Damien DESMARAIS

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

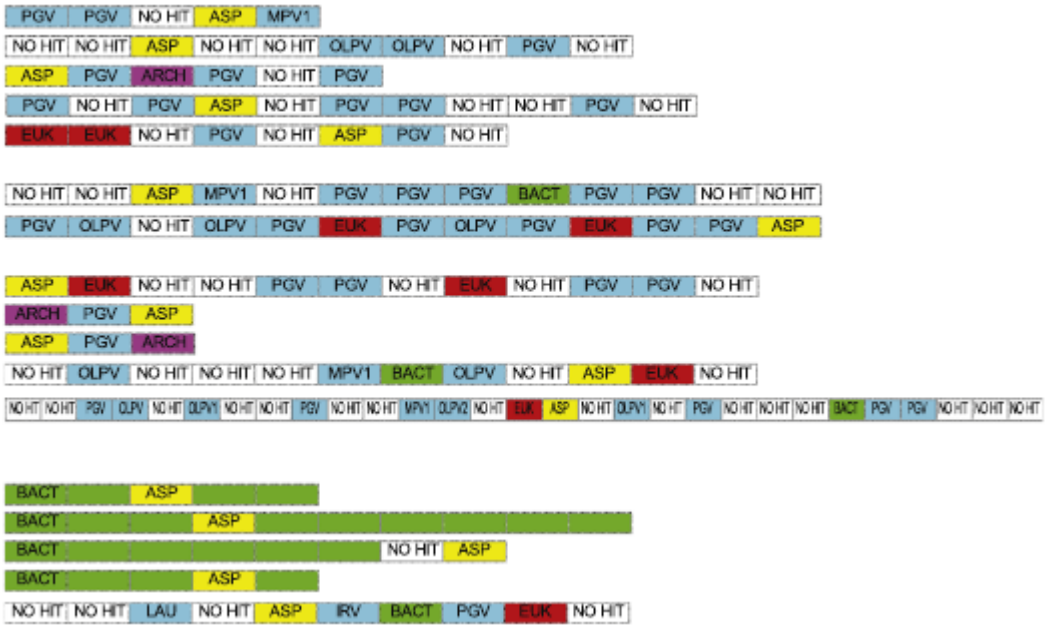
Sommaire

Problématique

Données de génomique bruitées

Séquençages de cellule unique entrainent souvent des problèmes d'annotation

Projet : Washingtonvirus

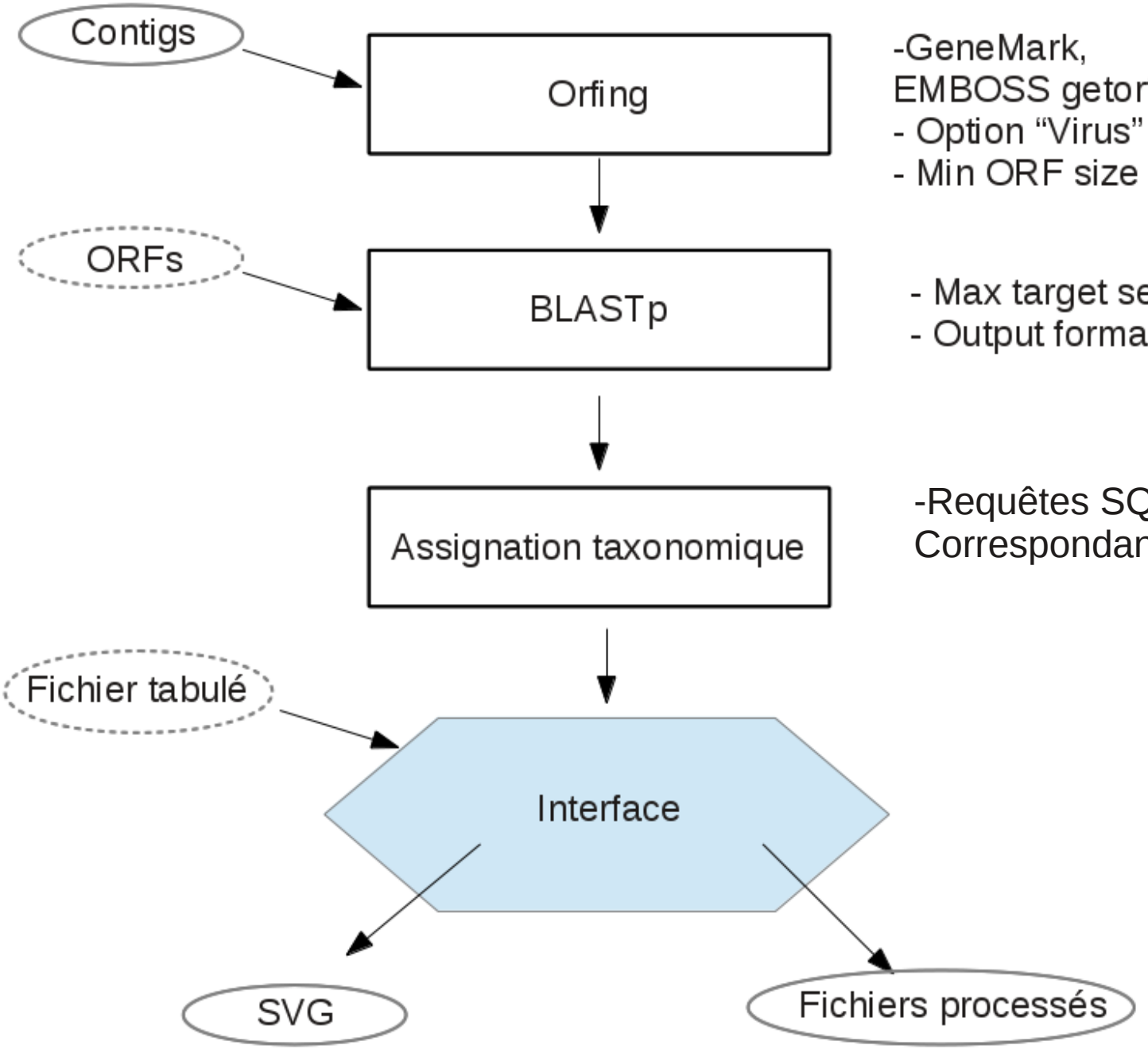


Données de métagénomique

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

Projet : Spongex

Traitement des données



-GeneMark, MetaGeneMark ou EMBOSS getorf
- Option “Virus” pour GeneMark
- Min ORF size pour getorf

- Max target sequences : 25
- Output format : tabulated file

-Requêtes SQL contre une table de Correspondance GI / taxId (NCBI)

Choix techniques

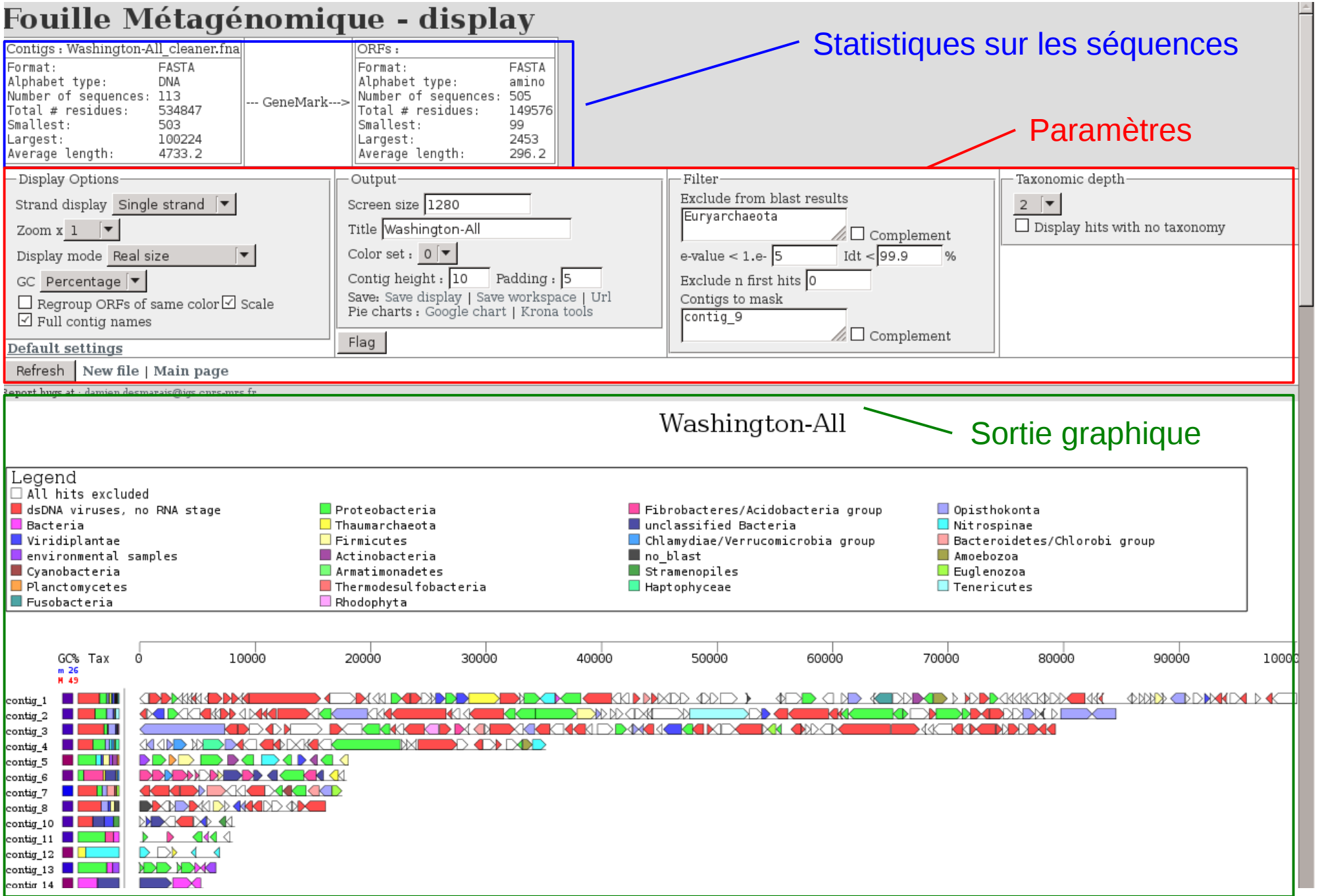
Langage : Perl (module CGI)

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

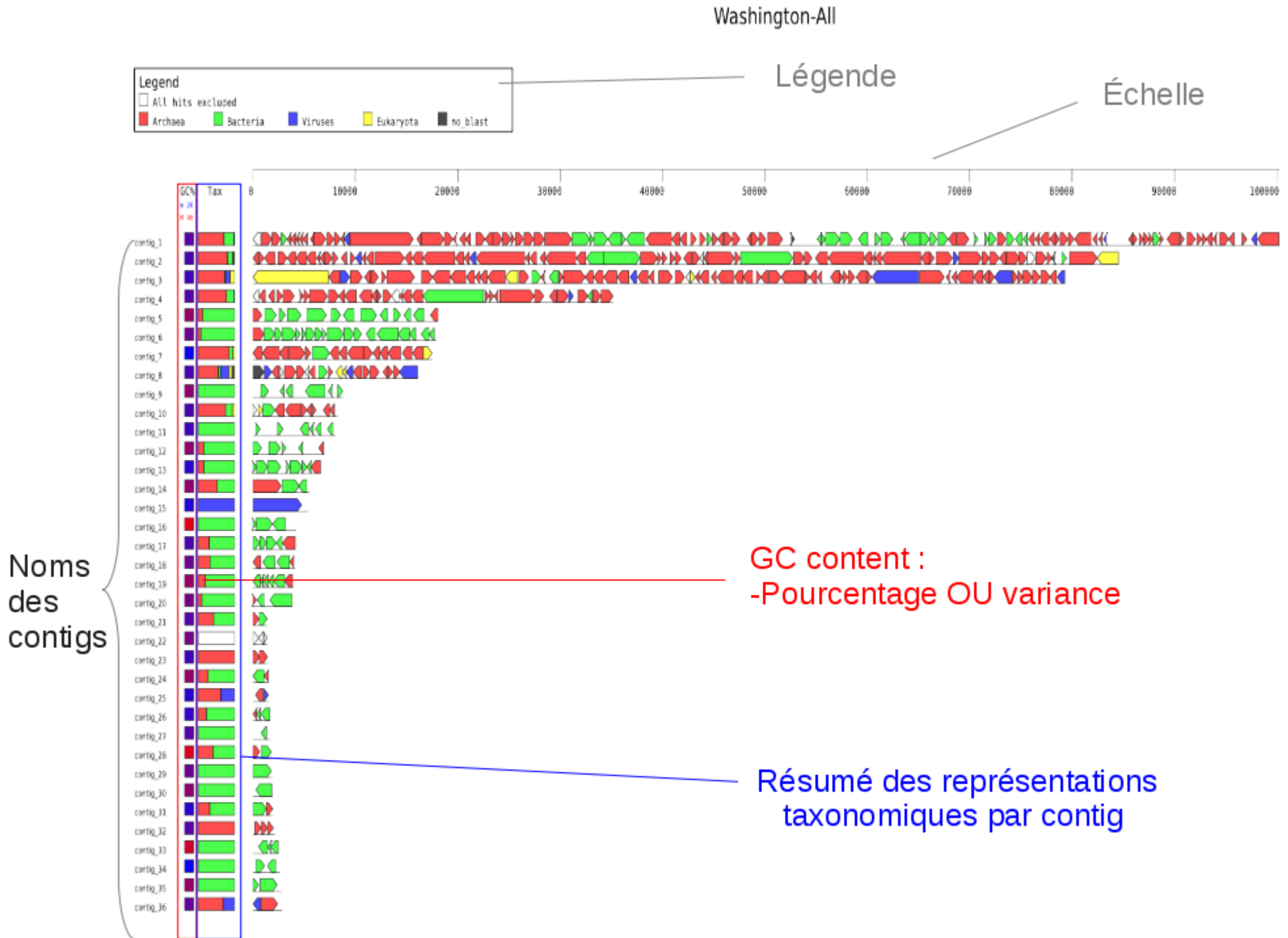
SVG

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

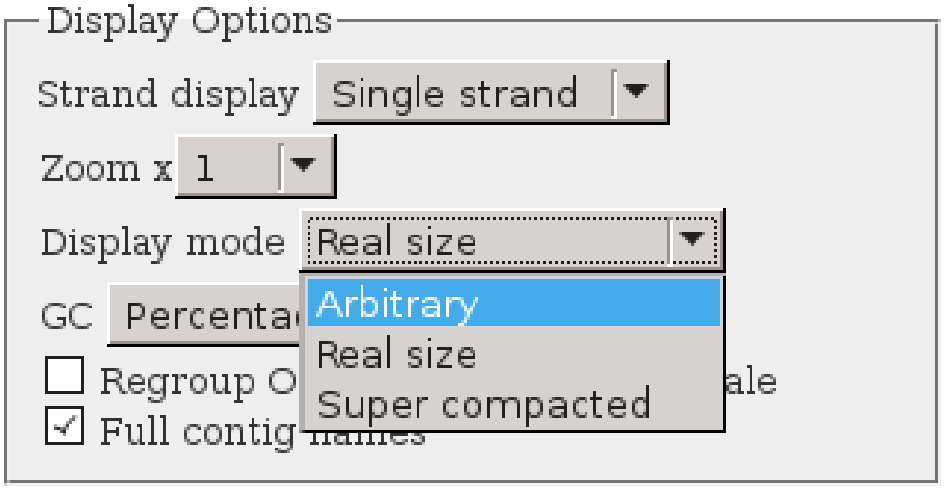
Vue d'ensemble



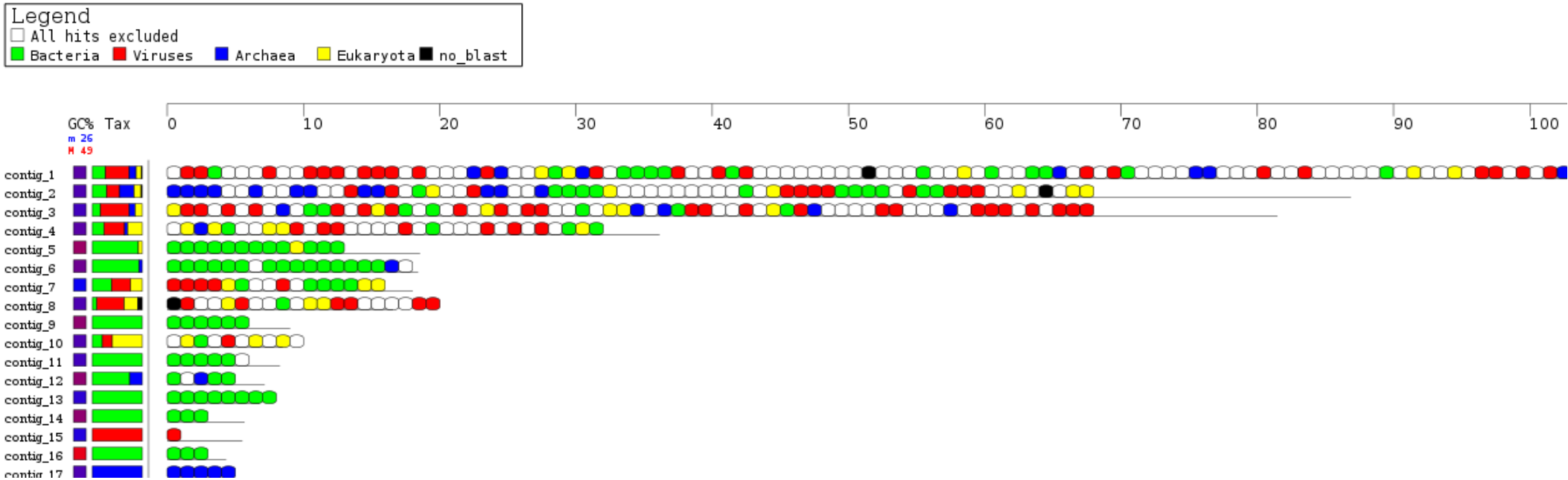
Détails de l'affichage



Mode d'affichage : Taille arbitraire



Washington-All



Mode d'affichage : Histogramme

Display Options

Strand display: Single strand

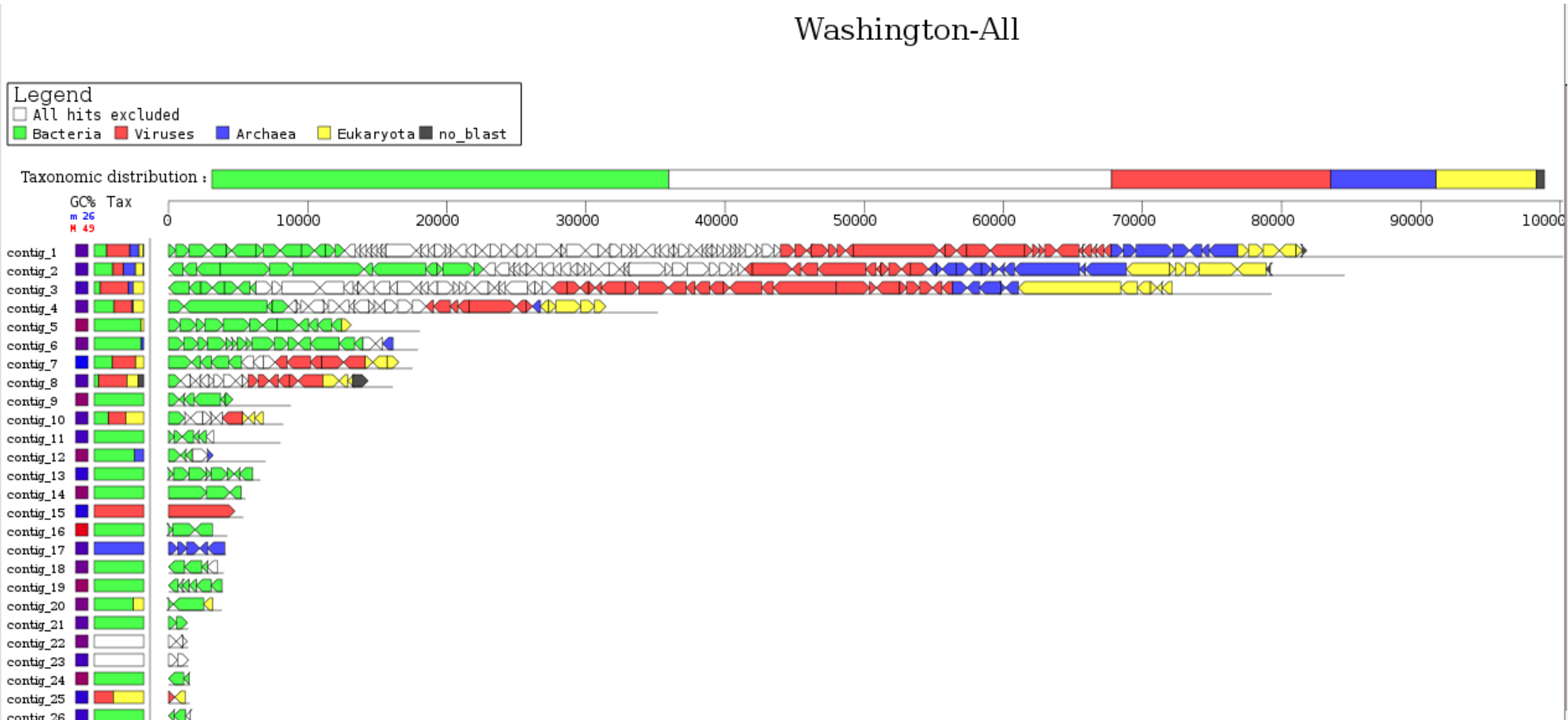
Zoom x: 1

Display mode: Real size

GC: Percentage

☑ Regroup ORFs of same color

☑ Full contig names



Options graphiques

Output

Screen size 1280

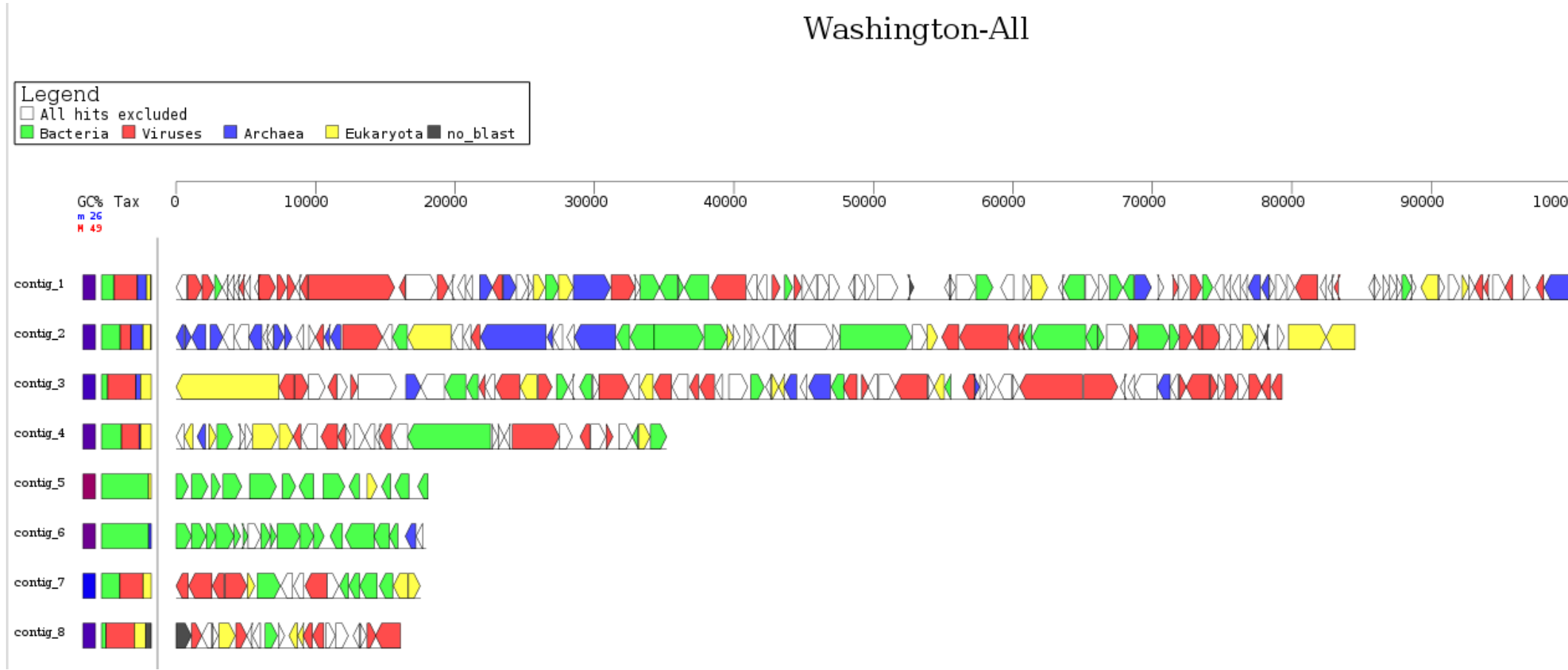
Title Washington-All

Color set : 0

Contig height : 20 Padding : 20

Save: Save display | Save workspace | Url

Pie charts : Google chart | Krona tools



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

Filter

Exclude from blast results

Euryarchaeota

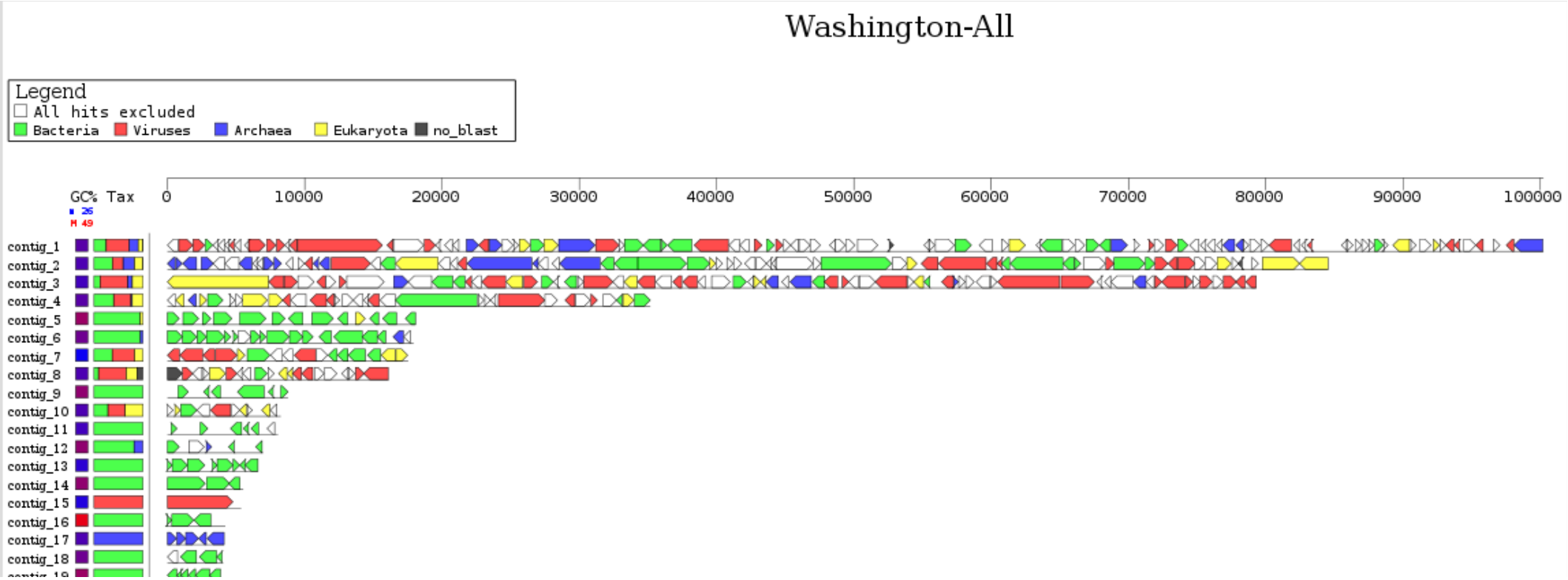
☐ Complement

e-value < 1.e- 5 Idt < 99.9 %

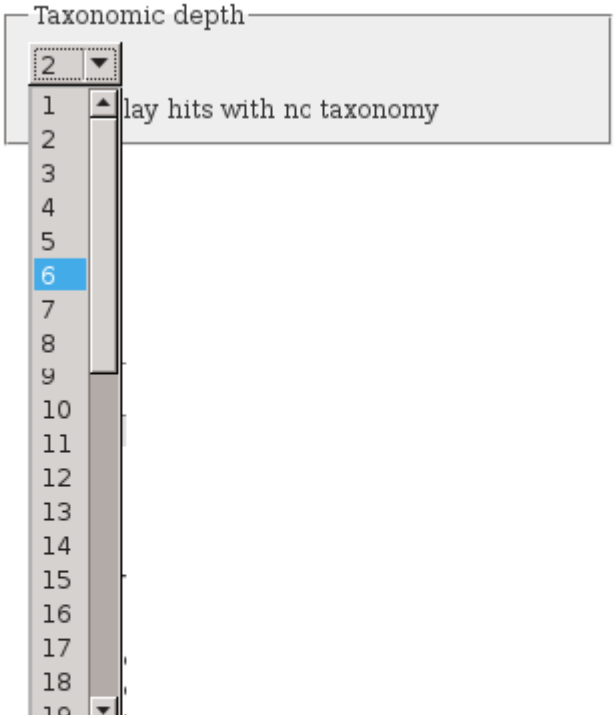
Exclude n first hits 0

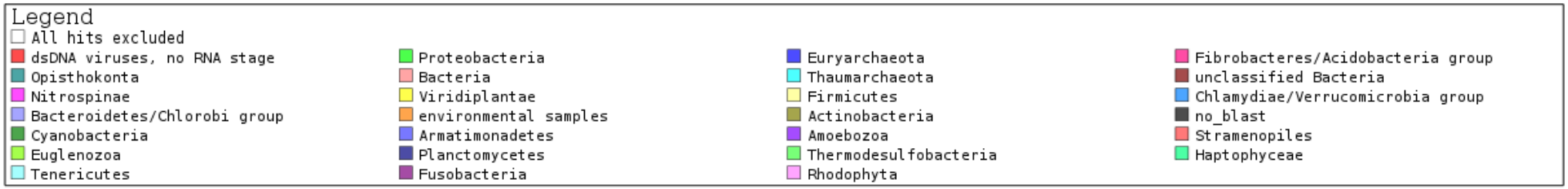
Contigs to mask

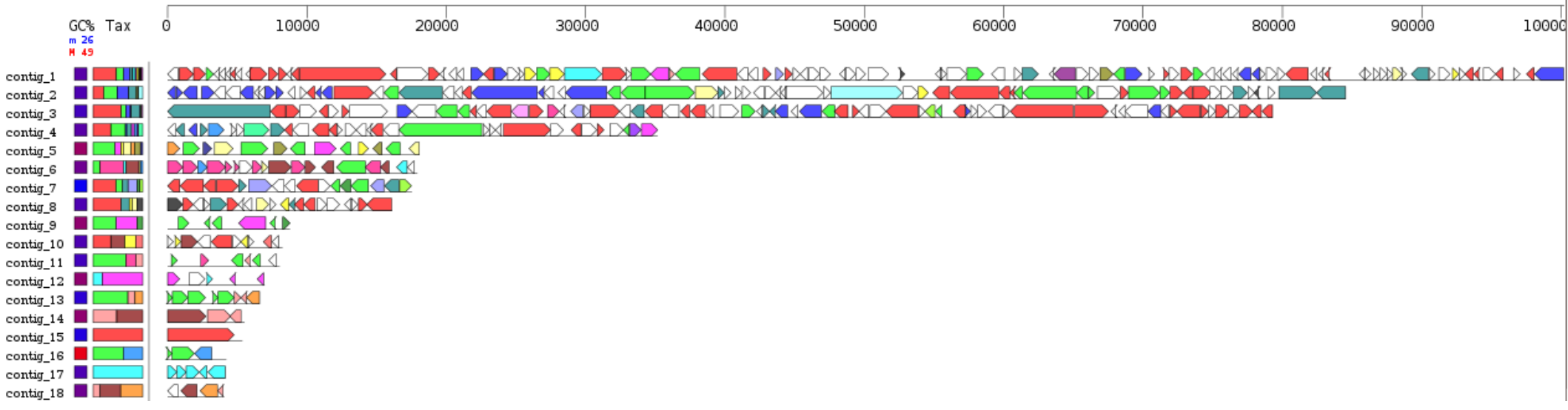
☐ Complement



Choix de la profondeur taxonomique

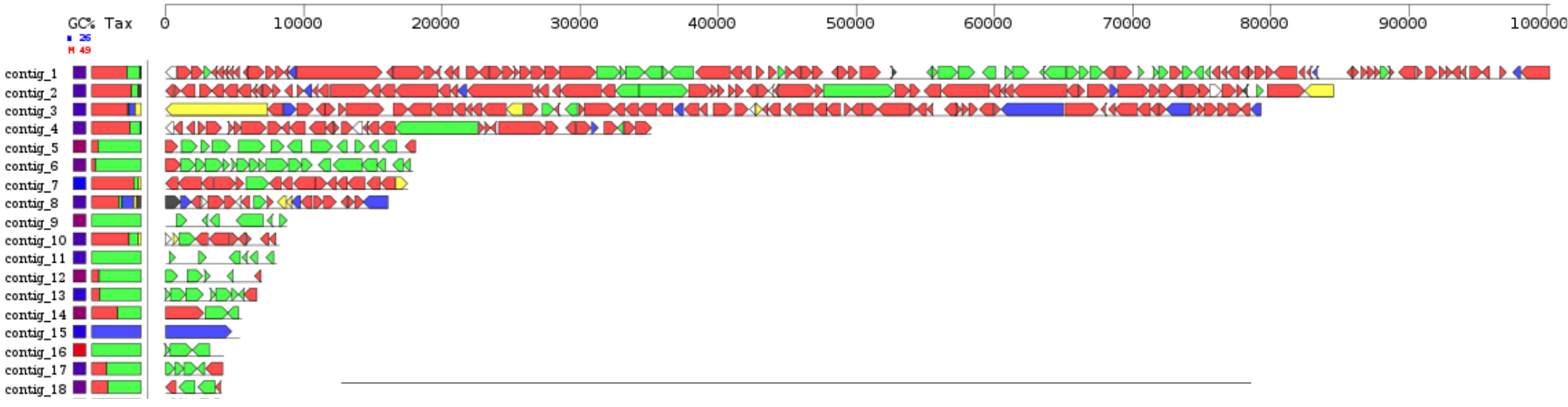




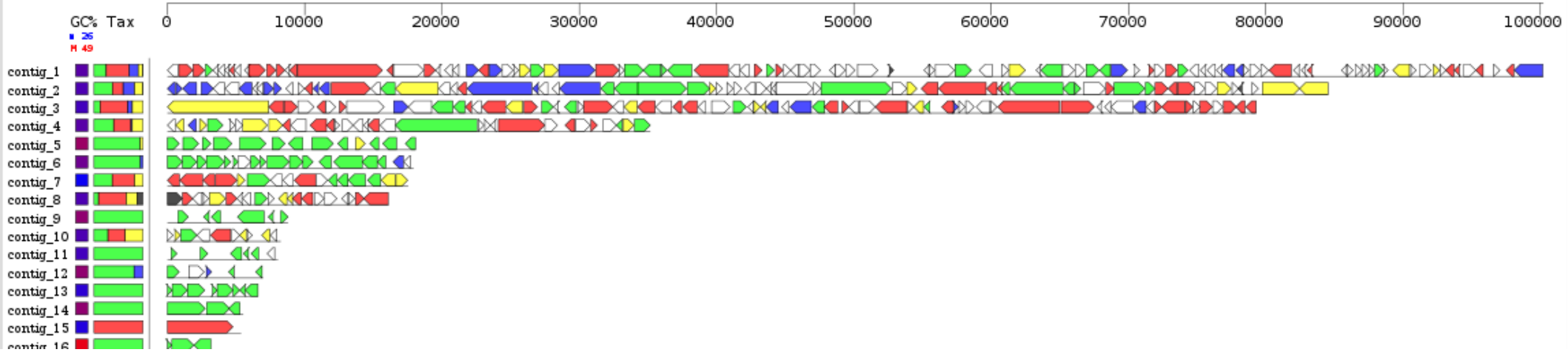


Assignation taxonomique erronée : Utilisation du filtre

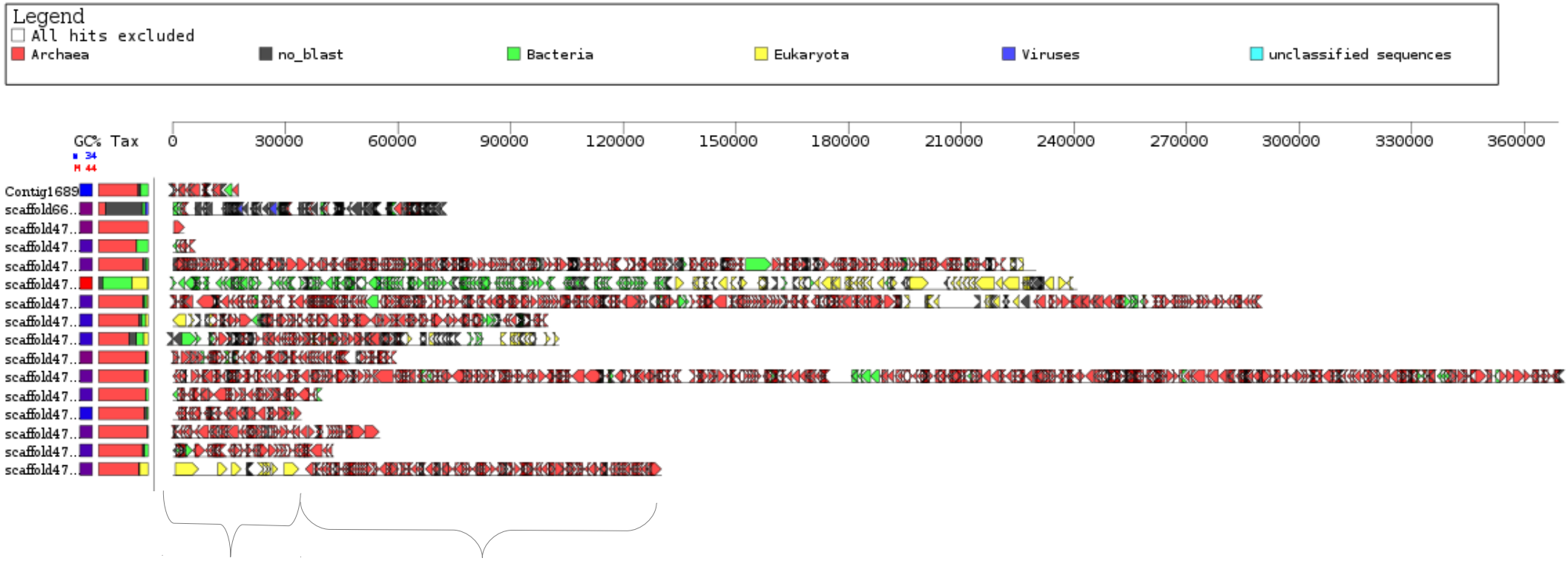
Affichage par défaut



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

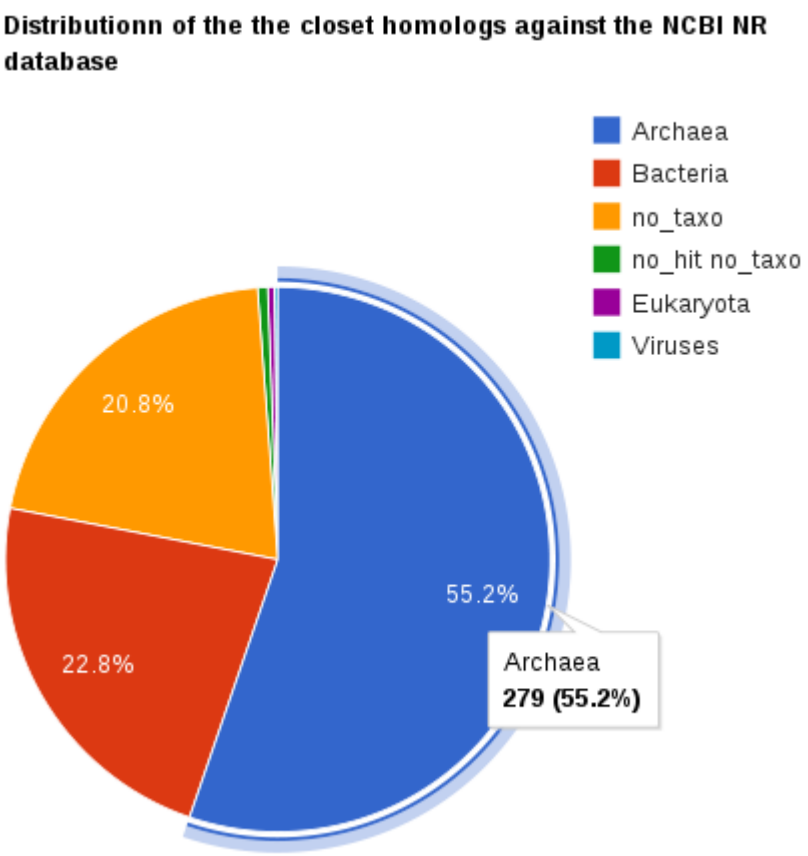
-Jean-Michel Claverie

-Olivier Poirot

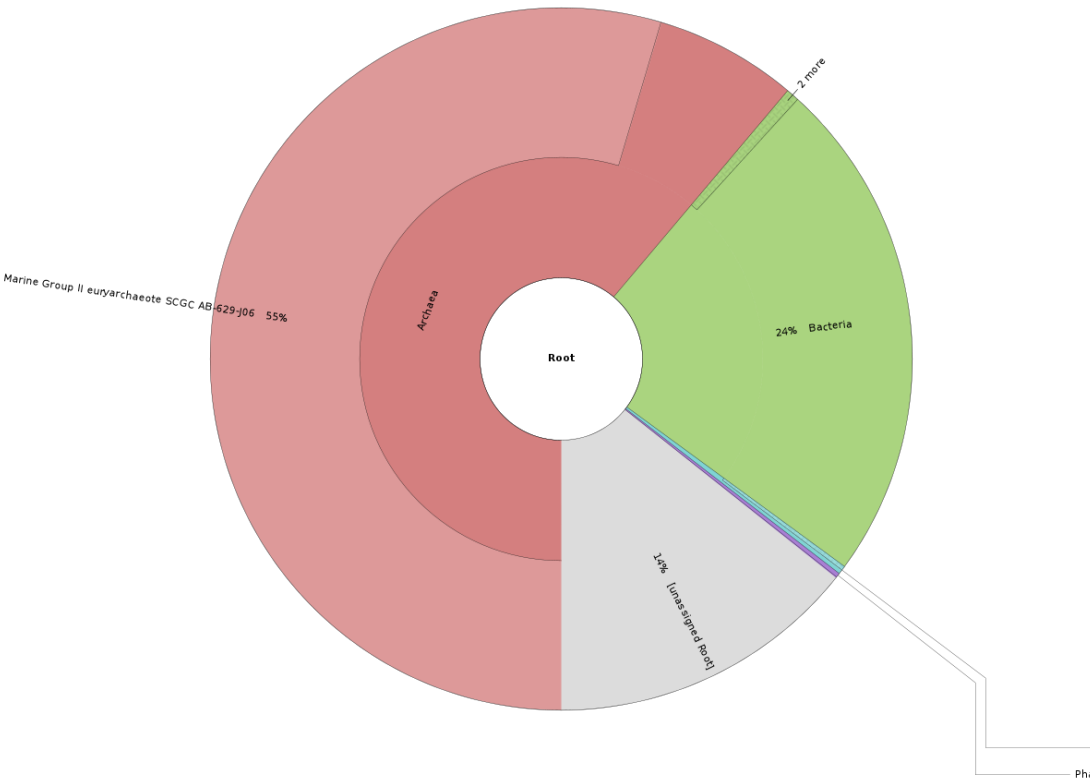
-Sebastien Santini

Annexe 1 : Graphiques

Google charts

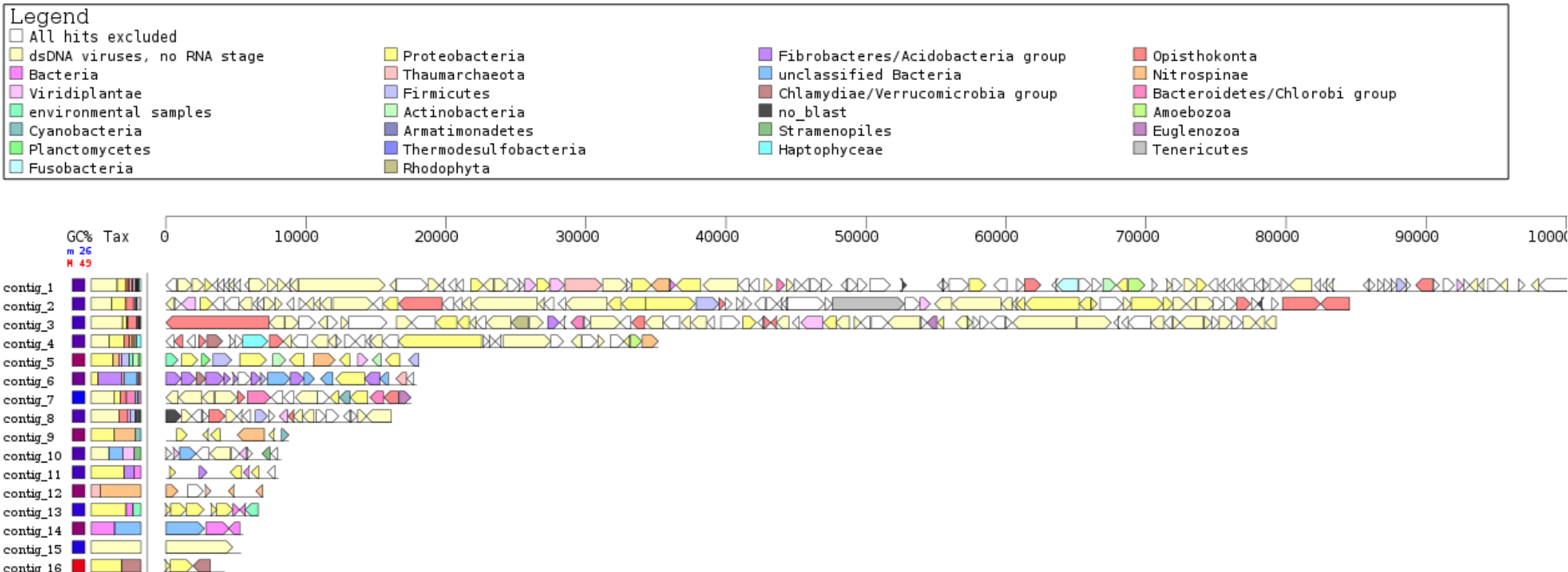


Krona tools



Annexe 2 : Colorsets différents

Washington-All



Annexe 3 : Affichage double brin