Elevage de précision :
L'analyse de données génétiques

François GUILLAUME
Chargé de méthodes à Evolution

Plan

- Illustration personnelle
- Analyses et perspectives

Vision du monde du généticien quantitatif

$$Y = G + E + e$$

- Une observation (Y)
- Un effet Génétique (Aléatoire)
- Des effets fixes

➔ L'estimation des effets génétiques possible grâce à :

- ➔ Déterminisme génétique favorable
- ➔ De larges bases de données
- ➔ Un nombre relativement faible d'individus statistiquement indépendants (Ne ~100 pour la Holstein mondiale)

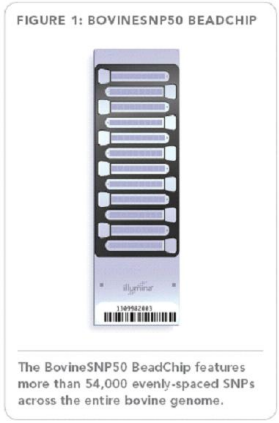
➔ Une valorisation de données d'élevages bâtie sur une construction logique robuste

La révolution génomique

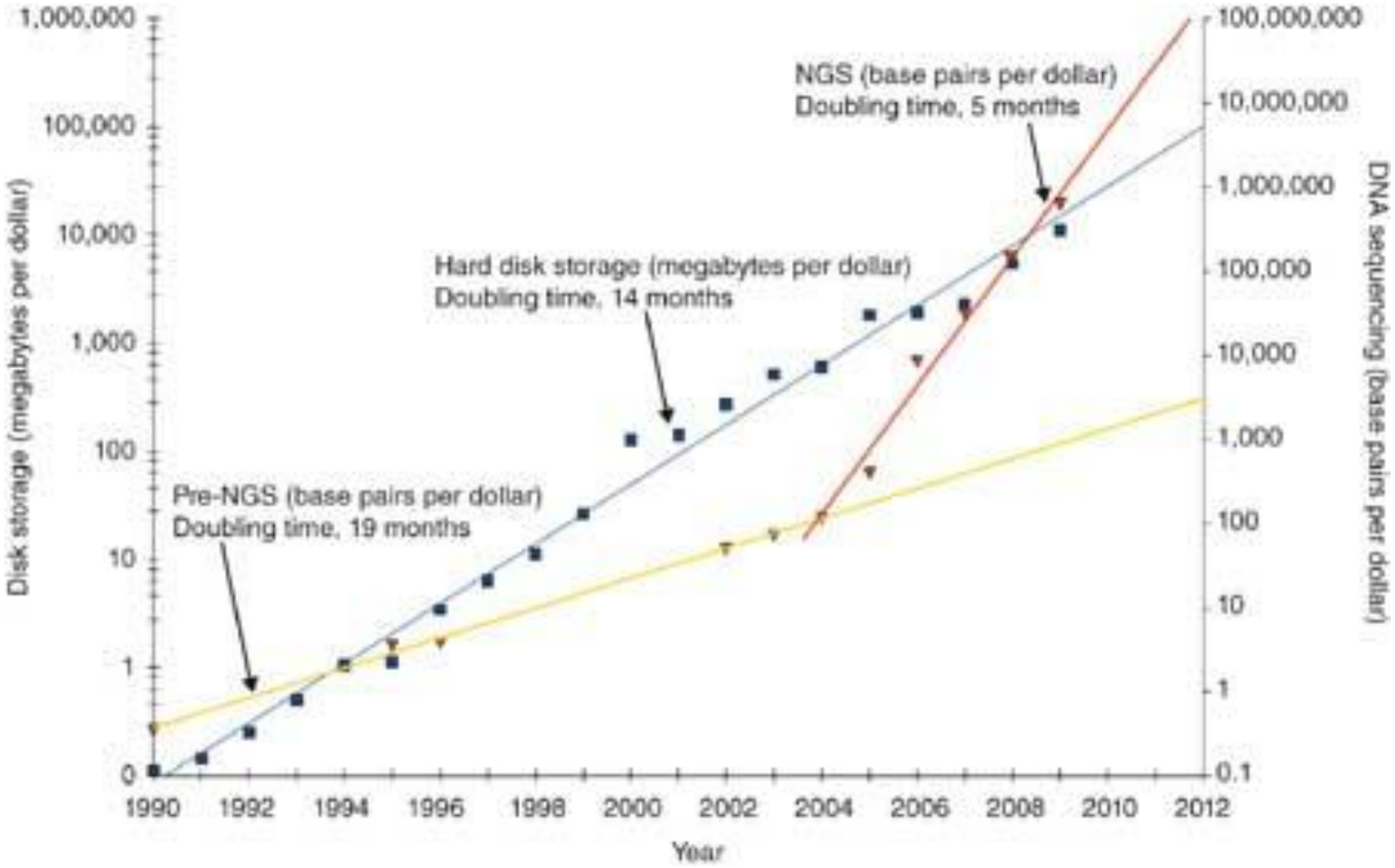
- Deux modèles de l'effet génétiques :
  - Si deux animaux sont apparentés à un certain degré **d**, ils partageront une proportion **p** de leur matériel génétique
  - Si deux individus partagent une proportion **p** de leur matériel génétique ils sont apparentés à un certains degré **d**

➔ La possibilité d'avoir des marqueurs génomiques a permis de substituer le second modèle au premier

- Une chute vertigineuse du coût de l'information génomique
  - 2001 : 43 marqueurs micro satellites
  - 2008 : 50 000 SNP
  - 2012 : Séquences complètes (~3Milliards de bases)



Evolution des coûts de typages



Nature Methods **7**, 495 - 499 (2010)
doi:10.1038/nmeth0710-495

La révolution génomique : conséquences

- Une augmentation parallèle des difficultés techniques
  - 2001 : Mainframe 24 CPU
  - 2008 : Cluster 400 CPU
  - 2012 : Cluster de calcul ~2000 CPU

  - 2001 : 1 animal quelques octets
  - 2008 : 1 animal 0,7 Mo
  - 2012 : 1 animal ~ 10Go fastq.gz / ~40 Go BAM / 0,5 Go Vcf

  - 2001 : I/O => direct, beaucoup d'analyse in-memory
  - 2008 : I/O : de quelques secondes à quelques jours
  - 2012 : I/O : Envois de disques durs pour transférer les données ;-)

➔Une obligation d'acquérir des compétences algorithmiques / informatiques

➔Les connaissances métiers sont régulièrement les facteurs clefs de réussites

BIG DATA en génétique : point de situation

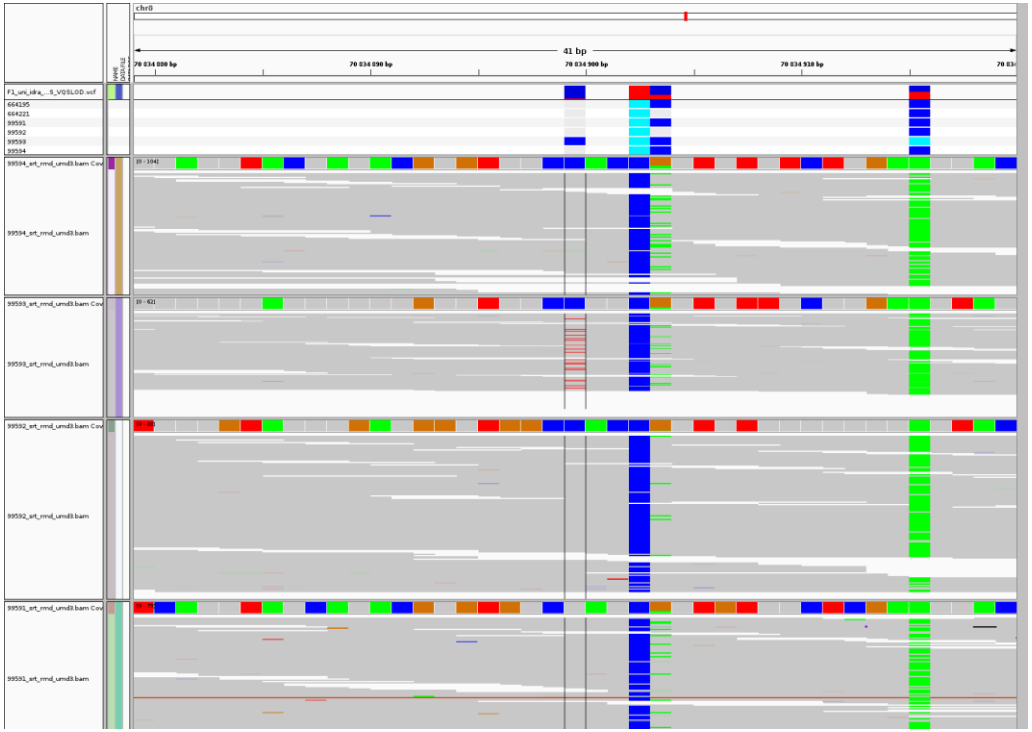
- Des réussites évidentes :
  - Augmentation de la précision des évaluations génomiques
  - Meilleure valorisation potentielle de phénotypes nouveaux
  - Identification de mutations causales à fort impact

…des nuances à apporter !

- Informativité réelle des données
- Dépendance des acquis antérieurs
- Simplicité des modèles utilisés

Informativité réelle

- Informativité réelle des données :
  - Gains moyens de précisions des évaluations génomiques permis par l'utilisation de séquence => Très faible
  - Efficacité de l'imputation 7000 SNP => 50 000 SNP (r2 >0,98)
  - Format de données de séquences VCF => Stockage de ce qui est « a priori » utile



Dépendance aux acquis historiques

- Génotypages :
  - Standard de fait en bovin (Illumina):
    - Qualité du traitement du signal
    - Qualité de la définition du problème de génotypage

- Phénotypes :
  - Réussites pour les phénotypes « historiques »
    - Attention aux évolutions potentielles !
    - Des opportunités de revisiter certains anciens phénotypes demeurent
  - Plus de difficultés sur les nouveaux phénotypes
    - Standards ?
    - Qualité / pertinence de la mesure

Simplicité des modèles

« Ce qui est simple est faux, ce qui ne l'est pas est inutilisable »

- Equilibre entre temps de calcul et complexité
  - Hypothèses simplificatrices pour réaliser des évaluations en un temps raisonnable
  - Problème d'estimabilité des termes de modèle trop complexes
- Difficultés de valider les phénomènes complexes et/ou rares
- Intérêts pour la filières d'effets complexes/rares

Take home message

- Les technologies générant des quantités importantes de données à bas coût sont de belles opportunités pour l'élevage

- La transformation de ces opportunités en succès dépend de :

  ➔ La capacité des hommes du métier à comprendre ces nouvelles technologies

  ➔ La capacité des hommes du métier à réinterroger les problèmes d'élevage pour identifier les nouvelles réponses permises par ces nouvelles technologies