SMBE Satellite meeting on De Novo Gene Birth

TITLES OF INVITED PRESENTATIONS

Diethard Tautz, Max-Planck Institute for Evolutionary Biology, Plön, Germany
Functional studies on de novo genes from mouse

Eve Syrkin Wurtele, Iowa State University, Ames, IA, USA
Compiling and compelling orphan genes

Christian Rödelsperger
Max Planck Institute for Biology, Tübingen, Germany
Evolutionary dynamics of novel genes in the shark-tooth nematode

Erich Bornberg-Bauer, University of Muenster, Germany
The Rise and Fall (or Assimilation) of de novo Genes in a Cellular Context

Alan Saghatelian, Salk Institute and Unversity of California in San Diego, USA
Microproteins that Control Cell Fate

Gisela Storz, National institute for Child Health and Development, Bethesda, MD, USA
How do genes-within-genes evolve?

Jorge Ruiz-Orera, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany
Evolutionary origins, roles and developmental regulation of human small open reading frames

Manyuan Long, University of Chicago, IL, USA
Phenotypic Effects of De Novo Genes in Drosophila and Underling Evolutionary Forces

David Begun, University of California, Davis, CA, USA
Investigating the population biology of de novo genes in Drosophila

CHC Cheng, University of Illinois at Urbana-Champaign, IL, USA
Repeat Sequences As Genetic Raw Material For New Genes Besides Antifreeze Proteins?

Anne-Ruxandra Carvunis, University of Pittsburgh School of Medicine, PA, USA
Systems approaches to decipher the mechanisms of evolutionary innovation

Yusuke Suenaga, Laboratory of Evolutionary Oncology, Chiba Cancer Center Research Institute, JP
Targeting de novo gene NCYM for cancer therapy

David Begun
Department of Evolution and Ecology
University of California, Davis

Investigating the population biology of de novo genes in Drosophila

Erich Bornberg-Bauer, University of Muenster, Germany

The Rise and Fall (or Assimilation) of de novo Genes in a Cellular Context

The emergence, maintenance and possible functionalisation of de novo protein coding genes poses a riddle to many fields of research, including genetics, molecular evolution and structural biology. We view this matter from as many angles as possible. Recently we have demonstrated, using data from several model species and modelling approaches, that their frequent emergence can be explained by a neutral model and that both emergence and loss can be parametrised in a simple stochastic model. We also provide some evidence that the fixation of at least some of the few de novo genes which survive over longer times can be explained by neutral model. Following some general consideration about how difficult it is to convert protein structures and/or to "find" them in vast sequences space, we conclude that the early days of functionalisation of de novo proteins may be largely neutral too. Most of the findings align with the "drift barrier hypothesis" and the model of "constructive neutral evolution" which views the adaptation of de novo proteins as a systemic process in which mutual dependencies between already existing proteins and the newcomers establish over time.

Anne-Ruxandra Carvunis, PhD
Associate Professor
Department of Computational and Systems Biology
Center for Evolutionary Biology and Medicine
University of Pittsburgh School of Medicine

Systems approaches to decipher the mechanisms of evolutionary innovation

All genomes contain genes whose sequences appear unique to a given species or lineage to the exclusion of all others. These “orphan” genes cannot be related to any known gene family; they are considered evolutionarily novel and are thought to mediate species-specific traits and adaptations. Many orphan genes have evolved through an enigmatic process called “*de novo* gene birth”. I will present a series of integrated computational and experimental analyses in budding yeast that begin to shed light on the molecular mechanisms of *de novo* gene birth. Serendipitously, these analyses reveal the existence of thousands of previously unsuspected translated elements in the yeast genome that appear to mediate beneficial phenotypes yet are evolutionarily transient. I will discuss the implications of these findings for our understanding of genome, cell and systems biology in the light of evolution.

Repeat Sequences As Genetic Raw Material For New Genes Besides Antifreez Proteins?

CHC Cheng, University of Illinois, Urbana-Champaign

Freezing water temperatures and floating ice crystals in the marine cryosphere are a deadly combination for teleost fishes, which compelled the evolution of a diversity of antifreeze proteins. Rarely is the selective pressure driving the evolution of an adaptive trait as clear and quantifiable, and the fitness consequence as visible and vital as these life-preserving proteins. Fish antifreezes evolved mostly through the classical mechanism of gene or domain duplication and sequence divergence, except for two. The antifreeze glycoproteins (AFGPs) of the cryonotothenioid fishes endemic to the Southern Ocean involved a pre-existing ancestor but with a partly *de novo* twist, and the convergently evolved AFGPs in the unrelated northern codfishes from entirely non-coding DNA. Both instances drew on one or a few tandem tripeptide (ThrAlaAla) coding elements and expanded it through repeated duplications into a long protein backbone that then becomes glycosylated with the ice-binding sugar side chains. On land where winter temperatures could plunge far colder than in the marine realm, various cold hardy insects have also evolved antifreeze proteins that are composed of longer repeats (11-12 amino acids). The encoding genes in most species are intronless, and the genetic origin/s of almost all of them have no known homologs, intimating potential instance/s of *de novo* origination. The repeat sequences of antifreeze proteins produce a structural fit with their ligand - ice crystal for interaction and binding to occur. Could simple or short repeat sequences be a class of fodder for broader novel gene generation remains uncertain.

Manyuan Long
University of Chicago

Phenotypic Effects of De Novo Genes in Drosophila and Underling Evolutionary Forces

Evolutionary dynamics of novel genes in the shark-tooth nematode

Christian Rödelsperger[1]

[1]Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Max-Planck-Ring 9, 72076 Tübingen, Germany

The morphological diversity of vertebrates is a beautiful example of phenotypic evolution. Although their archetypical body plan makes nematodes little less glamorous model organisms, they represent one of the most successful animal phyla and have adapted to almost all ecological niches including extreme abiotic environments and, in the case of parasitic nematodes, multiple host species. This makes them promising model systems to study molecular innovations. Our group focuses on the shark-tooth nematode *Pristionchus pacificus* which belongs to a clade that evolved teeth-like structures that allow them to predate and kill other nematodes. When its genome was sequenced, around one third of genes were classified as taxonomically-restricted orphan genes that lack homologs in other nematodes. Over the past years, we have characterized the origin and evolution of *P. pacificus* orphan genes using deep taxon phylogenomics. This revealed multiple mechanisms of new formation including divergence, *de novo* gene birth and mixed origin. While recent reports of thousands of noncanonical translated elements emphasize the potential impact of *de novo* gene birth on genomic novelty, we recently proposed to assess their evolutionary impact by studying their dynamics across multiple time scales and explicitly comparing these patterns with products of gene duplication which represents another major source of novelty. Comparative genomic and population-scale analysis in *P. pacificus* demonstrate a high abundance of *de novo* candidates relative to duplicates at shorter time-scales but opposite patterns across longer time-scales. Thus, both processes appear to operate at different time-scales whereby *de novo* genes may carry out more transient functions in fluctuating environments but more stable evolutionary changes are implemented by duplication events.

Evolutionary origins, roles and developmental regulation of human small open reading frames

Jorge Ruiz-Orera[1]
[1] Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany

In the last years, advances in genomics have revealed the translation of small open reading frames (sORFs) in untranslated regions from protein-coding genes and long non-coding RNAs. At the forefront of this field is the experimental data from ribosome profiling, a technique for the high throughput sequencing of ribosome-protected RNA fragments. While there are multiple catalogs of sORFs identified with ribosome profiling across several human organs and cell lines, the evolutionary properties of these sORFs and their encoded microproteins are yet poorly understood.

We explore the origins and evolutionary timelines of a set of 7,264 translated sORFs, as identified by seven ribosome profiling studies and hosted within GENCODE. By analyzing the structural conservation of these sORF frame structures across the genomes of 120 mammalian species, our analysis reveals that 6,506 sORF structures are specific to primates or humans. Notably, over half of these evolutionarily young sORFs originated de novo from ancestral non-coding sequences. We further elucidate the evolutionary dynamics of sORF translation by generating ribosome profiling data from adult hearts and induced pluripotent stem cell (iPS)-derived cardiomyocytes, across four primate species. We find that the translation of the majority of sORFs is evolutionarily recent and exhibits an increase throughout development. To shed light on the functional significance of these newly emerged sORFs, we conducted mass spectrometry-based interactome screens with motif resolution and additional cellular assays. We associate several microproteins encoded by evolutionarily young sORFs with vital cellular processes, such as splicing, translational regulation, and endocytosis.

In conclusion, we provide evidence that thousands of novel sORFs emerged de novo during human and primate evolution and we underscore the potential role of sORF-encoded microproteins in the origins of functional innovations and organ development.

Alan Saghatelian,
Salk Institute and Unversity of California in San Diego, USA

Microproteins that Control Cell Fate

Abstract: Microproteins are an emerging frontier in molecular biology and the study of these genes has revealed thousands of these proteins in the dark matter of the proteome. As we transition from discovery to characterization, we discovered several microproteins, some conserved others not, that can control cell fate. These microproteins can modulate the cells response to stress or help cells differentiate, and we will describe their discovery and characterization.

Gisela Storz

How do genes-within-genes evolve?

A major focus of research by my group is identifying and characterizing genes that were previously missed, particularly those encoding small RNAs and small proteins of less than 50 amino acids. Although they were overlooked, small RNAs and small proteins have important regulatory roles. Initially, we focused our searches on intergenic regions, but it is now becoming clear that these regulatory molecules also can be encoded in intragenic regions. I will discuss the roles of example genes-within-genes in bacteria as well as thoughts about their evolution.

Targeting de novo gene NCYM for cancer therapy

Yusuke Suenaga
Laboratory of Evolutionary Oncology, Chiba Cancer Center Research Institute

The de novo gene NCYM is located on the antisense strand of the MYCN oncogene and encodes an oncoprotein in humans. In human neuroblastomas, NCYM inhibits GSK3β to stabilize MYCN and promotes distant metastasis. MYCN directly binds to its own intron 1 region and stimulates both MYCN/NCYM transcription, thereby forming a positive feedback loop with NCYM. Among adult cancers, the expression levels of NCYM, but not that of MYCN, are associated with poor prognosis of cholangiocarcinoma; however, the roles of NCYM in the development of cancer have remained elusive. In this study, we found that NCYM promotes cholangiocarcinogenesis accompanied by autophagy activation. Based on the result that NCYM expression levels correlate with Ras signaling in cholangiocarcinoma, we first employed a mouse bile duct organoid carcinogenesis model that expresses Kras^{G12D}. The Kras-induced organoids with NCYM overexpression formed subcutaneous tumors in nude mice at a higher rate (75%) compared with that of the control mice with only the Kras mutation (25%). Long-read RNA sequencing revealed that genes with elevated mRNA expression were enriched in autophagy-related genes, and elevated LC3 protein was detected in the NCYM-overexpressing organoids. The induced autophagy in the organoids was further confirmed using DALGreen staining and electron microscopy. We next performed machine-learning-based screening for the compounds binding to NCYM and identified lysine acetate and curcumin analogues as NCYM inhibitor candidates. These compounds reduced cell proliferation of mouse and human cholangiocarcinoma organoids in a manner dependent on NCYM expression and inhibited autophagosome formation in the organoids. Using nuclear magnetic resonance, small-angle scattering, and atomic force microscopy, we analyzed the structural dynamics of complexes of these compounds with NCYM to understand the molecular mechanism of drug efficacy at the atomic level.

In the last part of my presentation, I will discuss the challenges in realizing NCYM-targeted cancer therapy and our attempts to solve them.

Abstract for SMBE satellite meeting 2023

Functional studies on de novo genes from mouse

Diethard Tautz#
Max-Planck Institute for Evolutionary Biology, Plön, Germany
#tautz@evolbio.mpg.de

The genome data from house mouse populations, subspecies and species are an excellent resource for identifying de novo evolved genes, which can be subjected to functional studies. I will discuss results from three different approaches: knockouts, fitness estimates in seminatural environments and overexpression in a heterologous cell system. The results suggest that de novo emerged genes do not need to accumulate adaptive mutations to become functional. This is in line with the finding that even the expression of random sequences in cells can positively influence the growth competitiveness of cells.

Compiling and compelling orphan genes.

Eve Syrkin Wurtele

Tens of thousands of unannotated genes of unknown function may reside within even well-researched genomes. I will discuss our development of high-throughput tools and approaches to identify and reveal function of *de novo* and *de regeneratio* orphan genes, and the application of these to maize, Arabidopsis, human, and SARS-CoV-2. Our experimental follow-up for two such *de novo* genes will be highlighted.

SMBE Satellite meeting on De Novo Gene Birth

TITLES OF CONTRIBUTED PRESENTATIONS

Zachary Ardern, Wellcome Sanger Institute, Hinxton, Cambridgeshire, United Kingdom
The non-canonical translatome in bacteria and beyond

Josué Barrera-Redondo, Department of Algal Development and Evolution, Max Planck Institute for Biology, Tübingen, Germany
Tracking down the evolution and functional integration of de novo emerged genes in the brown algae

Luuk A. Broeils, Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands
Evolutionarily young genes in pediatric cancer

Claudio Casola, Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA
Degradation determinants are more abundant in noncanonical than canonical proteins in human

Igor Fesenko, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA
Large-scale analysis of the cryptic microproteome in prokaryotes

Katherine Fleck, Department of Ecology and Conservation Biology, University of Connecticut, Storrs, USA
Novel gene evolution and 3D chromatin organization

Idan Frumkin, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

Random Proteins provide Evolutionary Novelty by interacting with Diverse Cellular Pathways

Anna Grandchamp, University of Muenster, Germany
Study of mutations underlying de novo gene emergence

Amir Karger, IT Research Computing, Harvard Medical School, Boston, USA
Improving gene age estimation: protein length, sequence similarity method, reference database composition and search stringency are essential to accurate gene age estimation

Victor Luria, Department of Neuroscience, Yale University, New Haven, USA
Novel genes enable protein structural innovation and function in the brain

Somya Mani, Institute for Basic Science – Center for Soft and Living Matter, South Korea
Population model characterizes genomic regions that are fertile for de novo gene birth

Kazuma Nakatani, Graduate School of Medical and Pharmaceutical Sciences, Chiba University, Chiba, Japan
NCYM, a human de novo evolved gene product, promotes tumorigenesis in a mouse cholangiocarcinoma organoid model

Anne O'Donnell-Luria, Broad Institute of MIT and Harvard, Cambridge, MA, USA
The role of non-canonical open reading frames in Mendelian disease

Chris Papadopoulos, Evolutionary Genomics Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona 08003, Spain
De novo gene intra-species diversity in Saccharomyces cerevisiae

Claire Patiou, Université de Lille, France
Unraveling the Origin, Evolution, and Role of de novo sORFs: A Case Study in the Arabidopsis Genus

Junhui Peng, Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA
The origin and structural evolution of de novo genes in Drosophila

April Rich, Joint Carnegie Mellon University-University of Pittsburgh Computational Biology PhD Program, University of Pittsburgh, Pittsburgh, PA, USA
Exploring transcriptional profiles of de novo ORFs using massively integrated coexpression analysis

Nicolas Svetec, Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA
The birth and function of de novo genes in the Drosophila brain

Emilios Tassios, Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", 34 Fleming Street, 16672, Vari, Greece
A large-scale analysis of genetic novelty in budding yeast

Vyacheslav Tretyachenko, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel
Mistranslation in protein function space exploration

Nikolaos Vakirlis, Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming"
34 Fleming Street, 16672, Vari, Greece
Ancestral Sequence Reconstruction as a tool to study de novo gene birth

Covadonga Vara, Evolutionary Genomics Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona 08003, Spain
Investigating de novo gene formation in human populations

Shengqian Xia, Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America
The evolution and knock-out lethality of stepwise de novo genes in Drosophila

Li Zhao, Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA
The dynamics and regulatory mechanisms of de novo gene expression

The non-canonical translatome in bacteria and beyond

Zachary Ardern, PhD
Wellcome Sanger Institute, Hinxton, Cambridgeshire, United Kingdom

Growing evidence from ribosome profiling studies suggests that in a typical bacterial genome more than one hundred open reading frames (ORFs) at least 30 codons long but not labelled as “genes” are translated as proteins under some growth conditions. Most of these ORFs are short, but in higher GC genomes a few approach the length of an average bacterial gene and are encoded overlapping known genes in an alternative reading frame e.g. in antisense. Very little is known about these ORFs and other ORFs of dubious coding status, sometimes annotated as “hypothetical proteins”. Here I bring together recent large-scale analyses of these ORFs concerning their translation, transcriptional regulation, and evolution, focusing on research in *E. coli*, while also comparing this with results from other bacterial species. These ORFs have mostly arisen as coding sequences within the species or genus (very recently compared to the large majority of “canonical” genes), and so are some of the best candidates for understanding “de novo” origin more generally.

Two aspects of potentially wide interest are: 1) the large majority of work on non-canonical translatomes has been in eukaryotes, but bacteria are important systems to understand for reasons including links to human health and disease and their experimental tractability which makes them excellent models for evolutionary systems biology; 2) alternative frame sequences are under-studied yet results from bacteria have implications for alternative frame coding in eukaryotes and also in diverse viruses.

Tracking down the evolution and functional integration of *de novo* emerged genes in the brown algae

Josué Barrera-Redondo[1], Jia Xuan Leong[1], Cátia Igreja[2], Susana M. Coelho[1].

[1] Department of Algal Development and Evolution, Max Planck Institute for Biology, Max-Planck-Ring 5, 72076, Tübingen, Germany.
[2] Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Max-Planck-Ring 5, 72076, Tübingen, Germany.

The case of *de novo* gene birth addresses some critical questions in evolutionary biology. So far, studies on *de novo* gene birth have focused in a handful of plants, animals and fungi. We therefore ignore if the processes behind *de novo* gene birth are conserved across the vast diversity of eukaryotes. Brown algae are the third most complex multicellular lineage in the planet. They have been evolving independently from animals and plants for more than a billion years, representing unique models to study the universality or uniqueness of biological processes throughout the tree of life. We used our newly developed pipeline GenEra to scan for taxonomically-restricted genes (TRGs) in 46 brown algal species, showing that ~25% of their genes represent brown algal TRGs. Interestingly, brown algal TRGs were overrepresented in the sex chromosomes. We focused on the model organism *Ectocarpus* for which several genomic resources are available and which can be used for experimental studies. We found 1852 TRGs in the reference genome of *Ectocarpus* that are restricted to the species, genus and family levels. RNA-seq data show that these TRGs are differentially expressed throughout the life cycle, suggesting they are functional. The use of synteny-based approaches allowed us to distinguish the TRGs that potentially evolved *de novo* in *Ectocarpus* and to search for enabling mutations. We are currently establishing a ribosomal profiling protocol in *Ectocarpus* to validate the translational activity of these genes and generate a list of candidate *de novo* genes for further experimental assays (e.g., CRISPR knock-outs).

Evolutionarily young genes in pediatric cancer

Luuk A. Broeils[1], Ana P. Pinheiro-Lopes[1], Sem A.G. Engels[1], Jip T. van Dinter[1], Jorge Ruiz-Orera[2], Jasper van der Lugt[1], Thomas G.P. Grünewald[3,4,5], Sebastiaan van Heesch[1]

1) Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584 CS Utrecht, Utrecht, The Netherlands
2) Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), 13125 Berlin, Germany
3) Hopp-Children's Cancer Center (KiTZ), Heidelberg, Germany
4) Division of Translational Pediatric Sarcoma Research (B410), German
Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
5) Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany

Unlike adult cancers, pediatric cancers typically exhibit a limited number of mutations but are often characterized by epigenetic dysregulation, resulting in abnormal chromatin accessibility. We hypothesized that this dysregulated chromatin state could lead to the transcription and translation of previously unrecognized genes with potential roles in tumor biology. While traditionally, conserved genes have been considered functionally important, we recognize that lack of conservation does not exclude a sequence from having a cellular role. Evolutionarily young genes are often expressed during early organismal development and some of them have evolved *de novo*. Given that childhood cancers typically arise during early developmental stages, these genes may play a role in tumorigenesis. To investigate this phenomenon, we have been investigating pediatric brain tumors, as they are often epigenetically dysregulated, and the brain has been found to be a hotspot for evolutionarily young genes. Additionally, we are investigating Ewing Sarcoma, a human-specific disease, where the EWSR1-FLI1 fusion protein has been found to drive the transcription (and possible translation) of novel genes, which could be human-specific genes that become active outside their physiological role. We conducted Ribo-seq analysis on 60 Ewing Sarcoma patient tumor tissues, mapping the data to custom-built transcriptomes based on in-house and public RNA-seq data of Ewing Sarcoma. By doing so, we found numerous novel genes and ORFs specifically expressed in Ewing Sarcoma. We also explored the evolutionary origin of these ORFs, finding many have emerged through de novo evolution. Future experiments will investigate if these ORFs encode microproteins with cellular roles.

Degradation determinants are more abundant in noncanonical than canonical proteins in human

Claudio Casola[1,2,3], Adekola Owoyemi[1], Nikolaos Vakirlis[4]

[1]Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA 77843
[2]Interdisciplinary Graduate Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, USA 77843
[3]Interdisciplinary Graduate Program in Genetics and Genomics, Texas A&M University, College Station, USA 77843
[4]Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari, Greece

The characterization of the complete repertoire of human proteins, a central goal of modern biology, has recently been upended by the discovery of thousands potential novel proteins via ribosome profiling. Determining the physiological activity, if any, of these noncanonical proteins (NCPs) has proven difficult. Preliminary evidence and overall protein degradation rates suggest that many noncanonical proteins may possess low stability in the cellular environment. This is in agreement with the "defective ribosomal products" (DRiPs) hypothesis, which posits that a large proportion of NCPs represent misfolded polypeptides or translation errors. Here, we tested this hypothesis by analyzing the frequency of multiple sequence degradation determinants eliciting either proteasomal removal or autophagy, using 91,003 canonical isoforms and 232,460 noncanonical proteins. Overall, noncanonical proteins were enriched for degradation-related features compared to all canonical proteins. Analyses of original and shuffled sequences showed evidence of stronger selective constraints against the accumulation of degradation signatures in canonical proteins. However, stability was significantly higher than expected by chance in NCPs with evidence of phenotypic effects when knocked-out in cell lines. The C-terminal tail hydrophobicity represents a particularly reliable proxy for degradation propensity with potential application in identifying functional noncanonical proteins. These findings underscore the critical role of degradation processes in regulating the life cycle of noncanonical proteins and demonstrate the power of degradation determinants in discriminating noncanonical proteins likely to represent biologically functional molecules.

Large-scale analysis of the cryptic microproteome in prokaryotes

Igor Fesenko, Svetlana A Shabalina, Eugene V Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Small open reading frames (smORFs) often represent evolutionarily young translatable units some of which encode functional microproteins. However, the coding potential and evolutionary history of the smORFs remain poorly understood and are often neglected in genomic and proteomic studies. To assess the potential of intergenic regions in microbial genomes as birth places of new genes, we performed a comprehensive comparative analysis of intergenic smORFs (15–70 codons) in 5668 *Enterobacteriaceae* genomes. When compared to annotated small proteins and shuffled smORFs, the intergenic smORFs have a substantially lower entropy, a higher overall hydrophobicity, and a much greater fraction of predicted transmembrane domains. These features suggest many intergenic smORFs encode short, membrane proteins. Comparative analysis of intergenic smORFs within and between enterobacterial genera suggested that these smORFs encode different types of cryptic microproteins. We clustered the sequences of intergenic smORFs and assessed their coding potential using various approaches, including analysis of evolutionary rates and trinucleotide periodicity. This analysis identified several tens of thousands of microprotein candidates. We further analyzed the predicted structures and evolutionary conservation of the putative microproteins, and compared the predictions with the available transcriptomic and proteomic data. These results demonstrate the high potential of intergenic regions in bacterial genomes for *de novo* generation of genes encoding microproteins.

Novel gene evolution and 3D chromatin organization

Katherine Fleck[1,†], Victor Luria[2,3†], Nitanta Garag[1], Amir Karger[4], Trevor Hunter[1], Daniel Marten[5,6,7], William Phu[5,6,7], Nenad Sestan[3], Anne H. O'Donnell-Luria[5,6,7*], Jelena Erceg[1,8,9*]

[1]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA

[2]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

[3]Department of Neuroscience, Yale School of Medicine, New Haven, CT 06510, USA

[4]IT-Research Computing, Harvard Medical School, Boston, MA 02115, USA

[5]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston 02115, USA

[6]Center for Mendelian Genomics, Eli and Edythe L Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[7]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[8]Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269, USA

[9]Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06030, USA

[†]These authors contributed equally to this work

*Correspondence

Genome organization may be intricately tied to regulating genes and associated cell fate decisions. Recent technological advances in mapping of chromosomal interactions and single-cell imaging have provided insights into chromatin organization at various levels including domains, loops, and boundaries. However, how the placements of genes of different evolutionary age in the 3D genome landscape relate to their biological role remains unclear. In this study, we examine the positioning and functional associations of human genes, grouped by their evolutionary age, within the 3D genome organization. We reveal that genes of different evolutionary origin have variable positioning relationships with both domains and loop anchors, but remarkably consistent associations with boundaries across cell types. The functional associations of each grouping of genes are primarily cell type-specific, however, those with recently evolved genes are sensitive to 3D genome architecture. Moreover, the sensitivity of such recent genes in diseased state is more pronounced in loop anchors compared to domains. We complement these findings with analysis of the expression from genes of differing evolutionary ages across cell types. Altogether, these distinct relationships between gene evolutionary age, their function, and positioning within 3D genomic features may contribute to understanding tissue-specific gene regulation in development and disease.

Random Proteins provide Evolutionary Novelty by interacting with Diverse Cellular Pathways

Idan Frumkin[1], Christopher Vassallo[1], and Michael Laub[1,2]

1 Department of Biology, Massachusetts Institute of Technology

2 Howard Hughes Medical Institute

How do organisms evolve new functions? While novelty usually emerges by modifying existing genes, "*de novo* gene birth" is another mechanism in which new genes originate from random sequences. However, it is poorly understood what functions such *de novo* genes serve and how they integrate into complex cellular systems.

To investigate whether proteins with random sequences can benefit cells, we screened a library of ~10^8 random proteins, with no homology to existing proteins, for the ability to promote survival of *E. coli* cells facing two threats: the endoribonuclease toxin MazF or the bacteriophage T4. We found hundreds of functional random proteins that inhibit MazF or prevent phage infection. We then revealed that these proteins modify cellular physiology by integration with pre-existing central homeostasis pathways, such as chaperones, proteases, or membranal signal transduction systems. This cellular remodeling by random proteins results in rapid MazF degradation or inhibition of viral cellular entry, which allow cells to overcome these harsh challenges.

Our work provides a mechanistic basis for how *de novo* gene birth can produce new, functional proteins integrated into complex cellular systems and benefit cells.

Study of mutations underlying de novo gene emergence

Novel genes are essential for evolutionary innovations and differ substantially even between closely related species. Recently, multiple studies across many taxa showed that some novel genes arise de novo, i.e. from previously non-coding DNA. In order to characterise the underlying mutations that allowed de novo gene emergence and their order of occurrence, homologous regions must be detected within non-coding sequences in closely related sister genomes. So far, most studies do not detect non-coding homologs of de novo genes due to incomplete assemblies and annotations, and long evolutionary distances separating genomes.

Here, we overcome these issues by searching for de novo expressed ORFs (neORFs), the not-yet fixed precursors of de novo genes that emerged within a single species. We sequenced and assembled genomes with long-read technology and the corresponding transcriptomes from inbred lines of Drosophila melanogaster, derived from 7 geographically diverse populations.

We found line-specific neORFs in abundance but few neORFs shared by lines, suggesting a rapid turnover. Gain and loss of transcription is more frequent than the creation of Open Reading Frames (ORFs), e.g. by forming new START- and STOP-codons. Consequently, the gain of ORFs becomes rate limiting and is frequently the initial step in neORFs emergence.

Furthermore, Transposable Elements (TEs) are major drivers for intra genomic duplications of neORFs, yet TE insertions are less important for the emergence of neORFs.

However, highly mutable genomic regions around TEs provide new features that enable gene birth.

In conclusion, neORFs have a high birth-death rate, are rapidly purged, but surviving neORFs spread neutrally through populations and within genomes.

Anna Grandchamp, University of Muenster, Germany

Study of mutations underlying de novo gene emergence

Improving gene age estimation: protein length, sequence similarity method, reference database composition and search stringency are essential to accurate gene age estimation

Amir Karger 1, Victor Luria 2,3,4, John W. Cain 5, Kee Young Nam 3, Daniel Marten 4, 6, Nenad Sestan 2, Anne O'Donnell-Luria 4,6, Marc Kirschner 3

1 Harvard Medical School, IT Research Computing, Boston, USA
2 Yale University, Department of Neuroscience, New Haven, USA
3 Harvard Medical School, Department of Systems Biology, Boston, USA
4 Harvard Medical School, Boston Children's Hospital, Division of Genetics and Genomics, Boston, USA
5 Harvard University, Department of Mathematics, Cambridge, MA, USA
6 Broad Institute of MIT and Harvard, Cambridge, MA, USA

Non-copying mechanisms generate evolutionarily new genes - *de novo* from intergenic DNA or long non-coding RNAs - which encode genuinely novel proteins. A given species' proteome, then, will contain proteins of many different ages. To understand how new genes appear and differ from ancient established genes, it is essential to have a general method to estimate gene age. We built upon the idea of gene age estimation by phylostratigraphy: a gene's protein sequence is used to query the largest possible database to determine the most distant species containing a sufficiently similar sequence. The minimal evolutionary birthdate of the gene is then estimated as the evolutionary age when the query species and the most distant species last shared a common ancestor. We systematically tested key parameters impacting age estimation. We determined a minimal protein length for reliable age estimation. We compared two widely used similarity search methods, BLASTP and HMMER, and found they have different sensitivities. We found that gene age estimates vary little between different databases, that the genomic coverage of every taxonomic node should be evaluated in reference databases, and that a large minimum number of hits is necessary to find the most distant species. Our investigations find parameter regimes for which gene age estimation is robust for most Eukaryotic protein-coding genes and enable systematically comparing the size and biophysical properties of genes of different evolutionary ages in all Eukaryotic species with good quality genomes.

Novel genes enable protein structural innovation and function in the brain

Victor Luria [1,2,3], Amir Karger [4], John W. Cain [5], Kee Young Nam [2], Daniel Marten [3,6], Anne O'Donnell-Luria [3,6], Marc Kirschner [2], Nenad Sestan [1]

1 Yale University, Department of Neuroscience, New Haven, USA
2 Harvard Medical School, Department of Systems Biology, Boston, USA
3 Boston Children's Hospital, Division of Genetics and Genomics, Harvard Medical School, Boston, USA
4 Harvard Medical School, IT Research Computing, Boston, USA
5 Harvard University, Department of Mathematics, Cambridge, MA, USA
6 Broad Institute of MIT and Harvard, Cambridge, MA, USA

How genuinely new protein-coding genes originate is a central question in biology. Many novel genes arising *de novo* from intergenic genomic DNA long considered to be "junk", or from long non-coding RNAs, were recently found in Eukaryotes. Novel genes are taxon-restricted, may encode structurally novel proteins with new protein domains and thus illuminate the emergence of protein structure. To understand how novel genes arise, we built a mathematical model based on gene and genome parameters, and dynamic factors such as mutation and recombination. Combining phylostratigraphy and proteogenomics, we identified novel genes in >100 eukaryotic genomes ranging from human to paramecium and evaluated their predicted biophysical properties. Compared to ancient proteins, novel proteins are shorter, more fragile, disordered and promiscuous, yet less prone to forming toxic aggregates. We experimentally measured structure content and protease resistance of novel human proteins, and showed novel genes function *in vivo* in zebrafish brains. We showed novel human genes have fewer regulatory elements than ancient genes but more than control intergenic open reading frames. Our GTEx RNA expression analysis shows novel human genes are expressed in many tissues. Our gnomAD mutational constraint analysis shows some novel human genes are functional. Using single-cell RNA-Seq and proteomics, we found novel genes are expressed in human brains at multiple ages. Thus, genomic sequence turnover generates many novel genes encoding short proteins with distinct structural features and functioning in the brain. Variation in large eukaryotic genomes having large intergenic "dark matter" regions continuously generates new protein structures and new functions.

Population model characterizes genomic regions that are fertile for *de novo* gene birth

Somya Mani and Tsvi Tlusty
Institute for Basic Science – Center for Soft and Living Matter, South Korea

A gene can be viewed as a complex assemblage that endows a DNA sequence with the ability to reliably express a functional product. Any evolutionary process that involves assembling these components of a gene from scratch appears, at first glance, tortuous. Yet, ample evidence indicates that *de novo* gene birth is very widespread among organisms. Within genomes, *de novo* genes are found to be enriched in the vicinity of established genes.

In this work, we ask about the features of genomic loci that make it suitable for *de novo* gene birth. We approached this question with a population genetic model, and we characterize genomic loci by the distribution of fitness effects (DFE) of mutations incident on them.

In our model, adaptation of non-genic loci required DFEs rich in beneficial mutations, and was very rare for loci with predominantly deleterious DFEs. Our study thus indicates that genomic loci associated with *de novo* gene birth could harbor DFEs that are more beneficial than expected. Equally, loci in the vicinity of established genes can be expected to have non-negligible expression levels due to read-through, and they are genetically linked to a conserved gene: These factors could increase the chances of adaptation and gene birth. Additionally, it has been empirically observed that even loci that show signs of adaptation in a population are frequently lost. Our simple model is able to recapitulate this phenomenon.

Satellite Meeting on De Novo Gene Birth

Title

NCYM, a human *de novo* evolved gene product, promotes tumorigenesis in a mouse cholangiocarcinoma organoid model

Authors

Kazuma Nakatani[1,2], Daisuke Muto[1,2], Akiko Endo[2], Hiroyuki Kogashi[1,2], Hidefumi Suzuki[3], Hidehisa Takahashi[3], Yoshitaka Hippo[1,2,4], Yusuke Suenaga[2]

[1]Graduate School of Medical and Pharmaceutical Sciences, Chiba University, Chiba, Japan

[2]Laboratory of Evolutionary Oncology, Chiba Cancer Center Research Institute, Chiba, Japan

[3]Department of Molecular Biology, Yokohama City University Graduate School of Medical Science, Kanagawa, Japan.

[4]Laboratory of Precision Tumor Model Systems, Chiba Cancer Center Research Institute, Chiba, Japan

Abstract (up to 250 words)

NCYM, a *cis*-antisense gene of *MYCN*, encodes a Homininae-specific protein that promotes the aggressiveness of human tumors. High *NCYM* expression is associated with activation of the Wnt/ß-catenin pathway and Ras signaling pathway, both of which play critical roles in the cholangiocarcinoma development. Although high expression of *NCYM* is associated with poor prognosis in human cholangiocarcinoma, the contribution of NCYM to tumorigenesis remains unknown. In this study, we demonstrated that NCYM promotes tumorigenesis in a mouse cholangiocarcinoma organoid model. We induced the expression of KrasG12D, a cancer-associated mutant, along with NCYM in mouse normal bile duct organoids and inoculated them in nude mice. As a result, organoids expressing both KrasG12D and NCYM formed tumors at a high frequency of 13/15 (86%), compared to organoids expressing KrasG12D mutation alone, which formed tumors at a frequency of 12/20 (60%). Gene ontology analysis revealed that genes upregulated by NCYM overexpression were associated with the "cell cycle." Measurement of open reading frame (ORF) dominance score, an index correlating with translation of RNA, revealed that bile duct organoids overexpressing NCYM had high ORF dominance transcripts compared to the control. These results suggest that NCYM

directly promotes tumorigenesis in cholangiocarcinoma through the global changes in translation of transcripts.

The role of non-canonical open reading frames in Mendelian disease

Anne O'Donnell-Luria[1,2,3], Daniel Marten[1,2], John Prensner[1,2,3], Amir Karger[3], Victor Luria[1,3,4]

[1]Boston Children's Hospital, Boston, MA, USA; [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [3]Harvard Medical School, Boston, MA, USA; [4]Yale University School of Medicine, New Haven, CT, USA

While the existence of non-canonical open reading frames is well-established, the significance of variants within their sequences remains a major challenge for the genetics research community. This topic becomes all the more complex given the generally poor understanding of which non-canonical ORFs are functional and how these might function. Therefore, developing approaches to understand these variants is an important next step in deciphering the genetic causes of human disease. Elegant work in human genetics spanning the past 25 years has demonstrated that some monogenic human diseases are caused by variants in unannotated protein-coding genes, including through the creation of a non-canonical ORF that generates a new translational start site, a frameshift, or loss of a termination codon. Rapid expansion of genome sequencing approaches in people affected or unaffected by disease enables deciphering variation within these unannotated genes. Here, we will examine the evidence supporting functional variants within non-canonical ORFs. We will emphasize both the challenges and promise in identifying and interpreting variants found in non-canonical ORFs, and we will explore the latest developments in investigating these variants via computational, evolutionary biology, and functional techniques.

ABSTRACT

De novo gene intra-species diversity in *Saccharomyces cerevisiae*

Chris Papadopoulos[1], José Carlos Montañés[1], M. Mar Albà[1]

[1]Evolutionary Genomics Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona 08003, Spain.

Whereas *de novo* gene birth is nowadays a well-established mechanism for the formation of new genes at relatively long evolutionary timescales, its importance in intra-species diversification has barely been considered. In previous studies, we have successfully used transcriptomics data from different species to identify hundreds of *de novo* genes in *S. cerevisiae*. In this study, we perform deep and strand-specific RNA-sequencing, together with de novo transcriptome assembly, of several widespread *Saccharomyces cerevisiae* strains, with the aim of studying the diversity of *de novo* genes within the yeast population. For this, we have developed a specific computational pipeline that combines orthology with genomic synteny, and which discriminates between transcripts that are fixed in the population and the ones that are only expressed in a subset of the strains. The results provide new light into the emergence of new genes and how they impact intra-species diversification.

Unraveling the Origin, Evolution, and Role of de novo sORFs: A Case Study in the Arabidopsis Genus

Claire Patiou, Sylvain Legrand, Flavia Pavan, Christelle Blassiau, Vincent Castric, Eléonore Durand

The emergence of new, advantageous phenotypes remains a captivating question in evolutionary biology. While various biological and environmental mechanisms contribute to this process, gene birth has emerged as a significant source of phenotypic innovation across the tree of life. Traditionally, evolution was viewed as a "tinkerer," relying on the reuse of pre-existing genetic material to generate novelty. However, recent research has shed light on the prevalence of "de novo" gene birth in a wide range of species.

Such young genes, often characterized as shorter than older ones, can initially be considered as short open reading frames (sORFs). Still, little is known about the abundance, evolutionary dynamics, and function of sORFs; especially among closely related species. Hence, our research focuses on species within the Arabidopsis genus, aiming to investigate the early stages of gene birth through an evolutionary approach to sORFs found in intergenic regions and non-coding RNAs.

To address these questions, we employ a comprehensive methodology centered around a short evolutionary timescale. Our approach combines long read RNA-seq, Ribo-seq, syntenic and population genetics analyses. Through these techniques, we aim to characterize the expression patterns, conservation profiles, evolutionary dynamics, origins and functions of sORFs. By doing so, we seek to provide a detailed understanding of the emergence of functionality from non-genic material.

The origin and structural evolution of *de novo* genes in *Drosophila*

Junhui Peng, Li Zhao

Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA

Understanding how genes originate and evolve is crucial to explaining the origin and evolution of novel phenotypes and, ultimately, the diversity of life. *De novo* gene origination from nongenic sequences has been proved as a relatively common mechanism for gene innovation in many species and taxa. These young proteins provide a unique set of candidates to study the structural and functional origination of proteins. Despite extensive efforts to identify *de novo* genes in different species, our understanding of their protein structures and how they originate and evolve are still very limited.

In this study, we applied highly accurate reference-free progressive whole genome alignments and identified 555 *de novo* gene candidates in *D. melanogaster* that originated within the *Drosophilinae* lineage. We found a gradual shift in sequence composition, evolutionary rates, and expression patterns with their gene ages, indicating possible gradual shifts of their functions. Interestingly, we found little overall changes in structural properties (structural disorder and probability of being transmembrane proteins) for proteins encoded by *de novo* genes along their evolutionary trajectories. Single-cell RNA-seq analysis in testis showed that although most *de novo* genes are enriched in spermatocytes, several young *de novo* genes are biased in the early spermatogenesis stage, indicating potentially important but less emphasized roles of early germline cells in the *de novo* gene origination in testis.

We further combined AlphaFold2, ESMFold predictions, and molecular dynamics (MD) simulations to study the 3D structures and structural evolution of the identified *de novo* gene candidates. Our results suggest that while many of these candidates are highly disordered, a small subset may be well-folded. Most of the potentially well-folded *de novo* gene candidates adopt known structural folds, but some may have novel structural folds. By using ancestral sequence reconstruction and structural modeling, we found that these potentially well-folded proteins are often born folded. We observed one case where disordered ancestral proteins become ordered within a relatively short evolutionary time frame, suggesting that even for evolutionarily young genes, protein structures evolve slowly. Altogether, we provide a systematic overview of the origination, evolution, and structural changes of *Drosophilinae*-specific *de novo* genes.

Nicolas Svetec*, Samuel Khodursky, Sylvia Durkin, Evan Witt, Junhui Peng, Alice Gadau, Christopher B Langer, Vivian Yan, Li Zhao
Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA
*Correspondence to: nsvetec@rockefeller.edu

The birth and function of de novo genes in the *Drosophila* brain.

Many studies have suggested that the testis is a hotbed for evolutionary innovations such as duplicated, orphan, and de novo genes. While de novo genes appear to adhere this rule and are clearly biased toward the testis, one intriguing aspect of their tissue distribution is their relative prevalence in the brain. While more brain-biased de novo genes are discovered, why the brain tissue is a hotbed for de novo genes and what function de novo genes achieve in the brain remain unanswered. To tackle this problem, we sought to use a comparative and population genomic approach to identify recently born *de novo* genes in the *Drosophila* brain. By comparing the expression patterns between sexes, we provide a hypothesis about the role of sexual selection in de novo gene birth in this somatic but sexually dimorphic tissue. Moreover, we use a breadth of genomic techniques to identify the cell types and characterize the chromatin states associated with de novo gene origination. Finally, I will detail our findings related to the functional characterization of very recently born brain de novo genes. Notably, we found that brain-expressed de novo genes are expressed in neuronal cell types and exhibit tightly regulated expression patterns through time and cell types.

A large-scale analysis of genetic novelty in budding yeast

Emilios Tassios, Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", 34 Fleming Street, 16672, Vari, Greece

Abstract

De novo gene emergence, although thought of as improbable for decades, is an important source of novelty for organisms, linked with phenotypic innovations and species-specific characteristics. Here, we conducted the largest scale computational investigation of de novo gene emergence to date, exploiting a rich dataset comprised from 332 budding yeast genomes, spanning the entire biodiversity of the Saccharomycotina subphylum. We were able to identify over 400,000 taxonomically restricted genes (TRGs) at different phylogenetic levels, from species-specific ones to conserved across the Saccharomycotina. This enabled us to reveal macro-evolutionary trends of gene and protein properties that hold across yeast lineages, including that GC% of genes do not change with age while intrinsic protein disorder consistently decreases. By employing synteny analysis, we isolated more than 10,000 de novo genes. Additionally, we found thousands of TRGs that have diverged beyond recognition and have properties contrasting those of de novo genes such as longer length and lower biosynthetic cost. Furthermore, we investigated the cryptic property of intergenic regions to encode transmembrane (TM) domains, if theoretically translated, more frequently than expected by chance, a finding previously reported in baker's yeast. This TM-forming enrichment is present genome wide and is not explained by the hydrophobic content of the sequences nor their size and composition. Finally, we found a correlation, across species, between this intergenic enrichment and the number of TM domains in evolutionarily young genes hinting towards a link to de novo emergence.

Abstract

Traditional view on organismal adaptation postulates that mutations in genotype drive the phenotypic variation which allows to sample diverse adaptive strategies towards the environment. In our work we explore different modes of adaptation where protein phenotypic variations arise directly from the frequent ribosomal mistranslations. Combination of mistranslations upon the genes might increase the adaptive capacity of the organism by sampling multiple phenotypic variations at the same time. Additionally, error sampling of phenotypic variants in combination with genetic variation might explain the rise of complex epistatic molecular traits emergence of which would be exceptionally rare otherwise. Here we present the significance of mistranslations on organismal fate on the example of *E. coli* TEM-1 beta lactamase enzyme. Protein was engineered in such a way that only mistranslation will rescue organisms from presented antibiotics. We report that mistranslation frequency directly correlates with organismal fitness allowing us to speculate towards alternative evolutionary pathways of protein function landscape exploration and potential discovery of novel activities.

Vyacheslav
Tretyachenko

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

Mistranslation in protein function space exploration

Ancestral Sequence Reconstruction as a tool to detect and study de novo gene emergence

Nikolaos Vakirlis[1*], Omer Acar[2] Vijay Cherupally[2] & Anne-Ruxandra Carvunis[2*]

Affiliations: [1] : Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari, Greece. [2] : Pittsburgh Center for Evolutionary Biology and Medicine, Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, United States.

*: Correspondence to: anc201@pitt.edu, vakirlis@fleming.com

Abstract

New genes can evolve from previously non-coding genomic regions through a process known as de novo gene emergence. Evidence suggests that this process is universal and has likely occurred throughout evolution and across the tree of life. Yet, confidently identifying de novo emerged genes remains challenging. Ancestral Sequence Reconstruction (ASR) is a promising approach for inferring whether a gene has emerged de novo or not as it an in theory enable to inspect whether its most likely ancestral sequences harbored protein-coding capacity. However, the use of ASR in the context of de novo emergence is still in its infancy and its capabilities, limitations and overall potential are largely unknown. Notably, it is difficult to formally evaluate the protein-coding capacity of ancestral sequences, particularly when new gene candidates are short. How well-suited is ASR as a tool for the detection and study of de novo genes? Here, we address this question by designing an ASR workflow incorporating different tools and sets of parameters and by introducing a formal criterion for predicting when protein-coding capacity originated. Applying this workflow on ~2,600 short, annotated yeast genes (<1,000 nucleotides), we found that ASR robustly predicts an ancient origin for most widely conserved genes, which constitute "easy" cases. For less robust cases, we calculated a randomization-based empirical P-value estimating whether the observed conservation between the extant and ancestral reading frame could be attributed to chance or selection. This formal criterion allowed us to pinpoint a branch of origin for most of the less robust cases, identifying 33 genes that can unequivocally be considered de novo originated since the split of the Saccharomyces, including 19 S. cerevisiae-specific ones. We find that the remaining, equivocal cases, may be explained by different evolutionary scenarios including rapid evolution, multiple losses, horizontal gene transfer as well as a very recent de novo origin. Overall, our findings suggest that ASR can be a valuable tool to study de novo gene emergence but should be applied with caution and awareness of its limitations.

Investigating *de novo* gene formation in human populations

Covadonga Vara[1], José Carlos Montañés[1], Chris Papadopoulos[1], Anikó Szegedi[1], Lucas Wange[1], M. Mar Albà[1,2]

[1]Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain.

[2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

De novo genes arise from previously non-genic sequences and, in contrast to other mechanisms of gene formation, such as gene duplication, the contribution of *de novo* genes to differences between human individuals and populations still remains largely unexplored. We hypothesize that the formation of *de* novo genes has a prominent role in intra-species differences. Thus, we are investigating the heterogeneity of *de novo* gene content in human populations and closely related great apes using ribosome profiling data from Lymphoblastoid Cell Lines (LCLs). We have used an already described catalog of human *de novo* genes, restricting the genes to ORFs with evidence of translation, and measured their translation levels in human individuals from different populations. Preliminary results show inter-individual diversity of human-specific (not shared with other close-by primate species) ORFs classified as *de novo.* Interestingly, the human-specific ORFs that are mostly shared within a population appear to have a higher expression than the less shared. We expect that this work will provide new clues to understand the process of *de novo* gene formation in primates.

Research article

The evolution and knock-out lethality of stepwise de novo genes in Drosophila

Shengqian Xia[1,†], Zihan Liang[2,6,†], Yuxin Peng[3,†], Yuan Gao[2], Zhicheng Wang[2], Yi Wei[2], Dylan Sosa[1], Unjin Lee[1], Laura Faulere[1], Hanim Nuru[1], Yiming Zhang[4], Chunyan Chen[5], Yong E. Zhang[5], Wei Zhang[4], Manyuan Long[1#], Jian Zu[3#], Li Zhang[2,#]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America
[2]Chinese Institute for Brain Research, Beijing, China
[3]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China
[4]State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, and School of Life Sciences, Peking University, Beijing, China
[5]Institute of Zoology, Chinese Academy of Sciences, Beijing, China
[6]Peking Union Medical College, Beijing, China

[†]These authors contributed equally to this work
[#]Corresponding Author

Abstract

Prevailing knowledge indicates *de novo* genes originating from non-coding sequences could have rapidly acquired essential functions in male fertility. However, firstly, the validity of irrefutable de novo genes remains controversial due to scarce stepwise evidence from non-coding to coding sequences. Secondly, it also remains uncertain that knock-out of these stepwise de novo genes could cause complete lethality. Additionally, orphan genes consist of fast-evolving old genes and *de novo* genes, are incorrectly annotated as *de novo* genes. Here, we identified 41 stepwise *de novo* genes in the *D. melanogaster* genome originating within 50 million years. These stepwise *de novo* genes are significantly deficit on the X chromosome and show an early-ORF-late-transcription pattern with less optimized for both transcription and translation. By machine learning classifiers built on them we predicted 31 putative *de novo* genes out of 89 orphan genes which means *de novo* genes accounts for one-third orphan genes. Both RNAi and CRISPR screening suggest that 5 out of 73 putative de novo genes show consistent essential phenotypes and one stepwise de novo gene CG43248 KO show complete lethal. Overall, our results provide a reliable dataset of *Drosophila* stepwise *de novo* genes and show that *de novo* genes comprises one-third orphan genes. Both *stepwise and* putative *de novo* genes can contribute to essential viability or fertility in *D. melanogaster*, which provides new insights into the origination and functionality of *de novo* genes.

The dynamics and regulatory mechanisms of de novo gene expression

Li Zhao

Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065

Understanding the origin and evolution of genes is crucial to explaining the origin and evolution of novel phenotypes and ultimately the diversity of life. Although it was previously thought that almost all new genes were derived from duplication-related processes, recent work has revealed that de novo genes, which are genes born from ancestrally non-genic sequences, also contribute to gene and functional innovation. However, the earliest steps in the birth process of de novo genes and how they are maintained in populations and species were largely unknown. Among these steps, the gain of expression and expressional regulation is crucial for the gain of a new gene, because it enables the sequence unit to be selected for or against as a gene. I will present our work on the origin of expression and expressional regulation. We used single-cell RNA sequencing to study the expressional dynamics. We used ATAC-seq and analytical tools to study the role of open chromatin conformation in the gain of gene expression. We also used deep learning to study the changes of chromatin state, providing hints on how poised sequences contribute to the gain of novel regulatory sequences and how they impact the expression of new genes.

SMBE Satellite meeting on De Novo Gene Birth

TITLES OF POSTERS

1 - Oyinoluwa O. Bola
Department of Biological Sciences; Sam Houston State University, Texas, USA
Construction of Fish Species Composition Surveillance Database of Lake Raven, Huntsville Texas using Environmental DNA

2 - Jianhai Chen
Department of Ecology and Evolution, The University of Chicago, Chicago, IL, USA
Novel protein-complex driven by de novo genes

3 - Jae Young Choi
Department of Ecology and Evolutionary Biology, University of Kansas
Transposition of the telomerase RNA underlies the Monkeyflower (Mimulus) telomere sequence diversification

4 - Lin Chou
Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
Recent tandem repeat expansion of a human de novo ribo-seq ORF encoded in an oncogenic miRNA MY

5 - D.M.K.C. Daundasekera
Department of Biology, Texas A&M University, College Station, TX, USA
Identifying serpentine adaptation genes by tracing evolutionary-genomic history of Streptanthus, Caulanthus and their allied genera (Brassicaceae)

6 - Luis Delaye
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Unidad Irapuato, Irapuato, Gto., México
The origin of a novel gene by overprinting in E. coli: an overview of its discovery 15 years later

7 - Md. Hassan uz-Zaman
Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA
Emergence and persistence of proto-genes in a long-term Escherichia coli evolution experiment

8 - Yunzhe Jiang
Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
Identification of horizontal gene transfer in Oryza sativa pan-genome

9 - Rishabh Kapoor
Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, MA, USA
Gene fusion and horizontal gene transfer

10 - Ashit Kumar Dutta
East Tennessee State University, Johnson City, TN, USA

T->C transitions disproportionally eliminate stop codons in Drosophila de novo ORFs

11 - Vinita Lamba
Department of Biological Sciences, University of Arkansas, Fayetteville, AK, USA
Exploring New Gene Formation and their Adaptive Significance in Antarctic Notothenioids

12 - UnJin Lee
Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY, USA
Origination of a Testis-Specific lncRNA

13 - Matthew A. Marano
Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA
The evolutionary origins of orphan genes

14 - Daniel Marten
Broad Institute of MIT and Harvard, Cambridge, MA, USA
Analysis of large human genomic datasets shows expression and mutational constraint increase with gene evolutionary age

15 - Mayra Mendoza
Department of Veterinary Integrative Biosciences, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX, USA
Assembly of repetitive sequences in long read-based chromosome-level Alpaca (Vicugna pacos) reference genome

16 - Seth O'Conner
Department of Biology UNC Chapel Hill, Chapel Hill, NC, USA
Investigation of Drosophila Melanogaster Gene Conservation Throughout 101 Drosophilid Genomes

17 - Sara Openheim
American Museum of Natural History, New York City, NY, USA
What's in a name? Strain names are poor predictors of cannabinoid concentration in commercially available Cannabis products

18 - Adekola Owoyemi
Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA
Winning Against Unbalanced Datasets in a Machine Learning De Novo Genes Prediction Algorithm in Angiosperms

19 - Nathan Rives
Department of Biological Sciences, University of Arkansas, Fayetteville, AK, USA
Evolutionary Origins and Mechanisms of Fish Antifreeze Protein in Unrelated Taxa: Insights into New Gene Birth

20 - Nozomu Saeki
University of Pittsburgh, Pittsburgh, PA, USA
Investigating the impact of *de novo* genes on ion transport and homeostasis in *Saccharomyces cerevisiae*

21 - Isabella Simon
Department of Biological Sciences, Sam Houston State University, Huntsville, TX, USA
Characterization of dennd5a/b during early embryonic development of zebrafish

22 - Jeffrey Vedanayagam
Developmental Biology Program, Sloan-Kettering Institute, New York, NY, USA
Meiotic drive and suppression: de novo protamine copies and hairpin RNAs fuel intragenomic arms races in the male germline

23 - Vighnesh Ghatpande
Department of Molecular Biosciences, University of Texas at Austin, Austin, TX, USA
Non-canonical translation events in preimplantation mouse development

24 - Aaron Wacholder
Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
A vast pool of short-lived de novo genes contribute to phenotype and fitness

25 - Yen-Wen Wang
Yale School of Public Health, Yale University, New Haven, USA
Origins of lineage-specific elements via gene duplication, relocation, and regional rearrangement in Neurospora crassa

26 - Chathuri Devmika Wickramasinghe
Department of Biology, The University of Texas at Arlington, Arlington, TX, USA
Recurrent co-domestication of PIF/Harbinger transposable element proteins in insects

Construction of Fish Species Composition Surveillance Database of Lake Raven, Huntsville Texas using Environmental DNA

Oyinoluwa O. Bola[1], Sabrina Perico[1] and Sharmin Hasan[1]
[1]Department of Biological Sciences; Sam Houston State University, Texas, USA

Precise fish species composition assessment plays a crucial role in the effort to preservation of biodiversity. Fish species composition estimation using eDNA is an established noninvasive method that causes minimal disturbance and harm to the natural environment. Existing databases of Huntsville Texas lack quantitative composition, geographical and time-series integrated data interpretation of local fish species. To solve the above problems, we aimed to construct a fish species composition surveillance database for a local freshwater lake, Lake Raven of Huntsville Texas using environmental DNA (eDNA). To collect eDNA, 1 L of water samples were collected from three different sites of Lake Raven. Followed by membrane filtration, eDNAs were extracted following standard extraction protocol. The extracted eDNAs are subjected to PCR amplification of the 12S rRNA gene (~170bp) using the universal MiFish-U primers. PCR-amplified fragments will be gel-purified and sequenced using the MiSeq platform. A homology search will be performed using blastn against the MitoFish database using the sequenced files as queries. Analyzing the sequencing data, we aim to visualize the fish species composition by making an interactive map. Thus this project aims to conduct a pilot study of a comprehensive understanding of the spatial and temporal distribution of fish species composition on a local scale which can further be used as a model to imply on a global scale.

Novel protein-complex driven by de novo genes

Jianhai Chen[1], Dong Wang[2, 3], Manyuan Long[1,*]

[1]Department of Ecology and Evolution, The University of Chicago, 1101 E 57th Street, Chicago, IL 60637;

[2]Division of Pharmaceutical Sciences, Skaggs School of Pharmacy & Pharmaceutical Sciences, University of California San Diego, La Jolla, 92093, California, USA

[3]Department of Cellular & Molecular Medicine, School of Medicine, University of California San Diego, La Jolla, 92093, California, USA

*Corresponding author: mlong@uchicago.edu;

Abstact: Recent studies on rice have showed concrete evidence that de novo genes can emerge from non-coding sequences with a stepwise mode to increase protein diversity. Given their recent origin, it's conceivable that many de novo proteins have not yet undergone extensive structural evolution. Consequently, they may exhibit a higher propensity for intrinsic structural disorder (ISD). Despite advancements in our understanding on ISD proteins, the relationship between ISD and de novo genes has been a subject of long-time intense debate. Based on our previously identified 175 de novo genes from rice genome and gene age dating for gene duplicates in this study, we revealed that 83.52% of de novo proteins are ISD in structure, which is sharply different from the genome-wide percentage of 34.99%. Both de novo genes and gene duplicates suggest a general pattern of higher degree of ISD in young genes than in old genes. Surprisingly, despite general reduction of ISD level during evolutionary time elapse, the fractions of ISD proteins in de novo genes are still very high after ~20 million years of evolution (82.11%), in comparison to duplicated genes at the same evolutionary timeframe (21.80%). Based on extensive RNASEQ gene co-expression analysis, we identified the highly correlated candidate partners for 130 de novo genes. Interestingly, around 45% of these de novo genes may interact with the protein serine-threonine kinases, suggesting that a high proportion of de novo genes may under the post-translational regulation process after gene birth. The protein-complex analyses revealed that most of these de novo genes could form compact protein ensemble and some novel protein-complexes could be formed, suggesting the potential regulatory role of these proteins in the world of protein complex and the diversity at the protein-protein interaction level.

Transposition of the telomerase RNA underlies the Monkeyflower (*Mimulus*) telomere sequence diversification

Surbhi Kumawat[1] and Jae Young Choi[1]
[1]Department of Ecology and Evolutionary Biology, University of Kansas

Telomeres are nucleoprotein complexes with a crucial role of protecting chromosome ends. Telomere consists of a TG-rich microsatellite sequence and dedicated telomere-binding proteins. Because of its vital functions components of the telomere are thought to be under evolutionary constraint. For instance, all vertebrates have an identical telomere sequence suggesting extreme purifying selection but in plants, telomere sequences deviate from this ultra-constraint and displays an enormous range of sequence diversity. What evolutionary mechanism is driving the sequence diversification is unknown. Here, we discovered in Monkeyflower (*Mimulus*) the telomere sequence is even variable between species of the same genus. We investigated the evolution of the *Mimulus* telomere sequence by studying the long noncoding telomerase RNA (TR). Telomere DNA sequence is determined by the TR template sequence, which is used by the telomerase for telomere DNA synthesis. We sequenced the transcriptomes of 16 *Mimulus* species and conducted a *de novo* transcriptomics analysis. TR was expressed in highly differentiating meristem tissue but not in fully developed tissues (*e.g.* leaf). We discovered three TR template sequence types; $(TTTAGGG)_n$, $(TTTCGGG)_n$, and $(TTTCGG)_n$, suggesting the *Mimulus* telomere has evolved through stepwise mutation process. Surprisingly, TR genomic regions were not syntenic with each other, indicating TR jumps into a new chromosomal location and evolved a telomere sequence change. We also discovered a first reported incidence of a species with two copies of TR, but only one was used for telomere DNA synthesis. In sum, *Mimulus* telomeres shed light on the processes underlying the dynamic plant telomere evolution.

Recent tandem repeat expansion of a human *de novo* ribo-seq ORF encoded in an oncogenic miRNA MYU

Lin Chou[*1,3,4], Shu-Ting Cho[*1,2,3,5], Anne-Ruxandra Carvunis[1,3]

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA; [2]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, 450 Technology Drive Rm. 426, Pittsburgh, PA, 15219, USA; [3]Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA; [4]Integrative Systems Biology Program, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA; [5]Joint CMU-Pitt PhD Program in Computational Biology, University of Pittsburgh, Pittsburgh, PA 15213, USA

*co-first authors

Expansion and contraction of tandem repeats (TRs) in proteins during evolution can often affect phenotypes, such as height and Huntington's disease in humans. However, not much is known about TR-containing genes that emerge *de novo* from non-genic sequences, especially their clinical relevance, and how these *de novo* genes and TRs affect each other's evolution. Here we analyzed the evolutionary history of orf143, a TR-containing translated human ORF that was previously identified using ribosome profiling (ribo-seq ORF). Orf143 is encoded in MYU, an oncogenic long non-coding RNA that binds to tumor-repressing microRNAs (miRNA). We found that orf143 contains six tandem repeats, with a repeat unit length of 11 amino acids, at its 3' end. Examination of the syntenic region in eutherian mammals and blast analyses revealed that orf143 are anthropoid-specific and that the orthologous start codon emerged in the ancestor of anthropoids through point mutations, suggesting the *de novo* origin of orf143. However, the orthologs in non-ape anthropoid species only cover the regions orthologous to the 5' (non-repetitive) portion of human orf143 because their stop codons separate the ORFs from the single repeat unit downstream. On the other hand, mutations made the downstream TR become the 3' region of the ORFs in apes. Notably, the incorporation of the TR into ORFs coincides with the TR expansion in the ape lineage. We hypothesized that the emergence of this ORF and its repeat expansion might suppress oncogenesis by facilitating the binding and retention of ribosomes on MYU, preventing MYU's binding of tumor-repressing miRNA.

Identifying serpentine adaptation genes by tracing evolutionary-genomic history of *Streptanthus*, *Caulanthus* and their allied genera (Brassicaceae)

D.M.K.C. Daundasekera, Elyssa R. Garza and Alan E. Pepper
Department of Biology, Texas A&M University

Abstract

Adaptation to extreme environments is an important problem in ecology and evolutionary biology. Serpentine soil, which has high concentrations of toxic heavy metals and low concentrations of essential plant nutrients, is an excellent model environment to study plant adaptations to harsh environments. The annual mustard, *Caulanthus amplexicaulis* var *barbarae* (CAB), which is serpentine tolerant, and its sister taxon, *Caulanthus amplexicaulis* var *amplexicaulis* (CAA), which is serpentine intolerant to study genetic mechanisms underlying serpentine tolerance. Several approaches (QTL analysis, coding sequence evolution, RNAseq) are being used in our laboratory to identify candidate genes for serpentine tolerance in CAB. In this study, we are using the gene ancestry of CAB and CAA as an additional tool to prioritize candidate genes.

To trace the ancestry of CAB and CAA, we are using ~30 species representing tribe Thelypodieae. Elucidation of the phylogenetic history of this group is challenging due possible introgression. Here, we are determining the evolutionary history of CAB and CAA by comparing highly resolved phylogenies from both organellar (chloroplast and mitochondria) and nuclear genomes. Comparison of plastid genomes has led to identification of natural selection in this phylogenetic group. Results from chloroplast and mitochondrial phylogenies show that the maternal lineage of CAB and CAA clade is likely a serpentine intolerant *Caulanthus* lineage. We are using gene-tree discordance between organellar and nuclear phylogenies to identify nuclear loci with paternal inheritance and explore the potential of using this information to prioritize candidate serpentine tolerance genes and test using synthetic biology and CRISPR/CAS9 mutagenesis approaches.

The origin of a novel gene by overprinting in *E. coli*: an overview of its discovery 15 years later

Luis Delaye[1]

[1]Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Unidad Irapuato, Cinvestav. Km 9.6 Libramiento Norte Carretera Irapuato-León C.P. 36824 Irapuato, Gto., México

E-mail: luis.delaye@cinvestav.mx

Here we make an historical assessment of the discovery and characterization of a novel gene in *Escherichia coli*. In 2008 we proposed that the gene *htgA* originated *de novo* by overprinting the gene *yaaW* in *E. coli* (Delaye et al. 2008). At that time, there was no proof of the existence of overlapped genes in bacterial genomes and the possibility of *de novo* gene origination was believed to be an extremely unlikely event (Jacob 1997). However, in 2009 a research group obtained experimental evidence of the existence of proteins coded by overlapped genes in *Pseudomonas fluorescens* Pf0-1 thus indicating that origin of genes by overprinting in bacteria is plausible (Kim et al. 2009). The functions of *htgA* and *yaaW* were further investigated by another research group by introducing strand specific stop codons (Fellner et al. 2013). They showed that *htgA* and *yaaW* have differential phenotypes. No further analyses have been performed on these overlapped genes ever since. However, the analyses described above showed early on that genes can originate *de novo* in bacterial genomes.

References

Delaye L, et al. The origin of a novel gene through overprinting in Escherichia coli. BMC Evol Biol. 2008 Jan 28;8:31.

Jacob F. Evolution and tinkering. Science. 1977 Jun 10;196(4295):1161-6.

Kim W, et al. Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens Pf0-1. PLoS One. 2009 Dec 24;4(12):e8455.

Fellner L, et al. Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaaW. FEMS Microbiol Lett. 2014 Jan;350(1):57-64.

Emergence and persistence of proto-genes in a long-term *Escherichia coli* evolution experiment

Md. Hassan uz-Zaman[1], Simon D'Alton[1], Jeffrey E. Barrick[1], and Howard Ochman[1]*

[1]Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA

*Corresponding author
E-mail: howard.ochman@austin.utexas.edu

Abstract

The phenomenon of *de novo* gene birth remains largely unexplored in bacteria despite the abundance of novel, lineage-specific genes in their genomes and the ease with which bacteria can be studied in an experimental context. We searched the historical record of the *Escherichia coli* Long-Term Evolution Experiment (LTEE) for the emergence of "proto-genes", cases where non-genic sequences have evolved stable transcription and/or translation. By analyzing RNA-seq and Ribo-seq data from LTEE populations, we found that non-genic regions frequently experience stable regulatory changes and identify numerous instances in which new mutations have produced novel RNAs and peptide species. Most proto-genes resulted from insertion-element activity or from chromosomal translocations that fused pre-existing regulatory sequences to previously silent regions of the genome. Prior to their gain of new expression, most of these regions lacked detectable upstream promoters or consistent expression in any of a wide range of growth conditions, suggesting that they represent genuine cases of *de novo* proto-gene emergence. Our findings show that even on experimental timescales, novel proto-genes can emerge and persist, thereby serving as potential substrates for new gene formation.

Identification of horizontal gene transfer in *Oryza sativa* pan-genome

Yunzhe Jiang[1,2,3], Kun Li[3], Chaochun Wei[3,*] and Mark Gerstein[1,2,4,5,6,*]

1. Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA
2. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA
3. Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
4. Department of Computer Science, Yale University, New Haven, Connecticut, USA
5. Department of Statistics & Data Science Yale University, New Haven, Connecticut, USA
6. Department of Biomedical Informatics & Data Science, Yale University, New Haven, Connecticut, USA

*Correspondence: ccwei@sjtu.edu.cn (C.W.), mark@gersteinlab.org (M.G.)

Horizontal gene transfer (HGT) is a well-documented phenomenon in prokaryotes, but its prevalence and mechanisms in eukaryotic genomes remain unclear. Our study leverages the pan-genome of Oryza sativa to systematically explore HGT events within a eukaryotic context. Unlike traditional reference genomes, pan-genomes encapsulate the full genetic diversity of a species, making them an ideal resource for identifying ongoing HGT events. Utilizing a method grounded in sequence composition bias and whole-genome alignments, we screened the Oryza sativa pan-genome against 1,179 eukaryotic genomes, identifying 521 and 587 non-redundant HGT events in the reference and novel genomes, respectively. To validate our findings, we employed 12 third-generation rice sequencing datasets (sequencing depth >100X) and nearly 13 TB of whole-genome sequencing data, corroborating the presence of HGTs in both reference and novel genomes. Our analysis delves into various dimensions of HGT, such as their length distribution, relationship with repetitive DNA elements, functional implications, and the role of various microorganisms (archaea, bacteria and viruses) as HGT intermediaries. This study is the first to systematically investigate HGT from the vantage point of the Oryza sativa pan-genome. Our results elucidate the distinctions between ongoing and historical HGT events, laying the groundwork for future investigations in other eukaryotic species.

Authors: Rishabh Kapoor and Cassandra Extavour

Gene fusion and horizontal gene transfer (HGT) are established mechanisms contributing to the emergence of new genes. Gene fusions occur throughout the tree of life and can generate novel functions by recombining domains with distinct functionalities While HGT was previously believed to be limited to transfers between prokaryotes, recent investigations have unveiled multiple instances of functional gene transfers from prokaryotes to eukaryotes. Notably, arthropods, a diverse phylum with stable associations with germline endosymbionts, have been found to experience numerous prokaryote-arthropod HGT events. However, the extent to which HGT followed by gene fusion can synergistically lead to novel genes with mixed evolutionary ancestry and the significance of post-transfer fusion in the co-option and neofunctionalization of transferred genes remain uncertain.

In this study, we present an in-depth examination of "chimeric HGT" genes, which consist of one or more horizontally transferred segments from non-metazoan sources within the same open reading frame (ORF) as segments of ancient metazoan ancestry. Leveraging sensitive homology detection and rigorous phylogenetic inference methods, we have developed a bioinformatics pipeline that enables the identification of HGT chimeras. We applied our pipeline to 237 high-quality annotated arthropod genomes and uncovered a high-confidence set of >100 chimeric HGTs across diverse arthropods. Ongoing bioinformatic and experimental investigations are shedding light on the diverse functions of HGT-chimeras, their evolutionary origins and patterns of molecular evolution. This comprehensive analysis advances our understanding of the underlying processes of gene birth from HGTs and their functional significance.

T->C transitions disproportionally eliminate stop codons in *Drosophila de novo* ORFs.

Ashit Kumar Dutta[1], Lev Yampolsky[1]
[1]*East Tennessee State University, Johnson City, Tennessee, 37604, USA.*

Theoretically, mutational biases and strand asymmetry should prevent elongation of a *de novo* ORF for the following two reasons. First, the C->T transition is the most common nucleotide substitution in many organisms, mostly due to deamination of cytosines. Second, deamination is much more likely in single-stranded DNA, so it will disproportionally affect the non-template (coding) strand in regions newly acquiring transcription. Because C->T transition can create stop codons but cannot eliminate them, this asymmetry should result in net gain of stops. Yet, we observed the inverse pattern, namely that stop codons in 103 *de novo Drosophila* genes were disproportionally eliminated by the opposite substitution T->C (contingency table using all substitutions in the gene region as a reference, Fisher's Exact Test, P<1.2E-6). Intriguingly, the corresponding opposite strand substitution, A->G, does not disproportionally eliminate stops, (adjusted for the target size), indicating that strand asymmetry plays a role in the mechanism that prefers the T->C transitions. We hypothesize that this mechanism may be the GC-biased gene conversion. However, we have not been able to detect any non-randomness of the location of newly elongated ORFs with respect to chromosomal region propensity for recombination and gene conversion, perhaps due to much fine grain recombinational hotspot proximity. An alternative hypothesis is that there is a yet unknown bias specific to single-stranded non-template strand that operated during transcription, thus affecting stop codons disproportionally.

Exploring New Gene Formation and their Adaptive Significance in Antarctic Notothenioids

Vinita Lamba, Xuan Zhuang

New genes have profound impacts on organisms, enhancing their potential for adapting to new environments by providing a diverse range of tinkering and novelties. Antarctic notothenioids (cryonotothenioids), the dominant fish group of the Southern Ocean (SO), are a classic example of adaptive radiation. The evolution of a new gene family, antifreeze glycoprotein (AFGP), which prevent freezing-related mortality, exemplifies the gain of adaptive traits in cryonotothenioids. This genomic innovation is directly associated with their thriving in the SO, making them an ideal model for investigating new genes evolution. Nevertheless, other new genes involved in their adaptation to the harsh environment remain largely unknown. Recent studies have revealed that transposable elements (TEs) play a pivotal role in genome size expansion in cryonotothenioids. However, it remains unclear whether these TEs have contributed to new genes formation. To determine the genetic origin of new genes and their roles in adaptation, we propose to identify new genes in cryonotothenioids compared to basal non-Antarctic notothenioids, employing phylostratigraphy and synteny-based comparative genomics approaches. We will analyze the genomes and transcriptomes of representative species, to uncover evolutionary processes underlying new gene emergence. We expect to discover new genes shared among cryonotothenioid species that are associated with cold adaptation, and lineage-specific new genes linked to the adaptation to their specific ecological niches. By examining the evolutionary mechanisms of these new genes, we aim to gain insights into the emergence of genetic novelties and determine whether there is a preferred pathway for new gene birth under strong natural selection.

Origination of a Testis-Specific lncRNA

UnJin Lee, Li Zhao

The Rockefeller University

While models of de novo gene origination are typically classified as being either “transcription-first” or “ORF-first,” a significant gap in both models is understanding the evolutionary mechanisms driving the transcriptional activation of previously inert promoter sequences. To explore the mechanisms underlying promoter activation, we report a newly characterized, repeat-rich, species-specific *de novo* lncRNA, *gumiho*, whose expression is restricted to the *D. melanogaster* testis and is associated with a 60%-100% fertility reduction in knockout males. Comparative analysis of the *gumiho* locus reveals that the species-specific activation of *gumiho* is associated with transcriptional activation of a cluster of 4 neighboring genes in the testis. This cluster includes a highly conserved, essential gene which is under the control of multiple functionally validated enhancer elements. Interestingly, the activation of the *gumiho* promoter appears to be the result of a single nucleotide substitution that alters the binding affinity from a *dif* motif to a *dl* motif. Consistent with “transcription-first” models, analysis of ChIP-Seq and ATAC-Seq data shows that *gumiho* contains a large number of transcription factor binding sites within a large region of accessible chromatin in *melanogaster* head tissue. Alternatively, consistent with “ORF-first” models, 615 of *gumiho*’s 856 bases contain a long open reading frame that appears to be lowly translated in male head tissue. While it is currently unclear whether the transcriptional activation of *gumiho* was driven by the *cis*-regulatory activity upstream of pre-existing neighboring genes or downstream *trans*-function of some part of *gumiho*’s ORF or lncRNA, our continuing work highlights the intrinsic complexities of understanding *de novo* gene origination.

The evolutionary origins of orphan genes

Matthew A. Marano[1], Adekola Owoyemi[2], Ze Fang[3], Rigoberto Ayala Zaragoza[4], Andres A. Neira[4], Alex Samano[5], Andres David Barboza Pereira[6], Ramya R. Bathala[4], Claudio Casola[1,2]

[1]Interdisciplinary Doctoral Degree Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, USA

[2]Department of Ecology and Conservation Biology, Texas A&M University, College Station, USA

[3]Department of Soil and Crops, Texas A&M University, College Station, USA

[4]Department of Biochemistry and Biophysics, Texas A&M University, College Station, USA

[5]Interdisciplinary Doctoral Degree Program in Genetics and Genomics, Texas A&M University, College Station, USA

[6]Department of Biology, Texas A&M University, College Station, USA

Young genes are key contributors to adaptation and phenotypic innovation. Genes lacking sequence homology outside a given taxon, known as orphan genes, are often used as a proxy for young, lineage-specific genes, because they can readily be identified using relatively simple bioinformatic strategies. Traditionally, orphan genes have been described as the result of two distinct mechanisms: duplication-divergence of pre-existing genes, and de novo gene birth from ancestrally noncoding DNA. However, several other molecular processes are known to generate novel genes that fall within the definition of orphan genes, including transposable element recruitment, overprinting and horizontal gene transfer. Here, we summarize the variety of mechanisms by which orphan genes are formed and elucidate the impact that genes with different origination modalities can have on genome evolution. Furthermore, we address how taxonomic, phylogenetic and sequence evolution biases can affect gene age estimates, leading to incorrect predictions about orphan genes. Finally, we propose guidelines to correct for biases in gene age inferences and to discriminate between orphan genes that originate via different mechanisms.

Title: *Analysis of large human genomic datasets shows expression and mutational constraint increase with gene evolutionary age*

Author list: Daniel Marten, William Phu, Amir Karger, Victor Luria, Anne O'Donnell-Luria

While evolutionarily novel genes arise *de novo* and were discovered experimentally by proteomics and ribosomal profiling, their expression and mutational constraint were not previously determined on a genome-wide scale. We leveraged two large databases – the Genome Aggregation Database (gnomAD) and the Genotype-Tissue Expression (GTEx) database – to investigate gene expression and constraint as a function of gene evolutionary age. We used GTEx (contains RNA-Seq data from 54 adult human tissues, 948 total donors) to compare expression between all single-isoform human protein-coding genes, including novel, unannotated, open reading frames (ORFs) identified from proteomics and Ribo-Seq, and control intergenic ORFs (without expression evidence) and intergenic non-ORF sequences. We found gene expression of novel ORFs is higher than of intergenic sequences but lower than of Ensembl-annotated genes. Notably, novel ORFs overlapping elements of annotated genes (introns, untranslated regions) have higher expression than those not overlapping annotated genes. We compared all genes by evolutionary age (estimated by phylostratigraphy) and found young genes – Primate-specific, Mammal-specific – have lower expression than ancient genes. To determine how constraint varies with gene age, we tested 18,059 annotated human genes with constraint information in gnomAD v2 (contains exome or genome data from >140,000 individuals in the general population). We found evolutionary age positively correlates to loss-of-function constraint and haplo-insufficiency. In summary, expression and constraint strongly depend on gene age, indicating a gene evolutionary maturation process exists, such that genes increase their expression and constraint with time. These approaches will enable us to prioritize genes with important functions in humans.

Assembly of repetitive sequences in long read-based chromosome-level Alpaca (*Vicugna pacos*) reference genome

Mayra Mendoza[1], Kylie Munyard[2], Terje Raudsepp[1], Brian W. Davis[1]
[1]Department of Veterinary Integrative Biosciences, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX, USA
[2]Faculty of Health Sciences, Curtin Medical School, Curtin University, Perth, Australia
[3]Department of Small Animal Clinical Sciences, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX, USA

A reference genome is indispensable to study the genetic health and structure of populations, genome evolution, morphological variation, and heritable disease. As such, we present *VicPac4,* a chromosome-level alpaca genome, generated by combining PacBio, Hi-C and Bionano technologies. The current assembly is 2,572 Mb, with 1200 contigs (N50 = 40.28 Mb) and 783 scaffolds (lengths: 6.7 to 127.7 Mb, N50 = 67.53 Mb). Telomeric TTAGGG repeats were positioned in 36 chromosome ends. We also analyzed the multicopy sequences of 18S-5.8S-28S rDNA, known as Nucleolar Organizer Regions (NORs). While rDNA sequences in the NOR are highly conserved across eukaryotes; their number, chromosomal location, and transcriptional units per locus show considerable variation between species and are of interest for comparative and evolutionary genomics. The expansion, contraction, and translocation of NOR within and between species has potential functional consequences, as evidenced by their association with pathologies such as Minute Chromosome Syndrome in infertile female alpacas. Yet, the number and location of NOR loci in camelids remains ambiguous. Here we applied dual-color fluorescence *in situ* hybridization to map NOR loci in four camelid species. Alpacas NOR mapped to chromosomes 12, 20, 21, 27, 28, 33, 34, 35; llamas 21, 27, 28, 31, 33, 34, 35, 36; dromedaries 4, 17, 23, 25, 27, 28, 29, 34, 36; and Bactrian camels 4, 23, 28, 36, with 6 more loci to be confirmed. The findings facilitate incorporation of these multicopy sequences into recent high-quality genome assemblies and provide valuable information for comparative genomics. Because all camelids have similar 2n=74 karyotypes, differences in NOR chromosomes can serve as molecular markers for identifying camelid species, subspecies, and interspecific hybrids.

Investigation of Drosophila Melanogaster Gene Conservation Throughout 101 Drosophilid Genomes

Seth O’Conner and Corbin Jones

Department of Biology UNC Chapel Hill

For most organisms, genome content is dynamic over evolutionary time. Newly sequenced genomes reveal that genes are gained and lost even among closely related species. The tempo and mode of these changes is, at best, only partially understood. Here, we describe when and how frequently gain/loss events of both protein coding and non-coding *Drosophila melanogaster* genes occur using 101 recently published, highly contiguous drosophilid genomes representing 93 different species. This work revealed that 14.4% of protein coding genes were missing from at least one of the 101 genomes, while 89.9% of ncRNA genes had at least one dropout. Furthermore, using publicly available transcriptome data in *D. melanogaster* to track expression trends in relation to gene conservation level showed that taxonomically restricted genes have a higher peak expression than conserved genes. Many of these taxonomically restricted genes appear to be secretion proteins with the highest peak expression in salivary glands. To see if this phenomenon was upheld in other eukaryotes, a similar dataset was produced for *Arabidopsis thaliana* protein coding genes. Surprisingly, the trend was flipped. Our results also suggest that while the use of closely related genomes lowers the chance of homology detection failure errors, it is still important to implement tools to verify true gene dropouts vs. failures to detect. Overall, our work illuminates the differing genome dynamics affecting protein coding and non-coding genes in *D. melanogaster* while providing a resource for studying gene-level diversification throughout many closely related drosophilid species.

What's in a name? Strain names are poor predictors of cannabinoid concentration in commercially available *Cannabis* products.

Sara J Oppenheim & Robert DeSalle
American Museum of Natural History, Invertebrate Zoology, New York City, NY, USA.

"'*Must* a name mean something?'" asks Alice in *Through the Looking*-Glass. She is talking about her own name, not about *Cannabis* strain names, but the question applies equally to all names. In the case of *Cannabis*, strain names are assumed to be good predictors of the phenotype--and phenotypic effect--of a gummy, vape, joint, or other *Cannabis* product. But if strains are not "real" then the THC content of a product labelled "Maui Wowie" may be no more like another Maui Wowie product than it is like an Alaskan Thunderf*ck product.

We are using a GWAS approach to examine the genetic basis of variation in THC:CBD ratios and other cannabinoid-related traits in over a thousand strains of commercially grown *Cannabis*. To date, we have identified many significant SNPs that do not coincide with any known candidate gene. Instead, we find a variety of other genes implicated, and some of these lack obvious orthologs in other plant genomes, suggesting that they may be fairly recent products of *de novo* gene birth processes.

During our GWAS work we have discovered that the admixture profiles of some samples with the same strain identity are highly variable. This intra-strain diversity of origin means that recreational and medicinal users of *Cannabis* may experience unexpected and mostly unpleasant phenotypic effects when they dose themselves with a product that is not at all what they expected. We present here our preliminary findings on the age-old question, "would a rose by any other name, smell as sweet?"

Winning Against Unbalanced Datasets in a Machine Learning *De Novo* Genes Prediction Algorithm in Angiosperms

Adekola Owoyemi[1], Alan E. Pepper[2,3,4], Thomas R. Ioerger[5] and Claudio Casola[1,2,4]

[1]Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA 77843
[2]Department of Biology, Texas A&M University, College Station, USA
[3]Interdisciplinary Doctoral Degree Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, USA 77843
[4]Interdisciplinary Graduate Program in Genetics and Genomics, Texas A&M University, College Station, USA 77843
[5]Department of Computer Science and Engineering, Texas A&M University, College Station, USA

De novo genes (DNGs) evolve from ancestrally noncoding DNA sequences and have been shown to form a substantial fraction of novel genes in eukaryotes. However, outside well-characterized model species, the prevalence of these genes remains unknown due to limitations in current methods to discover DNGs. We have recently shown that machine learning algorithms (MLAs) trained on simple DNA and protein sequence features can rapidly identify the vast majority of DNGs in angiosperm genomes. However, these models generated a high number of false positives, primarily due to the overwhelming number of ancestral (non-*de novo*) genes, or AGs. Here, we sought to explore novel MLA methods and develop strategies to improve precision in DNG identification. Using hundreds of previously identified DNGs from three angiosperm genomes, we trained and tested five types of MLAs: decision tree (DT), neural network (NN), random forest (RF), light gradient boost machine (LGBM), and naive bayes (NB). To address the unbalanced nature of DNG/AG datasets, we used balanced datasets obtained through a variety of resampling models derived from the synthetic minority over-sampling (SMOTE) algorithm. We found that some combinations of SMOTE with RF, LGBM and NN algorithms led to precision scores up to 97% (vs <90% without SMOTE) while maintaining recall levels above 80%. These results indicate that balanced DNG/AG datasets have the potential to significantly limit the impact of false positives in MLA-based DNG surveys.

Evolutionary Origins and Mechanisms of Fish Antifreeze Protein in Unrelated Taxa: Insights into New Gene Birth

Nathan Rives

To understand the process of gene birth and the emergence of novel traits, it is essential to investigate the evolutionary origins and mechanisms of new genes. The diverse fish antifreeze protein (AFP) provides a unique system for studying new gene and novel function evolution, as they have reasonably recent origins, and their lifesaving function of preventing freezing is clearly defined and linked to natural selection. In this study, we investigate the genomic origin and evolutionary process of type I AFP (AFPI) genes in three distinct fish lineages. We sequenced the complete genome of two AFPI-containing (AFPI$^+$), each belonging to separate lineages. We also incorporated data from other AFPI+ species available in public databases and included AFPI-lacking (AFPI-) outgroup species for comparison. We isolated the genomic loci containing *AFPI* or homologous regions and annotated the *AFPI* to illustrate their gene structure. By comparing the new gene and their homologous sequences in these genomes, we found that the *AFPI* in the three lineages exhibit dissimilar gene structure and their neighboring genomic regions do not share microsynteny, suggesting that AFPI independently evolved in each lineage. The *AFPI* in each AFPI$^+$ lineage was linked to homologous regions in outgroup species that contain the ancestral sequence. Interestingly, the near-identical *AFPIs* originated from different precursor genes in each lineage. They arose from functionally unrelated precursor gene, and de novo evolved coding regions with novel antifreeze functions. This study provides an illustrative example of how novel functions can arise and advance our understanding of new gene formation.

Investigating the impact of *de novo* genes on ion transport and homeostasis in *Saccharomyces cerevisiae.*

Nozomu Saeki, Anne-Ruxandra Carvunis

Department of Computational and Systems Biology, School of Medicine, University o Pittsburgh, Pittsburgh, PA 15213, USA

Ions, such as potassium, calcium, and iron, are vital for cellular function and must be obtained from the environment through various transport mechanisms. The efficiency of ion uptake can vary based on the characteristics of the extracellular environment, driving the evolution of ion transport and homeostasis mechanisms tuned to these environmental conditions. However, the genetic mechanisms underlying these adaptations are not fully understood. We focus on *de novo* genes—new genes that arise from previously non-coding sequences—as potential drivers of these evolutionary adaptations. A previous co-expression analysis in *Saccharomyces cerevisiae* found that 49% of tested noncanonical ORFs, predominantly *de novo* genes, are co-expressed with canonical ORFs, mainly conserved genes, involved in monoatomic ion transport (Rich et al., 2023). These findings led us to hypothesize that the emergence of *de novo* genes could be a crucial cellular strategy for modifying ion acquisition processes. To test this hypothesis, we employed two strategies: examining the effects of both knockout and overexpression of *de novo* genes on ion transport and homeostasis and performing an environmental screen to identify beneficial changes in ion homeostasis. Through this approach, we found that the absence of certain *de novo* genes was detrimental under high sodium conditions. Furthermore, we identified an association between specific ions and *de novo* genes. In summary, our research provides a deeper understanding of the evolution of ion homeostasis and highlights the significant role *de novo* genes play in critical physiological processes.

Characterization of *dennd5a/b* during early embryonic development of zebrafish

Isabella Simon[1], Alicia Mendoza[1] and Sharmin Hasan[1]

[1] Department of Biological Sciences, Sam Houston State University, Huntsville, TX, USA

During early embryonic development, the WNT family of evolutionarily conserved signaling molecules plays important roles in cell fate determination, cell proliferation, cell motility, and establishment of the primary axis. In the non-conical WNT pathway, the disheveled binding partner, *Daam1*, the disheveled associated activator of morphogenesis 1, is reported to be required for gastrulation and neural tube closure. In a screen for effector proteins of *Daam1*, a Guanine Exchange Factor (GEF), the *dennd5a* protein was identified. *dennd5a* can catalyze the conversion of inactive GDP-bound Rab proteins into active GTP-bound form. Mutation of *dennd5a* causes early infantile epileptic encephalopathy in humans though its role in early embryonic development remains largely unexplored. In this study, we are characterizing the role of zebrafish *dennd5a* and its paralogue, *dennd5b* during zebrafish (*Danio rerio*) development. Zebrafish *dennd5a/b* is temporally expressed maternally from the 1-cell stage to the Prim-6 stage. Our functional analysis of dennd5a/b using zebrafish embryos resulted in a compressed head, and a deformed tail phenotype suggesting an important role of *dennd5a/b* in development. Together with the expression and functional analysis, this study will better understand how *dennd5a/b* is expressed and functions during the early embryonic development of zebrafish.

Meiotic drive and suppression: *de novo* protamine copies and hairpin RNAs fuel intragenomic arms races in the male germline

Jeffrey Vedanayagam[1*], Marion Herbette[2], Ching-Jung Lin[1], Holly Mudgett[3], Caitlin McDonough-Goldstein[4], Stephen Dorus[4], Benjamin Loppin[2], Colin Meiklejohn[3], Raphaelle Dubruille[2], Eric C. Lai[1]

1. Developmental Biology Program, Sloan-Kettering Institute, New York NY 10065, USA
2. Claude Bernard Lyon I, 69622 Villeurbanne Cedex, France
3. School of Biological Sciences, University of Nebraska, Lincoln, NE 68588, USA
4. Center for Reproductive Evolution, Syracuse University, Syracuse NY

*Present address: Department of Neuroscience, Developmental and Regenerative Biology, University of Texas at San Antonio, San Antonio, TX 78249

Meiotic drivers “cheat” Mendel’s law of segregation to gain an unfair transmission advantage during reproduction. As a result, their harmful activities trigger opposing mechanisms of host suppression. We recently documented an expanding cohort of presumed meiotic drive factors located on X chromosomes of simulans-clade species (the *Dox* superfamily), which are in turn connected to hairpin RNA (hpRNA) suppressors that generate endogenous-siRNAs (*Nmy*/*Tmy* family). In the present study, we used *D. simulans* as a molecular genetic system to investigate meiotic drive and suppression by *Dox* loci and their hpRNA suppressors. Using CRISPR-mediated knockouts, we show that loss of these hpRNAs yields profound defects in male gametogenesis and reproductive performance. These include loss of Y-bearing sperm (resulting in near complete loss of sons by *nmy* mutants) or complete disruption of spermatogenesis (yielding sterility in *tmy* mutants). *nmy* and *tmy* mutants derepress distinct members of the *Dox* family, which in turn encode novel homologs of protamine. These HMG box factors are normally involved in compaction of sperm chromatin, providing a molecular basis for how Dox proteins may specifically derail spermatogenesis. Their causality is shown by phenotypic rescue of both hpRNA mutants with wild-X chromosomes that bear deletions of different *Dox* family loci. Overall, this work provides foundations to study the mechanistic basis of meiotic drive and suggests that RNAi-mediated defense of endogenous selfish genes may mediate reproductive isolation and speciation.

Non-canonical translation events in preimplantation mouse development

Vighnesh Ghatpande[1], Uma Paul[1], Can Cenik[1]
1. Department of Molecular Biosciences, University of Texas at Austin, Austin, TX, USA

In the early stages of vertebrate embryogenesis, post-transcriptional regulation plays a pivotal role in shaping the landscape of gene expression, particularly through translational control. Recent work led to the intriguing discovery of non-canonical short proteins, often referred to as microproteins, in model organisms like zebrafish that can modulate differentiation (Pauli et al. 2014. *Science*). These findings raise the prospect that similar micropeptides may also perform critical signaling functions in mammalian development. Until now, technological limitations have hindered the characterization of such microproteins in mammalian systems. We have recently pioneered a microfluidic isotachophoresis based technique (Ribo-ITP) that enables high quality and high coverage measurements of translation from single cells and embryos (Ozadam et al. 2023. *Nature*). Here, we used Ribo-ITP in single mouse embryos from 16-cell and 32- cell stages in combination with a translation initiation inhibitor to discover novel microproteins that are translated during these formative stages of preimplantation development. We developed a custom pipeline and combined it with existing approaches to detect >20,000 translated ORFs These events stem predominantly from upstream ORFs (uORF) in addition to those from downstream ORFs and long non-coding RNAs (lncRNA). The identified ORF encode small peptides less than 100 amino acids long with uncharacterized cellular functions. We are currently characterizing the function of these microproteins using mouse embryonic stem cells as a model system using high-throughput screens and targeted biochemical characterization. Our study reveals a rich landscape of microproteins are translated in early mouse development and we hypothesize that some of these may function as signaling peptides in dictating early cell fate decisions.

A vast pool of short-lived *de novo* genes contribute to phenotype and fitness

Aaron Wacholder[12], Saurin Bipin Parikh[123], Nelson Castilho Coelho[12], Omer Acar[124], Carly Houghton[124], Lin Chou[123], and Anne-Ruxandra Carvunis[12]

1. Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States
2. Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States
3. Integrative Systems Biology Program, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States
4. Joint CMU-Pitt Ph.D. Program in Computational Biology, University of Pittsburgh, Pittsburgh, PA, 15213, United States

Ribosome profiling experiments demonstrate that eukaryotic model organisms translate thousands of open reading frames (ORFs) outside of annotated coding sequences, but the significance of this “noncanonical” translation is poorly understood. Is noncanonical translation mostly biologically insignificant noise, or does it indicate thousands of never-studied proteins that play important biological roles? To address this question, we performed a comprehensive analysis of translation, evolution, and phenotype among unannotated ORFs in *Saccharomyces cerevisiae*.

Using a novel approach to aggregate ribosome profiling reads across hundreds of published experiments, we find that over 18,000 unannotated yeast ORFs are translated. Conducting a comparative genomics analysis at the species and population level, we find that more than 90% of unannotated translated ORFs are of recent *de novo* origin and show no signatures of purifying selection, suggesting membership in a class of ORF that experiences rapid evolutionary turnover. Despite their short lifespans, these *de novo* ORFs do affect phenotypes: employing a genetic screen in which start codons were disabled, we find that 14% of unannotated *de novo* ORFs provide fitness benefits in some environmental conditions, and overexpression screens show that many have gain-of-function phenotypes. Around 100 annotated yeast genes have evolutionary properties nearly identical to unannotated *de novo* ORFs, providing a window into their potential roles; these genes are involved in processes including DNA repair and post-transcriptional regulation.

Our results suggest that, though the great majority of *de novo* genes do not persist over evolutionary time, they nevertheless have major consequences for organism biology during their short lifespans.

Origins of lineage-specific elements via gene duplication, relocation, and regional rearrangement in *Neurospora crassa*

Yen-Wen Wang, Zheng Wang, Oded Yarden, and Jeffrey P. Townsend

The origin of new genes has long been a central interest of evolutionary biologists. However, their novelty evades reconstruction by the classical tools of evolutionary modeling. This evasion of insight from deep ancestral investigation necessitates intensive study of model species within well-sampled, recently diversified, clades. One such clade is the model genus *Neurospora*, members of which lack recent gene duplications, yet harbor clusters of lineage-specific genes (LSGs) adjacent to the telomeres. Several *Neurospora* species are comprehensively characterized organisms apt for studying the evolution of LSGs. Using gene synteny, we documented that 78% of *Neurospora* LSGs clusters are located in chromosomal regions featuring extensive tracts of non-coding DNA and duplicated genes. Here we report several instances of LSGs that are likely from regional rearrangements and potentially from gene rebirth. To broadly investigate functions of LSGs, we assembled transcriptomics data from 68 experimental data points and identified co-regulatory modules using Weighted Gene Correlation Network Analysis, revealing that LSGs are widely but peripherally involved in known regulatory machinery for diverse functions. The ancestral status of mas-1 and its neighbors was investigated in detail, suggesting that it arose from an ancient lysophospholipase precursor that is ubiquitous in lineages of the Sordariomycetes; mas-1 plays a role in cell-wall integrity and cellular sensitivity to antifungal toxins. Our discoveries illuminate a "rummage region" in the *N. crassa* genome that enables formation of new genes and functions to arise via gene duplication and relocation, followed by fast mutation and recombination facilitated by tandem repeats and deconstrained non-coding sequences.

Recurrent co-domestication of PIF/Harbinger transposable element proteins in insects

Chathuri Devmika Wickramasinghe, Dragomira N. Markova, Fatema B. Ruma, Claudio Casola, Ayda Mirsalehi and Esther Betrán

Abstract

Transposable elements (TEs) are selfish DNA sequences capable of moving and amplifying at the expense of host cells. Despite this, an increasing number of studies have revealed that TE proteins can contribute to the emergence of novel host proteins through molecular domestication. We previously described seven transposase-derived domesticated genes from the PIF/Harbinger DNA family of TEs in *Drosophila* and a co-domestication. PIF TEs distinguish themselves from other DNA transposons by the presence of two genes. We hypothesize that there should often be co-domestications of the two genes from the same TE because the transposase (gene 1) has been described to be translocated to the nucleus by the MADF protein (gene 2). After the exploration of hits to PIF transposases and analyses of their context and evolution in insect species genomes, we present evidence of six additional and independent PIF transposable elements proteins domestication events. Our results support that new gene origination through domestication of a PIF transposase is frequently accompanied by the co-domestication of a cognate MADF protein in insects. We propose a detailed model that predicts that PIF TE protein co-domestication should often occur from the same PIF TE insertion. We have been studying the function of four of these domesticated transposases in *Drosophila melanogaster*, named *Drosophila PIF Like Genes* (*DPLGs*) and reveal their potential regulatory functions.