PlantReSeq

12 mai 2009

E P G V
Etude du Polymorphisme des Génomes Végétaux
UR1279_DGAP

Contexte

- Valoriser les données issues du séquençage / re-séquençage par SOLEXA et 454
  - ➔ Mise en place / développement d'outils d'analyse
  - ➔ Alignements, assemblage, MDR , SNPs

- Organiser le stockage des données d'intérêt
  - ➔ Quelles données conserver ?
  - ➔ Où les stocker ?
  - ➔ Sous quelles formes ?
  - ➔ Comment les transférer ?

Coordination des ressources et besoins

CNG

. outils de séquençage SOLEXA

. outils d'analyse SOLEXA

. stockage temporaire des données « brutes »

. support commercial Illumina

CNS

. outils de séquençage 454

. outils d'analyse 454

. stockage temporaire des données « brutes »

EPGV
bio-informatique

Equipes « PlantReseq »

. définition des questions biologiques

. besoins bio-informatiques

. données de référence

URGI

. repository center

. outils d'analyse

Environnement informatique

- CNG
  - 2 processeurs Quad Core, 3GHz
  - 74 GB de RAM
  - 81 TB d'espace de stockage

- EPGV
  - PowerEdge T300 :
    - **Quad Core Xeon 2,5GHz**
    - **16 GB de RAM**
    - **4 TB (RAID5)**

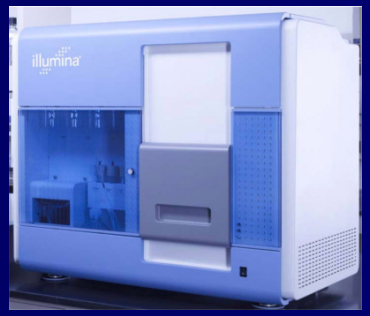
Solexa : caractéristiques sommaires

- GA II – pipeline version 1.0
- Séquençage SBS "Sequencing By Synthesis"

Cycles / Run	Durée	Données bruts (TB)	Données finales (TB)
36	3-4 jours	0,72 TB	0,54 TB
100	~ 9 jours	2,00 TB	1,50 TB

Solexa : acquisition, analyses et stockage des données

Echantillons



Stockage

temporaire

CNG

- analyses « Solexa »
- analyses complémentaires

EPGV

- analyses complémentaires
- stockage des données finales

URGI

stockage pérenne

rapport d'analyse

INRA

Blé : run Solexa

Chromosome 3B trié
995.000.000 bases

54.808.646 séquences
~ 2.10^9 bases
Profondeur : 1,98 X

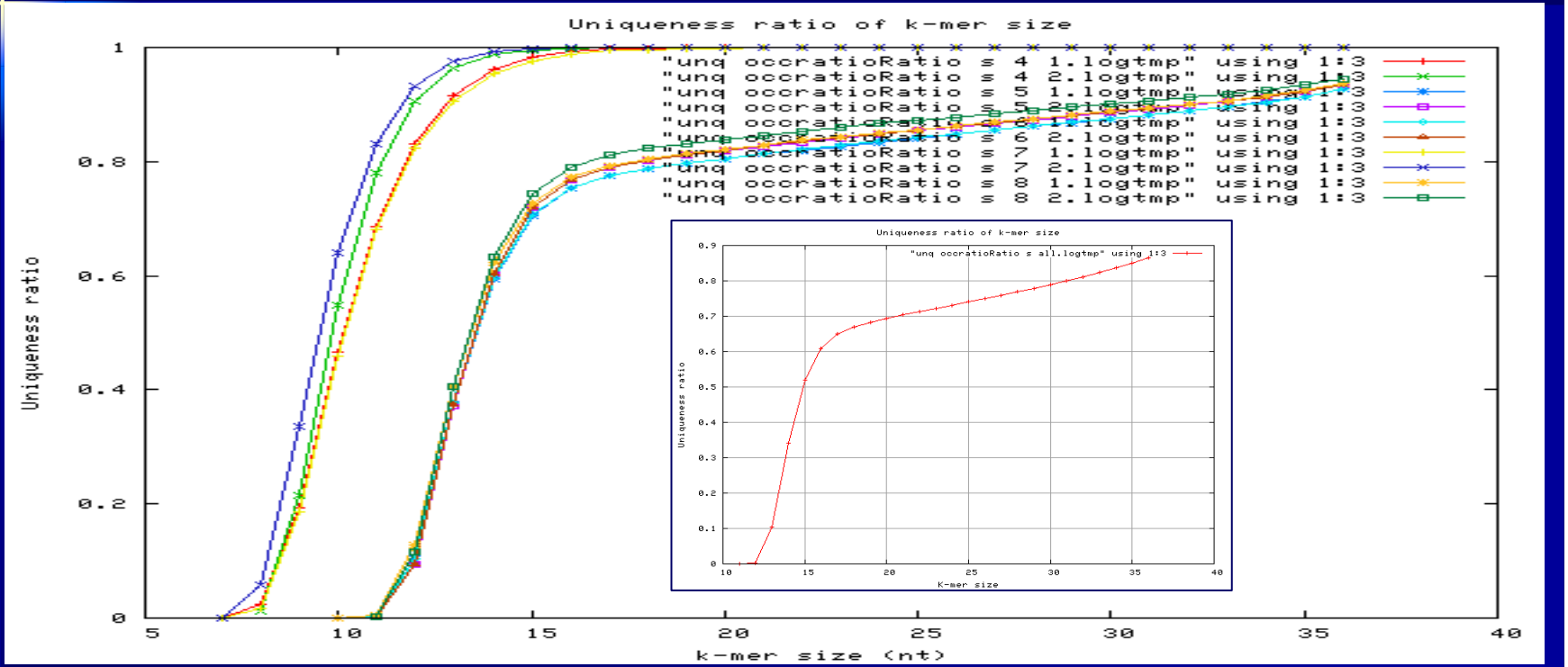
GA II, 36 cycles, PE
3 canaux retenus sur les 5

Blé : constitution d'un index MDR

- MDR : « Mathematically Defined Repeats »
  = occurrence de mots de taille donnée (k-mers) présents dans le jeu de séquences en entrée (short reads Solexa)

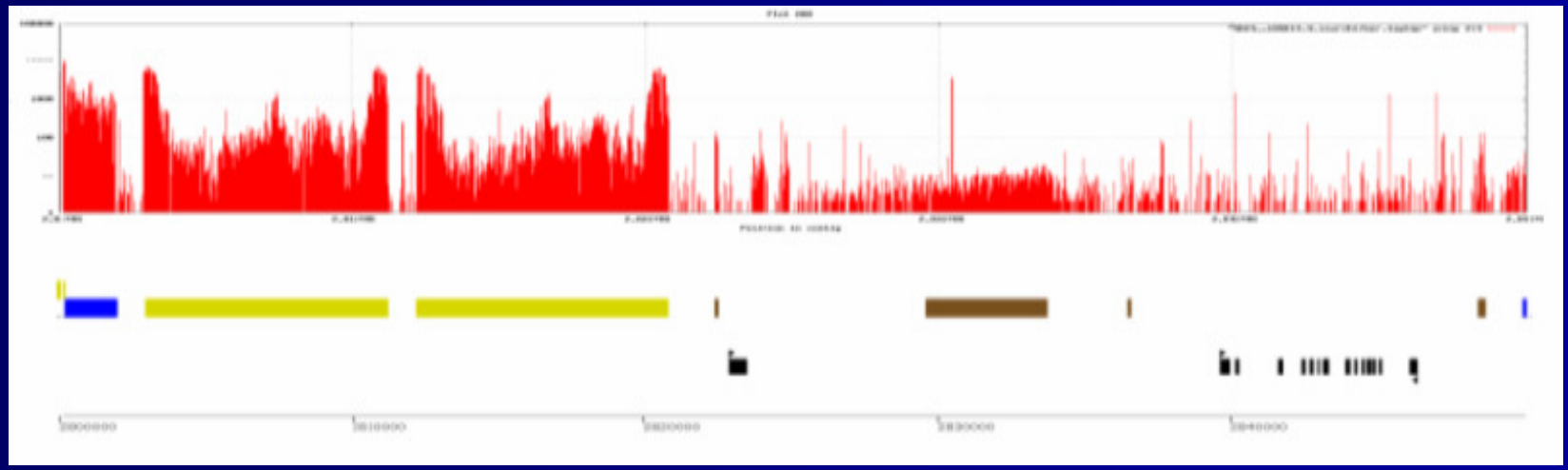
- Outil : Tallymer (package *GenomeTools* http://genometools.org/)
  A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.
  Kurtz S, Narechania A, Stein JC, Ware D.
  BMC Genomics. 2008 Oct 31;9(1):517.

Blé : constitution d'un index MDR



Ratio de motifs uniques en fonction de la taille du k-mer, pour l'ensemble des canaux 5, 6 et 8 retenus pour la création de l'index MDR.
Le ratio de motif unique est le ratio d'un k-mer présent une seule fois dans l'ensemble des k-mers . Le point d'inflexion de la courbe se situe pour k=17.

Blé : constitution d'un index MDR



Annotation de contigs : comparaison de l'annotation « MDR » (haut) avec l'annotation experte correspondante (bas).

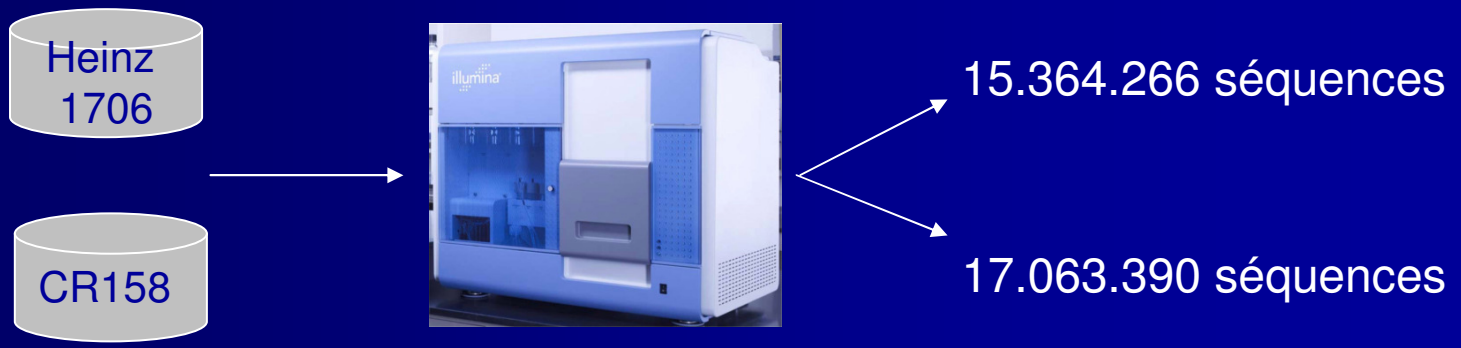
[Image : Frédéric Choulet]

Blé : constitution d'un index MDR

- En cours :
  - Création d'un catalogue des éléments répétés
  - Identification des séquences faiblement répétées

Tomate : run Solexa

198 fragments d'~10 Kb
issus de LR PCR
soit ~ 2 Mbases,
pour chacune de 2 accessions



GA II, 36 cycles, PE,
1 canal par accession

Tomate : prédiction de SNPs

- Outil : MAQ (http://maq.sourceforge.net/)
  Mapping short DNA sequencing reads and calling variants using mapping quality scores.
  Li H, Ruan J, Durbin R.
  Genome Res. 2008 Nov;18(11):1851-8. Epub 2008 Aug 19.

- Génome de référence : 198 fragments de LR PCR séquencés en Sanger correspondant à l'accession Heinz 1708

- Résultats attendus :
  - Heinz 1708 : ~ 0 SNPs (séquençage de l'accession de référence)
  - CR158 : ~ 1 SNPs toutes les 500pb, soit ~ 4.000 SNPs

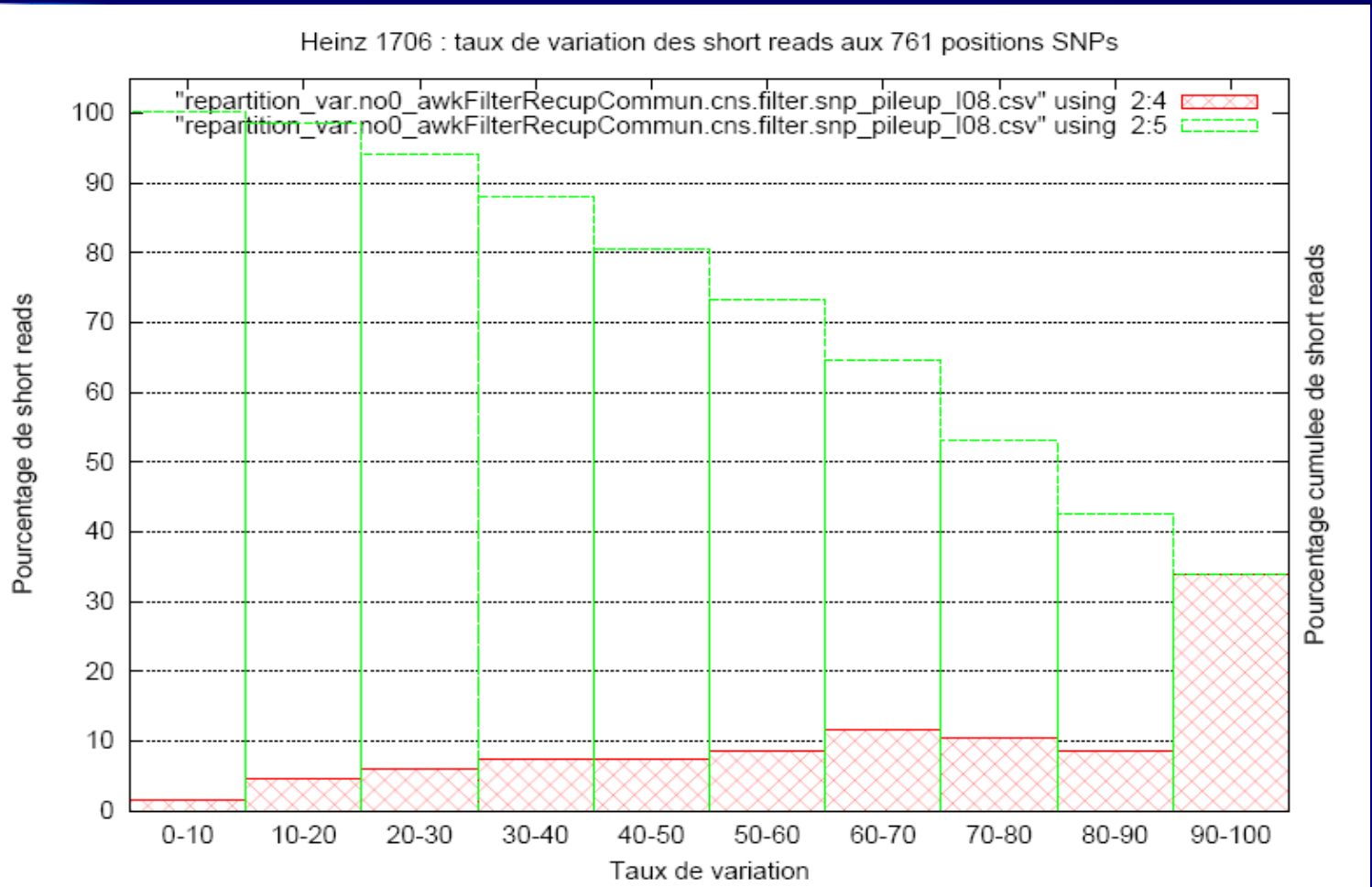
Tomate : prédiction de SNPs

- Résultats obtenus (en cours d'analyse) :

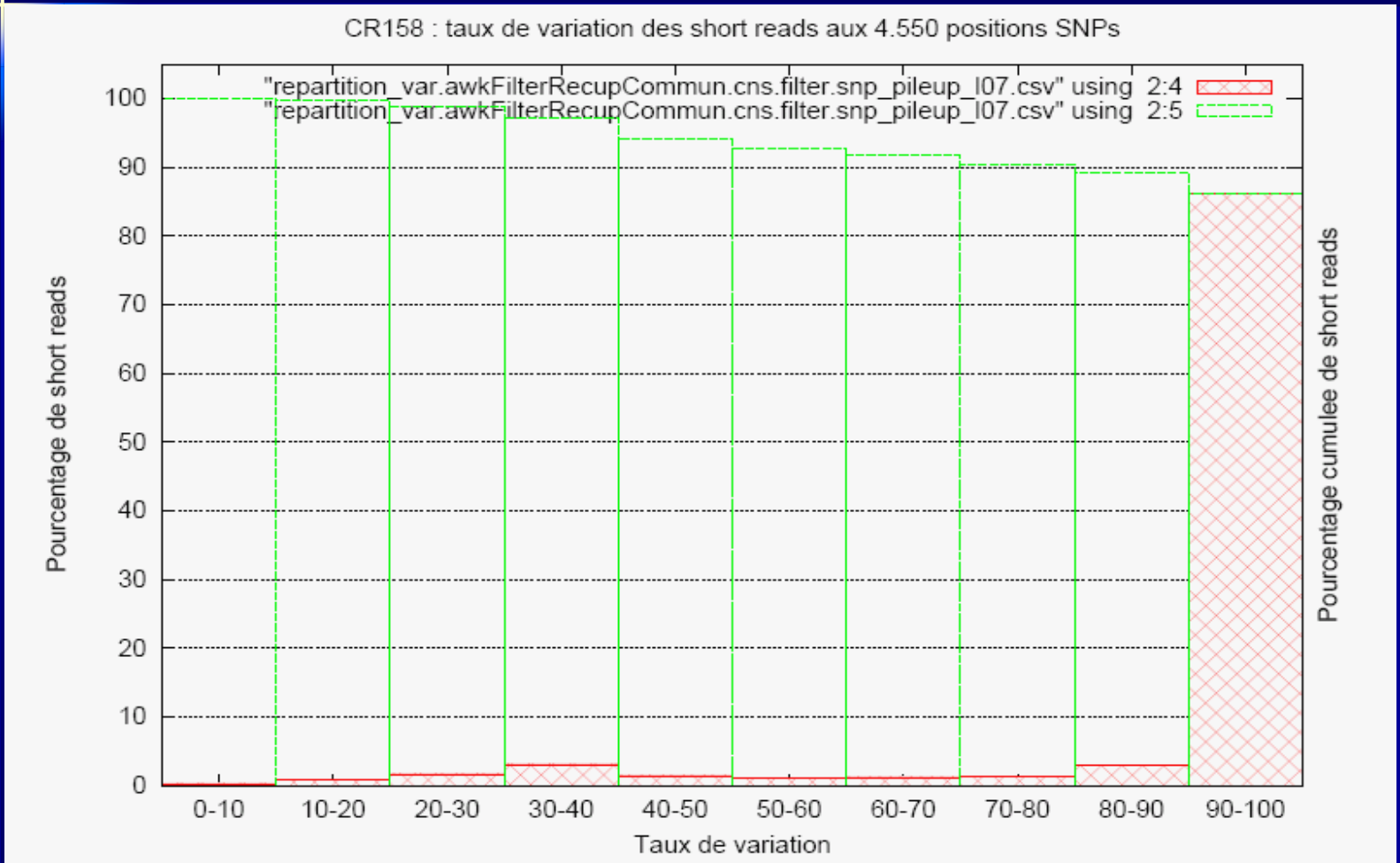
	Heinz 1708	CR158
Profondeur moyenne	~ 254	~ 274
Taux de couverture	94,53%	94,69%
SNPs prédits / référence	1.476	5.242
SNPs prédits / accession	799	4.565

➔ dont 677 communs

Tomate : prédiction de SNPs



Tomate : prédiction de SNPs



Tomate : assemblage *de novo*

- Outil : SHARCGS (http://sharcgs.molgen.mpg.de)
  SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing.
  Dohm JC, Lottaz C, Borodina T, Himmelbauer H.
  Genome Res. 2007 Nov;17(11):1697-706.

- Outil : Velvet (http://www.ebi.ac.uk/~zerbino/velvet/)
  Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
  Zerbino DR, Birney E.
  Genome Res. 2008 May;18(5):821-9.

Mise en perspective

- Suite du projet
  - Mise en place d'une procédure estimant la qualité des données analysées
  - Mise en place d'une procédure d'exportation des données « SNPs » vers GnpSNP

- Problématiques / difficultés
  - Volumétrie des données générées par les NSG : stockage, temps d'analyse
  - Interactions / interopérabilité entre les différentes plateformes (bio)informatiques
  - Évolution rapide des technologies et outils

Remerciements

Martha Gut et son équipe
Yannis Duffourd