How many SNPs are needed to provide an accurate estimate of lineage C introgression into black honey bees?

Henriques D[1], Chávez-Galarza J[1], Johnston JS[2], Rufino J[1], Pinto MA[1]

[1]Mountain Research Centre (CIMO), Polytechnic Institute of Bragança, Campus de Sta. Apolónia, Apartado 1172, 5301-855 Bragança, Portugal.

[2]Dept. of Entomology, Texas A&M University, College Station, Texas 77843-2475, USA.

Correspondence: Maria Alice Pinto, CIMO, Instituto Politécnico de Bragança, Portugal. E-mail: apinto@ipb.pt

Introduction

Beekeeping activities, especially queen trading, have shaped the distribution of honey bees subspecies in Europe, which have resulted in extensive introductions of two lineage C subspecies, *A. m. ligustica* and *A. m. carnica*, into western Europe. As a consequence, replacement and gene flow between native and commercial honey bees have been occurring at varying levels across western European populations. Previous studies have monitored introgression by using microsatellite and PCR-RFLP markers. However, single nucleotide polymorphism (SNP) markers are more advantageous as they provide a genome-wide coverage and higher quality data. In addition, SNPs are suitable for automated high throughput technologies allowing genotyping of hundreds to thousands of loci in many individuals. Here we depart from a set of 1183 SNPs to determine the minimum number of SNPs that provide an estimate of introgression of lineage C honey bees into black honey bees as accurate as that generated by the 1183 loci.

Samples

A total of 77 *A. m. mellifera* individuals from France (18), Denmark (10), the Netherlands (15), Switzerland (6), Scotland (10), Norway (10) and England (8) were collected. Samples of *A. m. carnica* (19) from Croatia and Serbia, and *A. m. ligustica* (17), from Italy, were included as reference populations of C-lineage.

SNP genotyping

1536 SNP loci were scored using Illumina's BeadArray Technology and the Illumina GoldenGate® Assay with a custom Oligo Pool Assay following manufacturer's protocols. The 77 individuals were genotyped using Illumina's Genome Studio software.

Data sets

Of the 1536 a total of 1183 SNPs were available for analysis after removing monomorphic loci (cutoff 2%) and non-calls. To obtain genomic position, each SNP's 100 bp flanking sequence was mapped to the honey bee Assembly 4.5 using BLAST in NCBI. Genomic position was ascertained using the Map Viewer tool available in NCBI.

Introgression was estimated for 14 data sets using different SNP combinations. The reference data set included the 1183 SNPs whereas the other 13 were built by sequentially and randomly eliminating SNPs within a range of 1 to 25 cM (Table 1).

Table 1. The 14 data SNP data sets

Data set	cM	Nº of SNPs
1	Full	1183
2	<1 cM	921
3	<2 cM	743
4	<3 cM	574
5	<4 cM	446
6	<5 cM	327
7	<6 cM	238
8	<7 cM	183
9	<8 cM	138
10	<9 cM	104
11	<10 cM	81
12	<11 cM	61
13	<13 cM	40
14	<25 cM	16

Introgression analysis

Introgression of C-lineage was inferred for each black honey bee individual by running STRUCTURE 2.3.3 (Pritchard *et al.* 2000) for the 14 SNP data sets. The probabilistic estimations of the admixture coefficient (Q) was generated by STRUCTURE using the following settings: admixture model and correlated allele frequency, 250 000 burn in steps, 750 000 MCMC iterations, 20 runs, K=2 clusters. CLUMPP 1.1.2 (Jakobsson and Rosenberg 2007) was used to compute the pairwise "symmetric similarity coefficient" between pairs of runs and to align the 20 runs. The means of the permuted results were plotted using DISTRUCT 1.1 (Rosenberg 2004).

Results

STRUCTURE analysis performed with the 1183 SNPs shows that introgression levels are variable across the 77 individuals sampled in the black honey bee range, with membership proportions (Q) in the yellow cluster as high as 0.69 in France and as low as 0.01 in Norway (Fig. 1).

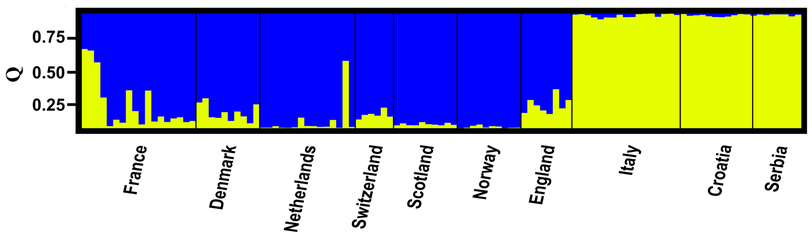

Fig. 1. C- lineage introgression estimates (Q) for the 77 individuals sampled in western Europe using 1183 SNPs

When the single group of the 77 individuals was examined, mean introgression is similar across the 14 SNP data sets being the highest Q value estimated by 16 SNPs (mean Q=0.16 as compared to 0.11-0.12). A similar pattern arises when the 77 individuals were examined by introgression classes. Mean introgression tends to differ more dramatically from the full data set when the number of SNPs is lower than 40. For example, for the class of the most introgressed individuals (Q> 0.50), mean Q varies between 0.62 and 0.63 when estimated by the first 10 SNP data sets (from 1183 to 104 SNPs) and it is 0.77 when estimated by 16 SNPs. While the mean introgression is similar across data set 1 to 13, dispersion increases considerably when the number of SNPs is lower than 40 (Fig. 2).

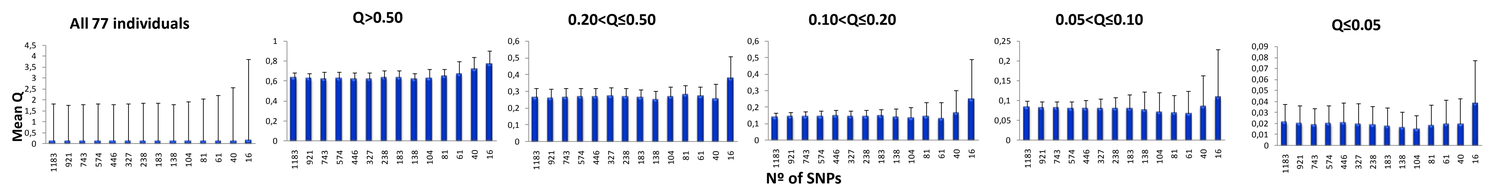

Fig. 2. Mean introgression (Q) and standard deviation for the 77 honey bees and for the same individuals separated by introgression classes (established using the full 1183 SNP data set)

A more detailed analysis at the individual level shows a clear trend of increased dispersion of overestimated and underestimated values around the 1183-SNP reference data set as the number of SNPs decreases (Fig. 3) .

Taking the 1183 data set as a reference, the deviation of individual estimates was lower than 5% for a number of SNPs higher than 327. As the number of SNPs decreases, the maximum deviation increases steadily. The most dramatic maximum deviation was observed for the 16-SNP data set, with an individual value that differed 53% from the reference (Fig. 4).

Furthermore, as the number of SNPs decreases, the number of individuals that exceed a 5% threshold increases considerably. That increase was particularly abrupt for a number of SNPs lower than 81 and 40 (Fig. 5).

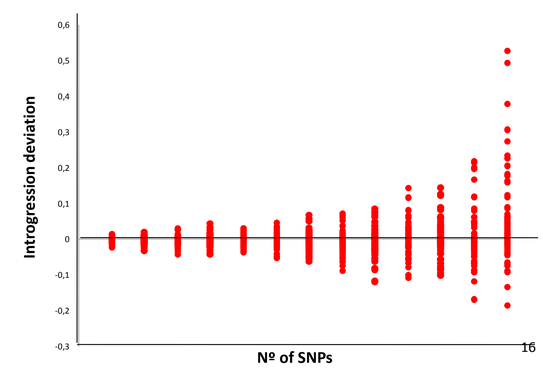

Fig. 3. Introgression deviation from the Q value estimated with the full SNP data set (marked by the "zero line") for each of the 77 individuals

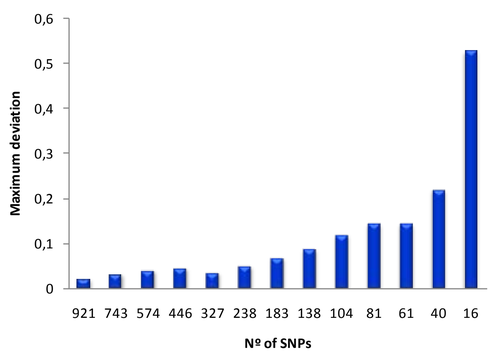

Fig. 4. Maximum Introgression deviation from the Q value estimated with the full SNP data set

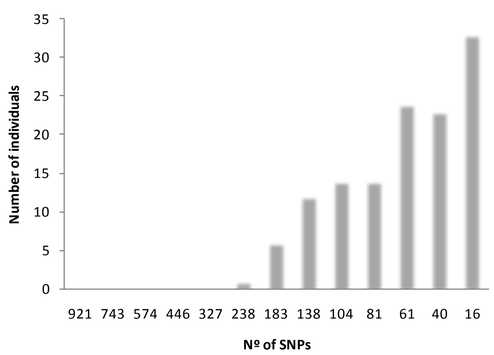

Fig. 5. Number of individuals, among the 77, whose introgression deviation, from the Q value estimated with the full SNP data set, exceeds a threshold of 5%

Conclusions

1. The analysis at the population level suggests that 40 SNPs provide introgression estimates similar to those generated by the 1183-SNP data set.
2. However, a more detailed analysis at the individual level suggests that a larger number of SNPS is needed to provide introgression estimates as accurate as those generated by the full data set.
3. According to the preliminary analysis conducted herein the minimum number of SNPs is 238.

Acknowledgements

We thank Per Kryger, Bjorn Dahle, Lionel Garnery, Pilar de la Rúa, Raffaele Dall' Olio, Norman Carreck, Gabriele Soland-Reckeweg and Romée van der Zee for providing the honey bee samples. DNA extractions and SNP genotyping was performed by Colette Abbey, who was deeply committed to this project. Julio Chávez-Galarza is supported by Fundação para a Ciência e Tecnologia through the scholarship SFRH/BD/68682/2010. This research was funded by Fundação para a Ciência e Tecnologia and COMPETE/QREN/EU through the project PTDC/BIA-BEC/099640/2008.

References

•Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23: 1801-1806.

•Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155:945-959.

•Rosenberg NA. 2004. DISTRUCT.: a program for the graphical display of population structure. *Molecular Ecology Notes*, 4:137-138.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR