Sequence Alignment-Based *In Silico* Pathogenicity Predictions for APC Missense Variants

Marc S. Greenblatt, M.D.

University of Vermont Larner College of Medicine

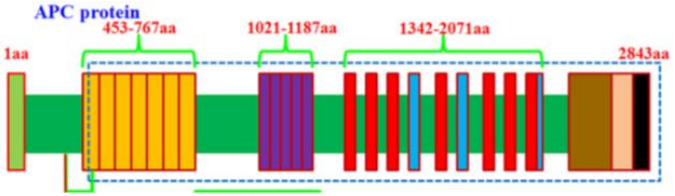
Acknowledgements and Disclosures

- University of Vermont
  - Alexander Karabachev
  - Dylan Martini
  - David Hermel
  - Dana Solcz
- Brown University
  - Indra Neil Sarkar
- Ambry Genetics
  - Tina Pesaran
  - Marcy Richardson

Introduction

- *APC* variants causing FAP are loss of function, usually frameshift or nonsense
- Genetic testing frequently reveals APC missense substitutions, usually classified as Variants of Unknown Significance

Missense pathogenic variants are rare



Many functional domains are repeated

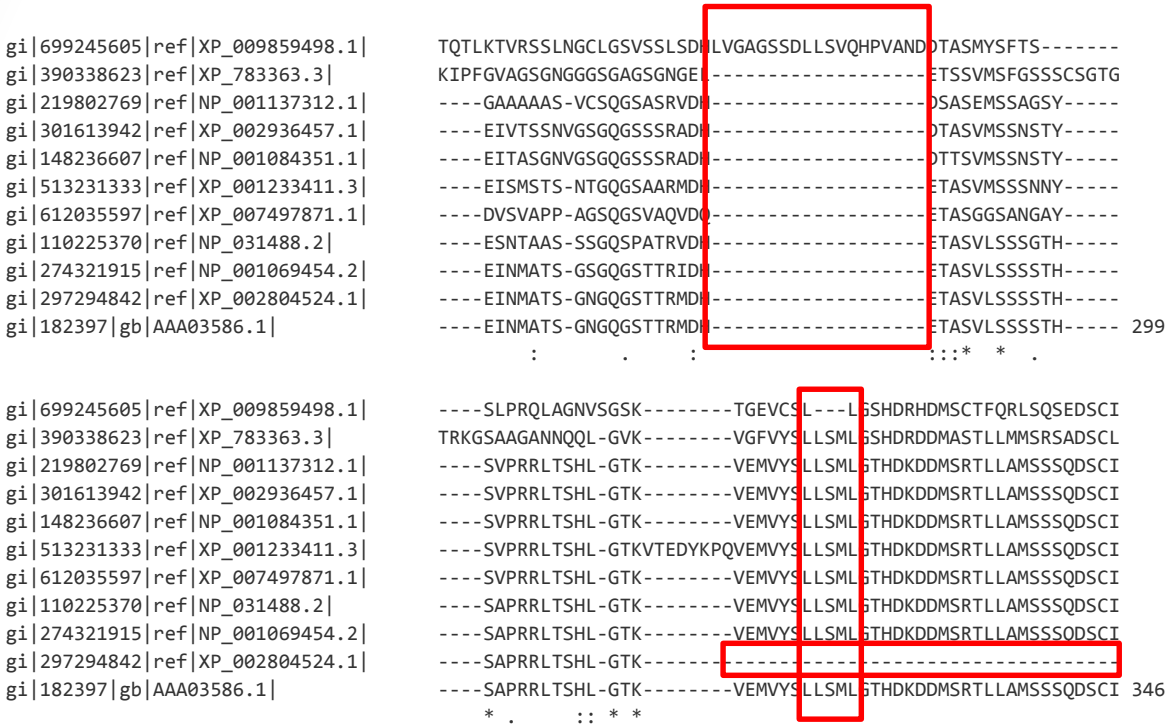
- *In silico* methods based on evolutionary sequence conservation are often used in hereditary cancer genes to help predict pathogenicity
- Unclear whether these *in silico* algorithms are useful in APC and other genes where non-missense mechanisms predominate- but they are often reported
- More data are needed to help interpret sequence-based methods for APC and less well-understood genes

Computational Analysis

- *In silico* tools generate predictions of pathogenicity using:
  - Physiochemical characteristics of the substitution
  - Evolutionary conservation
  - Crystal structure if known
- Validated for BRCA1/2, MMR genes, TP53, not other genes
- Start with Protein Multiple Sequence Alignment (PMSA)

```
Sea Squirt          EYKVEDTPACFTPRSAISDLPCEEEDD
Sea Urchin          TYCVEGTPGPISRCSSLSSLDLNEELD
Zebrafish           TYCVEDTPICFSRGSSLSSLSSEEDEM
Frog                TYCVEDTPICFSRGSSLSSLSSAEDEI
Chicken             TYCVEDTPICFSRCSSLSSLSSAEDEI
Opossum             TYCVEDTPICFSRCSSLSSLSSAEDEI
Mouse               TYCVEDTPICFSRCSSLSSLSSADDEI
Cow                 TYCVEDTPICFSRCSSLSSLSSAEDEV
Monkey              TYCVEDTPICFSRCSSLSSLSSAEDEI
Human               TYCVEDTPICFSRCSSLSSLSSAEDEI
                     * **.**  ::  *::*.*   ::
```

PMSAs Require Much Curation



Gaps or Insertions may represent:

- True deletions and insertions
- Misreading of intron-exon boundaries
- Other sequence anomalies from genome assembly

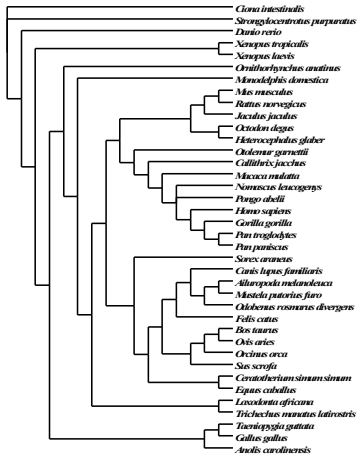
Our Goals

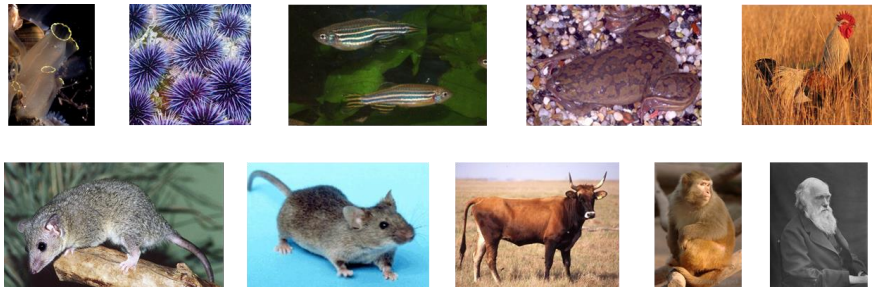
- Define the challenges of using *in silico* alignment-based tools for a protein with few pathologic missense variants.
- Create high quality PMSA for APC
- Generate predictions of pathogenicity for APC variants from *in silico* methods
- Compare to clinical classifications

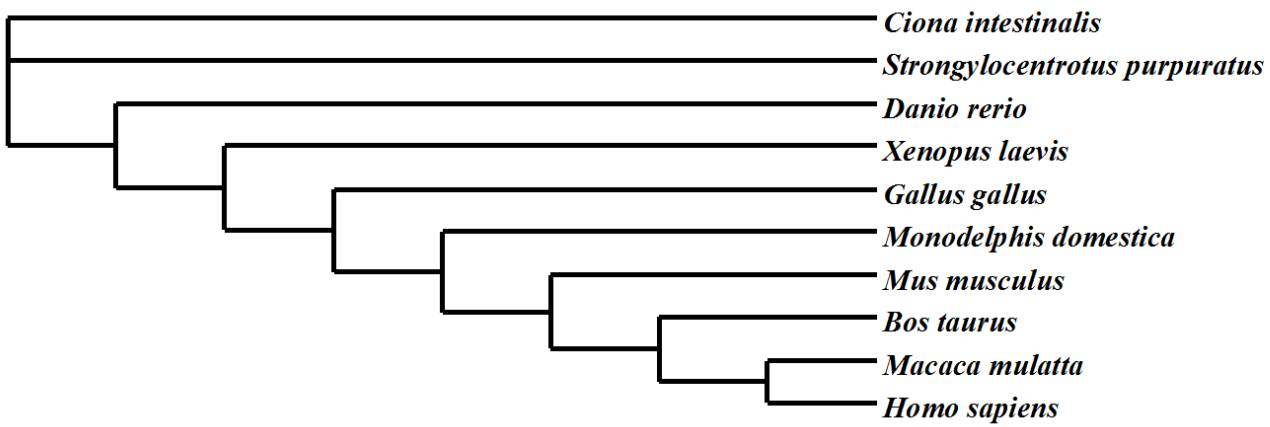
Methods for Phylogenetic Sequence Analysis To Validate Alignments Applied to APC

- Phylogenetic Trees constructed from 38 species, subsets

- ≥10 species is enough APC evolutionary variation for statistically significant predictions (Greenblatt 2003, Cooper 2003)







Curating the APC PMSA

- Created a PMSA using Clustal Omega
  - 10 species with evolutionary depth to sea squirt
  - Manually curated segments of the alignment (took some effort)
  - BLAST searches, removed exons that did not align with human
  - Located exon-intron boundaries, found omitted exons

```
Sea Squirt      SGRILTNLTYADNLNKVLLMNMRGLLETVRDQLQHESEEIQQAMASILRNLSWQADKEGR   479
Sea Urchin      AGMALTNLTFGDVTNKALLCSMKGCMKALVALLSAESEDLRQVAASVLRNLSWRADMASK   502
Zebrafish       AGMALTNLTFGDVANKATLCSMKGCMRAMVAQLKSESEDLQQVIASVLRNLSWRADVNSK   554
Frog            AGMALTNLTFGDVANKATLCSMKSCMRALVAQLKSESEDLQQVIASVLRNLSWRADVNSK   562
Chicken         AGMALTNLTFGDVANKATLCSMKGCMRALVAQLKSESEDLQQVIASVLRNLSWRADVNSK   560
Opossum         AGMALTNLTFGDVANKATLCSMKGCMRALVAQLKSESEDLEQVIASVLRNLSWRADVNSK   560
Mouse           AGMALTNLTFGDVANKATLCSMKGCMRALVAQLKSESEDLQQVIASVLRNLSWRADVNSK   558
Cow             AGMALTNLTFGDVANKATLCSMKGCMRALVAQLQSESEDLQQVIASVLRNLSWRADVNSK   560
Monkey          AGMALTNLTFGDVANKATLCSMKGCMRALVAQLKSESEDLQQVIASVLRNLSWRADVNSK   560
Human           AGMALTNLTFGDVANKATLCSMKGCMRALVAQLKSESEDLQQVIASVLRNLSWRADVNSK   560
                :*  *****:.*  **. * .*:. :.::   *. ***::.*. **:******:**  .:

Sea Squirt      DLLRQKGVVRTLTECVICAKGEGTLKAMLSALWNLSGHCPANREDICSVEGSLAFLVSTL   539
Sea Urchin      KALREAGAVVALMTCSLEVKKESTLKSVLSALWNLSAHCTENKADICAVNGALEFLVSSL   562
Zebrafish       KILREVGSVRALMECALEVQKESTLKSVLSALWNLSAHCTENKADICTVPGALAFLVSTL   614
Frog            KTLREVGSVKALMECALDVKKESTLKSVLSALWNLSAHCTENKADICSVDGALAFLVSTL   622
Chicken         KTLREVGSVKALMECALEVKKESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTL   620
Opossum         KTLREVGSVKALMECALEVKKESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTL   620
Mouse           KTLREVGSVKALMECALEVKKESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTL   618
Cow             KTLREVGSVKALMECALEVKKESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTL   620
Monkey          KTLREVGSVKALMECALEVKKESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTL   620
Human           KTLREVGSVKALMECALEVKKESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTL   620
                . **: * * :*  * : .: *.***::********.**  *: ***:* *:* ***.:*
```

APC Missense Variants Found in ClinVar

- From 2013 to 2018 the total number of reported APC missense variants increased from N= 47 → N= 1,988

- Missense variants in ClinVar, 5/2018, classifications:
  - Benign – 22 (1.1%)
  - Pathogenic – 9 (0.4%)
    - 2 somatic
    - 7 at splice sites→ **not** "Likely Pathogenic" due to missense
  - Uncertain significance – 1806 **(93.2%)**
  - Conflicting interpretations of Pathogenicity – 103 **(5.3%)**

- **98.5% of APC missense classifications in ClinVar are not clinically useful.**

In silico Predictions of Pathogenicity for APC Missense Variants

	ClinVar Benign Variants (N=22)	ClinVar Variants of Uncertain Significance (N=1903)
Method	**Total Accuracy**	**Predicted Neutral (%)**
REVEL	100%	N/A
A-GVGD	100%	82.5%
SIFT	95.5%	68.1%
PolyPhen2	81.8%	41.0%
MAPP	77.8%	25.0%

- Two likely pathogenic (LP) variants found in ClinVar p.S1028N, p.N1026S
- Classified as LP in July 2018 by Ambry Genetics using ACMG/AMP guidelines
- Evidence: protein features, segregation, published functional data
- Both located in the first 15-amino acid repeat of the β-catenin binding domain

Conclusions

- Creating a high quality PMSA is labor intensive and a limiting factor in using *in silico* predictive tools
- *In silico* methods that are excellent classifiers for variants of some hereditary cancer genes (BRCA1/2, MMR, TP53) are not as accurate when applied to APC and some other genes
- Some APC features may predict why *in silico* methods perform poorly; these features may be common in other genes
- Systematic study of these features can improve predictive algorithms, interpretation of inherited genetic variation