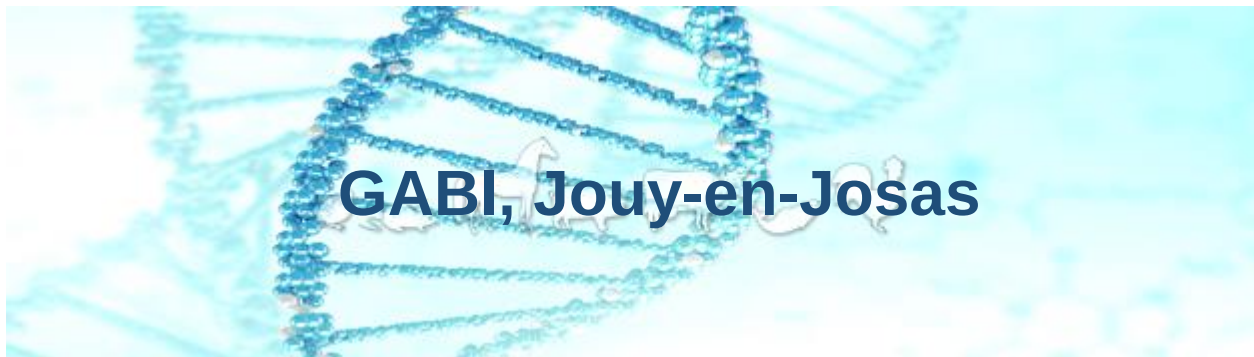


Le pangénome des bovins

Didier Boichard, Mekki Boussaha



Les choix en génomique bovine au cours des 30 dernières années

- Dans les années 90, un effort mondial pour améliorer une carte génétique, une carte physique et une carte d'irradiation uniques ; des panels de marqueurs partagés
- En 2006, le premier assemblage du génome bovin, progressivement amélioré, mais toujours du même animal : Dominette, de race Hereford, consanguine, et qui a fourni également une collection de tissus très complète.
- Assemblages successifs, d'utilisation quasi universelle, conduisant à la version actuelle ARS-UCD1.2.





Conséquences pratiques



- Tous les travaux s'effectuent par rapport à cette référence
- Exemple du séquençage d'un nouvel individu :
 - alignement sur cette référence
 - détermination des SNP, petits Indels, parfois les SV par rapport à cette référence
- Pb1 : les séquences nouvelles, absentes de Dominette, sont non considérées
- Pb2 : les cartes peuvent être inappropriées, d'où des erreurs d'imputation
- Pb3 : les annotations peuvent être erronées

Ca n'empêche pas de travailler...

- Effort interne : ~1400 séquences (12-15X), séquençage quasi systématique des taureaux d'IA récents (projet SeqOccln)
- Effort international dans le cadre du projet « 1000 bull genomes », (~5500 séquences partagées, env 120 M variants décrits et annotés)



- Imputation à grande échelle de la séquence en supposant la carte et donc l'assemblage unique

- Identification de variants causaux à grande échelle, en combinant imputation, GWAS, métaanalyses, annotation, et génotypage sur puce



Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals

Aniek C. Bouwman¹, Hans D. Daetwyler^{2,3}, Amanda J. Chamberlain², Carla Hurtado Ponce^{2,4}, Mehdi Sargolzaei^{5,6}, Flavio S. Schenkel⁵, Goutam Sahana⁷, Armelle Govignon-Gion⁸, Simon Boitard⁹, Marlies Dolezal¹⁰, Hubert Pausch^{2,11,12}, Rasmus F. Brøndum⁷, Phil J. Bowman², Bo Thomsen⁹, Bernt Gulbrandsen¹³, Mogens S. Lund⁷, Bertrand Servin¹³, Dorian J. Garrick¹⁴, James Reecy¹⁴, Johanna Vilkkii¹⁵, Alessandro Bagnato¹⁶, Min Wang^{2,3}, Jesse L. Hoff¹⁷, Robert D. Schnabel¹⁷, Jeremy F. Taylor¹⁷, Anna A. E. Vinkhuyzen^{18,19}, Frank Panitz⁹, Christian Bendixen⁹, Lars-Erik Holm⁹, Birgit Gredler²⁰, Chris Hozé^{8,21}, Mekki Boussaha⁸, Marie-Pierre Sanchez⁸, Dominique Rocha⁸, Aurelien Capitan^{8,21}, Thierry Tribout⁸, Anne Barbat⁸, Pascal Croiseau⁸, Cord Drögemüller²², Vidhya Jagannathan²², Christy Vander Jagt², John J. Crowley²³, Anna Biebert²⁴, Deirdre C. Purfield²⁵, Donagh P. Berry²⁵, Reiner Emmerling²⁶, Kay-Uwe Götz²⁶, Mirjam Frischknecht²⁰, Ingolf Russ²⁷, Johann Sölkner²⁸, Curtis P. Van Tassell²⁹, Ruedi Fries¹¹, Paul Stothard³⁰, Roel F. Veerkamp¹, Didier Boichard⁸, Mike E. Goddard^{2,4} and Ben J. Hayes^{2,31*}

Des premiers travaux sur les variants structuraux

A partir des lectures courtes Illumina

Analyses complexes combinant

- la profondeur de lecture
- L'information des lectures mated pairs
- l'analyse des split reads.

Outils d'efficacité modérée, avec à la fois des faux positifs et des faux négatifs
=> sélection des variants mis en évidence par plusieurs approches

 **PLOS** | ONE 2015

RESEARCH ARTICLE

Bilan :
Carte de SV fiable
mais très incomplète

Genome-Wide Study of Structural Variants in
Bovine Holstein, Montbéliarde and
Normande Dairy Breeds

Mekki Boussaha^{1,2*}, Diane Esquerré^{3,4,5}, Johanna Barbieri^{3,4,5}, Anis Djari⁶,
Alain Pinton^{3,4,5}, Rabia Letaief^{1,2}, Gérald Salin^{3,4,5}, Frédéric Escudie^{3,4,5}, Alain Roulet^{3,4,5},
Sébastien Fritz^{1,2,7}, Franck Samson⁸, Cécile Grohs^{1,2}, Maria Bernard⁶, Christophe Klopp⁶,
Didier Boichard^{1,2}, Dominique Rocha^{1,2}

Approches complémentaires

Gros SV détectables par génotypage puce

- intensité de fluorescence des signaux de puce
- anomalies de transmission mendélienne,
- DL fort entre marqueurs supposés indépendants

Qq exemples :

- Absence de cornes : 2 allèles, une insertion de 80kb ou une délétion de 200 bp
- Anomalie brachyspina : délétion d'une partie du gène FANCI
- Anomalie CDH : insertion d'un élément transposable dans le gène APOB
- Tête blanche : duplication dans le gène KIT
- PMS syndrome : délétion de ZEB2
- CNV dans GC : résistance aux mammites
- Etc...

Mais l'analyse des SV restait difficile jusqu'à l'avènement du séquençage lectures longues

Identification de délétions

Exemple du syndrome PMS

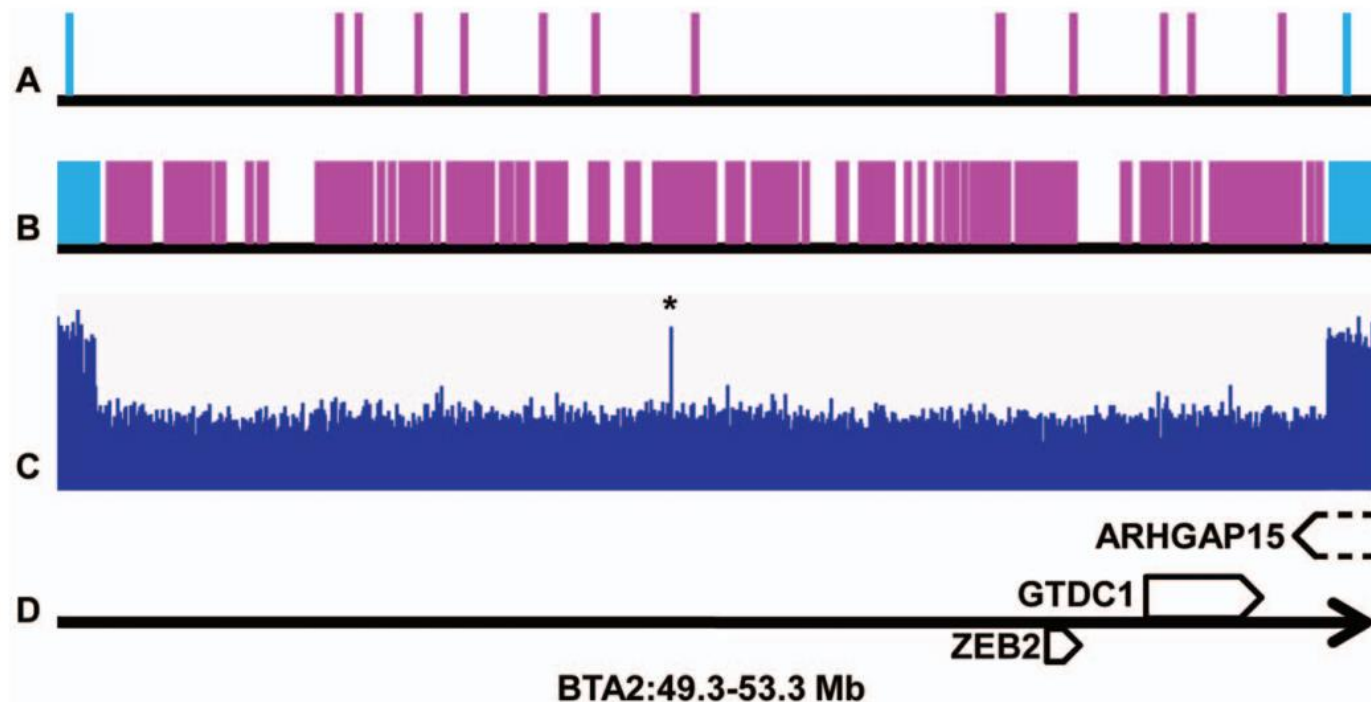
- Par profondeur de lecture
- Par anomalies mendéliennes

OPEN ACCESS Freely available online

PLOS ONE

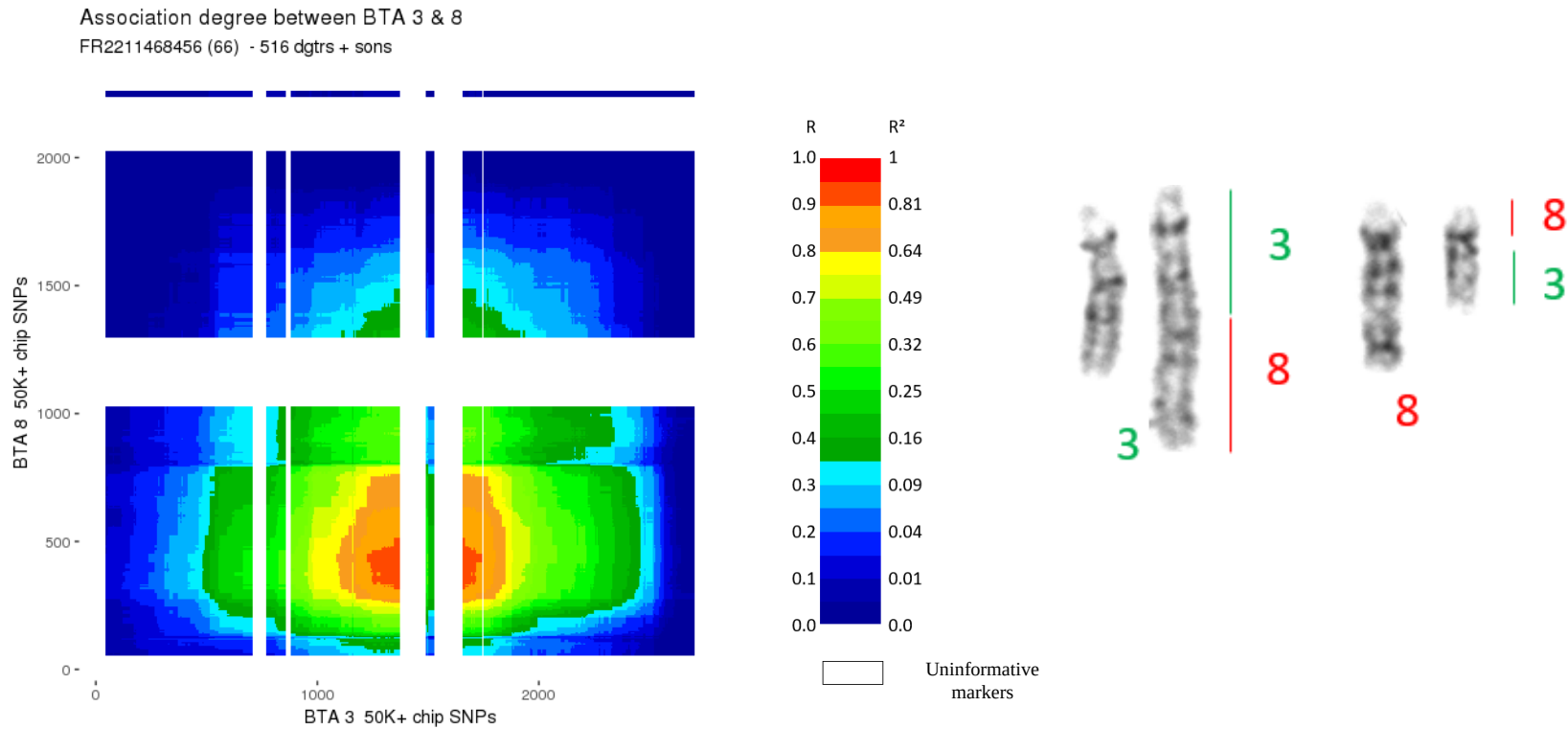
A 3.7 Mb Deletion Encompassing ZEB2 Causes a Novel Polled and Multisystemic Syndrome in the Progeny of a Somatic Mosaic Bull

Aurélien Capitan^{1,2*}, Aurélie Allais-Bonnet³, Alain Pinton⁴, Brigitte Marquant-Le Guienne⁵, Daniel Le Bourhis^{3,5}, Cécile Grohs¹, Stéphan Bouet¹, Laëtitia Clément⁵, Laura Salas-Cortes⁵, Eric Venot¹, Stéphane Chaffaux¹, Bernard Weiss¹, Arnaud Delpeuch⁶, Guy Noé⁷, Marie-Noëlle Rossignol⁷, Sarah Barbey⁸, Dominique Dozias⁸, Emilie Cobo⁸, Harmonie Barasc⁴, Aurélie Auguste³, Maëlle Pannetier³, Marie-Christine Deloche⁵, Emeline Lhuillier^{9,10}, Olivier Bouchez^{9,10}, Diane Esquerré^{9,10}, Gérald Salin^{9,10}, Christophe Klopp¹¹, Cécile Donnadiou^{9,10}, Céline Chantry-Darmon⁷, Hélène Hayes¹, Yves Gallard⁸, Claire Ponsart⁵, Didier Boichard¹, Eric Pailhoux³



Détection des translocations à partir des génotypes

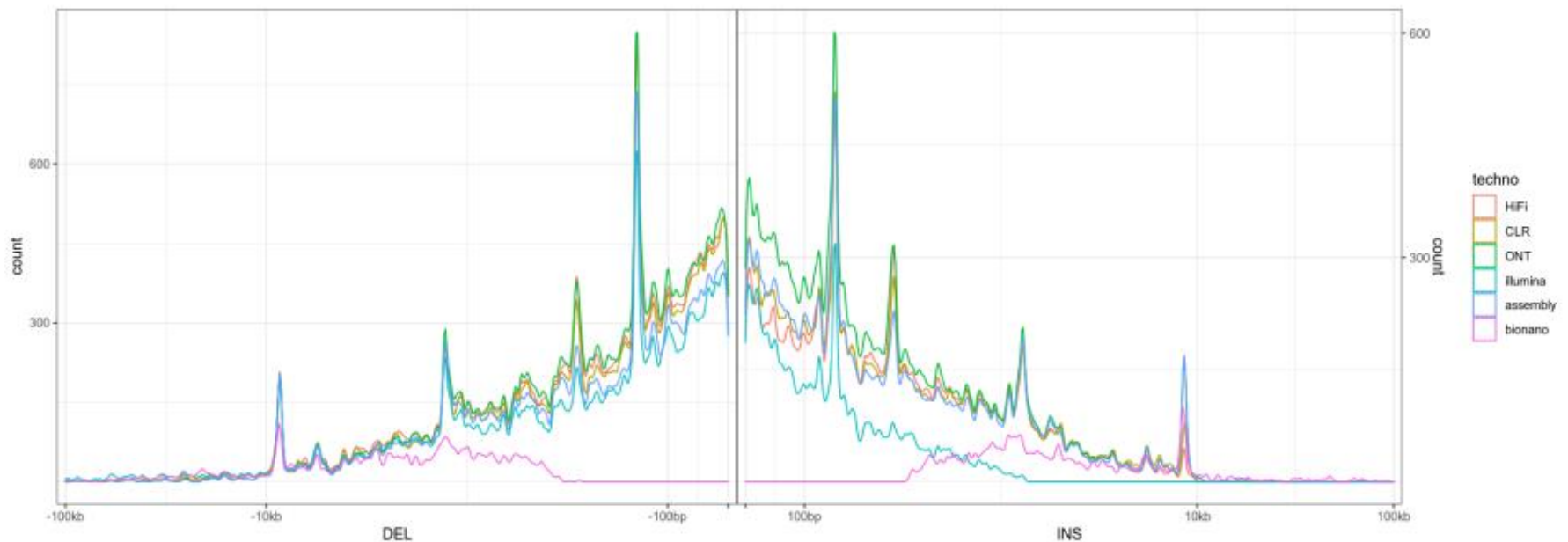
Recherche de déséquilibre de liaison entre chromosomes intra parent



(Jourdain et al, en préparation)

Détection de SV avec des lectures longues

Les lectures longues permettent une détection plus efficace et moins biaisée des SV



(Faraut, comm pers)

Les lectures longues et l'assemblage de novo

Les lectures longues permettent des assemblages *de novo* au prix d'un effort raisonnable

Deux assemblages peuvent être produits par individu hétérozygote

=> Notion de trio binning

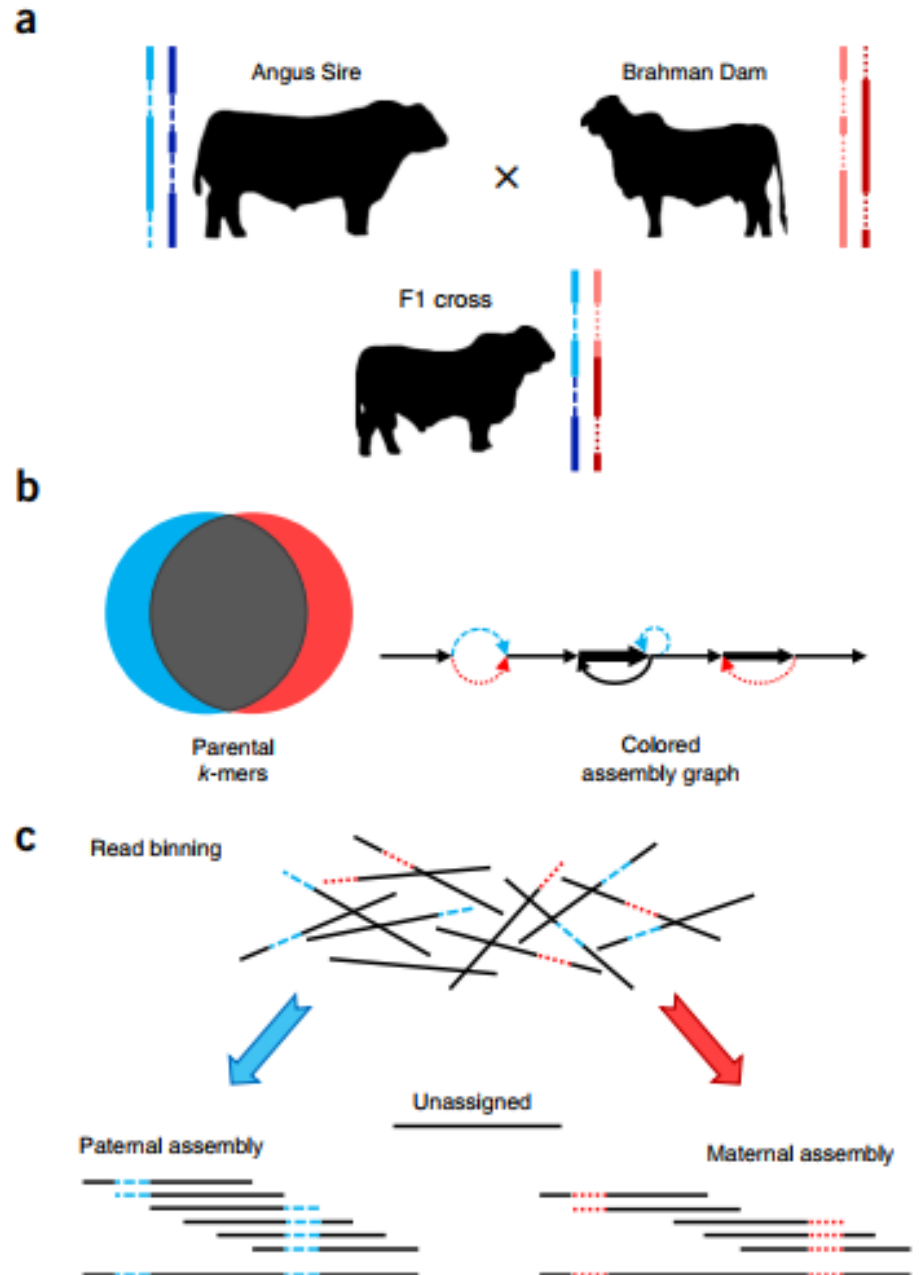
nature
biotechnology

De novo assembly of haplotype-resolved genomes with trio binning

Sergey Koren^{1,8}, Arang Rhie^{1,8} , Brian P Walenz¹, Alexander T Dilthey^{1,2} , Derek M Bickhart³ ,
Sarah B Kingan⁴ , Stefan Hiendleder^{5,6} , John L Williams⁵ , Timothy P L Smith⁷ & Adam M Phillippy¹

Le trio binning

Fonctionne d'autant mieux que les deux parents sont plus différents



Les premières études bovines

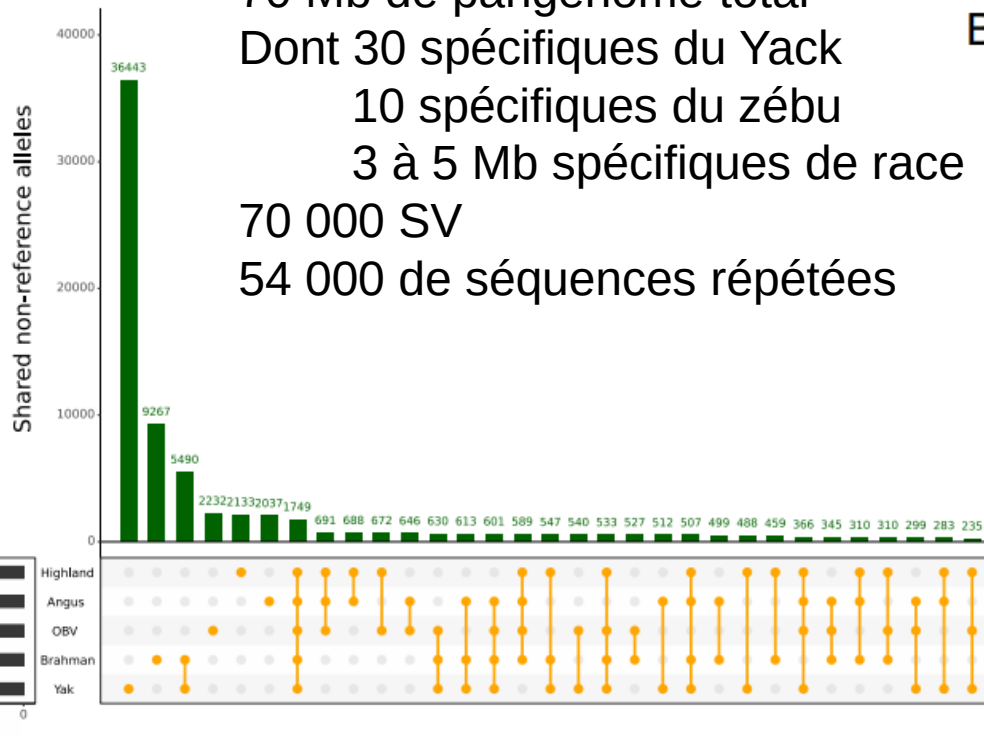
Novel functional sequences uncovered through a bovine multiassembly graph

PNAS

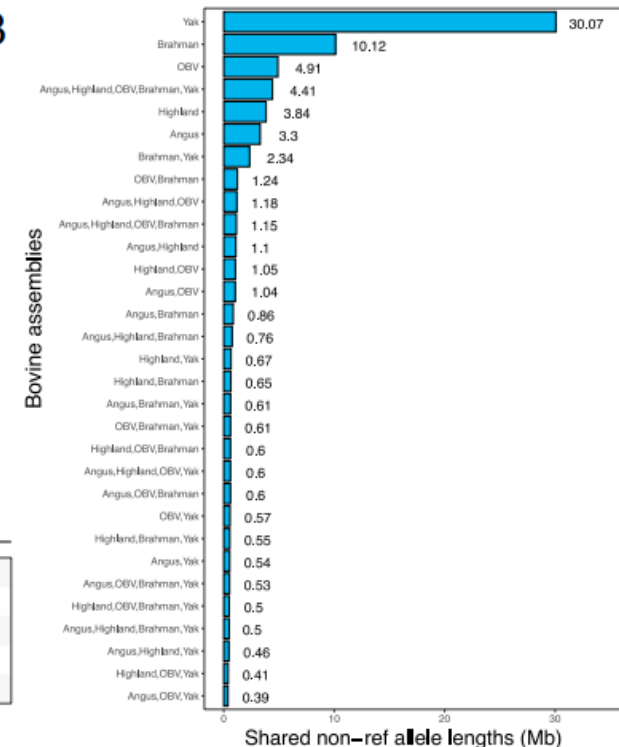
Danang Crysnanto^{a,1} , Alexander S. Leonard^a , Zih-Hua Fang^a , and Hubert Pausch^a 

^aAnimal Genomics, Eidgenössische Technische Hochschule (ETH) Zürich, 8315 Zürich, Switzerland

A



B



Les premières études bovines

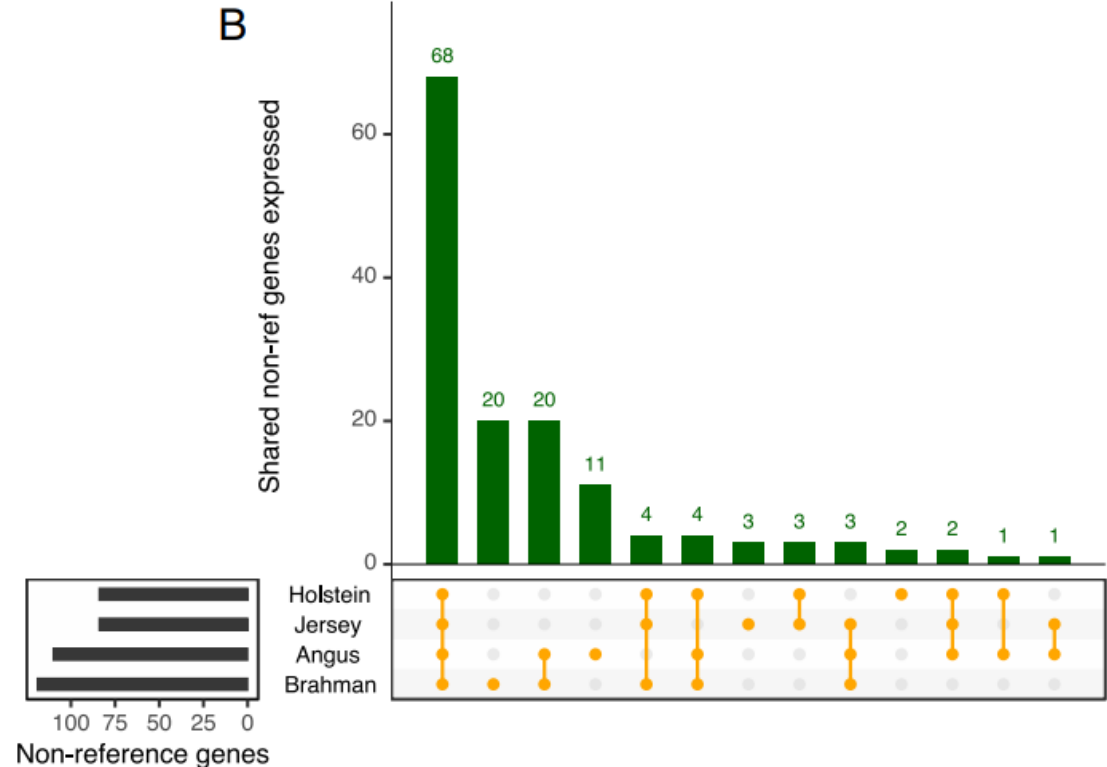
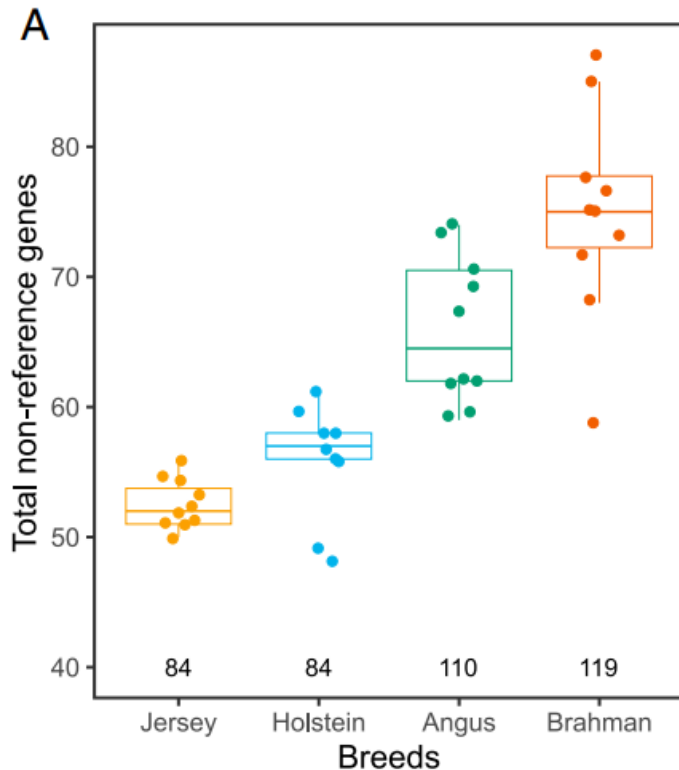
Analyse de 40 transcriptomes de foie

Gain de mapping de 0.06% (zébu) à 0.01% (référence)

263 gènes exprimés, absents de la référence Hereford (121 ribosomaux ou olfaction)

Dont plusieurs gènes de l'immunité

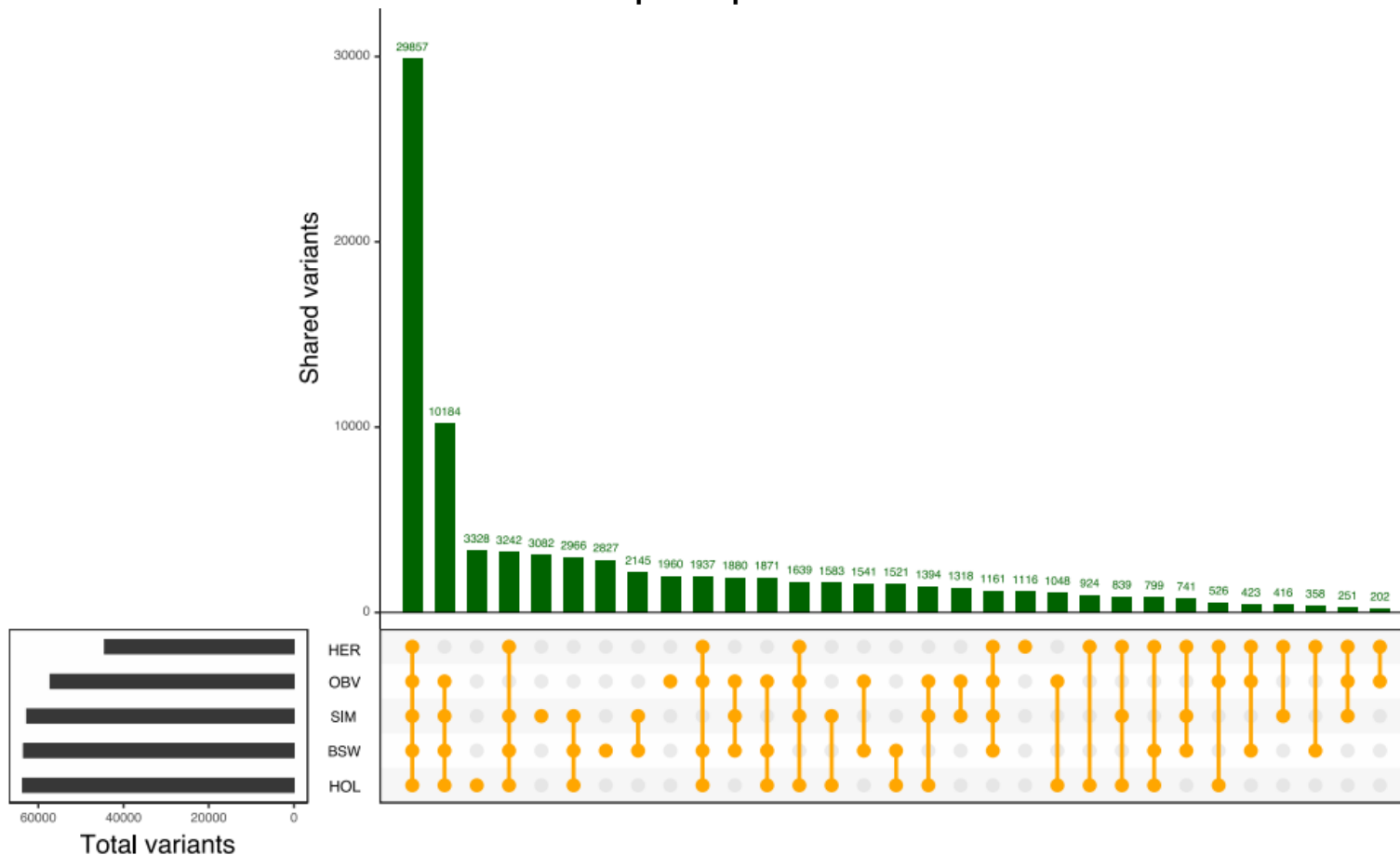
(NB : une petite partie du pangénome reflète une référence incomplète)



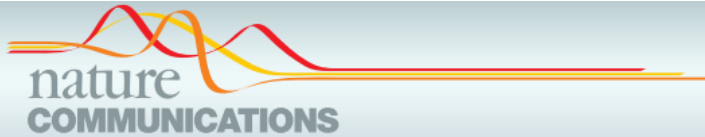
Nouveaux variants absents de la référence

Mapping de lectures courtes issues de séquençage génomique (45 animaux, 5 races)

0.01 (Hereford) à 0.024% (BS) de lectures alignées en plus
83 250 nouveaux variants, surtout intergéniques et introniques
54 variants classés avec un fort impact prédit



Pangénome africain



ARTICLE



<https://doi.org/10.1038/s41467-022-28605-0>

OPEN

A cattle graph genome incorporating global breed diversity

A. Talenti¹ , J. Powell¹ , J. D. Hemmink^{1,2,3,4} , E. A. J. Cook^{2,4} , D. Wragg^{1,3}, S. Jayaraman¹ , E. Paxton¹, C. Ezeasor⁵, E. T. Obishakin^{6,7} , E. R. Agusi^{6,7}, A. Tijjani^{8,9} , K. Marshall^{2,4}, A. Fisch¹⁰ , B. R. Ferreira¹⁰ , A. Qasim¹¹ , U. Chaudhry¹, P. Wiener¹, P. Toye^{2,4}, L. J. Morrison^{1,3}, T. Connelley^{1,3} & J. G. D. Prendergast^{1,3} 

N'Dama + Ankole + 294 animaux de races diverses

116 Mb absents de la référence au total

7.4 Mb pour l'Ankole (taurin x zébu)

4.6 Mb pour le N'Dama (taurin)

Différences les plus importantes

- Chromosomes sexuels (artefact)
- MHC



SeqOccln



Projet de séquençage de grande envergure

L'axe (A) Génome inclut un gros volet bovin

- 6 trios pour des assemblages haute qualité (PacBio HiFi)
- 150 séquences long reads (PacBio CLR), avec assemblages individuels
- 571 séquences short reads (dont les 150)



Projet cofinancé par le Fonds Européen de Développement Régional

Analyse des 6 trios

- Deux trios Charolais ont permis de tester de nombreuses méthodes et matériels
 - Tissus, méthodes d'extraction, Hi-C
 - Méthodes de séquençage : Pacbio Hifi et CLR, ONT, Bionano, 10X, Illumina...
 - Méthodes d'assemblage
- Un trio Holstein x Normand
- Un trio Montbéliard x Yack
- Un trio Blonde d'Aquitaine
- Un trio Aubrac



Soit 12 assemblages haplotypiques de 7 races, en cours d'analyse

- Trois autres prévus d'ici un an : Abondance, Tarentaise, Vosgienne

Analyse de la diversité des races

Race	Animaux
Holstein	25
Montbéliarde	25
Normande	25
Brune	5
Simmental	5
Abondance	10
Tarentaise	5
Vosgienne	4
Blonde d'A	10
Charolaise	10
Limousine	10
Aubrac	10
Flamande	3
Parthenaise	3
Total	150

PacBio CLR 30X

Aujourd'hui,
120 séquencés,
101 assemblés

Première analyse du pangénome en races Charolaise, Tarentaise et Vosgienne

Graphes multi-assemblage construits avec MINIGRAPH

Races			CHA	TAR	VOS
Pan-genome	Nb assemblages		6	5	4
	Longueur du graphe		2 524 730 836	2 551 978 622	2 554 024 324
	Longueur de la référence		2 489 385 779		
	Longueur du core genome		2 317 557 075	2 363 900 584	2 330 940 534
	Longueur de la partie flexible du graphe		35 345 057	62 592 843	64 638 545
SVs	Délétions Bi-alléliques		5 475	4 019	3 851
	Insertions Bi-alléliques		6 840	5 144	5 140
	Insertion Multi-alléliques		787	121	217
	Substitution de Sequence	AltInsertion Bi-al	6 823	10 445	10 241
		AltDeletion Bi-al	6 045	10 875	9 964
		AltDeletion Multi-al	10 390	4 630	6 223
	Total SVs		36 360	35 234	35 636

Caractéristiques des séquences non-REF (insertions)

Races	CHA	TAR	VOS
Longueur NRS (bp)	7 841 655	3 812 582	3 834 157
GC%	40,26	40.79	40.63
LINEs (%)	30,91	26,12	28.11
Séq répétées	33,75	28,65	30.76

Breeds	CHA	TAR	VOS
BlastX hits	1 864	1 190	1 194
ab initio gene model prediction	244	90	105

Projet international “Pangenome”

- Une initiative américaine (USDA), avec un financement pour le séquençage et la bioinformatique
Nous participons pour les races françaises
- Un assemblage pour chaque race au niveau mondial (1000-2000 races ?)
- Stratégie : Séquençage long read, analyse de trios.
- Une ressource dans le domaine public : graphe + outils d'analyse
- Un « white paper » en préparation
- En complément, des discussions préliminaires pour un nouveau consortium « 1000 bull genomes » en long reads ?

Remerciements

SeqOccln



Genotoul
GeT



Genotoul
Bioinfo



Projet cofinancé par le Fonds Européen de Développement Régional



APIS-GENE
Investir Innover Valoriser