Troisièmes rencontres autour de la plate-forme bio-informatique 18 octobre 2005

Apport de la combinaison des séquençages légers (1-2x) et des cartes d'hybrides d'irradiation denses : Application au génome canin

Christophe HITTE & Thomas DERRIEN

Génétique & Développement

UMR 6061 - CNRS
Université Rennes 1

1: 6x10.6 Lectures

2: 1,16x10.6 Contigs (ave:1,4 kb)

3: 5x10.5 Contigs > 3kb (ave : 8,6 kb)

4: 75 % couverture

5: 18 473 gènes canins (24 657 humains)

6: 150 000 microsatellites

7: 974 000 polymorphismes (SNP)

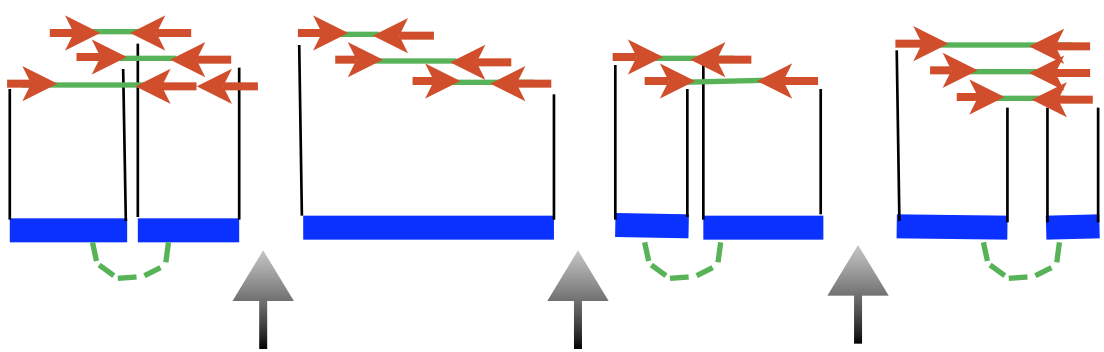
8: Élements répétés (SINE, LINE)

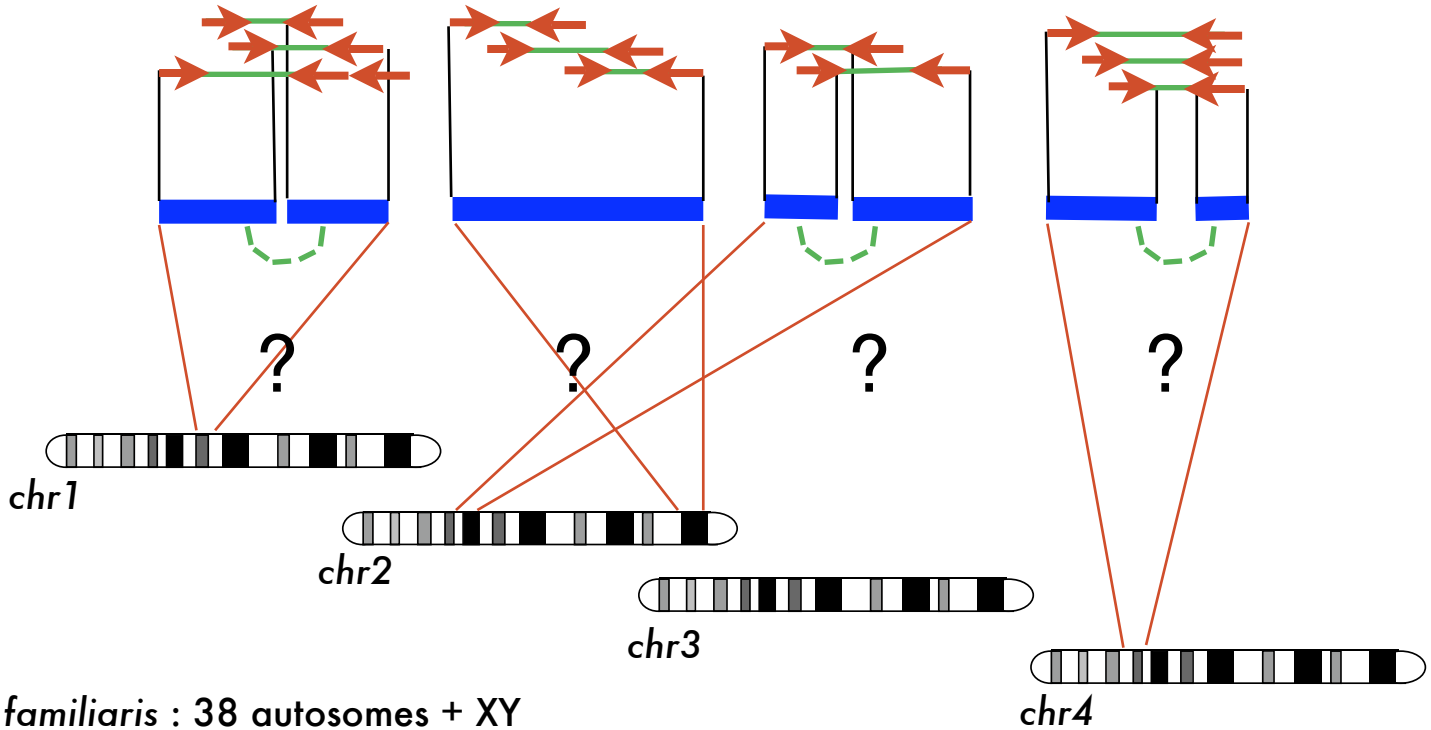
26 SEPTEMBER 2003 VOL 301 SCIENCE www.sciencemag.org

The Dog Genome: Survey Sequencing and Comparative Analysis

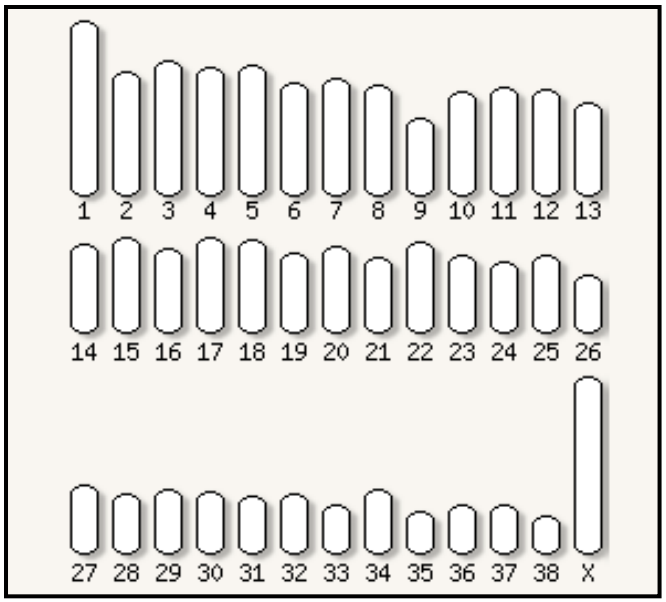
Ewen F. Kirkness,[1] Vineet Bafna,[2*] Aaron L. Halpern,[2*] Samuel Levy,[2*] Karin Remington,[2*] Douglas B. Rusch,[2*] Arthur L. Delcher,[1] Mihai Pop,[1] Wei Wang,[1] Claire M. Fraser,[1] J. Craig Venter[2]

A survey of the dog genome sequence (6.22 million sequence reads; 1.5× coverage) demonstrates the power of sample sequencing for comparative analysis of mammalian genomes and the generation of species-specific resources. More than 650 million base pairs (>25%) of dog sequence align uniquely to the human genome, including fragments of putative orthologs for 18,473 of 24,567 annotated human genes. Mutation rates, conserved synteny, repeat content, and phylogeny can be compared among human, mouse, and dog. A variety of polymorphic elements are identified that will be valuable for mapping the genetic basis of diseases and traits in the dog.





Canis familiaris : 38 autosomes + XY



Une carte de 10 000 gènes du génome canin

> Les objectifs :

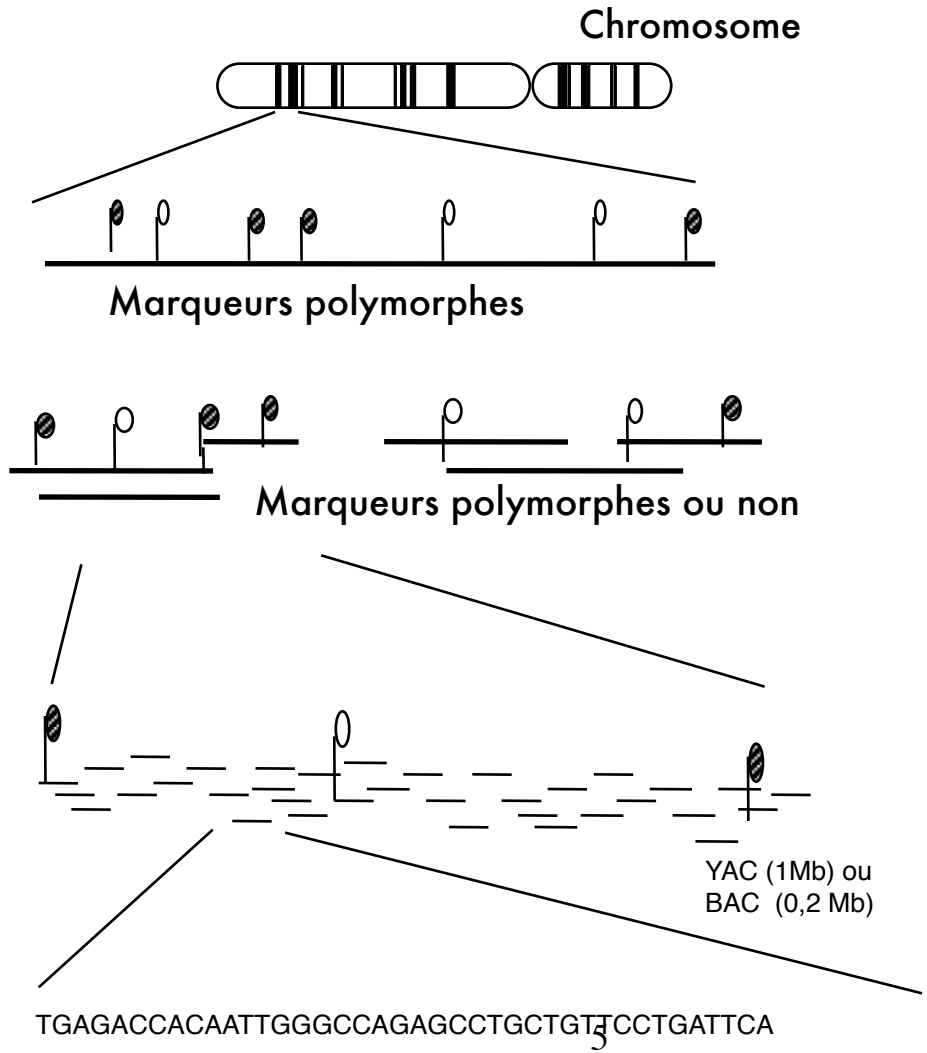
> cartographie dense et précise de 10 000 gènes canins
>> valorisation et optimisation des données de séquençage léger (1-2x)
>>> contribution à l'assemblage du séquençage profond (7.5x) du génome canin

Méthodes de Cartographie

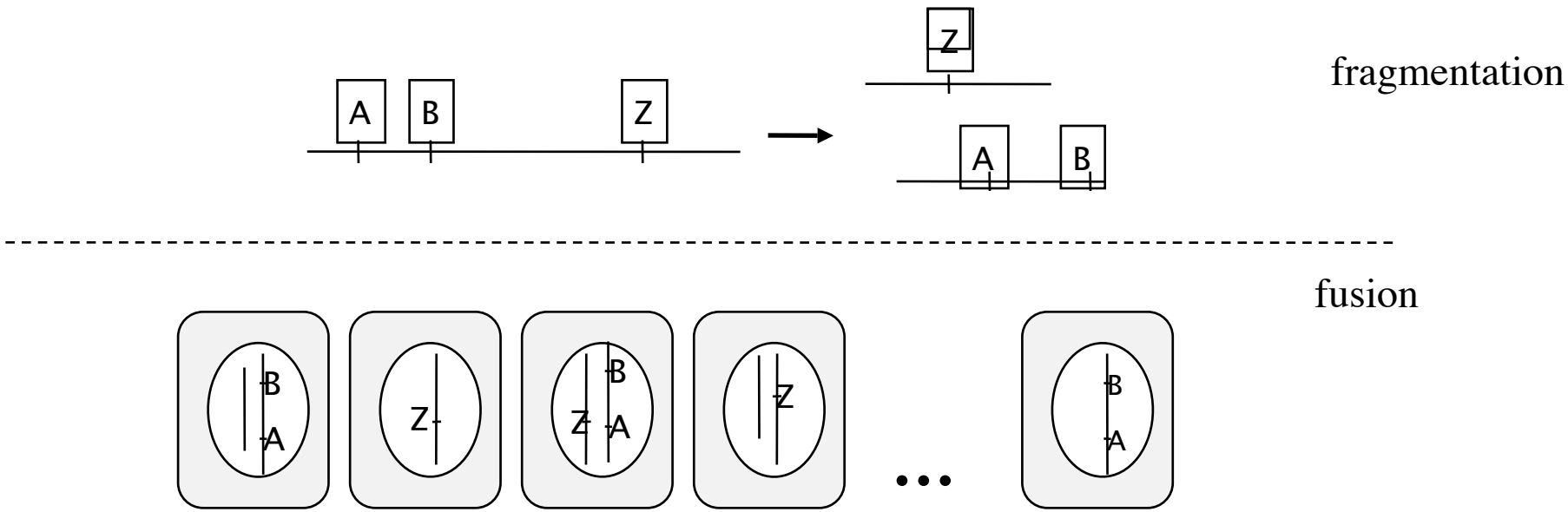
résolution de la cartographie :

Technique	Résolution
Caryotype	100 à 10 Mb
Cartographie génétique	1 Mb
Cartographie par hybrides d'irradiation (cartographie RH)	à volonté
Cartographie physique	50-1000 k
Séquençage	1 pb

Caractéristiques



Principe de l'analyse des données RH



A	**1**	**0**	**1**	**0**	**...**	**1**
B	**1**	**0**	**1**	**0**	**...**	**1**
Z	**0**	**1**	**1**	**1**	**...**	**0**

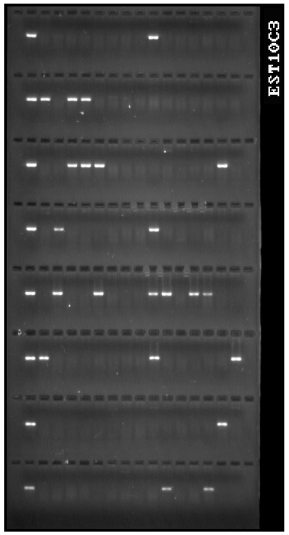
Cartographie sur hybrides irradiés : Format des données

fichier de vecteurs

```
A1A11T7   10000110000000100010000000001001110000100001011001000000010001010000000010
A1A12SP6  10000000010000000010010100000001011000000001000000000100000001000000000011
A1B12SP6  01000001010100000110000010000001000001000001000000010011000001011101010000
A1B2T7    12010100011001100111000100010001110000010101000000001101110001011000000011
A1B7SP6   00001110010000010100000100000001100000000001000000000001010000000000010011
A1C10T7   00000010000000000100000000000000100100010000000000000101010011000000010000
A1C11SP6  12000110210100000111000010100001000000000001001000000000001101001110000011
A1C12SP6  10000110010000000100011000010001111000000001101000010111000000100000011011
A1C3T7    00000010010000000111100000000001000000000000000000001000000010101000000011
A1C4SP6   10000010010000010000000110000000100010010000000000001011100000200010000011
A1C6T7    00000010000001010000000000000000000001000010000000000100100000000000000000
A2A11T7   10000110000000100010000000001001110000100001011001000000010001010000000010
A2A12SP6  10000000010000000010010100000001011000000001000000000100000001000000000011
A2B12SP6  01000001010100000110000010000001000001000001000000010011000001011101010000
A2B2T7    12010100011001100111000100010001110000010101000000001101110001011000000011
A2B7SP6   00001110010000010100000100000001100000000001000000000001010000000000010011
A2C10T7   00000010000000000100000000000000100100010000000000000101010011000000010000
A2C11SP6  12000110210100000111000010100001000000000001001000000000001101001110000011
A2C12SP6  10000110010000000100011000010001111000000001101000010111000000100000011011
A2C3T7    00000010010000000111100000000001000000000000000000001000000010101000000011
A2C4SP6   10000010010000010000000110000000100010010000000000001011100000200010000011
A2C6T7    00000010000001010000000000000000000001000010000000000100100000000000000000
```

Identification des marqueurs de gènes canin

> $6{,}22.10^6$ séquences du TIGR
  > 18 473 fragments de gènes canins identifiés
  > 13 000 sélectionnés avec un orthologue humain (1:1)
  (répartition uniforme sur le génome humain)
  > 10 500 marqueurs de gènes positionnés sur le génome canin



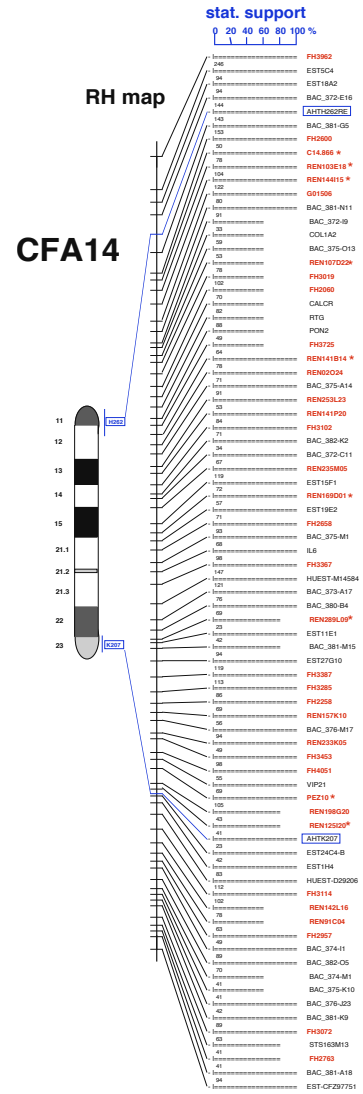
Objectifs du traitement statistique

>> 1: constitution de groupes de marqueurs liés entre-eux.

- densité en marqueurs
- capacité de liaison

>> 2: déterminer l'ordre linéaire des marqueurs.

- principe de l'algorithme
- puissance de calcul



Le traitement statistique des données

>> Pour N marqueurs –> N!/2 combinaisons d'ordre possibles.

2!/2 = 1

5!/2 = 60

10!/2=1 814 400

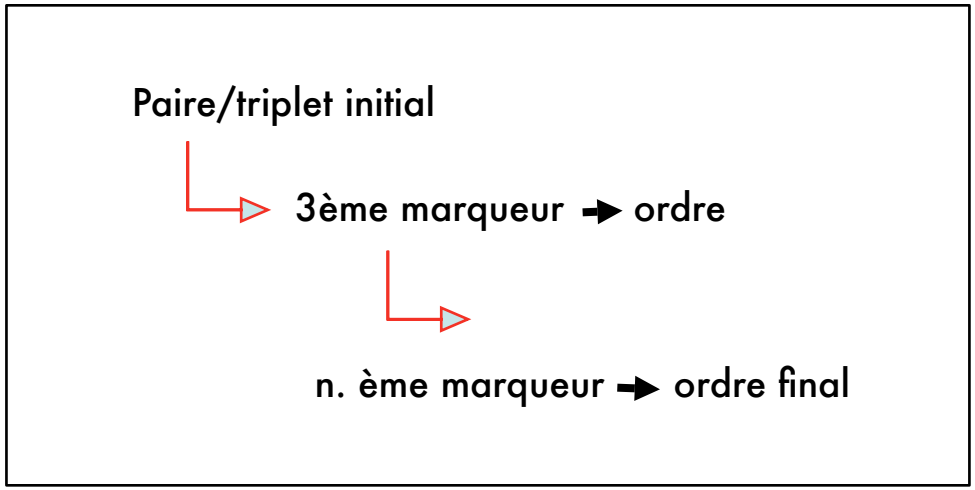
30!/2=132626429906095529318154240000000 (~$1{,}3 \times 10^{32}$)

Le traitement statistique des données

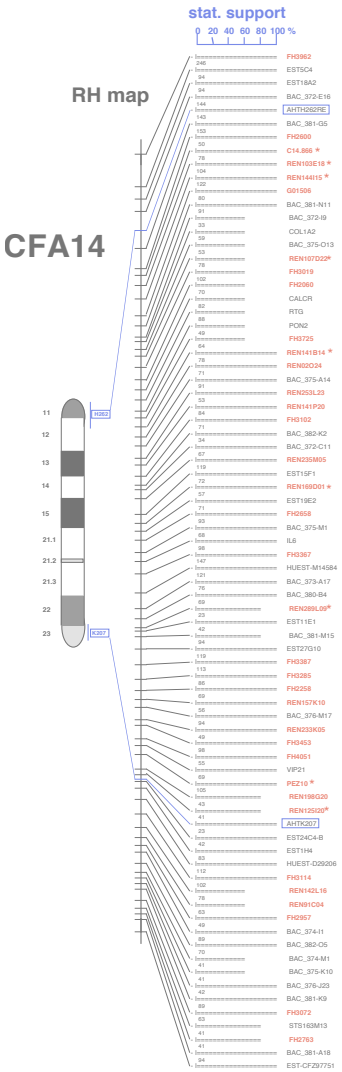
2: Ordonner les marqueurs

>> Les méthodes combinatoires non-paramétriques
principe de cassures obligatoires des chromosomes
(obligate chromosomes breaks -OCB-)
Déterminer l'ordre qui présente le moins de cassures chromosomiques

>> Les méthodes statistiques paramétriques
déterminer les permutations les plus vraisemblables entre paires de marqueurs.
Estimation du maximum de vraisemblance (Maximum Likelihood Estimation -MLE-)



Le traitement statistique des données : L'approche TSP



Construction de carte RH par approche TSP (traveling salesman problem approach)

>> Étant donné un nombre fini de villes, trouver le moyen le moins coûteux de visiter chaque ville une fois puis de revenir au point de départ.

> cartographie de 10 000 gènes du génome canin

Nombre de marqueurs : 10348

Nombre de marqueurs chomosome-spécifique* : 545

Nombre de positions uniques : 9081

Densité de marqueur : 1/275 kb

Couverture : 99 % [zones non représentées : 43 [taille 3-6 Mb] (dont 13 sont dépourvus de gènes)]

** marqueurs BAC préalablement positionnés par FISH et RH sur la carte intégrée FISH/RH (Breen M. et al, 2004)*

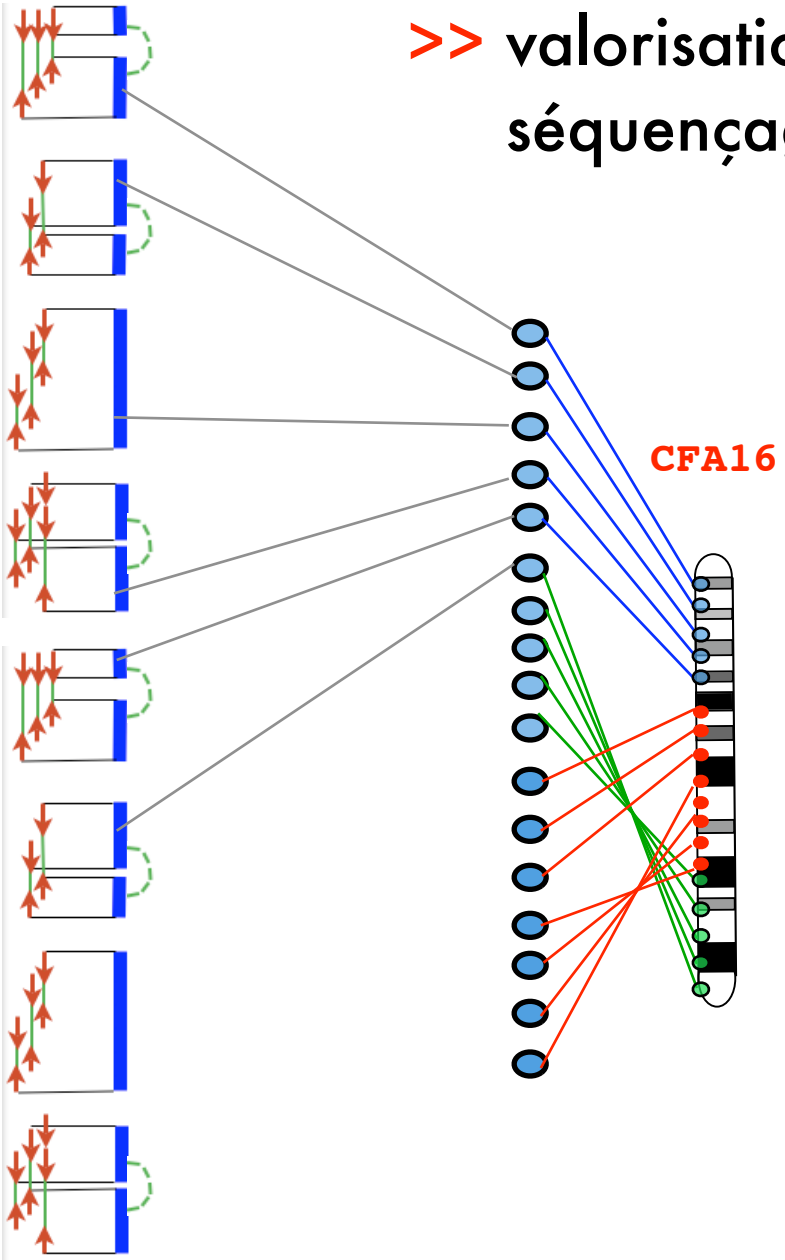
Nature Reviews Genetics **6**, 643-648 (2005); doi:10.1038/nrg1658

nature REVIEWS GENETICS

FACILITATING GENOME NAVIGATION: SURVEY SEQUENCING AND DENSE RADIATION-HYBRID GENE MAPPING

Christophe Hitte, Jennifer Madeoy, Ewen F. Kirkness, Catherine Priat, Travis D. Lorentzen, Fabrice Senger, Dan Thomas, Thomas Derrien, Christina Ramirez, Carol Scott, Gwenaelle Evanno, Barbara Pullar, Edouard Cadieu, Vinay Oza, Kristelle Lourgant, David B. Jaffe, Sandrine Tacher, Stéphane Dréano, Nadia Berkova, Catherine André, Panagiotis Deloukas, Claire Fraser, Kerstin Lindblad-Toh, Elaine A. Ostrander & Francis Galibert about the authors

>> valorisation et optimisation des données de séquençage léger (1-2x)

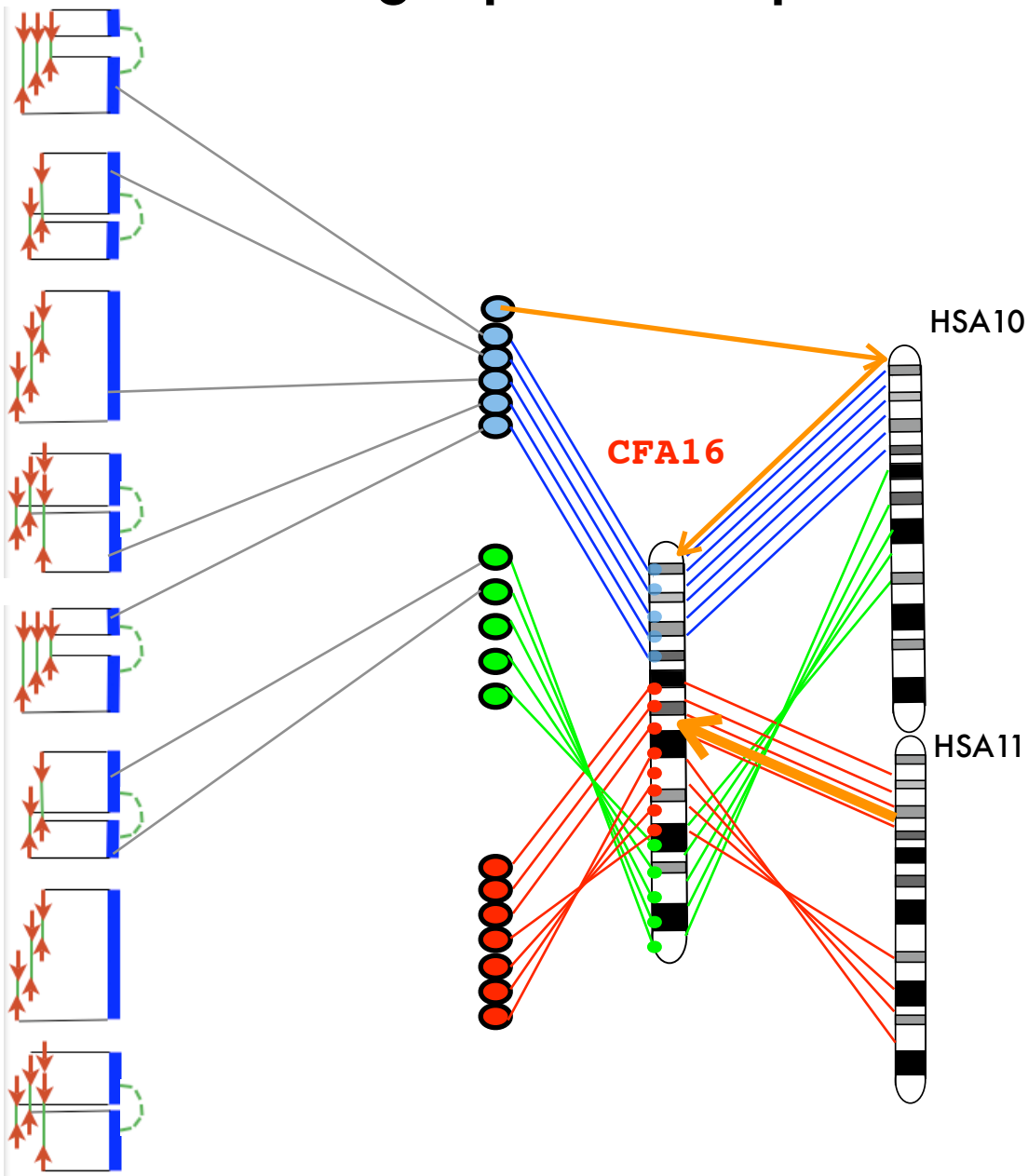


Approche combinée
Carte RH/séquençage léger

- >> 10 000 gènes
- >> 10 000 Contigs
- >> Les super-Contigs
  - Microsatellites
  - SNPs

valorisation des données de séquençage léger (1-2x)

>> Cartographie comparée Homme-chien

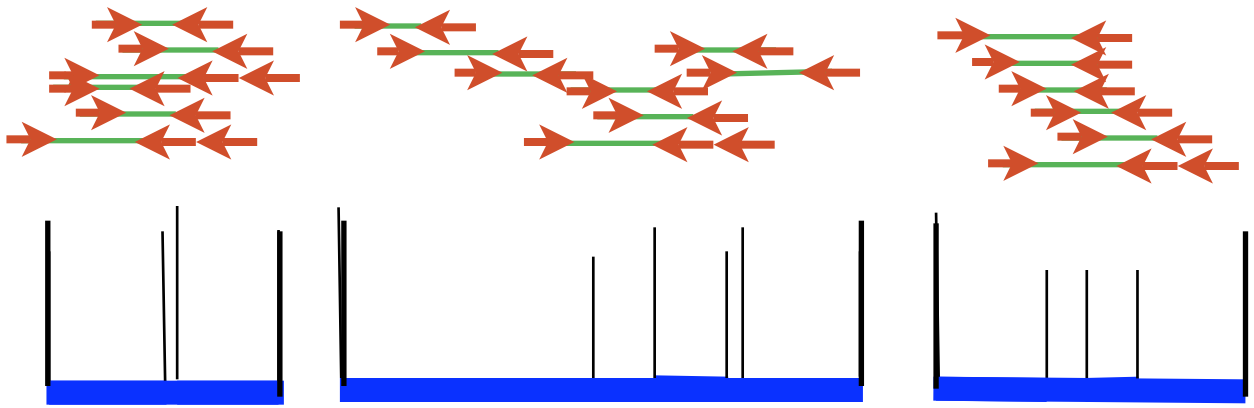


Approche combinée
Carte RH/séquençage léger

- >> 10 000 relations orthologie 1:1 (H/Ch)
- >> 260 segments conservés (> 500 kb)
- >> rupture de synténie

>>> Contribution à l'assemblage de la séquence (7.5x) du génome canin

1.5x
7.5x



>>> Contribution à l'assemblage de la séquence 7.5x du génome canin

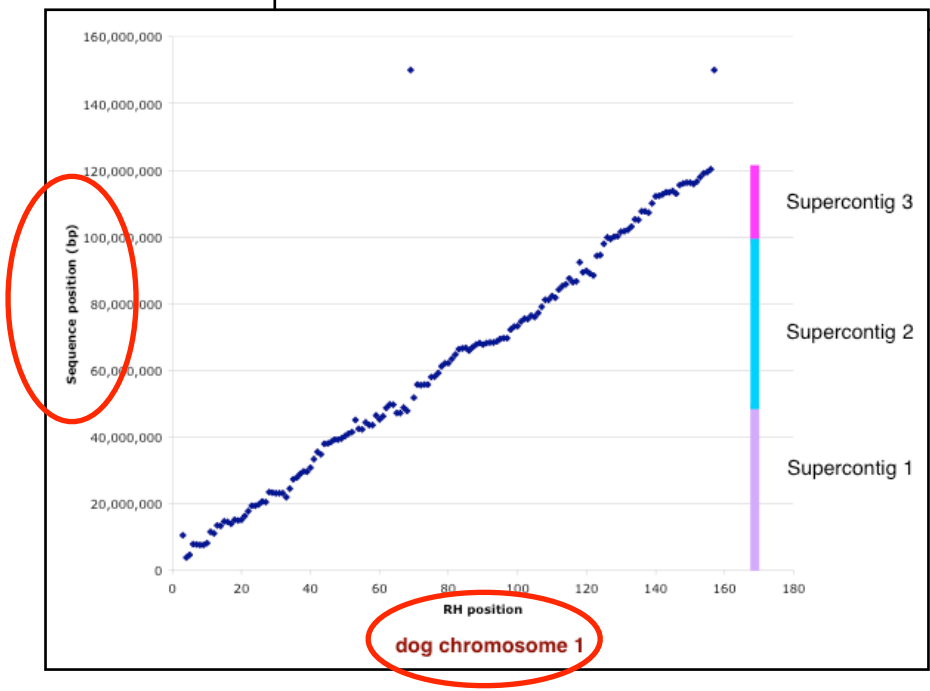
Dog Genome Sequencing Consortium

The dog genome project is a consortium led by:

- The Broad Institute - Plasmid and fosmid libraries and sequencing, assembly

The other members of the consortium, in alphabetical order are:

- Agencourt Bioscience - BAC-end sequencing
- Centre National de la Recherche Scientifique (CNRS) - RH map
- Children's Hospital Oakland Research Instituter (CHORI) - BAC libraries
- Fred Hutchinson Cancer Research Center (FHCRC) - Dog Selection
- North Carolina State University (NCSU) - FISH mapping

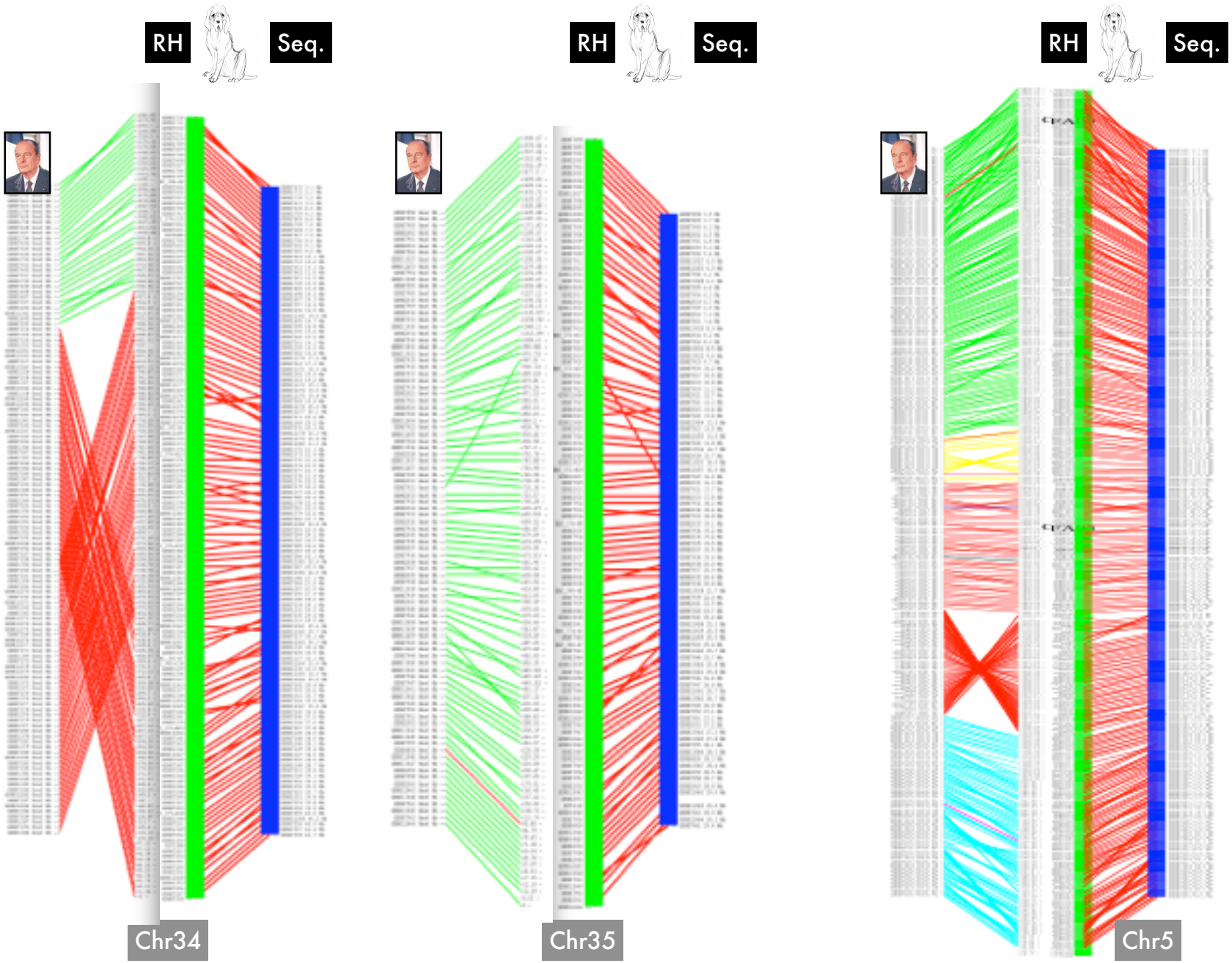


Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog

nature

Kerstin Lindblad-Toh[1*], Claire M Wade[1,2], Tarjei S Mikkelsen[1,3], Elinor K Karlsson[1,4], David B Jaffe[1], Michael Kamal[1], Michele Clamp[1], Jean L Chang[1], Edward J Kulbokas III[1], Michael C Zody[1], Evan Mauceli[1], Xiaohui Xie[1], Matthew Breen[5], Robert K Wayne[6], Elaine A Ostrander[7], Chris P Ponting[8], Francis Galibert[9], Douglas R Smith[10], Pieter J deJong[11], Ewen Kirkness[12], Pablo Alvarez[1], Tara Biagi[1], William Brockman[1], Jonathan Butler[1], Chee-Wye Chin[1], April Cook[1], James Cuff[1], Mark J Daly[1,2], David DeCaprio[1], Sante Gnerre[1], Manfred Grabherr[1], Manolis Kellis[1,13], Michael Kleber[1], Carolyne Bardeleben[6], Leo Goodstadt[8], Andreas Heger[8], Christophe Hitte[9], Lisa Kim[7], Klaus-Peter Koepfli[6], Heidi G Parker[7], John Pollinger[6], Stephen MJ Searle[14], Nathan B Sutter[7], Rachael Thomas[5], Caleb Webber[8], Broad Institute Genome Sequencing Platform[1], Eric S Lander[1,15*].

Validation Carte RH



Approche combinée Carte RH/séquençage léger

Séquençage profond 7-8x (98%)
- 36.10^6 seq
- annotation : 20 000 gènes
- identification : microsatellites
- > 2.10^6 SNP

>> coût : 40-50 x 10^6 $

Séquençage 2x (78%)
$6,22 .10^6$ seq
$1,16 .10^6$ contig
522,101 scaffold (3.8kb)
18,473 gènes

>> coût : 5 x 10^6 $

Cartographie RH
RH panel
10,000 gènes

>> coût : 1 x 10^6 $

Seq1.5x + RH mapping
- 90% fragments de syntenie >500kb
- Positions pour 10,000 gènes
- déduction pour ~25 % du génome
- identification de microsatellites

Mammalian Genome Project

Data release summary

Organism	Sequence complete	Sequence in progress	Assembly complete	Assembly in progress
Elephant	√		√	
Armadillo	√		√	
Rabbit	√		√	
Tenrec	√		July 2005	√
Guinea Pig	√		Sept. 2005	√
Cat	July 2005	√		
Common shrew	July 2005	√		
Hedgehog		Sept. 2005		

An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing

Elliott H. Margulies*†, Jade P. Vinson†‡, NISC Comparative Sequencing Program*§¶, Webb Miller‖, David B. Jaffe#, Kerstin Lindblad-Toh#, Jean L. Chang#, Eric D. Green*§, Eric S. Lander#, James C. Mullikin*§**, and Michele Clamp§*

*Genome Technology Branch and §NISC, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; ‡Broad Inst of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02141; and ‖Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802

Contributed by Eric S. Lander, December 30, 2004

alignement 2 espèces : identifier éléments conservés > 50 bp

alignement 16 espèces (2x): identifier éléments conservés > 8 bp

D = Distance évolutive (taux de substitution)
cumulée entre espèces

1: H/souris D = 0.450
2: H/souris/rat D = 0.530
16 : .... D = 3.8

Conclusions-perspectives

1: Utilisation des données de cartographie comparée dans l'histoire évolutive des génomes mammifères. [Murphy WJ. et al.]

2: Contribution de la synténie dans l'approche de construction des cartes génomique : Développement d'un algorithme de cartographie combinant l'apport de la synténie et de l'approche TSP

3: Développement d'un outil de construction de cartographie comparée : AutoGRAPH (Thomas Derrien)

Génétique & Développement
UMR 6061 • CNRS • UNIVERSITÉ RENNES 1

UNIVERSITE DE RENNES 1

CNRS

Francis Galibert
Christophe Hitte
Catherine André
Catherine Priat
Fabrice Senger
Edouard Cadieu
Thomas Derrien
Gwenaelle Évano
Kristelle Lourgant
Sandrine Tacher
Stéphane Dréano

TIGR THE INSTITUTE FOR GENOMIC RESEARCH

Claire M. Fraser &
Ewen F. Kirkness

NATIONAL INSTITUTES OF HEALTH

Elaine A. Ostrander

Travis Lorentzen
Heidi G. Parker
Vinay Oza
Barbara Pullar
Jennifer Madeoy
Christina Ramirez

NC STATE UNIVERSITY

Matthew Breen

Rachel Thomas
Allyson Scott

The Wellcome Trust Sanger Institute

The Sanger Institute is a genome research institute primarily funded by the Wellcome Trust. Our purpose is to further the knowledge of genomes, particularly through large scale sequencing and analysis.
Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK

Panos Deloukas
Dan Thomas
Carol Scott

Broad Institute

The Eli & Edythe L. Broad Institute is
Harvard University and affiliate

Kerstin Lindblad-Toh

NATIONAL INSTITUTES OF HEALTH

Alejandro Schaffer
Richa Agarwala

génopole OUEST

Hugues Leroy

Génétique & Développement
UMR 6061 • CNRS • UNIVERSITÉ RENNES 1

Nicolas Soriano