מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

מאת
רן אשכנזי

שעתוק במהופך של גנים באמצעות הרטרוטרנספוזון Ty
והשפעתו על אבולוציה של שמרים

Thesis for the degree
Master of Science

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

By
Ran Ashkenazi

Reverse transcription of genes by the Ty retrotransposon and its impact on yeast evolution

מנחה:
פרופ' יצחק פלפל

Advisor:
Prof. Yitzhak Pilpel

אייר תשפ"ג

April 2023

Abstract:

The possibility for Lamarckian modes of evolution rests on the notion that phenotypic changes can either be inherited to the next generation or be converted into the genome. While the first option can be subserved by diverse epigenetic means, the latter requires a back flow of information, e.g., from RNA to genomic DNA. This reverse flow can be realized by the molecular process of reverse transcription (RT). Retrotransposons propagate in genomes via RT and here we examine the extent to which they can induce this process in other genes. RT can create new copies of reverse transcribed genes or replace existing copies with mature RNA via homologous recombination. This can propagate transcription and RT errors back into the genome and cause intron loss and therefore, can significantly impact gene evolution. Specifically, highly expressed genes are more likely to undergo RT, making it a potential agent of Lamarckian inheritance.

Ty elements in yeast are retrotransposons that form virus-like particles (VLPs) within the cytosol in which cellular mRNA can be contained, and potentially be reverse transcribed and subsequently incorporated into the genome. High throughput sequencing assays were performed on Ty1 VLPs in different conditions to determine their mRNA and cDNA contents. Here, we analyze these results and form a list of VLP enriched and depleted genes and of reverse transcribed genes.

We see that VLP depleted RNAs tend to be evolutionarily conserved while VLP enriched RNAs exhibit a high mutation rate. We also see that VLP enrichment and depletion are highly associated with specific RNA localization and are correlated with RNA half-life. We see that VLP depletion is associated with specific gene functions and GO categories.

We searched for evidence of cDNA within VLP layer DNA by testing for coding region bias using different methods. We find that, indeed, there appears to be cDNA within the VLP fraction, and we find the strongest evidence for it when cells were starved for nitrogen, a condition in which Ty1 retrotransposition is known to be induced. These data of DNA in the VLP allow us to characterize the genes which undergo RT.

We performed an evolution experiment on yeasts with induced Ty1, comparing their evolution to yeast without induced Ty1. We find a larger variation from the ancestor in the strains with induced Ty and are planning to further proceed with the experiment to better understand the observed difference.

Together my findings establish the molecular – cellular basis for potential RT of a selection of yeast genes and suggest potential evolutionary implications of the process.

Table of Contents

List of abbreviations

cDNA – Complementary DNA.

ER – Endoplasmatic reticulum.

GO – Gene Ontology.

LTR – Long terminal repeat.

Minus – The Ty-Minus strain.

N – Nitrogen starvation condition.

PC – Principal component.

PCA – Principal component analysis.

Plus – The Ty-Plus strain.

RP – Ribosomal protein.

RT – Reverse transcription.

SGD – Saccharomyces Genome Database.

TOTAL – RNA taken from the whole cell.

UMI – Unique molecular identifier.

VLP – Virus-like particle.

YP – YPD control condition.

Introduction

Lamarckian evolution is the notion that changes in an organism's attributes which are acquired during its lifetime can be transmitted through inheritance. This seemingly contrasts with the accepted notion of Darwinism, according to which inherited material is changed randomly and then passed on to the offspring, with fitness-raising characteristics retained in the population through natural selection. Though natural

selection and random changes are still considered the main drivers of evolution, different mechanisms such as epigenetics are known to facilitate a form of Lamarckian evolution. While epigenetics can propagate phenotypic changes for some generations, reversal of information flow from the phenotype to the genotype, if were possible, could have resulted in indefinite inheritance of acquired traits (Figure 1)

.

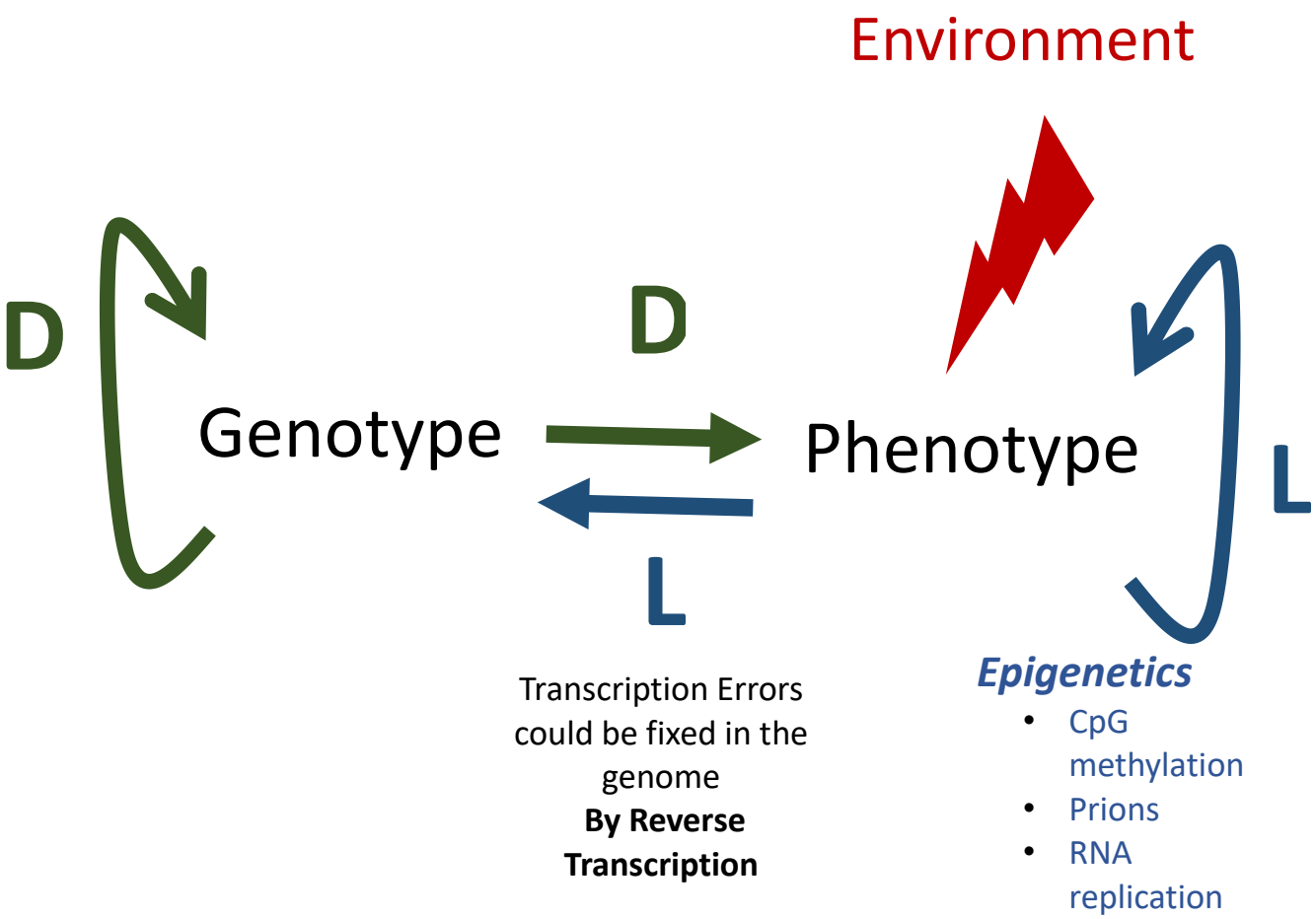


Figure 1. While epigenetics can propagate phenotypic changes caused by the environment, reverse transcription has the capacity to fixate such phenotypic changes, e.g., transcription errors in the genotype. D stands for Darwinian evolution and L stands for Lamarckian-like evolution.

The only known mechanism to reverse the flow of information from DNA is reverse transcription (RT) – conversion of RNA into DNA. Since RNA is part of the phenotype

and is modified constantly under the current environmental conditions in cells, its RT followed by incorporation into the genome can serve as a mechanism by which phenotypic changes enter the genome.

Ty1 is a long terminal repeat (LTR) retrotransposon in *S. Cerevisiae*. It generates new copies via RT of a genomic RNA (gRNA) intermediate [1]. It is one of 5 families of LTR retrotransposons in *S. Cerevisiae* - Ty1-5, which are the only known transposable elements in its nuclear genome [1]. Ty elements form virus like particles (VLPs) within the cytosol which contain the retroelement mRNA and the protein machinery required for its RT and integration into the genome [1]. Ty1 contains 2 overlapping ORFs – GAG and POL. The GAG ORF encodes the Gag protein product which comprises the VLP's nucleocapsid and the POL ORF encodes the Gag-Pol protein product which is subsequently cleaved to 3 catalytic proteins: protease, reverse transcriptase, and integrase. Protease cleaves the Gag and Gag-Pol proteins. Reverse transcriptase reverse transcribes the Ty gRNA – forming cDNA and destroying the gRNA in the process. Integrase integrates the cDNA to the genome, creating a new copy of it within the genome [1]

Ty1 can insert sequences into the genome via integration or homologous recombination [2]. While integration can create new copies of a given sequence in the genome, recombination can overwrite older copies and insert changes the RNA sequence underwent during and after its transcription.). There is past evidence of the Ty1 integrase integrating non-Ty1 DNA into the genome and forming retrosequences [3][4][5][6]. This process, when applied to non-Ty genes, has a potential to impact their evolution either by creating new gene copies or via modification of an existing copy by recombination, as it can potentially fixate within the genome any change the RNA encoded by the gene undergoes during or past its transcription, including modification, splicing, polyadenylation, transcription errors and editing. Potentially even more important, the most profound information that is stored in the population of RNA transcripts of a gene is their actual expression and expression level at a given environmental condition. So that even random and blind-to-gene-identity RT could

return into the genome, potentially as an additional copy, RNAs that were expressed more at a given condition.

One way in which genes are modified during their transcription is the accumulation of transcription errors. RNA polymerases have an error rate between 5.0e-6 and 1.6e-05 in yeast which is several orders of magnitude higher than the error rate of DNA replication [7]. RT of genes can allow these errors to enter the genome, potentially substantially increasing the mutation rate of transcribed genes at a given condition.

Another way in which RNA can be modified via RT is through intron loss. During post transcriptional processing, RNAs undergo splicing, during which their introns are excised from the sequence. RT of the processed sequence followed by incorporation into the genome can result in intron loss within the reverse transcribed gene [8]. Interestingly, *S. Cerevisiae* and all other hemiascomycetous yeast, have experienced significant intron loss, with only ~5% of their protein coding genes containing introns on average [8]. One of the most plausible explanations for intron loss is via homologous recombination of reverse transcribed complementary DNA (cDNA), made from the template of a mature intronless RNA, with the genomic locus containing the gene [8]. Therefore, if RT of genes by Ty is highly prevalent, it's likely to be a significant cause for the intron loss in yeast, via homologous recombination or via creation of intronless copies.

Transposition and expression of Ty, similarly to transposition of retroelements in many other organisms, are induced by stress [9]. Many different stresses were shown to induce Ty transposition e.g., ionizing radiation [10], different types of nutrient starvation [9], [11], DNA damage [12], [13], mutagens [14] and telomere erosion [5]. Ty is tightly regulated by the host cell, with evidence of over 200 different host factors regulating retrotransposition [1]. Another important regulator of Ty is the mediator complex whose different subunits can activate or repress Ty post transcriptionally with a difference in retrotransposition of up to an accumulative 5 orders of magnitude [15]. At least 1 mediator subunit has been shown to be potentially regulated by stress, which implies it might be a part of the stress-based upregulation of transposition as well [16]. This tight regulation that results in increased retrotransposition in stressful conditions, can imply

Ty's activation and mutational properties are used by the cell when mutation-based innovation is required.

In some cases, the expression of Ty in stress conditions can rescue cells and promote adaptation. Specifically, Ty was shown to mobilize subtelomeric y' elements in survivors of telomere inactivation, rescuing the cells form telomere erosion by elongating the telomeres through duplication of the Y' elements [5]. The same study, showed that some genes, in this case Y' elements, are enriched in Ty1 VLPs [5]. Moreover, it has been previously shown that when comparing yeasts with varying copy numbers of Ty1, there is selection against strains that do not have any copies of Ty, implying they are fitter [17]. Chromosomal rearrangements induced by Ty were shown to drive adaptation to High copper concentrations [18].

We propose another method by which Ty1 can help promote adaptation during stress – via the RT of genes. We hypothesized that Ty1 has the potential to facilitate Lamarckian-like evolution by RT of genes during stress conditions, as stresses activate specific genes, they are more likely to have higher transcription and, therefore, a higher chance to enter the Ty virus-like particle (VLP), as most RNAs were thought to enter the VLP in proportion to their abundance [5]. This can facilitate transcription coupled mutagenesis in genes specific to stress. As some genes can be especially enriched in Ty VLPs [5], RT of genes can be a mutagenic process which selectively affects different genes in general, enabling a variable mutation rate throughout the genome and throughout growth conditions.

It was shown that VLPs could be extracted via centrifugation and separation on a sucrose step gradient [19]. We have established the use of this protocol in the lab as well [20]. We see that layers in the gradient with a sucrose content of ~60% contain evidence for VLPs, i.e., RT activity, Ty RNA and Gag protein while being depleted for ribosomes (Figure 2)[20]. Usage of this protocol allows us to extract VLP DNA and RNA from the VLP layer - allowing us to peek into the VLP and see which sequences are reverse transcribed within it.

Using our ability to look into the VLP, we would like to understand the extent of the phenomenon of Ty RT of genes. We would like to characterize the genes which enter

Ty VLPs and undergo RT within them. We would like to understand what characterizes the RNAs which undergo RT by Ty. We would also like to know how entrance to the VLP and subsequent RT impacts gene evolution and organism/genome evolution.

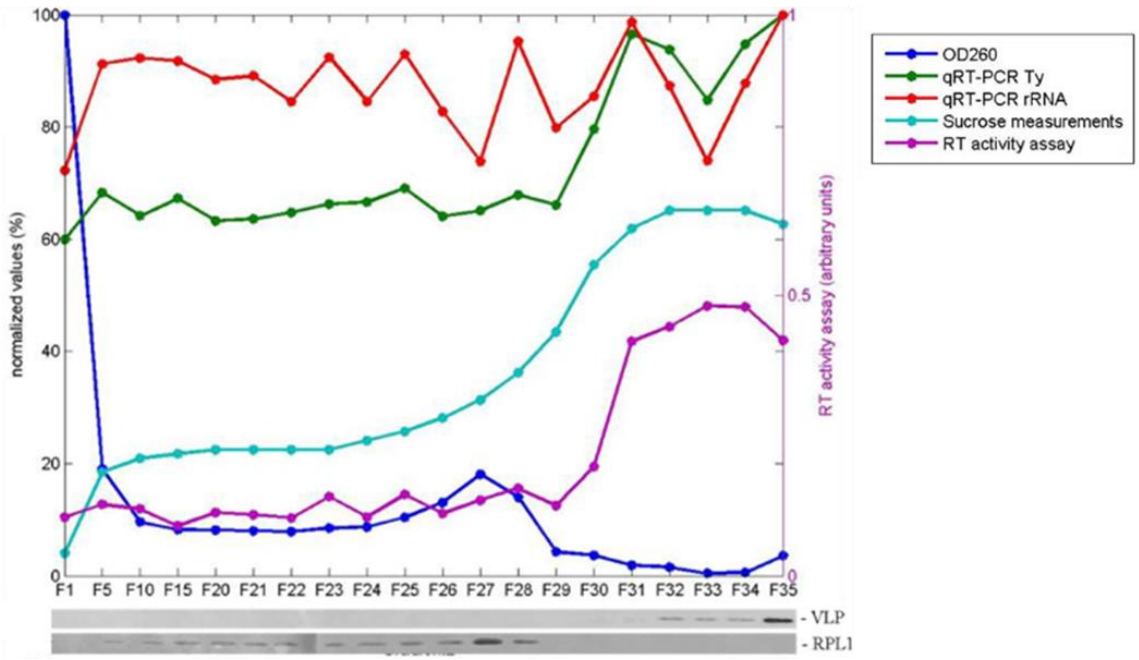


Figure 2 VLP extraction using a sucrose step gradients from Kaminski, 2015 [20]. Several different VLP and ribosome detecting assays were used on the different layers and their results were normalized. VLP detecting assays include RT activity and qRT-PCR Ty. Ribosomal detecting assays include OD260 and qRT-PCR rRNA. Sucrose content was measured for each gradient layer as well. The bottom panel shows western blot analyses of VLP and ribosomal protein (RPL1) from the different layers

.

Goals

Characterize VLP enriched and depleted genes in the RNA and DNA levels.

Understand which gene characteristics can determine VLP enrichment/depletion at the RNA and DNA levels.

Characterize the impact VLP entry in the RNA and DNA levels can have on a given gene's evolution and on genome evolution in general.

Materials & Methods

Sequencing experiment

Data was taken from an experiment performed in the lab by Yonat Gurvich, a former postdoc in the lab. VLPs were separated from the cytosol using a sucrose gradient and a fraction containing VLPs was extracted. From the VLP containing fractions RNA and DNA were extracted. RNA was extracted from the whole cell as well. RNA and DNA samples were then sequenced.

Two sequencing datasets were generated from Yonat's experiment – Illumina RNA and DNA-seq and MARS-seq of RNA [21]. The illumina RNA and DNA-seq dataset included VLP-RNA, VLP-DNA and whole cell RNA (called here TOTAL) samples. The MARS-seq dataset included VLP-RNA and TOTAL-RNA samples.

The samples include two strains: a VLP-increased strain called Ty-Plus (Plus) and a control strain called Ty-Minus (Minus). The strains were created from the lab strain RM11-1A that does not contain any copies of Ty1. The strain was evolved to reduce clumping. Minus is the unaltered strain (after the evolution) and Plus was created by adding an inducible copy of Ty1 to the Minus, downstream of the GAL promoter. The different conditions are termed as "N" and "YP". Samples termed "N" were subjected to stress – they were grown in a low nitrogen medium and were sampled during the stationary phase. Samples termed as "YP" were the control for stress – grown on a rich YPGal medium and were sampled during the exponential phase. [Unpublished methods].

RNA and DNA sequence data alignment:

RNA and DNA Seq: The reads from RNA and DNA-seq dataset were aligned using the STAR program, which was used to count reads in genes and exon-exon junctions as well [22]. The genome used for the alignment is the S288C strain reference genome version R64-3-1 which was taken from the Saccharomyces Genome Database (SGD) [23]. The annotation file for the alignment was also based on the S288C reference genome version R64-3-1 from the SGD[23].

Counting reads mapped to transposons is nontrivial, as the alignment software is generally not designed to map many short reads to many repeats of highly similar elements [24]. As we would like to properly count the reads mapped to the many similar copies of sequences in the different Ty families, we masked all retroelements from the genome similarly to the method used in [25] to count Ty reads. Ty genes and LTRs were masked using the repeatmasker software [26]. Based on the SGD annotation [23], sequences annotated as a transposable element, retrotransposon gene or LTR were compiled to a single fasta file. This file was then used as a base for repeatmasker to mask sequences across the genome. A representative gene from each Ty family was then added to the genome in an artificial "Ty chromosome". This single repeat copy thus serves as the unique address for the mapping of all repeat-originated reads.

The GeTMM [27] read count normalization algorithm was used for normalizing cDNA and RNA read counts for graphing purposes.

MARS-Seq: The MARS-Seq dataset alignment was performed using a pipeline based on the UTAP pipeline for MARS-Seq alignment [28]. Unique molecular identifiers (UMIs) were first extracted from the sequence files using the UMI extraction tool from the UMI-tools suite [29]. The annotation gtf file used for alignment was based on SGD's genome and 3' UTR annotation [23]. For each gene with an annotated 3' UTR, the end was defined as the end of the longest annotated UTR of this gene. Otherwise, if there was no annotated UTR, the end coordinate from the original SGD annotation was used. A new annotation file was generated which included only the last 1000 bases of each gene from the previously defined end coordinates. The reads were then aligned using STAR [22]. The aligned sequences were then "deduplicated" based on UMIs using UMI-tools' deduplication tool [29]. After deduplication, read counts in the genetic features were counted using featureCounts [30]. Read counts were then corrected for UMI clashing using a dedicated python script from the UTAP pipeline [28].

Differential gene expression analysis and VLP enriched and depleted RNA characteristics

Differentially expressed gene analysis of VLP RNAs: The DESeq2 package in R was used for differential gene expression analyses which were used to define VLP

enriched and VLP depleted genes [31]. PCA was calculated using the count data after applying a regularized log transformation to it using the “rlog” function in the DESeq2 package [31]. Differentially expressed genes were determined via directly comparing the two given groups in DESeq except for the VLP/TOTAL fraction comparison between the Plus and Minus strains which was done using a likelihood ratio test (LRT) comparison. Estimated log fold changes were shrunk for visualization purposes using the ‘ashr’ shrinkage method in the DESeq2 package [31], [32]. VLP-enriched and VLP-depleted RNAs were determined based on the comparison of VLP to TOTAL in the Plus strain samples.

GO category analysis: GO (Gene Ontology) category enrichment analysis was performed using the goseq package [33] on enriched genes found in the MARS-seq dataset. The genes and GO categories were hierarchically clustered using the complete-linkage method and the Euclidean distance metric. Distances between the binary vectors of enriched genes within each GO category were used for the clustering.

dN and dS calculation: dN and dS – rates of non-synonymous and synonymous substitution were calculated for each gene relative to the phylogenetic tree of other 1,011 *S. cerevisiae* strains [34]. Strain CEG from the Taiwanese outgroup clade was used as a proxy for the common ancestor and dN dS of all genes were calculated for all strains versus this ancestor not including the Taiwanese strains[34]. Homologs of the each of the ancestor strain’s genes found in the reference (S288C) genome for each of the remaining 1,008 strains were taken from the database. Genes were then aligned using clustalOmega [35], [36]. Pairs of genes and strains were then filtered for viability of dN/dS calculation - only genes that had a copy in the ancestor-proxy could be used and only pairs that didn’t have an insertion or deletion from the reference of ancestor-proxy strain were taken. Values of dN and dS were then calculated using the yn00 program in the PAML package [37] using the biopython bioPAML module [38]. Comparison of dN and dS of all genes to the VLP depleted and enriched gene sets were done using a one-sided Wilcoxon test hypothesizing VLP enriched genes have a higher dN and dS and that VLP depleted genes have a lower dN and dS.

RNA localization and set overlaps: RNA localization data was taken from a dataset which used sequencing of RNAs in cells that express Endoplasmatic reticulum (ER) and mitochondria anchored marker enzymes [39]. By counting the markers added RNA reads, the authors put each gene in tiers based on how localized or depleted they are from a given organelle. A union of all tiers was used to define the RNA localization enriched and depleted gene sets. The R package GeneOverlap was used to calculate and visualize set overlap p-value using Fisher's exact test. P-body RNA localization data was taken from a dataset which used crosslinking followed by affinity purification[40]. The authors exposed cells to different stress conditions. Following the application of stress, cells were enriched for using tagged P-body factors Dcp2p, Scd6p. Technically, mRNAs that can be crosslinked to either Dcp2p, Scd6p or their interaction partners during stress were enriched. The union of RNAs enriched in P-bodies in the 3 different stress conditions was used for the list of P-body localized RNAs [40].

RNA half-life: I used 3 RNA half-life datasets from *S. Cerevisiae* [41]–[43]. All 3 datasets are based on experiments using metabolic labeling by 4s-thiouracil. For each of our VLP-enriched and VLP-depleted gene sets, we used a two-sided Wilcoxon test to compare the given set with all the genes not included within the set.

cDNA measurement in VLP DNA samples

Genomic DNA control: Genomic DNA sequencing data which was used as control was taken from the ancestor strain of a project previously done in the lab, published earlier [44]. Sequencing data was mapped to the S288C genome using STAR, which was used to count splice junction containing reads as well [22].

Exon-intron ratio: Calculation of exon-intron coverage ratio was done on the average of read counts on intervals of equal length within introns and exons. For each intron-containing gene, the segment length l was defined as the length of the shortest of the introns and exons. All introns of the gene were then divided to $n - l$ overlapping intervals of length l where n is the length of the given intron or exon. Reads within the segments were then counted using featureCounts [30] and a mean of exonic reads and intronic reads was taken for each gene. The exonic-intronic coverage ratio was then calculated by dividing the exonic mean with the intronic mean of each gene with an

added pseudocount of 1 to the numerator and denominator. To compare the genomic DNA exon-intron ratio to the VLP DNA exon-intron ratio, the ranks of exon-intron ratios from all samples were taken. The ranks of the genomic DNA were then compared to the ranks of the VLP DNA samples using a two-sided Wilcoxon test.

In-gene and out-gene coverages and comparison: Comparison of coverage within and outside of the gene was done by calculating in-gene coverage - taking the sum of read coverages from segments of 200bp directly upstream and downstream of the center of each gene. The center was defined as the center of the longest exon in genes with multiple exons. The out-gene coverage was calculated as the sum of read coverages of 200 bp segments directly upstream of the start of the 5' UTR of each gene and directly downstream of the 3' UTR of each gene. In case a gene didn't have an annotated UTR, the end or start of the coding region was used instead. UTR annotation was taken from the SGD [23]. The in-gene/out-gene ratio was calculated by dividing the in-gene coverage with the out-gene coverage with an added pseudocount of 1 to the numerator and denominator.

Evolution experiment

Strains and media: Raffinose medium was prepared from 2% Raffinose, 6.7 g/L Yeast nitrogen base without amino acids and with ammonium sulfate, and amino acids. Low Nitrogen- Galactose medium was prepared from 2% Galactose, 0.67% Yeast Nitrogen Base without Amino Acids and Ammonium Sulfate (YNB), 5 µM of ammonium sulfate and amino acids: His, Met, Leu and URA (0.15 gr mix in ratios 1:1:4:1 gr; respectively). Media were made according to the protocol used in the sequencing experiment [Unpublished methods].

Dilution frequency determination: An initial growth experiment was performed on the unevolved strains to determine growth rate and assess the amount of time required for the strains to reach saturations. Two repeats of starters of the unevolved strains were grown for 72 hours, reaching a plateau after ~35 hours in both repeats in both strains. Following this experiment, we've decided to dilute the evolution once every 2 days.

Serial dilution experiment: Ty-Plus and Ty-minus yeast strains were evolved using serial dilution in 24-well plates. Each plate contained 3 repeats of the same strain and one blank repeat. Strains were evolved in a low nitrogen YPGal medium to induce the Gal promoter. The Frozen strains were thawed on a YP agar plate. Starters were taken from single colonies on the plates and grown on an SC Raffinose medium. Starters were then serially diluted with a dilution frequency of 48 hours in plates containing Low Nitrogen-Galactose medium. The plates were kept in a shaking incubator set to 30ºC for a total of 42 dilution cycles, which amount to 84 days and 294 generations.

Results

VLP enriched and depleted RNAs

RNA of VLP and whole cell (referred to as "TOTAL" herein) in different conditions was extracted and sequenced in an experiment previously performed in the lab. RNA was extracted from 2 different strains (see methods), a high TY strain with an inducible copy of a Ty1 construct (called "Ty-Plus" or "Plus" herein) and an otherwise identical low Ty strain (called "Ty-Minus" or "Minus" herein). Both strains were tested in two different

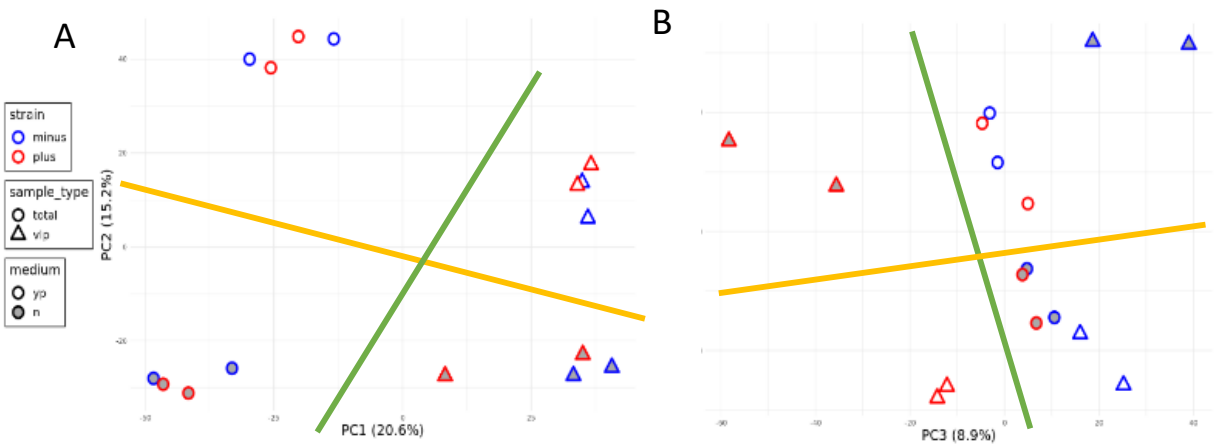
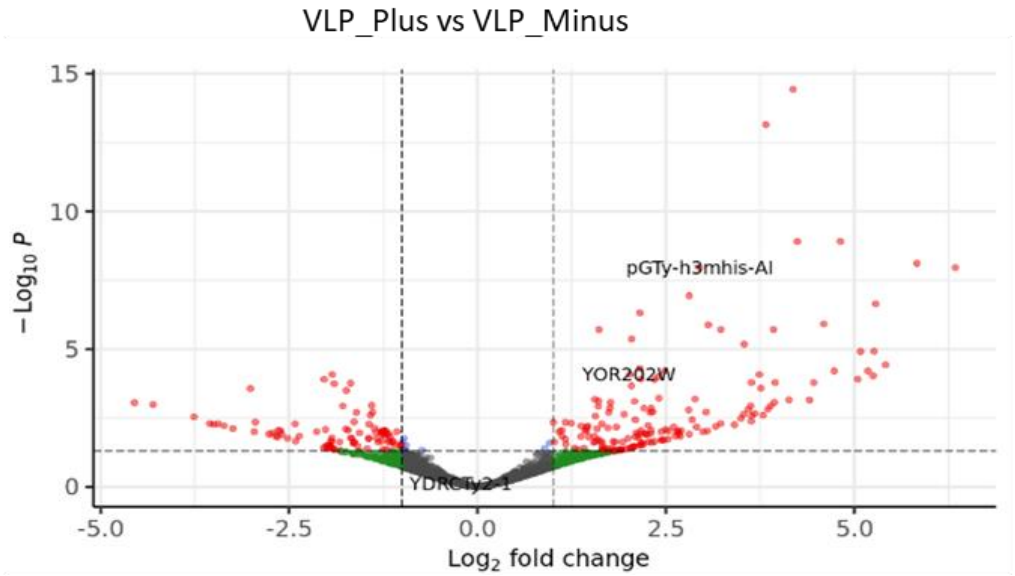


Figure 3: Principal component analysis of RNA- seq data from the VLP fraction and the whole cell. Different colors, shapes and fill colors represent the different strains (VLP plus and minus) different sample types (VLP fraction and whole cell) and different growth conditions (n – stationary on low nitrogen media; yp – exponential on YPGal medium). A. principal components 1,2. Green line illustrates the separation of VLP from TOTAL and orange line illustrates the separation of N and YP. B. principial components 3,4. Green line illustrates the separation of VLP plus and VLP minus while the orange line illustrates separation of N and YP which is opposite in VLP and TOTAL.

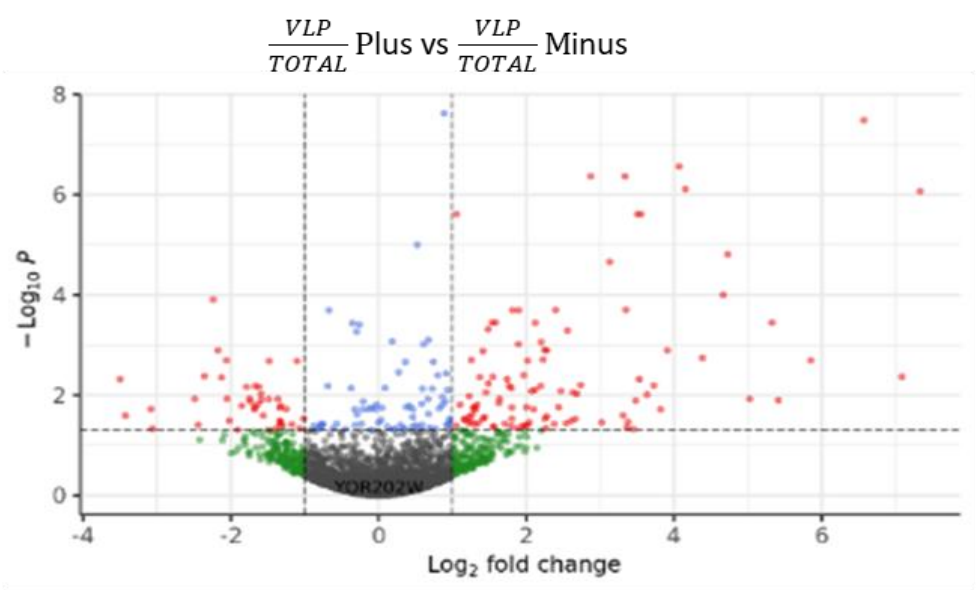
conditions – a YPGal medium in exponential growth phase (called the “YP condition”), and a stress condition in which cells were grown on a low nitrogen medium in stationary phase (called “N-starvation”, or “N” herein). The RNA-Seq reads were then aligned to the *S. cerevisiae* strain S288C genome.

To visualize the differences between samples, we’ve performed dimensionality reduction of the mRNA samples from the VLP fraction (VLP) and the whole cell (TOTAL) using Principal component analysis (PCA) (Figure 3). We see that the four first principal components (PCs) of mRNA PCA separate the samples according to their groupings. PC1 separates VLP RNA from cytosol RNA, while PC2 separates the YP grown samples from the N-starvation samples. PC3 separates the VLP Ty-Plus samples from the VLP Ty-Minus samples, placing the TOTAL samples between them. PC4 separates N-starvation and YP in VLP and TOTAL but groups N VLP with YP TOTAL and N TOTAL with YP condition VLP unlike PC 2 which groups all N-starvation

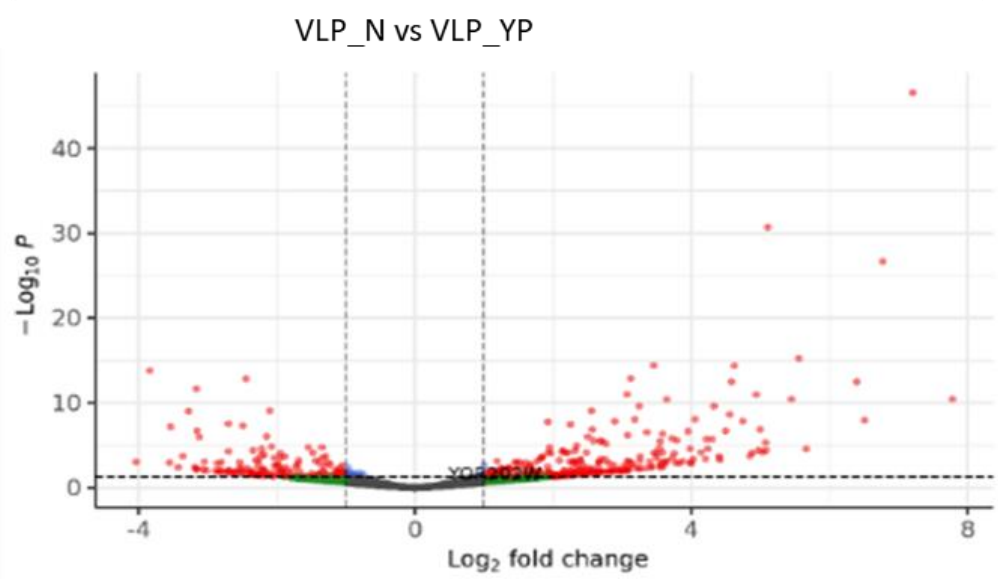
A


VLP_Plus vs VLP_Minus
15
10
5
0
$-\text{Log}_{10}\,P$
pGTy-h3mhis-AI
YOR202W
YDR033W-1
-5.0
-2.5
0.0
2.5
5.0
Log_2 fold change

B


$\frac{VLP}{TOTAL}$ Plus vs $\frac{VLP}{TOTAL}$ Minus
8
6
4
2
0
$-\text{Log}_{10}\,P$
YOR202W
-4
-2
0
2
4
6
Log_2 fold change

C


VLP_N vs VLP_YP
40
30
20
10
0
$-\text{Log}_{10}\,P$
-4
0
4
8
Log_2 fold change

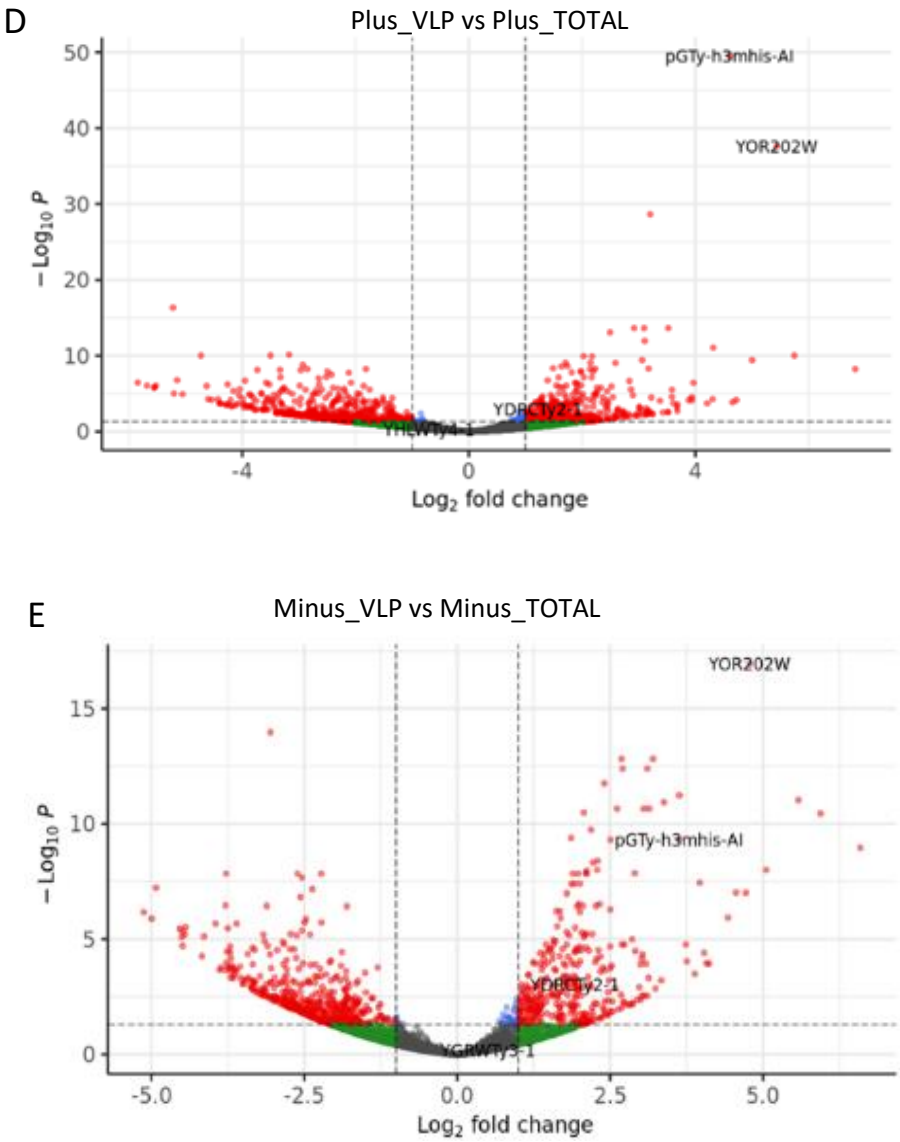


Figure 4: Differential expression analysis of RNA samples. Shown are volcano plots in which the y-axis is the enrichment p-value and the x-axis is the log2 fold change estimated by DESeq2. Labels on the plot represent the locations of The Ty1 construct (pGTy-h3mhis-AI), His3 (YOR202W) and other Ty elements in the case they were assigned a log-fold change by DESEQ. Enriched genes are determined by an adjusted p-value threshold of 0.05. A. (VLP plus) / (VLP minus). 165 genes enriched in PLUS and 97 genes enriched in MINUS. B. (VLP/ TOTAL plus) / (VLP/TOTAL minus). 144 genes enriched in PLUS and 69 genes enriched in MINUS. C. (VLP N) / (VLP YP). 254 genes enriched in n and 142 genes enriched in YP. D. (plus VLP) / (plus TOTAL). 374 genes enriched in VLP and 617 genes enriched in TOTAL E. (minus VLP) / (minus TOTAL). 369 genes *enriched in VLP and 534 genes enriched in TOTAL.*

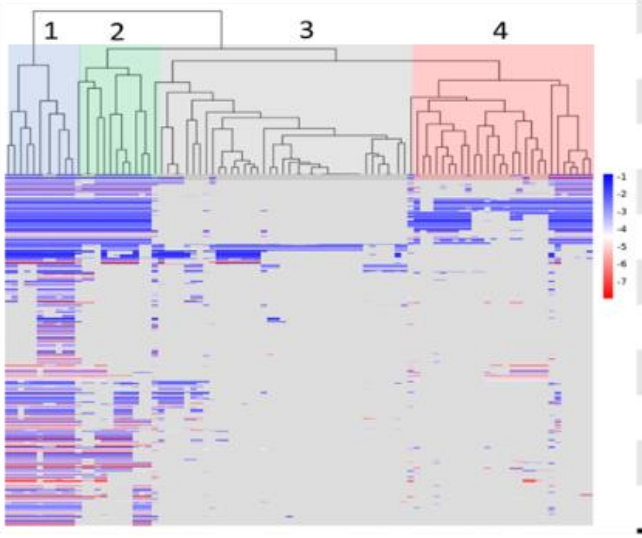
together and all YP condition together. Notably, the VLP samples are more well-

separated along this axis.

We wanted to examine these differences in the resolution of specific genes, so we performed differential expression analysis using the DESeq2 package [31] to characterize VLP enriched and VLP depleted genes and to examine the differences in these genes in different conditions. We compared between different sets of samples (Figure 4). We first examined, as positive controls, the two genes whose mRNA is supposed to be enriched in the VLP, especially in the Ty-Plus strain. The first of these genes is Ty1, and the second is the HIS3 gene, as HIS3 is part of the artificial Ty1 construct within the Ty-Plus strain's genome. Indeed, we find that the mRNAs of the Ty1 and HIS3 genes are enriched in Plus VLP relative to Minus VLP and in VLP vs TOTAL in both Plus and Minus strains. Interestingly, Ty2, which is a VLP forming retrotransposon closely related to Ty1 but distinct from it [1], is enriched in VLP relative to TOTAL in both Plus and Minus strains, implying the VLP fraction of the sucrose gradient includes Ty2 VLPs as well and that both Ty-Plus and Ty-Minus strain express Values of the assays are normalized. Ty2. Surprisingly, no Ty gene, nor His3 are differentially expressed in the comparison of the VLP/TOTAL fraction in Ty-plus vs the VLP/TOTAL fraction in Ty-minus. The strongest signal of differential expression we see, when considering the number of differentially expressed genes in both sides are when comparing between VLP and TOTAL samples in the Ty-Plus strain, with 374 genes enriched in VLP and 617 genes enriched in TOTAL (VLP depleted). We see another strong signal when comparing VLP to TOTAL in the minus strain with 369 genes enriched in VLP and 534 genes enriched in TOTAL (VLP depleted). Strangely, the Ty1 construct is highly enriched in VLP vs TOTAL in the minus-strain and HIS3 is the most

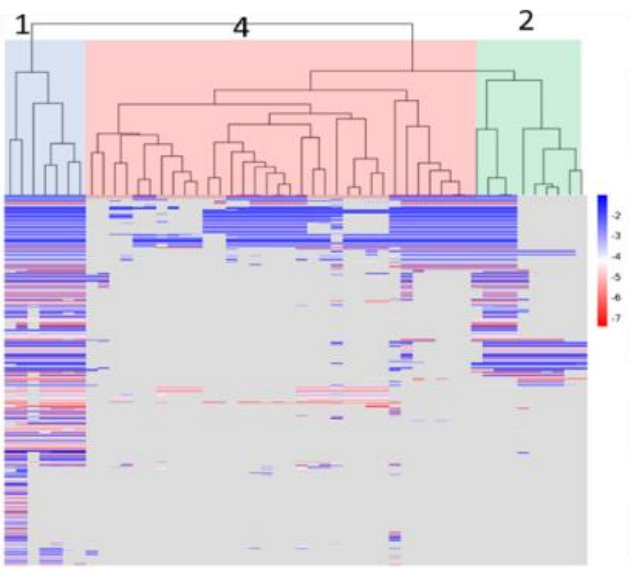
highly enriched gene in that comparison, even though the VLP-minus strain doesn't contain copies of Ty1 or the Ty1-HIS3 construct within its genome.

A



Enriched GO category	P-value
Intracellular	7.1e-11
Cellular metabolic process	4.1e-10
Metabolic process	8.7e-10
Organonitrogen compound metabolic process	2.6e-08
Biosynthetic process	5.3e-08
Organic substance biosynthetic process	1.6e-07
ATP metabolic process	1.5e-09
Mitochondrial protein complex	4.7e-07
Mitochondrion	1.5e-06
Cytosolic ribosome	3.0e-20
Ribosome	4.6e-20
Ribosomal subunit	9.6e-20

B



Enriched GO category	P-value
Organonitrogen compound metabolic process	2.4e-09
Cytoplasm	7.5e-07
Cellular metabolic process	3.4e-05
Organic substance biosynthetic process	1.1e-05
Cellular biosynthetic process	1.2e-05
Biosynthetic process	1.9e-05
Cytosolic ribosome	2.6e-26
Ribosome	1.9e-23
Cytoplasmic translation	2.1e-23

Figure 5: enriched GO categories in VLP depleted genes. Each subplot contains a binary heatmap of enriched GO categories in genes which are depleted in VLP plus relative to TOTAL plus. The value in each cell represents the log-fold depletion and greyed out cells are for genes that are not in each go category. Depleted genes that are not in any enriched go category are excluded. Each subplot contains a representative GO category table as well, containing the 3 most significantly enriched GO categories of each main cluster and their enrichment p-value. A. Enriched GO categories in the VLP Plus depleted genes. Enriched GO categories in the VLP Minus depleted genes.

Now, as we have lists of VLP enriched and depleted genes, we would like to know if genes that are related to certain pathways or cellular components are more or less likely to be encapsulated in VLPs at the mRNA level. Using the goseq [33] package, we performed GO category analysis on the VLP enriched and VLP depleted genes in the Plus and Minus strains. We found no enriched GO categories in the VLP enriched gene sets in the Plus strain and one enriched go category in the VLP enriched genes in the minus strain - GO:0005515 (protein binding). We found many enriched GO categories in the VLP depleted gene sets in Plus and Minus as shown in Figure 5 and Supplementary figure 1- 50 enriched GO categories were found in the minus strain and 92 enriched GO categories were found in the plus strain. 45 GO categories overlap between the sets of enriched GO categories in VLP depleted genes in the plus and minus strains. For each of these GO category sets, we clustered the GO categories based on the depleted genes that correspond to each category.

In Figure 5A and Supplementary figure 1 we see that the GO categories enriched in the VLP depleted genes in the Ty-Plus strain, when clustered according to the VLP depleted genes within them, are broadly divided into 4 categories which vary by function. Cluster 1 is a generalist cluster containing broad categories such as “organelle”, “cellular process”, “intracellular”, and “cellular anatomical entity”. Most VLP-depleted genes in the Ty-Plus strain include genes that correspond to GO categories in cluster 1 and it includes the most highly differentially expressed genes (shown in Figure 5A). Cluster 2 is another generalist cluster which mostly corresponds to different types of biosynthetic processes and to broad categories e.g., “gene expression” and most depleted genes correspond to some GO categories in this cluster. Cluster 3 corresponds to different metabolic processes and the mitochondria and processes related to it and contains terms such as “mitochondrion”, “ATP biosynthetic process” and “respiratory chain process”. Notably, a small subset of depleted genes corresponds to categories within this cluster. Cluster 4 contains processes and structural elements corresponding to the cytosolic ribosome and includes go terms such as “ribosome”, “rRNA processing” and “cytoplasmatic translation”, “cytosol” and “translation”. A substantial minority of depleted genes corresponds to categories within this cluster.

In, Figure 5B and Supplementary figure 1 we see that the GO categories enriched in the VLP depleted genes in the Ty-Minus strain are clustered in similar groupings to the enriched categories in the VLP depleted genes in the Ty-Minus strain, containing clusters analogous to the previously described clusters, 1, 2 and 4. Most genes within the enriched GO categories correspond to categories in cluster 1, with sizeable minorities corresponding to categories in clusters 2 and 4. Notably, the main difference between the enriched categories in Ty-Plus VLP depleted RNAs compared to the enriched categories in the Ty-Minus VLP depleted RNAs is that the Ty-Minus VLP depleted RNAs are not significantly enriched in most GO categories in the mitochondria related cluster 3 – or any mitochondria related categories.

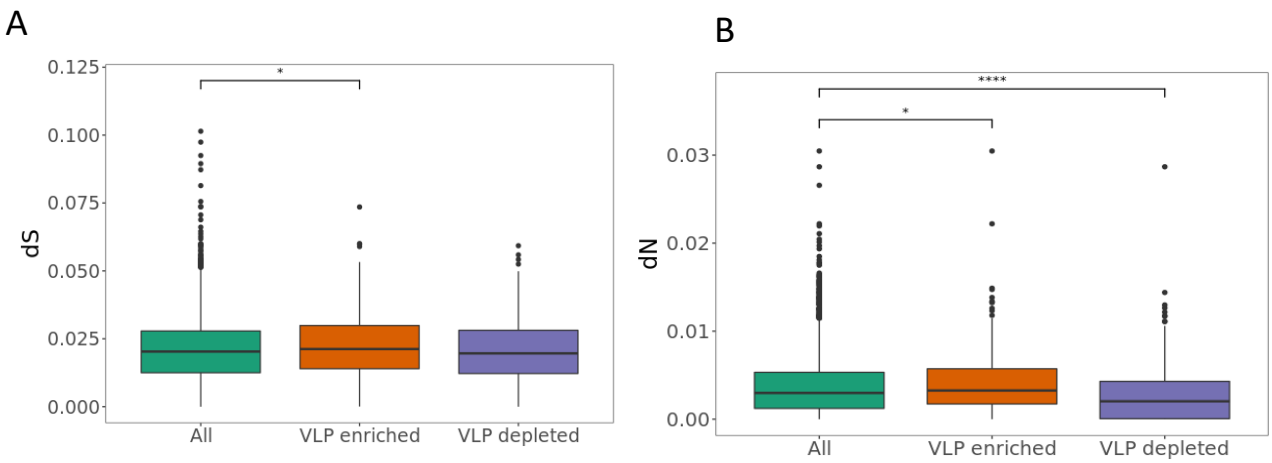


Figure 6: dN and dS of VLP enriched and VLP depleted genes. Each line represents a comparison done using a one-sided Wilcoxon test comparing the given gene list with the set of all genes. *: p<0.05 ****:p<0.0001. A. dS of all genes, VLP enriched and VLP depleted genes. B. dS of all genes, VLP enriched and VLP depleted genes.

Having established a list of VLP enriched and VLP depleted genes, we would like to know whether entrance to the VLP is somehow correlated with a change in the rate of evolution of genes. One of the most important drivers of evolution is mutation. Here we will examine the possibility of mutagenic effect of genes' reverse transcription (RT). To do so we examine whether VLP inclusion of RNA is correlated with mutation rate, implying a possible correlation with a difference in evolution between VLP enriched and depleted genes. We calculated dN (rate of nonsynonymous mutation) and dS (rate of synonymous mutation) of genes in the *S. Cerevisiae* genome across a comprehensive

database of *S. Cerevisiae* strains [34]. A seen in Figure 6A, we find that VLP enriched genes have significantly higher dS while VLP depleted genes don't significantly vary from the rest of the genes. In Figure 6B we see that VLP enriched genes have significantly higher dN than the rest of the genes and VLP depleted genes have much significantly lower dN than the rest of the genes. In summary, we see that VLP enriched genes have a higher mutation rate while VLP depleted genes tend to be more conserved.

There is evidence indicating that VLP formation is localized to specific cellular foci called T-bodies or retrosomes that are related to P-bodies [45] and that its nucleation by RNA is localized in the endoplasmatic reticulum (ER) [46]. As VLPs are formed and nucleated in specific loci, we hypothesized that cellular RNAs that are localized within these areas might be more likely to be encapsulated in VLPs. We tested this hypothesis by examining the RNA localization of VLP enriched and depleted genes. We took databases of RNA localization from two different experiments. One of which used anchored tagging enzymes followed by RNA-seq[47]. The cellular locales measured in this experiment include the Endoplasmatic reticulum (ER) and the mitochondrial outer

A

	VLP enriched genes(374)	VLP depleted genes (617)
ER enriched (1096)	91, 2.13e-06	73, N.S.
ER depleted (1165)	37, N.S.	286, 1.26e-76
Mitocohondria enriched (590)	32, N.S.	60, N.S.
Mitocohondria depleted (464)	29, N.S.	107, 8.64e-23
P-body enriched (1544)	109, 2.78e-04	143, N.S.

B

	VLP enriched genes (374)	VLP depleted genes (617)
Intron containing genes (348)	5, N.S.	74, 7.41e-14
Ribosomal genes (158)	0, N.S.	64, 2.76e-28
Intron containing ribosomal proteins (90)	0, N.S.	47, 5.68e-27
Ribosomal proteins without introns (68)	0, N.S.	17, 4.58e-05
Non-ribosomal intron containing genes (258)	5, N.S.	27, N.S.

Figure 7: Set overlaps of VLP enriched and VLP depleted genes with different gene sets. Each outer cell contains a gene set followed by its size in brackets. Each inner cell in the table contains the size of the overlap followed by a significant p-value for set overlap calculated using Fisher's exact test. Any nonsignificant p-value is shown as N.S. A. Set overlap of VLP enriched and VLP depleted genes with sets genes with specific cellular RNA localization. B. Set overlaps of VLP depleted genes with ribosomal protein genes and genes containing introns.

membrane. This experiment provides a dataset of sub-cellular organelle localization of mRNAs in yeast. The other experiment tested for P-body RNA localization. The authors used crosslinking followed by affinity maturation [40]. They defined P-body associated RNAs as RNAs that can be crosslinked to the P-body factors Dcp2p, Scd6p or their binding partners during stress [40]. We examined if the mRNAs in each organelle are either enriched or depleted from the VLP. Figure 7A shows the enrichment or depletion of VLP-enriched and VLP-depleted genes with different RNA localizations. We see that VLP-depleted and VLP-enriched RNAs appear to have characteristic cellular localizations – ER localized mRNAs are significantly enriched in the VLP, as seen in both sets of VLP enriched and depleted genes. In contrast, mitochondria localized mRNAs are significantly depleted from the VLP, as seen in both sets of VLP enriched and depleted genes. P-body enriched genes appear to have significant overlap with VLP enriched genes as well.

As RT of genes is a potential mechanism of intron loss [8], we wanted to see whether VLP encapsulation of mRNAs is correlated with intron presence in genes. In Figure 7B, we see that intron containing genes are significantly enriched within the set of VLP-depleted genes. As we've seen in Figure 5 and Supplementary figure 1, ribosomal proteins (RPs) are enriched in the set of VLP depleted genes as well. RPs and intronic genes have a significant overlap [48]. Specifically, 93 of the 282 RPs (as defined by genes in "GO:0022626" - cytosolic ribosome GO category, not including the 12 rRNA genes) have introns. Therefore, we tested whether the significant overlap of VLP depleted genes with genes containing introns is due to the overlap with ribosomal proteins. We see that ribosomal proteins which contain introns and ribosomal proteins which don't contain introns significantly overlap with VLP depleted genes while non-ribosomal proteins which contain introns don't have significant overlap with VLP-depleted genes. Yet, RPs without introns show a much weaker tendency to avoid VLP encapsulation. We also note that RPs are among the most conserved proteins in the yeast genome [49], and as shown above, in Figure 6B, conserved genes tend to avoid VLP encapsulation. We thus conclude that the combination of the two factors, i.e. being an RP, and having an intron, together predispose the mRNA of a gene to avoid VLP encapsulation.

Another RNA attribute that might affect VLP inclusion is RNA half-life, as RNAs with a

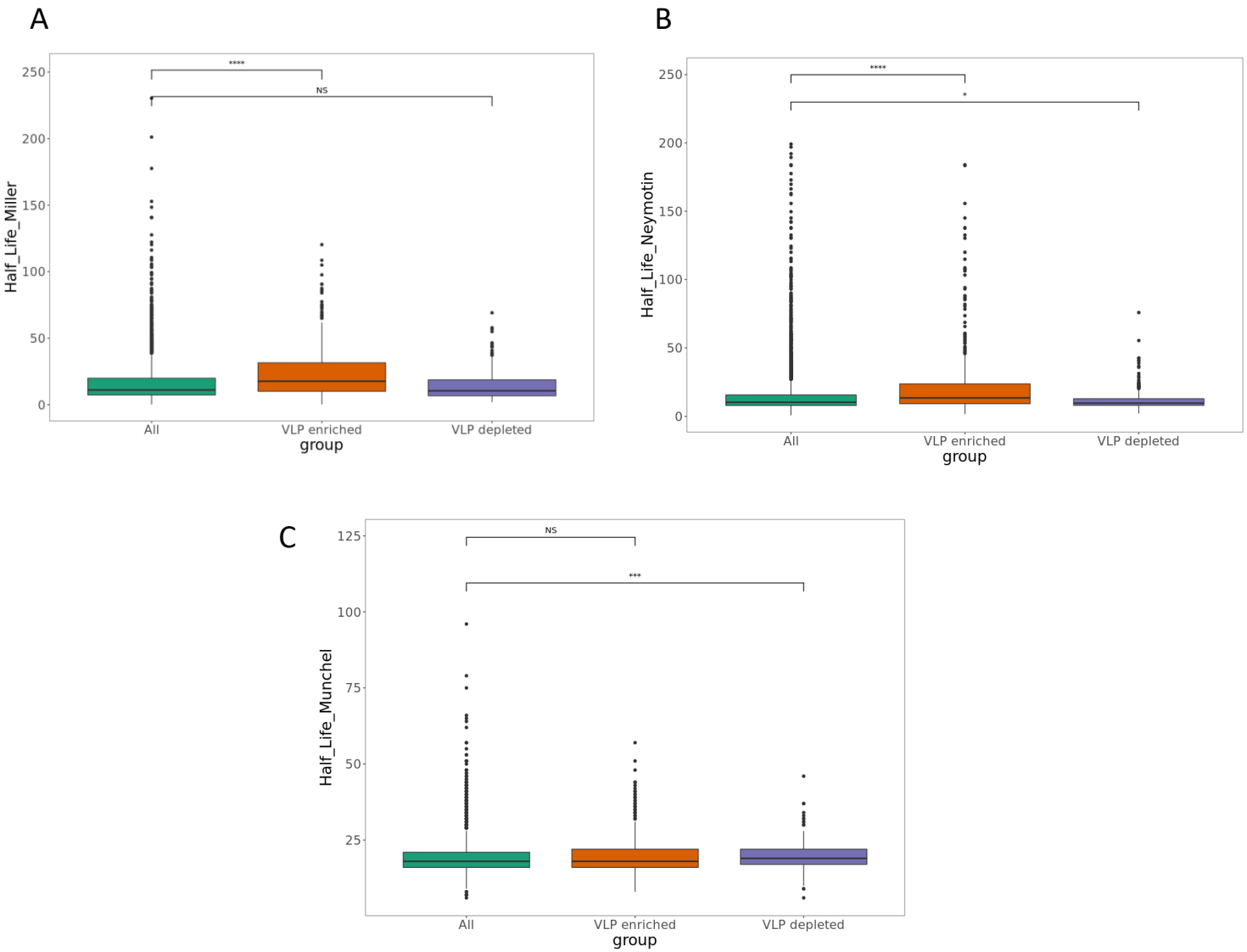


Figure 8: half-lives of VLP-enriched and VLP depleted genes. Above the boxplots are p-values from a two-sided Wilcoxon test of each subset (VLP enriched or VLP depleted) compared with all genes. A. Dataset from Miller et al. B. Dataset from Neymotin et al. C. Dataset from Munchel et al. D. Scatterplots comparing the different half-life measurements. Shown on the plots are Pearson correlations between plotted measurements and their significance.

longer cellular half-life perhaps have a higher chance to be encapsulated in the VLP within their lifetime. To test if long lived RNAs are more likely to be encapsulated we've examined 3 different RNA half-life datasets [41]–[43] which are derived from different experiments using metabolic labeling by 4s-thiouracil. We examined the half-lives of VLP depleted and VLP enriched genes and compared them to the half-lives of all other genes according to the 3 datasets.

According to the Miller and Neymotin datasets (Figure 8A and Figure 8B respectively), VLP depleted genes have significantly longer half-lives and, specifically in the Neymotin dataset, VLP enriched genes have significantly shorter half-lives. According to the Munchel dataset (Figure 8C), VLP enriched genes have a significantly longer half-life.

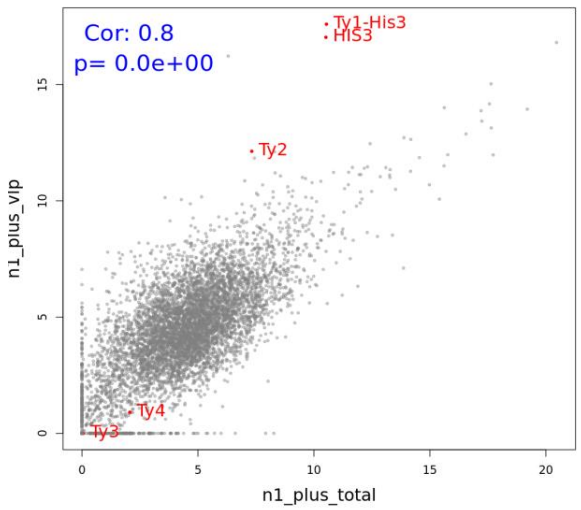
In Figure 8D we see that one of the three half-life datasets is not in agreement with the other two. The Miller and Neymotin measurements and Neymotin and Munchel measurements are significantly correlated but have significant disagreement, while the Miller and Munchel datasets are not significantly correlated. This partial agreement between the three databases explains why we obtain conflicting results concerning the half-life of VLP enriched and depleted genes.

cDNA measurement in VLP DNA

In parallel to quantifying mRNAs in the VLP, we also attempted to sequence and quantify complementary DNA (cDNA) from the VLP. This was done by direct DNA sequencing of the cellular fraction that contains the VLP. Yet, DNA extraction from VLPs has a risk of contamination from the genome's DNA, as we extract DNAs from living cells. We need to determine whether our VLP layer DNA (referred to as VLP DNA herein) data represents VLP cDNA or whether it is contaminated significantly by genomic DNA to the point it mostly represents genomic DNA. We determine this by testing for bias towards transcribed and coding regions – as we assume cDNA that is made from transcribed RNAs has such a bias relative to genomic DNA.

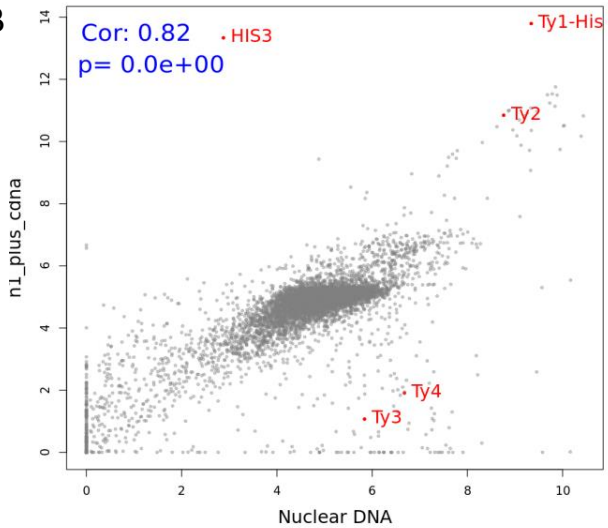
One way we tested for coding region bias is by looking at reads from genes with introns as they contain regions within them that are retained or eliminated from the mature mRNA. We note that as the VLP is present in the cytosol, and not in the nucleus, and since mature mRNA in the cytosol is devoid of introns, we obviously expect to see a high coverage of reads from exons compared to introns, in mRNA. If we see higher read coverage in exons compared to introns in the VLP fraction DNA too, that could serve as a strong indication that this DNA is actually a cDNA, having formed from mRNA through RT. An additional prediction of the hypothesis that some of our VLP-derived DNA is cDNA that was formed via RT of cellular mRNAs is that we should observe reads that span exon-exon junctions, with no intron in between. A third prediction of this hypothesis is that read counts of transcribed regions should be higher than read counts of flanking regions next to genes. While the first two of these hypotheses can only be

A



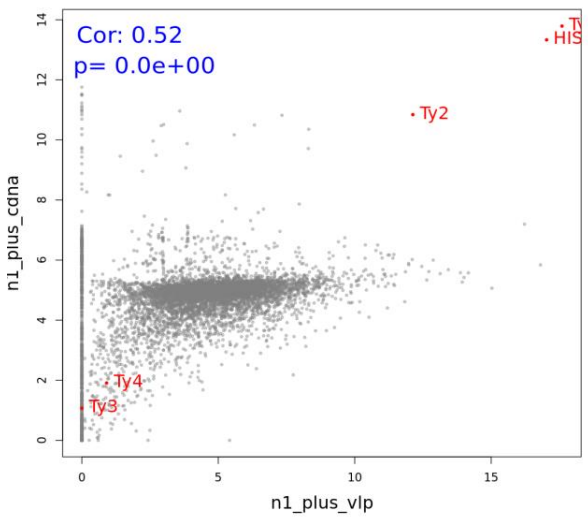
	corr	pval
yp1 minus	0.75	0.0e+00
yp2 minus	0.74	0.0e+00
yp1 plus	0.80	0.0e+00
yp2 plus	0.73	0.0e+00
n1 minus	0.83	0.0e+00
n2 minus	0.85	0.0e+00
n1 plus	0.80	0.0e+00
n2 plus	0.84	0.0e+00

B



	corr	pval
yp1 minus	0.81	0.0e+00
yp2 minus	0.79	0.0e+00
yp1 plus	0.80	0.0e+00
yp2 plus	0.80	0.0e+00
n1 minus	0.82	0.0e+00
n2 minus	0.81	0.0e+00
n1 plus	0.82	0.0e+00
n2 plus	0.81	0.0e+00

C



	corr	pval
yp1 minus	0.40	1.3e-278
yp2 minus	0.44	0.0e+00
yp1 plus	0.46	0.0e+00
yp2 plus	0.47	0.0e+00
n1 minus	0.48	0.0e+00
n2 minus	0.48	0.0e+00
n1 plus	0.52	0.0e+00
n2 plus	0.53	0.0e+00

Figure 9: Scatterplots comparing DNA and RNA reads of chosen samples (n1 plus) with tables containing Pearson correlations and p values of a given comparison in all samples. A. cellular RNA (x axis) vs VLP layer RNA (y axis). B. VLP layer RNA (x axis) vs VLP layer DNA (y axis). C. genomic DNA (x axis) vs VLP layer DNA (y axis)

examined for the 348 genes that have introns, the third could be tested with respect to all genes in the genome.

To assess cDNA content in the VLP DNA sample, we've tested for bias towards of coding regions in VLP DNA relative to a genomic DNA control, as we assume there is a strong bias towards transcribed sequences in cDNA, while genomic DNA does not have such a bias or has a significantly smaller bias. As a control, we took a sequenced sample of Genomic DNA from *S. Cerevisiae* from a different experiment done in the lab [44]. We aligned it to the same reference genome, and we compared it to VLP DNA, which we compare with the VLP RNA samples as well. In Figure 9 and Supplementary figure 2, we see that while VLP RNA is greatly correlated (Pearson's R=0.8) with TOTAL RNA (as seen in Figure 9A), it shows a lower correlation (Pearson's R=0.52) with VLP DNA (as seen in Figure 9B). VLP DNA samples, though, do seem to be much more highly correlated (Pearson's R=0.82) with a genomic DNA sample (from a different strain and experiment, seen in Figure 9C). This implies significant genomic contamination in VLP DNA. Encouragingly, we do see extremely significant enrichment of Ty1 and HIS3 in VLP DNA samples (seen in Figure 9C and Supplementary figure 2),

which imply that VLP DNA samples contain cDNA as well. HIS3 is highly enriched in all Plus samples. Ty1 is highly enriched in all N Plus samples.

For each intron containing gene, we calculate an exon-intron coverage ratio. In order to prevent read length and segment length-based biases, we calculate the coverage ratio based on an average of read counts on intervals of equal length within the introns and exons of each gene. Reads within the segments were then counted and a mean of exonic reads and intronic reads was taken for each gene. An exonic-intronic ratio was then calculated using these means. We can see that compared to the sequencing data from genomic DNA, cDNA samples have a significantly higher exon-intron coverage

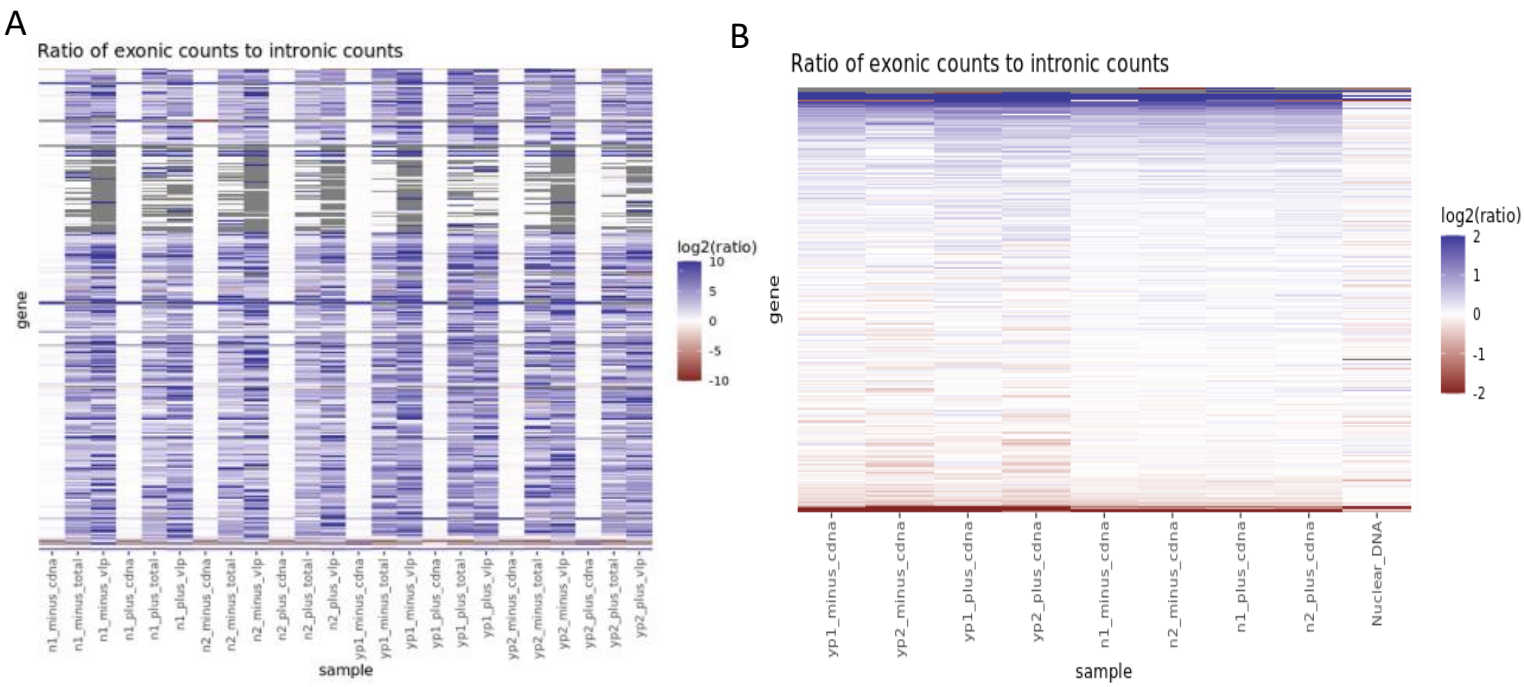
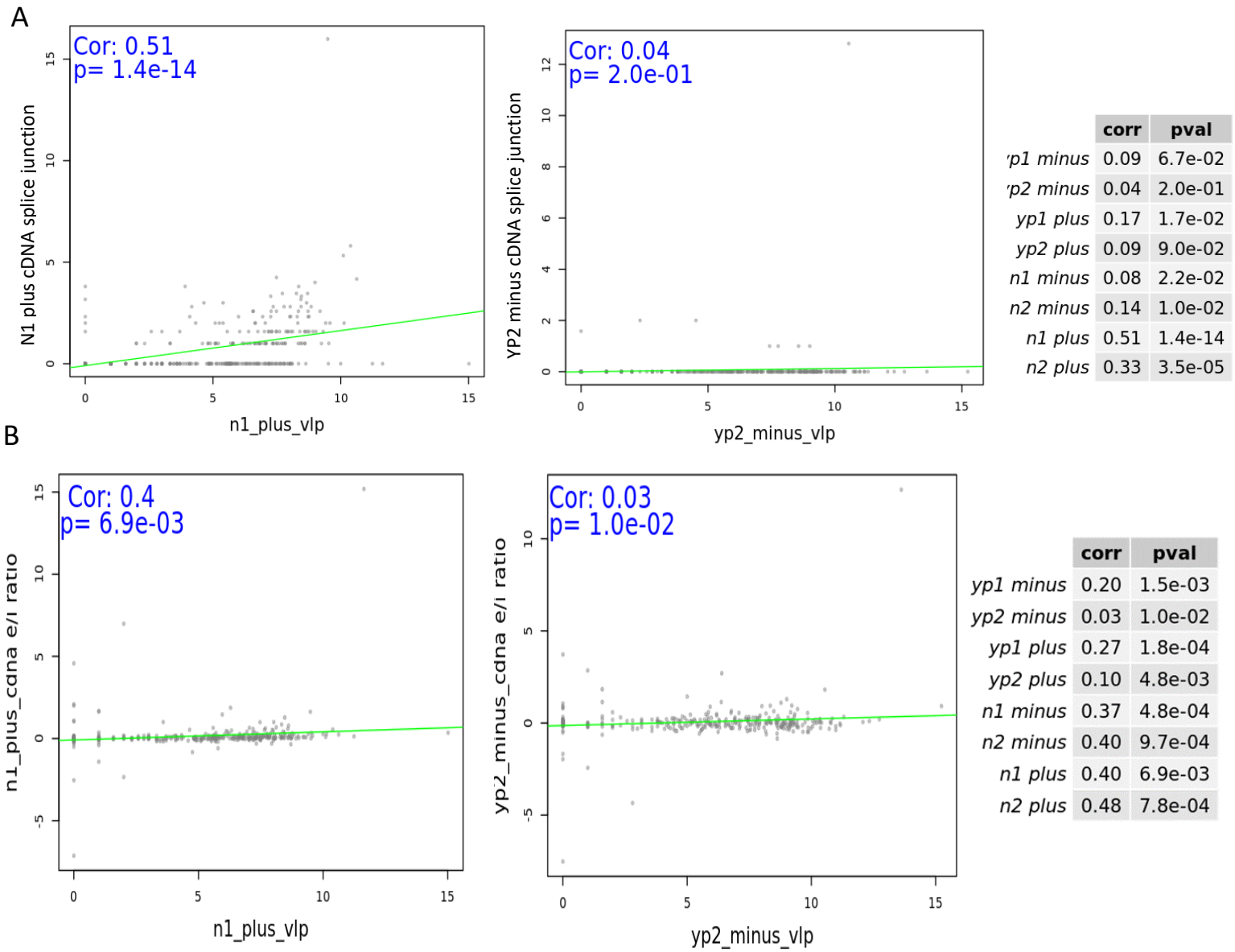


Figure 10: Ratios of exonic to intronic counts of genes containing introns in VLP DNA and RNA samples. A. Log ratios of exon/intron counts in RNA and VLP DNA samples. Values are truncated to -10 to 10. B. Log ratios of exon/intron ratio in VLP DNA and genomic DNA. Values are truncated to -2 to 2.

ratio. Nonetheless, compared to VLP and TOTAL RNA samples, exon/intron read ratio of cDNA samples is lower, Figure 10. Exon-intron ratios in VLP DNA are significantly higher than the exon-intron ratios in genomic DNA (Wilcoxon p=6.05*e^-12, using a nonparametric test that takes all genes into account). In Figure 11B, we see that



	corr	pval
'p1 minus	0.09	6.7e-02
'p2 minus	0.04	2.0e-01
yp1 plus	0.17	1.7e-02
yp2 plus	0.09	9.0e-02
n1 minus	0.08	2.2e-02
n2 minus	0.14	1.0e-02
n1 plus	0.51	1.4e-14
n2 plus	0.33	3.5e-05

	corr	pval
yp1 minus	0.20	1.5e-03
yp2 minus	0.03	1.0e-02
yp1 plus	0.27	1.8e-04
yp2 plus	0.10	4.8e-03
n1 minus	0.37	4.8e-04
n2 minus	0.40	9.7e-04
n1 plus	0.40	6.9e-03
n2 plus	0.48	7.8e-04

Figure 11: Correlation of exon/intron coverage ratios and splice junction reads in VLP DNA with VLP RNA reads. Representative sample scatter plots (N1 plus and YP2 minus) with a table containing spearman correlations and significance. Green lines are linear fit. A. VLP RNA reads (x-axis) vs VLP DNA splice junction reads (y axis). B. VLP RNA reads (x-axis) vs VLP DNA exon/intron coverage ratio (y-axis).

exon/intron ratios of genes are significantly correlated with VLP RNA counts in the given samples as well.

Our second prediction of the hypothesis that some of the VLP DNA is cDNA is that we should expect to see reads that span exon-exon splice junctions as well. We make two

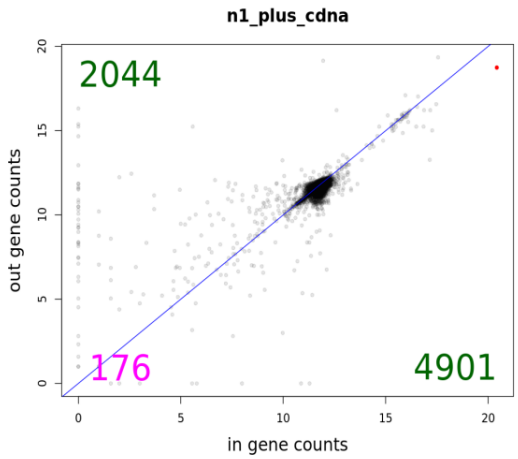
important observations. First, for 148 of the 348 intron containing genes we observed reads that span exon-exon junctions with 132 of these genes having observed reads in the “N1 Plus” sample, in contrast with 8 genes having such reads in the genomic DNA sample. This is a clear indication that some of their VLP DNA is indeed cDNA that probably is derived from RT. Secondly, In the two repeats of the Nitrogen starvation

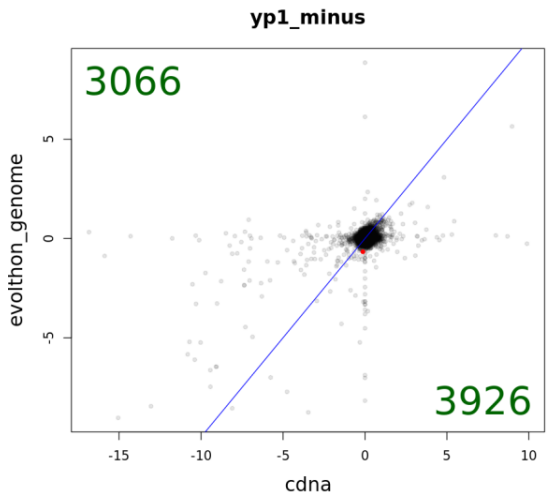
A

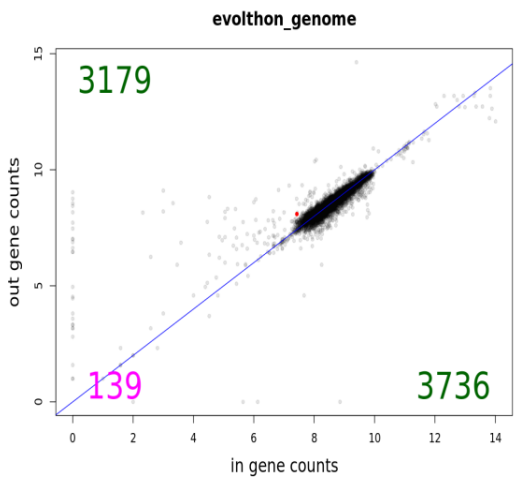
	coding < noncoding	coding > noncoding	zeros
yp1_minus_cdna	2733	4197	185
yp2_minus_cdna	3262	3657	187
yp1_plus_cdna	2618	4303	184
yp2_plus_cdna	3061	3843	186
n1_minus_cdna	2112	4837	180
n2_minus_cdna	1975	4969	182
n1_plus_cdna	2044	4901	176
n2_plus_cdna	1828	5116	180
evolthon_genome	3179	3736	139

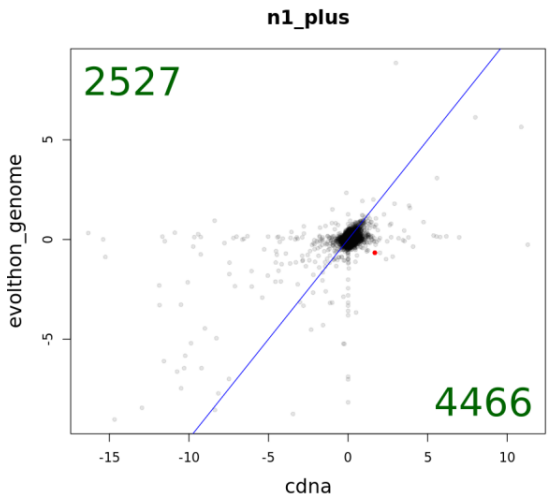
B

	VLP DNA < genomic DNA	VLP DNA > genomic DNA
yp1_minus_cdna	3066	3926
yp2_minus_cdna	3475	3515
yp1_plus_cdna	2930	4063
yp2_plus_cdna	3308	3681
n1_minus_cdna	2704	4289
n2_minus_cdna	2415	4578
n1_plus_cdna	2527	4466
n2_plus_cdna	2262	4730









C

	yp < n	yp > n
1_minus_cdna	3783	3171
2_minus_cdna	4530	2420
1_plus_cdna	3797	3160
2_plus_cdna	4429	2521

D

	minus < plus	minus > plus
yp1_cdna	3664	3283
yp2_cdna	3811	3131
n1_cdna	4007	2948
n2_cdna	4025	2927

E

	rep1 < rep2	rep1 > rep2
yp_minus_cdna	2961	3986
yp_plus_cdna	3072	3873
n_minus_cdna	4304	2648
n_plus_cdna	4303	2653

Figure 12: Comparison of coding coverages, noncoding coverages and coding/noncoding coverage ratio in VLP DNA samples. Blue line is the x=y diagonal. Numbers in the bottom-right and top-left corners count numbers below and above the diagonal, respectively. A. noncoding vs coding coverage of genes in VLP DNA samples vs a genomic DNA sample. B. coding/noncoding coverage ratio between genomic DNA (y axis) and different vs VLP DNA samples (x axis). C. coding/noncoding coverage ratio comparison between n and YP VLP DNA samples. D. coding/noncoding coverage ratio comparison between plus and minus VLP DNA samples. E. coding/noncoding coverage ratio comparison between repeat 1 and repeat 2 (y axis) VLP DNA samples.

experiment there exists a positive and significant correlation between the number of on-exon spanning reads and the extent of mRNA expression of the gene in the VLP (seen in Figure 11A).

We then moved to our third prediction, that is not restricted only to the intron containing genes, but that rather applies to all genes in the genome, that if some VLP DNA is cDNA then we should observe high read counts in the transcribed regions of a gene compared to its flanking vicinity. For that we analyzed the difference of read coverage within genes vs read coverage of sequences directly adjacent to the genes but that are not included within their transcribed sequences. We did this by examining the coverage of segments in gene centers and comparing them with equally sized segments immediately 3' and 5' from the 3' UTR and 5' UTR, respectively. For the in-gene

coverage, we took the sum of read coverage from segments of 200 bp directly upstream and downstream of the center of each gene. The out-gene coverage was calculated as the sum of read coverages of 200 bp segments directly upstream of the start of the 5' UTR of each gene and directly downstream of the 3' UTR of each gene.

In Figure 12A we see that in many VLP DNA samples, and specifically in the Nitrogen starvation experiment VLP DNA samples, genes tend to have higher coverage within them vs the coverage of directly adjacent sequences. Curiously, in the genomic DNA sample we see more genes with a higher in-gene coverage as well with genes ~3700 having a higher in-gene coverage compared to ~3000 genes with a higher out-gene coverage. Yet reassuringly, this difference is significantly more pronounced in the VLP layer DNA of N condition samples which have 4800~-5100~ genes with higher within-gene coverage compared to ~1800-~2100 genes with higher out-gene coverage.

Thus, each gene was given a score of the ratio of read counts mapped to the gene's center compared to its upstream and downstream flanking regions. This score for each gene in each sample was calculated by dividing the in-gene coverage with the out-gene coverage. When we compare this ratio score of VLP DNA with the same ratio in the genomic DNA sample (as seen in Figure 12B). We see that the Nitrogen starvation VLP DNA samples have a higher ratio score ratio relative to the genomic DNA. In contrast, as once again consistent with lower activation of the Ty element in the YP condition, we do not observe such a strong difference in the ratio score between VLP and genomic DNA in that (YP) condition. Further, we see that the ratio scores are higher in the Ty-Plus strain compared to Ty-Minus in both feeding conditions (as seen in Figure 12D). Notably, there is a significant number of genes with higher ratio in YP repeat 1 compared to repeat 2 and in N repeat 2 compared to repeat 1 (as seen in Figure 12E).

Evolution experiment of Ty plus and Ty minus strains

To determine whether there's a direct evolutionary effect of RT of genes on gene evolution, we performed an in-lab evolution experiment, comparing the rate of evolution of the Ty-Plus and Ty -Minus strains. We evolved both in an inducing nitrogen-poor Galactose medium. We used this condition for two reasons. Firstly, this condition is an example of a stress condition which induces transcription and retrotransposition of the

Ty element [11], and as we see above Figures 11-12, mRNAs of many more genes are encapsulated in the VLP in this condition. Second, as nitrogen starvation is a stress, we wanted to see if and how the activity of the Ty may enhance adaptation to this stress.

We used the methodology of serial dilution in which 10^6 cells are inoculated each day

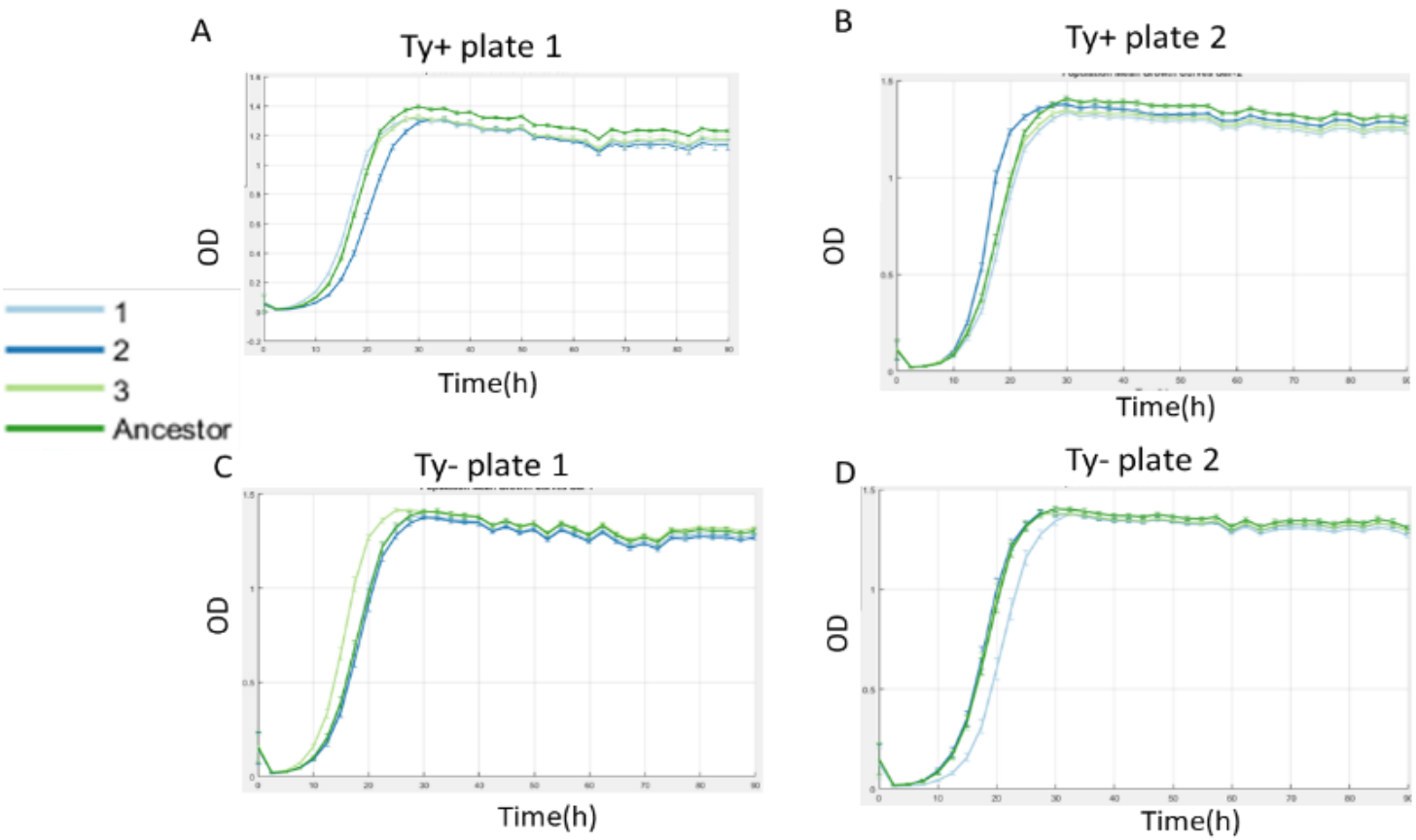


Figure 13: Growth curves of evolution experiment plates of Plus and Minus strains compared to their unevolved ancestors. Shows time vs OD. A. Gal Plus plate #1 (repeats 1-3). B. Gal Plus plate #2 (repeats 4-6). C. Gal Minus plate #1 (repeats 1-3). D. Gal Minus plate #2 (repeats 4-6).

in each well of a 24-well plate and are allowed to grow for 7 generations, thus reaching ~1.2*10^8 cells. Of these, ~0.01 of the cells are transferred to a fresh medium, thus keeping roughly constant population size. I have carried out the experiment for the two strains, in 6 repeats, with each 3 repeats on 2 different plates for a total duration of 84 days, or 294 generations. The different repeats were compared to their respective ancestor strains (Ty-plus or Ty-minus) as seen in Figure 13.

Surprisingly, we see that in all Plus samples, the ancestor has the highest final OD and that only some of the evolved repeats have managed to eclipse it in growth rate, with some evolved repeats having a visibly lower growth rate than their ancestor. In The Ty-Plus plate #1 (Figure 13A), repeat 1 has a higher growth rate relative to the ancestor, repeat 3 has a similar growth rate, and repeat 2 has a lower growth rate than the ancestor. Repeats 1-3 all have a similar carrying capacity which is significantly lower than the ancestor's. In Ty-Plus plate #2 (Figure 13B), repeat 2 has a higher growth rate relative to the ancestor, and repeats 1,3 have a similar growth rate to the ancestor. Repeats 1, 3 have a similar carrying capacity as well, while repeat 2's carrying capacity is higher. All 3 repeats have a lower carrying capacity than the ancestor's. In The Ty-Minus plate #1, repeat 3 has a higher growth rate relative to the ancestor while the other samples grow in a similar rate. The ancestor and repeat 3 has the highest carrying capacity, while repeats 1-2 have a lower carrying capacity. In Ty- plate #2 the ancestor has a similar growth rate to repeats 2, 3 and repeat 1 has a lower growth rate than it. The ancestor has the highest carrying capacity, yet carrying capacity appears to be close between repeats 1-3 and the ancestor.

Noticeably, carrying capacity difference between the ancestor and the evolved repeats appears to be generally higher in the Plus strain relative to the Minus strain.

Discussion

VLP enriched and depleted RNAs

Using our RNA sequencing data, we managed to characterize VLP depleted and VLP enriched genes in the Ty-Plus strain in the two growth conditions. We note several characteristics that are correlated with VLP enrichment and VLP depletion based on these characteristics. It appears that genes that are depleted from VLPs are more functionally conserved at the amino acid level (lower dN) and genes with higher mutation rate (both dN and dS) tend to be more VLP enriched. This implies VLP entrance has a mutagenic effect while genes that are subject to selective pressure tend to avoid encapsulation in the VLPs.

From the GO category analysis, we see that genes that fall into certain functions tend to be VLP depleted while VLP enriched genes don't have any significantly enriched GO categories. This implies that genes with certain functions are less likely to be encapsulated in VLPs but no such symmetry exists for genes more likely to be encapsulated in VLPs. Perhaps this is because VLP depleted genes tend to be more conserved and genes with critical functions, which are more likely to be conserved, are more likely to form pathways and be annotated to belong to functional GO categories, or that certain genes in certain crucial pathways evolved to avoid VLP encapsulation as to not increase their mutation rate.

Many genes in yeast experienced intron loss during evolution [8] and only ~5% of its genes contain introns [23]. As reverse transcription (RT) is a highly probable mechanistic explanation for intron loss [8] we tested whether this explanation works with our VLP encapsulation data. We find that genes with introns are indeed significantly depleted from the VLP. Perhaps the same mechanism that bars these genes from entering VLPs is a possible explanation for their intron retention. Or perhaps, these genes are barred from entering VLPs to conserve functionally important introns they contain. In any case the avoidance of intron containing genes from VLP encapsulation attests indirectly to the RT and possible genome integration of cDNA of genes whose mRNA do enter the VLP.

Ribosomal proteins (RPs) appear to be significantly VLP depleted. This might be due to the essentiality of their functions and conservation. While RPs with introns have a more significant overlap with VLP depleted genes than RPs that don't contain introns, non-ribosomal genes with introns are not significantly VLP depleted. This raises a "chicken and egg question" regarding RPs, their intron prevalence and intron loss and its relation to VLP inclusion. Are RPs with introns specifically excluded from VLPs to conserve their functionally relevant introns [50]? Did RPs retain their introns throughout *S. Cerevisiae's* evolution unlike most other yeast genes due to being excluded from VLPs for other reasons e.g., their RNA localization or an evolutionary need to limit their mutation rate, as RPs are very highly conserved [49].

ER enrichment of VLP enriched genes and depletion of VLP depleted genes is in concordance with previous studies. There is evidence that Ty VLPs are formed in specific subcellular compartments formed by phase separation called T-bodies or retrosomes[41]. There is evidence that Ty RNA nucleates the T-bodies via the endoplasmatic reticulum (ER)[42]. Perhaps, other RNAs that are localized in the ER are more likely to reach the Ty VLP via the same path. This is interesting as if we consider the mutagenic potential of VLP encapsulation and following RT, it implies cellular localization can affect the mutation rate of genes.

Interestingly, while DCP2 is distinct from T-bodies [51], RNAs that associate with P-bodies and with Dcp2 or Scd6p specifically, and have significant overlap with VLP enriched RNAs. Perhaps both are bound to the overlapping proteins between P-bodies and T-bodies [1] or, specifically, to Scd6p.

Half-life datasets appear to exhibit substantial disagreement between them. From two of the three datasets, we see that VLP enriched RNAs tend to have a longer half-life and one of those two datasets implies that VLP depleted genes have a shorter half-life on average. Perhaps, as RNAs exist for a longer time within the cell, they have a higher chance of being encapsulated at some point within the VLP before degrading. The third data set contrasts the other two, as according to it, VLP depleted genes have a longer half-life. As RNA half-life appears to vary drastically between conditions and measurements, perhaps it would be worthwhile to test it specifically within the conditions of the VLP experiment.

Mitochondrial genes are correlated with VLP depletion, yet RNA that localize to the mitochondria's outer membrane don't appear to overlap with VLP depleted genes but Interestingly, RNAs depleted from the mitochondria's outer membrane do overlap with VLP depleted genes.

Interestingly, there are many similarities in the difference between VLP and TOTAL in the Ty-Plus strain to the same difference in the Ty-Minus strain. This is surprising as previous results from the lab [20] indicate that the VLP enriched layer from the sucrose gradient centrifugation separation doesn't contain VLP RT activity in the Ty-Minus strain. Worryingly, his might imply that the RNA differences we see between the VLP

fraction, and the cell total are due to background differences that might be related to the sucrose gradient separation (perhaps other organelles or parts thereof are localized with the VLP fraction after centrifugation in the sucrose gradient). Otherwise, as the minus strain does appear to contain copies of Ty2, and Ty2 is enriched in minus- strain VLP relative to minus-strain TOTAL, perhaps the VLP fraction in the minus-strain is comprised of Ty2 VLPs. In that case, perhaps the experiment should be repeated with a control strain that contains no copies of Ty, to properly differentiate the VLP layer from the whole cell and removing any other possible background effects from the comparison.

Strangely, we see that Ty1 and His3 are both highly enriched in VLP RNA in the Minus strain compared to TOTAL in the Minus strain. This is strange as the Ty-Minus strain shouldn't contain the Ty1-HIS3 construct or any copy of Ty1. Ty1 is extremely highly expressed in Ty-plus VLP samples, so perhaps these counts are a result of index hopping, as VLP reads were multiplexed together before Illumina sequencing and were sequenced via single-index sequencing [52].

cDNA measurement in VLP DNA

We find signals confirming cDNA presence in VLP – we see exon-exon splice junction reads, we see higher read counts in transcribed regions (in gene) relative to untranscribed regions (out gene), with higher in-gene/out-gene coverage ratios in VLP DNA than genomic DNA. We see higher exon-intron read count ratios in all VLP DNA samples when comparing to genomic DNA as well.

We see the strongest signals confirming cDNA, considering both exon splice junction reads and the comparison of in-gene to out-gene read counts, in the N Plus samples - VLP-plus strain samples subjected to nitrogen starvation. This makes sense, as transcription of Tys and retrotransposition are highly induced in the Plus strain and, in normal conditions, by nitrogen starvation [11]. When comparing the samples which vary from the control by a single retrotransposition inducing factor: the N minus samples (VLP-minus strain samples subjected to nitrogen starvation) and YP Plus samples (VLP-plus strain samples in the YP condition) we see a stronger potential cDNA signal (as seen in exon-exon splice junction, exon-intron ratios, and in-gene to out-gene

comparison) in the N Minus samples relative to the YP Plus. As the GAL promoter is one of the strongest inducible promoters in *S. Cerevisiae* [53] and inducing Ty using the gal promoter is known to induce Ty VLPs [1], we can assume we strongly induce Ty1 expression in the YP plus samples. Therefore, the bottleneck in cDNA content is perhaps due to post transcriptional regulation. The mediator complex is a strong regulator of Ty1 retrotransposition in the post transcriptional level - decreasing the rate of viable VLPs which can reverse transcribe RNAs [15]. As there is evidence that subunits of the mediator complex might be regulated by stress [16], it has a potential of being a significant part of the post-transcriptional regulation of RT that maintains the bottleneck in RT that prevents us from seeing a stronger DNA signal in the YP Plus samples.

A concern we have is that some of the differences we see between the VLP DNA samples, and the genomic DNA are due to mapping artifacts. The RM11-1a strain we used as a basis for the Ty-Minus and Ty-Plus strains for the experiment is different from the S288C lab strain whose genome we used for mapping (due to having much better annotations and being of higher quality), while the dataset we used for our genomic DNA is from the S288C strain. Transcribed areas and exons tend to be more conserved than introns and untranscribed areas. Therefore, mapping of the cDNA reads might be biased against introns and untranscribed areas. This bias could skew our results towards finding a cDNA signal where there is none. To properly get rid of this bias we are planning to sequence DNA from the plus and minus strain, align it like we did for the cDNA and use it as our genomic control instead.

Evolution experiment

Many samples in the evolution experiment appear to have barely evolved at all in terms of growth rate – with a growth rate comparable to the ancestor and perhaps even “devolved” with a lower growth rate compared to the ancestor. This is quite concerning and might imply a longer evolution experiment is required. The carrying capacity of the strains decreased after evolution as well, with the ancestor measurements having the highest maximum OD during the growth experiment. Perhaps the decrease in capacity is an evolutionary strategy employed to conserve resources in a starved environment.

In some ecological niches, *S. Cerevisiae* population can adopt different life strategies categorized as “ants”, which reproduce quickly and have a higher carrying capacity with smaller cell sizes or ”grasshoppers” which reproduce slowly and have a lower carrying capacity with larger cell sizes [54]. Perhaps, the evolved yeasts adapted a “grasshopper” strategy which can potentially help them handle long term starvation [54].

In that case, it appears that evolved Ty Plus strains have a lower carrying capacity relative to the ancestor than the Ty Minus strains which reach a similar capacity to the ancestor. One notable exception to that statement is repeat 2 in Ty plate 2 which, while having a carrying capacity only slightly smaller than the ancestor, has a higher growth rate and perhaps adopted a different strategy to adapt to the nitrogen starvation stress. This implies Ty Plus strains have better adapted to the condition or, at least, have managed to grow more different than their ancestor relative to the strains evolved from the Ty Minus strain. We will proceed with evolving these strains for more generations and attempt to understand, via sequencing, the extent to which they evolved differently.

Acknowledgements:

I would like to thank Prof. Tzachi Pilpel for welcoming me to his amazing lab, for his infectious enthusiasm and optimistic outlook, and for many fruitful and stimulating discussions.

I would like to thank Dr. Orna Dahan for her frequent and indispensable guidance, and for her expertise in anything wet or yeasty.

I would like to thank Yonat Gurvich and Sivan Kaminski Strauss for laying the experimental foundation for this work, without whom this work would not be possible.

I would like to thank all the other lab members for their intelligent insights, for always lending a hand when needed, and for making the lab a welcoming and cooperative environment.

I would like to thank my friends and family for their constant support and especially my partner, Naama Barak for her immeasurable empathy and understanding.

References

[1] M. J. Curcio, S. Lutz, and P. Lesage, " The Ty1 LTR-Retrotransposon of Budding Yeast, Saccharomyces cerevisiae ," *Microbiol Spectr*, vol. 3, no. 2, Apr. 2015, doi: 10.1128/microbiolspec.mdna3-0053-2014.

[2] L. K. Derr, J. N. Strathern, and D. J. Garfinkel, "RNA-Mediated Recombination in S. cerevisiae," 1991.

[3] A. A. Fried, M. Kiechle, H. G. Maxeiner, R. H. Schiestl, and F. Eckardt-Schupp, "Ty1 integrase overexpression leads to integration of non-Ty1 DNA fragments into the genome of Saccharomyces cerevisiae," *Molecular Genetics and Genomics*, vol. 284, no. 4. pp. 231–242, Oct. 2010. doi: 10.1007/s00438-010-0561-4.

[4] P. H. Maxwell and M. J. Curcio, "Retrosequence formation restructures the yeast genome," *Genes Dev*, vol. 21, no. 24, pp. 3308–3318, Dec. 2007, doi: 10.1101/gad.1604707.

[5] P. H. Maxwell, C. Coombes, A. E. Kenny, J. F. Lawler, J. D. Boeke, and M. J. Curcio, " Ty1 Mobilizes Subtelomeric Y' Elements in Telomerase-Negative Saccharomyces cerevisiae Survivors ," *Mol Cell Biol*, vol. 24, no. 22, pp. 9887–9898, Nov. 2004, doi: 10.1128/mcb.24.22.9887-9898.2004.

[6] J. Schacherer, Y. Tourrette, J. L. Souciet, S. Potier, and J. de Montigny, "Recovery of a function involving gene duplication by retroposition in Saccharomyces cerevisiae," *Genome Res*, vol. 14, no. 7, pp. 1291–1297, Jul. 2004, doi: 10.1101/gr.2363004.

[7] J.-F. Gout *et al.*, "The landscape of transcription errors in eukaryotic cells," 2017.

[8] K. B. Hooks, D. Delneri, and S. Griffiths-Jones, "Intron evolution in Saccharomycetaceae," *Genome Biol Evol*, vol. 6, no. 9, pp. 2543–2556, Sep. 2014, doi: 10.1093/gbe/evu196.

[9] G. Servant *et al.*, "Tye7 regulates yeast Ty1 retrotransposon sense and antisense transcription in response to adenylic nucleotides stress," *Nucleic Acids Res*, vol. 40, no. 12, pp. 5271–5282, Jul. 2012, doi: 10.1093/nar/gks166.

[10] C. Sacerdot, G. Mercier, A. L. Todeschini, M. Dutreix, M. Springer, and P. Lesage, "Impact of ionizing radiation on the life cycle of Saccharomyces cerevisiae Ty1 retrotransposon," *Yeast*, vol. 22, no. 6, pp. 441–455, Apr. 2005, doi: 10.1002/yea.1222.

[11] A. Morillon, M. Springer, and P. Lesage, "Activation of the Kss1 Invasive-Filamentous Growth Pathway Induces Ty1 Transcription and Retrotransposition in Saccharomyces cerevisiae," 2000. [Online]. Available: http://www.mips.biochem.mpg.de/

[12] T. Mcclanahan and K. Mcentee, "Specific Transcripts Are Elevated in Saccharomyces cerevisiae in Response to DNA Damage," 1984.

[13] V. A. Bradshaw and K. McEntee, "DNA damage activates transcription and transposition of yeast Ty retrotransposons," 1989.

[14] L. Staleva Staleva and P. Venkov, “Activation of Ty transposition by mutagens,” 2001.

[15] A. C. Salinero, E. R. Knoll, Z. I. Zhu, D. Landsman, M. J. Curcio, and R. H. Morse, “The Mediator co-activator complex regulates Ty1 retromobility by controlling the balance between Ty1i and Ty1 promoters,” *PLoS Genet*, vol. 14, no. 2, Feb. 2018, doi: 10.1371/journal.pgen.1007232.

[16] M. J. Curcio, “Border collies of the genome: domestication of an autonomous retrovirus-like transposon,” *Current Genetics*, vol. 65, no. 1. Springer Verlag, pp. 71–78, Feb. 11, 2019. doi: 10.1007/s00294-018-0857-1.

[17] C. M. Wilke and J. Adams, “Fitness Effects of Ty Transposition in Saccharomyces cerevisiae,” 1992.

[18] S. L. Chang, H. Y. Lai, S. Y. Tung, and J. Y. Leu, “Dynamic Large-Scale Chromosomal Rearrangements Fuel Rapid Adaptation in Yeast Populations,” *PLoS Genet*, vol. 9, no. 1, Jan. 2013, doi: 10.1371/journal.pgen.1003232.

[19] D. J. Eichinger and J. D. Boeke, “The DNA Intermediate in Yeast Tyl Element Transposition Copurifies with Virus-like Particles: Cell-Free Tyl Transposition,” 1966.

[20] S. Kaminski, “Reverse Transcription as a potential molecular mechanism for Lamarckian evolution,” Weizmann Institute of Science, 2015.

[21] H. Keren-Shaul *et al.*, “MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing,” *Nat Protoc*, vol. 14, no. 6, pp. 1841–1862, Jun. 2019, doi: 10.1038/s41596-019-0164-4.

[22] A. Dobin *et al.*, “STAR: Ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[23] J. M. Cherry *et al.*, “Saccharomyces Genome Database: The genomics resource of budding yeast,” *Nucleic Acids Res*, vol. 40, no. D1, Jan. 2012, doi: 10.1093/nar/gkr1029.

[24] Y. Jin, O. H. Tam, E. Paniagua, and M. Hammell, “TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets,” *Bioinformatics*, vol. 31, no. 22, pp. 3593–3599, May 2015, doi: 10.1093/bioinformatics/btv422.

[25] M. Carr, D. Bensasson, and C. M. Bergman, “Evolutionary Genomics of Transposable Elements in Saccharomyces cerevisiae,” *PLoS One*, vol. 7, no. 11, Nov. 2012, doi: 10.1371/journal.pone.0050978.

[26] M. Tarailo-Graovac and N. Chen, “Using RepeatMasker to identify repetitive elements in genomic sequences,” *Current Protocols in Bioinformatics*, no. SUPPL. 25. 2009. doi: 10.1002/0471250953.bi0410s25.

[27] M. Smid *et al.*, “Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample

comparisons," *BMC Bioinformatics*, vol. 19, no. 1, Jun. 2018, doi: 10.1186/s12859-018-2246-7.

[28] R. Kohen *et al.*, "UTAP: User-friendly Transcriptome Analysis Pipeline," *BMC Bioinformatics*, vol. 20, no. 1, Mar. 2019, doi: 10.1186/s12859-019-2728-2.

[29] T. Smith, A. Heger, and I. Sudbery, "UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy," *Genome Res*, vol. 27, no. 3, pp. 491–499, Mar. 2017, doi: 10.1101/gr.209601.116.

[30] Y. Liao, G. K. Smyth, and W. Shi, "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014, doi: 10.1093/bioinformatics/btt656.

[31] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol*, vol. 15, no. 12, Dec. 2014, doi: 10.1186/s13059-014-0550-8.

[32] M. Stephens, "False discovery rates: A new deal," *Biostatistics*, vol. 18, no. 2, pp. 275–294, Apr. 2017, doi: 10.1093/biostatistics/kxw041.

[33] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack, "Open Access METHOD Gene ontology analysis for RNA-seq: accounting for selection bias GOseq GOseq is a method for GO analysis of RNA-seq data that takes into account the length bias inherent in RNA-seq," 2010. [Online]. Available: http://genomebiology.com/2010/11/2/R14

[34] J. Peter *et al.*, "Genome evolution across 1,011 Saccharomyces cerevisiae isolates," *Nature*, vol. 556, no. 7701, pp. 339–344, Apr. 2018, doi: 10.1038/s41586-018-0030-5.

[35] F. Sievers *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Mol Syst Biol*, vol. 7, 2011, doi: 10.1038/msb.2011.75.

[36] M. Goujon *et al.*, "A new bioinformatics analysis tools framework at EMBL-EBI," *Nucleic Acids Res*, vol. 38, no. SUPPL. 2, May 2010, doi: 10.1093/nar/gkq313.

[37] Z. Yang, "PAML 4: Phylogenetic analysis by maximum likelihood," *Mol Biol Evol*, vol. 24, no. 8, pp. 1586–1591, Aug. 2007, doi: 10.1093/molbev/msm088.

[38] E. Talevich, B. M. Invergo, P. J. Cock, and B. A. Chapman, "SOFTWARE Open Access Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython," 2012. [Online]. Available: http://www.biomedcentral.com/1471-2105/13/209

[39] H. C. Medina-Munoz, C. P. Lapointe, D. F. Porter, M. Wickens, and M. Rosbash, "Records of RNA locations in living yeast revealed through covalent marks," *PNAS*, vol. 117, no. 38, pp. 23539–23547, 2020, doi: 10.1073/pnas.1921408117/-/DCSupplemental.

[40] C. Wang, F. Schmich, S. Srivatsa, J. Weidner, N. Beerenwinkel, and A. Spang, "Context-dependent deposition and regulation of mRNAs in P-bodies", doi: 10.7554/eLife.29815.001.

[41] C. Miller *et al.*, “Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast,” *Mol Syst Biol*, vol. 7, 2011, doi: 10.1038/msb.2010.112.

[42] B. Neymotin, R. Athanasiadou, and D. Gresham, “Determination of in vivo RNA kinetics using RATE-seq,” *RNA*, vol. 20, no. 10, pp. 1645–1652, Oct. 2014, doi: 10.1261/rna.045104.114.

[43] S. E. Munchel, R. K. Shultzaberger, N. Takizawa, and K. Weis, “Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay,” *Mol Biol Cell*, vol. 22, no. 15, pp. 2787–2795, Aug. 2011, doi: 10.1091/mbc.E11-01-0028.

[44] S. K. Strauss *et al.*, “Evolthon: A community endeavor to evolve lab evolution,” *PLoS Biol*, vol. 17, no. 3, Mar. 2019, doi: 10.1371/journal.pbio.3000182.

[45] F. Malagon and T. H. Jensen, “ T-body formation precedes virus-like particle maturation in S. cerevisiae ,” *RNA Biol*, vol. 8, no. 2, pp. 184–189, Mar. 2011, doi: 10.4161/rna.8.2.14822.

[46] J. H. Doh, S. Lutz, and M. J. Curcio, “Co-translational Localization of an LTR-Retrotransposon RNA to the Endoplasmic Reticulum Nucleates Virus-Like Particle Assembly Sites,” *PLoS Genet*, vol. 10, no. 3, 2014, doi: 10.1371/journal.pgen.1004219.

[47] H. C. Medina-Munoz, C. P. Lapointe, D. F. Porter, M. Wickens, and M. Rosbash, “Records of RNA locations in living yeast revealed through covalent marks,” *PNAS*, vol. 117, no. 38, pp. 23539–23547, 2020, doi: 10.1073/pnas.1921408117/-/DCSupplemental.

[48] M. Spingola, L. Grate, D. Haussler, and M. Ares, “Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae,” 1999. [Online]. Available: http://www.cse.ucsc.edu/research/compbio/yeast_

[49] C. Pál, B. Papp, and L. D. Hurst, “Letter to the Editor Highly Expressed Genes in Yeast Evolve Slowly,” 2001. [Online]. Available: https://academic.oup.com/genetics/article/158/2/927/6049680

[50] J. Parenteau *et al.*, “Introns within ribosomal protein genes regulate the production and function of yeast ribosomes,” *Cell*, vol. 147, no. 2, pp. 320–331, Oct. 2011, doi: 10.1016/j.cell.2011.08.044.

[51] F. Malagon and T. H. Jensen, “ The T Body, a New Cytoplasmic RNA Granule in Saccharomyces cerevisiae ,” *Mol Cell Biol*, vol. 28, no. 19, pp. 6022–6032, Oct. 2008, doi: 10.1128/mcb.00684-08.

[52] M. Costello *et al.*, “Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms,” *BMC Genomics*, vol. 19, no. 1, May 2018, doi: 10.1186/s12864-018-4703-0.

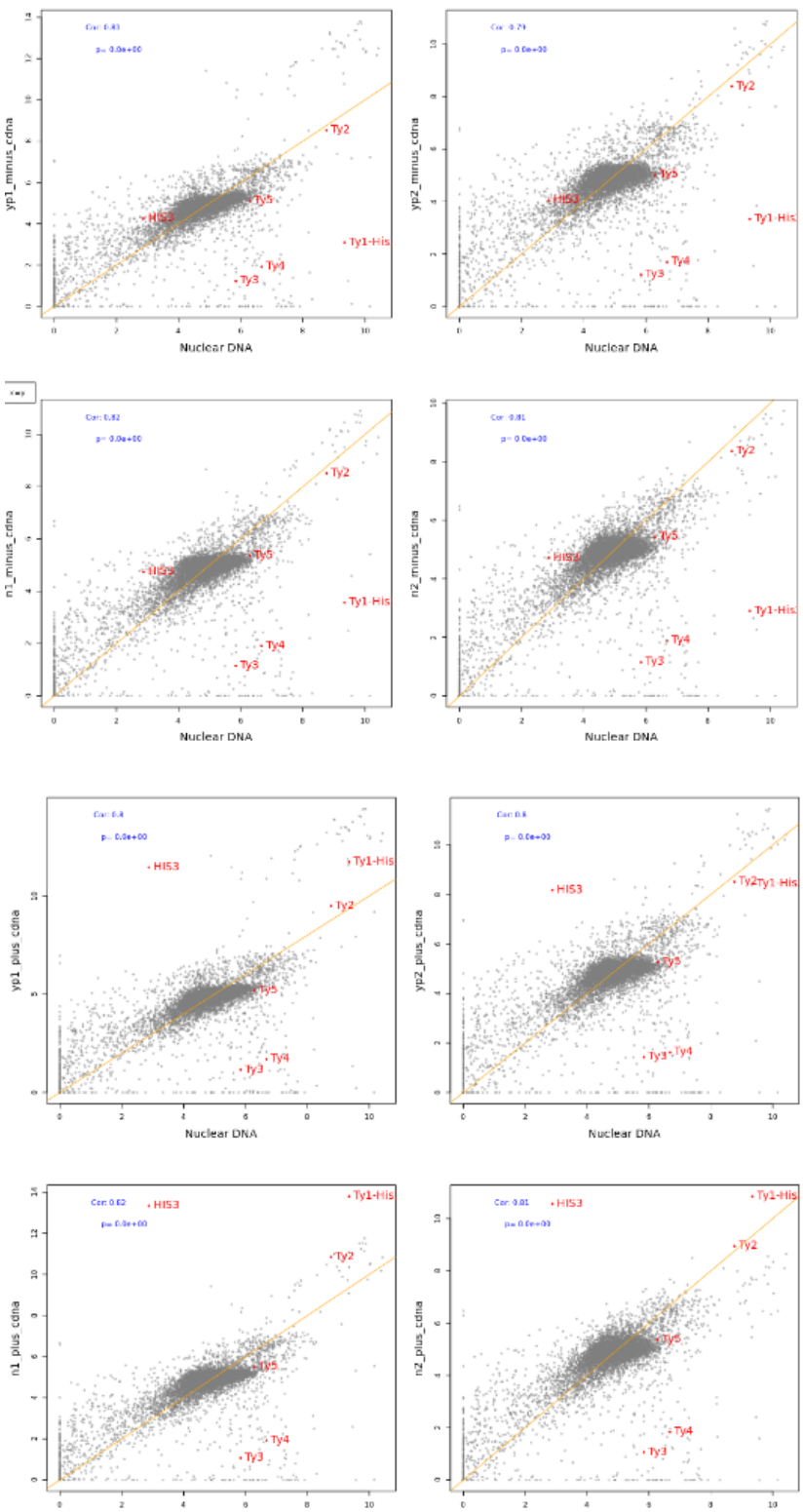
[53] M. N. Štagoj, A. Comino, and R. Komel, “Fluorescence based assay of GAL system in yeast Saccharomyces cerevisiae,” *FEMS Microbiol Lett*, vol. 244, no. 1, pp. 105–110, Mar. 2005, doi: 10.1016/j.femsle.2005.01.041.

[54] A. Spor, T. Nidelet, J. Simon, A. Bourgais, D. De Vienne, and D. Sicard, “Niche-driven evolution of metabolic and life-history strategies in natural and domesticated populations of Saccharomyces cerevisiae,” *BMC Evol Biol*, vol. 9, no. 1, 2009, doi: 10.1186/1471-2148-9-296.

Supplementary figures:

Enriched in VLP-Depleted RNAs in Ty-Plus	Enriched in VLP-Depleted RNAs in Ty-Minus
cellular metabolic process: 4.1e-10, metabolic process: 8.7e-10, nitrogen compound metabolic process: 8.0e-04, primary metabolic process: 9.0e-06, organic substance metabolic process: 6.8e-05, cytoplasm: 9.5e-08, intracellular organelle: 2.2e-06, organelle: 4.1e-06, cellular process: 2.7e-02, intracellular: 7.1e-11, cellular anatomical entity: 9.9e-04, protein-containing complex: 2.8e-07, non-membrane-bounded organelle: 4.9e-04, intracellular non-membrane-bounded organelle: 4.9e-04, gene expression: 3.8e-03, cellular nitrogen compound metabolic process: 2.7e-02, cellular nitrogen compound biosynthetic process: 3.3e-05, cellular biosynthetic process: 1.0e-06, biosynthetic process: 5.3e-08, organic substance biosynthetic process: 1.6e-07, organonitrogen compound metabolic process: 2.6e-08, cellular protein metabolic process: 4.0e-05, protein metabolic process: 9.1e-04, mitochondrion: 1.5e-06, small molecule metabolic process: 8.0e-04, carboxylic acid metabolic process: 3.4e-03, oxoacid metabolic process: 5.0e-03, organic acid metabolic process: 9.0e-03, generation of precursor metabolites and energy: 2.7e-03, oxidation-reduction process: 1.8e-03, oxidoreductase activity: 1.9e-02, mitochondrial matrix: 1.9e-02, ligase activity: 6.9e-06, nucleotide metabolic process: 6.5e-03, nucleoside phosphate metabolic process: 7.0e-03, purine ribonucleotide metabolic process: 5.3e-04, ribonucleotide metabolic process: 4.1e-03, purine-containing compound metabolic process: 4.4e-03, purine nucleotide metabolic process: 4.1e-04, ribose phosphate metabolic process: 1.6e-03, mitochondrial protein complex: 4.7e-07, proton transmembrane transport: 2.0e-03, proton transmembrane transporter activity: 7.5e-03, proton-transporting two-sector ATPase complex: 4.7e-03, ion channel activity: 4.4e-02, mitochondrial proton-transporting ATP synthase complex: 4.7e-03, proton-transporting ATP synthase complex: 4.7e-03, proton channel activity: 3.4e-04, proton-transporting ATP synthase activity, rotational mechanism: 3.4e-04, purine nucleoside triphosphate metabolic process: 3.8e-02, purine ribonucleoside triphosphate metabolic process: 1.6e-02, purine ribonucleoside triphosphate biosynthetic process: 9.1e-03, purine nucleoside triphosphate biosynthetic process: 9.1e-03, ATP synthesis coupled proton transport: 6.8e-04, ATP biosynthetic process: 6.8e-04, energy coupled proton transport, down electrochemical gradient: 6.8e-04, electron transport chain: 4.7e-02, oxidative phosphorylation: 8.4e-04, cytochrome complex: 2.2e-02, respiratory chain complex: 1.1e-02, mitochondrial respirasome: 2.8e-02, ATP metabolic process: 1.5e-09, inner mitochondrial membrane protein complex: 4.3e-06, RNA binding: 1.3e-03, ribonucleoprotein complex: 1.4e-12, cytosolic large ribosomal subunit: 1.1e-07, large ribosomal subunit: 2.1e-06, structural molecule activity: 2.8e-13, ribosome: 4.6e-20, ribosomal subunit: 9.6e-20, structural constituent of ribosome: 1.0e-19, cytosolic ribosome: 3.0e-20, cytoplasmic translation: 4.9e-19, small ribosomal subunit: 8.2e-11, cytosolic small ribosomal subunit: 9.4e-11, preribosome: 3.6e-02, ribosomal small subunit biogenesis: 2.4e-03, maturation of SSU-rRNA: 1.4e-02, maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA): 1.4e-02, ribosome biogenesis: 8.9e-08, ribonucleoprotein complex biogenesis: 8.0e-07, ncRNA metabolic process: 1.2e-04, ncRNA processing: 1.2e-02, rRNA metabolic process: 9.6e-05, rRNA processing: 4.0e-04, cytosol: 5.0e-09, organonitrogen compound biosynthetic process: 1.2e-15, translation: 1.2e-18, peptide biosynthetic process: 2.2e-19, amide biosynthetic process: 1.2e-15, peptide metabolic process: 2.7e-19, cellular amide metabolic process: 7.7e-1	cytoplasm: 7.5e-07, intracellular: 9.0e-04, organonitrogen compound metabolic process: 2.4e-09, cellular metabolic process: 3.4e-05, metabolic process: 1.1e-04, primary metabolic process: 2.9e-03, organic substance metabolic process: 4.3e-03, ATP metabolic process: 3.6e-02, nucleotide biosynthetic process: 4.9e-02, ribosome assembly: 3.8e-02, ribonucleoprotein complex subunit organization: 4.3e-02, cytosolic small ribosomal subunit: 2.1e-11, small ribosomal subunit: 8.9e-09, preribosome: 3.0e-02, ribosomal small subunit biogenesis: 7.5e-05, maturation of SSU-rRNA: 7.5e-04, maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA): 3.4e-03, cytosolic large ribosomal subunit: 1.1e-12, large ribosomal subunit: 1.2e-09, cytoplasmic translation: 2.1e-23, cytosolic ribosome: 2.6e-26, ribosome: 1.9e-23, structural molecule activity: 1.4e-12, ribosomal subunit: 2.3e-21, structural constituent of ribosome: 3.2e-21, ribonucleoprotein complex: 1.7e-13, ribosome biogenesis: 1.9e-07, ribonucleoprotein complex biogenesis: 2.2e-07, RNA binding: 5.9e-04, rRNA processing: 7.6e-03, rRNA metabolic process: 2.1e-02, ncRNA metabolic process: 1.5e-02, ncRNA processing: 1.5e-02, cytosol: 3.8e-05, organonitrogen compound biosynthetic process: 4.7e-20, cellular amide metabolic process: 8.6e-16, amide biosynthetic process: 1.3e-15, peptide metabolic process: 4.9e-15, translation: 4.3e-16, peptide biosynthetic process: 8.7e-16, cellular nitrogen compound biosynthetic process: 4.7e-02, cellular biosynthetic process: 1.2e-05, organic substance biosynthetic process: 1.1e-05, biosynthetic process: 1.9e-05, small molecule metabolic process: 1.0e-04, organic acid metabolic process: 5.0e-05, carboxylic acid metabolic process: 2.9e-05, oxoacid metabolic process: 9.2e-05, cellular amino acid metabolic process: 1.6e-04, alpha-amino acid metabolic process: 4.7e-04

Supplementary figure 1. Table containing all enriched GO categories and their enrichment p-value for each VLP depleted set. They are colored and ordered according to their clusters.



Supplementary figure 2. Scatterplots comparing read counts for all genes of the genomic DNA control (x axis) vs VLP layer DNA (y axis) of all the VLP DNA samples. Ty genes and His3 are marked on all plots. The orange line is x=y.