2. TP: annotation de variantes génomiques

2.0 Description du projet

Dans cet exercice, vous suivez une série d'étapes typiques dans l'analyse de variantes génétiques par séquençage haut débit. Notamment, on veut analyser les génomes de lignées de levure, interrogés dans le cadre du projet de *Saccharomyces Genome Resequencing Project* (SGRP). Le projet combine des données de séquençage capillaire (ABI 370⋆, séquences longues, couverture bas) et de séquençage haut débit (séquences courtes, couverture haut). Vous aurez besoin de quelques gigaoctets d'espace sur le disque.

Ce TP est destiné à des équipes de 2 ou 3.

2.1 Tâches (20 points)

Assurez-vous que vous avez le génome de référence *Saccharomyces cerevisiæ* dans un fichier nommé `Scer.fa` (on a besoin de version 3, identique au fichier "Ultra-scaffolds" stocké à saccharomycessensustricto.org).

a. Choix de lignées. Étudiez le manuel de l'utilisateur de SGRP [https://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp_manual.pdf]. Tableaux 2 et 3 sur pages 8–9 montrent les lignées échantillonnées dans le projet :

Strain	Location	Notes	Seqd	Aligd	IGA
273614N	RVI, Newcastle UK		0.93	0.87	
322134S	RVI, Newcastle UK		0.98	0.91	
378604X	RVI, Newcastle UK		1.05	0.99	
BC187	Napa Valley, USA	spore from UCD2120	0.67	0.63	
DBVPG1106	Australia		0.69	0.64	
DBVPG1373	Netherlands		1.28	1.22	
DBVPG1788	Finland		1.18	1.08	

Légende : la lignée 273614N provient de Newcastle, avec un génome couvert par séquences longues à 0.93X (colonne «Seqd»), ajusté à 0.87X après alignement (colonne «Aligd») ; etc. La colonne «IGA» montre la couverture par séquences courtes.

Q62.5	London, UK		1.26	1.16	
Q69.8	London, UK				27.83
Q74.4	London, UK				1.77

Récemment, la deuxième phase du projet (**SGRP2**) a été complétée [Bergström et al. (2014) "A high-definition view of functional genetic variation from natural yeast genomes" *Molecular Biology and Evolution*, DOI :10.1093/molbev/msu037]. Les séquences Illumina générées dans le cadre du projet sont disponibles à ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/SGRP2/input/strains/.

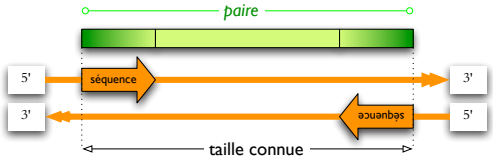
► Choisir (1) une lignée (**lignée** L_1) avec Aligd > 0 (séquençage capillaire de SGRP), et (2) une lignée (**lignée** L_2) avec données Illumina dans SGRP2. $L_1 = L_2$ est permis.

b1. Télécharger les données ABI (1 point). ► Télécharger les séquences longues (lignée L_1) : elles se trouvent dans le répertoire ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/latest/, emballées dans `cere_reads.tgz`

ou `para_reads.tgz` (fichiers de 500 Mo). Après extraction, vous trouverez les séquences dans le fichier `cere/strains/`L_1`/renamed.fastq` (ou `para/strains/`L_1`/renamed.fastq`).

b2. Télécharger les données Illumina (1 point). ► Télécharger les séquences courtes (lignée L_2) : il y a deux fichiers FASTQ à sauvegarder dans le répertoire de chaque lignée (p.e., `97A_Sc_Y55_1.fastq` et `97A_Sc_Y55_2.fastq` pour la lignée Y55).

c1. Séparation de séquences appariées (3 points). Le fichier `renamed.fa` de **b1** contient des séquences appariées (*mate pairs*), reconnues par les suffixes `.p1k` et `.q1k`.



► Mettez les séquences dans deux fichiers FASTQ : l'un pour des séquences `p1k`, l'autre pour des séquences `q1k`. Il faut assurer que toutes les séquences ont des paires : si le membre p1k ou q1k manque, enregistrez une séquence «nulle» dans son fichier : n fois la nucléotide 'N' avec qualité 0 (' `!`'). (La i-ème séquence dans l'un fichier doit former une paire avec la i-ème séquence dans l'autre fichier pour tout i.)

```
@273614N-10a02.p1k bases 1 to 50 (faked)
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+273614N-10a02.p1k
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```

d1. Alignement de séquences longues (5 points). Téléchargez et installez le logiciel SHRiMP [http://compbio.cs.toronto.edu/shrimp/]. Étudiez le fichier README. ► Lancer SHRiMP avec les séquences longues (les deux fichiers FASTQ de **c1** spécifiés avec arguments `-1` et `-2`) et sauvegarder le résultat dans un fichier SAM. Autres arguments à considérer :

`--qv-offset` Notre encodage est Phred+33, où qualité 0 est `!`, qualité 1 est `"` etc. (Consultez Wikipedia pour le format FASTQ.)

`--read-group` Les outils plus tard ont besoin d'un identificateur de «groupe» pour les séquences indiquant leur provenance. L'identificateur est votre choix mais il doit être unique. Le plus simple ici est d'utiliser le nom de la lignée (p.e., `--read-group=273614N,27361N`).

`--longest-read` La séquence la plus longue dépasse la valeur de défaut de SHRiMP. Choisissez une valeur assez grande pour les séquences.

`--insert-size-dist` La distribution de taille de clones pour les *mate pairs* a une moyenne de 4500 pb et un écart-type de 1000 pb [page 7, manuel SGRP].

`--isize` C'est la rangée de distance permise entre les membres d'un *mate pair* : ajustez le maximum comme nécessaire (p.e., 10000 pb).

d2. Alignement de séquences courtes (5 points). Téléchargez et installez le logiciel bwa [http://bio-bwa.sourceforge.net]. Étudiez la page de référence à http://bio-bwa.sourceforge.net/bwa.shtml. ► Faire les étapes d'alignement avec bwa : (1) indexage de la référence (`bwa index Scer.fa`), (2) alignement soit par la commande `mem` (avec les deux fichiers FASTQ), soit par la combinaison de `aln` (avec les deux fichiers séparamment) et `sampe`. Lancez l'alignement avec l'option `-R` pour spécifier le *read group* :
`-R '@RG\tID:IL29_4505\tSM:Y55\tPL:Illumina'` (ID est l'identificateur du groupe, SM donne l'échantillon — la lignée ici — et PL spécifie la plate-forme technologique).

e. Création de fichiers BAM (2 points). On a besoin de trier les alignements selon leur position au long des chromosomes. Téléchargez et installez le logiciel samtools [http://samtools.sourceforge.net]. Étudiez le mode d'emploi à http://samtools.sourceforge.net/samtools.shtml. ► Performez les étapes suivantes avec chacun des deux fichiers SAM produits en **d1** et **d2**. (1) Transformer dans le format BAM avec la commande `samtools view -b -h -o foo.bam -S foo.sam`. (2) Trier et indexer par `samtools sort` et `samtools index`.

f. Annotation de variantes (3 points). La dernière tâche est de déterminer les variantes (*variant calling*). Utiliser la commande `samtools mpileup` et l'outil `bcftools` (ce dernier vient avec samtools). Le plus simple est de se servir d'un tube (*pipe*) :

```
samtools mpileup -u -f Scer.fa foo.bam bar.bam | bcftools view -vcg - > foobar.vcf
```

► Comptez le nombre de variantes dans le génome entier. ► Créez un fichier VCF qui montre les variantes des lignées choisies dans deux régions intergéniques et deux introns (les régions analysées dans Devoir 1).

2.2 Exercices à option (≤ 20 points boni)

a. Automatisation (12 points boni). ► Implanter un ensemble d'outils qui permet d'exécuter §2.1.c1-d1-e (pour une lignée avec séquences longues) ou §2.1.b2-d2-e (pour une lignée avec séquences courtes) par la spécification simple de la lignée.

b. Annotation complète (8 points boni). ► Produire un fichier VCF qui montre les variantes de plusieurs lignées dans les introns de *S. cerevisiæ* (et non pas dans le génome entier). Utilisez la liste de tous les introns de Devoir 1. Notez que samtools permet de spécifier des régions pour l'analyse.

c. Amélioration de l'analyse (8 points boni). ► Ajuster les paramètres ou utiliser d'autres outils dans les étapes **d1**, **d2** ou **f**. Pour ce dernier, considérez FreeBayes [https://github.com/ekg/freebayes] ou GATK [http://www.broadinstitute.org/gatk/] qui permettent de spécifier la ploïdie (ici, on travaille avec des génomes haploïdes). Quantifiez la différence entre les VCFs résultant à l'aide du logiciel vcftools [http://vcftools.sourceforge.net].

2.3 Logistique

Remise. Remettez un rapport (format libre, en texte ou PDF) incluant un journal de ce que vous avez fait exactement : URLs de fichiers téléchargés, noms de fichiers locaux, paramètres à la ligne de commande, options choisies dans les logiciels. Donnez les statistiques suivantes : (1) nombre de séquences et nombre de bases par jeu de données téléchargé, (2) nombre de bases alignées à §2.1.d1 et d2, et (3) nombre de variantes [lignes] trouvées dans le génome entier à §2.1.f. Soumettez tout le code que vous avez implanté. Finalement, soumettez aussi le (petit) fichier VCF de §2.1.f avec les variantes dans quelques régions génomiques (intergéniques et introniques).

Mettez tous les fichiers dans un seul archive tar, et soumettez comme pièce jointe dans un e-mail à csuros@iro.umontreal. . . .

Date limite. Soumettez avant 21 :00 le 27 mars.