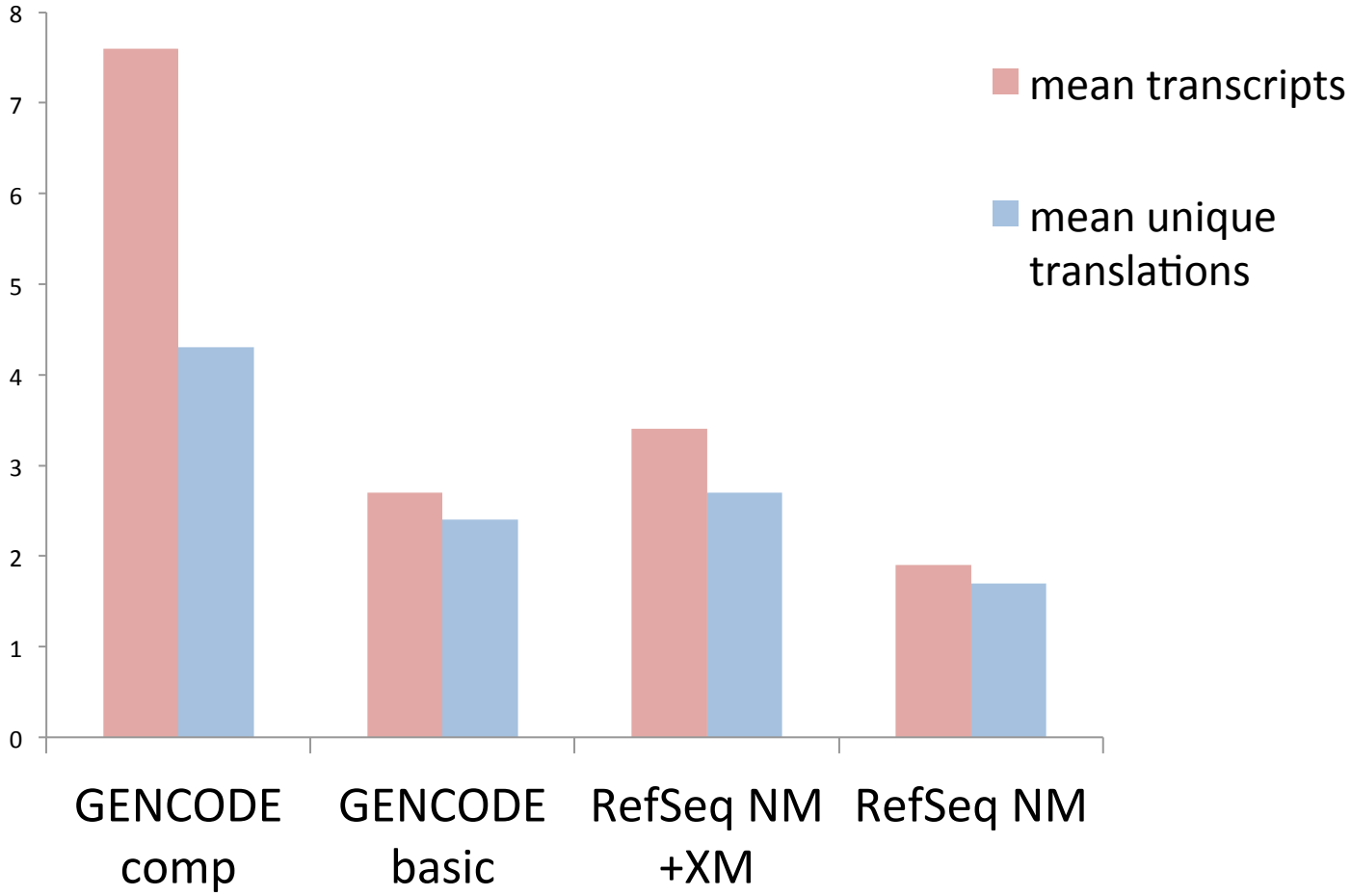
GENCODE Basic set reduces transcript complexity



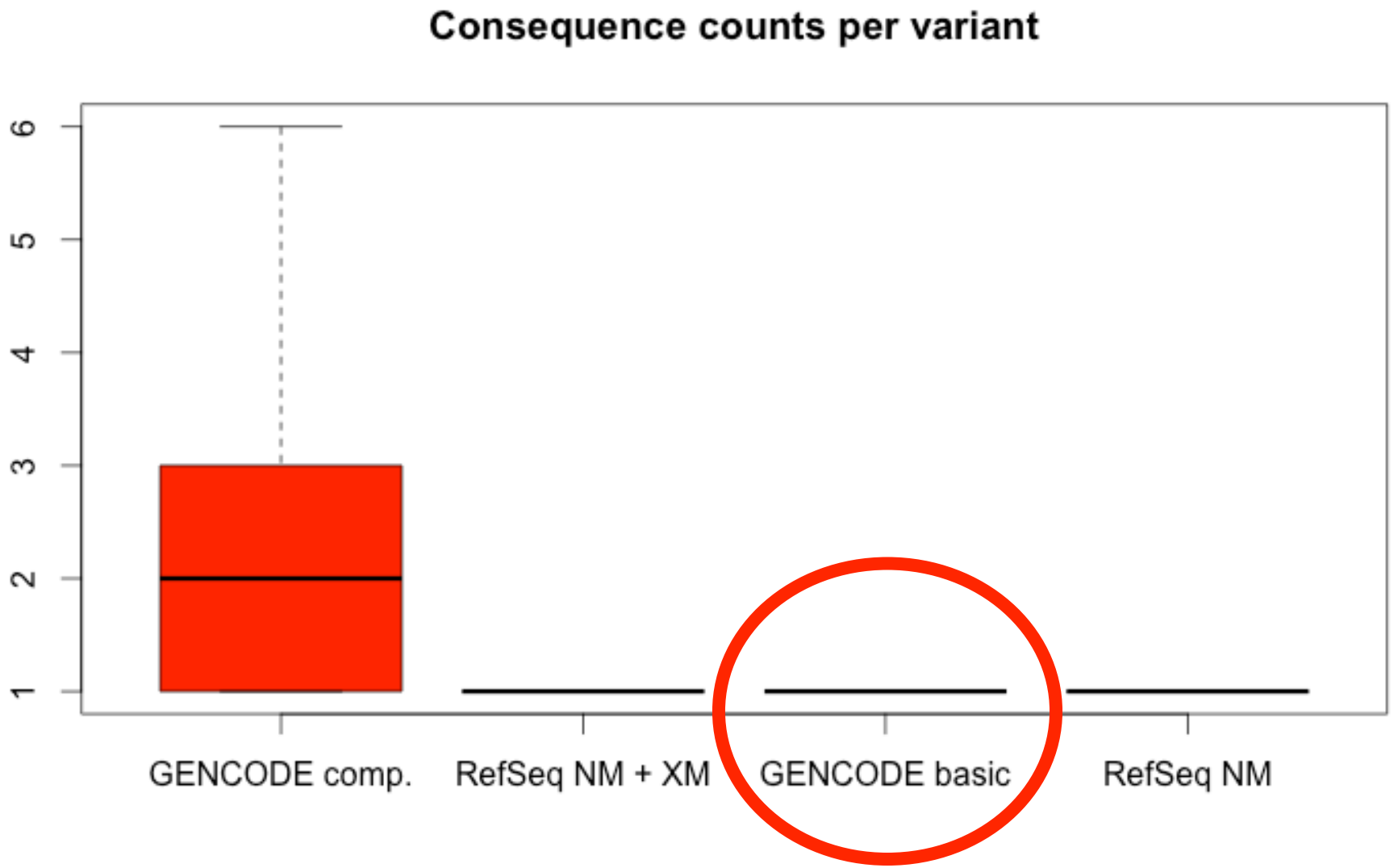
At protein-coding locus GENCODE basic transcript set contains:

- only protein-coding transcripts
- only transcripts with a full-length CDS (ATG to stop)

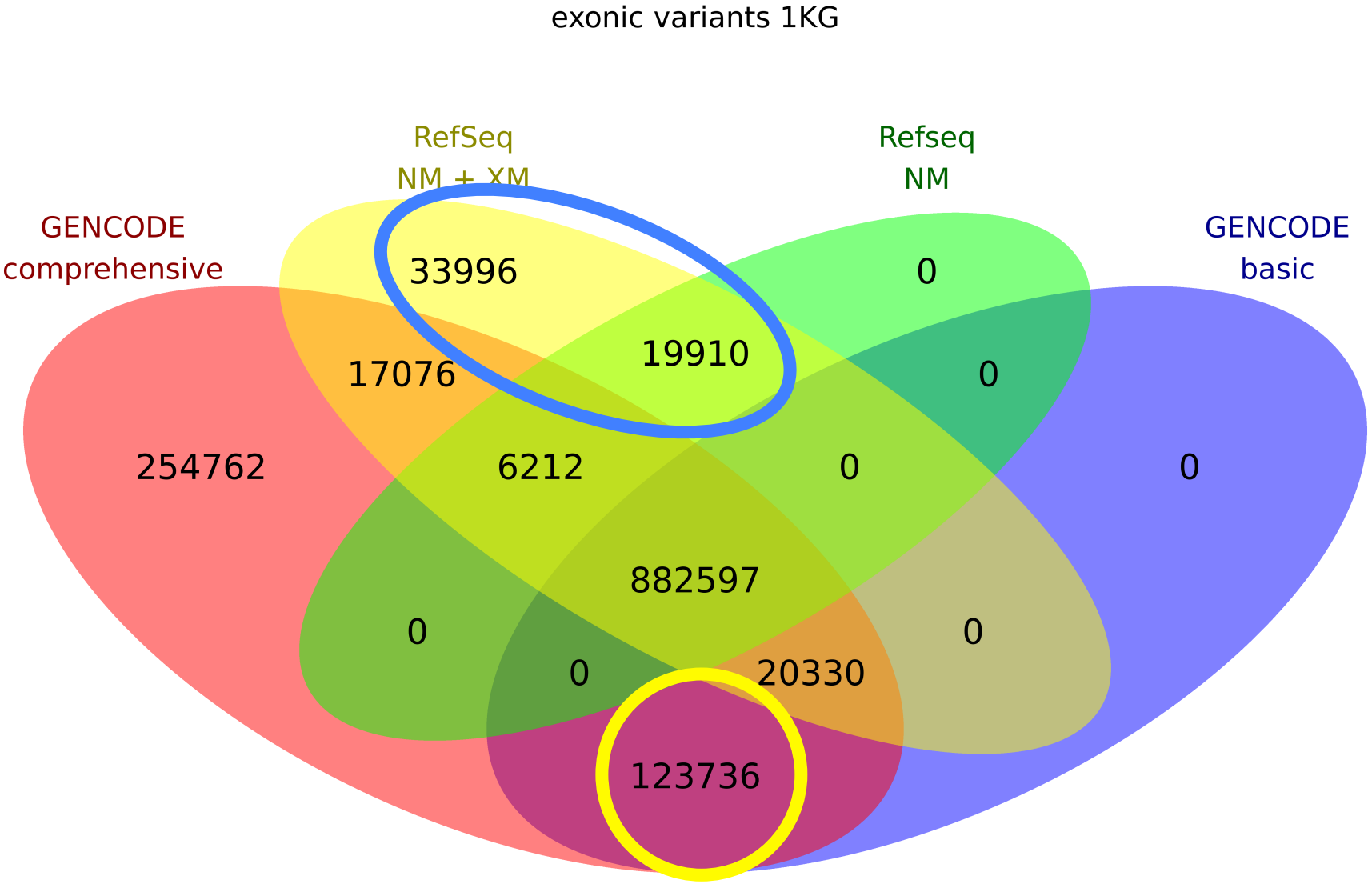
GENCODE Basic shares characteristics with RefSeq genesets



GENCODE basic reduces complexity in variant consequence calling



GENCODE basic still gives greater coverage of variation than RefSeq



Transcript level quantification unreliable

Accurate Identification and Analysis of Human mRNA Isoforms Using Deep Long Read Sequencing

Hagen Tilgner,* Debasish Raha,† Lukas Habegger,‡ Mohammed Mohiuddin,§ Mark Gerstein,‡ and Michael Snyder*,1

*Department of Genetics, Stanford University, Stanford, California 94305, †Department of Molecular, Cellular and Developmental Biology, and ‡Program in Computational Biology and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06120, and §Roche, Branford, Connecticut 06405

Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, The RGASP Consortium, Tim J Hubbard, Roderic Guigó, Jennifer Harrow & Paul Bertone

Affiliations | Contributions

Nature Methods **10**, 1177–1184 (2013) | doi:10.1038/nmeth.2714

Received 31 March 2013 | Accepted 23 September 2013 | Published online 03 November 2013

But can we identify most highly expressed transcripts?

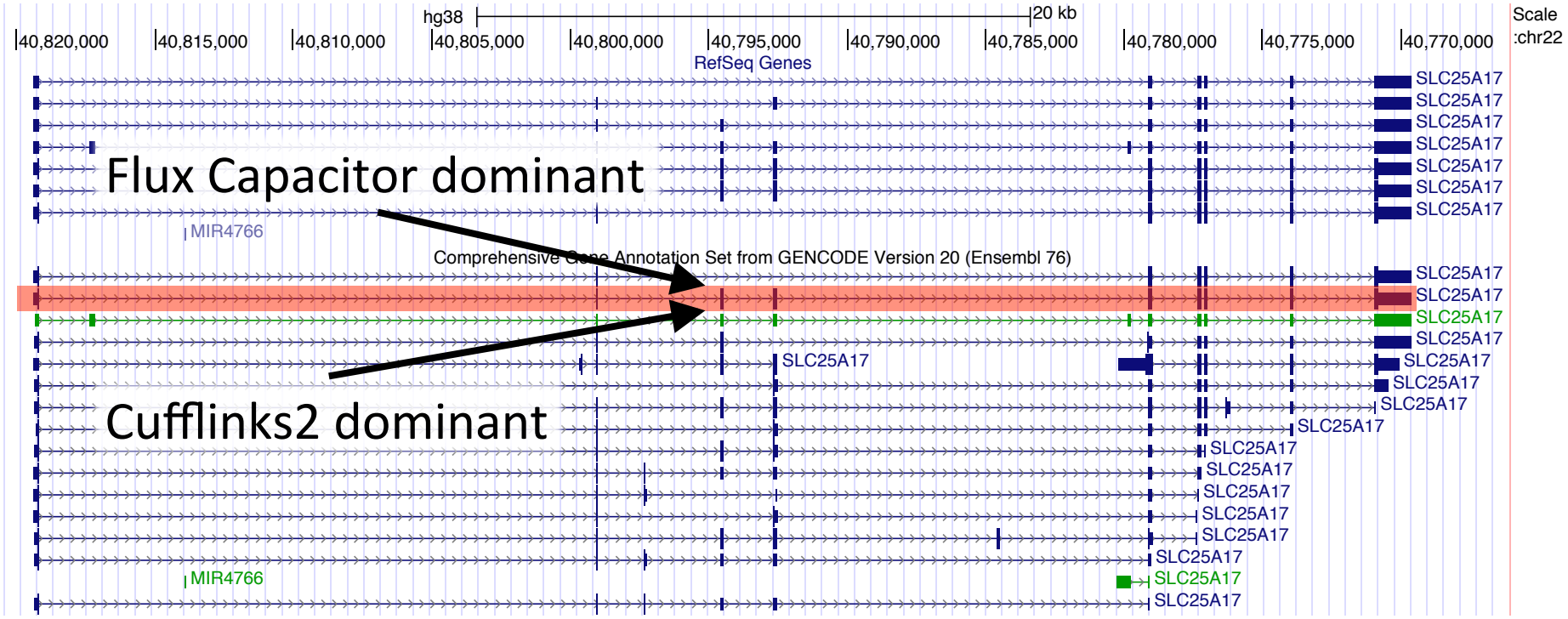
Gonzàlez-Porta *et al. Genome Biology* 2013, **14**:R70
http://genomebiology.com/2013/14/7/R70

RESEARCH Open Access

Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene

Mar Gonzàlez-Porta[1], Adam Frankish[2], Johan Rung[1], Jennifer Harrow[2] and Alvis Brazma[1*]

Identifying dominant transcripts to reduce transcript complexity



Transcript quantification with Flux Capacitor and Cufflinks2

- both find a dominant transcript (DT) for only 50-60% of all genes
- where both declare a DT agree ~85% of cases (5x dominance)
- where agree DT, 75%-90% are in GENCODE basic
- where disagree on DT, up to 50% are in GENCODE basic (>0x dominance)

Summary

- GENCODE aims to capture as much biological complexity as possible
- This facilitates consistent discovery of genic variants but complicates their interpretation
- GENCODE basic transcripts are intended to bridge this

Acknowledgements

WTSI

Adam Frankish

Jen Harrow *et al.*

Anacode team

Annotools team

EBI

Graham Ritchie

Robert Petryszak

Bronwen Aken *et al.*

CNIO, Madrid

Alfonso Valencia

Michael Tress

CRG, Barcelona

Roderic Guigo

Dimitri Pervouchine

Barbara Uszczynska

University of Lausanne

Alex Reymond *et al.*

UCSC

Mark Diekhans

Rachel Harte

MIT

Manolis Kelis *et al*.

Yale

Mark Gerstein *et al*.

CSHL

Tom Gingeras *et al.*

Accessing GENCODE

Gencodegenes.org: http://www.gencodegenes.org/data.html

UCSC Browser: http://genome.ucsc.edu/cgi-bin/hgGateway

ENSEMBL Browser: http://www.ensembl.org/Homo_sapiens/Info/Index

Vega Browser: http://vega.sanger.ac.uk/Homo_sapiens/Info/Index