DE LA RECHERCHE À L'INDUSTRIE

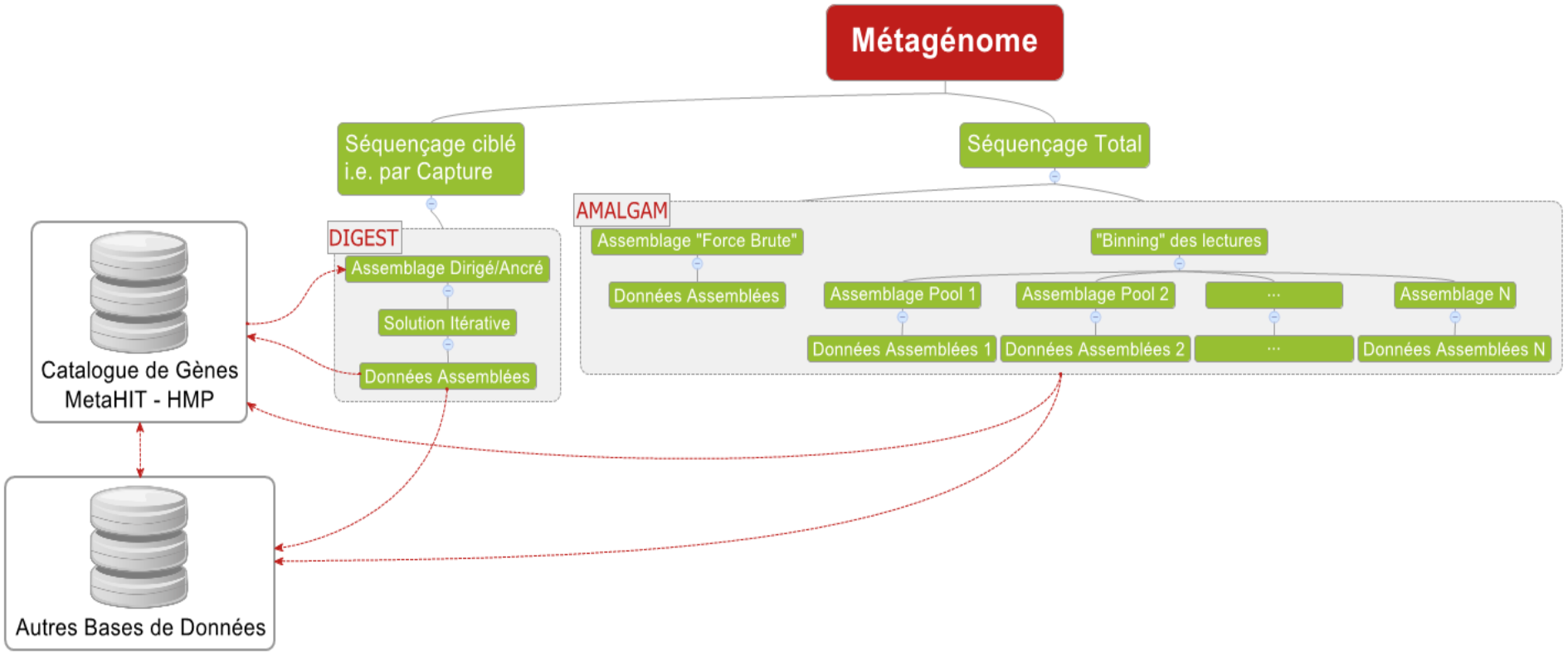
Développements bioinformatiques pour l'analyse de données métagénomiques – AMALGAM, vers un outil d'assemblage automatique

FG_WP2.7.2

S. Cruveiller, M. Séjourné

Assemblages de Métagénomes

Au moins 2 angles d'attaque selon qu'on dispose de données ciblées ou de données totales...



Arnaud Felten - CDD France Génomique

(Mars 2013 - Décembre 2014)

Le pipeline DIGEST

Directed **I**terative **G**ene **E**xtension by **S**equencing capture **T**echnology

Objectif: Compléter les gènes partiels du catalogue du microbiome intestinal humain

Microbiome intestinal humain : Un organe à part entière

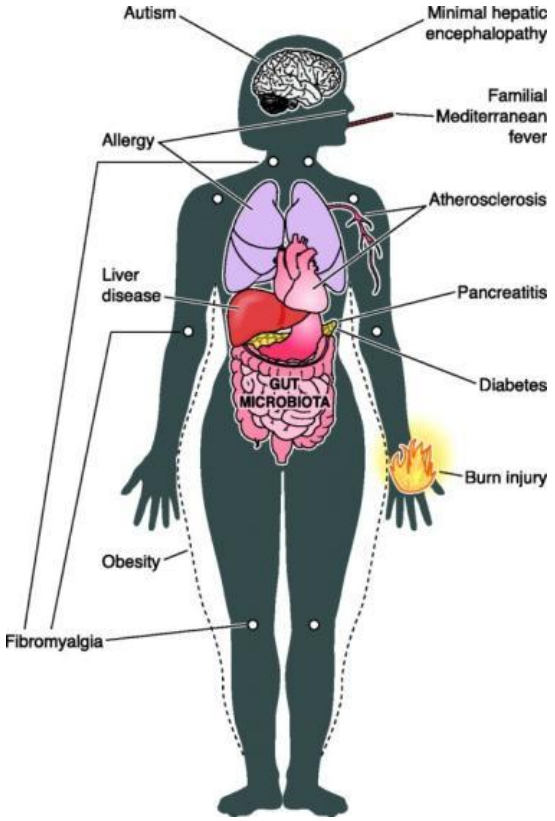
- 100 trillion (10^14) micro-organismes
  - **10 X cellules humaines**
  - +2kg

- A l'interface des aliments et de l'épithélium intestinal

- **Connu seulement à 30%** [Qin et al, 2010]

- **Souvent associé à des maladies chroniques:**

➢Maladie de Crohn [Seksik et al., 2003; Sokol et al.,2006,2008, 2009]
➢ Colites [Sokol et al., 2008; Martinez et al., 2008]
➢ Obésité [Ley et al., 2007; Kalliomäki et al., 2008]
➢ diabètes de Type-2 [Cani and Delzenne, 2009]
➢ diabètes de Type-1 [Dessein et al., 2009; Wen et al., 2008]
➢ Allergies [Kirjavainen et al., 2002; Björkstén, 2009]
➢ Autisme [Finegold et al., 2002; Paracho et al., 2005]
➢ Cancer colorectal [Mai et al., 2007; Scanlan et al., 2008]
➢ Maladies du Foie [Gunnarsdottir et al. 2003]
➢ Maladies cardiovasculaires [Wang et al. 2011]



I. Sekirov et *al*, *Gut Microbiota in Health and Disease, Physiol Rev* 90: 859-904, 2010

Projets METAHIT et HMP

MetaHIT

- Qin et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature, page 59-65, volume 464, 2010.
- http://www.metahit.eu/

HMP

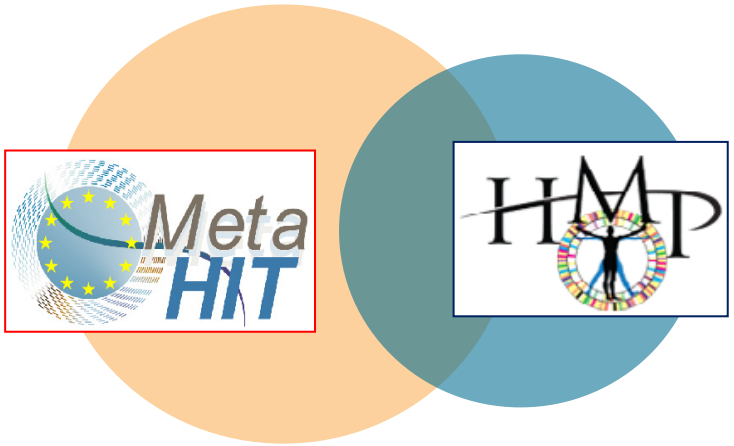
- Turnbaugh et al., The Human Microbiome Project. Nature, page 804-810, volume 449, 2007. Nature, page 804-810, volume 449, 2007.
- http://www.hmpdacc.org/

The Human Microbiome Project

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon

A strategy to understand the microbial components of the human genetic and metabolic landscape and how they contribute to normal physiology and predisposition to disease.

METAHIT et HMP GENES SET



9 millions **7 millions**

12 millions gènes non redondants

56%

genes partiels
(START et/ou STOP sont manquants)

Junhua Li *et al., Integrated reference gene catalog for human gut microbiome*, unpublished

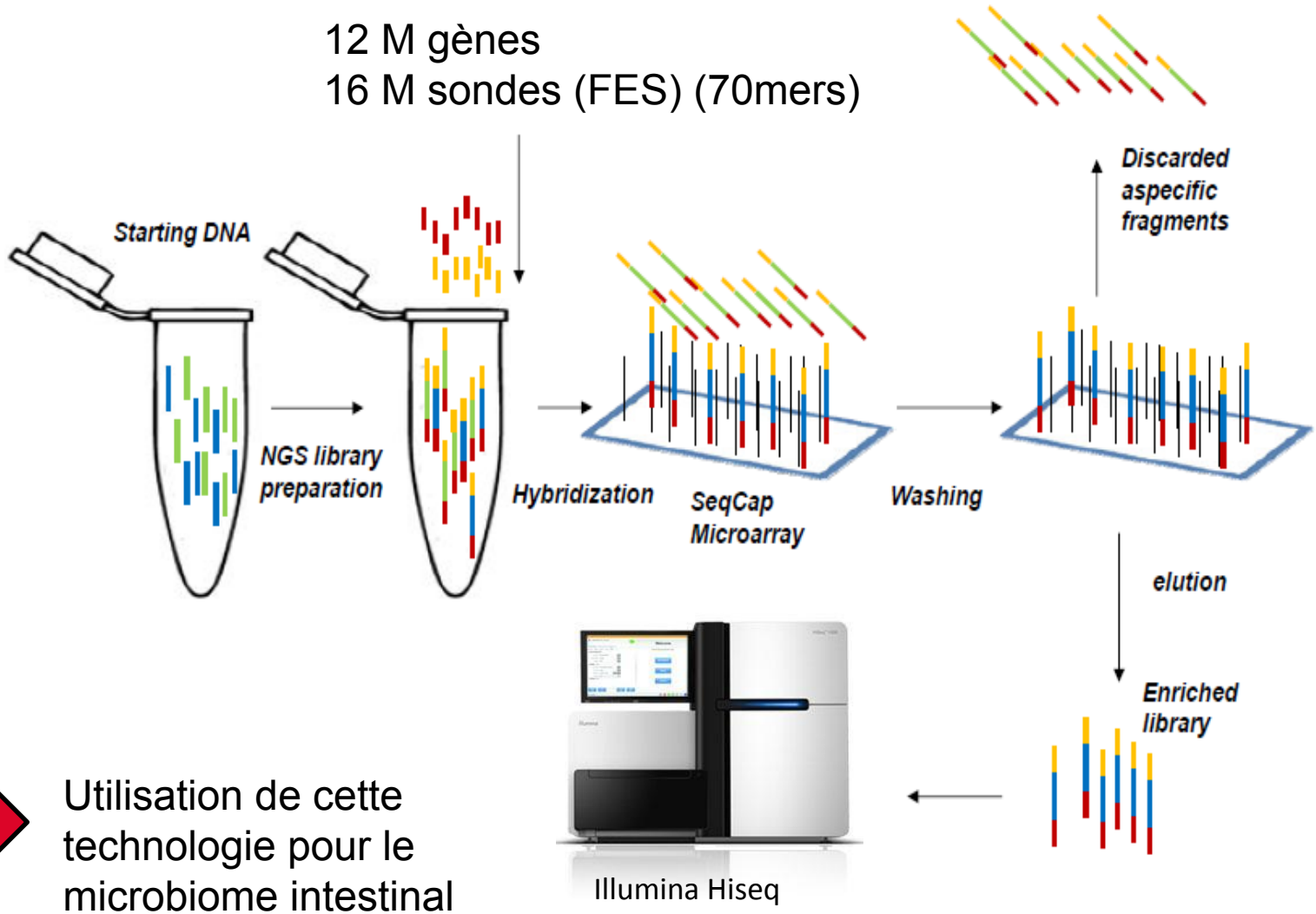
Une stratégie en 2 étapes...

1) cibler les gènes d'intérêts → **capture technology**
   ≠ whole shotgun

2) workflow bioinformatique→ **DIGEST**
   Directed Iterative Gene Extension by Sequencing capture Technology

Séquençage par capture

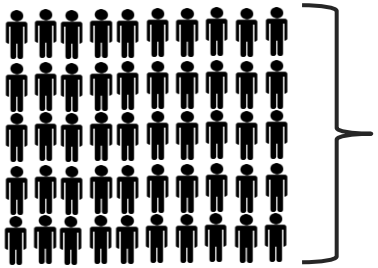
12 M gènes
16 M sondes (FES) (70mers)



Utilisation de cette technologie pour le microbiome intestinal humain

DATA SET produite

16.10^6 Sondes



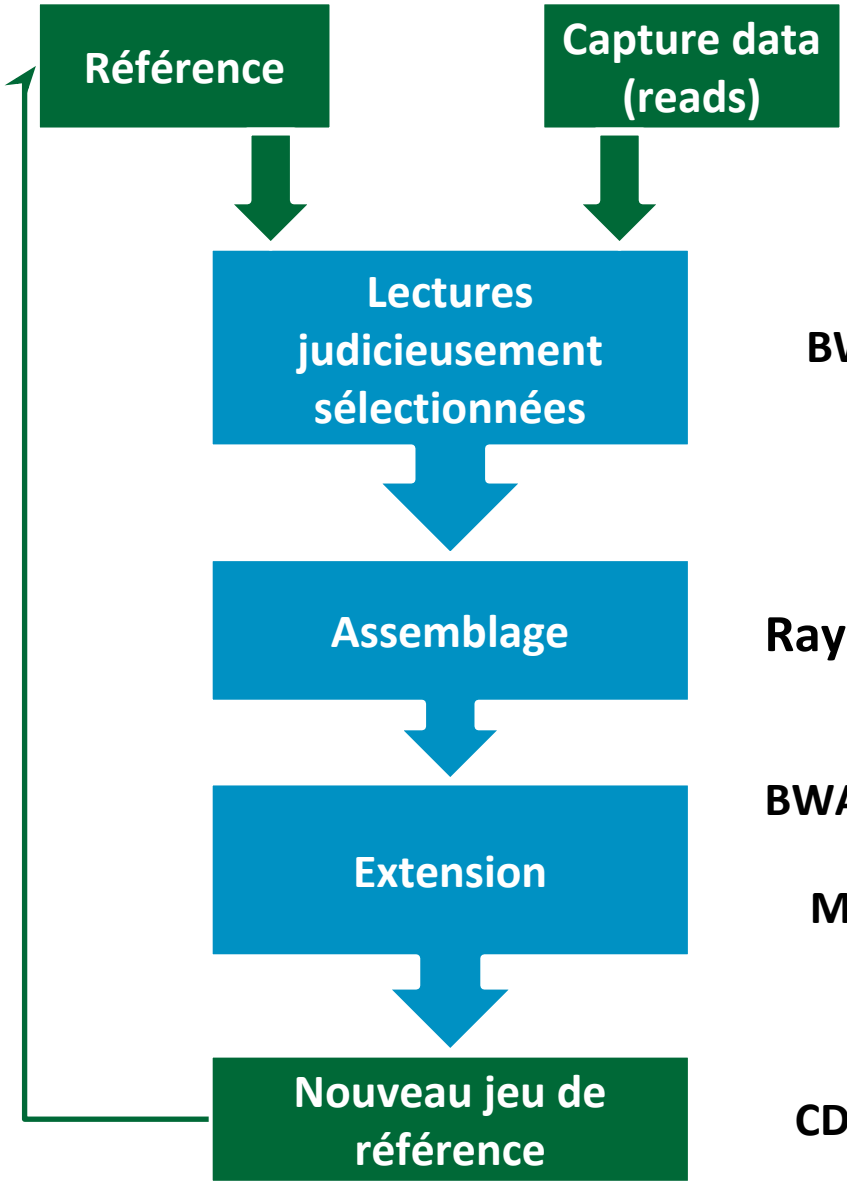
50 Individus séquencés par **capture**

4 x ¼ run par individu
paired-end (2x100bp)
Illumina Hiseq

2.10^{10} Lectures

6To data

Stratégie DIGEST



BWA

Li and Durbin, **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics*, page 1754-1760, volume 25, 2009

Ray Meta

Boisvert et al., **Ray Meta: scalable de novo metagenome assembly and profiling**. *Genome Biology*, page R122, volume 13, 2012.

BWA-MEM

Heng Li, **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. 2013.

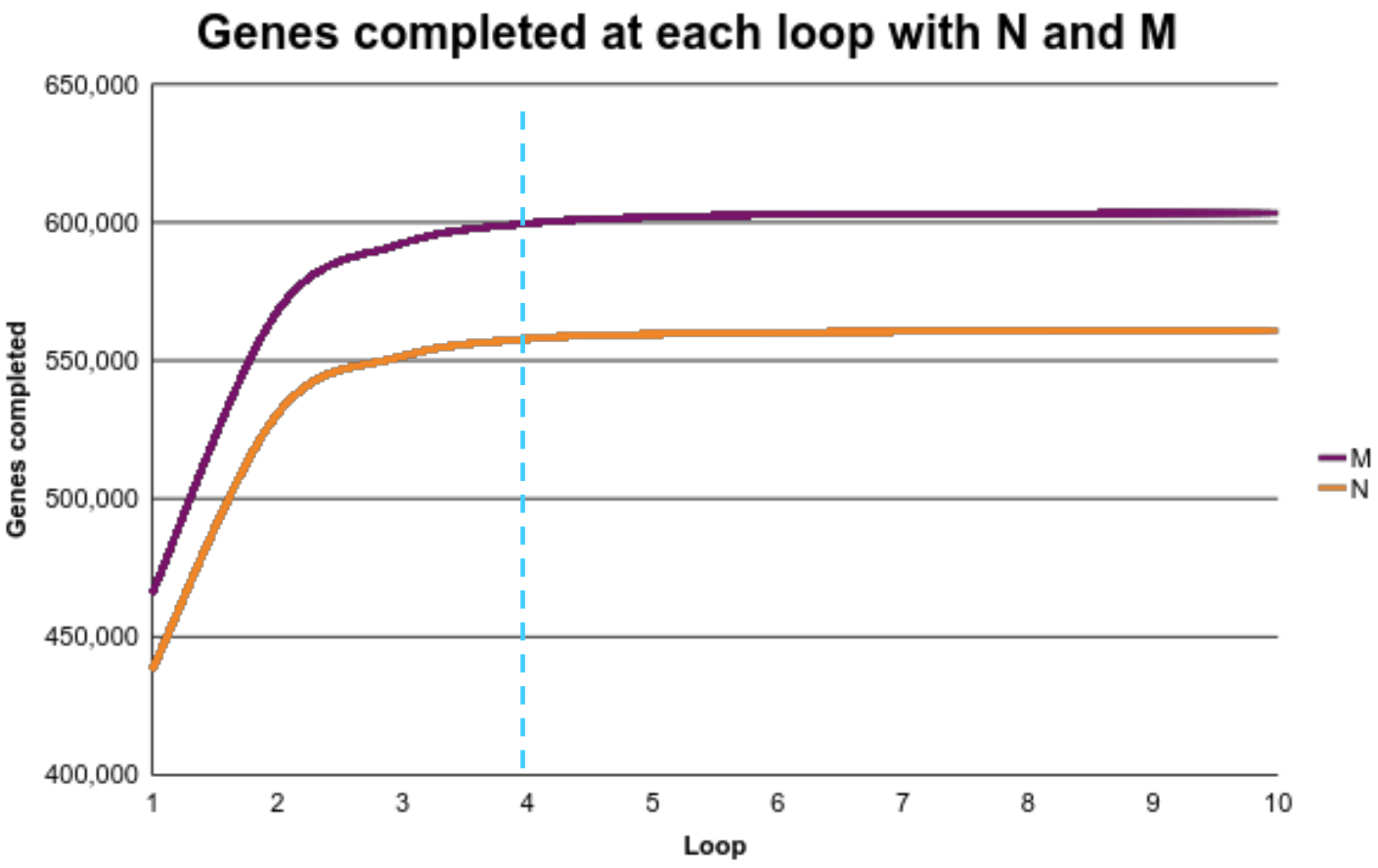
MGA

Noguchi et al**., MetaGene: prokaryotic gene finding from environmental genome shotgun sequences**. *Nucleic Acids Research*, page 5623-5630, volume 34, No. 19, 2006.

CD-HIT

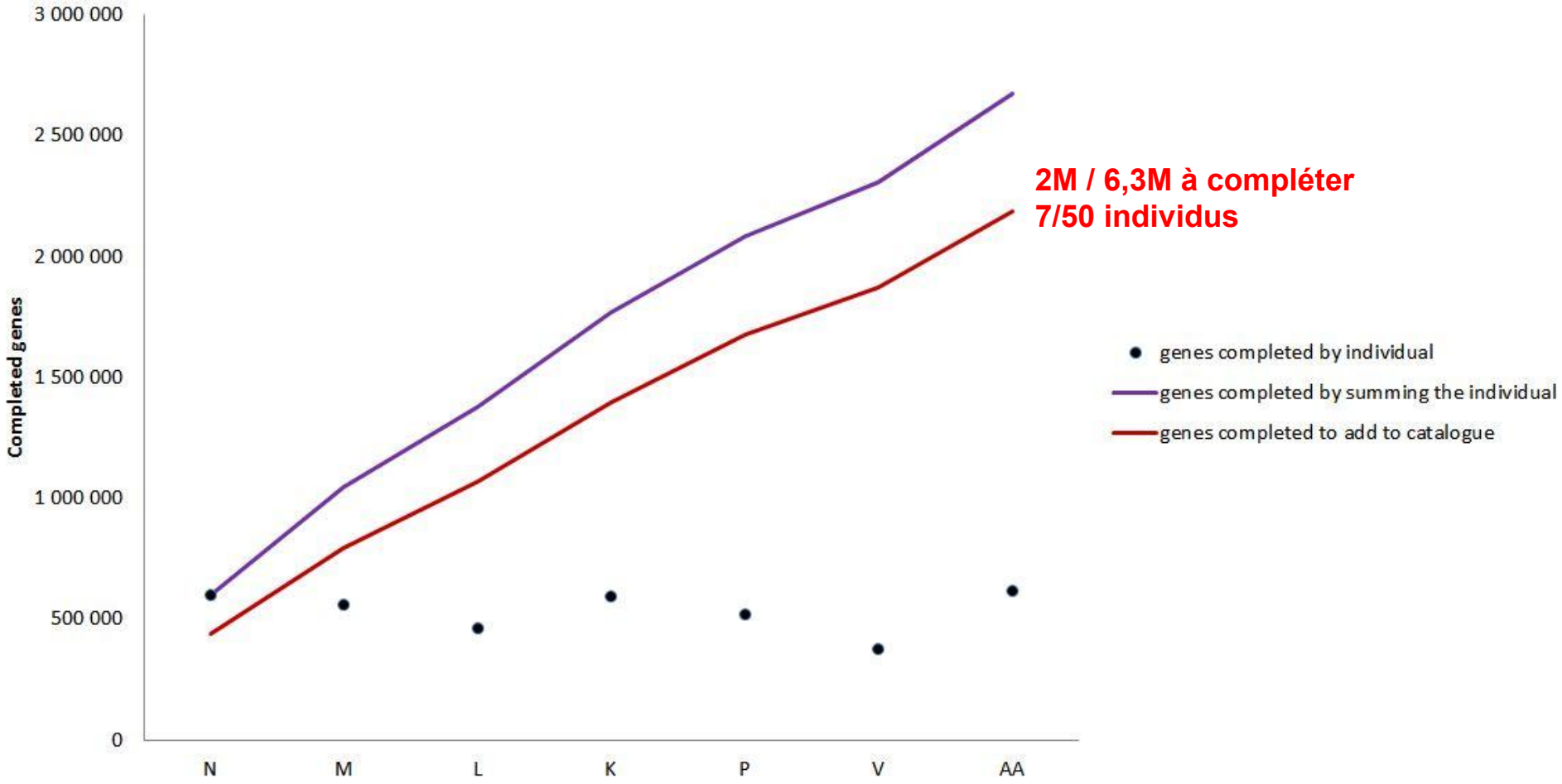
Limin Fu et al., **CD-HIT: accelerated for clustering the next generation sequencing data**. *Bioinformatics*,, 28 (23): 3150-3152, 2012.

Condition d'arrêt de DIGEST: test sur 2 individus



2 ou 3 Itération sur 1 individu suffisent pour exploiter un maximum d'information

Résultats actuels

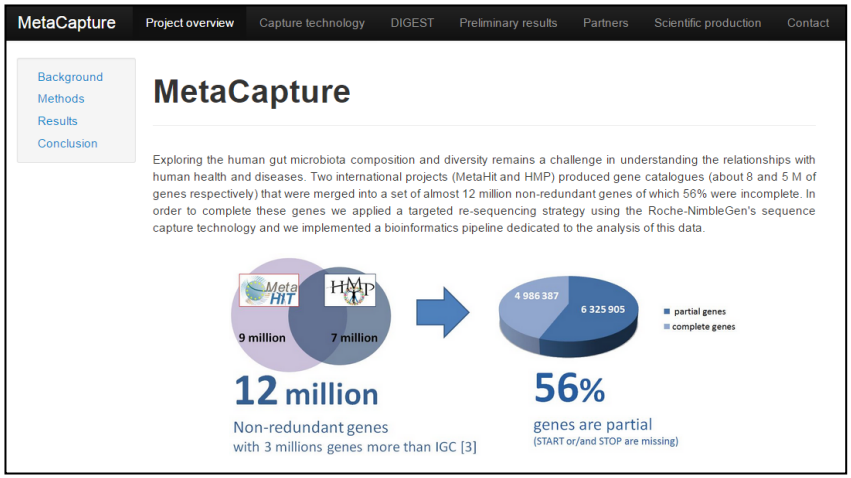


Beaucoup d'individus restent à analyser pour compléter intégralement le catalogue de gènes initial → absence de plateau

Performances

1 Individuals processed by **DIGEST** → **72H** computing → **1To** results → **~ 500 000** genes completed

7 Individuals → **2million** genes completed → **25% genes** initially incompleted has been completed with 10% of the data



DIGEST available for others genes completion projects at :

http://www.genoscope.cns.fr/projects/metacapture/index.html

Bilan & Perspectives

Bilan

- Pipeline fonctionnel
- DIGEST porté sur les clusters de calcul du CCRT
- Disponible pour la communauté

Perspectives

- Bioanalyse des données restantes (43 individus) pour compléter le catalogue de gènes
- Optimisation du pipeline pour réduire les temps de cacluls
- Optimisation du pipeline pour améliorer de manière qualitative les résultats produits (limitation des chimères)

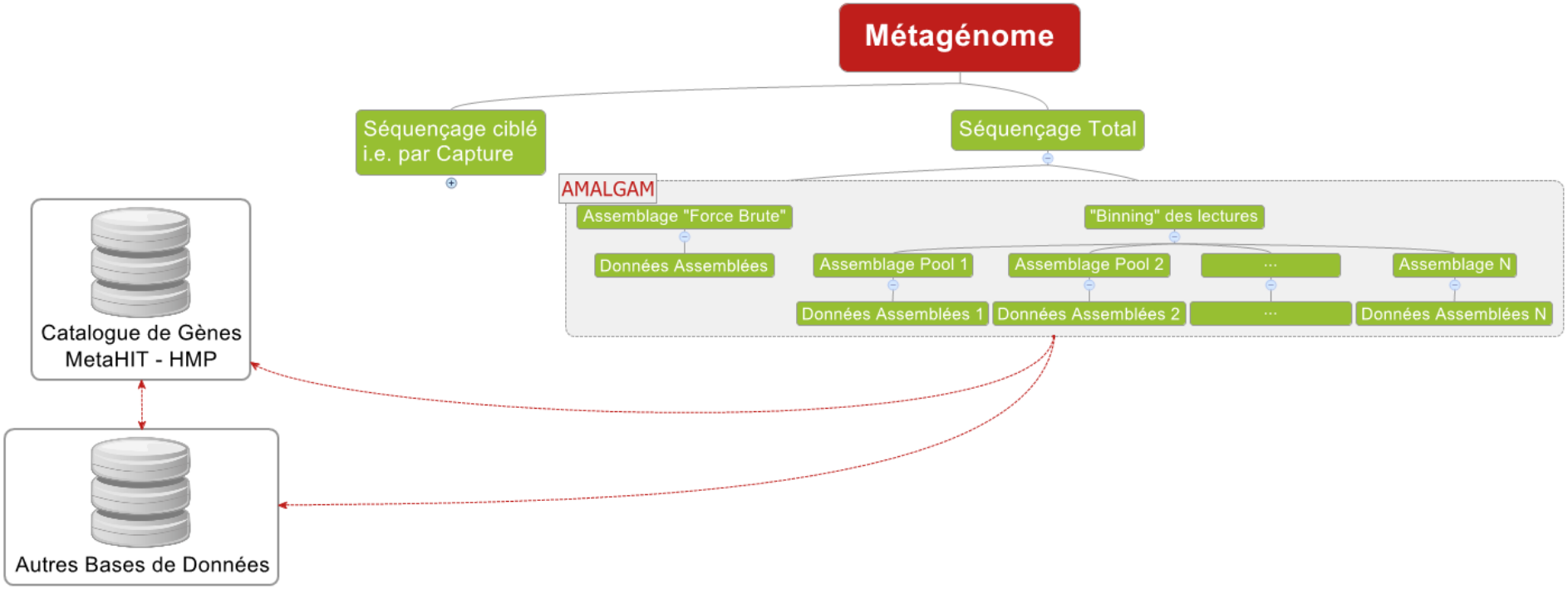
Marine Séjourné - CDD France Génomique

(Mars 2015 - ...)

Assembleurs en métagénomique

ÉTAT DE L'ART

Assemblage de Métagénomes



Limites :

➔ Assemblage brut :

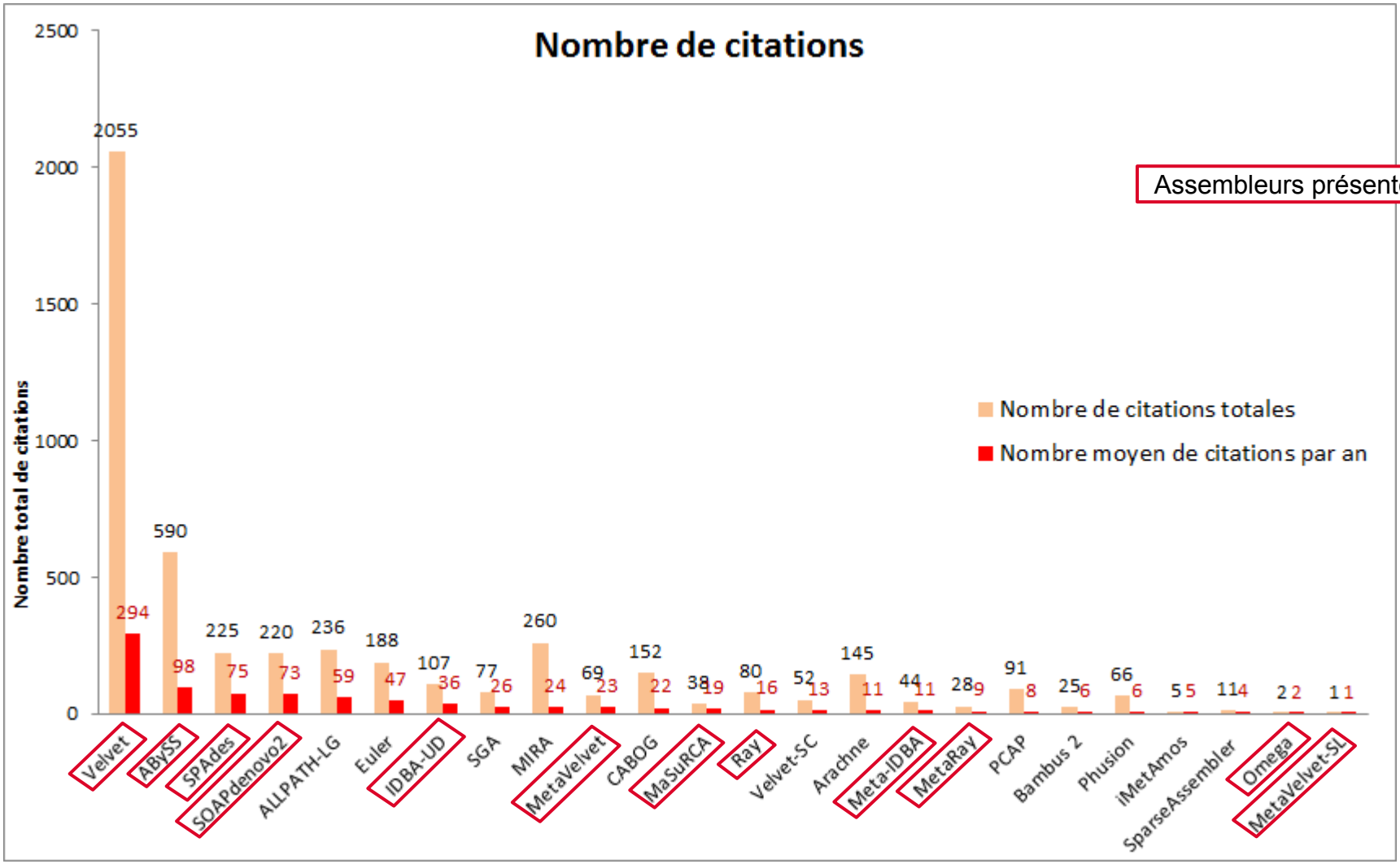
Information hétérogène, bruitage en raison d'une multitude de scaffolds.

➔ "Binning" des lectures :

Sur quels critères définir les pools ? (genres, génomes).

Exclusion des lectures connues ?

Citation des Assembleurs



Quelques assembleurs ont des variantes pour l'assemblage de métagénomes.

Certains assembleurs reviennent fréquemment dans les comparaisons.

Quelques assembleurs ...

Nom de l'assembleur	Type d'approche	Modification de l'approche	Licence	Spécialité
ABySS (2009)	Graphe de de Bruijn	Simplification du Graphe	Commercial mais gratuit pour les académics et non-commerciaux	Utilisation sur un cluster
MaSuRCa (2013)	Hybride (Graphe de de Bruijn et Overlap-Layout-Consensus)	Modification de CABOG	Open source	Transformation des nombreuses lectures courtes en peu de longues lectures
Meta-IDBA (2011)	Graphe de de Bruijn		Open source	Séparation du graphe de de Bruijn en fonction des espèces
IDBA-UD (2012)	Graphe de de Bruijn		Open source	Assemblage avec des profondeurs de séquençage différente
Velvet (2008)	Graphe de de Bruijn		Open source	Assemblage de courtes lectures
MetaVelvet (2012)	Graphe de de Bruijn		Open source	Séparation du graphe de de Bruijn en fonction des espèces
MetaVelvet-SL (2015)	Graphe de de Bruijn		Open source	Utilisation de l'apprentissage supervisé pour détecter les noeuds chimériques
Omega (2014)	Graph de chevauchement		Open source	
Ray (2009)	Hybride (Graphe de de Bruijn - "greedy assembler")		Open source	Assemblage de données de différentes technologies
Ray Meta (2012)	Graphe de de Bruijn	Ajout de couleur dans les noeuds	Open source	Couplé à une assignation taxonomique
SOAPdenovo2 (2012)	Graphe de de Bruijn		Open source	Assemblage des génomes de grande taille
SPAdes (2012)	Graphe de de Bruijn		Open source	
Newbler	Overlap-Layout-Consensus		Commercial	Premier assemblage commercial (associé à 454)

Marine Séjourné - CDD France Génomique

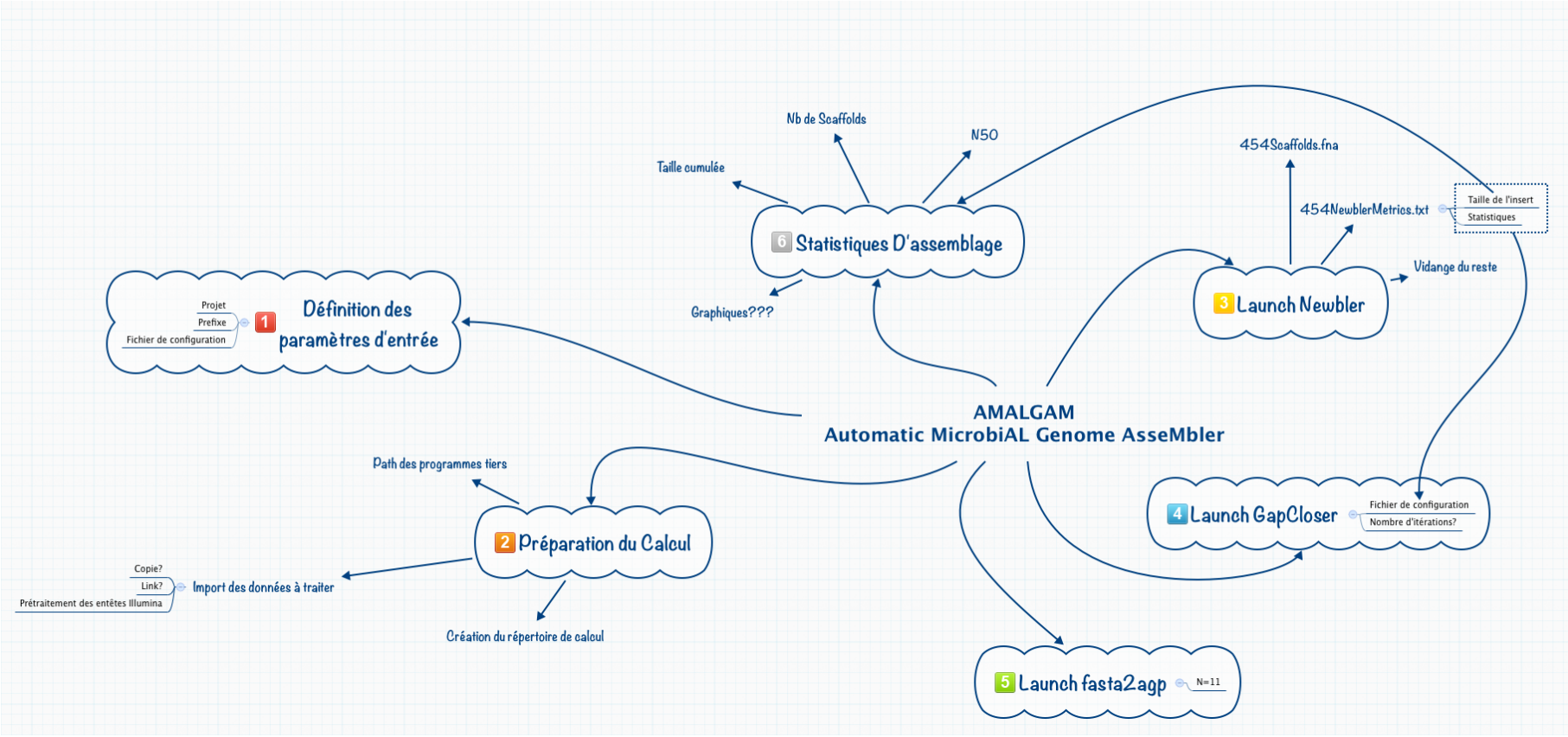
(Mars 2015 - ...)

Présentation d'AMALGAM

Automatic MicrobiAL Genome AsseMbler

Objectif: Proposer un workflow d' assemblage générique pour assembler des métagénomes

Méthodologie - AMALGAM Automatic MicrobiAL Genome AsseMbler



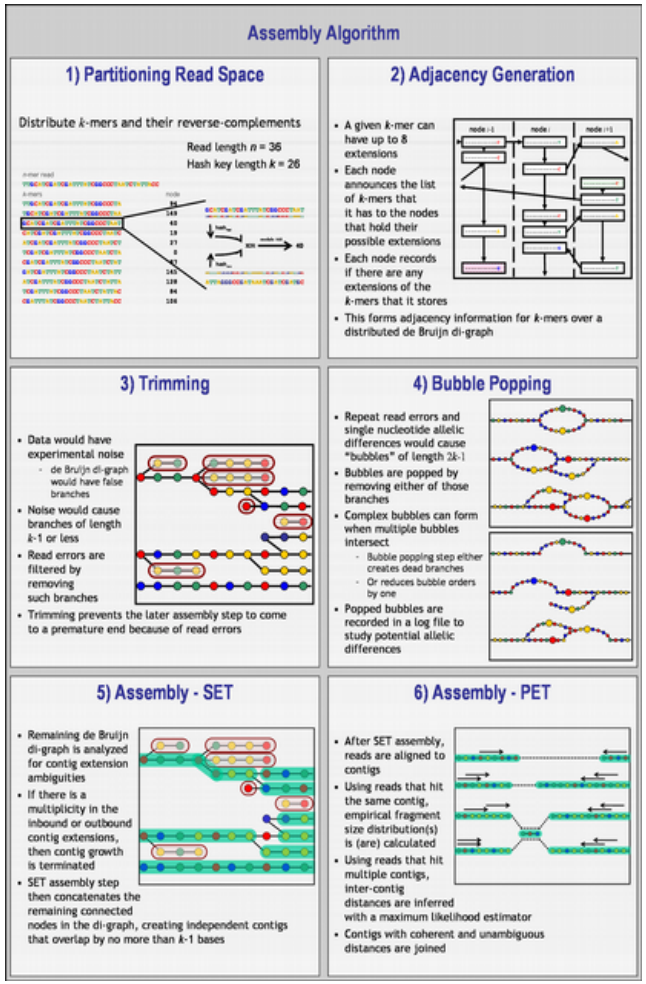
AMALGAM = Script Python prépare un fichier de configuration ou utlise un fichier établi

- lance l'assembleur (Newbler, 454-Roche LifeSciences)
- lance le GapCloser (package SOAP)
- lance le fasta2agp (IG-made script qui relate l'ordonnancement des scaffolds/contigs)
- Effectue une partie des statistiques d'assemblage (via QUAST)

Assembleur 454-ROCHE NEWBLER vs ABYSS

ABYSS ASSEMBLER

(M.Smith Genome Science Center)



NEWBLER ASSEMBLER

(454 – Roche LifeSciences)

1. pre-computes K-mer content across all reads, selects overlap candidates that share K-mers, and computes alignments using the K-mers as alignment seeds

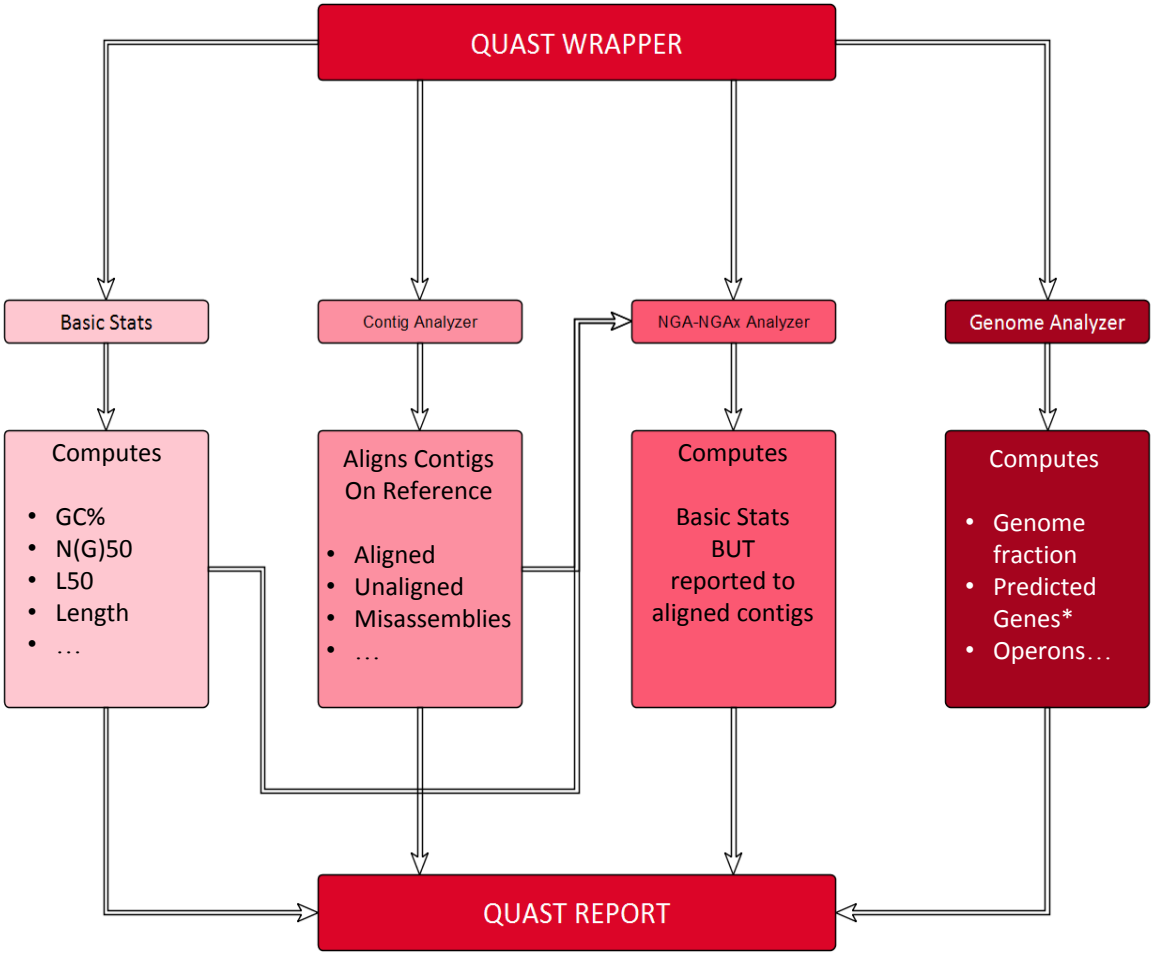
2. Construction and manipulation of an overlap graph leads to an approximate read layout

3. Multiple sequence alignment (MSA) determines the precise layout and then the consensus sequence.

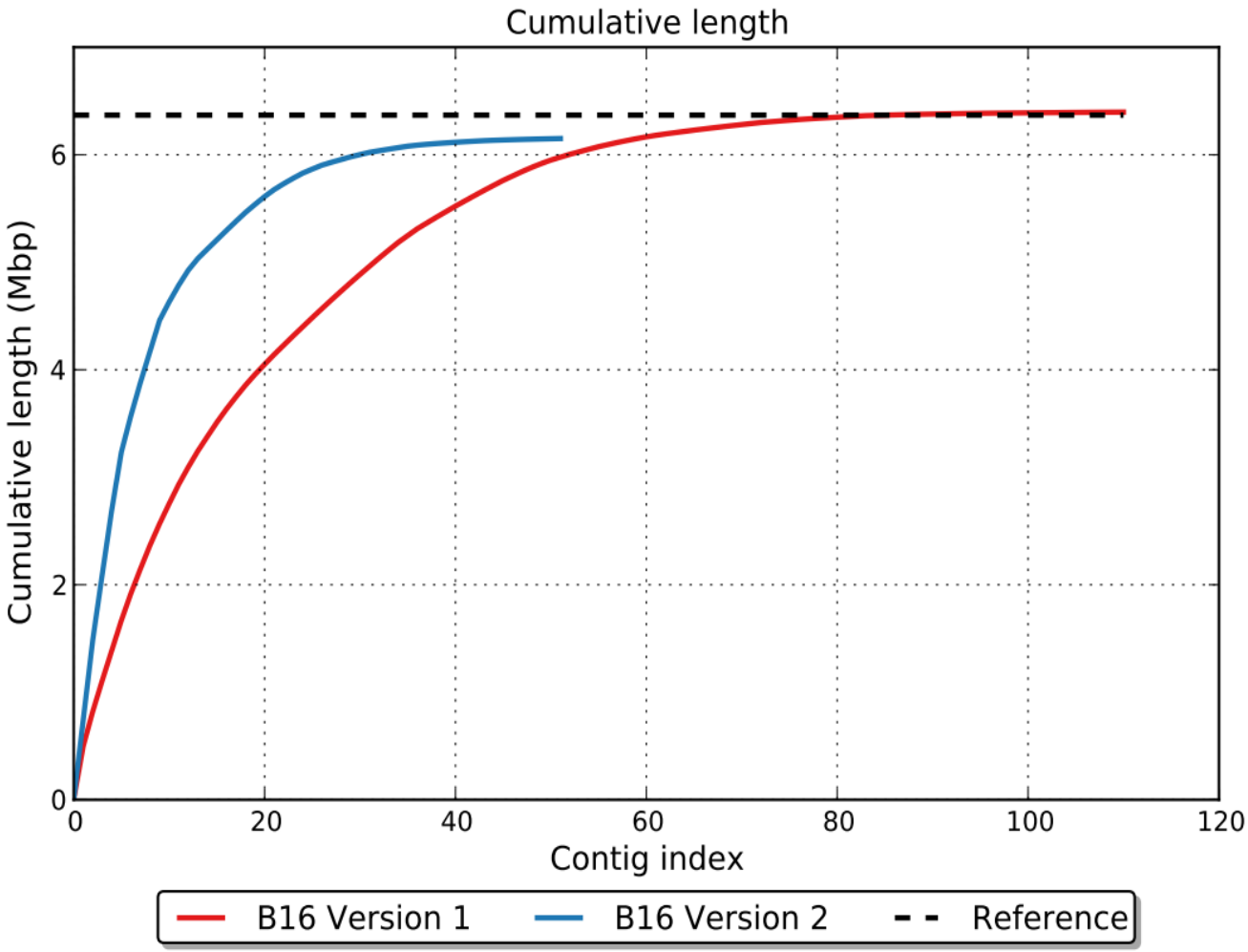
Comparaison des assemblages...

QUAST = Quality Assessment Tool for Genome Assemblies

(Gurevich *et al.*, Bioinformatics (2013))



Assembly length ...



Version 1: ABYSS
Version 2: Newbler

NEWBLER vs ABYSS…

Genome	Feature	Version 1 (ABYSS)	Version 2 (Newbler)
***Ralstonia solanacearum* B16**	*Total Length*	6465365	6150538
	Contigs	707	51
	Contigs (>1kb)	96	51
	N50	**149857**	**531604**
	N75	**79486**	**169923**
	L50	13	5
	L75	29	10
	GC %	66.34	66.50
	RGF	97.213	96.069
***Ralstonia solanacearum* G16**	*Total Length*	6409625	6136973
	Contigs	770	54
	Contigs (>1kb)	180	54
	N50	78580	462783
	N75	40762	220086
	L50	23	5
	L75	53	10
	GC %	66.31	66.46
	RGF	96.788	95.997
***Cupriavidus taiwanensis* LMG19425**	*Total Length*	7022375	6930426
	Contigs	386	29
	Contigs (>1kb)	85	29
	N50	218663	516938
	N75	121249	284931
	L50	10	6
	L75	21	11
	GC %	21	11
	RGF	**8.495**	**7.101**
***Cupriavidus taiwanensis* LMG19431**	*Total Length*	6620596	6488410
	Contigs	476	66
	Contigs (>1kb)	126	66
	N50	129649	350523
	N75	62080	142201
	L50	16	7
	L75	34	15
	GC %	66.98	67.06
	RGF	94.265	94.077
***Cupriavidus oxalaticus* LMG2235**	*Total Length*	6797551	6681945
	Contigs	331	36
	Contigs (>1kb)	110	36
	N50	135975	296831
	N75	74583	162437
	L50	15	8
	L75	32	16
	GC %	66.87	66.92
	RGF	NE	NE

N50 x 4 et N75 x2

Améliorations V2 versus V1

- Nombre plus faibles de contigs
- Contigs plus grands

Bilan et Perspectives

BILAN

➔ Newbler propose des résultats nettement meilleurs qu'ABYSS mais:

- les tests ont été pratiqués sur des génomes isolés
- le soft reste une technologie propriétaire
- les temps d'exécution sont longs (algorithme OLC)

PERSPECTIVES

➔ Tester Newbler en condition Métagénomique
➔ Le challenger par d'autres assembleurs open source (benchmark)

MAIS AVANT SE POSER LA VRAIE QUESTION...

Faut-il tenter d'assembler 1 fois n génomes ou n fois 1 "génome" pour obtenir un résultat probant/valable?

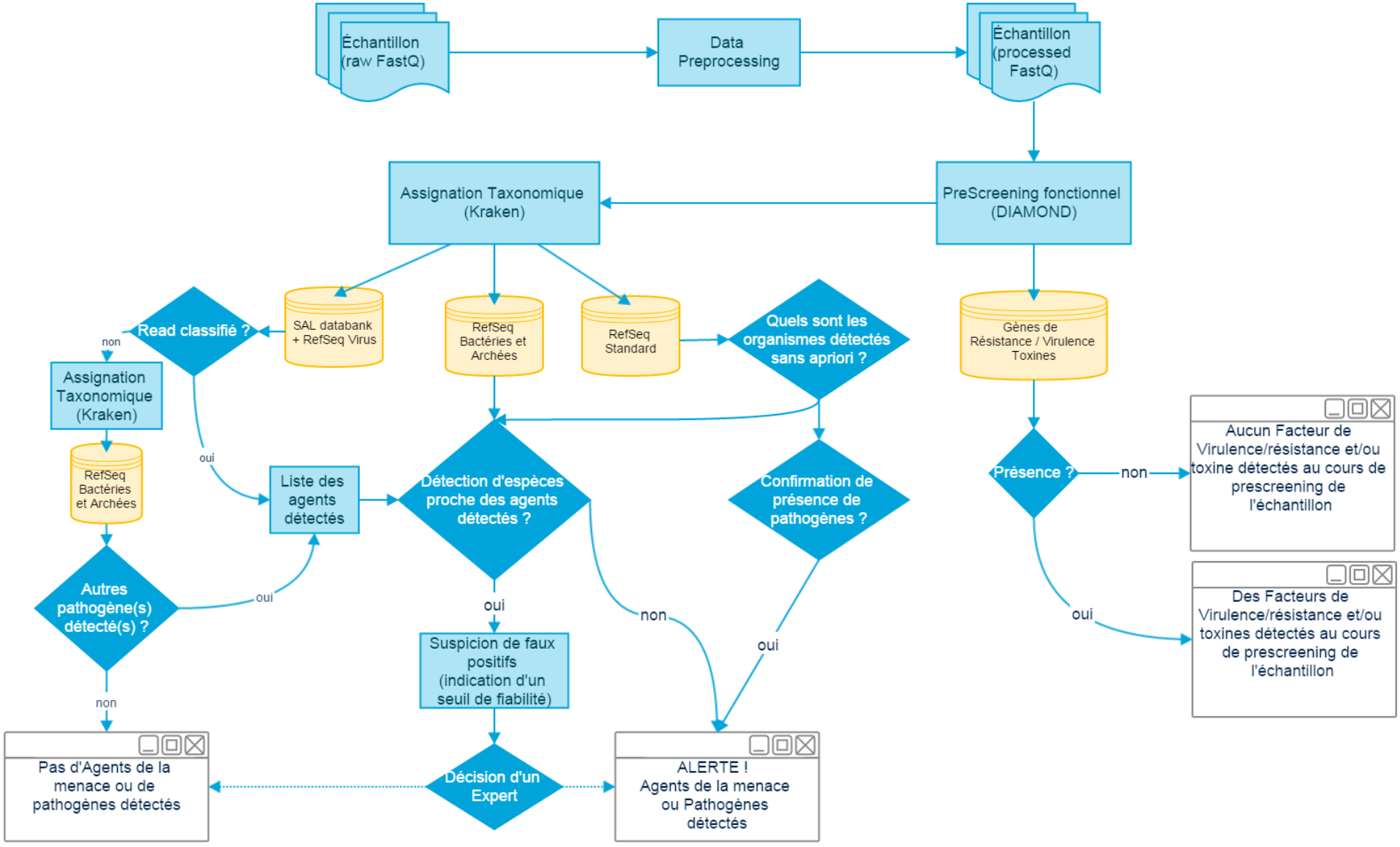
Marine Séjourné - CDD France Génomique

(Mars 2015 - ...)

Plateforme d'analyse métagénomique

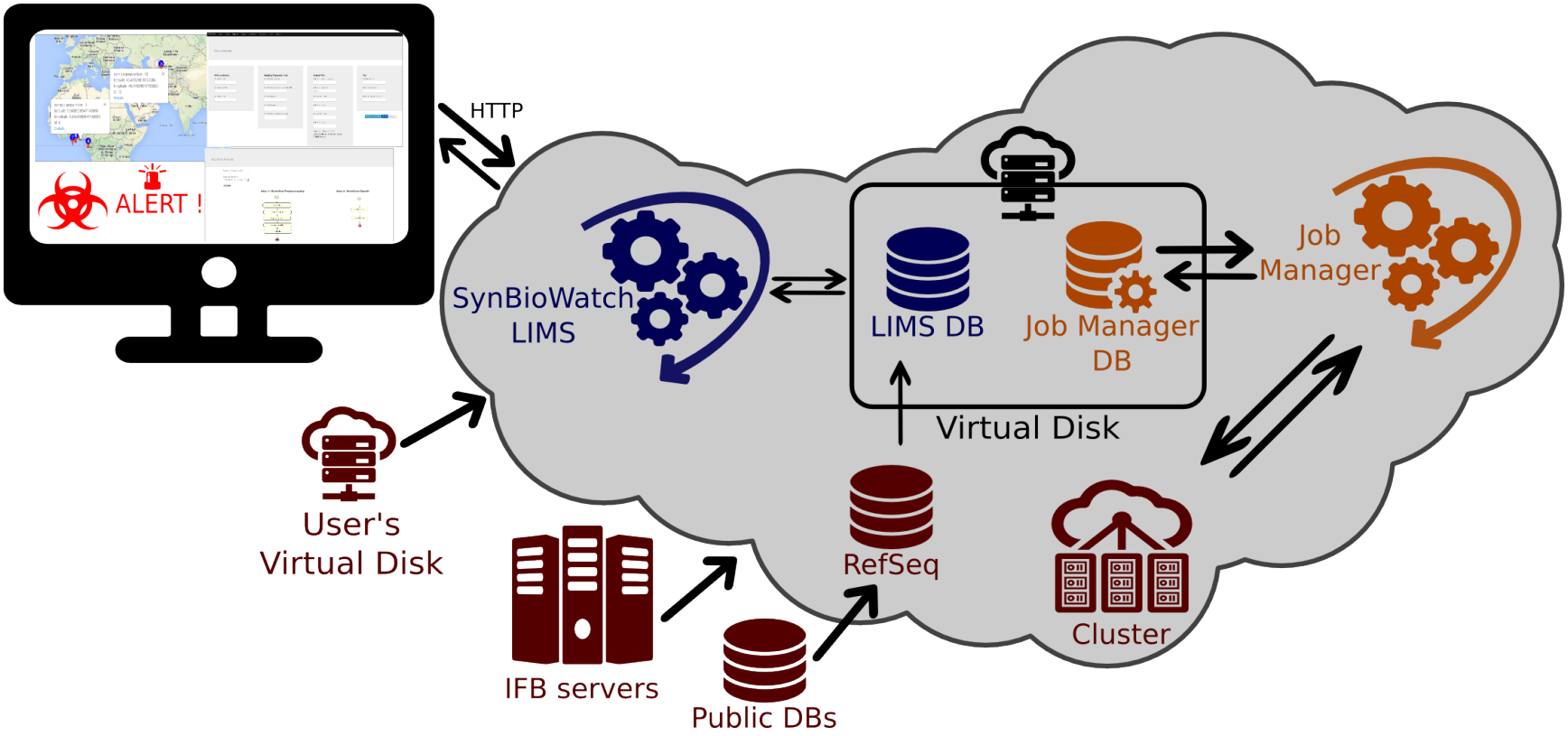
Objectif: Informer sur le développement de l'analyse d'un métagénome au sein de l'équipe.

SynBioWatch : Détection et Identification de pathogènes dans un échantillon environnemental



SynBioWatch : Déploiement au sein du Cloud IFB

- Machine Virtuelle
  - CentOS 6.6 64 bits
  - 4 CPUs / 8 Go de RAM
  - 3 Serveurs web => 2 interfaces web
  - 2 bases de données



Merci de votre attention

Remerciements:

France Génomiqu

LABGeM