

Assignation taxonomique et détection de chimères dans les données génomiques complexes et les métagénomomes: prototype d'un outil interactif

Sommaire

I. Introduction

- 1) Problématique : Les données métagénomiques
- 2) Choix techniques

II. Présentation du programme

- 1) Vue d'ensemble
- 2) Explication des fonctions

III. Réponses aux problèmes biologiques

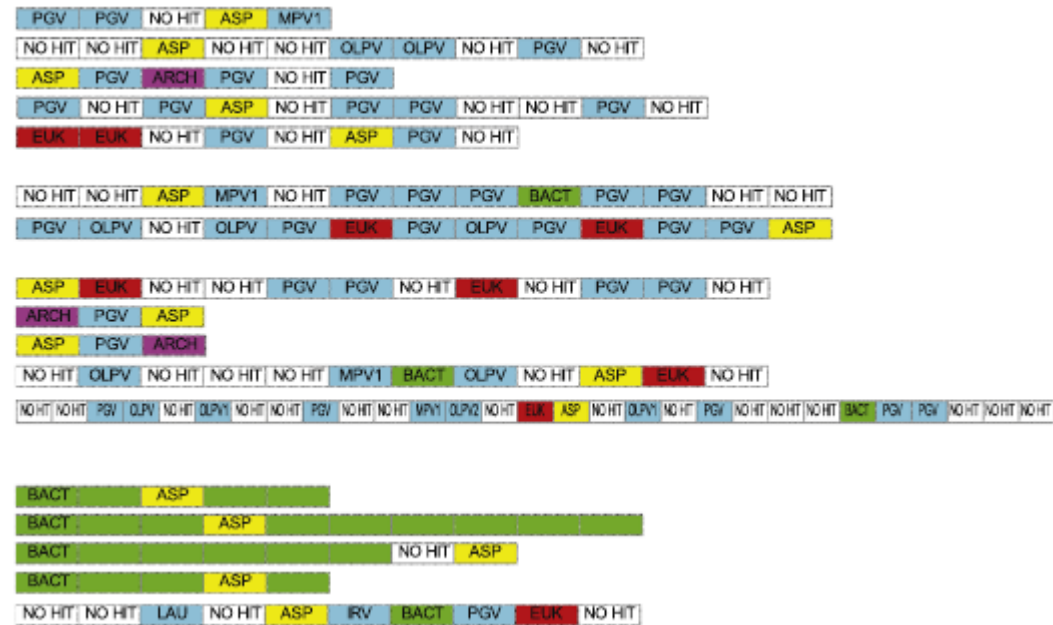
- 1) Assignation taxonomique erronée
- 2) Détection du chimerisme

Problématique

Données de génomique bruitées

Séquençages de cellule unique entraînent souvent des problèmes d'annotation

Projet : Washingtonvirus



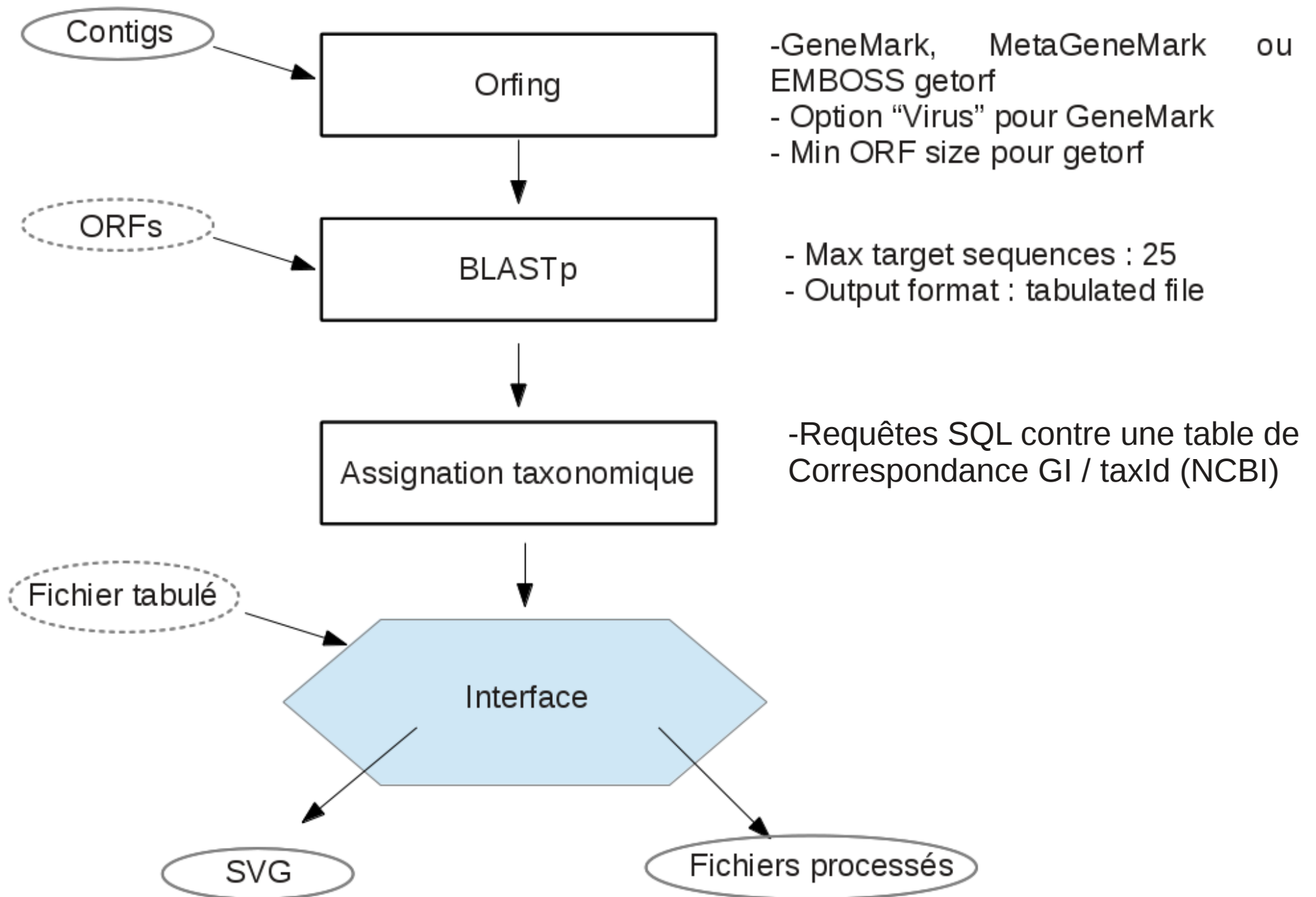
Données de métagénomique

Microbiome environnementaux :

Nombreuses espèces différentes (symbiontes, parasites, ADN libre dans le milieu)

Projet : Spongex

Traitement des données



Choix techniques

Langage : Perl (module CGI)

- Langage très efficace pour les traitements de chaînes de caractères
- Possibilité de faire des pages web simplement
- Nombreuses bibliothèques et liens entre langages

SVG

- Langage de dessin vectoriel en XML
- Flexible
- Contient des fonctions natives : interactivité

Vue d'ensemble

Fouille Métagénomique - display

Contigs : Washington-All_cleaner.fna	
Format:	FASTA
Alphabet type:	DNA
Number of sequences:	113
Total # residues:	534847
Smallest:	503
Largest:	100224
Average length:	4733.2

--- GeneMark --->

ORFs :	
Format:	FASTA
Alphabet type:	amino
Number of sequences:	505
Total # residues:	149576
Smallest:	99
Largest:	2453
Average length:	296.2

Statistiques sur les séquences

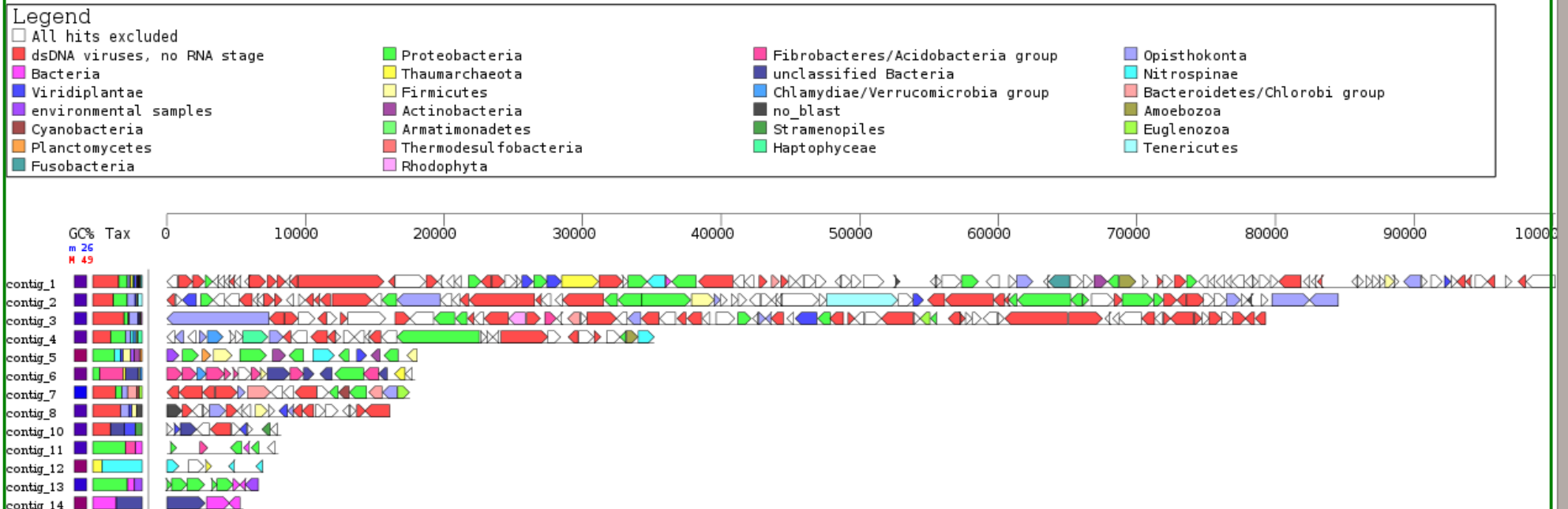
Paramètres

Display Options Strand display Single strand Zoom x 1 Display mode Real size GC Percentage <input type="checkbox"/> Regroup ORFs of same color <input checked="" type="checkbox"/> Scale <input checked="" type="checkbox"/> Full contig names	Output Screen size 1280 Title Washington-All Color set : 0 Contig height : 10 Padding : 5 Save: Save display Save workspace Url Pie charts : Google chart Krona tools Flag	Filter Exclude from blast results Euryarchaeota <input type="checkbox"/> Complement e-value < 1.e- 5 Idt < 99.9 % Exclude n first hits 0 Contigs to mask contig_9 <input type="checkbox"/> Complement	Taxonomic depth 2 <input type="checkbox"/> Display hits with no taxonomy
---	--	---	--

Default settings
[Refresh](#) [New file](#) [Main page](#)

Washington-All

Sortie graphique



Détails de l'affichage

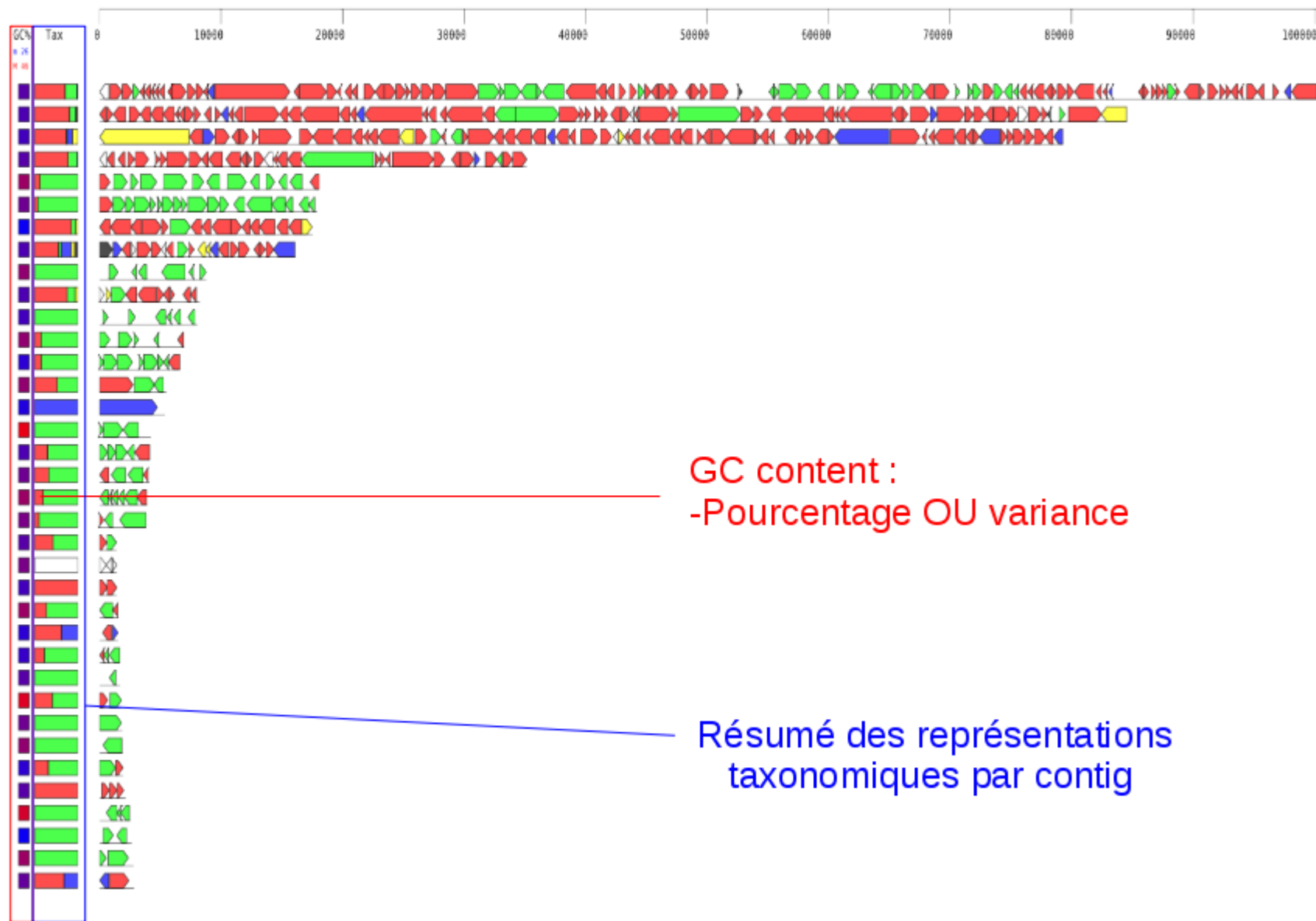
Washington-All



Légende

Échelle

Noms
des
contigs



Mode d'affichage :
Taille arbitraire

Display Options

Strand display Single strand ▾

Zoom x 1 ▾

Display mode Real size ▾

GC Percentage Arbitrary Real size Super compacted

☐ Regroup O

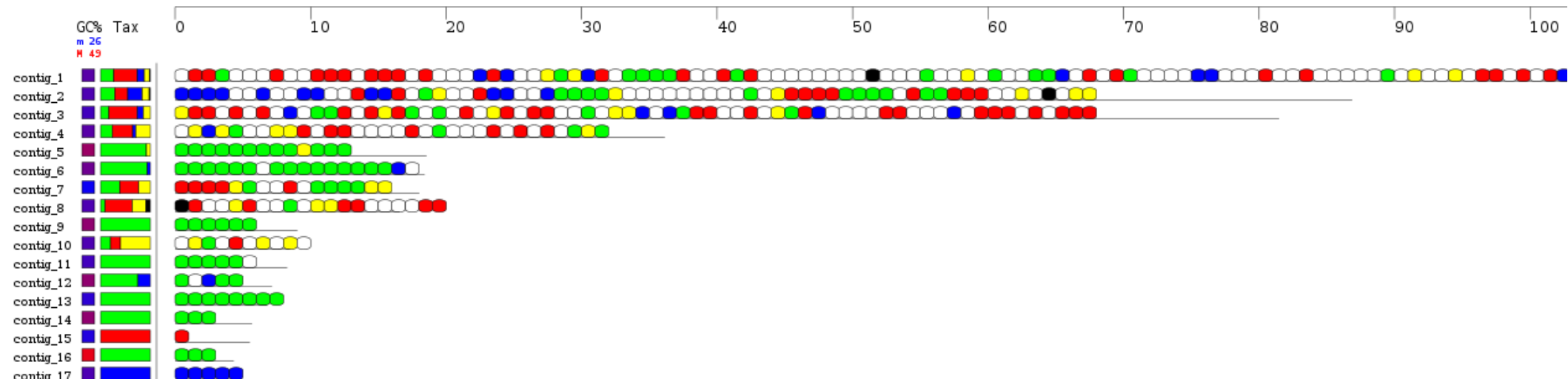
☒ Full contig names

Washington-All

Legend

☐ All hits excluded

☒ Bacteria ☒ Viruses ☒ Archaea ☒ Eukaryota ☒ no_blast



Mode d'affichage : Histogramme

Display Options

Strand display Single strand ▾

Zoom x 1 ▾

Display mode Real size ▾

GC Percentage ▾

☒ Regroup ORFs of same color

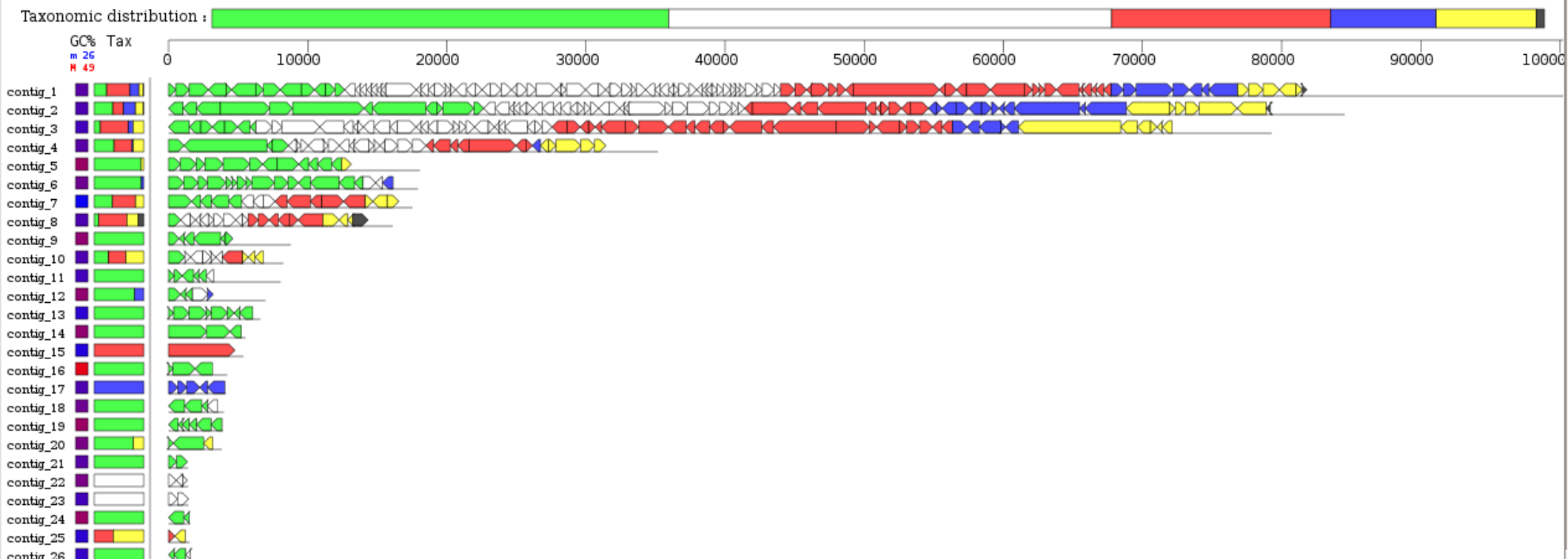
☒ Full contig names

Washington-All

Legend

☐ All hits excluded

☒ Bacteria ☒ Viruses ☒ Archaea ☒ Eukaryota ☒ no_blast



Options graphiques

Output

Screen size

Title

Color set : ▼

Contig height : Padding :

Save: [Save display](#) | [Save workspace](#) | [Url](#)

Pie charts : [Google chart](#) | [Krona tools](#)

Washington-All

Legend

☐ All hits excluded

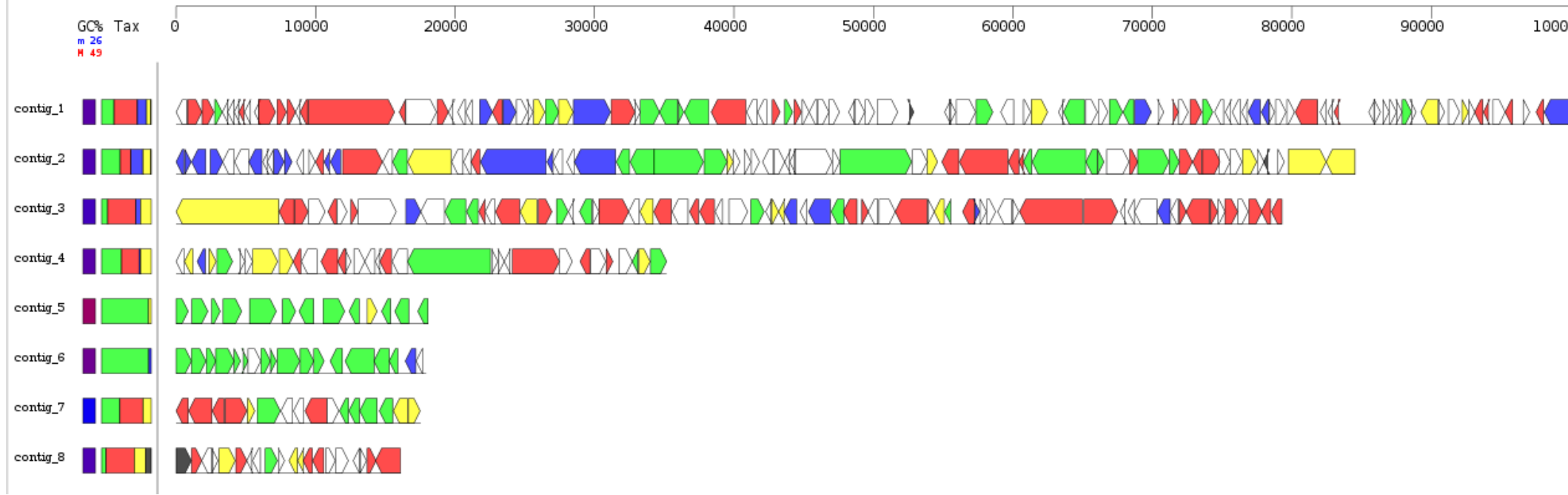
☒ Bacteria

☒ Viruses

☒ Archaea

☒ Eukaryota

☒ no_blast



Filtres :
Taxonomie, pourcentage d'identité
et e-value

Filter

Exclude from blast results

☐ Complement

e-value < 1.e- Idt < %

Exclude n first hits

Contigs to mask

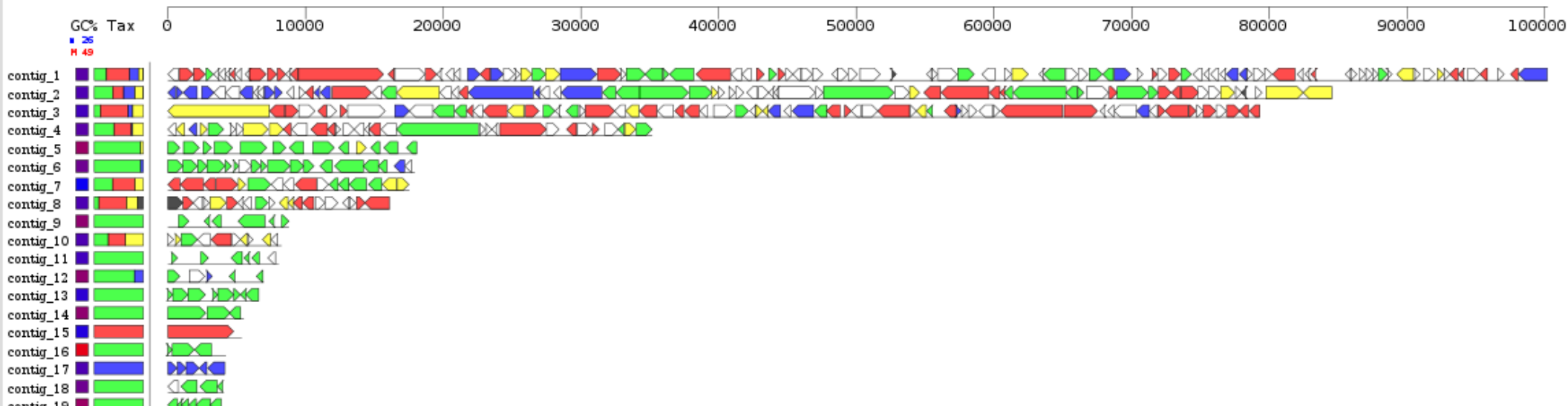
☐ Complement

Washington-All

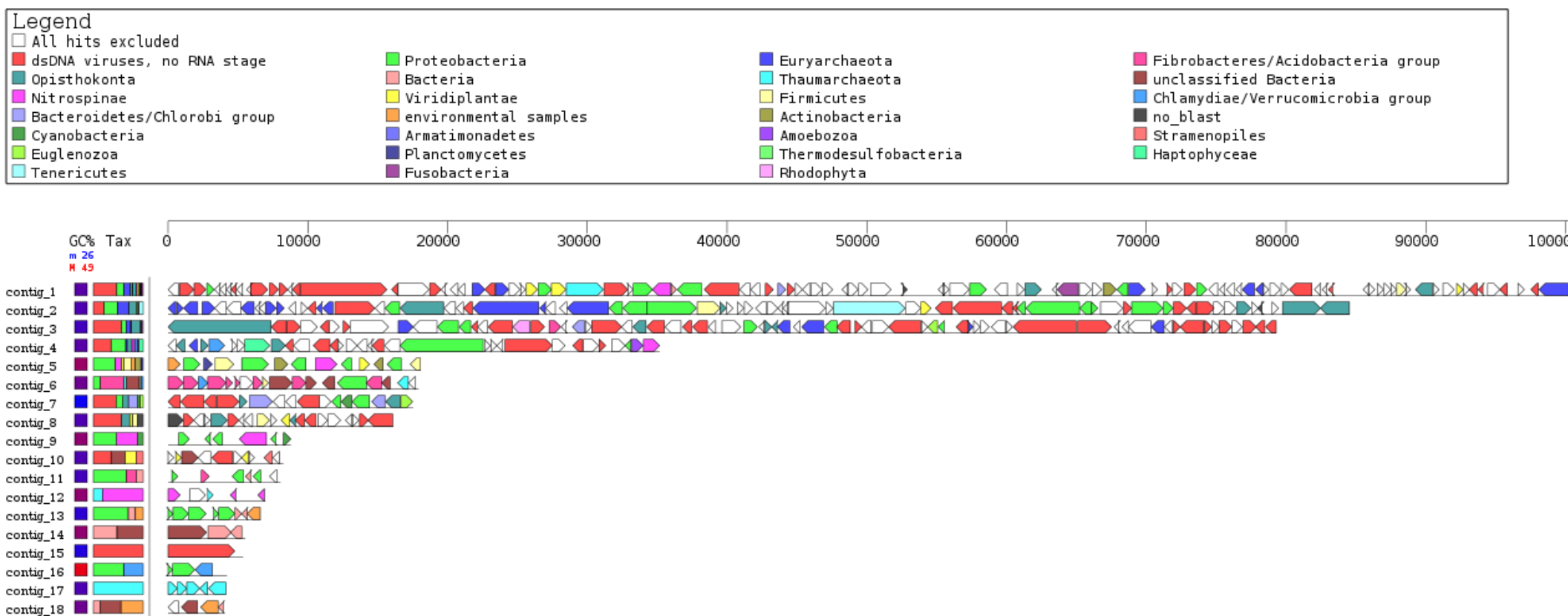
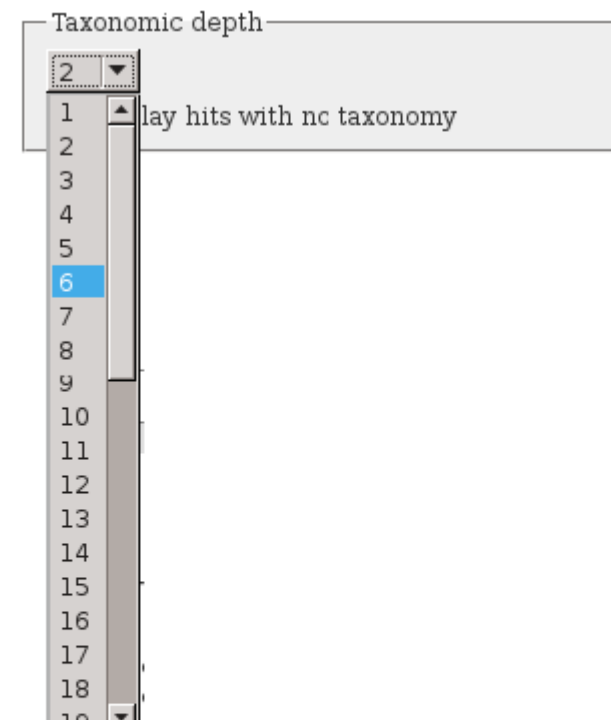
Legend

☐ All hits excluded

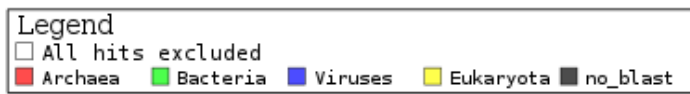
☐ Bacteria ☐ Viruses ☐ Archaea ☐ Eukaryota ☐ no_blast



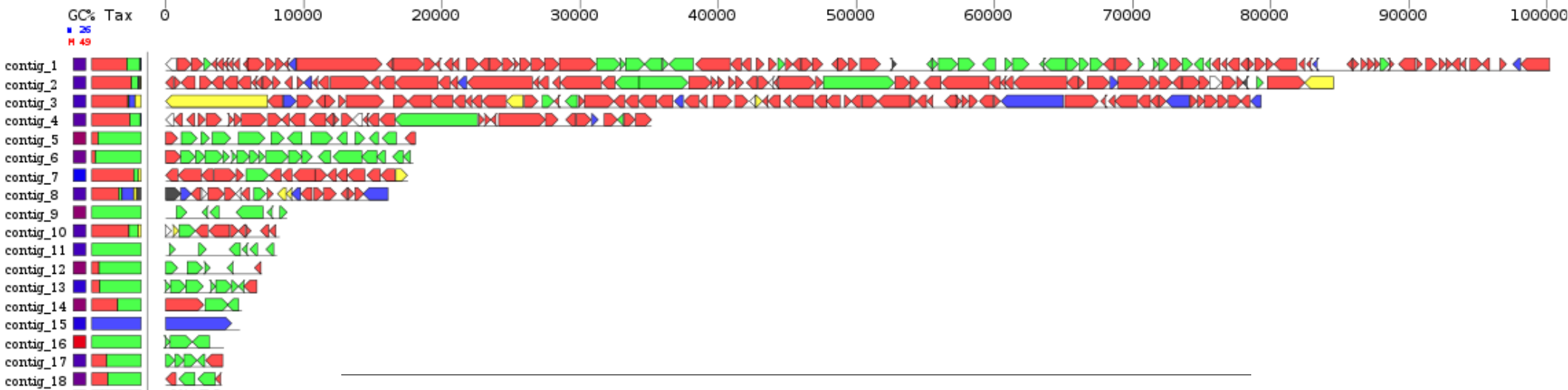
Choix de la profondeur taxonomique



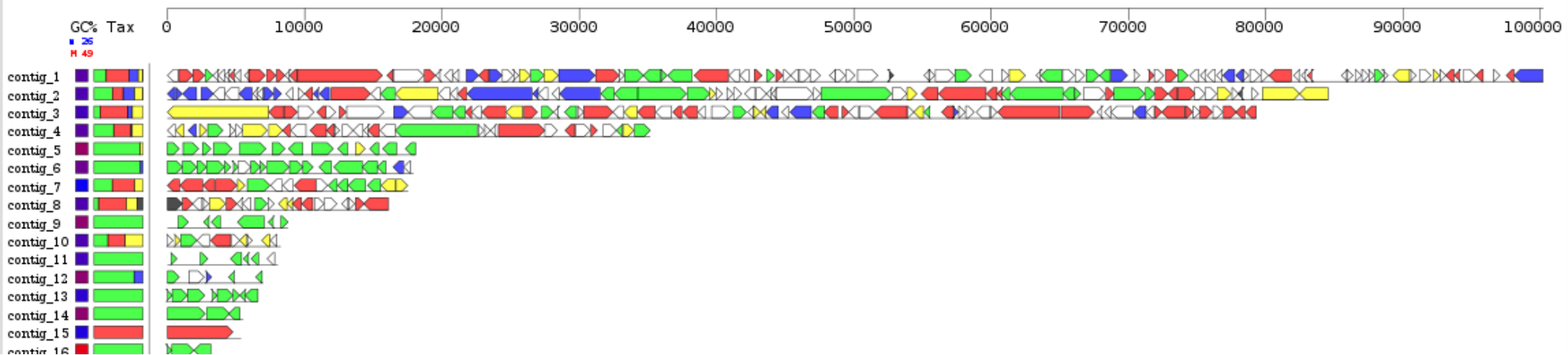
Assignation taxonomique erronée : Utilisation du filtre



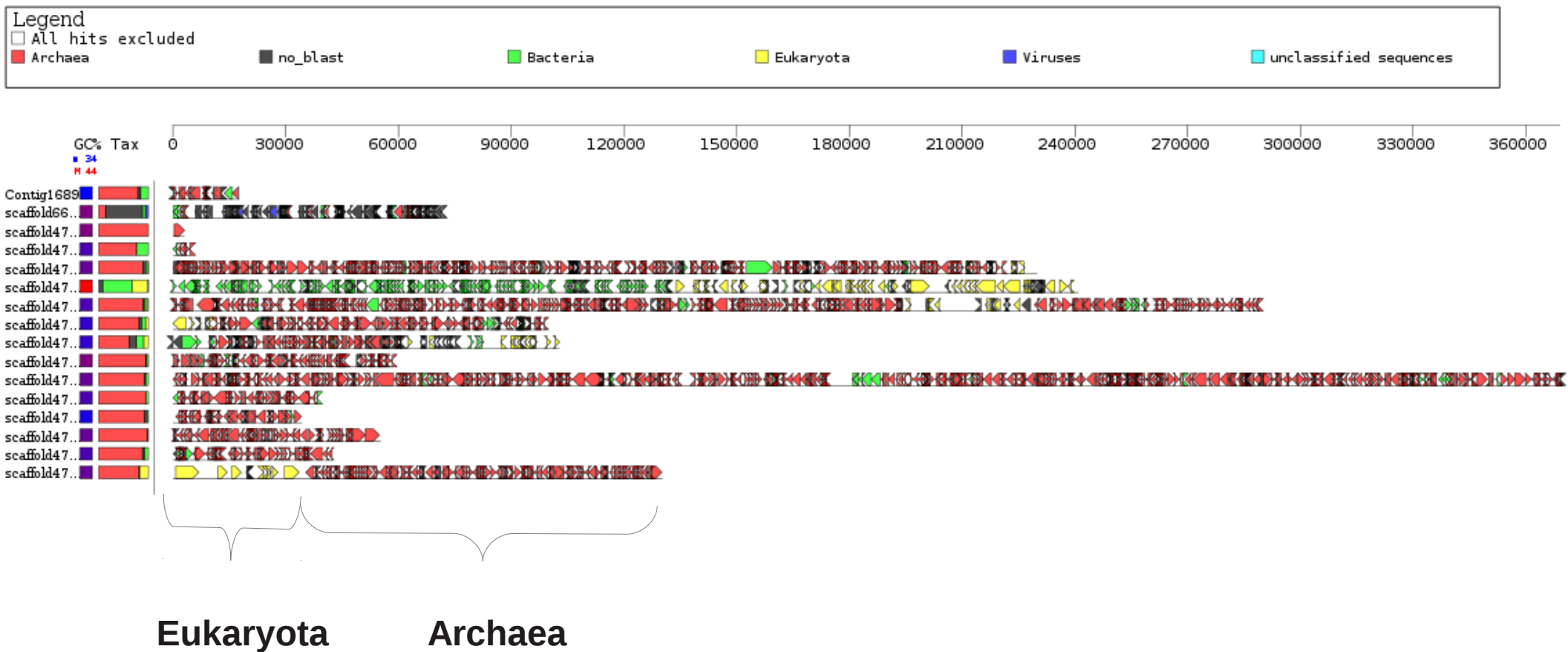
Affichage par défaut



Application du filtre sur 100% d'identité



Détection des chimères



Conclusion

Etat du projet

- Traitement des données : ~25 minutes pour 505 ORFs (sur 32 noeuds de 32go de RAM)
- Phase alpha terminée pour les fonctions existantes
- Nécessité de retours par les utilisateurs (phase beta)

En projet

- Algorithme plus complexe pour assigner une taxonomie à chaque contig
- Graphisme de représentation taxonomique pour chaque contig
- Tri des contigs

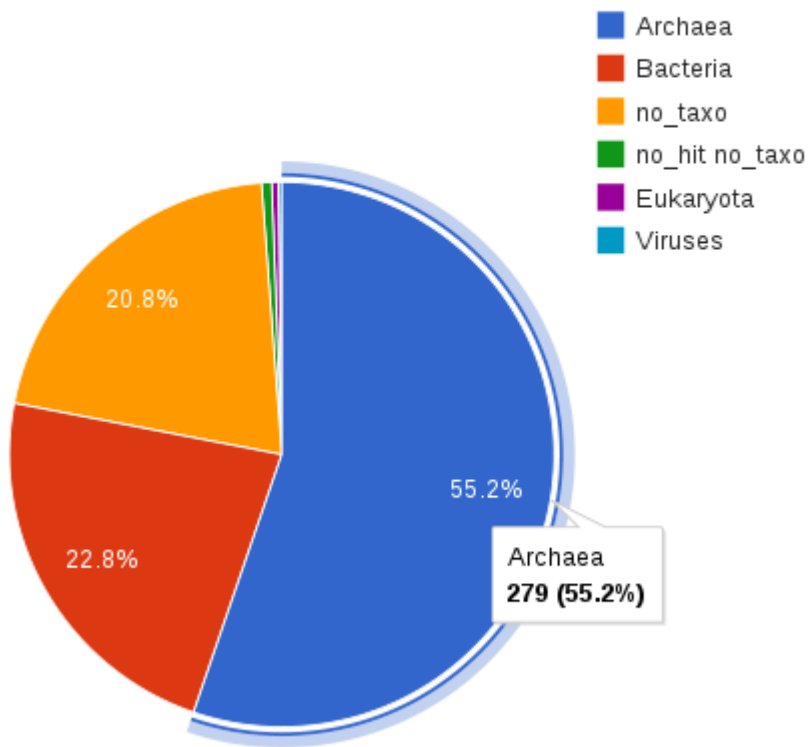
Remerciements

- France génomique
- Jean-Michel Claverie
- Olivier Poirot
- Sebastien Santini

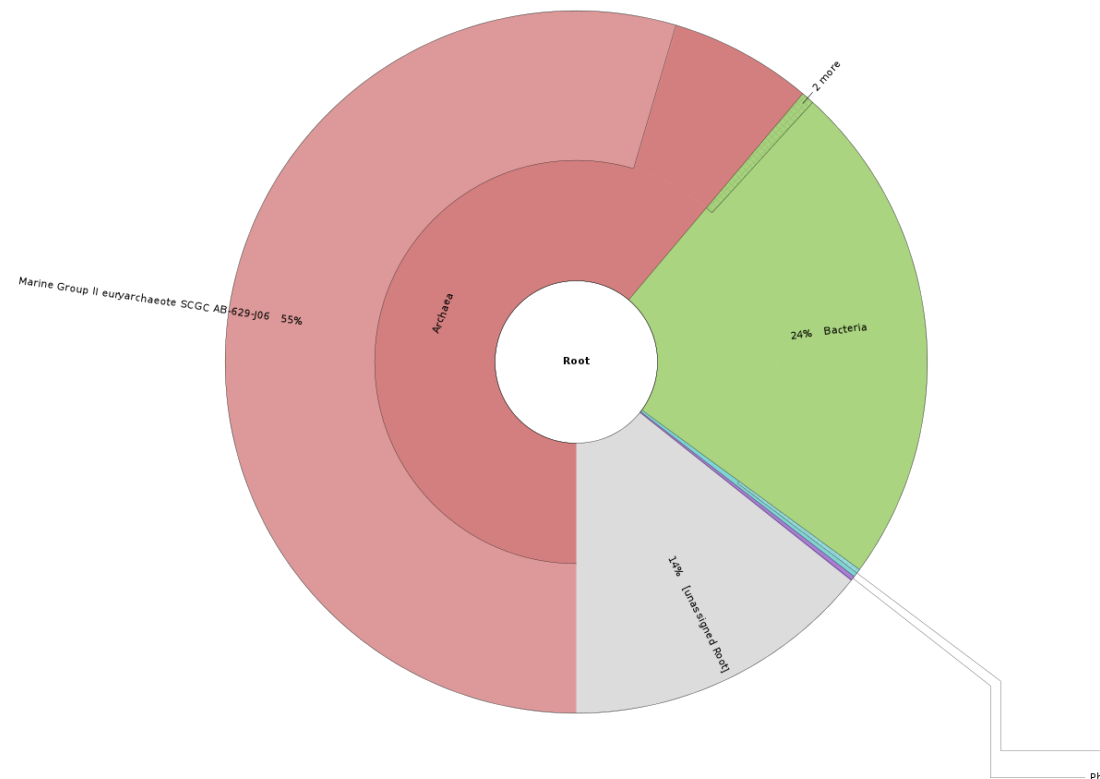
Annexe 1 : Graphiques

Google charts

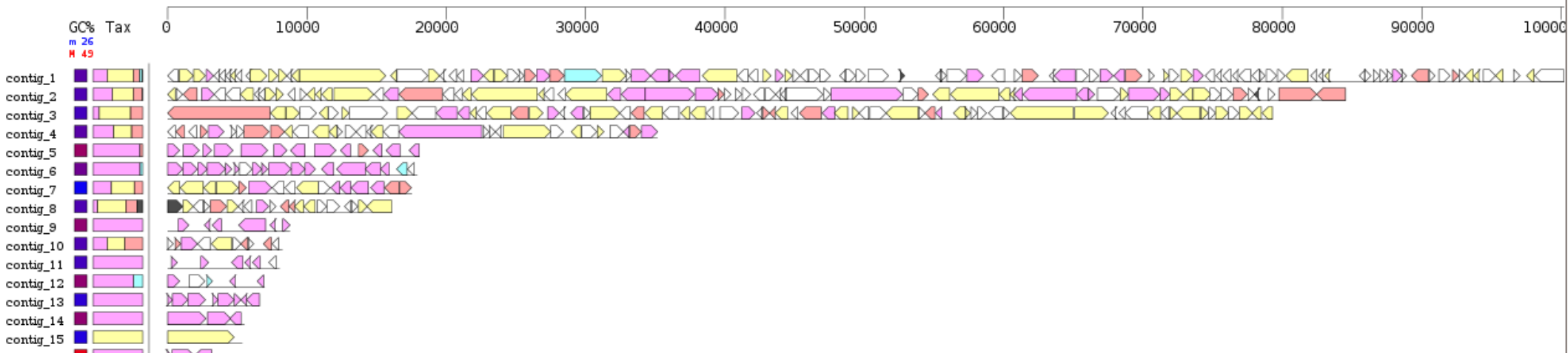
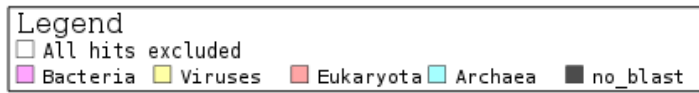
Distributionn of the the closet homologs against the NCBI NR database



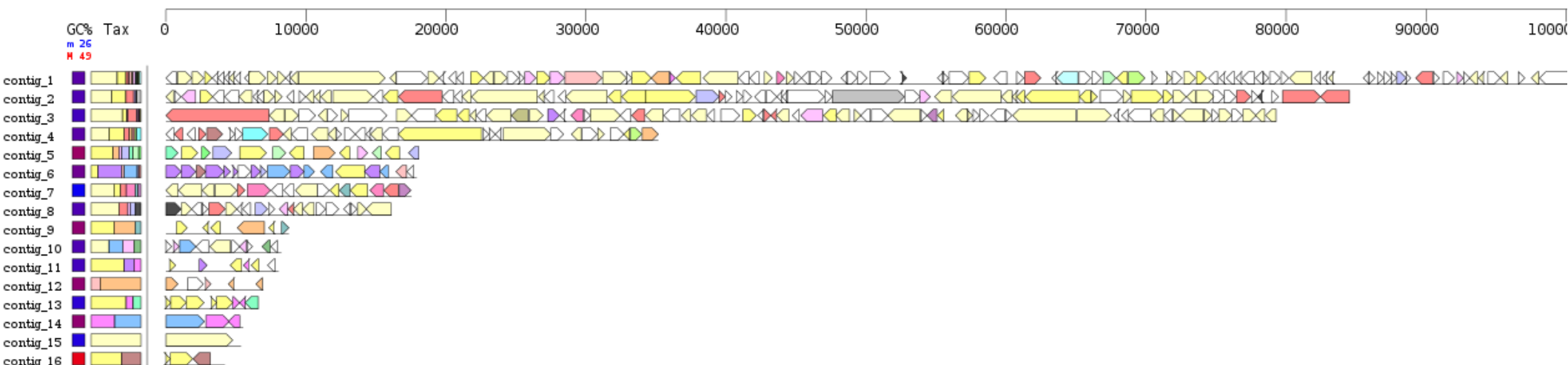
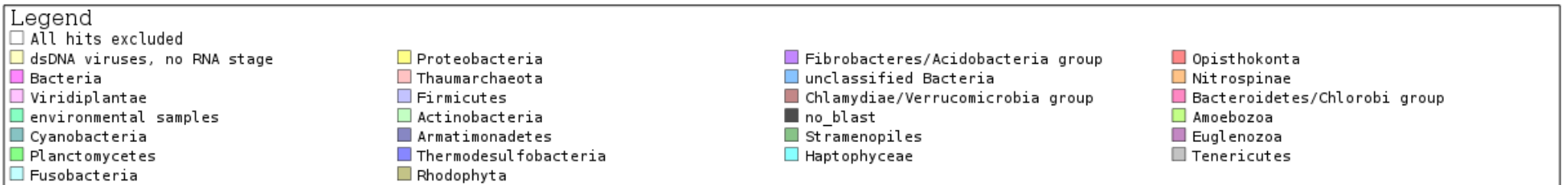
Krona tools



Annexe 2 : Colorsets différents



Washington-All



Annexe 3 : Affichage double brin

Washington-All

