Genetic analysis of disease-related short tandem DNA repeat instability in yeast

Berit Haitjema: s3821226

Research Project Behaviour and Neurosciences

Supervisor: Daniele Novarina (ERIBA, UMCG)

Abstract

DNA repeat-instability causes up to 50 human diseases, most of which are currently not treatable or preventable. These pathologies originate from DNA repeat expansion in the human genome, which can lead to interference with important cellular processes and to cell death and disease. In this pilot experiment, we aimed to find the optimal repeat length for 2 of the 13 known disease-causing repeats, namely GAA/TTC and TGAAA/TTTCA, in yeast. The results of this pilot experiment will be used to determine the optimal repeat length for a genome wide screening done by the Chang research group (EIRBA). With this result genes involved in the decrease and increase in DNA instability in yeast can be identified. We found that the strains that were tested did not yet show a gross chromosomal rearrangement (GCR) frequency in the desired range. Therefore, different repeat lengths should be used for these two DNA repeat types in both orientations to obtain more suitable GCR frequency results. The results of the following genome wide screening could be used to identify important genetic factors for repeat expansion diseases (REDs) in humans, and hopefully lead to promising treatment options.

Introduction

DNA repeat-instability causes up to 50 human diseases and most of these are currently not treatable or preventable *(Khristich and Mirkin, 2020)*. Nowadays, 13 different types of DNA repeats are known that cause a variety of human diseases, named repeat expansion diseases (REDs).

These DNA repeats can be defined as an increased number of repeats in the genomic DNA *(Mirkin, 2007)*. This expansion is caused by their unusual structural properties, which can eventually lead to complications in cellular processes such as replication, repair and recombination. Interference in these processes can result in altered gene expression depending on the type, length and location of the repeat, which can lead to different pathologies.

Thus, among these different repeat types and REDs, pathogenetic mechanisms differ. However, all REDs are initialized by the instability of the repeat sequence at the DNA level *(Khristich and Mirkin, 2020)*. It is therefore important to understand the mechanisms preceding DNA repeat instability and the cellular pathways preventing this instability, to eventually be able to understand and cure these diseases.

Currently the Chang research group (ERIBA) is working on exploring the causes of short tandem repeats instability in yeast. They are doing so by inserting a tract of each repeat in both orientations into the yeast genome and creating knockouts for all yeast genes. With this, they want to find which genes are involved in increasing or decreasing the stability of the DNA, using high throughput screening. To do this in all 13 types of repeats known, they need to know which repeat length of each repeat type in each orientation is optimal to experiment on.

Therefore, we conducted a pilot study for 2 types of repeats in both orientations. The GAA/TTC repeats, which are expanded in Duchenne muscular dystrophy *(Kekou et al., 2016)* and Friedreich ataxia *(Campuzano et al., 1997)*, together with the TGAAA/TTTCA repeats, which are expanded in several types of familial adult myoclonic epilepsy *(Lagorio et al., 2019)*. Our aim is to find the optimal repeat length of the GAA/TTC and TGAAA/TTTCA sequences to be used for the genome wide screening done by the Chang research group (ERIBA). To examine this, we obtained different repeat lengths for the sequences and determined their gross chromosomal rearrangement (GCR) frequency in yeast. GCRs include translocation, inversions, amplifications and deletions of a large part of the chromosome *(Chen and Kolodner, 1999)*. Therefore, determining the GCR frequency is an established method of determining instability of the DNA in yeast.

Yeast is deployed for this research for various reasons. First of all, yeast is an eukaryotic organism, which makes its cellular processes similar to those of human cells (*Botstein and Fink, 2011*). This enables us to model cellular mechanisms accurately. Next to this, it has a fast division rate of approximately 90 minutes, can be easily genetically manipulated and is inexpensive (*Duina et al., 2014)*. These characteristics are advantageous since a GCR event in yeast is quite unique (1 in 10^7), thus many cells are needed to do an accurate measure of the GCR frequency.

The method of determining this GCR frequency relies on two genes in the genome of the yeast, namely URA3 and CAN1 *(Srivatsan et al., 2018)*. Cells that have lost the part of the chromosome containing these two genes are resistant to both canavanine (CAN) and 5-fluoroorotic acid (5-FOA), meaning that cells that have undergone a GCR event will be able to grow in an environment containing CAN and 5-FOA. This allows us to visualize their survival number and subsequently determine the GCR frequency for the yeast strains containing different repeat sequences and lengths. The GCR frequency we want to obtain is around 30-50% GCRs for it to be functional for the following genome wide screen of the repeat sequences.

For this pilot study we used the following repeat lengths. For the GAA repeat these are 80, 95 and 160 repeats and for the TTC repeat these are 81, 121 and 160 repeats. For the TGAAA repeat these are 40, 41 and 43 repeats and for the TTTCA repeat these are 40, 41 and 42 repeats. Previous results from the ERIBA yeast lab from the UMCG have shown that these repeat lengths seem to be in the right range for this pilot and therefore the following genome wide screening. We therefore expected to find a useful length to be in this range.

Using our yeast experimental system to measure the optimal repeat length for the following genome wide screening, we found that the strains that were tested did not have a GCR frequency in the optimal range for the following genome wide screening. We therefore recommend using different repeat lengths to elongate this pilot experiment, and with this, try to find better fitting GCR frequencies for the different repeats in different orientations to be used in further research.

Methods

Plasmid extraction from E.coli

E. coli cells were grown on LB medium containing 100 µg/mL ampicillin, expressing one of the plasmids (Supplementary table S1). Incubation was done at 30°C to ensure stability of the repeats. A miniprep was conducted to extract the plasmids from *E. coli*, using the NucleoSpin® Plasmid kit from Macherey-Nagel.

Yeast transformation

The yeast cells containing genetic markers for the GCR assay (CZY105) (Supplementary table S2) were streaked for single colonies, which were then used for transformation with the digested plasmids and PCR-amplified repeat fragments. Yeast cells that were successfully transformed were selected on YPD medium containing Hygromycin. Survivors were streaked for single colonies, from which several colonies for every length and repeat sequence were subsequently used to make temporary storage patches on YPD medium containing Hygromycin. Incubation was done at 30°C.

PCR

Genomic DNA was extracted from single colonies for PCR. $TTTCA_{42}$-HphMX and $TGAAA_{42}$-HphMX were amplified from pSO25 and pSO14 respectively, using primers pAG32_PRB1_fwd and pAG32_PRB1_rvs and Q5 polymerase, according to the New England BioLabs protocol (Supplementary table S3, S4). To verify that the repeat sequences were inserted at the *PRB1* locus, and to prepare samples to be sent for sequencing, primers HygB fwd and FR_PRB1_insertion_Rvs and Taq polymerase were used according to the New England BioLabs protocol (Supplementary table S3, S5).

To check the GAA repeat lengths, colony PCR was performed according to the New England BioLabs protocol, using Taq polymerase and primers expansion-FOR and expansion-REV (Supplementary table S3, S6). To check the TTC repeat lengths, colony PCR was performed according to the New England BioLabs protocol, using Taq polymerase and primers expansion-FOR and expansion-REV2 (Supplementary table S3, S7) that anneal near the repeat sequences in order to assess the repeat length more accurately. PCR clean-up was performed using the Nucleospin® Gel and PCR clean-up kit.

Digestion

Plasmids pDN41.1, pDN64.1, pDN59.1, pDN40.3, pDN63.1, and pDN58.1 were digested using SalI to linearize the repeat fragments. We digested 5 µg of plasmid with 100 units of SalI for 2 hours at 37 °C, followed by 20 minutes of heat activation at 65 °C.

Verification of repeat sequence and lengths

Integration of repeats at the *PRB1* locus was verified on a 1% agarose gel (5 µl ethidium bromide per 100 mL agarose solution). PCR products from GAA and TTC repeats were verified on a 2% agarose gel (5 µl ethidium bromide per 100 mL agarose solution).

Genomic DNA was extracted from storage patches for diagnostic PCR to verify that the repeats are integrated at the *PRB1* locus in the transformants, and that the repeats have the desired length. The PCR product was then analyzed by Sanger sequencing, using 2 µl of HygB fwd primer per sample, to determine the exact length and sequence of the repeats.

Patch-and-replica-plating

A patch-and-replica-plating experiment was performed to estimate the GCR frequencies of the strains containing the repeat sequences. First, to exclude GCRs before the start of the experiment, patches were made on SD-URA plates of the selected strains, as well as three positive control strains containing the following repeat sequences: CAG_{74}, CAG_{121}, and CAG_{161}. The patches were grown for 1 day at 30°C, and then replica-plated on 5-FOA/CAN plates to be grown for 4 days at 30°C. GCR frequencies were then estimated by counting the number of colonies per patch, as every colony stems from one GCR event.

GCR assay via robotic pinning

The strains containing the repeat sequences were streaked for single colonies on SD-URA. The selected colonies were manually arrayed in 96-format on SD-URA rectangular plates and grown for 1 day at 30°C. With the ROTOR HDA pinning robot, the colonies were arrayed from 96-format to 384-format, and from 384-format to 1536-format on SD-URA RotoR non-selective plates. In between pinning steps, the plates are incubated for 1 day at 30°C. The colonies were robotically replica-pinned onto new SD-URA RotoR plates and imaged.

The colonies were then robotically replica-pinned in 1536-format onto selective 5-FOA/CAN RotoR plates and incubated for 2 days at 30°C. These colonies were then again robotically replica-pinned in 1536-format onto selective 5-FOA/CAN RotoR plates and incubated for 1 day at 30°C. This selective plate was then also imaged. To calculate the GCR frequency for each repeat sequence and length, colonies on the non-selective and selective plates were counted manually. All imaging was done using a ScanMaker 9800XL scanner (MicroTek International, Inc.), scanning in 16-bit grayscale at 300-dpi resolution with the Transparent Media Adapter.

Results

Transformation of CZY105 yeast with repeat sequences of different lengths

The GAA and TTC sequences were digested with SalI to obtain linear repeat fragments of different sizes. Preliminary data suggested that slight variations in repeat lengths for TGAAA and TTTCA sequences, would already result in notable differences in GCR frequencies. Therefore, $TGAAA_{42}$-HphMX and $TTTCA_{42}$-HphMX were amplified using PCR in order to obtain different repeat lengths, due to expansions and contractions in the repeat tract. After transformation, yeast cells (CZY105) containing genetic markers for the GCR assay (Supplementary table S2), were grown on YPD+Hygromycin plates to select for successful transformants. Colony growth was observed for all sequences and repeat lengths, however, for GAA_{121} and GAA_{160} very little colony growth was observed.

To then verify that the repeat fragments were integrated into the correct location in the genome, the *PRB1* locus, a diagnostic PCR was performed on extracted genomic DNA. Integration and repeat lengths of all transformants were analyzed by Sanger sequencing. From this, we selected the following repeat lengths for the different sequences: $TGAAA_{40}$, $TGAAA_{41}$, $TGAAA_{43}$, $TTTCA_{41}$, $TTTCA_{42}$, $TTTCA_{43}$, GAA_{80}, GAA_{95}, and TTC_{81}.

However, for the longer repeat sequences (roughly longer than 300 base pairs) the sequencing results became unreliable. Therefore PCR was performed on transformants containing GAA_{121}, GAA_{160}, TTC_{121}, and TTC_{160}, which were then validated on a 2% agarose gel.

From this, we selected the following repeat lengths for the different sequences: TTC_{121} and TTC_{160}. The GAA repeat sequences on the gel were shorter than required, therefore we only continued with GAA_{95}, which was the result of the transformation with pDN58.1 and was verified by sequencing. To be able to continue the experiment with three different repeat lengths for GAA as well, we used a GAA_{160} which originated from a different background (DNY188), which was available at Chang's lab (ERIBA).

Verification of repeat sequences and lengths of selected strains to proceed with the GCR assay

To verify the repeat lengths of the different yeast strains before starting the GCR assay, PCR was performed. As a positive control, the plasmids were also amplified with PCR using the same primers. For the GAA and TTC repeat sequences, we validated the length of the PCR product by 2% agarose gel electrophoresis (Figure 1). We expected to see bands with the following sizes: 312 bp (GAA_{81}), 432 bp (GAA_{121}), 549 bp (GAA_{160}), 304 bp (TTC_{81}), 424 bp (TTC_{121}), 541 bp (TTC_{160}). Even though not all bands were at the expected locations in the gel based on the DNA ladder, the samples did match the size of the controls, suggesting that the number of repeats is the same.

Due to unreliable sequencing data, we were unable to verify that the TGAAA and TTTCA repeat sequences in the different strains were still of the same size before starting the GCR assay.

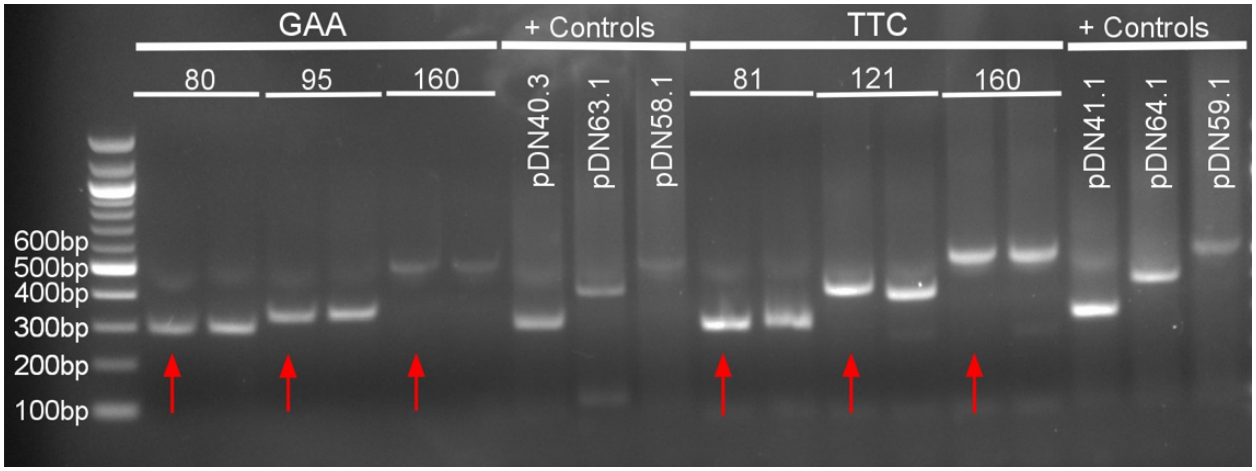

Figure 1: 2% agarose gel electrophoresis for verification of the repeat lengths before proceeding with the GCR assay. Repeat sequence lengths were of the expected size based on the bands of the controls on the gel. We proceeded the experiment with the first sample (red arrow) for every repeat length. A 100bp DNA ladder (NEB) was used.

Estimation of GCR frequency via patch-and-replica-plating

To obtain a rough estimation of the GCR frequency of each strain, we performed patch-and-replica-plating. From this we could observe whether there is a difference in colony growth between the patches (Figure 2), as a proxy for the GCR frequency. Here, each colony originates from a single GCR event, and therefore a higher number of colonies equals a higher GCR frequency. For the CAG repeat controls, an increased colony growth is observed for only CAG_{161} and not for the CAG_{121} and CAG_{74}. For the TTC repeats, an increase in colony growth is observed in TTC_{160} compared to TTC_{121} and TTC_{81} on the selective plate. For the other sequences, there is no clear difference in colony growth between the different repeat lengths.

TTC 81	TTTCA 41	TGAAA 40	GAA 80
TTC 121	TTTCA 42	TGAAA 41	GAA 95
TTC 160	TTTCA 43	TGAAA 43	GAA 160

CAG 161	CAG 121	CAG 74

Figure 2: Quick and dirty estimation of GCR frequency via patch-and-replica-plating. Patches were replica plated on CAN 5-FOA plates. On the left we see the CAN 5-FOA plate where we can observe increased colony growth for TTC_{160} and CAG_{161}. On the right we see a schematic overview of the patches on the plate.

Measuring the length-dependent GCR frequency via robotic pinning

To obtain a more precise GCR frequency, we measured the GCR frequency of all strains via robotic pinning. Colonies were first grown on nonselective SD-URA plates and arrayed from 96-format to 384-format and then to 1536-format (Figure 3A). All colonies were of equal size on the nonselective plates, except for GAA_{160}, which were visibly smaller (Figure 3B, 4B). This could indicate that this strain grows slower.

Subsequently, the colonies were replica-pinned on selective CAN 5-FOA plates (Figure 3B). Surprisingly, the colonies on both selective plates (Figure 3B, 4B) were not all of the same size. The controls (CAG_{121}, CAG_{161}) indicate that a longer repeat length leads to a higher GCR frequency (Figure 3C, 3D, 4C, 4D). A similar pattern is observed for the TTC repeat lengths, where repeat length correlates with GCR frequency (Figure 3C). For the GAA repeat lengths, the difference in GCR frequency for GAA_{80} and GAA_{95} is rather large, but this increase in GCR frequency is not observed when the repeat length is increased to GAA_{160} (Figure 3D).

The GCR frequency did not differ considerably between the different lengths for TTTCA (Figure 4C). For the TGAAA repeat lengths, we see a minor decrease in GCR frequency as the repeat length increases (Figure 4D).

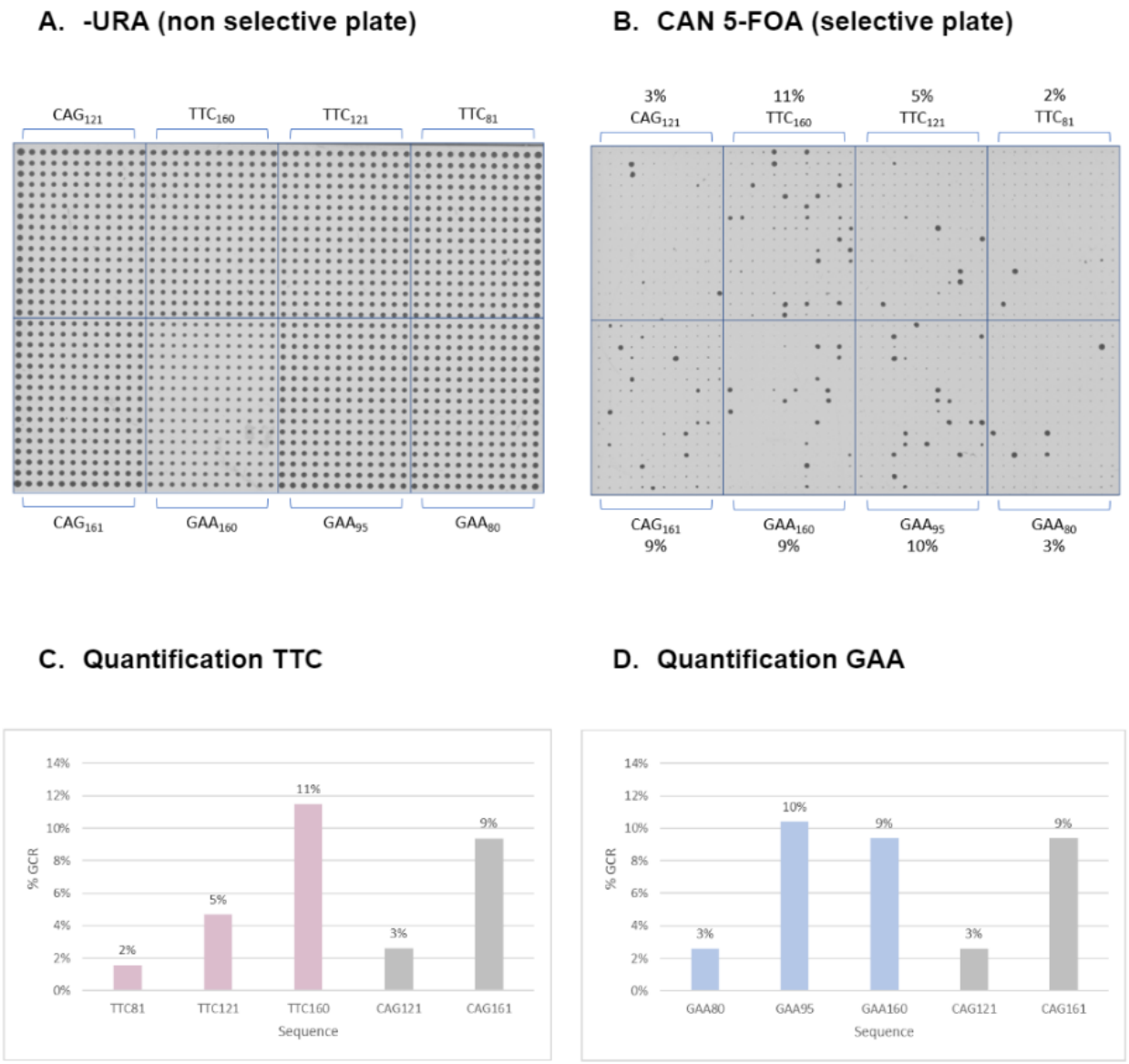

Figure 3: Result of GCR assay for different TTC and GAA repeat containing strains. **(A)** Colonies were grown on a non selective -URA plate to exclude GCRs before the start of the experiment. From the plate we can see that GAA_{160} colonies grew less than the other strains. **(B)** Colonies were grown on CAN 5-FOA plates to select colonies that have undergone a GCR event. Percentages of colonies that grew are displayed together with respective strains. Observed colonies are of different sizes. **(C)** Graphical representation of percentage of TTC repeat colonies that have undergone a GCR event. Here, a longer repeat sequence results in a higher GCR frequency. The CAG controls match this observation. **(D)** Graphical representation of the percentage of GAA colonies that have undergone a GCR event. Here, an increase in repeat length from 80 to 95 repeats already gives a high increase in GCR frequency. This increase is not visible at GAA_{160} repeats, which might be caused by its slower growth frequency.

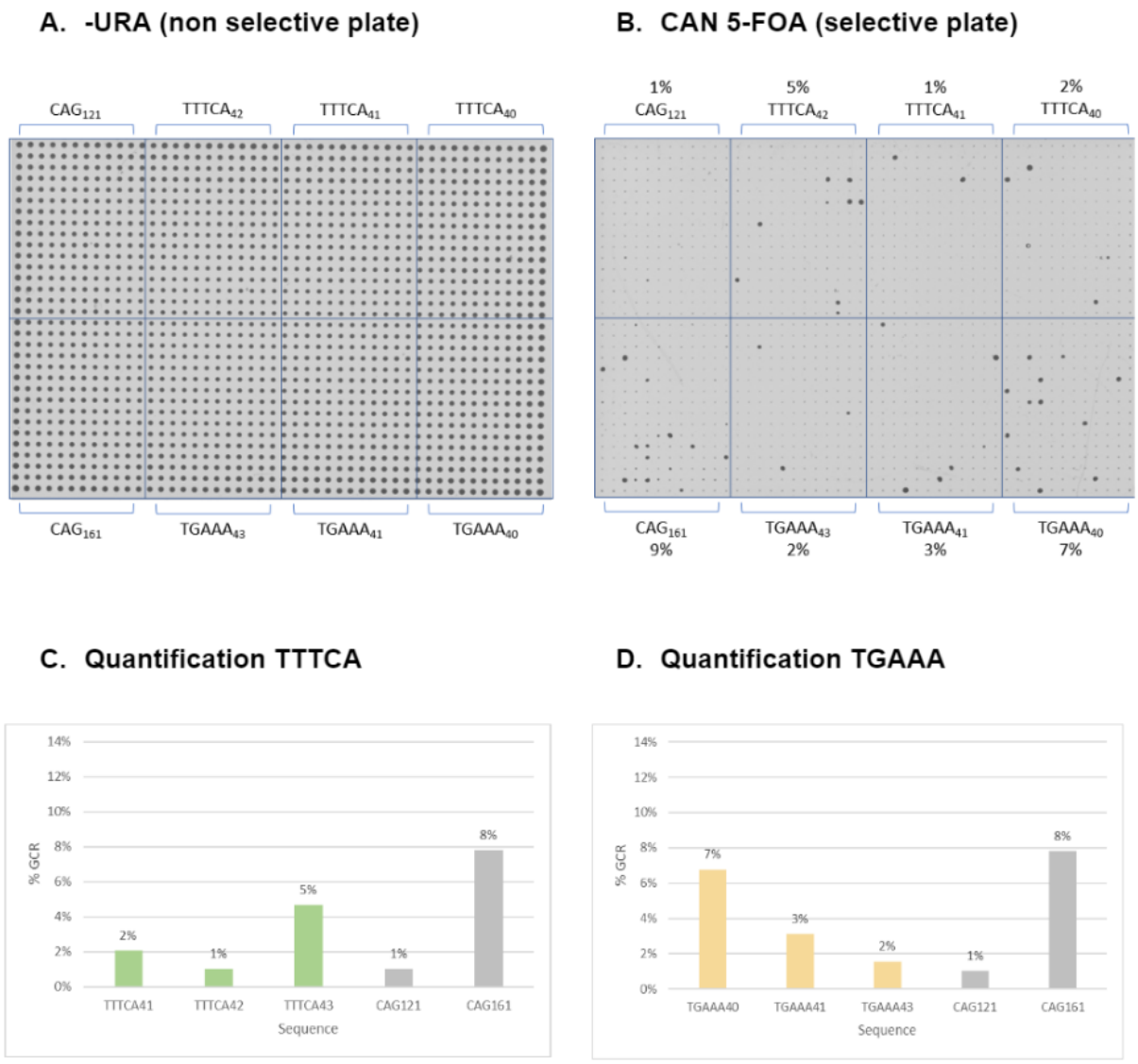

Figure 4: Result of GCR assay for different TTTCA and TGAAA repeat containing strains. **(A)** Colonies were grown on a non selective -URA plate to exclude GCRs before the start of the experiment. From the plate we can see that all colonies grew similarly. **(B)** Colonies were grown on CAN 5-FOA plates to select colonies that have undergone a GCR event. Percentages of colonies that grew are displayed together with respective strains. Observed colonies are of different sizes. **(C)** Graphical representation of percentage of TTTCA repeat colonies that have undergone a GCR event. Here, the GCR frequency does not seem to differ considerably between different repeat lengths. **(D)** Graphical representation of the percentage of TGAAA colonies that have undergone a GCR event. Here, an increase in repeat sequences seems to give a decrease in GCR frequency. However, because of the small differences, this could be due to random variability. For the CAG controls, the GCR frequency increases as the repeat length increases. This pattern is not observed for the TTTCA and TGAAA repeat containing strains.

Discussion

In this pilot study, we tried to find a suitable repeat length for the different strains to be used for the future genome wide screening done by the Chang research group (ERIBA). Previously, researchers found that REDs are initialized by the instability of the repeat sequence at the DNA level *(Khristich and Mirkin, 2020)*, which indicates the importance of finding the genetic origins of this DNA instability.

First, we transformed yeast (CZY105) with different repeat sequences and lengths. This transformation succeeded for most strains, but unfortunately not for GAA_{121} and GAA_{160}. We then verified Integration in the right site in the genome and size. After this we decided to, instead of GAA_{121} and GAA_{160}, use GAA_{95} and from another yeast background (DNY188) GAA_{160} for the GCR assay. The selected repeat sequences and lengths were verified again before the start of the GCR assay to examine whether the repeat size had remained stable. This was done by gel electrophoresis for the GAA and TTC sequences, and these sequences were still of the previously established size. For the TGAAA and TTTCA sequences Sanger sequencing was performed. However, the sequencing failed and due to time constraints we could not determine these sequence lengths anymore. We can therefore not be certain of the total lengths of these repeat sequences.

Patch-and-replica plating was performed to obtain a rough estimation of the GCR frequency for each strain. This indicated that for the CAG repeats (control) a longer repeat sequence resulted in a higher GCR frequency, and this was also observed for the TTC repeats. For the other strains, no clear differences were yet observed here. Since we observed partially different results for the GCR assay, this method does not seem reliable to get an accurate measurement of the GCR frequency.

To assess the GCR frequency more precisely, a GCR assay via robot pinning was performed. For the TTC repeat we observed that a longer repeat sequence results in a higher GCR frequency. This result is consistent with the patch-and-replica plating observations. For 160 TTC repeats, we obtained a GCR frequency of 11%. However, with this GCR frequency it would only be possible to observe whether gene knockouts cause an increase in GCRs. Therefore, longer repeat lengths should be tested for TTC if possible, to be able to include a decrease in GCRs for the genome wide screening. For the GAA repeat we observed a steeper correlation between repeat length and GCR frequency. This is especially evident from the increase of 8o to 95 repeats, where the GCR frequency increases from 3% to 10%. This trend does not continue for the 160 repeats, but this is likely due to the lower growth frequency that was observed for this strain, which might result from the different background of this strain. It would therefore be interesting to test the CZY105 yeast with 160 GAA repeats, as this data suggests that this could lead to a high increase in GCR frequency, and this might produce a good length to be used for genome wide screening.

Additionally, we observed a clear difference in GCR frequency resulting from different repeat lengths for the two orientations. This is also mentioned in earlier research by Alexandra N. Khristich and Sergei M. Mirkin, where the researchers mention that there is a clear orientation-dependent instability for some repeat types (*Khristich and Mirkin, 2020)*. However, recent research suggests that there is almost no difference in repeat contraction frequency for the two orientations of the GAA/TTC repeats (*Khristich et al., 2020)*.

For the TGAAA and TTTCA strains we observed no considerable differences in GCR frequency, and the differences seem mostly due to random variability. This indicates that different lengths should be tested for these repeat sequences, and specifically more divergent lengths. A suggestion would be to include 60 and 80 repeats of these strains in both orientations as well, as this might result in more diverse GCR frequencies. From this a suitable length could be selected to be used for the genome wide screening.

It was also observed that some colonies on the selective GCR plate grew larger than other colonies. This could be due to colonies not dying immediately after being pinned on a selective plate, but being able to survive for some time and therefore have more time to undergo a GCR event. This might indicate that the plates should be incubated longer to be able to observe more colonies. Next to this, source mixing with the robot for the selective GCR plate might also increase colony growth and should be done for next robot pinnings. Another technical issue is that for the longer repeat sequences (>121 repeats), it was not possible to get an accurate sequencing result. Therefore these lengths were verified by using gel electrophoresis, but this method is less precise. Lengths verified with this method could therefore have slightly longer or shorter repeat tracts, but from the results shown here we do not expect this to have a significant effect on the GCR frequency. Next to this the GCR assay is a fragility assay, used as a proxy for repeat instability. However, fragility and instability do not always fully overlap and therefore it is important to consider using a different assay to test DNA instability, but this is not yet generated.

In conclusion, this pilot experiment should be elongated with different repeat lengths to determine the optimal repeat size per sequence. In this way, suitable GCR frequencies can be selected that can be used to determine the effect of different genes on the stability of the DNA repeats in yeast. The results of this genome wide screening could then be used in identifying important genetic factors for REDs in humans, and hopefully be a starting point for promising therapeutic treatment.

References

1: Khristich, A. N., & Mirkin, S. M. (2020). On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *Journal of Biological Chemistry*, *295*(13), 4134-4170.

2: Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature*, *447*(7147), 932-940.

3: Kekou, K., Sofocleous, C., Papadimas, G., Petichakis, D., Svingou, M., Pons, R. M., Vorgia, P., Gika, A., Kitsiou-Tzeli, S., & Kanavakis, E. (2016). A dynamic trinucleotide repeat (TNR) expansion in the DMD gene. *Molecular and cellular probes*, *30*(4), 254–260.

4: Campuzano, V., Montermini, L., Lutz, Y., Cova, L., Hindelang, C., Jiralerspong, S., Trottier, Y., Kish, S. J., Faucheux, B., Trouillas, P., Authier, F. J., Dürr, A., Mandel, J. L., Vescovi, A., Pandolfo, M., & Koenig, M. (1997). Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Human molecular genetics*, *6*(11), 1771–1780.

5: Lagorio, I., Zara, F., Striano, S., & Striano, P. (2019). Familial adult myoclonic epilepsy: a new expansion repeats disorder. *Seizure*, *67*, 73-77.

6: Chen, C., & Kolodner, R. D. (1999). Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. Nature genetics, 23(1), 81-85.

7: Srivatsan, A., Putnam, C. D., & Kolodner, R. D. (2018). Analyzing genome rearrangements in Saccharomyces cerevisiae. Genome Instability: Methods and Protocols, 43-61.

8: Duina, A. A., Miller, M. E., & Keeney, J. B. (2014). Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. *Genetics*, *197*(1), 33-48.

9: Botstein, D., & Fink, G. R. (2011). Yeast: an experimental organism for 21st century biology. *Genetics*, *189*(3), 695-704.

10: Khristich A. N., Armenia J. F., Matera R. M., Kolchinski A. A., and Mirkin S. M. (2020) Large-scale contractions of Friedreich's ataxia GAA repeats in yeast occur during DNA replication due to their triplex-forming ability. *Proc. Natl. Acad. Sci. U.S.A*. 117, 1628–1637

Appendix

Table S1. Plasmids used in this pilot study.

Plasmid	Repeat sequence + length	Selection marker
pS025	$TTTCA_{42}$	Ampicillin resistance
pS014	$TGAAA_{42}$	Ampicillin resistance
pDN41.1	TTC_{81}	Ampicillin resistance
pDN64.1	TTC_{121}	Ampicillin resistance
pDN59.1	TTC_{160}	Ampicillin resistance
pDN40.3	GAA_{81}	Ampicillin resistance
pDN63.1	GAA_{121}	Ampicillin resistance
pDN63.1	GAA_{160}	Ampicillin resistance

Table S2. Yeast strains used in this pilot study.

Strain name	Genotype
CZY105	*Matα hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + $TTTCA_{41}$	*Matα prb1::HPH:*$TTTCA_{41}$ *hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + $TTTCA_{42}$	*Matα prb1::HPH:*$TTTCA_{42}$ *hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + $TTTCA_{43}$	*Matα prb1::HPH:*$TTTCA_{43}$ *hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + $TGAAA_{40}$	*Matα prb1::HPH:*$TGAAA_{40}$ *hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + $TGAAA_{41}$	*Matα prb1::HPH:*$TGAAA_{41}$ *hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + $TGAAA_{43}$	*Matα prb1::HPH:*$TGAAA_{43}$ *hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*

CZY105 + TTC_{81}	*Matα prb1::HPH:TTC_{81} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + TTC_{121}	*Matα prb1::HPH:TTC_{121} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + TTC_{160}	*Matα prb1::HPH:TTC_{160} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + GAA_{80}	*Matα prb1::HPH:GAA_{80} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
CZY105 + GAA_{95}	*Matα prb1::HPH:GAA_{95} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
DNY188	*Matα prb1::HPH:GAA_{160} hxt13Δ::URA3 ARG4 CAN1 lys2Δ ADE2 leu2-3,112 his3-11,15 ura3-1 TRP1 RAD5*
DNY107	*Matα prb1::HPH:CAG_{74} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
DNY178	*Matα prb1::HPH:CAG_{121} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*
DNY160	*Matα prb1::HPH:CAG_{161} hxt13::URA3 mfa1::MFA1pr-HIS3 his3Δ1 ura3Δ0 lyp1Δ leu2Δ0 metl5Δ0*

Table S3. Primer sequences.

Primer name	**Sequence**
pAG32_PRB1_fwd	CAAACTTAAGAGTCCAATTAGCTTCATCGCCAATAAAAAAACAA ACTAAACCTAATTCTAACAgggttaattaaggcgcgccagatct
pAG32_PRB1_Rvs	TTGTAACCTCGAGACGCCTAAGGAAAGAAAAAGAAAAAAAAA AGCAGCTGAAATTTTTCTAAAcactatagggagaccggcagatccgl
HygB fwd	TCTTGACCAACTCTATCAGA
FR_PRB1_insertion_rvs	CGCAAATATGTAGTAATACGTGG
Expansion_FOR	ttctaaatgtcctgcaggcg
Expansion_REV	gtcctgcagttaatTAAGGTC
Expansion_REV2	tgctgtcgattcgaaggatc

Table S4. PCR program for amplification of TTTCA42-HphMX and TGAAA42-HphMX from pSO25 and pSO14, respectively.

Cycle step	Temperature	Time	Cycles
Initial denaturation	98°C	30 seconds	1
Denaturation	98°C	10 seconds	30
Annealing + extension	72°C	1 minute and 20 seconds	
Final extension	72°C	5 minutes	1
Hold	10°C	∞	

Table S5. PCR program for verification of sequence length and correct integration at the PRB1 locus.

Cycle step	Temperature	Time	Cycles
Initial denaturation	94°C	30 seconds	1
Denaturation	94°C	30 seconds	35
Annealing	45°C	30 seconds	
Extension	68°C	1 minute and 30 seconds	
Final extension	68°C	5 minutes	1
Hold	10°C	∞	

Table S6. Colony PCR program to check GAA repeats size.

Cycle step	Temperature	Time	Cycles
Initial denaturation	94°C	30 seconds	1
Denaturation	94°C	30 seconds	35
Annealing	51°C	30 seconds	
Extension	68°C	40 seconds	
Final extension	68°C	5 minutes	1
Hold	10°C	∞	

Table S7. Colony PCR program to check TTC repeats size.

Cycle step	Temperature	Time	Cycles
Initial denaturation	94°C	30 seconds	1
Denaturation	94°C	30 seconds	35
Annealing	47°C	30 seconds	
Extension	68°C	40 seconds	
Final extension	68°C	5 minutes	1
Hold	10°C	∞	