DNA Replication in Growth Conditions that Mimic the Natural Habitat of *Haloferax volcanii*

Bryn McCulloch, BSc

Student ID: 20215139

Thesis submitted to the University of Nottingham

for the degree of Master of Research

October 2020

26112 words

Table of Contents

Chapter One: Origin independent replication of *Haloferax volcanii* in conditions that better mimic its natural environment

Chapter Two: An Origin independent replication prediction tool

Abstract

The initial aim of the project was to assess origin-independent replication in *Haloferax volcanii (Hfx. volcanii)*. DNA replication is initiated at specific sites on the chromosome called origins. Origins are assumed to be an essential feature of all cells, because they serve as binding sites for proteins that recruit the DNA replication machinery. In work published by Hawkins *et al,* (2013), it was demonstrated that mutants of *Hfx. volcanii* lacking all replication origins are viable; in fact, they grow faster than the wild-type and have no obvious cellular defects. By contrast, deletion of origins from Eukaryotes and Bacteria leads to cell death or profound growth defects.

The question addressed in this project was whether the accelerated growth of *Hfx. volcanii* cells in the absence of replication origins is due to an artefact created by rich laboratory media conditions. This may explain why replication origins have not been eliminated by natural selection, as in the natural habitat of *Hfx. volcanii*, the wild-type strain would have an evolutionary advantage. To test this, a growth competition assay was modified to use fluorescent proteins and flow cytometry. It was predicted that in low nutrient media, the growth advantage of origin-deleted mutants will be minimised or eliminated, as these phenotypes are not witnessed in a natural environment.

However, due to the outbreak of the COVID-19 pandemic, the project was altered to examine which factors are required for an organism to replicate without origins. A bioinformatic approach was chosen, adapting previously created tools to better fit a large data set and to predict the ability of 85 species to survive without origins. The bioinformatic pipeline involved a principal component analysis, which would take into account for any given species their respective nucleotide skew indices, spectral ratios, information gene linkage, co-orientation of core genes with DNA replication, and types of DNA polymerase genes located near origins. The results suggested several new candidate species for further experimentation and potential directions for improvement of the origin independent replication prediction tool.

Covid-19 statement

Covid-19 interrupted my Master of Research because I was unable to access the laboratory to complete planned experiments and data collection. This impacted on the ability to write up the project, as the results obtained were less than expected. As I am high risk and due to the maximum number of people allowed within the lab after reopening, I was not able to return to finish the planned project. Hence, I could not achieve the planned outcomes of my MRes project. In an attempt to mitigate this, a bioinformatics variation on the project was designed. However, planning a suitable and viable project took several weeks and required me to learn a new set of bioinformatics skills, which could not be fully supported by my supervisors who mostly work in a wet lab setting. The conceptual divide in the two research projects has also made my thesis lack the coherency it would otherwise have had.

As previously mentioned, I am high risk, and because of the lab occupancy size under Covid-19 regulations I was therefore unable to return to the lab to finish the initial project. A new project was designed that was based on bioinformatics (Chapter two). This resulted in increased stress levels as I was unable to have face to face supervisor meetings to discuss issues with the project that required a steep learning curve, and which was not the expertise of my supervisors to begin with. During this time, I had to relocate due to housing issues, which resulted in a brief period where I could not work.

For further details see the Covid-19 statement form in the appendix.

Acknowledgements

Firstly. I'd like to thank my supervisors Prof. Thorsten Allers and Dr. Ambika Dattani for the opportunity to work on this project and the help and support provided throughout.

I would also like to give my thanks to the other members of the Aller's lab particularly Vicky, Patricia and Laura for their various aid throughout the year.

Special acknowledgements to Jon for all his work making media and for generally ensuring the lab continued to run smoothly. As well as the other laboratory technicians working in QMC who aided in the day to day running of the laboratory.

I'd like to thank Caroline Adlam as well as Yuri Wolf's and Ian Duggin's Labs for their work which directly contributed to aspects of this project, which would not have been possible otherwise.

Finally, a huge thank you to Abbie for the support and putting up with me over the year.

Chapter One: Origin independent replication of *Haloferax volcanii* in conditions that better mimic its natural environment.

1.Introduction

1.1 Archaea and the origin of life.

For most of the 20th Century, organisms were grouped into two apparently distinct domains within the tree of life: Prokaryotes containing all Bacteria and Archaea, and Eukaryotes consisting of the remaining plant, Fungus and animal life. Now scientists have included Protista, a group of organisms that do not fit into the previously mentioned groups into the Eukarya domain and created a third domain designated for Archaea.

The two-domain belief was based on morphological and physiological traits until the mid-1970's, when Carl Woese and George Fox (1977) amongst other microbiologists revolutionised phylogenetic taxonomy with the use of RNA sequence analysis of the 16S ribosomal component. They suggested a third domain of life separate from those which had been previously established, which they called Archaebacteria. In the following years, the close relationship between Archaea and Eukaryotes was further established leading to the previously named Archaebacteria being moved from the bacterial domain to its own domain on the tree of life known as Archaea (see Figure 1.) (Woese *et al,* 1990).

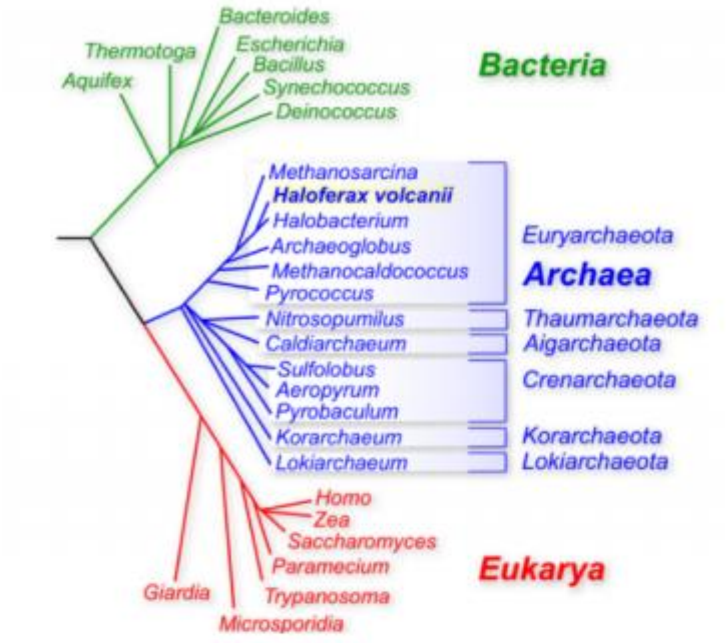


Figure 1. The three-domain tree of life proposed by Woese based on 16S rRNA sequencing. Defining three separate lineages; Bacteria in green, Archaea in blue and Eukarya in red (Adapted from Allers and Mevarech, 2005)[4].

Since the introduction of Archaea into the tree of life, various advancements have occurred that have improved our understanding of the group and where it fits on the tree of life. This was largely contributed to by the increased use and development of genetic analysis strategies, in particular cultivation-independent techniques for genome sequencing. Winker and Woese (1991) had suggested the introduction of two archaeal kingdoms, the *Crenarchaeota* and *Euryarchaeota,* in the years following the addition of archaeal domain to the tree of life. These kingdoms were defined as different based solely on the small subunit of rRNA sequences. Since then, the placement of new lineages can no longer be inferred on a singular gene and instead multiple genes have been analysed simultaneously, including: *recA* and *gyrB* alongside 16S rRNA sequences (Yoon *et al*, 2017).

Between the years 2006 and 2011, three new lineages were added to the archaeal domain, *Korarcheota* (Auchtung *et al,* 2006), *Thaumarchaeota* (Brochier-Armanet *et al,* 2008) and *Aigarchaeota* (Nunoura *et al,* 2011). These three domains alongside *Crenarchaeota* form a superphylum referred to as 'TACK' and have also been provisionally designated the kingdom name of *Proteoarchaeota*. Since the addition of this superphylum other lineages have been discovered and suggested to branch within this group, such as *Bathyarchaeota* and *Geoarchaeota* (Barns *et al,* 1996). Now the archaeal domain consists of at least four major supergroups; the Euryarchaeota, TACK, Asgard and DPANN Archaea (Koonin, 2015).

Despite advancements in phylogeny, Archaea have demonstrated morphological similarity with Bacteria, including their chromosomal organisation and lack of intracellular compartments (Londei, 2005). But Archaea are characteristically different to Bacteria in numerous ways, such as the absence of peptidoglycan, a component of the cell wall utilised by most Bacteria. Archaea share features with the Eukaryotic domain, such as the subunit structure of DNA-dependent RNA polymerases (Huet *et al,* 1983) and the use of similar machinery for both initiation of DNA replication and DNA repair (Kelman and White 2005:,O'Donnell *et al,* 2013). Additionally, Archaea present a range of peculiar metabolisms and physiologies including: methanogenesis in Methanogens (Fox *et al*, 1977), sulphur metabolism in Sulfobales, alongside numerous other thermophilic and halophilic archaeal groups (Rother and Metcalf, 2005).

The similarity between Archaea and Eukaryotes, and the extreme environmental conditions inhabited by Archaea, have led some members of the scientific community to believe that the eukaryotic domain may have originated from an ancient Archaea species, present on Earth when atmospheric oxygen levels were low (termed the last universal common ancestor or

LUCA). Recent evidence proposed a two-domain tree of life in which Eukaryotes in fact originate from the archaeal domain, with the most likely candidate being a member of the recently cultivated Asgard Archaea (Embley and Williams, 2015). The species '*Candidatus Prometheoarchaeum syntrophicum*', an Archaeon in the Lokiarchaeota phylum (a part of the 'Asgard' superphylum), has been recently cultivated and has been suggested to be the closest living archaeal relative of the eukaryotes . The cultivation of this species has also led to the creation of a new hypothetical model for eukaryogenesis (Zaremba-Niedzwiedzka *et al,* 2017: Imachi *et al*, 2020).

1.2. Halophiles

Halophiles are extremophilic organisms adapted to high salinity environments (DasSarma and DasSarma, 2015), in areas such as salt flats, salt mines and polar aqueous environments where the salt can readily dissolve forming brine (Javor, 2012). Hypersaline environments are defined as environments which surpass the salt concentration of the sea, which resides at 3.5% (w/v) (Díaz-Cárdenas *et al,* 2017). As the chemical composition of the environment naturally fluctuates over time , as does the salt concentration, as a result halophiles have adapted to be tolerant to a range of salinities and can be categorised based on these ranges (Margesin and Schinner, 2001). Slight halophiles thrive in a salinity range of 2% to 5% NaCl, moderate halophiles in 5% to 20% NaCl and extreme halophiles in 20% to 30% NaCl (Kates *et al,* 1993). Environments containing high salt concentrations apply high osmotic pressure on organisms (Wood, 2015), which would cause osmosis of cytoplasmic fluid from non-halophilic cells not specialised to combat this pressure, resulting in cell death (Oren, 2011). Halophiles combat this in various ways; slightly and moderate halophiles (principally Bacteria) use 'compatible solutes' to reduce osmotic pressure. This involves synthesising sugars and amino acids into the cell's cytoplasm, increasing the solute concentration in the cell and preventing osmosis of cytoplasmic fluid (Roberts, 2005). Extreme halophiles (principally Archaea) deal with this pressure in a different manner, instead accumulating salts such a potassium chloride (KCl) in the cell to prevent osmosis; this requires specialised cell machinery, involving the adaptation of proteins to function in molar salt concentrations, therefore halophiles using this technique do not survive in lower salinity conditions (Oren, 2008).

1.3. *Haloferax volcanii*

Hfx. volcanii is a fast-growing, easy to cultivate, haloarchaeon. Haloarchaea are one of the largest groups of archaea found within the Euryarchaeota phylum, *Hfx. volcanii* belongs to the *Haloferax* genus alongside 21 other species, notably closely related to *Hfx*. m*editerranei* with an 86.6% nucleotide reference identity (Naor *et al,* 2012). The *Haloferax* genus are most commonly found in hypersaline environments such as oceanic environments containing high Salt concentrations such as the Dead sea and the great Salt Lake. *Hfx. volcanii was* originally isolated from the Dead Sea and from a saltern in Alicante, Spain (Mullakhanbhai, 1975). Its morphology is that of a flat crisp like shape with red pigmentation caused by the presence of carotenoids.

Hfx. volcanii is most commonly cultured at 45 °C in an aerobic atmosphere with the presence of 1.7 to 2.5 M sodium chloride (NaCl) in the lab. Under these conditions, in liquid media, the generation time is around 2 to 3 hours (Zhou *et al,* 2008).

The genome of *Hfx. volcanii* exhibits extensive polyploidy with a genome copy number of 20 per cell. A high-GC content can also be exhibited at approximately 65%. As a whole, the genome of wildtype *Hfx. volcanii* is 4.2 Mb consisting of a main chromosome (2.85 Mb) and three mini chromosomes; pHV1, pHV3, and pHV4 with sizes of; 86 Kb, 442 Kb and 690 Kb respectively (Norais *et al,* 2007). In addition, a 6 Kb plasmid (pHV2) was cured from the laboratory strain. In the course of generating the laboratory strain H26, the pHV4 mini chromosome has also been incorporated (inadvertently) into the main 2.85 Mb chromosome. A whole genome sequence of *Hfx. volcanii* is also available (Hartman *et al,* 2010).

1.4. Haloferax volcanii genetic toolbox

Since its discovery *Hfx. volcanii* has emerged as an important archaeal model. An extensive repertoire of genetic, biochemical and molecular tools has been developed for this archaeon, including selectable markers, gene-deletion constructs, expression vectors, and CRISPR Cas systems (Allers and Mevarech, 2005: Gophna *et al,* 2017).

Selectable markers

Several *Hfx. volcanii* strains have been manipulated in the laboratory to allow the use of selectable markers. Antibiotic resistant selectable markers, which were identified as mutants of the essential *gyrB and hmgA* genes; however, they suffer from the acquisition of antibiotic resistance via homologous recombination due to the closely matched homology. The mutated *gyrB* gene allows for resistance to novobiocin and the *hmgA* gene from mevinolin (Allers *et al,* 2004). More commonly, the following selectable markers are used: *pyrE2, trpA, leuB* and *hdrB.* These are involved in the corresponding biosynthesis pathways: uracil, tryptophan, leucine and thymidine, respectively (Allers *et al,* 2004; Bitan-banin *et al*, 2003; Ortenberg Rozenblatt-Rosen and Mevarech, 2000).

Transformation

Hfx. volcanii can be transformed with plasmid DNA that has been demethylated (*dam-)*, allowing for easy manipulation of strains. This is required due to a restriction endonuclease (Mrr*)* in *Hfx. volcanii* that targets methylated DNA (*dam+)* resulting in a 10 fold drop in transformation efficiency (Holmes *et al,* 1991). The use of a *dam- Escherichia coli* host, which are unable to methylate at GATC sites, as a shuttle vector can be used to effectively avoid this barrier. (see Table 12 for *dam- E.coli* strains utilised for this method). The deletion of *mrr* in a strain of *Hfx. volcanii* allows for direct transformations of methylated DNA (*dam+)* as the cell is unable to recognise and degrade methylated DNA (Allers *et al,* 2010). Linear DNA transformations can be done in this way, but the efficiency is approximately 100-fold less than the use of circular plasmid DNA, so should be avoided where possible. Transformation of *Hfx. volcanii* requires removal of the S-layer, a layer of glycoproteins on the surface layer of the cell, via treatment with ethylenediaminetetraacetic acid (EDTA) (Cline *et al,* 1989).

Gene deletion and replacement

The pop-in/pop-out system has been developed in *Hfx. volcanii* to carry out gene deletion/knockout events. This mechanism utilises the *pyrE2* marker. Strains deleted for *pyrE2 (ΔpyrE2)* are transformed with a plasmid containing a deletion construct for a desired gene, this construct typically contains a selection marker for example a *trpA* marker in place of the targeted gene and the *pyrE2* marker. Successful transformants, namely pop-ins, will grow on media absent of uracil. The uracil selection can then be removed to select for pop-outs (see Figure 2.). This is done on Hv Ca media with the addition of 5-FOA (5-fluoroorotic acid) which is toxic to strains which have retained the *pyrE2* marker (Bitan-Banin *et al,* 2003).

The addition of a marker such as *trpA* allows for direct selection of deletion mutants as well as forcing the gene deletion event by providing an additional selection for pop-outs, particularly when the gene targeted for deletion is near-essential. In the absence of a marker such as *trpA*, the pop-out will more readily revert to a wild type state than result in a deletion of the desired gene (Allers *et al,* 2004).

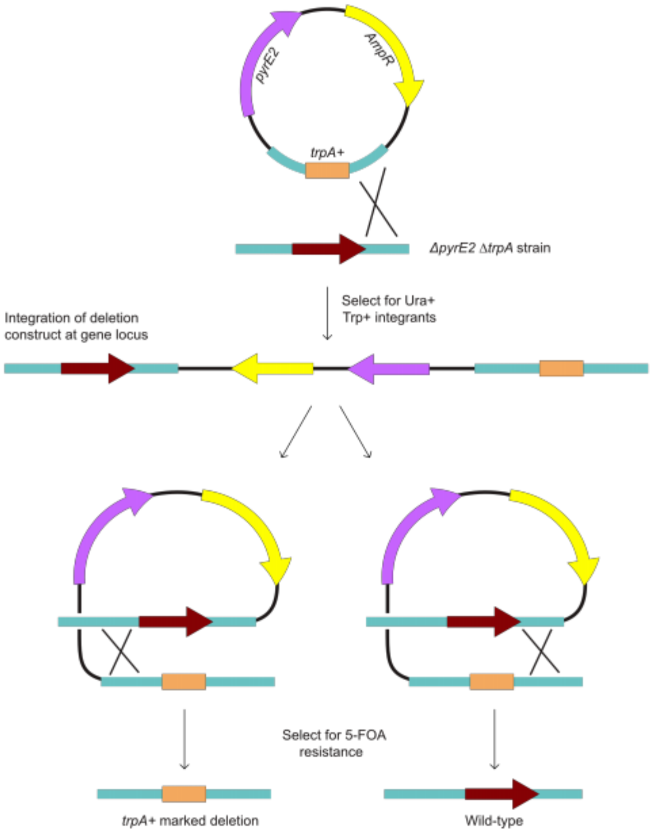


Figure 2. Gene deletion construct utilising the pop-in/pop-out method. Δ*pyrE2* strains are transformed with *pyrE2* deletion construct. Pop-ins can be selected for on media lacking uracil. Removing the uracil selection allows for pop-outs. The resulting recombination between homologous regions can be upstream or downstream of the desired gene. Pop-outs can then be selected for by plating on 5-FOA.The result will either be a gene deletion or wild type (Bitan-Banin *et al,* 2003)^{36}.

Reporter genes

There are two main reporter gene tools utilised for *Hfx. volcanii*. The first, most commonly used for growth competition assays is the *bgaH* ß-galactosidase gene (Holmes and Dyall-Smith, 2000). This allows for blue/white screening under x-gal treatment; blue colonies for cells with an active ß-galactosidase gene and white colonies for an inactive gene. This can be used to identify between strains grown in pairwise competition against one another in liquid media (Delmas *et al,* 2009).

The second method uses fluorescent proteins, most commonly a green fluorescent protein (GFP) that has been modified for high salinity conditions via amino acid substitutions for use in the hypersaline cytoplasm of *Hfx. volcanii* (Crameri *et al*, 1996; Reuter and Maupin-Furlow, 2004; Duggin *et al*, 2015).

1.5.DNA Replication and Repair

DNA Replication

It is of fundamental importance that complete and accurate DNA replication occurs for all life. The replication of DNA is therefore strictly regulated as this process must occur prior to cell division for proper inheritance of genetic information by the next generation. The process of DNA replication is broken down into three stages; initiation, elongation and termination, all of which can be observed in all forms of life with slight differences to methodology (Kornberg and TA, 1980) (See Figure 3.).

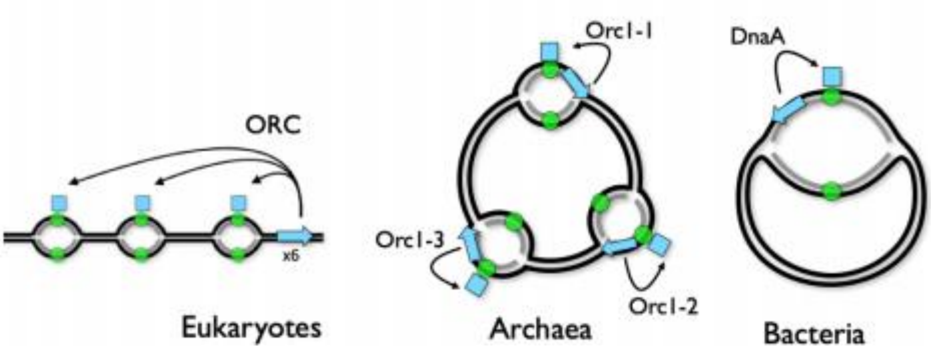


Figure 3. Summary of replication initiation for each domain of life. Eukaryotes possess linear chromosomes with multiple origins where ORC (origin recognition complex) controls initiation. Archaea initiated replication from multiple origins also controlled by ORC and Bacteria initiate replication on their circular chromosomes from a singular origin using DnaA.

Initiation

Origin-dependant replication

The initiation of DNA replication typically occurs at specific regions defined as origins of replication**.** These are sequences where the DNA unwinds allowing for loading of replication machinery and synthesis of new DNA. Organisms such as *E. coli*, a bacterium with a circular chromosome, undergo concurrent rounds of replication that are initiated from a single origin of replication known as *oriC* (O'Donnell, Langston and Stillman, 2013).

Eukaryotes have multiple origins of replication across a linear genome. In this case, initiation is dependent on the origin recognition complex (ORC), a protein complex composed of individual proteins known as Orc1-6 (Bell and Dutta, 2002). ORC recruits a replication factors known as Cdc6 and Cdt1. These function as a helicase loader and recruit MCM (a replicative helicase), which is part of the CMG complex alongside Cdc45 and GINS (Makarova, Koonin and Kelman, 2012). All these complexes and factors together then allow DNA replication at an origin to initiate.

Archaeal organisms can have either a singular origin of replication or several. For example, *Hfx. volcanii* has three origins *oriC-1, 2* and *3* and an additional origin on the integrated mini chromosome pHV4 known as *ori-pHV4*. As a result, the laboratory strain H26 four origins, including that from integrated pHV4 (see Figure 4). On the other hand, P*yrococcus abyssi* has a singular origin of replication (Matsunaga *et al,* 2003). All Archaea appear to have at least one homologue of Orc1 or Cdc6, similar to their eukaryotic counterpart, and the archaeal *orc1/cdc6* genes are typically located next to their cognate origins. Orc1 proteins bind DNA to origin recognition boxes (ORBs) and recruit MCM, which forms a CMG complex similar to the eukaryotic method of replication in which CMG is essential. However, not all the Orc1/Cdc6 homologs are involved in DNA replication, in *Hfx. volcanii* at least two are known to have no role in DNA replication (Norais *et al*, 2007). It has also been suggested that many archaeal Orc1 or Cdc6 proteins have overlapping functions. The archaeal replication machinery shares similarities with both Bacteria such as *E. coli* by utilising a DNA unwinding element (DUE), and with Eukaryotes as several archaeal replicative proteins share sequence homology with eukaryotic counterparts (Ausiannikava and Allers, 2017).

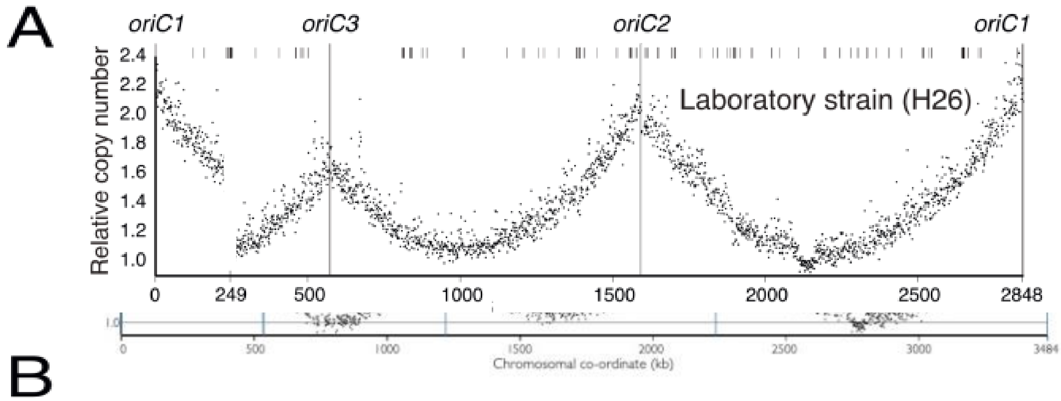


Figure 4. Replication profiles for *H. volcanii*. A.) relative copy number plotted against chromosomal coordinates for the main chromosome, showing *oriC1-3*. B.) The relative copy number against chromosomal coordinates for the integration of the pHV4 origin between *oriC1 and oriC3*. Figure for laboratory strain H26 adapted from (Hawkins *et al,* 2013)[1].

After the recruitment of the initiator proteins, replicative helicases and replication factors, bi-directional synthesis of DNA is initiated at the origin resulting in two replication forks as shown by Figure 5.

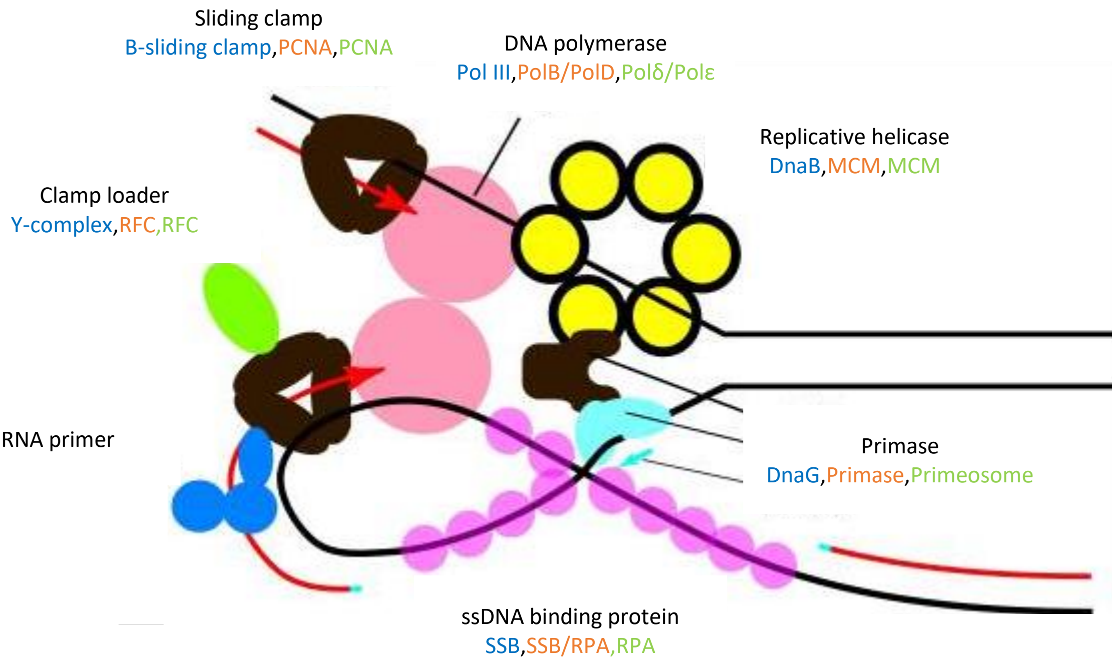


Figure 5. Components of a replication fork for each of the three domains of life. Bacteria in blue, Archaea in orange and Eukaryotes in green. It should be noted that Bacterial DnaB helicase is located on the lagging strand template whereas both eukaryotic and archaeal MCM is located on the leading strand (as shown above). Adapted from Barry and Bell (2006).

Elongation

The elongation stage can then be further sub-divided into priming and DNA synthesis. DNA primase synthesises RNA primers on both the leading and lagging strands. DNA polymerase is then able to elongate the RNA primers and carry out DNA synthesis. This occurs differently on each strand. On the (5’-3’) strand the synthesis is continuous and on the (3’-5’) strand it is discontinuous forming Okazaki fragments which are later linked via a DNA ligase.

Bacteria use the primase DnaG in combination with DnaB a hexametric helicase, while Eukaryotes use a heterodimer of PriS and PriL subunits along with Pol α and an accessory B subunit forming the Pol α/primase complex; Archaea species use homologs of PriS and PriL primase subunits, but without the use of Pol α and its B subunit (Böhlke *et al,* 2002).

DNA synthesis then occurs from the RNA primers synthesised by the various primases. Bacteria DNA synthesis is carried out by the C-family DNA polymerase III. Eukaryotes contain two B-family DNA polymerases; Pol epsilon and Pol delta. All Archaea contain a B-family DNA polymerase called PolB, with some species containing an additional PolD which is composed of DP1 and DP2 subunits (MacNeill, 2001).

Termination

The third and final stage, termination, halts the process of DNA replication. As Bacteria have circular genomes, termination occurs opposite the initiation site (*oriC)* when the replication forks meet at the *Ter* site; these *Ter* sequences are bound to by a terminator protein known as *Tus,* which block the replication fork from travelling past them resulting in termination of the replicative process (Duggin *et al*, 2008).

In Eukaryotes, the termination site is less clearly defined. Termination occurs when two replication forks collide and as a result are ligated together. This process occurs randomly between two origins of replication (Eydmann *et al,* 2008). In a similar manner, Archaea do not appear to have defined sites of termination. It is likely that termination occurs in a manner similar to that of Eukaryotes when forks from multiple origins collide. In *Hfx. volcanii* termination has been shown to occur over a broad range of regions within the genome (Hawkins *et al*, 2013).

DNA Damage and repair

DNA is constantly exposed to damage from a variety of different sources, this damage must be repaired in order to prevent mutation and potential loss of normal cell functionality. These sources can be categorised into endogenous and exogenous sources. The former is a result of normal metabolic processes in the cell which may produce harmful by-products such as the reactive oxygen radicals or through mistakes such as errors during DNA replication. Exogenous damage is caused by irradiation or exposure to chemical mutagens. This can be highly cytotoxic and result in single or double-strand breaks (DSB).

Most forms of life contain mechanisms to repair DNA by reversing chemical changes, this process is known as direct repair. There are three types of direct repair: DNA ligation, photoreactivation and reversal of methylation (Friedberg, 2003). Excision repair is a universally conserved cut-and-patch process that includes; base excision repair, nucleotide excision repair and mismatch repair.

The base excision repair pathway rectifies small DNA lesions that arise from various sources including; oxidation, deamination, alkylation and methylation. Components used in this pathway are largely conserved across all domains of life (Sartori and Jiricny, 2003), the characteristic initial step is cleavage by a lesion-specific glycosylase of the N-glycosidic bond between the damaged base and the phosphodiester backbone.

Nucleotide excision repair is a more versatile DNA damage removal pathway that is used to repair bulky helix-distorting lesions, via the excision of short nucleotide segment. Defects in this repair pathway can results in a predisposition to cancer. The components of this pathway are less conserved than base excision repair across domains, with bacterial and eukaryotic proteins involved showing little homology (de Laat, Jaspers and Hoeijmakers, 1999). Homologues of bacterial UvrABC nucleotide excision repair enzymes are found in some archaeal species, including *Hfx. volcanii*.

Mismatch repair is used to replace mismatched bases which most commonly arise as a result of replication errors. This is vital in maintaining genome stability and is a highly conserved mechanism across most species of Bacteria and Eukaryotes. However, the proteins used for mismatch repair in Bacteria and Eukaryotes can be observed in only a limited set of archaeal species (Schaaper, 1993).

Homologous recombination

The most relevant method of DSB repair to this project is homologous recombination. This method is utilised across Bacteria, Eukaryotes and Archaea and has been observed to be a highly accurate method of repair which uses a homologous DNA molecule as a template for repair (White, 2011). This process can be split into three stages: pre-synapsis, synapsis and post-synapsis(Figure 6).

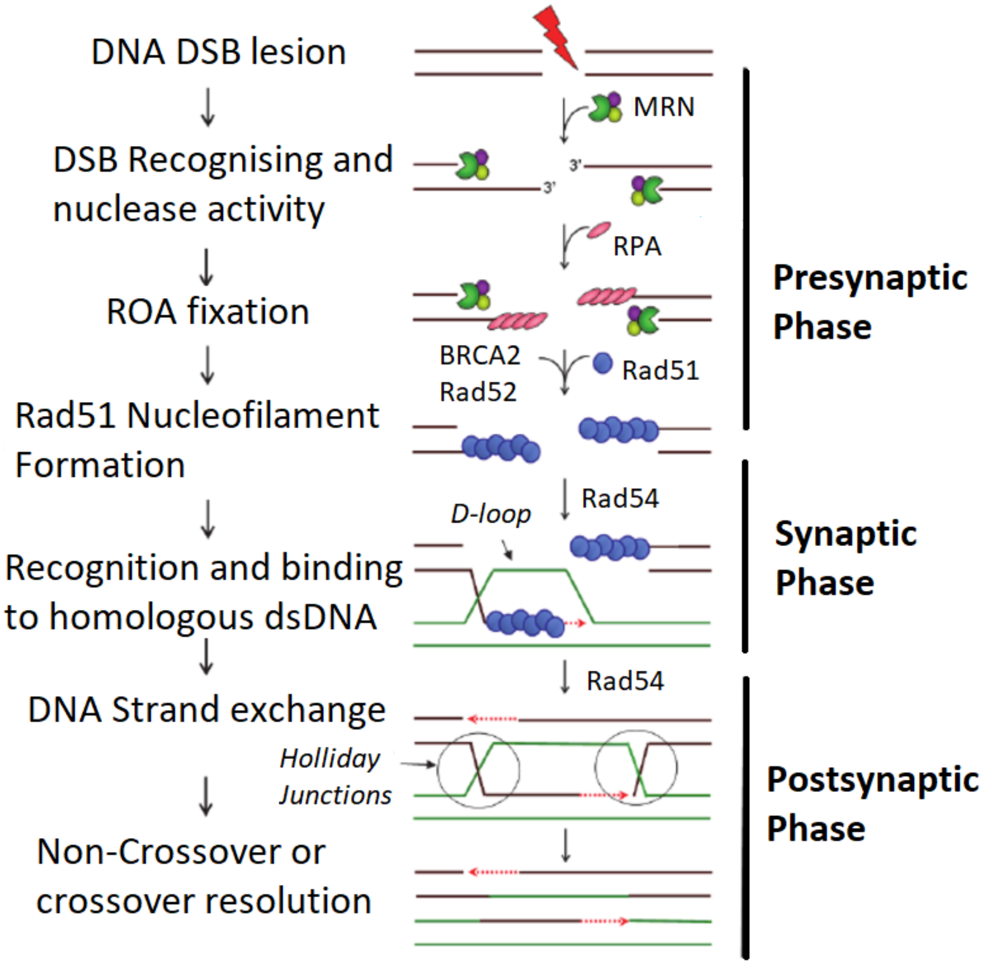


Figure 6. Representation of the homologous recombination repair mechanism for a double stranded break in an organism from the Eukaryotic domain (Popova *et al*, 2011).

Pre-Synapsis

During pre-synapsis, double stranded DNA (dsDNA) is resected in a 5'-3' direction generating ssDNA with 3' overhangs. Recombinases are then loaded onto these 3' ssDNA overhangs. The machinery used for these differ between domains. Bacteria uses the recombinase RecA. Eukaryotes use the recombinase Rad51, and Archaea the recombinase RadA.

The pre-synapsis stage differs between domains. Bacterial pre-synapsis utilises both RecBCD and RecFOR pathways in order to initiate replication. Initiation via either of these two pathways results in the creation of single stranded DNA (ssDNA) which RecA, bacterial recombinase, can then be loaded onto. Once RecA is loaded it forms a right-handed filament on the ssDNA with six RecA molecules and 18 nucleotides per turn, then the nucleoprotein filament invades the homologous DNA molecule forming a D-loop. During synapsis the RecA filament slides along the dsDNA molecule in search of a homologous sequence (Ragunathan, Liu and Ha, 2012: Rocha, Cornet and Michel, 2005).

Pre-synapsis end resection in Eukaryotes is carried out by the Mre11-Rad50 complex. This complex varies between species for example yeast use a Mre11-Rad50-Xrs2 complex known as MRX whereas mammals use a Mre11-Rad50-Nbs1 complex (MRN). The Mre11-Rad50 component is however conserved. If extensive end resection is required, Exo1 exonuclease or Dna2 nuclease are recruited to process the DNA ends (Bonetti *et al,* 2018).

In Eukaryotes, Rad51 is the homologue of bacterial RecA. Rad51 is loaded onto ssDNA that arise from disruptions to DNA replication or resection of 5' double strand breaks. The loading of Rad51 replaces RPA, a single stranded binding protein. This process is aided by various recombination mediators including; BRCA2, Rad52, Rad54 and Swi5-Sfr2. Similar to the bacterial RecA, Rad51 is loaded onto ssDNA in a right-handed filament with six Rad51 molecules to 18 nucleotides per turn. This filament stretches the ssDNA aiding in an efficient search for homology (Chen *et al,* 2008: Klapstein *et al, 2004*).

Archaea also possess the Mre11 and Rad50 homologous recombination initiation proteins found in eukaryotes. These can be commonly observed in thermophilic Archaea within an operon that also encodes a helicase HerA and a 5' to 3' nuclease NurA. In some species of Archaea, it has been shown that the Mre11-Rad50 complex generates 3' overhangs that allow the HerA-NurA complex to initiate end resection (White, 2011: Constantinesco *et al*, 2004). Contrasting to this, in *Hfx. volcanii* the Mre11-Rad50 complex is suggested to delay the repair of double strand breaks by the homologous recombination pathway. Hence, in *Hfx. volcanii* the Mre11-Rad50 complex could be acting as a control mechanism for entry into the homologous recombination pathway (Delmas *et al,* 2009).

Similar to eukaryotes and Bacteria, the archaeal RPA homologue binds to the 3' ssDNA overhangs after end resection and is replaced by the archaeal recombinase RadA, which forms a nucleoprotein filament formation on the ssDNA. This is process is aided by RadB a paralogue of RadA which acts as a recombinase mediator protein that assists in the formation of the RadA filament (Wardell *et al,* 2017). The deletion of RadA or RadB in *Hfx. volcanii* results in growth, DNA repair and recombination defects; however, the deletion of RadB to a lesser extent (Guy *et al*, 2006).

Synapsis

During the synapsis stage, DNA strand exchange occurs. Strand exchange is catalysed by the following recombinases; RecA in Bacteria, Rad51 in eukaryotes and RadA in Archaea. The recombinase nucleoprotein filament catalyses the interaction between the invading ssDNA and the homologous dsDNA template. In eukaryotes the homology search is assisted by Rad54 and Rhd54 allowing the sliding of ssDNA along the dsDNA template. Once homology is found, the recombinase catalyses strand invasion and D-loop (displacement look) V formation (Kil *et al*, 2000). The DNA synthesis initiated at the site of strand exchange has also been suggested to be able to restart DNA replication.

Post-synapsis

Several different pathways exist for the processing of recombination intermediate products generated by strand exchange. The resolution of these intermediates creates either a crossover product where genetic exchange has occurred, or a non-crossover product which is known as a gene conversion.

Holliday junctions are branched DNA structures that contain four double-stranded arms (McKinney *et al,* 2003). In Bacteria, these junctions are resolved by the RuvABC complex, which is made up of RuvA, RuvB and RuvC, the first two are highly conserved in bacterial species. RuvA constrains the Holliday junction allowing for the helicase RuvB to catalyse the relocation of the branch (Eggleston and West, 2000). RuvC plays a role in making dual symmetric incisions across the Holliday junction intermediate at targeted specific sequences (Iwasaki *et al*, 1991). The resulting cleavage allows for direct ligation of nicked duplexes. As RuvC is less highly conserved, some species utilise RusA instead (Chan *et al,* 1998).

Eukaryotic resolution of Holliday junctions is significantly more complex than the bacterial pathway as multiple Holliday junction resolution pathways are utilised and these vary from species to species. There are many endonucleases suggested to play a role in resolution of Holliday junctions and the resolution process is complex and multi-stepped in eukaryotes, involving a series of sequential nicking stages of the homologous recombination intermediates by the corresponding endonucleases (Schwartz and Heyer, 2011).

In Archaea the Holliday junction endonuclease is Hjc, which has been shown to have similar resolving properties to the bacterial RuvC. The resolvase Hjc cuts the Holliday junction symmetrically (Bolt, Lloyd and Sharples, 2001). In addition to Hjc, *Hfx. volcanii* contains the structure-specific nuclease/helicase Hef. Neither Hjc or Hef are essential in *Hfx. volcanii*. However, when one is deleted the other becomes essential. It has been shown that when *hef* is deleted in combination with *radA*, a highly deleterious effect is observed. By constrast, the deletion of *hjc* and *radA* results in a similar phenotype to that of a single *radA* deletion. It has therefore been suggested that Hjc acts exclusively in homologous recombination whereas Hef acts in a pathway that is able to bypass homologous recombination (Lestini *et al*, 2010).

Recombination-dependent replication

Although DNA replication is typically origin-dependent, several archaeal (and some bacterial) species have been found to initiate replication in the absence of origins of replication. However, bacterial cells that are deleted for *dnaA* tend to show severe growth defects (Kogoma, 1997). When *Hfx. volcanii* has all origins deleted, *radA* a highly conserved gene, from the RecA family of recombinases that is involved in homologous recombination in Archaea, becomes essential (Hawkins *et al,* 2013). This suggests that homologous recombination is an alternative mechanism for initiation of DNA replication (Michel and Bernander, 2014). The *Hfx. volcanii* origin deleted mutant grows 7.5% faster than its wild type counterpart, demonstrating a survival advantage in origin-independent replication. As a result , it has been suggested that origins could be selfish genetic elements which ensure their own replication (Hawkins *et al*, 2013). This observed result appears counter-intuitive to current understanding of evolution. Similar results can also be seen in other archaeal species such as *Thermococcus kodakarensis*, suggesting that this observation in *Hfx. volcanii* is not just a one-off anomalous result (Gehring *et al*, 2017).

DNA replication and nutrient availability

Previous research has linked nutrient availability to DNA replication, via a process of nutritional control. As initiation of replication is coordinated with cell growth and division it is therefore responsive to nutrient availability (Wang *et al,* 2007). It has also been shown that *dnaA* translation in bacteria decreases as nutrients become increasingly scarce. This is an origin associated gene in bacteria. As nutrient availability has been shown to link to these types of genes it is hypothesised that it may have an affect on an organism's ability to replicate independently of the origin of replication (Leslie *et al,* 2015). This is further supported by the lack of observable origin independent replication in a wild type strain.

1.6. Aims

This Chapter aims to:

- Assess recombination-dependent replication in *Hfx. volcanii* in relation to laboratory conditions;
- Observe the effect of nutrient-poor conditions on the growth of *Hfx. volcanii* replicating via recombination-dependent versus origin-dependent replication;
- Test whether the observed growth advantage in *Hfx. volcanii* origin deleted strains is due to rich laboratory media?
- Create a real time growth competition assay using fluorescent proteins to better measure growth between competing strains.

2. Materials and Methods

2.1. Materials

Strains

A multiple microorganisms with various strains were utilised or created in the study, to produce a suitable set of strains for use in a flow cytometry growth competition assay. These strains can be seen in Table 1 and Table 2. The first showing all *Hfx. volcanii* strains, the second showing all *E.coli* strains utilised in the transformation of mutants.

Table 1. *Haloferax volcanii* strains utilised or created in this study

Strain	Reference	Genotype
H26	Allers *et al*, 2004	*ΔpyrE2*
H53	Allers *et al*, 2004	*ΔpyrE2, ΔtrpA*
H54	Delmas *et al,* 2009	*ΔpyrE2, bgaHa*
H121	Allers *et al*, 2004	*ΔpyrE2, ΔtrpA, Δlhr*
H431	Bailey, 2005 unpublished	*ΔpyrE2, Δdna2*
H678	Mullakhanbhai and Larsen, 1975	*Wild type*
H779	Norais *et al,* 2007	*ΔpyrE2, ΔtrpA, Δlhr2*
H781	Norais *et al,* 2007	*ΔpyrE2, ΔtrpA, Δrad25C, Δrad25D*
H1546	Hawkins *et al*, 2013	*ΔpyrE2, ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+*
H2085	Allers 2015 unpublished	*ΔpyrE2, ΔtrpA, Δhel308*
H3696	Lever *et al*, 2017	*ΔpyrE2, ΔtrpA, Δhel308, ΔradB*
H5047	This study	*bgaHa*
H5048	This study	*ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+*
H5119	This study	*bgaHa, pyrE2pyrE2+::[ΔpyrE2pyrE2 psyn.GFP]*
H5120	This study	*bgaHa, pyrE2+::[ΔpyrE2 psyn.mCherry]*
H5121	This study	*bgaHa, pyrE2+::[ΔpyrE2 psyn.mTurq]*
H5122	This study	*bgaHa, pyrE2+::[ΔpyrE2 psyn.YPet]*
H5123	This study	*bgaHa, pyrE2+::[ΔpyrE2 psyn.mScarlet]*
H5124	This study	*ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,pyrE2+::[ΔpyrE2 psyn.GFP]*
H5125	This study	*ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,pyrE2+::[ΔpyrE2 psyn.mCherry]*
H5126	This study	*ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,pyrE2+::[ΔpyrE2 psyn.mTurq]*
H5127	This study	*ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,pyrE2+::[ΔpyrE2 psyn.YPet]*
H5128	This study	*ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,pyrE2+::[ΔpyrE2 psyn.mScarlet]*
H5150	This study	*ΔpyrE2, bgaHa, mCherry*
H5152	This study	*ΔpyrE2, bgaHa, mTurq*

H5154	This study	*ΔpyrE2, bgaHa, mScarlet*
H5156	This study	*ΔpyrE2, ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,GFP*
H5158	This study	*ΔpyrE2, ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,mCherry*
H5160	This study	*ΔpyrE2, ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,mTurq*
H5163	This study	*ΔpyrE2, ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,YPet*
H5164	This study	*ΔpyrE2, ΔtrpA, ΔoriC1, ΔoriC2, ΔoriC3, Δori-pHV4-2::trpA+,mScarlet*
H5165	This study	*ΔpyrE2, bgaHa, GFP*
H5167	This study	*ΔpyrE2, bgaHa, YPet*

Table 2. *Escherichia coli* strains utilised or created in this study

Strain	Genotype	Use
XL-Blue MRF	*endA1, gyrA96 (NalR), lac [F' proAB lacIqZΔM15 tn10 (TetR)], Δ(mcrA)183, Δ(mcrCBhsdSMR-mrr)173, recA1, relA1, supE44, thi-1*	*Dam+* cloning strain for blue/white screening methodologies. Also deficient for restriction endonuclease and recombination
N2338 (GM121)	*F-, ara-14, dam-3, dcm-6, fhuA31, galK2, galT22, hsdR3, lacY1, leu-6, thi-1, thr-1, tsx-78*	*Dam-* mutant used for DNA preparation for *Haloferax volcanii* transformations. (Allers *et al*, 2004).

Plasmids

All plasmids created or utilised throughout this study can be seen in Table 3, these were used throughout the process of strain generation for the proposed new competition assay method. Table 4 shows all oligonucleotides utilised in the construction of some of these plasmids.

Table 3. plasmids utilised or created in this study

Name	Use	dam^{-} strain	Notes
pTA51	Making a pyrE2 deletion	n/a	1.7 kb flanking sequences of *Hfx. vol* pyrE2 cloned into pBR-Nov cut with asp718 + HindIII created by Bitan-Banin *et al* (2003) See Figure. 9
pTA593	restoring pyrE2	n/a	Clone of pyrE2 via PCR of H9 genomic DNA created by Delmas *et al* (2009)
pTA2377	Integration of GFP fluorescent protein at the pyrE2 locus	n/a	pHVID4 Glink GFP created by Duggin *et al (*2015)
pTA2378	Integration of mCherry fluorescent protein at the pyrE2 locus	n/a	pHVID6 Glink mCherry created by Duggin *et al (*2015)
pTA2379	Integration of mScarlet fluorescent protein at the pyrE2 locus	n/a	pHVID7 Glink mScarlet created by Duggin *et al (*2015)
pTA2412	Integration of YPet fluorescent protein at the pyrE2 locus	n/a	pHVID8 Glink YPet created by Duggin *et al (*2015)
pTA2413	Integration of mTurq fluorescent protein at the pyrE2 locus	n/a	pHVID9 Glink mTurq created by Duggin *et al (*2015)
pTA2502	Used as a selective marker for the insertion and promotion of fluorescent proteins	n/a	pTA51 cut with BamHI-HF and inserted with *p.syn* (a strong synthetic promoter from oligos 02235 and 02236) at the pyrE2 locus See Figure. 10
pTA2508	Vector to facilitate the use of mCherry in *Hfx. vol.(dam+)*	pTA2511	PCR amplified mCherry with BamHI and NdeI inserted in pTA2502 under the promoter *p.syn*. See Figure. 11
pTA2531	Vector to facilitate the use of GFP in *Hfx .vol. (dam+)*	pTa2539	PCR amplified GFP with BamHI and NDEI inserted in pTA2502 under the promoter *p.syn*. **See Figure**. 12
pTA2532	Vector to facilitate the use of mScarlet in *Hfx. vol. (dam+)*	pTA2538	PCR amplified mScarlet with BamHI and NdeI inserted in pTA2502 under the promoter *p.syn*. **See** Figure. 13

pTA2533	Vector to facilitate the use of mTurq in *Hfx. vol. (dam+)*	pTA2537	PCR amplified mTurq with BamHI and NdeI inserted in pTA2502 under the promoter *p.syn*. **See** Figure. 14
pTA2534	Vector to facilitate the use of YPet in *Hfx. vol. (dam+)*	pTA2536	PCR amplified YPet with BamHI and NdeI inserted in pTA2502 under the promoter *p.syn*. See Figure. 15

Oligonucleotides

Table 4. Oligonucleotides utilised in this study

Name	**Sequence (5’-> 3’)**	**Notes**
Fluo_GFP_F_NdeI	GGCTCCCATATGAGTAAAGGAGAAGAACTTTTCAC	Used for PCR amplification of GFP with Fluo_R_BamHI. For the creation of pTA2531 and pTA2539
Fluo_mCherry_F_NdeI	GGCTCCCATATGGTCTCGAAGGGCGAGGAGGACAA	Used for PCR amplification of mCherry with Fluo_R_BamHI. For the creation of pTA2508 and pTA2511
Fluo_YPet_F_NdeI	GGCTCCCATATGTCGAAGGGCGAGGAGCTCTTCAC	Used for PCR amplification of YPet with Fluo_R_BamHI. For the creation of pTA2534 and pTA2536
Fluo_mTurq_F_NdeI	GGCTCCCATATGGTCTCGAAGGGCGAGGAGCTCTT	Used for PCR amplification of mTurq with Fluo_R_BamHI. For the creation of pTA2533 and pTA2537
Fluo_mScarlet_F_NdeI	GGCTCCCATATGGTCTCGAAGGGCGAGGCCGTCAT	Used for PCR amplification of mScarlet with Fluo_R_BamHI. For the creation of pTA2532 and pTA2538
Fluo_R_BamHI	GCTGGGGATCCACCGCGCCGAAAAATGCGATGGTC	Used as the reverse primer for all fluorescent PCR reactions
p.synF_BglII	GATCTGAGAATCGAAACGCTTATAAGTGCCCCCCGGCTAGAGAGATCATATGTTTTAGATCTA	Used to create pTA2502 with p.synR_BglII
p.synR_BglII	GATCTAGATCTAAAACATATGATCTCTCTAGCCGGGGGGCACTTATAAGCGTTTCGATTCTCGA	Used to create pTA2502 with p.synF_BglII

2.2. Media

Haloferax volcanii media

All liquid and solid media listed below were stored in the following way: liquid media were kept at room temperature in the dark in order to reduce photodegradation of tryptophan in the broth. Solid media, in the form of agar plates, were stored at 4 °C in sealed bags to reduce desiccation. Before use, plates were dried for approximately 30 minutes to remove water precipitation from storage. The types of media utilised, and their component solutions are as follows:

Component solutions;

- Salt water (30%): 4 M NaCl, 148 mM $MgCl_2.6H_2O$, 122 mM $MgSO_4.7H_2O$, 94 mM KCl, 20 mM Tris.HCl pH7.5.
- Salt water (18%): Made from dilution of 30% salt water with dH20. Add 3 mM of $CaCl_2$ after autoclaving.
- Trace elements: 1.82 mM $MnCl_2.4H_2O$, 1.53 mM $ZnSO_4.7H_2O$, 8.3 mM $FeSO_4.7H_2O$, 200 µM $CuSO_4.5H_2O$. Filter sterilised and stored at 4°C
- 10 x YPC (enough for 10 bottles of media): 5% yeast extract (Difco), 1% peptone (Oxoid), 1% casamino acids, 17.6 mM KOH.
- 10 x CA (enough for 10 bottles of media): 5% casamino acids, 17.6 mM KOH.
- KPO_4 Buffer: 308 mM K_2HPO_4, 192 mM KH_2PO_4, net pH of 7.0
- Hv-Ca salts: 362 mM $CaCl_2$, 8.3% v/v of trace elements, 615 µg/ml thiamine 77 µg/ml biotin.
- Hv-min salts: 0.4 M NH_4Cl, 0.25 M $CaCl_2$, 8% v/v of trace element solution. Stored at 4°C.
- Hv-min carbon source: 10% DL-lactic acid Na_2 salt, 8% succinic acid Na_2 salt·$6H_2O$, 2% glycerol, pH to 7.0 with NaOH. Filter sterilised.

For minimal media components were altered then added to media at the same ratios. This is utilised for media without or with reduced carbon or nitrogen sources for example.

The following types of media were utilised in the study:

- Hv-YPC agar: 1.6% agar (Bacto), 18% SW, microwave to dissolve agar then add, 1 x YPC and autoclave. $CaCl_2$ added prior to pouring.
- Hv-YPC broth: 18% SW, 1 x YPC, autoclave then add 3 mM $CaCl_2$.
- Hv-Ca agar: 1.6% agar (Bacto), 18% SW, microwave to dissolve agar then add, 1 x Ca and autoclave, 0.84% v/v of Hv-Ca salts, 0.002% v/v of KPO_4 buffer (pH 7.0) added prior to pouring.
- Hv-Ca+ broth: 18% SW, 30 mM Tris.HCl pH 7.0. Autoclave then add the following when cool; 1 x Ca, 2.5% v/v of Hv-Min carbon source, 1.2% v/v of Hv-Min Salts, 0.002% v/v of KPO_4 buffer (pH 7.0), 444 nM biotin, 2.5 µM thiamine.

- Hv-min agar: 1.6% agar (Bacto), 18% SW microwave to dissolve agar then add 30 mM Tris.HCL pH7.0. Autoclave then add, 2.5% v/v of Hv-Min carbon source, 1.2% v/v of Hv-Min Salts, 0.002% v/v of KPO_4 buffer (pH 7.0), 2.5 µM thiamine and biotin
- Hv-min broth: 18% SW, 30 mM Tris.HCL pH7.5. Autoclave then add, 2.5% v/v of Hv-Min carbon source, 1.2% v/v of Hv-Min Salts, 0.002% v/v of KPO_4 buffer (pH 7.0), 2.5 µM thiamine and biotin

Other supplements such as tryptophan (trp) uracil (ura), 5-Fluoroorotic acid (5-FOA) can be added to the media after autoclaving if required for selection or other purposes (Allers *et al,* 2010).

Escherichia coli **Media**

Media used for cultivation of *E. coli* is listed below;

- LB (lysogeny broth): 1% tryptone (Bacto), 0.5% yeast extract (Difco), 170 mM NaCl, 2 mM NaOH, pH 7.0. Autoclave and then pour.
- LB agar: 300 ml of LB broth, 1.5% agar (Bacto). Autoclave and then pour.

All *E. coli* media is sterilised via autoclave at 121 °C and stored at room temperature until the addition of supplements such as ampicillin which is added to a final concentration of 50 µg/ml. Media with supplementation is then stored at 4°C.

Other Chemicals and Enzymes

All chemicals unless specified otherwise were purchased from Sigma, enzymes were purchased from New England Biolabs (NEB) and Primers from Eurofins. See relevant methods for specific details on reagents used.

2.3 Methods

2.3.1. General Microbiology

Growth and Storage of *Haloferax volcanii*:

Cultures were plated onto solid media using a sterile serological pipette from glycerol stocks (80% glycerol in 6% saltwater added as 20% v/v to liquid cultures and flash frozen on dry ice before being stored at - 80°C). These plates were then grown in a static incubator (LEEC) at 45 °C for approximately 5 days depending on the strain being cultured, unhealthy genotypes may take longer. Liquid cultures were inoculated from cultures grown on solid media using a sterile platinum loop. Small cultures (<10 ml) were grown at 45 °C with an 8-rpm rotation overnight. Larger liquid cultures up to 600 ml of culture were grown at 45 °C overnight in a shaking incubator (Innova 4330 floor-standing incubator) at 110 rpm. All *Hfx. volcanii* cultures were stored at room temperature for short term use or frozen in glycerol as described above for long term.

Transformation of *Haloferax volcanii:*

Transformations methods for *Hfx. volcanii* utilising PEG600 allow for easy and efficient transformations (Cline *et al*, 1989). This involves passing DNA through a *dam- E. coli* host strain prior to the transformation process itself. This is required as *Hfx. volcanii* encodes for a restriction endonuclease known as Mrr which targets and breaks down methylated DNA. *Hfx. volcanii* strains deficient in Mrr can be directly transformed with *dam+* plasmid DNA. A *Hfx. volcanii* culture in 5-10 ml of Hv-YPC broth (+Thy if required) was grown over night at the conditions mentioned earlier until cell growth reached A_{650}=0.6-0.8. Cells were then transferred to a 15 ml round bottomed tube and pelleted by centrifugation at 3300 *xg* for 8 minutes. The supernatant removed and the cells resuspended in 1 ml buffered spheroplasting solution (1 M NaCl, 27 mM KCl, 50 mM Tris.HCl pH 8.5, 15% sucrose). Cells were then transferred to a sterile 2 ml round bottomed tube and pelleted once again. The supernatant once again removed, and cells were resuspended in 400-800 µl buffered spheroplasting solution. A 200 µl aliquot per transformation was then transferred to a new 2 ml round bottomed tube. A 20 µl drop of EDTA pH 8.0 (Ethylenediaminetetraacetic acid) was pipetted onto the side of the tube before gently inverting and being left to incubate at room temperature for 10 minutes. Transforming DNA (5 µl 0.5 M EDTA pH 8.0, 15 µl unbuffered spheroplasting solution (1 M NaCl, 27 mM KCl, 15% sucrose, pH 7.5) and 10 µl DNA (~1-2 µg) was then added in a similar manner to the EDTA and left to incubate for a further 5 minutes at room temperature. Then 250 µl of PEG600 (60% Polyethylene Glycol 600: 150 µl PEG600, 100 µl unbuffered spheroplasting solution) was added to the side of the tube and mixed by gentle inverting and allowed to incubate for 30 minutes at room temperature. Following this 1.5 ml of spheroplast dilution solution (23% SW, 15% sucrose, 37.5 mM $CaCl_2$.) was added, mixed in the same manner as previous steps and incubated for 2 minutes again at room temperature. The mixture was then centrifuged at 3300 cg for minutes at 25 °C to form a pellet. The pellet was then transferred whole into a sterile 4 ml tube containing 1 ml regeneration solution (18% SW, 1 x YPC, 15% sucrose, 30 mM $CaCl_2$). The cells were left to recover undisturbed at 45 °C for 90 minutes. The pellet was then resuspended via gently tapping on the side of the tube and left to incubate for a further 3-4 hours at 45 °C and an 8rpm rotation. Cells were once again transferred to a 2 ml round bottomed tube and pelleted at 3300 *xg* for 8 minutes at 25 °C. The supernatant removed and the pellet resuspended in 1 ml of transformation dilution solution. Serial dilutions were made and 100 µl of each dilution was plated on to appropriate media and left to grow at 45 °C for 5 days.

Growth and storage *Escherichia coli*:

Cultures were plated onto solid media using a sterile serological pipette from glycerol stocks (80% glycerol added as 20% v/v to liquid cultures and flash frozen on dry ice before being stored at – 80 °C). Cultures plated on solid media were then grown in a static incubator (LEEC) at 37 °C overnight. Similarly, to *Hfx. volcanii* cultures, liquid *E. coli* cultures were inoculated from cultures grown on solid media using a sterile loop. Small cultures (<10 ml) were grown at 37 °C with an 8-rpm rotation overnight and large quantities in an Innova 4330 floor-standing incubator at 110 rpm and 37 °C overnight. All cultures were stored at 4 °C for short term storage and frozen in a glycerol stock as aforementioned for long term.

Transformations *Escherichia coli:*

Electrocompetent cells must first be prepared before a *E. coli* transformation can be conducted. These are prepared for two different *E. coli* strains; XL-1 Blue (*dam+)* and N2338(*dam*-). A 5 ml culture was grown overnight 37 °C with an 8-rpm rotation and an appropriate antibiotic selection. Cells are then diluted 1/100 in LB broth supplemented with the selected antibiotics before being grown at 37 °C with an 8-rpm rotation to an optical density (A_{650}) of 0.5-0.8. Cells were then pelleted in a centrifuge (Eppendorf 5417R) at 4 °C and 6000 *xg* for 12 minutes. The supernatant was then discarded, and the pellet resuspended in an equal volume of 1 mM ice cold HEPES (pH 7.5). This process was then repeated using two thirds the volume and then one third of the volume of 1 mM HEPES (pH 7.5). Following this 0.1 volume of 1 mM HEPES and 0.001 volume 1 mM HEPES but with the addition of 10% glycerol to both steps. Cells were then aliquoted into 100 µl cultures and frozen on dry ice before being stored at -80°C.

Then 1-2 µg of DNA was suspended in 5 µl of dH_2O and added to 40 µl of electrocompetent cells, keeping this on ice. An electroporation cuvette (GENEFLOW 1 mm gap) was added to the ice to chill while the DNA and cells were mixed via gently pipetting up and down. Once chilled the cuvette was filled and placed in an *E. coli* gene pulser (BioRad) and pulsed at 1.8kV. 1 ml of SOC broth (2% tryptone (Bacto), 0.5% yeast extract (Difco), 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose) was immediately added. Samples were then incubated at 37 °C with an 8-rpm rotation for 1 hour, allowing for cell recovery before being plated onto LB + ampicillin agar plates and grown over night at 37°C.

2.3.2. DNA Extraction

Plasmid extraction from *Escherichia coli*:

Plasmid DNA extraction from *E. coli* was conducted using NucleoSpin plasmid mini and Nucleobond Xtra midi kits from Macherey-Nagel. The protocol was followed as describe by the manufacturer's guidelines. For minipreps 1-2 ml of *E. coli* culture in LB broth + ampicillin was used and eluted with 30 µl elution buffer. Whereas for midi preps 300 ml culture was used and 200 µl TE was used for resuspending. The DNA was then ethanol precipitated and resuspended in 200 µl of TE (10 mM Tris.HCl pH 8.0, 1 mM EDTA) before being stored at -20°C.

Genomic DNA extraction by spooling from *Haloferax volcanii*:

A 5 ml culture was grown overnight in Hv-YPC (+Thy) at 45 °C until A_{650}= 0.6-0.8. 2 ml of culture was then transferred to a 2 ml round-bottomed tube and centrifuged at 3300 *xg* for 8 minutes at 25 °C. The supernatant was removed, and the pellet resuspended in 200 µl of ST buffer (1 M NaCl, 20 mM Tris.HCl pH 7.5.) 200 µl of lysis solution (100 mM EDTA pH 8.0, 0.2% SDS.) was added, the tube was then mixed via inversion and the cell lysate overlaid with 1 ml of 100% EtOH. DNA was spooled at the interface using a capillary tip until the liquid was homogenous and clear. The spool of DNA was then washed twice in 100% EtOH and allowed to air dry. The DNA was then suspended in 500 µl of TE and isopropanol prepped before being centrifuged at 11,000 *xg* for 5 minutes then washed in 1 ml 70% EtOH and dried thoroughly to remove excess EtOH. The pellet was then resuspended in 100-500 µl of TE and stored at 4 °C.

2.2.3. Nucleic Acid Manipulation

Polymerase chain reaction amplification:

DNA amplification was carried out using either Q5 Hotstart or Onetaq (NEB). These enzymes are suitable for genomes with a high GC content hence their selection with Q5 Hotstart being used for high fidelity amplifications. All PCR reactions were carried out using a Techne Tc-512 thermocycler(Tables 5 and 6).

Table 5. PCR components for each enzyme

OneTaq	Q5 Hot Start
200 µM of dNTP's	200 µM of dNTPs
0.5 µM of each primer	0.5 µM of each primer
10ng of template DNA	1ng of genomic DNA or plasmid DNA template
1 x OneTaq GC Buffer	1 x Q5 Reaction Buffer
0.025 U/µl OneTaq	1 x Q5 High GC Enhancer
-	0.02 U/µl Q5 Hotstart

Table 6. PCR cycle conditions for each enzyme

Step	OneTaq	Q5 Hot Start
Initial Denaturation	94 °C for 30 seconds	98 °C for 30 seconds
Denaturation	94 °C for 30 seconds	98 °C for 10 seconds
Annealing	Tm °C for 30 seconds	Tm °C for 10-30 seconds
Extension	68 °C for 60 seconds per kb	72 °C for 30 seconds per kb
Final extension	68 °C for 5 minutes	72 °C for 5 minutes

Annealing temperatures for primers (Tm ^{0}C) in Table 6 were calculated using the following equation; $81.5 + (16.6\text{x}log_{10}[Na^{+}]) + (0.41x\%GC) - (100 - \%homology) - \left(\frac{600}{l}\right)$ where % GC is the percentage of guanine and cytosine in the primer, % homology is the percentage of homology shared between the primer and the template and L is the length the primer in bases.

Annealed Oligos:

To anneal, 20ul of appropriate oligos were added (p.synF_BglII and p.synR_BglII for *p.syn)* to 10ul NEBuffer 2 (10 mM Tris-HCl, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTT, pH 7.9) and 50ul dH_2O for a 100ul reaction. This was then boiled (100°C) for 15 minutes and allowed to cool to room temperature, then stored at 4 °C overnight. Annealed oligos can then be used in the ligation process (see below).

Restriction Digests:

Restriction digest conditions varied depending on the enzyme and type of DNA used. If two enzymes were required, buffers were selected in which both enzymes had at least 75% activity. All enzymes and buffers were purchased from New England Biolabs (NEB). In all reactions, enzymes added did not exceed 10% of the reaction mixture. See Table 7 for more detail.

Table 7. Restriction digest components for each DNA type

	Mini prepped DNA	Maxi prepped DNA	PCR Genomic DNA
DNA	1-2 µg	5 µg	Approximately 41 µl
Buffer (10x)	2	2	5
Enzyme 1	1	1	1

Enzyme 2 (if needed)	1	1	1
SDW	Enough to make up to total volume	Enough to make up to total volume	Enough to make up to total volume
Total volume	20 µl	20 µl	50 µl

Dephosphorylation of vector DNA:

The ability of vector DNA to self-ligate was prevented via the dephosphorylation of 5' phosphate groups. This was performed using a mix of Shrimp alkaline phosphatase (rSAP; NEB). 1 µl of rSAP was added to any digest that required dephosphorylation and the digest was then left to incubate for 30 minutes at 37 °C before being heat inactivated for 10 minutes at 65 °C.

DNA ligations and ethanol precipitation:

Ligations of DNA were conducted using T4 DNA ligase. 1 µl of T4 ligase and 5 µl of T4 ligase buffer were added to each 50 µl ligation reaction along with ratio of >3:1 insert to vector DNA the rest of the volume was then made up with sterile distilled water. Ligations were carried out at 15 °C overnight before being ethanol precipitated for transformation. To ethanol precipitate the DNA, 2 volumes of 100% EtOH and 1/10 volume of 3M sodium acetate (pH 5.2) were added to the DNA and incubated at -20 °C for minimum of 1 hour. Samples were then centrifuged at 20,000 *xg* and 4 °C for 30 minutes. The supernatant was then removed, and the pellets were washed in 400 µl of 70% EtOH followed by another centrifugation step at 20,000 *xg* and 4 °C for 10 minutes. The supernatant was once again removed, and the pellets air dried before being resuspended in sterile dH$_2$0.

Agarose Gel Electrophoresis:

Casting and running of agarose gels utilised both TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA pH 8.0.) and TAE (40 mM Tris, 20 mM acetic acid, 1 mM EDTA pH 8.0.) with TBE being used for the majority of gels and TAE only being used for southern blotting or where high quality resolution gels were required. Gels were made using agarose powder (SeaKem Lonza) and the appropriate buffer (TBE or TAE). Gel loading dye (50 mM Tris.HCl, 100 mM EDTA pH 8.0, 15% Ficoll (w/v), 0.25% Bromophenol Blue (w/v), 0.25% Xylene Cyanol FF (w/v).) was added to DNA samples up to a final concentration of 1 x. All samples and molecular markers (1 kb NEB ladder or 100 bp NEB ladder) were loaded into the gel. TBE gels were run at 110V for approximately 1 hour. The larger 25cm TAE gels were run over night for 16 hours at 50V with a buffer circulation pump in place. For visualisation of bands, gels were stained with SYBR Safe (Invitrogen) at a final concentration of 0.5 x or for southern blots with ethidium bromide to a concentration of 0.5 µg/ml. To extract DNA from agarose gels, sample lanes were protected using foil while the appropriate band was removed with a scalpel and placed into a 2 ml round bottomed tube. DNA was visualised using a UV transilluminator (UVP inc.) and then purified (See nucleic acid purification section).

Nucleic acid purification:

Nucleic acids from ligations, restriction digests, dephosphorylation of DNA and PCR products were purified using Macherey-Nagel DNA purification kits. The protocol was followed according to the manufacturer's instructions.

2.3.4 Genetic manipulation of *Haloferax volcanii*

Creating a gene replacement construct:

All gene replacement constructs were made by inserting a gene of interest and its flanking regions, alongside an inducible promoter or tag into a plasmid such as pTA51 or a similar derivative. The protocol for this may vary from construct to construct but the premised remains the same. More detail on each construct created will be given in the relevant chapters. See Figure 7 for a summary of the method and Bitan-Banin *et al* (2003) for further details.

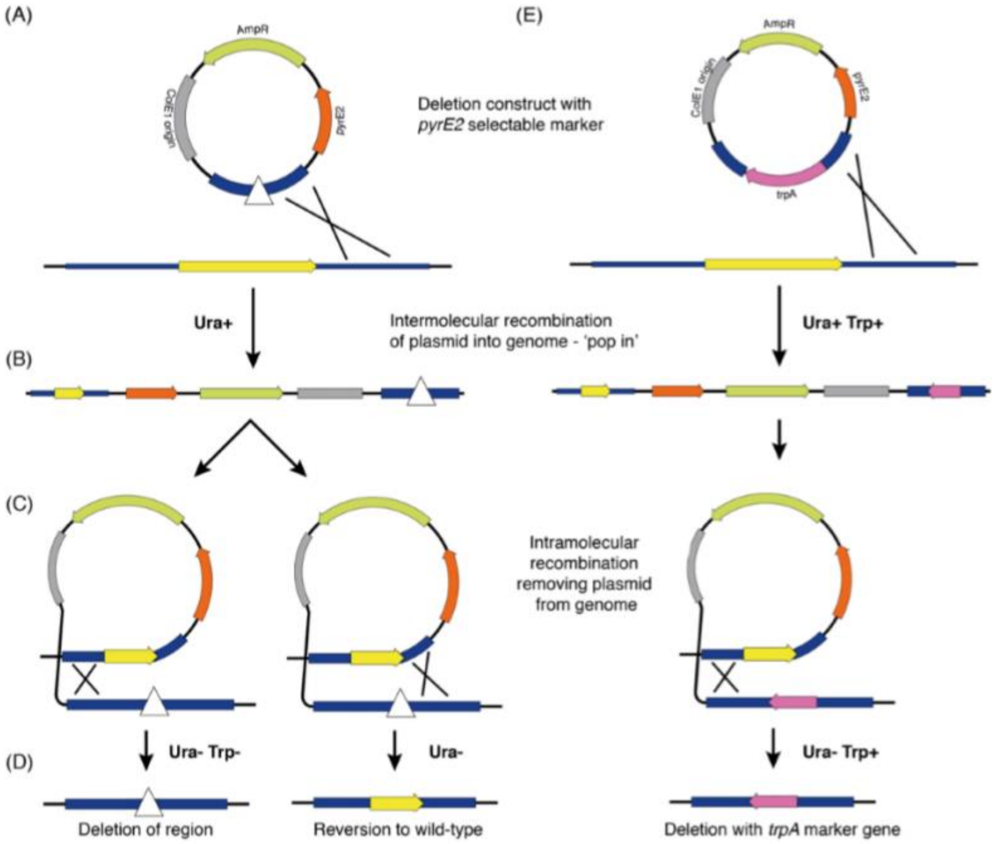


Figure 7. Example gene deletion construct. (A) A Δ*pyrE2* strain is transformed with a pyrE2+ deletion construct. (B) Pop-in colonies plated on ura+ media as a selective pressure. Cells then undergo pop-out when the uracil selection is removed which can be screened for by plating on 5-FOA. (C) Recombination either occurs upstream or downstream as indicated by the X and the direction of the gene arrows, the left diagram being upstream and the right downstream. (D) The gene is either deleted or reverts to its wild type. (E) shows an alternative method using a *trpA* (tryptophan marker) to directly select for the deletion of a gene.

2.3.5 Southern blots, Colony lifts and Radiation

Southern Blotting/Vacuum Transfer:

Hfx. volcanii DNA is first purified as previously mentioned, this DNA is then digested with appropriate restriction enzymes cutting either side of the region of interest. The digested DNA was then separated using a 200 ml 0.75% TAE agarose gel run at 50 V for 16 hours with buffer circulation. The gel was then post stained with ethidium bromide (see agarose gel electrophoresis methods for details) for 30 minutes while gently shaking before being visualised. The gel was then acid nicked for 20 minutes in 0.25M HCl, followed by a 10-minute wash step in sterile dH_20. The DNA was then denatured in a denaturing solution (1.5 M NaCl, 0.5 M NaOH.) for 45 minutes. A membrane (Amersham Hybond-XL) was then soaked in dH_20 for 5 minutes before being equilibrated in denaturing solution for 2 minutes. The vacuum transfer was conducted using a Vacugene XL gel blotter and a Vacugene pump (Pharmacia Biotech) for 1 hour and 1 minutes at 50 mBar. Post transfer the membrane was washed in 2 x SSPE (20 x SSPE: 3 M NaCl, 230 mM NaH_2PO_4, 32 mM EDTA, pH 7.4) for 30 seconds and air dried. DNA was then crossed linked using 120mJ/cm^2 of UV.

Colony lift:

In order to perform a colony lift agar plates were chilled at 4 °C for 30 minutes to ensure the agar had hardened. An 82 mm filter was then rolled onto the surface of the plate from the centre outwards and positions were asymmetrically marked on the filter using a needle. After 2 minutes of allowing the filter to rest on the agar, the filter was removed with forceps and placed colony side up on blotting paper soaked in 10% SDS (sodium dodecyl sulphate) this was left to incubate for 10-15 minutes. The filter was then removed and placed on blotting paper soaked in denaturing solution (1.5 M NaCl, 0.5 M NaOH.) colony side up for 15 minutes before being placed on another piece of blotting paper soaked in neutralising solution (1.5M NaCl, 0.5M Tris.HCl,1 mM EDTA) for 5 minutes. This step was then repeated with fresh blotting paper and neutralisation solution.

The filter was then washed in 2x SSPE (20 x SSPE: 3 M NaCl, 230 mM NaH_2PO_4, 32 mM EDTA, pH 7.4) for 30 seconds before being dried and crosslinked with UV.

Hybridisation with radioactive probes:

Membranes from colony lifts or Southern blots were first pre-hybridised for >3 hours at 65 °C in 40 ml of pre-hybridisation solution (6 x SSPE, 1% SDS. 5 x Denhardt's solution, 200 µg/ml salmon sperm DNA, boiled for 5 minutes at 100 °C prior to addition). Radiolabelled DNA probes were then made using 50ng of DNA and 0.74 MBq of [α-^{32}P] dCTP (Perkin Elmer). The DNA was denatured at 100 °C for 5 minutes before being incubated with HiPrime (a random priming mix from Roche) for 15-20 minutes at 37 °C. The radiolabelled probe was then purified using a BioRad P-30 column and mixed with 10 mg/ml salmon sperm DNA. This mix was then denatured at 100 °C for 5 minutes before being quenched on ice. For Southern blots 3 µl of 1 µg/ml 1 kb ladder was also added to the radiolabelled mix. The pre-hybridisation solution was then discarded and replaced with 30 ml of hybridisation solution (6 x SSPE, 1% SDS, 5% dextran sulphate) the probe DNA was added alongside the membranes and were left to incubate 65 °C overnight. The membranes were then washed twice with 50 ml low stringency wash solution (2 x SSPE, 0.5% SDS.), first for 10 minutes then for 30 minutes. This was then followed by another two washes using high stringency wash solution (0.2 x SSPE, 0.5% SDS) each for 30 minutes. The membranes were air-dried before being encased in plastic wrap and exposed to a phosphorimager screen for 24 hours before being visualised on a phosphor-imager.

2.3.6 Competition Assays and Flow Cytometry

Blue/white competition assay:

Blue white competition assays allow for analysis of growth rates of two different strains in a competing environment. This method utilises the *bgaH* ß-galactosidase gene which allows for the detection of blue colonies upon treatment with X-gal (Holmes and Dyall-Smith, 2000). A 5 ml culture of YPC (+Thy) was set up from colonies grown over night on solid media. This culture was grown over night at 45 °C and 8 rpm rotation. A 5 µl, 10 µl and 20 µl aliquot was then transferred to three fresh 10 ml YPC (+Thy) cultures which were left to grow over night once again. On the third day when the A_{650}=0.4 serial dilutions of the cells were made ranging from 10^0 to 10^{-6}. 100 µl of 10^{-5} cells were plated on YPC. Then another 10 ml of YPC (+Thy) was inoculated with 100 µl of 10^{-4} cells from both WT and mutant strains and left to incubate for 2 days. This process of diluting plating and inoculating was repeated every 2 days from day 3 to 11. After the inoculated YPC plates had been growing for 5 days they were sprayed with X-Gal and incubated overnight. Then the number of blue and white (red in the case of *Hfx. volcanii*) colonies were counted and recorded.

Fluorescent imaging:

Single colonies of culture were taken using a sterile inoculation loop and mixed with 1 ml 18% SW before being loaded into a 48 well plate (Corning Inc). Images were then taken under the frequencies labelled cy3 (548nm to 561nm) and cy2(488nm to 506nm) of light using a typhoon phosphor-imager as these were the best fitted frequencies available. Cells were serial diluted if required.

2.3.7 Gradient plates

To generate a nutrient gradient across a plate square plates were first poured with a 7° gradient with Hv-Ca/min (+ the desired concentration of supplements) to create a wedge. Once set the plates were then poured over with 43 ml of Hv-Ca/min without the added supplements to form a gradient tapering to zero (*Hawkins et al,* 2013 and Figure 8).

A 5 ml culture of *Hfx. volcanii* strains were grown at 45 °C at 8rpm rotation in Hv-YPC until A_{650}=0.6-0.8. These were then diluted and regrown in fresh Hv-YPC until A_{650}=1.0. Serial dilutions in 18% SW were prepared to 10^{-4}. Sterile paint brushes were then soaked in 18% SW before being dipped in culture and painted in a straight line across the plate. The brush was then re-dipped and painted across the same line in the opposite direction. The plates were then left to dry and incubated at 45 °C for 5 days.

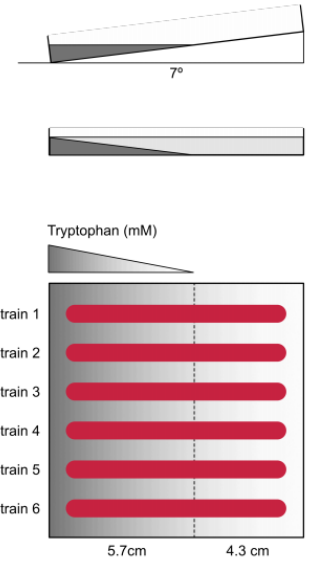


Figure 8. Gradient plates. Plates were pouted with 17 ml of Hv-min with the addition of any desired supplements on a 7° slant to form a tapered wedge. Once set the plate was placed flat and the wedge was covered with 43 ml Hv-min lacking the supplement. Strains were then painted across the plate (Lever, 2019)[82].

3.Results

3.1. Generation of fluorescent marked strains

Fluorescent marked strains were generated to replace the *bgaH* beta gal reporter method used in blue/white screening. This was required due to the large amount of time needed to complete a blue/white screen, making it an inefficient method for testing large quantities of minimal growth conditions in a competitive manner, with limited time. The fluorescent proteins; GFP, mCherry, mScarlet, YPet and mTurq were chosen for use. Firstly, as the fluorescent markers had already been adapted for use in *Hfx. volcanii* via a series of amino acid substitutions allowing for use in a halophilic cytoplasm (Duggin *et al,* 2015). Secondly, because of the range of absorption and emission frequencies across these proteins. This range increases the chance that two markers needed for the competition assay will be easy to distinguish from each other and able to be detected via flow cytometry. The *pyrE2* locus was targeted for use in the pop-in/pop-out method using uracil and 5-FOA as selection to create *Hfx. volcanii* strains.

3.1.1 Plasmid construction

Gene replacement constructs were made by first inserting *p.syn*, a 43 bp strong constitutive synthetic promoter based on the *Hfx. volcanii* consensus tRNA promoter sequence, into the pyrE2 *locus of* pTA51 to create pTA2502 (Large *et* al, 2007: Haque *et al,* 2019). This promoter was chosen to ensure detectable expression of fluorescent markers and was inserted via annealed oligos. The fluorescent marker from the plasmids provided by Duggin *et al* (2015) were then amplified via PCR to ensure enough fluorescent marker DNA was present for successful ligation.

pTA51

The *pyrE2* deletion construct pTA51 was created prior to the study by Bitan-Banin *et al* (2003). This plasmid was selected for use as a *pyrE2* deletion construct due to the restriction digest sites present matching those of the fluorescent protein plasmids from Duggin *et al* (2015) (see Table 3).The selection of *pyrE2* was due to this gene being deleted in the majority of the laboratory strains available and so 5-FOA selection can be used to ensure *pyrE2* replacement with the fluorescent marked genes. Novobiocin was used for selection of dam$^-$ colonies containing the desired plasmid. See Figure 9 for the construct map.

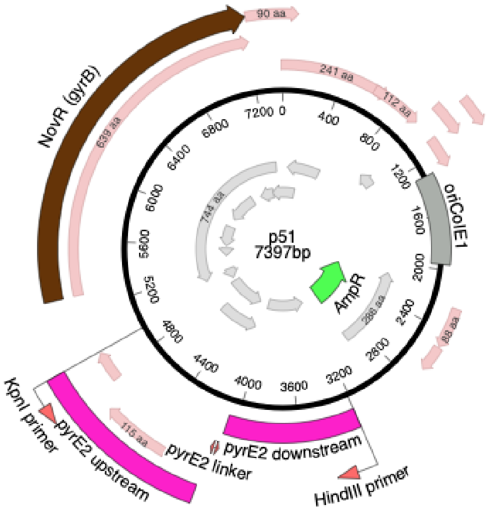


Figure 9. pTA51. The *pyrE2* deletion construct by Bitan-Banin *et al* (2003).

pTA2502

The construct pTA2502 was created during this study from pTA51 by the insertion of the *p.syn, a* strong synthetic promoter via annealed oligos (Haque *et al,* 2019) at the *pyrE2* deletion region, used as a basis for the insertion of all fluorescent markers. (Figure 10, See Table 4 for oligonucleotide details).

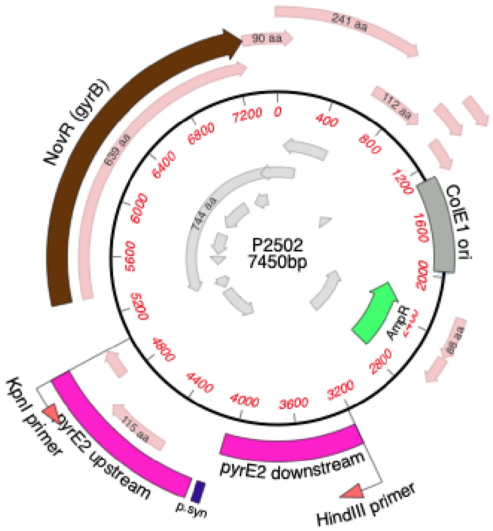


Figure 10. pTA2502. The *pyrE2* deletion construct with the addition of *p.syn*, a strong synthetic promoter (Haque *et al,* 2019).

pTA2508

The construct containing the mCherry fluorescent marker was created using PCR-amplified mCherry from pTA2378 (Duggin *et al,* 2015) which was inserted into pTA2502 downstream of the *p.syn* promoter.

This plasmid was passed through a dam^{-} host to create pTA2511 with novobiocin being used for selection of *dam*$^{-}$ colonies containing the desired plasmid. See Figure 11 for the construct map (See Table 2 for *E.coli* hosts and Table 4 for oligonucleotide details used for PCR amplification.)

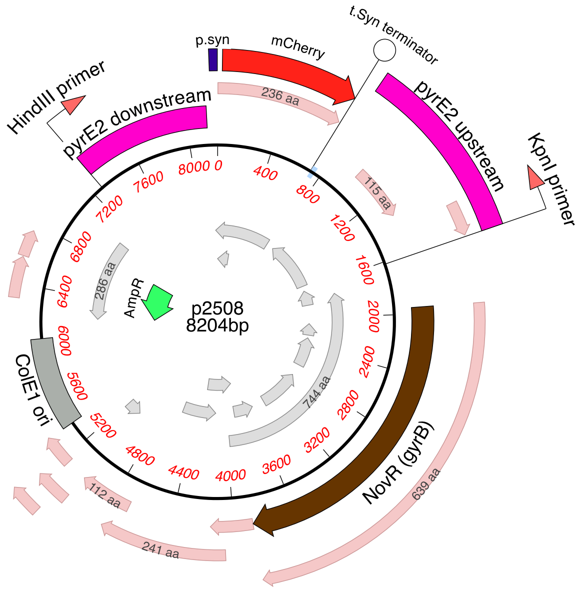


Figure 11. pTa2508. The *pyrE2* deletion construct with the addition of *p.syn* (Haque *et al,* 2019), a strong synthetic promoter and the mCherry fluorescent marker.

pTA2531

The construct containing the GFP fluorescent marker was created using PCR amplified GFP from pTA2377 (Duggin *et al,* 2015) which was then inserted into BamHI/NdeI digested pTA2502 downstream of *p.syn*. Before being passed through a dam$^-$ host to create pTA2539 with novobiocin being used for selection of *dam*$^-$ colonies containing the desired plasmid. See Figure 12 for the construct map and Table 4 for oligonucleotide details.

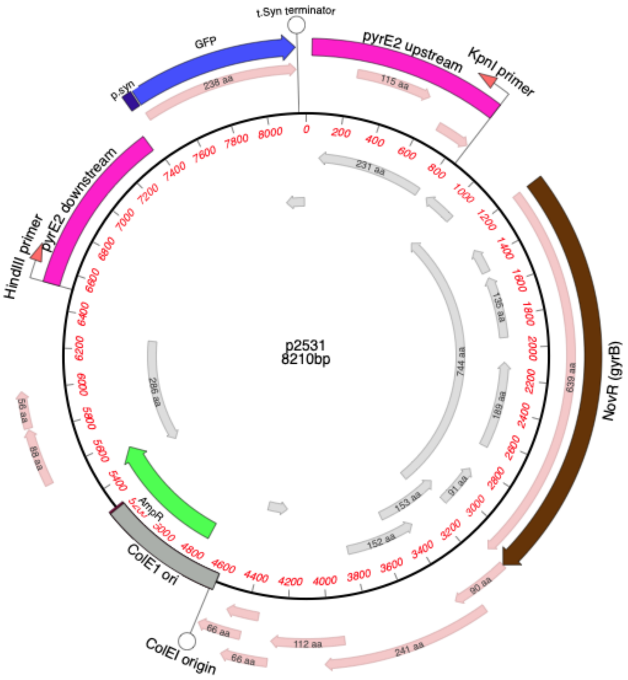


Figure 12. pTA2531. The *pyrE2* deletion construct with the addition of *p.syn* (Haque *et al,* 2019)[84], a strong synthetic promoter and GFP fluorescent marker.

pTA2532

The mScarlet fluorescent marker construct was created using PCR amplified mScarlet from pTA2379 (Duggin *et al,* 2015). This was then inserted into pTA2502 downstream of *p.syn*. The resulting plasmid was then passed through a dam$^-$ host to create pTA2538. Novobiocin was used for selection of *dam*$^-$ colonies containing the desired plasmid (See Table 2 for *E.coli* hosts and Table 4 for oligonucleotide details used for PCR amplification.). See the Figure 13 for the construct map.

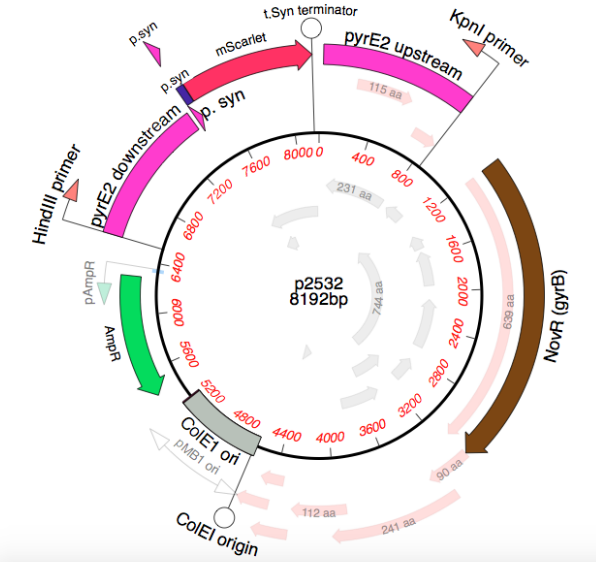


Figure 13. pTa2532. The *pyrE2* deletion construct with the addition of *p.syn* (Haque *et al,* 2019), a strong synthetic promoter and mScarlet fluorescent marker.

pTA2533

The mTurq fluorescent marker construct was created using PCR amplified mTurq from pTA2413 (Duggin *et al,* 2015) which was then inserted into pTA2502 downstream of *p.syn*. This was then passed through a dam^{-} host to create pTA2537 (See Table 2 for *E.coli* hosts and Table 4 for oligonucleotide details used for PCR amplification). Similarly to the previously mentioned constructs novobiocin was used for selection of colonies containing the desired plasmid. See Figure 14 for the construct map.

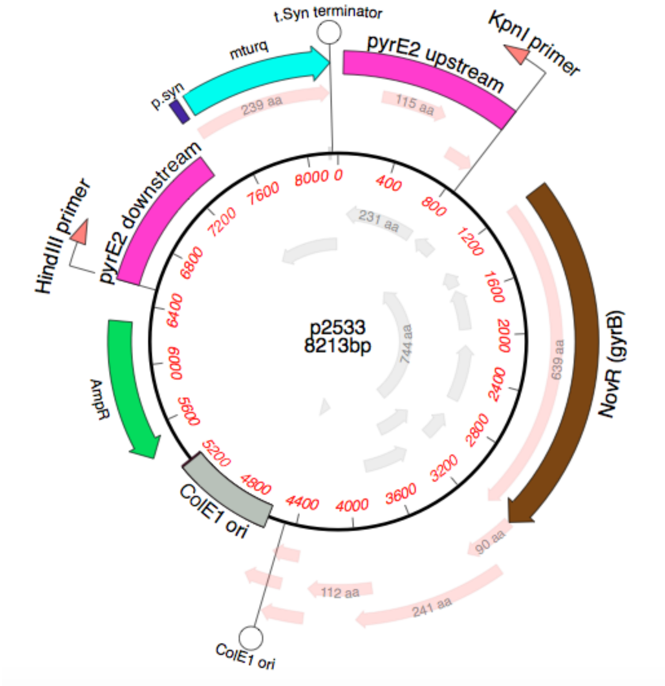


Figure 14. pTA2533. The *pyrE2* deletion construct with the addition of *p.syn* (Haque *et al,* 2019), a strong synthetic promoter and mTurq fluorescent marker.

pTA2534

The YPet fluorescent marker construct was created using PCR amplified YPet from pTA2412 (Duggin *et al,* 2015) this was then inserted into pTA2502 downstream of *p.syn*. Before being passed through a dam$^-$ host to create pTA2536 novobiocin was used for selection of colonies containing this plasmid (See Table 2 for *E.coli* hosts and Table 4 for oligonucleotide details used for PCR amplification). See Figure 15 for the construct map.

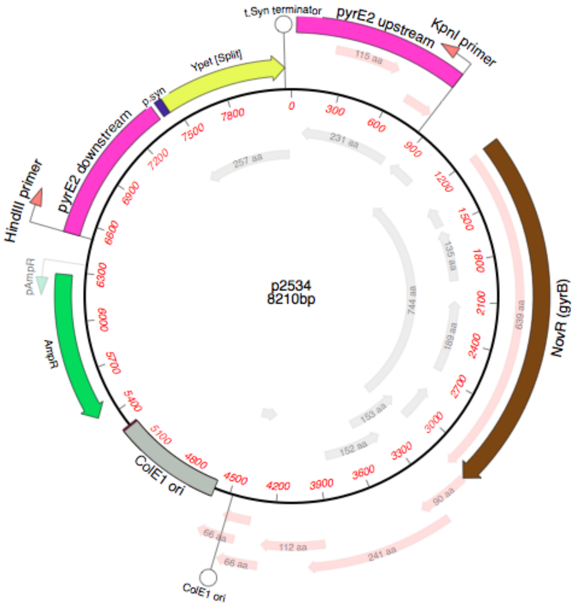


Figure 15. pTA2534. The *pyrE2* deletion construct with the addition of *p.syn* (Haque *et al,* 2019), a strong synthetic promoter and YPet fluorescent marker.

All plasmids were checked by restriction digest and confirmed via DNA sequencing using the dideoxy chain termination method (Sanger *et al*, 1977) by the Deep sequencing unit of the University of Nottingham.

3.2. Generation of *Haloferax volcanii* Strains

Hfx. volcanii strains were generated via transformation with the *dam*$^{-}$ plasmids. Strains H5047 (ori+) and H5048 (Δori) of *Hfx. volcanii*, were transformed with each fluorescent marker construct (pTA2511, pTA2539, pTA2538, pTA2537 and pTA2536) resulting in a wild type (ori+) and origin-deleted strain (Δori) with each fluorescent marker at the pyrE2-deleted locus (see Table 1). The pop-in/pop out method was used to create the deletion. As *Hfx. volcanii* strains may be mero-diploid, meaning they may have a mixture of deleted and wild type alleles present in different chromosomal copies, the resulting strains were also confirmed using a Southern blot with a probe made from pTA2502 digested with SmaI, which will hybridise with a 1157 bp band and a 6293 bp band (see Figure 16).

Strain	Locus	Desired Band Size (Bp)
H1546	Δ*yrE2*	1156,2998,3297
H678	*PyrE2*	1169,631
H5156	*GFP*	1948,6262
H5165	*GFP*	1948,6262
H5158	*mCherry*	1766,224
H5150	*mCherry*	1766,224
h5163	*YPet*	1086,33,321,333,225
h5167	*YPet*	1086,33,321,333,225
h5160	*mTurq*	1088,32,320,332,225
h5152	*mTurq*	1088,32,320,332,225
h5164	*mScar*	1755,225
h5154	*mScar*	1755,225

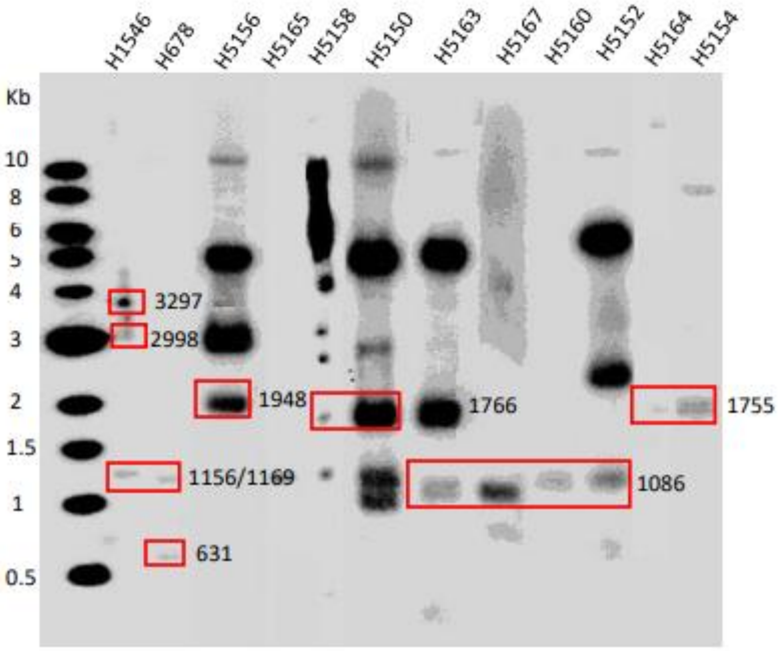


Figure 16. Southern blot confirmation of integration of fluorescent markers into wild type and origin deleted mutants. The table shows the strain names, fluorescent markers present and the size of expected bands. The red boxes highlight these bands on the Southern blot. H5165 has no highlighted bands as the desired band was not present.

Strain H5165 had an incorrect banding pattern and therefore had not successfully taken up the GFP marker. As this is only the case for the wild-type strain, the GFP origin deleted mutant strain could be used for the downstream assays alongside any of the other wild type strains for the competition assay containing a different fluorescent marker. In addition, the Southern blot had a lot of background; this would have been repeated if time permitted.

3.3. Imaging signals and flow cytometry

The fluorescently marked strains were tested for their ability to fluoresce under varying frequencies of light using a phosphor-imager. The fluorescent markers with the most contrasting emission frequencies were chosen for use in the growth competition assay as the southern blot was not clear. Emission frequencies are shown in Figure 17.

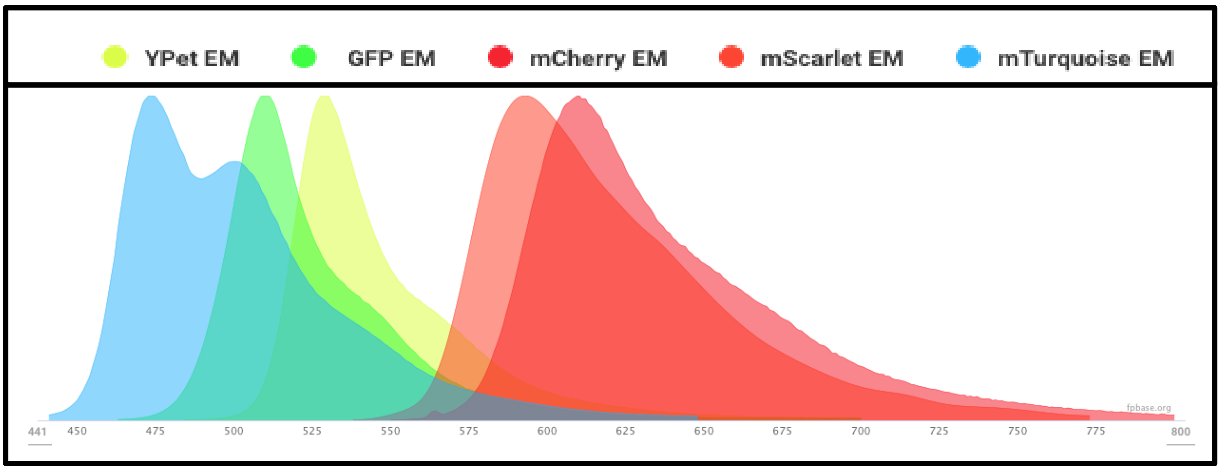


Figure 17. Emission frequencies of each fluorescent marker, GFP and mCherry show the least overlap in the emission spectrum and have distinct peaks so were selected as the best candidates for the competition assay.

The fluorescence tests from the phospho-imager showed strongest signals for GFP, mCherry and YPet strains. See Figure 18 for excitation at Cy2, 488nm to 506nm and for excitation at Cy3, 548nm to 561nm. Dark wells show fluorescence of the relevant strain under these excitation wavelengths. Strong fluorescence was observed for GFP under both excitation wavelengths. mCherry showed moderate amount of fluorescent close to that of GFP at a Cy3 excitation. Ypet showed weak fluorescent for both ori+ and Δori strains under Cy2 excitation. No fluorescence was observed for the other proteins tested under these conditions,

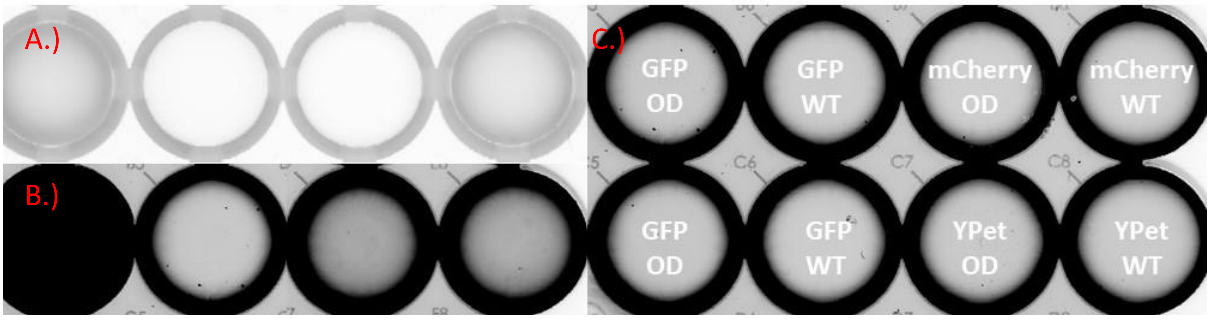


Figure 18. A.) Fluorescent imaging under Cy3 excitation frequencies (548nm to 561nm). B.) Fluorescent imaging under Cy2 excitation frequencies (488nm to 506nm). Dark wells show absorption of wavelengths. C.) A key representing the location of mutant and wild type strains. Under Cy3 excitation, Weak absorption is observed for GFP origin deleted strains and mCherry wild type strains. Under Cy2 excitation Strong absorption was observed for GFP mutants and weak absorption for both wild type and origin deleted YPet strains.

Flow cytometry:

Results for flow cytometry were not able to be collected in the time available before lockdown due to Covid-19. However, the growth competition assay was to be tested first using the pairs with the strongest fluorescence, namely GFP and mCherry at an excitation wavelength of approximately 550nm, or GFP and YPet with a wavelength of approximately 490nm.

The similarity in colour of GFP and YPet was expected to make this a less efficient pairing than GFP and mCherry. As it is required that the two markers selected for the assay can be distinctly distinguished on the flow cytometer. The emission and excitation wavelengths need to differ by a level in which overlap does not occur in the detection of emission wavelengths. Hence, GFP and mCherry is expected to be a better pair for use, although signal strength may be higher for the GFP YPet combination. It may be the case that the signal strength is too weak for cells to be detected or that the cells could not be distinguished from one another. In this case, the established blue/white beta-galactosidase screening method would have been used to test media conditions using a competition assay.

3.4. Media

Preliminarily media testing utilised the gradient plate methods a relatively easy way of testing many stains in parallel (Hawkins *et al,* 2013) (see methods). These tests were conducted in an attempt to establish a promising starting point for minimal media conditions to be used in the growth competition assay. The aim of these tests was to find a condition that reduced growth significantly in a range of *Hfx. volcanii* strains with varying genotypes but that did not show lethality. These conditions would then be used as a priority for the flow cytometry competition assay as it is predicted that origin deleted strains may perform less efficiently than wild type strains under these sub-optimal conditions. All gradient plates were created using Hv-min (see table 9) (1.6% agar (Bacto), 18% SW microwave to dissolve agar then add 30 mM Tris.HCL pH7.0 (Allers *et al,* 2010). This was autoclaved with the subsequent addition of 2.5% v/v of Hv-Min carbon source, 1.2% v/v of Hv-Min Salts, 0.002% v/v of KPO_4 buffer (pH 7.0), 2.5 µM thiamine and biotin) with the appropriate changes to the nutrient source of interest. (See methods, Media 2.2 for further details of components.)

Table 8. *Haloferax volcanii* strains and their corresponding genotypes used for gradient plate tests

Strain	Genotype
H26	*ΔpyrE2*
H53	*ΔpyrE2, ΔtrpA*
H121	*ΔpyrE2, ΔtrpA, Δlhr*
H779	*ΔpyrE2, ΔtrpA, Δlhr2*
H787	*ΔpyrE2, ΔtrpA, Δlhr, Δlhr2*
H431	*ΔpyrE2, ΔtrpA, Δdna2*
H2085	*ΔpyrE2, ΔtrpA, Δhel308*
H3691	*ΔpyrE2, ΔtrpA ,Δlhr, Δlhr2, Δhel308::trpA+*

Hfx. volcanii strains; (Table 8) H26,H53,H121,H779,H787,H431,H2085 and H3691 were chosen for initial media screening by semi-random selection. The strains include a range of different genotypes with some known to grow more slowly than others. All strains are involved in DNA recombination and repair studies.

As salts and trace elements are widely accepted to be close to natural conditions and remain relatively stable, these were not altered. Hence, the initial screens focused on alterations to phosphate, carbon and nitrogen sources in Hv-min media (Dyall-Smith, 2015), conditions which are presumed to change as natural fluctuations in the environment occur such as changes in light levels or the strength of currents.

Table 9. Summary of media alterations

Components

Plate	Carbon Source			Salts	KPO_4 Buffer
	Lactic Acid	Succinic Acid	Glycerol	NH_4Cl	
Control	10%	8%	2%	90 µM	1 µM
Lactic acid absent media	0%	8%	2%	90 µM	1 µM
Succinic acid absent media	10%	0%	2%	90 µM	1 µM
Glycerol absent media	10%	8%	0%	90 µM	1 µM
Phosphate	10%	8%	2%	90 µM	1 µM
$\frac{1}{10}$ **Phosphate**	10%	8%	2%	90 µM	0.1 µM
Nitrate	10%	8%	2%	90 µM	1 µM
$\frac{1}{2}$ **Nitrate**	10%	8%	2%	45 µM	1 µM
$\frac{1}{4}$ **Nitrate**	10%	8%	2%	22.5 µM	1 µM

Carbon sources:

The use of gradient plates allowed for rapid screening of multiple strains simultaneously in a manner which is easy to observe (see methods Figure 9.).

When glycerol was absent from the Hv-min media, a significant decrease in growth for all strains was observed across the plate as the concentration of glycerol decreased. The one exception to this was H2085 which had little to no growth across all repeats and controls. This is probably because this strain is deleted for *hel308* so is known to be slow growing (Gamble-Milner, 2016). H3691 was also observed to be slower growing than other strains across all repeats and controls. The clear reduction in growth as the glycerol concentration dropped suggests this as a strong candidate for further screening in the flow cytometry assay as there is a noticeable decrease in growth without significant lethality to cells. See Figure 19.

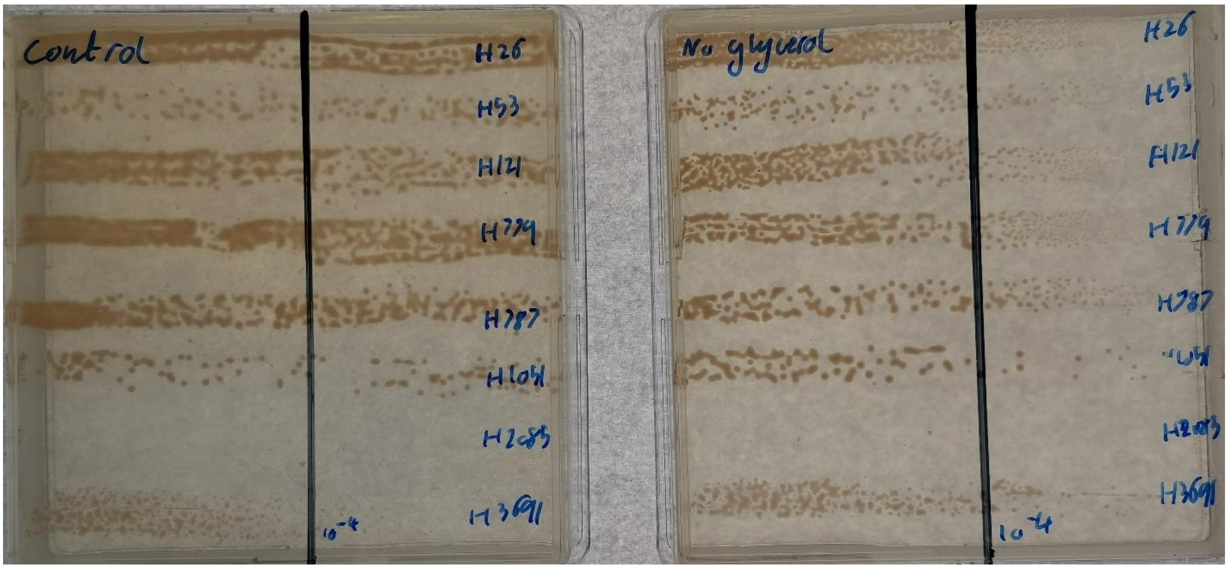


Figure 19. Gradient plates for selected *Hfx. volcanii* strains. The left showing growth under normal glycerol conditions of HV-min media and the right showing a glycerol concentration reducing from left to right, the far right having no glycerol.

Gradient plates lacking lactic acid showed a similar effect to glycerol plates with all strains showing a reduction in growth as the concentration decreased. The effect of reduced or no lactic acid concentration appeared slightly stronger than the effects of removing glycerol, although this cannot be quantified from a gradient plate. Once again, H2085 showed no growth on all repeats and H3696 showed increased growth on one of the lactic acid gradient plates repeats (see Figure 20).

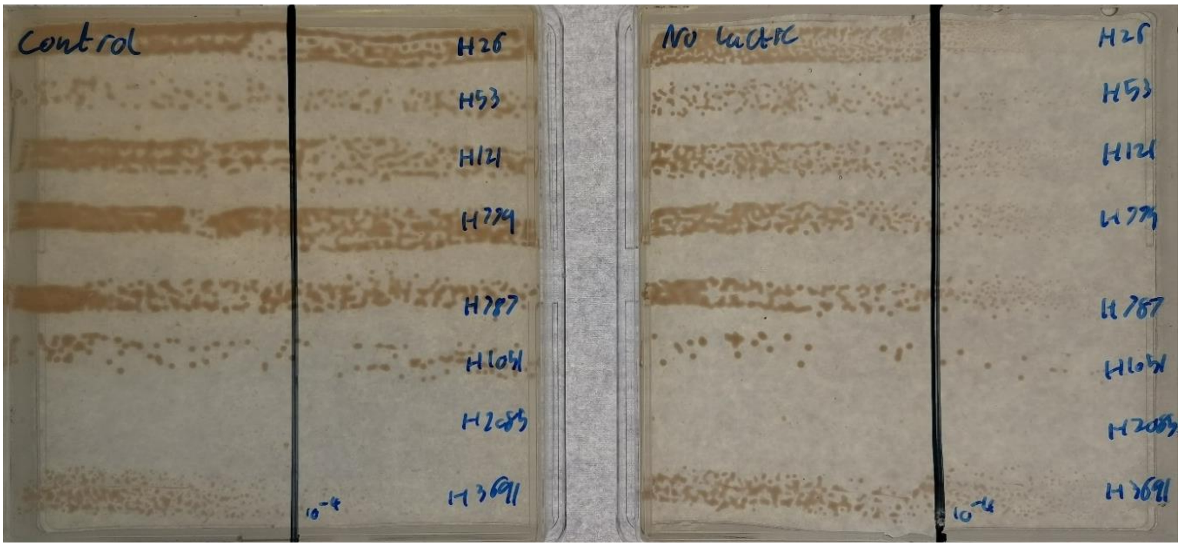


Figure 20. Gradient plates for selected *Hfx. volcanii* strains. The left showing growth under normal lactic acid conditions of HV-min media and the right plate showing lactic acid concentration reducing from left to right, the far right having no lactic acid.

Succinic acid gradient plates once again yielded similar results to the other two carbon sources that were reduced, with growth decreasing as concentration did. Strain H3691 grew better on the succinic acid gradient plate in comparison to the control for one of the repeats (Figure 21).

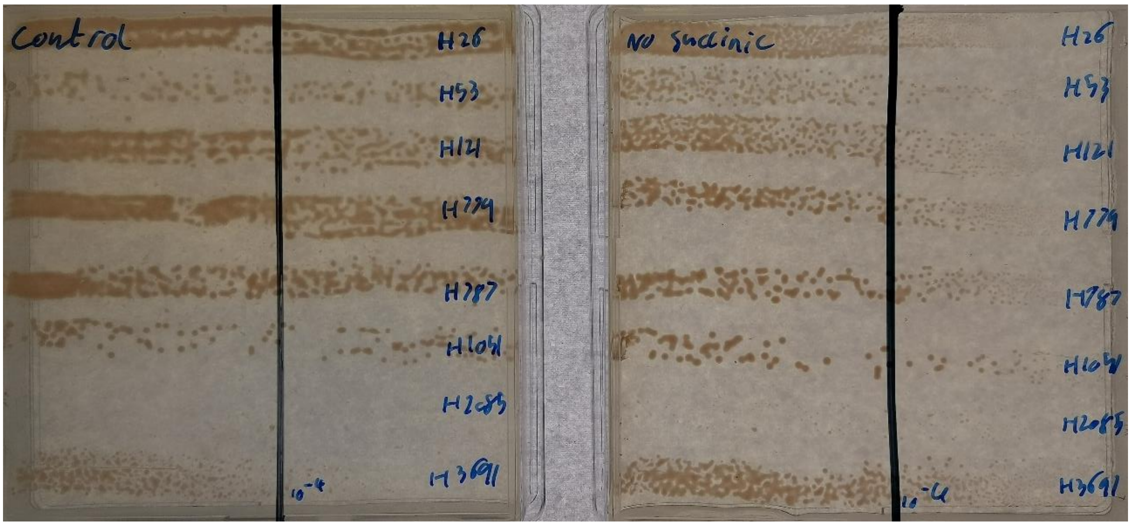


Figure 21. Gradient plates for selected *Hfx. volcanii* strains. The left showing growth under normal succinic acid conditions of HV-min media and the right showing succinic acid concentration reducing from left to right, the far right having no succinic acid.

Phosphate source:

Phosphate gradient plates were created for two different gradients, one ranging from 100% of the standard phosphate source (1 µM KPO_4 buffer) in Hv-min media to no phosphate source (see media section of methods). The other ranging from 10% of the normal level of phosphate in Hv-min media (0.1 µM KPO_4 buffer) to no phosphate. The gradient plate ranging from the normal phosphate level to zero showed a slight decrease in growth as the concentration of phosphate is decreased (see Figure 22.A). Strain H2085 showed no growth on either plate. The gradient plate with 10% the normal phosphate levels showed significantly weaker growth across the plate in comparison to the control. (see Figure 22.B).

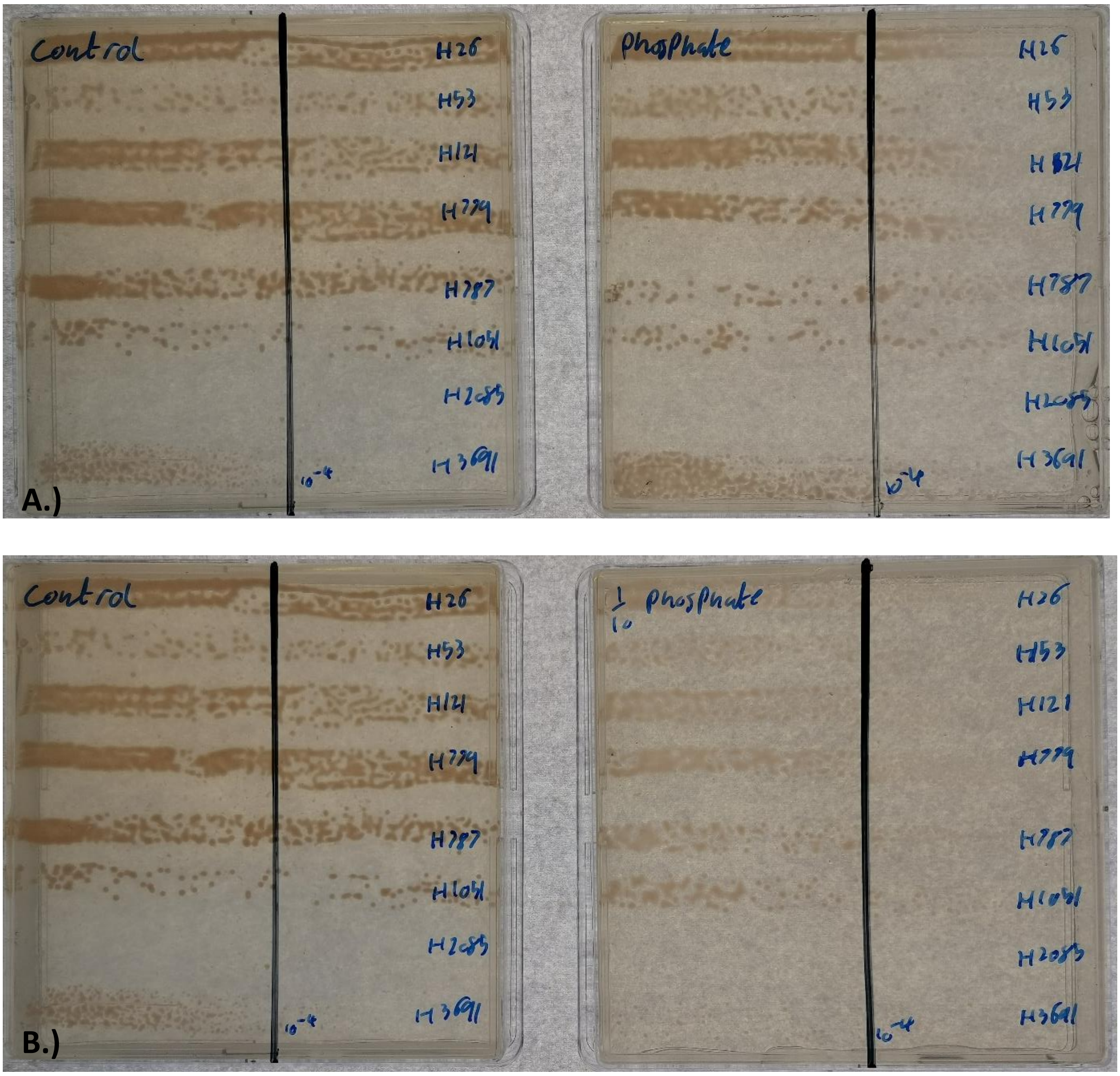


Figure 22. Gradient plates for selected *Hfx. volcanii* strains. **A.)** The left (control) showing growth under normal phosphate conditions of HV-min media (no gradient) and the right showing Phosphate concentration reducing from left to right, the far right having no phosphate. **B.)** The left plate showing controlled conditions normally used in Hv-min media with no gradient and the right plate showing a phosphate concentration ranging from 1/10th of the normal level on the left decreasing in concentration toward the right where no phosphate is present.

Nitrogen source:

Ammonium chloride gradient plates were created for two different concentrations, the first set ranging from half the standard concentration of ammonium chloride (45 µM NH_4Cl) and the second from ¼ of the standard concentration (22.5 µM NH_4Cl). The control plate has the normal nitrogen levels (90 µM NH_4+). The 50% NH_4Cl (45 µM NH_4Cl) gradient plates showed weak growth for all strain in comparison to the controls. Strains H53 and H121 showed weaker growth than other strains as the concentration decreased. H2085 showed no growth on any repeats (See Figure 23.A). The second set of plates with a 25% NH_4Cl (22.5 µM NH_4Cl) gradient showed extremely weak growth with a lack of pigmentation at even the highest concentration range for all strains. H2085 showed no growth (see Figure 23.B).

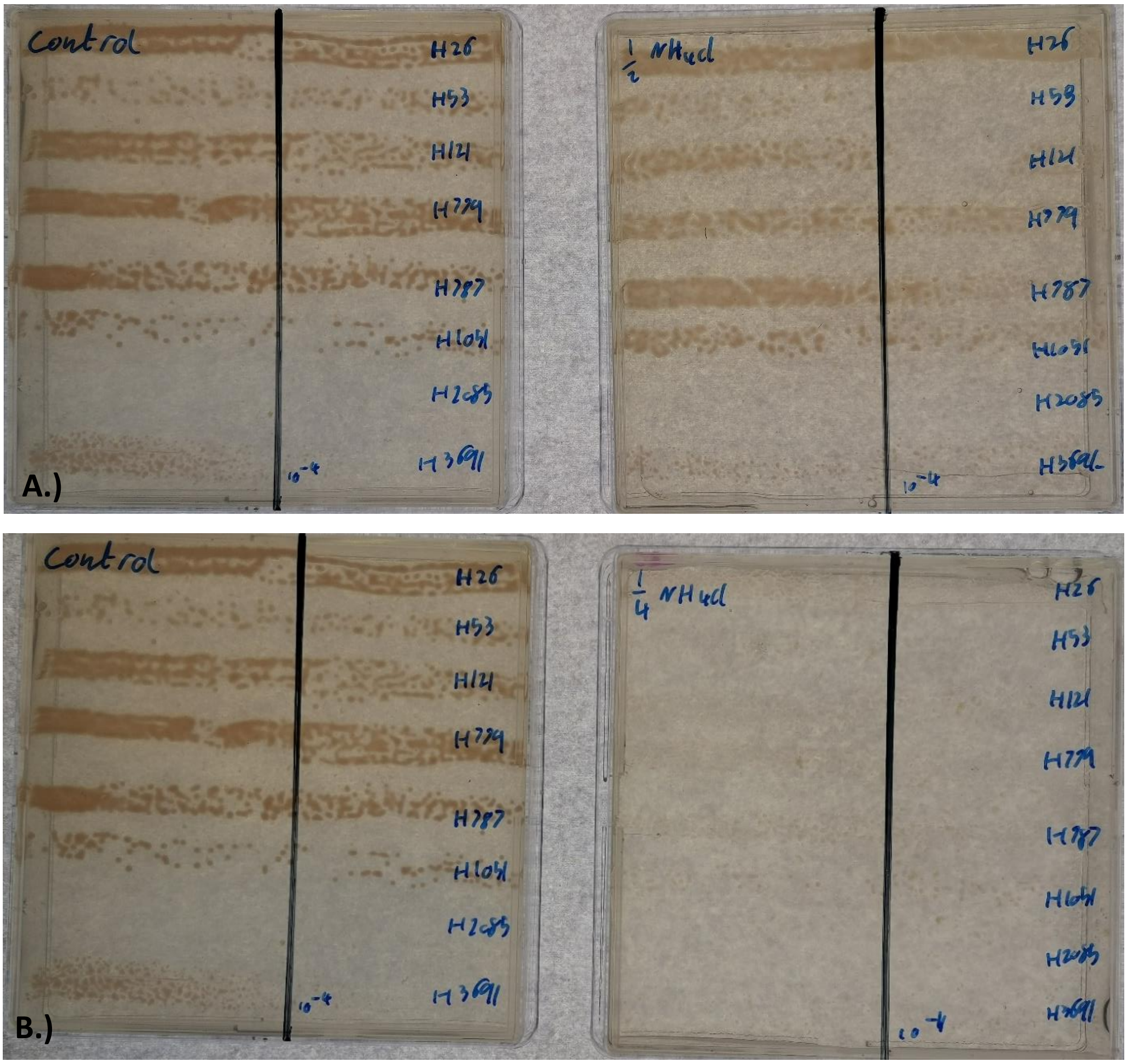


Figure 23. Gradient plates for selected *Hfx. volcanii* strains. **A.)** The left (control) showing growth under normal nitrogen conditions of HV-min media (no gradient) and the right showing ½ of normal nitrogen concentrations reducing from left to right, the far right having no nitrogen source. **B.)** The left plate showing controlled conditions normally used in Hv-min media with no gradient and the right plate showing a nitrogen concentration ranging from 1/4 of the normal level on the left decreasing in concentration toward the right where no ammonium is present.

4.Discussion

This study created the basis for a new real time growth competition of *Hfx. volcanii*, utilising fluorescent marker constructs. Several strains were successfully generated and confirmed via Southern blot. However, only a few of these strains showed fluorescence as measured by phosphor-imager. There are several reasons that this could have occurred. Firstly, the high background in the Southern blot may have masked the true result. Secondly that the limited wavelength options provided by the phosphor-imager may have failed to excite the fluorescent proteins in most of the strains. The excitations range available was designed for fluorescent markers Cy2 and Cy3, which have excitation wavelengths of 489-506 nm and 548-561 nm, respectively. This explains the fluorescence in GFP strains as the standard excitation wavelength is 488 nm. The maximum excitation for mCherry is 587 nm which is supported by the weak fluorescence observed at Cy3 settings and lack of at the lower wave lengths of Cy2 (Duggin *et al,* 2015).

Although the fluorescence detected on the imager at Cy3 was not as strong as the signal for GFP at Cy2 settings, this pairing still looks like a promising candidate for use in the real time growth competition assay. If Covid-19 had not halted this project, GFP and mCherry would have been tested using the flow cytometer at an excitation wavelength of approximately 550 nm, and GFP and YPet would have been tested at a wavelength of approximately 490nm. The wavelengths were selected in order to achieve maximum fluorescence of both proteins simultaneously. The similarity in the colour of GFP and YPet was expected to make this a less efficient pairing than GFP and mCherry although signal strength may be higher for this combination (Duggin *et al,* 2015).

For all media conditions tested using the gradient plate assay, strain H2085 showed little to no growth. While there is the possibility that this is due to the change in media conditions, it is perhaps more likely due to the strain being deleted for Hel308, an ATP-dependent helicase which when deleted is known to show a negative impact on DNA repair and hence growth (Gamble-Milner, 2016). In order to test this, plates could be repeated with a longer incubation time. However, as the results were unanimous across many conditions, this was not performed. H3691 unusually had equal or higher amounts growth on several plates that were nutrient poor, when compared to control strains. This could be a due to a methodological error in inoculating the plates as this method introduces the potential for accidental bias based on the painting of the strains onto the plates, or that this genotype is less affected by the reduction in these nutrients in some way. As the only difference between H3691 and H787 that does not show this difference in growth is the *Δhel308::trpA+* genotype it is likely this is in some way responsible.

Other than the aforementioned strains, all other *Hfx. volcanii* strains tested showed similar results. The reduction and removal of any of the three carbon sources (glycerol, lactic acid and succinic acid) yielded similar outcomes. As each carbon source decreased in concentration, growth diminished until a point where there was no observable cell growth. This suggests that with slight decreases in one carbon source *Hfx. volcanii* can continue to survive by utilising the other sources until a point where the concentration is too low for growth to continue. Previous research has shown the importance of these carbon sources (Buckley *et* a*l*, 2020). The quantities of these available in the natural environment are not known, raising an interesting question about whether carbon source conditions used in the lab are at all reflective of natural conditions. However, dissolved carbon levels have been recorded for various hypersaline environments, from which *Hfx. volcanii* does not originate. For example Zachara *et al* (2016) showed dissolved carbon levels in a heliothermic hypersaline lake to range up to 0.04 mol/L. This level of carbon is higher than the utilised level in standard Hv min media (see chapter 2.2 Media). Although the exact composition of the dissolved carbon in this environment is unclear. As the results suggest depletion of a single carbon source can cause significant detriment to growth perhaps, the ratio of carbon sources is a more important factor than quantity alone.

Phosphate conditions ranging from the normal phosphate levels showed similar results to the carbon source plates. More interestingly 1/10th phosphate plates showed weak growth with a lack of developed pigmentation. Weak growth was also observed for ½ NH_4Cl plates with similar pigmentation to those of the 1/10th phosphate plates. 1/4 NH_4Cl conditions showed minimal growth for all strains. However, this also shows an ability of *Hfx. volcanii* to grow in very low concentrations of phosphate and nitrogen. Natural levels of phosphate in the Jordan river, which resides very close to the dead sea have been shown to range from 9-85 µg P/L which is significantly lower than the normal Hv min media conditions. The lowest recorded phosphate level at 9 µg P/L. This aligns with the growth observed in the phosphate plates suggesting minimal phosphate is required for survival and highlighting the excessive use of nutrients in lab media. Stiller and Nissenbaum, (*1999*) also reports nitrogen levels to range from 0.35 mg N/L to 3.2 mg N/L. The maximum recorded value here is again significantly lower nitrogen level than the 1.2% v/v of KPO_4 buffer added to standard Hv min media. These preliminary media tests suggest both phosphate and nitrogen as likely candidates for further screening.

4.1. Future work:

Due to Covid-19, a great deal of work is required to refine the results collected so far. Firstly, flow cytometry needs to be conducted to confirm the ability to use the fluorescent marker strains created. This would then be used to further explore media conditions, starting with the previously tested conditions that show a promising reduction in growth rate without being lethal to the cells. A wider range of media conditions may also need to be screened in order to confidently decide whether the growth difference between wild type and origin deleted cells is due to media conditions. Regardless of the media results, if fluorescence between strains can be distinguished on the flow cytometer, the creation of a more efficient real-time growth competition assay could prove invaluable for general *Hfx. volcanii* research; this assay would extend to other species with slight modifications, allowing for real time growth comparisons under almost any condition that can be created in liquid media. This would be a significant improvement on the established but time-consuming method of blue/white screening currently used (see methods). Additional experiments are then dependant on whether there is a negative or positive result from the competition assay screening.

It is possible that no media condition is found that creates a growth difference between the *ori+* and *Δori* strains either reducing the growth gap or removing it entirely. Then this may support previous theories about origin dependent replication being a selfish concept and raises interesting questions about our current perceptions of the evolutionary process. However, the absence of a singular nutrient condition cannot guarantee that there is not a physiological cause due to the complexity of nutrient requirements for halophilic Archaea. It does perhaps imply a genetic cause for the growth difference that occurs in the absence of origins. Another suggestion for why this result occurs is that it is metabolically more efficient to replicate without origins. This is seen in *T. kodakarensis* which has been shown to replicate via origin dependent replication even in the presence of an origin (Gehring *et al*, 2017). Although this result has only been observed in laboratory conditions and not in vivo. It has also been observed that GC skew around the origin of replication can still be seen in *T. kodakarensis* which shows evidence of origin usage somewhere along the evolutionary timescale (Gehring *et al*, 2017). However, there are still other numerous other potential causes that would need to be investigated, such as the interaction between origin deleted mutants and other species in the natural microbiome.

If a media linked phenotype is found, where reduction in a nutrient source reduces the observed growth advantage of origin deleted mutants, the importance of laboratory conditions being as close to those present in the natural environment is highlighted even if this results in slower growth rates generally. The requirement of origins in low nutrient conditions could suggest several things about the survival of this species in its natural habitat and the role of replicative origins. Firstly, nutrient conditions in a natural environment for *Hfx. volcanii* (the Dead Sea) are likely to fluctuate frequently whereas lab conditions remain stable. This is an important factor for consideration, if this result is found to be a media linked phenotype, it implies that the origin of replication may play a more significant role by initiating replication in a manner that is more resource efficient. This is one factor that would be interesting to further investigate should the result be a media related phenotype. A starting point for this being RNA-seq experiments for both *ori+* and *Δori* strains under nutrient deficient conditions to observe any changes in regulation of genes around the origins of replication. As the growth difference has decreased between the *ori+* and *Δori* strains it is assumed that some regulatory genes will be upregulated in the *ori+* strain that are not in the *Δori* strain, allowing the cell to survive in *ori+* cell to outperform its origin deleted counterpart. This would then allow for identification of key regulator genes involved in origin dependent DNA replication and cellular growth under conditions that at deficient and to some degree better mimic those of a natural setting. These genes could then be deleted from both backgrounds for further study.

Chapter Two: An Origin independent replication prediction tool

5. Introduction

5.1. Recombination dependent replication

The vast majority of organisms replicate via origin-dependent replication, where DNA replication is initiated at specific sites on the chromosome referred to as origins (see introduction of chapter one). These are AT-rich sites which are generally in close proximity to a DNA replication initiator gene. In Bacteria this is always *dnaA* and in Archaea it nearly always *cdc6,* also known as *orc1*. In *Sulfolobus* species, the *whip* gene is used instead at one of the three origins (Samson *et al*, 2013). A small number of Archaea and Bacteria have been found to survive despite the deletion of origins of replication. When this origin independent replication occurs, different protein complexes appear to be essential (Michel and Bernander, 2014) and replication appears to occur at more uniform levels across the genome rather than at specific locations (Hawkins *et al,* 2013). It is suggested that the ability to survive without origins is possible because the homologous recombination pathway takes over replication in the absence of origins. Hence recombination-dependent DNA replication (RDR) becomes dependent on homologous recombination machinery, such as RadA. This was previously discussed in section 1.5 (DNA and repair).

5.2. Bioinformatic prediction tools

A range of bioinformatic tools have been developed to predict the locations of origins of DNA replication. These methods are discussed below;

GC skew

GC skew methods were originally designed for circular bacterial chromosomes, where replication starts at an origin (*ori*) and continues bidirectionally until it reaches the replication terminus (*ter*), it is assumed that the length of each arm between the *ori* and the *ter* is equidistant. GC skew refers to where excess of guanine (G) over cytosine (C) is present on one DNA strand, and where an excess of C over G is present on the other strand; it is plotted on a graph that segregates the genome into regions with sliding windows of a specific size. The maximum and minimum GC skew points are correlated with the loci of the *ori* and *ter*, respectively. GC skew and other nucleotide disparities accumulate over time due to the different mutational spectra of continuous versus discontinuous DNA replication. Therefore, maximum GC skew is found at the transition from leading to lagging strand DNA synthesis, namely the origin. The strength of GC skew can be calculated via the GC skew index (GCSI). This calculates strength by combining Fourier power spectral analysis with the Euclidean distance between the maximum and minimum values of a cumulative skew vector vector (Arakawa and Tomita, 2007)[90]. This will be covered in more detail later.

Z-curves

Z-curves are a refinement of the GC skew method. This is a 3-dimensional plot that represents three independent distributions that describe nucleotide disparity in a DNA sequence. These distributions are; purine versus pyrimidine (RY), amino versus keto (MK) and strong hydrogen bonding versus weak hydrogen bonding (WS) respectively along the genome (Zhang and Zhang 2003).

Genetic linkage of Origins to DNA replication initiator genes

Genetic linkage of origins of replication to their corresponding initiator genes is another form of prediction method utilised. This is normally conducted alongside the GC skew and Z-curve methods (Wu *et al,* 2014). The link between cognate initiator genes such as *oriC* and *dnaA* in Bacteria or *oriC* and *cdc6* (*orc1*) in Archaea and the origin of replication can be utilised in determining the location of an origin. A strong correlation is observed between the initiator genes and origins of replication in the majority of archaeal species. This was first observed in *Pyrococcus abyssi* by Myllykallio *et al* (2000) who predicted the location of *oriC* in *P. abyssi* via the GC skew method and noted that *oriC* is flanked with the *cdc6* gene and several eukaryotic-like DNA replication genes. A similar organisation was also observed in two other *Pyrococcus species*, *P. horikoshii* and *P. furiosus*, and as a result it was suggested that origin organisation is highly conserved (Luo *et al,* 2014). It should also be noted that typically the link between *dnaA* and bacterial origins of replication is weaker than the link between Cdc6 and the archaeal origins. However, there are archaeal exceptions to this correlation, particularly in *Sulfolobus* where one Cdc6 protein is not encoded by a gene adjacent to an origin of replication and instead the origin lies beside the gene for a crenarchaeal-specific protein known as WhiP (Dao *et al,* 2019: Samson *et al,* 2013).

This study has utilised these predictions tools in order to create a novel tool which is able to predict the ability of a species to replicate in the absence of origins of replication. The basis of this tool being that genomic features such as GC skew arise due to the frequent or obligatory use of origins of replication over evolutionary timescales, and will therefore be less pronounced in species where the origins are not essential or seldom used, such as *Hfx. volcanii* and *T. kodakarensis*. This logic not only applies to genomic features on a nucleotide level but to features on the gene-level such as linkage of initiator genes with origins, location and direction of transcription of core genes and in linkage of DNA polymerases genes to the origin. The factors influencing this prediction tool will be discussed below in further detail Adlam (2018).

Skew indices

It was first observed that the frequencies of adenine to thymine (known as AT skew) and guanine to cytosine (known as GC skew) change suddenly at the origin and terminus of replication in bacterial species (Lobry, 1996). This was determined by analysing DNA sequences of 3 different species and calculating AT and GC skew frequencies using a sliding window. This showed evidence for asymmetric mutation pressure resulting in nucleotide bias on the leading strand in comparison to the lagging strand resulting in GC skew (Lobry and Sueoka, 2002). At the origin of replication, a sharp transition occurs between the leading and lagging strand resulting in a sudden change in GC skew. This change can be used to predict the loci of an origin of replication in various organisms (Lobry, 1996).

In addition, in some organisms this nucleotide skew profile results in high levels of noise such as seen in *Hfx. volcanii* which are known to be able to replicate via RDR. This increase in noise is likely caused by multiple origins present in some archaeal species, similar manner to Eukaryotes which have a large quantity of origins with variable usages, only some of which are used for replication. The spectral ratio is the signal to noise ratio obtained by a fast Fourier transform of a GC skew, this can be as the measurement of clarity, hence the inverse of spectral ratio and can be defined as the noise of the skew. A 50 kb AT-rich prophage sequence was found to be responsible for an erratic peak in the *Hfx. volcanii* signal (Norais *et al*, 2007). The first version of the tool created by previous members of the Allers' lab, aimed to determine whether non-native regions arising from lateral gene transfer scrambled disparity signals. Hence, whether a gene was native or not, the strength of the skew and the amount of noise were also compiled into the previous prediction tool. However, it was determined that non-native regions did not alter the composition of nucleotide disparity plots, with the exception of the previously mentioned 50 kb AT-rich prophage sequence in *Hfx. volcanii* (Adlam 2018).

DNA Polymerases

It has been shown that in Archaea, DNA polymerase B (polB*)* is the most common replicative polymerase. However, this DNA polymerase is not essential in species such as *Thermococcus kodakarensis*, whereas DNA polymerase D (polD*)* is essential (Čuboňová *et al*, 2013). It was suggested that *polD* could act in recombination dependent replication, explaining why *polB* is not essential; if *polB* is deleted, replication then occurs via recombination dependent replication. Similar results have also been seen in *Hfx. volcanii.* It was therefore suggested that the synteny between replication initiation factors and the type of polymerase present was observed to be useful in predicting RDR (Maurer *et al,* 2018: Hogrel *et al*, 2020).

Clusters of Orthologous groups of proteins

Clusters of orthologous groups of proteins (COGs) have been previously used to define types of genes. The proteins in these groups have been previously defined by Tatusove *et al (*2000) as proteins across three or more species which are more similar to each other than to other proteins in the species respective genomes. An arCOG database previously published by Galperin *et al* (2015) is available for use and has recently been updated to include a more comprehensive set of arCOGS . Archaeal Clusters of orthologous groups of proteins (arCOGs) are an Archaea-specific subset of COGs.

Information linkage

The arCOGS/COGS have been divided into functional groups each assigned with a letter code. The most relevant classes to predicting RDR being information storage and processing gene classes (A,B,J,K, and L) which are suggested to be enriched round origins of replication as they are all involved in DNA replicative processes. Hence, analysis of information linkage of core genes around an origin associated gene can be assessed using these arCOGs/COGs. It is expected that the ratio of information storage and processing genes (mentioned above) to other genes is higher around an origin-associated gene, in comparison to the rest of the genome.

Co-orientation of Core Genes

These arCOGS/COGS were used in version 1 of the prediction tool for the calculation of co-orientation of core genes with the direction of DNA replication (Adlam, 2018). The premise was that in *E. coli*, the two replichores (halves of the chromosome in relation to the *ori* and ter) of the chromosome are co-orientated with the transcription of highly expressed core genes, and this particularly the case around the origin of replication. It is suggested that this occurs in order to prevent head-on collisions between replication and transcription machinery. This is noticeable in Archaea but seen to a lesser extent with archaeal genes. For example, the archaeal *Cdc6/orc1* gene which is found next to the origin is always orientated away from the origin of replication. In a similar manner, rRNA genes are very highly expressed in Archaea and are always orientated away from the origin (Paul *et al*, 2013: Pomerantz and O'Donnell, 2010: Dimude *et al,* 2016). Hence, if origins are not essential and are used infrequently, co-orientation is expected to be reduced as head-on collisions of replicative and transcriptive machinery are less likely.

Principal component analysis

The previous bioinformatic tool by Adlam (2018) compiled the above factors and calculated the contribution of each variable on a principal component analysis (PCA) graph. The resulting system allowed 15 archaeal species and 10 bacterial species to be grouped on the PCA in accordance to whether they are predicted to be able to replicate via RDR, cannot replicate via RDR or if the results were unclear. Eukaryotes were excluded from the tool as they do not exhibit nucleotide skew, this is most likely due to eukaryotes possessing multiple origins and their usage of them being variable; this results in a scrambled nucleotide skew signal.

The PCA was designed so that species which could replicate without origins had low numerical values in comparison to those species that could not (See Figure 24). This was suggested to be the case particularly on the x-axis as all species known to replicate lay towards the left side of the PCA, confirming the concept of the prediction tool. The PCA showed several species as promising candidates for being able to survive without origins including; *H. marismortui, P. abyssi, H. hispanica, H. borinquense, N. maritimus, Synechococcus sp., T. gammatolerans, A. fulgidus* and *H. lacusprofundi.*

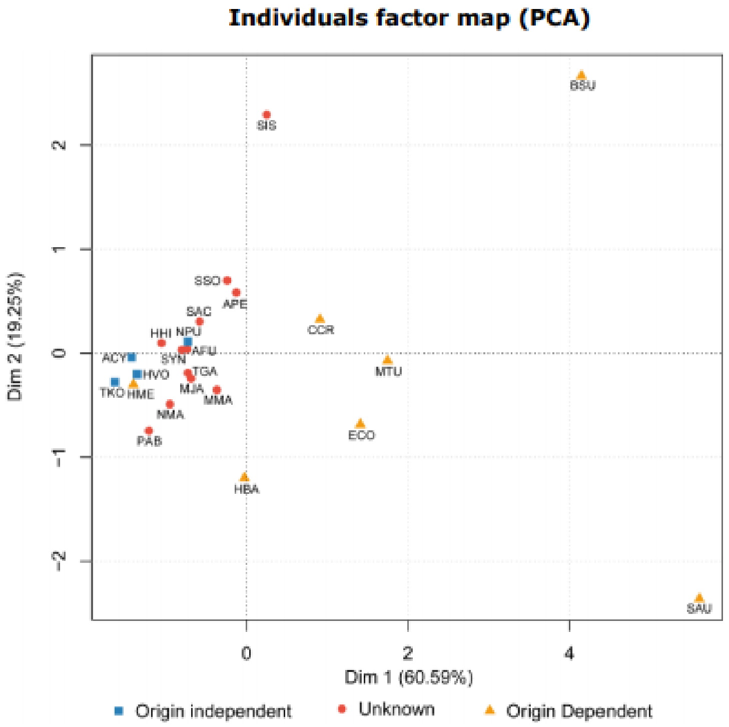


Figure 24. Principal component analysis from Adlam (2018), categorising species by their ability to replicate without origins. Species in the unknown group (organisms which have yet to be tested for the ability to replicate without origins) shown by red circles that lie to the bottom left were suggested to be likely candidates for origin independent species. Three letter species codes correspond to the first letter of the genus and the first two letters of the species name. For a list of species and abbreviations see Table 12-15.

5.3 Aims

This section firstly aims to support the acquisition of knowledge needed in order to use and develop a previously created bioinformatics prediction tool. As the current tool requires a large amount of manual calculation and input, it is impractical for use on a large scale. The Aim of this project is to update the tool using the most recent findings in the field particularly new and updated arCOG and COG databases. Then to adapt the tool for more convenient use on large data sets allowing the community to quickly assess species they are interested in, for potential origin independent replicative processes. This holds relevance as to date very few species are known to be able to undergo this form of replication. Previously suggested adaptations to refine the tool by Adlam (2018) will also be implemented where appropriate. This project aims to provide the first steps towards a comprehensive tool to allow for large scale screening of origin independent replication in Bacteria and Archaea.

6. Materials and Method

6.1. Materials

A range of species were chosen for use in the prediction tool encompassing a wide range of taxonomic groups, within the Bacterial and Archaeal domains. Species were selected semi-randomly with any species with insufficient availability of genomic information being removed from the study. Yeast and fungi were avoided due to their genetic make-up being unsuitable for the following prediction methods. (See Table 10 for species selected)

Table 10. Genomes used to predict origin independent replication.

Species	Strain	Group	NCBI Reference
Acidilobus saccharovorans	345-15	Archaea	NC_014374.1
Acidobacterium capsulatum	ATCC 51196	Bacteria	NC_012483.1
Aeropyrum camini	SY1	Archaea	NC_022521.1
Aeropyrum pernix	K1	Archaea	NC_000854.2
Anabaena cylindrica	PCC 7122	Bacteria	NC_019771.1
Aquifex aeolicus	VF5	Bacteria	NC_000918.1
Archaeoglobus fulgidus	DSM 4304	Archaea	NC_000917.1
Bacillus aerophilus	232	Bacteria	NZ_CP026008.1
Bacillus amyloliquefaciens	DSM 7	Bacteria	NC_014551.1
Bacillus subtilis	168	Bacteria	NC_000964.3
Bacteroides thetaiotaomicron	7330	Bacteria	NZ_CP012937.1
Caulobacter crescentus	CB15	Bacteria	NC_002696.2
Desulfovibrio vulgaris	RCH1	Bacteria	NC_017310.1
Desulfurobacterium thermolithotrophum	DSM 11699	Bacteria	NC_015185.1
Escherichia coli	K-12 sub-strain MG1655	Bacteria	NC_000913.3
Fusobacterium nucleatum	NCTC10562	Bacteria	NZ_LN831027.1
Gloeobacter kilaueensis	JS1	Bacteria	NC_022600.1

Gloeobacter violaceus	PCC 7421	Bacteria	NC_005125.1
Granulicella mallensis	MP5ACTX8	Bacteria	NC_016631.1
Haloarcula hispanica	ATCC 33960	Archaea	NC_015948.1
Halobacterium salinarum	NRC-1	Archaea	NC_002607.1
Haloferax alexandrinus	WSP1	Archaea	NZ_CP048738.1
Haloferax gibbonsii	ARA6	Archaea	NZ_CP011947.1
Haloferax mediterranei	ATCC 33500	Archaea	NC_017941.2
Haloferax volcanii	DS2	Archaea	NC_013967.1
Halopiger xanaduensis	SH-6	Archaea	NC_015666.1
Haloquadratum walsbyi	DSM 16790	Archaea	NC_008212.1
Halorhabdus tiamatea	SARL4B	Archaea	NC_021921.1
Halorhabdus utahensis	DSM 12940	Archaea	NC_013158.1
Halorubrum lacusprofundi	ATC 49239	Archaea	NC_012029.1
Haloterrigena turkmenica	DSM 5511	Archaea	NC_013743.1
Hyperthermus butylicus	DSM 5456	Archaea	NC_008818.1
Halogeometricum borinquense	DSM 1151	Archaea	NC_014729.1
Haloarchula marismortui	ATCC43049	Archaea	NC_006396.1
Halomicrobium mukohataei	DSM 12286	Archaea	NC_013202.1
Halovivax ruber	DSM18193	Archaea	NC_019964.1
Methanobacterium formicicum	MB9	Archaea	NZ_LN734822.1
Methanocaldococcus fervens	AG86	Archaea	NC_013156.1
Methanococcus jannaschii	DSM 2661	Archaea	NC_000909.1
Methanoregula boonei	6A8	Archaea	NC_009376.1
Methanoregula formicica	SMSP	Archaea	NC_019943.1
Methanosarcina mazei	Go1	Archaea	NC_003901.1
Methanothermococcus okinawensis	IH1	Archaea	NC_015636.1
Mycobacterium tuberculosis	H37Rv	Bacteria	NC_000962.3
Neisseria meningitidis	MC58	Bacteria	NC_003112.2
Nitrosopumilus maritimus	SCM1	Archaea	NC_010085.1
Nitrosphaera viennensis	EN76	Archaea	NZ_CP007536.1
Nostoc punctiforme	PCC 73102	Bacteria	NC_010628.1
Pyrobaculum arsenaticum	DSM 13514	Archaea	NC_009376.1
Pyrobaculum islandicum	DSM 4184	Archaea	NC_008701.1
Pyrococcus abyssi	GE5	Archaea	NC_000868.1
Pyrococcus furiosus	DSM 3638	Archaea	NC_003413.1
Pyrococcus yayanosi	CH1	Archaea	NC_015680.1
Pyrolobus fumarii	1A	Archaea	NC_015931.1
Rickettsia prowazekii	Str. Chernikova	Bacteria	NC_017049.1
Salinicoccus halodurans	H3B36	Bacteria	NZ_CP011366.1
Staphylococcus aureus	NCTC 8325	Bacteria	NC_007795.1
Staphylothermus hellenicus	DSM 12710	Archaea	NC_014205.1
Staphylothermus marinus	F1	Archaea	NC_009033.1

Streptomyces coelicolor	A3(2)	Bacteria	NC_003888.3
Sulfolobus acidocaldarius	DSM 639	Archaea	NC_007181.1
Sulfolobus islandicus	L.S.2.15	Archaea	NC_012589.1
Sulfolobus solfataricus	P2	Archaea	NC_012589.1
Sulfolobus tokodaii	Str. 7	Archaea	NC_003106.2
Sulfuracidifex tepidarius	IC-007	Archaea	NZ_AP018930.1
Sulfurihydrogenibium azorense	Az-Fu1	Bacteria	NC_012438.1
Synechococcus sp.	PCC 6312	Bacteria	NC_019680.1
Thermococcus barophilus	MP	Archaea	NC_014804.1
Thermococcus celer	VU13	Archaea	NZ_CP014854.1
Thermococcus chitonophagus	Isolate 1	Archaea	NZ_LN999010.1
Thermococcus gammatolerans	EJ3	Archaea	NC_012804.1
Thermococcus gorgonarius	W-12	Archaea	NZ_CP014855.1
Thermococcus kodakarensis	KOD1	Archaea	NC_012804.1
Thermococcus litoralis	DSM 5473	Archaea	NC_022084.1
Thermococcus pacificus	P-4	Archaea	CP015102.1
Thermococcus peptonophilus	OG-1	Archaea	NZ_CP014750.1
Thermococcus profundus	DT 5342	Archaea	NZ_CP014862.1
Thermococcus radiotolerans	EJ2	Archaea	NZ_CP015106.1
Thermococcus siculi	RG-20	Archaea	NZ_CP015103.1
Thermococcus thioreducen	OGL-20P	Archaea	NZ_CP015105.1
Thermoproteus tenax	Kra 1	Archaea	NC_016070.1
Thermoproteus uzoniensis	768-20	Archaea	NC_015315.1
Thermosphaera aggregans	DSM 11486	Archaea	NC_014160.1
Thermotoga maritima	MSB8	Bacteria	NC_023151.1
Thermovibrio ammonificans	HB-1	Bacteria	NC_014926.1
Treponema pallidum	Subsp. Pertenue str. SamoaD	Bacteria	N_016842.1

6.2. Methods

Bioinformatic predictions

Nucleotide disparity plots:

Combined nucleotide disparity plots, showing GC, AT, RY and MK disparity alongside Z-curves (a 3D representation of RY, MK and WS disparity across three axes) were created using a custom function in MATLAB (MATLAB R2020a, The 208 MathWorks, Inc., Natick, Massachusetts, United States) based on equations from (Zhang and Zhang, 2005; Hartman *et al*, 2010) (Table 11). The function creates a 4 x n zero matrix where n is the length of the selected genome sequence. Then each row in the matrix is assigned to a DNA base (A, C, G and T). The selected genome was then run through the matrix and whenever a base is present a value of 1 is assigned. This is then summed cumulatively for each row and used to calculate disparity.

Table 11. Equations to calculate each type of disparity.

Disparity	**Equation**
GC skew	Gn – Cn = (xn – yn)/2
AT skew	An – Tn = (xn + yn)/2
Purine/pyrimidine (RY)	xn = (An + Gn) – (Cn + Tn)
Amino/keto (MK)	yn = (An + Cn) – (Gn + Tn)
Weak/strong hydrogen bonds (WS)	zn = (An + Tn) – (Cn + Gn)

Skew index:

The skew index for all types of nucleotide disparity were calculated using MATLAB in a similar manner to methods used to quantify strength of GC skew (Arakawa and Tomita, 2007). The maximum skew value was selected for principal component analysis alongside the corresponding spectral ratio, whereas the inverse of the spectral ratio was used as a quantitative measure of noise for disparity graphs. Each species was assigned a numerical code which could be used for rapid skew profile analysis via the MATLAB code, as the input for each species could simply increase by a single increment after being recorded.

arCOG/COG analysis:

The arCOG database by Galperin *et al (*2019) was accessed and filtered removing any unclassified arCOGs and their corresponding COGS, as well as all super clusters present in the data base. The remaining arCOGS/COGs were collated in Microsoft Excel (2019) with their ID's, genomic loci and functional classes. The functional classes were then divided into information storage and processing classes (A, B, J, K and L) or other.

Information gene linkage:

Information gene linkage to origins was calculated by counting the number of information and storage processing genes present in a 25-gene window either side of the origin-associated gene. This was then compared against numbers across the whole genome using the integrated microbial genomes and microbiomes tool JGI (The Regents of the University of California, 2020). The numbers were collated on Excel (2019) and a χ^2 test on a 2x2 contingency table was used with a 1 degree of freedom and a *p* value of <0.05 to assess significance. The χ^2 value was chosen for the principal component analysis over the *p* value.

Co-orientation of core genes:

The percentage of core genes co-orientated with the direction of DNA replication was calculated, assuming the two arms between the *ori* and *ter* were of equal length. Core genes were defined as genes that corresponded with 453 arCOGS/COGs from the previously mentioned filtered data base (Galperin *et al,* 2015;2019). All genomes were screened for the selected list of core genes (using integrated microbial genomes and microbiomes tool JGI) and the genomes (Pelve *et al*, 2012) split into 100 kb windows using Excel (2019). The number of core genes in each direction were counted and a weighted average for each window calculated. The sum of weighted averages was chosen for use in the principal component analysis.

Linkage to DNA polymerase genes

The protein sequences of all origin associated genes such as *dnaA, cdc6/orc1* and *whip* were queried using the Absynte tool (Despalins *et al*, 2011) and scored for linkage with replicative DNA polymerase genes. The following rulings were used for scoring: if an origin associated gene was linked to DNA pol III (in Bacteria) it was assigned a value of 1, if it was not linked to any replicative polymerases a value of 0 was assigned; a score of -1 was given for each subunit of the archaeal DNA pol II (*polD)* linked to the origin associated gene.

Principal component analysis (PCA):

The contribution of each variable represented on the principal component analysis axis, can be seen in figure 25. The size of and shade of blue show a visual representation of the level of contribution measured by the $\cos^2$ value previously measured by Adlam (2018). It was found that co-orientation of core genes and strength of skew made the largest contributions to the first axis. It can also be observed that SI and information linkage made the largest contributions to the second axis. However, as the first axis represents the potential ability of an organism to survive without DNA replications it is suggested that this axis is the most significant for the purposes of this tool.

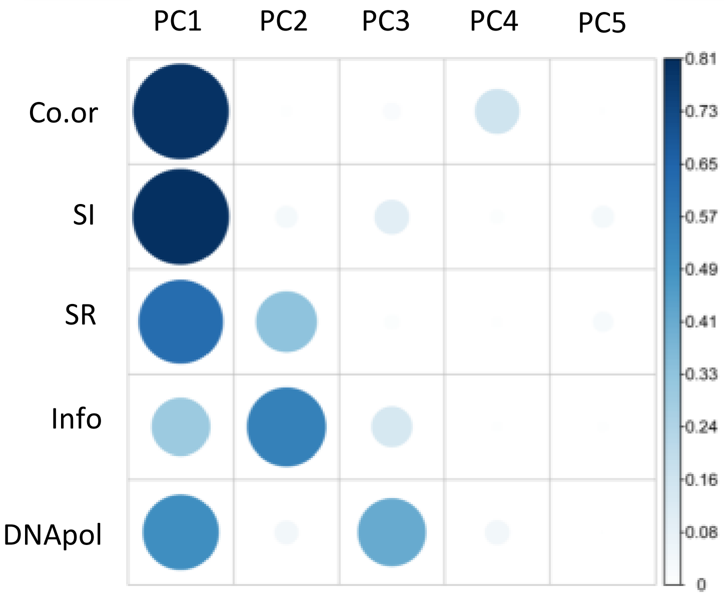


Figure 25. Visual representation of the contributions of each variable to the principal component analysis. The size of contribution being represented by the shade of blue. Figure adapted from Adlam (2018).

The statistical software Rstudio (RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.) was used to perform principal component analysis. Based on the previous findings by Adlam (2018) the variables used for this were as follows: Co-orientation of core genes, Skew index (SI),Spectral ratio (SR) Information linkage and Linkage to DNA polymerases. The statistics were chosen in a manner so that all values are likely to be low for organisms that can survive without origins of replication. Hence the spectral ratio and χ^2 statistic being chosen over the noise value and the p value. The organisms can then be grouped into species that are likely able to survive without origins, unable to survive without origins or unknown based on their positions on the PCA.

7.Results

7.1. Nucleotide skew Indices

The nucleotide skew index profiles show varying levels of noise and signal strength but can be somewhat grouped based on the clarity of the profile. The first group having relatively 'clean' profiles with minimal noise and strong skew signals occurring near their origin associated genes. This can be seen in several species as summarised in the Table 12. Two examples of a clean skew profiles can be seen in Figures 26 and 27. It should be noted that the skew can vary widely based on the type of nucleotide disparity being considered, hence the maximal skew was used for analysis of groups. *T. litoralis* was not assigned to a group due to an issue with the reference sequence containing large amounts of unassigned nucleotides which interfered with the code used.

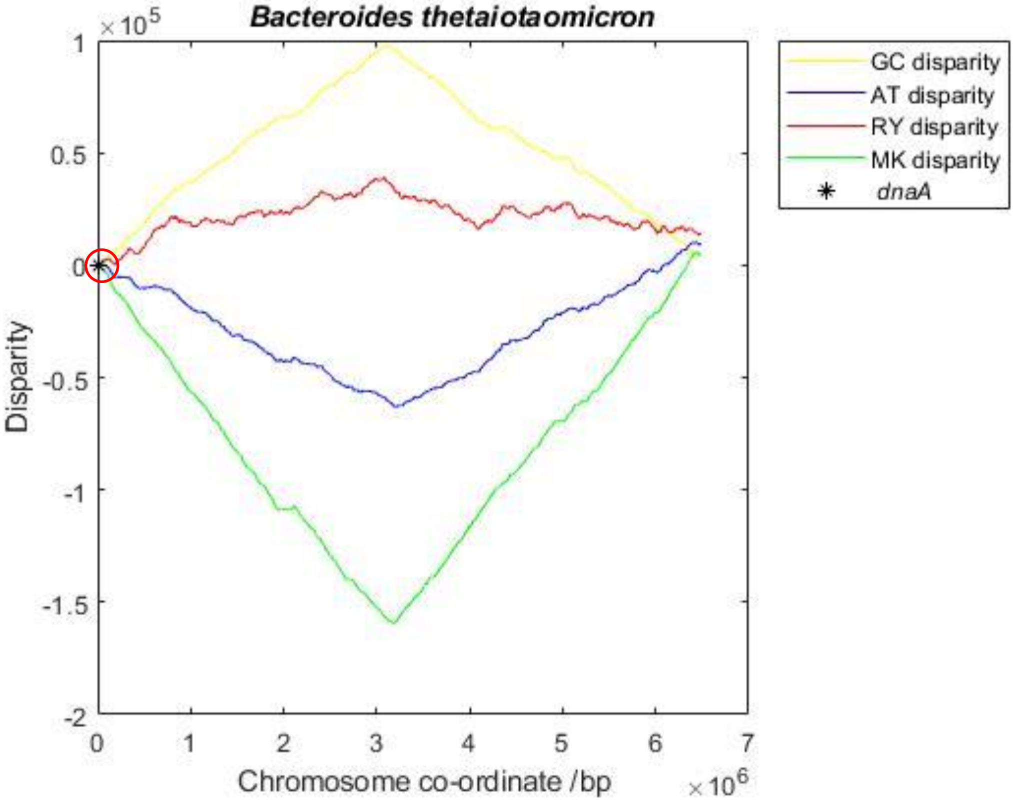


Figure 26. Example of a clean profile, with minimal noise and clear peaks. See key for disparity types and the location of origin-associated genes (Circled in red). Origin-associated genes appearing at a peak or trough in the disparity curve are potentially involved in origin independent replication, these have been circled in red.

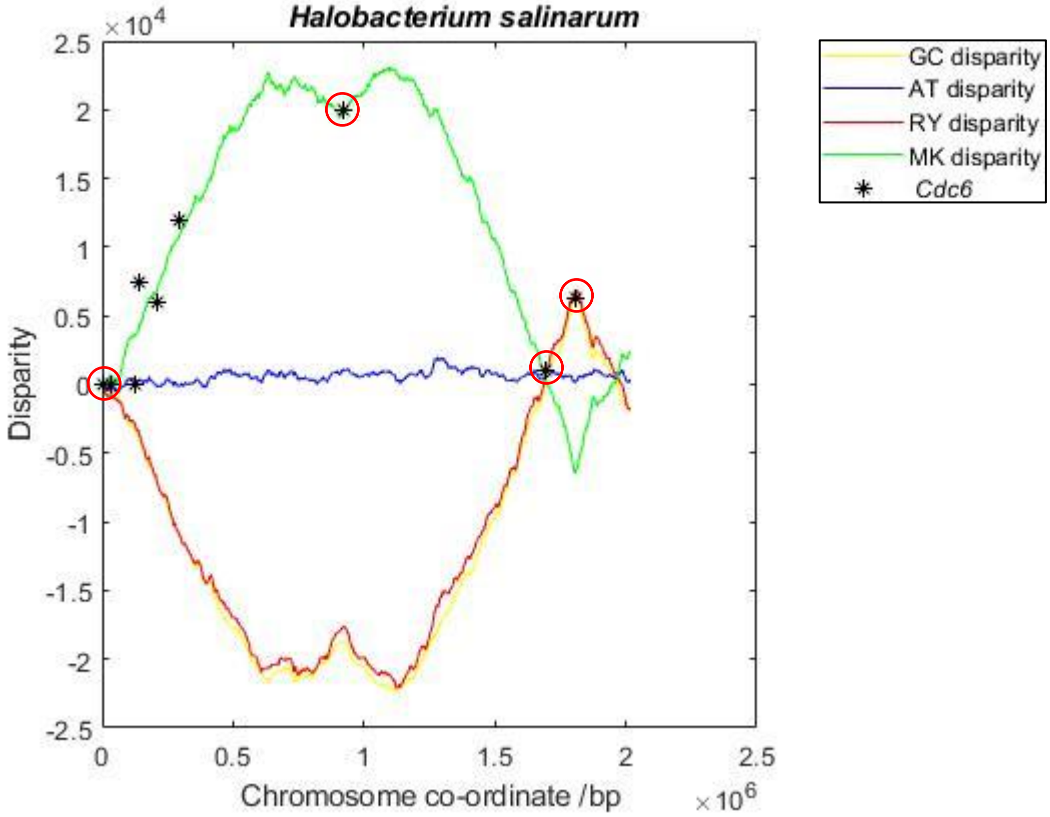


Figure 27. Another example of a clean profile, minimal noise and clear peaks occurring from some of the origin-associated genes. See key for disparity types and origin associated genes. Five origin associated genes can be seen at clear peaks or troughs (2 located closely together at the start of the graph and the remaining 3 being spread across the curve) in the disparity curve suggesting these may play a role in replication (circled in red). The other 4 origin-associated genes are likely not involved in replicative processes.

Table 12. Summary of all species grouped into the 'clean' skew profile group

Species	3 Letter Code	Organism Type	Skew grouping
A. capsulatum	ACA	Bacteria	Clean
B. aerophilus	BAE	Bacteria	Clean
B. amyloliquefaciens	BAM	Bacteria	Clean
B. subtilis	BSU	Bacteria	Clean
C. crescentus	CCR	Bacteria	Clean
D. vulgaris	DVU	Bacteria	Clean
E. coli	ECO	Bacteria	Clean
G. mallensis	GMA	Bacteria	Clean
H.salinarum	HSA	Archaea	Clean
M. mazei	MMA	Archaea	Clean
M. tuberculosis	MTU	Bacteria	Clean
N. maritimus	NMA	Archaea	Clean
N. meningitidis	NME	Bacteria	Clean
R. prawazekii	RPR	Bacteria	Clean
S. aureus	SAU	Bacteria	Clean
S. coelicolor	SCO	Bacteria	Clean
S. halodurans	SHA	Bacteria	Clean
H. ruber	HRU	Archaea	Clean
B. thetaiotaomicron	BTH	Bacteria	Clean
T. palidium	TPA	Bacteria	Clean
F. nucleatum	FNU	Bacteria	Clean

The second observable group has a moderate level of noise and varying strengths of skew signals. In general, the skew indices are lower than the 'clean profile' groups. There also appears to be less correlation between peaks and origin associated genes. For details of species in this group see Table 13. Examples of 'moderate' skew profile species are shown in Figures 28 and 29.

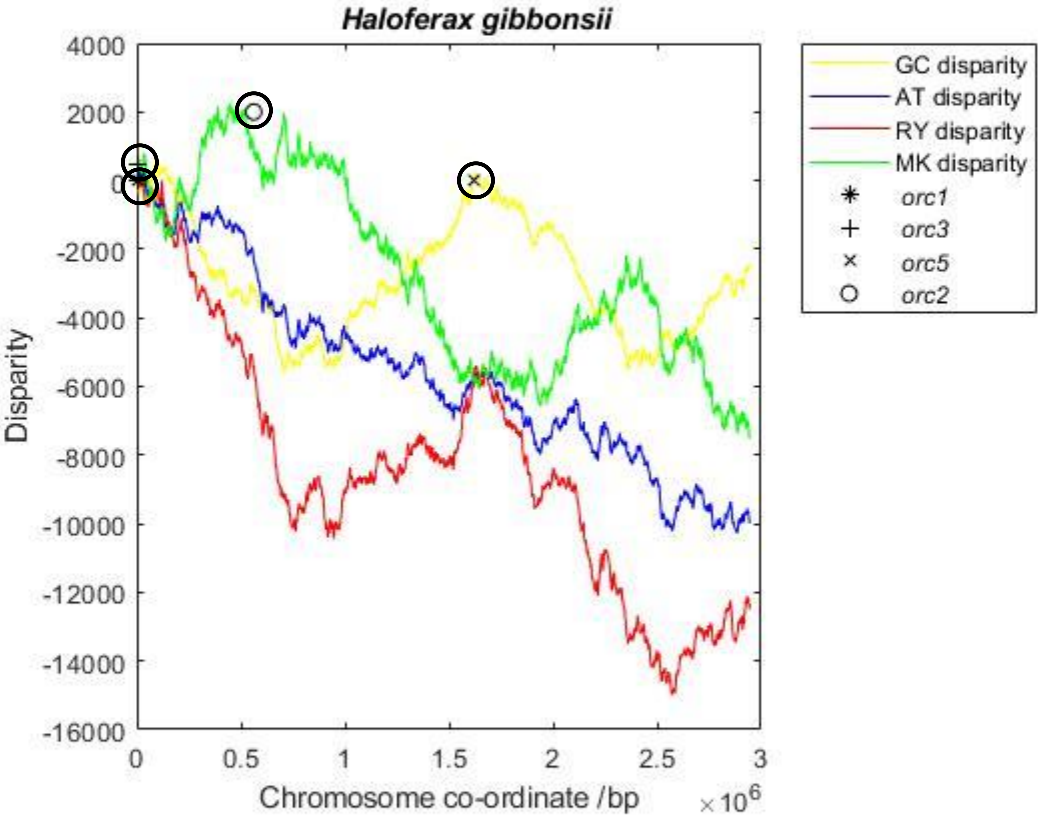


Figure 28. Example of a skew profile in the 'moderate' group. Clear peaks can be observed near origin-associated genes, however there is also a moderate level of noise present. See key for disparity types and origin associated genes. All origin-associated genes highlighted occur near peaks so may play a role in replication.

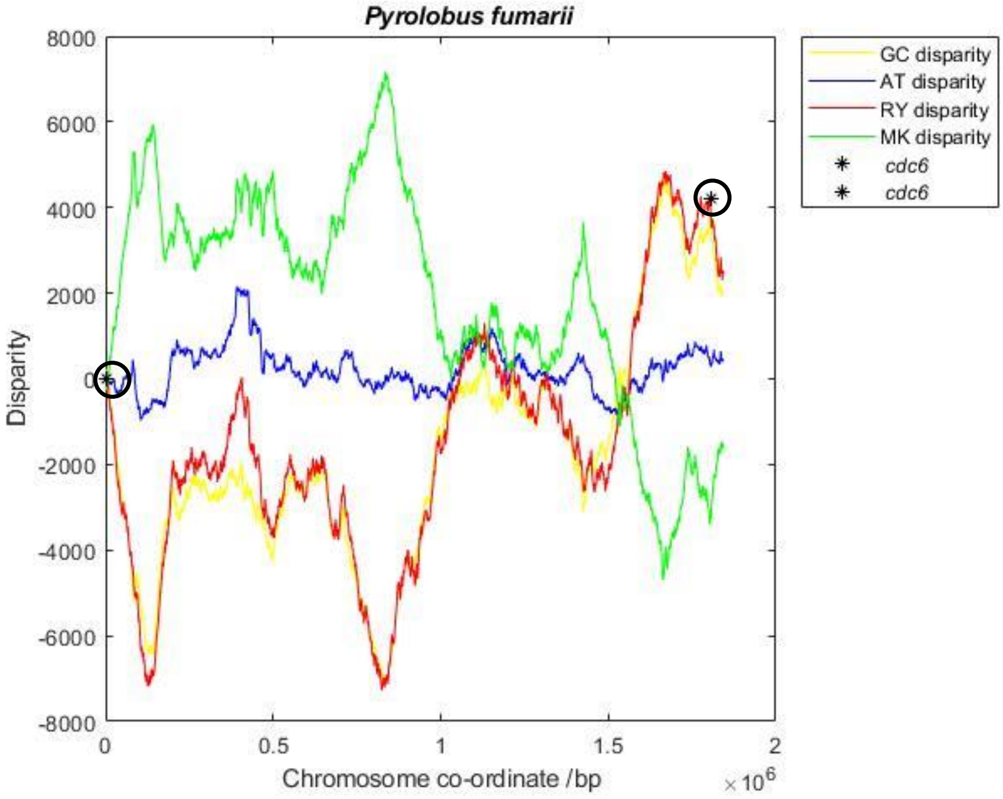


Figure 29. A second example of a skew profile in the 'moderate' group. Clear peaks can once again be observed near origin-associated genes alongside a moderate level of noise present. See key for disparity types and origin-associated genes. All origin-associated genes highlighted occur near peaks so may play a role in replication.

Table 13. Summary of species within the 'moderate' skew profile group.

Species	3 Letter Code	Organism Type	Skew grouping
A. pernix	APE	Archaea	Moderate
A. camini	ACA	Archaea	Moderate
A. saccharovorans	ASA	Archaea	Moderate
D. thermolithotrophum	DTH	Bacteria	Moderate
G. kilaueensis	GKI	Bacteria	Moderate
G. violaceus	GVI	Bacteria	Moderate
H. hispanica	HHI	Archaea	Moderate
Hfx. mediterranei	HME	Archaea	Moderate
Hfx. volcanii	HVO	Archaea	Moderate
Hfx. gibbonsii	HGI	Archaea	Moderate
H. mukohataei	HMU	Archaea	Moderate
M. jannaschii	MJA	Archaea	Moderate
M. formicicum	MFO	Archaea	Moderate
P. arsenaticum	PAR	Archaea	Moderate
P. fumarii	PFU	Archaea	Moderate
P. yayanosi	PYA	Archaea	Moderate
S. tokodaii	STO	Archaea	Moderate
S. azorense	SAZ	Bacteria	Moderate
Synechococcus sp	SYN	Bacteria	Moderate
T. chitonophagus	TCH	Archaea	Moderate
T. gammatolerans	TGA	Archaea	Moderate
T. kodakarensis	TKO	Archaea	Moderate
T. maritima	TMA	Bacteria	Moderate
T. pacificus	TPC	Archaea	Moderate
T. peptonophilus	TPE	Archaea	Moderate
T. profundus	TPR	Archaea	Moderate
T. thioreducen	TTH	Archaea	Moderate
T. aggregans	TAG	Archaea	Moderate
T. barophilus	TBA	Archaea	Moderate
T. tenax	TTE	Archaea	Moderate

The third noticeable group is those species which have a high level of noise masking the potential peaks at origin associated genes. These skew graphs are impossible or difficult to use for predicting whether an origin associated genes is involved in replication. See Table 14 for more details on these species. Figures 30 and 31 show example skew graphs for species included in this category.

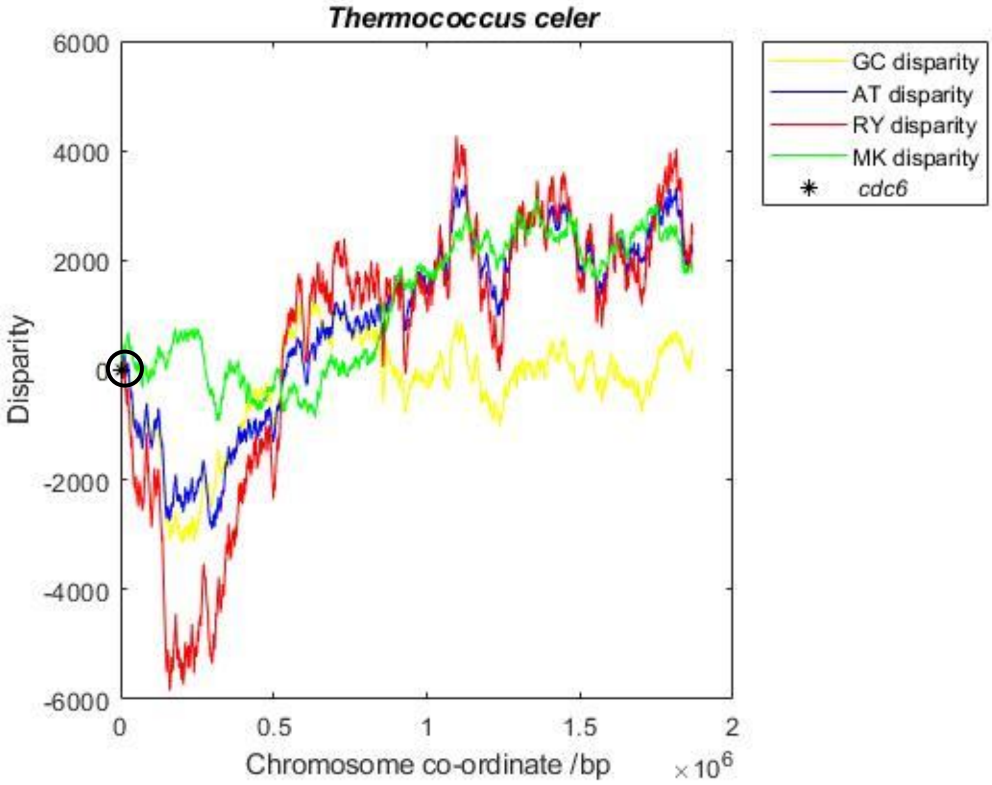


Figure 30. An example of a profile in the 'high noise' group. High levels of noise mask any clear peaks near the origin-associated gene. The role of origin-associated genes in replication cannot be predicted.

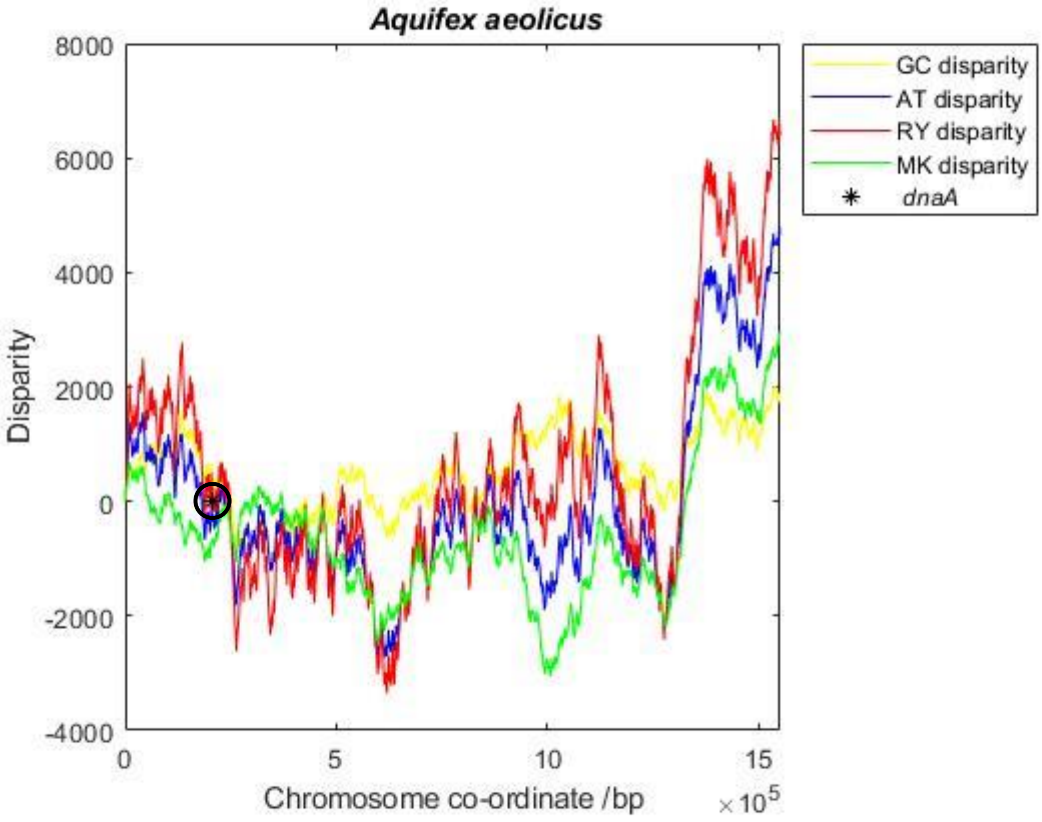


Figure 31. Another example of a profile in the 'high noise' group. High levels of noise mask any clear peaks near the origin-associated gene. The role of origin associated genes in replication cannot be predicted.

Table 14. Summary of species within the 'High noise' skew profile group.

Species	3 Letter Code	Organism Type	Skew Grouping
T. celer	TCE	Archaea	High Noise
A. aeolicus	AAE	Bacteria	High Noise
A. cylindrica	ACY	Bacteria	High Noise
A. fulgidus	AFU	Archaea	High Noise
H. butylicus	HBU	Archaea	High Noise
M. fervens	MFE	Archaea	High Noise
M. formicica	MFA	Archaea	High Noise
M. okinawensis	MOK	Archaea	High Noise
N. punctiforme	NPU	Bacteria	High Noise
N. viennensis	NVI	Archaea	High Noise
H. utahensis	HUT	Archaea	High noise
P. islandicum	PIS	Archaea	High Noise

The final grouping ('Other') contains all species that do not fit one of the categories. This can include; high noise with clear peaks, low noise and no peaks or uniform profiles with low noise and little to no peaks. Alongside any other unusual skew profiles that were not able to be grouped with the rest. Examples of these profiles can be seen in Figures 32 and 33 and all species grouped this way in Table 15.

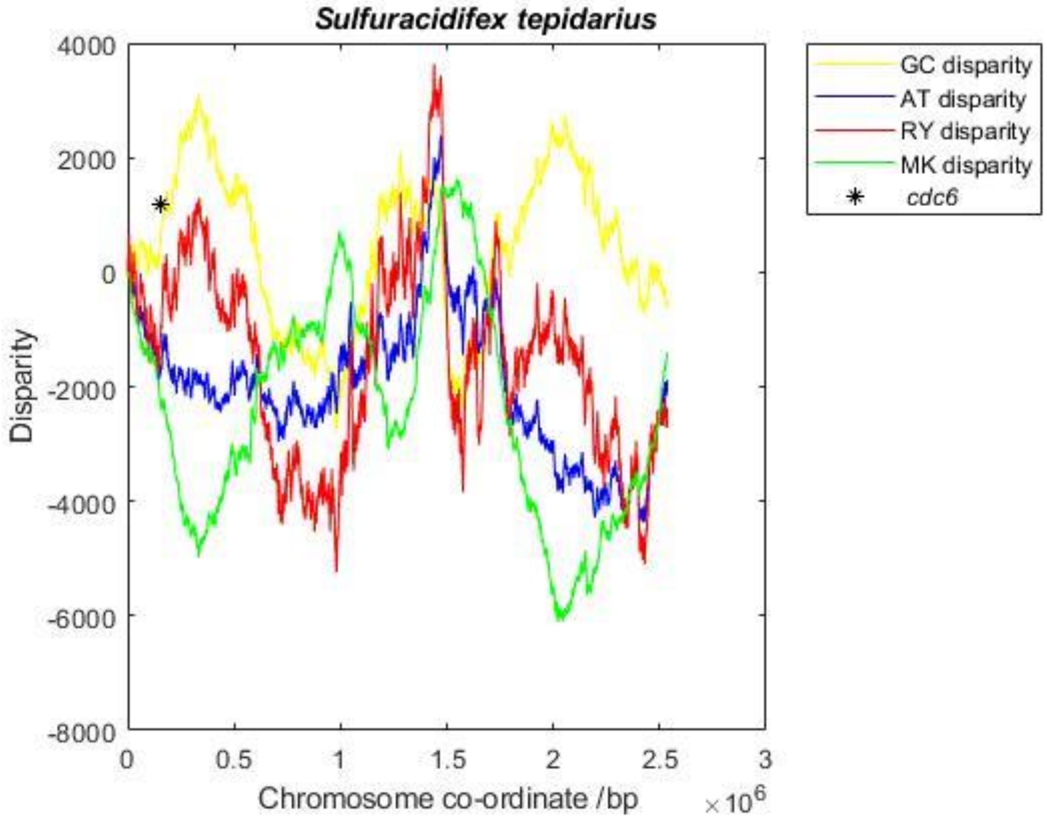


Figure 32. An example of species placed into the 'other' skew profile group, as trends do not fit in with the previously mentioned groups. High levels of noise can be observed in the profile but with clear peaks near the origin-associated gene.

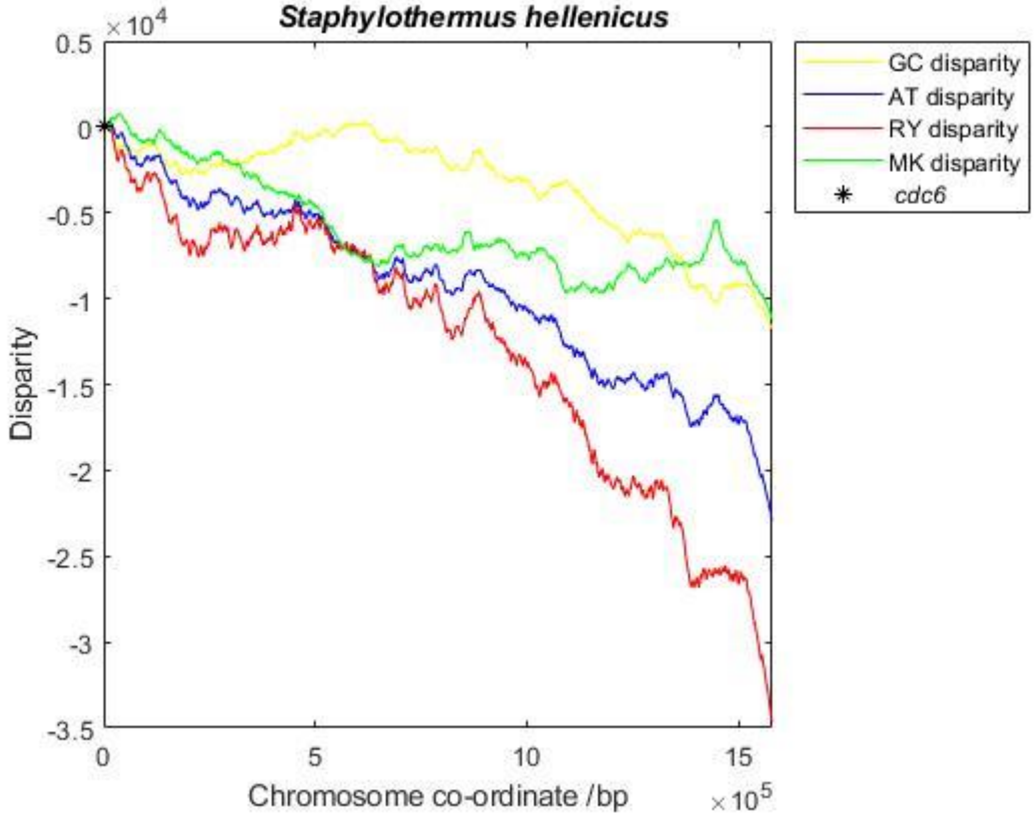


Figure 33. Another example of a species placed into the 'other' skew profile group, as trends do not fit in with the previously mentioned groups. Low levels of noise can be observed in the profile with minimal sized peaks.

Table 15. Summary of species within the 'Other' skew profile group.

Species	3 Letter Code	Organism Type	Skew Grouping
H. lacusprofundi	HLA	Archaea	Other
H. tiamatea	HTI	Archaea	Other
H. turkmenica	HTU	Archaea	Other
H. walsbyi	HWA	Archaea	Other
H. xanaduensis	HXA	Archaea	Other
Hfx. alexandrinus	HAL	Archaea	Other
M. boonei	MBO	Archaea	Other
P. abyssi	PAB	Archaea	Other
P. furiosus	PFU	Archaea	Other
S. acidocaldarius	SAC	Archaea	Other
S. islandicus	SIS	Archaea	Other
S. solfataricus	SSO	Archaea	Other
S. hellenicus	SHE	Archaea	Other
S. marinus	SMA	Archaea	Other
S. tepidarius	STE	Archaea	Other
T. ammonificans	TAN	Bacteria	Other
T. gorgonarius	TGO	Archaea	Other
T. radiotolerans	TRA	Archaea	Other
T. siculi	TSI	Archaea	Other
T. uzoniensis	TUZ	Archaea	Other

7.2. Z-Curves:

The Z-curves mimic what can be seen in the skew disparity plots above. This is to be expected as Z-curves are suggested to be a refinement of the nucleotide skew method (Zhang and Zhang, 2005). Hence, similar patterns can be observed in some species which yield Z-curves with clear, well-defined V shapes corresponding to the peaks observed on the disparity plot. Conversely, organisms from the 'High noise' group show very few clear and distinguishable features (Figure 34).

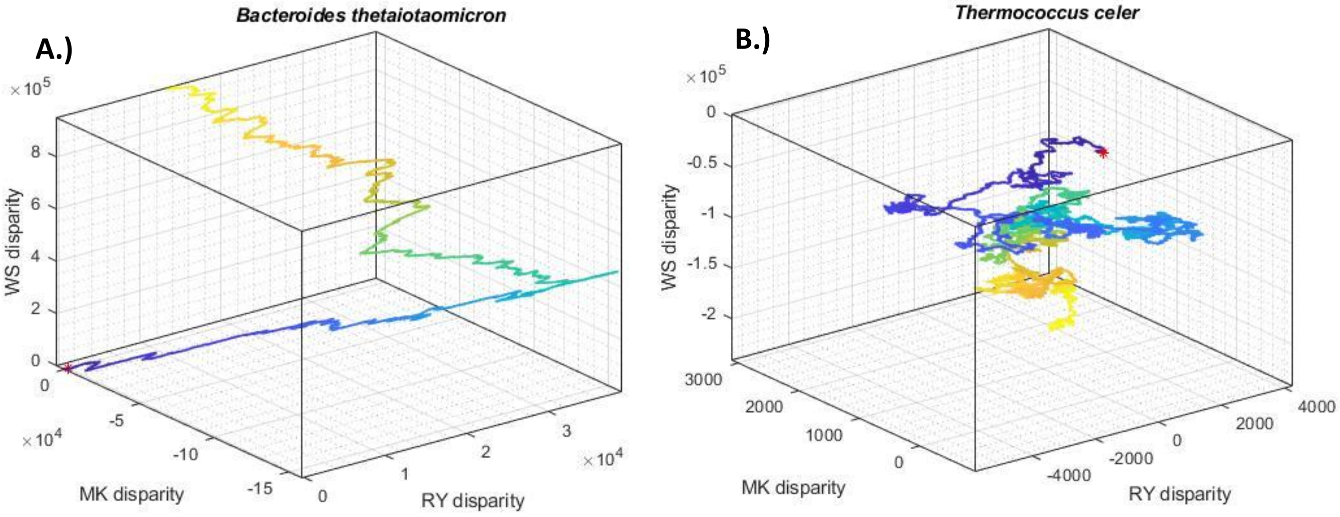


Figure 34. Comparison of Z-curves at opposite ends of the noise spectrum. A.) *B. thetaiotaomicron* with the a 'clean' skew profile. B.) *T. celer* with a 'high noise' skew profile.

7.3. Linkage of information processing genes to origins:

Linkage of origin-associated genes with information storage and processing genes commonly occurs in species where origins of replication are essential and may be a predictor of origin usage. Several species were found to have significant ($p < 0.05$) linkage for at least one origin-associated gene when tested 2x2 contingency χ^2 test.

7.4. Principal component analysis (PCA):

Principal component analysis graphs were plotted for each group mentioned in the skew indices section above as well as for all species. In each group, three positive control species were also plotted which have been experimentally proven to be able to replicate without origins of replication (*Hfx. volcanii, T. kodakarensis* and *A. cylindrica*) and three negative controls where origins are (near-) essential (*E. coli, Hfx. mediterranei* and *C. crescentus*). The PCA for all species can be seen in Figure 34. It should be noted that due to PCA scaling issues and for improved clarity, *D. thermolithotrophum* and *R. prowazekii* were omitted from Figure 35.

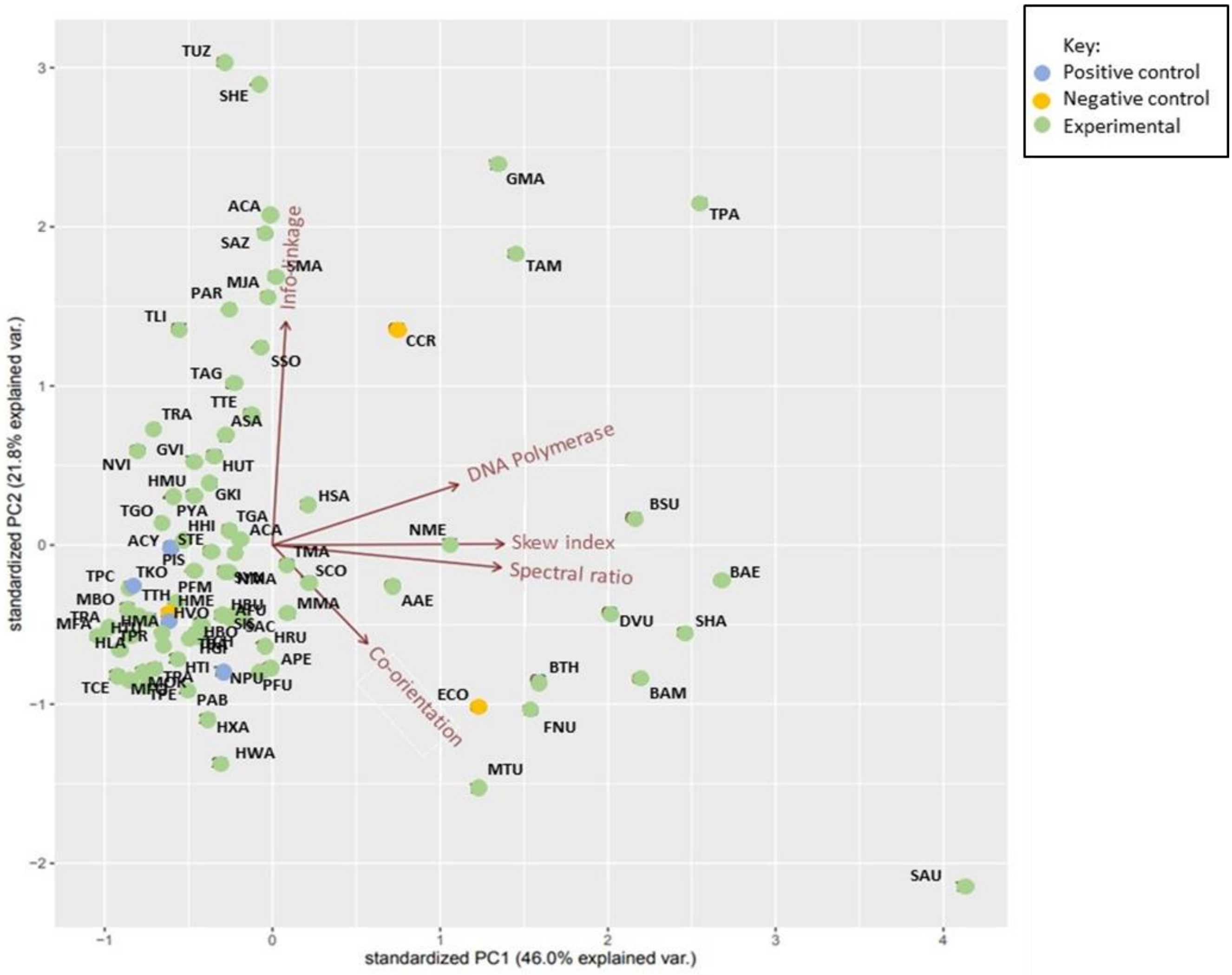


Figure 35. PCA showing positive and negative controls as well as all experimental species except for *D. thermolithotrophum* and *R. prowazekii.* Species which may be able to replicate without origins are likely to be plotted in the bottom left with low values, in proximity to those species known to replicate via origin independent replication. Three letter species codes can be seen in Tables 12-14.

The PCA in Figure 33 shows a large portion of species in the bottom left hand side of the PCA, where origin-independent replicating species are predicted to lie. However, due to the quantity of species on the PCA and the scaling, it is difficult to conclude from this PCA which species can or cannot replicate without origins. Hence, additional PCAs were created for each type of skew group mentioned previously.

The PCA for species in the ‘clean’ group showed the positive controls grouping in the top left of the PCA close to *N. maritimus* and *H. salinarum* (See Figure 36). Several other species could be grouped such as, *H. ruber*, *M. mazei* and *S. coelicolor.*

A. capsulatum is located at the bottom of left of the PCA close to *C. crescentus*, a species which is unable to replicate without origins. Therefore *A. capsulatum* can be suggested to be unable to replicate via RDR alongside other species located on the out skirts of the PCA such as *S. aureus.*

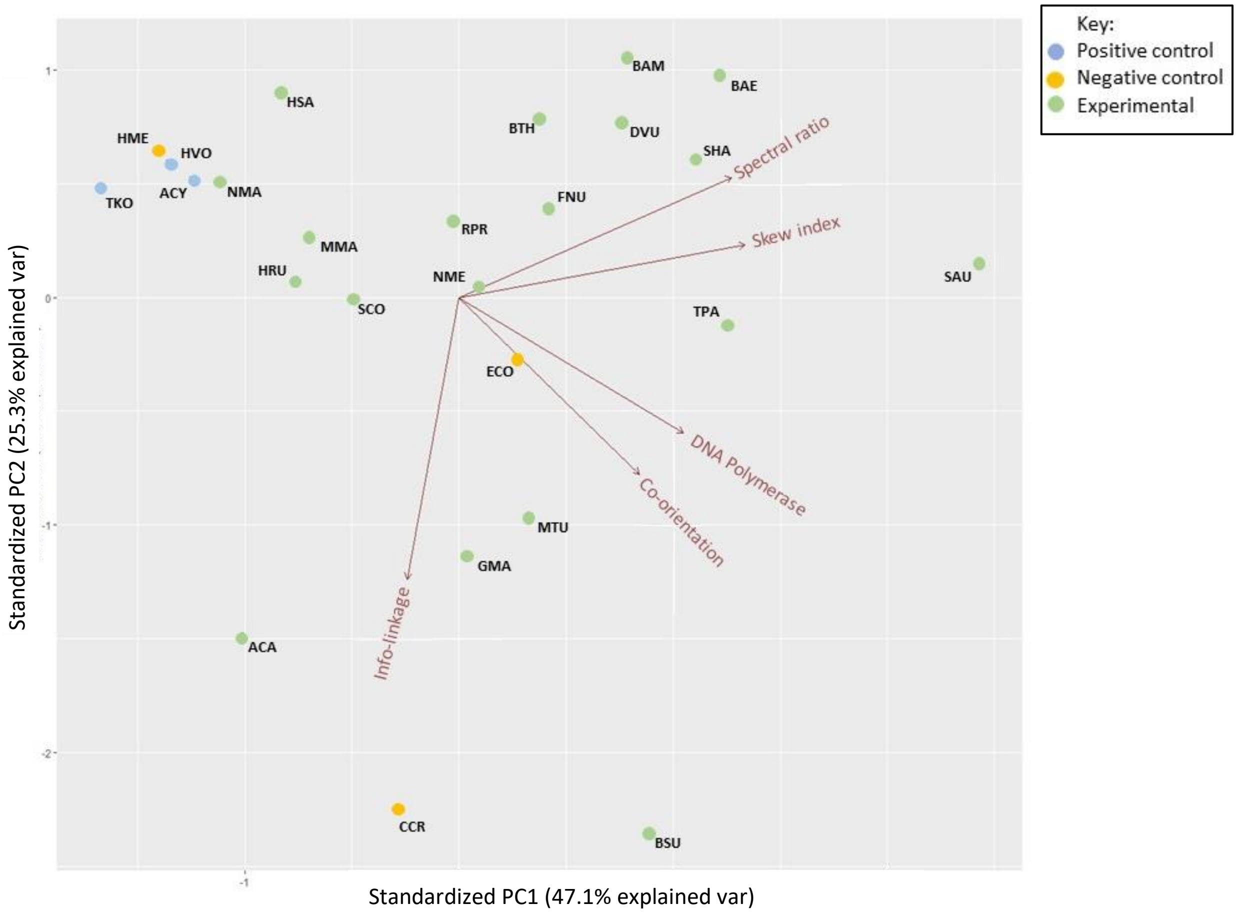


Figure 36. Principal component analysis for the species which had a 'clean' skew curve. Blue plots show control species experimentally proven to replicate via origin-independent replication. Orange plots show control species proven to be unable to replication via origin-independent replication and green plots show experimental species in which this study aims to predict.

The PCA for the 'moderate' group is largely clustered around the top left, with a few isolated species, two of which are negative controls. In general, this group has the most promise for being able to replicate without origins based on their skew profiles. There is a slight divide within the main cluster of species with a small cluster occurring above 0 for PC2 and the rest clustering just below 0. All positive controls lie in the smaller cluster of these two (Figure 37).

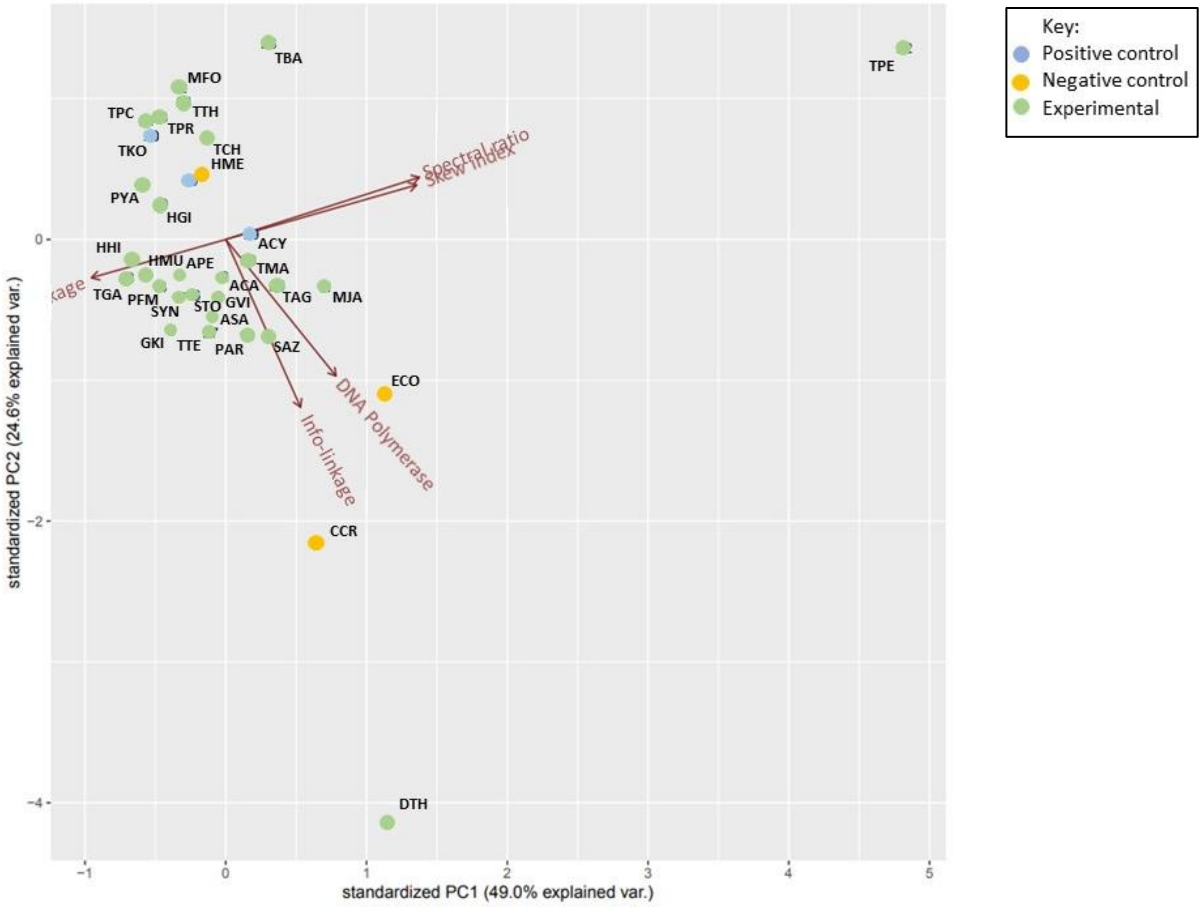


Figure 37. Principal component analysis for the species which had a 'moderate' skew curve. Blue plots show control species experimentally proven to replicate via origin-independent replication. Orange plots show control species proven to be unable to replication via origin-independent replication and green plots show experimental species in which this study aims to predict.

The 'high noise' group showed some degree of clustering towards the bottom left of the PCA, although not as tightly as on the other graphs. Two of the three negative controls are located far from any other species, as are *A. aeolicus*, *H. utahensis* and *N. viennensis* (Figure 38)**.**

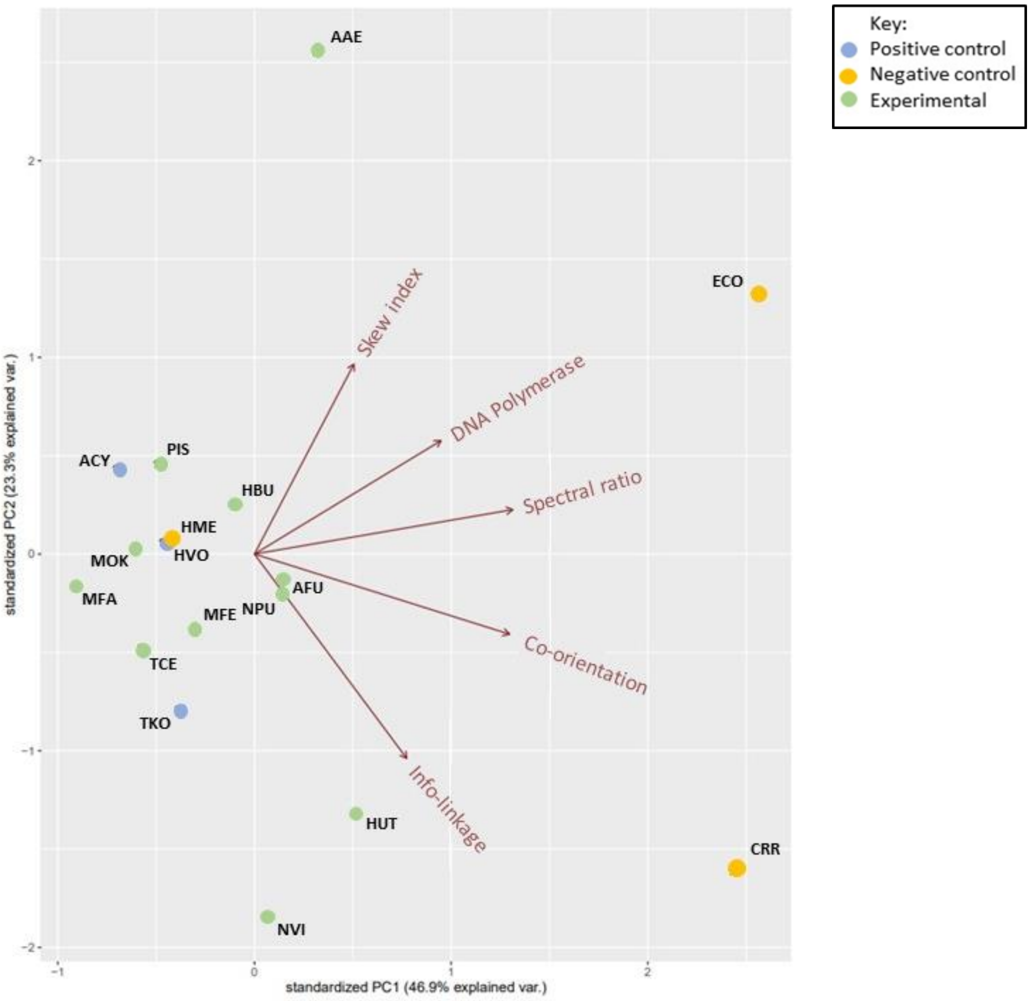


Figure 38. Principal component analysis for the species which had a 'high noise' skew curve. Blue plots show control species experimentally proven to replicate via origin independent replication. Orange plots show control species proven to be unable to replication via origin independent replication and green plots show experimental species in which this study aims to predict.

The 'other' group shows the majority of species clustered to the left centre of the PCA alongside two of the positive controls. A few species including the third positive control inhabit the top left of the graph in a sparse pattern. Once again, two of the negative controls lie far away from any other species as well as *T. ammonificans*, which can be assumed to be unable to replicate via RDR (Figure 39). Species clustering around *T. kodakarensis* in the bottom left corner are likely candidates for origin-independent replication as this species has shown the greatest use of RDR.

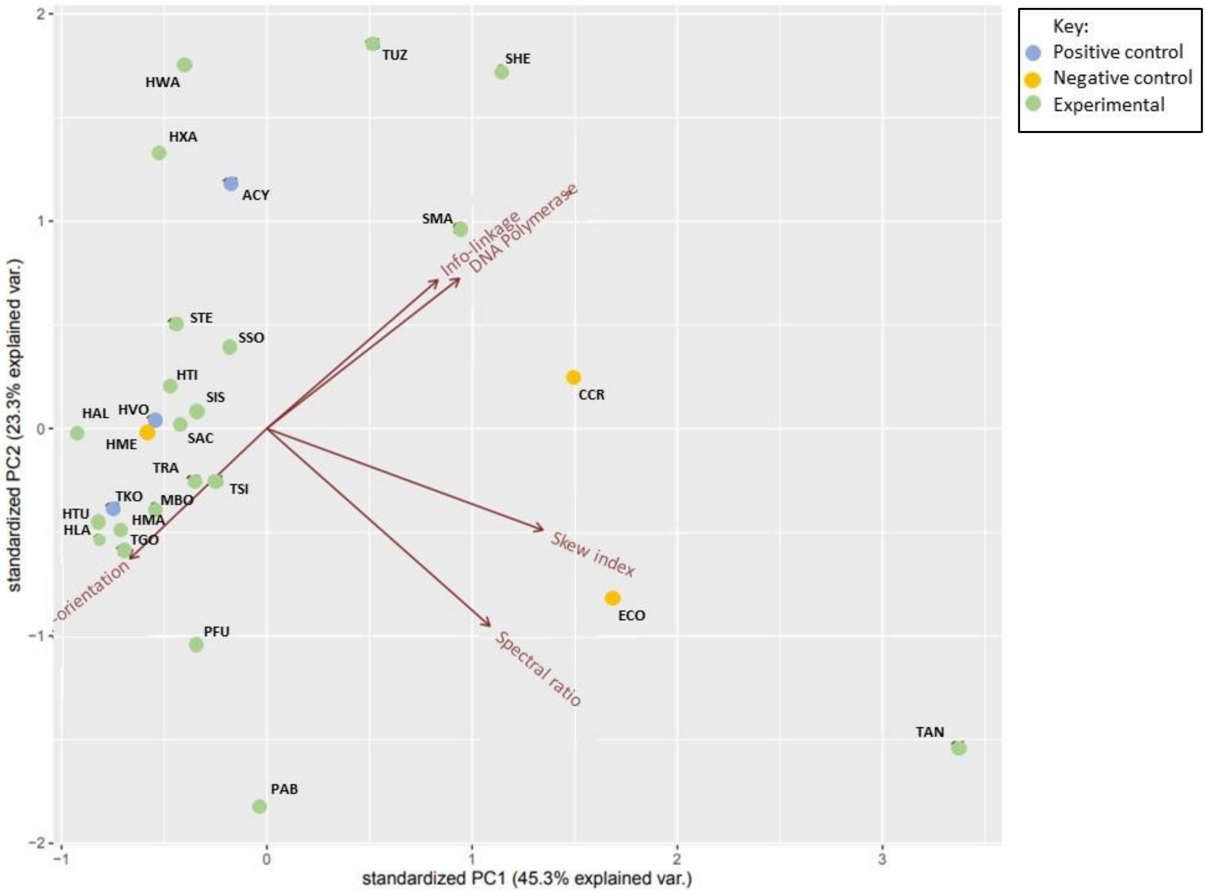


Figure 39. Principal component analysis for the species which had an 'other' skew curve. Blue plots show control species experimentally proven to replicate via origin independent replication. Orange plots show control species proven to be unable to replication via origin independent replication and green plots show experimental species in which this study aims to predict.

In summary, a large proportion of the species examined, particularly archaeal species, cluster in the region expected for organisms that can replicate via origin-independent replication, in the vicinity of those species that have been experimentally proven to be able to replicate in this manner. The model predicts that species most likely to survive in the absence of origins of replication include: *N. maritimus*, *Hfx. gibbonsii*, *P. yayanosi*, *A. fulgidus*, *M. okinawensis*, *M. formicica*, *T. celer*, *M. fervens*, *H. marismortui*, *T. gorgonarius*, *H. lacusprofundi*, *H. turkmenica* and *M. boonei*. These species hold great promise for experimental verification of their ability to replicate via RDR. It is also noted that *Hfx. mediterranei* is a clear outlier within the negative controls, as predictions suggest it should be able to replicate without origins.

However, this species has been shown experimentally to require origins, it has been found that when they deleted three active origins, a dormant origin became active and was able to replicate the entire chromosome. These dormant origins can be used as a back-up if the replication fork has stalled for any reason and may be beneficial in harsh intracellular or extracellular conditions (Yang *et al*, 2015). Therefore, when considering the locations of other species in relation to positive controls, the chance of other false positives such as *Hfx. mediterranei* cannot be ruled out.

There are several species that the PCA strongly predicts to be dependent on origins, the most notable being: *T. ammonificans*, *D. thermolithotrophum, R. prowazekii, S. hellenicus, T. uzoniensis, A. aeolicus, B. subtilis S. aureus, T. peptonophilus* and *T. pallidum*. Several other species lie in locations in the PCA where their ability to replicate without origins is unlikely but could not be confidently predicted.

8. Discussion

The purpose of this study was to adapt a previously developed prediction tool to better fit larger data sets and predict whether a wider range of species can survive in the absence of origins. The tool has shown that species which are able to survive without origins cluster towards the left of the PCA graphs created whereas species which cannot tend to lie more to the right. This has suggested numerous species which may be able to replicate without origins, in particular Archaeal species. The grouping of species from left to right along the x axis suggests that factors contributing to the distribution on this dimension are perhaps more important than those contributing to the y axis, which display a more sporadic pattern. It is therefore inferred that co-orientation of core gene, skew indices and linkage to DNA polymerases are most important to this prediction tool, while information gene linkage and spectral ratio are less so. It is possible that the information gene linkage value was less important due to the method utilised for the study. A more robust method of calculating information storage and processing genes around the origin and better knowledge of unclassified genes for specific species may be required.

The modifications to the tool to allow for better use for larger data sets, since some variables were removed that were previously suggested to be of little to no impact on predictive results; aspects of code were also updated to reduce manual input. This is largely reflected in the PCAs: previously tested species such as the control species, alongside several others including *H. marismortui, P. abyssi, H. hispanica, H. borinquense, N. maritimus, Synechococcus sp., T. gammatolerans, A. fulgidus and H. lacusprofundi*, all lie in similar locations along the x axis. The position on the y axis does however differ. The one reason for this could be that non-native genes with high information linkage are causing variance in this axis, this was one factor that the original tool screened out, but this part of the pipeline was removed due to the manual and time-consuming process required. The original tool also reported that in all the species previously tested, only one *Hfx. volcanii* had an origin associated gene in a non-native area. That being said this study tested a significantly larger range of species hence it is plausible that a number of them had origin-associated genes in non-native areas. Native and non-native regions can be determined using a codon adaptive index (CAI). This measures codon bias, as organisms typically favour certain codons in translational use. Translationally favoured codons are frequently found in highly expressed genes. Adlam (2018) used the CAI statistic instead as a measure of codons that were found frequently across the genome. A CAI statistic of 1 would correspond with codon usage that matches that of the rest of the genome (native). Whereas a CAI statistic of 0 suggests completely different usage and the presence of a non-native gene (Adlam 2018: Xia, 2007). This method was chosen over the standard CAI technique as a way to assess lateral gene transfers. However, the issue with this as previously mentioned is the high amount of manual input and time required to assess a genome. This is not an issue for species such as *Hfx. volcanii* where good data exists on native and non-native regions, but this is not the case for many other species. As a result, predictions of native/non-native genes for large data sets of less studied organisms remains a problem to be solved. In order to improve the tool, perhaps a better method of screening native/non-native genes is required that does not rely on codon adaptive index screening for all genes within a genome.

Although the changes to this tool remove some of the main issues with its utilisation for larger data sets, the core issues still exist. These include issues with skew indices for species with multiple origins of replication. The work by Arakawa and Tomita (2007) suggested there should be singular maximum and minimum skew for each species, but this is not the case when multiple origins of replication are considered. Skew profiles on nucleotide disparity graphs seem to contradict skew index calculations. In almost all cases GCSI and ATSI are calculated to be the maxima skew although on nucleotide disparity plots this is not the case. Due to the Euclidean distance value calculated within the skew index code, singular maxima and minima are provided for species with one origin. However, when several origins are involved as is the case with many Archaea species, several maxima and minima are calculated and only the overall maximum value and minimum value are considered. This could cause significant issues in predictions, especially for those with large quantities of origins.

As mentioned above, the second core issue lies in the calculation of linkage between the origin and information storage and processing genes. This study updated arCOGs/COGS to the most recent data set (Galperin *et al,* 2015;2019). However, many genes are still unclassified or grouped in the functional classes R and S which have vague non-descript functions, which may or may not fall into the information storage and processing category once better understood; this is unlikely as scientific understanding of replication is advanced but cannot be ruled out as a potential issue. Secondly, the creation of a specific tool to count relevant genes around the origin-associated gene would progress this tool immensely as the current system relies on a combination of manual counting and locating of the origin-associated genes on the integrated microbial genomes and microbiomes tool JGI (The Regents of the University of California, 2020). The introduction of such a tool would not only remove any potential for human error but also allow for increased automation further improving the capacity for larger data sets.

8.1. Summary

The improved bioinformatic tool created in this study allows for larger quantities of species to be screened in a shorter amount of time, however a relatively large amount of manual input is still required in some areas, especially for information gene linkage calculations. The modified tool has been used to predict origin usage in 85 species, excluding species which had compatibility issues with the tool or lacked the required information for analysis. These species show several promising candidates which may be able to survive without origins, as well as several species which can confidently be suggested to be origin-dependent, when compared with known example from literature. Therefore, this tool may can provide direction for future *in vivo*_studies. In order to improve the confidence of predictions, an increased sample size of known species is required to aid in differentiation between tightly clustered groups of species.

8.2. Future work

Next steps for this tool should include alterations to the calculation of linkage of origin associated genes with information and storage processing genes. This would allow for improved accuracy and would be less time-consuming. The addition of multiple maxima and minima calculations to skew index code would account for species with multiple origins of replication and result in increased accuracy for archaeal species; perhaps two separate pipelines should be created, one for archaeal species and one for Bacteria. Finally, the reintroduction of screening for non-native origin associated genes, which may interfere with other calculations, should be considered. This will however make the tool more time-consuming to use and less appropriate for large scale screening, hence it is proposed that this be conducted after the initial species of interest have been identified. Alternatively, an automated method for eliminating non-native genes could be developed.

References

1. Hawkins, M., Malla, S., Blythe, M.J., Nieduszynski, C.A. and Allers, T., 2013. Accelerated growth in the absence of DNA replication origins. Nature, 503(7477), pp.544-547.
2. Woese, C.R. and Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences, 74(11), pp.5088-5090.
3. Woese, C.R., Kandler, O. and Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences, 87(12), pp.4576-4579.
4. Allers, T. and Mevarech, M., 2005. archaeal genetics—the third way. Nature Reviews Genetics, 6(1), pp.58-73.
5. Winker, S. and Woese, C.R., 1991. A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. Systematic and Applied Microbiology, 14(4), pp.305-310.
6. Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H. and Chun, J., 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. International Journal of Systematic and Evolutionary Microbiology, 67(5), p.1613.
7. Auchtung, T.A., Takacs-Vesbach, C.D. and Cavanaugh, C.M., 2006. 16S rRNA phylogenetic investigation of the candidate division “Korarchaeota”. Applied and Environmental Microbiology, 72(7), pp.5077-5082.
8. Brochier-Armanet, C., Boussau, B., Gribaldo, S. and Forterre, P., 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nature Reviews Microbiology, 6(3), pp.245-252.
9. Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.J., Hattori, M., Kanai, A., Atomi, H. and Takai, K., 2011. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Research, 39(8), pp.3204-3223.
10. Barns, S.M., Delwiche, C.F., Palmer, J.D. and Pace, N.R., 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proceedings of the National Academy of Sciences, 93(17), pp.9188-9193.
11. Koonin, E.V., 2015. Archaeal ancestors of eukaryotes: not so elusive anymore. BMC Biology, 13(1), pp.1-7.
12. Londei, P., 2005. Evolution of translational initiation: new insights from the Archaea. FEMS Microbiology Reviews, 29(2), pp.185-200.

13. Huet, J., Schnabel, R., Sentenac, A. and Zillig, W., 1983. Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. The EMBO Journal, 2(8), pp.1291-1294.
14. Kelman, Z. and White, M.F., 2005. archaeal DNA replication and repair. Current Opinion in Microbiology, 8(6), pp.669-676.
15. O'Donnell, M., Langston, L. and Stillman, B., 2013. Principles and concepts of DNA replication in Bacteria, Archaea, and eukarya. Cold Spring Harbor Perspectives in Biology, 5(7), p.a010108.
16. Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S. and Woese, C.R., 1977. Classification of methanogenic Bacteria by 16S ribosomal RNA characterization. Proceedings of the National Academy of Sciences, 74(10), pp.4537-4541.
17. Rother, M. and Metcalf, W.W., 2005. Genetic technologies for Archaea. Current Opinion in Microbiology, 8(6), pp.745-751.
18. Embley, T.M. and Williams, T.A., 2015. Steps on the road to eukaryotes. Nature, 521(7551), pp.169-170.
19. Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U. and Stott, M.B., 2017. Asgard Archaea illuminate the origin of eukaryotic cellular complexity. Nature, 541(7637), pp.353-358.
20. Imachi, H., Nobu, M.K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M. and Matsui, Y., 2020. Isolation of an archaeon at the prokaryote–eukaryote interface. Nature, 577(7791), pp.519-525.
21. DasSarma, S. and DasSarma, P., 2015. Halophiles and their enzymes: negativity put to good use. Current Opinion in Microbiology, 25, pp.120-126.
22. Javor, B.J., 2012. Hypersaline environments: microbiology and biogeochemistry. Springer Science & Business Media.
23. Díaz-Cárdenas, C., Cantillo, A., Rojas, L.Y., Sandoval, T., Fiorentino, S., Robles, J., Ramos, F.A., Zambrano, M.M. and Baena, S., 2017. Microbial diversity of saline environments: searching for cytotoxic activities. AMB Express, 7(1), p.223.
24. Margesin, R. and Schinner, F., 2001. Potential of halotolerant and halophilic microorganisms for biotechnology. Extremophiles, 5(2), pp.73-83.
25. Kates, M., Kushner, D.J. and Matheson, A.T. eds., 1993. The biochemistry of archaea (archaebacteria). Elsevier.
26. Wood, J.M., 2015. Bacterial responses to osmotic challenges. Journal of General Physiology, 145(5), pp.381-388.
27. Oren, A., 2011. Thermodynamic limits to microbial life at high salt concentrations. Environmental Microbiology, 13(8), pp.1908-1923.

28. Roberts, M.F., 2005. Organic compatible solutes of halotolerant and halophilic microorganisms. Saline Systems, 1(1), p.5.

29. Oren, A., 2008. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Systems, 4(1), p.2.

30. Mullakhanbhai, M.F. and Larsen, H., 1975. *Halobacterium volcanii* spec. nov., a Dead Sea halobacterium with a moderate salt requirement. Archives of Microbiology, 104(1), pp.207-214.

31. Zhou, G., Kowalczyk, D., Humbard, M.A., Rohatgi, S. and Maupin-Furlow, J.A., 2008. Proteasomal components required for cell growth and stress responses in the haloarchaeon *Haloferax volcanii.* Journal of Bacteriology, 190(24), pp.8096-8105.

32. Norais, C., Hawkins, M., Hartman, A.L., Eisen, J.A., Myllykallio, H. and Allers, T., 2007. Genetic and physical mapping of DNA replication origins in *Haloferax volcanii*. PLoS Genet, 3(5), p.e77.

33. Hartman, A.L., Norais, C., Badger, J.H., Delmas, S., Haldenby, S., Madupu, R., Robinson, J., Khouri, H., Ren, Q., Lowe, T.M. and Maupin-Furlow, J., 2010. The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. PLOS One, 5(3), p.e9605.

34. Gophna, U., Allers, T. and Marchfelder, A., 2017. Finally, Archaea get their CRISPR-Cas toolbox. Trends in Microbiology, 25(6), pp.430-432.

35. Allers, T., Ngo, H.P., Mevarech, M. and Lloyd, R.G., 2004. Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the leuB and trpA genes. Applied and Environmental Microbiology, 70(2), pp.943-953.

36. Bitan-Banin G, Ortenberg R, Mevarech M., 2003. Development of a gene knockout system for the halophilic archaeon *Haloferax volcanii* by use of the *pyrE* gene. Journal of Bacteriology, 185(3), pp.772-778.

37. Ortenberg, R., Rozenblatt-Rosen, O. and Mevarech, M., 2000. The extremely halophilic archaeon *Haloferax volcanii* has two very different dihydrofolate reductases. Molecular Microbiology, 35(6), pp.1493-1505.

38. Allers, T., Barak, S., Liddell, S., Wardell, K. and Mevarech, M., 2010. Improved strains and plasmid vectors for conditional overexpression of His-tagged proteins in *Haloferax volcanii*. Applied and Environmental Microbiology, 76(6), pp.1759-1769.

39. Holmes, M.L., Nuttall, S.D. and Dyall-Smith, M.L., 1991. Construction and use of halobacterial shuttle vectors and further studies on *Haloferax* DNA gyrase. *Journal of Bacteriology*, *173*(12), pp.3807-3813.

40. Cline, S.W., Lam, W.L., Charlebois, R.L., Schalkwyk, L.C. and Doolittle, W.F., 1989. Transformation methods for halophilic archaebacteria. Canadian Journal of Microbiology, 35(1), pp.148-152.

41. Holmes, M.L. and Dyall-Smith, M.L., 2000. Sequence and expression of a halobacterial β-galactosidase gene. Molecular Microbiology, *36*(1), pp.114-122.

42. Delmas, S., Shunburne, L., Ngo, H.P. and Allers, T., 2009. Mre11-Rad50 promotes rapid repair of DNA damage in the polyploid archaeon *Haloferax volcanii* by restraining homologous recombination. PLOS Genet, 5(7), p.e1000552.

43. Crameri, A., Whitehorn, E.A., Tate, E. and Stemmer, W.P., 1996. Improved green fluorescent protein by molecular evolution using DNA shuffling. Nature Biotechnology, 14(3), pp.315-319.

44. Reuter, C.J. and Maupin-Furlow, J.A., 2004. Analysis of proteasome-dependent proteolysis in *Haloferax volcanii* cells, using short-lived green fluorescent proteins. Applied and Environmental Bicrobiology, 70(12), pp.7530-7538.

45. Duggin, I.G., Aylett, C.H., Walsh, J.C., Michie, K.A., Wang, Q., Turnbull, L., Dawson, E.M., Harry, E.J., Whitchurch, C.B., Amos, L.A. and Löwe, J., 2015. CetZ tubulin-like proteins control archaeal cell shape. Nature, 519(7543), pp.362-365.

46. Kornberg, D.N.A. and TA, D., 1980. Replication. San Francisco: W H. Freeman.

47. Bell, S.P. and Dutta, A., 2002. DNA replication in eukaryotic cells. Annual Review of Biochemistry, 71(1), pp.333-374.

48. Makarova, K.S., Koonin, E.V. and Kelman, Z., 2012. The CMG (CDC45/RecJ, MCM, GINS) complex is a conserved component of the DNA replication system in all Archaea and eukaryotes. Biology Direct, 7(1), p.7.

49. Matsunaga, F., Norais, C., Forterre, P. and Myllykallio, H., 2003. Identification of short 'eukaryotic'Okazaki fragments synthesized from a prokaryotic replication origin. EMBO Reports, 4(2), pp.154-158.

50. Ausiannikava, D. and Allers, T., 2017. Diversity of DNA replication in the Archaea. Genes, 8(2), p.56.

51. Barry, E.R. and Bell, S.D., 2006. DNA replication in the Archaea. Microbiology and Molecular Biology Reviews, 70(4), pp.876-887.

52. Böhlke, K., Pisani, F.M., Rossi, M. and Antranikian, G., 2002. archaeal DNA replication: spotlight on a rapidly moving field. Extremophiles, 6(1), pp.1-14.

53. MacNeill, S.A., 2001. Understanding the enzymology of archaeal DNA replication: progress in form and function. Molecular Microbiology, 40(3), pp.520-529.

54. Duggin, I.G., Wake, R.G., Bell, S.D. and Hill, T.M., 2008. The replication fork trap and termination of chromosome replication. Molecular Microbiology, 70(6), pp.1323-1333.

55. Eydmann, T., Sommariva, E., Inagawa, T., Mian, S., Klar, A.J.S. and Dalgaard, J.Z., 2008. Rtf1-mediated eukaryotic site-specific replication termination. Genetics, 180(1), pp.27-39.

56. Friedberg, E.C., 2003. DNA damage and repair. Nature, 421(6921), pp.436-440.

57. Sartori, A.A. and Jiricny, J., 2003. Enzymology of base excision repair in the hyperthermophilic archaeon *Pyrobaculum aerophilum*. Journal of Biological Chemistry, *278*(27), pp.24563-24576.

58. de Laat, W.L., Jaspers, N.G. and Hoeijmakers, J.H., 1999. Molecular mechanism of nucleotide excision repair. Genes & Development, *13*(7), pp.768-785.

59. Schaaper, R.M., 1993. Base selection, proofreading, and mismatch repair during DNA replication in *Escherichia coli*. Journal of Biological Chemistry, 268(32), pp.23762-23765.

60. White, M.F., 2011. Homologous recombination in the Archaea: the means justify the ends. Biochemical Society Transactions, 39(1), pp. 15-19.

61. Popova, M., Henry, S. and Fleury, F., 2011. Posttranslational modifications of Rad51 protein and its direct partners: Role and effect on homologous recombination–mediated DNA repair. DNA Repair, 1, pp.143-160.

62. Ragunathan, K., Liu, C. and Ha, T., 2012. RecA filament sliding on DNA facilitates homology search. Elife, *1*, p.e00067.

63. Rocha, E.P., Cornet, E. and Michel, B., 2005. Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLOS Genet*, *1*(2), p.e15.

64. Bonetti, D., Colombo, C.V., Clerici, M. and Longhese, M.P., 2018. Processing of DNA ends in the maintenance of genome stability. Frontiers in genetics, 9, pp.390.

65. Chen, Z., Yang, H. and Pavletich, N.P., 2008. Mechanism of homologous recombination from the RecA–ssDNA/dsDNA structures. Nature, *453*(7194), pp.489-494.

66. Klapstein, K., Chou, T. and Bruinsma, R., 2004. Physics of RecA-mediated homologous recognition. Biophysical Journal, 87(3), pp.1466-1477.

67. Constantinesco, F., Forterre, P., Koonin, E.V., Aravind, L. and Elie, C., 2004. A bipolar DNA helicase gene, herA, clusters with rad50, mre11 and nurA genes in thermophilic Archaea. Nucleic Acids Research, 32(4), pp.1439-1447.

68. Wardell, K., Haldenby, S., Jones, N., Liddell, S., Ngo, G.H. and Allers, T., 2017. RadB acts in homologous recombination in the archaeon *Haloferax volcanii*, consistent with a role as recombination mediator. DNA Repair, 55, pp.7-16.

69. Guy, C.P., Haldenby, S., Brindley, A., Walsh, D.A., Briggs, G.S., Warren, M.J., Allers, T. and Bolt, E.L., 2006. Interactions of RadB, a DNA repair protein in Archaea, with DNA and ATP. Journal of Molecular Biology, 358(1), pp.46-56.

70. Kil, Y.V., Baitin, D.M., Masui, R., Bonch-Osmolovskaya, E.A., Kuramitsu, S. and Lanzov, V.A., 2000. Efficient strand transfer by the RadA recombinase from the hyperthermophilic archaeon *Desulfurococcus amylolyticus*. Journal of Bacteriology, 182(1), pp.130-134.

71. McKinney, S.A., Déclais, A.C., Lilley, D.M. and Ha, T., 2003. Structural dynamics of individual Holliday junctions. Nature Structural Biology, 10(2), pp.93-97.

72. Eggleston, A.K. and West, S.C., 2000. Cleavage of holliday junctions by the *Escherichia coli* RuvABC complex. Journal of Biological Chemistry, 275(34), pp.26467-26476.

73. Iwasaki, H., Takahagi, M., Shiba, T., Nakata, A. and Shinagawa, H., 1991. Escherichia coli RuvC protein is an endonuclease that resolves the Holliday structure. The EMBO Journal, 10(13), pp.4381-4389.

74. Chan, S.N., Vincent, S.D. and Lloyd, R.G., 1998. Recognition and manipulation of branched DNA by the RusA Holliday junction resolvase of *Escherichia coli*. Nucleic Acids Research, 26(7), pp.1560-1566.

75. Schwartz, E. K. and Heyer, W. D. (2011) 'Processing of joint molecule intermediates by structure-selective endonucleases during homologous recombination in eukaryotes', Fdel Chromosoma, 120(2), pp. 109–127.

76. Bolt, E. L., Lloyd, R. G. and Sharples, G. J. (2001) 'Genetic analysis of an archaeal Holliday junction resolvase in *Escherichia coli*', Journal of Molecular Biology, 310(3), pp. 577–589.

77. Lestini, R., Duan, Z. and Allers, T., 2010. The archaeal Xpf/Mus81/FANCM homolog Hef and the Holliday junction resolvase Hjc define alternative pathways that are essential for cell viability in *Haloferax volcanii*. DNA repair, 9(9), pp.994-1002.

78. Kogoma, T., 1997. Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. Microbiology and Molecular Biology Reviews, 61(2), pp.212-238.

79. Michel, B. and Bernander, R., 2014. Chromosome replication origins: do we really need them?. Bioessays, 36(6), pp.585-590.

80. Gehring, A.M., Astling, D.P., Matsumi, R., Burkhart, B.W., Kelman, Z., Reeve, J.N., Jones, K.L. and Santangelo, T.J., 2017. Genome replication in *Thermococcus kodakarensis* independent of Cdc6 and an origin of replication. Frontiers in Microbiology, 8, p.2084.

81. Wang, J.D., Sanders, G.M. and Grossman, A.D., 2007. Nutritional control of elongation of DNA replication by (p) ppGpp. *Cell*, *128*(5), pp.865-875.

82. Leslie, D.J., Heinen, C., Schramm, F.D., Thüring, M., Aakre, C.D., Murray, S.M., Laub, M.T. and Jonas, K., 2015. Nutritional control of DNA replication initiation through the proteolysis and regulated translation of DnaA. *PLoS Genetics*, *11*(7), p.e1005342.

83. Bailey, L., 2005 (unpublished).

84. Allers, T., 2015 (unpublished).

85. Lever, R.J., 2019. Genetic and Biochemical analysis of the hel308 helicase in the archaeon *Haloferax volcanii.* PhD Thesis, University of Nottingham, Nottingham.

86. Large, A., Stamme, C., Lange, C., Duan, Z., Allers, T., Soppa, J. and Lund, P.A., 2007. Characterization of a tightly controlled promoter of the halophilic archaeon *Haloferax volcanii* and its use in the analysis of the essential cct1 gene. Molecular Microbiology, 66(5), pp.1092-1106.

87. Haque, R.U., Paradisi, F. and Allers, T., 2019. *Haloferax volcanii* as immobilised whole cell biocatalyst: new applications for halophilic systems. Applied Microbiology and Biotechnology, 103(9), pp.3807-3817.

88. Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74(12), pp.5463-5467.

89. Dyall-Smith, M., 2015. The Halo Handbook v7.3. 10.13140/RG.2.1.1750.5441.

90. Gamble-Milner, R., 2016. Genetic analysis of the Hel308 helicase in the archaeon *Haloferax volcanii* (Doctoral dissertation, University of Nottingham).

91. Stiller, M. and Nissenbaum, A., 1999. Geochemical investigation of phosphorus and nitrogen in the hypersaline Dead Sea. *Geochimica et Cosmochimica Acta*, *63*(19-20), pp.3467-3475.

92. Buckley, R.J., Kramm, K., Cooper, C.D., Grohmann, D. and Bolt, E.L., 2020. Mechanistic insights into Lhr helicase function in DNA repair. Biochemical Journal, 477(16), pp.2935-2947.

93. Samson, R.Y., Xu, Y., Gadelha, C., Stone, T.A., Faqiri, J.N., Li, D., Qin, N., Pu, F., Liang, Y.X., She, Q. and Bell, S.D., 2013. Specificity and function of archaeal DNA replication initiator proteins. Cell Reports, 3(2), pp.485-496.

94. Arakawa, K. and Tomita, M., 2007. The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection. Evolutionary Bioinformatics, 3, p.117693430700300006.

95. Zhang, C.T., Zhang, R. and Ou, H.Y., 2003. The Z curve database: a graphic representation of genome sequences. Bioinformatics, 19(5), pp.593-599.

96. Wu, Z., Liu, J., Yang, H. and Xiang, H., 2014. DNA replication origins in Archaea. *Frontiers in microbiology*, *5*, p.179.

97. Myllykallio, H. and Forterre, P., 2000. Mapping of a chromosome replication origin in an archaeon: response. Trends in Microbiology, 8(12), pp.537-539.

98. Luo, H., Zhang, C.T. and Gao, F., 2014. Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. Frontiers in Microbiology, 5, p.482.

99. Dao, F.Y., Lv, H., Wang, F., Feng, C.Q., Ding, H., Chen, W. and Lin, H., 2019. Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. Bioinformatics, 35(12), pp.2075-2083.

100. Adlam, C, 2018. Life Without DNA Replication Origins: What is Required? MSci thesis, University of Nottingham, Nottingham.

101. Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of Bacteria. Molecular Biology and Evolution, 13(5), pp.660-665.

102. Lobry, J.R. and Sueoka, N., 2002. Asymmetric directional mutation pressures in Bacteria. Genome Biology, 3(10), pp.research0058-1.

103. Čuboňová, L., Richardson, T., Burkhart, B.W., Kelman, Z., Connolly, B.A., Reeve, J.N. and Santangelo, T.J., 2013. archaeal DNA polymerase D but not DNA polymerase B is required for genome replication in *Thermococcus kodakarensis*. Journal of Bacteriology, 195(10), pp.2322-2328.

104. Maurer, S., Ludt, K. and Soppa, J., 2018. Characterization of copy number control of two *Haloferax volcanii* replication origins using deletion mutants and haloarchaeal artificial chromosomes. Journal of Bacteriology, 200(1).

105. Hogrel, G., Lu, Y., Alexandre, N., Bossé, A., Dulermo, R., Ishino, S., Ishino, Y. and Flament, D., 2020. Role of RadA and DNA Polymerases in Recombination-Associated DNA Synthesis in Hyperthermophilic Archaea. Biomolecules, 10(7), p.1045.

106. Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research, 28(1), pp.33-36.

107. Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V., 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Research, 43(D1), pp.D261-D269.

108. Paul, S., Million-Weaver, S., Chattopadhyay, S., Sokurenko, E. and Merrikh, H., 2013. Accelerated gene evolution through replication–transcription conflicts. Nature, 495(7442), pp.512-515.

109. Pomerantz, R.T. and O'Donnell, M., 2010. What happens when replication and transcription complexes collide?. Cell Cycle, 9(13), pp.2537-2543

110. Dimude, J.U., Midgley-Smith, S.L., Stein, M. and Rudolph, C.J., 2016. Replication termination: containing fork fusion-mediated pathologies in *Escherichia coli.* Genes, 7(8), p.40.

111. Zhang, R. and Zhang, C.T., 2005. Identification of replication origins in archaeal genomes based on the Z-curve method. Archaea, 1(5), pp.335-346.

112. Galperin, M.Y., Kristensen, D.M., Makarova, K.S., Wolf, Y.I. and Koonin, E.V., 2019. Microbial genome analysis: the COG approach. Briefings in Bioinformatics, 20(4), pp.1063-1070.

113. Pelve, E.A., Lindås, A.C., Knöppel, A., Mira, A. and Bernander, R., 2012. Four chromosome replication origins in the archaeon *Pyrobaculum calidifontis*. Molecular Microbiology, 85(5), pp.986-995.

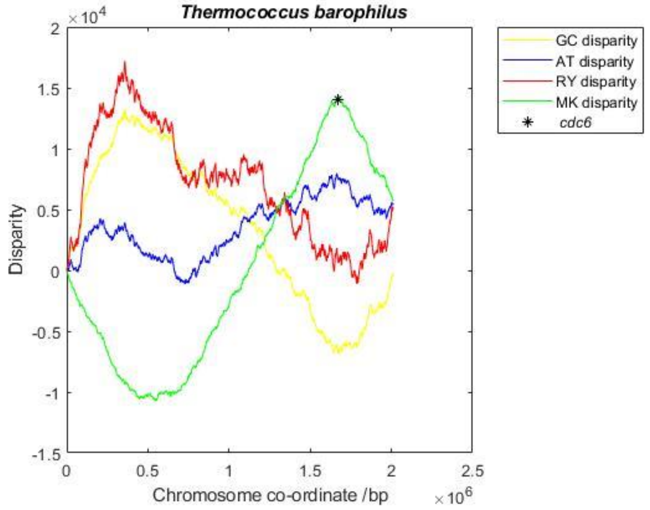
114. Despalins, A., Marsit, S. and Oberto, J., 2011. Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. Bioinformatics, 27(20), pp.2905-2906.

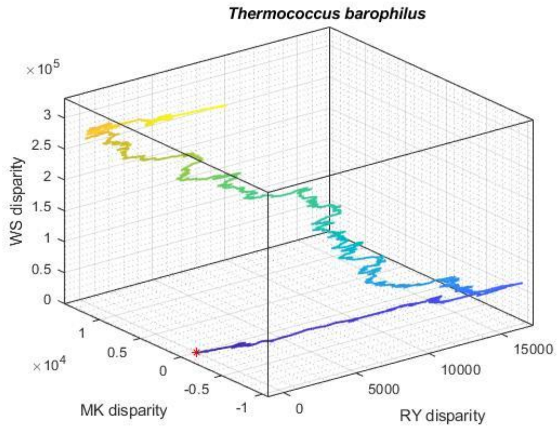
115. Yang, H., Wu, Z., Liu, J., Liu, X., Wang, L., Cai, S. and Xiang, H., 2015. Activation of a dormant replication origin is essential for *Haloferax mediterranei* lacking the primary origins. Nature communications, 6(1), pp.1-11.

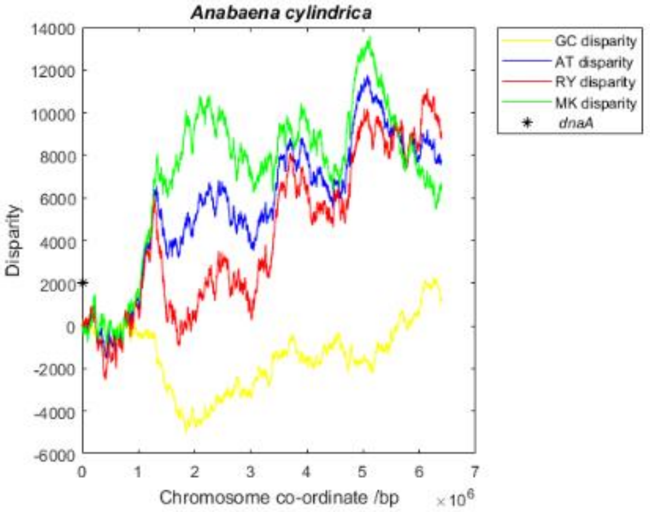
116. Xia, X., 2007. An improved implementation of codon adaptation index. Evolutionary Bioinformatics, 3*(1)*, p.53-58.

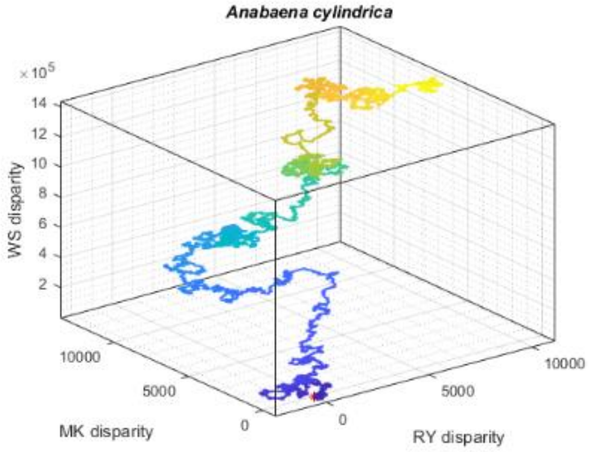
Appendix

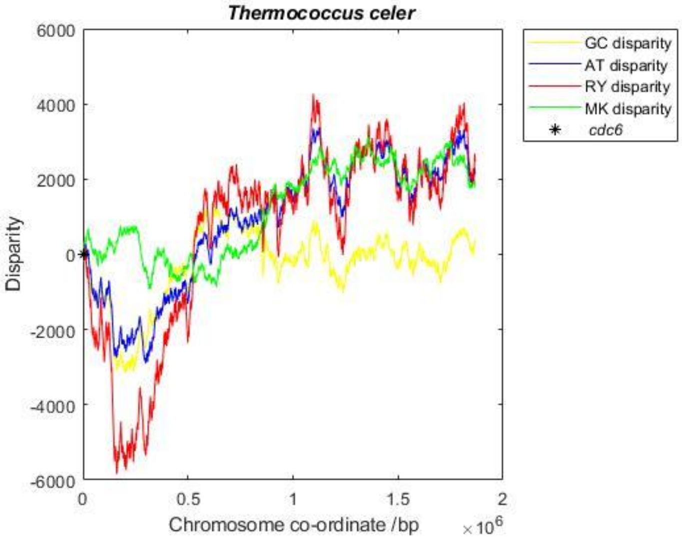
Additional nucleotide skew graphs and z-curves

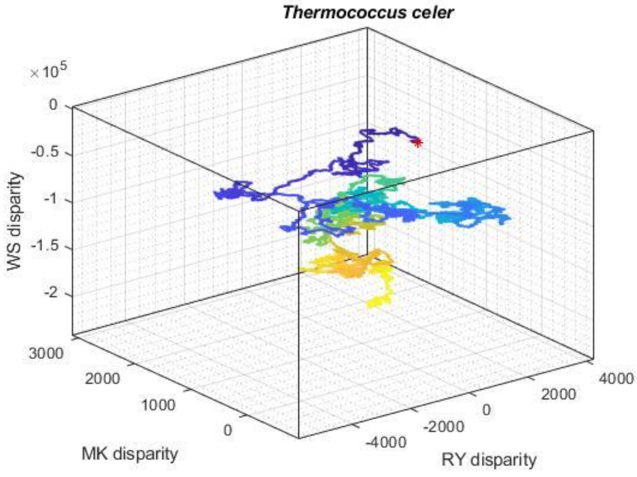


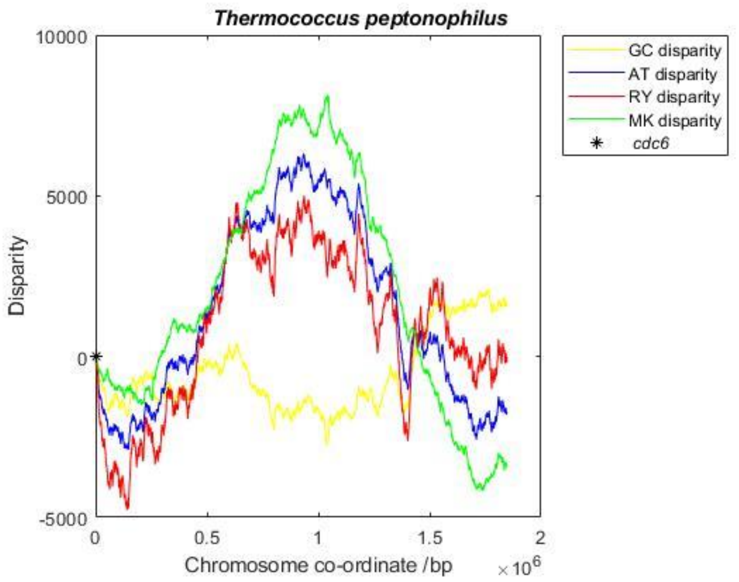


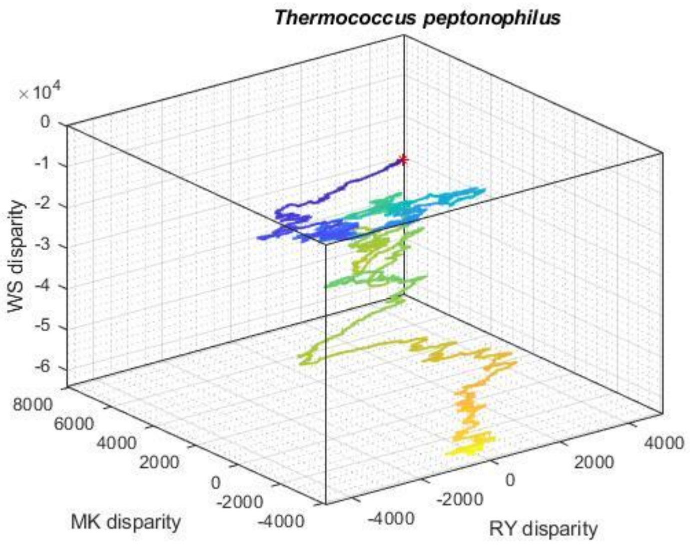


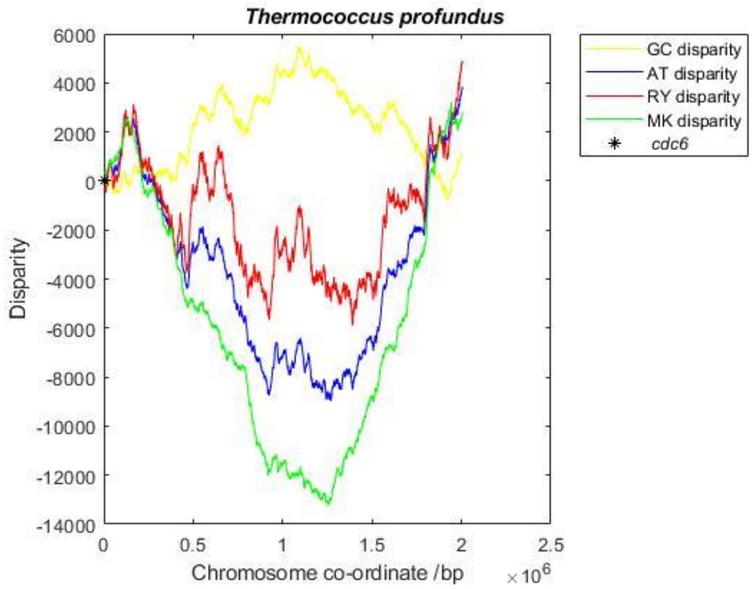


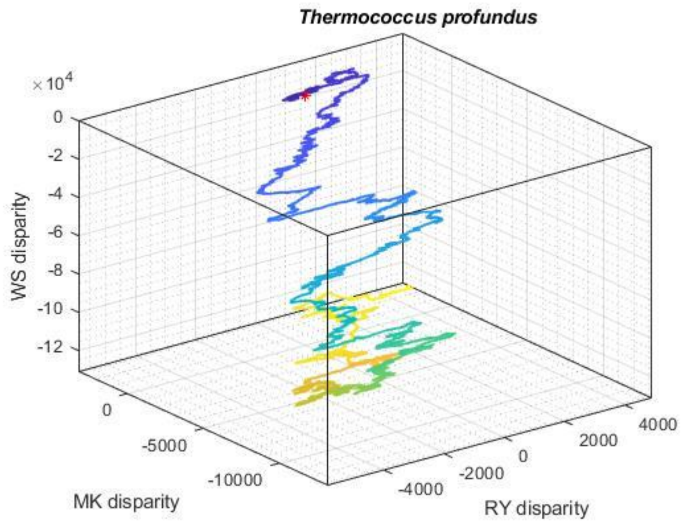


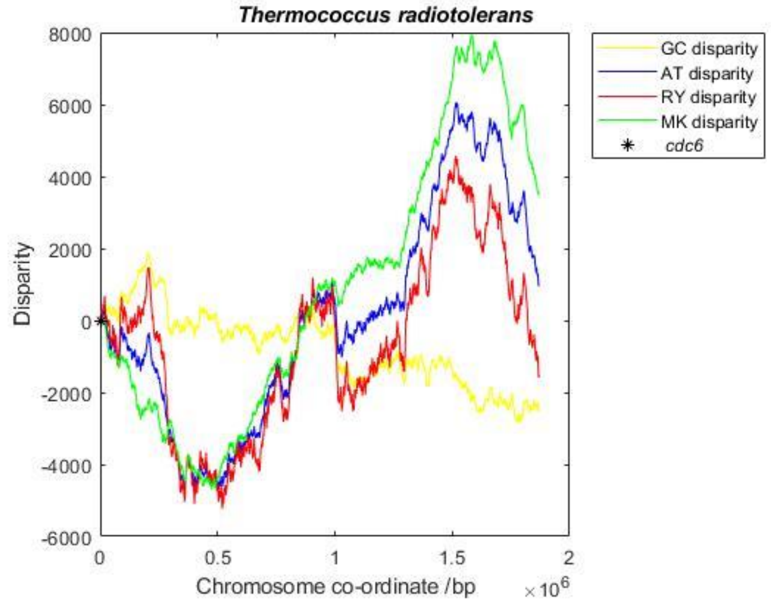


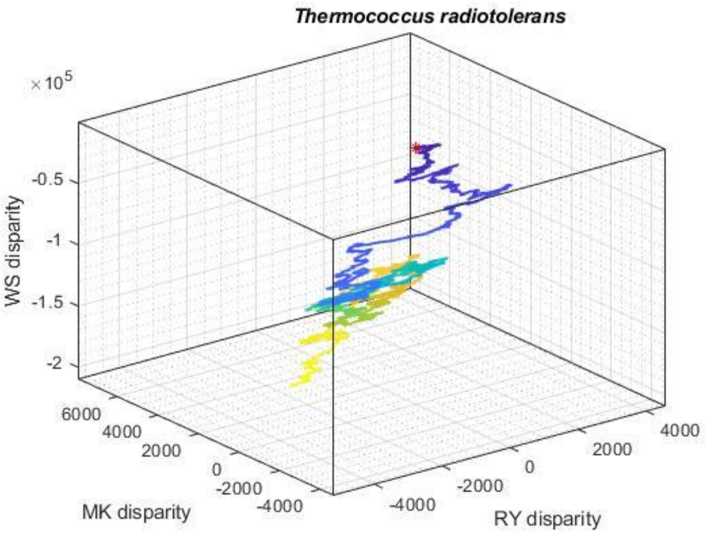

Thermococcus peptonophilus
10000
5000
0
-5000
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

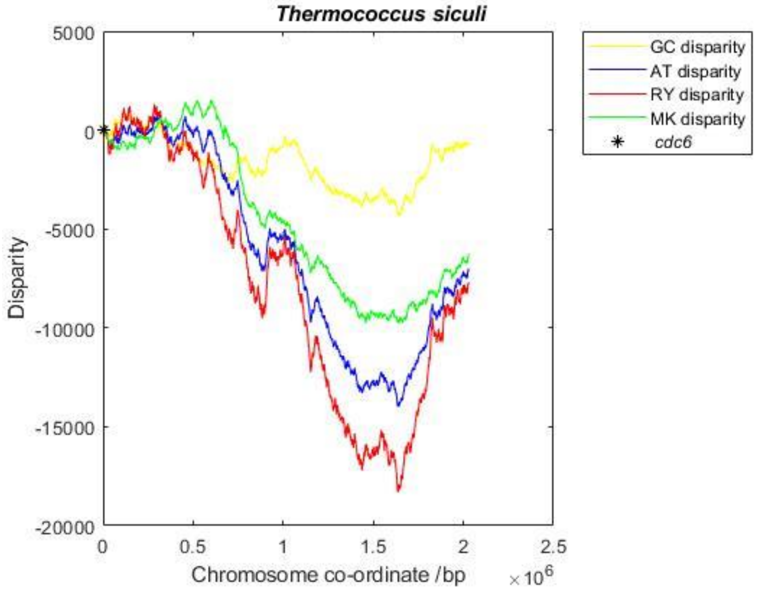

Thermococcus peptonophilus
×10^4
0
-1
-2
-3
-4
-5
-6
WS disparity
8000
6000
4000
2000
0
-2000
-4000
MK disparity
-4000
-2000
0
2000
4000
RY disparity


Thermococcus profundus
6000
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
Disparity
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

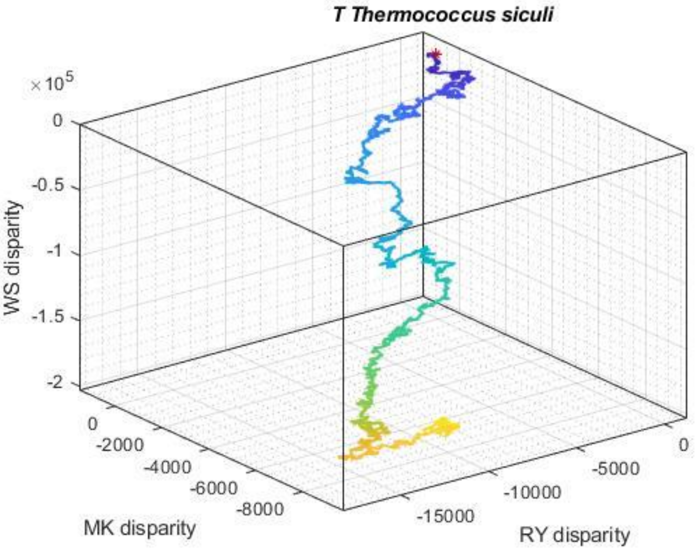

Thermococcus profundus
×10^4
0
-2
-4
-6
-8
-10
-12
WS disparity
0
-5000
-10000
MK disparity
-4000
-2000
0
2000
4000
RY disparity


Thermococcus radiotolerans
8000
6000
4000
2000
0
-2000
-4000
-6000
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

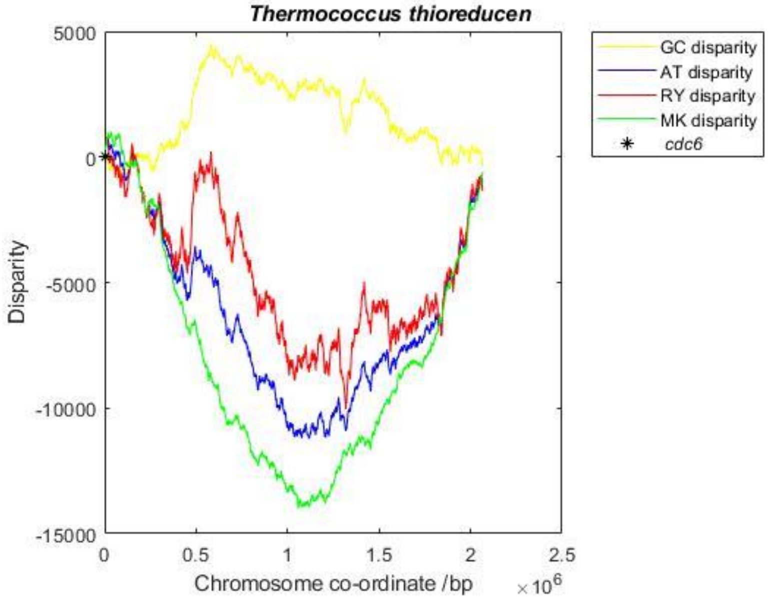

Thermococcus radiotolerans
×10^5
-0.5
-1
-1.5
-2
WS disparity
6000
4000
2000
0
-2000
-4000
MK disparity
-4000
-2000
0
2000
4000
RY disparity



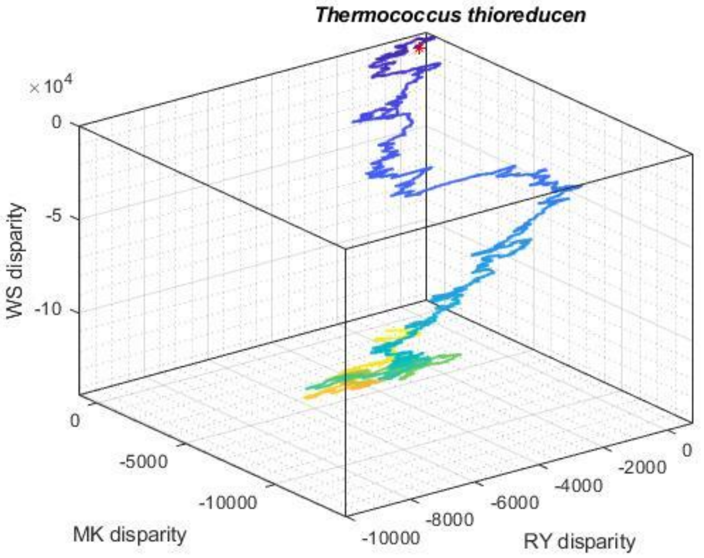
Thermococcus siculi



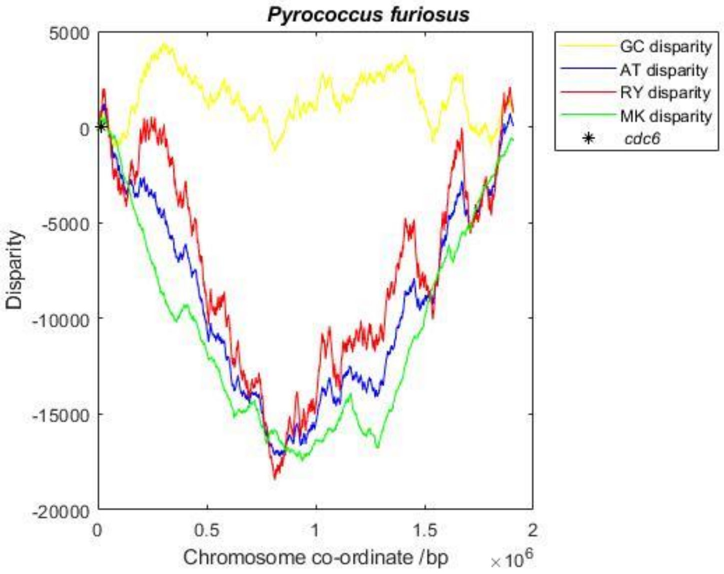
T Thermococcus siculi



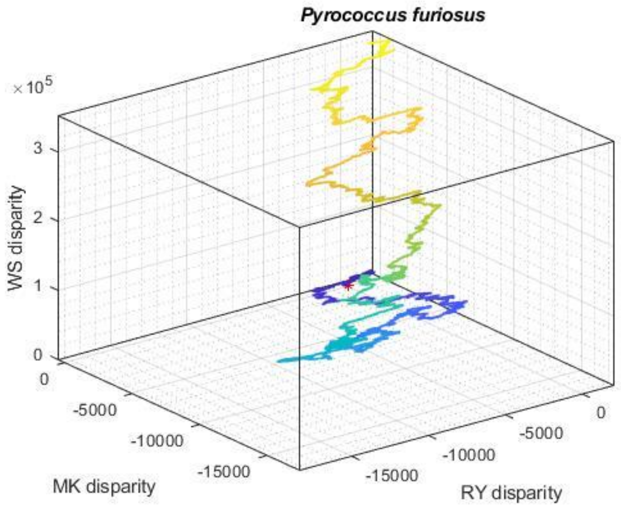
Thermococcus thioreducen



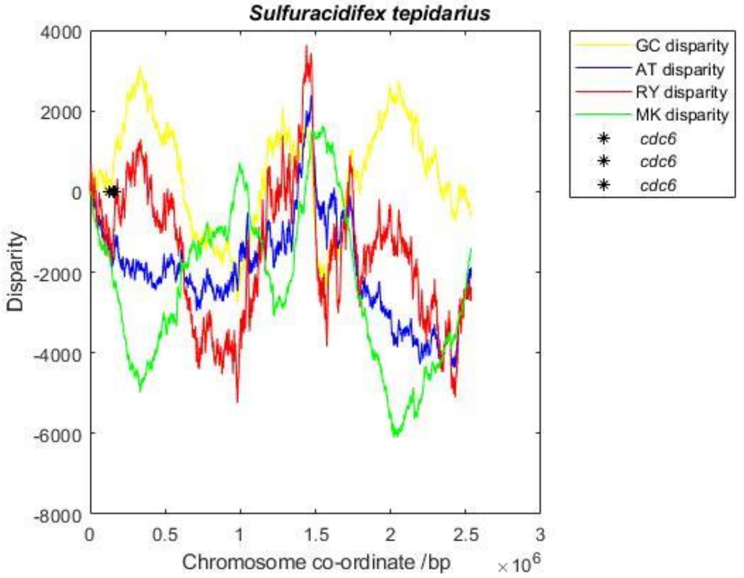
Thermococcus thioreducen

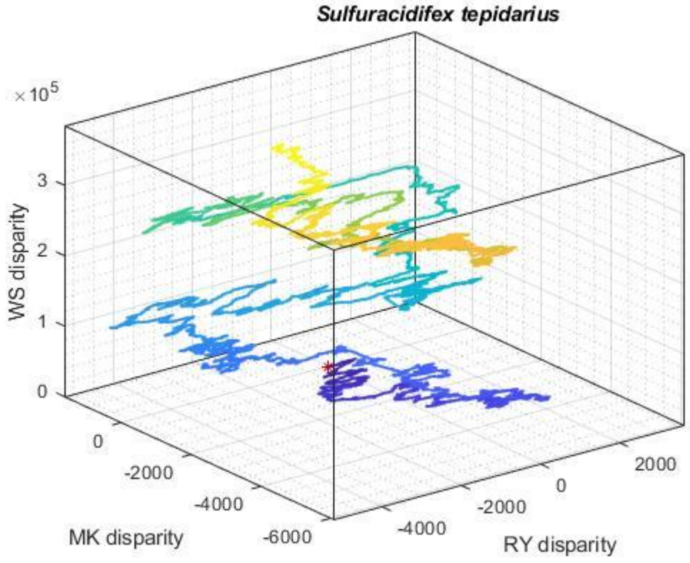


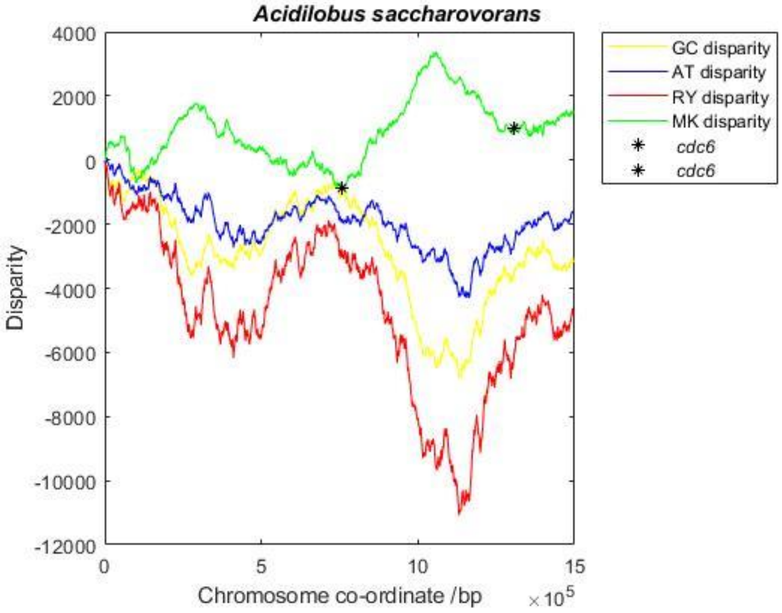
Pyrococcus furiosus

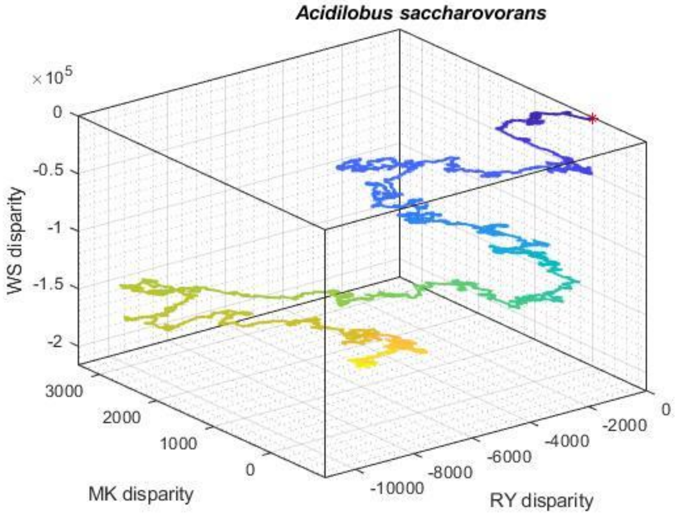


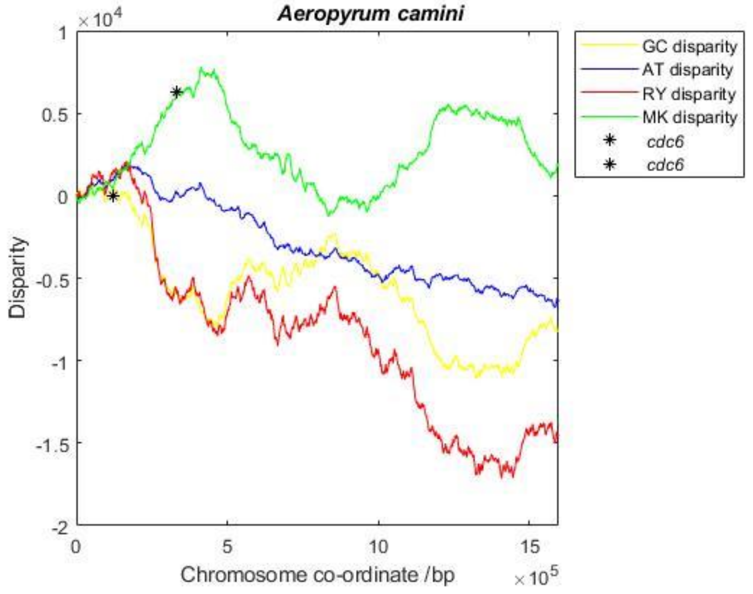
Pyrococcus furiosus

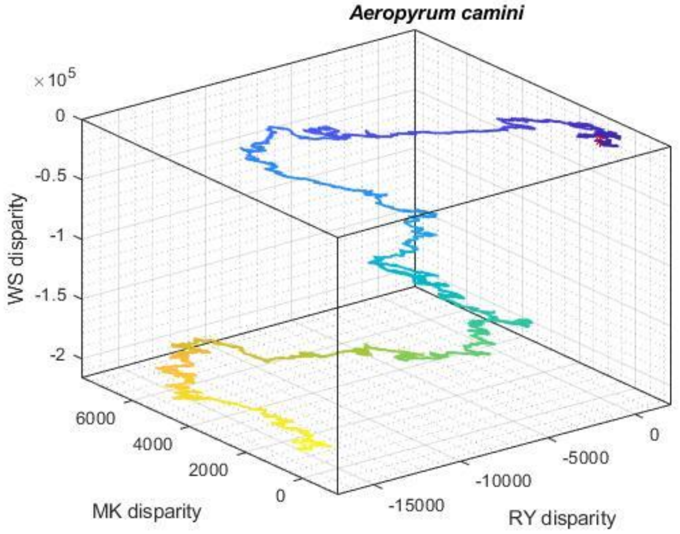

Sulfuracidifex tepidarius
Disparity
Chromosome co-ordinate /bp
×10^6
4000
2000
0
-2000
-4000
-6000
-8000
0
0.5
1
1.5
2
2.5
3
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6
cdc6

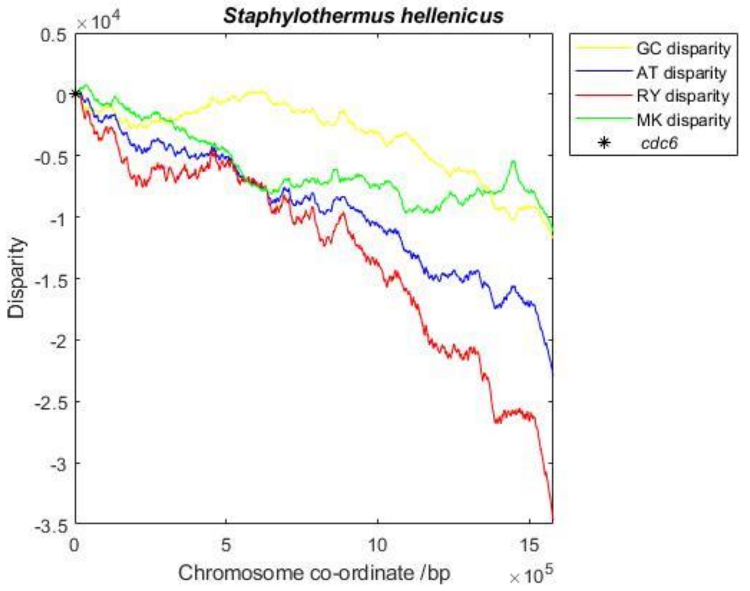

Sulfuracidifex tepidarius
×10^5
WS disparity
3
2
1
0
0
-2000
-4000
-6000
MK disparity
2000
0
-2000
-4000
RY disparity

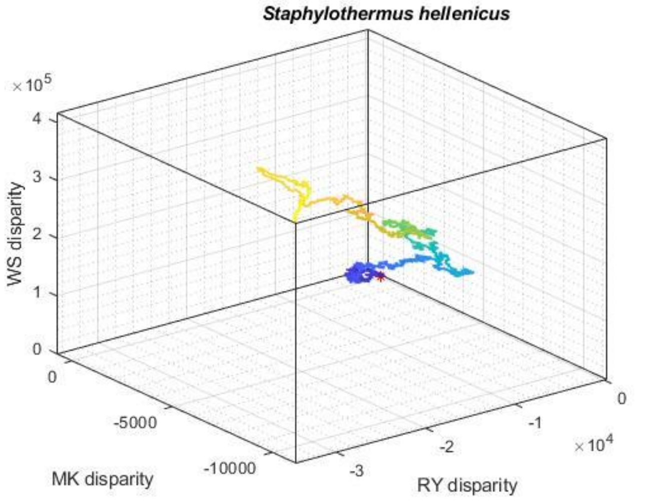

Acidilobus saccharovorans
Disparity
Chromosome co-ordinate /bp
×10^5
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
0
5
10
15
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

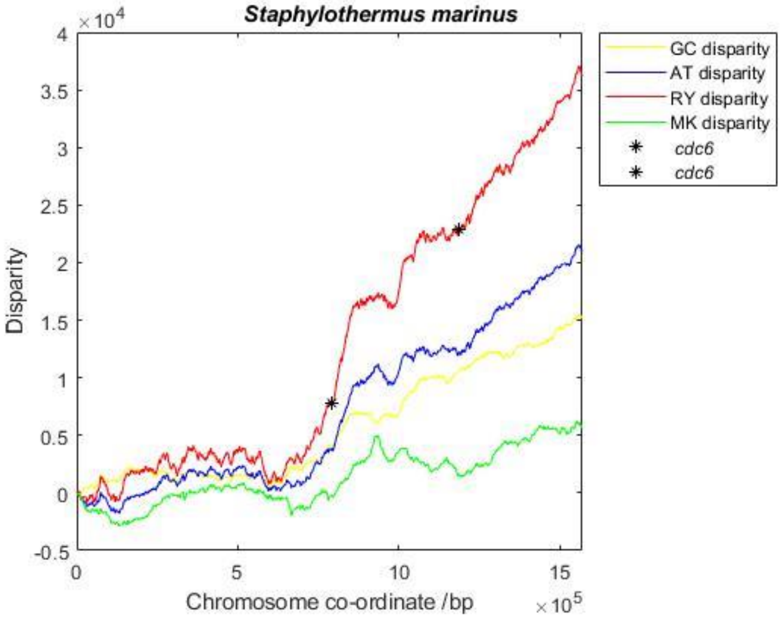

Acidilobus saccharovorans
×10^5
WS disparity
0
-0.5
-1
-1.5
-2
3000
2000
1000
0
MK disparity
0
-2000
-4000
-6000
-8000
-10000
RY disparity

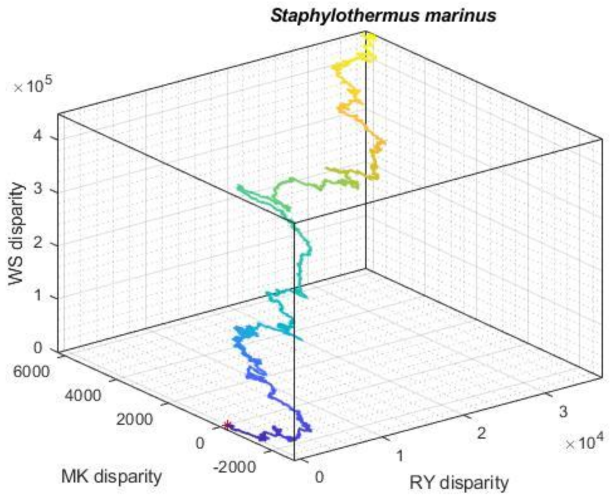

Aeropyrum camini
×10^4
Disparity
Chromosome co-ordinate /bp
×10^5
1
0.5
0
-0.5
-1
-1.5
-2
0
5
10
15
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

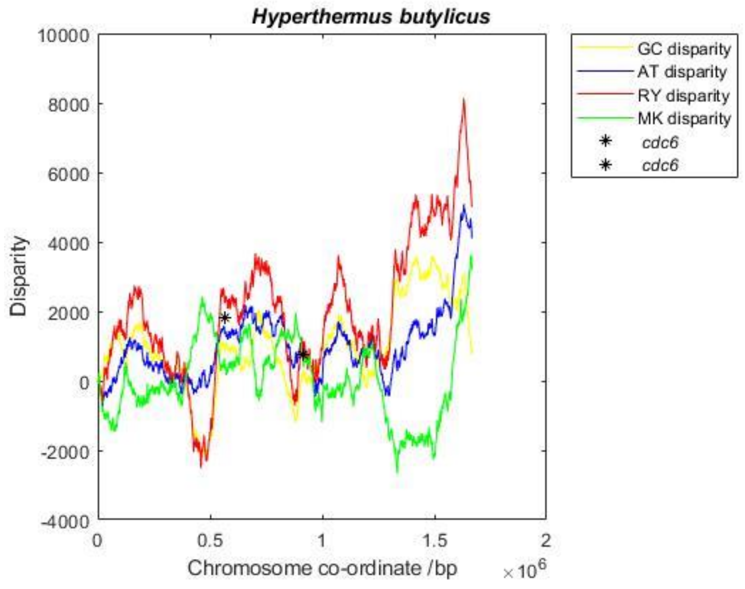

Aeropyrum camini
×10^5
WS disparity
0
-0.5
-1
-1.5
-2
6000
4000
2000
0
MK disparity
0
-5000
-10000
-15000
RY disparity

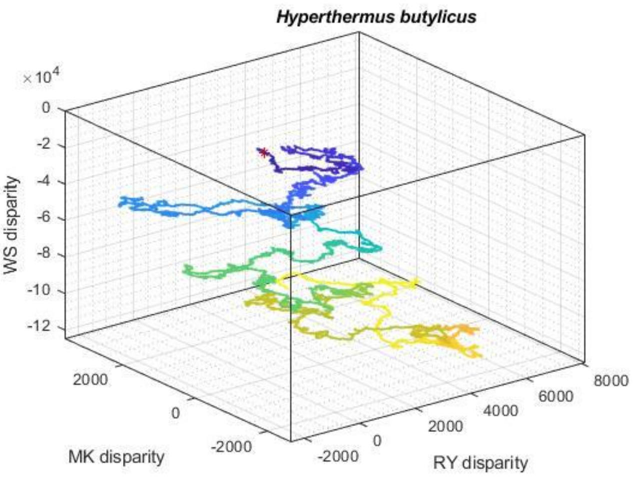

Staphylothermus hellenicus
×10^4
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
Disparity
0
5
10
15
Chromosome co-ordinate /bp
×10^5
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

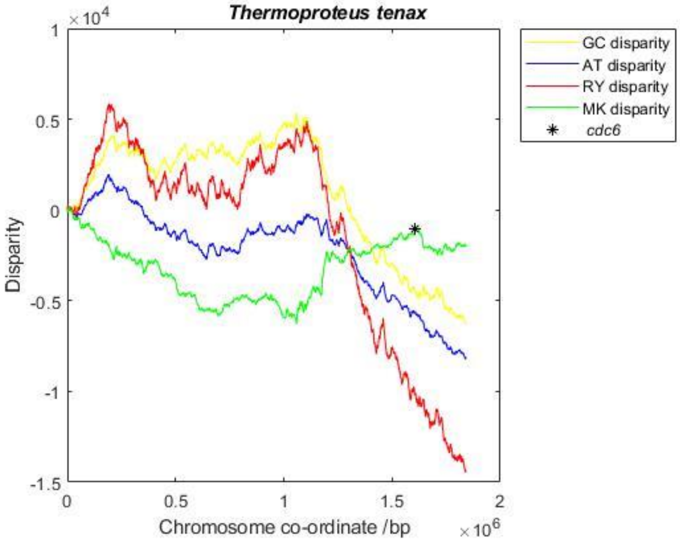

Staphylothermus hellenicus
×10^5
4
3
2
1
0
WS disparity
0
-5000
-10000
MK disparity
0
-1
-2
-3
×10^4
RY disparity

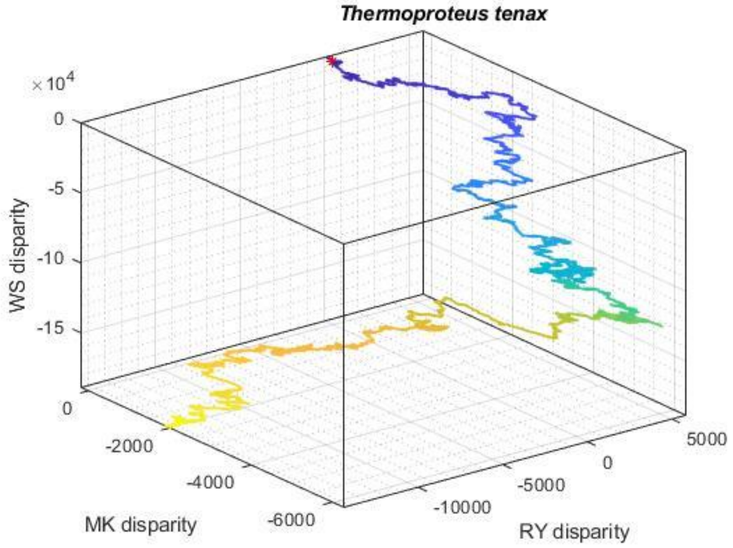

Staphylothermus marinus
×10^4
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
Disparity
0
5
10
15
Chromosome co-ordinate /bp
×10^5
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

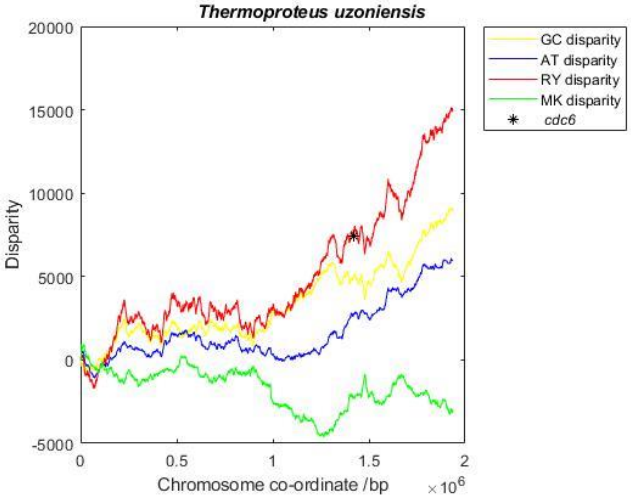

Staphylothermus marinus
×10^5
4
3
2
1
0
WS disparity
6000
4000
2000
0
-2000
MK disparity
0
1
2
3
×10^4
RY disparity

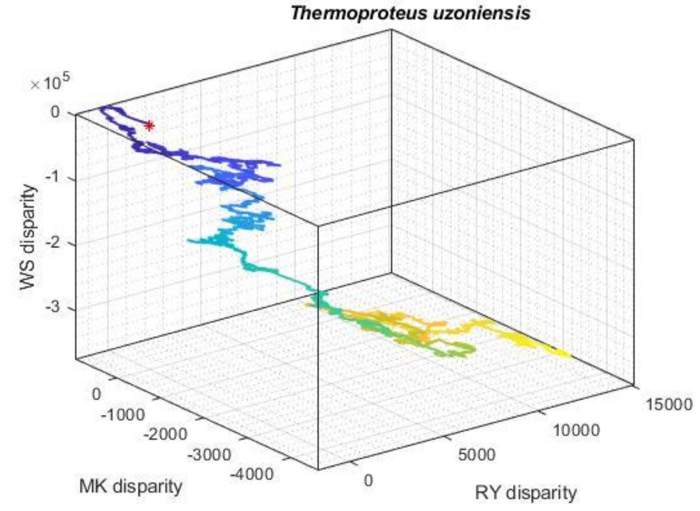

Hyperthermus butylicus
10000
8000
6000
4000
2000
0
-2000
-4000
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

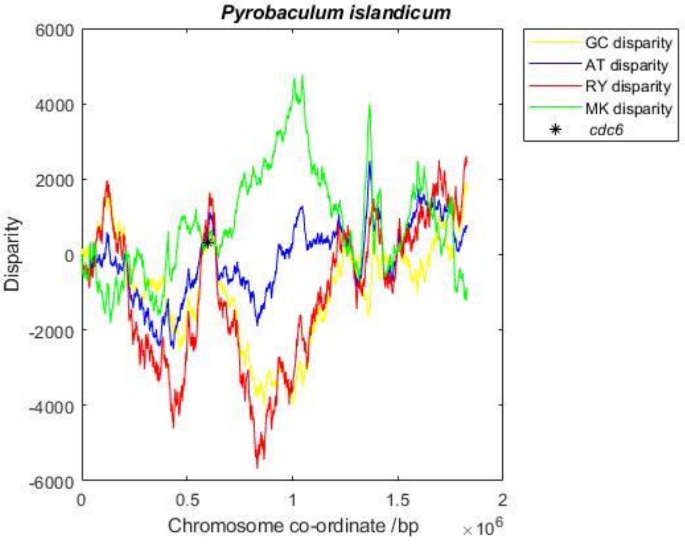

Hyperthermus butylicus
×10^4
0
-2
-4
-6
-8
-10
-12
WS disparity
2000
0
-2000
MK disparity
-2000
0
2000
4000
6000
8000
RY disparity

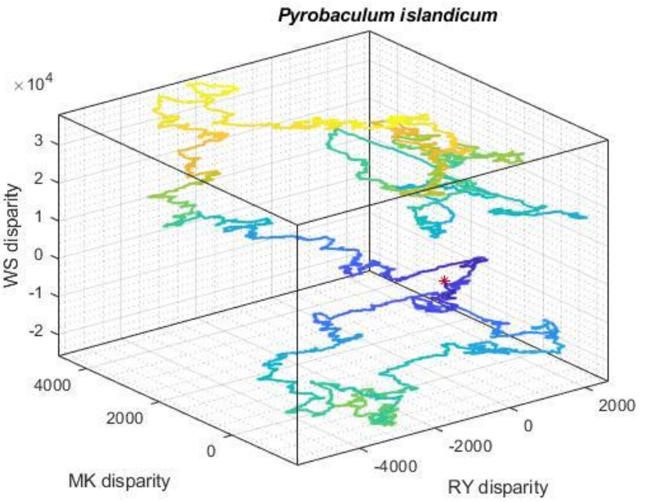

Thermoproteus tenax
×10^4
1
0.5
0
-0.5
-1
-1.5
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

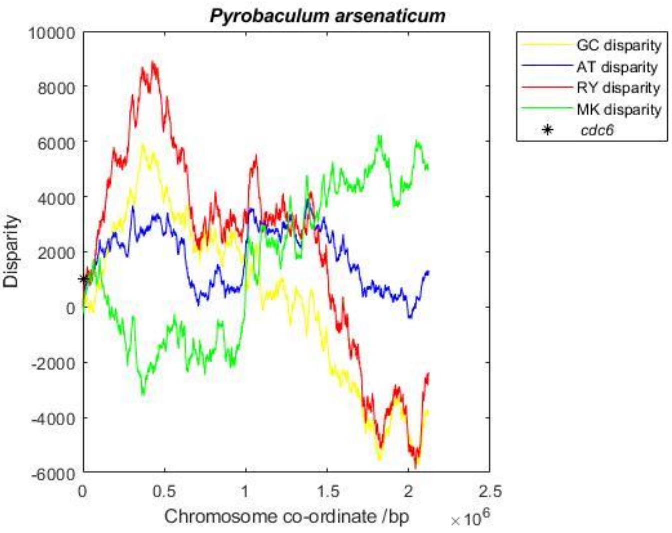

Thermoproteus tenax
×10^4
0
-5
-10
-15
WS disparity
0
-2000
-4000
-6000
MK disparity
5000
0
-5000
-10000
RY disparity

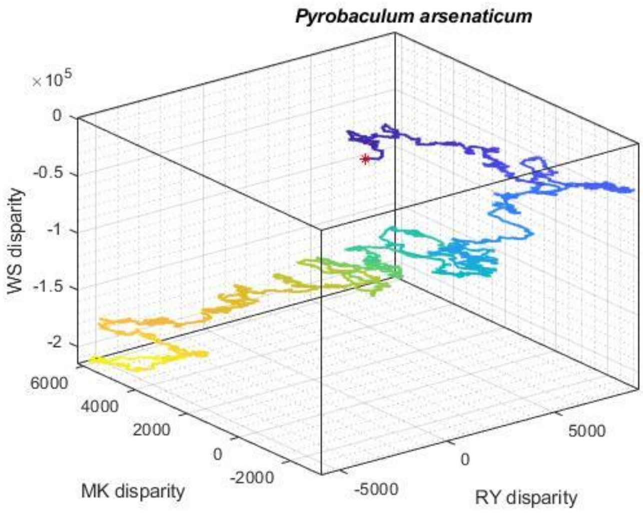

Thermoproteus uzoniensis
20000
15000
10000
5000
0
-5000
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

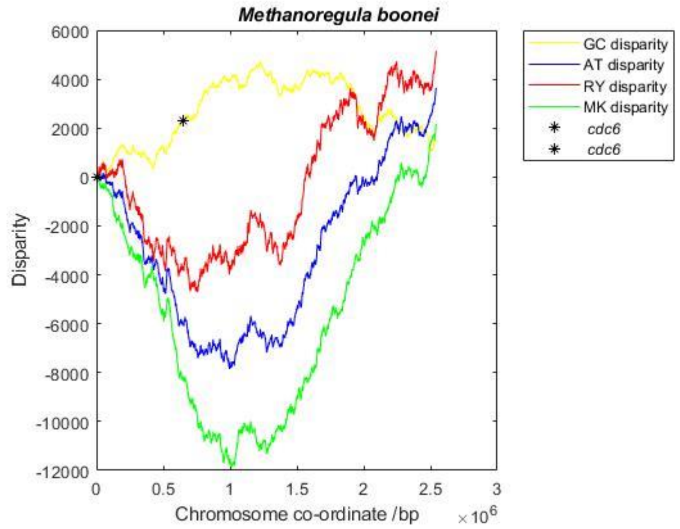

Thermoproteus uzoniensis
×10^5
0
-1
-2
-3
WS disparity
0
-1000
-2000
-3000
-4000
MK disparity
15000
10000
5000
0
RY disparity

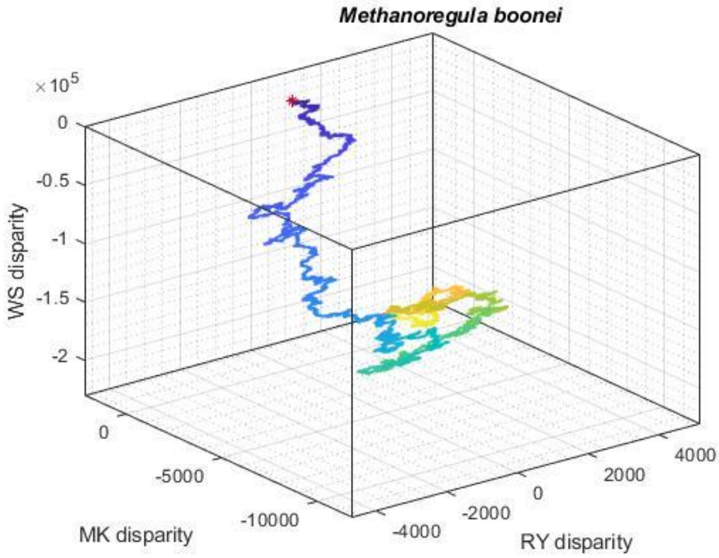

Pyrobaculum islandicum
6000
4000
2000
0
-2000
-4000
-6000
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

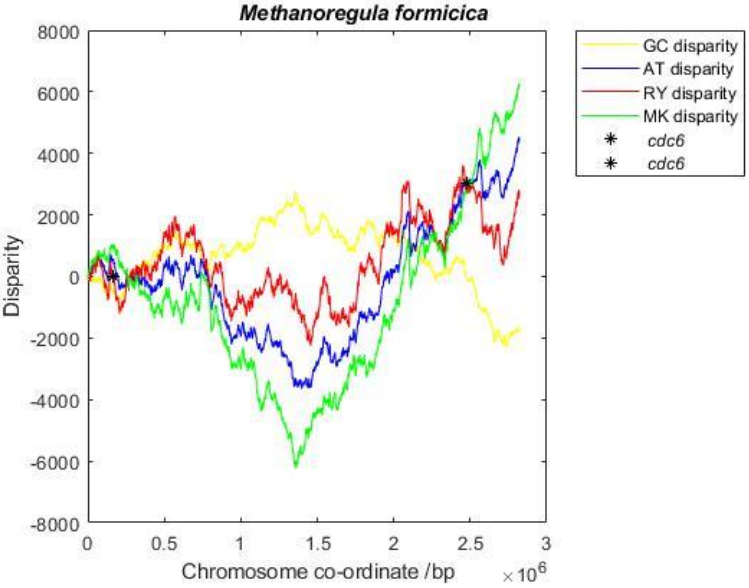

Pyrobaculum islandicum
×10^4
3
2
1
0
-1
-2
WS disparity
4000
2000
0
MK disparity
2000
0
-2000
-4000
RY disparity

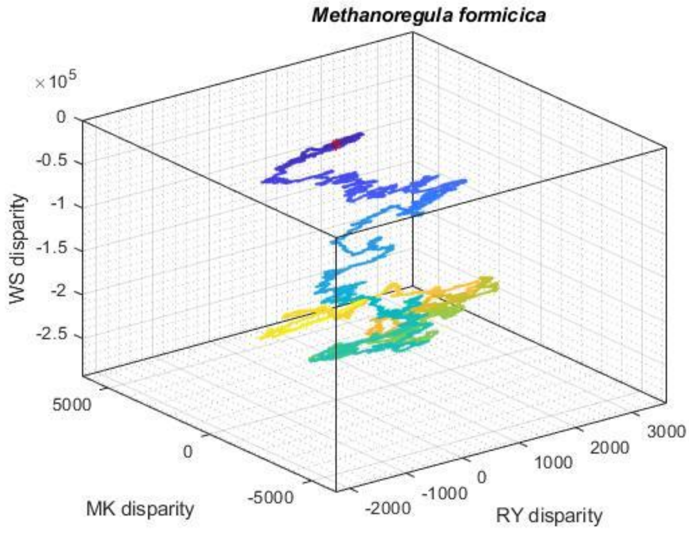

Pyrobaculum arsenaticum
Disparity
10000
8000
6000
4000
2000
0
-2000
-4000
-6000
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
× 10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

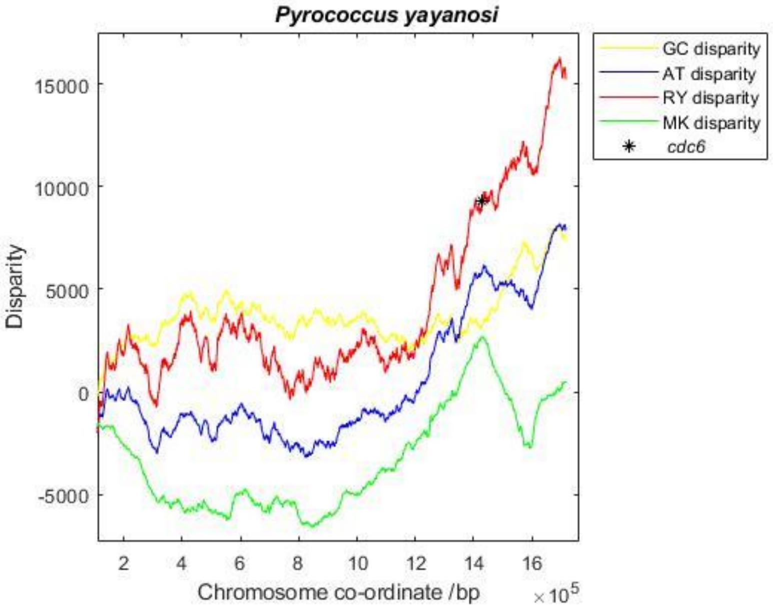

Pyrobaculum arsenaticum
× 10^5
0
-0.5
-1
-1.5
-2
WS disparity
6000
4000
2000
0
-2000
MK disparity
-5000
0
5000
RY disparity

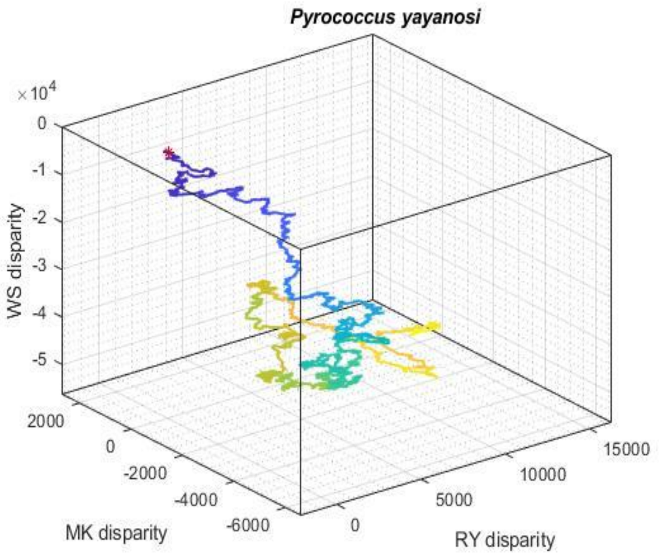

Methanoregula boonei
Disparity
6000
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
× 10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

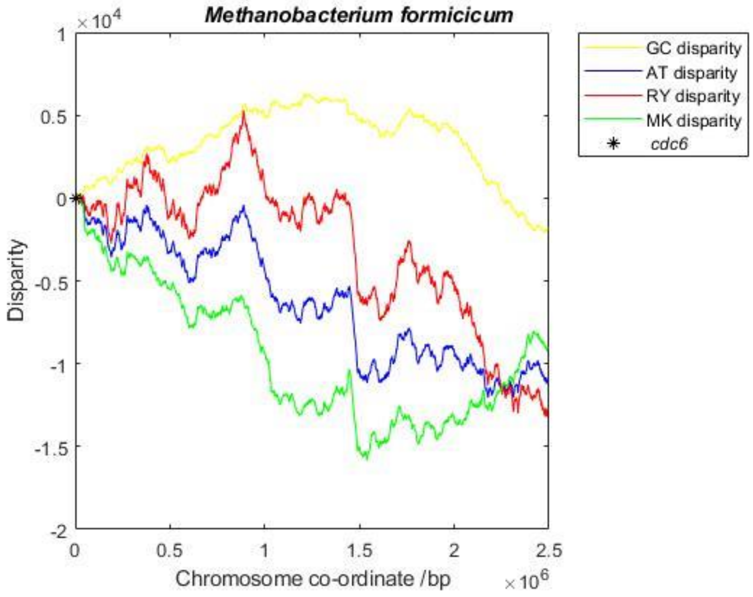

Methanoregula boonei
× 10^5
0
-0.5
-1
-1.5
-2
WS disparity
0
-5000
-10000
MK disparity
-4000
-2000
0
2000
4000
RY disparity

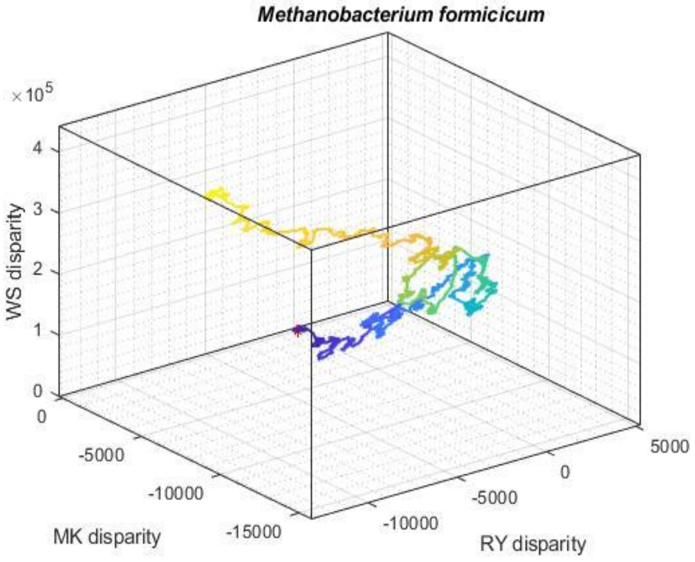

Methanoregula formicica
Disparity
8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
× 10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

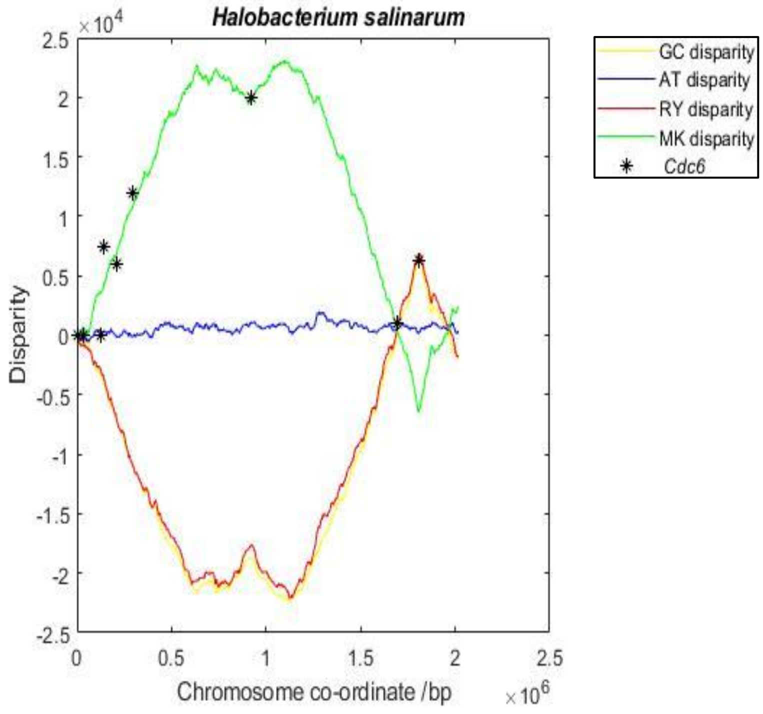

Methanoregula formicica
× 10^5
0
-0.5
-1
-1.5
-2
-2.5
WS disparity
5000
0
-5000
MK disparity
-2000
-1000
0
1000
2000
3000
RY disparity

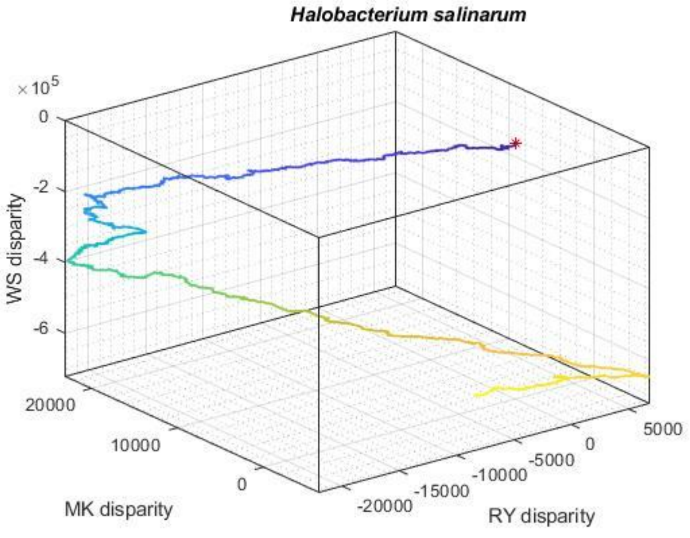

Pyrococcus yayanosi
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
Disparity
15000
10000
5000
0
-5000
2
4
6
8
10
12
14
16
Chromosome co-ordinate /bp
× 10^5

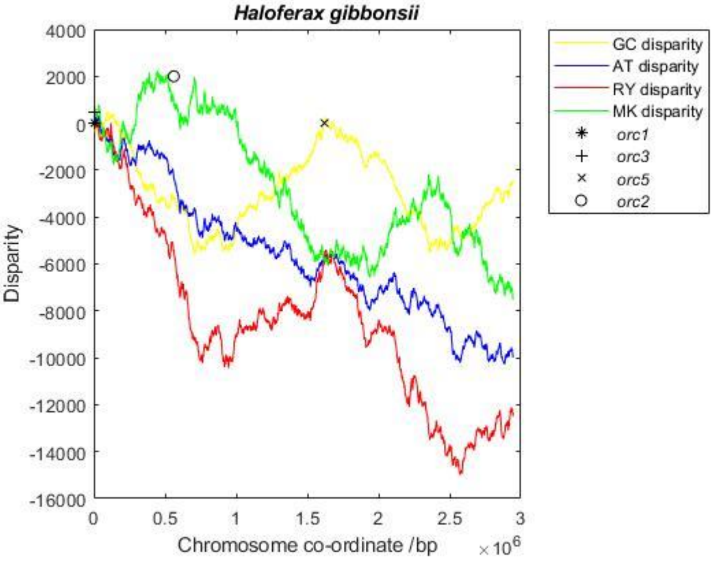

Pyrococcus yayanosi
× 10^4
0
-1
-2
-3
-4
-5
WS disparity
2000
0
-2000
-4000
-6000
MK disparity
0
5000
10000
15000
RY disparity

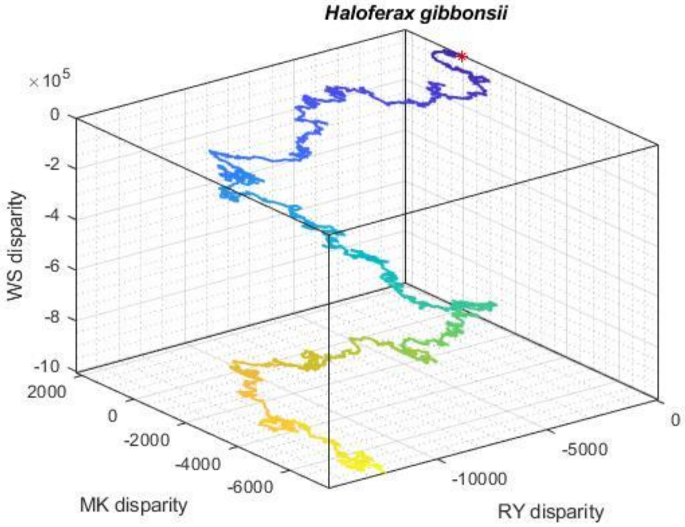

Methanobacterium formicicum
× 10^4
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
× 10^6

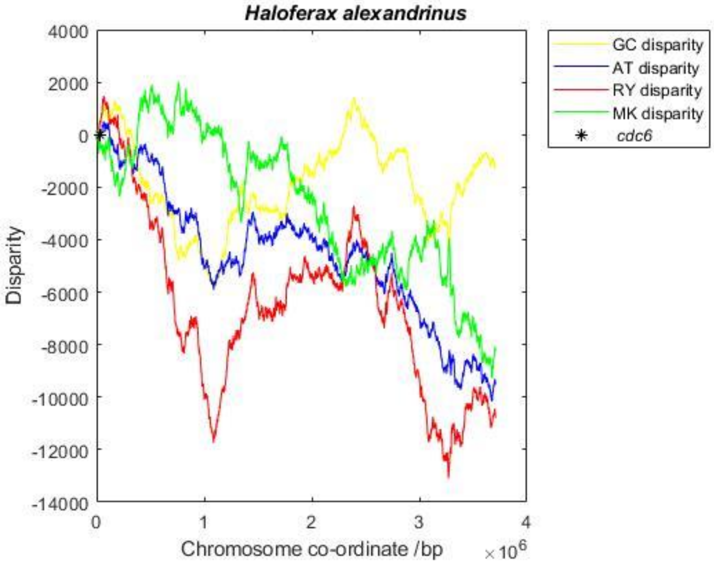

Methanobacterium formicicum
× 10^5
4
3
2
1
0
WS disparity
0
-5000
-10000
-15000
MK disparity
-10000
-5000
0
5000
RY disparity

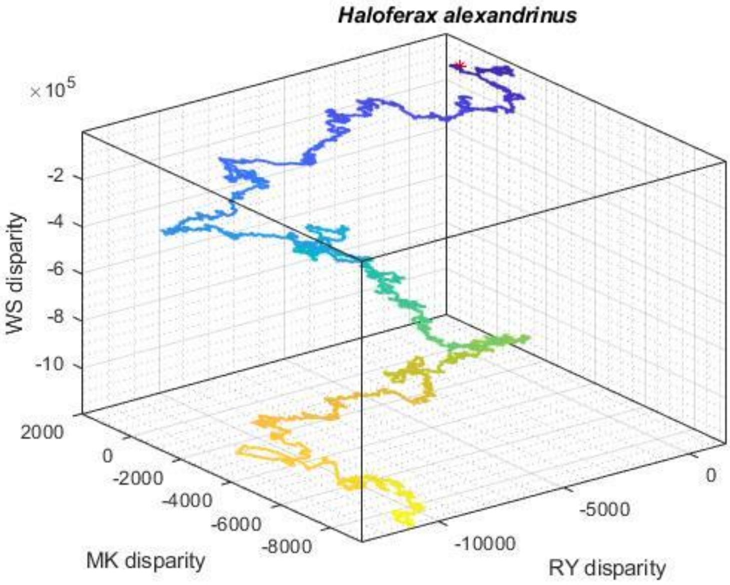

Halobacterium salinarum
× 10^4
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Disparity
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
× 10^6

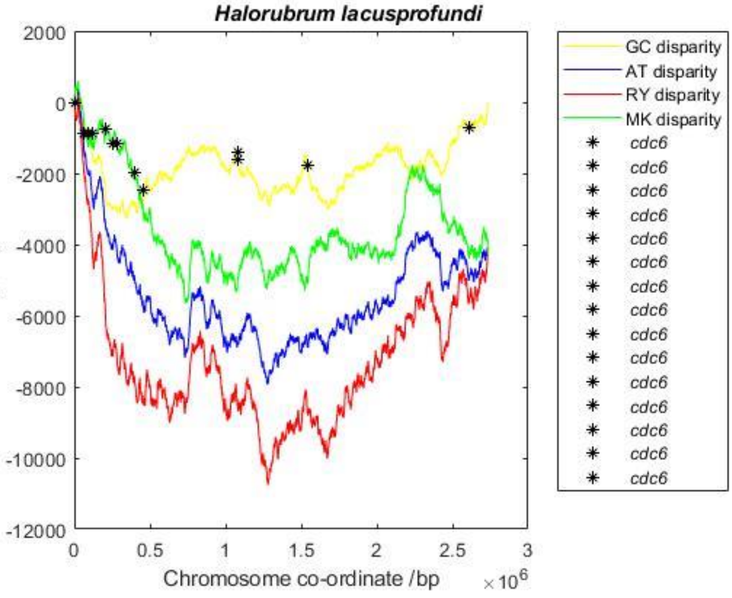

Halobacterium salinarum
× 10^5
0
-2
-4
-6
WS disparity
20000
10000
0
MK disparity
-20000
-15000
-10000
-5000
0
5000
RY disparity

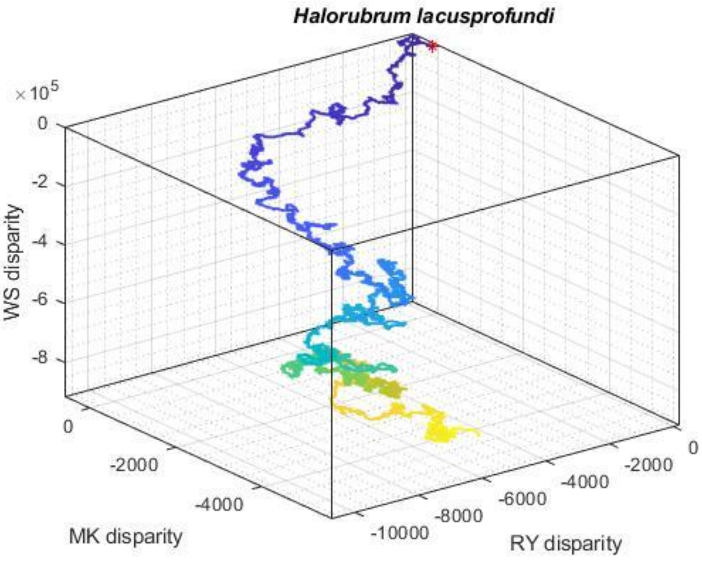

Haloferax gibbonsii
Disparity
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
-16000
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
orc1
orc3
orc5
orc2

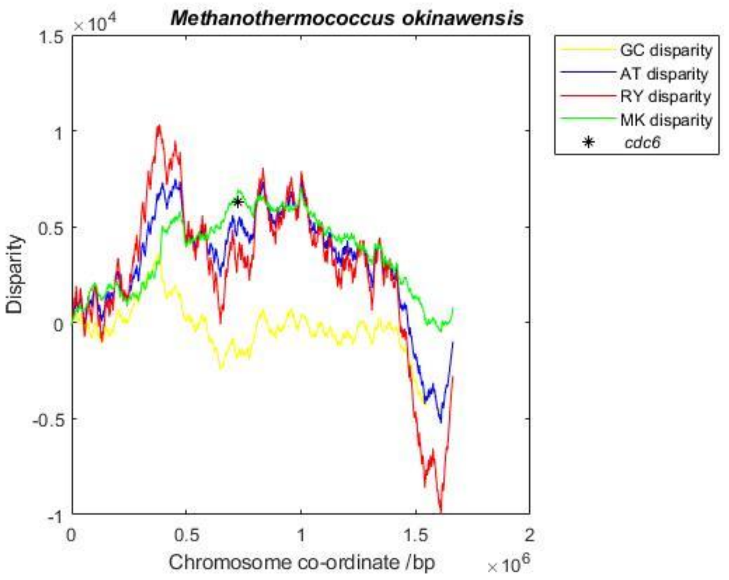

Haloferax gibbonsii
×10^5
0
-2
-4
-6
-8
-10
WS disparity
2000
0
-2000
-4000
-6000
MK disparity
0
-5000
-10000
RY disparity

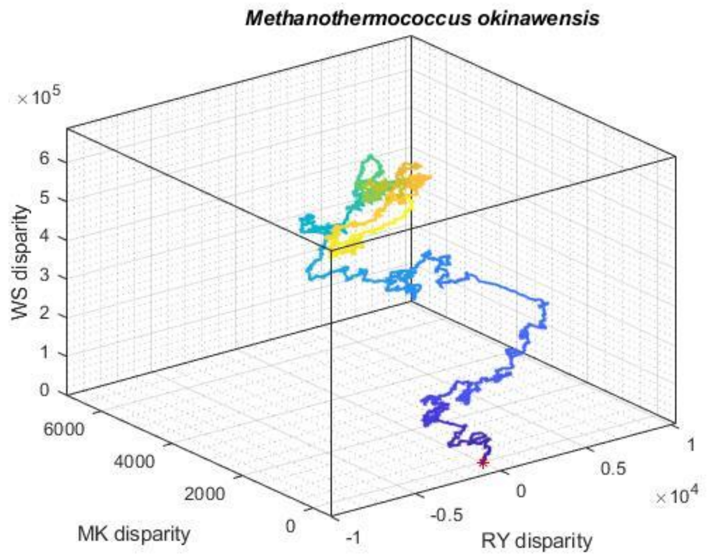

Haloferax alexandrinus
Disparity
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
0
1
2
3
4
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

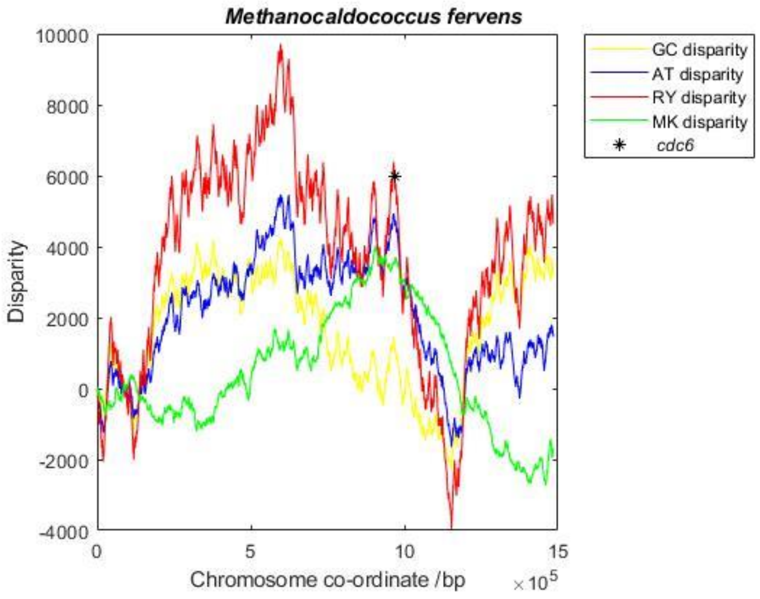

Haloferax alexandrinus
×10^5
-2
-4
-6
-8
-10
WS disparity
2000
0
-2000
-4000
-6000
-8000
MK disparity
0
-5000
-10000
RY disparity

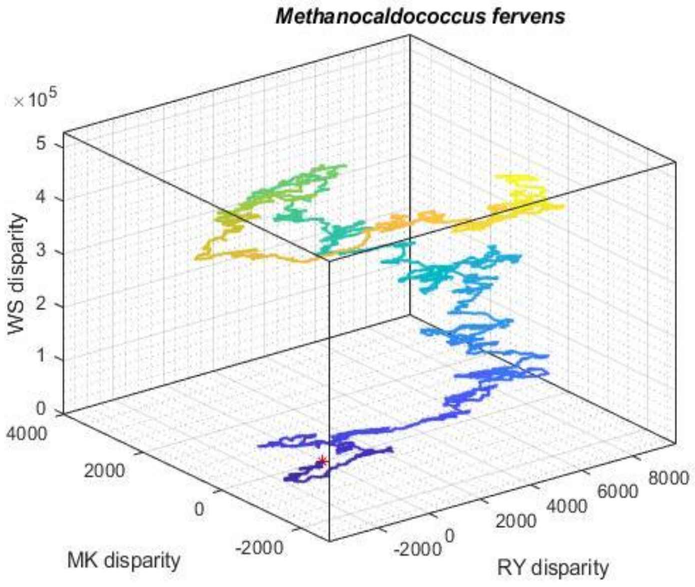

Halorubrum lacusprofundi
Disparity
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6
cdc6

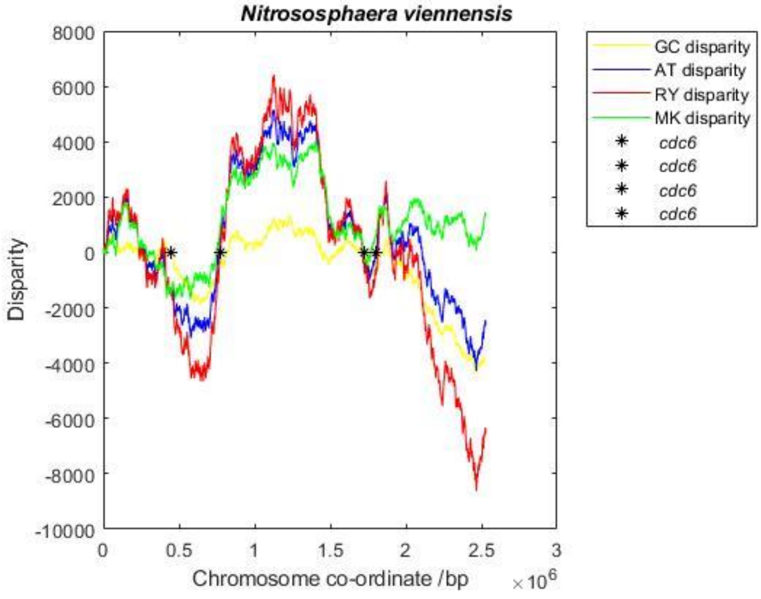

Halorubrum lacusprofundi
×10^5
0
-2
-4
-6
-8
WS disparity
0
-2000
-4000
MK disparity
0
-2000
-4000
-6000
-8000
-10000
RY disparity

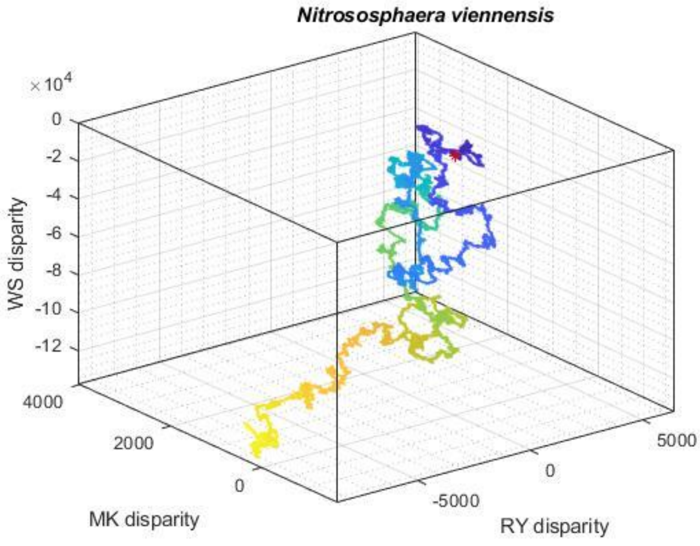

Methanothermococcus okinawensis
×10^4
1.5
1
0.5
0
-0.5
-1
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

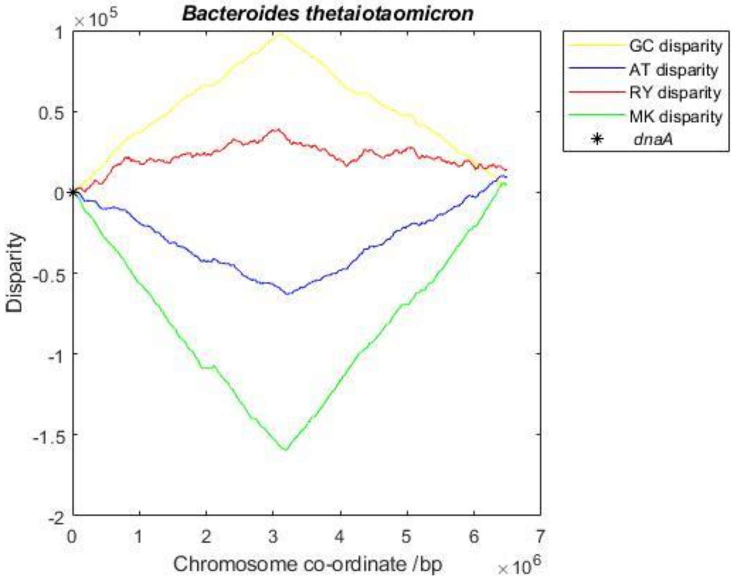

Methanothermococcus okinawensis
×10^5
6
5
4
3
2
1
0
WS disparity
6000
4000
2000
0
MK disparity
-1
-0.5
0
0.5
1
×10^4
RY disparity


Methanocaldococcus fervens
10000
8000
6000
4000
2000
0
-2000
-4000
Disparity
0
5
10
15
Chromosome co-ordinate /bp
×10^5
GC disparity
AT disparity
RY disparity
MK disparity
cdc6

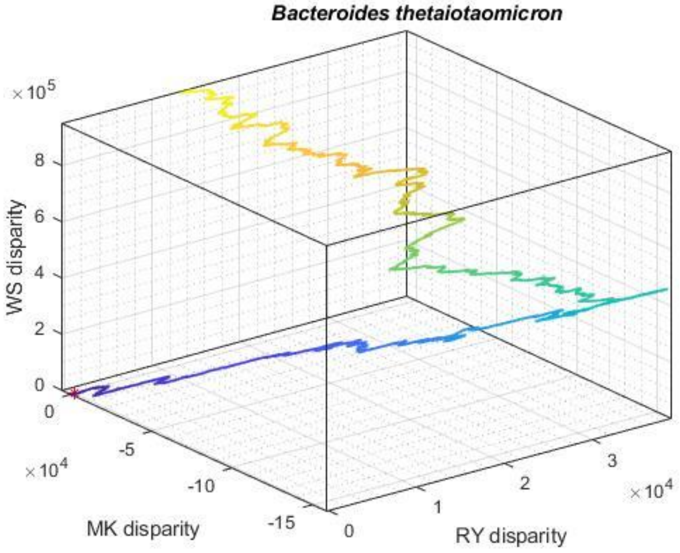

Methanocaldococcus fervens
×10^5
5
4
3
2
1
0
WS disparity
4000
2000
0
-2000
MK disparity
-2000
0
2000
4000
6000
8000
RY disparity


Nitrososphaera viennensis
8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
-10000
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6
cdc6
cdc6

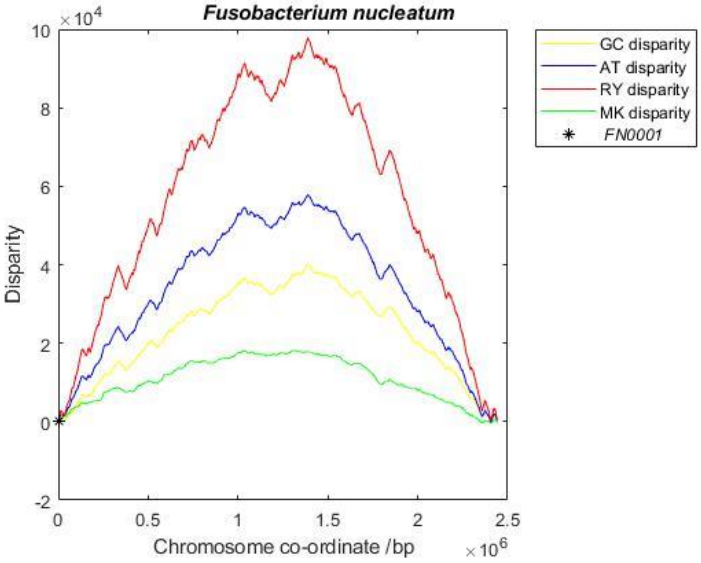

Nitrososphaera viennensis
×10^4
0
-2
-4
-6
-8
-10
-12
WS disparity
4000
2000
0
MK disparity
-5000
0
5000
RY disparity



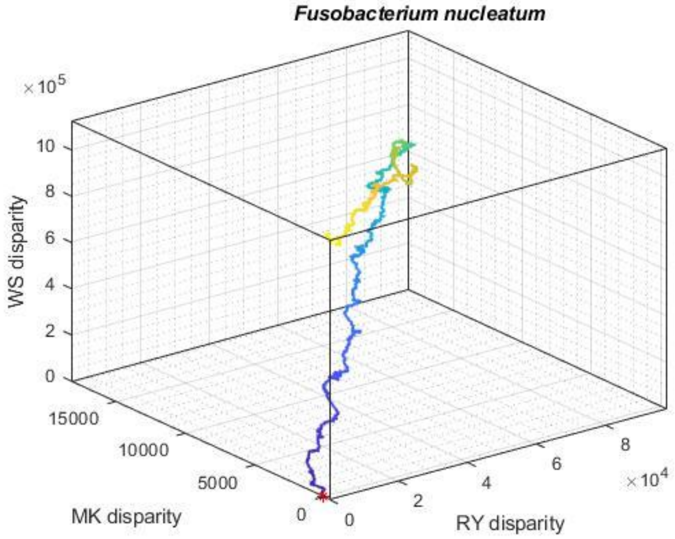
Bacteroides thetaiotaomicron
×10^5
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
0
1
2
3
4
5
6
7
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA



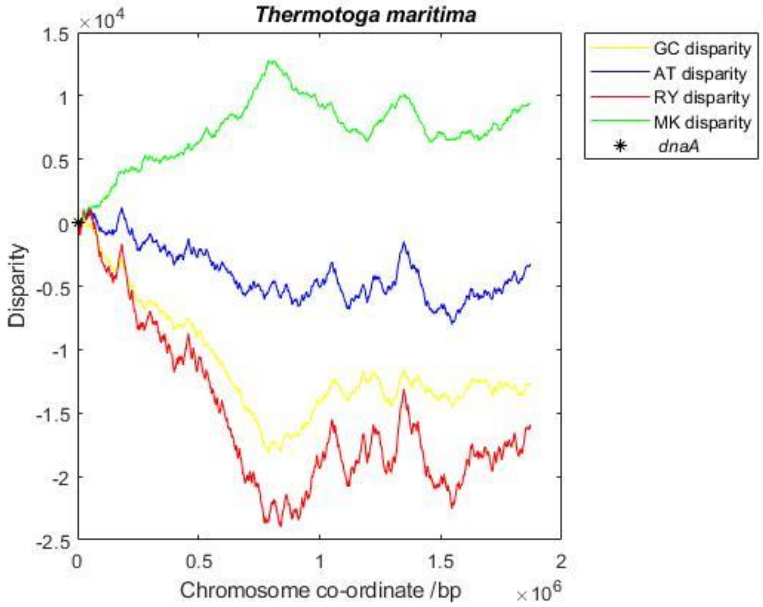
Bacteroides thetaiotaomicron
×10^5
8
6
4
2
0
WS disparity
0
-5
-10
-15
×10^4
MK disparity
0
1
2
3
×10^4
RY disparity



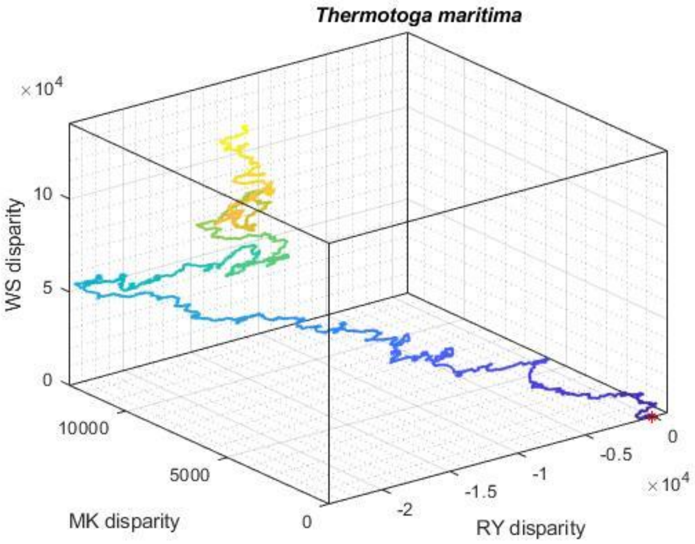
Fusobacterium nucleatum
×10^4
10
8
6
4
2
0
-2
Disparity
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
FN0001



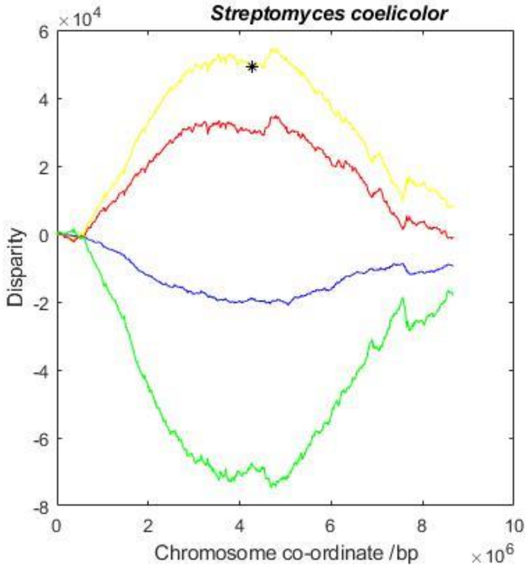
Fusobacterium nucleatum
×10^5
10
8
6
4
2
0
WS disparity
15000
10000
5000
0
MK disparity
0
2
4
6
8
×10^4
RY disparity



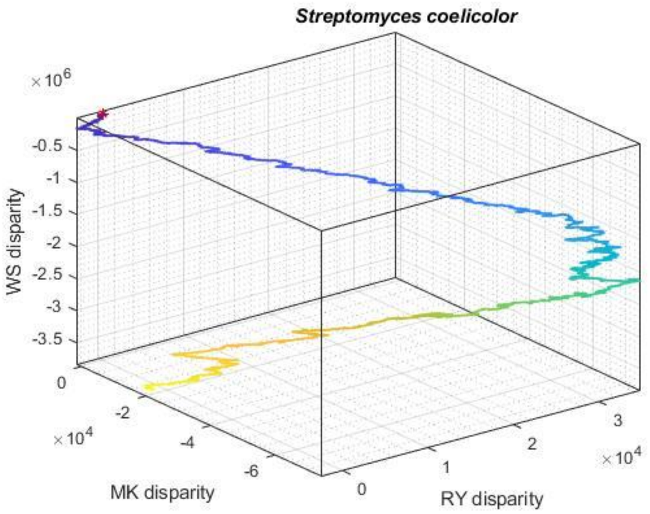
Thermotoga maritima
×10^4
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA



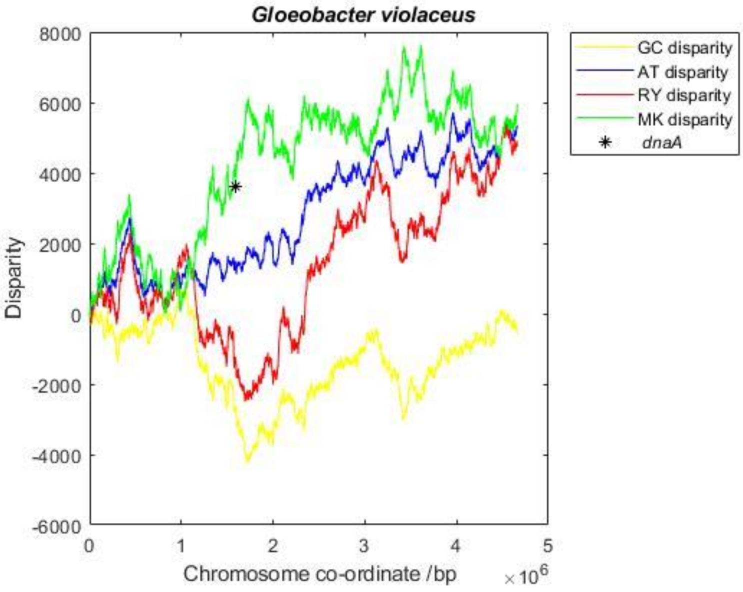
Thermotoga maritima
×10^4
10
5
0
WS disparity
10000
5000
0
MK disparity
0
-0.5
-1
-1.5
-2
×10^4
RY disparity



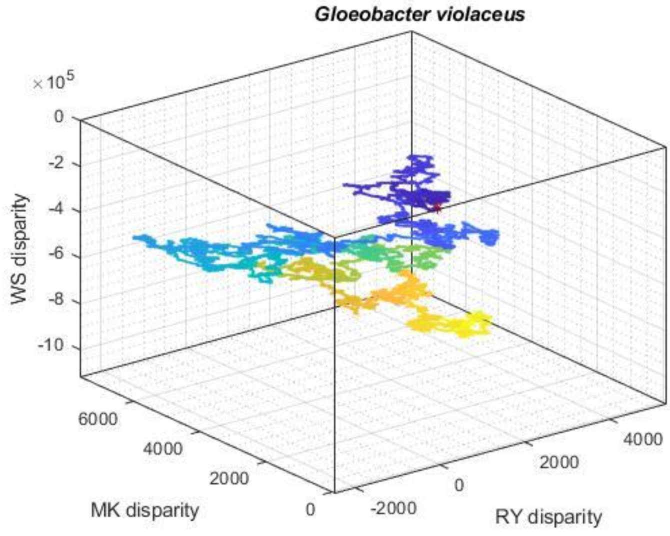
Streptomyces coelicolor



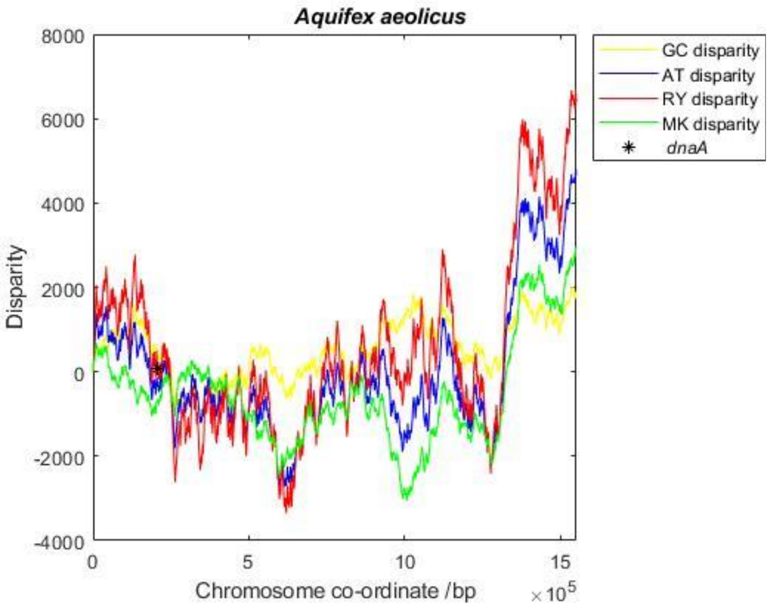
Streptomyces coelicolor



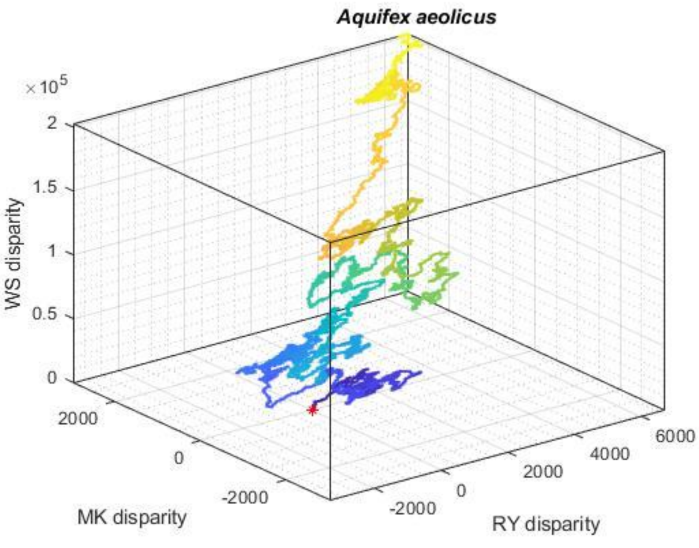
Gloeobacter violaceus



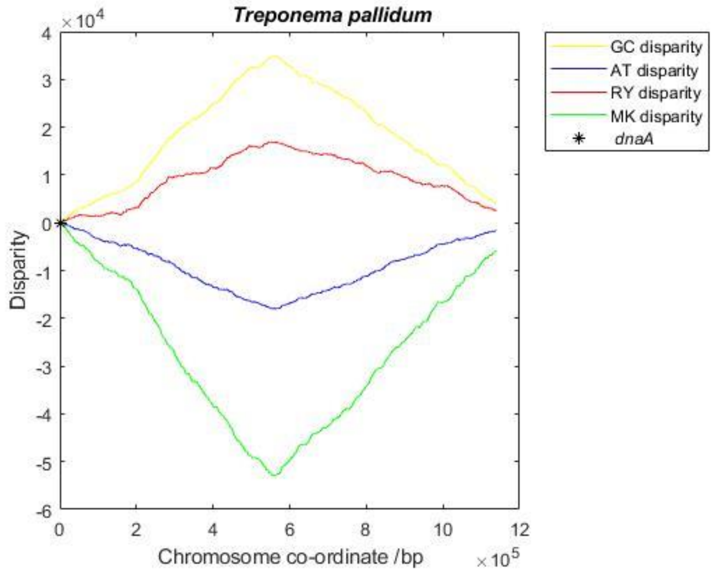
Gloeobacter violaceus

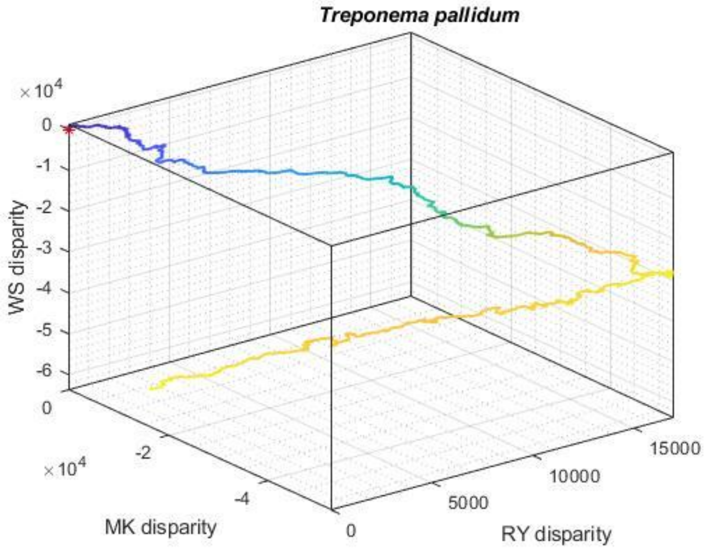


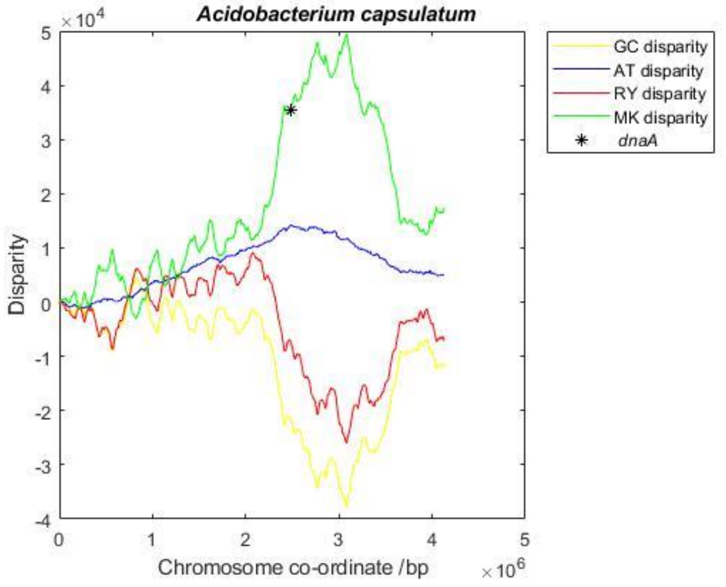
Aquifex aeolicus

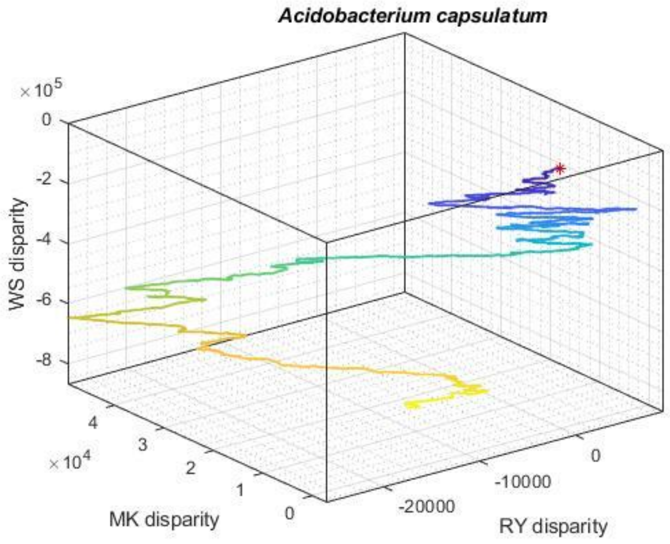


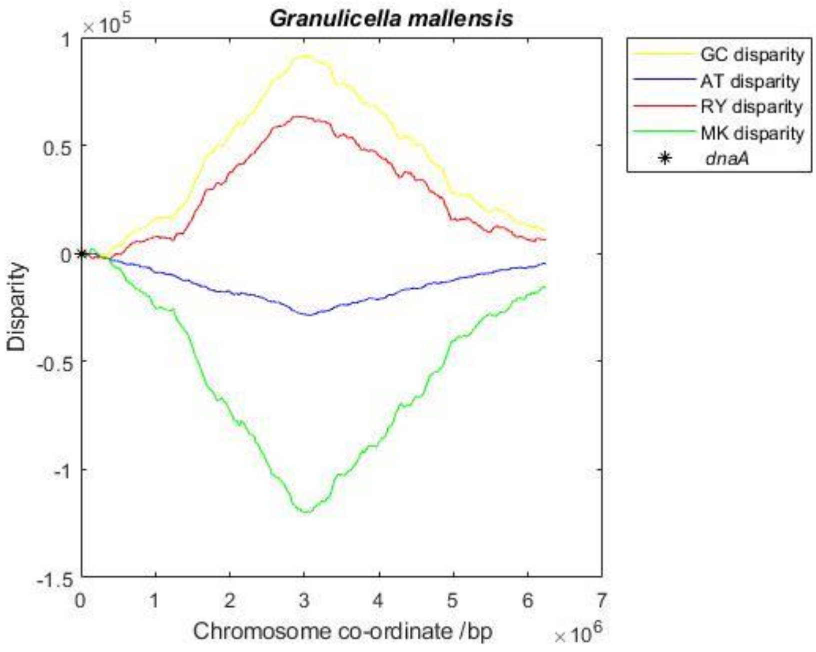
Aquifex aeolicus

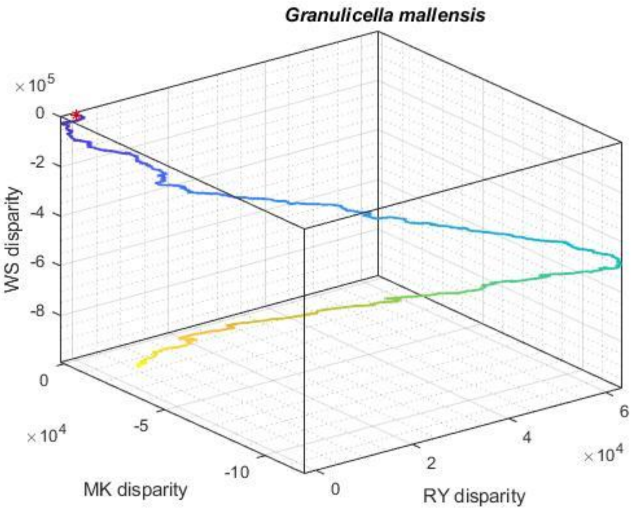

Treponema pallidum

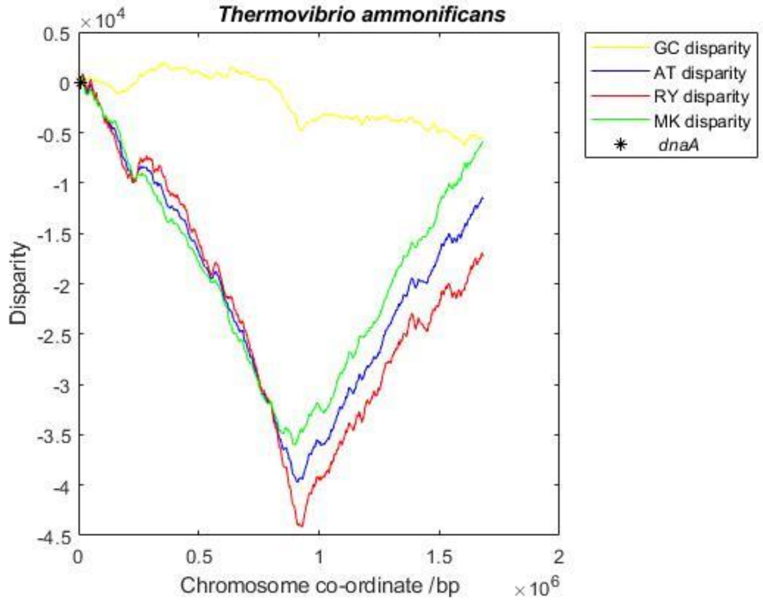

Treponema pallidum

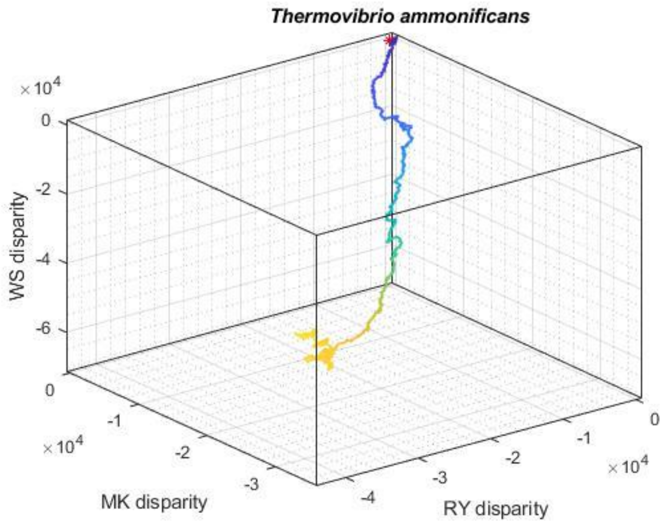

Acidobacterium capsulatum

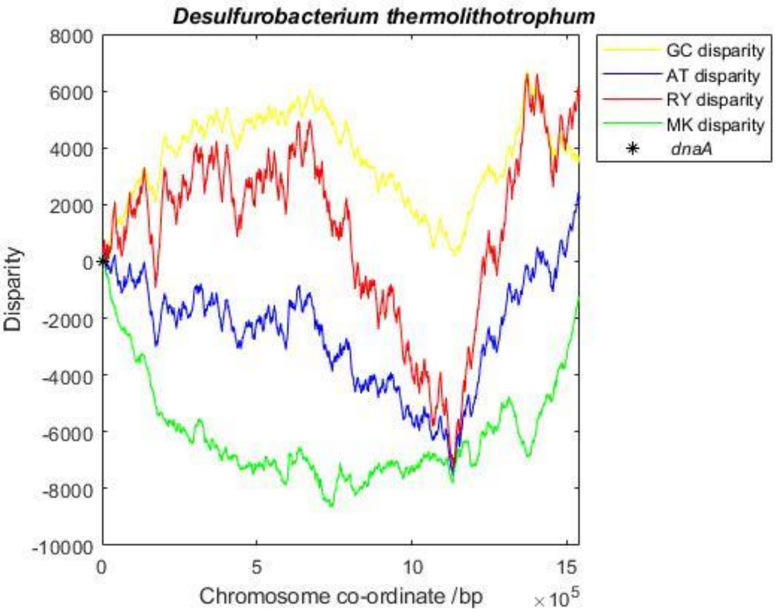

Acidobacterium capsulatum

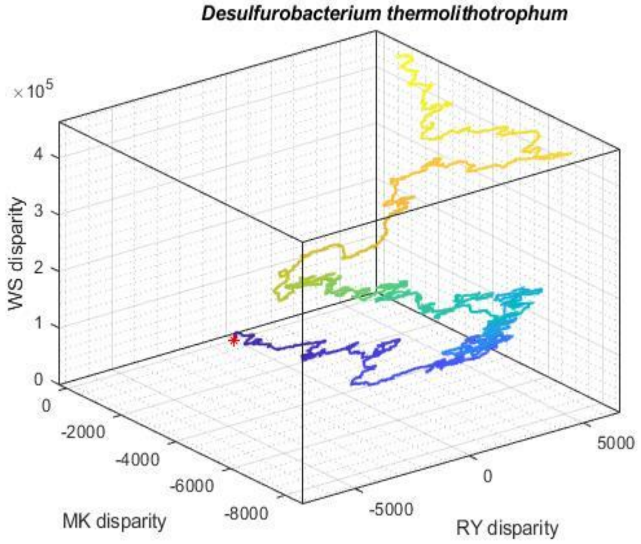

Granulicella mallensis

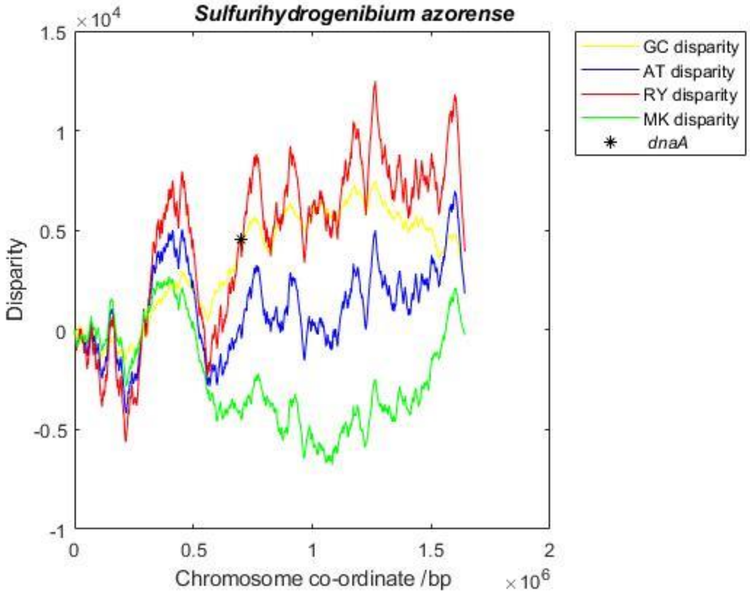

Granulicella mallensis

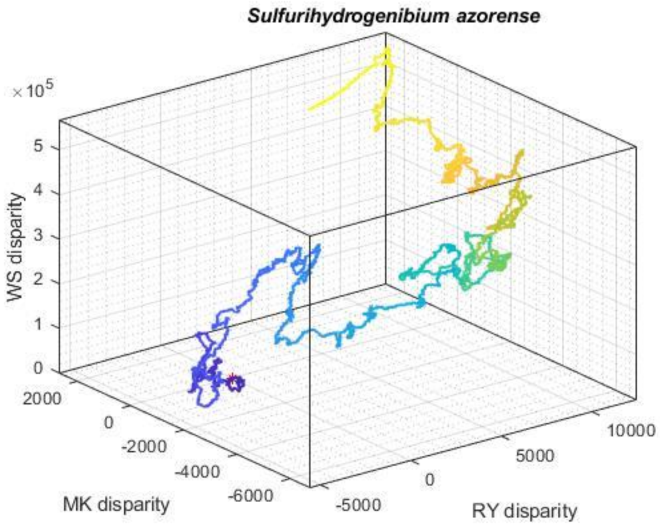

Thermovibrio ammonificans
×10^4
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

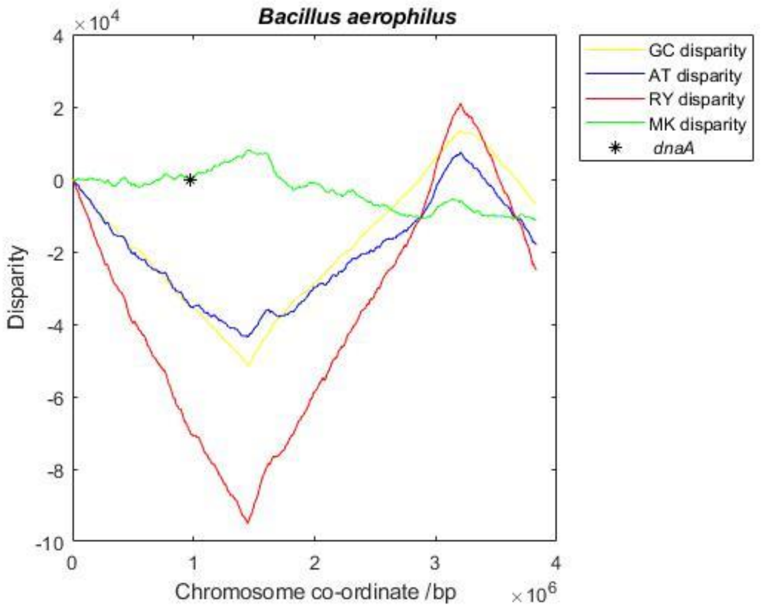

Thermovibrio ammonificans
×10^4
0
-2
-4
-6
WS disparity
0
-1
-2
-3
×10^4
MK disparity
-4
-3
-2
-1
0
×10^4
RY disparity


Desulfurobacterium thermolithotrophum
8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
-10000
Disparity
0
5
10
15
Chromosome co-ordinate /bp
×10^5
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

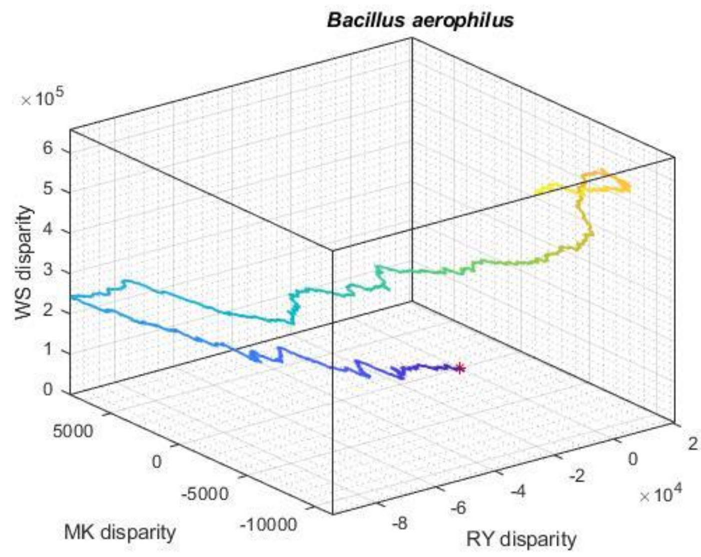

Desulfurobacterium thermolithotrophum
×10^5
4
3
2
1
0
WS disparity
0
-2000
-4000
-6000
-8000
MK disparity
-5000
0
5000
RY disparity


Sulfurihydrogenibium azorense
×10^4
1.5
1
0.5
0
-0.5
-1
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

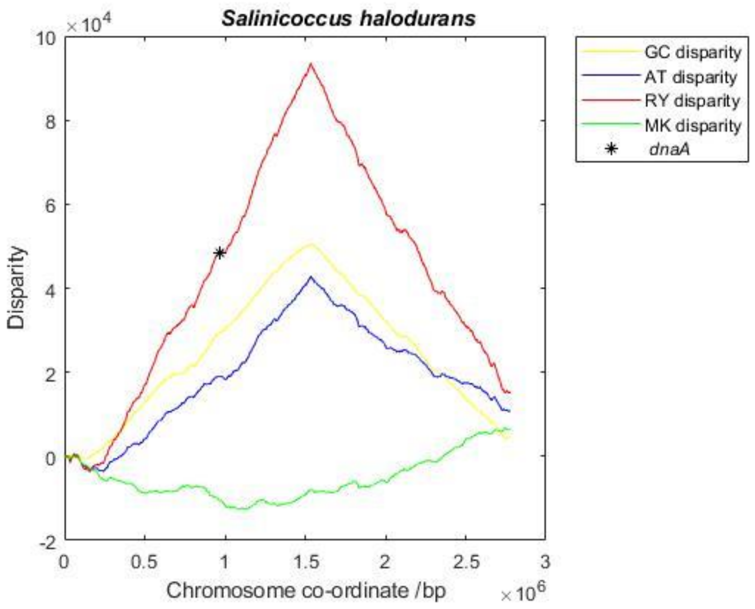

Sulfurihydrogenibium azorense
×10^5
5
4
3
2
1
0
WS disparity
2000
0
-2000
-4000
-6000
MK disparity
-5000
0
5000
10000
RY disparity



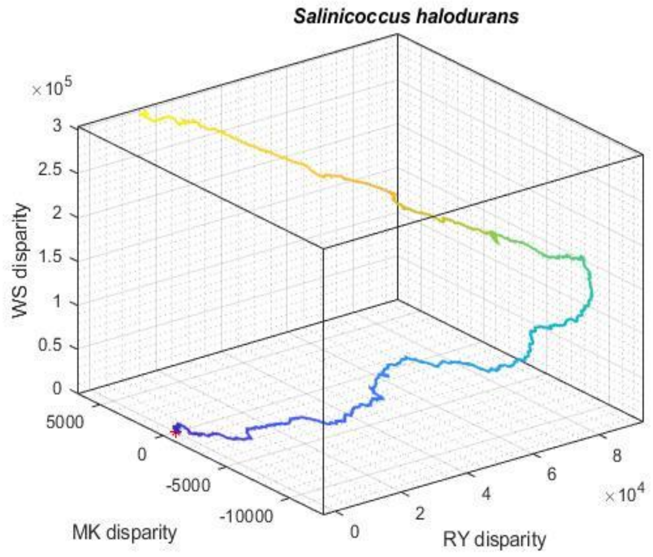
Bacillus aerophilus



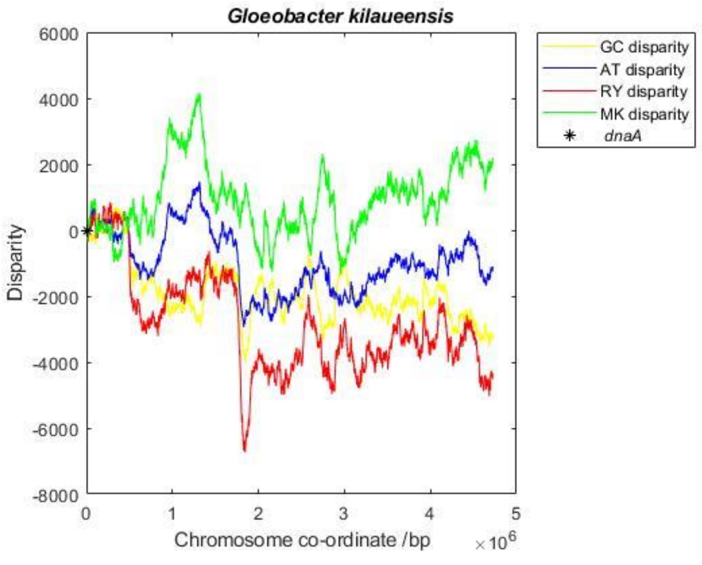
Bacillus aerophilus



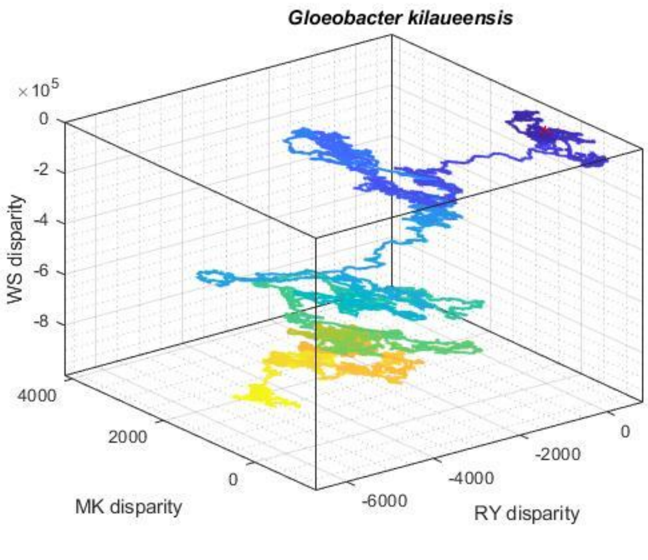
Salinicoccus halodurans



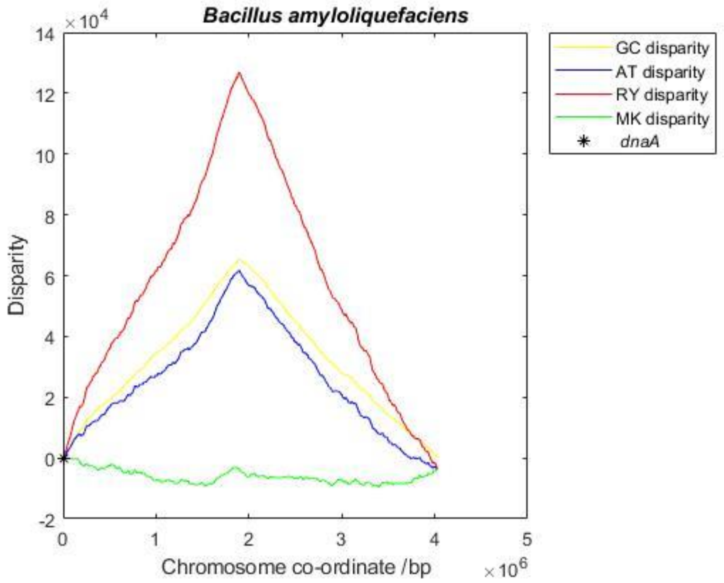
Salinicoccus halodurans

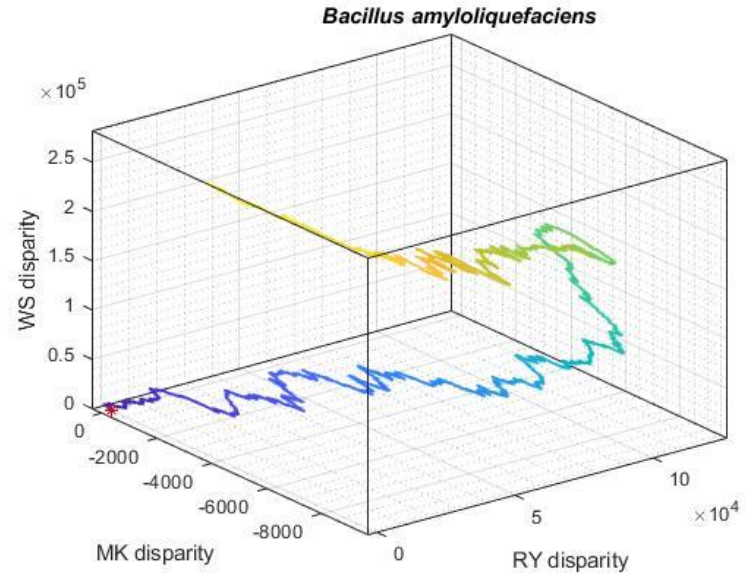


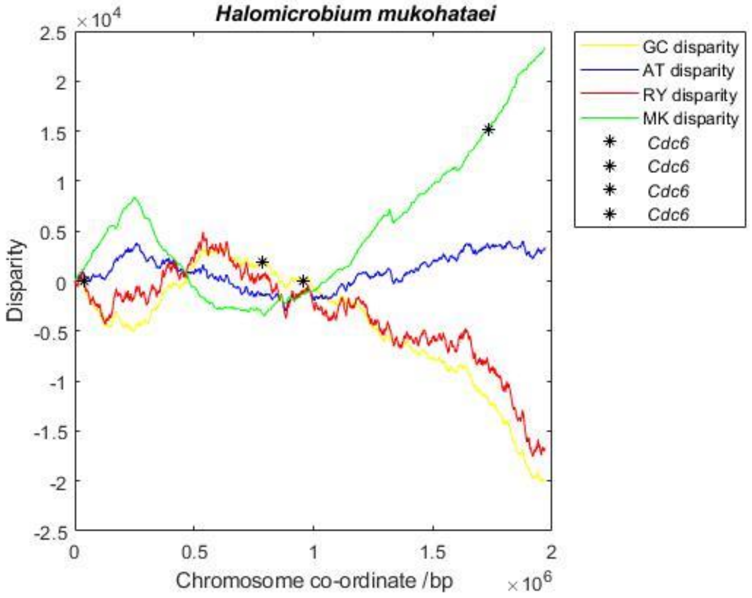
Gloeobacter kilaueensis

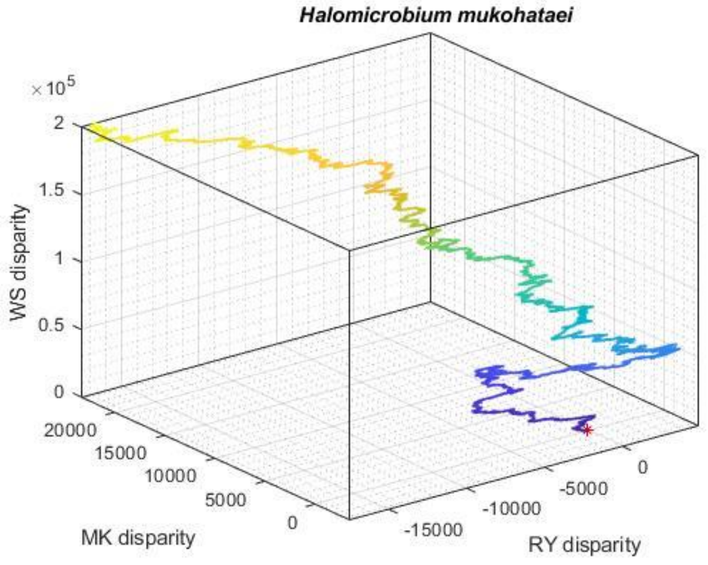


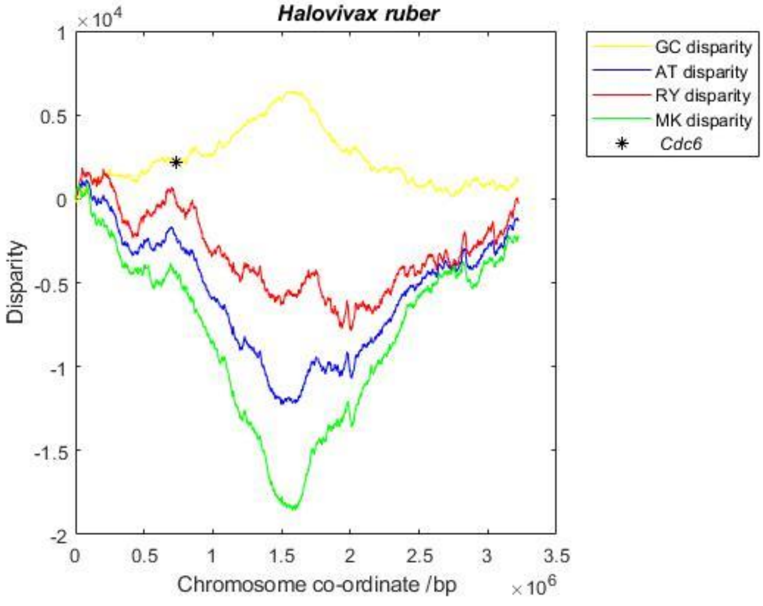
Gloeobacter kilaueensis

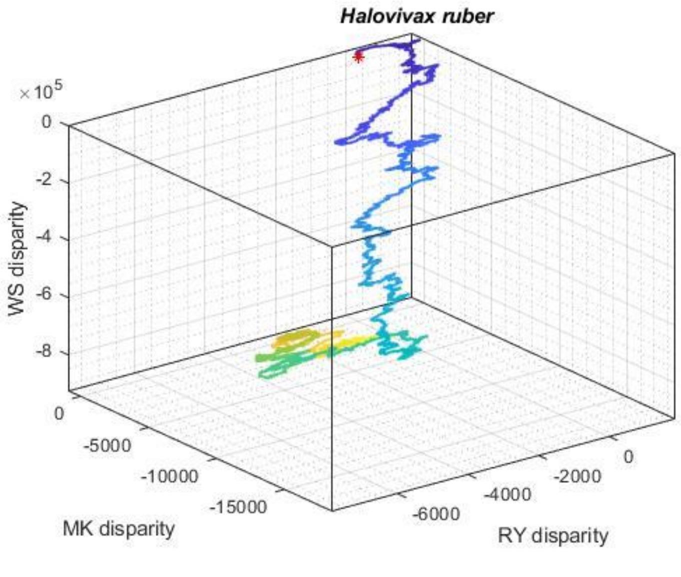

Bacillus amyloliquefaciens
×10^4
14
12
10
8
6
4
2
0
-2
Disparity
0
1
2
3
4
5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

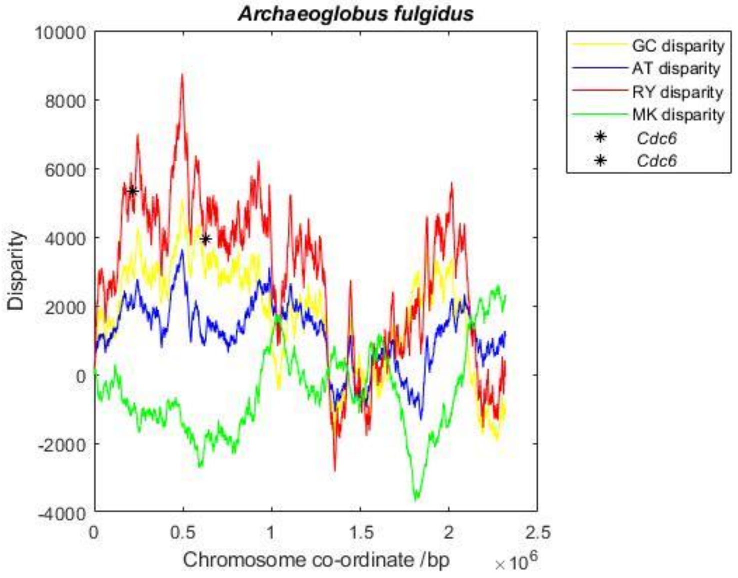

Bacillus amyloliquefaciens
×10^5
2.5
2
1.5
1
0.5
0
WS disparity
0
-2000
-4000
-6000
-8000
MK disparity
0
5
10
×10^4
RY disparity

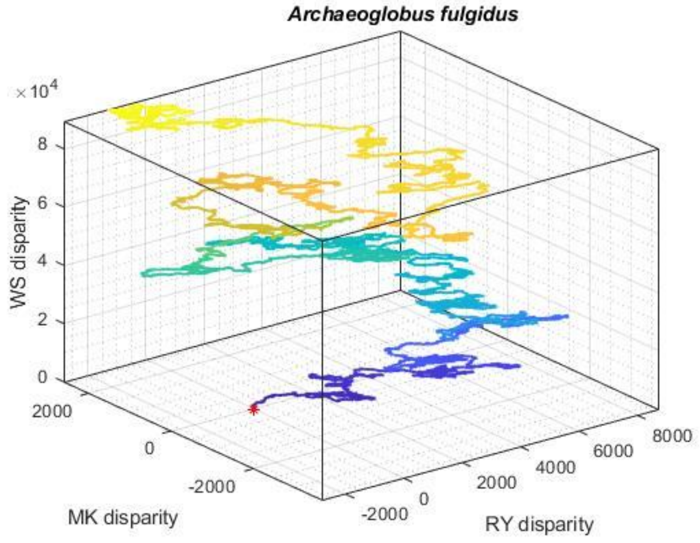

Halomicrobium mukohataei
×10^4
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6

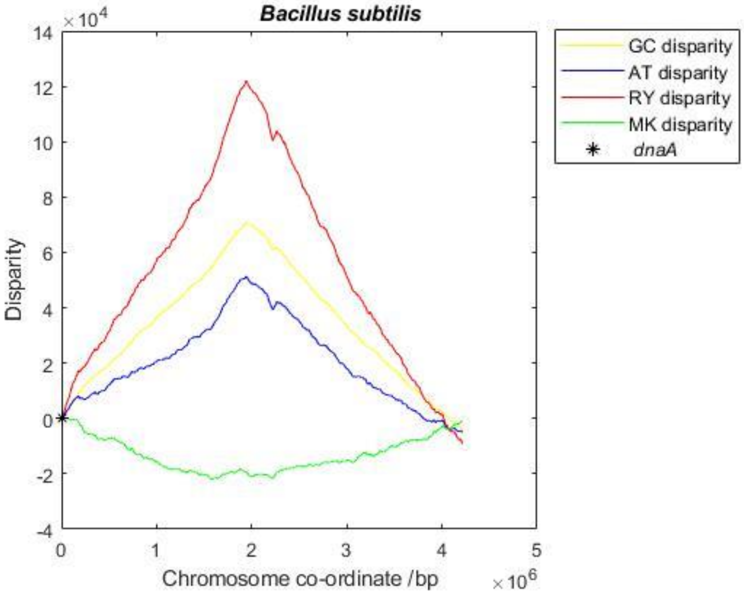

Halomicrobium mukohataei
×10^5
2
1.5
1
0.5
0
WS disparity
20000
15000
10000
5000
0
MK disparity
-15000
-10000
-5000
0
RY disparity

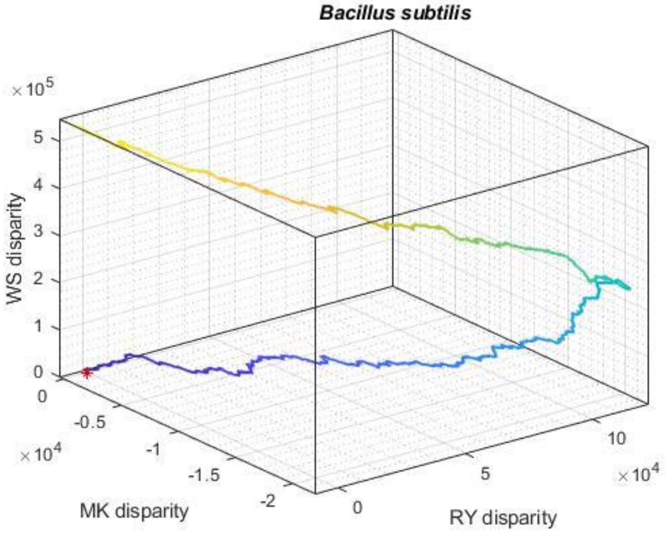

Halovivax ruber
×10^4
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
0
0.5
1
1.5
2
2.5
3
3.5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6

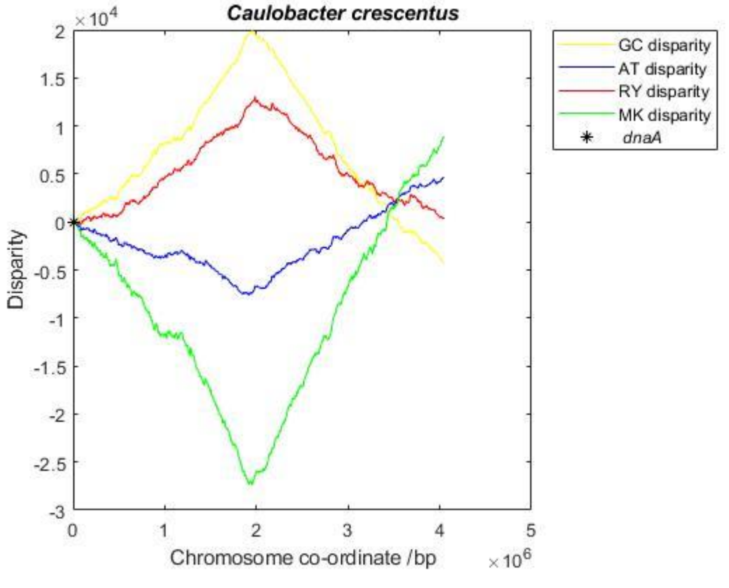

Halovivax ruber
×10^5
0
-2
-4
-6
-8
WS disparity
0
-5000
-10000
-15000
MK disparity
-6000
-4000
-2000
0
RY disparity

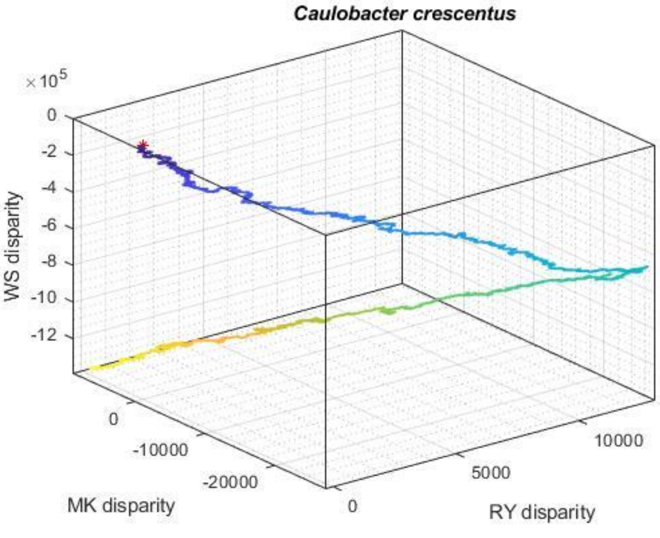

Archaeoglobus fulgidus
Disparity
Chromosome co-ordinate /bp
×10^6
10000
8000
6000
4000
2000
0
-2000
-4000
0
0.5
1
1.5
2
2.5
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6

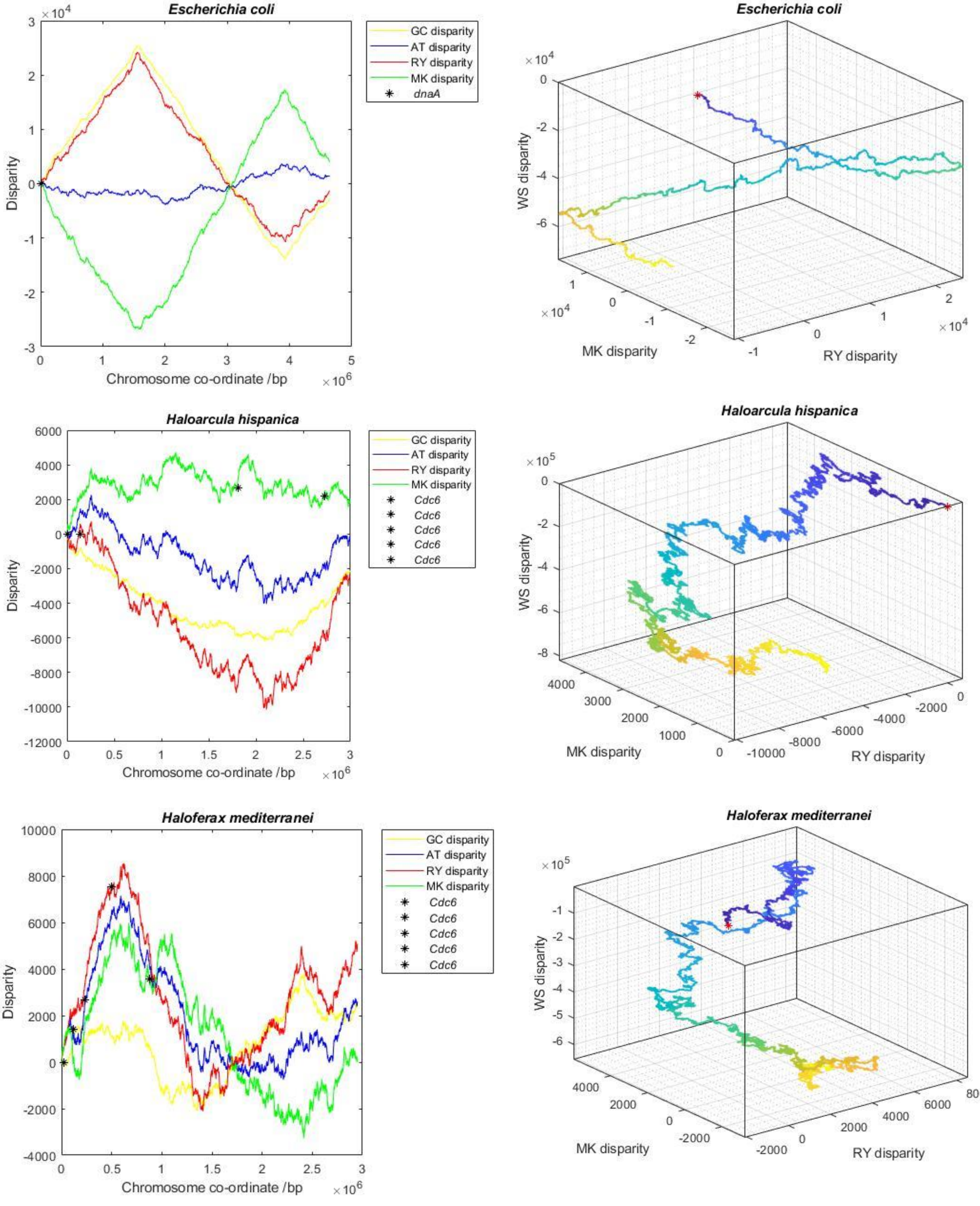

Archaeoglobus fulgidus
×10^4
WS disparity
8
6
4
2
0
2000
0
-2000
MK disparity
-2000
0
2000
4000
6000
8000
RY disparity

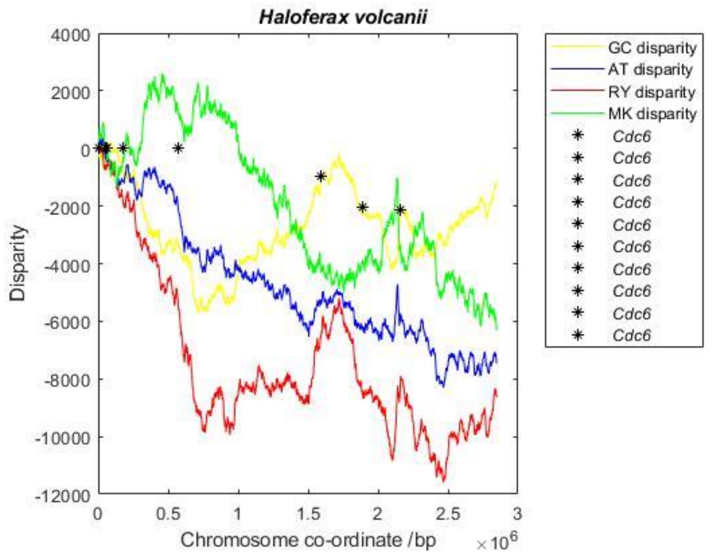

Bacillus subtilis
×10^4
14
12
10
8
6
4
2
0
-2
-4
Disparity
0
1
2
3
4
5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

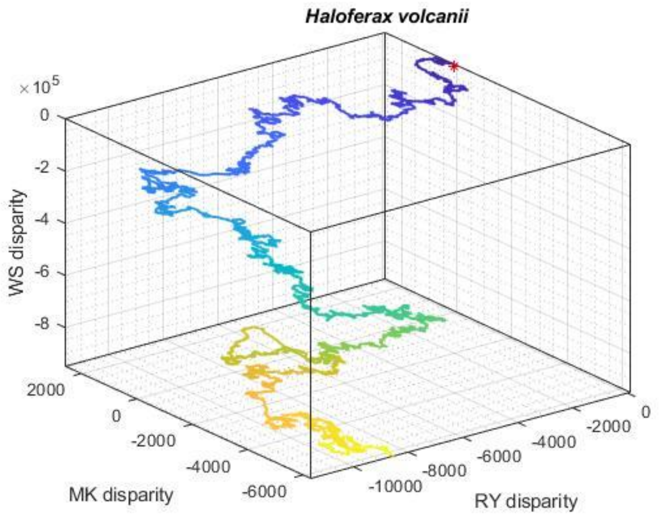

Bacillus subtilis
×10^5
WS disparity
5
4
3
2
1
0
0
-0.5
×10^4
-1
-1.5
-2
MK disparity
0
5
10
×10^4
RY disparity

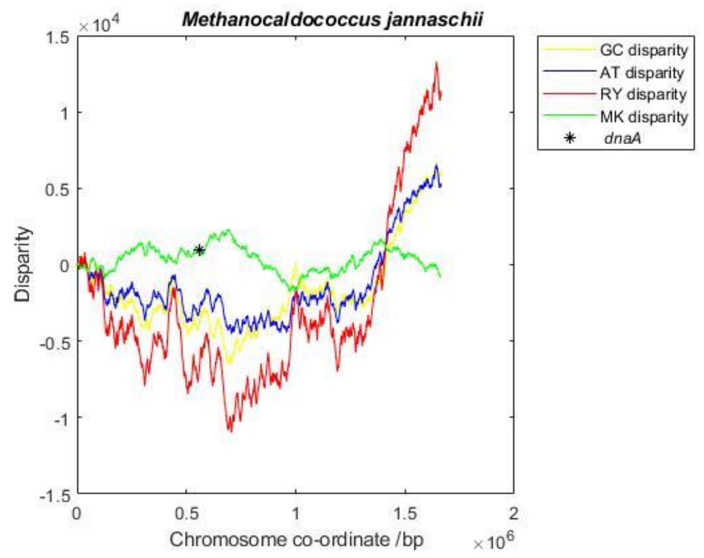

Caulobacter crescentus
×10^4
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
Disparity
0
1
2
3
4
5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

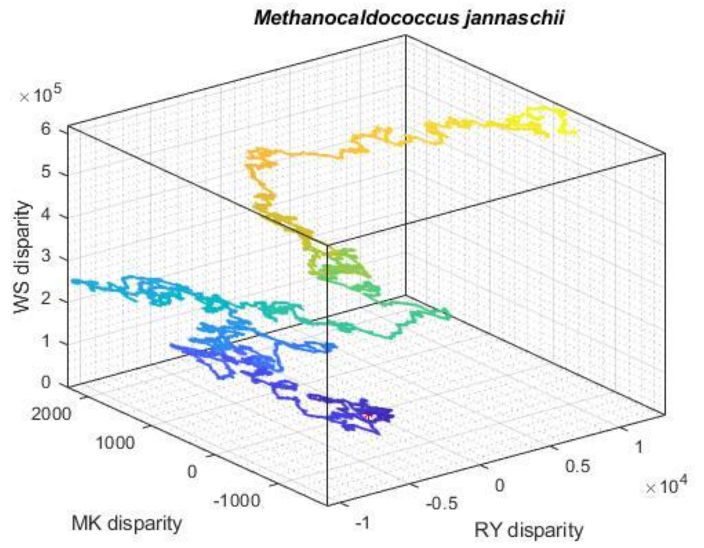

Caulobacter crescentus
×10^5
WS disparity
0
-2
-4
-6
-8
-10
-12
0
-10000
-20000
MK disparity
0
5000
10000
RY disparity

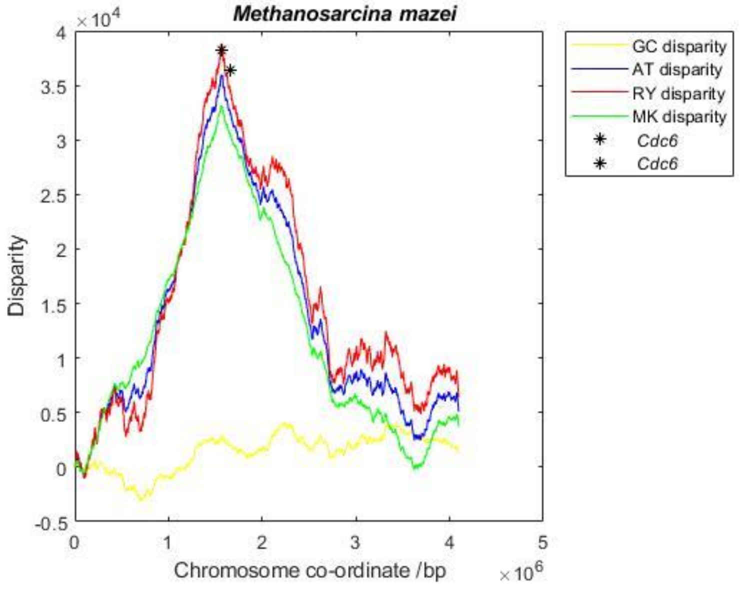

Escherichia coli
×10^4
3
2
1
0
-1
-2
-3
Disparity
0
1
2
3
4
5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA
Escherichia coli
×10^4
0
-2
-4
-6
WS disparity
1
0
-1
-2
×10^4
MK disparity
-1
0
1
2
×10^4
RY disparity
Haloarcula hispanica
6000
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Haloarcula hispanica
×10^5
0
-2
-4
-6
-8
WS disparity
4000
3000
2000
1000
0
MK disparity
-10000
-8000
-6000
-4000
-2000
0
RY disparity
Haloferax mediterranei
10000
8000
6000
4000
2000
0
-2000
-4000
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Haloferax mediterranei
×10^5
-1
-2
-3
-4
-5
-6
WS disparity
4000
2000
0
-2000
MK disparity
-2000
0
2000
4000
6000
8000
RY disparity

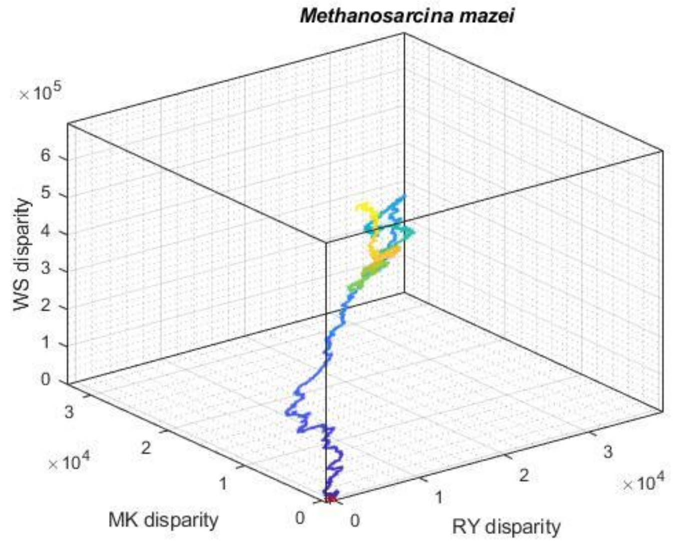

Haloferax volcanii
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6

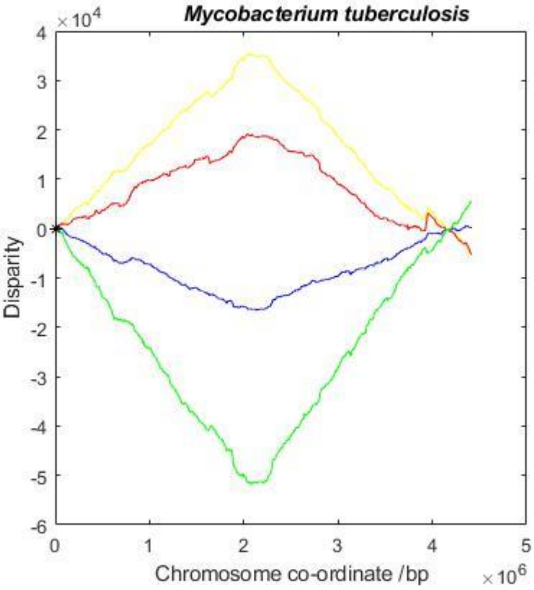

Haloferax volcanii
×10^5
0
-2
-4
-6
-8
WS disparity
2000
0
-2000
-4000
-6000
MK disparity
0
-2000
-4000
-6000
-8000
-10000
RY disparity

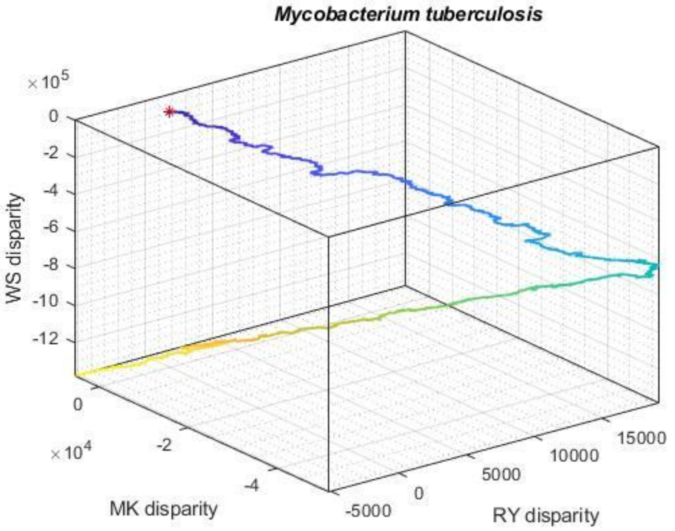

Methanocaldococcus jannaschii
×10^4
1.5
1
0.5
0
-0.5
-1
-1.5
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

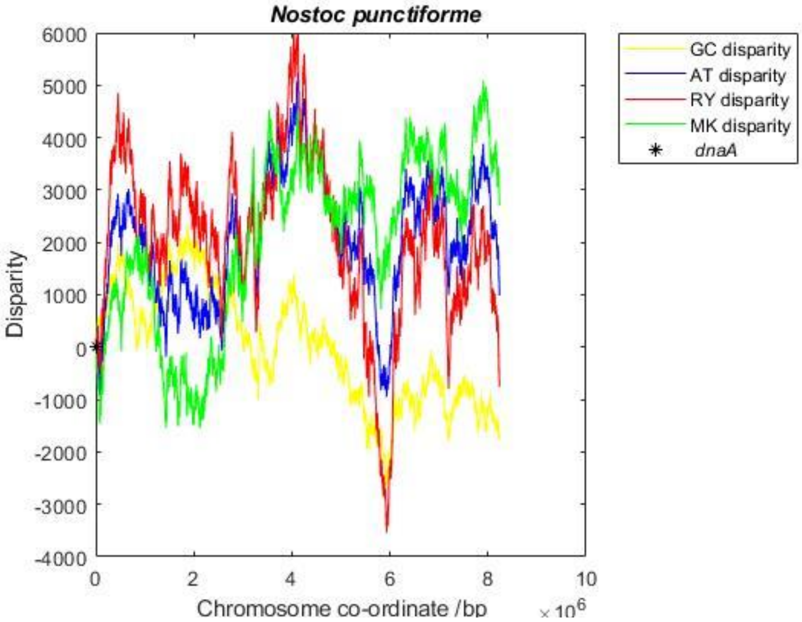

Methanocaldococcus jannaschii
×10^5
6
5
4
3
2
1
0
WS disparity
2000
1000
0
-1000
MK disparity
-1
-0.5
0
0.5
1
×10^4
RY disparity

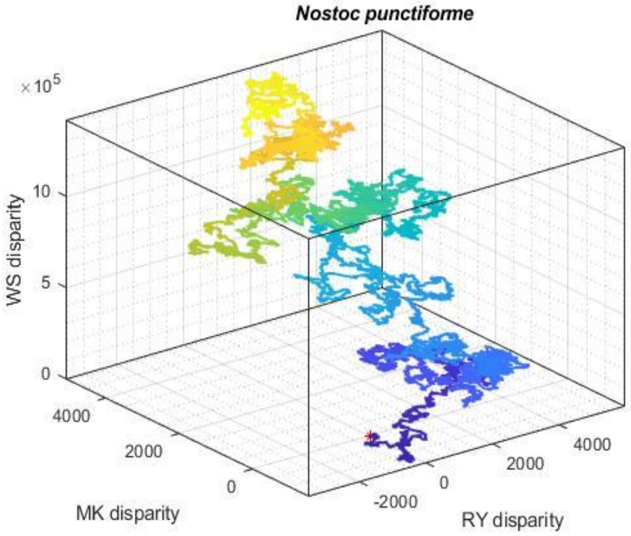

Methanosarcina mazei
×10^4
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
Disparity
0
1
2
3
4
5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6

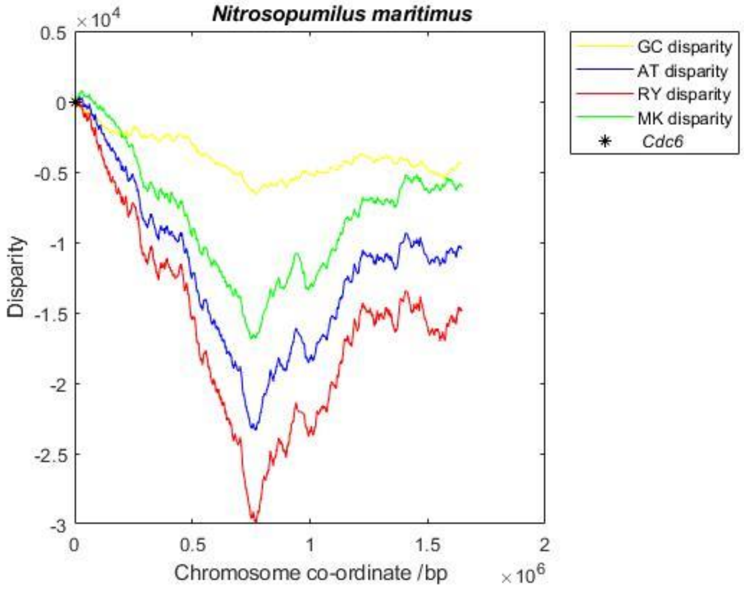

Methanosarcina mazei
×10^5
6
5
4
3
2
1
0
WS disparity
3
2
1
0
×10^4
MK disparity
0
1
2
3
×10^4
RY disparity

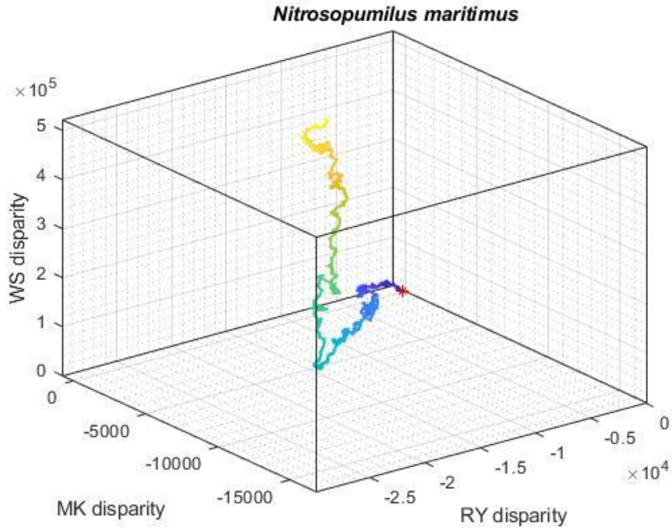

Mycobacterium tuberculosis
×10^4
4
3
2
1
0
-1
-2
-3
-4
-5
-6
Disparity
0
1
2
3
4
5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

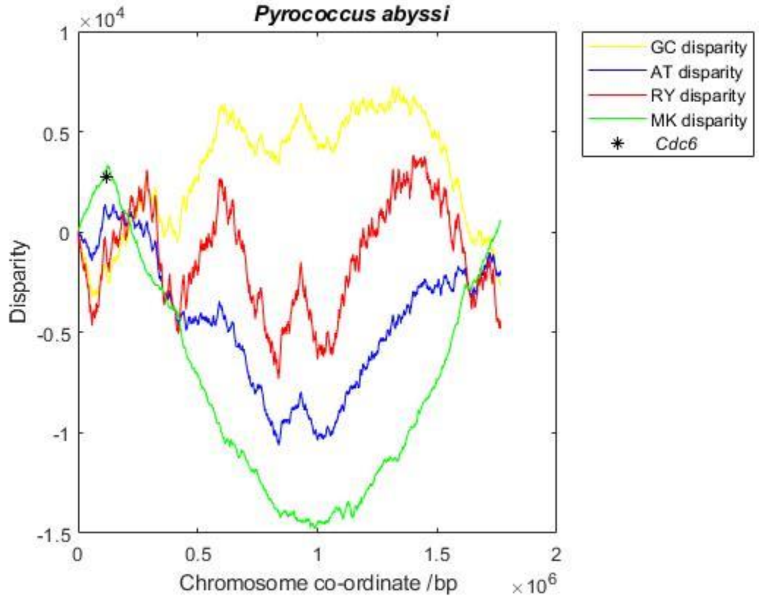

Mycobacterium tuberculosis
×10^5
0
-2
-4
-6
-8
-10
-12
WS disparity
0
×10^4
-2
-4
MK disparity
-5000
0
5000
10000
15000
RY disparity

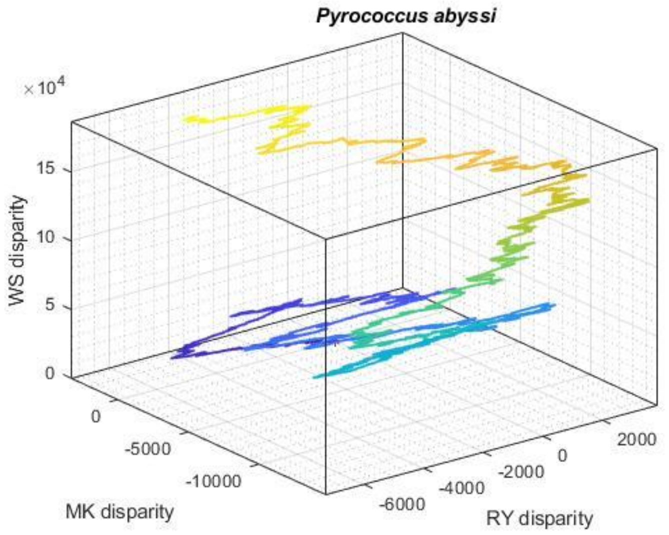

Nostoc punctiforme
6000
5000
4000
3000
2000
1000
0
-1000
-2000
-3000
-4000
Disparity
0
2
4
6
8
10
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

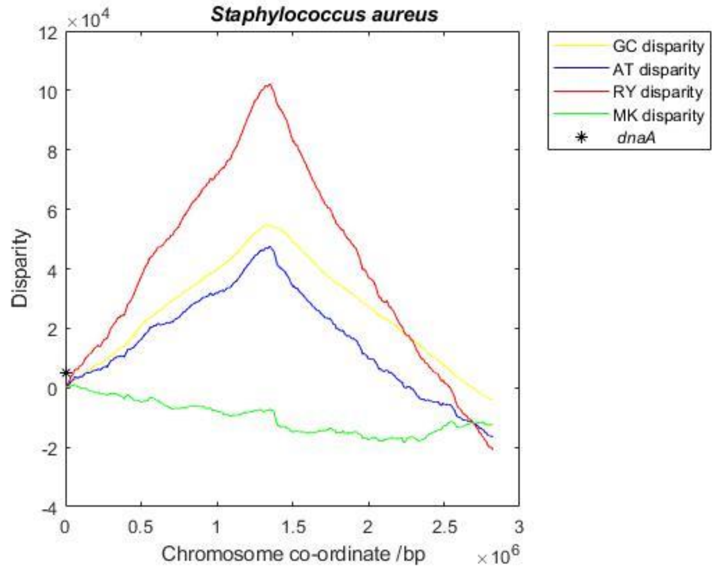

Nostoc punctiforme
×10^5
10
5
0
WS disparity
4000
2000
0
MK disparity
-2000
0
2000
4000
RY disparity

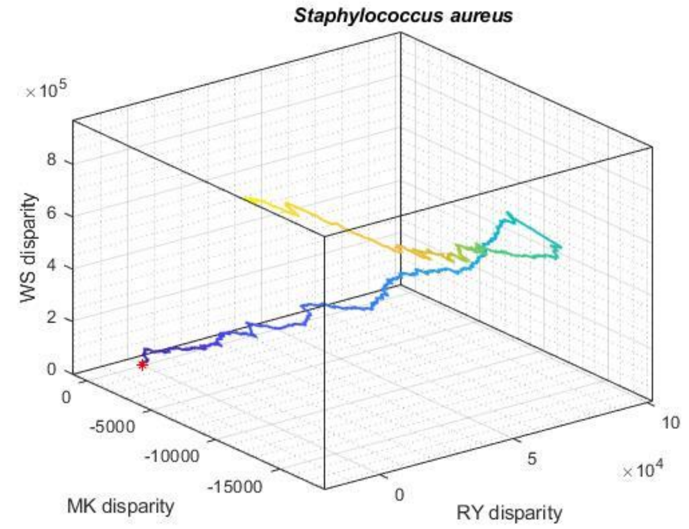

Nitrosopumilus maritimus
×10^4
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6

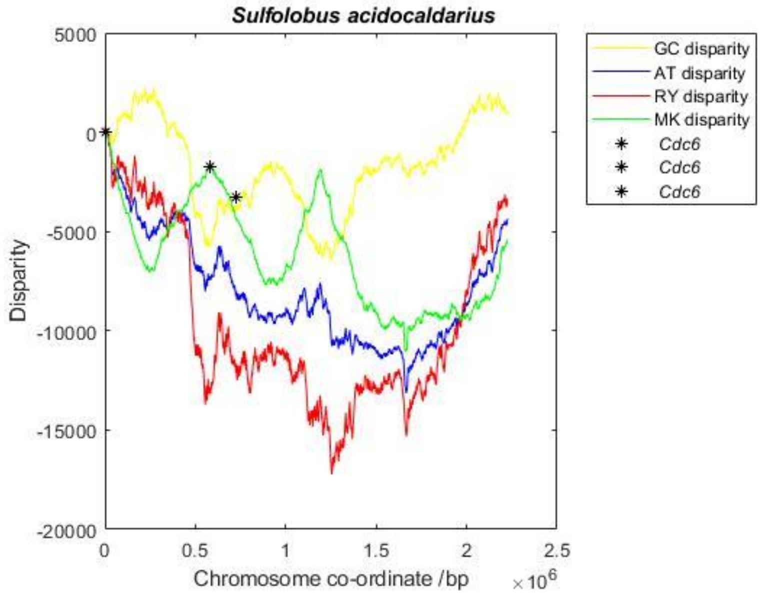

Nitrosopumilus maritimus
×10^5
5
4
3
2
1
0
WS disparity
0
-5000
-10000
-15000
MK disparity
0
-0.5
-1
-1.5
-2
-2.5
×10^4
RY disparity

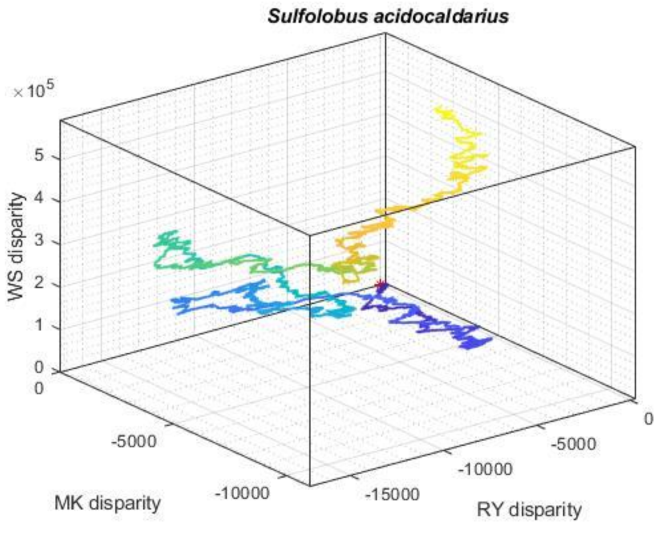

Pyrococcus abyssi

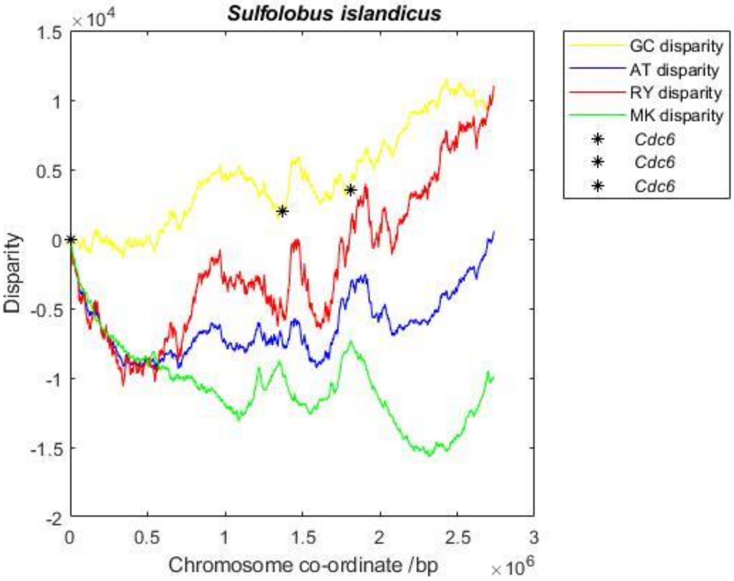

Pyrococcus abyssi

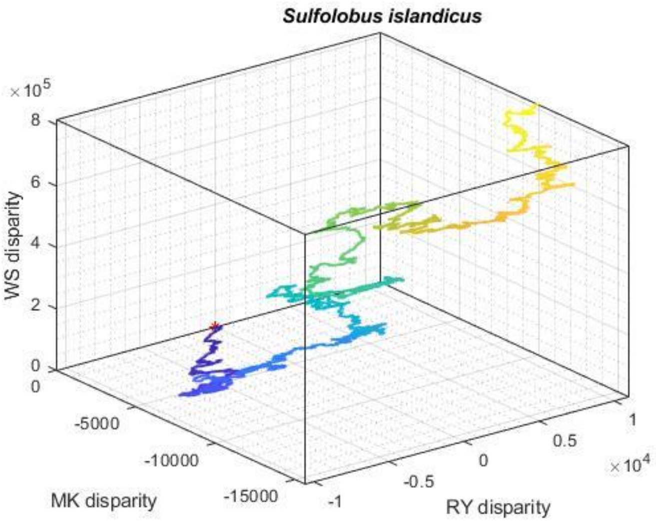

Staphylococcus aureus

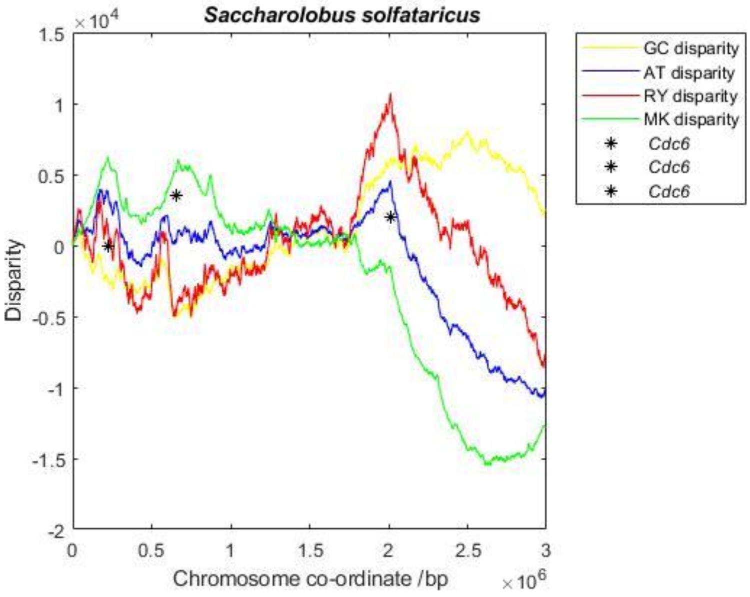

Staphylococcus aureus

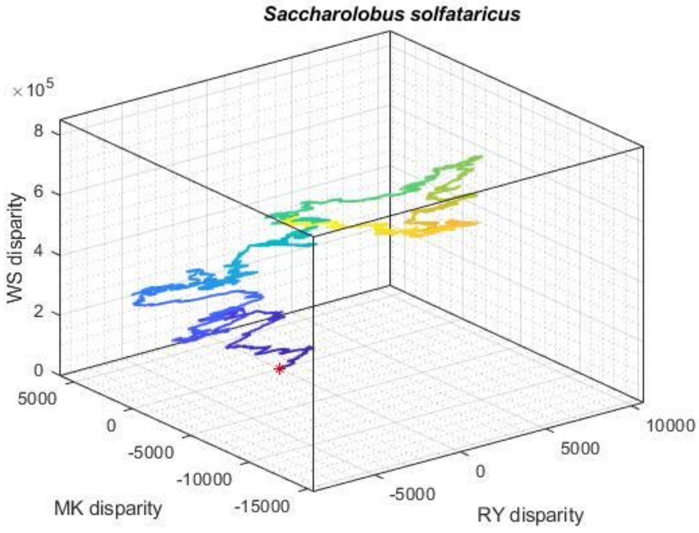

Sulfolobus acidocaldarius

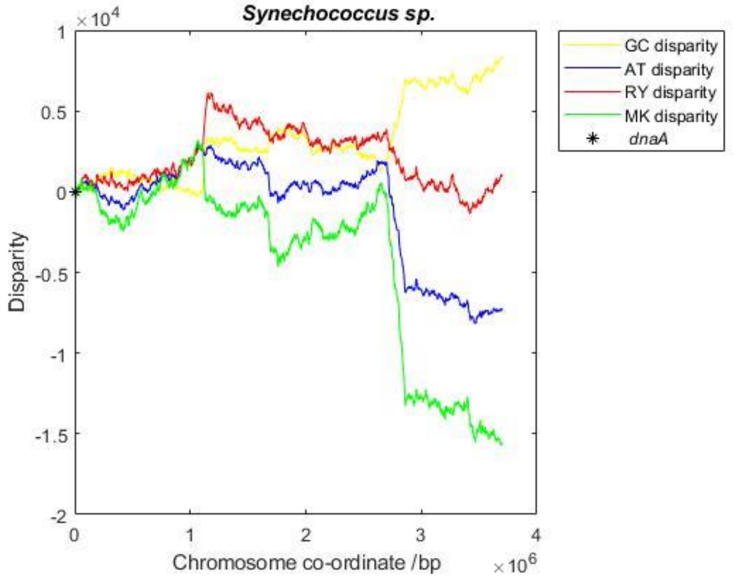

Sulfolobus acidocaldarius

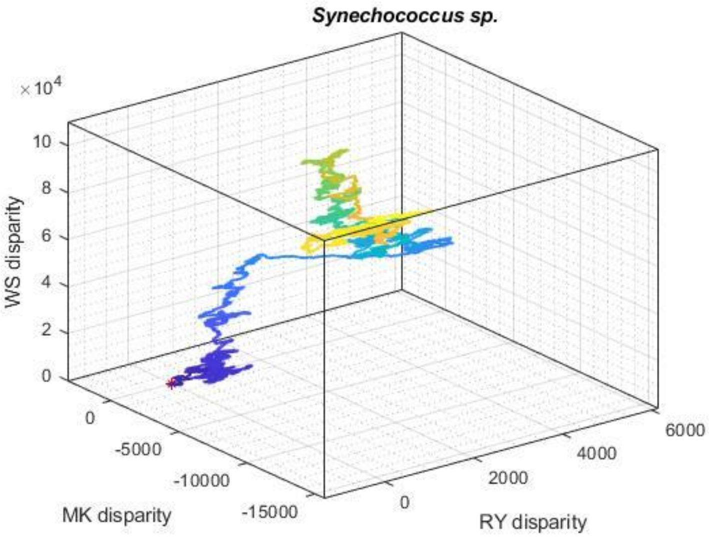

Sulfolobus islandicus
×10^4
1.5
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6

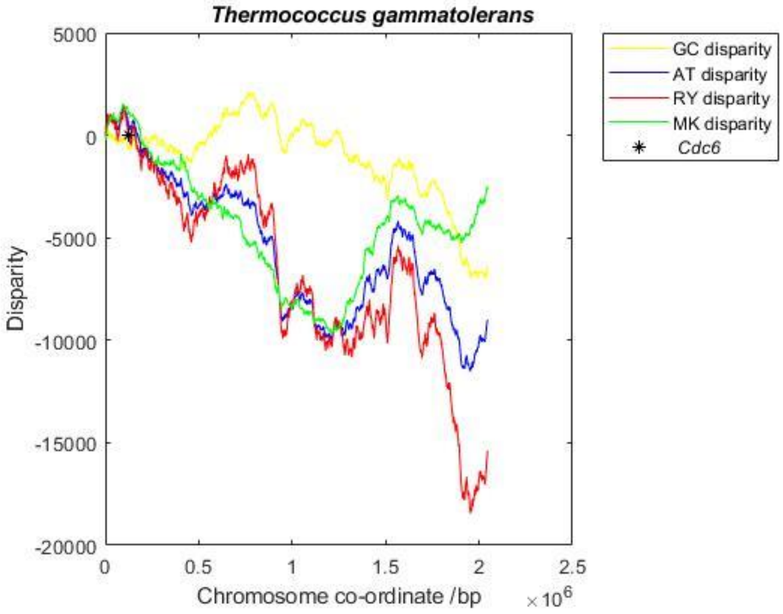

Sulfolobus islandicus
×10^5
8
6
4
2
0
WS disparity
0
-5000
-10000
-15000
MK disparity
-1
-0.5
0
0.5
1
×10^4
RY disparity


Saccharolobus solfataricus
×10^4
1.5
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6

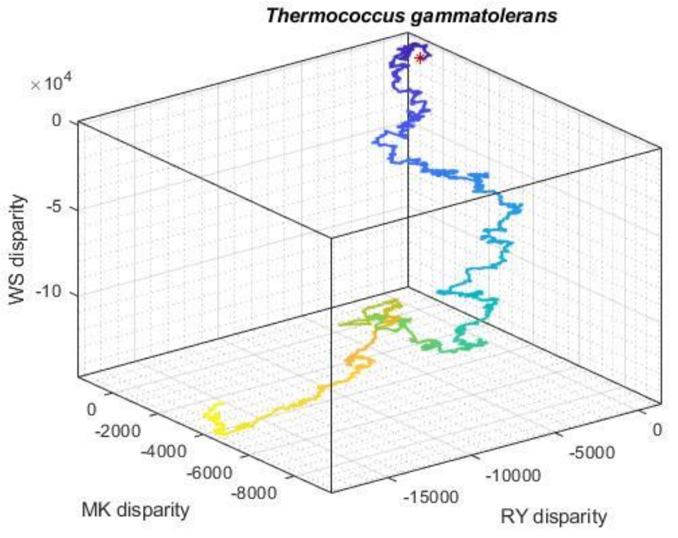

Saccharolobus solfataricus
×10^5
8
6
4
2
0
WS disparity
5000
0
-5000
-10000
-15000
MK disparity
-5000
0
5000
10000
RY disparity


Synechococcus sp.
×10^4
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
0
1
2
3
4
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

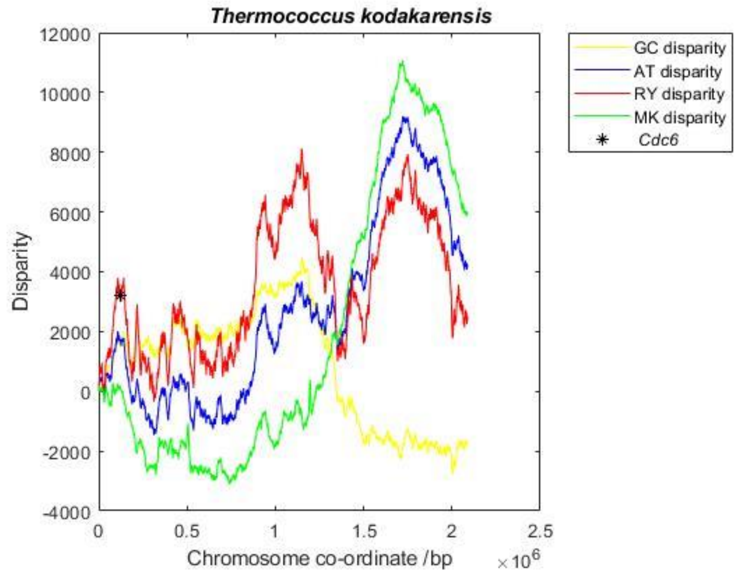

Synechococcus sp.
×10^4
10
8
6
4
2
0
WS disparity
0
-5000
-10000
-15000
MK disparity
0
2000
4000
6000
RY disparity



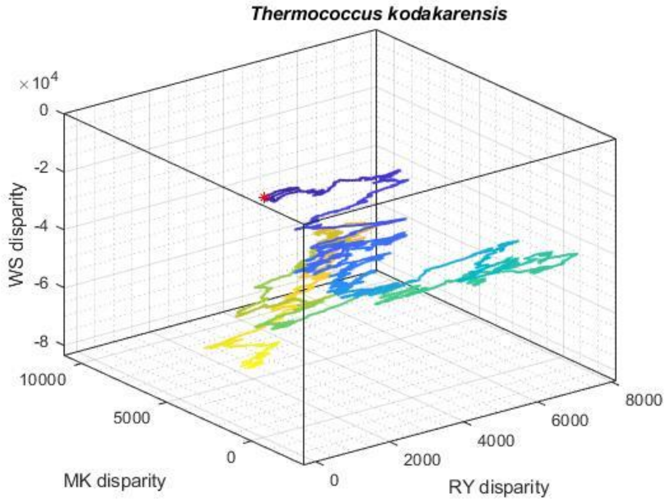
Thermococcus gammatolerans
5000
0
-5000
-10000
-15000
-20000
Disparity
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6



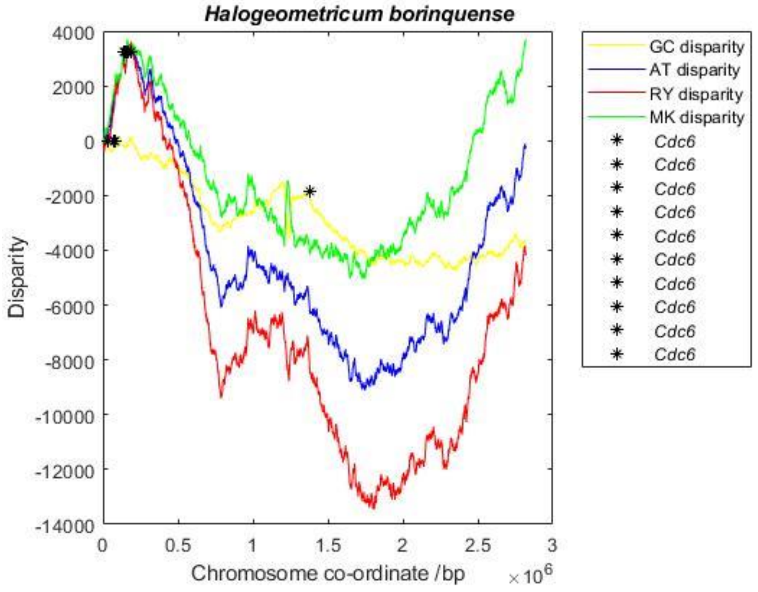
Thermococcus gammatolerans
×10^4
0
-5
-10
WS disparity
0
-2000
-4000
-6000
-8000
MK disparity
0
-5000
-10000
-15000
RY disparity



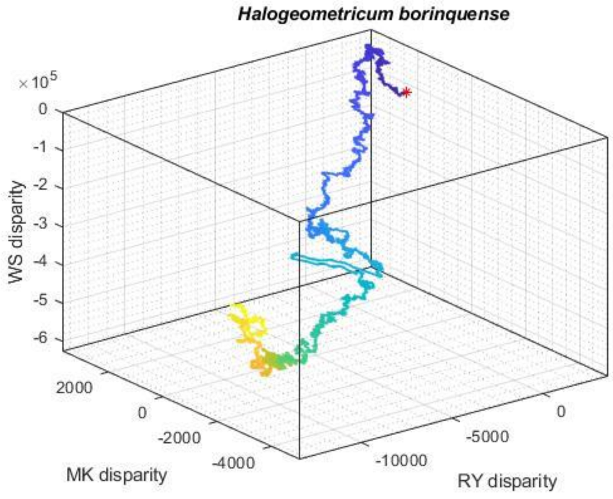
Thermococcus kodakarensis
12000
10000
8000
6000
4000
2000
0
-2000
-4000
Disparity
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6



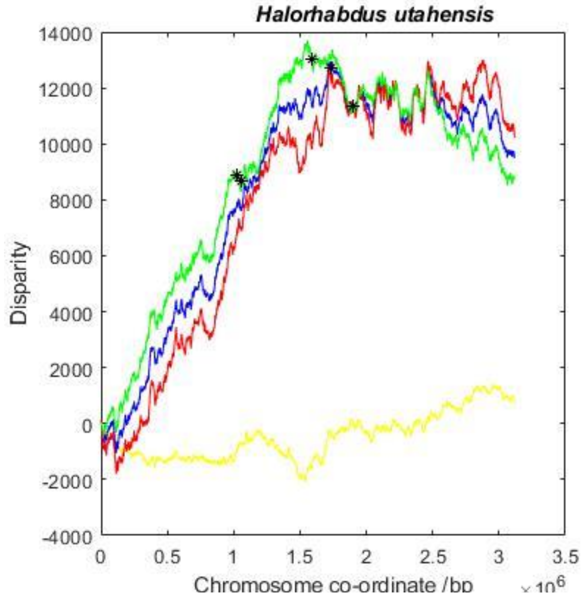
Thermococcus kodakarensis
×10^4
0
-2
-4
-6
-8
WS disparity
10000
5000
0
MK disparity
0
2000
4000
6000
8000
RY disparity

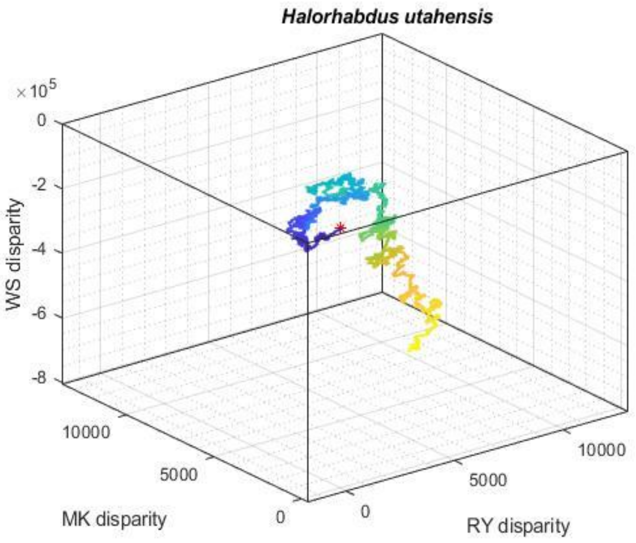


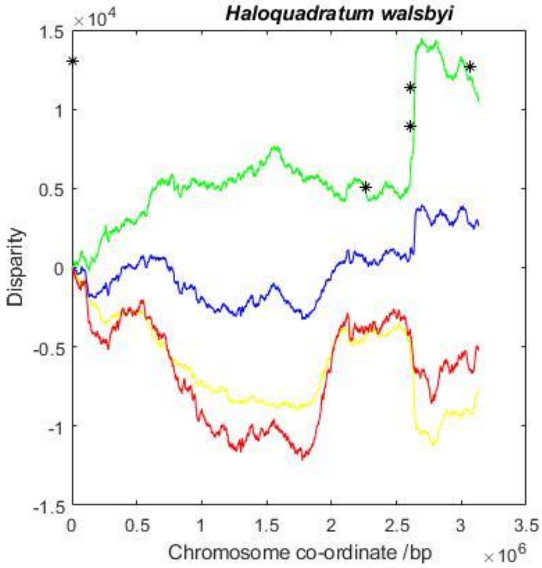
Halogeometricum borinquense
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6

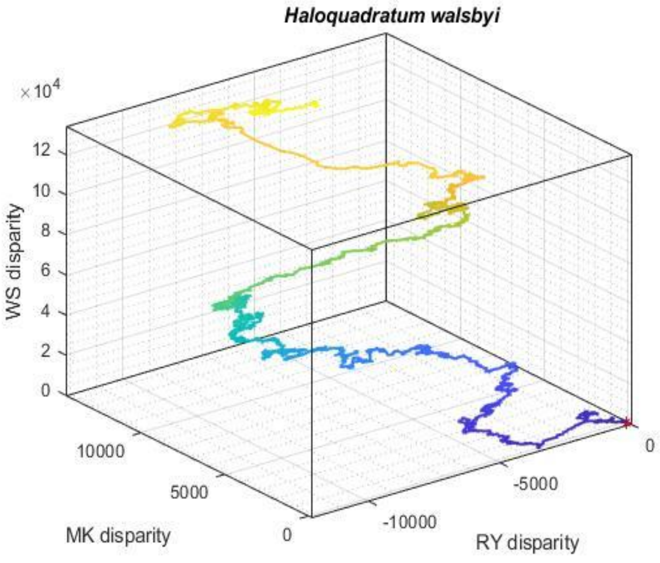


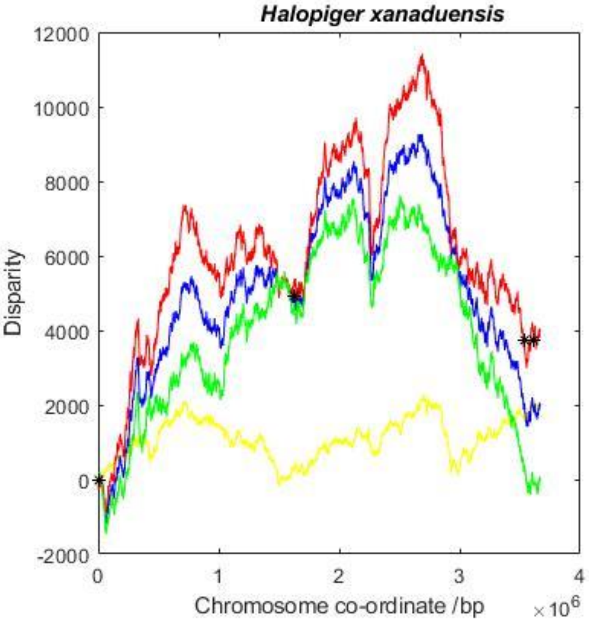
Halogeometricum borinquense
×10^5
0
-1
-2
-3
-4
-5
-6
WS disparity
2000
0
-2000
-4000
MK disparity
0
-5000
-10000
RY disparity

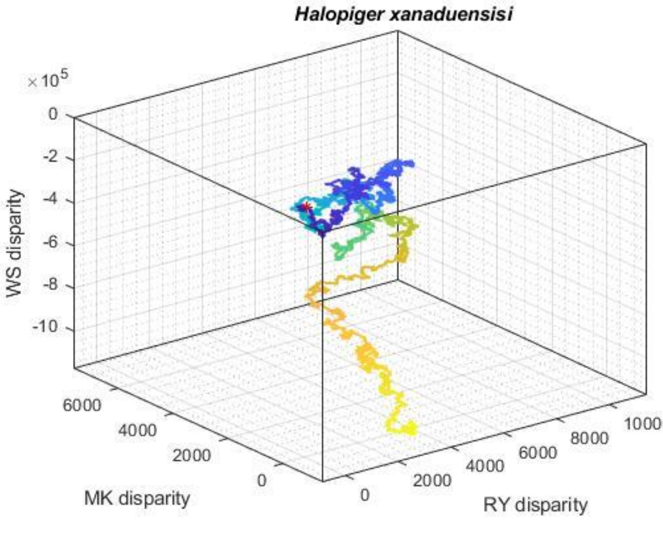

Halorhabdus utahensis
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
Disparity
14000
12000
10000
8000
6000
4000
2000
0
-2000
-4000
0
0.5
1
1.5
2
2.5
3
3.5
Chromosome co-ordinate /bp
×10^6

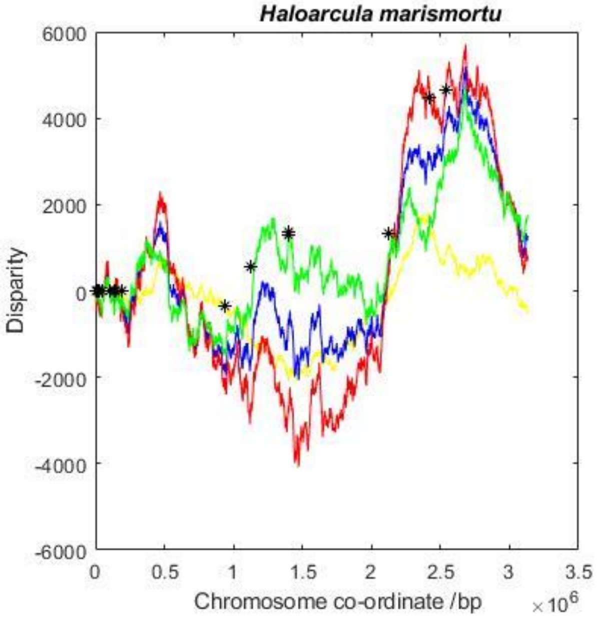

Halorhabdus utahensis
×10^5
0
-2
-4
-6
-8
WS disparity
10000
5000
0
MK disparity
0
5000
10000
RY disparity

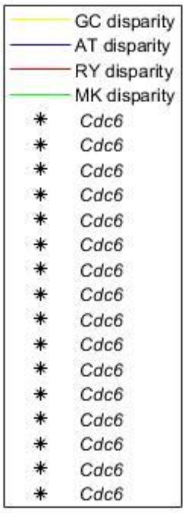

Haloquadratum walsbyi
×10^4
1.5
1
0.5
0
-0.5
-1
-1.5
Disparity
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6
Cdc6
0
0.5
1
1.5
2
2.5
3
3.5
Chromosome co-ordinate /bp
×10^6

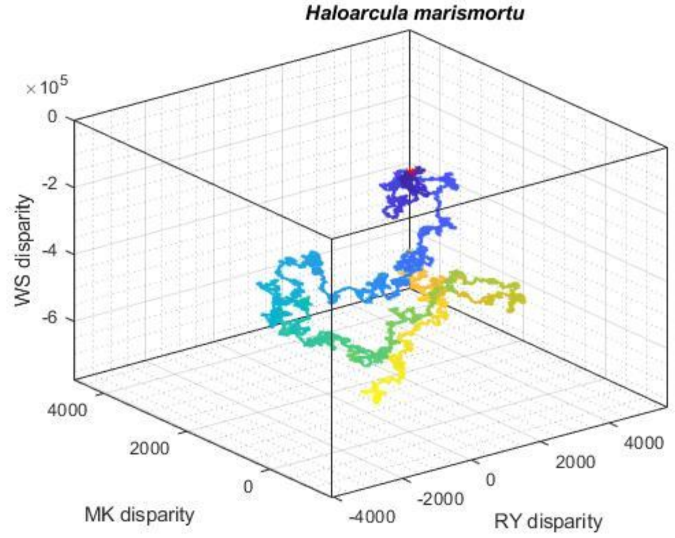

Haloquadratum walsbyi
×10^4
12
10
8
6
4
2
0
WS disparity
10000
5000
0
MK disparity
-10000
-5000
0
RY disparity

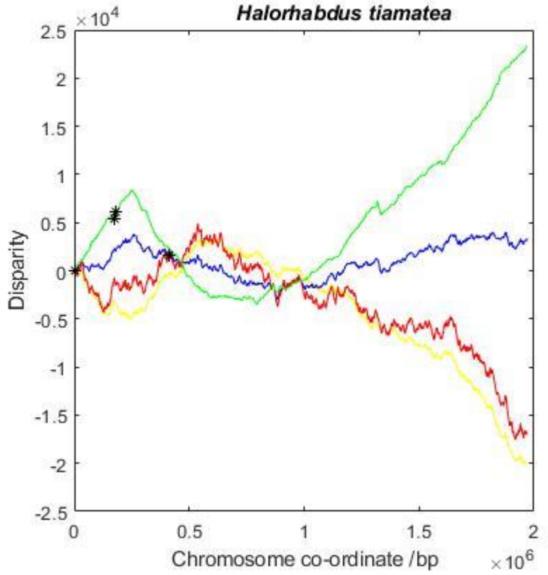

Halopiger xanaduensis
GC disparity
AT disparity
RY disparity
MK disparity
Cdc6
Cdc6
Cdc6
Cdc6
12000
10000
8000
6000
4000
2000
0
-2000
Disparity
0
1
2
3
4
Chromosome co-ordinate /bp
×10^6

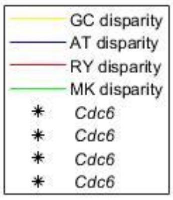

Halopiger xanaduensisi
×10^5
0
-2
-4
-6
-8
-10
WS disparity
6000
4000
2000
0
MK disparity
0
2000
4000
6000
8000
10000
RY disparity

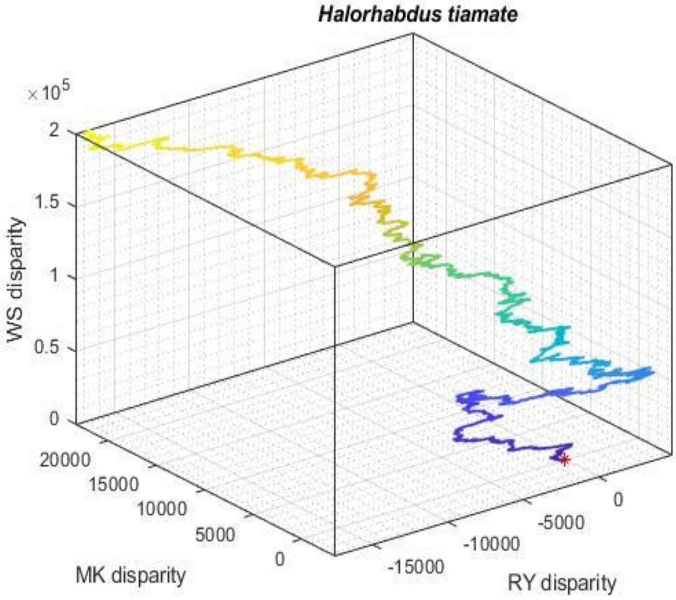

Haloarcula marismortu
6000
4000
2000
0
-2000
-4000
-6000
Disparity
0
0.5
1
1.5
2
2.5
3
3.5
Chromosome co-ordinate /bp
× 10^6

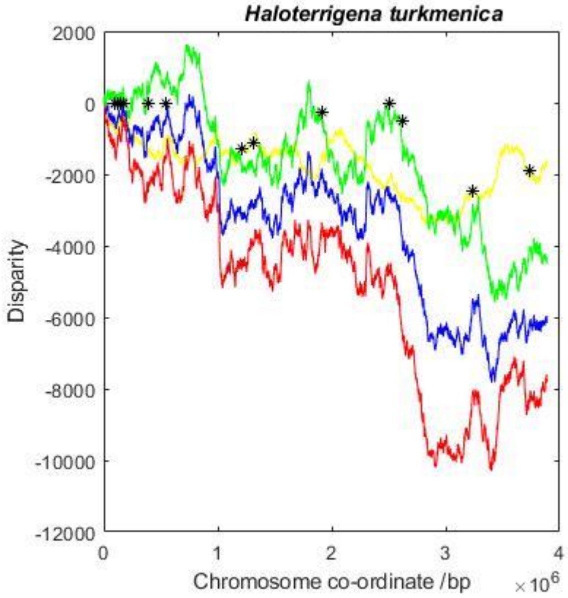

GC disparity
AT disparity
RY disparity
MK disparity
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6

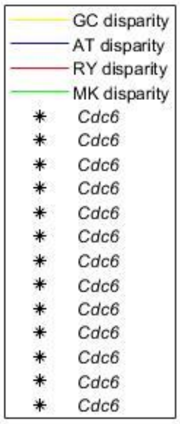

Haloarcula marismortu
× 10^5
0
-2
-4
-6
WS disparity
4000
2000
0
MK disparity
-4000
-2000
0
2000
4000
RY disparity

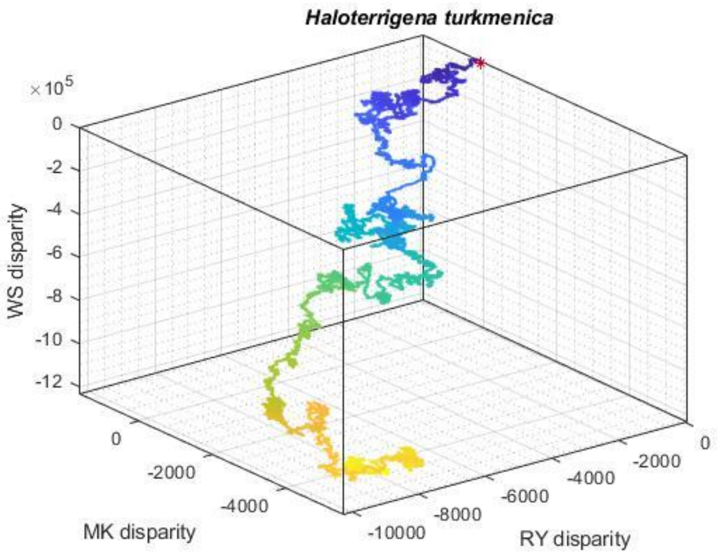

Halorhabdus tiamatea
× 10^4
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
× 10^6

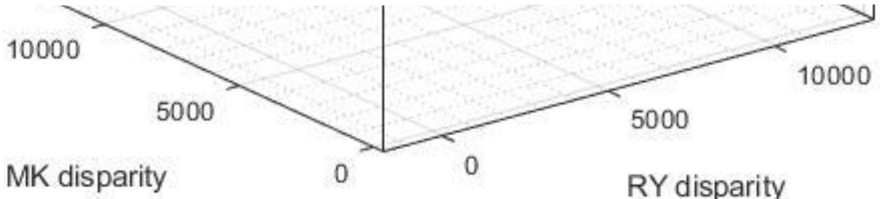

GC disparity
AT disparity
RY disparity
MK disparity
* Cdc6
* Cdc6
* Cdc6
* Cdc6

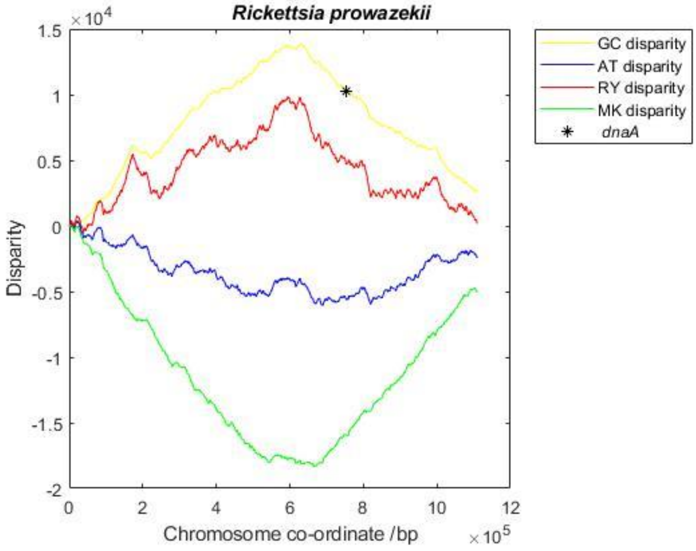

Halorhabdus tiamate
× 10^5
2
1.5
1
0.5
0
WS disparity
20000
15000
10000
5000
0
MK disparity
-15000
-10000
-5000
0
RY disparity

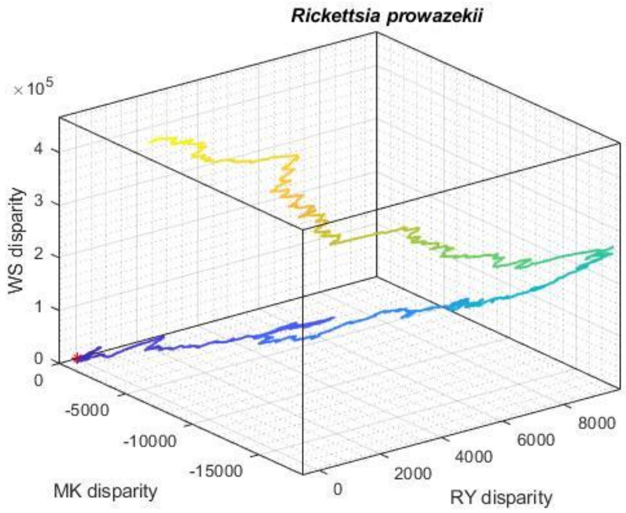

Haloterrigena turkmenica
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
Disparity
0
1
2
3
4
Chromosome co-ordinate /bp
× 10^6

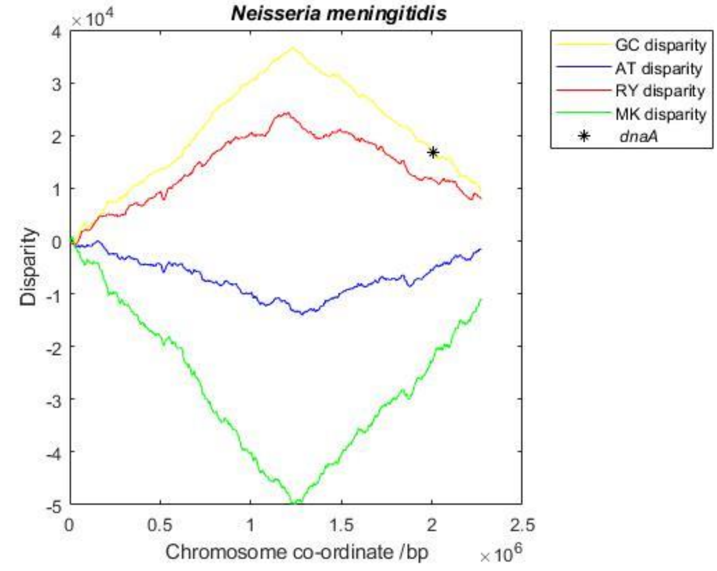

GC disparity
AT disparity
RY disparity
MK disparity
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6
* Cdc6

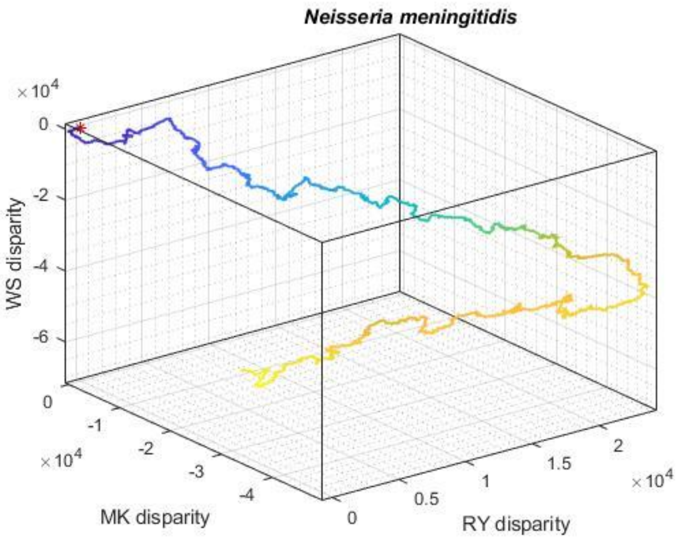

Haloterrigena turkmenica
× 10^5
0
-2
-4
-6
-8
-10
-12
WS disparity
0
-2000
-4000
MK disparity
-10000
-8000
-6000
-4000
-2000
0
RY disparity

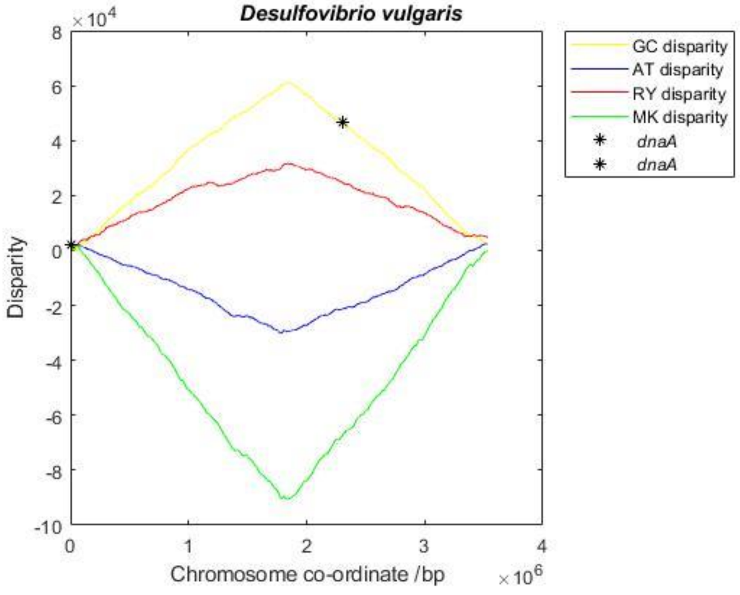

10000
5000
MK disparity
0
0
5000
10000
RY disparity

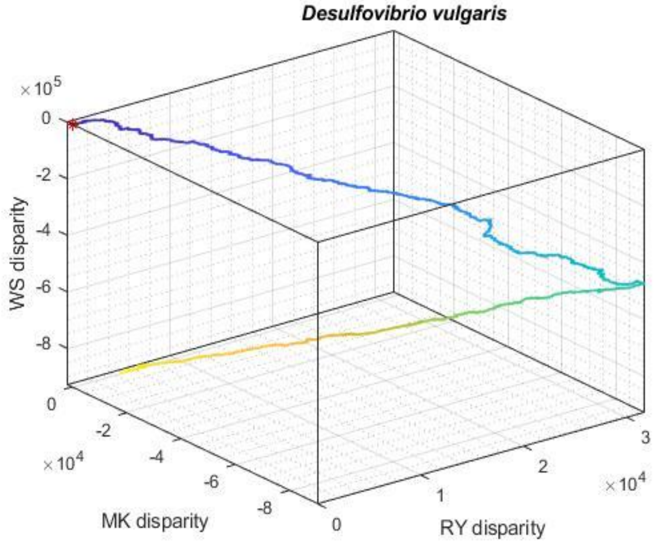

Rickettsia prowazekii
×10^4
1.5
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
0
2
4
6
8
10
12
Chromosome co-ordinate /bp
×10^5
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

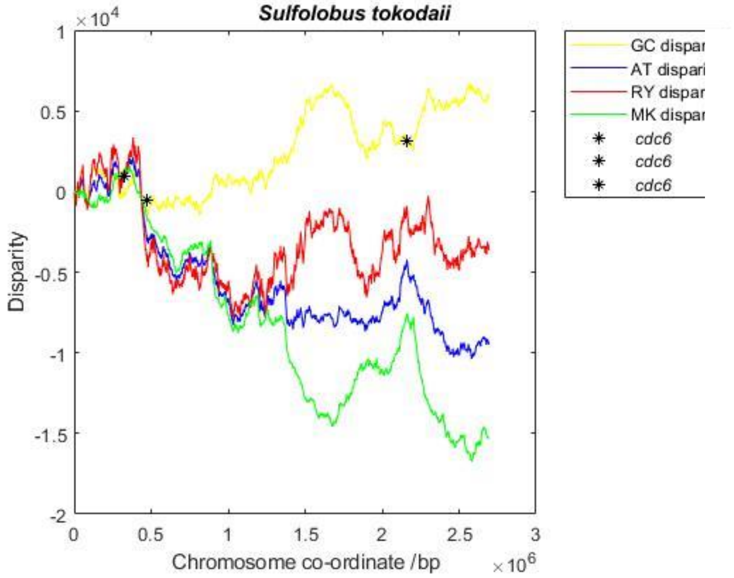

Rickettsia prowazekii
×10^5
4
3
2
1
0
WS disparity
0
-5000
-10000
-15000
MK disparity
0
2000
4000
6000
8000
RY disparity

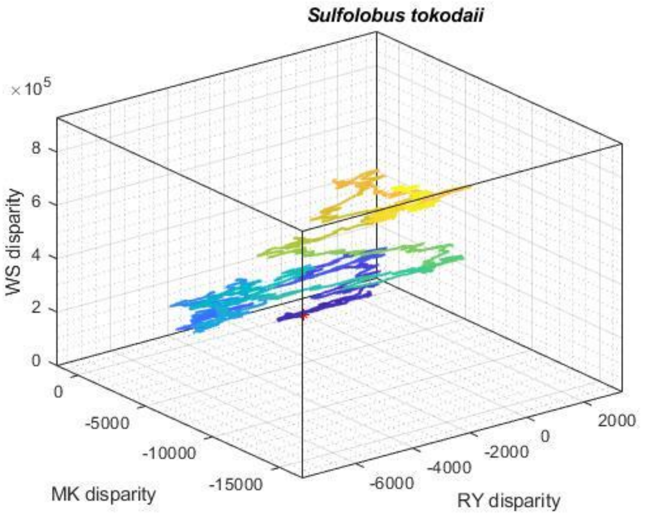

Neisseria meningitidis
×10^4
4
3
2
1
0
-1
-2
-3
-4
-5
Disparity
0
0.5
1
1.5
2
2.5
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA

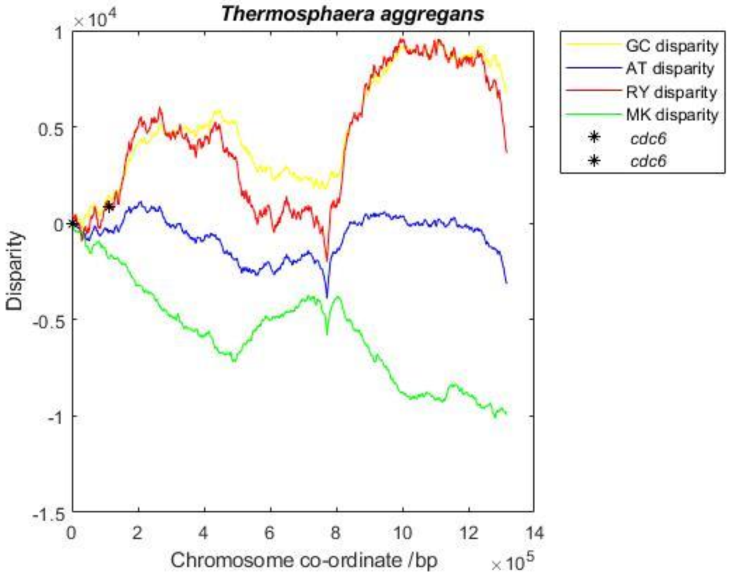

Neisseria meningitidis
×10^4
0
-2
-4
-6
WS disparity
0
-1
-2
-3
-4
×10^4
MK disparity
0
0.5
1
1.5
2
×10^4
RY disparity

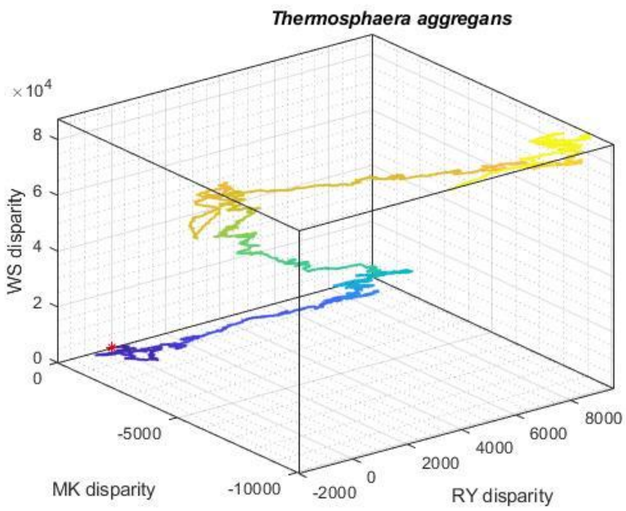

Desulfovibrio vulgaris
×10^4
8
6
4
2
0
-2
-4
-6
-8
-10
Disparity
0
1
2
3
4
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
dnaA
dnaA

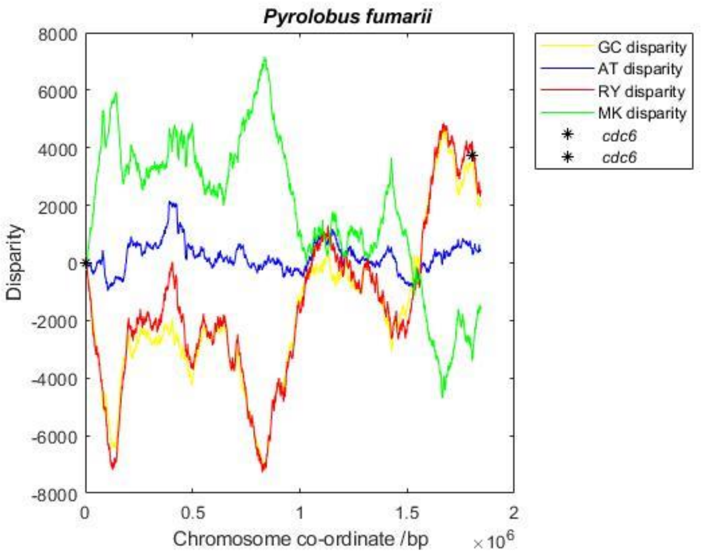

Desulfovibrio vulgaris
×10^5
0
-2
-4
-6
-8
WS disparity
0
-2
-4
-6
-8
×10^4
MK disparity
0
1
2
3
×10^4
RY disparity

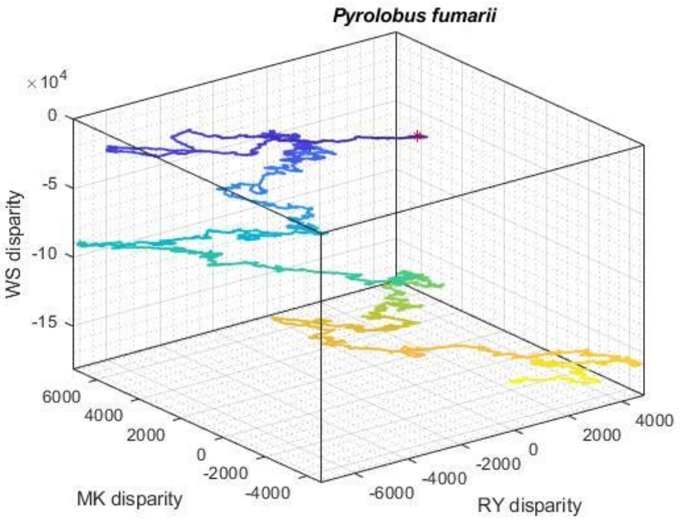

Sulfolobus tokodaii
×10^4
1
0.5
0
-0.5
-1
-1.5
-2
Disparity
0
0.5
1
1.5
2
2.5
3
Chromosome co-ordinate /bp
×10^6
GC dispar
AT dispari
RY dispar
MK dispar
cdc6
cdc6
cdc6

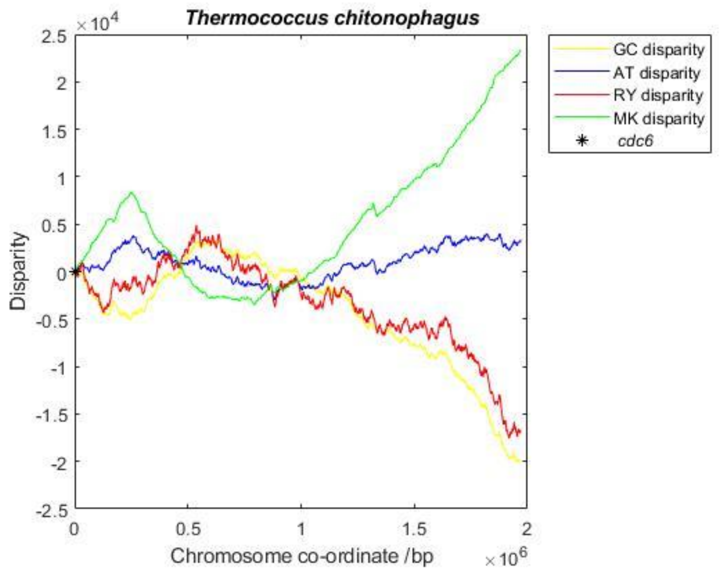

Sulfolobus tokodaii
×10^5
8
6
4
2
0
WS disparity
0
-5000
-10000
-15000
MK disparity
2000
0
-2000
-4000
-6000
RY disparity

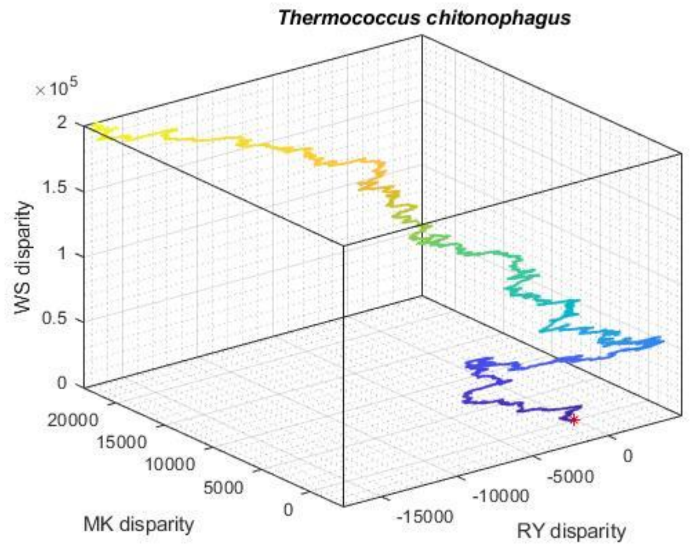

Thermosphaera aggregans
×10^4
1
0.5
0
-0.5
-1
-1.5
Disparity
0
2
4
6
8
10
12
14
Chromosome co-ordinate /bp
×10^5
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

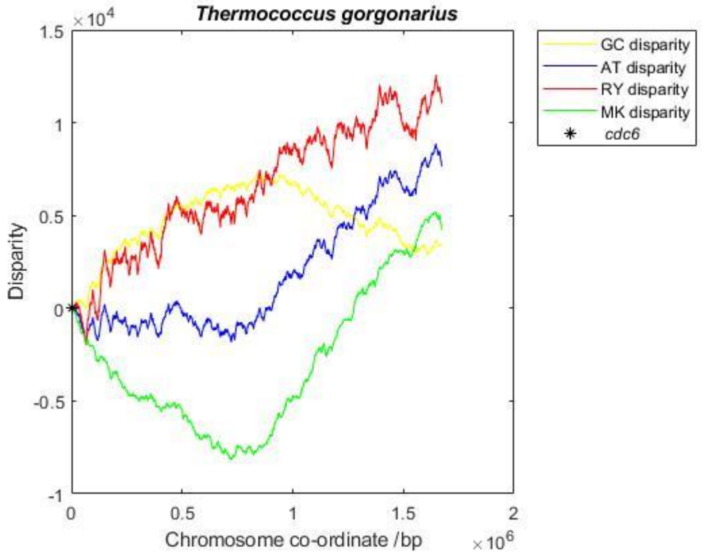

Thermosphaera aggregans
×10^4
8
6
4
2
0
WS disparity
0
-5000
-10000
MK disparity
8000
6000
4000
2000
0
-2000
RY disparity

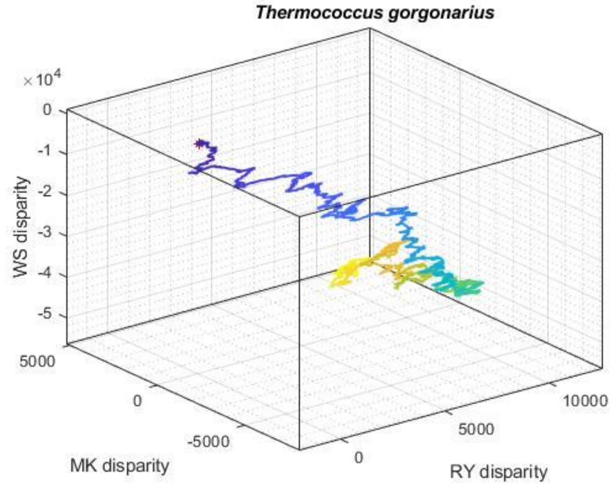

Pyrolobus fumarii
8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
×10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6
cdc6

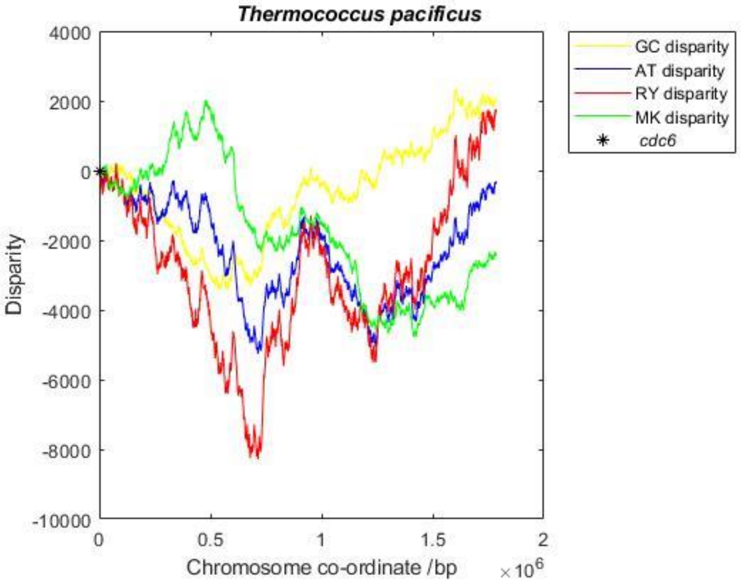

Pyrolobus fumarii
×10^4
0
-5
-10
-15
WS disparity
6000
4000
2000
0
-2000
-4000
MK disparity
4000
2000
0
-2000
-4000
-6000
RY disparity

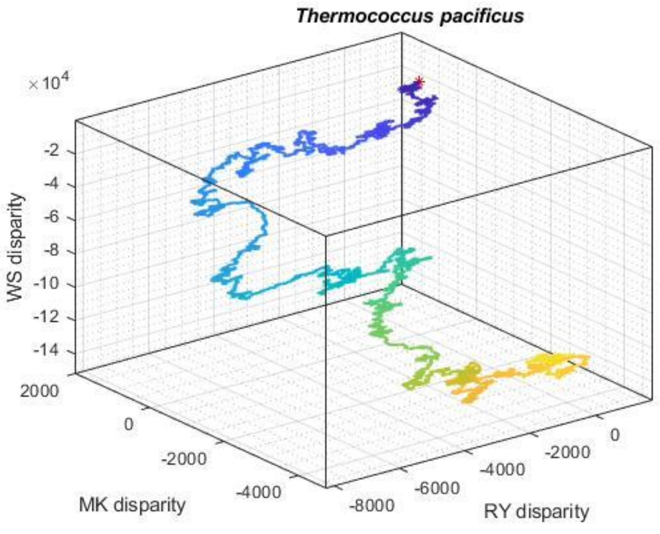

Thermococcus chitonophagus
× 10^4
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
× 10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6


Thermococcus chitonophagus
× 10^5
2
1.5
1
0.5
0
WS disparity
20000
15000
10000
5000
0
MK disparity
-15000
-10000
-5000
0
RY disparity


Thermococcus gorgonarius
× 10^4
1.5
1
0.5
0
-0.5
-1
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
× 10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6


Thermococcus gorgonarius
× 10^4
0
-1
-2
-3
-4
-5
WS disparity
5000
0
-5000
MK disparity
0
5000
10000
RY disparity


Thermococcus pacificus
4000
2000
0
-2000
-4000
-6000
-8000
-10000
Disparity
0
0.5
1
1.5
2
Chromosome co-ordinate /bp
× 10^6
GC disparity
AT disparity
RY disparity
MK disparity
cdc6


Thermococcus pacificus
× 10^4
-2
-4
-6
-8
-10
-12
-14
WS disparity
2000
0
-2000
-4000
MK disparity
-8000
-6000
-4000
-2000
0
RY disparity

MATLAB CODE

GC skew indices

```
%Import file into MATLAB
fastaread('56.fasta')
%Extract all CDS and transpose them
Sequence = {ans.Sequence}.';
%Join all CDS together
catcdsorg = horzcat(Sequence{:});
ormorg = {catcdsorg}
%This function calculates codon weights across the whole genome
strlength(one)
x = strlength(one)
x = sum(strlength(one))/4096
a = 1:4095;
startpoints = a*x
Y = round(startpoints)
startpoints = Y
b = 1:1
c = [b startpoints(:,1:4095)]
startpoints = c
endpoints = startpoints - 1;
endpoints(:,1) = []
e = sum(strlength(one))
f = [endpoints(:,1:4095) e]
endpoints = f
length(one)
normorg = cell(4096,1)
strlength(ormorg)
for i = 1:4096;
    newStr = extractBetween(ormorg,startpoints(i),endpoints(i));
    normorg(i) = newStr;
end

%Apply the above function to each of the 4096 genomic chunks
GCwinfun = @GCwindow
GCwin = cellfun(GCwinfun,normorg)
find(isnan(GCwin));
GCwin(isnan(GCwin))=0
%Calculate the cumulative sum of the GCwin values
cumGCwin = cumsum(GCwin);
%Calculate dist
dist = abs(max(cumGCwin))+abs(min(cumGCwin))
%Calculate fast Fourier transform of GCwin
GCfft = fft(GCwin);
%Calculate power spectrum
PS = abs(GCfft).^2;
%Calculate spectral ratio (first value is ignored)
SR = PS(2)/mean(PS(3:4096))
%Calculate GC skew index
GCSI = (SR/6000 + dist/600)/2

%Apply the above function to each of the 4096 genomic chunks
ATwinfun = @ATwindows
ATwin = cellfun(ATwinfun,normorg)
find(isnan(ATwin));
ATwin(isnan(ATwin))=0
F = fillmissing(ATwin,'constant',0)
%Calculate the cumulative sum of the ATwin values
cumATwin = cumsum(ATwin);
```

```
%Calculate dist
dist2 = abs(max(cumATwin))+abs(min(cumATwin))
%Calculate fast Fourier transform of ATwin
ATfft = fft(ATwin);
%Calculate power spectrum
PS = abs(ATfft).^2;
%Calculate spectral ratio (first value is ignored)
SR2 = PS(2)/mean(PS(3:4096))
%Calculate GC skew index
ATSI = (SR2/6000 + dist2/600)/2

%Apply the above function to each of the 4096 genomic chunks
RYwinfun = @RYwindows
RYwin = cellfun(RYwinfun,normorg)
find(isnan(RYwin));
RYwin(isnan(RYwin))=0
F = fillmissing(RYwin,'constant',0)
%Calculate the cumulative sum of the RYwin values
cumRYwin = cumsum(RYwin);
%Calculate dist
dist3 = abs(max(cumRYwin))+abs(min(cumRYwin))
%Calculate fast Fourier transform of RYwin
RYfft = fft(RYwin);
%Calculate power spectrum
PS = abs(RYfft).^2;
%Calculate spectral ratio (first value is ignored)
SR3 = PS(2)/mean(PS(3:4096))
%Calculate GC skew index
RYSI = (SR3/6000 + dist3/600)/2

%Apply the above function to each of the 4096 genomic chunks
RYwinfun = @MKwindows
RYwin = cellfun(MKwinfun,normorg)
find(isnan(MKwin));
RYwin(isnan(MKwin))=0
F = fillmissing(MKwin,'constant',0)
%Calculate the cumulative sum of the RYwin values
cumMKwin = cumsum(MKwin);
%Calculate dist
dist4 = abs(max(cumMKwin))+abs(min(cumMKwin))
%Calculate fast Fourier transform of RYwin
MKfft = fft(MKwin);
%Calculate power spectrum
PS = abs(MKfft).^2;
%Calculate spectral ratio (first value is ignored)
SR4 = PS(2)/mean(PS(3:4096))
%Calculate GC skew index
MKSI = (MK4/6000 + dist4/600)/2

%Summary of values
dist
SR
GCSI

dist2
SR2
ATSI

dist3
SR3
```

RYSI

dist4
SR4
MKSI

Nucleotide disparity graphs and z curves

```
function result=genomeplot(seq)
OPEN SPECIES FASTA FILE (eg. 87 CORRESPONDS TO Halogeometricum borinquense)
fastaread('87.fasta')
%Extract all CDS and transpose them
Sequence = {ans.Sequence}.';
%Join all CDS together
catcdsorg = horzcat(Sequence{:});
sequence = catcdsorg
% This function plots a Z-curve as described by Zhang and Zhang(2004)
% The shading uses code written by Walter Roberson source =
https://uk.mathworks.com/matlabcentral/answers/285872-shading-with-plot3
% This function also plots combined AT, GC, MK and RY
% disparity
% Create a zero matrix with 4 rows for the 4 bases and as many
% columns as there are bases
seqmat = zeros(4,length(Sequence))
% Scan along each base in the sequence, if a is present the value 1
% will be assigned to the first row, for c 1 to the second row etc. If
% the input sequence contains a character that is not a/c/t/g then an
% error message is displayed
result = true;
for c = 1:length(sequence)
switch lower(sequence(c))
case 'a'
    seqmat(1,c) = 1;
    case 'c'
        seqmat(2,c) = 1;
    case 'g'
        seqmat(3,c) = 1;
    case 't'
        seqmat(4,c) = 1;
    otherwise
        result = false;
end
end
if result == false
    error('An error has occurred - check all bases are a/c/g/t')
end
 %An = cumulative occurrence numbers of adenine bases
 An = cumsum(seqmat(1,:));
 %Cn = cumulative occurrence numbers of cytosine bases
 Cn = cumsum(seqmat(2,:));
 %Gn = cumulative occurrence numbers of guanine bases
 Gn = cumsum(seqmat(3,:));
 %Tn = cumulative occurrence numbers of thymine bases
 Tn = cumsum(seqmat(4,:));
 %Purine vs. pyrimidine distribution (RY)
 xn = (An+Gn)-(Cn+Tn);
 save('xn.mat','xn')
 %Amino vs. keto distribution (MK)
 yn = (An+Cn)-(Gn+Tn);
 save('yn.mat','yn')
 %Weak vs. strong hydrogen bond distribution (WS)
 zn = (An+Tn)-(Cn+Gn);
 save('zn.mat','zn')
 %AT disparity
 AT = (xn+yn)/2;
 save('AT.mat','AT')
 %GC disparity
 GC = (xn-yn)/2;
 save('GC.mat','GC')
 %Define n
 n = 1:length(sequence);

 %Create a 3D plot of xn,yn,zn (Z-curve)
  figure('Name','Z-curve')
  r = sqrt(xn.^2 + yn.^2 + zn.^2);
  g = patch('Vertices', [xn(:), yn(:),zn(:); nan nan nan], 'Faces',
(1:length(xn)+1).', 'FaceVertexCData', [r(:); nan], 'EdgeColor', 'interp',
'Marker','.','MarkerSize',0.001);
   hold on
   %SPECIES SPECIFIC PLOTTING OF ORC/CDC6
   plot3(-2,-12,-40,'r*')
   view(3)
  axis tight
  grid on
  grid minor
  box on
  ax = gca;
  ax.BoxStyle = 'full';
   %SPECIES SPECIFIC NAME
   title('\it               Halogeometricum borinquense')
   xlabel('RY disparity')
   ylabel('MK disparity')
   zlabel('WS disparity')
   %Plot GC, AT, RY and MK disparity on the same graph
    figure('Name','Combined disparity')
    plot(n,GC,'y','DisplayName','GC disparity')
    hold on
    plot(n,AT,'b','DisplayName','AT disparity')
    plot(n,xn,'r','DisplayName','RY disparity')
    plot(n,yn,'g','DisplayName','MK disparity')
    %SPECIES SPECIFIC MARKING OF CDC6/ORC
        plot(36836,0,'k*','DisplayName','\it Cdc6')
        hold off
        %SPECIES SPECIFIC NAME
        title('\it           Halogeometricum borinquense')
        xlabel('Chromosome co-ordinate /bp')
        ylabel('Disparity')
        legend('show','Location','northeastoutside')
end
```

COVID19 Impact Statement 2020
General Use

The University of Nottingham aims to support all our PGRs to complete their degrees within their period of registered study, by meeting our Doctoral Outcomes. We recognise, and aim to take into account, personal circumstances that may affect a PGR's ability to achieve this.

This Impact Statement should be used to record details and capture evidence of the impact that the COVID pandemic has had on your research progress for use in your annual review process, thesis examination and may be useful if you need to make a future request for an extension to your registered study as a result of the COVID pandemic.

We strongly encourage you to discuss the completion of this form with your supervisors. If you prefer, you can alternatively discuss the form with an appropriate member of PGR support staff such as your DTP/CDT Director or Manager, DTP/CDT Welfare Officer, School Postgraduate Student Advisor, School PGR Director or other member of the Welfare team, or the Researcher Academy Faculty Lead (formerly Associate Dean for the Graduate School).

To ensure that you cover the full impact of the COVID-19 pandemic on you and your research **since March 15th 2020**, please complete all relevant sections of the form. You can be very brief but please include all relevant information even in note or bullet form.

If you apply for an extension you will need to answer similar questions to those on this form and should find that you can draw in the responses you have captured in this document. You will need show how/whether your work to date already meets some of the University and QAA Doctoral Outcomes, and clarify which doctoral outcomes are not currently met and how your plan will enable you to meet these (Appendix 1).

Background Information – your details			
Family Name:	McCulloch	First Name(s)	Bryn James
ID:	20215139	School:	Life sciences
Please identify any relevant funder(s)/sponsors	n/a	Dates of impact: (the date from which the impact has had an effect).	March 2020
Start date	September 2019	Current end date	October 2020
Programme length (3, 3.5, 4 years) and full time or part time	1 year Full time		

The primary areas of impact:

Please tick all that are relevant for the ways in which you have been affected by the COVID pandemic and the resulting effect(s) on you and/or your research progression. You can give more details on these impacts, if you wish, on the next page.

Note: We will ask you to explain whether and how you have been able to manage or reduce any of these impacts in Section 2, on p.5.

The ways in which you have been affected (choose all that apply)

☐ additional/new caring responsibilities (including illness of someone for whom you are a carer)
☐ new illness, accident or hospitalisation, including any mental health problems
☒ being at higher risk of coronavirus
☒ increased anxiety and/or stress
☐ lack of access to mental health support (if needed);
☒ re-location
☐ death or illness of a partner/close relative*
☐ personal financial impact;
☐ impacts related to any protected characteristics*
☐ military or other service (e.g. NHS) that has not already been accommodated
☐ parental leave that has not already been accommodated
☐ redeployment to work in another area (e.g. COVID) where this has not already been accommodated.
☐ other events not on this list that are specifically related to the COVID pandemic (please describe below)

The ways in which your research activity has been affected
(for each that applies, please also indicate whether you have tried to mitigate the effect in this area).

	Was any mitigation possible?
☒ disruption/interruption of planned activities	No
☒ access to facilities/archives/lab/equipment/field sites etc	No
☒ postponement of critical activities where alternatives are not available	Yes/No
☐ access to other research resources including financial impact	Yes/No
☒ ability to achieve a planned outcome/ milestone/deliverable	Yes
☐access a research partner, including research-related placements	Yes/No
☐ an impact on your supervisory team that has affected your supervision or progress*	Yes/No
☐ other (please describe below)	Yes/No

*We are collecting this information in order to fully understand how you have been affected. Any information that you give here will only be used as information to inform us and will not be shared with anyone other than the teams considering the cases for extension and collating information for submission to UKRI.

1. DESCRIBING THE IMPACT

For example you could write a short clear description of the nature of the impacts or problems that you face/have faced, make making this description as brief, and specific as possible. You could also give more detail on the nature of the impacts on your research progress.

We understand that personal and research impacts will be related, so if it helps you could structure the content in line with the impacts you identified in the tick boxes above.

Section 1, additional guidance

The impact **on you**:
Due to being high risk, and because of the lab housing size under Covid-19 regulations I was therefore unable to return to the lab to finish activities. A new project was designed which was based on bioinformatics. This resulted in increased stress levels as I was unable to have face to face supervisor meetings to discuss issues with the project which required a steep learning curve, which is not the expertise of my supervisors to begin with. During this time I had to relocate due to housing issues which resulted in a brief period where I could not work.

The impact **on your research**:
Unable to access the laboratory to complete planned experiments and data collection. Which impacted on the ability to write up the project as results collected were less than expected. As I am high risk and due to the maximum number of people allowed within the lab after reopening, I was not able to return to finish the planned project. Hence, I could not achieve the planned outcomes of my masters. In an attempt to mitigate this a bioinformatics spin on the project was designed although coming up with a suitable and viable project took several weeks out of my time line and this project required me to learn a new set of bioinformatics skills which could not be fully supported by my supervisors who mostly work in a wet lab setting. The divide in research projects also made my thesis lack the coherency it may otherwise of had.

2. ACTIONS TAKEN TO MINIMISE THE IMPACT

a) How have you tried to mitigate the risk to your project?

Please **briefly** explain how you are trying/tried to minimise the impact of the situation on your research activities and progress. **With reference to the time between the COVID pandemic, national lockdown and the end of your registered period of study, if you have not tried to alter your plans to lessen the impact of this on your research progress, it's particularly important to explain here why you have taken/took this decision.**

For example,

- have you discussed how to do this with your supervisors?
- have you considered different ways to get the research done, such as changing your research plans to alter the order in which you do different elements?
- have you altered your research design, for example to conduct research online, or using other digital resources?
  what constraints or barriers did you have to try to remove, modify or overcome?
- **If you have not tried to alter your plans at all, why not?**

Try to show how/whether your work to date already meets some of the University and QAA Doctoral Outcomes, clarify which doctoral outcomes are not currently met and how your plan will enable you to meet these.
up to 200 words

Section 2 additional guidance

Alternative digital projects were discussed with my supervisors, after starting several ideas and finding out they would not be plausible due to the impact of Covid-19 on collaborators. A bioinformatics tool was decided on and a new set of objectives made. This involved the learning of numerous bioinformatic software's and statistical coding in order to complete the project which was time consuming and reduced the quantity of work able to be completed. Several short courses were completed to aid in this progress.

b) List the aspects of your research plan that you have managed to achieve or progress during the period of impact.

Original plan
- Create plasmid constructs
- Make strains
- Briefly test fluorescent readings of said strains
- Briefly test media conditions

3. NEXT STEPS

Please **list** what you have done/planned to do, in order to continue to lessen the impact on your research **once you are/were able to** resume the specific activities listed in Section 1

For example, what plans did you have to make sure that elements of your research that you have been unable to undertake due to the University closure restart quickly, or to efficiently complete the work you started during the closure?

up to 200 words

Section 3 additional guidance

Things intended for the original project which could not be done:
-confirm the fluorescent via further testing
-repeat southern blots for fluorescent strains
-Create a growth competition assay using flow cytometry of fluorescently tagged strains
-Use this assay to compare wild type and mutant strains in various nutrient deficient conditions, in search of a condition where the mutant performs worse than the wild type (currently it grows 7.5% faster)
- This would have utilised a bottom up and top down approach breaking down and building up on each component of the current laboratory media, assessing a huge range of conditions

-depending on the findings of this perform RNA Seq experiments to assess gene regulation in these strains in the deficient conditions compared to under normal lab conditions

-Assess Lhr helicase activity in *H.volcanii* under media deficient conditions using Mitomycin C assays.

Instead the project was change and the following was completed

- Learnt bioinformatics software such as MATlab and Rstudio aswell as online databases
- Adapt a previously created tool to suit large data sets of species
- Collect data on 5 different genetic factors on over 80 species
- Run through various bioinformatic pipelines
- Visualised these Factors in an appropriate way
- Complied all 5 factors in a principal component analysis to predict if they can survive without origins
  This has not been done before for many of the species, and several of factors analysed were also unknown for some species.

4. EVIDENCE

List any evidence that you have to demonstrate the impact you have detailed in section 1.

Please also provide here:

- a brief bullet list of the doctoral work completed prior to COVID-19 impact
- a revised research plan **that shows how the requested length of extension is justified by the work that remains to be done to enable you to meet the Doctoral Outcomes**;
- only if available, a previous work plan for comparison

up to 200 words

Section 4 additional guidance

Work completed prior

- Create plasmid constructs
- Make strains
- Briefly test fluorescent readings of said strains
- Briefly test media conditions

Revised project aims

- Learn how to use the previous bioinformatics prediction tool and software
- Perform predictions for a larger data set of species
- Adapt the tool to allow for more convenient use with a large amount of species
- Refine the tool based on suggested changes from previous work (Adlam, 2018)[83]
- Update the findings using the most current arCOG/COG data
- Predict the ability of a numerous species to survive without origins using the tool

☒ **I confirm that I have completed this form after/in discussion with**:
(indicate all those that apply, discussion with only one person is required)

☒ Primary supervisor/other supervisor ☐ SPSA ☐ School PGR Director ☐ DTP/CDT Director
☐ DTP/CDT Manager ☐ DTP/CDT Welfare Officer ☐ other member of the Welfare Team
☐ Researcher Academy Faculty Lead (RAFL, aka Associate Dean of the Graduate School)

RAFLs are: Prof A Grabowska (MHS), Dr L Bradnock (Arts), Prof R Graham (Science) and Dr N Porter (Eng), Prof. L. Cohen (Social Sciences)

Appendix 1.

University of Nottingham Criteria for award of PhD and other qualifications at Doctoral Level

(i) the creation and interpretation of new knowledge, through original research or other advanced scholarship, of a quality to satisfy peer review, extend the forefront of the discipline, and merit publication;
(ii) a systematic acquisition and understanding of a substantial body of knowledge which is at the forefront of an academic discipline or area of professional practice;

(iii) the general ability to conceptualise, design and implement a project for the generation of new knowledge, applications or understanding at the forefront of the discipline, and to adjust the project design in the light of unforeseen problems;
(iv) a detailed understanding of applicable techniques for research and advanced academic enquiry.

Typically, holders of the qualification will be able to:
(a) make informed judgements on complex issues in specialist fields, often in the absence of complete data, and be able to communicate their ideas and conclusions clearly and effectively to specialist and non-specialist audiences;
(b) continue to undertake pure and/or applied research and development at an advanced level, contributing substantially to the development of new techniques, ideas, or approaches; and will have:
(c) the qualities and transferable skills necessary for employment requiring the exercise of personal responsibility and largely autonomous initiative in complex and unpredictable situations, in professional or equivalent environments.

Appendix 2.

Additional Guidance notes.

What to include:

Section 1, Describing the impact. Please limit the information on this form to impacts that have occurred, and only extend this forwards to future impacts that can be predicted to result from current impacts. If future plans might be disrupted you should show how you plan to adjust the project or use other means to mitigate the risk that this presents. This form will continue to be available on the R&I sharepoint or through the Graduate School and you can use it if needed to record longer-term or future impacts of COVID-19 on your work over the coming months.

Please do not feel that you have to write a large amount in any of the sections of this form. Your statement of impact can be brief and to the point, please see the sample form also available to view alongside this form.

Please only include research activities that you had planned to undertake during the Lockdown/University Closure, and the periods immediately before and after this, if relevant. For example, if you had planned a period of research activity at another organisation before or after lockdown that has had to be cancelled, or postponed and cannot be rescheduled within your registered period of study.

Section 16 of the Postgraduate Regulations describes the usual acceptable and unacceptable circumstances for extensions

16. Acceptable and Unacceptable Circumstances (for extension to Thesis Pending):

The following circumstances may result in an extension being granted:

- Exceptional personal circumstances (eg illness, hospitalisation, accident) if significantly impacting on the writing-up process (or resubmission/minor corrections process relating to paragraph 37 below)
- Maternity
- Paternity
- Death of a close relative, or illness of a close relative where the student is the carer
- Illness or death of a partner
- Prolonged jury service
- Expeditions for sport of national significance (providing the extension is acceptable to the student’s funding body)
- Requirement for a student to undertake military service.

The following are examples of circumstances which would not normally warrant an extension:

- Taking up employment during the thesis pending period (or resubmission/minor corrections process relating to paragraph 37 below)
- voluntary service overseas.

Section 2, Action taken. Please list the people with whom you have discussed your research plans and what advice and support you have had in adjusting your activities to mitigate any risk to the progress of your research. You are not obliged to consult or discuss the completion of this form with your supervisors, but we encourage you to do so, before finalising the form. Include if and how your plans have changed as a result of either these discussions or your own planning.

It may be that you feel that you have experienced COVID-related impacts on your research but you have decided not to alter your research plans in any way. If this is the case, we would like to understand the reasons why you have decided that this is the best course of action for you.

Please also detail the things that you have managed to achieve or move forwards under the current conditions, even if you feel that you haven't managed to achieve as much as you planned. Please show how your achievements relate to your previous and future research plans.

Section 3, next steps. It's important to plan both how to deal with a current or emergent situation that disrupts your research, and also how to get back into 'normal' working once you are able to do so. These plans should include how you will get everything back on track, getting started and up and running as quickly as possible. What can/could you be doing now to make sure there are no added delays in resuming 'normal' activity?

If there is anything that is still presenting you with a problem, and that is likely to continue to be a problem once things change, please record it here. Give information on why this might be an ongoing concerns and give brief information on discussions you have had to try and solve the problem.

Section 4 Other (please specify below)**, documents and evidence:** We advise you to support your case with evidence wherever possible, but we recognise that there may be circumstances in which evidence is not available to you. Under such circumstances please explain the case in a way that includes the reason why you cannot provide supporting evidence.

Your future/revised plans do not need to be complicated, nor in Gantt chart form unless this is a planning method that you already use. A simple table of milestones, deadlines, and outputs is sufficient.

Privacy and confidentiality: We encourage everyone to discuss the information contained in the form, and its completion with a member of the PGR support staff in the University, particularly with your supervisors. We do however recognise that there may be aspects of this form that you might wish to keep confidential, and so you could alternatively discuss things with your SPSA, your School PGR Director(s), your DTP/CDT Director, Manager or Welfare Officer, or if none of these other supports available to you is appropriate, the Researcher Academy Faculty Lead (Arts - Dr L Bradnock, Science - Prof R Graham, MHS - Prof A Grabowska, Engineering - Dr N Porter, Social Science – TBC).

For use in thesis assessment: We suggest that you save a copy of this form, with any confidential material redacted, and include it with your submitted thesis, as a record of how you have managed and mitigated the impact of the COVID pandemic on your achievements during this time.

The Researcher Academy Faculty Leads are the Faculty representatives with responsibility for our PGRs. They have oversight of PGR support and activities at Faculty level, and they also work closely with the Graduate School/Researcher Academy. They can advise and support you in completing this form, if there is no-one else that you feel comfortable with, in sharing this information. They should not however be the first person that you approach, as it would be best to discuss this with someone that you know and who knows you, if possible.

The Researcher Academy Faculty Leads are: Prof A Grabowska (MHS), Dr L Bradnock (Arts), Prof R Graham (Science) and Dr N Porter (Eng), Prof. L Cohen (Social Sciences)