Assemblage de novo de génomes humains en quelques heures

Sébastien Boisvert[1,2], François Laviolette[3] et Jacques Corbeil[1,2]

[1] Département de médecine moléculaire, Université Laval, Québec, Québec, Canada
[2] Centre de recherche en infectiologie, Centre hospitalier universitaire de Québec, Pavillon CHUL, Québec, Québec, Canada
[3] Département d'informatique et de génie logiciel, Université Laval, Québec, Québec, Canada

Journée annuelle de la recherche
Faculté de médecine, Université Laval
7 juin 2011

Doctorat débuté en janvier 2010

Prénom	Sébastien
Nom	Boisvert
Évalué par un jury	Oui
Local	Grand Hall
Heure	13h00 à 14h30
Numéro d'affiche	132
Comité	N

Largeur: 128 cm, hauteur: 96 cm

Introduction

- technologies de séquençage à haut débit d'ADN (454, Illumina, SOLiD) -> révolutionné les sciences génomiques

- liste des polymorphismes d'un seul nucléotide et des petites délétions et insertions pour un individu

- variations structurales (délétions, insertions, duplications en tandem ou entrelacées, inversions & translocations) -> un rôle dans certaines maladies (obésité, cancers, autisme, ou autre)

- assemblage de novo: assembler les millions de courtes séquences en séquences contiguës sans utiliser une référence

- assemblage de novo permet d'obtenir, en théorie, toutes les variations structurales; mais les outils permettant ce type d'analyses sont en développement

Matériel et méthodes

- technologies de séquençage massivement parallèles -> basées sur deux principes
- le séquençage par synthèse
- l'utilisation d'un support solide qui permet d'analyser des millions de réactions simultanément (**Figure 1**)

- longueur des séquences obtenues: entre 50 et 100 nucléotides pour la technologie Illumina

- nous développons un assembleur massivement parallèle de novo de génomes appelé Ray

- utilise l'interface de passage de messages pour permettre à beaucoup d'ordinateurs de communiquer
- calcul colossal collectif avec le colosse -- le superordinateur de l'Université Laval.

- chaque séquence brisée en sous-séquences successives (voir la **Figure 2**)

- traiter toutes les séquences avec cette approche permet de construire un graphe contenant toutes les sous-séquences dans les données avec des flèches liant celles qui sont successives (**Figure 3a, 3b, 3c**)

- régions répétées (**Figure 3d, 3e**)

- simulation de séquences du génome de la bactérie Streptococcus pneumoniae (numéro GenBank: AE007317)
- 6 milliards de séquences Illumina de 75 nucléotides en paires (génome humain; fragments de ~190 nucléotides; # SRA010766; 431 gigaoctets)

- comparaison de notre assemblage à la référence du génome humain (version: hg18; avec BLAT)

Figure 1 | Séquençage des extrémités de fragments d'ADN.

a, ADN chromosomique. **b**, Fragments d'ADN obtenus par sonication et réparation des extrémités. **c**, Ligation d'adapteurs aux fragments d'ADN. **d**, Hybridation sur support solide. **e**, Les séquences des extrémités de tous les fragments sont obtenues par amplification en pont et séquençage par synthèse en parallèle.

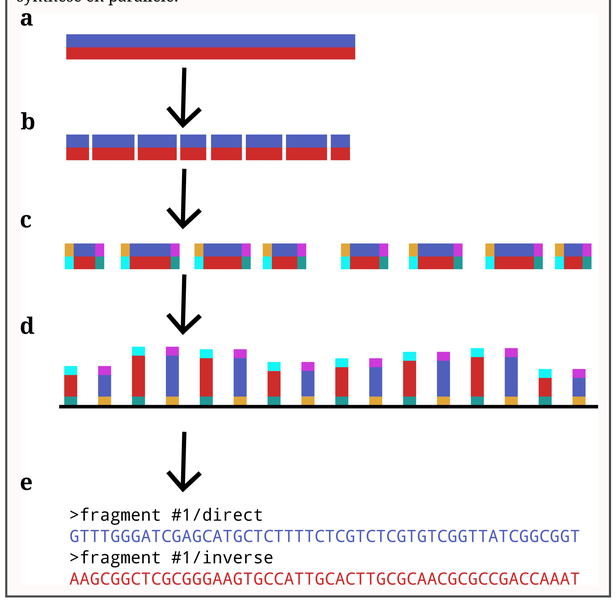


Figure 2 | Transformation d'une séquence.

a, Une séquence d'ADN. **b**, Représentation en sous-séquences. **c**, Relation entre deux sous-séquences successives.

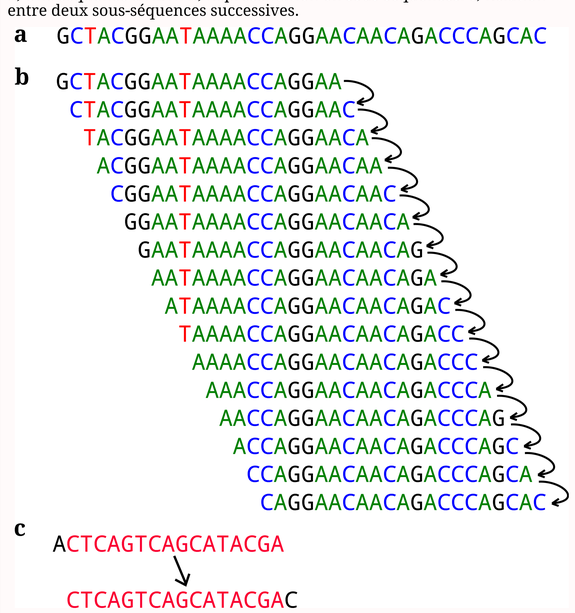
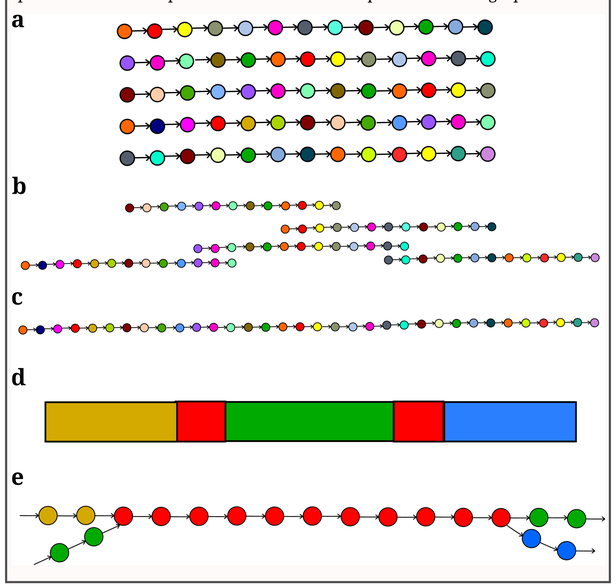


Figure 3 | Redondance de l'information dans les séquences.

a, Séquences représentées en sous-séquences consécutives. **b**, Chevauchements des sous-séquences. **c**, Information contenue dans les chevauchements. **d**, Régions répétées dans un génome. **e**, La région répétée a plusieurs entrées et plusieurs sorties dans la représentation en graphe.



Résultats

- redondance des sous-séquences avec et sans erreurs de séquençage (**Figure 4**; données de S. pneumoniae)

- génome humain assemblé en 1 758 326 séquences (longueur moyenne: 1333); totalisant 2 345 582 359 nucléotides

- 5 heures sur 512 coeurs de calcul

- distribution de la redondance des sous-séquences pour le génome humain (**Figure 5a et 5b**)

- distribution de la longueur des séquences assemblées (**Figure 5c et 5d**)

- comparaison de notre assemblage avec la référence du génome humain (**Tableau 1**)

Figure 4 | Distribution de la redondance des sous-séquences avec et sans erreurs de séquençage.

a, Distribution sans erreurs de séquençage. **b**, Dérivée numérique sans erreurs de séquençage. **c**, Distribution avec erreurs de séquençage. **d**, Dérivée numérique avec erreurs de séquençage.

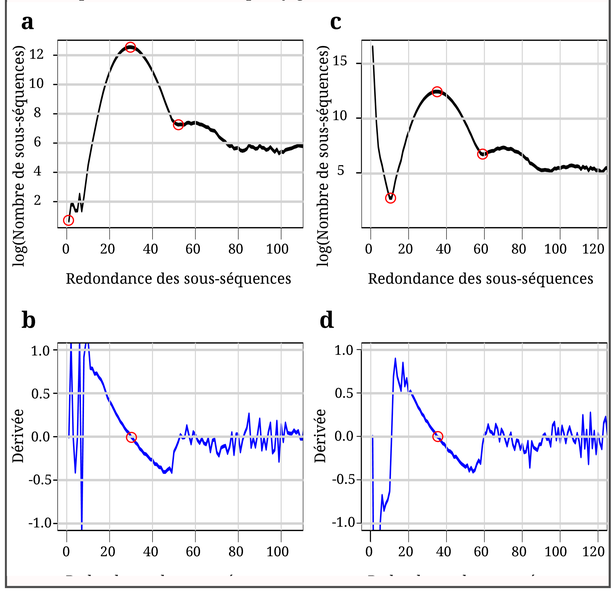


Figure 5 | Sous-séquences provenant du génome humain.

a, Distribution de la redondance pour un génome humain. **b**, Nombre cumulatif de sous-séquences. Le nombre cumulé final est 5 128 089 556. **c**, Distribution de la longueur des séquences assemblées **d**, Longueur cummulative des séquences assemblées. Le cumulatif final est 2 345 582 359.

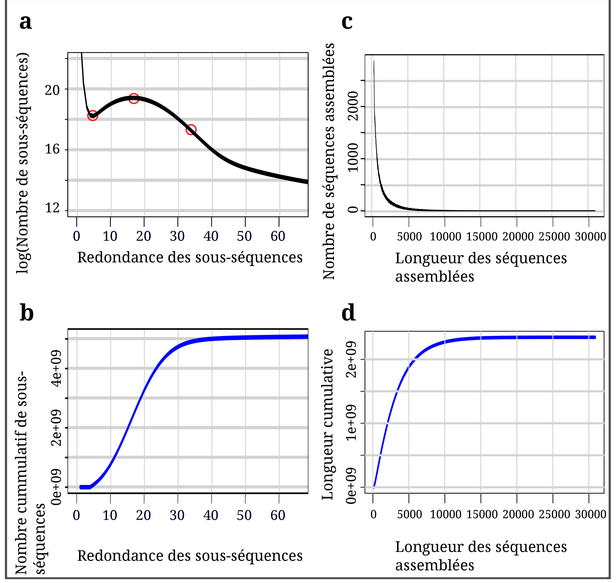


Tableau 1 | Différences entre le génome assemblé de novo et la référence du génome humain.

Purines: A, G. Pyrimidines: C, T.

Substitution	Nombre	Délétion	Nombre	Insertion	Nombre
G->A	130808	1	32553	1	30516
C->T	129933	2	8238	2	7041
T->C	120423	3	3291	3	3145
A->G	120033	4	3892	4	3683
C->A	32306	5	1210	5	1169
G->T	32132	6	785	6	610
C->G	31795	7	318	7	266
G->C	31755	8	465	8	345
A->C	30066	Total	50752	Total	46775
T->G	29601				
A->T	26474				
T->A	26443				
Total	741769				

Discussion

- avec Ray, 6 milliards séquences assemblées en 5 heures avec 512 coeurs de calcul.
- ALLPATHS-LG du Broad Institute of MIT and Harvard = 3 semaines pour assembler un génome humain
- Ray est donc plus rapide

- assemblage de novo surtout utile pour les organismes dont le génome est totalement inconnu

- assemblage de novo -> utile pour l'analyse de métagénomes

- logiciel (et les paramètres) utilisé pour comparer -> ne permet pas de détecter des grandes insertions ou délétions

- il faut donc utiliser un autre logiciel (comme Exonerate) pour faire la comparaison

- Ray: logiciel libre, utilisateurs partout dans le monde

- nous allons participer à l'Assemblathon 2 http://assemblathon.org

Disponibilité

- http://denovoassembler.sf.net

- Ray 1.4.0 = 24523 lignes de code en C++

- http://github.com/sebhtml/ray

Remerciements

- SB est boursier au doctorat des IRSC
- FL est financé par le CRSNG
- JC est financé par les IRSC et a une chaire de recherche du Canada
- Calcul Canada et le CLUMEQ pour les ressources de calcul

Références

- Alkan, Coe & Eichler (2011) Nature Reviews Genetics doi:10.1038/nrg2958
- Batzer & Deininger (2002) Nature Reviews Genetics doi:10.1038/nrg798
- Boisvert et al. (2010) J. Comp. Biol. doi:10.1089/cmb.2009.0238
- Boisvert et al. (2011) RECOMB Satellite Workshop on Massively Parallel Sequencing
- Flicek & Birney (2009) Nature Methods doi:10.1038/nmeth.1376
- Pevzner et al. (2001) PNAS doi:10.1073/pnas.171285098
- Shendure & Ji (2008) Nature Biotechnology doi:10.1038/nbt1486.
- Simpson et al. (2009) Genome Research doi:10.1101/gr.089532.108
- Zerbino & Birney (2008) Genome Research doi:10.1101/gr.074492.107