Analysis of the *Euplotes* genome using proteomics approaches

Fang Xie[1], Vladislav A. Petyuk[1], Alexey V. Lobanov[2], Anton A. Turanov[2], Lawrence A. Klobutcher[3], Heather M. Brewer[1], Marina A. Gritsenko[1], Ronald J. Moore[1], David G. Camp II[1], Vadim N. Gladyshev[2], and Richard D. Smith[1]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Brigham & Women's Hospital, Harvard Medical School, Boston, MA; [3]University of Connecticut Health Center, Farmington, CT

Pacific Northwest
NATIONAL LABORATORY

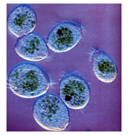
Overview

- **Purpose**: to characterize the *Euplotes crassus* proteome to enhance the genome annotation
- **Methods**: multiple proteolysis and fractionation approaches; high mass accuracy measurements; optimized data analysis
- **Results**: unprecedentedly high coverage and confidence in gene annotation, direct support for the unusual genetic code and translational mechanisms in *Euplotes crassus*

Introduction

- The ciliated protozoan *Euplotes crassus* has unusual genetic features that attract many investigators' attention. These features include dual amino acid coding (the codon UGA codes for both cysteine and selenocysteine),[1] the occurrence of gene-sized chromosomes in the macronucleus, and the frequent translational frameshifting (a directed change in translational reading frames).[2]
- Previous genetic studies have indicated several genes in *Euplotes* that require a +1 frameshift to express a functional protein, especially genes encoding proteins with enzymatic functions.
- Here, we employed proteomics approaches to identify proteins encoded in the *Euplotes* genome and characterize its genetic code and protein synthesis strategies.

Methods



Resuspend the *Euplotes crassus* cell pellet in the lysis buffer

↓

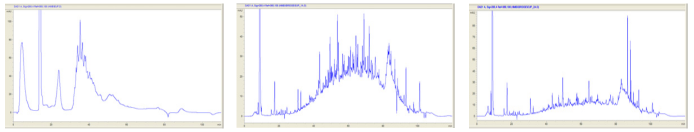
Pulverize in liquid N_2, melt, and sonicate for 1 min

↓

Reduction-alkylation

Split into three aliquots

Digestion	trypsin	trypsin	Glu-C
Fractionation	SCX	High-pH RP	High-pH RP



25 fractions collected

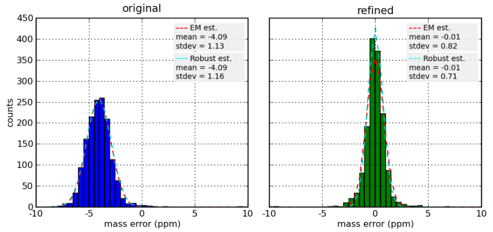
96 fractions collected, and then combined into 24 fractions per sample in a rolling fashion

↓

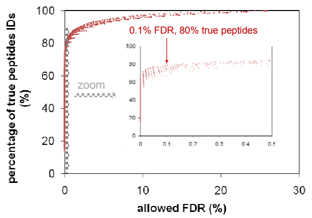
Low-pH RPLC on-line coupled with LTQ-Orbitrap MS

↓

SEQUEST search and data analysis



Apply DtaRefinery to decrease the width of the ΔM distribution histogram and thus the allowable deviation of the parent ion mass for true peptide identifications.



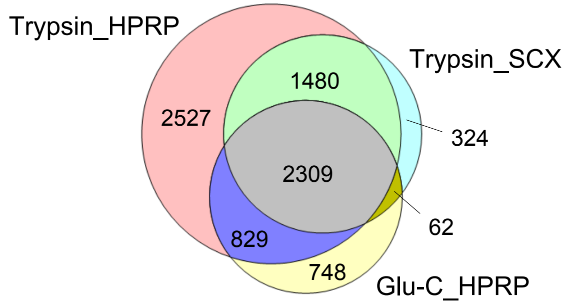
Maximize the number of peptide identifications by searching multiple combinations of ΔM, XCorr, and ΔCn parameters within the following ranges: ΔM: 0.5 → 10, XCorr: 0 → 4.9, ΔCn: 0 → 0.49.

Results

Exhaustive protein database

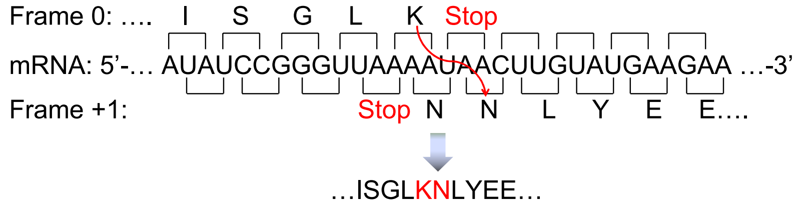
- The polypeptide sequence database was constructed by translating the *E. crassus* genome sequence in all 6 coding frames from stop-to-stop codons.
- In cases when the AAATAA is predicted to function as a frameshift, the pre- and post-shift polypeptide sequences were merged.
- The resulting polypeptide FASTA file resulted in ~3.7M entries (≥7 aa long), i.e.,150- to 200-fold more than the number of proteins in the closely related *Tetrahymena* ciliate genome (24K entries) or human (18K in Swiss Prot v15.10).
- The two orders of magnitude increase in search space resulted in a proportional increase in the number of false identified peptides and proteins, and the need for approaches providing confident identification of low frequency events such as frameshifts.

High proteome coverage with multiple proteolysis and fractionation methods



High proteome coverage allowed confident identification of unique proteins that resulted from unusual genetic features in *E. crassus*, e.g., proteins produced from frameshifting.

Identification of proteins resulting from +1 translational frameshifting



Frame 0: …. I S G L K Stop

mRNA: 5'-… AUAUCCGGGUUAAAAUAACUUGUAUGAAGAA …-3'

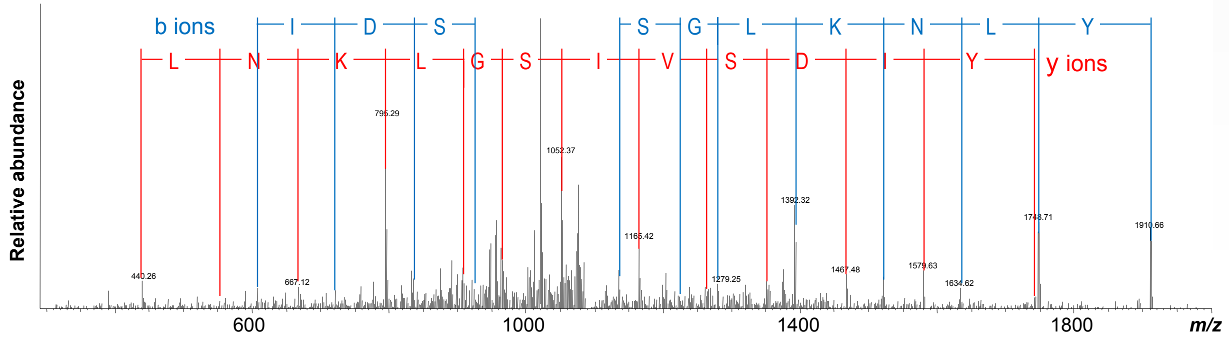
Frame +1: Stop N N L Y E E….

…ISGLKNLYEE…

Because it preserves the amino acid information at the C-terminus of K, Glu-C proteinase is suitable for identifying predicted AAATAA frameshifts.

MS/MS spectrum of 1094.54 *m/z* ion (z = +2)

ID: E.IDDTYIDSVISGLKNLYEE.L, XCorr = 4.11, ΔCn = 0.44



Conclusions

- Significant coverage (>60%) of the *E. crassus* proteome achieved by using different enzymatic digestion and fractionation approaches
- Identified 8279 unique proteins (with 4512 represented by at least two peptides; FDR <0.1%)
- Detected 5 out of 8 selenoproteins predicted from the DNA sequence
- Demonstrated the use of UGA to code for cysteines in proteins and detected no evidence that this codon terminates protein synthesis
- Confirmed +1 frameshifting events by identifying peptides that corresponded to two ORFs in the same gene
- Next step: look for proteins resulting from other translational frameshifting events, e.g., AAATAG frameshifts

Acknowledgements

This work was funded by the NIH National Center for Research Resources (RR18522). Samples were analyzed using capabilities developed under the support of the Proteomics National Center for Research Resources and the U.S. Department of Energy Biological and Environmental Research (DOE/BER).

Significant portions of the work were performed in the Environmental Molecular Science Laboratory (EMSL), a DOE/BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the DOE by Battelle under contract DE-AC05-76RLO-1830.

References

1. Turanov, A. A., *et al.* (2009) *Science* 323: 259-261.
2. Klobutcher, L. A. (2005) *Eukaryot Cell* 4: 2098-2105.

CONTACT: Fang Xie, Ph.D.
Biological Sciences Division, K8-98
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
E-mail: fang.xie@pnl.gov