SNP detection in French populations by whole genome sequencing

Alain Vignal, David Wragg, Benjamin Basso, Jean Pierre Bidanel, Yves Le Conte

Characterisation of French honey bee populations by whole-genome sequencing

“sequencing drones circumvents issues of heterozygosity”

Alain Vignal[1], **David Wragg**[1], **Benjamin Basso**[2], **Jean-Pierre Bidanel**[3], **Yves Le Conte**[4]
[1]INRA, GenPhySE, Toulouse; [2]ITSAP, PrADE, Avignon; [3]INRA, GABI, Jouy-en-Josas; [4]INRA, PrADE, Avignon
Correspondance to david.wragg@toulouse.inra.fr or alain.vignal@toulouse.inra.fr

A collaboration between l'INRA and l'Institut de l'abeille (ITSAP) is currently in progress, the SeqApiPop project, in which 1000 drones representing 1000 colonies will be paired-end sequenced at 6X coverage on the Illumina HiSeq platform. The aim of which is to characterise the genetic diversity of the domestic black bee, *Apis mellifera mellifera*, in France.

Heterozygosity is a major challenge to high-quality variant calling in next-generation sequence data, typically requiring high-depths of coverage to circumvent. This places considerable economic constraints on population studies of humans and livestock species, which are diploid in nature and typically have large genome sizes (> 2 Gb). The haplodiploid nature of honeybees, and their small genome size (< 250 Mb), permits large population studies to be conducted economically by whole-genome sequencing of haploid drones, although to date no research group to our knowledge has capitalised on this opportunity. Here we present an overview of the bioinformatics framework and preliminary analyses following its application to 60 drones.

METHODS

One drone was sequenced per colony from a total of 60 colonies representing two distinct populations; one used in the **production of honey** (**JFM**), and one for the **production of royal jelly** (**OTH**). Reads were mapped to the reference (Amel4.5) using *BWA-MEM*[1], variants called (**Fig 1A**), and missing genotypes imputed within-population using *BEAGLE*[2] (**Fig 1B**). Population data were then merged (**Fig 1C**) and converted from VCF to *Plink*[3] format using *Plink v1.9*[4]. Subsequent analysis was conducted in R[5] to detect signatures of selection, and *ADMIXTURE*[6] to estimate ancestry.

RESULTS

Alignment metrics indicate a mean depth of coverage across all data of **7X**. On average, reads mapped to **92%** of the genome, of which **69%** was considered callable (depth ≥ 3X). The average number of variants identified per caller, and the percentage retained after combining with *BAYSIC*[7] were: *GATK*[8] 952K (87%), *Pileup*[9] 846K (94%), *Platypus*[10] 492K (98%); resulting in a final set of **830K SNPs per individual**. The combined data resulted in **4.27M SNPs overall**, of which an average of **7% were imputed** per sample. **4M SNPs mapped to LG1-16 and MT**, of which 34K were unique to JFM and 19K unique to OTH.

Prior to running *ADMIXTURE* the data was pruned to remove SNPs in high linkage disequilibrium ($r^2 > 0.3$) in 50-SNP windows (10-SNP step size) using Plink v1.9. Resulting in **665K SNPs for *ADMIXTURE* analysis**, the results of which are presented in **Fig 2**. The optimal *K*, inferred from the cross-validation error, was observed to be 2 and this correctly partitions the two populations. However it is interesting to note with increasing *K* values the difference in admixture levels between the two populations. An identity-by-state (IBS) matrix calculated in Plink using the same dataset reported a **mean IBS of 0.79 in JFM** and **0.84 in OTH**.

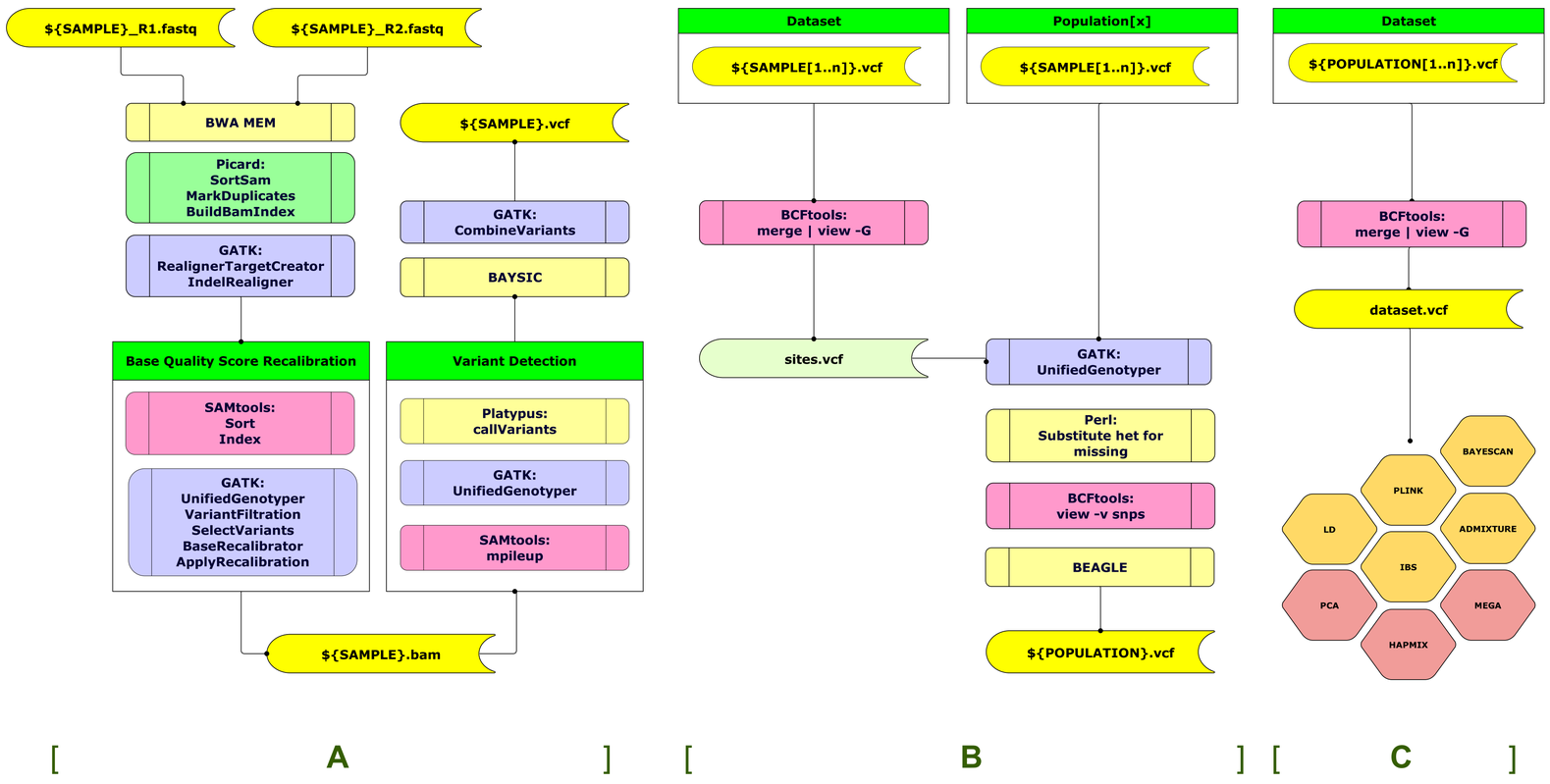

Fig 1 | *SeqApiPop framework*

Reads are mapped and processed according to best practices (**A**). Variants are called using GATK, Platypus and Pileup, and then combined using BAYSIC which estimates their likelihood of being genuine. SNP sites are identified across all samples. Within-population, samples are re-genotyped at all sites and missing values imputed (**B**). Multiple populations are merged into a single dataset for downstream analysis (**C**).

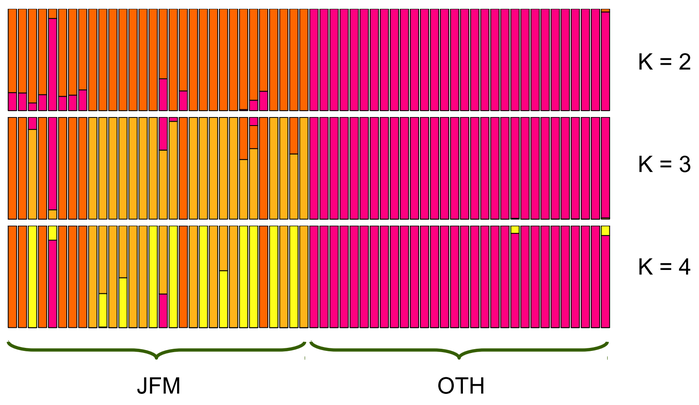

Fig 2 | *ADMIXTURE* plot for K 2 to 5

Signatures of selection were detected by estimating F_{ST} using a sliding window approach (**Fig 3**). Windows overlapped by 75%, and significance thresholds were marked at the 99th and 99.9th percentiles. A number of regions exceed the significance thresholds, the most notable of which is an interval on LG1 (**Fig 3B**)which corresponds with a previously published quantitative trait loci (QTL; *pln2*)[11]. The peak of this interval (LG1:18.51-18.67 Mb) contains an odorant receptor gene, suggesting it to be a strong candidate for honey production. Two less substantial intervals were detected on LG11 (**Fig 3C**) in proximity to the major royal jelly protein gene complex.

CONCLUSIONS

Sequencing of haploid drones circumvents issues of heterozygosity, enabling a lower depth in variant calling. An assessment of this strategy in 60 bees, following basic tests to measure LD, admixture, and to detect signatures of selection indicates that it will be effective for larger studies.

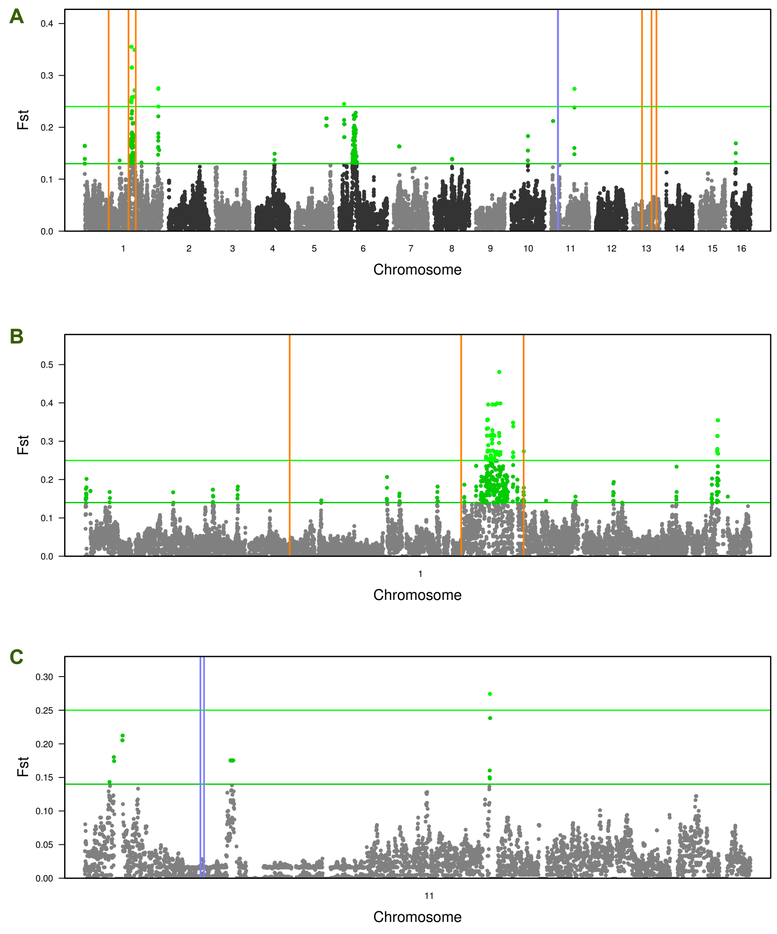

Fig 3 | Signatures of selection

Genome-wide F_{ST} based analyses using sliding windows with 75% overlap. Significance thresholds indicate 99th and 99.9th percentiles. Chromosomes 1-16 illustrated following analysis using 40 Kb windows (**A**), whilst individual plots for chromosome 1 (**B**) and 11 (**C**) were generated using 10 Kb windows. Red vertical lines indicate quantitative trait loci linked to pollen hoarding[11], whilst blue vertical lines indicate the boundaries of the major royal jelly protein gene complex.

REFERENCES
1. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013). at <http://arxiv.org/abs/1303.3997>
2. Browning, B. L. & Browning, S. R. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet.* **84,** 210–223 (2009).
3. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
4 Purcell, S & Chang, C. PLINK v1.90ble. at <https://www.cog-genomics.org/plink2>
5. R Development Core Team. R: A Language and Environment for Statistical Computing. **1,** 409 (2011).
6. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* (2009). doi:10.1101/gr.094052.109
7. Cantarel, B. L. *et al.* BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics* **15,** 104 (2014).
8. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
9. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).
10. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **advance online publication,** (2014).
11. Page, R. E., Fondrk, M. K. & Rueppell, O. Complex pleiotropy characterizes the pollen hoarding syndrome in honey bees (Apis mellifera L.). *Behav. Ecol. Sociobiol.* **66,** 1459–1466 (2012).

UMR1388 GenPhySE
31326 Castanet Tolosan, France
https://www-lgc.toulouse.inra.fr/intranet