*Our research*
*– your profit*

matís

# ComGen software for comparative genomics

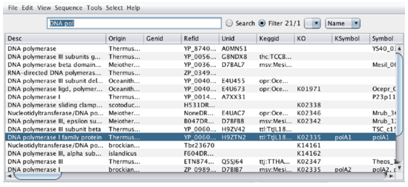
Sigmar Karl Stefánsson[1,2], Snædís Björnsdóttir[1,2], Ólafur H. Friðjónsson[1], Sólveig Pétursdóttir [1] and Guðmundur Ó. Hreggviðsson[1,2]

[1] Matís, Reykjavík, Iceland, [2] Faculty of Life and Environmental Sciences, University of Iceland, Reykjavík Iceland

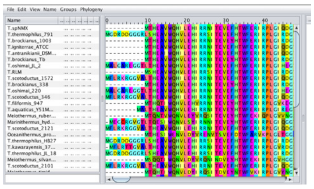
Horizontal gene transfer is a major force of bacterial evolution. We have performed extensive comparative genomic analysis of thermophilic bacteria that belong to the genus *Thermus*. These bacteria are very effective in DNA uptake and genetic elements such as plasmids, transposons and prophages were found at many of the numerous shuffling sites within their genomes. Genome rearrangements were examined using a new software which identifies orthologs and uses both paralog clustering and novel visualization techniques to extract information from massive sequence data. The software identifies the most likely order of contigs of assembled sequences and predicts genetic islands using conventional methods and genome comparison. Analyses of synteny and gene presence-absence as well as of the convergent evolution of duplicated genes are used for the construction of genome phylogenies. Gene loss and gene gain are estimated on the basis of observed geographical and environmental patterns in the phylogeny of individual genes. Metabolic mapping in the context of multiple genomes give the gene sets required for core cellular functions and potentially a deeper insight into the impact of slight genome differences. The software is written in Java and is downloadable as such. It also has a web interface with public and private access. Our results from genome comparison within the genus *Thermus* exemplify how genome rearrangement occur in nature.

## Main screen

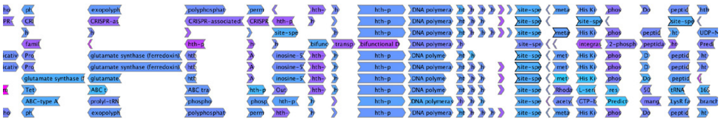
The main screen is split into two tables, the above gene table and the below function table. The tables are linked and double-clicking on an entry highlights the corresponding row in the other table. A search textbox accompanies both tables for easy search and filtering. The function table allows for the selection of specific metabolic pathways. The resulting selection can consequently be viewed in a gene atlas. The sequence editor supports different color coding, multiple sequence alignment and the generation of phylogenetic trees. The gene aligner view shows the difference of small scale regions between strains.



| Desc | Origin | Genid | Refid | Unid | Keggid | KO | KSymbol | Symbol |
|---|---|---|---|---|---|---|---|---|
| DNA polymerase | Thermus... | | YP_8740... | A0MN51 | | | | YS40_0 |
| DNA polymerase III subunits g... | Thermus... | | YP_0056... | G8NDX8 | thc:TCCB... | | | |
| DNA polymerase beta domain... | Meiother... | | YP_0036... | D7BAL7 | msv:Mesi... | | | Mesil_0 |
| RNA-directed DNA polymeras... | Thermus... | | ZP_0349... | | | | | |
| DNA polymerase III subunit del... | Oceanith... | | YP_0040... | E4U455 | opr:Oce... | | | |
| DNA polymerase ligd, polymer... | Oceanith... | | YP_0040... | E4U673 | opr:Oce... | K01971 | | Ocepr_ |
| DNA polymerase I | Thermus... | | YP_0014... | A7XX31 | | | | P23p11 |
| DNA polymerase sliding clamp... | scotoduc... | | H531DR... | | | K02338 | | |
| Nucleotidyltransferase/DNA po... | Meiother... | | NoneDR... | E4UAC7 | opr:Oce... | K02346 | | Mrub_3 |
| DNA polymerase III, epsilon su... | Meiother... | | B047DR... | D7BFB8 | msv:Mesi... | K02342 | | Mrub_1 |
| DNA polymerase III subunit beta | Thermus... | | YP_0060... | H9ZV42 | ttl:TtJL18... | | | TSC_c1 |
| DNA polymerase I family protein | Thermus... | | YP_0060... | H9ZTN2 | ttl:TtJL18... | K02335 | polA1 | polA1 |
| Nucleotidyltransferase/DNA po... | brockian... | | Tbr23670 | | | K14161 | | |
| DNA polymerase III, alpha sub... | islandicus | | F604DR... | | | K14162 | | |
| DNA polymerase III | Thermus... | | ETN874... | Q5SJ64 | ttj:TTHA... | K02347 | | Theos_1 |
| DNA polymerase I | brockian... | | ZP_0989... | D7BIB7 | msv:Mesi... | K02335 | polA2 | polA2, |

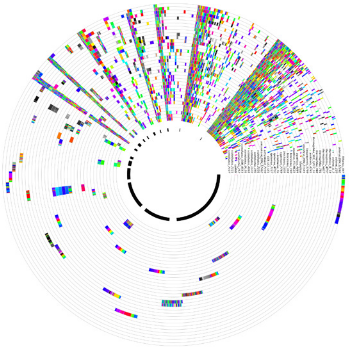


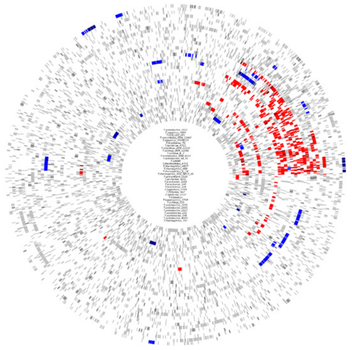
T.spNMX
T.thermophilus_791
T.brockianus_1003
T.igniterrae_ATCC
T.antranikianii_DSM...
T.brockianus_Tb
T.oshimai_JL_2
T.RLM
T.scotoductus_1572
T.brockianus_338
T.oshimai_220
T.scotoductus_346
T.filiformis_947
T.aquaticus_Y51M...
Meiothermus_ruber...
Marinithermus_hyd...
T.scotoductus_2121
Oceanithermus_pro...
T.thermophilus_HB27
T.kawarayensis_37...
T.thermophilus_JL_18
Meiothermus_silvan...
T.scotoductus_2101



## Genetic islands

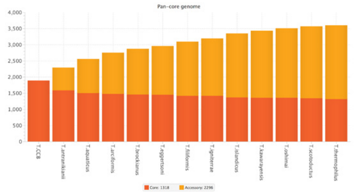
The gene permutation atlas is used to view sharing of genes between strains. The core genome has been removed and the genes are ordered according to the size of each permutation group. The image to the right shows that the largest group of unique genes within the genus *Thermus* is observed in *Thermus filiformis*. The largest group of genes shared between more than one strain is a prophage found in *Thermus brockianus* and "Thermus eggertssonii". The synteny colors shows the relative gene locations. The image to the right shows the same genes in their correct order for each species, which reveals genetic islands as lanes in the circle. Prophage genes are colored blue and plasmid genes are colored red.

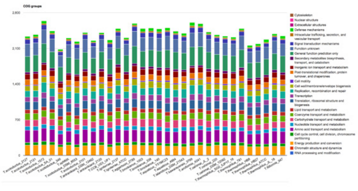




## Comparative genomics

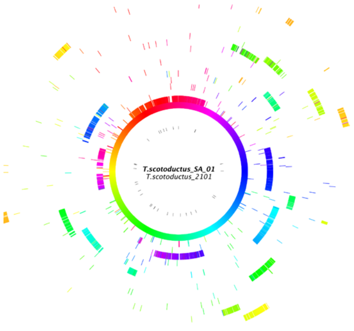
Genomes inserted into the software are automatically annotated. The genes are grouped together using criteria fitting the data, such as the 50-50 rule for inter-genus genomes. The results are stored in an easily portable zip formatted file. The software supports automatic and manual connection of contigs based on comparative data. The software generates basic comparative genome statistics such as a pan-core graphs, GC%, GC skew, COG-groups and genome size column plots.

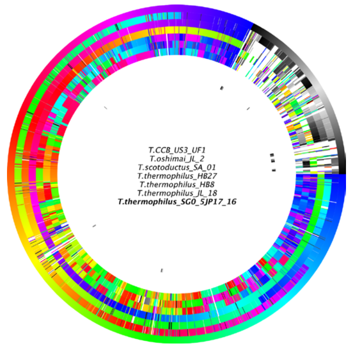




## Synteny gradients

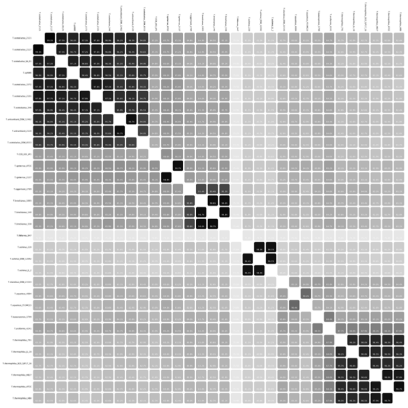
The synteny gradient colors are used to visualize genome rearrangements. Plasmids are shown in grayscale. The starting point is on the right (blue) indicating the *ori*. The image to the left shows each contig as intact lanes, where uniform synteny color indicates that the order of the genes is preserved between the two strains. The image to the right shows four *Thermus thermophilus* genomes and three genomes from other *Thermus* species. The inverted areas could reveal errors in the published genomes. The smaller case (shown with an arrow) shows a region flanked by 23S region on both sides.

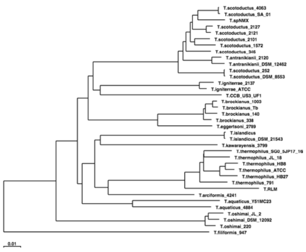




## Phylogeny

Multiple genes within and between genomes are compared using phylogenetic methods such as sequence concatenation and majority rule consensus. Gene duplication sites are used to support the prediction for rooting phylogenetic trees. The ANI matrix shows the average nucleotide identity between all species in a phylogenetic order. The ANI matrix shows that bacterial species are not always clearly defined. An ANI value of 90% was calculated for two strains of Thermus aquaticus while ANI values of at least 95% were calculated for strains that belonged to Thermus antranikianii and Thermus scotoductus. The phylogenetic tree viewer supports circular and radial trees and color coded nodes. The web link is http://webconnectron.appspot.com/Treedraw.html




T.scotoductus_4063
T.scotoductus_SA_01
T.spNMX
T.scotoductus_2127
T.scotoductus_2121
T.scotoductus_2101
T.scotoductus_1572
T.scotoductus_346
T.antranikianii_2120
T.antranikianii_DSM_12462
T.scotoductus_252
T.scotoductus_DSM_8553
T.igniterrae_2137
T.igniterrae_ATCC
T.CCB_US3_UF1
T.brockianus_1003
T.brockianus_Tb
T.brockianus_140
T.brockianus_338
T.eggertsoni_2789
T.islandicus
T.islandicus_DSM_21543
T.kawarayensis_3799
T.thermophilus_SG0_5JP17_16
T.thermophilus_JL_18
T.thermophilus_HB8
T.thermophilus_ATCC
T.thermophilus_HB27
T.thermophilus_791
T.RLM
T.arciformis_4241
T.aquaticus_Y51MC23
T.aquaticus_4884
T.oshimai_JL_2
T.oshimai_DSM_12092
T.oshimai_220
T.filiformis_947
0.01