Les données transcriptomiques

Véronique

* Fiche Transcriptomique

BUT : fiche orientée description de données, dans forgeMIA WP2 et doc partagé pour modifications

Définition du Transcriptome

Le transcriptome est défini comme l'ensemble des ARN présents

- dans une cellule ou une population de cellules,
- un tissu ou un organisme entier

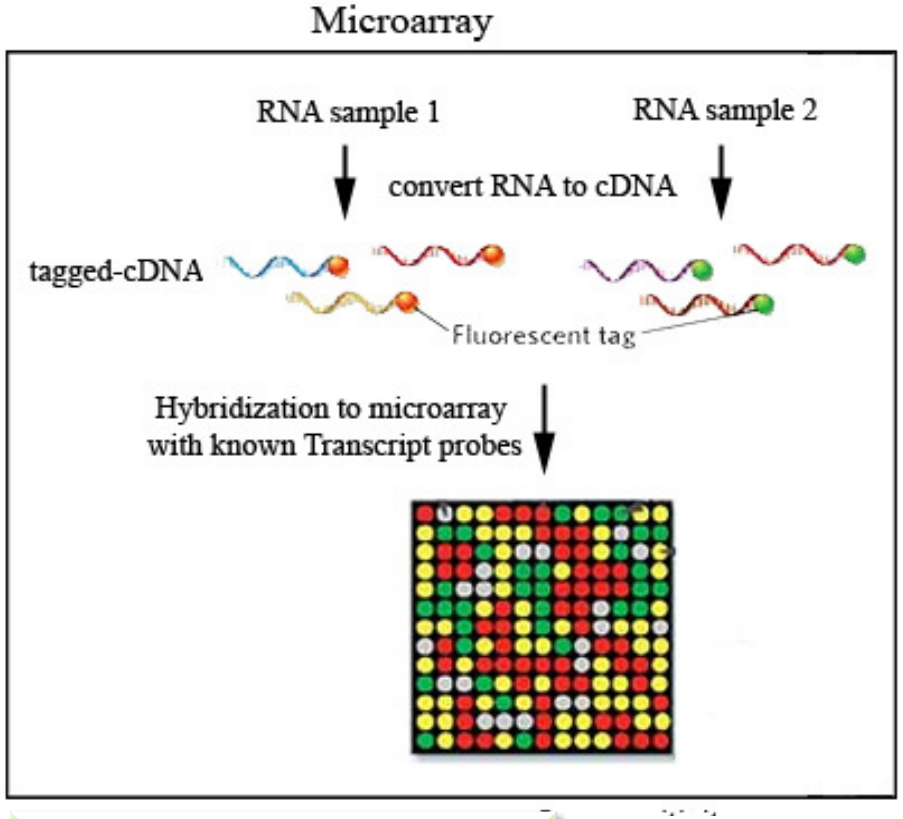
Il peut donc contenir

- tous les ARN (ARNm+ARNnc),
- ou cibler les ARN codant pour les protéines (eukaryotes).

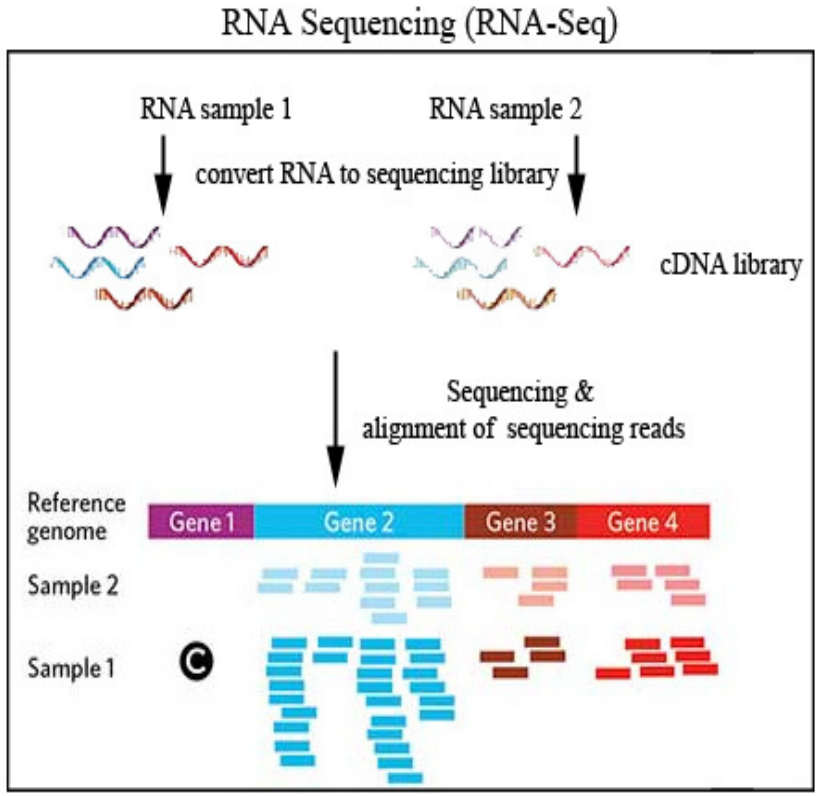
L'analyse du transcriptome consiste à définir la nature et la quantité d'ARN. L'identification et la quantification des ARN permettront de connaître

- les gènes ou opérons exprimés
- et les variations d'expression suivant les conditions ou organes étudiés.

* 2 principales technologies



Hybridation (1990)



Séquençage (2006)

Différences majeures entre RNA-Seq et puces à ADN

- Pas d'obligation d'utiliser une annotation…mais indispensable pour faire les analyses !
- Pas besoin d'avoir toutes les comparaisons dès le début (design complet en théorie !)

* Fiche : tableau général simple de description de l'expérience

	Description
Echantillons	source: espèce, variété, génotype sauvage/mutant; RNA extraction: organisme entier, organe/tissu/culture cellulaire;stade de développement/âge
Conditions expérimentales	type d'exp: comparaison de traitement, génotype, cinétique facteur(s): si présent, nom + quantité
Protocoles	Puces à ADN: type de puces, sondes, marquage, protocoles ; RNA-Seq: construction des librairies
Données	Objectifs: mesurer le niveau d'expression de chaque gène par échantillon ; Puces: des intensités ; RNA-Seq: des comptages ou liste de gènes

* Fiche : tableau des données

Données	technique RNA-Seq/Puces	Types de données	Description	Mesures
intensité de fluorescence	Puce, mesure du niveau d'expression d'un gène ou sonde	FLOAT: sous forme de table pour tous les gènes	Mesure quantitative d'intensité (moyenne) de signal	mesure brute par échantillon; mesure normalisée pour un ensemble de réplicats technique
comptage	RNA-Seq, mesure du niveau d'expression d'un gène	ENTIER: sous forme de table pour tous les gènes	nombre de reads associées à chaque gène après mapping	données brutes ou données normalisées pour l'ensemble des réplicats
LogRatio	RNA-Seq et Puces	FLOAT (LOG) sous forme de table + Pvalue associée	mesure d'expression différentielle entre 2 conditions	comparaison des données normalisées (intensités/comptages) entre 2 groupes d'échantillons; ex: traités/non traités

* Conclusion sur les données manipulées

- **Les objets manipulés sont les gènes, opérons, transcrits**
- **Ces données se présentent sous forme de matrice Gènes-Opéron-Trancrits versus Échantillons-Comparaisons**

2 solutions pour les données à exploiter

1- On part des résultats finaux, des logRatio pour savoir dans quel sens varie tel ou tel gène ou opéron (sur-exprimé ou sous-exprimé). On peut avoir une notion de quantité de variation avec le logRatio

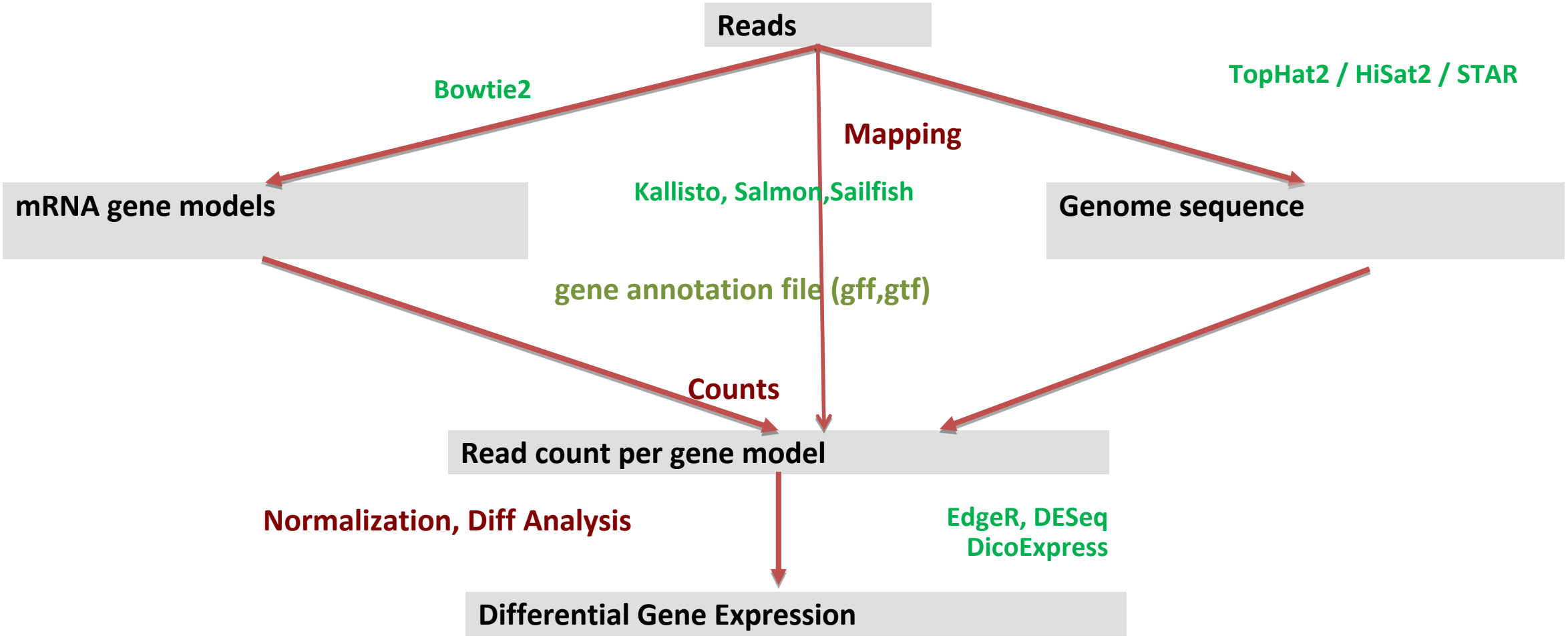
- Avantage : on va directement on but du projet SysMics, se base sur les variations
- Inconvénient : toutes les données récupérées ne sont pas analysées avec les mêmes méthodes au sein même du omic

2- On part des données brutes pré-analysées : les intensités brutes ou comptages brutes, et on refait les analyses différentielles qui répondent à notre question avec les mêmes méthodes (voir WP2 RFLOMICs et article DicoExpress)

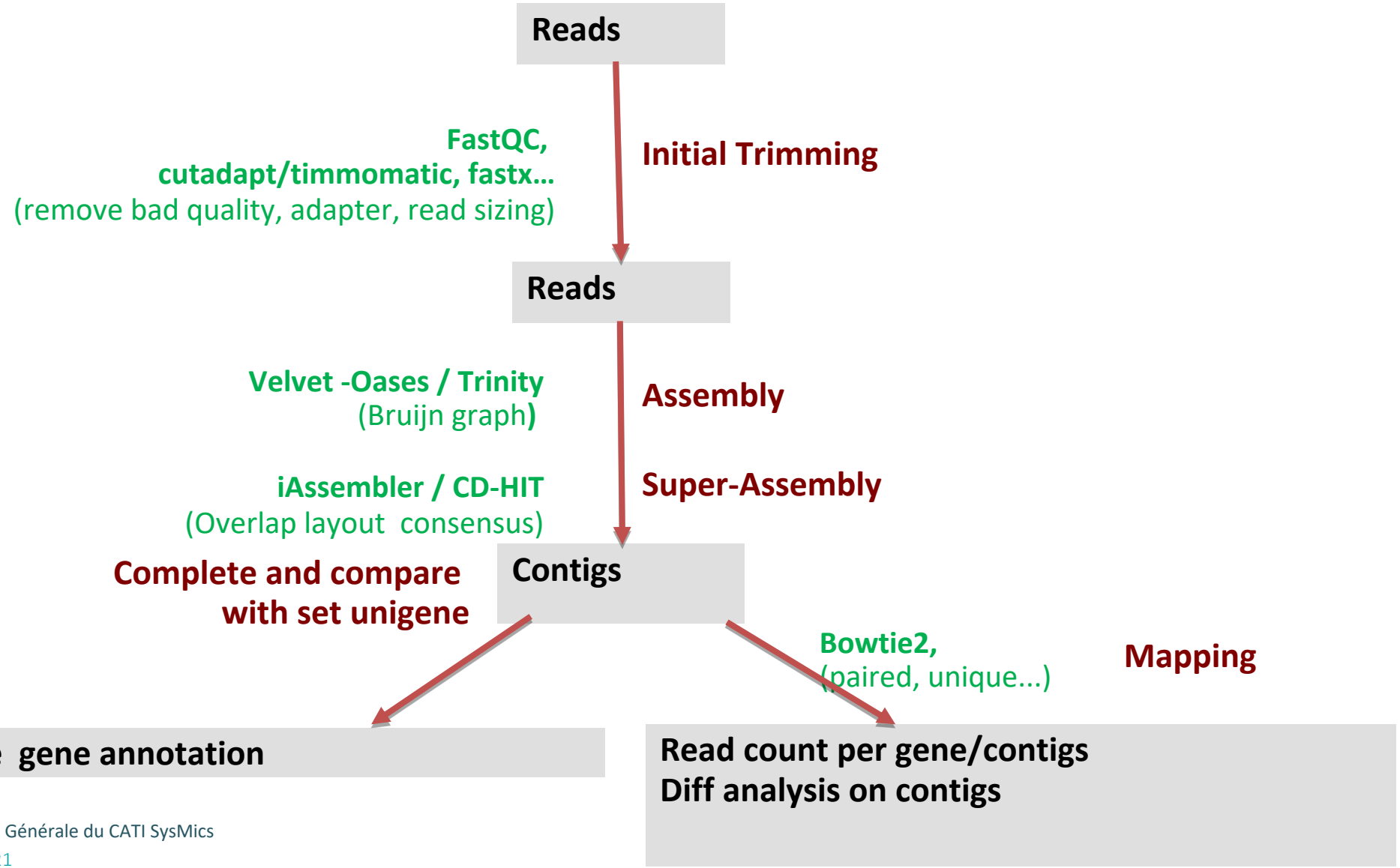
- Avantage : toutes les données sont analysées avec les mêmes outils
- Inconvénient : demande à refaire les analyses et donc bien comprendre les métadonnées et le design expérimental

Rajout de 3 diapos sur les outils utilisés et approches pour le RNA-seq, diapos liées à la fiche transcriptome

1st strategy : mapping RNA-Seq against a genome (transcripts or genome)



2nd strategy : de novo Assembly of RNAseq (without genome)



Goal : obtain counts by gene

Reads

Read count per gene model / contigs

Type of counts : raw count (number of assigned reads), estimated counts,
normalized counts TMM method
(size of library & gene/transcript)

Diff Analysis with different factors (= conditions)
P-value, FDR

Define biological replicats (normalization of sample names)