27 mai 2010, Douxième journée annuelle de la recherche
Pavillon Ferdinand-Vandry, Faculté de médecine, Université Laval
10 minutes, 14h30-14h45, Salle 1811

Ray: assemblage parallèle de génomes à partir de séquences Illumina très courtes en paires.

Sébastien Boisvert[1,2],
François Laviolette[3,4],
Mario Marchand[3], et
Jacques Corbeil[1,2]

1 Faculté de médecine, Université Laval, Canada
2 Centre de recherche en infectiologie, Canada
3 Faculté des sciences et de génie, Université Laval, Canada
4 Dept. of Computer Science, University College London, United Kingdom

UNIVERSITÉ LAVAL

La vision

« En 20 ans, nous avons réussi à mettre sur le marché des tests qui identifient les microbes en une heure plutôt qu'en 48 h. Aujourd'hui, je rêve de remplacer la microbiologie pasteurienne par **la microbiologie à base d'ADN.** »

- Dr. Michel G. Bergeron, O.Q., M.D., FRCPC
  professeur titulaire
  directeur du Centre de recherche en infectiologie
  Grand Diplômé de l'Université Laval

Rapport annuel 2008-2009, Faculté de médecine, Université Laval.

L'ADN

La structure de l'ADN a été décrite en 1953.

Watson & Crick, **Nature**, 1953
> http://dx.doi.org/doi:10.1038/171737a0

L'ADN encode le fonctionnement des organismes vivants.

http://www.nature.com/nature/dna50/archive.html

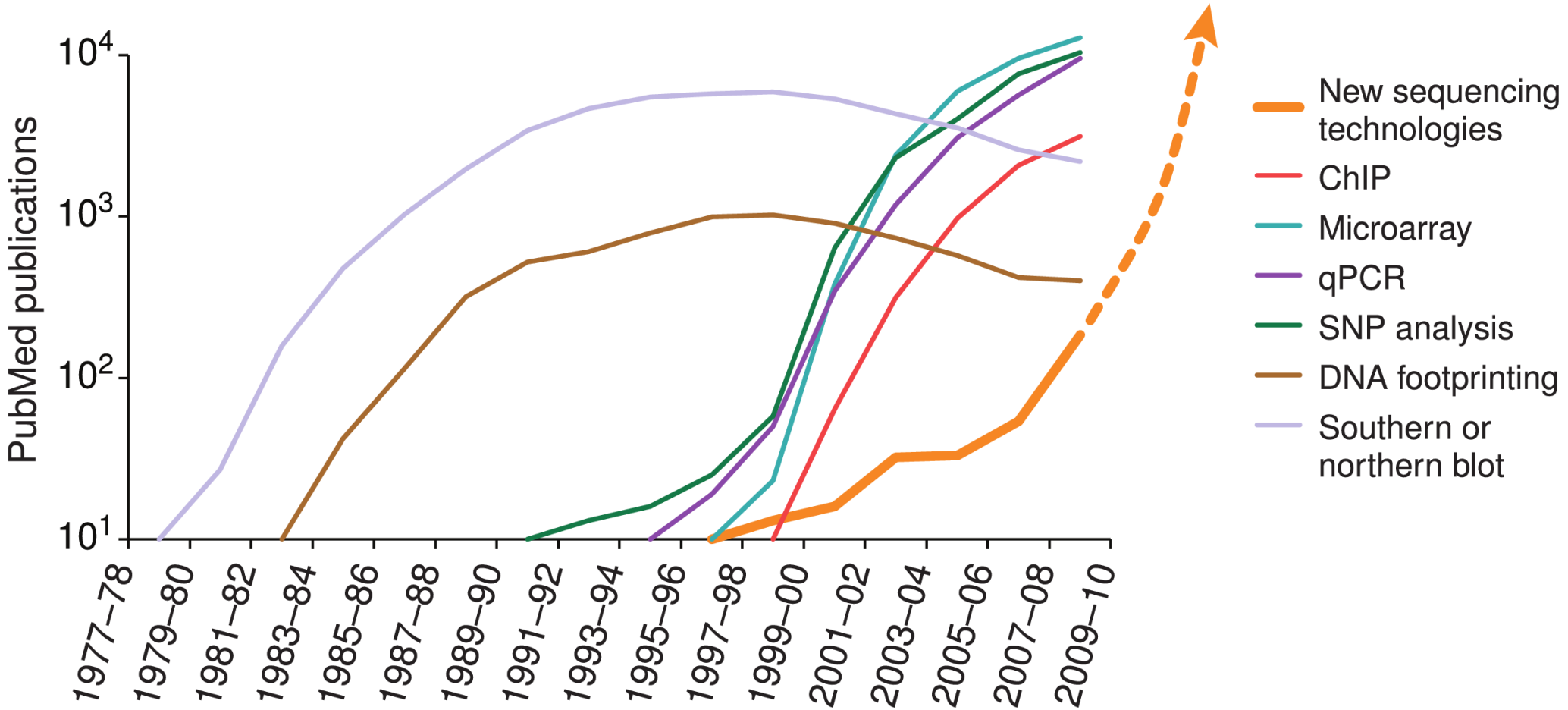
La séquence de lettre (A,T,C, et G) de l'ADN est appellée génome — lequel contient quelques centaines de milliers de lettres (mycoplasme) jusqu'à quelques milliards (humain).

Médecine technologique

Des technologies pour
1) **détecter** et
2) **séquencer** l'ADN
ont été développées depuis 1953.

Fan, Chee & Gunderson, **Nature Reviews Genetics**, 2006
> http://dx.doi.org/doi:10.1038/nrg1901

Nombre de publications avec les mots clés (en anglais) de technologies de détection et de séquençage d'acides nucléiques au fils des années.



Les nouvelles technologies de séquençage (en anglais: New sequencing technologies) sont très utilisées!

Kahvejian, Quackenbush & Thompson, **Nature Biotechnology**, 2008 > http://dx.doi.org/doi:10.1038/nbt1494

Détecter l'ADN

Est-ce que mon corps contient de l'ADN de *C. difficile*?

Séquencer l'ADN

Si mon corps contient de l'ADN de *C. difficile*:

Quelle est la variabilité génétique de ma population de *C. difficile*?

Son génome encode-t-il des mécanismes de résistance?

Quelles sont ses failles qui peuvent aider la médecine moderne à le vaincre?

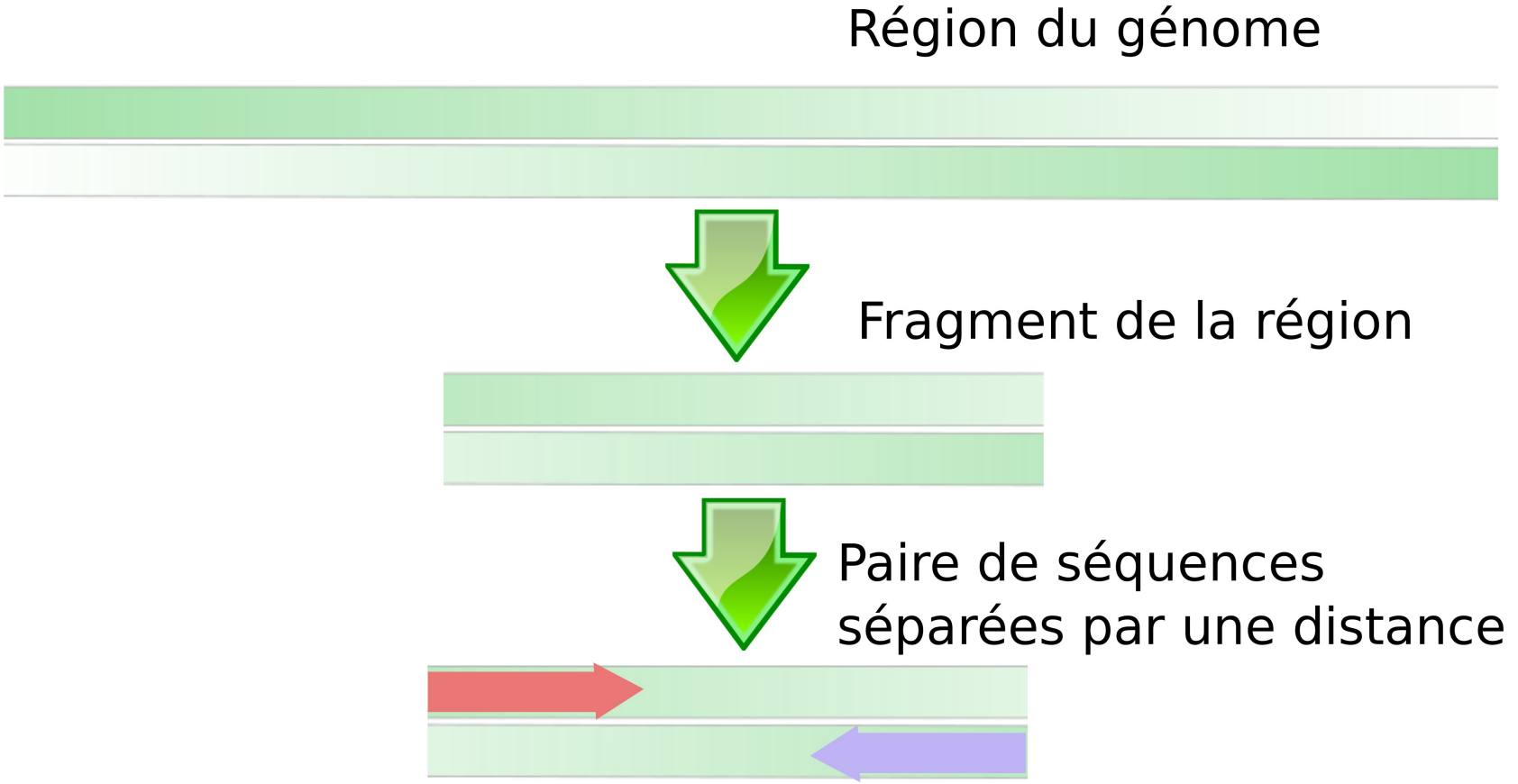
Assemblage de génome

calcul de la séquence du génome à partir des données obtenues lors du séquençage de l'ADN

Avant tout, l'assemblage est une tâche (bio-)informatique

Pop, **Briefings in Bioinformatics**, 2008
> http://dx.doi.org/doi:10.1093/bib/bbp026

Séquencer un génome



De nos jours, on obtient des millions de paires de séquences numériques en moins d'une semaine.

Bentley et al., **Nature**, 2008
> http://dx.doi.org/doi:10.1038/nature07517

Représentation informatique d'une séquence d'ADN

GCTACGGAATAAAACCAGGAACAACAGACCCAGCAC

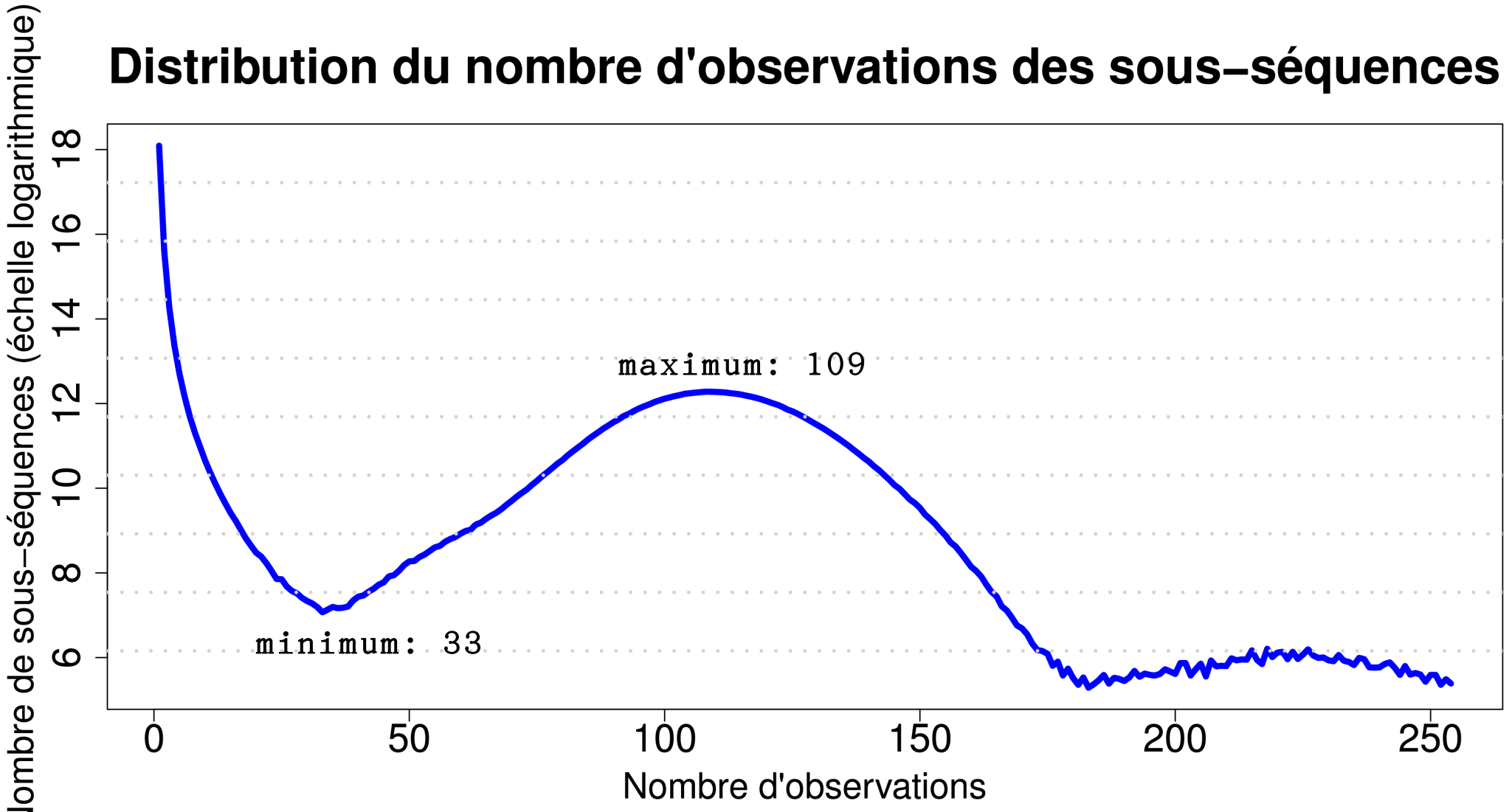
(séquence de 36 lettres)

(sous-séquences de 21 lettres)

Deux sous-séquences liées par une flèche sont identiques si l'on enlève la première lettre de la première et la dernière de la deuxième!

Flicek & Birney, **Nature Methods**, 2009
> http://dx.doi.org/doi:10.1038/nmeth.1376

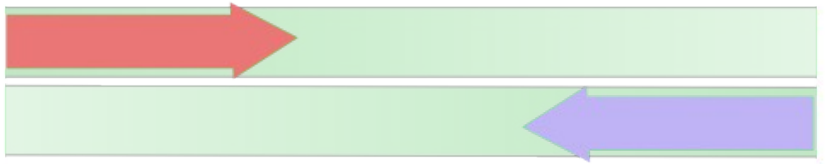
Observer les sous-séquences dans les données

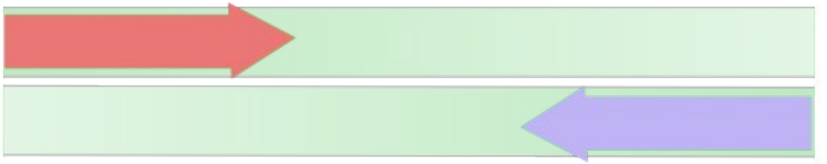


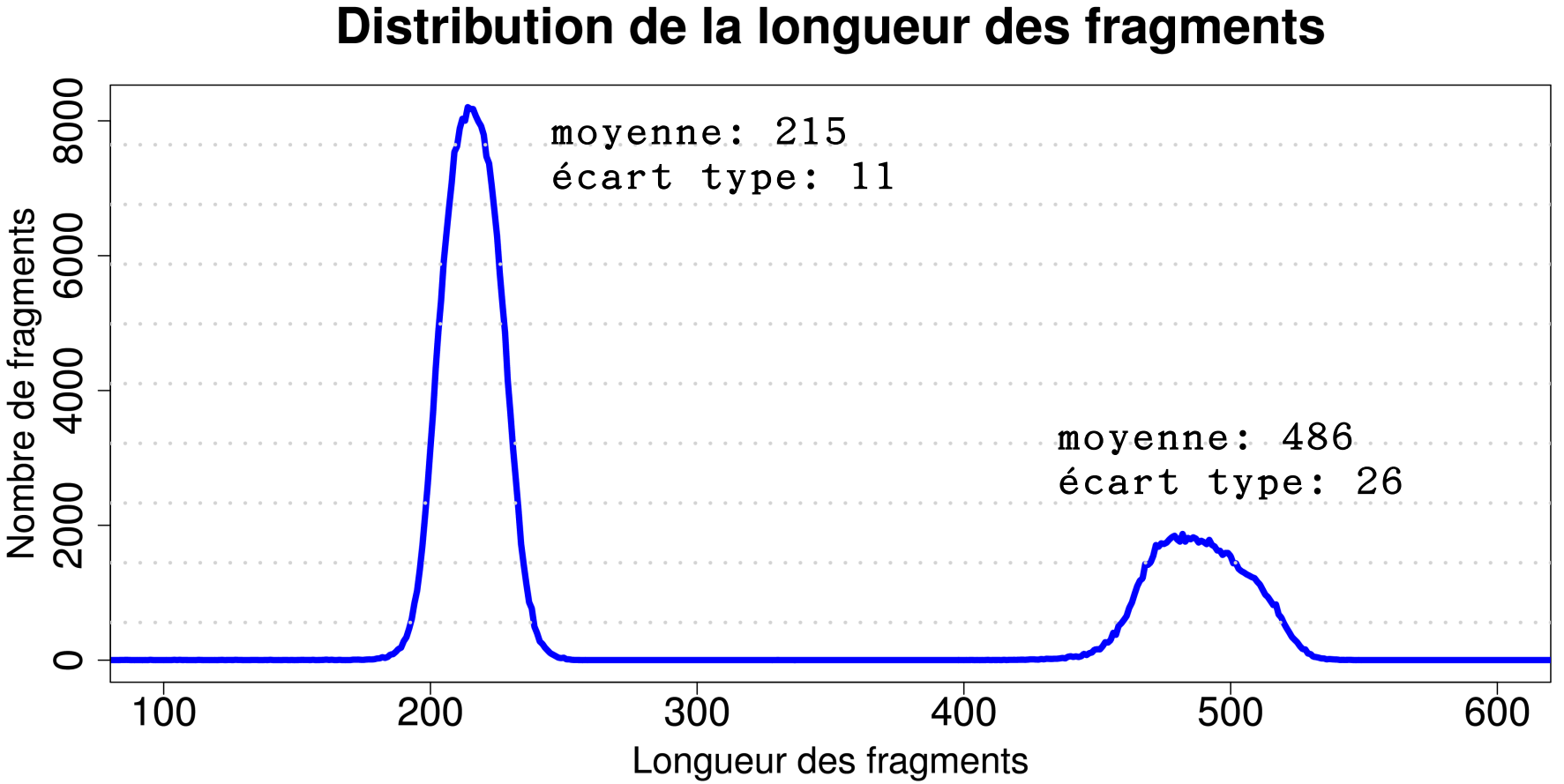
Les sous-séquences avec très peu d'observations sont des erreurs.

(résultats avec des séquences du génome de *E. coli* K12 MG1655 obtenue avec la technologie Illumina)

Distance physique entre les séquences d'une paire

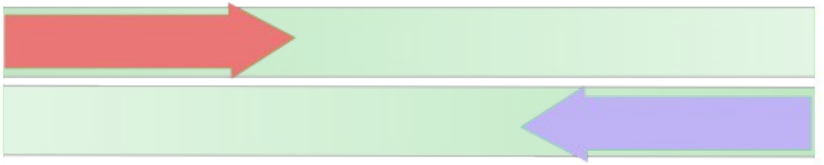






(résultats avec des séquences du génome de *E. coli* K12 MG1655 obtenue avec la technologie Illumina)

Ray



Mon project de doctorat:

Ray est un assembleur pour séquences très courtes en paires.

En calcul parallèle, les processeurs sont numérotés — on les appelle les rangs.

Ray distribue les paires de séquences sur les rangs.

Les sous-séquences sont aussi distribuées sur les rangs.

Ray distribue donc les données et les calculs.

À partir de millions de paires de séquences Illumina, Ray génère 126 séquences qui couvrent 98.17% du génome d'*E. coli* en 47 minutes sur 25 processeurs.

Comparaisons

Assembleur	Séquences	Nucléotides	Longueur moyenne	N50	Longueur maximale	Couverture du génome	Séquences erronées	Changements de nucléotide	Insertions et délétions	Temps d'exécution
EULER-SR	1761	4099614	2328	3429	19094	0.8708	13	1173	7355	143m27.647s
Velvet	83	4542631	54730	125611	311586	0.9657	28	456	895	31m51.199s
Ray	126	4590771	36434	72499	174569	0.9817	0	2	4	47m18.992s
ABySS	154	4661190	30267	56703	174288	0.9840	0	233	9	44m53.714s

> Avec ces paires de séquences Illumina, Ray produit moins de séquences (bon), moins d'erreurs, et effectue tout cela sur plusieurs processeurs Intel, AMD, ou autres.

> Ray, l'assembleur développé au Centre de recherche en infectiologie de l'Université Laval, est meilleur pour produire des séquences de génomes bactériens.

En plus, Ray est un logiciel libre!

http://denovoassembler.sf.net/

http://genome.ulaval.ca/corbeillab

Remerciements

Sébastien Boisvert est un étudiant boursier des
Instituts de Recherche en Santé du Canada (IRSC)
doctorat: 200910GSD-226209-172830 et
maîtrise: 200902CGM-204212-172830

Jacques Corbeil est titulaire de la
Chaire de recherche du Canada en génomique médicale.

Nous remercions la Fondation canadienne pour l'innovation pour le financement de l'infrastructure (ls30.genome.ulaval.ca) et le CLUMEQ et Calcul Canada pour l'accès à colosse.clumeq.ca.

Nous remercions également la communauté des logiciels libres GNU, Linux, et Open-MPI.

Financé par:

IRSC CIHR
Instituts de recherche en santé du Canada
Canadian Institutes of Health Research

Venez voir mon affiche!