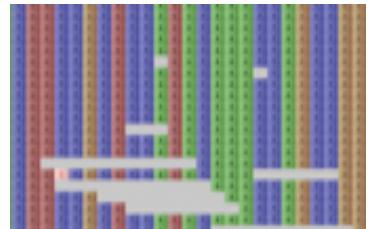
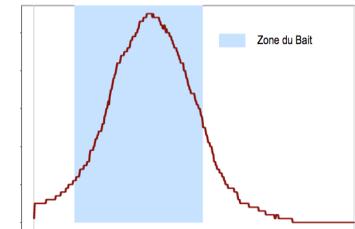
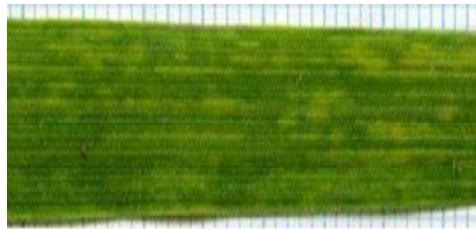


INTRODUCTION A LA BIOINFORMATIQUE

Tunis, Septembre 2015



INRA / Supagro : Y. Holtz, V. Ranwez, J. David, S. Santoni

NGS : des reads aux SNPs

- Introduction NGS
- Les reads : la donnée initiale du bio-informaticien
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

La bioinformatique ? Kezako ?

« Cela va de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, l'assemblage de génome et la reconstruction d'arbres phylogénétiques (phylogénie) »

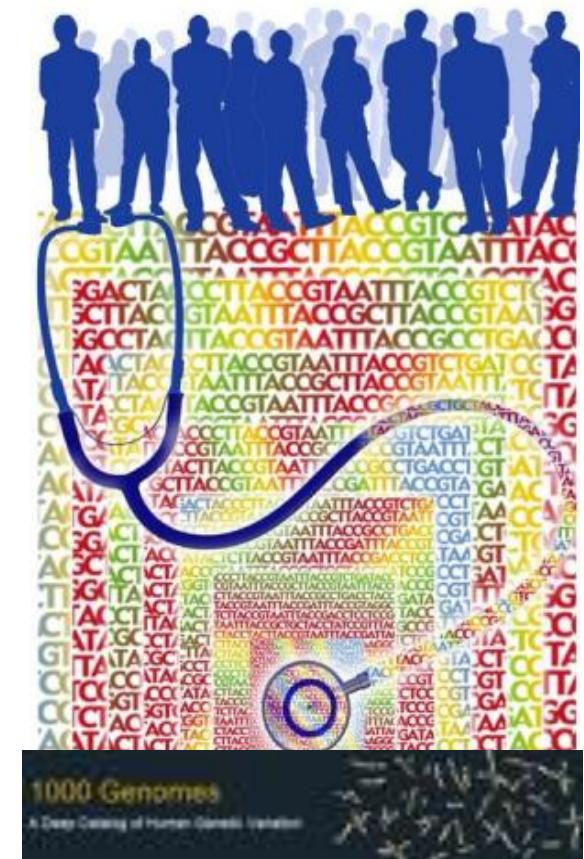


- Répondre à des questions **biologiques** nécessitant **l'informatique**
- Très souvent lié à la gestion de données de séquençages **NGS** : New Generation Sequencing

Applications : Santé humaine

Les NGS peuvent permettre de lier une **maladie** à une **zone** du génome.

Projet 1000 génomes :
Reséquençage massif 1000 génomes, 2500 génomes etc.

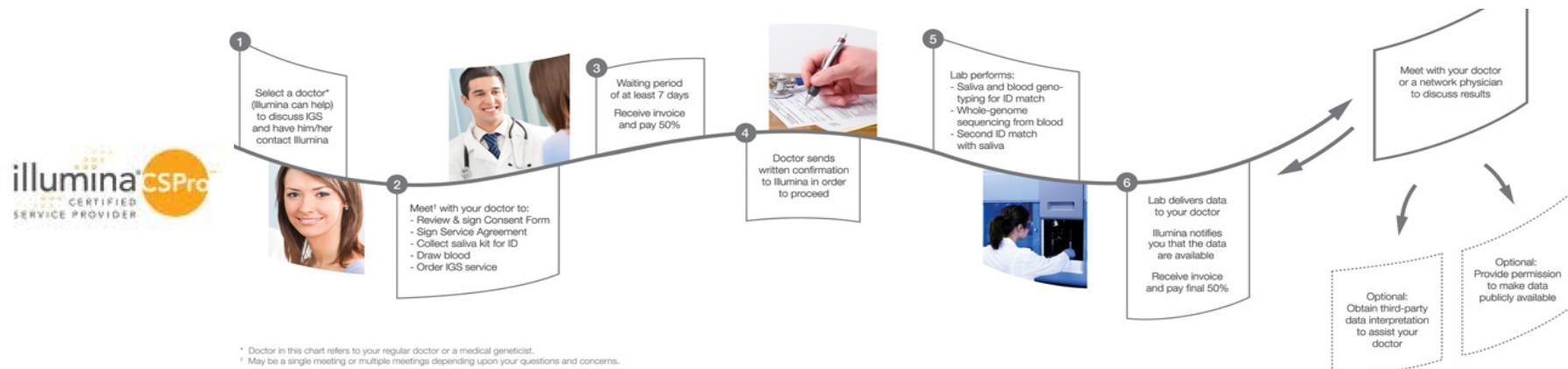


Applications : Santé humaine

Médecine personnalisée ☺

Assurance personnalisée ☹

De Nombreuses questions médicales, juridiques et **éthiques**



Applications : Analyse de diversité

- Etude de la diversité inter et intra espèces
 - Phylogénie, flux de gènes, protection de la biodiversité

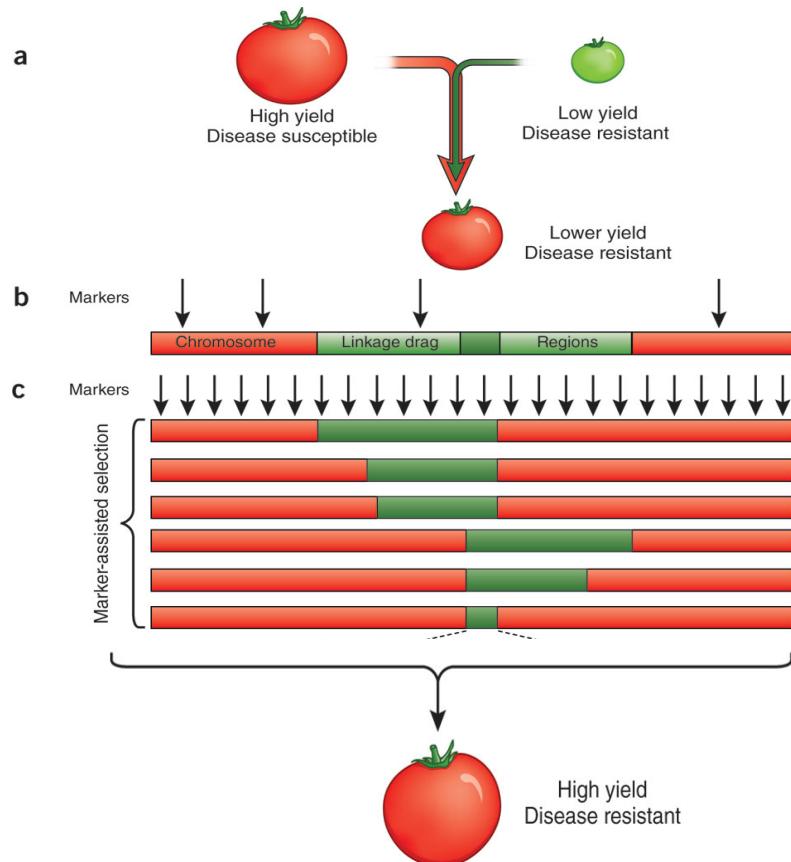
The image shows a repeating pattern of green leaves and flowers, possibly hibiscus, overlaid on a grid of DNA sequence data. The DNA sequence is composed of the letters G, A, T, C, and C, arranged in a grid format. The green floral pattern is centered on the grid, creating a decorative effect.

<http://genome10k.soe.ucsc.edu/>



<http://www.1001genomes.org/>

Applications : Marqueurs Moléculaires

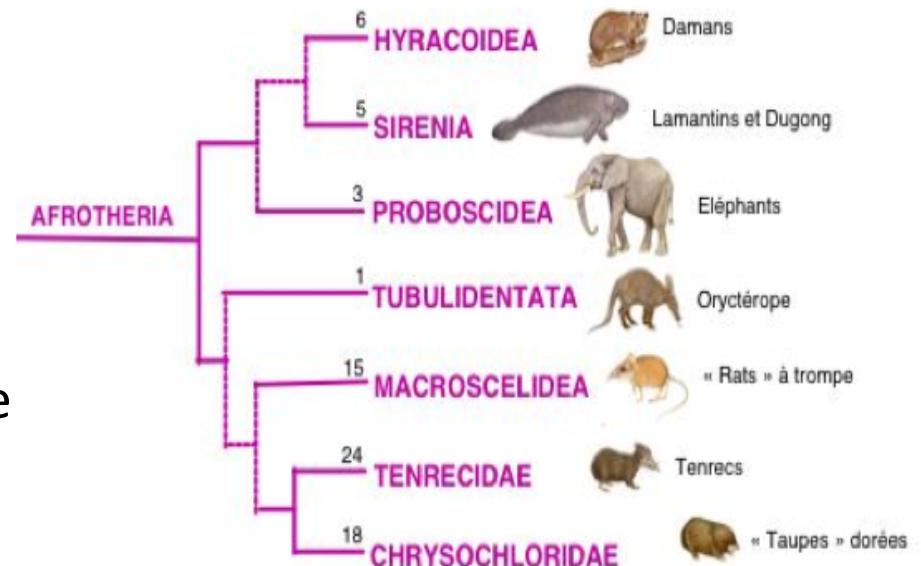


Identification de **marqueurs génomiques** sur l'ensemble du génomes : **SNP**.

Permet ensuite de trouvé des lien entre **génotype** et **phénotype** → utilisation de **sélection assistée par marqueur (SAM)** !

Autres Applications

- Etude de la **structure** de protéines.
- Analyse **phylogénétiques**
- Etude de l'**expression** des gènes
- Autre : Police criminelle ? Analyse d'ADN ancien ? ...



<http://www.lirmm.fr/~vberry/Recherche/>

La bioinfo : comment faire ?



Cluster de Calcul

Ligne de Commande



Interface graphique

```
[ holtz davem-210 ~/Desktop/TMP_CAPTURE3 ] scp holtz@marmadais.cirad.fr_P_3_FEVRIER/MAPPING_ON_BREAD_WHEAT/RESULTAT_MAPPING_BWA_MEM/*13.bam .  
holtz@marmadais.cirad.fr's password:  
resultat_re_mapping.tag_13.bam  
[ holtz davem-210 ~/Desktop/TMP_CAPTURE3 ]  
[ holtz davem-210 ~/Desktop/TMP_CAPTURE3 ] scp holtz@marmadais.cirad.fr_P_3_FEVRIER/MAPPING_ON_BREAD_WHEAT/mart* .  
holtz@marmadais.cirad.fr's password:  
mart_export_wheat_unspliced_gene.txt  
mart_export_wheat_unspliced_gene.txt.amb  
mart_export_wheat_unspliced_gene.txt.ann  
mart_export_wheat_unspliced_gene.txt.bwt  
mart_export_wheat_unspliced_gene.txt.pac  
mart_export_wheat_unspliced_gene.txt.sa  
[ holtz davem-210 ~/Desktop/TMP_CAPTURE3 ] scp holtz@marmadais.cirad.fr_P_3_FEVRIER/MAPPING_ON_BREAD_WHEAT/RESULTAT_MAPPING_BWA_MEM/*13.bai .  
holtz@marmadais.cirad.fr's password:  
resultat_re_mapping.tag_13.bai
```

Pourquoi un serveur de calcul ?



To Do List
4 Ko

Ce PPT
4,1 M

Un film
600 M

Sortie de
séquenceur
d'une pop
de 180
individus
263 G

Des données spécifiques Des outils spécifiques !



L'environnement Linux

```
[ holtz cc2-admin /NAS/davem_data/EPO ]
[ holtz cc2-admin /NAS/davem_data/EPO ]
[ holtz cc2-admin /NAS/davem_data/EPO ]
[ holtz cc2-admin /NAS/davem_data/EPO ] ls
all_cleaned_reads  global_conf_mapping  prev_conf_mapping.bak  SAUVEGARDE_DATA_YAN
diff_conf_mapping  prev_conf_mapping   raw_data              sauvegarde juillet
[ holtz cc2-admin /NAS/davem_data/EPO ]
```

Exemple

Permet de gérer de **grandes quantités** de données très **rapidement**.

Permet d'utiliser tous les **programmes, scripts** etc... développés par d'autres personnes

Mais une période d'**apprentissage** : le language « **bash** »

Interface : Galaxy



Simple d'utilisation (interface graphique)

Installable et Utilisable depuis n'importe quel poste de travail via internet

Connexion direct avec le cluster de calcul

Enchainement de brique (programmes) de calcul.

Mais Limité aux programmes installés, et aux options disponibles

OUTILS

The screenshot shows the Galaxy bioinformatics platform interface. On the left, a sidebar titled "OUTILS" lists various tools categorized under "TOOLS" and "UNTESTED TOOLS". An orange circle highlights the "Tools" section. In the center, the main workspace displays the "South Green bioinformatics platform" logo and a welcome message. A dashed orange box contains two informational messages: one about finding team-made tools using the search function, and another about tool requests. On the right, a sidebar titled "DONNÉES" shows a history of analysis runs, each with a preview icon, a name, and three circular icons. An orange circle highlights the "History" section.

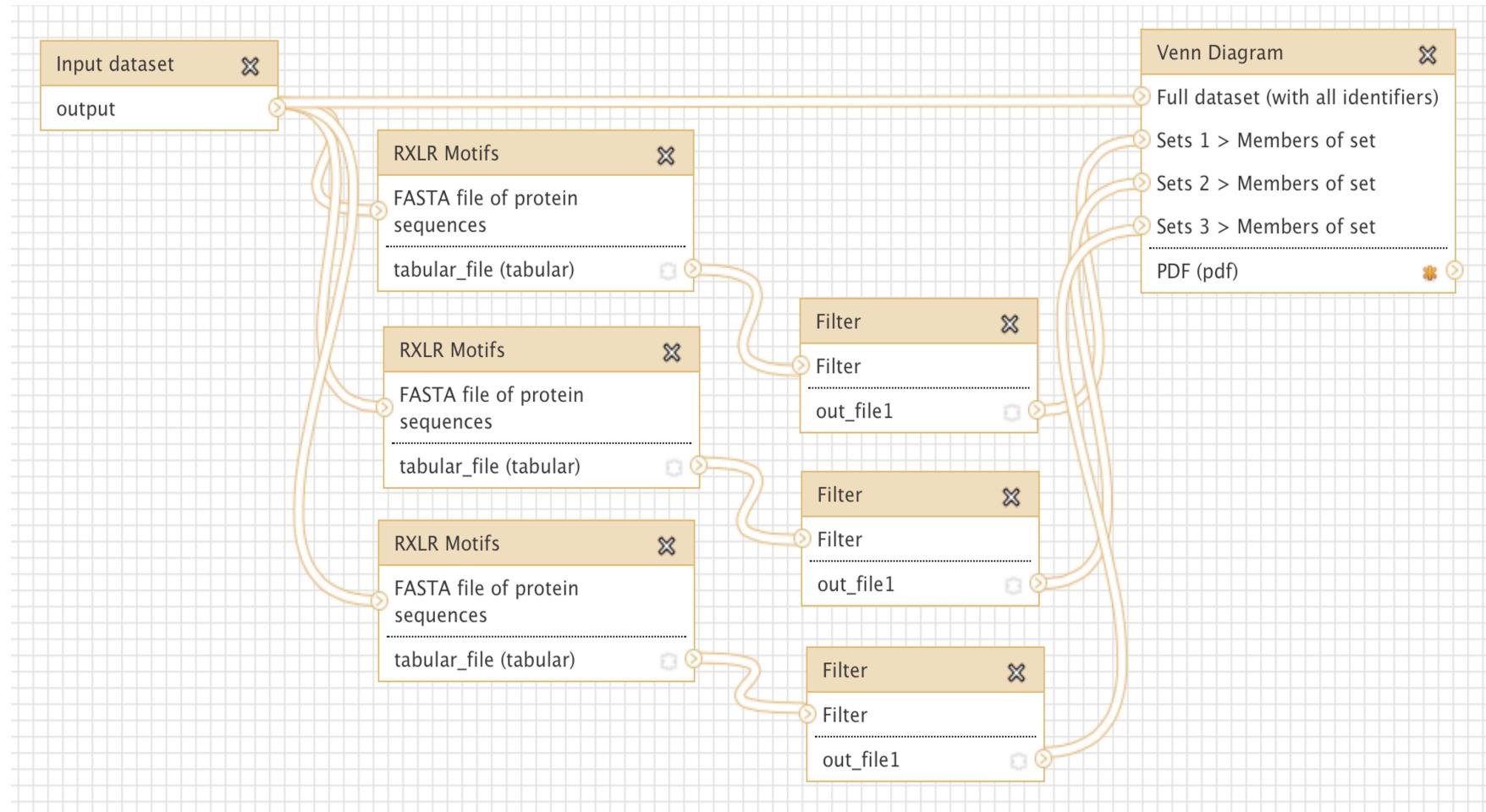
TOOLS

- Tools
- search tools
- Recently Used
- Get Data
- Send Data
- TOOLS
- Convert Formats
- Evolution
- ESTtik
- Filter and Sort
- Gene/Protein prediction
- SAT
- NGS: Quality Control
- NGS: Mapping
- NGS: SAM/BAM Manipulations
- NGS: SNP Detection
- Protein Structures
- Sequence comparisons
- UNTESTED TOOLS
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences

DONNÉES

| History | Options |
|------------------------------|----------|
| Unnamed history | 4.0 MB |
| 20: PhyML on data 18 | ✖️ ✖️ ✖️ |
| 19: PhyML on data 18 | ✖️ ✖️ ✖️ |
| 18: Fasta2Phyphil on data 16 | ✖️ ✖️ ✖️ |
| 17: Gblocks on data 15 | ✖️ ✖️ ✖️ |
| 16: Gblocks on data 15 | ✖️ ✖️ ✖️ |
| 15: MAFFT on data 14 | ✖️ ✖️ ✖️ |
| 14: new.fasta | ✖️ ✖️ ✖️ |

Interface : Galaxy



Interface : iPlant Collaborative

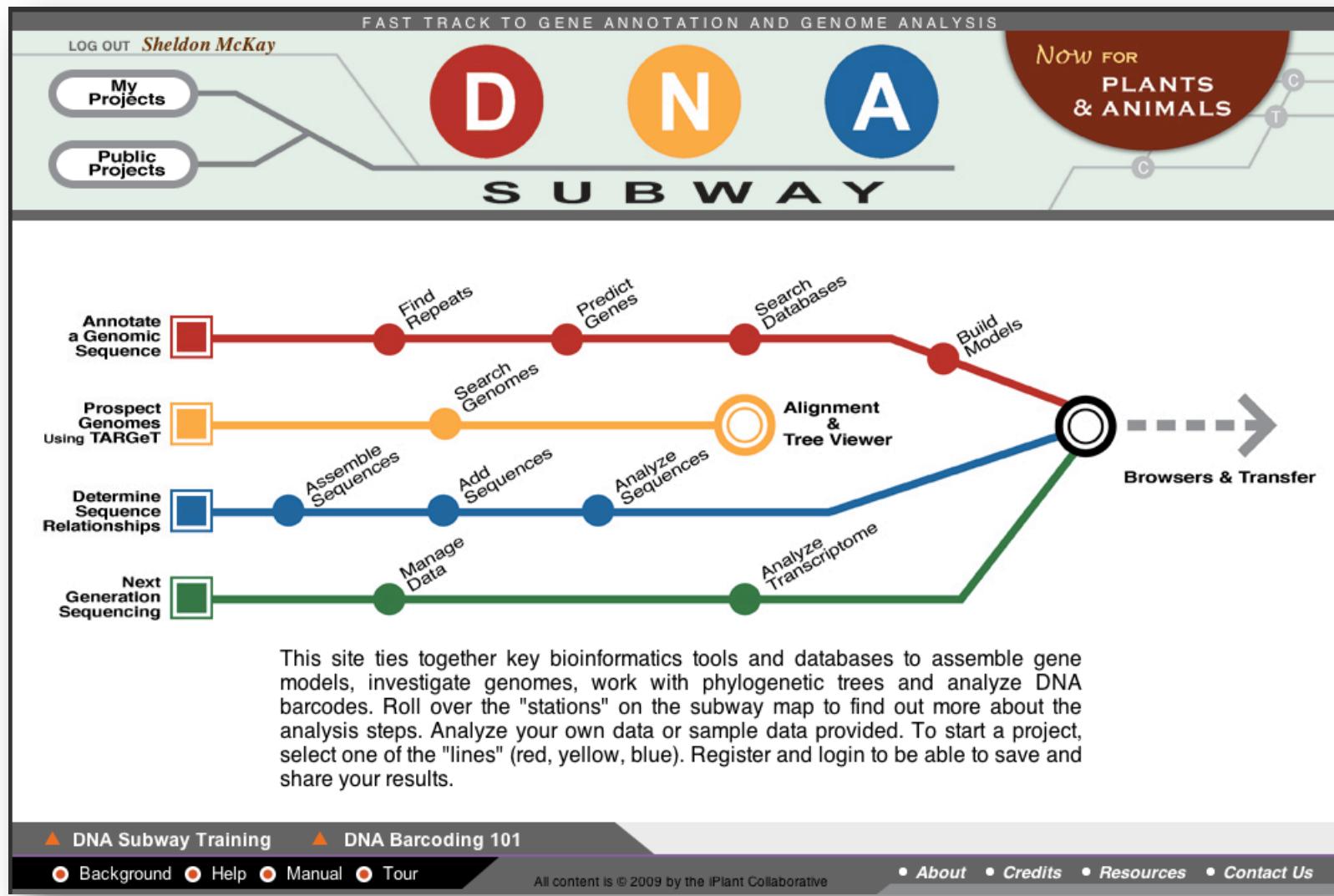
Simple d'utilisation

Enchainement de **brique** (programmes) de calcul.

4 grandes parties : annotation / recherche de gène apparentés / phylogénie / RNAseq.



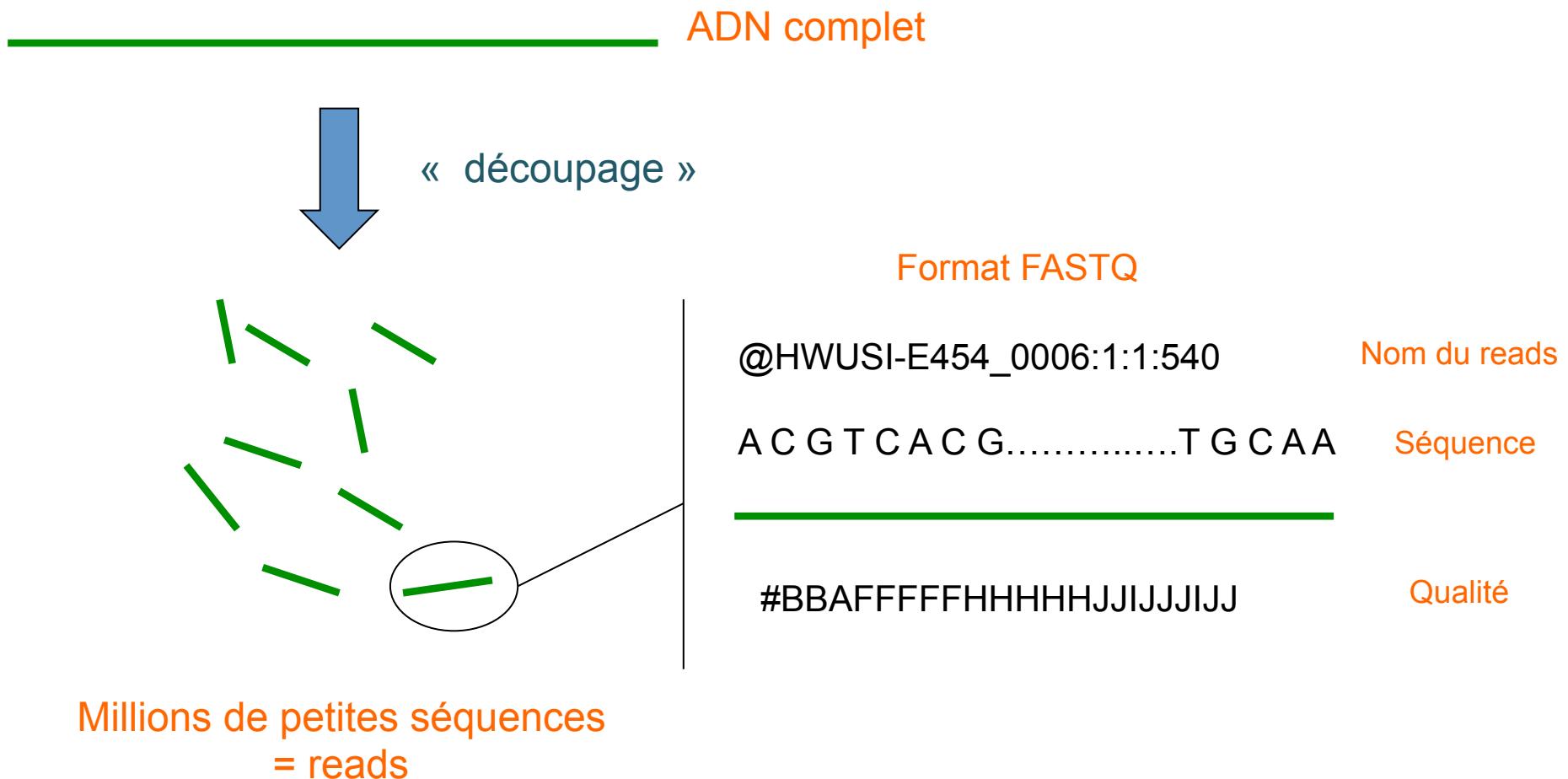
Interface : iPlant Collaborative



NGS : des reads aux SNPs

- Introduction NGS
 - Les reads : la donnée initiale du bio-informaticien
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

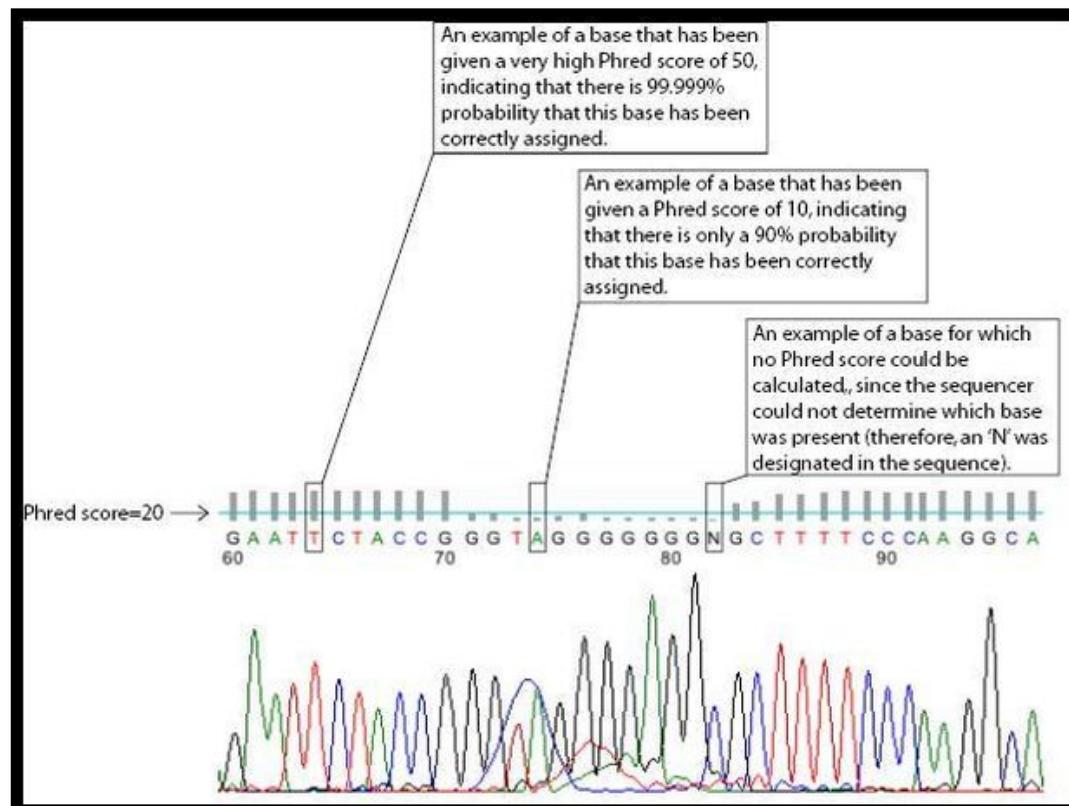
Qu'est ce qu'un Read ??



Qualité d'un nucléotide

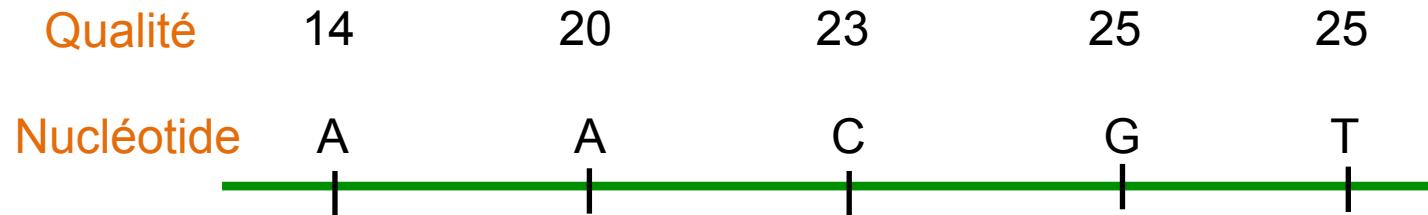
- Pour chaque nucléotide, un indice de qualité Q
- Q est obtenu à partir de la probabilité P que la base soit erronée : **Q = -10 log(P)**

| Q | P | Base call accuracy |
|----|-------------|--------------------|
| 10 | 1 in 10 | 90 % |
| 20 | 1 in 100 | 99 % |
| 30 | 1 in 1000 | 99.9 % |
| 40 | 1 in 10000 | 99.99 % |
| 50 | 1 in 100000 | 99.999 % |



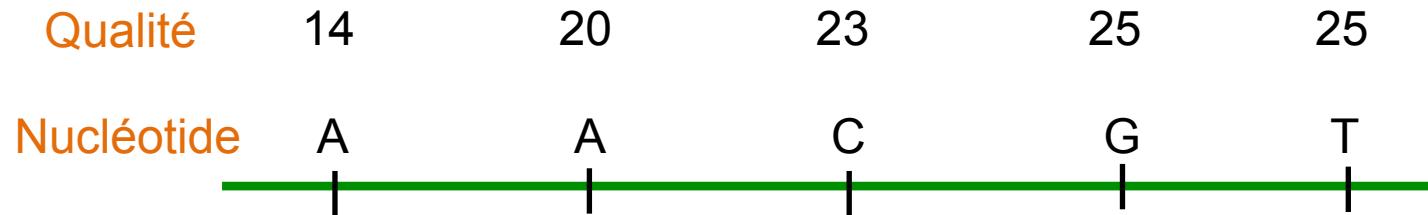
http://en.wikipedia.org/wiki/Phred_quality_score

Comment stocker cette information ?



Comment stocker cette information de manière compacte ???

Comment stocker cette information ?

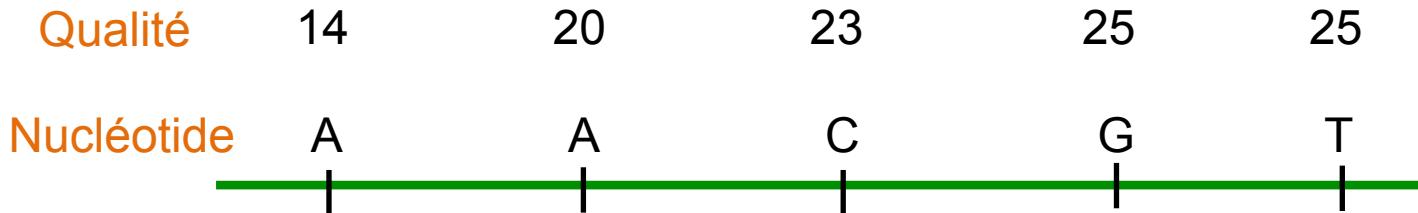


Comment stocker cette information de manière compacte ???

ACCGT
14 20 23 25 25

De manière encore plus
compacte ???

Comment stocker cette information ?



Comment stocker cette information de manière compacte ???

ACCGT
14 20 23 25 25

De manière encore plus
compacte ???

ACCGT
#BBAF

Table ASCII

| | | | | | | | | | | | | | | | |
|-----|----------------------------|-----|----|-----|---|-----|-------|-----|---|-----|---|-----|---|-----|---|
| 000 | NUL | 033 | ! | 066 | B | 099 | c | 132 | à | 165 | Ñ | 198 | ã | 231 | b |
| 001 | Start Of Header (SOH) | 034 | " | 067 | C | 100 | d | 133 | â | 166 | ¤ | 199 | Ã | 232 | þ |
| 002 | Start Of Text (STX) | 035 | # | 068 | D | 101 | e | 134 | ã | 167 | º | 200 |   | 233 |   |
| 003 | End Of Text (ETX) | 036 | \$ | 069 | E | 102 | f | 135 |   | 168 |   | 201 |   | 234 |   |
| 004 | End Of Transmission (EOT) | 037 | % | 070 | F | 103 | g | 136 |   | 169 |   | 202 |   | 235 |   |
| 005 | Enquiry | 038 | & | 071 | G | 104 | h | 137 |   | 170 |   | 203 |   | 236 |   |
| 006 | Acknowledge (ACK) | 039 | * | 072 | H | 105 | i | 138 |   | 171 |   | 204 |   | 237 |   |
| 007 | Bell | 040 | (| 073 | I | 106 | j | 139 |   | 172 |   | 205 | = | 238 | - |
| 008 | Backspace (BS) | 041 |) | 074 | J | 107 | k | 140 |   | 173 |   | 206 |   | 239 | - |
| 009 | Horizontal Tab | 042 | * | 075 | K | 108 | l | 141 |   | 174 |   | 207 |   | 240 | - |
| 010 | Line Feed (LF) | 043 | + | 076 | L | 109 | m | 142 |   | 175 |   | 208 |   | 241 |   |
| 011 | Vertical Tab | 044 | , | 077 | M | 110 | n | 143 |   | 176 |   | 209 |   | 242 | - |
| 012 | Form Feed (FF) | 045 | - | 078 | N | 111 | o | 144 |   | 177 |   | 210 |   | 243 |   |
| 013 | Carriage Return (CR) | 046 | . | 079 | O | 112 | p | 145 |   | 178 |   | 211 |   | 244 |   |
| 014 | Shift Out | 047 | / | 080 | P | 113 | q | 146 |   | 179 |   | 212 |   | 245 |   |
| 015 | Shift In | 048 | 0 | 081 | Q | 114 | r | 147 |   | 180 |   | 213 |   | 246 |   |
| 016 | Data Line Escape (DLE) | 049 | 1 | 082 | R | 115 | s | 148 |   | 181 |   | 214 |   | 247 |   |
| 017 | DC 1 (XON) | 050 | 2 | 083 | S | 116 | t | 149 |   | 182 |   | 215 |   | 248 |   |
| 018 | DC 2 | 051 | 3 | 084 | T | 117 | u | 150 |   | 183 |   | 216 |   | 249 | - |
| 019 | DC 3 (XOFF) | 052 | 4 | 085 | U | 118 | v | 151 |   | 184 |   | 217 |   | 250 | - |
| 020 | DC 4 | 053 | 5 | 086 | V | 119 | w | 152 |   | 185 |   | 218 |   | 251 | - |
| 021 | Negative Acknowledge (NAK) | 054 | 6 | 087 | W | 120 | x | 153 |   | 186 |   | 219 |   | 252 |   |
| 022 | Synchronous Idle | 055 | 7 | 088 | X | 121 | y | 154 |   | 187 |   | 220 |   | 253 |   |
| 023 | End Of Transmission Block | 056 | 8 | 089 | Y | 122 | z | 155 |   | 188 |   | 221 |   | 254 |   |
| 024 | Cancel | 057 | 9 | 090 | Z | 123 | { | 156 |   | 189 |   | 222 |   | 255 | - |
| 025 | End Of Medium | 058 | : | 091 | [| 124 |] | 157 |   | 190 |   | 223 |   | | |
| 026 | Substitute | 059 | , | 092 | \ | 125 | } | 158 |   | 191 |   | 224 |   | | |
| 027 | Escape (ESC) | 060 | < | 093 |] | 126 | - | 159 |   | 192 |   | 225 |   | | |
| 028 | File Separator | 061 | = | 094 | ^ | 127 | (DEL) | 160 |   | 193 |   | 226 |   | | |
| 029 | Group Separator | 062 | > | 095 | _ | 128 |   | 161 |   | 194 |   | 227 |   | | |
| 030 | Record Separator | 063 | ? | 096 | ` | 129 |   | 162 |   | 195 |   | 228 |   | | |
| 031 | Unit Separator | 064 | @ | 097 | a | 130 |   | 163 |   | 196 |   | 229 |   | | |
| 032 | SPACE(SP) | 065 | A | 098 | b | 131 |   | 164 |   | 197 | + | 230 |   | | |

Préparation des reads : nettoyage

Suppression ou masquage des bases de mauvaise qualité
(extrémité souvent)



- Programmes : Cutadapt, Trimmomatic

Préparation : Suppression Tag et Adaptateurs

Organisation d'un read :

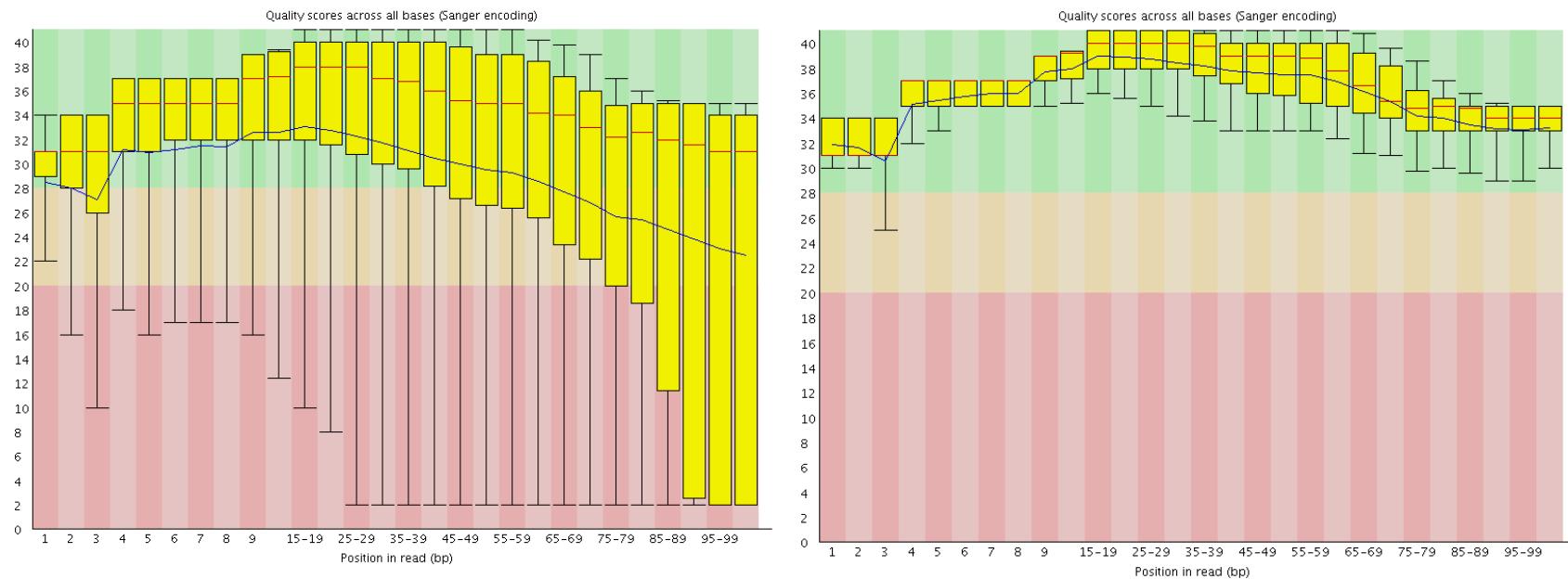


- Programmes : Cutadapt, Trimmomatic

Visualisation des reads : FastqC

Permet de s'assurer de la qualité des reads avant toute analyse. **Démonstration.**

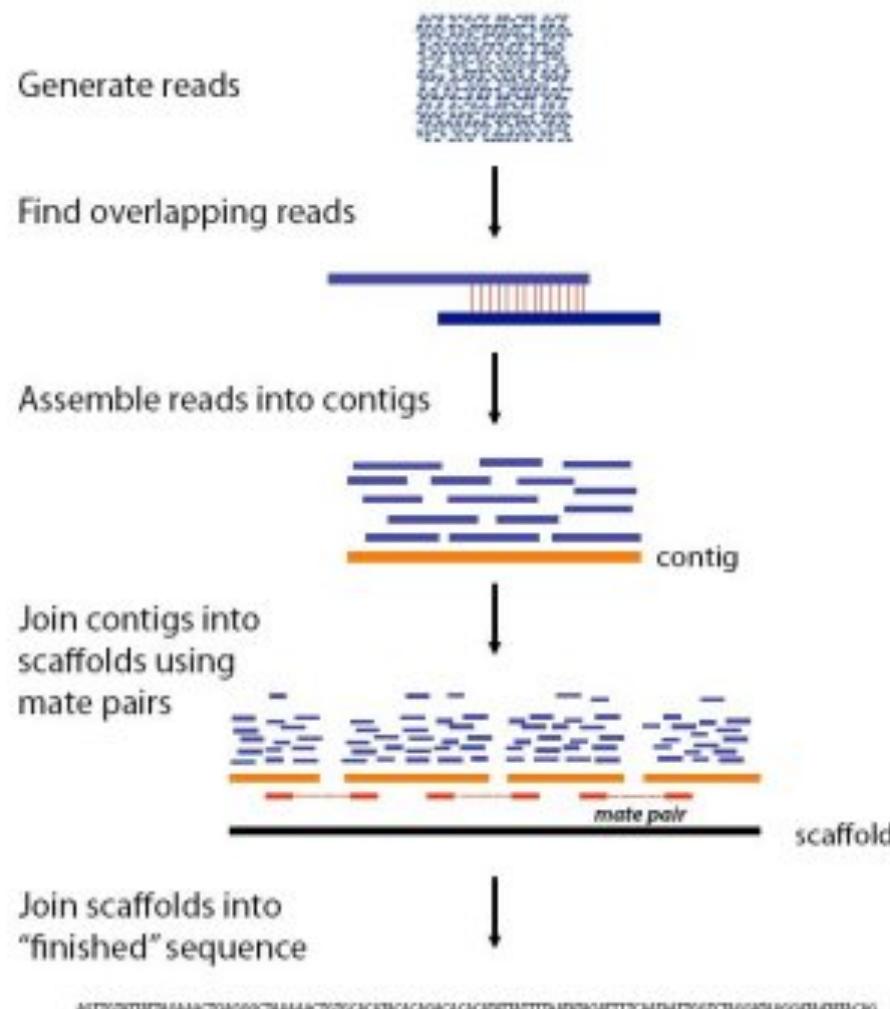
Exemple : Riz avant et après nettoyage



NGS : des reads aux SNPs

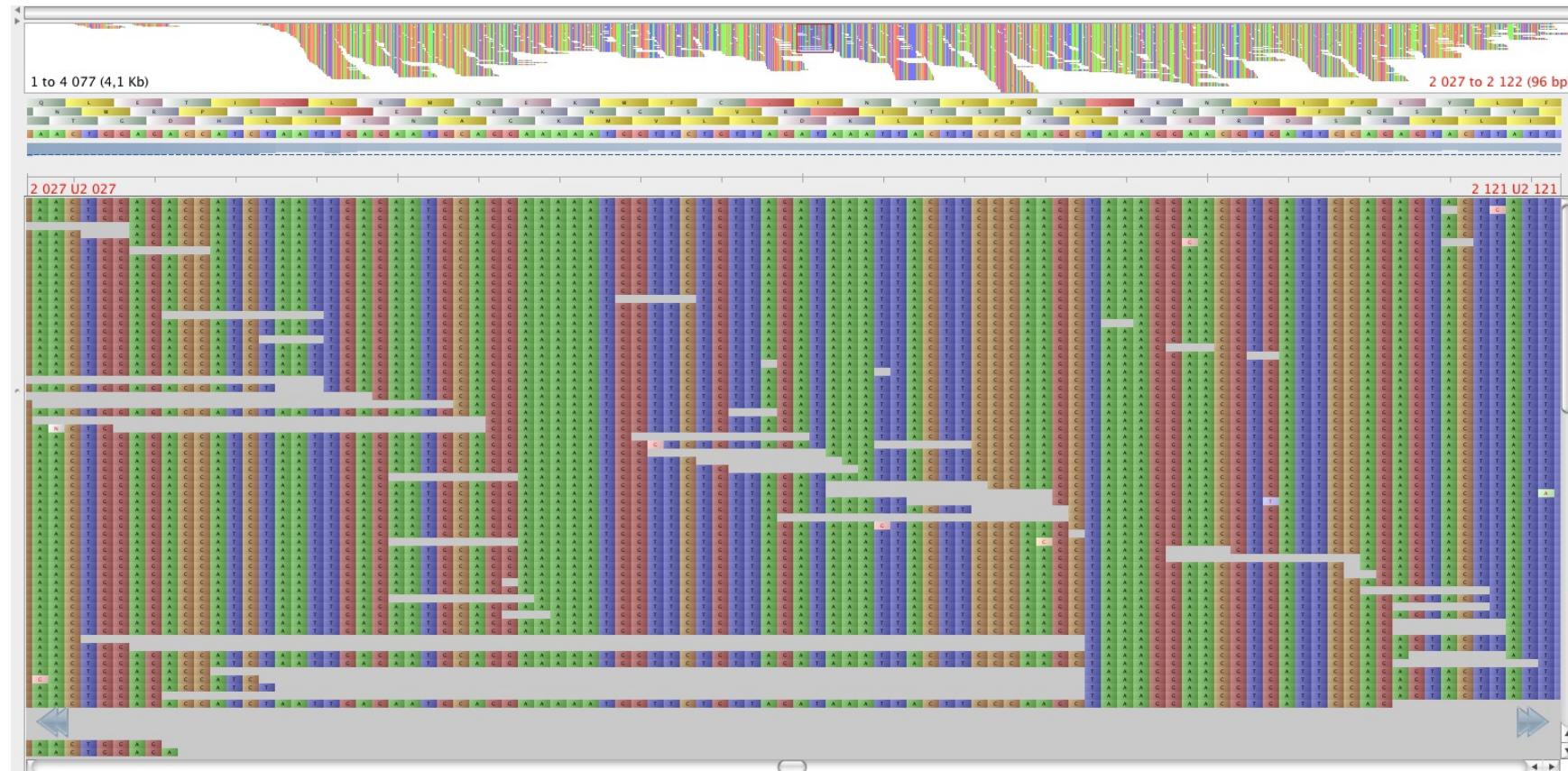
- Introduction NGS
- Les reads : la donnée initiale du bio-informaticien
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

Intuition, problématique



АКТУАЛЬНЫЕ ПРОБЛЕМЫ МАСТЕРСТВА И СОВРЕМЕННОСТИ КАК АСПЕКТЫ СОВРЕМЕННОЙ МАСТЕРСТВОВАНИЯ

Intuition, problématique



Programme : Tablet

Exemple

- Supposons que l'on séquence un tout petit fragment de génome avec des reads de 10pbs. On obtient 3 reads :

CTCCCTGTCA

GTCATCTGTC

ACCCTCCCTG

- Comment retrouver le génome de départ ? Quelle est la meilleure solution ?

Exemple

- Supposons que l'on séquence un tout petit fragment de génome avec des reads de 10pbs. On obtient 3 reads :

CTCCCTGTCA
GTCATCTGTC
ACCCTCCCTG

- Comment retrouver le génome de départ ? Quelle est la meilleure solution ?
 - Augmenter le **chevauchement**
 - Diminuer la **taille** de la séquence finale

CTCCCTGTCA
ACCCTCCCTG
GTCATCTGTC
CTCCCTGTCAACCCTCCCTGTCACTCTGTC

Solution 1

ACCCTCCCTG
CTCCCTGTCA
GTCATCTGTC
ACCCTCCCTGTCACTCTGTC

Solution 2

Premières approches gloutonnes

Consiste à construire le génome petit à petit en ajoutant un read après l'autre, en maximisant le chevauchement.

Exemple :

CTCCCTGTCA
GTCATCTGTC
ACCCTCCCT

Cas 1

CTCCCTGTCA



Premières approches gloutonnes

Consiste à construire le génome petit à petit en ajoutant un read après l'autre, en maximisant le chevauchement.

Exemple :

CTCCCTGTCA
GTCATCTGTC
ACCCTCCCT

Cas 1

GTCATCTGTC

CTCCCTGTCA

Premières approches gloutonnes

Consiste à construire le génome petit à petit en ajoutant un read après l'autre, en maximisant le chevauchement.

Exemple :

CTCCCTGTCA
GTCATCTGTC
ACCCTCCCT

Cas 1

GTCATCTGTC

CTCCCTGTCA

ACCCTCCCT

Premières approches gloutonnes

Consiste à construire le génome petit à petit en ajoutant un read après l'autre, en maximisant le chevauchement.

Exemple :

CTCCCTGTCA
GTCATCTGTC
ACCCTCCCT

Cas 1

GTCATCTGTC

CTCCCTGTCA

ACCCTCCCT

Cas 2

ACCCTCCCT

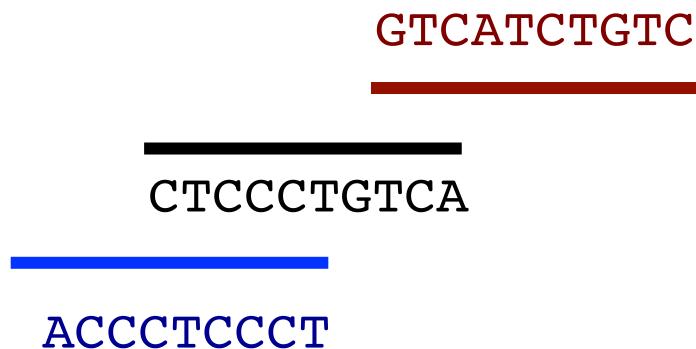
Premières approches gloutonnes

Consiste à construire le génome petit à petit en ajoutant un read après l'autre, en maximisant le chevauchement.

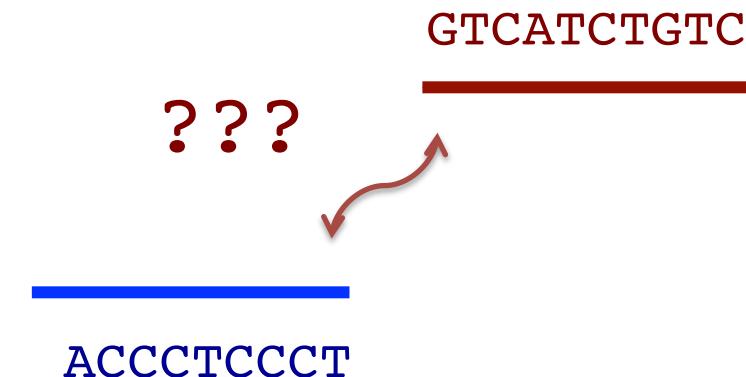
Exemple :

CTCCCTGTCA
GTCATCTGTC
ACCCTCCCT

Cas 1



Cas 2



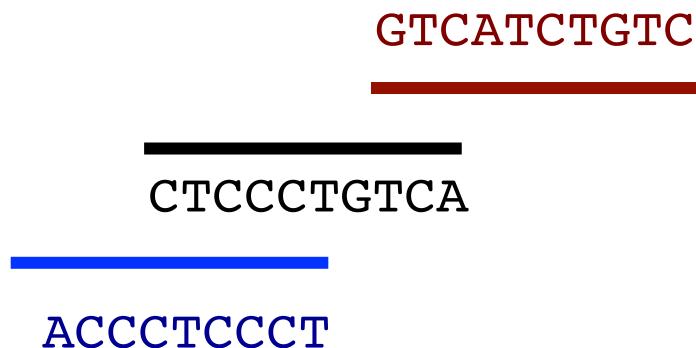
Premières approches gloutonnes

Consiste à construire le génome petit à petit en ajoutant un read après l'autre, en maximisant le chevauchement.

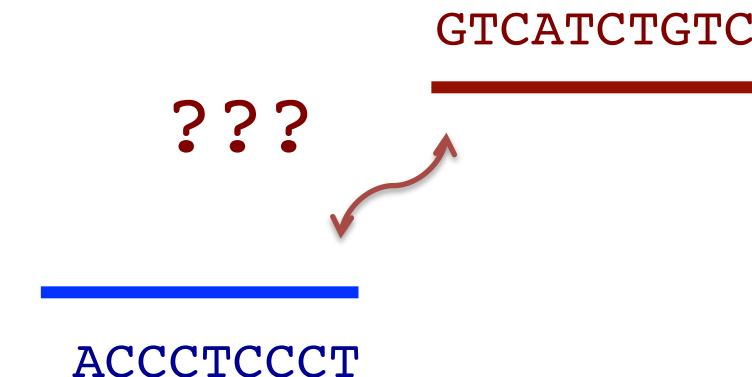
Exemple :

CTCCCTGTCA
GTCATCTGTC
ACCCTCCCT

Cas 1



Cas 2

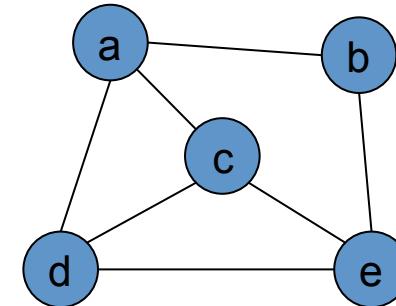


Problème : influence de l'initialisation !

Approche par graphe de chevauchement

Un graphe $G = (V, E)$ est composé de :

- V : un ensemble de sommets/nœuds (Vertices)
- E : un ensemble d'arcs/arêtes (Edges)



- Chaque reads peut être placé sur un nœud.
- Les reads qui se chevauchent sont reliés par une arête.
- On se rapporte à un problème d'algo **connu** mais **complexe** : trouver un chemin qui passe par toutes les arêtes.

Approche par graphe de chevauchement

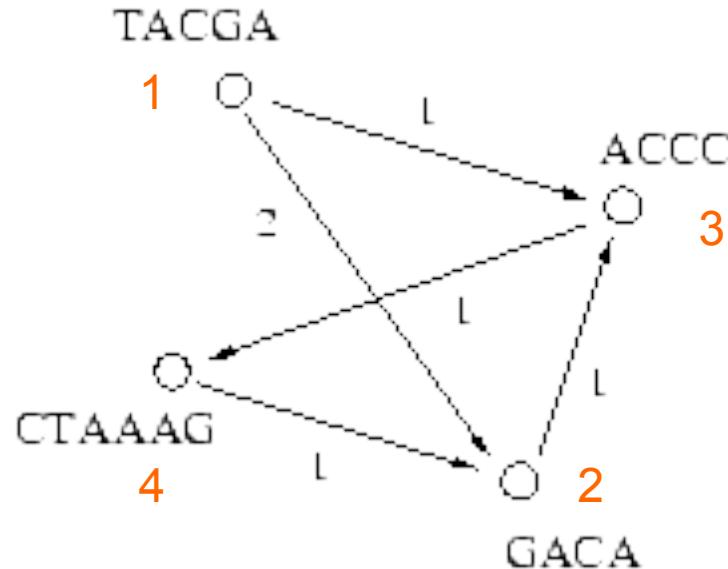
Exemple :

TACGA

ACCC

CTAAAG

GACA



→ On cherche un chemin qui passe **une et une seule fois** par chaque **sommet** et qui soit de poids maximal

TACGA
GACA CTAAAG
ACCC

Approche par graphe de chevauchement

Limites:

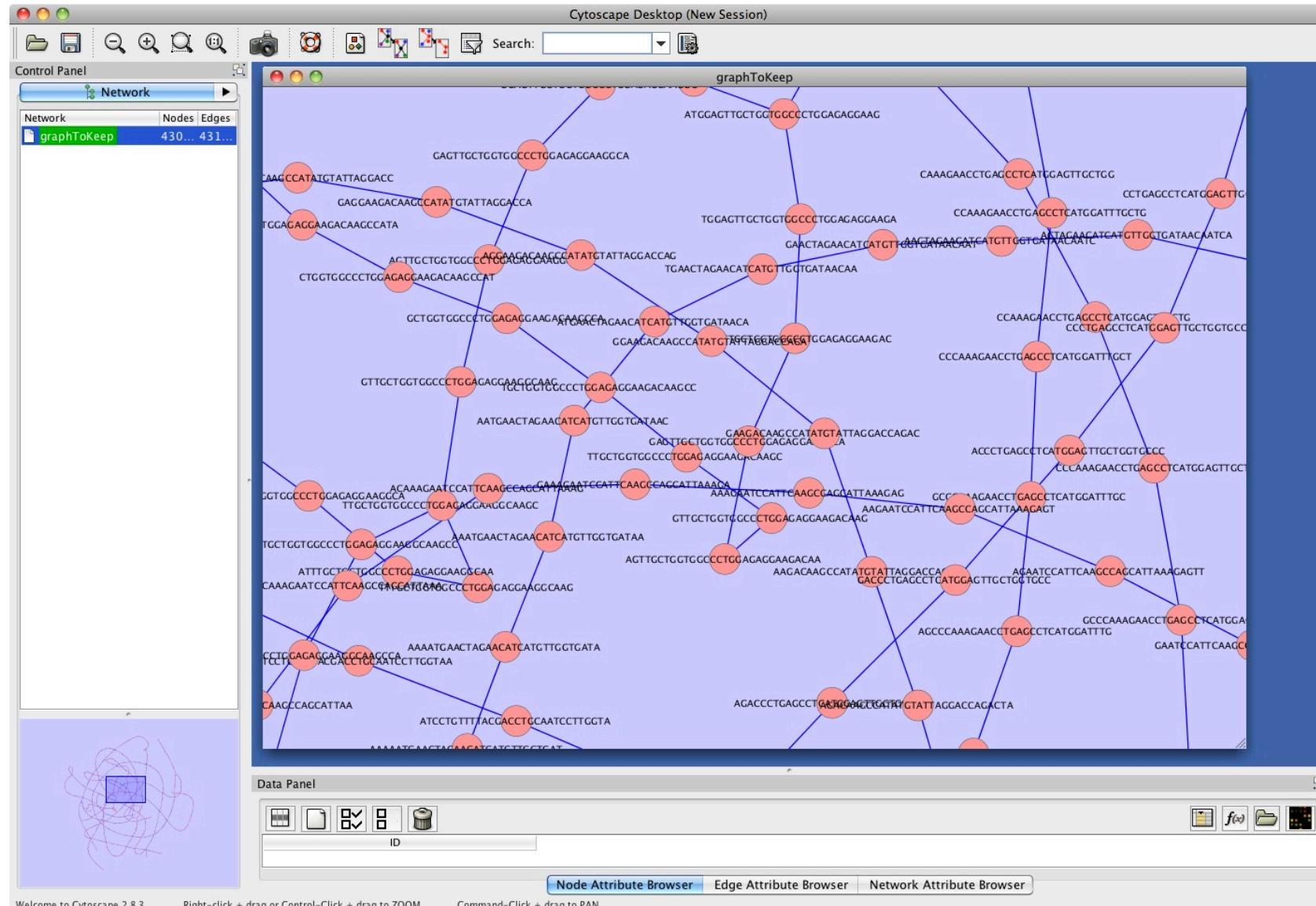
- Combien de comparaisons pour 1 million de reads ?
- La recherche d'un chemin passant une et une seule fois par chaque nœud est un problème difficile

| | | | | | | | |
|---------|------------|------------|----------------|------------|--------------------|------------|------------------------|
| $2^0 =$ | 1 | $2^{10} =$ | 1 024 | $2^{20} =$ | 1 048 576 | $2^{30} =$ | 1 073 741 824 |
| $2^1 =$ | 2 | $2^{11} =$ | 2 048 | $2^{21} =$ | 2 097 152 | $2^{31} =$ | 2 147 483 648 |
| $2^2 =$ | 4 | $2^{12} =$ | 4 096 | $2^{22} =$ | 4 194 304 | $2^{32} =$ | 4 294 967 296 |
| $2^3 =$ | 8 | $2^{13} =$ | 8 192 | $2^{23} =$ | 8 388 608 | $2^{33} =$ | 8 589 934 592 |
| $2^4 =$ | 16 | $2^{14} =$ | 16 384 | $2^{24} =$ | 16 777 216 | $2^{34} =$ | 17 179 869 184 |
| $2^5 =$ | 32 | $2^{15} =$ | 32 768 | $2^{25} =$ | 33 554 432 | $2^{35} =$ | 34 359 738 368 |
| $2^6 =$ | 64 | $2^{16} =$ | 65 536 | $2^{26} =$ | 67 108 864 | $2^{36} =$ | 68 719 476 736 |
| $2^7 =$ | 128 | $2^{17} =$ | 131 072 | $2^{27} =$ | 134 217 728 | $2^{37} =$ | 137 438 953 472 |
| $2^8 =$ | 256 | $2^{18} =$ | 262 144 | $2^{28} =$ | 268 435 456 | $2^{38} =$ | 274 877 906 944 |
| $2^9 =$ | 512 | $2^{19} =$ | 524 288 | $2^{29} =$ | 536 870 912 | $2^{39} =$ | 549 755 813 888 |

Idée :

- décomposer les reads en fragments de taille k fixée à l'avance : k-mer.
- Reads de taille 10 : combien de 7-mer ?
- K-mer de taille 10, combien de possibilités ?

Approche par graphe de chevauchement



Résultat de l'assemblage : le format fasta

Concrètement, on utilise des programmes connus : Abyss, Velvet, Cap3 ...

Au final on obtient un **génome** ou un **transcriptome** de référence que l'on va pouvoir utiliser pour la suite du pipeline.

Le format utilisé est le format **FASTA** :

```
>Cluster_36012|Contig1|less_than_5_individuals|original
AGCACACGGGCCCTGCCAAGCGCAGCAGCCCCAACACTCTTGCATCATCCAGGTTTTAGGTCAACCTGCTCGAT
GCACTGCCTGCCAGGATAATGAACAGGATCCAAGTGTGTCGTCCACCTCTTAGGAGAGTTCTATTATCGGCTGCA
AGTCGCCCTGCCACACCAGATTGTCCACCGCGGTGCCCTCCAAACAATGGATGACCTAGGTCTGGAGGTAGGAGACAGGT
CTGCAGCCTCTACCGTCGTTACCGGAATGCACCGGAGCTGATCCACGGATCAGGACGCCGGGACAAGTCGAGCCACG
>Cluster_36013|Contig1|less_than_5_individuals|original
AACAGCACCCATTATTATCAATACAAACCAAACCCATGGCCTGGGCAGGTTGAGAGGGCGGCAGCTAGTTCAAAGTTAAGGG
GAGGCACTTGCGGTGGCGGAGGCAGGCGCAGCCTCCTCTCAGGTCTCAAACCTGCCATGGTACGCTGCAGCTCGTC
TAGTACACCCCGAAGACGAACCCCGTCAGCCGCCACCACCGTTCTCGTCTGGGCCAGGCTGCCGAACCCCTGC
>Cluster_36014|Contig1|original
CGAATACTACCAAATGGAAATTTAGACAGATGGCTACACCACAACCCATGGAAGATGCTGAAAAGAGGACATTACATCTCC
AGCATCATCACTTGAATACCTCACCAACATAAGCCAATGCCAGTTATTCACTGTGATCTTAAGCCAAGCAATGTTCTCCTTG
TTTGGGCTTGCAGGTTGTACATCAAGACTCGGAGAAATCAACTAGTTGGGCATCAATGAGAGGCACAATAGGCTATGCCG
```

NGS : des reads aux SNPs

- Introduction NGS
- Les reads : la donnée initiale du bio-informaticien
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

Où trouver un génome de référence

Ensembl : biomart Database (Démonstration)

NCBI

Base de données spécifiques (IWGSC par exemple)

Problématique

On dispose d'un génome/transcriptome de référence

On veut positionner les reads dessus

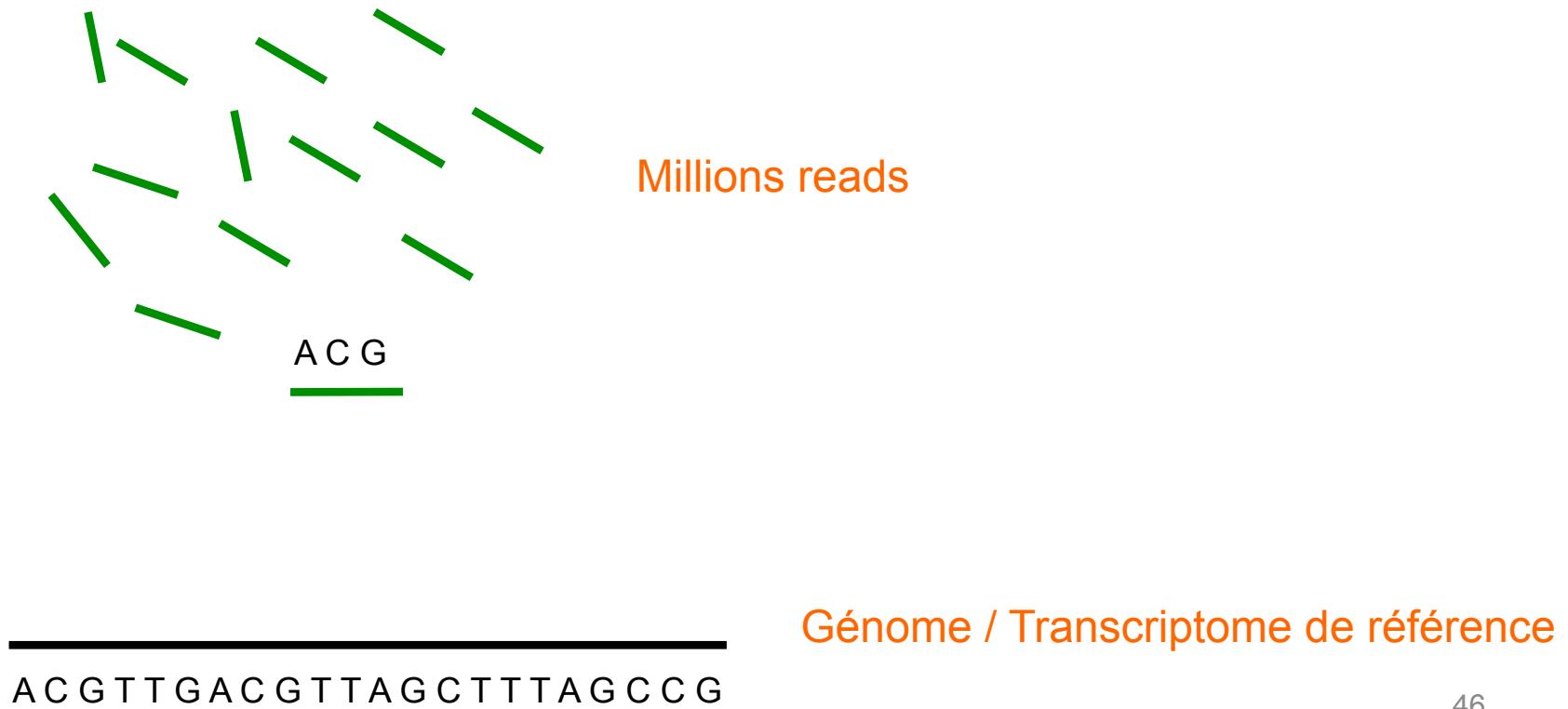
ACGTTGACGTTAGCTTAGCCG

Génome / Transcriptome de référence

Problématique

On dispose d'un génome/transcriptome de référence

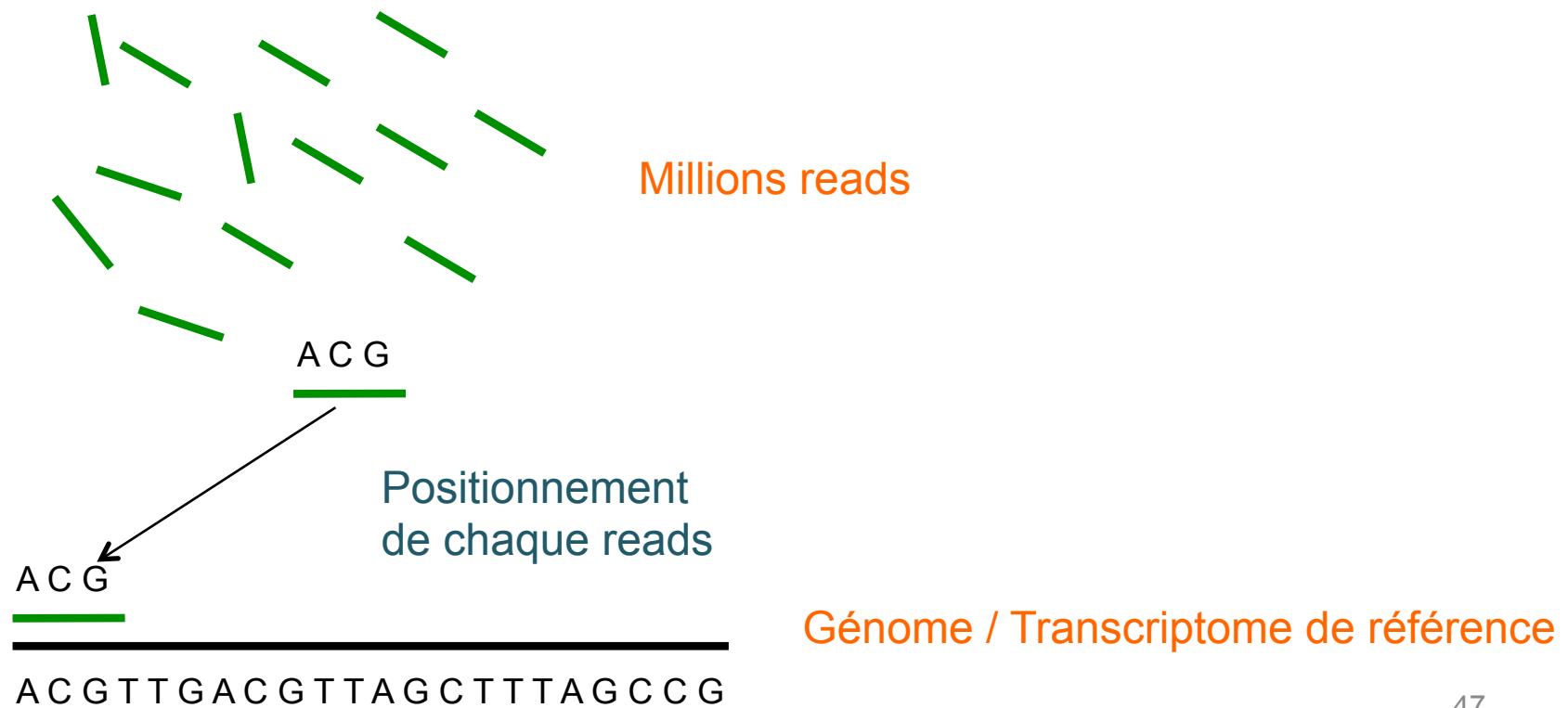
On veut positionner les reads dessus



Problématique

On dispose d'un génome/transcriptome de référence

On veut positionner les reads dessus



Problématique

On dispose d'un génome/transcriptome de référence

On veut positionner les reads dessus

Reads positionnés !

→ Mais concrètement, comment faire ??



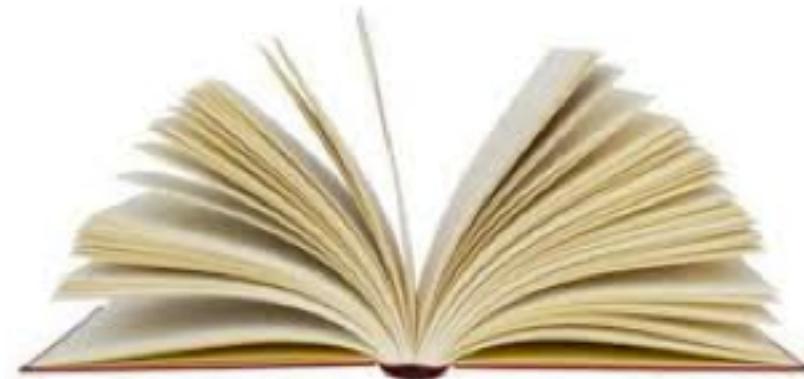
Approches Brute

-- Solution 1 --

Pour chaque read, Parcourir tout le génome et tester l'égalité read/portion génome.

→ Trop loooooong !!

Cela reviendrait à chercher dans un livre de 5000 pages où le mot « éléphant » apparaît 😞



Approches par graines et arbre de suffixe

-- Solution 2 --

Indexer le génome !!! → créer une sorte de « table des matières » :

- Découpage du génome en k-mer
- Stocker la position de chaque k-mer dans le génome !

Reads de 60 pb

- Prendre des K-mer de 60 ?
⇒ Lourd à stocker
- Prendre des K-mer de 20 ?
⇒ Tolérance aux erreurs
⇒ Mieux, mais reste lourd à stocker

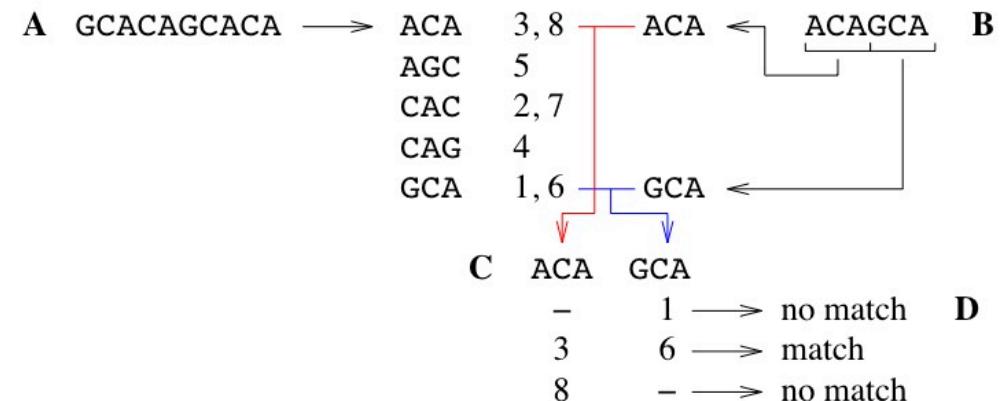
416 Index

Copyrighted Material

- chimpanzee 36, 76, 105, 107, 312, 291, 303
chlorophyll 173
chloroplasts 3, 11, 48, 92–93
chromatin 14–15
remodeling 14–15
immunoprecipitation 384
chronic myeloid leukaemia 34
cis-regulatory region 396
cladistics 148, 150
classification 73
clone 55, 66, 202, 214
CLUSTAL-W 238
clustering 148–149
clustering coefficient 406
CODIS (Combined DNA Index System) 225
co-evolution 138
co-expression patterns 385
codon usage 50
coiled-coil 354
Collins, F. 21
collusion-induced dissociation 323
Combined DNA Index System (CODIS) 225
common ancestor 75, 145, 156
last universal (LUCA) 156
comparative genomics 107, 377
complementary base pairing 20
complexity 82, 368
computational 373, P and NP 374
dynamic 371
static 371
computer science 43–44, 233, 373
concanavalin A 316
conformation-sensitive gel electrophoresis 220, 238
conformational angles 310
conformational change 360
Commonwealth Scientific Industrial and Research Organization (CSIRO) 87, 285
conjugation, bacterial 92
constitutive mutant 396
contig 30, 214
contig map 38
core of protein family 326
cost of DNA sequencing 23, 208
Cox, T.M. 229
CpG islands 54
creatine kinase 356
Crick, F.H.C. 9, 14, 19, 21, 228, 317
Critical Assessment of Structure Prediction (CASP) 337–338
crossing-over 29
Cryo-electron microscopy 382
CSIRO (Commonwealth Scientific Industrial and Research Organization) 87, 285
cyanobacteria 285
cyanobacteria, photosynthesis 173
cystic fibrosis 38, 124
cytochrome c 114, 117
DALI 233
Darwin, C.R. 3, 15–17, 73, 138, 142, 147, 346
databases 247
Dayhoff, M.O. 45, 240, 251
Delbrück, M. 18
deletion 13, 32–33, 95
deletion loop 34
delphinidin 285
depression 111
deuterostome 76
developmental expression pattern 5
changes in *Drosophila melanogaster* 281
desomethasone 272
diagnosis 269
diary 274–275, 402
dideoxynucleoside triphosphate 204–205, 207
dihydrokamferol 285
diphosphoglycerate (DPG) 334
directed evolution 346–347
diseases, protein aggregation associated 350
disulphide bridges 309
ditrosterol 318
DNA damage 402
DNA fingerprinting 37, 260, 221ff
DNA packaging 389
DNA polymerase 204
DNA sequencing 202–203
automated 207
cost of 208
mass spectrometry 209–210
Maxam–Gilbert method 207
pyrosequencing 209
Sanger method 207
DNA structure 18–20
 docking problem 332
dog
genome 182
domestication 182
breeds 185
domain 101, 312
domain recombination networks 385
dosage compensation 70
dot plot 235–238, 239
double helix 3

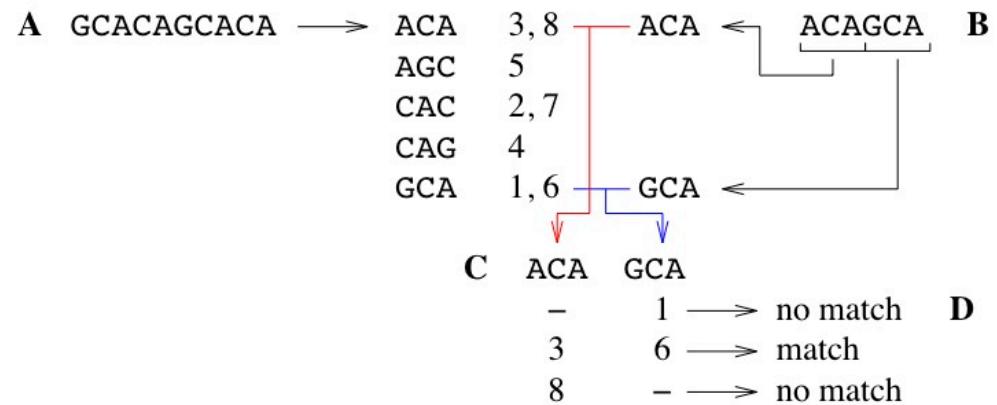
Approches par graines et arbre de suffixe

- Exemple
 - Génome GCACAGCACA
 - Reads ACAGCA
 - Indexation des 3-mers

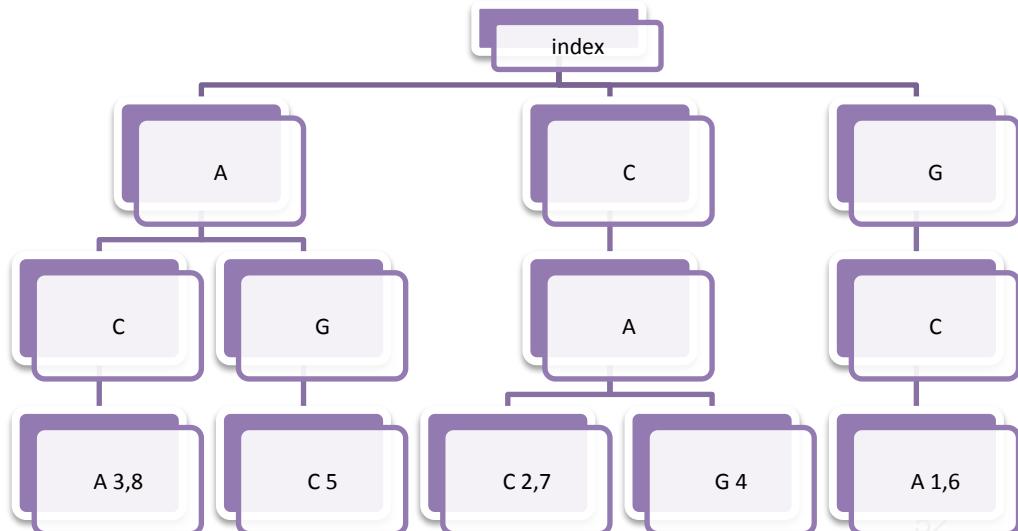


Approches par graines et arbre de suffixe

- **Exemple**
 - Génome GCACAGCACA
 - Reads ACAGCA
 - Indexation des 3-mers



Encore plus compressé :
12 lettres au lieu de 15 :
→ 20% de gain



Approches par table de suffixe

- Chercher dans un livre de 500 pages où le mot « séquence » apparaît 😞
- Chercher la définition de « séquence » dans un dictionnaire 😊. Car le dictionnaire est Ordonné !
- Idée : Indexer le génome sur lequel on mappe en une table de suffixe classé par ordre alphabétique !

Exemple : mapping sur le génome **CATTATTAGGA** ?

Approches par table de suffixe

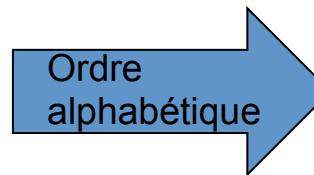
CATTATTAGGA

- 1 CATTATTAGGA
- 2 ATTATTAGGA
- 3 TTATTAGGA
- 4 TATTAGGA
- 5 ATTAGGA
- 6 TTAGGA
- 7 TAGGA
- 8 AGGA
- 9 GGA
- 10 GA
- 11 A

Approches par table de suffixe

CATTATTAGGA

- 1 CATTATTAGGA
- 2 ATTATTAGGA
- 3 TTATTAGGA
- 4 TATTAGGA
- 5 ATTAGGA
- 6 TTAGGA
- 7 TAGGA
- 8 AGGA
- 9 GGA
- 10 GA
- 11 A



- 1 11 A
- 2 8 AGGA
- 3 6 ATTAGGA
- 4 2 ATTATTAGGA
- 5 1 CATTATTAGGA
- 6 10 GA
- 7 9 GGA
- 8 7 TAGGA
- 9 4 TATTAGGA
- 10 6 TTAGGA
- 11 3 TTATTAGGA

Positions des occurrences de TTA, de GGT ?

Gain Important ?

Concrètement, comment réaliser un mapping ?

- Fichiers nécessaires en entrée :
 - Ensembles de fichiers « **Fastq** » compressés.
 - **Référence** sur laquelle on mappe au format **.fasta**
- Principaux programmes existant :
 - **Bwa**, **Bwa mem**
 - **Bowtie**
 - **Cushaw**
- Etapes clés :
 - Commande 1 : Crédit de l'**Index**
 - Commande 2 : **Mapping** des reads **R1** + des reads **R2**
 - Commande 3 : **Fusion** des 2 résultats.
 - Commande 4 : **Nettoyage** des résultats du mapping.

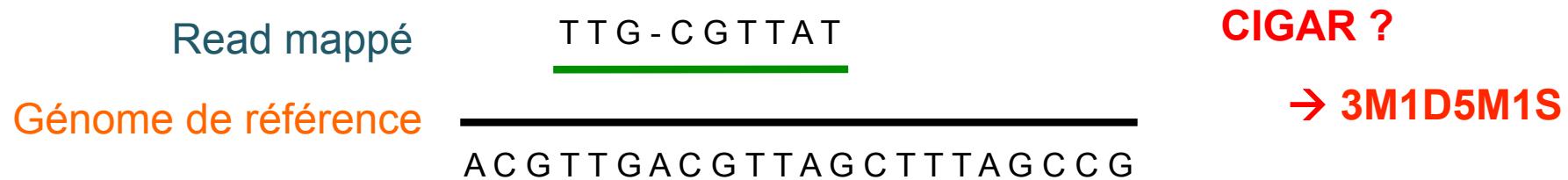
Format de sortie d'un mapping

- Les fichiers « SAM » (Sequence Alignment Map)
 - Fichier au format texte très lourd
 - Une ligne par read
 - Chaque ligne donne les résultat du mapping de ce read :
 - Nom du read
 - Nom du contig sur lequel a eu lieu le mapping
 - Position du reads sur la référence
 - Qualité du mapping
 - Description « CIGAR » du mapping etc.
- Les fichier « BAM »
 - Idem mais compressé !
 - On passe de l'un a l'autre en utilisant les « **samtools** »

Format de sortie d'un mapping :

- Format « CIGAR »

- Description des évènements associés à un mapping
- M=match or mutation | I=insertion | D=Délétion | S=substitution
- Exemple



■ Calcul de la qualité

- ⇒ Qualité = Phred scaled posterior probability = Q
- ⇒ Elle représente la probabilité que le positionnement du reads soit faux = P
- ⇒ $P = 10^{-Q/10}$
- ⇒ Exemple : si j'obtiens une qualité de 10, quelle est la proba que ce positionnement soit faux ?

Format de sortie d'un mapping :

- Format « CIGAR »

- Description des évènements associés à un mapping
- M=match or mutation | I=insertion | D=Délétion | S=substitution
- Exemple



■ Calcul de la qualité

- ⇒ Qualité = Phred scaled posterior probability = Q
- ⇒ Elle représente la probabilité que le positionnement du reads soit faux = P
- ⇒ $P = 10^{-Q/10}$
- ⇒ Exemple : si j'obtiens une qualité de 10, quelle est la proba que ce positionnement soit faux ?

Comment caractériser le mapping?

Samtools « Flagstat »

- Rappelle le nombre total de reads
- Donne le nombre de reads qui ont mappés
- Permet de voir si la référence utilisée était adéquate

Samtools IDXstat

- Donne pour chaque contig le nombre de reads attribués
- Très utile pour analyser l'expression des gènes (RNA-Seq)

NGS : des reads aux SNPs

- Introduction NGS
- Les reads : la donnée initiale du bio-informaticien
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

Mais qu'est ce qu'un SNP ?

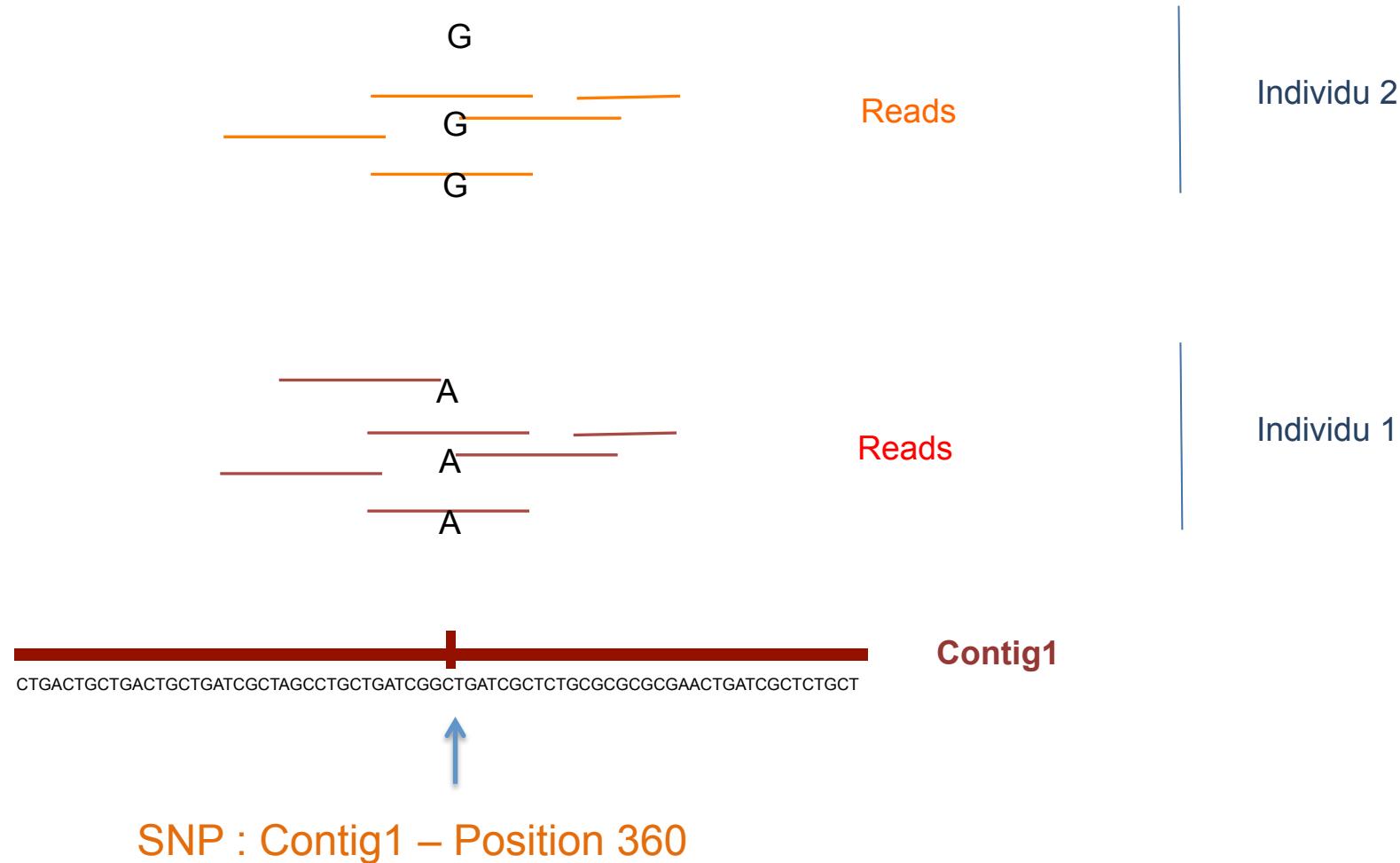
CTGACTGCTGACTGCTGATCGCTAGCCTGCTGATGGCTGATCGCTCTGCGCGCGCAACTGATCGCTTGCT

Contig1

Mais qu'est ce qu'un SNP ?



Mais qu'est ce qu'un SNP ?



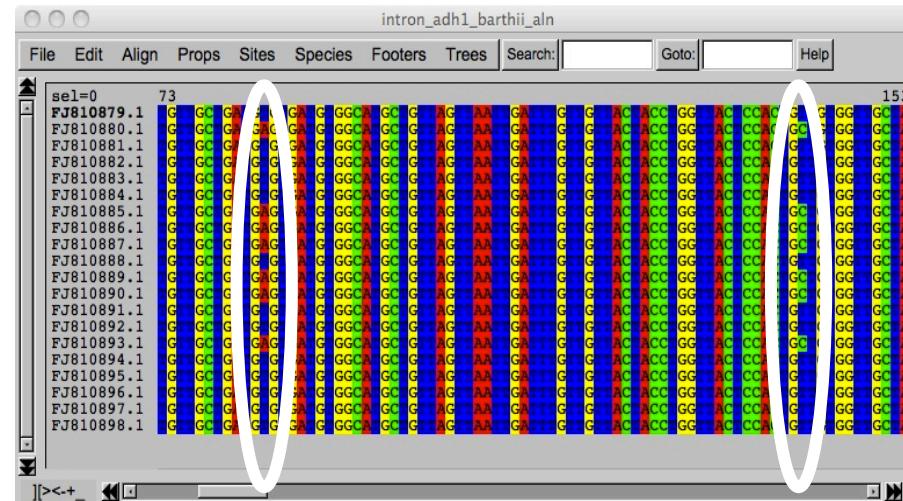
Approche par séquence consensus

On fait une séquence consensus par individus. Et on compare ces séquences



Approche par séquence consensus

Limite de l'approche ?



Indiv 1
TACAGTACGATC...

TACAGT
 AGTACGATC
 ATCAC
 CAGCACCG
 ACATCA

Indiv 2
TACAGTACGATC...

ACAGT
 AGTACGATC
 AGTACG
 ATCA
 TACAGCACCG
 ACATCA
 TCAC

| | |
|---|-------------|
| A | 02050400100 |
| C | 00300030001 |
| G | 00003002000 |
| T | 10002000010 |

Indiv 1 : TACAGTACGATC...
Indiv 2 : TACAGTACGATC...

| | |
|---|-------------|
| A | 03060600100 |
| C | 00300040001 |
| G | 00004003000 |
| T | 10003000010 |

Approche probabiliste

→ On donne le génotype le **plus probable** d'un génotype à une position. (Obtention d'un fichier VCF (GATK) ou ALR (Reads2SNP))

| | Individu 1 | | Individu 2 | | Individu 3 | |
|------------|--------------|---------|--------------|---------|-------------|--------|
| Contig1@23 | 14[10/2/0/0] | AA 0.91 | 14[1/35/0/0] | CC 0.99 | 14[1/2/0/0] | AC 0.6 |

Approche probabiliste

- On donne le génotype le **plus probable** d'un génotype à une position. (Obtention d'un fichier VCF (GATK) ou ALR (Reads2SNP))
- On filtre ensuite en fixant un seuil de **couverture** et de **probabilité** !

| | Individu 1 | | Individu 2 | | Individu 3 |
|------------|--------------|---------|--------------|---------|--------------------|
| Contig1@23 | 14[10/2/0/0] | AA 0.91 | 14[1/35/0/0] | CC 0.99 | 14[1/2/0/0] AC 0.6 |

Objectif final : une matrice de génotypage

Format final d'une table de génotypage : allèle A pour le parent 1, B pour le parent 2.

| SNP name | ind1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 10 | ... |
|-----------------|------|---|---|---|---|---|---|---|----|-----|
| Contig1@23 | A | A | B | B | A | A | B | B | - | AA |
| Contig18@42 | B | A | - | A | B | B | B | A | B | AA |
| Contig143@123 | A | A | B | B | A | A | A | B | B | - |
| Contig11123@294 | B | A | B | B | A | A | A | B | B | - |
| Contig11200@243 | A | A | A | B | A | A | A | A | B | AA |
| Contig2900@2 | B | A | - | A | B | B | B | A | B | AA |

Prêt à être utiliser pour différents types d'analyses !

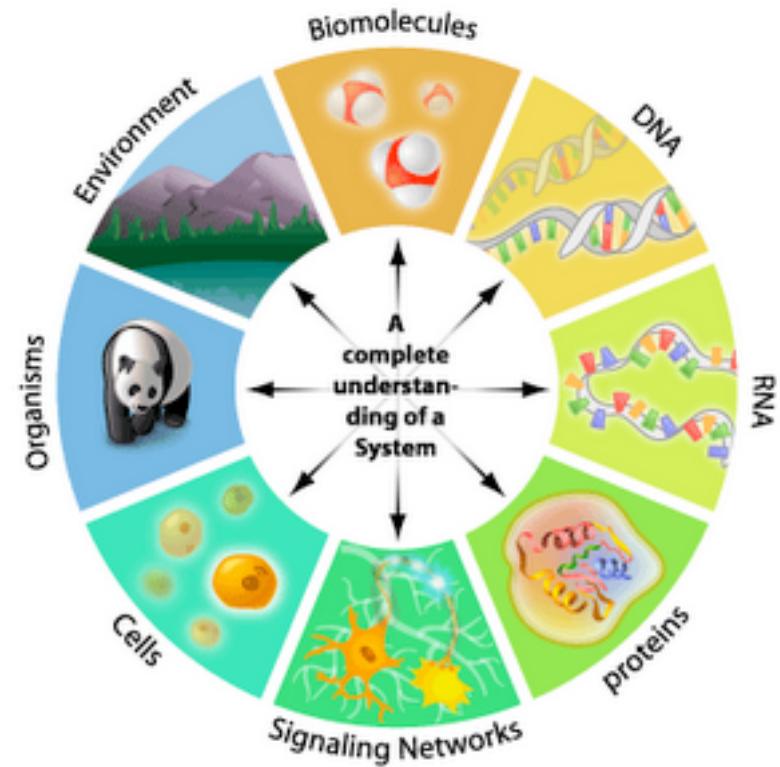
- Recherche de QTL ?
- Cartographie ?
- Structure de la pop ?
- ...

NGS : des reads aux SNPs

- Introduction NGS
- Les reads : la donnée initiale du bio-informaticien
- Assemblage de-novo
- Mapping sur un génome connu
- Détection de SNP
- Conclusions et discussions

La bioinfo c'est aussi ...

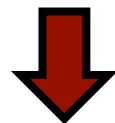
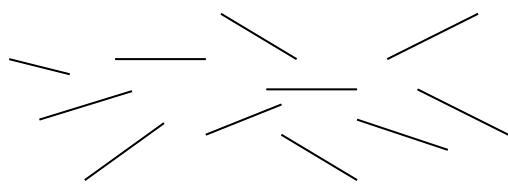
- Etude de l'expression
- Phylogénie
- Recherche de gènes apparentés
- Annotation des gènes
- Structure des protéines
- Analyses de réseaux
- Réalisation de cartes génétiques
- Recherche de QTLs
- Lien gène – ARN – Protéines
- Création d'algorithme
- Blast de séquences
- Etude de la structure d'une population
- Etude de l'évolution d'un espèce
- ...



Toujours des cas particuliers !

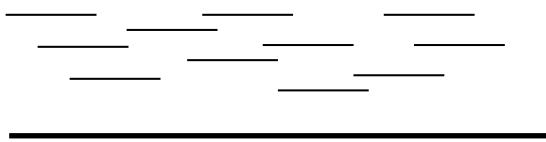
Diploidy

Transcripts of Gene 1



Assembling

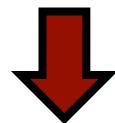
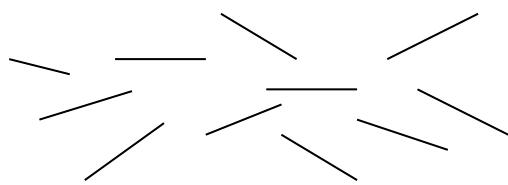
Contig of Gene 1



Toujours des cas particuliers !

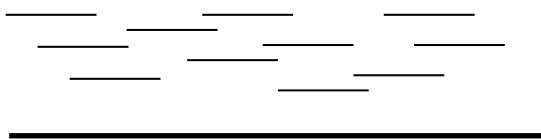
Diploidy

Transcripts of Gene 1



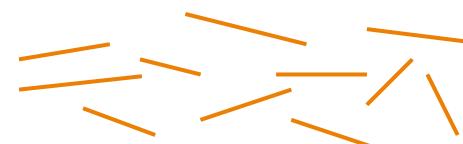
Assembling

Contig of Gene 1

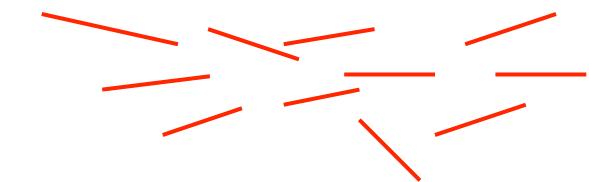


Polypliody

Transcripts of Gene 1
Genome A



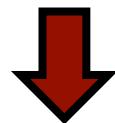
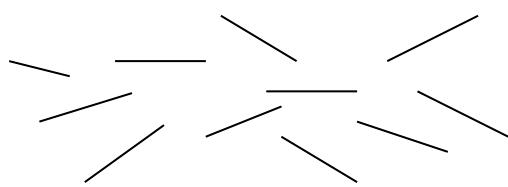
Transcripts of Gene 1
Genome B



Toujours des cas particuliers !

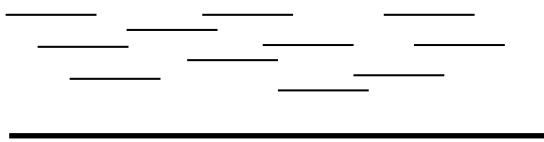
Diploidy

Transcripts of Gene 1



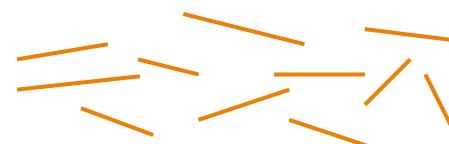
Assembling

Contig of Gene 1

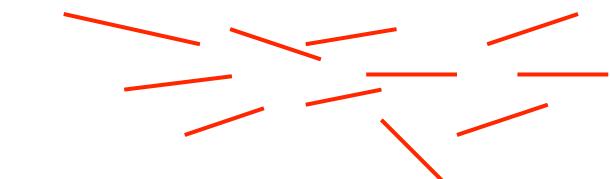


Polyploidy

Transcripts of Gene 1
Genome A



Transcripts of Gene 1
Genome B



Assembling

Chimeric Contig of Gene 1



Avant un projet NGS ?

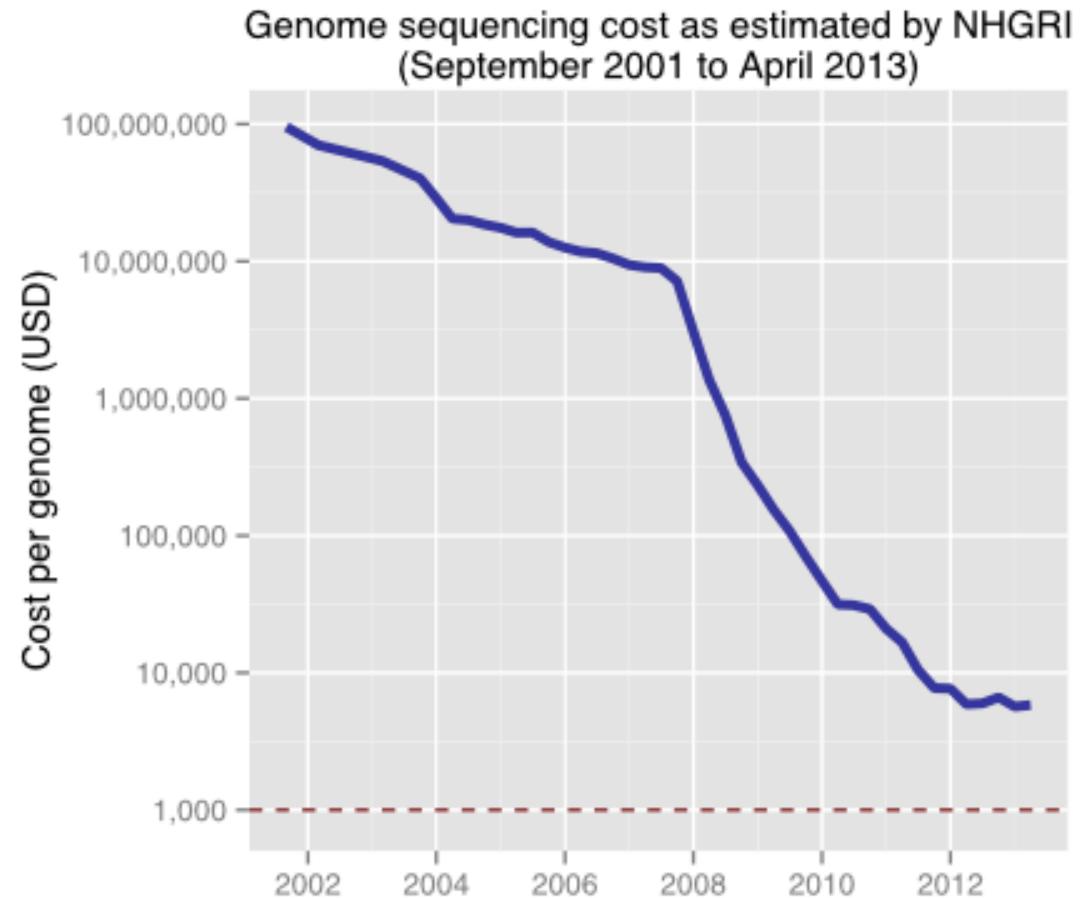
- Séquencer pour quoi faire ?
 - Clarifier la question à laquelle on veut répondre
- Séquencer qui ?
 - Choix des **individus**, du type **de tissu**, de la **date** de prélèvement
 - Garder les individus vivants ou au moins de l'ADN (**re-séquençage**)
- Séquencer quoi ?
 - **Génome** ? **Transcriptome** ? Les deux ?
- Séquencer combien ?
 - Nombre d'individus => puissance **statistique**
 - Nombre de reads par individu (**profondeur**) => fiabilité, chevauchement
 - Choix de la technologie ...
- Evaluer les ressources nécessaires
 - **CPU**, stockage et ... **temps humain** (CDD, partenaires, sous-traiter)

Un domaine d'avenir ?

Evolution des technologies de séquençages



Evolution de la demande en bioinformaticiens !!



Le plus simple a été fait ?

