

Les pan-génomés et questions transverses aux domaines du vivant

Johann Joets

UMR Génétique Quantitative et Evolution
Institut Diversité Evolution & Ecologie du Vivant
INRAE - Université Paris-Saclay - CNRS - AgroParisTech

Jusqu'au début du siècle, il est admis que tous les individus d'une espèce possèdent le même répertoire de gène

Intraspecific violation of genetic colinearity and its implications in maize

Huihua Fu* and Hugo K. Dooner*^{†‡}

*The Waksman Institute, Rutgers University, Piscataway, NJ 08855; and [†]Department of Plant Science, Rutgers University, New Brunswick, NJ 08901

Communicated by Michael Freeling, University of California, Berkeley, CA, May 1, 2002 (received for review March 5, 2002)

Although allelic sequences can vary extensively, it is generally assumed that each gene in one individual will have an allelic counterpart in another individual of the same species. We report here that this assumption does not hold true in maize. We have sequenced over 100 kb from the *bz* genomic region of two different maize lines and have found dramatic differences between them. First, the retrotransposon clusters, which comprise most of the repetitive DNA in maize, differ markedly in make-up and location relative to the genes in the *bz* region. Second, and more importantly, the genes themselves differ between the two lines, demonstrating that genetic microcolinearity can be violated within the same species. Our finding has bearing on the underlying genetic basis of hybrid vigor in maize, and possibly other organisms, and on the measurement of genetic distances.

Comparative genetic mapping has revealed a remarkable degree of synteny or conservation of gene order among closely related plant species (1). The extensive conservation of gene content and order among grass chromosomes has even led

regions of the maize genome (available at http://www.nsf.gov/bio/dbi/dbi_pgr.htm). We find that the *bz* regions of these two maize lines differ not only in the make-up and pattern of interspersed retrotransposon clusters, but also, surprisingly, in the density and content of their genes.

Materials and Methods

BAC Isolation. The isolation of the two adjacent *NotI* *Bz-McC* BAC clones, which comprise a 230-kb contig of the *bz* region, has been described (16). Those BAC clones were isolated from a maize line carrying the *Bz-McC* allele that had been introgressed into the genetic background of the inbred W22 by repeated backcrossing. The *Bz-McC* allele, named after B. McClintock, is the progenitor of the *bz-m2(Ac)* unstable mutation described by her (17) and most likely derives from a New England flint variety. The *Bz-B73* BAC clones were isolated from a commercially available BAC library of the Corn Belt inbred B73 (Genome Systems, St. Louis), by using a probe from the *stc1* gene, which is immediately distal to *bz* in the *Bz-McC* BAC

Jusqu'au début du siècle, il est admis
que tous les individus d'une espèce
possèdent le même répertoire de gène.

Mais ce n'est pas vrai chez le maïs !

Intraspecific violation of genetic colinearity and its implications in maize

Huihua Fu* and Hugo K. Dooner**

*The Waksman Institute, Rutgers University, Piscataway, NJ 08855; and [†]Department of Plant Science, Rutgers University, New Brunswick, NJ 08901

Communicated by Michael Freeling, University of California, Berkeley, CA, May 1, 2002 (received for review March 5, 2002)

Although allelic sequences can vary extensively, it is generally assumed that each gene in one individual will have an allelic counterpart in another individual of the same species. We report here that this assumption does not hold true in maize. We have sequenced over 100 kb from the *bz* genomic region of two different maize lines and have found dramatic differences between them. First, the retrotransposon clusters, which comprise most of the repetitive DNA in maize, differ markedly in make-up and location relative to the genes in the *bz* region. Second, and more importantly, the genes themselves differ between the two lines, demonstrating that genetic microcolinearity can be violated within the same species. Our finding has bearing on the underlying genetic basis of hybrid vigor in maize, and possibly other organisms, and on the measurement of genetic distances.

regions of the maize genome (available at http://www.nsf.gov/bio/dbi/dbi_pgr.htm). We find that the *bz* regions of these two maize lines differ not only in the make-up and pattern of interspersion of their retrotransposon clusters, but also, surprisingly, in the density and content of their genes.

Materials and Methods

BAC Isolation. The isolation of the two adjacent *NotI* *Bz-McC* BAC clones, which comprise a 230-kb contig of the *bz* region, has been described (16). Those BAC clones were isolated from a maize line carrying the *Bz-McC* allele that had been introgressed into the genetic background of the inbred W22 by repeated backcrossing. The *Bz-McC* allele, named after B. McClintock, is the progenitor of the *bz-m2(Ac)* unstable mutation described by

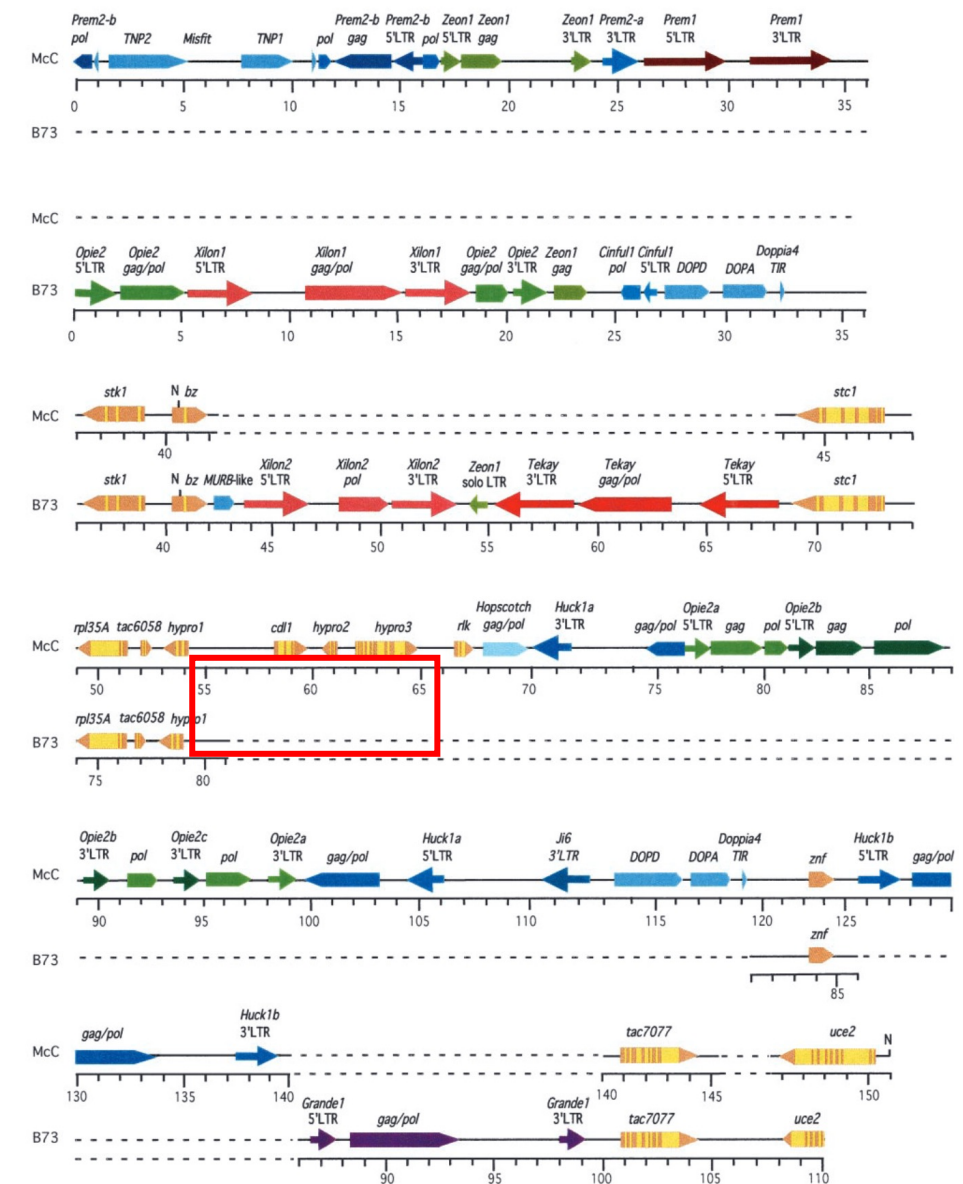


Fig. 1. Organization of genes and transposons in the *bz* genomic region of two different maize lines, McC and B73. The proximal end is at 0 kb; the distal end is at 151 kb in McC and at 110 kb in B73. Genes are shown as pentagons pointing in the direction of transcription; exons are in bronze, and introns are in yellow. Each transposon is in a different color. To facilitate the identification of interrupted retrotransposons, LTRs and encoded proteins of the same element are in the same color. Where a sequence in one line cannot be aligned with a sequence at the corresponding position in the other line, a dotted line is used to show the lack of alignment. The paired sequences of McC and B73 can be aligned only at the genes that they have in common: *stk1*, *bz*, *stc1*, *rpl35A*, *tac6058*, *hypro1*, *znf*, *tac7077*, and *uce2*. The McC chromosome contains three retrotransposon blocks, proximal to *stk1*, *znf*, and *tac7077*, respectively. The B73 chromosome

Des cas de gènes présents
uniquement chez certains individus
étaient connus depuis les années 90.

Notamment parmi les gènes de
résistance aux maladies.

Structure of the *Arabidopsis* *RPM1* Gene Enabling Dual Specificity Disease Resistance

Murray R. Grant,* Laurence Godiard,*† Esther Straube,
Tom Ashfield, Jürgen Lewald, Airlie Sattler, Roger W. Innes,
Jeffery L. Dangl‡

Plants can recognize pathogens through the action of disease resistance (*R*) genes, which confer resistance to pathogens expressing unique corresponding avirulence (*avr*) genes. The molecular basis of this gene-for-gene specificity is unknown. The *Arabidopsis thaliana* *RPM1* gene enables dual specificity to pathogens expressing either of two unrelated *Pseudomonas syringae* *avr* genes. Despite this function, *RPM1* encodes a protein sharing molecular features with recently described single-specificity *R* genes. Surprisingly, *RPM1* is lacking from naturally occurring, disease-susceptible *Arabidopsis* accessions.

Plants express sophisticated genetic systems to recognize pathogens. Complex *R*-gene loci have been defined in a variety of plant-pathogen interactions. Genetic analyses suggest that each *R*-encoded specificity responds to a single *avr*-dependent pathogen signal (1). The recent cloning of four *R* genes (2–7) and accumulation of cloned

bacterial and fungal *avr* genes (8) provide limited insight into the structural basis of specificity, although most models predict that *R* gene products interact with specific pathogen signal molecules produced in an *avr*-dependent manner (9).

RPM1 was identified in *A. thaliana* accession Col-0 as conferring resistance to *P. syringae* isolates expressing the *avrRpm1* gene (10, 11). Functional homologs of *RPM1* exist in pea, bean, and soybean (12). Resistance in *A. thaliana* to the *P. syringae* *avrB* gene also mapped to the *RPM1* interval (initially termed *RPS3*) (13), and genetic analyses of *A. thaliana* mutants have suggested that *RPM1* conferred resistance to *P. syringae* expressing either *avrRpm1* or *avrB* (14). Because the sequences of *avrB* and *avrRpm1* are unrelated (12, 15), *RPM1* appears to determine a dual specificity.

RPM1 was mapped by restriction fragment length polymorphism (RFLP) analysis

M. R. Grant, L. Godiard, E. Straube, J. Lewald, Max-Delbrück Laboratory, Carl von Linné Weg 10, 50829 Köln, Germany.

T. Ashfield, A. Sattler, R. W. Innes, Department of Biology, Indiana University, Bloomington, IN 47405, USA.

J. L. Dangl, Max-Delbrück Laboratory, Carl von Linné Weg 10, 50829 Köln, Germany, and Department of Biology, University of North Carolina, Chapel Hill, NC 27599–3280, USA.

*These authors contributed equally to this work.

†Present address: Laboratoire de Biologie Moléculaire INRA-CNRS, B.P. 27, 31326 Castanet Tolosan Cedex, France.

‡To whom correspondence should be addressed at Department of Biology, Coker Hall 108, CB 3280, University of North Carolina, Chapel Hill, NC 27599–3280, USA.

SCIENCE • VOL. 269 • 11 AUGUST 1995

843

Evolutionary dynamics of full genome content in *Escherichia coli*

Howard Ochman¹ and Isaac B. Jones

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA

¹Corresponding author
e-mail: hochman@email.arizona.edu

The evolutionary history of the entire *Escherichia coli* chromosome was traced by examining the distribution of the ~4300 open reading frames (ORFs) from *E. coli* MG1655 among strains of known genealogical relationships. Using this framework to deduce the incidence of gene transfer and gene loss, a total of 67 events—37 additions and 30 deletions—were required to account for the distribution of all genes now present in the MG1655 chromosome. Nearly 90% of the ORFs were common to all strains examined, but, given the variation in gene content and chromosome size, strains can contain well over a megabase of unique DNA, conferring traits that distinguish them from other members of the species. Moreover, strains vary widely in their frequencies of deletions, which probably accounts for the variation in genome size within the species.

Keywords: bacterial chromosomes/deletions/*Escherichia coli*/genome evolution/lateral gene transfer

Introduction

The analysis of total gene content, as recovered from complete genomic sequences, has begun to elucidate the manner in which bacterial genomes change over an evolutionary timescale. Initial comparisons of sequenced genomes have demonstrated that relatively few genes are shared by all bacterial species (Mushegian and Koonin, 1996; Watanabe *et al.*, 1997; de Rosa and Labedan, 1998; Huynen and Bork, 1998; Snel *et al.*, 1999), indicating that bacterial genomes are very dynamic and subject to repeated events of gene acquisition and loss (Doolittle, 1999; Jain *et al.*, 1999). However, the set of organisms sequenced to date cannot fully reveal the rate and pattern of genetic events that shape their genomes. In most comparisons, the organisms are too distantly related and have sustained so many changes that their evolutionary histories have been erased. Even in cases where complete sequence information is available for two closely related or conspecific strains, e.g. *Helicobacter pylori* (Alm and Trust, 1999) or *Chlamydia* spp. (Read *et al.*, 2000), it is usually not possible to establish: (i) the rate and order of events that led to the present day chromosome compositions; (ii) whether the differences in gene content are due to the acquisition of horizontally transferred genes or to the loss of ancestral sequences; or even (iii) whether

shared genes are ancestral to the species as a whole or were later acquired by a more recent common ancestor.

An alternative approach that does not rely upon information from other sequenced organisms has been developed to establish the ancestry and ages of genes within a bacterial genome (Lawrence and Ochman, 1997). This method assesses the history of a gene by examining its degree of departure from sequence characteristics prevalent in the genome as a whole and, hence, is independent of both the number and the phylogenetic distribution of other completely sequenced organisms. Application of these procedures has revealed that bacteria vary widely in the amount of horizontally acquired sequences currently present in their genomes (Ochman *et al.*, 2000). For example, in *Escherichia coli* strain MG1655 (Blattner *et al.*, 1997), it was estimated that nearly 18% of its 4290 open reading frames (ORFs) were introduced in at least 234 transfer events and that the majority of the acquired genes appeared in this lineage relatively recently (Lawrence and Ochman, 1998). Coupled with the fact that natural isolates of *E. coli* may differ by nearly a megabase in genome size (Richmond *et al.*, 1999; Tao *et al.*, 1999), these findings suggest that gene content is highly variable within this species, with divergent strains having gained and deleted very different arrays of genes.

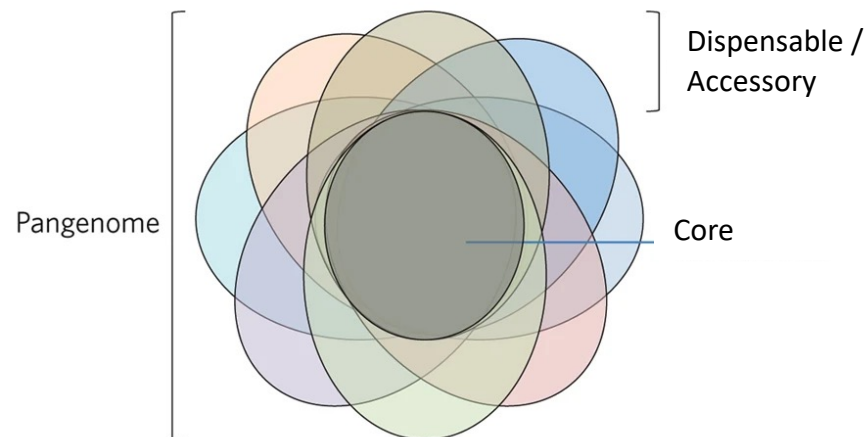
To investigate the dynamics of genome content over an evolutionarily relevant timescale, we applied a whole genome approach to trace the distribution of every gene from the sequenced *E. coli* MG1655 chromosome among natural strains of varying degrees of relatedness, known ancestries and representing the range of chromosome sizes in the species at large. By comparing lineages of varying degrees of relatedness, it is possible to reconstruct individual events of gene transfer and loss occurring over the entire *E. coli* genome, to estimate the amount of unique DNA within each strain and to determine the relative ages of every gene on the MG1655 chromosome. Ages of genes inferred from their phylogenetic distributions and derived from sequence characteristics (Lawrence and Ochman, 1997, 1998)—and hence the evolutionarily effective rates of gene transfer and loss calculated by both approaches—were in many cases found to be similar. Repeated events of gene acquisition and the concomitant loss of sequences have created a situation in which divergent lineages of *E. coli* possess, along with the basal set of genes, a unique complement of genes (and their encoded traits) that distinguishes them from other members of the species. Tracing the distribution of gene transfer events showed that certain strains are prone to deletions, which might account for the wide variation in genome size within the species.

Au regard de leur contenu en gène, les premiers génomes complets bactériens ont offerts quelques surprises.

evolutionary timescale. Initial comparisons of sequenced genomes have demonstrated that relatively few genes are shared by all bacterial species (Mushegian and Koonin, 1996; Watanabe *et al.*, 1997; de Rosa and Labedan, 1998; Huynen and Bork, 1998; Snel *et al.*, 1999), indicating that bacterial genomes are very dynamic and subject to repeated events of gene acquisition and loss (Doolittle, 1999; Jain *et al.*, 1999). However, the set of organisms

in the MG1655 chromosome. Nearly 90% of the ORFs were common to all strains examined, but, given the variation in gene content and chromosome size, strains can contain well over a megabase of unique DNA, conferring traits that distinguish them from other members of the species. Moreover, strains vary

Le concept de pan-génome est forgé en 2005 par Tettelin et Fraser



Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”

Hervé Tettelin^{a,b}, Vega Masignani^{b,c}, Michael J. Cieslewicz^{b,d,e}, Claudio Donati^c, Duccio Medini^c, Naomi L. Ward^{a,f}, Samuel V. Angiuoli^a, Jonathan Crabtree^a, Amanda L. Jones^g, A. Scott Durkin^a, Robert T. DeBoy^a, Tanja M. Davidsen^a, Marirosa Mora^c, Maria Scarselli^c, Immaculada Margarit y Ros^c, Jeremy D. Peterson^a, Christopher R. Hauser^a, Jaideep P. Sundaram^a, William C. Nelson^a, Ramana Madupu^a, Lauren M. Brinkac^a, Robert J. Dodson^a, Mary J. Rosovitz^a, Steven A. Sullivan^a, Sean C. Daugherty^a, Daniel H. Haft^a, Jeremy Selengut^a, Michelle L. Gwinn^a, Liwei Zhou^a, Nikhat Zafar^a, Hoda Khouri^a, Diana Radune^a, George Dimitrov^a, Kisha Watkins^a, Kevin J. B. O'Connor^h, Shannon Smithⁱ, Teresa R. Utterback^j, Owen White^a, Craig E. Rubens^g, Guido Grandi^c, Lawrence C. Madoff^{e,j}, Dennis L. Kasper^{e,j}, John L. Telford^c, Michael R. Wessels^{d,e}, Rino Rappuoli^{c,k,l}, and Claire M. Fraser^{a,b,k,m}

^aInstitute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; ^bChiron Vaccines, Via Fiorentina 1, 53100 Siena, Italy; ^cDivision of Infectious Diseases, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115; ^dHarvard Medical School, Boston, MA 02115; ^eCenter of Marine Biotechnology, University of Maryland Biotechnology Institute, 701 East Pratt Street, Baltimore, MD 21202; ^fChildren's Hospital and Regional Medical Center, 307 Westlake Avenue N, Seattle, WA 98109; ^gThe Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218; ^hJ. Craig Venter Institute, 5 Research Place, Rockville, MD 20850; ⁱChanning Laboratory, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115; and ^jGeorge Washington University Medical Center, 2300 Eye Street NW, Washington, DC 20037

Contributed by Rino Rappuoli, August 5, 2005

The development of efficient and inexpensive genome sequencing methods has revolutionized the study of human bacterial pathogens and improved vaccine design. Unfortunately, the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species and also limits genome-wide screens for vaccine candidates or for antimicrobial targets. We have generated the genomic sequence of six strains representing the five major disease-causing serotypes of *Streptococcus agalactiae*, the main cause of neonatal infection in humans. Analysis of these genomes and those available in databases showed that the *S. agalactiae* species can be described by a pan-genome consisting of a core genome shared by all isolates, accounting for ~80% of any single genome, plus a dispensable genome consisting of partially shared and strain-specific genes. Mathematical extrapolation of the data suggests that the gene reservoir available for inclusion in the *S. agalactiae* pan-genome is vast and that unique genes will continue to be identified even after sequencing hundreds of genomes.

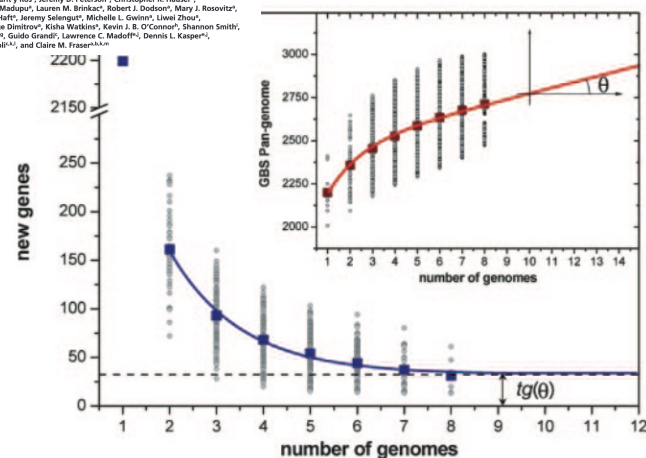
genome sequence of the type Ia strain A909 and draft genome sequences (8× sequence coverage) of five additional strains, representing the five major serotypes. Comparative analysis of the six newly sequenced genomes and the two genomes already available in the databases suggests that a bacterial species can be described by its “pan-genome” (pan, from the Greek word *παν*, meaning whole), which includes a core genome containing genes present in all strains and a dispensable genome composed of genes absent from one or more strains and genes that are unique to each strain. Surprisingly, unique genes were still detected after eight genomes were sequenced, and mathematical extrapolation predicts that new genes will still be found after sequencing many more strains. Thus, the genomes of multiple, independent isolates are required to understand the global complexity of bacterial species. Analysis of multiple GBS genomes was found to be instrumental for the development of vaccines (8) and for the functional characterization of important genetic determinants (9).

Materials and Methods

Quelle est la taille du pan-génome ? Vers l'infini et au-delà ?

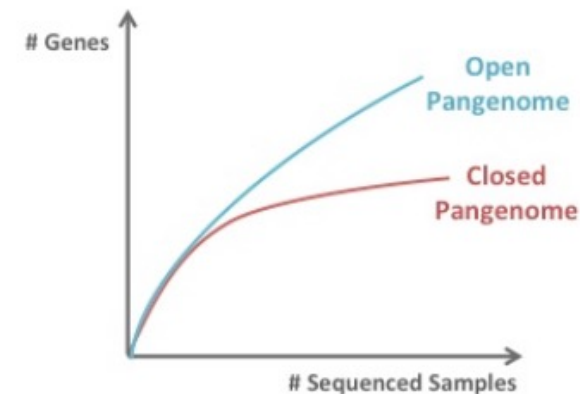
Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin^{1,2}, Vega Masignan^{1,2}, Michael J. Cislewicz^{1,2,3,4}, Claudio Donati¹, Duccio Medini¹, Naomi L. Ward¹, Samuel V. Angioli¹, Jonathan Crabtree¹, Amanda L. Jones¹, A. Scott Dunkin¹, Robert T. DeBoy¹, Tanja M. Davidson¹, Marirosa Mora¹, Maria Scarselli¹, Immaculada Margarit y Ros¹, Jeremy D. Peterson¹, Christopher R. Hauser¹, Jaldeep P. Sundaram¹, William C. Nelson¹, Ramana Madupu¹, Lauren M. Brinkac¹, Robert J. Dodson¹, Mary J. Roovitor¹, Steven A. Sullivan¹, Sean C. Daugherty¹, Daniel H. Haft¹, Jeremy Selengut¹, Michelle L. Gwinn¹, Liwei Zhou¹, Nikhat Zafar¹, Hoda Khouri¹, Diana Radu¹, George Dimitrov¹, Kisha Watkins¹, Kevin J. B. O'Connor¹, Shannon Smith¹, Teresa E. Ullrich¹, Owen White¹, Craig E. Rubens¹, Guido Grandi¹, Lawrence C. Madoff¹, Dennis L. Kasper¹, John L. Telford¹, Michael R. Wesels^{1,2}, Rino Rappuoli^{1,2,3,4}, and Claire M. Fraser^{1,2,3,4}

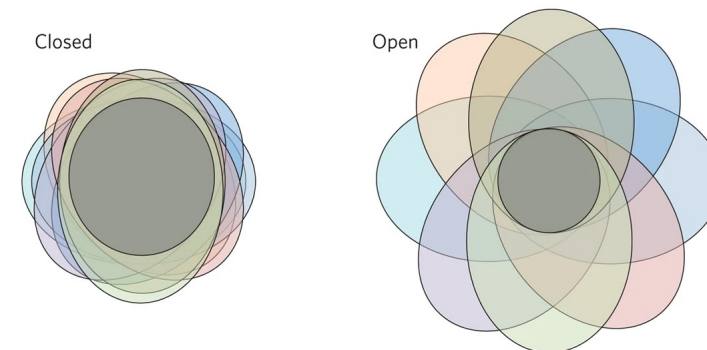


levels of genomic diversity (data not shown). As in the case of GBS, each additional GAS genome added an average of 27 new genes to the pool, leading to an open pan-genome. In the case of *B. anthracis*, the number of specific genes added to the pan-genome dropped to zero after the addition of a fourth strain. This result probably reflects the fact that *B. anthracis* is a highly clonal, recently evolved species in which genome variability is associated only with virulence plasmids (23, 24). Alternatively, the sequenced strains may belong to the same evolutionary clade and may not adequately represent the *B. anthracis* species.

Les pangénomes, mise à jour et exploitation en agriculture.



Tranchant-Dubreuil, C., Rouard, M. & Sabot, F.
Annual Plant Reviews online. 1–25 (wiley, 2019).



McInerney, J., McNally, A. & O'Connell, M. Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 17040 (2017).

Des pan-génomes ouverts ou fermés ? Une question ouverte ?



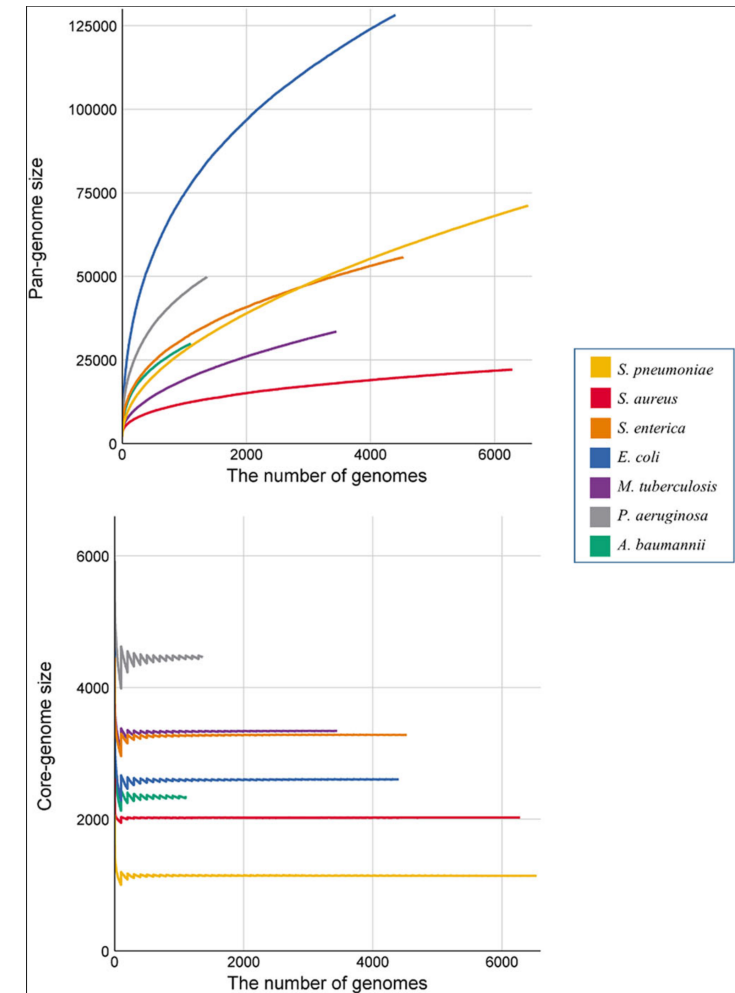
ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fmicb.2019.00634



Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size

Sang-Cheol Park¹, Kihyun Lee², Yeong Ouk Kim³, Sungho Won^{1,3,4,*} and Jongsik Chun^{3,5,*}

S. aureus was predicted to have a closed pan-genome according to a power law model, whereas other studies reported that its pan-genome repertoire generated from 64 strains is likely to increase indefinitely as more strains are analyzed (Tettelin et al., 2008; Bosi et al., 2016). Such inconsistencies might be attributable to an insufficient number of genomes used to explain general species properties and/or sampling bias in genome sequencing, such as those associated with clinically significant pathogens.



Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size

Sang-Cheol Park,¹ Kihyun Lee,² Yeong Ouk Kim,³ Sungho Won,^{1,3,4,*} and Jongsik Chun^{3,5,*}

Attention à l'échantillonnage de la diversité !

Le ratio de taille entre core et dispensable génome est très variable d'un règne et d'une espèce à l'autre

	Species	Ploidy	~Genome size	Number of individuals	Core genome (% genes/ gene clusters)	Core gene/ gene cluster detection method	Core genome (% sequence)	Additional information ^a	Refs
Bacteria	<i>Chlamydia trachomatis</i>	Haploid	1.04 Mb	85	80	HGC	N/A	Allopatric	[10]
	<i>Rickettsia prowazekii</i>	Haploid	1.1 Mb	10	8	HGC	N/A	Allopatric	[10]
	<i>Mycobacterium tuberculosis</i>	Haploid	4.4 Mb	168	78	HGC	N/A	Allopatric	[10]
	<i>Yersinia pestis</i>	Haploid	4.7 Mb	36	56	HGC	N/A	Allopatric	[10]
	<i>Bacillus anthracis</i>	Haploid	5.2 Mb	50	51	HGC	N/A	Allopatric	[10]
	<i>Streptococcus agalactiae</i>	Haploid	2.2 Mb	57	28	HGC	N/A	Sympatric	[10]
	<i>Streptococcus pneumoniae</i>	Haploid	2.2 Mb	52	31	HGC	N/A	Sympatric	[10]
	<i>Haemophilus influenzae</i>	Haploid	1.8 Mb	55	33	HGC	N/A	Sympatric	[10]
	<i>Escherichia coli</i>	Haploid	4.6 Mb	633	8	HGC	N/A	Sympatric	[10]
	<i>Clostridium botulinum</i>	Haploid	3.9 Mb	46	5	HGC	N/A	Sympatric	[10]
	<i>Prochlorococcus marinus</i>	Haploid	1.8 Mb	10	18	HGC	N/A	Sympatric	[10]

Le ratio de taille entre core et dispensable génome est très variable d'un règne et d'une espèce à l'autre

	Species	Ploidy	~Genome size	Number of individuals	Core genome (% genes/ gene clusters)	Core gene/ gene cluster detection method	Core genome (% sequence)	Additional information ^a	Refs
Fungi	<i>Saccharomyces cerevisiae</i> (baker's yeast)	Mostly diploid	12 Mb	1011	63	ORF sequence similarity	N/A	Single cell organism Domesticated and wild	[27]
	<i>Candida albicans</i>	Diploid	15 Mb	34	91	Synteny	N/A	Single cell organism Animal commensal	[26]
	<i>Cryptococcus neoformans</i>	Haploid	19 Mb	25	81	Synteny	N/A	Single cell soil organism	[26]
	<i>Aspergillus fumigatus</i>	Haploid	29 Mb	12	83	Synteny	N/A	Environmental filamentous fungus	[26]
	<i>Parastagonospora</i> spp.	Haploid	40 Mb	33	~40	HGC	N/A	Plant pathogen	[29]
	<i>Zymoseptoria tritici</i>	Haploid	40 Mb	5	58	HGC	N/A	Plant pathogen	[28]

Le ratio de taille entre core et dispensable génome est très variable d'un règne et d'une espèce à l'autre

	Species	Ploidy	~Genome size	Number of individuals	Core genome (% genes/ gene clusters)	Core gene/ gene cluster detection method	Core genome (% sequence)	Additional information ^a	Refs
Plants	<i>Brachypodium distachyon</i> (stiff brome)	Diploid	355 Mb	54	35	HGC	N/A	~5% Outcrossing Not domesticated	[33]
	<i>Brassica oleracea</i>	Diploid	650 Mb	10	81	Read mapping	N/A	>95% Outcrossing Crop plant	[34]
	<i>Brassica napus</i> (oil seed rape)	Tetraploid	1.1 Gb	53	62	Read mapping	N/A	~30% Outcrossing Crop plant	[35]
	<i>Brassica rapa</i>	Diploid	490 Mb	3	87	Syteny	N/A	>95% Outcrossing Crop plant	[96]
	<i>Capsicum</i> (pepper)	Diploid	3.5 Gb	383	56	Read mapping	N/A	10%–90% Outcrossing Crop plant	[36]
	<i>Glycine soya</i> (soybean)	Tetraploid (diploidized)	1 Gb	7	49	HGC	80	2%–20% Outcrossing Wild crop relative	[39]
	<i>Helianthus annuus</i> (sunflower)	Diploid	3 Gb	493	83	Sequence similarity	N/A	Mostly self-incompatible (modern varieties self-compatible)	[42]
	<i>Medicago truncatula</i>	Diploid	465 Mb	15	33	HGC	58	1%–35% Outcrossing Non-crop legume	[37]

Les pangénomes, mise à jour et exploitation en agriculture.

Le ratio de taille entre core et dispensable génome est très variable d'un règne et d'une espèce à l'autre

	Species	Ploidy	~Genome size	Number of individuals	Core genome (% genes/ gene clusters)	Core gene/ gene cluster detection method	Core genome (% sequence)	Additional information ^a	Refs
Animals	<i>Mytilus galloprovincialis</i> (mussel)	Diploid	1.6 Gb	16	75	Read mapping	N/A	Marine filter-feeder	[55]
	<i>Homo sapiens</i> ^b (human)	Diploid	3.2 Gb	2	86 Additional genes found	N/A	19–40 Mb additional sequence predicted	Asian and African genomes	[50]
	<i>H. sapiens</i> ^b	Diploid	3.2 Gb	910	N/A	N/A	296 Mb additional sequence found	Individuals of African descent	[51]
	<i>H. sapiens</i>	Diploid	3.2 Gb	185	97	Read mapping	N/A	Han Chinese Individuals	[3]
	<i>Sus scrofa</i> ^b (pig)	Diploid	2.7 Gb	9	1737 Additional genes found	N/A	137 Mb of additional sequence found	European and Chinese breeds	[54]

Le concept de pan-génome est applicable à des espèces de tout l'arbre du vivant.
 Le pan-génome comprends une fraction substantielle de gènes accessoires chez les plantes et les micro-organismes
 alors que, jusqu'à présent, le core génome domine largement chez les animaux.

Quels sont les facteurs influençant la taille du pan-génome ?

Quelles sont les forces évolutives qui façonnent le pan-génome ?

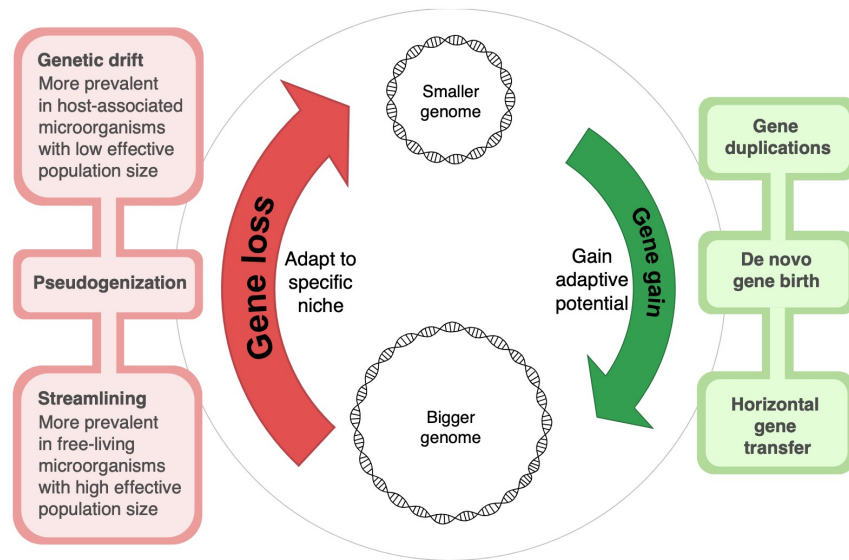
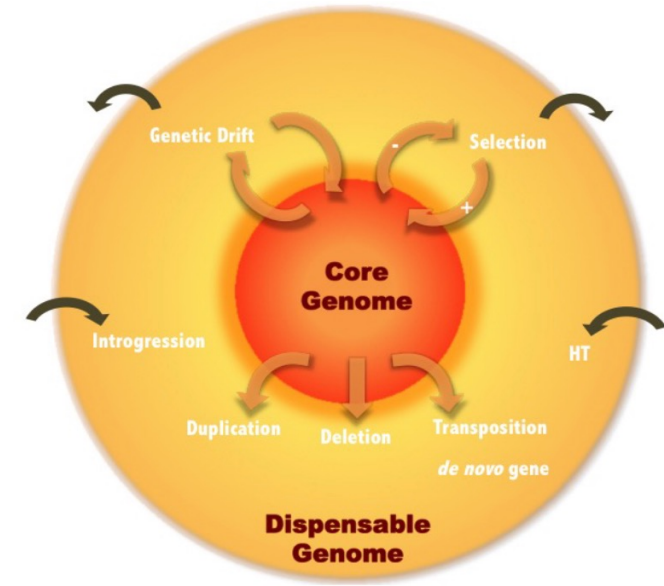


Figure 4. Conceptual figure of the evolutionary forces driving the expansion and reduction of genome sizes. Gene loss is represented with a bigger arrow because it dominates the evolutionary history we know based on extant microorganisms.



Tranchant-Dubreuil, C., Rouard, M. & Sabot, F.
Annual Plant Reviews online. 1–25 (wiley, 2019).

Origine des gènes accessoires

Bactéries : principalement transferts de gènes horizontaux

Eucaryotes : duplication (locale ou génome complet) ou naissance de gènes *de novo*

Quels sont les facteurs influençant la taille du pan-génome ?

La taille effective de population est corrélée à la taille du pan-génome

Prokaryote genome fluidity is dependent on effective population size

Nadia Andrea Andreani^{1,2,4}, Elze Hesse³ and Michiel Vos²

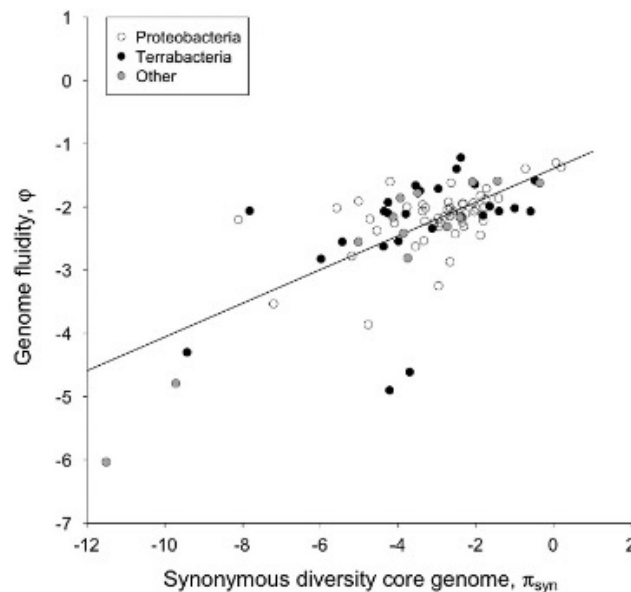


Figure 1 The genome fluidity statistic ϕ as a function of synonymous core genome nucleotide variation π for 90 free-living prokaryote species on a ln-ln scale. White dots: Proteobacteria, black dots: Terrabacteria (Actinobacteria, Firmicutes and Cyanobacteria), grey dots: other taxa.

La domestication et la sélection réduisent la taille du pan-génome

ARTICLES

<https://doi.org/10.1038/s41588-019-0410-2>

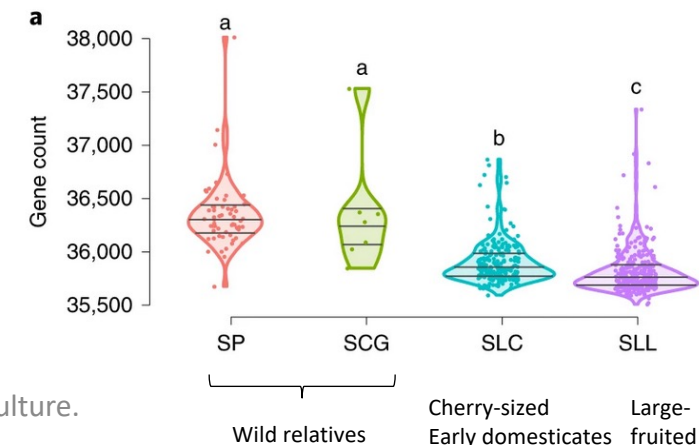
nature
genetics

The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor

Lei Gao^{1,10}, Itay Gonda^{1,2,10}, Honghe Sun¹, Qiyue Ma¹, Kan Bao¹, Denise M. Tieman³, Elizabeth A. Burzynski-Chang⁴, Tara L. Fish⁵, Kaitlin A. Stromberg¹, Gavin L. Sacks^{1,4}, Theodore W. Thannhauser⁵, Majid R. Foolad⁶, Maria Jose Diez⁷, Jose Blanca⁷, Joaquin Canizares⁷, Yimin Xu¹, Esther van der Knaap⁸, Sanwen Huang⁹, Harry J. Klee³, James J. Giovannoni^{1,5*} and Zhangjun Fei^{1,5*}

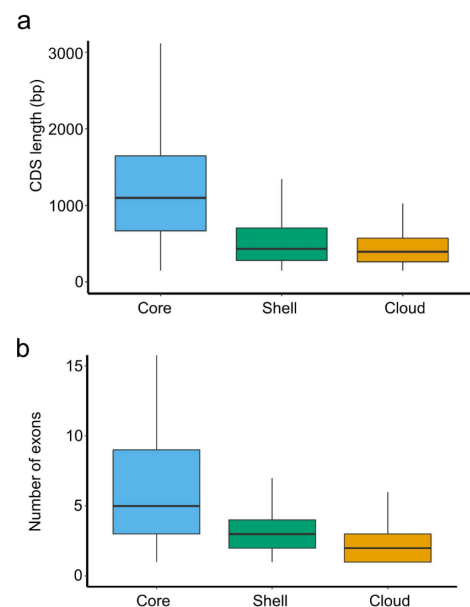
Selection of gene PAVs during tomato breeding

Genomes of wild accessions (SP and SCG) encoded significantly more genes than SLC, whereas SLC contained significantly more genes than SLL (Fig. 2a), suggesting a general trend of gene loss during tomato domestication and subsequent improvement. Furthermore, more genes were lost during domestication than improvement

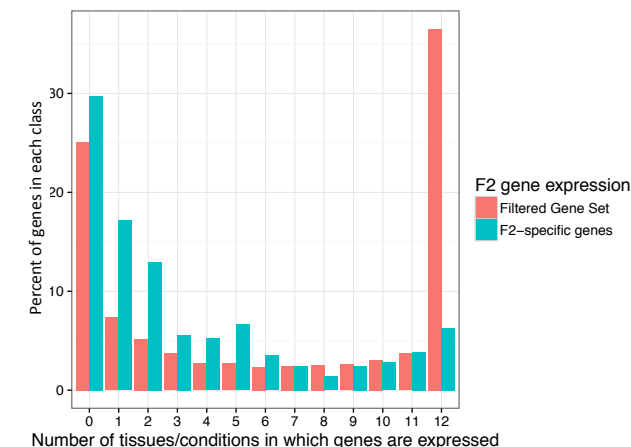
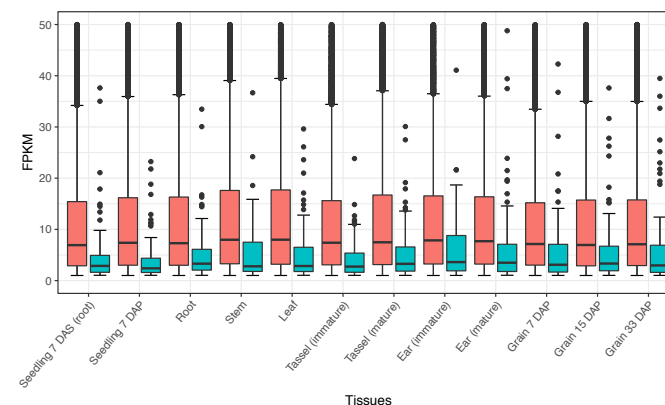


Les gènes du « flexible genome » sont différents de ceux du « core genome »

Ils sont plus courts, moins fortement exprimés et dans moins de conditions



a CDS length. Core genes are significantly longer than shell and cloud genes (p -value = 2.2×10^{-16} , Wilcoxon signed rank, two-sided). **b** Number of exons. Core genes have significantly more exons than shell and cloud genes (p -value = 2.2×10^{-16} , Wilcoxon signed rank, two-sided). Sample size: core, $n = 15,867$; shell, $n = 28,026$; cloud genes, $n = 186$. In the box plots, center lines represent the median, the bottom and top of boxes represent the first and third percentiles, whiskers show the data that lie within the 1.5 interquartile range of the first and third quartiles.



Les gènes du « flexible genome » sont différents de ceux du « core genome »

Leurs séquences sont moins bien conservées et ce sont plus souvent des gènes orphelins

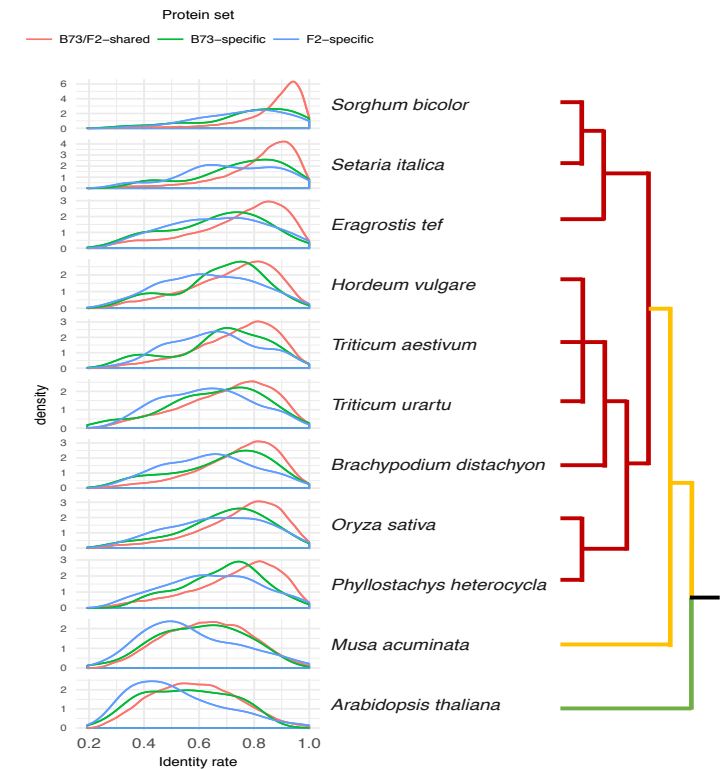
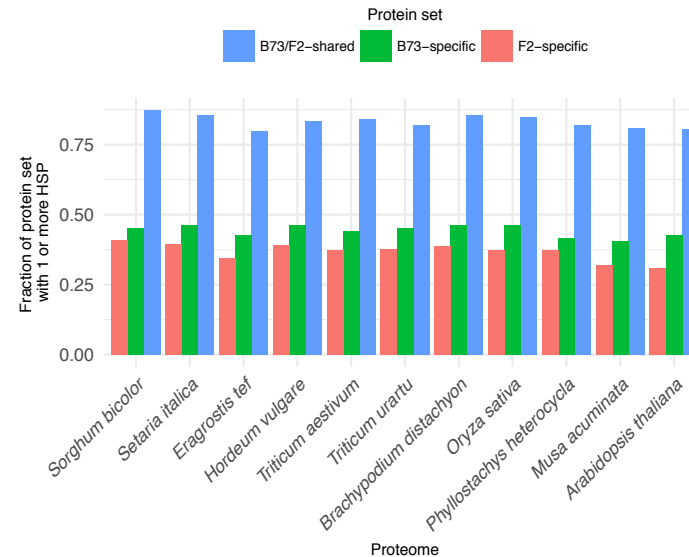
RESEARCH ARTICLE

Open Access



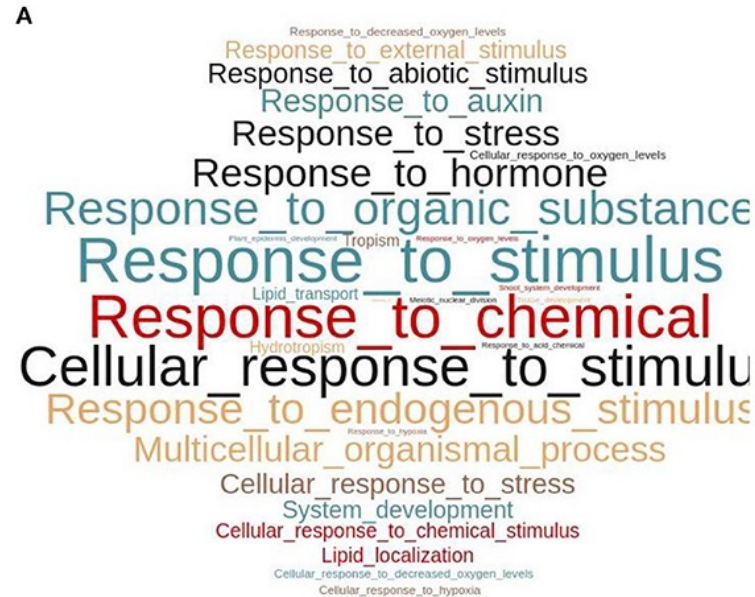
Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants

Aude Darracq^{1*}, Clémentine Vitte^{1*}, Stéphane Nicolas¹, Jorge Duarte², Jean-Philippe Pichon², Tristan Mary-Huard^{1,3}, Céline Chevalier¹, Aurélie Bérard⁴, Marie-Christine Le Paslier⁴, Peter Rogowsky², Alain Charcosset¹ and Johann Joets¹



Ils sont aussi beaucoup plus souvent de fonction inconnue !

Des gènes pouvant s'absenter sont-ils utiles ?



Sorghum Pan-Genome Explores the Functional Utility for Genomic-Assisted Breeding to Accelerate the Genetic Gain

Pradeep Ruperao¹, Nilesh Thirunavukkarasu², Prasad Gandham³, Sivasubramani Selvanayagam⁴, Mahalingam Govindaraj⁵, Balousa Nibbel⁶, Eric Manyasa⁷, Rajeev Gupta⁸, Roma Rani Das⁹, Damaris A. Odeny¹⁰, Harish Gandhi¹¹, David Edwards¹², Santosh P. Deshpande¹³ and Abhishek Rathore¹⁴

La majorité des gènes accessoires sont de fonction inconnue

Quelque soit le règne considéré, les fonctions biologiques sur-représentées dans le génome accessoire sont généralement la *réponse à l'environnement*, la *communication* et dans une moindre mesure le *développement*

Le génome accessoire semble globalement conférer *une capacité d'adaptation à différents environnements et à leur variations*.

Le génome accessoire est-il adaptatif ou neutre ?

correspondence

Are pangenomes adaptive or not?

To the Editor — Most types of bacteria can rapidly take up and lose genes, resulting in different strains carrying unique selections of accessory genes drawn from the species' 'pangenome'. It is well established that acquiring accessory genes can have adaptive, neutral or deleterious effects on fitness, but the relative prevalence of these different types of changes remains an important open question in evolutionary microbiology¹.

In a recent Perspective in *Nature Microbiology*, McInerney et al.² argue that accessory genes mostly have beneficial effects on fitness. Although we cannot disprove this scenario, we find that their arguments are inconclusive. They argue that it is unlikely that most gene content changes are deleterious. First, based on a comparative genomics analysis, they suggest that selfish genetic elements (which are most likely to be deleterious) are not the dominant class of gene content changes. However, at least two other studies have found that selfish elements do dominate the accessory genome^{3,4}. Furthermore, this in itself is not an argument that any of the more commonly

and not gene loss, and this might not always be the case⁵.

McInerney et al. make the case for the general advantageousness of accessory genes based on three lines of evidence. The empirical evidence does not convince us — one study cited by McInerney et al. found that experimental deletion of accessory genes is mostly deleterious, suggesting that the presence of these genes is mostly advantageous. However, in this study⁶, every deletion involved multiple genes, including core genes. This makes these results inconclusive, as deletion of beneficial genes could have been accompanied by deletion of multiple neutral or slightly deleterious genes. On a theoretical note, they argue that species with large effective population size (N_e) have large pangenomes, as these species experience more effective selection. Although it is true that species with larger N_e are expected to have a lower proportion of deleterious polymorphism, they still are expected to carry a higher absolute number of such non-adaptive changes. The argument made by McInerney et al. for the

note that gene transfer processes can be generally advantageous even when the majority of gene content changes are (slightly) deleterious, or that they can be deleterious even when occasionally giving rise to adaptations. It is difficult to think of data that will resolve the question of whether accessory genomes are dominated by deleterious, neutral or adaptive genetic variation. Purely sequence-based analysis is hampered by the fact that gene content changes cannot be divided into classes with different selective expectations, as is the case for nucleotide variation (that is, synonymous and non-synonymous polymorphism). Experiments combining careful genome editing to add or delete accessory genes could be useful, but are contingent on being able to perform sensitive fitness assays in relevant environments. By whatever approach, we hope that researchers take up the gauntlet to provide new data that could help resolve this important evolutionary puzzle. □

Michiel Vos¹* and Adam Eyre-Walker^{2*}
¹Department of Biology, University of Oxford, Oxford, UK
²Department of Biology, University of Oxford, Oxford, UK

correspondence

The population genetics of pangenomes

To the Editor — Two short papers recently attempted to explain prokaryotic pangenomes in light of population genetic theory, reaching surprisingly different conclusions. McInerney, McNally and O'Connell¹ define pangenomes as "the entire collection of gene families that are found in a given species" and conclude that "pangenomes are the result of adaptive, not neutral, evolution", whereas Andreani, Hesse and Vos² conclude the opposite, that "accessory gene turnover is for a large part dictated by neutral evolution". How can these views be reconciled? Going forward, how can we best use population genetic theory to understand pangenomes?

Both papers rely on neutral or nearly neutral population genetic theory to make their arguments. Using a dataset of 90 bacterial and archaeal species, Andreani et al. used a correlation analysis to demonstrate

genetic diversity, and also respond more efficiently to natural selection. Thus, some of the correlation between N_e and pangenome fluidity observed by Andreani et al. can be explained by neutral evolution, and some by niche adaptation.

While I applaud the use of population genetic theory to explain pangenomes, the theory relies on clearly defined populations — a concept that does not apply well to pangenomes as they are shaped by horizontal gene transfer (HGT), which can occur across population (or species) boundaries. McInerney et al. state that "HGT is a form of mutation and can be treated as such in models of pangenome evolution". This is a very strong assumption, which does not necessarily work. They go on to explain some basic population theory: "A truly neutral allele in a population of size N will have an initial frequency of $1/N$. If the

apply the theory on a gene-by-gene basis rather than a species-by-species basis, as different mobile genes will have different values of N_e . This will allow us to answer questions such as: do genes with high N_e tend to occur in very many niches? And which genes are associated with which ecological niches? The fact that Andreani et al. do observe a strong correlation between N_e and pangenome fluidity using species rather than genes as units suggests that most named species are indeed coherent units. If HGT is more frequent within than between species^{3,10}, most genes within a species would tend to have a similar N_e . However, distinguishing between neutral and adaptive interpretations of pangenomes will require a gene-focused and ecological approach. □

B. Jesse Shapiro

Chez les bactéries

To the Editor — Two short papers recently attempted to explain prokaryotic pangenomes in light of population genetic theory, reaching surprisingly different conclusions. McInerney, McNally and O'Connell¹ define pangenomes as "the entire collection of gene families that are found in a given species" and conclude that "pangenomes are the result of adaptive, not neutral, evolution", whereas Andreani, Hesse and Vos² conclude the opposite, that "accessory gene turnover is for a large part dictated by neutral evolution". How can these views be reconciled? Going forward, how can we best use population genetic theory to understand pangenomes?

Chez les eucaryotes

La question est complexifiée par des facteurs additionnels :

- La redondance fonctionnelle
- L'héritage possible de long haplotypes contenant plusieurs gènes accessoires de fonction différentes.

Dans tous les cas

- Le fait qu'une région génomique soit considérée comme essentielle, adaptative, délétère ou neutre peut évoluer avec le temps ou l'environnement.
- La majorité des gènes accessoires sont de fonction inconnue.

Les pangénomes, mise à jour et exploitation en agriculture.

Les éléments transposables peuvent représenter une fraction significative et hypervariable du génome

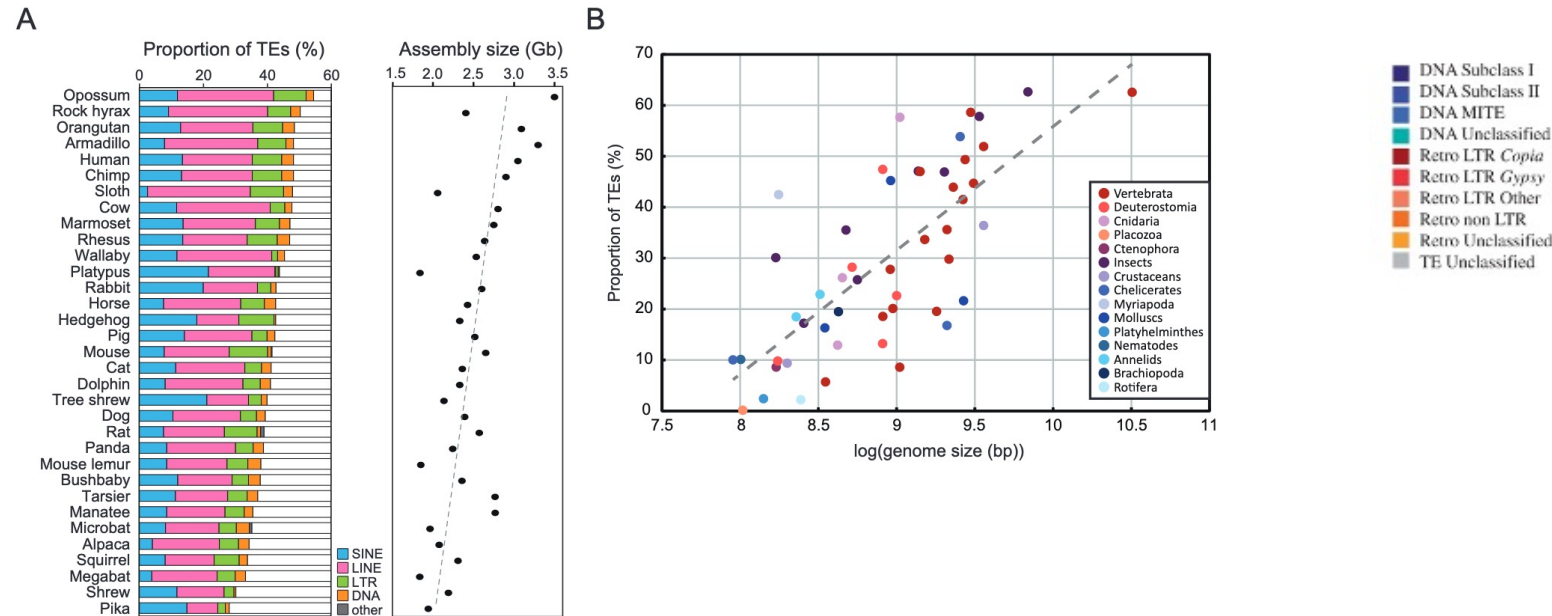
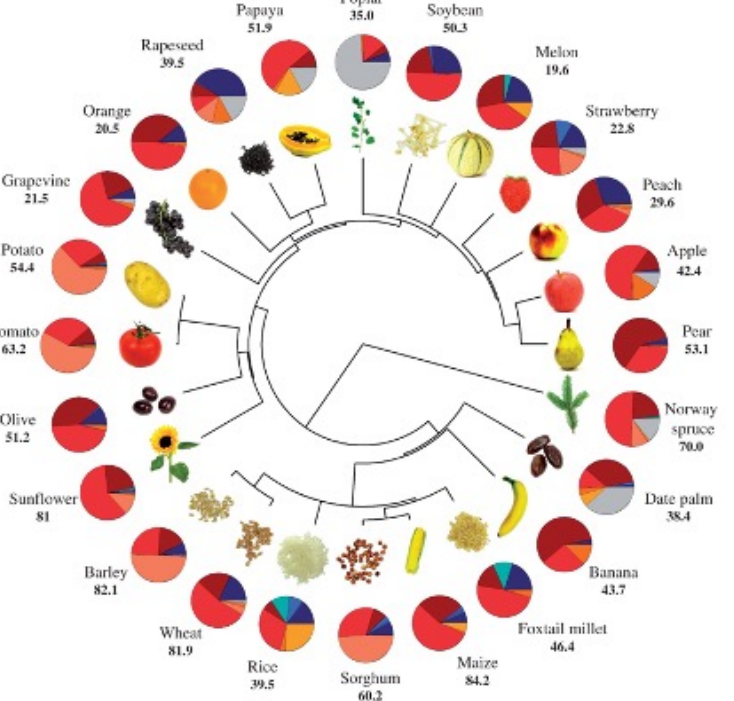


Fig. 3. TEs and genome sizes in animals. (A) The proportion of each class of TEs in mammalian genomes and the genome assembly sizes. The species are arranged in descending order of their TE content. TE content data were obtained from the RepeatMasker web-site (<http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html>). (B) Correlation between genome size and the proportion of TEs in animals. The data were retrieved from the original reports of genome sequencing projects. Vertebrata: human, mouse, opossum, platypus, chicken, alligator, painted turtle, lizard, frog, axolotl, coelacanth, *Takifugu*, Nile tilapia, Atlantic salmon, zebrafish, spotted gar and elephant shark; Deuterostomia: lamprey, lancelet, tunicate, sea urchin and acorn worm; Cnidaria: coral, *Hydra* and starlet sea anemone; Placozoa: *Trichoplax*; Ctenophora: comb jelly; Insects: mountain pine beetle, silkworm, honey bee, fruit fly, locust, dampwood termite, *Blattella germanica* and American cockroach; Crustaceans: amphipod and water flea; Chelicerates: tick, spider mite and velvet spider; Myriapoda: centipede; Molluscs: owl limpet, snail and octopus; Platyhelminthes: tapeworm; Nematodes: *Caenorhabditis elegans*; Annelids: freshwater leech and marine polychaete; Brachiopoda: brachiopod; Rotifera: rotifer.

Genes Genet. Syst. (2019) 94, p. 269–281

Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation

Hiddenori Nishihara*



BRIEFINGS IN FUNCTIONAL GENOMICS. VOL. 13. NO. 4. 276–295

The bright side of transposons in crop evolution

Clémentine Vitte, Margaux-Alison Fustier, Karine Alix and Maud I. Tenailon

Advance Access publication date 27 March 2014

Les pangénomes, mise à jour et exploitation en agriculture.

Chez les plantes le concept de pan-génome a d'abord été lié à la variabilité structurale produite par les éléments transposables.

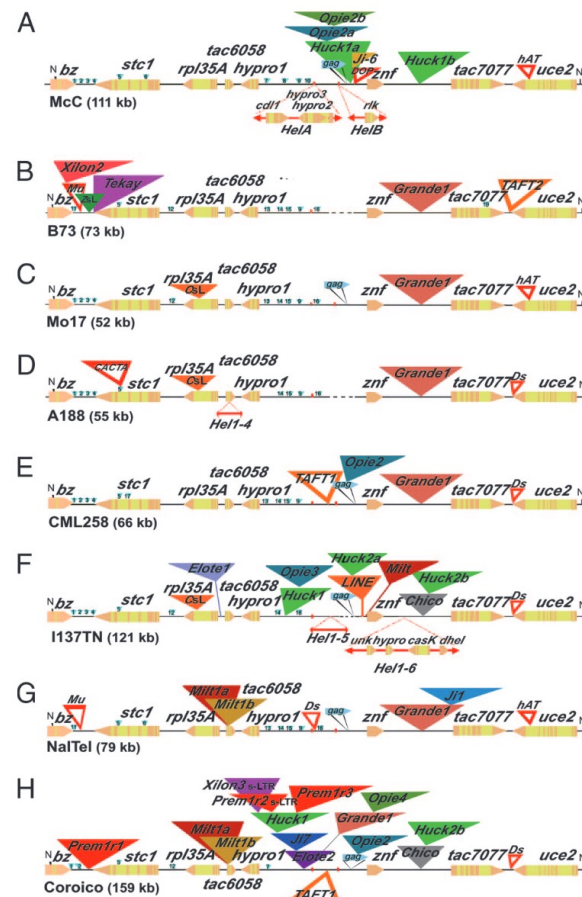
NAS

Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus

Qinghua Wang* and Hugo K. Dooner**

*The Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, NJ 08855; and **Department of Plant Biology, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901

Edited by Susan R. Wessler, University of Georgia, Athens, GA, and approved August 1, 2006 (received for review April 17, 2006)



De 25 à 85% des
séquences ne sont pas
partagées à un locus



Transposable elements and the plant pan-genomes

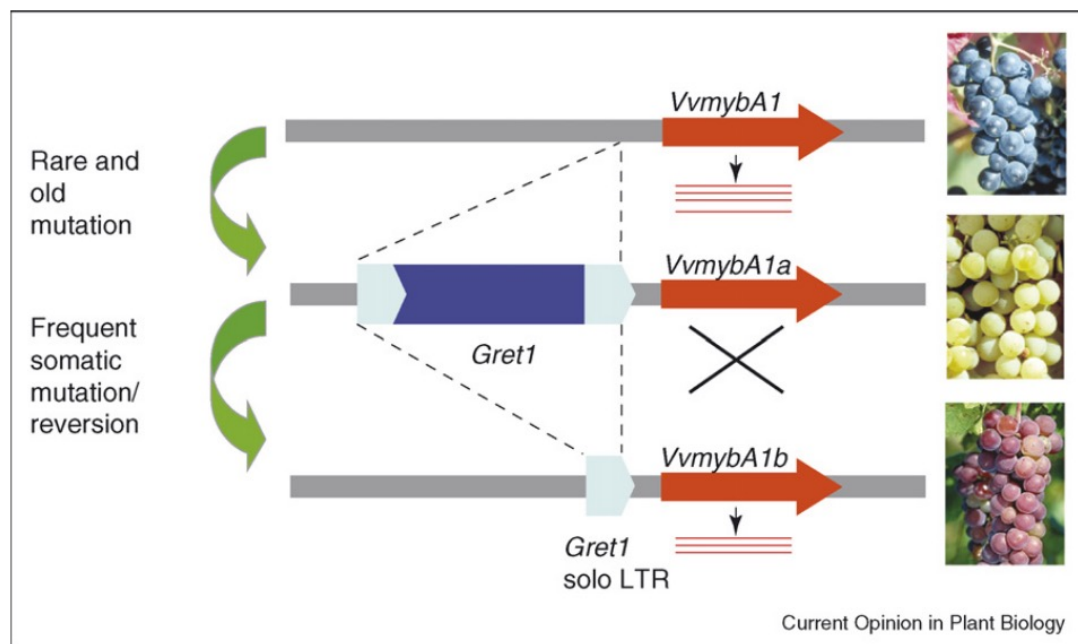
Michele Morgante^{1,2}, Emanuele De Paoli¹ and Slobodanka Radovic¹ 2007Full text provided by www.sciencedirect.com
ScienceDirect

« The **core genome** comprises both single-copy sequences (including most if not all genes) and transposable elements that are found among all individuals in a certain genomic location.

The **dispensable genome** is made up mostly of transposable elements of different types that, although present in multiple copies in each individual, can be found in a specific location only in some of them. A gene-like fraction can also be found in the dispensable maize genome. »

Les pangénomes, mise à jour et exploitation en agriculture.

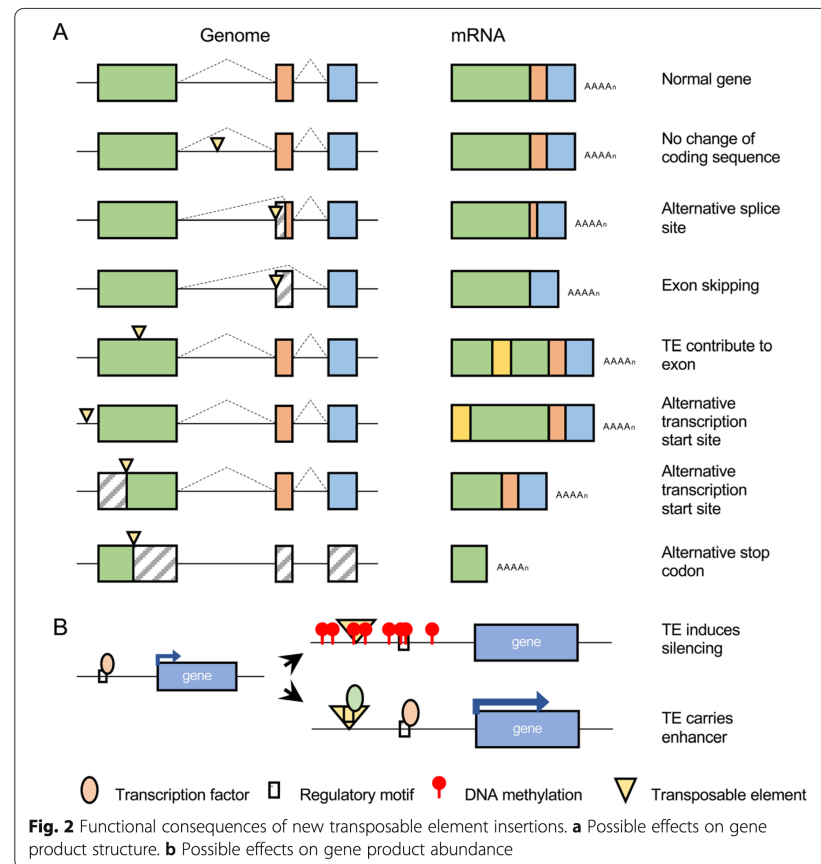
Les éléments transposables peuvent affecter la régulation des gènes de plusieurs manières.



Transposable elements and the plant pan-genomes

Michele Morgante^{1,2}, Emanuele De Paoli¹ and Slobodanka Radovic¹

Full text provided by www.sciencedirect.com



Les éléments transposables peuvent affecter la régulation des gènes de plusieurs manières.

Les éléments transposables sont une source d'enhanceurs* tissus-spécifiques

*= CRE Cis-Regulatory Element

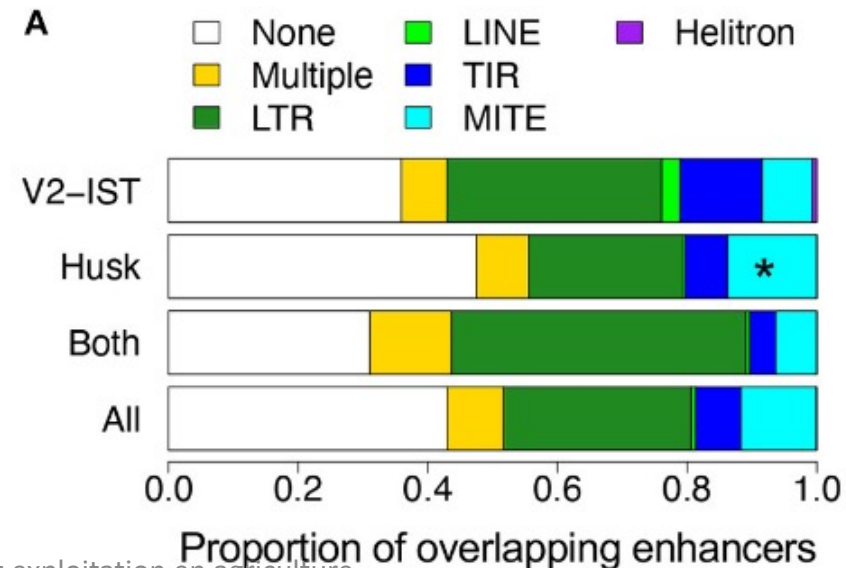
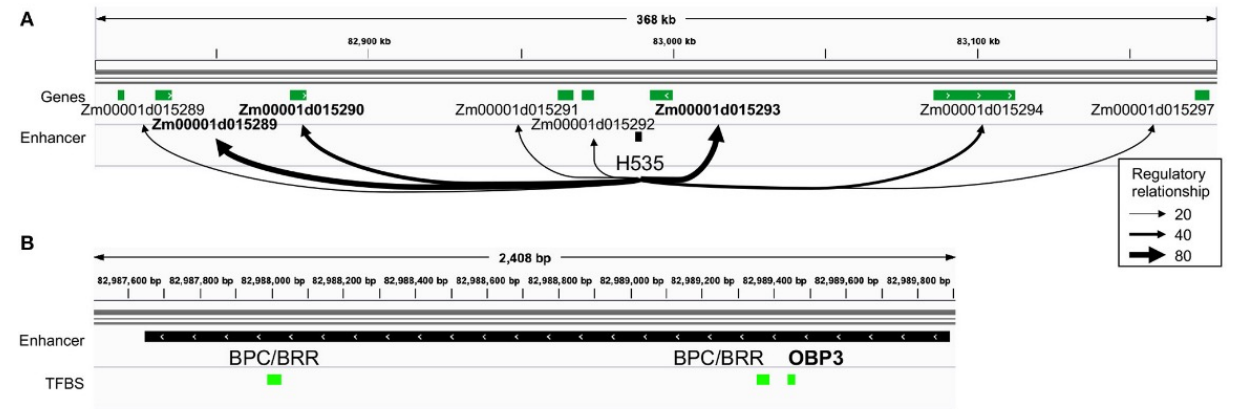
ORIGINAL RESEARCH article

Front. Genet., 11 January 2021 | <https://doi.org/10.3389/fgene.2020.606285>



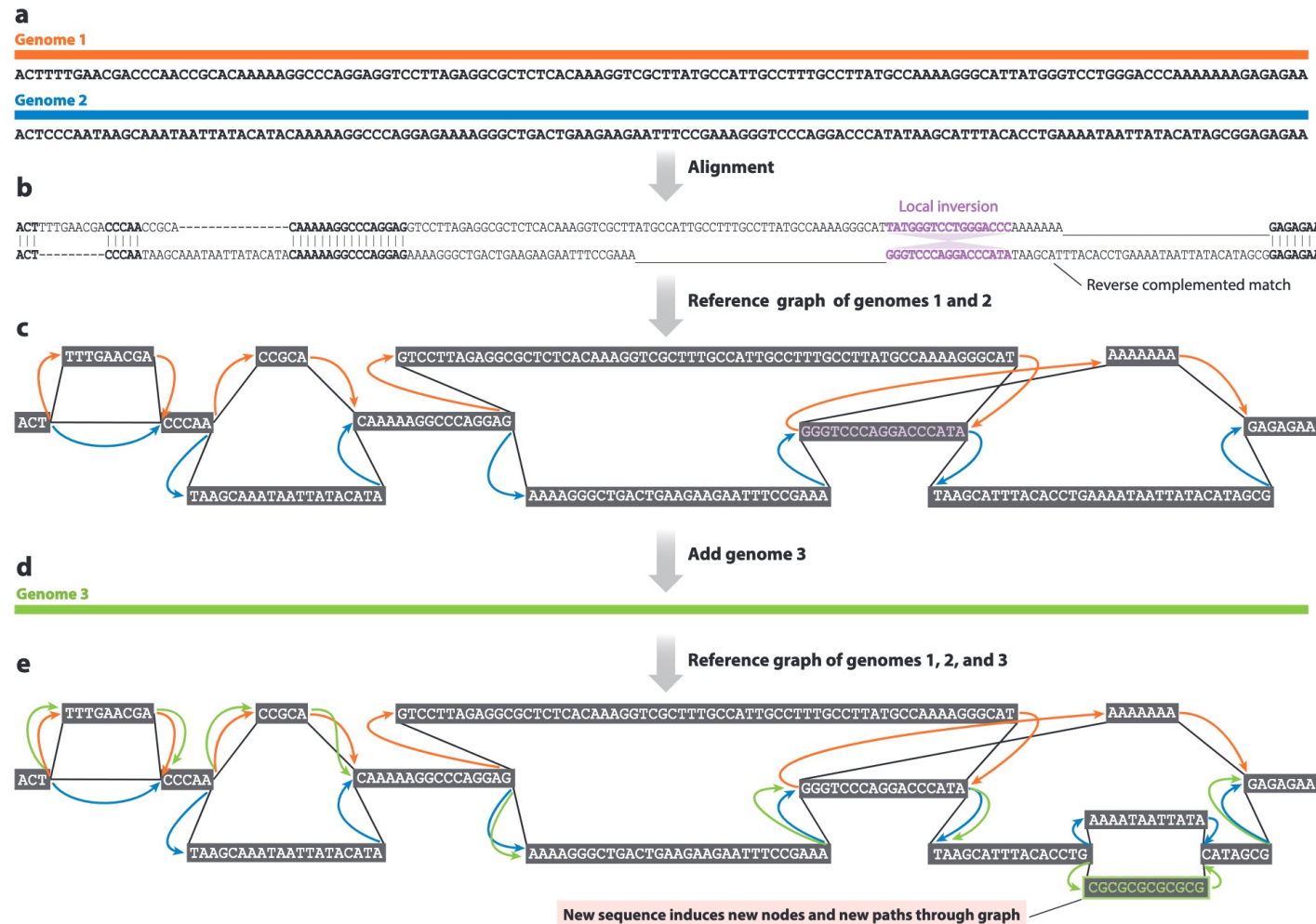
Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks

Maud Fagny^{1†}, Marieke Lydia Kuijjer^{2,3}, Maïke Stam⁴, Johann Joets⁵, Olivier Turc⁵, Julien Rozière^{1,6,7}, Stéphanie Pateyron^{6,7}, Anthony Venon¹ and Clémentine Vitte^{1*}



Le graphe de variation

Généralisation du concept de pan-génome à l'ensemble des variations de séquence observées dans une espèce.



Les pangénomes, mise à jour et exploitation en agriculture.

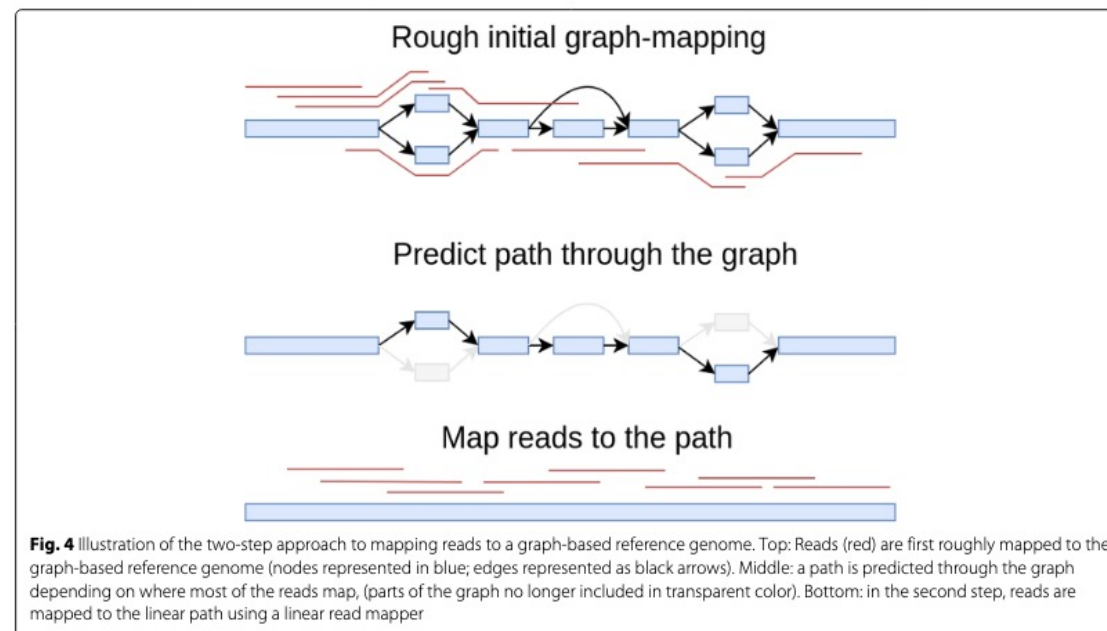
Annual Review of Plant Biology
Plant Pan-Genomics
Comes of Age

Li Lei,¹ Eugene Goltsman,¹ David Goodstein,¹
Guohong Albert Wu,¹ Daniel S. Rokhsar,^{1,2}
and John P. Vogel^{1,3}

Le graphe de variation

Généralisation du concept de pan-génome à l'ensemble des variations de séquence observées dans une espèce.

En alignant des « short reads » directement sur le graphe il est possible de génotyper facilement de nombreux individus



Grytten et al. *BMC Genomics* (2020) 21:282
<https://doi.org/10.1186/s12864-020-6685-y>

BMC Genomics

METHODOLOGY ARTICLE

Open Access

Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods



Exploring and Exploiting Pan-genomics for Crop Improvement

Yongfu Tao^{1,4}, Xianrong Zhao^{1,4}, Emma Mace^{1,2}, Robert Henry³ and David Jordan^{1,*}

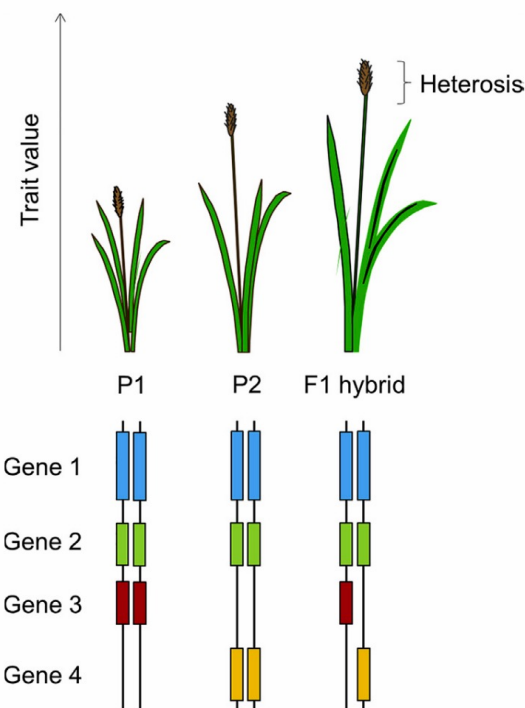


Figure 4. Complementarity of Genes between Parental Lines Drives Heterosis.

Heterosis refers to the superior performance of F_1 hybrids compared with their parental lines. The dominance hypothesis is one of three main theories explaining heterosis. According to the dominance hypothesis, complementarity of genes between parental lines leads to heterosis.

High-resolution genetic mapping of maize pan-genome sequence anchors

Fei Lu¹, Maria C. Romay¹, Jeffrey C. Glaubitz¹, Peter J. Bradbury², Robert J. Elshire¹, Tianyu Wang³, Yu Li³, Yongxiang Li³, Kassa Semagn⁴, Xuecai Zhang⁵, Alvaro G. Hernandez⁶, Mark A. Mikel^{6,7}, Ilya Soifer⁸, Omer Barad⁸ & Edward S. Buckler^{1,2}

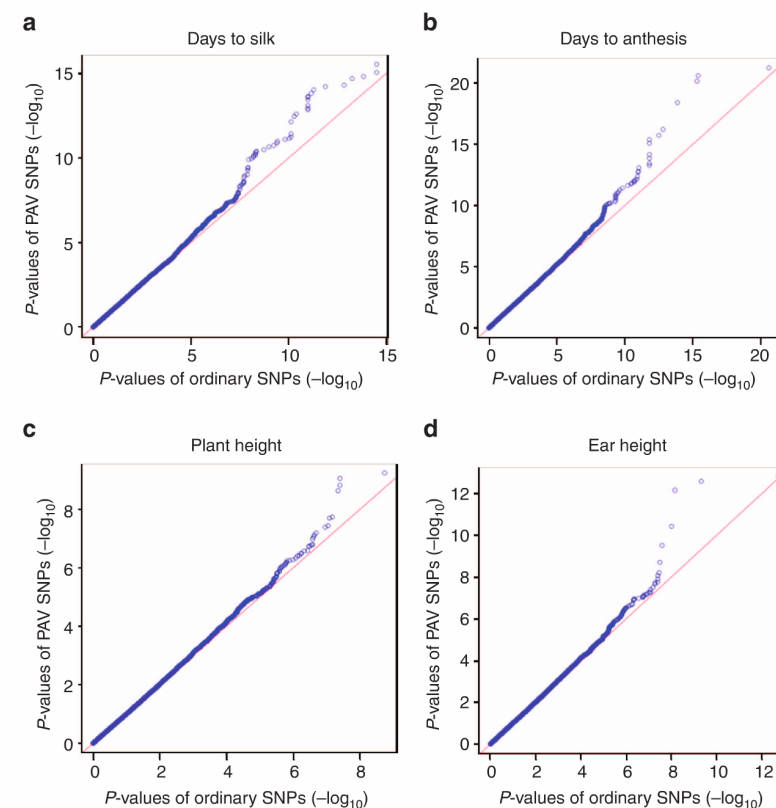


Figure 6 | QQ plots showing PAV SNPs are enriched for the significant GWAS hits. For the large amount of insignificant SNPs, the PAV SNPs and ordinary SNPs did not show differentiated distributions; while for those significant GWAS hits, PAV SNPs exhibited enriched associations with traits.

Les pangénomes, mise à jour et exploitation en agriculture.

Exploring and Exploiting Pan-genomics for Crop Improvement

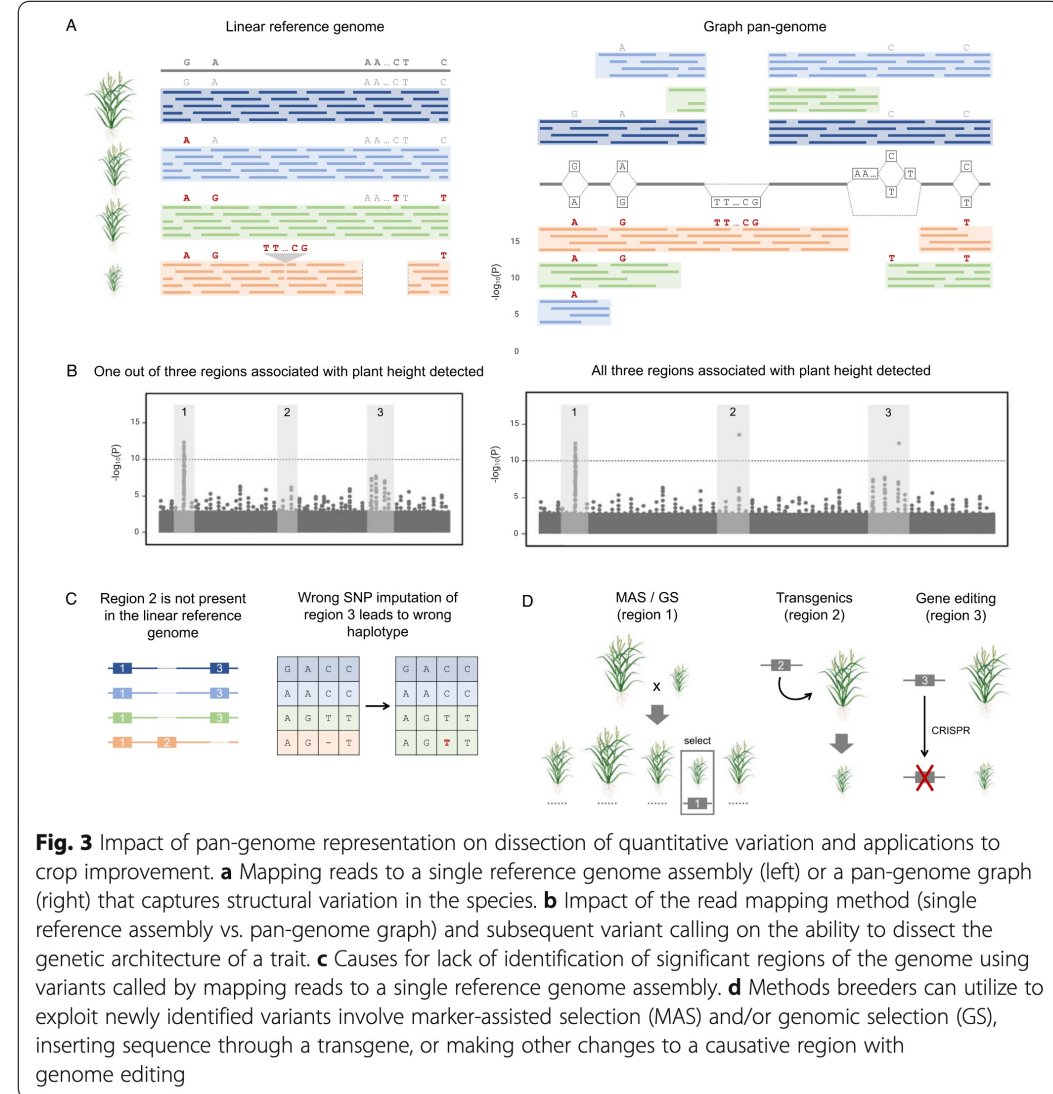
Yongfu Tao^{1,4}, Xianrong Zhao^{1,4}, Emma Mace^{1,2}, Robert Henry³ and David Jordan^{1,*}

Species	Gene (locus)	Trait	Type	Reference
Rice	<i>GW5/qSW5/GSE5</i>	Grain size	PAV	Liu et al., 2017a ; Shomura et al., 2008 ; Duan et al., 2017
Rice	<i>GL7</i>	Grain size	CNV	Wang et al., 2015
Rice	<i>SGDP7</i>	Grain size, grain number, yield	CNV	Bai et al., 2017
Rice	<i>Pikm1-TS</i>	Blast resistance	PAV	Ashikawa et al., 2008
Rice	<i>Pikm2-TS</i>	Blast resistance	PAV	Ashikawa et al., 2008
Rice	Sub1A	Submergence tolerance	PAV	Xu et al., 2006
Rice	<i>Pup1</i>	Phosphorus-starvation tolerance	PAV	Gamuyao et al., 2012
Rice	<i>SNORKEL1</i>	Deep water response	PAV	Hattori et al., 2009
Rice	<i>SNORKEL2</i>	Deep water response	PAV	Hattori et al., 2009
Rice	<i>qPE9-1</i>	Plant architecture	PAV	Zhou et al., 2009
Rice	<i>Pi21</i>	Blast disease	PAV	Fukuoka et al., 2009
Rice	<i>Sc</i>	Hybrid male sterility	CNV	Shen et al., 2017
Rice	<i>DPL1/DPL2</i>	Hybrid male sterility	PAV	Mizuta et al., 2010
Rice	<i>S27/S28</i>	Hybrid male sterility	PAV	Yamagata et al., 2010
Rice	<i>OsSh1</i>	Shattering	PAV	Lin et al., 2012
Maize	<i>KRN4</i>	Kernel row number	PAV	Liu et al., 2015
Maize	<i>ZmCCT10</i>	Photoperiod sensitivity	PAV	Yang et al., 2013
Maize	<i>TB1</i>	Apical dominance	PAV	Studer et al., 2011
Maize	<i>Vgt1</i>	Flowering time	PAV	Salvi et al., 2007
Maize	<i>qHSR1</i>	Resistance to head smut	PAV	Zuo et al., 2015
Maize	<i>Scmv1</i>	Resistance to sugarcane mosaic virus	SV	Liu et al., 2017b
Maize	<i>ZmCCT9</i>	Photoperiod sensitivity	PAV	Huang et al., 2018
Maize	<i>MATE1</i>	Aluminum tolerance	CNV	Maron et al., 2013
Sorghum	<i>LGS1</i>	Resistance to <i>Striga</i>	PAV	Gobena et al., 2017
Sorghum	<i>Sh1</i>	Shattering	PAV	Lin et al., 2012
Sorghum	<i>SbMATE</i>	Aluminum tolerance	PAV	Magalhaes et al., 2007

De nombreux gènes accessoires ou variants d'éléments transposables sont impliqués dans des caractères d'intérêt agronomiques.

Pourquoi la pan-génomique peut-elle contribuer à l'amélioration génétique ?

Elle permet d'augmenter la puissance des approches de génétique quantitative (QTL, GWAS, Genomic prediction)



Della Coletta, R., Qiu, Y., Ou, S. et al. Genome Biol 22, 3 (2021)

Le core genome permet de mieux identifier les gènes partagés et essentiels

Identification de nouvelles cibles vaccinales ou de nouveaux agents antimicrobiens en ciblant les produits de gènes partagés par l'ensemble des souches d'un pathogène.

Diminuer les pertes d'animaux d'élevage liées à la létalité embryonnaire induite par certaines mutations de gènes essentiels.

Meilleure compréhension des interactions hôte-pathogène.

Donner d'autres définitions de l'espèce ? *B. cereus* and *B anthracis* sont considérées comme deux espèces indépendantes alors qu'elle ne diffèrent que par une partie du génome accessoire (porté par deux plasmides).