
Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses.

2009. Cui et al. *PNAS*

Presenter: Vikas Rao Pejaver

10/14/2009

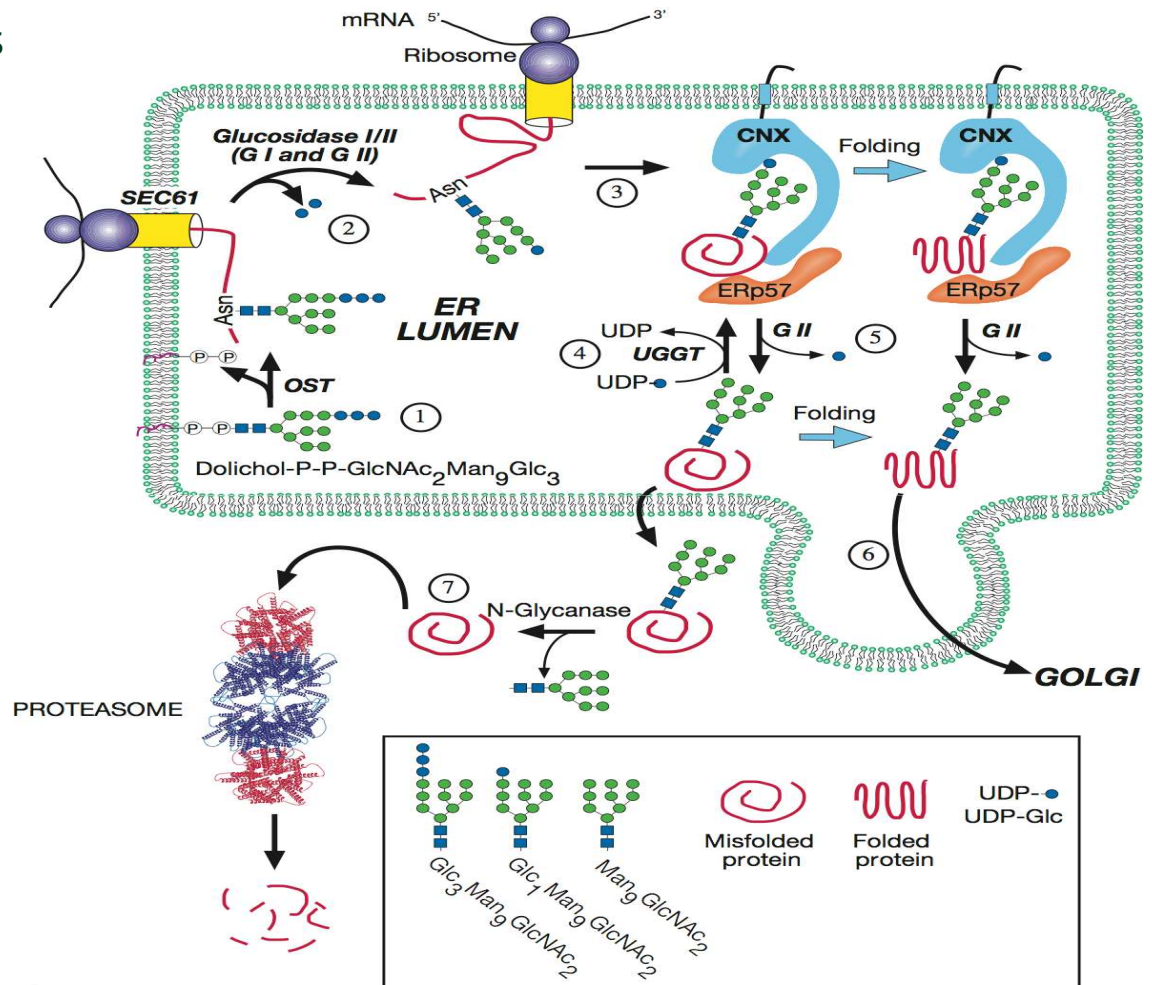
Outline

- Background
 - Motivation
 - Results
 - Methods
 - Conclusion
 - Points for Open Discussion
-

Background

N-glycan-dependent QC

- N-glycans help glycoproteins fold by creating a series of checkpoints that dictate whether certain membrane and secreted proteins are to be degraded or not
- Modification of the physical properties of glycoproteins by providing noncharged, bulky, hydrophilic groups that keep the protein in solution during folding
- In vertebrates, addition of the N-glycan occurs on incompletely folded polypeptides
- In bacteria, N-glycosylation can occur both on nascent polypeptides and on fully mature proteins



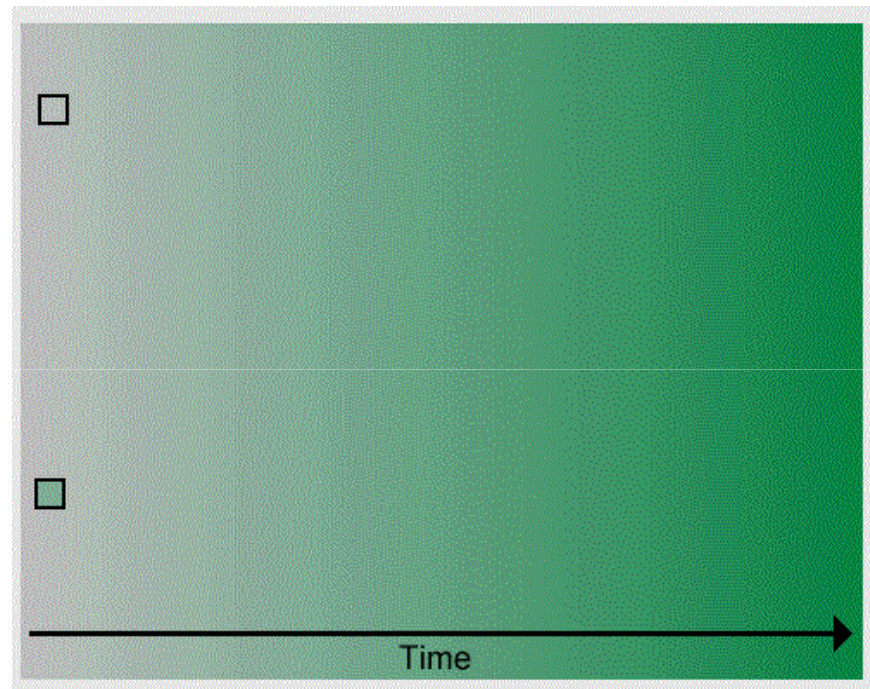
Basic terminology

- **Sequons:**

A sequon is a sequence of three consecutive amino acids in a protein that can serve as the attachment site to an N-glycan

Generally, **Asn-X-Ser** or **Asn-X-Thr** (sometimes **Asn-X-Cys**), where X is any amino acid except proline

- **Darwinian selection:**



Individuals that are more capable or 'fit' leave more offspring and this varying reproductive success of individuals based on their different genetic constitutions is natural selection

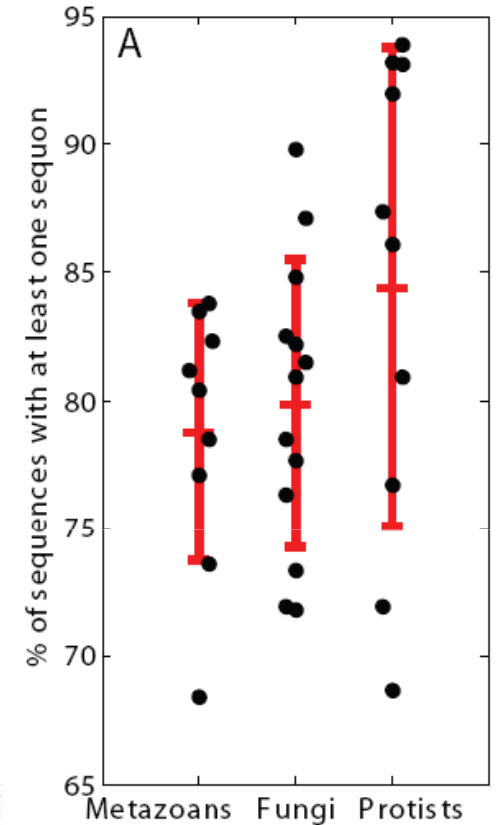
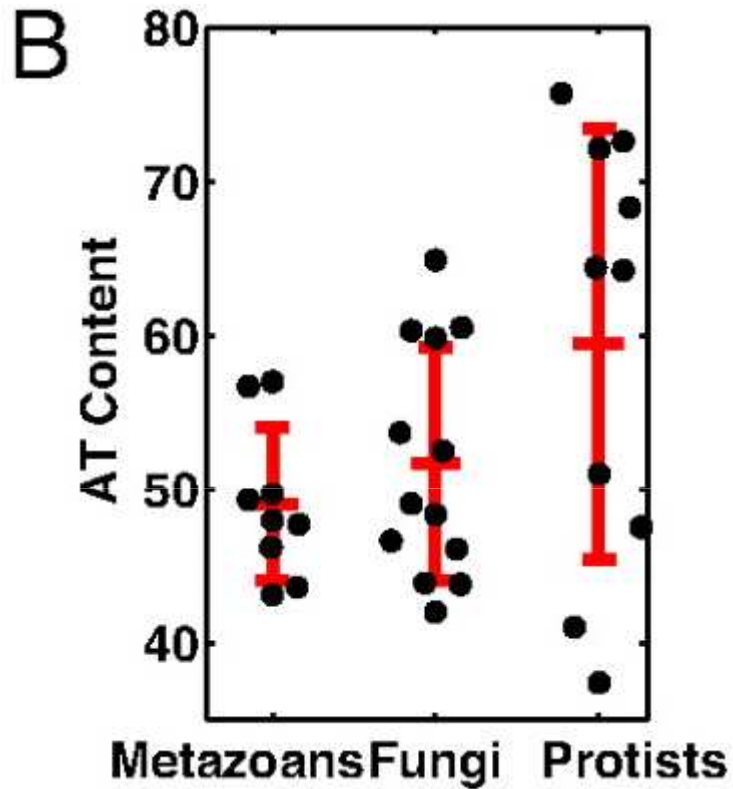
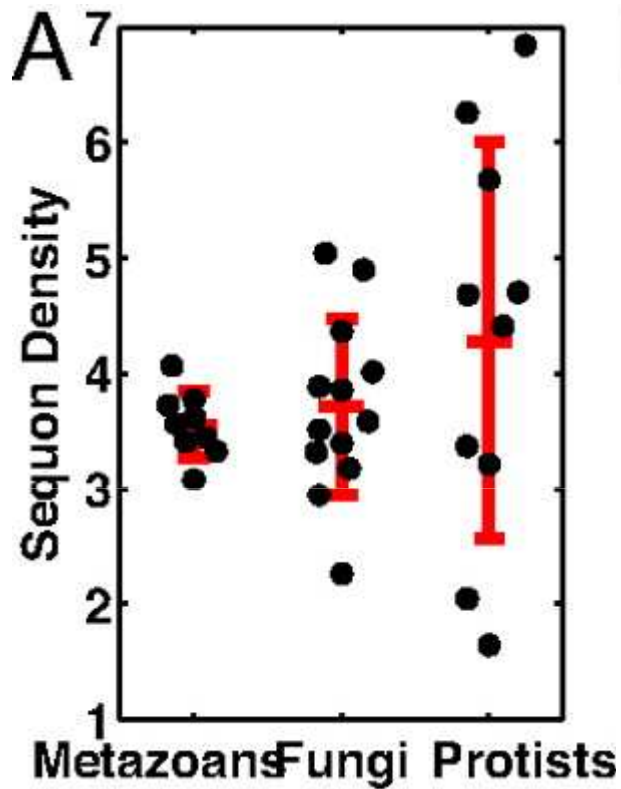
Motivation

A lot of related questions!

1. Is there a relation between N-glycan length and the density of sequons in their secreted proteins? If yes, how are they related?
 2. How are AT content (which is variable among protists) and sequon abundance related?
 3. Is there Darwinian selection for sequons in secreted and membrane proteins (vs cytosolic proteins)? If so, who's responsible:
 - i. Increased AT content?
 - ii. Increased Asn, Ser & Thr and decreased Pro in secreted proteins?
 - iii. Increased conditional probability that the above three amino acids will be present in sequons rather than elsewhere in the above proteins?
 4. What is the explanation behind the occurrence of high densities of sequons in influenza viruses (hemagglutinin - HA) and HIV (gp120)?
-

Results

Understanding the nature of the data:

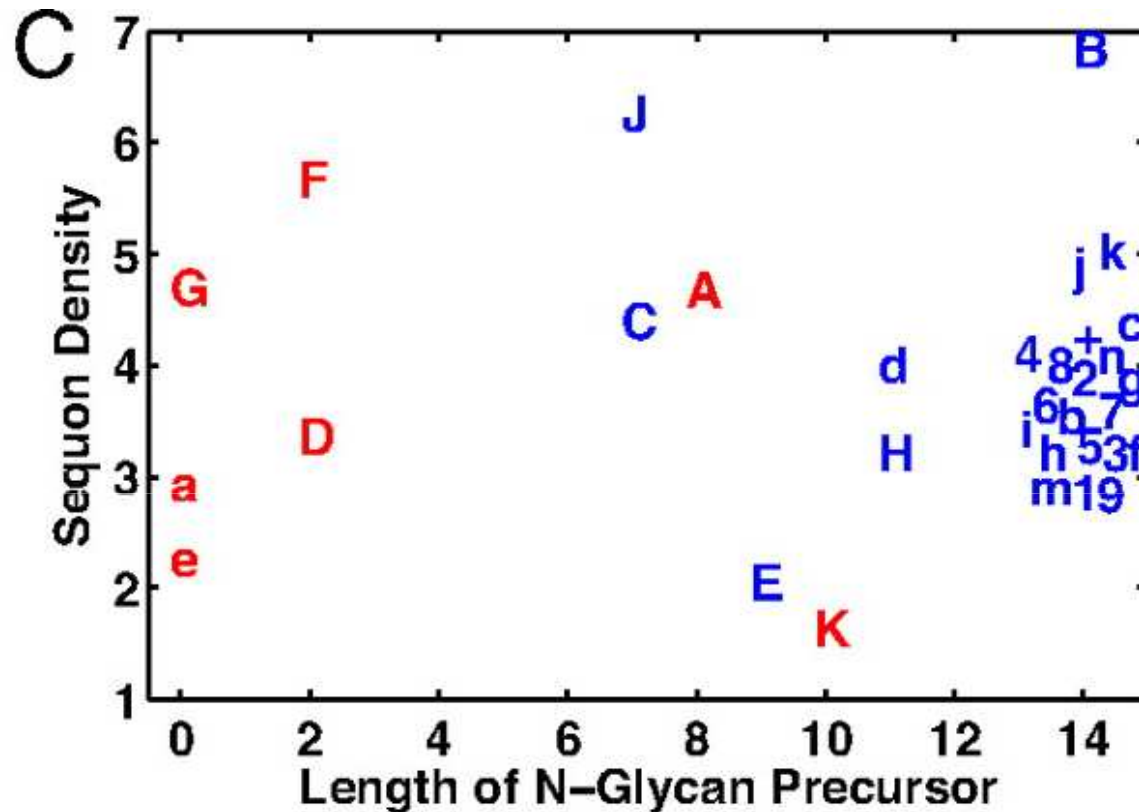


Density of sequons varies >4-fold among protists and this variability is observed in AT content as well

What next?

1. Is there a relation between N-glycan length and the density of sequons in their secreted proteins? If yes, how are they related?
 2. How are AT content (which is variable among protists) and sequon abundance related?
 3. Is there Darwinian selection for sequons in secreted and membrane proteins (vs cytosolic proteins)? If so, who's responsible:
 - i. Increased AT content?
 - ii. Increased Asn, Ser & Thr and decreased Pro in secreted proteins?
 - iii. Increased conditional probability that the above three amino acids will be present in sequons rather than elsewhere in the above proteins?
 4. What is the explanation behind the occurrence of high densities of sequons in influenza viruses (hemagglutinin - HA) and HIV (gp120)?
-

Answer to question 1:

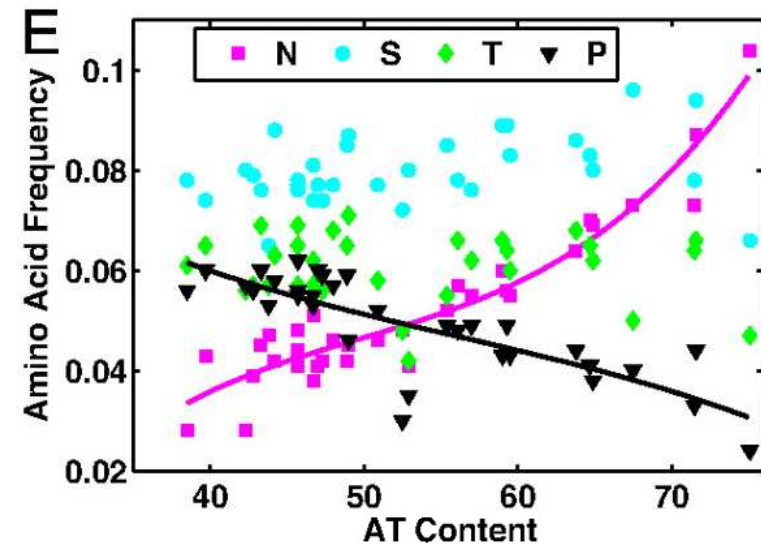
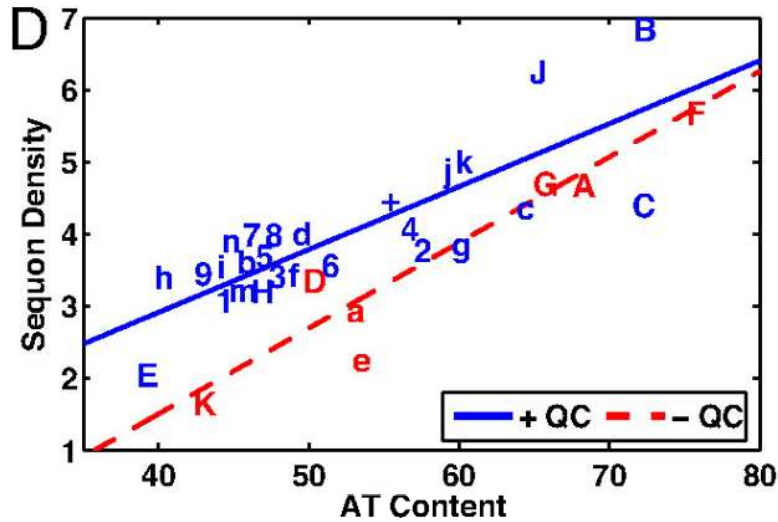


Effect of glycan length on sequon density seems to be indirect as it seems that eukaryotes with longer N-glycans are more likely to employ N-glycan-dependent QC folding

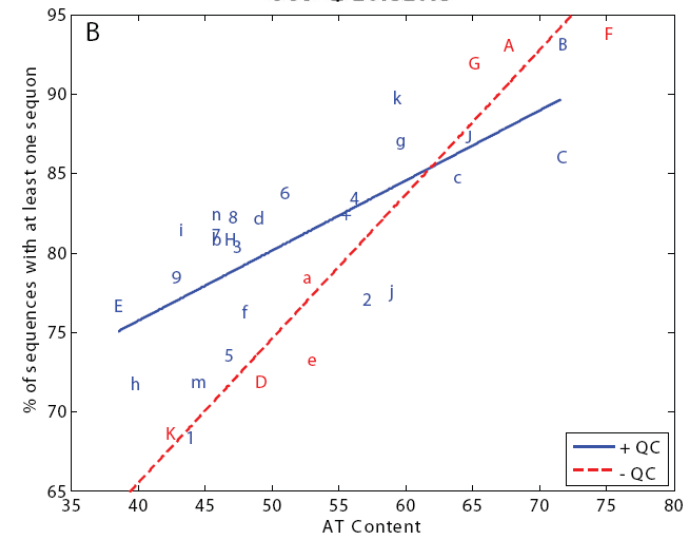
What next?

1. Is there a relation between N-glycan length and the density of sequons in their secreted proteins? If yes, how are they related?
 2. How are AT content (which is variable among protists) and sequon abundance related?
 3. Is there Darwinian selection for sequons in secreted and membrane proteins (vs cytosolic proteins)? If so, who's responsible:
 - i. Increased AT content?
 - ii. Increased Asn, Ser & Thr and decreased Pro in secreted proteins?
 - iii. Increased conditional probability that the above three amino acids will be present in sequons rather than elsewhere in the above proteins?
 4. What is the explanation behind the occurrence of high densities of sequons in influenza viruses (hemagglutinin - HA) and HIV (gp120)?
-

Answer to question 2:



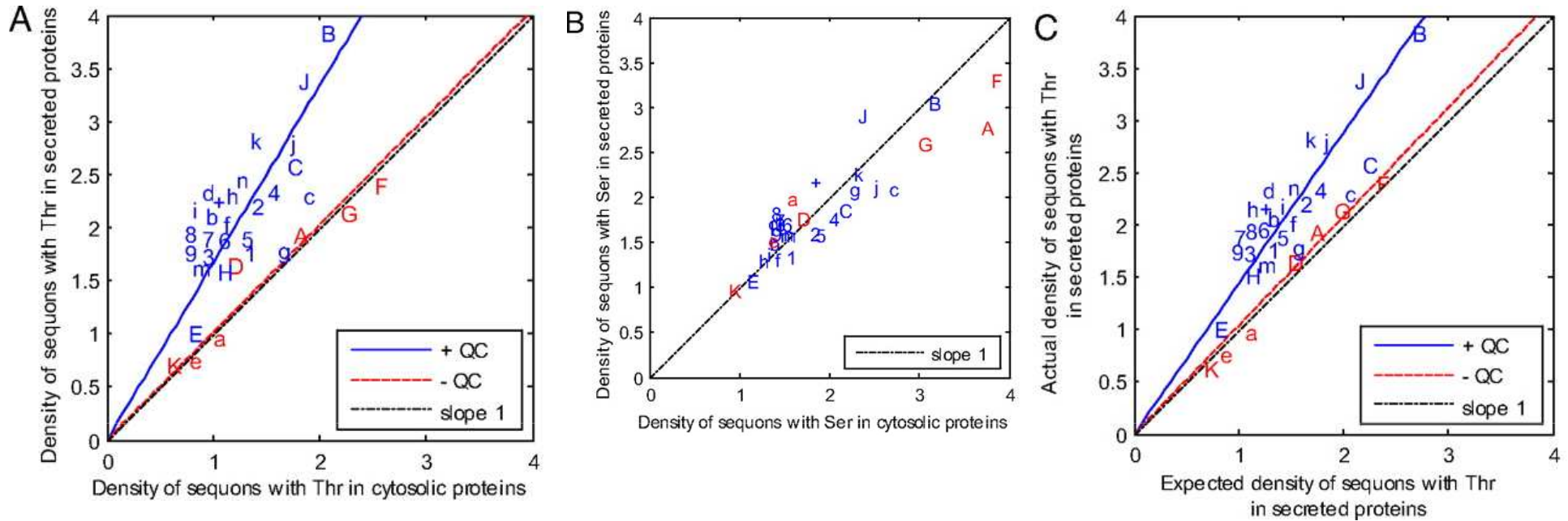
For each increase of AT content in coding regions by approx. 10%, there is an additional sequon per 500 amino acids



What next?

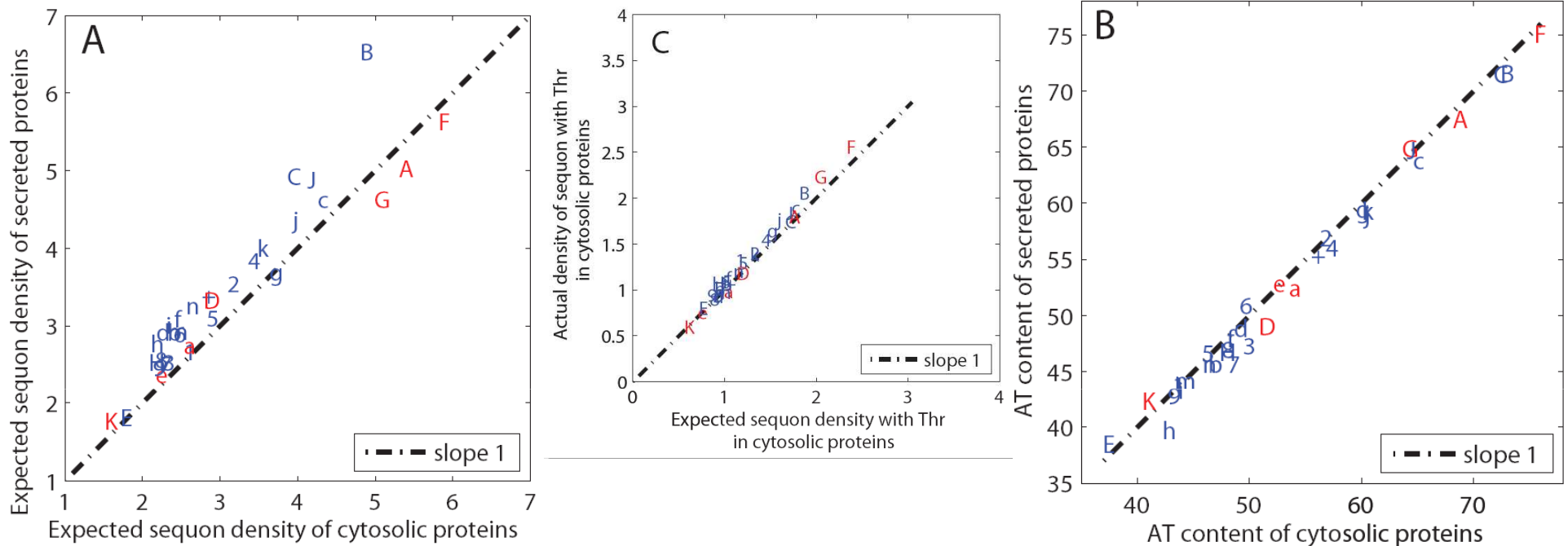
1. Is there a relation between N-glycan length and the density of sequons in their secreted proteins? If yes, how are they related?
 2. How are AT content (which is variable among protists) and sequon abundance related?
 3. Is there Darwinian selection for sequons in secreted and membrane proteins (vs cytosolic proteins)? If so, who's responsible:
 - i. Increased AT content?
 - ii. Increased Asn, Ser & Thr and decreased Pro in secreted proteins?
 - iii. Increased conditional probability that the above three amino acids will be present in sequons rather than elsewhere in the above proteins?
 4. What is the explanation behind the occurrence of high densities of sequons in influenza viruses (hemagglutinin - HA) and HIV (gp120)?
-

Answer to question 3:



Darwinian selection for sequons with Thr in the secreted and membrane proteins of eukaryotes with N-glycan-dependent QC of proteins folding is based mainly on an increased conditional probability that Asn and Thr will be present in sequons rather than elsewhere in these proteins

Answer to question 3:

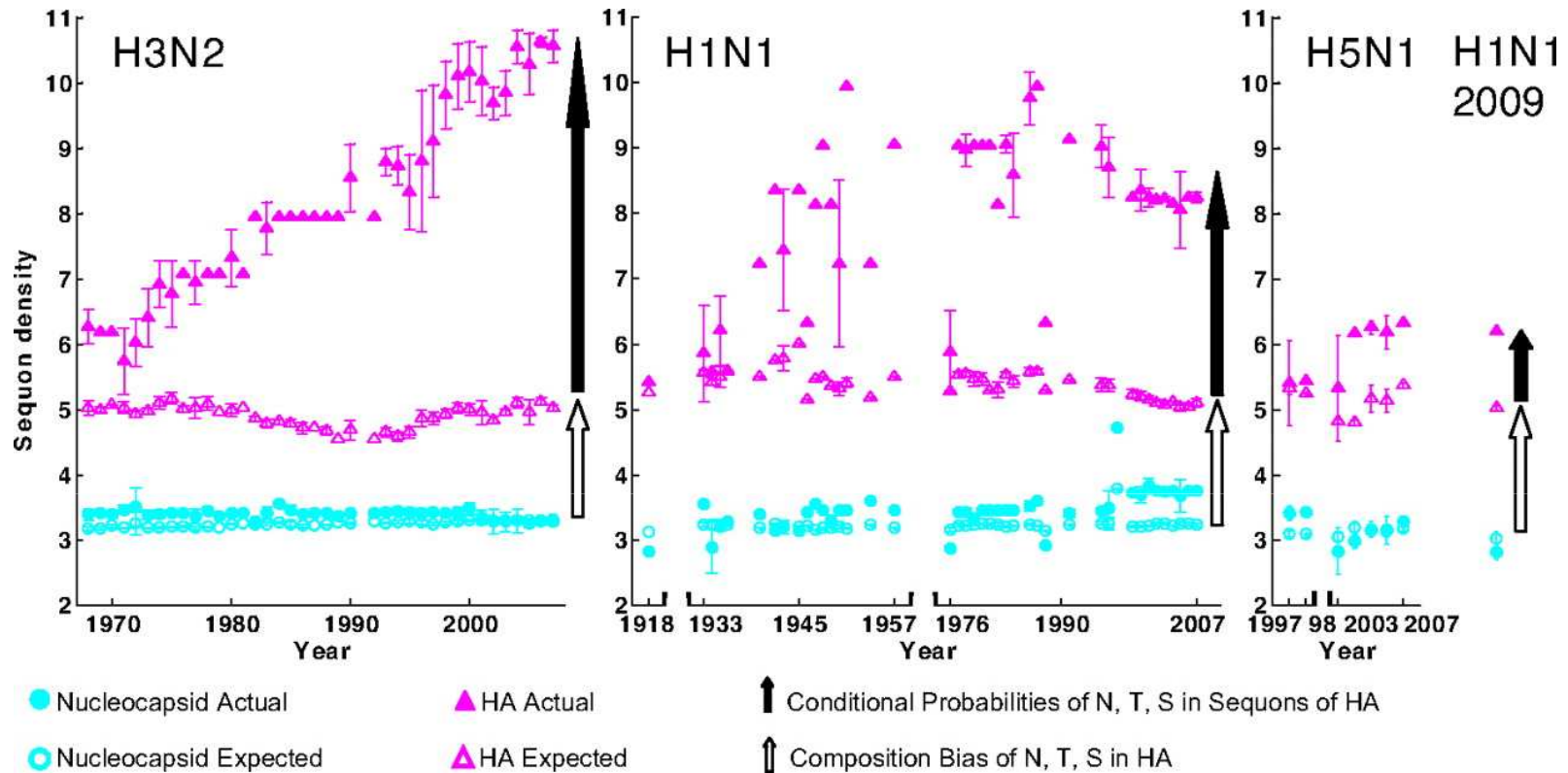


Amino acid bias and AT content contribute little to positive selection for sequons with Thr

What next?

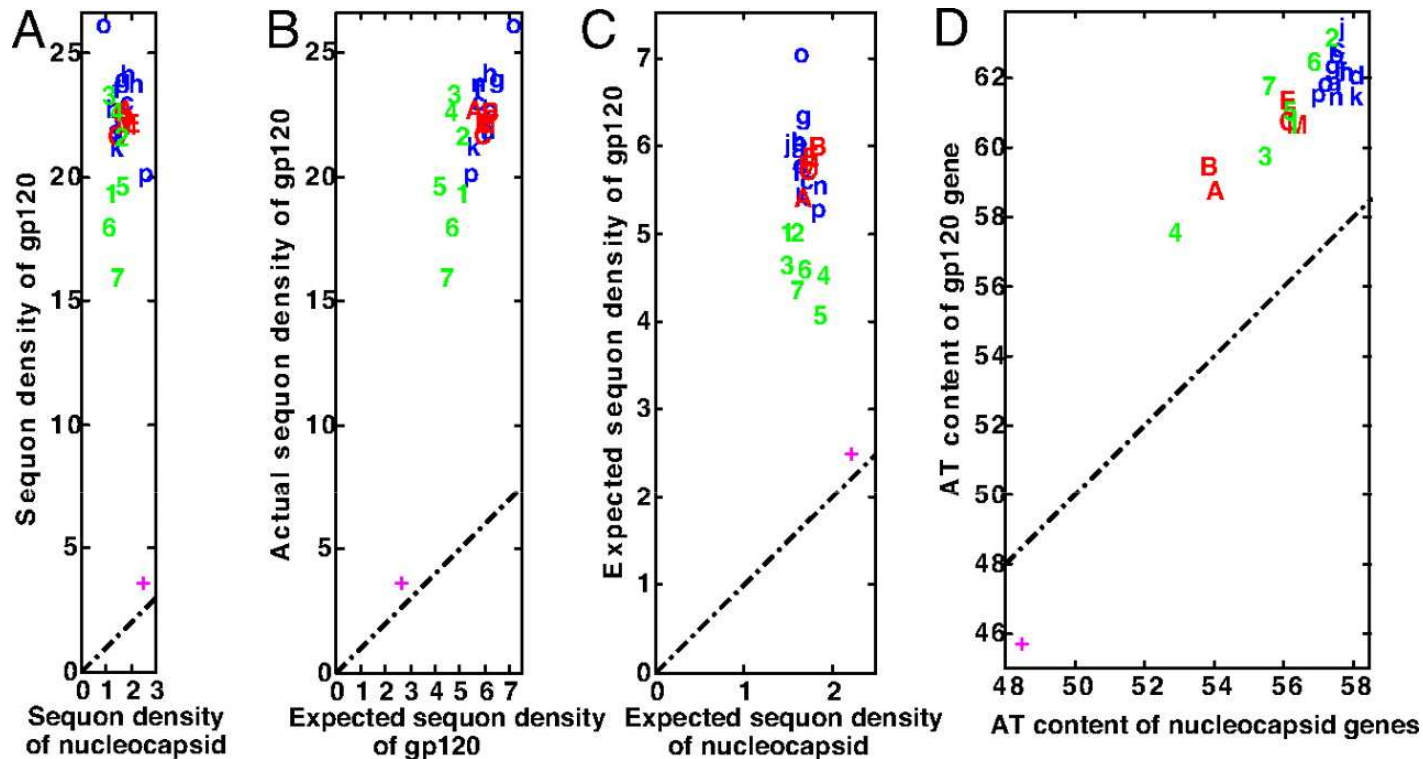
1. Is there a relation between N-glycan length and the density of sequons in their secreted proteins? If yes, how are they related?
 2. How are AT content (which is variable among protists) and sequon abundance related?
 3. Is there Darwinian selection for sequons in secreted and membrane proteins (vs cytosolic proteins)? If so, who's responsible:
 - i. Increased AT content?
 - ii. Increased Asn, Ser & Thr and decreased Pro in secreted proteins?
 - iii. Increased conditional probability that the above three amino acids will be present in sequons rather than elsewhere in the above proteins?
 4. What is the explanation behind the occurrence of high densities of sequons in influenza viruses (hemagglutinin - HA) and HIV (gp120)?
-

Answer to question 4:



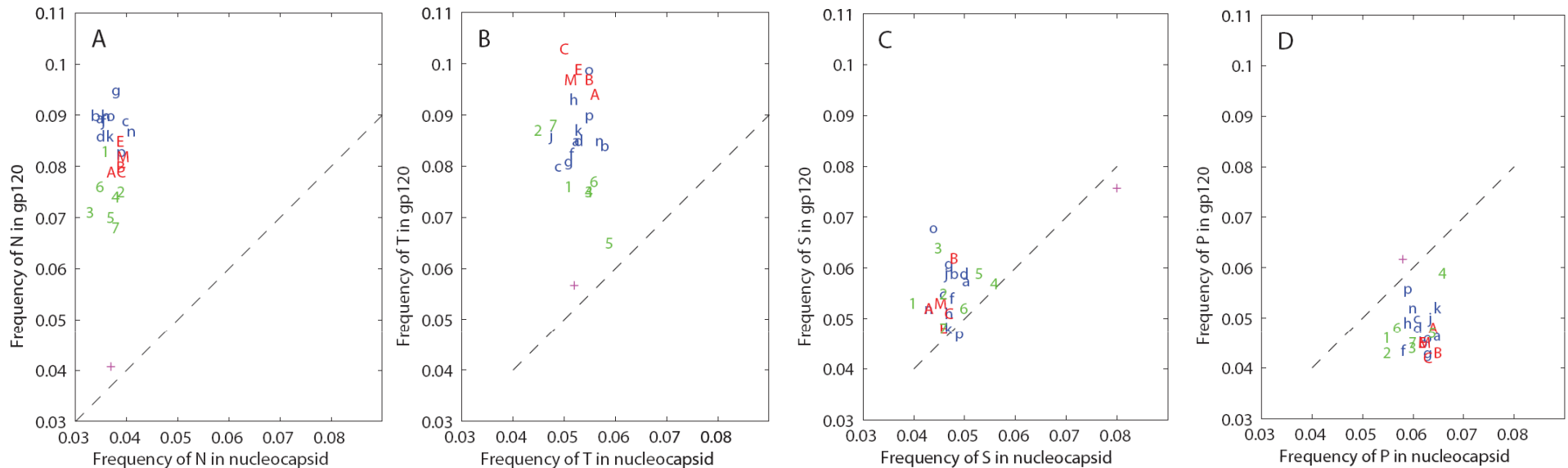
Increasing sequon densities of HA of A/H3N2 (Left) and A/H1N1 (Center) strains of influenza virus with antigenic drift results from an increased conditional probability that Asn, Thr, and Ser will be present in sequons rather than elsewhere in HA

Answer to question 4:



Very strong positive selection for sequons with Thr and Ser in gp120 of HIV and other retroviruses results from an increased conditional probability that Asn, Thr, and Ser will be present in sequons rather than elsewhere for selection in gp120, amino acid composition bias, and changes in AT content

Answer to question 4:



Breakdown of amino acid composition bias which contributes to the high density of sequons in gp120 of HIV and other retroviruses
Individual acids include Asn, Ser, Thr and Pro

Methods

Key aspects

- **Sequences:** Secreted and membrane proteins only (nucleocytosolic proteins used as control)
 - **Identification of QC systems:** Presence of UGGT, glucosidase II, calnexin or calreticulin and ERGIC-53 (Reference: *Schizosaccharomyces* proteins)
 - **Glycan lengths:** Presence of Alg enzymes in genomes through PSI-BLAST (Reference: *Saccharomyces cerevisiae*)
 - **AT content:** cDNA sequence for each protein
 - **Actual sequon densities:**
 - Sets of secreted and nucleocytosolic proteins concatenated into a single long sequence
 - From individual proteins without concatenation
 - **Expected sequon densities:** Calculated based on the frequencies of Asn, Ser, Thr and Pro in the set of secreted proteins for each organism
 - **Comparisons:**
 - Secreted vs. Cytosolic – Mann-Whitney rank-sum test
 - Actual vs. Expected – Wilcoxon matched-pairs test
-

Conclusions

Conclusions

- AT content has great effect on sequon density and thus on N-glycosylation of glycoproteins (Cause not confirmed)
 - Proof for Darwinian selection on N-glycan-associated QC system through a mechanism where there is an increased likelihood that Asn and Thr will be present in sequons rather than elsewhere in secreted and membrane proteins
 - N-glycan-associated QC of glycoprotein folding has nearly doubled the sequon densities of tens of thousands of secreted proteins of diverse eukaryotes
 - Similar observations in HIV and influenza viruses have been made before but explanation provided here
 - Sequon densities of viral envelope proteins appear to be far greater than those required to fold most host proteins
 - Important implications in the study of viral diseases as their N-glycans are known to be essential in pathogenesis and host entry
-

Points for open discussion

Points for open discussion

- ‘Causality with regard to AT content and sequon density’?
- Red Queen effect?
- Can we computationally predict the presence or the absence of QC systems based on secreted and membrane protein sequences?

$$P(QC | N\&T \text{ in sequons}) = \frac{P(N\&T \text{ in sequons} | QC) \cdot P(QC)}{P(N\&T \text{ in sequons}) + P(N\&T \text{ not in sequons})}$$

- Would bacterial sequences show these trends? Can the results be related to pathogenesis as well?
 - What do you think of the story of this paper in the context of the studies with viruses? Was the motivation clear there?
-

Thank you
