Thesis for the degree
Master of Science

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

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Sivan Kaminski

מאת
סיון קמינסקי

שעתוק במהופך כמנגנון מולקולרי אפשרי בדרך לאבולוציה למרקיסטית

Reverse Transcription as a potential molecular mechanism for Lamarckian evolution

Advisor:
Prof. Yitzhak Pilpel

מנחה:
פרופ' יצחק פלפל

December 2015

כסלו תשע"ו

Abstract

In evolution, there is a tradeoff between high mutation rate that causes rapid adaptation to new environments and low mutation rate that stabilizes population and preserves genes that were adapted before. There are few mechanisms that enable high mutation rate in expressed genes while keeping un-transcribed regions with background, low mutation rate such as Transcription-associated mutagenesis (TAM). In this work we suggest reverse transcription (RT) as a novel mechanism that can locally increase mutation rate of transcribed genes. In this project we aimed to better understand the process of reverse transcription in yeast, and the Ty-mediated effects on evolution using experimental and theoretical methods. Ty elements in yeast are retrotransposons elements that enable the reverse transcription of an mRNA and its re-incorporation to the genome. The Ty element creates a Virus Like Particle (VLP) within which the reverse transcription process takes place. As previous studies detected some of the genome's mRNAs encapsulated within the VLP, an intriguing possibility is that non-Ty genes might be able to "hitch-hike" on the VLP and get to be reverse transcribed and hence enter into the genome in a potentially modified way. To more systematically characterize all the mRNA content of the VLP we have developed a protocol for the isolation of the particles from the cell and have established a set of criteria to measure, optimize and ensure its high purity. Indeed we have succeeded in purifying and isolating VLPs from yeast cells and the samples are now ready for both RNA and cDNA sequencing of the VLP contents. In addition, we have conducted a lab evolution experiment *in vivo* and *in silico* to assess the effects of reverse transcription on evolution. For that we have created three yeast strains that encode and express various versions of the Ty element and evolved each over time in the lab. The lab evolution experiments have shown that strains that harbored reverse transcription capacity have adapted more rapidly than compared to a strain that evolved without reverse transcription. Yet further investigation indicated that the plasmid has integrated into the genome, an event that by itself could account for the improved fitness. This indicates that a next phase of the project will require evolving a strain in which the plasmid is genomic from the beginning. The evolution *in silico*, which also compared between species that evolve with and without reverse transcription capabilities have shown that in some cases, characterized by computationally harder optimization tasks, RT is preferred while in other cases it is unfavorable. In contrast, competition experiments done using the simulation have not shown this trend, and have generated puzzling results which are not clear to us. This project has shown that reverse transcription has potential beneficial effects on evolution that are not completely understood. By sequencing the mRNA and cDNA of the VLPs as well as characterize better the results of the evolution experiment we hope to get better understanding of the reverse transcription-mediated evolution.

Table of Contents

1. **Introduction**

In the field of evolutionary biology, the main theory that explains adaptation of populations to new environments is the natural selection theory, suggested by Darwin.

The natural selection theory by Darwin is based on three concepts; 1) random mutations in space (locus) and time; The mutations are beneficial or not, 2) selection (the fittest individuals succeed better than other), known as natural selection, and 3) inheritance of those mutations to the next generation such that beneficial mutations will increase in frequency in the next generation [1]. It is important to note that in Darwinian evolution the environment has no effect on mutagenesis, but it merely acts as the selection power.

A different theory was suggested earlier by Lamarck which suggests a non-random process in adaptation of species. Lamarck's theory, also known as Inheritance of Acquired Characters is based on couple of concepts; 1) the use of a particular organ would lead to its gradual functional improvement and 2) inheritance of those improvements through generations. [1]. The giraffe example is the most known example of Lamarckian evolution, in which a giraffe tries to reach high leaves; in response, and because of the extensive use of the neck, the neck will elongate due to mutations in "neck length" genes. These genes will be passed on to the next generation which could elongate it further and so on until a perfection of the trait.

The "long neck" example would be explained by Darwin by suggesting that random mutation occurred in the population; in one (or more) of the individuals this mutation caused a longer neck that allowed this individual to reach higher leaves, thus, live longer and have more offspring with a long neck.

Two main differences between Darwinian and Lamarckian theories puts them on opposites edges; the randomness of mutations and their nature; while in Lamarckian evolution mutations occur preferentially in used systems in the organism (genes, organs etc.), at a given environment, and are beneficial in nature, in Darwinian evolution mutations occur at random position in the genome, and can be beneficial, neutral or deleterious. Therefore, any deviation from total randomness into directed mutagenesis, or from beneficial-only mutations to beneficial-or-not mutations, cannot be considered as Lamarckian or Darwinian evolution. It is therefore possible to think of evolution as a spectrum between Lamarckian and Darwinian evolution (figure 1).

There are few biological mechanisms that are suggested to be on the spectrum between Darwinian and Lamarckian evolution; Horizontal Gene Transfer (HGT) in which bacteria uses foreign DNA from the environment and incorporate it into the genome [2], is mainly Darwinian since the position of the

mutation (where in the genome it enters, and the function of the foreign DNA) and the timing of the event are random, but the fact that the environment (i.e. other organisms' genomes) provides the "mutation" for the receiving cell deviates from the pure Darwinian evolution. Another example can be seen in transcription-associated mutagenesis (TAM). It was shown that in some cases, like in cancer cells, highly transcribed genes accumulate more DNA mutations in the genome, likely due to the openness of the chromatin in highly expressed genes [3]. This phenomenon is a massive deviation from Darwinian evolution since it implies that mutations are biased in time (i.e. environmental conditions), and space (i.e. locations along the genome). A third potential mechanism is the reverse transcription (RT) that has not been discussed before in the context of Lamarckian evolution, and is the focus of this project (figure 1).

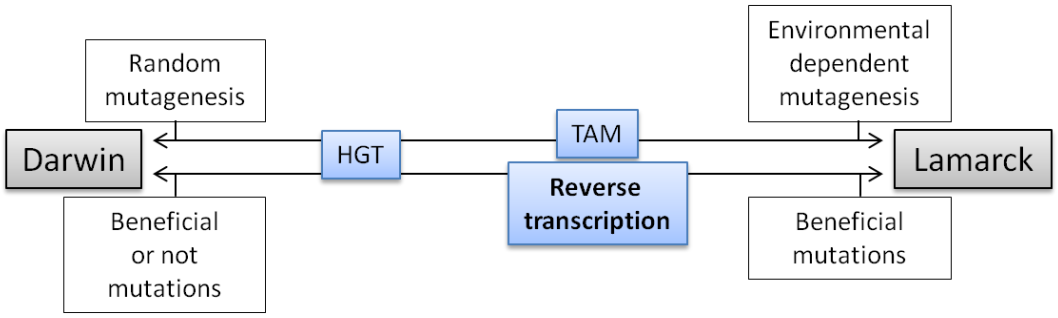

FIGURE 1. DARWIN'S THEORY AND LAMARCK'S THEORY ARE FOUND ON TWO EDGES OF AN EVOLUTIONARY SPECTRUM

Darwin's and Lamarck's theories differ in two main concepts. While Lamarck state that the environment is the cause of mutagenesis, Darwin's state that the environment role in evolution is merely by selection of the fittest. Moreover, in Lamarck's theory mutations are beneficial, while in Darwinian evolution mutations can be beneficial or not. Few mechanism can contribute to non-Darwinian evolution such as the HGT that presents a small deviation from Darwinian evolution and the TAM and reverse transcription that present a bigger step from Darwinian toward Lamarckian evolution. The reverse transcription mechanism is the focus of this project.

In the Central Dogma of biology, DNA is being replicated by the DNA polymerase and transcribed into RNA by the RNA polymerase; the RNA is being translated into protein. In this process, the RNA polymerase has couple orders of magnitude higher mutation rate than the DNA polymerase [4]. The reverse transcription process uses a reverse transcriptase (RT) enzyme to reverse transcribe RNA into DNA; the DNA can be then integrated into the genome by other enzymes such as the Integrase (Intp). There are two interesting aspects of RT affecting evolution.

First, RT mutagenesis occurs on expressed genes only; gene expression in the cell is not constant and is affected by environmental changes [5]. Since transcription, as well as RT increases mutation rate of expressed genes only, and the expressed genes is environment-dependent, RT-dependent evolution seems to be a Lamarckian process. On the other hand, the mutations occur in random position in the mRNA, and can be beneficial, deleterious or neutral such as in Darwinian evolution. Therefore, RT-dependent evolution is a step between Darwinian and Lamarckian evolution.

Second, mutation rate should be kept in a balance between high mutation rate that will allow organisms to adapt to new environments, and low mutation rate that will allow maintaining those adaptations for long periods of time [6]. Since the RNA polymerase has orders of magnitude higher mutation rate than the DNA polymerase, we suggest that RT can locally increase mutation rate of expressed genes while allowing constant and low mutation rate in the genome, and therefore settle the tradeoff.

Saccharomyces cerevisiae, in this respect, can be used as a model organism for RT-dependent evolution. *S. cerevisiae* has a reverse transcription capability in the form of a retroelement called Ty. There are 5 families of Ty elements in yeast (called Ty1-Ty5) each of them is found in multiple copies in the genome and all together comprise about 3% of its genome [7]. The most researched and known Ty is Ty1.

Ty1 has ~30 copies in *S. cerevisiae* genome; it is composed of 2 ORFs, the TyA ORF, also known as the Gag ORF and TyB ORF found in the +1 frame of TyA and is known as the Pol ORF. TyA ORF contains a structural protein (gag) that is used to form a capsid. TyB is the catalytic ORF containing 3 proteins, RT, integrase (Intp) and protease. The protease is auto-cleaved from a single TyB peptide, and catalyzes the cleavage of the Intp and the RT. The Ty1 life cycle begins with transcription and translation of both ORFs. The gag proteins form a Virus Like Particle (VLP) that encapsulates the Ty1's mRNA and all Ty1's proteins as well as a $tRNA^{Met}$ that is used as a primer for reverse transcription. In the VLP, a reverse transcription process of the Ty1 mRNA takes place followed by the import of the cDNA into the nucleus and its integration into the genome by the Intp (figure 2) [7]–[9].

In order for the RT to affect evolution through high mutation rate of expressed genes, the RT has to reverse transcribe cellular mRNAs in addition to the Ty's mRNA; we assume that mRNAs will be reverse transcribed only if they are encapsulated in the VLPs. It is therefore important to know if cellular mRNAs can be encapsulated in the VLPs and if they can be reverse transcribed. Curcio and Garfinkel have shown that the *HIS3*'s mRNA is being reverse transcribed by the Ty element. The fact that this gene was reverse transcribed indicates that it was encapsulated in the VLP [10]. It is therefore possible

to assume that other mRNAs except for the Ty's mRNA can be encapsulated and reverse transcribed in the VLPs. Indeed, Maxwell et al have shown that the VLPs contain other mRNA including the Ty1, *HIS3* and other mRNAs, as well as the Y' element (i.e. subtelomeric regions including a helicase ORF that take part in telomere's maintenance) [11].

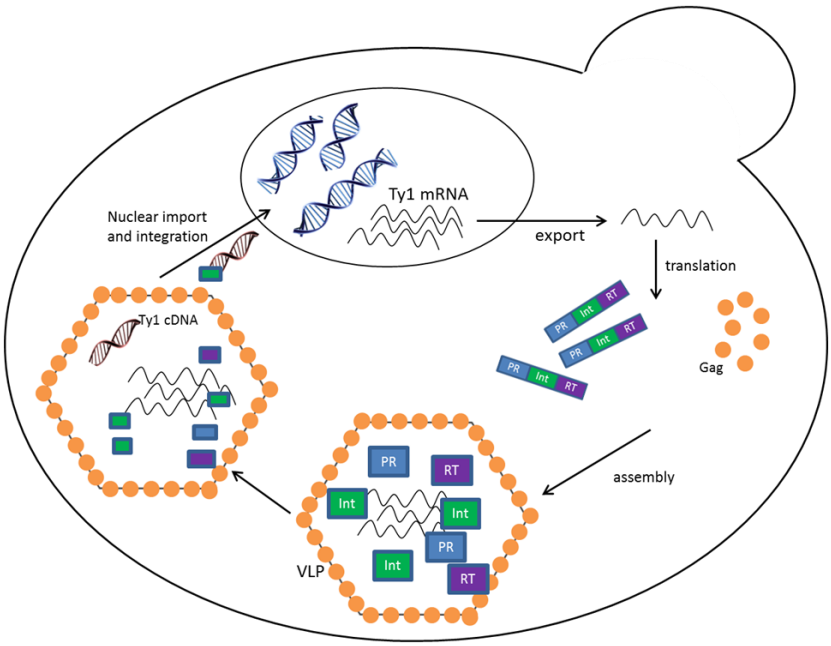

FIGURE 2. SCHEMATIC REPRESENTATION OF THE TY LIFE CYCLE IN *S. CEREVISIAE*

The Ty life cycles includes the following steps: transcription, export to the cytoplasm, translation of Ty proteins, assembly of the VLPs, proteolytic cleavages of the Ty's polypeptide, reverse transcription, transport to nuclear and integration to the genome.

In this project, we would like to expand Maxwell experiment and identify which mRNAs are encapsulated in the VLPs using RNA-seq, and in addition which of them are being reverse transcribed into cDNA, using DNA-seq, thus understanding more the encapsulation and reverse transcription processes and its gene specific preferences, if any. By knowing which of the genome's genes are being reverse transcribed we could understand in which environmental conditions the RT is most likely to affect evolution. For example, since the *HIS3* mRNA is being reverse transcribed, the evolution of cells toward low Histidine levels can be mediated through RT.

Another goal of this project is to understand the effects of RT on evolution such as changing dynamics of adaptation or leading to different solutions. To address this notion, we initiated a lab evolution experiment using yeast cells that express Ty and yeast cells that do not express Ty. At the end of the evolution the differences between the cells that have evolved with expressed Ty and the cells that evolved without it will be identify using fitness evaluations and DNA-seq. In addition, lab evolution *in silico* was performed by using a simulation that simulates evolution with and without RT. Using the simulation we aimed to further understand the effects of RT on evolution in general.

2. Results

2.1 Experimental Design

A Ty-less strain of *S. cerevisiae*, RM-11a, was used for both the VLP extraction and lab evolution experiments. Cells were transformed with one of three plasmids to conduct a three strains system. A Ty-containing plasmid (pGal-Ty) was kindly given to us by Pascale Lesage's lab. The plasmid contains a Ty element under the regulation of Gal1 promoter (3a), enabling expression of Ty on galactose-containing media, and inhibition of Ty expression on glucose-containing media.

cDNA that is made by the RT can be incorporated into the genome via one of two mechanisms; 1) Integration into the genome by the Integrase (Intp), creating an extra copy of a gene or 2) Homologous recombination, replacing an existing copy. To distinguish between the two possibilities, a second plasmid (pGal-Ty*Int*) was constructed in the lab by Ernest Mordret who cloned a linker sequence into the *INT* gene of the Ty causing it to be non-functional (the linker sequence is based on [12]) (3b). In addition, a pGal-control plasmid was made by deleting the full Ty element using RF cloning (3c). This plasmid was used as a control strain for both the evolution and the VLP identification experiments. For more information about the strains, plasmids and cloning see *Materials and Methods*.

In order to validate that there is no Ty expression from either plasmids on glucose, and no Ty expression from the Ty-control strain on either glucose or galactose, an RT-PCR analysis on total RNA from cells hrboring one of the 3 plasmids was done using primers specific to Ty1. Figure 4 shows that no expression of Ty can be detected in any of the three strains on glucuse, and no Ty-expression is detected in the pGal-control strain on either glucose or galactose. As expected, both pGal-Ty and pGal-Ty*Int* express Ty on galactose.

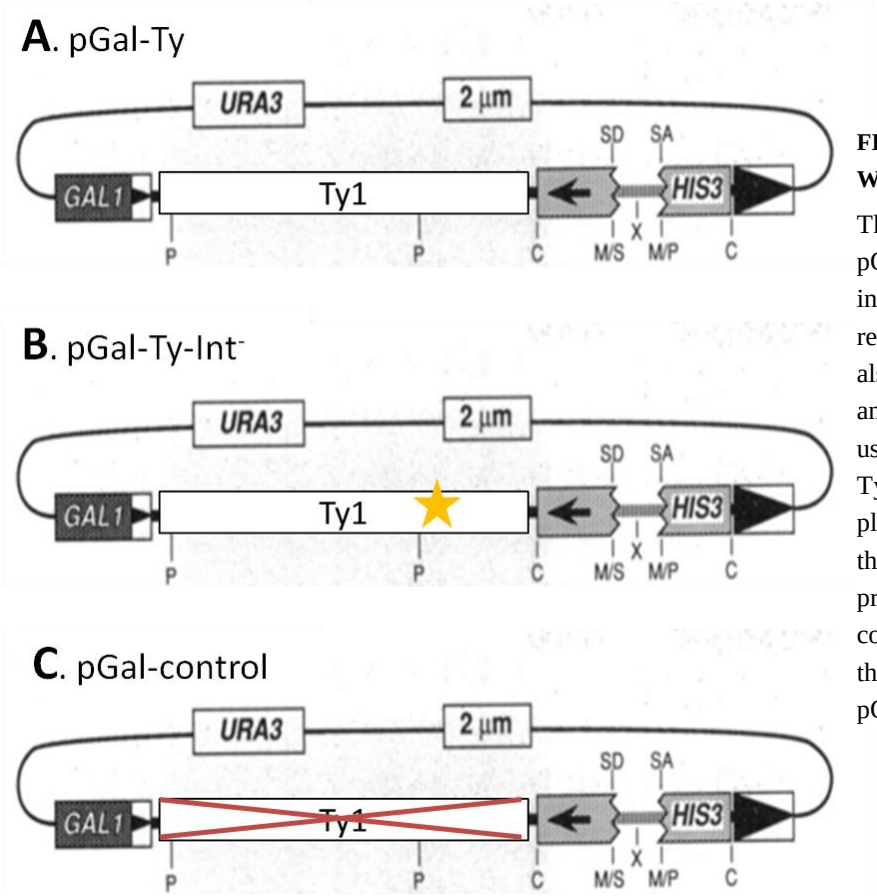

FIGURE 3 . THREE PLASMIDS' SYSTEM WAS USED IN THIS PROJECT

The three plasmids are all based on A) pGal-Ty plasmid, a 15kb, 2µm plasmid that includes the Ty element under the regulation of Gal promoter. The plasmid also contains a *URA3* gene for selection and a *HIS3* gene with an artificial intron to use as a marker for integration. B) pGal-Ty*Int* plasmid, identical to the pGal-Ty plasmid with an additional linker sequence that was cloned into the Integrase gene preventing the Integrase activity. C) pGal-control plasmid, contains a full deletion of the Ty element by RF cloning from the pGal-Ty plasmid, serves as control.

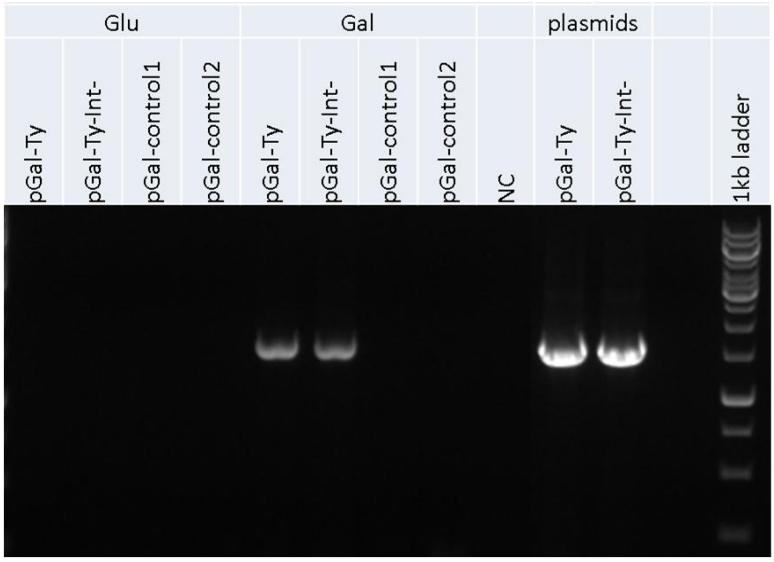

FIGURE 4 . PGAL-TY AND PGAL-TY*INT* EXPRESS TY ON GALACTOSE-CONATINING MEDIA, WHILE THE PGAL-CONTROL DOESN'T EXPRESS TY AT ALL

Total RNA was extracted from all three strains (two colonies were picked after cloning from the control strain, and the RNA extraction was done on both of them. pGal-control1 was used later on as the pGal-control strain) and was reverse transcribes *in vitro* to form cDNA. cDNA was amplified by PCR using primers specific to Ty. PCR product was ran on agarose gel (1%) to verify expression. The plasmids themselves were used as template for a positive control of the PCR reaction. For negative control DDW was used instead of a template. There is an empty lane between the plasmids lanes and the ladder. All strains do not show bands on the Glu-containing media, while the pGal-Ty and pGal-Ty*Int* shows a band on the Galactose-containing media.

2.2 VLP isolation and Identification

To find out, e.g. by deep sequencing, which mRNAs and cDNAs are encapsulated in the VLPs, the VLPs have to be extracted and separated from all cellular organelles, mRNA and DNA. A VLP-extraction protocol was adapted from Eichinger and Boeke [8], on yeast strain RM11-a that harbors either the pGal-Ty plasmid or the pGal-control plasmid.

Cells were grown under conditions that induce RT activity (i.e. GAL-containing media, 22°c); the cells were lysed, briefly centrifuged and fractionated on a sucrose step gradient. Two biological repeats were done on RM-11a cells harboring the pGal-Ty plasmid; one repeat was done on cells harboring the pGal-control plasmid. The sucrose step gradient was manually fractionated into ~35-40 fractions (~1ml per fraction) (see *Materials and Methods* for more information).

We describe below a set of indications and measures we've established in the lab to allow detecting the gradient fraction(s) which likely contain the VLP.

2.2.1 Analysis of the pGal-Ty Gradient

First, we measured OD260 to find where the majority of nucleic acids reside. This was done to evaluate the integrity of the step gradient since it is known that the free mRNAs are found in low sucrose concentration and the ribosome forms a peak between the 30% and 70% sucrose. Figures 5-6 (blue line) shows the expected results, and resembles the results of Eichinger and Boeke's gradient [8] that fraction 27 shows maximal levels of OD260.

Since the VLPs contain the mRNA and the proteins of the Ty, three assays were done in order to identify the VLP-containing fractions and to make sure that those fractions contain mature and active VLPs; qRT-PCR in order to find the Ty's mRNA-containing fractions, western blotting in order to find the gag protein-containing fractions and a reverse transcriptase activity (RT activity) assay to find the active VLP-containing fractions (figure 5, green line, bottom panel and purple line respectively).

In addition, the sucrose concentration of each of the fractions was determined using a refractometer in order to verify that the gradient did not deteriorate during the experiment (figure 5, cyan line). It can be seen that the sucrose concentrations form a gradient; the sucrose concentration in the beginning of the gradient is ~20-30% sucrose and in the last fractions is ~65% sucrose, as expected from a 20-30-70% sucrose step-gradient after centrifugation and some dilution.

Reassuringly, as figure 5 shows we found that the three methods peak at the same fractions, the last 4-5 fractions of the gradient, which are made of ~65% sucrose. It can be seen from the above analysis that

using the sucrose step gradient we've managed to identify the VLP-containing fractions, and show that they are not where the majority of nucleic acids reside (the OD260 peak).

In addition to finding the VLP-containing fractions, we wanted to find the ribosomes-containing fractions in the gradient and verify that the VLP fractions are not "contaminated" with ribosome. This is an essential condition since we plan to follow with RNA-seq that will be done on the VLP rich fractions. The sequencing will give us all the mRNA found in these fractions. If there are ribosome "contaminants" in the VLP-containing fractions, the RNA-seq will catch both the mRNA encapsulated in the VLPs as well as mRNAs that are being translated by the ribosomes. In order to address this question, we performed qRT-PCR using primers for the 18S rRNA, and a western blot analysis using an anti-ribosome antibody (anti-RPL1) kindly provided by François Lacroute (figure 5, red line and bottom panel respectively).

While the ribosomal qRT-PCR doesn't show any peak (meaning that the rRNA is found equally in all fractions) the western analysis shows that the ribosomes are found in most fractions, probably due to polyribosomes of various degrees of ribosome density per mRNA, but are depleted from the VLP-containing fractions thus no contamination of translated mRNA is expected in the sequencing of the VLP.

In conclusion we are assured that we found the VLP-containing fractions since we have two biological repeats which show the same results that were described in this part (Figure 5a and 5b).

2.2.2 Analysis of the pGal-control Gradient

The same three methods were used on the control step gradient since the corresponding VLP-containing fractions from this gradient will be used as a control for the RNA-seq experiment.

Figure 6 clearly shows that none of the methods peak in this gradient, meaning that there are indeed no VLPs in this gradient (green line, purple line, bottom panel).

The absence of the VLPs in this gradient is not due to technical issues since the gradient shows the correct sucrose concentration that resembles the initial concentration in the sucrose step gradient (figure 6, cyan line) and since the anti-RPL1 western shows the presence of ribosomes in most fraction above 20% (bottom panel) as in the VLP-gradient.

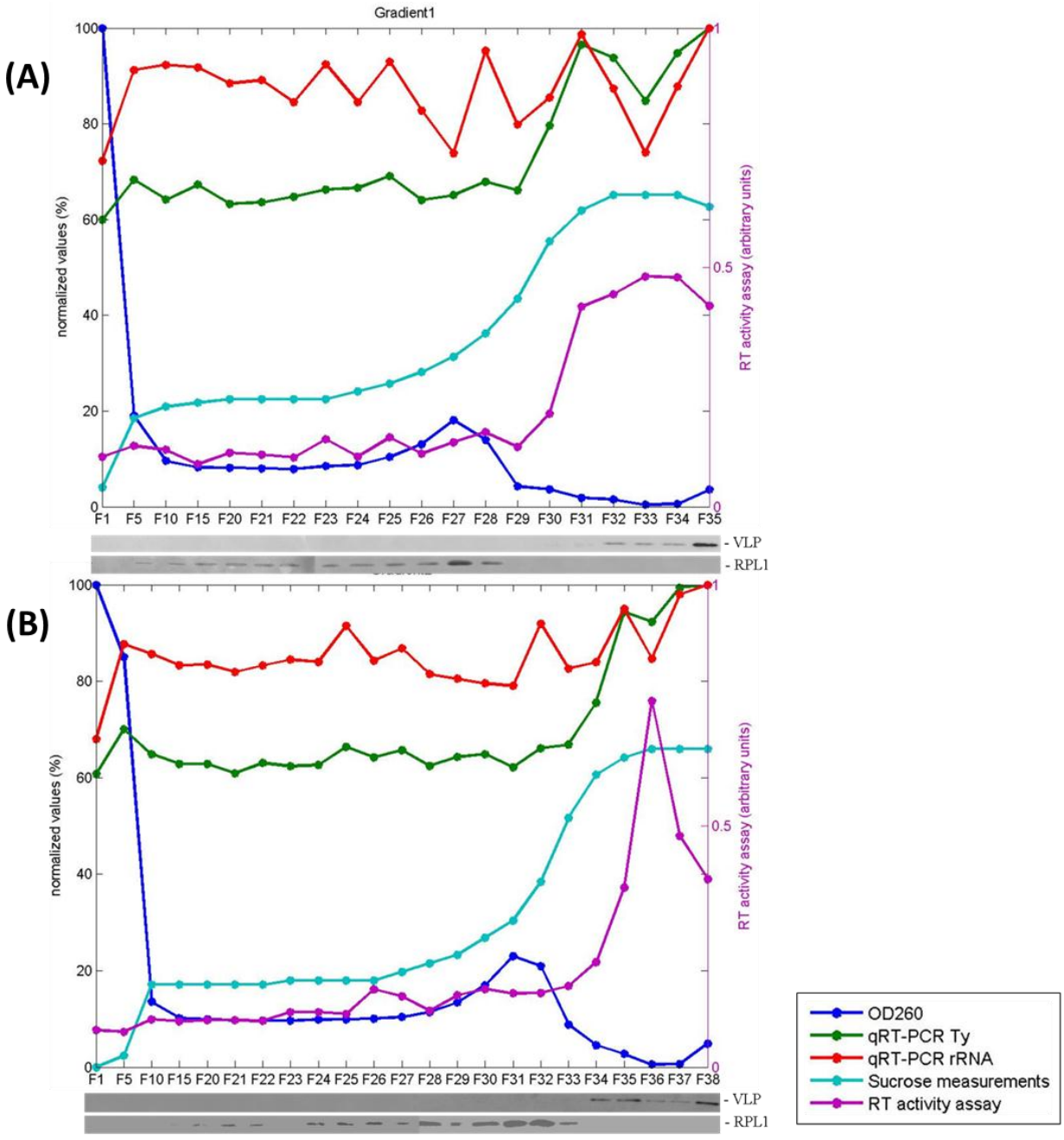

FIGURE5. VLPS ARE FOUND IN ~60% SUCROSE FRACTIONS OF THE PGAL-TY STRAIN, WHILE RIBOSOMES ARE DEPLETED FROM THESE FRACTIONS.

A sucrose gradient was made to isolate VLPs. The gradient was manually fractionated to (A) 34 fractions or (B) 38 fractions,. Multiple assays were done on each of the fractions to identify VLPs containing fractions, as well as ribosomes containing fractions. OD 260 measurements (blue line) were done on fractions; qRT-PCR analysis using both Ty primers and ribosomes primers (green and red lines, respectively) were done on total RNA extracted from the fractions; each of those assays results were normalized. RT-activity assay was also done on fractions (purple line). Western analysis using both anti-VLPs and anti-rbosomes were done on fractions (bottom panel). The sucrose percentage in the fractions was determined using a refractometer (cyan line).

(A) First repeat, (B) second repeat

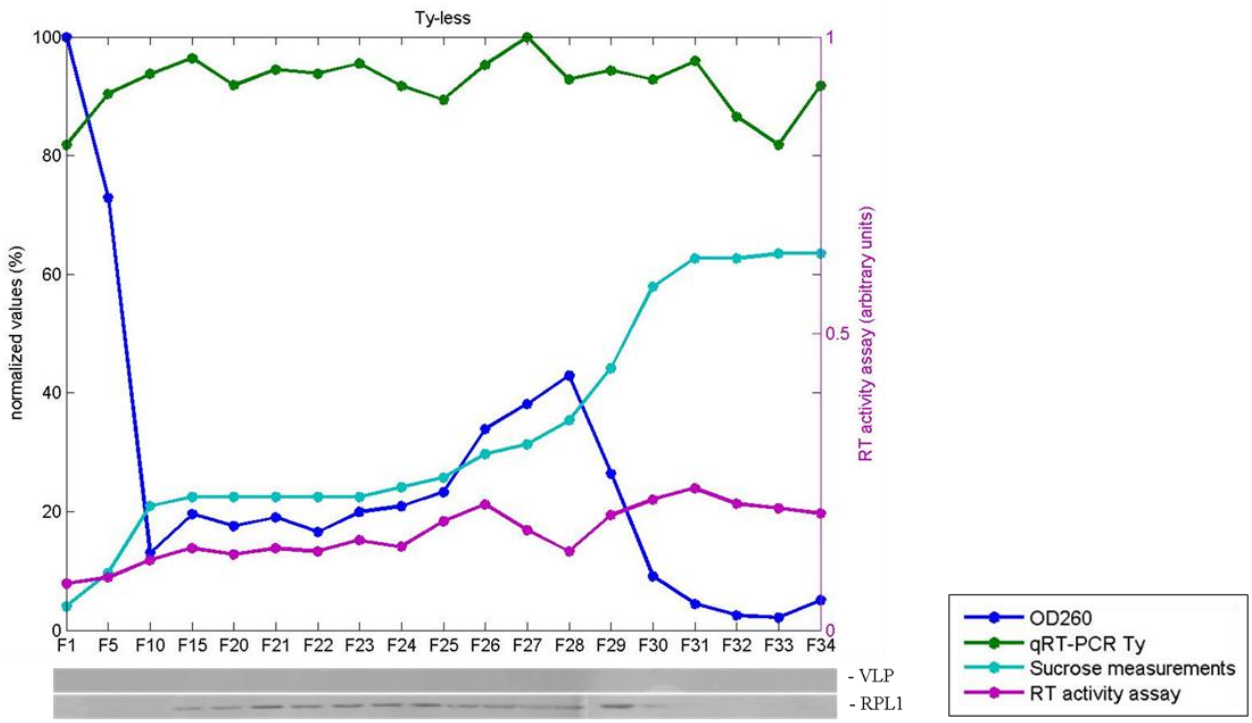

FIGURE6 . NO VLPS ARE FOUND IN THE PGAL-CONTROL STRAIN FRACTIONS

Multiple assays were done on each of the fractions to identify VLPs containing fractions, as well as ribosomes containing fractions. OD 260 measurements (blue line) were done on fractions; qRT-PCR analysis using Ty primers (green line, respectively) were done on mRNA extracted from the fractions using Trizol reagent; each of those assays results were normalized. RT-activity assay was also done on fractions (purple line). Western analysis using both anti-VLPs and anti-rbosomes were done on fractions (bottom panel). The sucrose percentage in the fractions was determined using a refractometer (cyan line).

2.2 Lab Evolution in vivo

The aim of this chapter was to examine the effect of the presence of an active Ty-element on the evolutionary dynamics of yeast cells in the lab. Cells evolving in the presence of an evolvability means, such as the RT machinery (or other agents such as a competence system in bacteria that allows HGT, or a means to obtain TAM) can show modified rate of adaptation and/or different solutions to the same challenge. In addition, the evolvability means can cause faster adaptation by finding a general solution that will help the cells in a variety of environments, thus growing the cells in a specific media will increase their fitness in other environments as well.

Therefore, a lab evolution experiment was conducted, using 3 different cultures of RM-11a each harboring one of the three plasmids discussed in the *Experimental design* part (pGal-Ty, pGal-Ty*Int* or pGal-control, referred to as the Int^+ strain, Int^- strain and the control strain, respectively). We have used the Int^+ strain and the Int^- strain in order to distinguish between the effects of high mutation rate induced by the RT and RT-mediated gene duplication events.

Lab evolution was done according to the established methodology of evolution *in vivo* as described in [13]. In this procedure cells were grown under certain conditions (in this project, liquid S-Gal-Ura media to avoid plasmid loss and 22˚c and Galactose to allow Ty activity) and are diluted regularly into fresh media. Dilution in this project was not constant per day; instead, cells were counted daily and diluted as they reached $5 * 10^7$ cells to avoid population collapse. Each strain was evolved in three biological repeats; cell were grown for 230 generations and frozen every 28 generations. The conditions were chosen to induce Ty activity, but as can be seen in figure 7a, the conditions are not the ideal conditions to yeast growth leading to a long Lag phase (48h) of the ancestral strains.

Fitness evaluation and the assays for plasmid presence were done on all repeats in four time points (generations 0, 140, 196 and 225) to evaluate which of the strains has adapted the most and to shed light on a possible mechanism, respectively.

2.2.1 Fitness evaluation by growth experiment

In order to assess RT effects on evolution we wanted to evaluate and compare the fitness of the strains. One way of assessing fitness during evolution experiments is to compare the growth of the evolved strains to the strain's ancestor. Therefore, growth assays were done for all strains on four evolutionary time points. Cells from generations 0 (ancestors), 140, 196 and 225 were grown for ~48 hours on S-Gal-Ura media, in 22˚c. The cells were then diluted (1:50) into fresh media while their growth was

monitored. Fitness of the strains was evaluated based on the yield (i.e. the cell density measured by OD600 in stationary phase) and the growth rate given by the growth experiment.

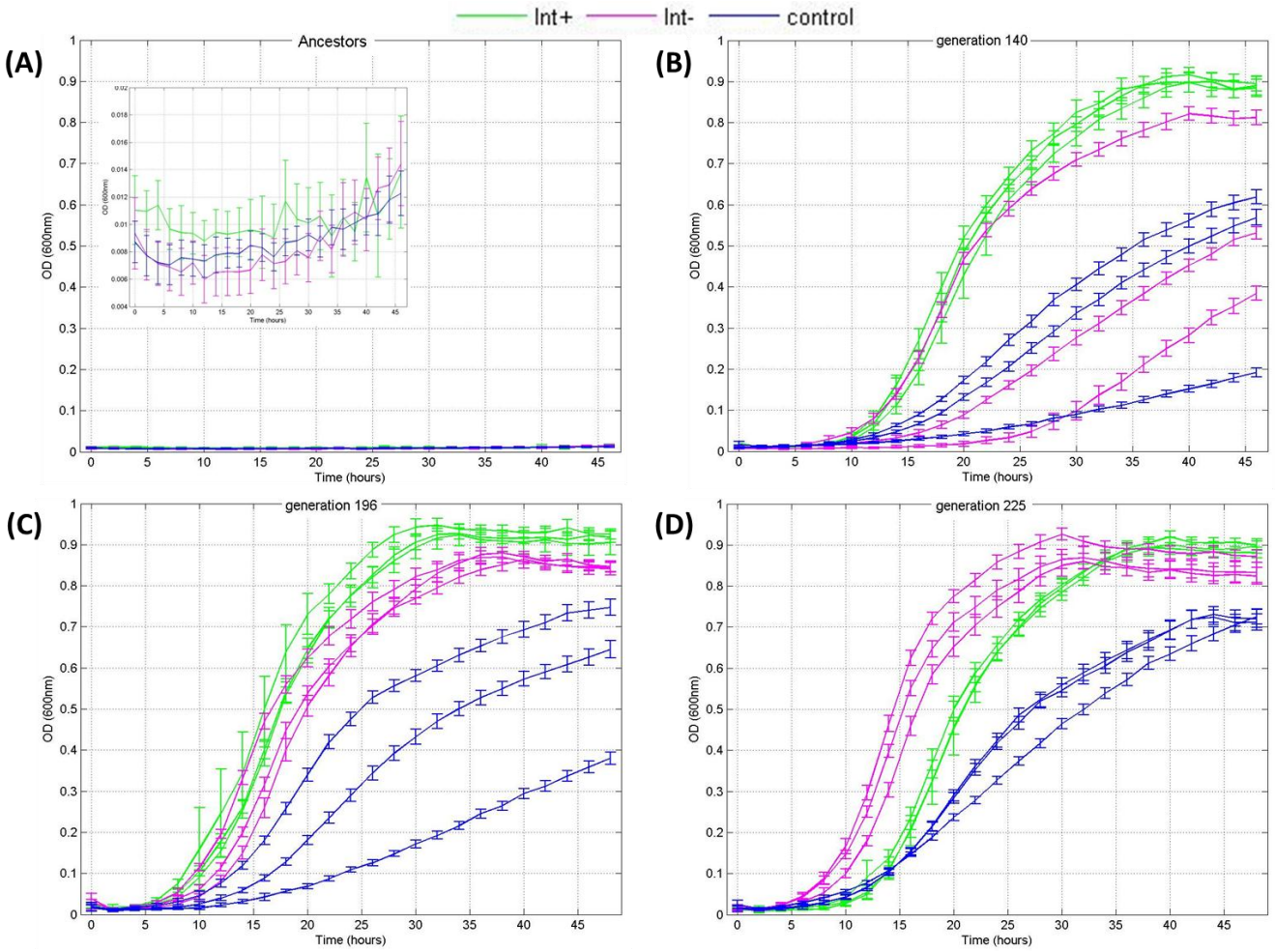

FIGURE 7. LAB EVOLUTION EXPERIMENT SHOWS RAPID EVOLUTION OF THE PGAL-TY STRAIN

Growth experiment was done in four evolutionary time points (A) ancestor (inset- zoom in), (B) generation 140, (C) generation 196 and (D) generation 225.

The pGal-Ty strain had adapted rapidly while the pGal-Ty*Int* strain had adapted slower, and the pGal-control strain had adapted the least, showing almost no improvement from generation 140 to 225.

Green – Int^+ strain; Magenta – Int^- strain; Blue – control strain.

As seen in figure 7, we observed a striking difference in the evolution dynamics of the different strains. All 3 repeats of the Int+ strain has adapted rapidly, reaching stationary phase within the 48 hours experiment on generation 140 and keeps having similar growth curves as in 140 on generations 196 and 225; it also has the highest yield of all strains (figure 7, green line). The control strain's fitness is the lowest in all time points not reaching stationary phase during the course of the growth experiment (45

hrs) on generations 140 and 196. On generation 225 the control strain reaches stationary within the course of the experiment, but having the lowest yield (figure 7, blue lines). Two out of three repeats of the Int^- strain have similar growth curves to the control strain on generation 140, but the fitness improved in generation 196 resembles the Int^+'s fitness (figure 7, magenta lines).

The growth curves were further analyzed by a software developed by Yoav Ram from Lilach Hadany's lab[14]. The software fits a mathematical model to the growth curve and determines several parameters such as Lag phase duration, maximal growth rate, yield and others. The lag phase duration in this experiment is not precise thus only the calculated maximal growth rate and yield to are used to evaluate cells' fitness.

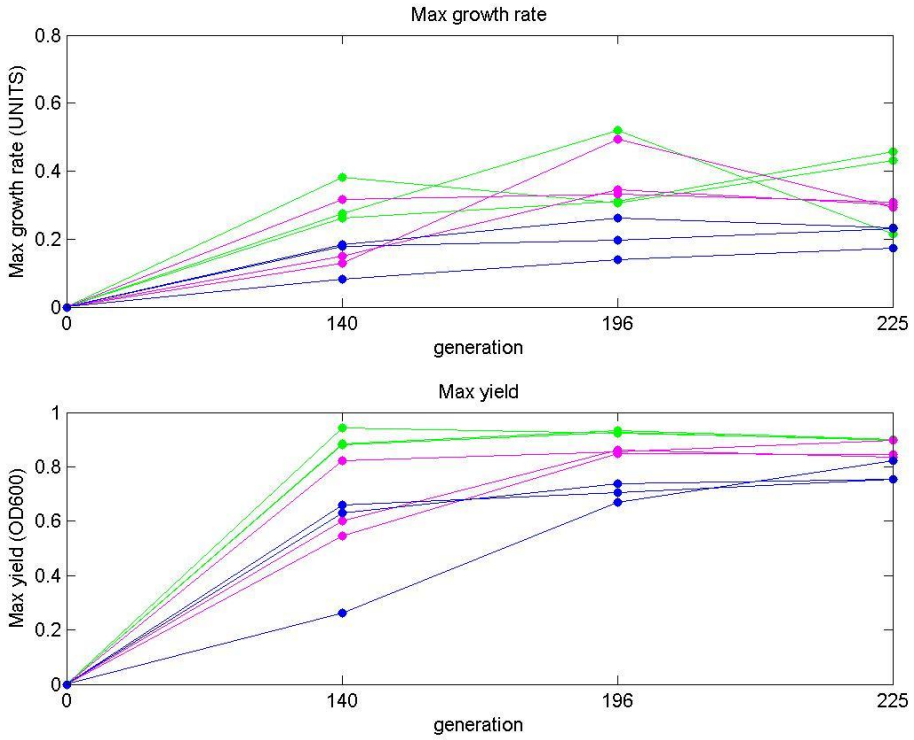

FIGURE 8. CURVEBALL ANALYSIS OF GROWTH EXPERIMENT

The Growth Experiment were analyzed using Curveball software that extracted parameters such as A) maximal Growth Rate and B) Maximal Yield. A) The maximal growth rate of pGal-Ty and pGal-Ty*Int* looks roghly the same, while the pGal-control growth rate is the lowest throughout the experiment. B) The maximal yield divide the repeats into three groups, the pGal-Ty with the highest yield, the pGal-Ty*Int* with an intermediate yield and the pGal-control with the lowest yield

Figure 8 shows the parameter extracted by the software. The Int^+ strain has the highest growth rate and yield in all evolutionary time points (figure 8a and 8b, respectively, green lines), except for the ancestor.

Taken together with the growth dynamics shown in figure 7, the Int^+ strain has adapted the most, reaching the highest fitness within the shortest amount of time. Analysis of the control strain shows that it has the lowest growth rate (figure 8a, blue line) and lowest yield (figure 8b, blue line) throughout the evolution, making it the least adapted strain, having the lowest fitness in all time points. As for the Int^- strain, The growth rate and yield shows that it had resembled the control strain on generation 140, but reaching the Int^+ fitness on generation 196, 225 (figure 8, magenta lines); thus the Int^- strain is the intermediate adapted strain, reaching the same fitness as the Int^+ strain but in longer time period. Interestingly, one of the Int^- strain's repeats resembles the Int^+ strain's growth curve, maximal growth rate and maximal yield on generation 140 (figure 7, asterisk). The resemblance in fitness of these repeat throughout the evolution may indicate a similar mechanism of adaptation for the asterisk label repeat and the Int^+ strain.

2.2.2 Mechanism of adaptation

The plasmid used in this project is a 15kb plasmid, with a high copy number. We therefore thought that carrying it might constitute a burden on the cells. Indeed, growth experiments were done to compare the fitness of the RM-11a strain harboring one of the pGal-Ty or the pGal-Ty*Int* plasmids and the fitness of the RM-11a strain harboring a simpler, smaller plasmid (pYes) (data not shown). The growth experiments revealed tremendous fitness decreased in the pGal-Ty and pGal-Ty*Int* strains in comparison to the pYes plasmid, indicating that the plasmid presents a significant burden on the cells.

Therefore, one hypothesis for the rapid adaptation of the pGal-Ty strain and of the pGal-Ty*Int* was that the cells managed to lose the plasmid while being able to grow on SC-Ura by integrating the *URA3* gene into the genome. To address this hypothesis we have tested each of the repeats in all of the same evolutionary time points as the growth rates, for the existence of the plasmid. Cells were grown and patched on non-selective media plate that enables rapid loss of the plasmid; cells were then replicated on selective media (SC-Ura) plate (examples for the rich media plates and the SC-Ura replicate plates can be seen in figure 9). Cells that have integrated the *URA3* gene into their genome will be able to grow on selective media plate after growing on non-selective media, since the *URA3* gene is an integral part of their genome and therefore cannot be lost. Percentages of colonies grown on selective media out of all colonies (represents percentage of colonies that have integrated the *URA3* into their genome and lost the plasmid) are shown in table 1. All ancestral strains didn't grow on the selective media, validating that all strains have started with an *ura* genotype and having the *URA3* gene on the plasmid. In addition, it

shows that the Int^+ strain, having the highest fitness according to the growth experiments, has lost the plasmid prior to generation 140, having 100% colonies growing on the selective media from this time point on. The control strain, which has the lowest fitness according to growth experiment, did not grow on selective media at all time points (or grew in very small percentages), indicating that this strain has kept the plasmid throughout the lab evolution. The Int^- strain shows an interesting trend; on generation 140, the 2 evolution lines with the low fitness didn't grow on selective media since they lost the plasmid while the one repeat with the high fitness (figure 7b, asterisk) did. From generation 196 on, all Int^-'s repeats grew on selective media. In summary, the results indicate that high fitness corresponds to the loss of the plasmid, while low fitness corresponds to plasmid existence, as hypothesized.

The fact that the Int^+ and Int^- strains have lost the plasmid, have lead us to think that the full plasmid (including the URA3 gene) was integrated into the genome. Therefore, a PCR was done on all repetition of all strains of generation 225. Cells were taken from the patches and lysed in 20mM NaOH. The lysate was used to perform PCR using primers against the backbone of the plasmid. Figure 10 shows that although the patch experiment revealed that the Int^+ and Int^- strains have lost the plasmid, the backbone of the plasmid still exist in the cells, meaning that the full plasmid was integrated into the genome. The control strain, which lost the ability of growing on SC-Ura after growing on YPD shows no band since the plasmid is not integrated into the genome.

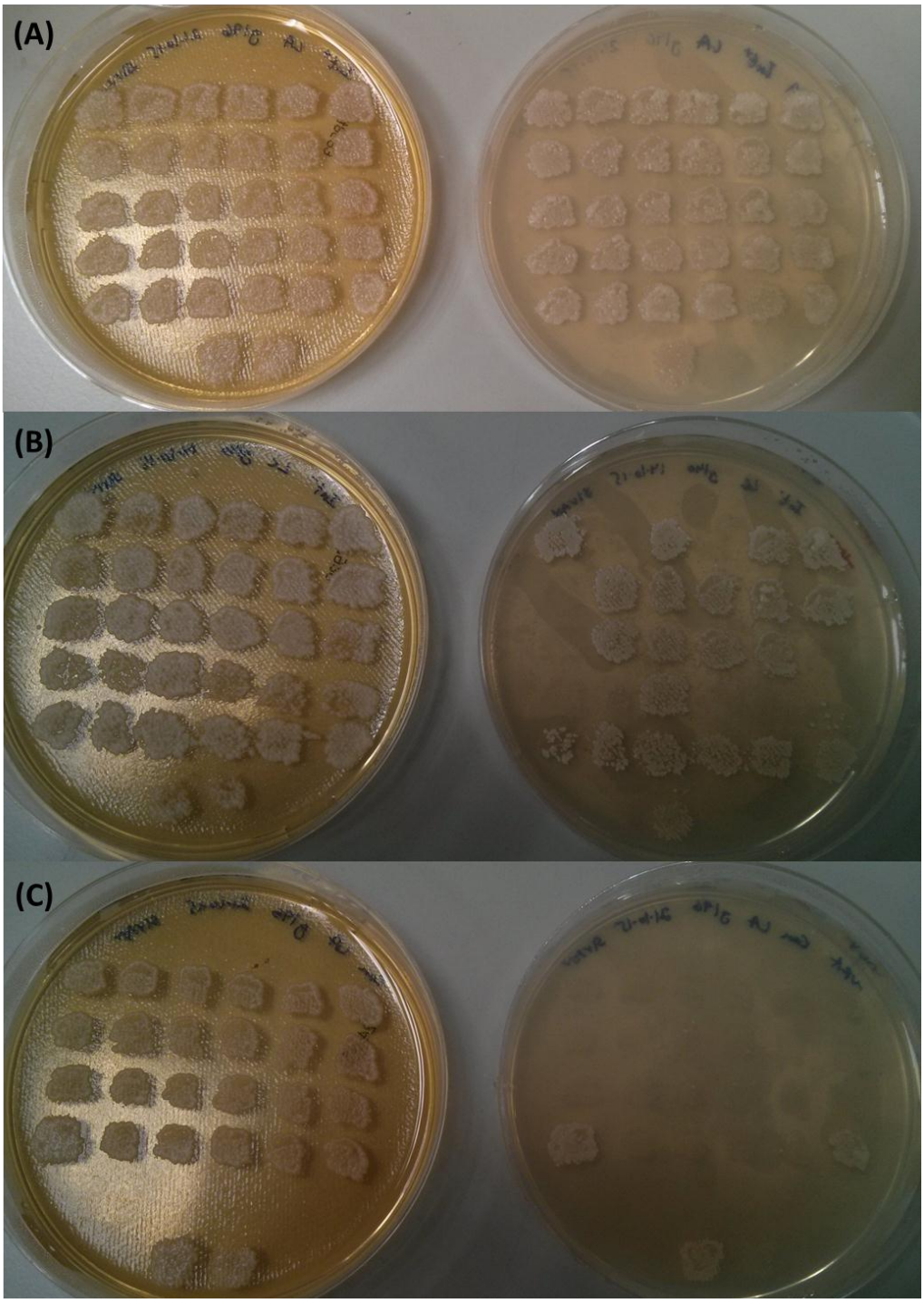

FIGURE9 . PATCH AND REPLICATE EXPERIMENT SHOWS THAT FEW OF THE STRAIN HAVE LOST THE PLASMID DURING THE EVOLUTION

Yeast single colonies (~30) were patched on YPD plates and replicate to a SC-Ura plates. Percentage of patches that grew on SC-URA represent percentage of cells that have lost the plasmid. (A) YPD plate (left) and the replicate SC-Ura plate (right) of an Int+ strain (B) YPD plate (left) and a SC-URA plate of an Int- strain and (C) YPD plate (left) and the replicate SC-Ura plate (right) of a control strain.

TABLE1 . NUMBER AND PERCENTAGE OF CELLS THAT HAVE LOST THE PLASMID DURING THE EVOLUTION IN ALL STRAINS AND ALL TIME POINTS

generation	Strain	repetition	#colonies Total	#colonies on Ura-	%colonies w/o plasmid
Ancestor	Int+	-	24	0	0.0
	Int-		24	0	0.0
	control		24	0	0.0
140	Int+	1	30	30	100.0
		2	30	30	100.0
		3	30	30	100.0
	Int-	1	30	1	3.3
		2	30	20	66.7
		3	30	3	10.0
	control	1	30	0	0.0
		2	30	0	0.0
		3	30	0	0.0
196	Int+	1	30	30	100.0
		2	30	30	100.0
		3	30	30	100.0
	Int-	1	24	22	91.7
		2	24	24	100.0
		3	24	24	100.0
	control	1	24	2	8.3
		2	24	1	4.2
		3	24	0	0.0
225	Int+	1	24	24	100.0
		2	20	20	100.0
		3	24	24	100.0
	Int-	1	30	30	100.0
		2	30	30	100.0
		3	24	24	100.0
	control	1	24	1	4.2
		2	24	0	0.0
		3	24	0	0.0

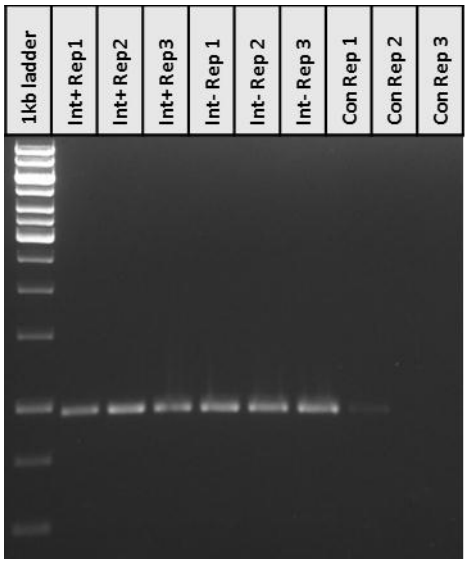

FIGURE10 . THE INT+ AND INT- STRAINS HAVE INTEGRATED THE PLASMID INTO THE GENOME, WHILE THE CONTROL STRAIN HAS NOT

PCR was done on patches from the replicate experiment for each of the repetition using primers against the backbone of the plasmid. PCR product was run on agarose gel. The Int+ and Int- strain contains a genomic copy of the plasmid, while the control strain does not.

2.3 Evolution in silico

In addition to the lab evolution experiment, a simulation was written, simulating and comparing evolution with and without reverse transcription. The simulation is based on a genetic algorithm (GA) which generates solutions to optimization problems using the natural selection mechanism [15]. GAs contains some basics biological concepts such as DNA mutations, fitness evaluation and selection. The fitness in the simulation is determined by the NK model, a mathematical model that generates a score for each sequence based on the length of the sequence and the epistasis between different positions in the sequence [16]. In here, the fitness is determined based on the genes' sequence in the transcriptome. The selection in the simulation is randomly picking organisms for reproduction, with probability of selection being weighted for each individual according to their fitness: high fitness is linearly proportional with probability of being chosen to the next generation. While in classical GA, DNA mutation are the only source of variation, this simulation has an additional couple of steps: RNA mutations and reverse transcription (figure 11). The RNA mutation's steps differ from the DNA mutation step in two main concepts; 1) the rate of mutation in the RNA mutation steps is ~10 times higher, 2) the RNA mutations will not be inherited to the next generation (unless an RT event will occur as well). The RT step randomly choses an RNA molecule from the transcriptome and integrates it into the genome. The integration part currently supports integration via homologues recombination only (i.e. not insertion of the "cDNA" as an additional gene allowed). Using the simulation we can assess the dynamics of evolution processes with and without RT in variety of conditions. For more details on the simulation structure and parameters, see *Materials and Methods* and figure 11.

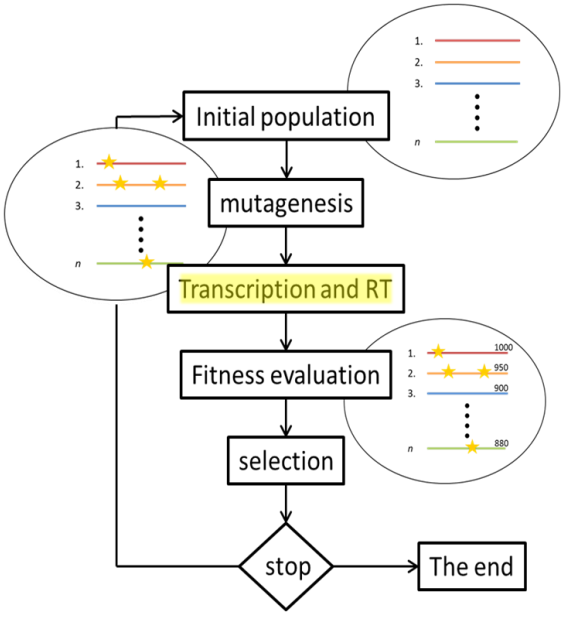

FIGURE 11. FLOW CHART OF GENETICS ALGORITHMS (GA) INCLUDING THE RNA MUTATION AND RT STEPS

Schematic representation of the different steps of GA. The simulation begins with an initial population then the population is going through mutagenesis, fitness evaluation and selection to form a new initial population for the next step. In this simulation there are additional steps of RNA mutagenesis and RT (labeled in yellow).

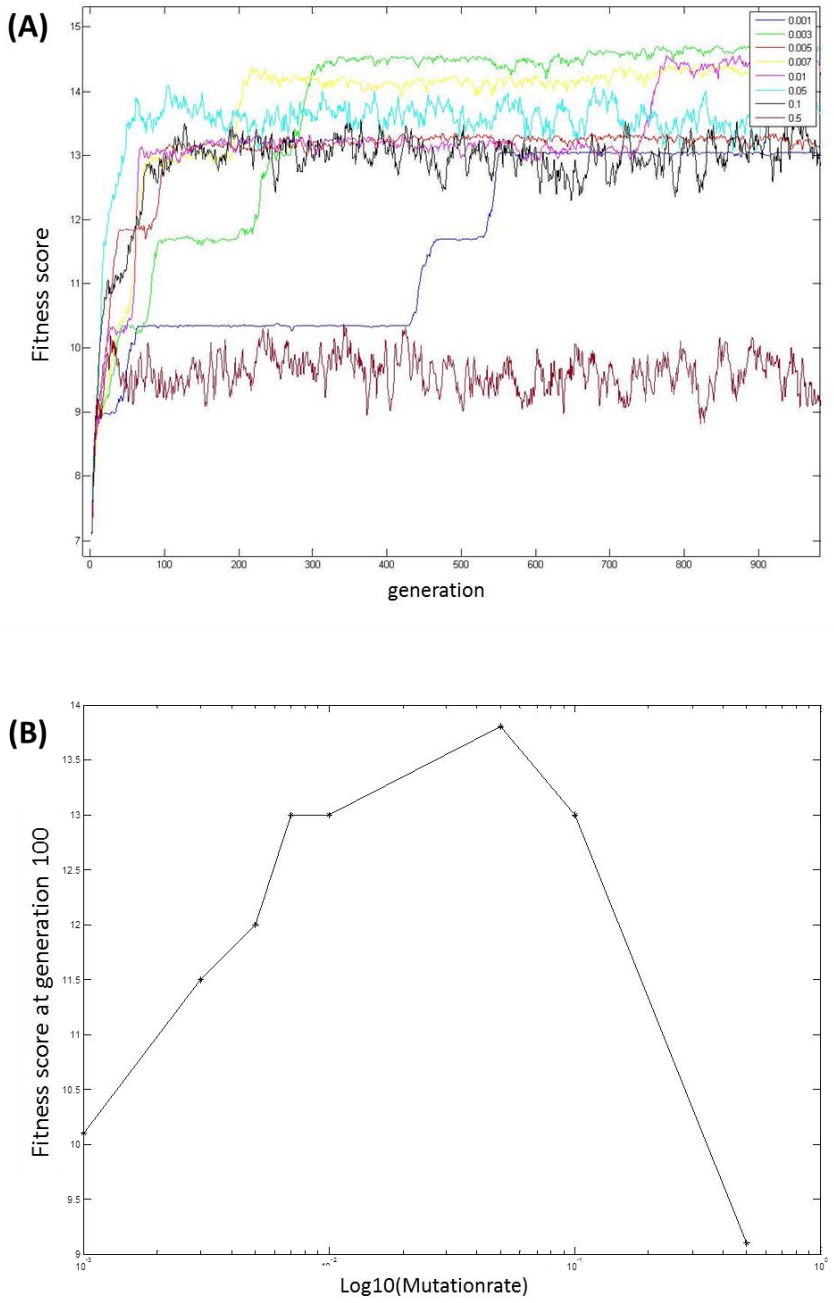

FIGURE 12. SIMULATION RESULTS FOR DIFFERENT MUTATION RATE SHOW RAPID ADAPTATION IN HIGH MUTATION RATE

(A) Eight mutation rates were evaluated in the simulation. (B) the fitness score on generation 100 was taken from each of the mutation rates plots and plotted as a function of the mutation rate.

High mutation rate are correlated with reaching to the highest fitness faster, up to a threshold where the population crushes ((A), brown and black lines, and (B)). The evolution process is stairs shaped.

2.3.1 Evaluating the Simulation

In order to test the simulation, we wanted to check how different mutation rates affect the rate of fitness improvement; thus, the simulation was ran on initial population of 100 individuals, each with a 2 genes' genome (30nt each gene) for 1000 generations with different mutation rates (figure 12a) (each of the lines in the figure is the average fitness values of 50 runs). The fitness values from generation 100 were plotted against the mutation rate in figure 12b showing clearly that as the mutation rate increases, the population usually reaches the maximal fitness faster. This trend is happening up to a point where the mutation rate is too high and the population collapses (figure 12a, brown line and 12b). The concluded optimal mutation rate is 0.05 mutations per nt which is 1.5 mutations per genome. Another interesting aspect of evolutionary dynamics using the simulation can be seen in figure 12a, where we see that the population fitness increases in a step-like shape; the population's fitness reaches a plateau in a local maximum and then increases to the next local maximum and so on until it reaches the global maximum (the maximal can be calculated since the NK model's parameter are known. In this case the maximal fitness is 15). This kind of evolution dynamics was reported from a long lab evolution experiment *in vivo* dome in Lenski's lab [17].

2.3.2 Assessing Evolution with and without RT in different Evolutionary landscapes

The NK model enables to generate different evolutionary landscapes, depending on the N and K values, and on the fitness that each of the positions contributes to the total fitness (the "rule", see *Materials and Methods*). Evolutionary landscapes were initially described by Wright [18] which suggested to visualize the relationship between genotypes and reproductive success. It is assumed that every genotype has a well-defined fitness. This fitness is the "height" of the landscape. Genotypes which are very similar are said to be "close" to each other, while those that are very different are "far" from each other. The set of all possible genotypes, their degree of similarity, and their related fitness values is then called a fitness landscape. The landscapes may have multiple local maximums (the landscape is then referred to as "rugged") or few ("smooth").

We thought that RT can affect evolution differently in rugged or smooth landscapes since the RT can mediate exploration of the fitness landscape, without affecting genes that were adapted in other in landscapes which can be more helpful in rugged landscapes that present a harder problem to the algorithm. On the other hand, in smooth landscapes, the solution is much easier to achieved, thus the increase mutation rate of expressed regions might be less needed and even detrimental.

We have therefor generated three landscapes (3 “rules” in the same NK parameters, N = 10 and K = 1), that were assessed using a greedy algorithm to be rugged, intermediate, and smooth, i.e. hard, intermediate and easy to optimize (Data not shown) landscapes. The simulation simulated 50 repeats on the evolution processes in each landscape, in each condition (with and without RT). The effective mutation rate (i.e. the DNA mutagenesis plus the RT-mediated mutagenesis) were set to 1/genome length (calculated to be the optimal mutation rate using figure 12). Environmental condition change occurred every 100 generation causing differential gene expression in the simulation.

The results are shown in figure 13. The insets show the average fitness of the 50 runs over 5000 generations. The red line represents the strain that has no RT capabilities (w/o RT), while the blue line represent the strain that has RT capabilities (w/ RT).

The evolution process was analyzed such that in each generation the no-RT strain’s fitness was divided by the with RT strain’s fitness (blue line, figure 13a,13b,13c). A with RT strain equal to no RT strain line was plotted for convenience in black (figure 13a,13b,13c). For each generation a Wilcoxon test was performed to check if the difference between the two strains is significant at a significance level of 0.05, the significantly different generations are labeled in red over the black line. The same analysis was done in all three landscapes, rugged (figure 13a), intermediate (13b) and smooth (13c).

The results suggest that the RT is beneficial in rugged landscapes, but deleterious in the smooth landscapes.

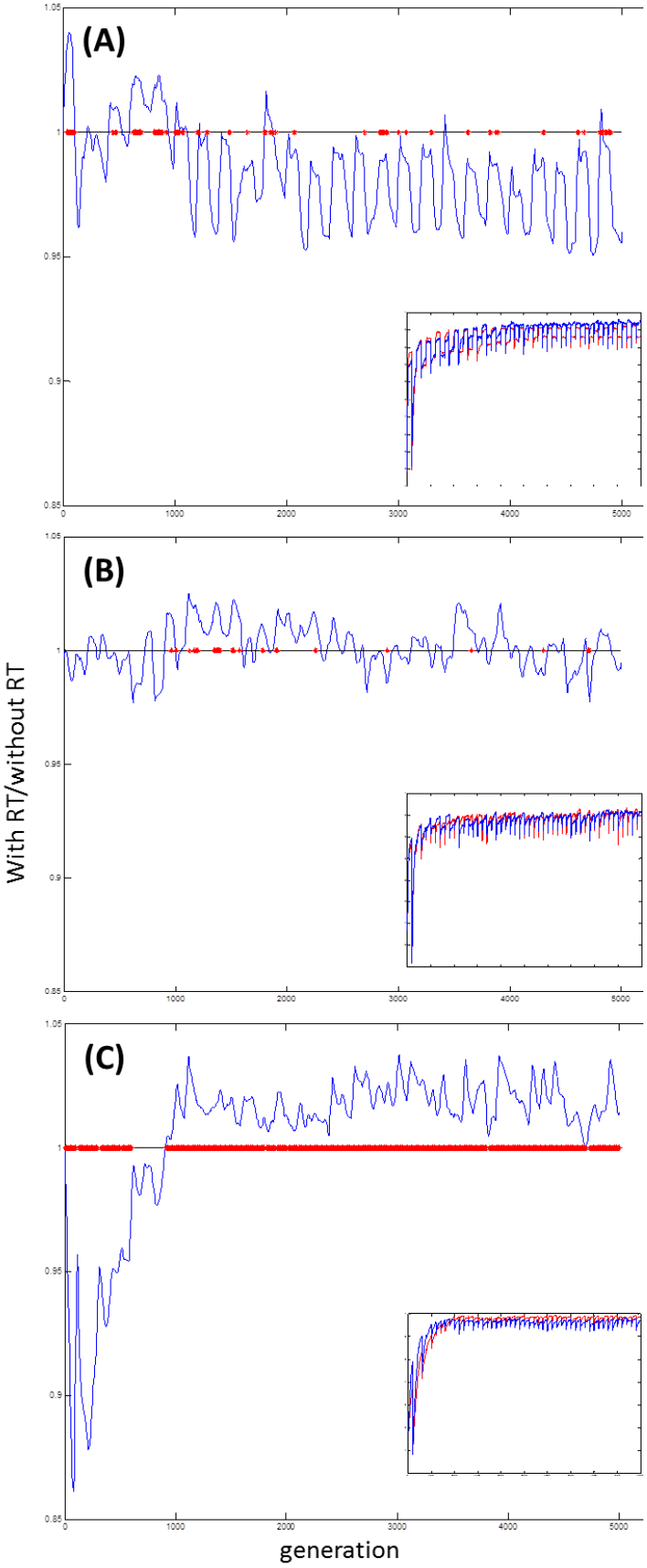

FIGURE 13. RT IS BENEFICIAL IN RUGGED LANDSCAPE AND DELETERIOUS IN SMOOTH LANDSCAPES

In each generation the fitness of the population w/o RT was divided in the fitness of the population w/ Rt (blue lines). The w/ RT equal to w/o RT line is plotted in black. Red asterisks on the line represent generations in which the difference between evolution w/o RT and evolution w/ RT are significantly difference (Wilcoxon, alpha = 0.05).The evolution process is shown in the insets (blue line – w/ RT and red line – w/o RT).

The evolution was done in three landscapes: (A) rugged, (B) intermediate and (C) smooth.

2.3.3 Competition experiment using the simulation

The simulation was expanded to support a competition experiment *in silico.* A competition experiment *in vivo* is an experiment in which two sub-populations (or more) are grown in the same tube for long periods of time until one of the sub-populations takes over the population. In this kind of experiment it is easy to determine which sub-population has higher fitness than the other, based on the notion that the higher fitness strain has higher probability of taking over the population. Equation 1 shows a formula that is a special case of a general formula given in [19] in which the two sub-populations start from an equal amount of cells. The formula generates the probability of fixation as a function of the relative fitness of the populations (r = fitness(A)/fitness(B), where A and B are the two sub-populations). In the simulation, the competition is currently done between two sub-populations that are selected on the same step; in each generation the selection process is blind to the identity of the individuals and takes into consideration only their fitness.

EQUATION 1

$$p(fixation) = \begin{cases} \dfrac{1-\frac{1}{\sqrt{r}}}{1-\frac{1}{r}} & | \, r \neq 0,1 \\ 0 & | \, r = 0 \\ 0.5 & | \, r = 1 \end{cases}$$

In these simulation competition parameters can be also set –rates of DNA mutations, RNA mutations or RT rate of the first competitor while, currently, the competition will be executed such that the second competitor is a "WT strain", having the same parameters as the first competitor except for the parameter that the competition is held upon (for example, in order to compare strains with and without RT, the first competitor has DNA mutation, RNA mutation and RT, while the second competitor has the same mutation rates of the DNA and RNA but lacking the RT activity). Each sub population is comprised of 50 individuals. In addition the ancestors of the first competitor are identical to those of the second competitor. Since the competition is always done in reference to a sub-population with 0 in the chosen parameter, the percentage of times the first competitor won can be thought of as an approximation of

fitness of the population, meaning that higher percentage of winning is higher fitness, as can be deduced from equation 1.

Figure 14 shows three competition experiments. Figure 12 showed a phenomenon where higher mutation rate resulted in better adaptation up to an optimum. We wanted to see if the competition mode will reproduce this observation (by means of higher percentage of winning of the strain with optimal mutation rate). Thus, a competition experiment was done using two subpopulations; the first competitor was featured each time a different mutation rate, while the second population had no mutation; in each rate 100 repetitions were ran. Figure 14a shows that increasing mutation rate increase the probability of wins by the first competitor up to a point. When rate of mutation exceeds an optimum (in this case, ~1 mutation/genome length), the subpopulation with the mutations loses the competition. We wanted to find out if RNA mutations follow the same rule as DNA mutations, therefore, multiple competition experiments were done, competing a subpopulation with RNA mutations to a subpopulation that has no RNA mutations, neither of the 2 subpopulations had DNA mutations in this competition. Interestingly, figure 11b shows that RNA mutations are neutral, not beneficial in lower-middle rates (~0.5 mutation/nt), but also are not deleterious in the high rates (even 100 mutations per nt in the RNA is not deleterious). Since the fitness is determined based on the transcriptome, we hypothesized that RT can change the results of the RNA mutations, since the RT will integrate some of the RNAs into the genome, thus beneficial RNAs will not only affect the fitness of the current individual but will be passed on to the next generation while deleterious mutations in the RNA might be integrated into the genome but will probably not be selected for. Therefore, a competition using RT was done. Both subpopulations had minimal mutation rate in the DNA and RNA but the first competitor had an additional step of RT, allowing mutations in the RNA to be integrated into the genome creating a Lamarck-like evolution. Surprisingly, even the additional step of RT didn't change the neutral effect of the RNA mutations. Increasing the rate of the RT didn't change the neutral effect of the RT over the no-RT competitor (figure 14c). The results of these competition experiments indicate that at the currently examined rate of RT, on the currently examined background rate of DNA mutations the RT competitor is equal in fixation probability (and hence fitness) to the competitor that has no RT activity.

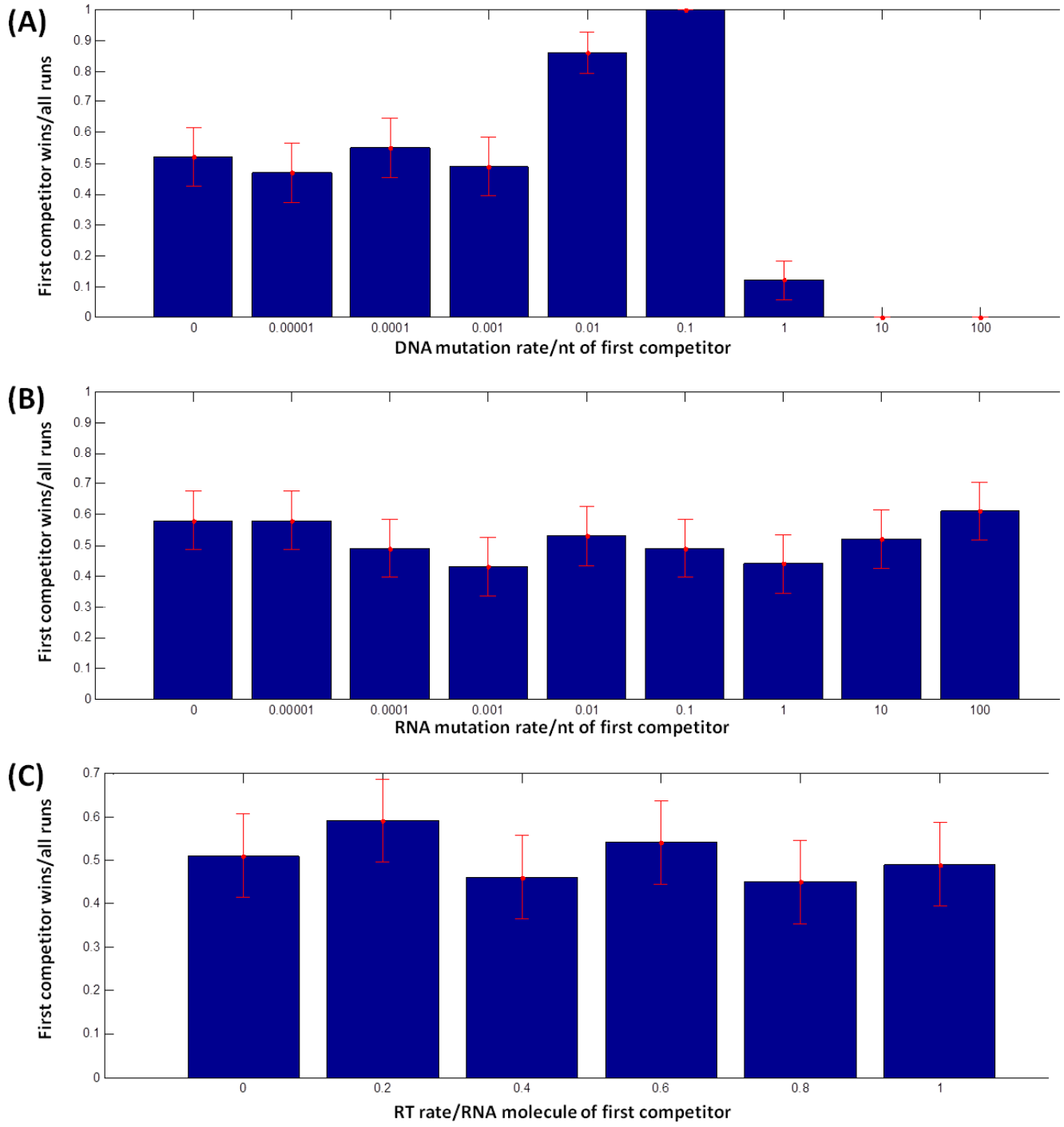

FIGURE14 . COMPETITION EXPERIMENTS DONE USING THE SIMULATION

Two subpopulation were competed one against the other, the first subpopulation had a A) DNA mutation rate, B) RNA mutation rate or C) RT rate as indicated by the x axis values. Each competition was ran 100 times. Each bar represent fraction of times that the first subpopulation won the compeition. Errorbar represent the 95% CI.

3.1 Materials and Methods for the VLP Identification and Lab Evolution

3.1.1 Strains, Plasmids, Primers and Media

Yeast strain RM11-a (*MATa leu2Δ0 ura3-Δ0 HO::kanMX*) which is a Ty-less strain was kindly sent to us by Pscale lesage's lab and was used in both the VLP identification and lab evolution experiments.
Four plasmids were used in this project;

pGal-Ty – a ~15kb plasmid that was kindly sent to us by Pascale Lesage's lab, the plasmid is described in [10], map can be seen in *figure 1a* therein. The plasmid is a 2µ plasmid and it contains a URA3 marker and a Ty element under the regulation of a Gal promoter. The plasmid also contains a *HIS3* gene, with an artificial introns enabling quantification of transposition events (not used here).

pGal-Ty*Int* - based on the pGal-Ty plasmid, with an additional linker sequence in the Integrase gene, which disrupts the integrase activity therefore preventing integration of the cDNA into the genome by the Int. This plasmid was done by Ernest Mordret in the lab.

pGal-control – based on the pGal-Ty plasmid. Using RF cloning the entire Ty element was deleted, resulting in a ~9kb plasmid. The primers used for the RF cloning are:

F-CCTGGCCCCACAAACCTTCAAATGAGAGCAATCCCGCAGTCTTCAGT;

R-ACTGAAGACTGCGGGATTGCTCTCATTTGAAGGTTTGTGGGGCCAGG.

The deletion was verified with both Sanger sequencing and Ty expression as discussed in the *Experimental Design* part.

pYES – The plasmid is a 2µm plasmid, with a *URA3* marker (Kindly provided by the Schuldiner's lab). This plasmid was used as a control in order to evaluate the pGal-Ty plsmid's growth defect.

Standard yeast media (**YPD)** were prepared according to a standard protocol [20].

SC-Ura – media composed of nitrogen base, amino acid and 2% Glucose (according to [20]).

SC-Ura, Galactose/Raffinose – media composed of nitrogen base and amino acid (according to [20]). Galactose (for the induction of Ty) or Raffinose (for inactivation of Ty) were added after autoclaved for a final concentration of 2%.

3.1.2 VLP extraction

Single colonies of RM11-a harboring the pGal-Ty or the pGal-control plasmid were picked from SC/-ura/glucose plates (approximately 10^8 cells), inoculated into 500 ml of liquid SC-ura+l% raffinose media, and shaken overnight at 30°c, Galactose was then added to a final concentration of 2% and the

cultures were shaken for 72hr at 22°c until cell density reached ~1*10^7. Cells were harvested by centrifugation, washed with 10 ml of distilled sterile water, and re-suspended in 5ml of cold buffer B/Mg (10mM HEPES-KOH (pH 7.8) 15mM KCI, 3mM DTT, 10µg/ml aprotinin , 5mM $MgCl_2$). All subsequent steps were carried out on ice or at 4°C. Cells were lysed by adding 8g of cold, nitric acid-washed glass beads and vortexed at 4°C for 5 min intervals, alternating with 1 min incubations on ice repeated 4 times. Cell lysis was monitored by phase-contrast microscopy. Glass beads were separated from the lysate by puncturing the tubes and collecting the lysate into new tubes. The lysate was centrifuged at 4000rpm, 4°C, for 10min. The supernatant (approximately 8 ml) was layered onto a sucrose step gradient composed of 5ml of 70% sucrose in buffer B and containing 10 mM EDTA (buffer B/EDTA), 5ml of 30% Sucrose in buffer B/EDTA and 20 ml of 20% sucrose in buffer B/EDTA in a Beckman SW28 polyallomer tube. The gradients were centrifuged for 3hr at 25,000 rpm, 4°C, and manually fractionated from the top (~1ml per fraction). OD260 was measured manually using Nanodrop on each of the fractions. Fractions were kept at -80 until further use. Additional assays (qRT-PCR, western analysis and RT-activity assay) were done on fractions 1,5,10,15, and 20 to end to identify the VLP-containing fractions.

3.1.3 RNA extraction and qRT-PCR

Total RNA was extracted from each of the fractions mentioned using Bio-Tri RNA reagent according to manufacture protocol and used as a template for quantitative RT–PCR using light cycler 480 SYBR I master (Biosystems)(LightCycler 480 system) according to the manufacture instructions. The absence of genomic DNA in RNA samples was checked by real-time PCR by using the RNA in the qRT-PCR. A blank (No Template Control) was also incorporated in each assay. The qRT-PCR was done using two sets of primers to identify levels of Ty's mRNA and rRNA. The Ty primers' sequences are: Ty-F-CGCTACACACGTCATCGACAT; Ty-R-GCGAGAATCATTCTTCTCATCACT; the rRNA primers are against the 18S subunit of the ribosomes are their sequences are: rRNA-F-TGGCGAACCAGGACTTTTAC; rRNA-R-CCGACCGTCCCTATTAATCAT.

3.1.4 Western analysis

20µl from each of the fractions were mixed with 60µl, 4X sample buffer and boiled for 10 minutes. 20µl of the boiled fraction was loaded on a 10% SDS-gel (give the composition of upper and lower). Proteins were transferred onto nitrocellulose membrane using a semi-dry (BioRad) protocol. Membrane was then

blocked for 1hr while shaking in room temperature in PBS-5% milk. First antibody (anti-VLP ab) was kindly sent to us by Jef Boeke's lab. Membrane was incubated in 1% milk-PBS + first antibody (1:10,000) in 4˚c overnight while shaking. 1hr incubation in room temperature was done for the secondary antibody, anti-rabbit-HRP (1:20000). Membrane was than stripped using DDW and NaCl (100mM). The stripped membranes were used again, for anti-ribosome (anti-RPL1, dilutiton 1:2500) antibody starting from the blocking step and continuing regularly the second antibody was the anti-rabbit-HRP (1:20000).

3.1.5 RT activity assay

A Retro-Sys kit by innovagen was used to quantify RT-activity of the fractions. 20µl from each of the fraction was used in the kit. Manufacture's protocol was followed with the following exceptions; incubation with the Alkaline phosphatase enzyme was done 3 times, every two hours, the 2 hour measurement was used in figures 5 and 6.

3.1.6 Lab Evolution Set up

Lab evolution was done on minimal SC-Ura (see *Strains, plasmids and media*). Evolution was done in 22˚c which is the optimized temperature to allow transposition events of Ty1 in yeast [21]. Lab evolution was done on RM11-a strain harboring one of three plasmids pGal-Ty, pGal-Ty*Int*, or pGal-control, each of the strains' evolution were done in 3 replicates. Cells were grown in liquid media, 1.2ml cultures in a 24-well plate. Cells were counted daily and were diluted (1:120) when reached to $5*10^7$ cells/ml (~7 generations); dilution was done roughly 1 per week at the beginning of the lab evolution, and 1 every two days at the end of the evolution. Cells were frozen in 30% Glycerol and are kept in -80˚c every 4 dilutions.

3.1.7 Growth Experiments

Cells were grown in 4ml of SC-Ura in 22˚c for 4 days, to reach deep-stationary phase. Cells were then diluted into fresh minimal media (1:50). Growth experiment were done in 96-well plates, 150µl per well. Two strains arranged in checkerboard pattern were analyzed in each plate – the ancestral strain as a control and the evolved strain. Plates were put in an incubator set to 22˚c. OD600 was measured every 2 hours for ~50 hours by a plate reader (infinity). All measurements were done automatically using a Hamilton robotic system. The results of the growth experiment were analyzed using a matlab-GUI

created in the lab by Avihu Yona, and using a softwate ("Curveball") created by Yoav Ram from Lilach Hadany's lab at Tel-Aviv University [14].

3.1.8 Plasmid Existence Check

Cells were grown on YPD plates in 30˚C to form colonies. Single colonies were patched on YPD plates, 30 colonies in a plate and incubated in 30˚C for 24 hours. Replicates were done for all plates on SC-Ura plates and on YPD as a control plates. Total number of colonies were determines by number of colonies that grew on the rich medium plates, i.e. the control plates; number of colonies without the plasmid was determines by the number of colonies that grew on both the control plates and the minimal media plates.

3.1.9 Colony PCR

From the Int^+ and Int^- strains, one patch that grew on Sc-Ura per evolution repetition was taken to colony PCR, and one patch per repetition that didn't grew on Sc-Ura from the control strain was taken. Cells were lysed in 20mM NaOH (50µl) and incubated for 20 minutes in 100°c. 2µl of the lysate was used as a template for the PCR. The PCR was conducted using the F: AAGCCTGACTCCACTTCCCG and R: GTGGCCAGGACAACGTATACTC primers using the iProof master mix (2X) in a 20µl total volume reaction.

Products were ran in a 1% agarose gel.

3.2 In silico Evolution

The evolution *in silico* is an evolution simulation based on a genetic algorithm (GA) [15]. The simulation incorporates genetics and evolutionary concepts such as mutations, selection and drift.

The simulation gets an initial population of *n* haploid individuals with an *m* genes genome (in current runs of the simulation $n = 100$, $m = 2$), gene expression (*g*) ranges between 0 and 1 and differ between genes and conditions. The genome of the population is being mutated with a constant rate throughout the simulation of *R* ($0<R<100$ mutations/bp). A transcriptome and an RNA mutation step followed by reverse transcription step were added over the basic flow of GA (figure 11). The RNA mutation rate (*TxR)* ranges between 0 and 100 mutation/RNA bp, and is usually set to 10 times the DNA mutation rate; the RT rate (*RTR)* ranges between 0 to 1 RT events/RNA molecule. Number of RNA molecules is 10 times the expression level (*g*) of the gene (equation 2). *g* can be affected by "environmental changes", meaning that the simulation supports different *g*'s values throughout the simulation and thus mimics environmental changes that in turn change gene expression.

The simulation supports changeable rates of each of the above parameters; the user can set and change all of the parameters above at the beginning of the simulation.

The mutagenesis part (DNA mutagenesis, RNA mutagenesis and RT) is followed by fitness evaluation and selection.

The fitness evaluation in this simulation is based on the NK model [16]. In short, this model calculates fitness for strings of characters (of length *N)* of any alphabet (in this case, binary string of 0 and 1). The fitness of a given string (*FS*) is the sum of contributions from each locus (*f(S)*) in the string (equation 3a), while the contribution of each locus in general depends on the value of *K* other loci (equation 3b). The complexity of a fitness landscape in this model is mainly a combination of the *N* and *K* values [16]. Yet, in addition we realized that complexity also varies among the various possible rules at a given pair of N and K values. Thus at N=10 and K=1 (our current setting) we identified a "rugged" and a "smooth" rule, i.e. rules that a simple greedy algorithm (not shown) solves hardly or easily. The fitness score of an individual (*FI*) is based on the fitness score of each of the genes, and their expression level (equation 3c).

The selection is done, as before, using a weighted random procedure based on the individuals' fitness, meaning that individuals' with higher fitness have higher probability of being selected to the next generations. The procedure, in short, is based on cumulative sum of normalized fitness scores (CSNF) of the population. An even distributed random number is drawn *r* (*0<=r<=1)* and the first individual with a CSNF score larger than *r* is chosen to the next generation.

The simulation has two versions, one that simulates the evolution of single population, and a second version that enables a competition between 2 sub-populations. For now, the simulation supports 3 types of competitions, w/ and w/o DNA mutations, w/ and w/o RNA mutations and w/ and w/o RT. The type of competition is user defined.

Equations for this part:

EQUATION 2

$$\#RNAmolecules = g * 10$$

EQUATION 3

a. $FS = \sum_i f(S_i))$

b. $f(S_i) = \sum_{j=1}^{k} S_j$

c. $FI = \sum_m (g_m * FS_m)$

4. Discussion

In this work we have examined for the first time the potential role of reverse transcription as an evolvability agent that if expressed in cells might serve for Lamarckian evolution.

We have started to explore this notion in three avenues: 1. we have evolved yeast cells with and without the RT activity and could observe accelerated evolution in cells that have the RT activity. On the second rout we have built the biochemical basis for a deep sequencing analysis of the VLP – the cell's factory of Ty-mediated RT activity. We have been able to set the parameters to isolate the VLP and we are now ready to sequence their content. This analysis will shed light on the natural capacity of the cell's transcriptome to "hitch hike" on the Ty machinery, and use it to reverse transcribe its own genes. On the third line of research we have set up a computer simulation that assesses the effect of RT on evolution. We have implemented several selection conditions and have examined each of them on simple optimization tasks at various degrees of difficulty. We have found a modest advantage of the RT-based genetic algorithm on hard tasks compared to easier one.

4.1 Discussion for the VLP identification part

As mentioned before, the RT could cause high mutagenesis on expressed genes only if other mRNAs (but the Ty mRNA) are being reverse transcribed in the VLP. It is known that some cytoplasmic mRNAs are being reverse transcribed [11] but the extent of the phenomena is not clear.

Maxwell et al [11] have extracted VLPs from *S. cerevisiae* cells and found using micro-arrays that the VLP contains ~1500 genes. The paper has shown that Y' element are being reverse transcribed by the Ty1 element, by tagging the Y' element with *HIS3AI* gene. If *HIS3AI* is being reverse transcribe, a copy without the AI will be found in the genome. Maxwell et al. have shown that Y' element tagged with the *HIS3* gene, without the AI had appeared in the genome after an evolution experiment. Furthermore, the tagged Y' element had a poly(A) tract followed by the Ty1 sequence. They have deduced that the Y' element are primed with the Ty1 element to allow reverse transcription. The priming of the Ty1 in the VLPs is done using a $tRNA^{Met}$, thus in addition to them being encapsulated in the VLPs the mRNAs will have to be primed, either by $tRNA^{Met}$, the Ty1 or by another, yet unknown, priming mechanism.

We have further analyzed Maxwell et al. microarray data. We have noticed that the transcriptome consists of two distinct sets of genes (figure 15): one of which shows preferential enrichment within the VLP. We were surprised to see that the rRNAs are found in the enriched subset; we hypothesized that it means that ribosomes and VLPs are found in the same fraction of the sucrose step gradient, and that

translated mRNAs can be found in those fractions even though they are not encapsulated in the VLPs. Our VLP extraction protocol appears to promise that we might not face such contamination is the VLP-positive fractions appear to be devoid of ribosomes (Figure 5).

In general it seems that we have managed to purify the VLPs from whole cell lysate since two biological repetitions show the same peak in three methods aimed to verify the presence of Ty1's mRNA, VLPs and RT activity (figure 5, green line, cyan line and bottom panel respectively). The fact that the three methods have peaked in the same fractions suggests that we have managed to isolate the complete and mature VLPs, and not immature or inactive VLPs or VLPs' proteins. The Ty-less strain did not show this peak (figure 6) contributing to our assurance that the peak in the VLP-gradient represents the VLPs. The depletion of the ribosome from the VLP-containing fraction is encouraging in that respect.

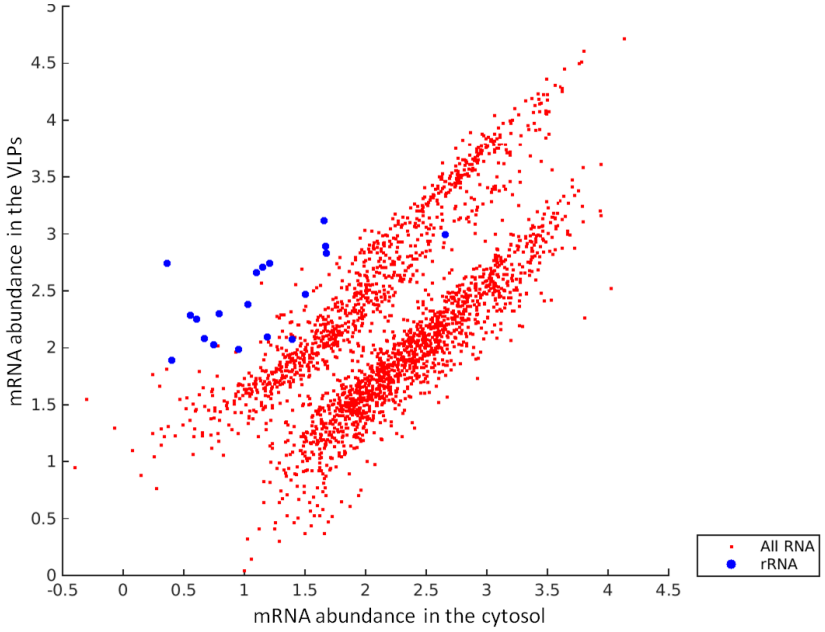

FIGURE 15. ANALYSIS OF VLP MICROARRAY DATA SHOWS TOW CLOUDS

The analysis was done on data adapted from [11]. The upper cloud represents mRNAs that are encapsulated in the VLPs while the bottom cloud represents the mRNAs that are depleted from the VLPs. In blue are rRNAs that are found mainly in the upper cloud

We are now planning on performing RNA-seq on the VLP-containing fraction that will allow us to find all mRNAs that are encapsulated in the particle. Furthermore, we will sequence both mRNA (RNA-seq) and cDNA (DNA-seq) in the VLP so we could learn a lot about the mechanism of the VLP encapsulation and reverse transcription processes. We will compare the mRNA abundance found in the VLP fractions to the mRNA abundance in whole cell lysate. This comparison will allow us to know if

the encapsulation is passive (mRNAs are encapsulated based on their concentration in the cytosol) or a controlled process (specific mRNAs are being encapsulated). The preliminary results of Maxwell et al. (Figure 15) show a mixture of both: on one hand highly expressed genes are more likely to be shown in the VLP, on the other hand there seem to be a privileged population of similarly expressed genes that are more likely to be included. We search for multiple properties among such genes (e.g. motifs and their secondary structure), but we could not find obvious features shared among the VLP enriched or depleted genes (not shown). By sequencing the cDNA we will learn if the reverse transcription is a passive (all mRNAs in the VLPs are being reverse transcribed) or a selective controlled process.

This experiment is far from being over; in the near future we want to extract the mRNA and cDNA from those fractions and sequence them separately in order to answer our original question of which mRNAs and cDNAs are in the VLPs. The sequencing will allow us to know if encapsulation of mRNA in the VLPs is a controlled process resulting in different mRNAs in the cytoplasm and in the VLPs, and if reverse transcription of mRNA in the VLPs is controlled resulting in differences between mRNAs found in VLPs and cDNA in VLPs.

4.2 Discussion for the Lab Evolution experiment

The results of the lab evolution have shown that the Int^+ strain has adapted rapidly to the experimental conditions, while the control strain has adapted the least and the Int^- strain has ultimately adapted as much as the Int^+ strain but more slowly. These results suggest that RT acts to induce adaptation and thus can affect evolutionary processes, and that the Intp plays major role in this induction.

The RT can affect evolution in two main ways; 1) Direct - the RT will reverse transcribe cytoplasmic mRNAs followed by cDNA integration to the genome. The fact that the RNA polymerase's mutation rate is higher than that of the DNA polymerase can affect evolution. Even if there are no transcription mutations in the gene, the RT can mediate gene duplication, or gene movement to another locus; this mechanism is the main hypothesis to this project. One such gene to be reverse transcribed could be the *URA3* gene itself, see below. 2) Indirect - the RT reverse transcribes itself, and then the Ty1's cDNA is being integrated into random positions in the genome and causes mutations.

As discussed in the results part, the plasmid causes impaired growth of the cells; the cells could not simply lose the plasmid since the plasmid contains the *URA3* gene that the cells need in order to grow on SC-Ura. The results show correlation between the loss of the plasmid and improved growth. There are two main ways to interpret the results of this experiment, with respect to the two ways it can affect evolution; 1) The *URA* gene from the plasmid was encapsulated in the VLP, reverse transcribed and

integrated to the genome in the strains that had the RT. Since the control strain has no Ty, its fitness improvement showed the slowest dynamics. The difference between the other two strains is due to the fact that in the pGal-Ty*Int* the Ura's cDNA can't integrate to the genome via integration, but only via homologues recombination, a process that takes longer especially since the *URA3* gene has no homologous sequence in this strain's genome. 2) Random integration of the Ty enabled homologues recombination between the plasmid and the genome, leading to integration of the full plasmid into the genome. In the Int^+ strain, the Ty was reverse transcribed and integrated into the genome by the Intp, creating a 6kb homologues region between the plasmid and the genome, making the recombination of the plasmid into the genome quite easy. In the control strain, the plasmid doesn't contain Ty1, so this 6kb homologues region is not formed and the plasmid cannot be integrated into the genome. In the Int^- strain, a Ty cDNA is being made, but since the Integrase is not active, it is being integrated to the genome via non-specific recombination such as NHEJ, this process is likely to take more time, but once it occurred the plasmid can integrate into the genome based on homologous recombination, making the fitness improvement slower than the Int^+'s fitness improvement. The PCR analysis done on the evolved strain has shown that although the plasmid was "lost" after growth on rich media, it was still found in the cells. This result means that the full plasmid was integrated into the genome, maintaining the *URA3* gene after growth on rich media but decreasing the cost of having such a large plasmid in the cells, fitting to the second mechanism that was described.

In either way, the fact that Ty can influence evolution (in this case, due to the indirect mechanism) is important for evolutionary knowledge since retrotransposons are found in most species, including human [22], [23]; it is also known that cancer cells activate the human retrotransposones in the process of becoming cancerous [24]. Understanding the possibilities that RT presents to the cells and its mechanisms can shed light on multiple evolutionary processes including cancerous processes.

4.3 Discussion for the Evolution in silico part

The simulation was meant to be used to assess the effects of RT in a theoretical manner, finding situations and conditions in which RT is beneficial or deleterious. The simulation worked well regarding the DNA mutations, since it predicted that the best mutation rate in the DNA is ~1 mutation per genome per generation. This result are compatible with previous work in which mutation rate of most species is roughly 1/L per generation, where L is genome length [19], the result also predicts the deleterious effect of the increase in DNA mutation rate above the 1/L threshold (figure 14a).

The competition simulation has resembled the results of the regular simulation, showing that the ideal mutation rate is ~1/genome length.

The simulation enables us to see that in difficult tasks, i.e. rugged landscapes, the RT has a beneficial effect on evolution, while in easy tasks the RT may cause a deleterious effect. The fact that the RT is beneficial in rugged hard problems makes sense since the RT enables adaptation of expressed genes, and keeps un-expressed (and adapted earlier) to be kept with a basal mutation rate. In addition, although the effective mutation rate (DNA mutation rate and transcription mutation rate) is the same in the with and without RT strains, in the with RT strains the mutations occur more in the expressed genes, enabling bigger changes in the genotype that cause bigger changes in the phenotype and thus preventing the population from being stuck in a local maximum. The fact that the RT is deleterious in the smooth landscape can be explained by the same token, the fact that the expressed genes have higher mutation rate may bring them to "over the threshold" point in the easy problems.

However, the results regarding the RNA mutation and RT are puzzling. The competition experiment, which showed that the RNA mutation rate is not beneficial or deleterious (figure 11b), implies that there is no need for the RNA polymerase to be accurate, which is not the situation in nature. This contradiction may be due to the fact that in nature the mRNA is mainly a message to create proteins thus mutated mRNA will result in mutated protein that might not fulfill its function and reducing the fitness of the individual, while in the simulation the RNA is the final step and the fitness is determined by it.

Even more puzzling is the fact that adding the RT to the simulation is also not beneficial nor deleterious in all RT rates (figure 11c). Assuming that RNA mutations themselves cannot affect the simulation since the mutations do not pass to the next generation, using RT on no DNA mutation and RNA mutation background was expected to give a result resembling the DNA mutation experiment results.

In the future we will study and develop conceptually simulation until we will understand what these results stand for. Once the simulation will run as expected, we will expand the simulation to support integration via the Intp and not only using homoluges recombination. Furthermore, we will increase population size, number of genes and their length, the complexity of the NK model and complexity of the environment to try and understand more about RT-mediated evolution.

Acknowledgments

I would like to thank my mentor, Prof. Yitzhak Pilpel, for his continuous support and caring. I have learned a lot about science and on much more from our discussions, Thank you!

I wish to thank all my lab members for many fruitful discussions. You taught me of the endless ways science can be performed and made this year full of stories and memories.

Special thanks go to Orna Dahan, Hila Gingold, Ruthy Towers & Tammy Biniashvili- for all your help at all times.

I would also like to thank Gerst's lab, to Nadav Segev and Dmitry Zabezhinsky for helping me in all the biochemical assays done in this project such as the sucrose gradient and the western blot analysis.

Finally, I will be forever grateful to my close family and friends for their tremendous support and guidance throughout my life.

Bibliography

[1] E. V Koonin and Y. I. Wolf, "Is evolution Darwinian or/and Lamarckian?," *Biol. Direct*, vol. 4, no. 1, p. 42, Jan. 2009.

[2] H. Ochman, J. G. Lawrence, and E. A. Groisman, "Lateral gene transfer and the nature of bacterial innovation.," *Nature*, vol. 405, no. 6784, pp. 299–304, May 2000.

[3] S. Jinks-Robertson and A. S. Bhagwat, "Transcription-associated mutagenesis.," *Annu. Rev. Genet.*, vol. 48, pp. 341–59, Jan. 2014.

[4] G. I. Lang and A. W. Murray, "Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae.," *Genetics*, vol. 178, no. 1, pp. 67–82, Jan. 2008.

[5] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes," *Mol. Biol. Cell*, vol. 11, no. 12, pp. 4241–4257, Dec. 2000.

[6] P. Eigen, M. ; McCaskill, J. & Schuster, "THE MOLECULAR QUASI-SPECIES," *Adv. Chem. Phys.*, vol. 75, pp. 149–263, 1989.

[7] D. J. Garfinkel, "Genome evolution mediated by Ty elements in Saccharomyces.," *Cytogenet. Genome*

Res., vol. 110, no. 1–4, pp. 63–9, Jan. 2005.

[8] D. J. Eichinger and J. D. Boeke, "The DNA intermediate in yeast Ty1 element transposition copurifies with virus-like particles: Cell-free Ty1 transposition," *Cell*, vol. 54, no. 7, pp. 955–966, Sep. 1988.

[9] J. F. Roth, "The yeast Ty virus-like particles.," *Yeast*, vol. 16, no. 9, pp. 785–95, Jun. 2000.

[10] M. J. Curcio and D. J. Garfinkel, "Single-step selection for Ty1 element retrotransposition," *PNAS*, vol. 88, no. 3, pp. 936–940, 1991.

[11] 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. 9887–98, Nov. 2004.

[12] "In-frame linker insertion mutagenesis of yeast transposon Tyl: phenotypic analysis." [Online]. Available: http://ac.els-cdn.com/0378111994905185/1-s2.0-0378111994905185-main.pdf?_tid=cb34469c-71b0-11e5-aca9-00000aacb361&acdnat=1444744164_0e19af56d41c27e112feb7c80db5fdd5. [Accessed: 13-Oct-2015].

[13] E. S. F and R. E. Lenski, "EVOLUTION EXPERIMENTS WITH MICROORGANISMS: THE DYNAMICS AND GENETIC BASES OF ADAPTATION," *Nat. Rev.*, vol. 4, pp. 457–469, 2003.

[14] R. Yoav, D.-G. Eynat, O. Uri, B. Maayan, and H. Judith, Berman and Lilach, "Predicting competition results from growth curves." [Online]. Available: http://biorxiv.org/content/biorxiv/early/2015/07/23/022640.full.pdf. [Accessed: 26-Oct-2015].

[15] M. Mitchell, *An Introduction to Genetic Algorithms*. MIT press, 1996.

[16] S. A. ; W. E. D. Kauffman, "The NK Model of Rugged Fitness Landscapes And Its Application to Maturation of the Immune Response," *J. Theor. Biol.*, vol. 141, no. 2, pp. 211–245, 1989.

[17] L. T. M. R.E;, "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations," *PNAS*, vol. 91, pp. 6808–6814, 1994.

[18] S. Wright, "The Roles of Mutations, Inbreeding, Crossbreeding and Selection in Evolution," *Proceeding of the sixth*, pp. 355–366, 1932.

[19] M. A. Nowak, *Evolutionary Dynamics*. 2006.

[20] G. R. Christine Guthrie;Fink, *Guide to Yeast Genetics and Molecular and Cell Biology, Part C | 978-0-12-182254-5 | Elsevier*. 2002.

[21] W. V. Paquin CE, "Temperature Effects on the Rate of Ty Transposition," *Science (80-.).*, vol. 226, pp. 53–55, 1984.

[22] H. H. K. Jr., "Mobile Elements: Drivers of Genome Evolution," *Science (80-.).*, vol. 303, pp. 1626–1632, 2004.

[23] E. S. et al Lander, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[24] W. a. Schulz, "L1 retrotransposons in human cancers," *J. Biomed. Biotechnol.*, vol. 2006, pp. 1–12, 2006.