Introduction to genome annotation

EBAII Assembly & Annotation - Roscoff June 2024

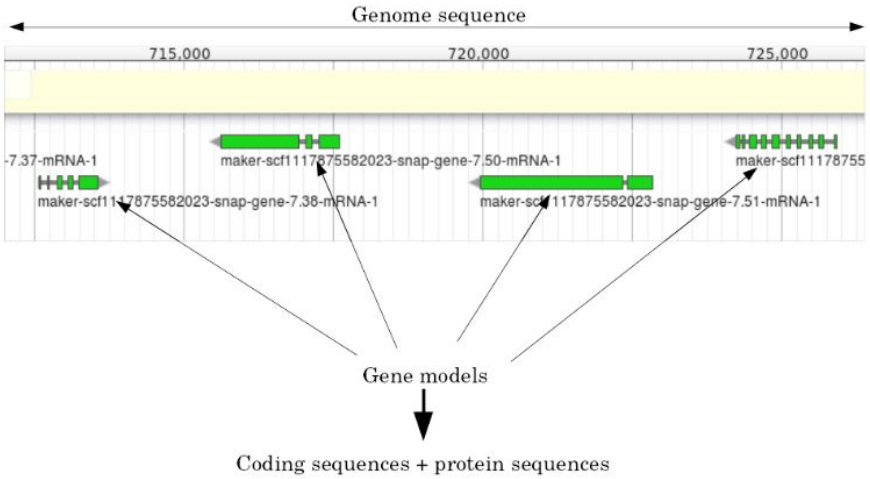
Romane LIBOUBAN
GenOuest platform, Rennes

Annotation des génomes

- **Annotation structurelle**
  - identifier et localiser la position des éléments génomiques le long du génome
  - *Types d'éléments :*
    - gènes (introns, exons, UTR, CDS)
    - régions régulatrices (par exemple, promoteur)
    - ARNnc (ARNr, ARNt, etc.)
    - éléments répétés
    - pseudogènes et paralogues

- **Annotation fonctionnelle**
  - attribuer des fonctions à ces éléments génomiques

Annotation structurelle : Pourquoi ?



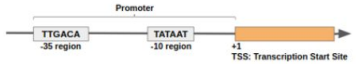
Annotation structurelle : Pourquoi ?

- Comprendre la complexité des génomes et leur fonctionnement
- Base pour d'autres analyses, par ex :
  - Données transcriptomiques (comptage des lectures à l'intérieur des exons)
  - Détection de variants (SNP, indels, ...) et leurs effets
  - Épigénomique (ChIPSeq, FAIRESeq, …)
- Comparaison avec d'autres espèces
  - Présence/absence/mutations de gènes
  - Réduction ou élargissement de la famille
  - Variants structurels

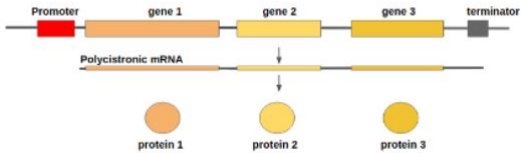
Annotation of Prokaryotic and Eukaryotic genomes

Gènes procaryotes

- Promoter
  - -35 région
  - TATA box
  - Initiation site (TSS)



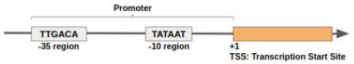
- Operons
  - Prometer
  - Some genes
  - A terminator



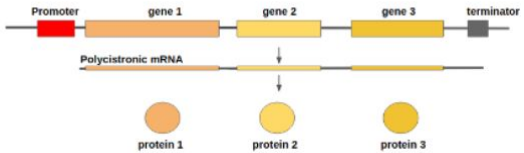
Annotation of Prokaryotic and Eukaryotic genomes

Gènes procaryotes

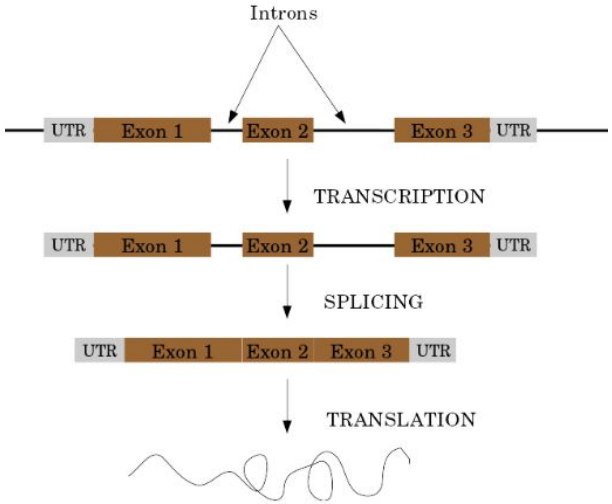
- Promoter
  - -35 région
  - TATA box
  - Initiation site (TSS)



- Operons
  - Prometer
  - Some genes
  - A terminator

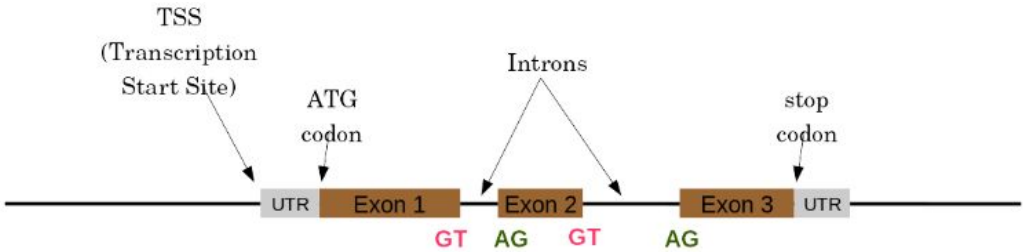


Gènes eucaryotes



Résumé

- Problème très difficile
  - Motifs courts, variables et peu spécifiques
  - Besoin de données pour étayer les prédictions

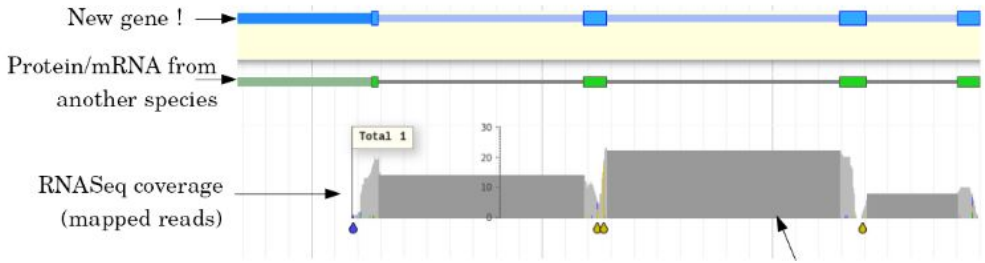


Annotation structurelle automatique

- Algorithmes et logiciels
- Marquer des structures spécifiques
- Avantages
  - Rapide : permettant de traiter des grandes données
  - Echelle : idéal pour des volumes massifs
- Inconvénients
  - Qualité de l’annotation dépend de l’algorithme
  - Sur le choix de l’algorithme

Evidence

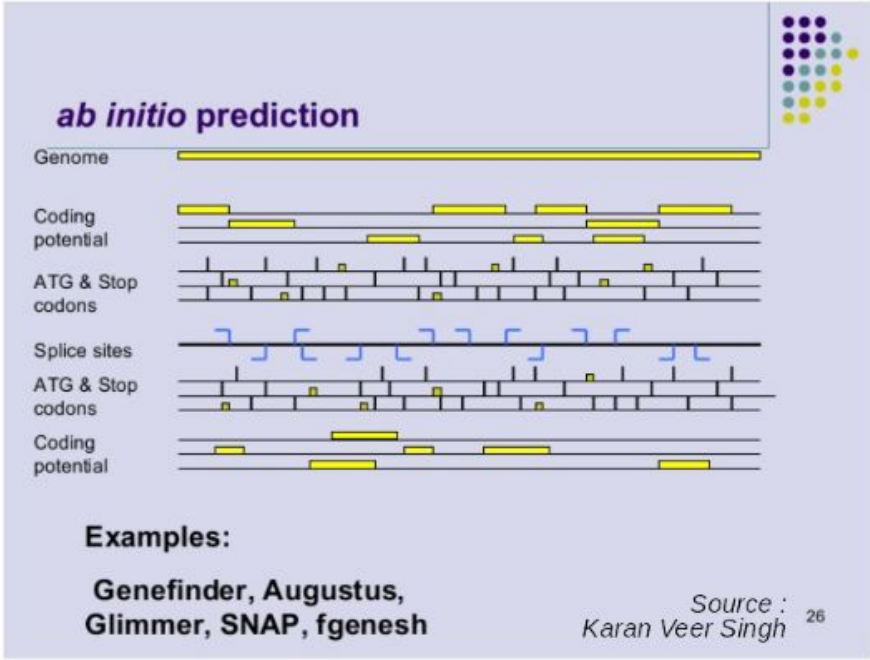
- Alignements de lectures RNAseq
- Alignements des EST ou des transcriptions
  - Même espèce
  - Espèces étroitement apparentées
- Alignement des protéines
  - Espèces étroitement apparentées



/!\ gènes nouveaux ou très éloignés, et gènes inexprimés

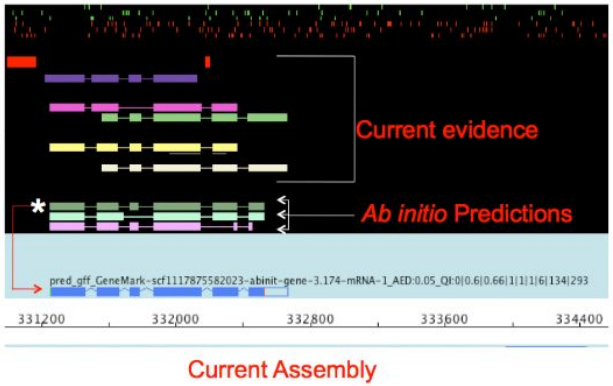
Prédiction *ab initio*

- Séquences du génome
- Modèle statistique
  - Étape d'entraînement pour attribuer un score
    - Prédiction de gènes
  - Répété
    - entraîner, prédire, sélectionner les meilleurs gènes, recycler, etc



Data reconcilliation

- Intégrer les évidences + prédictions ab initio
- Consensus
- Améliorer la précision et la fiabilité des données
- Pipelines automatisés
    - Maker, Braker, Braker3, Funannotate, Pasa, etc



Helixer

- Approche nouvelle et différente
  - Utilisation de GPU
  - Temps d'exécution plus rapide
  - Prédictions de gènes sans preuves
- Annotation basée sur le développement et l’utilisation d’un modèle d’apprentissage profond inter-espèces
- Annotation ab-initio des gènes principaux entre espèces de grands génomes eucaryotes

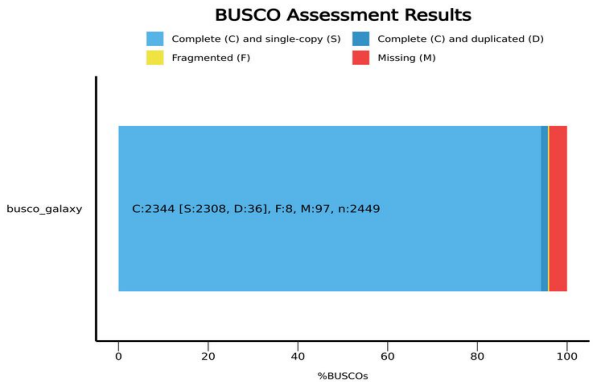
Évaluation de l'annotation : métriques

- Nombre de gènes
- Nombre moyen d'exons
- Longueur moyenne des gènes
- Longueur moyenne des protéines
- ...

Évaluation de l'annotation : BUSCO

→ Benchmarking Universal Single-Copy Orthologs

- Ensembles de gènes ayant des orthologues à copie unique dans toutes les espèces d'un clade (insectes, plantes, bactéries, …)
- Ces gènes sont censés être essentiels pour l'organisme
  - Leur présence indique une bonne annotation
- Résultats
  - Gènes trouvés
  - Des gènes fragmentés
  - Gènes dupliqués
  - Absents



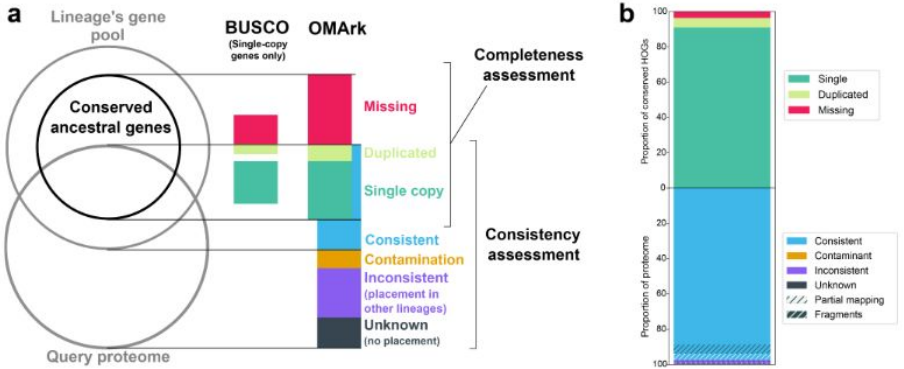
Évaluation de l'annotation : Compleasm

- Implémentation de BUSCO
  - Plus rapide
  - Plus précis
  - Neng Huang, Heng Li, compleasm: a faster and more accurate reimplementation of BUSCO, *Bioinformatics*, Volume 39, Issue 10, October 2023, btad595, https://doi.org/10.1093/bioinformatics/btad595
- Résultats similaires à BUSO
  - Gènes trouvés
  - Des gènes fragmentés
  - Gènes dupliqués
  - Absents

Évaluation de l'annotation : OMArk

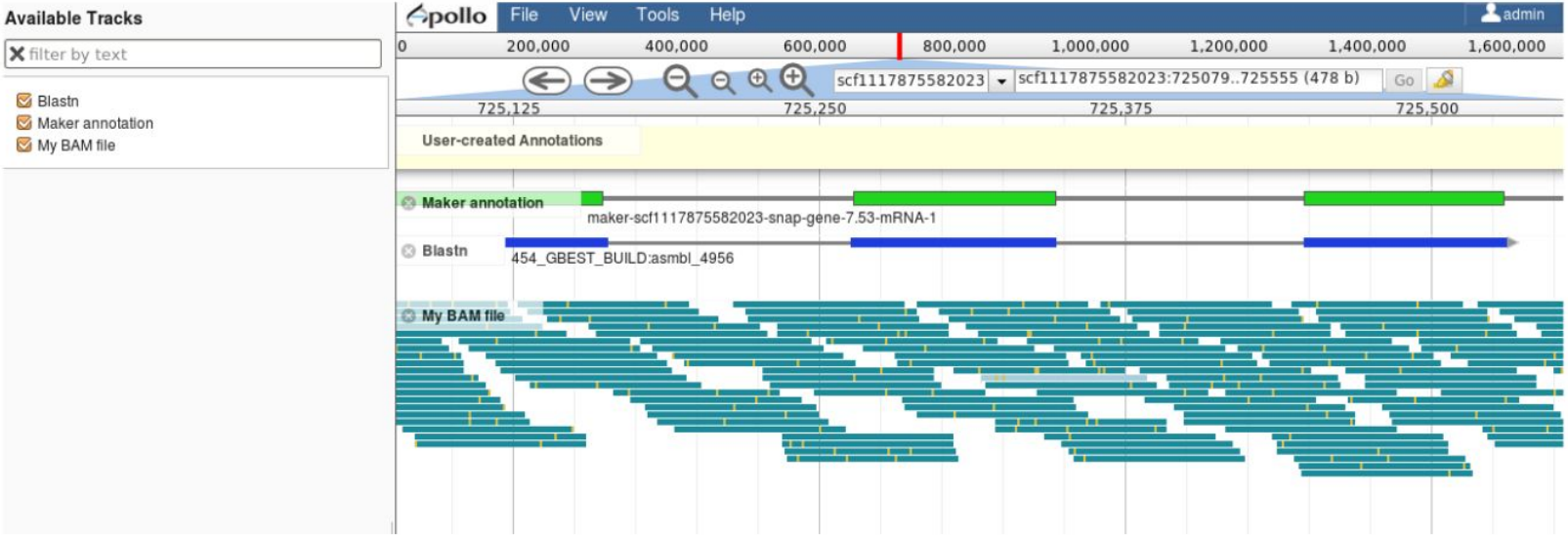
- Assigner les protéines aux HOGs en utilisant la composition k-mer
  - HOGs = Hierarchical Orthologous Groups
  - Base de données OMA
- Différences par rapport à BUSCO
  - Complétude : prend également en compte les gènes conservés dans des copies multiples
  - Cohérence : vérifie si (toutes) les protéines correspondent à la lignée auto-détectée ou non

Nevers, Y., Warwick Vesztrocy, A., Rossier, V., Train, C. M., Altenhoff, A., Dessimoz, C., & Glover, N. M. (2024). Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology*, 1-10.



Visualisation

- Genome Browser : JBrowse, UCSC, etc
- JBrowse
  - GTN : https://training.galaxyproject.org/training-material/topics/visualisation/tutorials/jbrowse/tutorial.html



Éléments répétés

- Transposons, rétrotransposons, régions de faible complexité
- Insertion
  - Influencer l’expression des gènes
- Important
  - Fournir des informations sur la structure et l’évolution des génomes
  - Stabilité chromosomique

Éléments répétés

- Transposons, rétrotransposons, régions de faible complexité
- Insertion
- Important
- Prediction pipelines:
  - RepeatMasker
  - RepeatModeler
  - Red (Repeat Element Detector)
  - REPET
- Bases de données d'éléments répétés
  - Can be used by pipelines
  - Dfam
  - RepBase (non free)

Eléments exotiques

- Exemples : tRNA, rNRA, ncRNA, etc
- Des outils dédiés à la prédiction
  - Aragorn
  - tRNAscance
- Rôle crucial
  - Diversité génétique
  - Evolution
  - Comprendre les mécanismes complexes de la génétique

Résumé sur l'annotation structurelle automatique

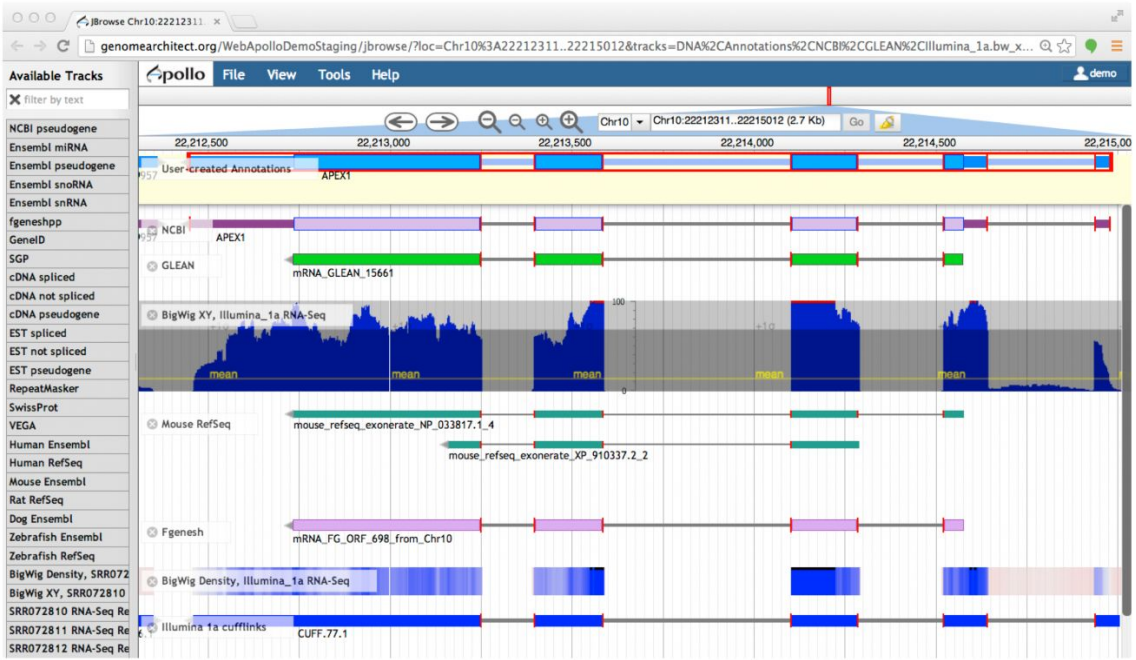
- Problème difficile
- Pipelines automatisés
- Besoin de preuves
- Jamais parfait
  - Gènes manquants/incomplets
  - Gènes divisés/fusionnés
  - Pseudogènes

Annotation manuelle

- Réalisée par des experts de certaines familles de gènes
  - Examiner
  - Marquer les gènes
- Limites
  - Pas d'experts pour tous les gènes
  - Annoter que ce qui se trouve dans la séquence
  - Mauvais assemblage ⇒ Mauvaise annotation
  - Coût en terme de temps
  - besoin d'un environnement convivial

Editeurs - Apollo, Artemis, etc

- Tutoriel :
  - Prokaryotes : https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/apollo/tutorial.html
  - Eukaryotes : https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/apollo-euk/tutorial.html



Etapes

- Vérifier la structure (exons, introns, start, stop, etc)
- Rechercher les isoformes
- Garantir des conventions de dénomination cohérentes
- Ajouter des annotations fonctionnelles
    - Basées sur des homologies avec d’autres espèces

Comparaison annotation automatique et manuelle

Annotation automatique	Annotation manuelle
- Très rapide - Précision dépend de l'algorithme - Suit les règles définies dans l'algorithme	- Plus lente - Plus coûteuses (temps et ressources humaines) - Précision dépend des ressources - Plus flexible → gérer les exceptions

Annotation fonctionnelle

- Collecter les informations sur la fonction des gènes identifiés
  - Fonctions biologiques
  - Régulation, expression, etc
- Les sources de données
  - Expériences en laboratoire
  - Affectation manuelle (Apollo)
  - Affectation automatique

Méthodes

- Recherche de similarité/homologie
- Recherche de modèles
- Orthologies
- Comparaison avec des bases de données
  - GenBank, NR : banques de données de séquences
  - IntroPro : banque de données de motifs (sites actifs, familles de protéines, etc)
  - EggNOG : banque de données de relations ontologiques et annotation fonctionnelle

Blast

- Comparer des séquences
- Trouver des régions de similarité entre séquences
  - Identifier des homologues
  - Annoter des gènes
  - Prédire la fonction des séquences
- Base de données énorme → bonne chance pour avoir un match
- Risque
  - Propagation de la protéine putative
  - Diffusion d’annotations à faible niveau de preuve

InterProScan

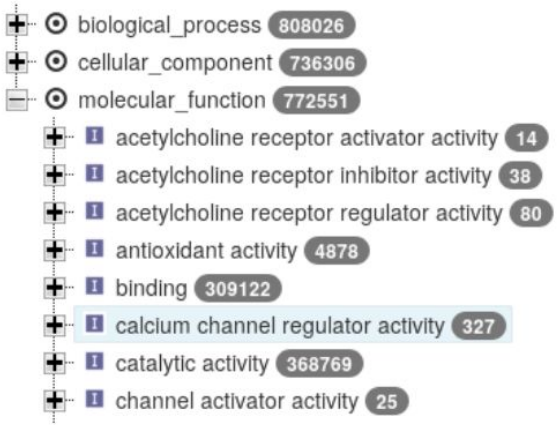
- Pour chaque protéine ou CDS de l’annotation
- Identifier des domaines fonctionnels et des motifs
- Termes d’ontologie génétique disponibles pour les domaines

EggNOG

- Pour chaque protéine de l’annotation
  - Rechercher des correspondances avec des groupes d’orthologie connus
- Attribuer le nom du gène correspondant et l’annotation fonctionnelle

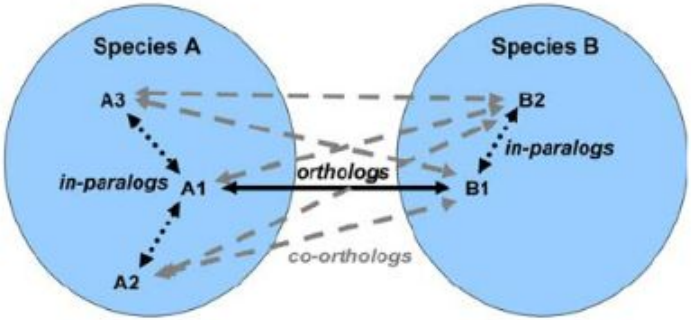
Gene Ontology

- Système de classification hiérarchique destiné à standardiser la représentation des attributs des gènes
  - Fonction moléculaire
  - Processus biologique
  - Composant cellulaire



Orthologie

- Pour chaque gène annoté
  - Recherche de gènes orthologues chez des espèces apparentées
  - Recherche de paralogues
- Méthode bioinformatique
  - Filtrage des meilleurs hits
  - Regroupement
  - OrthoFinder, OrthoMCL, etc



Visualisation

- Bases de données génomiques (NCBI, FlyBase, etc)
- Autres sites
  - Données de référence (assemblage, annotation, etc)
  - Des interfaces pour visualiser ces données
  - Des interfaces d'interrogation
    - https://bipaa.genouest.org/is/

Comparer les annotations

- Nécessaire de choisir entre différents résultats sur une même séquence du génome
- Comparer les statistiques générales
  - Nombre de gènes, nombre moyen d'exons, etc
- Comparer le contenue des gènes
  - Alignement des structures génétiques
  - Annotation fonctionnelle
  - Sur/Sous-représentation des fonctions
- Outils : AEGeAN, Funannotate comparer

Bilan - Annotation du génome

- Difficile
- Automatique
  - Besoins d’évidence
  - Jamais parfait
  - Pipelines automatisés
- Manuelle
  - Lent
  - Nécessite des experts et des évidences