

Croisement de données

Bastien Job, BiGR, INSERM/IGR
Claire Toffano-Nioche, I2BC, Paris-Saclay
Matthias Zytnicki, MIAT, INRA

Contexte

Objectifs du cours

- Permettre de croiser des données génomiques de types différents
- Renforcer les notions vues précédemment
- Se familiariser avec bedtools
- Proposer un plan expérimental

Ce que le cours n'est pas

- Une réponse à une vraie question biologique
- Une méthode statistiquement valide

Le croisement de données

Qu'est-ce ?

- Une comparaison des positions ou intervalles génomiques:
 - un variant par rapport à des gènes d'intérêt,
 - des pics de ChIP-Seq par rapport à des gènes différentiellement exprimés, etc.

À quel moment est-ce valide ?

- Lorsque vous cherchez des co-occurrences.
- Lorsque vous donnez des distributions (de distance).

À quel moment est-ce douteux ?

- Lorsque les résultats sont présentés comme significatifs.

Problème 1

Question scientifique

Est-ce que mes gènes différentiellement exprimés contiennent plus de promoteurs actifs qu'attendu ?

Données (humaines)

- une liste de gènes différentiellement exprimés (herpes simplex virus type 1 infected HeLa cells with knockdown of beta-2-microglobulin)
- un fichier d'annotation (GRCh38.94, de Ensembl)
- les sites de fixation de H3K4me3 (UCSC genome browser)

`/shared/projects/ebaii2020/atelier_croisement/data`

À vous

Quel est le protocole ?

Un protocole possible

1. Extraire les intervalles génomiques des gènes différentiellement exprimés
2. Comparer les intervalles génomiques des gènes avec les régions H3K4me3
3. Compter le nombre de chevauchements
4. Trouver des gènes non-différentiellement exprimés
5. Comparer ces intervalles génomiques avec les régions H3K4me3
6. Compter le nombre de chevauchements
7. Comparer les nombres de chevauchements

Problème 1

Question scientifique

Est-ce que mes gènes différentiellement exprimés contiennent plus de promoteurs actifs qu'attendu ?

Données (humaines)

- une liste de gènes différentiellement exprimés (herpes simplex virus type 1 infected HeLa cells with knockdown of beta-2-microglobulin)
- un fichier d'annotation (GRCh38.94, de Ensembl)
- les sites de fixation de H3K4me3 (UCSC genome browser)

Données

gènes

ENSG00000004846

ENSG00000005981

ENSG00000006747

ENSG00000015568

ENSG00000047457

ENSG00000050628

ENSG00000057149

ENSG00000058085

ENSG00000071575

annotation

1	havana	pseudogene	11869	14409	.	+
---	--------	------------	-------	-------	---	---	---	-----

1	havana	lnc_RNA	11869	14409	.	+
---	--------	---------	-------	-------	---	---	---	-----

1	havana	exon	11869	12227	.	+
---	--------	------	-------	-------	---	---	---	-----

sites de fixation

22	16192349	16192565	region_1
----	----------	----------	----------

22	16846630	16870710	region_2
----	----------	----------	----------

22	17067019	17067283	region_3
----	----------	----------	----------

Avant toute chose...

Créez-vous un répertoire

```
cd ~
```

```
mkdir tp_croisement
```

```
cd tp_croisement
```

Emplacement des données

- Les données du TP sont sur

`/shared/projects/ebaii2020/atelier_croisement/data`

- Nous utiliserons également les banques fournies par l'infrastructure

`/shared/bank/homo_sapiens/GRCh38/gff3/Homo_sapiens.GRCh38.94.gff3`

Étape 1

Extraire les intervalles génomiques des gènes différentiellement exprimés

Données

- Liste de gènes
- Fichier d'annotation, au format GFF3

Résultat

- Gènes différentiellement exprimés, au format GFF3

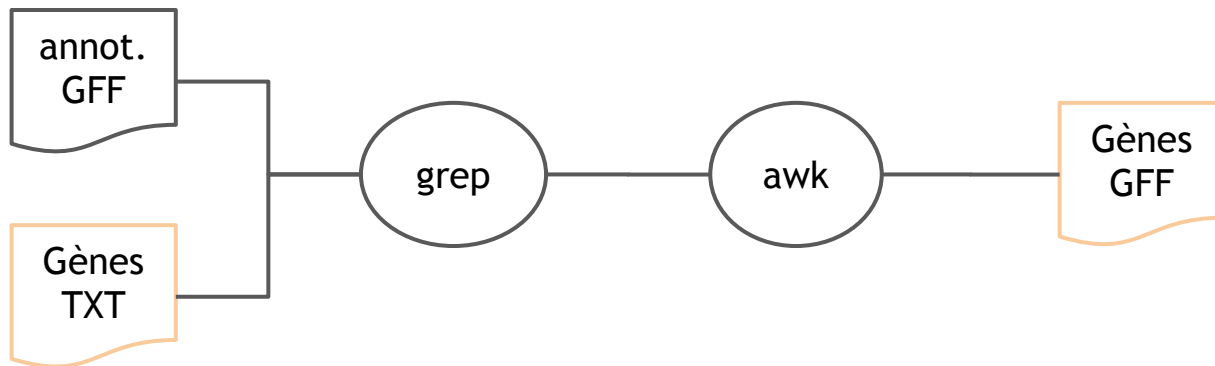
À vous

- Chercher une liste de mots dans un texte

```
grep -f liste_genes.txt annotation.gff > résultat.gff
```

- Se restreindre aux gènes

```
awk '($3 == "gene")' entrée.gff > sortie.gff
```



Étape 2

Comparer les intervalles génomiques des gènes avec les régions H3K4me3

Données

- Gènes différentiellement exprimés, au format GFF3
- Sites de fixation de H3K4me3, au format BED

Résultat

- Gènes différentiellement exprimés contenant des sites de fixation, au format GFF3

L'outil: bedtools

- Site web: <https://bedtools.readthedocs.io/>
- Format général

bedtools operation -a fichier1.bed [-b fichier2.bed] [options] -o resultat.bed

- Opérations les plus utilisées: intersect flank closest...
- Citation:

BEDTools: a flexible suite of utilities for comparing genomic features,
Aaron R. Quinlan, Ira M. Hall, 2010, Bioinformatics.

- Alternatives: gtftk, S-MART, etc.

À vous

- Cherchez et chargez les bedtools

`module avail`

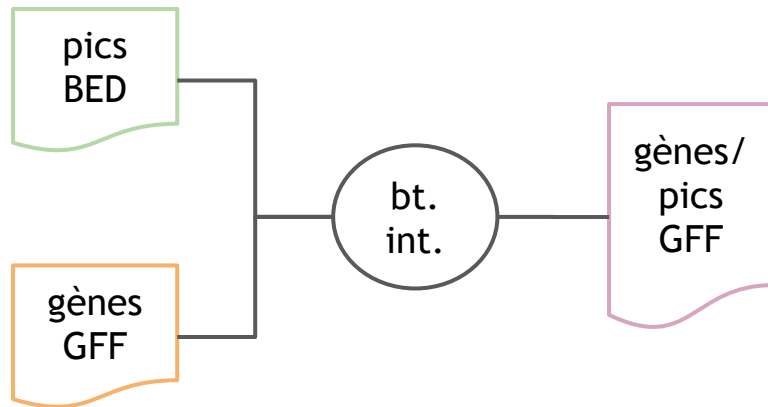
`module load bedtools/2.29.2`

- Voir la liste des opérations

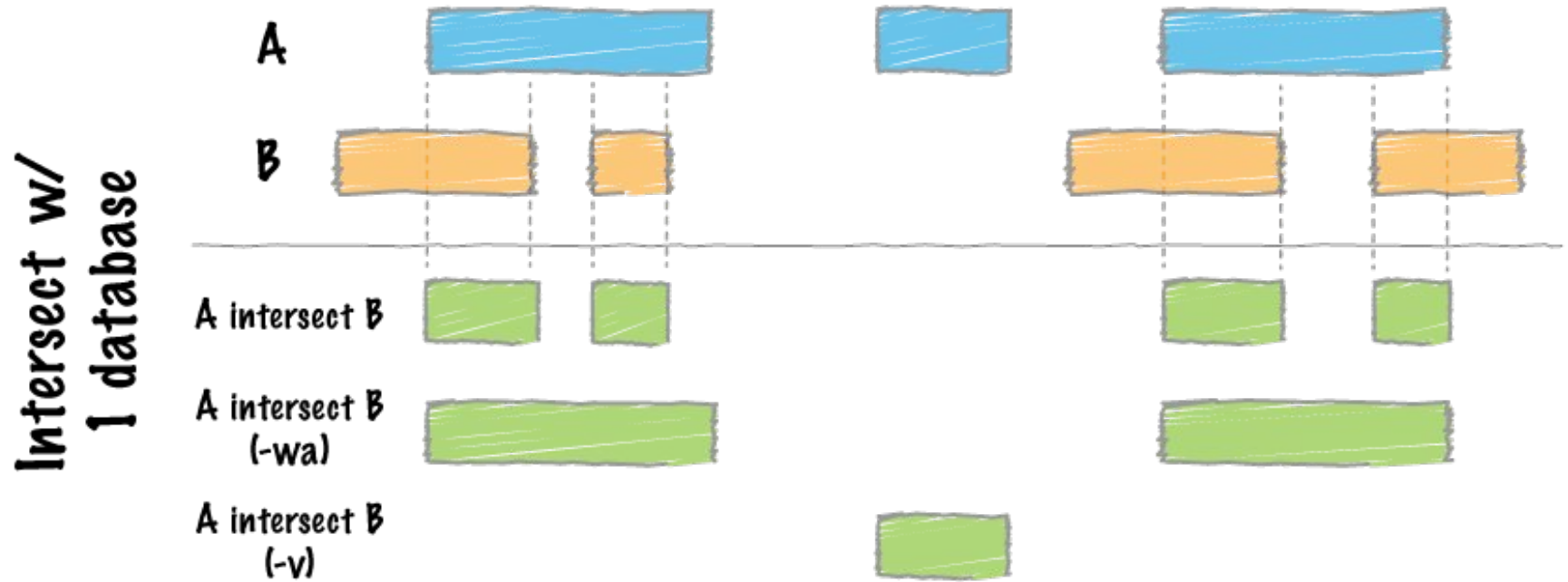
`bedtools`

- Voir la liste des options pour une opération

`bedtools intersect -h`



bedtools intersect



Étape 3

Compter le nombre de chevauchements

Données

- Gènes différentiellement exprimés contenant des sites de fixation, au format GFF3

Résultat

- Un nombre

Étape 4

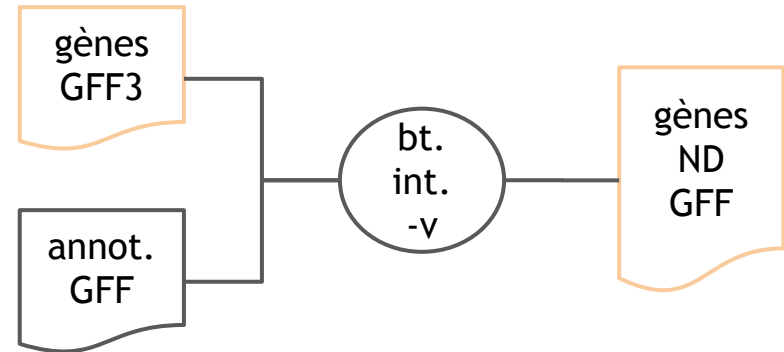
Trouver des gènes non-différentiellement exprimés

Données

- Gènes différentiellement exprimés, au format GFF3
- Fichier d'annotation, au format GFF3

Résultat

- Gènes non différentiellement exprimés, au format GFF3



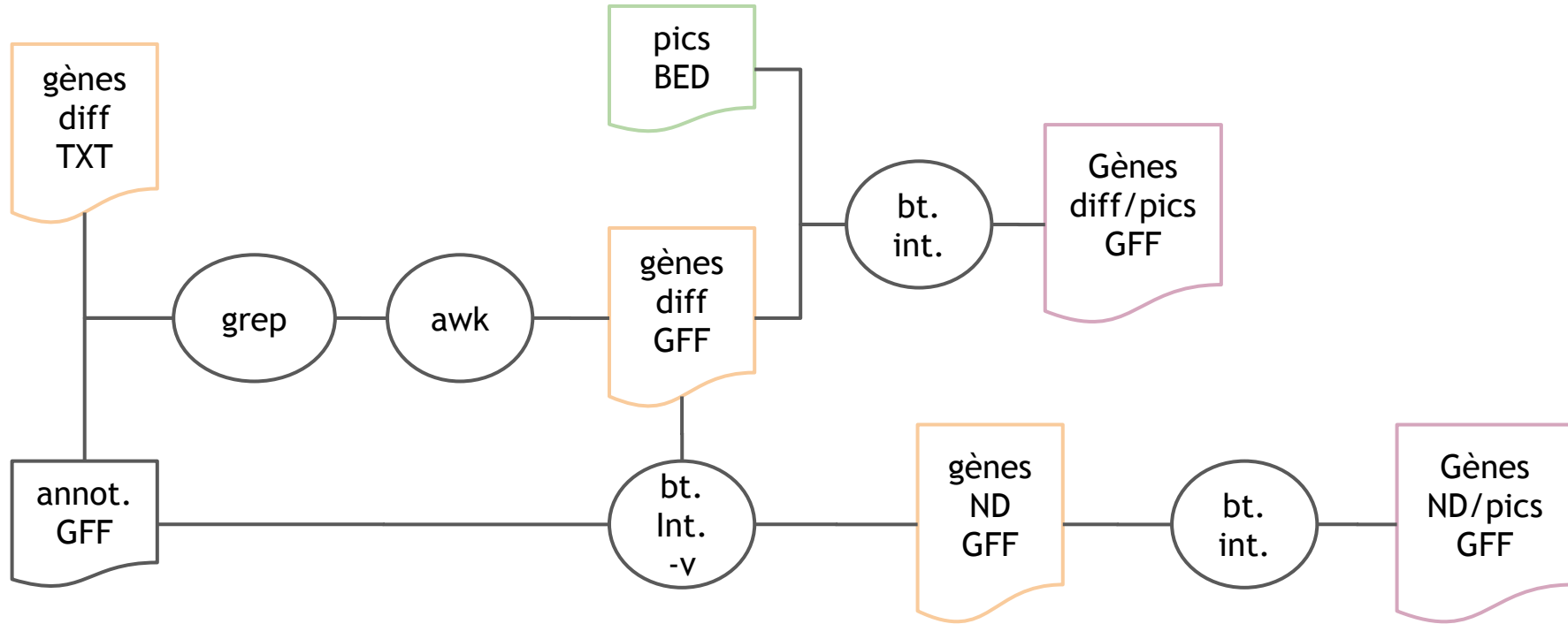
Étapes 5-7

Comparer ces intervalles génomiques avec les régions H3K4me3

Compter le nombre de chevauchements

Comparer les nombres de chevauchements

Workflow



Questions

- Le résultat était-il attendu ?
- Qu'est-ce que le *background* ?

Problème 2

Question scientifique

Quels sont les variants qui sont dans les promoteurs de mes gènes différentiellement exprimés ?

Données

- les gènes différentiellement exprimés, en GFF
- une liste de variants, au format VCF:

```
/shared/mfs/data/projects/ebaii2020/atelier_croisement/data/common  
_all_20180418_div.vcf
```

Données

variants

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	10177	rs367896724	A	AC	.	.	RS=367896724;RSPPOS=10177;dbSNPBuildID=138;SSR=0;SAO=0;VP=0x050000020005170026000200;GENEINFO=DDX11L1:100287102;WGT=1;VC=DIV;R5;ASP;VLD;G5A;G5;KGPhase3;CAF=0.5747,0.4253;COMMON=1;TOPMED=0.76728147298674821,0.23271852701325178
1	10352	rs555500075	T	TA	.	.	RS=555500075;RSPPOS=10352;dbSNPBuildID=142;SSR=0;SAO=0;VP=0x050000020005170026000200;GENEINFO=DDX11L1:100287102;WGT=1;VC=DIV;R5;ASP;VLD;G5A;G5;KGPhase3;CAF=0.5625,0.4375;COMMON=1;TOPMED=0.86356396534148827,0.13643603465851172
1	10616	rs376342519	CCGCCGTTGCAAAGGCGCGCCG	C	.	.	RS=376342519;RSPPOS=10617;dbSNPBuildID=142;SSR=0;SAO=0;VP=0x050000020005040026000200;GENEINFO=DDX11L1:100287102;WGT=1;VC=DIV;R5;ASP;VLD;KGPhase3;CAF=0.006989,0.993;COMMO

Un protocole possible

1. Extraire la région située à 2kb en amont des gènes
2. Trouver tous les variants chevauchant les régions trouvées précédemment

Étape 1

Extraire la région située à 2kb en amont des gènes

Nouvelle opération

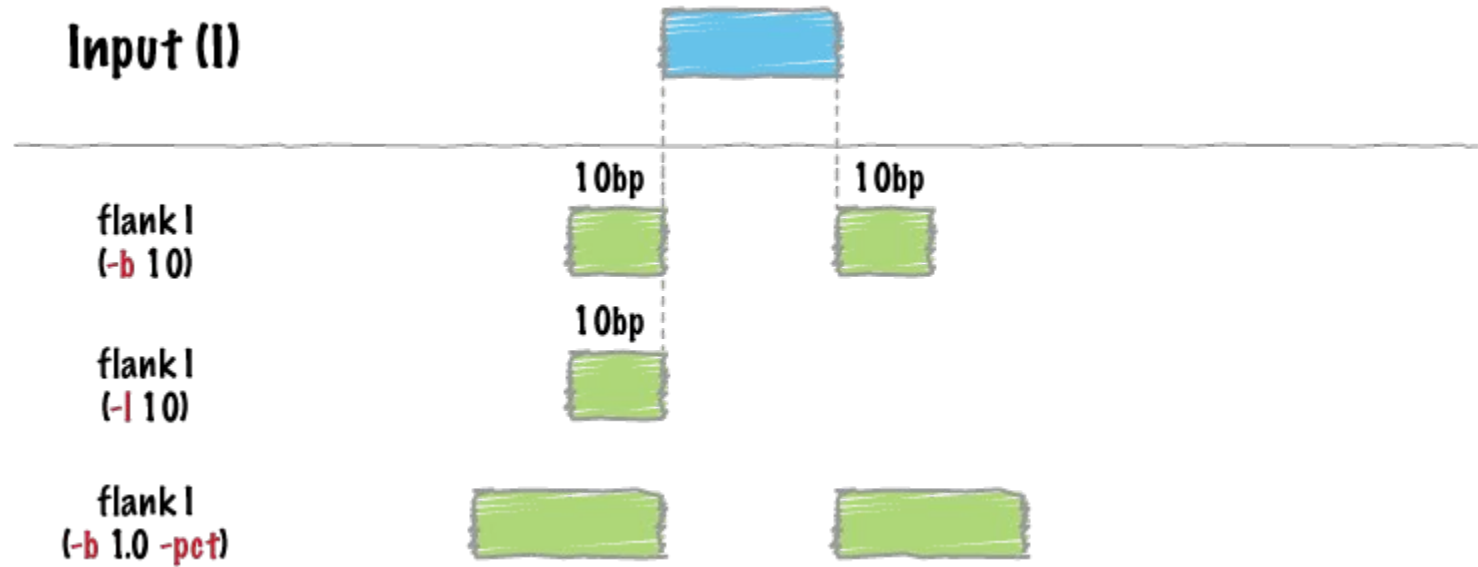
- `bedtools flank`
- il nécessite un fichier genome, pourquoi ?

Données:

- Taille des chromosomes:

`/shared/mfs/data/projects/ebaii2020/atelier_croisement/data/chrs.len`

bedtools flank



Et les options -r, -s, -pct...

Étape 2

Trouver tous les variants chevauchant les régions trouvées précédemment

Nouvelle option

- `bedtools intersect -loj`

Question complémentaire :

- Combien y a-t-il de SNP par promoteur ?

Problème 3

Question scientifique

Quels sont les gènes différentiellement exprimés les plus proches de mes pics ChIP qui contiennent une mutation ?

Données

- les gènes différentiellement exprimés
- les sites de fixation de H3K4me3
- une liste de variants, au format VCF

Un protocole possible

1. Trouver les pics qui contiennent une mutation
2. Trouver le gène le plus proche de chaque région précédemment trouvée

Étape 1

Trouver les pics qui contiennent une mutation

Rien de nouveau ici...

Étape 2

Trouver le gène le plus proche de chaque région précédemment sélectionnée

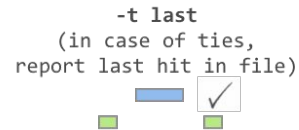
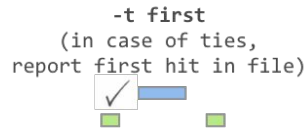
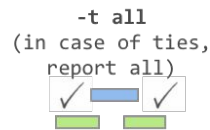
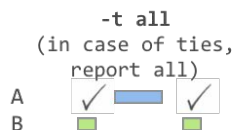
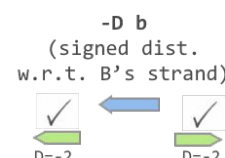
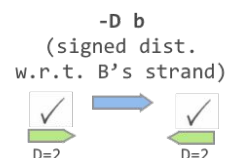
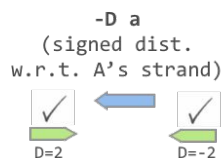
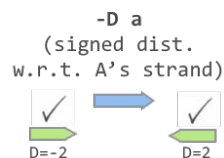
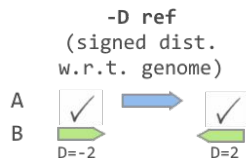
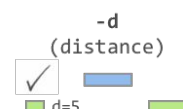
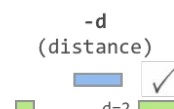
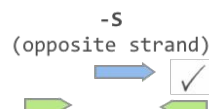
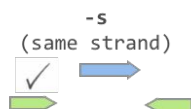
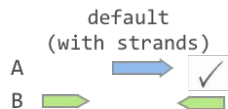
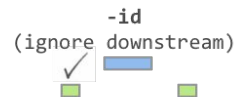
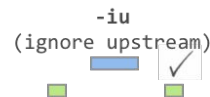
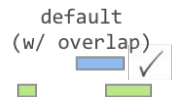
Nouvelle opération

- `bedtools closest`
- il demande que les annotations soient triées

Note

`bedtools closest` requires that all input files are presorted data by chromosome and then by start position (e.g., `sort -k1,1 -k2,2n in.bed > in.sorted.bed` for BED files).

bedtools closest



Étapes supplémentaires

- Limiter la recherche à 2kb.
- Se restreindre aux gènes dont les pics sont en amont d'un gène.
- Trouver les gènes en doublon.

Conclusion

- La majeure partie des questions que l'on se pose sur des données génomiques sont solubles avec un peu de bash, awk, et bedtools.
- bedtools propose *beaucoup* d'opérations
 - Conversions: BED / BAM / WIG / FASTA
 - Opérations binaires: intersect, closest, coverage, subtract
 - Opération unaires: merge, flank, shift, shuffle
 - Et autres...



Un croisement, ça ne donne pas de stats pertinentes !