

Fiche de données transcriptomes

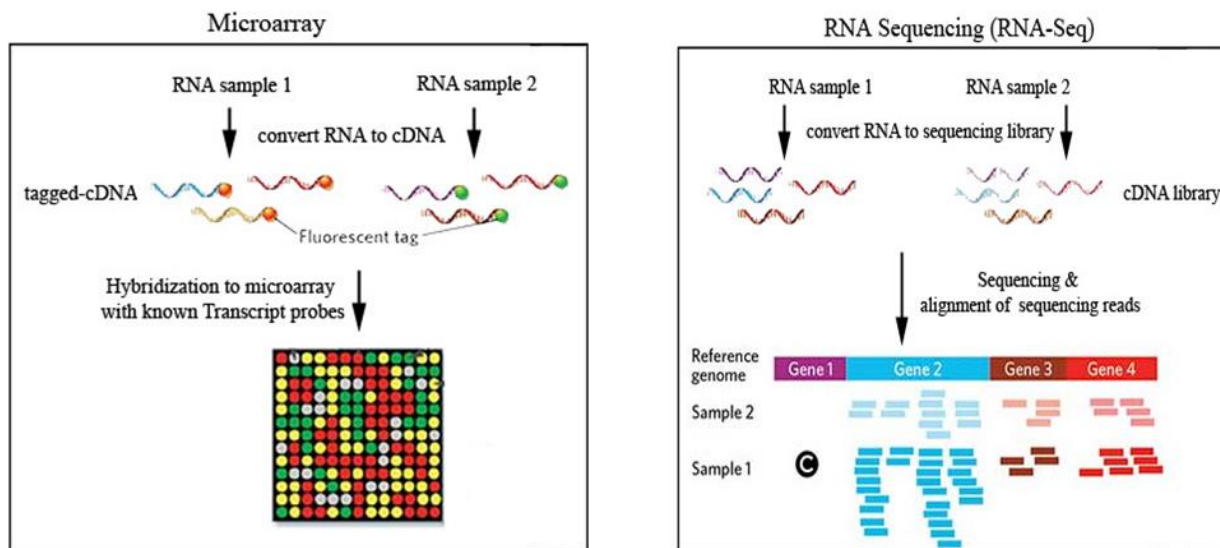
Objectif: décrire les types de données transcriptomiques manipulés lorsqu'on veut exploiter ces données pour différentes analyses ou faire de l'intégration d'omics. Cette fiche est volontairement orientée donnée et non pas description expérimentale.

1. Définition du Transcriptome

Le transcriptome est défini comme l'ensemble des ARN présents dans une cellule, un tissu ou un organisme entier. Il peut donc contenir tous les ARN (ARNm+ARNnc) ou cibler les ARN codant pour les protéines. 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 exprimés et les variations d'expression suivant les conditions ou organes étudiés.

2. Description des données Transcriptomes

2 principales technologies : les puces (microarray) et avec les nouvelles technologies de séquençage le RNA-Seq



Voir 4 diapos en fin de fiche pour exemples d'analyses de données sur PF Transcriptome IPS2

Tableau global sur fiche transcriptome

Tableau général décrivant brièvement ce qu'on attend comme description globale d'une expérience transcriptome.

	Description
Echantillon	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

Tableau sur les données transcriptomiques

Tableau qui décrit les types de données générés dans la majorité des expériences transcriptomes.

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

- RQ: ces données se présentent sous forme de matrice Gènes/Échantillons

3. Dépôts internationaux en transcriptomique :

2 standards internationaux :

- MIAME pour les puces (Minimum Information About a Microarray Experiment) avec 6 critères critiques (type de raw ; données finales = normalisées ; annotation des échantillons ; ...)
- MINSEQE pour RNA-Seq (Minimum Information about a high-throughput nucleotide SEQuencing Experiment), dernière version pour le RNA-Seq en 2012

Les articles contenant des données transcriptomes doivent contenir un accès à un dépôt international (GEO/ArrayExpress) pour la description des expériences et les données brutes (séquences=reads) sont déposées à SRA ou ERA-EBI. Le dépôt dans SRA ne permet pas de déposer les descriptions des expériences mais uniquement les séquences. Il y a un lien de GEO à SRA et inversement.

- **Au NCBI, expériences transcriptomiques décrites dans GEO (Gene Expression Omnibus) et données brutes de séquençage (fastq) dans archive SRA (Sequence Read Archive)**
- **A l'EBI, expériences transcriptomiques décrites dans ArrayExpress <https://www.ebi.ac.uk/arrayexpress/> et données brutes de séquençage (fastq) dans archive ENA (European Nucleotide Archive)**

4. Références associées et lien vers des projets :

- Lambert I, Paysant-Le Roux C, Colella S, Martin-Magniette M-L (2020) DiCoExpress: a tool to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models. Plant Methods 16:68. <https://doi.org/10.1186/s13007-020-00611-7>
- Rigai G., Balzergue S., Brunaud V., Blondet E., Rau A., Rogier O., Caius J., Maugis-Rabusseau C., Soubigou-Taconnat L., Aubourg S., Lurin C., Martin-Magniette M-L., Delannoy E. (2016). Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis. Brief. Bioinform. 1-12. doi:10.1093/bib/bbw092.
- Exemple avec un projet RNA-seq via GEO accès [GSE119344](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119344) ; **4 Diapos** décrivant les analyses types sur projets RNA-Seq et données générées : fichier associé 4diapos_forRNAseqData
- Exemple avec un projet puce à ADN via CATdb [4plex Physco 2014-05](https://www.ebi.ac.uk/biostudies/studies/S-EBI-GEO-SRA-ERA-ENACATdb-4plex-Physco-2014-05)

Project management for RNA-Seq

New Project
Biological question

Sample descriptions
Conditions, Treatments

RNA extraction
Library construction
Sequencer

COUNTS BY GENE

DATABASE



Technology evolution : NextSeq500

Name normalization

Project : NGS2017_03_Silicium
Samples : root_Si_1, root_NoSi_2

Explore bioinformatic tools :

- New tools
- Set of tests, references

INTERNATIONAL STANDARD



Sequencer

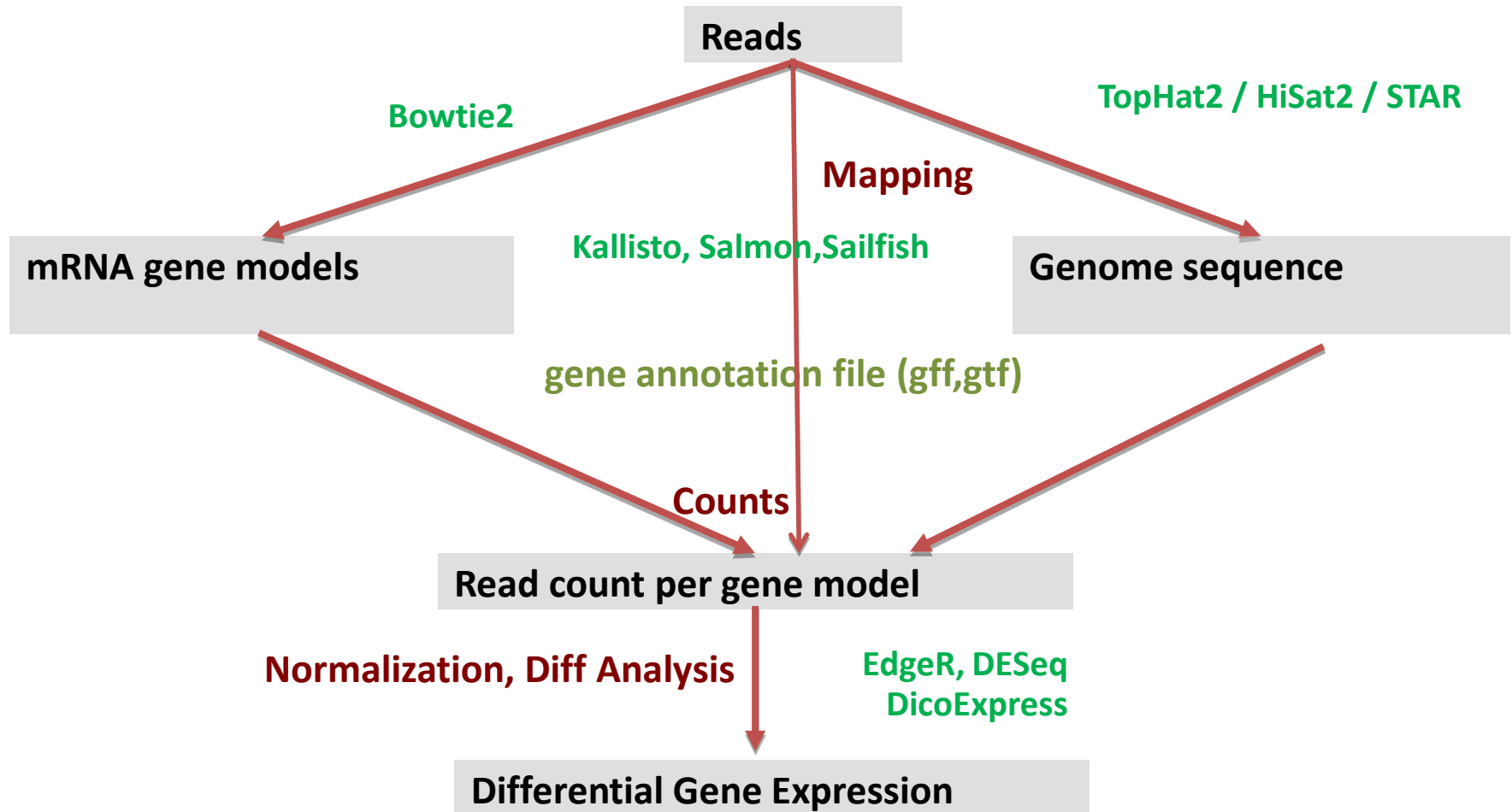
Pre-processing

Mapping/Assembly

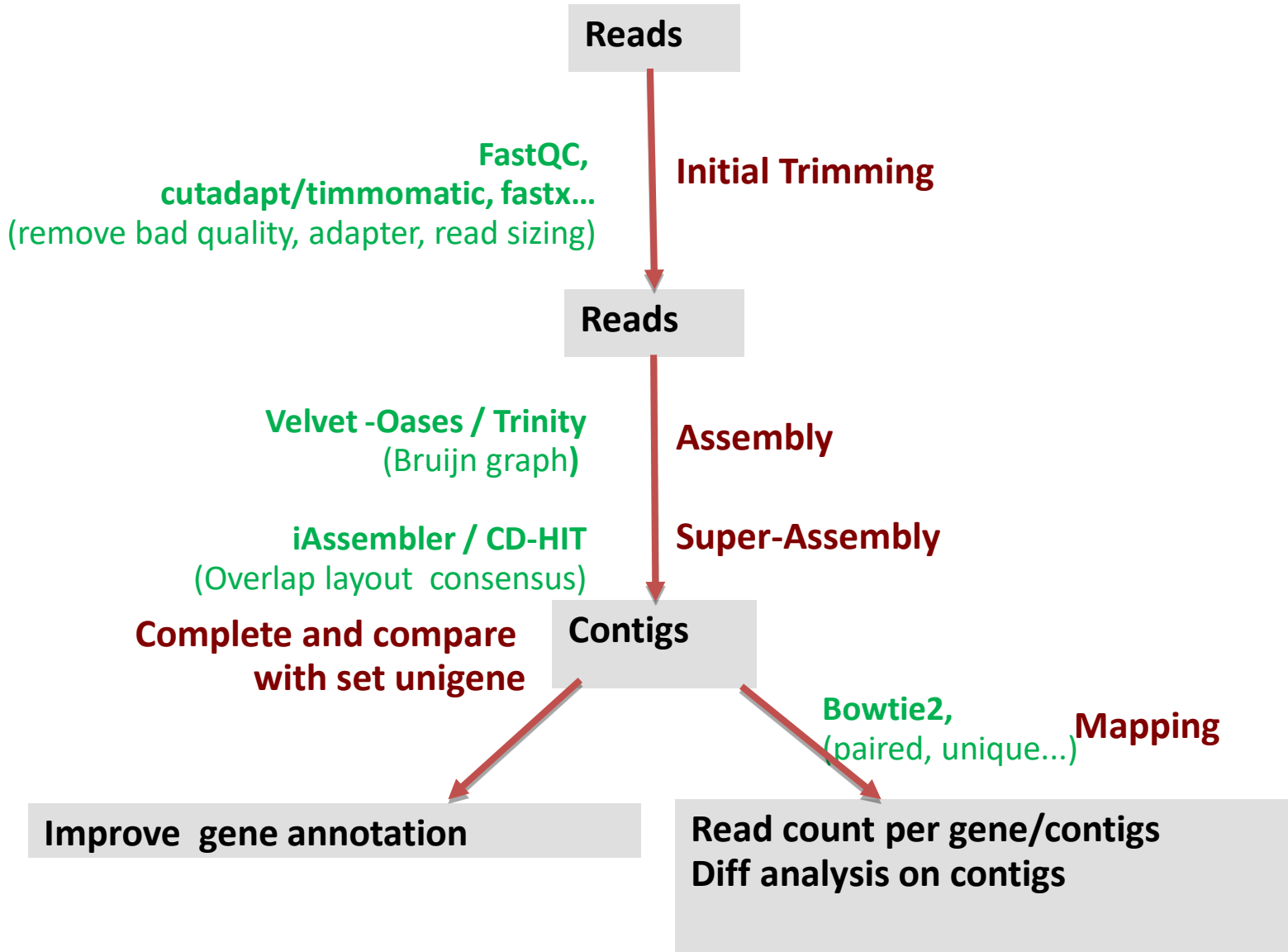
DIFF ANALYSIS
COUNTS BY GENE

TOOLS USAGE
(MONITORING)

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)