The genome of Apis mellifera: dialog between linkage mapping and sequence assembly

Michel Solignac, Lan Zhang, Florence Mougel, Bingshan Li, Dominique Vautrin, Monique Monnerot, Jean-Marie Cornuet, Kim C Worley, George M. Weinstock, Richard A. Gibbs

Correspondence

The genome of *Apis mellifera*: dialog between linkage mapping and sequence assembly

Michel Solignac*, Lan Zhang†, Florence Mougel*, Bingshan Li†, Dominique Vautrin*, Monique Monnerot*, Jean-Marie Cornuet‡, Kim C Worley†, George M Weinstock† and Richard A Gibbs†

Addresses: *Laboratoire Evolution, Génomes et Spéciation, Centre National de la Recherche Scientifique, 91198 Gif-sur-Yvette cedex, France and University of Paris Sud, 91405 Orsay, France. †Human Genome Sequencing Center, Baylor College of Medicine, Alkek 1519, One Baylor Plaza, Houston, TX 77030, USA. ‡Centre de Biologie et de Gestion des Populations, INRA, CS 30016 Montferrier-sur-Lez, 34988 Saint-Gély-du-Fesc, France.

Correspondence: Michel Solignac. Email: solignac@legs.cnrs-gif.fr

Published: 19 March 2007

Genome ***Biology*** 2007, **8:**403 (doi:10.1186/gb-2007-8-3-403)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/3/403

Abstract

Two independent genome projects for the honey bee, a microsatellite linkage map and a genome sequence assembly, interactively produced an almost complete organization of the euchromatic genome. Assembly 4.0 now includes 626 scaffolds that were ordered and oriented into chromosomes according to the framework provided by the third-generation linkage map (AmelMap3). Each construct was used to control the quality of the other. The co-linearity of markers in the sequence and the map is almost perfect and argues in favor of the high quality of both.

Most eukaryotic genome sequencing projects are preceded by the construction of physical, genetic and/or cytological maps. For the honey bee genome project there was no physical map, and because of the low resolution of the cytogenetic map, the meiotic map was the only resource for organizing the sequence assembly on the chromosomes. The first generation map AmelMap1 comprised 541 markers on 24 linkage groups for 16 chromosomes [1,2]. Saturation was achieved by addition of 601 markers prepared from cDNAs [3] and bacterial artificial chromosomes (BACs) [4] sequences. AmelMap2 was not published, but was used by the Human Genome Sequencing Center at Baylor College for the first assembly of the *Apis mellifera* genome in January 2004. From that time a dialog was set up between the map and sequence projects that became interactive, each taking advantage of the progress of the other. The density of the third-generation map, AmelMap3, was doubled and contributed greatly to the ultimate assembly (version 4.0, March 2006) of the honey bee genome [5].

AmelMap3 comprises 2,008 microsatellite markers (see Additional data file 1) and is 4,000 cM long (M.S, F.M, D.V M.M and J-M.C, unpublished work). Improvements in the map between the second and third generation resulted exclusively from addition of markers designed from the sequence: 587 from previously placed scaffolds in assemblies 1.1 and 2.0 to reduce long genetic distances, orient scaffolds and homogenize the marker density along and among chromosomes and 436 in 379 large unplaced scaffolds (GroupUn) which efficiently increased the fraction of the sequence integrated in chromosomes in the later assemblies (Tables 1 and 2). Chromosomes were oriented by half-tetrad analysis [6]. This orientation was later confirmed by positioning telomeric regions [7] and cytogenetic analysis [5].

Great care was taken to eradicate errors in the final versions (AmelMap3, assembly 4.0). For single markers with uncertain chromosomal positions, new markers were designed; in three cases, the scaffold moved and in two cases the marker did not amplify the expected product. In three cases, two blocks of markers on the same scaffolds mapped to two different positions; adding

Table 1

Improvements between assembly versions 1.1 (January 2004) and 4.0 (March 2006)

Map version	AmelMap2		AmelMap3	
Number of markers	1,050*		2,013†	
Assembly version	1.1		4.0	
	Length (Mb)	Percentage	Length (Mb)	Percentage
Total mapped sequence	110	53%	186	79%
Total unmapped sequence (GroupUn)	96	47%	49	21%
Total scaffold length (Mb)	206	-	235	-

Although the size of the assembled genome increased by 29 Mb (12% of the version 4.0 genome) as a result of additional sequencing reads and better assembly, a total of 76 Mb of sequence (32% of the genome) was mapped to chromosomes with longer scaffolds and additional markers in AmelMap3 compared with AmelMap2. *The number of markers used for the assembly differs from that given in the text (1,142). Markers without accession numbers (92) were omitted. †After the freeze of assembly 4.0, some markers were added and others removed from the AmelMap3, which now comprises 2,008 markers.

Table 2

Number of consistently mapped scaffolds

Assembly version	3.0	4.0
Total number of scaffolds	9,863	9,868
Consistently mapped scaffolds	431	626
Number of scaffolds broken	2	2
Number of scaffolds with inconsistency ignored	7	2

The increase of the number of mapped scaffolds (195) between version 3.0 and 4.0 of the genome assembly is less than the total number of unplaced scaffolds (379) in version 3.0 that were mapped in version 4.0 because many scaffolds were merged into previously mapped scaffolds or combined with other previously unmapped scaffolds.

markers narrowed the region responsible for the chimerism in which the assembly had to be split. Most of the remaining discrepancies were local marker misordering, eradicated by correction of genotyping errors detected by double crossovers.

A few trivial differences persist between the latest versions of the map and the assembly. Sixteen small scaffolds were reversed and the order of eight groups of short scaffolds will also be revisited. This is attributable to the fact that the last map improvements occurred after the freeze of the version 4.0 assembly. Four unresolved discrepancies remain: the map positions of two short scaffolds (1.43 and 3.37), orientation of a long scaffold (10.30) and remnants in a false position of the break of scaffold 6.37. This generally excellent co-linearity pleads in favor of the quality of the two constructions. If some mistakes remain within scaffolds, they should be below the level of resolution of the map (average 93 kb).

This agreement could seem to be a circular argument as the map is the framework of the assembly. This is not the case. The genetic map and sequence scaffolds have been constructed independently. The maps were calculated with a version of the software Cartha-Gène [8] that does not use physical information and the assembly did not use the map to construct the scaffolds but only to organize them. The eradication of errors in the map, even if it used the sequence to detect them and helped their resolution, was based on genetic methods (controls or addition of genotypes).

To evaluate the final control of correctness, the scaffolds that contained at least three markers with two non-null genetic distances were selected. The number of markers flanking non-null distances was 1,319 (that is, two-thirds of the total) and they showed only four local and unresolved mistakes (0.3 %). In addition, the 387 markers that are at a null genetic distance within scaffolds are always clustered in the sequence. This accurate co-linearity within scaffolds may be considered indicative of that between scaffolds, which cannot be tested in this way. In the mouse, a very detailed genetic map existed before the sequence of the genome, but of the 12,000 markers, only 2,605 were considered as 'unambiguously' mapped and were used to assess the accuracy of the assembly [9]; most of the conflicts (1.8% of chromosomal misassignment and 0.7% of local misordering) were attributable to mapping errors. For the rat genome, the radiation hybrid map was consistent for 98% of markers with the genetic maps and for 96% with the genome sequence [10].

Among the 626 honey bee scaffolds, 320, representing a physical length of 152 Mb, are oriented (Table 3); the other half were too short to be oriented genetically; they represent only 18.4% of the physical length. Among them, 113 scaffolds forming 44 blocks are not ordered relative to one another (due to null genetic

Table 3

Total number of scaffolds mapped in the honey bee genome and corresponding physical length of each of the 16 chromosomes

	Number of scaffolds			Physical length (in base pairs)		
Linkage group	Unoriented	Unordered	Total	Unoriented	Oriented	Total
1	37	11 (4)	83	4,324,756	21,509,334	25,834,090
2	21	4 (2)	43	2,072,401	11,899,776	13,972,177
3	15	8 (3)	39	1,707,550	10,013,970	11,721,520
4	13	2 (1)	27	1,741,230	9,215,460	10,956,690
5	13	2 (1)	33	1,898,448	11,002,244	12,900,692
6	30	4 (2)	55	3,630,628	11,408,455	15,039,083
7	28	15 (6)	47	3,141,542	7,407,431	10,548,973
8	26	16 (7)	47	2,825,708	8,063,515	10,889,223
9	11	2 (1)	26	1,566,427	8,266,480	9,832,907
10	22	7 (3)	45	2,686,951	7,755,626	10,442,577
11	22	7 (3)	42	3,091,854	9,380,123	12,471,977
12	16	4 (1)	30	1,527,861	8,331,149	9,859,010
13	4	0	21	399,867	8,866,870	9,266,737
14	10	3 (1)	25	990,212	7,786,449	8,776,661
15	28	22 (6)	42	2,097,987	6,011,700	8,109,687
16	10	6 (3)	21	745,354	5,327,518	6,072,872
Total	306	113 (44)	626	34,448,776	152,246,100	186,694,876
				18.4 %	81.6 %	

Unordered scaffolds are a subset of unoriented scaffolds (number of blocks of unordered scaffolds between brackets).

distances). The unoriented scaffolds are nevertheless placed on chromosomes, but their orientation is random.

Missing sequences in the gaps are probably very short, as suggested by short interscaffold genetic distances. Manual superscaffolding of the five smallest chromosomes (12-16) [11], mainly achieved through relaxing matching criteria, conserved the general structure of the map, included 178 GroupUn scaffolds in the gaps and reduced the 139 scaffolds to 25 superscaffolds by the addition of only 5.5% of the sequence length. For all chromosome arms, the telomeric regions are reached and the centromeric regions are close to being so [5,7]. Consequently, most of the euchromatic sequence of the chromosome arms is now organized and perhaps only 5% is not included in the assembly.

It may be asked if a genetic map alone provides sufficient information to organize an assembly. The large genetic length of the honey bee genome (about 4,000 cM) compared to its relatively small physical size (about 230 cM) was assuredly a great advantage because it suffices to genotype small families to observe recombination between markers at a short physical distance. The same resolution in organisms with shorter maps (that is, most organisms, if not all [12]), would require a larger genotyping effort in terms of the number of individuals, but it might be limited to a few markers within the largest scaffolds to get a reasonable picture of the genome organization.

Additional data files

Additional data file 1, a list of the primers used for mapping is available with this article online.

Acknowledgements

This work was funded by grants to R.A.G. from NHGRI, NIH (1 U54 HG02051 and 1 U54 HG003273) supporting L.Z., B.L., K.C.W., R.A.G., G.M.W., and to M.S. from FEOGA and to Katherine Aronstein from USDA supporting M.S., F.M., D.V., M.M., and J.-M.C.

References

1. Solignac M, Vautrin D, Loiseau A, Mougel F, Baudry E, Estoup A, Garnery L, Haberl M, Cornuet J-M: **Five hundred and fifty microsatellite markers for the study of the honeybee (*Apis mellifera* L.) genome.** *Mol Ecol Notes* 2003, **3:**307-311.
2. Solignac M, Vautrin D, Baudry E, Mougel F, Loiseau A, Cornuet JM: **A microsatellite-based linkage map of the honeybee, *Apis mellifera* L.** *Genetics* 2004, **167:**253-262.
3. Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB, Robinson GE: **Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee.** *Genome Res* 2002, **12:**555-566.
4. Tomkins JP, Luo M, Fang GC, Main D, Goicoechea JL, Atkins M, Frisch DA, Page RE, Guzman-Novoa E, Yu Y, *et al.*: **New**

genomic resources for the honey bee (*Apis mellifera* L.): development of a deep-coverage BAC library and a preliminary STC database. *Genet Mol Res* 2002, **1:**306-316.

5. Consortium HGS: **Insights into social insects from the genome of the honeybee *Apis mellifera*.** *Nature* 2006, **443:**931-949.
6. Baudry E, Kryger P, Allsopp M, Koeniger N, Vautrin D, Mougel F, Cornuet JM, Solignac M: **Whole-genome scan in thelytokous-laying workers of the cape honeybee *(Apis mellifera capensis)*: Central fusion, reduced recombination rates and centromere mapping using half-tetrad analysis.** *Genetics* 2004, **167:**243-252.
7. Robertson HM, Gordon KH: **Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis mellifera*.** *Genome Res* 2006, **16:**1345-1351.
8. Schiex T, Gaspin C: **CARTHAGENE: constructing and joining maximum likelihood genetic maps**. *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**258-267.
9. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
10. Kwitek AE, Gullings-Handley J, Yu J, Carlos DC, Orlebeke K, Nie J, Eckert J, Lemke A, Andrae JW, Bromberg S, *et al.*: **High-density rat radiation hybrid maps containing over 24,000 SSLPs, genes, and ESTs provide a direct link to the rat genome sequence.** *Genome Res* 2004, **14:**750-757.
11. Robertson HM, Reese J, Milshina N, Agarwala R, Solignac M, Walden KK, Elsik C: **Manual superscaffolding of honey bee (*Apis mellifera*) chromosomes 12-16: implications for the draft genome assembly version 4, gene annotation, and chromosome structure.** *Insect Mol Biol*, in press.
12. Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, *et al.*: **Exceptionally high levels of recombination across the honey bee genome.** *Genome Res* 2006, **16:**1339-1344.